Comparative Analysis of Regression Regularization Methods for Life Expectancy Prediction

Nataliya Boyko and Olena Moroz

Lviv Polytechnic National University, Profesorska Street 1, Lviv, 79013, Ukraine

Abstract

L1-, L2-, ElasticNet - regularizations of classification and regression were investigated in the course of work. The purpose of the scientific work is to explore different methods of regularization for life expectancy prediction, namely L1 -, L2, and ElasticNet regularization, to implement them programmatically and to draw conclusions about the results. First of all, the WHO Statistics on Life Expectancy dataset was analyzed, prepared and cleaned. It was checked if the data types match the attributes of dataset. A linear regression model was created using the scikit-learn library. After her training, the weights of the model features were obtained and it was observed that the weights at strongly correlated features were greater than the rest. To eliminate the problem of multicollinearity, 3 regularization methods were applied and compared.

Keywords 1

regression regularization; linear regression, ElasticNet regularization, multicollinearity, algorithm, machine learning, medicine

1. Introduction

The work is devoted to a comprehensive study of the regularization of regression for life expectancy prediction.

Regularization in machine learning is a way to reduce the complexity of a model by adding some additional constraints to the problem condition. The purpose of using regularization [1, 5]:

- correct an incorrect task
- prevent retraining
- save resources

It is known that regression models have the predisposition to relearn. If the model is too heavy and there is not enough data to determine its parameters, you can get some model that will describe the training sample very well, but will generalize to the test sample much worse. There are several ways to solve this problem:

- Take more data

Disadvantage: very often this solution is not available, because additional data costs extra money

- Use fewer features

Disadvantage: this requires a large number of subsets of features. However, the total number of subsets that are meant to be sorted increases very rapidly in accordance with the increasing dimension of the problem. A complete search is often unavailable.

- Limit the weight of the features

Disadvantage: this method is often ineffective. Retraining can only be done to a certain extent.

MoMLeT+DS 2021: 3rdInternational Workshop on Modern Machine Learning Technologies and Data Science, June 5, 2021, Lviv-Shatsk, Ukraine

EMAIL: nataliya.i.boyko@lpnu.ua (N. Boyko); olena.moroz.kn.2017@lpnu.ua (O. Moroz)
ORCID: 0000-0002-6962-9363 (N. Boyko); 0000-0003-4012-5142 (O. Moroz)

Given the shortcomings of the above methods, the use of regularization in order to prevent model retraining is a relevant and effective way to solve the problems of classification, regression and learning of deep neural networks [2, 8].

Heavy weights are a measure of complexity and a sign of model retraining. Therefore, the modern approach to reducing the generalization error is to use a larger model with the use of regularization during training, which keeps the weight of the model small. This method leads to faster optimization of the model and increase overall performance.

The purpose of the scientific work is to explore different methods of regularization for life expectancy prediction, namely L1, L2, and ElasticNet regularization, to implement them programmatically and to draw conclusions about the results [10,15].

2. Description of Linear regression

Linear regression - a model of the dependence of the variable x on one or more other variables (features, factors) with a linear dependence function, which has the following form [8]:

$$\alpha(x) = w_0 + \sum_{j=1}^{d} w_j x_j \tag{1}$$

Advantages of linear regression:

Speed and simplicity of obtaining the model.

- Interpretation of the model. The linear model is transparent and understandable to the analyst. The obtained regression coefficients can be used to judge how one or another factor affects the result, to make additional useful conclusions on this basis.
- Wide applicability. A large number of real processes in economics and business can be described with sufficient accuracy by linear models.
- Study of this approach. Typical problems (for example, multicollinearity) and their solutions are known for linear regression, tests of estimation of static significance of the received models are developed and implemented.

Linear regression quality metrics:

— Mean square error [7, 17]:

$$MSE(\alpha, X) = \frac{1}{l}\sum_{i=1}^{l}(\alpha(x_i) - y_i)^2 \tag{2}$$

- It is easy to optimize because it has a derivative at all points
- Strong penalty for outliers

— Average absolute error [4]:

$$MSE(\alpha, X) = \frac{1}{l}\sum_{i=1}^{l}\left|\alpha(x_i) - y_i\right| \tag{3}$$

- It is more difficult to optimize because it has no derivative at zero
- High endurance to outliers

— Coefficient of determination [12]:

$$R^2(\alpha, X) = 1 - \frac{\sum_{i=1}^{l}(\alpha(x_i) - y_i)^2}{\sum_{i=1}^{l}(y_i - \hat{y})^2} \tag{4}$$

This metric explains what proportion of variance in the entire target vector the linear model is able to explain. That is, it reflects the proportion of variety of responses that the model is able to predict [10-13].

For smart models $0 \le R^2 \le 1$, when

R^2 – is an ideal model;

R^2 – the quality of the model coincides with the optimal constant algorithm (returns the average answer for the entire training sample);

$R^2 < 1$ – the quality of the model is worse than the constant;

3. L1-regularization

L1 - Lasso regularization helps to improve the generalization of test data by selecting the most important factors that most strongly influence the result. Factors with a small value get the value of zero and do not affect the final result. In fact, they only help to predict noise in the training data set [17, 19].

General formula:

$$L1 = \sum_{i=1}^{N} (y_n - \hat{y}_n)^2 + \lambda \sum_{j=1}^{D} |w_j| , \tag{5}$$

where λ is the regularization coefficient.

The larger the value of λ, the more features are converted to zero and the simpler the model becomes.

The parameter can be reset to zero if:

- It has a value close to zero
- Its value changes greatly when changing the sample (large variance)
- Its removing has the least effect on changing the value of the error function

Thus, L1 - regularization contributes to the sparseness of the function, when only a few factors of the model are not equal to zero. This can completely eliminate some features and mitigate the multicollinearity and complexity of the model. The disadvantage of this approach may be the complexity of the optimization process, because the L1-regularizer is not smooth (has no derivative at zero).

Multicollinearity is the presence of a linear relationship between features in a sample [20]. That is, the existence of a vector of values of a certain feature on all objects, which is expressed through vectors of other features. In this case, regardless of the selected object, the result of the sum of the products of the coefficients on the value of the features will be equal to zero.

$$\alpha_1 x_i^1 + ... + \alpha_d x_i^\alpha = 0$$

or

$$\langle \alpha, x_i \rangle = 0 .$$

Therefore, the problem with multicollinearity is that it leads to an infinitely large number of optimal algorithms, many of which have large values of weights, but not all generalize the information well. As a result, it leads to retraining of the model.

The L1 - regularization method is better suited for cases where most of the model parameters are not necessary and their values can be neglected [16].

The Lasso regression problem (LASSO, Least Absolute Shrinkage and Selection Operator) corresponds to the problem of a priori distribution of Laplace by coefficients.

4. L2 - regularization

L2 - regularization (English Ridge regularization, Tikhonov regularization) does not allow retraining of the model by prohibiting disproportionately large weights. This leads to the selection of parameters whose values do not deviate much from zero [17, 5].

General formula:

$$L2 = \sum_{i=1}^{N} (y_n - \hat{y}_n)^2 + \lambda \sum_{j=1}^{D} w_j^2 , \tag{6}$$

where λ is the regularization coefficient.

Model optimization [7,9]:

$$Q(w, X) + \lambda \|w\|^2 \to \min_w , \tag{7}$$

where $Q(w, X)$ – is the loss function equivalent to the conditional optimization problem:

$$\begin{gathered} Q(w, X) \to \min_{w} \\ \|w\|^2 \le C \end{gathered}, \tag{8}$$

where C – is a constant that normally limits the vector of weights.

λ controls the error function and the regularization penaltu. If the value of λ is large, the weights will go to zero. If the value of λ is small or equal to zero, then the weights will tend to minimize the loss function [9-12].

By adding a constant multiplied by the sum of the squares of the weights, we change the initial loss function and add a penalty for large weights. The square penalty makes the loss function strongly convex, and therefore it has a unique minimum.

This method is suitable when most of the variables in the model are useful and necessary. Also, the addition of L2 - regularization does not complicate the optimization process (eg gradient descent) because this regularizer is smooth and convex [19-20].

The Tikhonov regression problem corresponds to the problem of normal a priori distribution on coefficients and has an analytical solution [17]:

$$w_* = (X^T X + \lambda I)^{-1} X^T y, \tag{9}$$

where λI – is a diagonal matrix in which the values of λ are on the diagonal.

5. ElasticNet regularization

ElasticNet regularization is a linear combination of L1 and L2 regularizations. This method uses the advantages of both methods at once. The fact that the variables do not turn into zero, as in L1 - regularization, makes it possible to create conditions for a group effect with a high correlation of variables [15].

General formula:

$$L_{EN} = \sum_{i=1}^{N} (y_n - \hat{y}_n)^2 + \lambda_1 \sum_{j=1}^{D} |w_j| + \lambda_2 \sum_{j=1}^{D} w_j^2 \tag{10}$$

The method of elastic net is most often used when the model has a lot of parameters, but whether they are necessary or can be neglected beforehand is unknown.

In particular in the following cases [3-6]:

- Cancer prediction
- Metric training
- Portfolio optimization

6. Analysis and preparation of the selected dataset

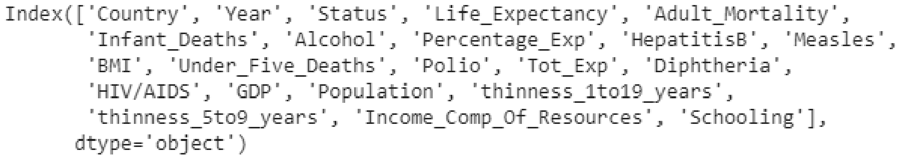
The WHO Statistics on Life Expectancy dataset was selected for software implementation [1]. This dataset contains information collected by the World Health Organization and the United Nations to track factors that affect life expectancy.

Dataset attributes:

- Country - the country
- Year - year
- Status - development status (currently being developed / already developed)
- Life expectancy - life expectancy
- Adult Mortality - mortality rate for adults of both sexes (probability of death from 15 to 60 years per 1000 population)
- Infant death - the number of infant deaths per 1,000 population
- Alcohol - per capita alcohol consumption (15+) (in liters of pure alcohol)
- percentage expenditure - health care expenditure as a percentage of gross domestic product per capita (%)

- Hepatitis B - immunization coverage against hepatitis B (HepB) among one-year-old children (%)
- Measles - measles - the number of reported cases per 1000 population
- BMI - the average body weight of the entire population
- under-five deaths - the number of deaths under the age of five per 1,000 population
- Polio - anti-polio coating (Pol3) among one-year-old children (%)
- Total expenditure - national health expenditure as a percentage of total public expenditure (%)
- Diphtheria - coverage by immunoprophylaxis against tetanus and pertussis (DTP3) among one-year-old children (%)
- HIV / AIDS - deaths per 1,000 live births HIV / AIDS (0-4 years)
- GDP - gross domestic product per capita (in US dollars)
- Population - the population of the country
- thinness 1-19 years - prevalence of weight loss among children and adolescents aged 10 to 19 years (%)
- thinness 5-9 years - the prevalence of weight loss among children aged 5 to 9 years (%)
- Income composition of resources - human development index by composition of resource income (index ranges from 0 to 1)
- Schooling - number of years of schooling

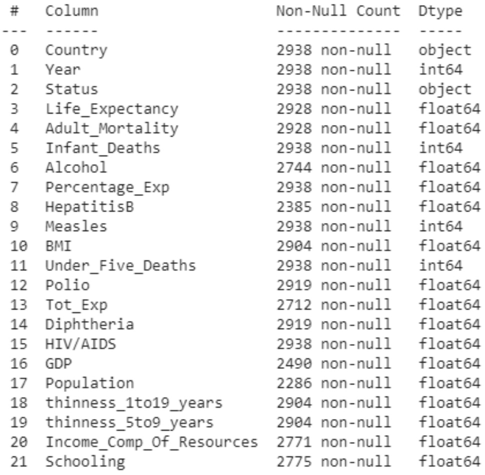
The selected dataset was cleaned of data, namely (Figure 1): some columns were renamed because they contained spaces:



```
Index(['Country', 'Year', 'Status', 'Life_Expectancy', 'Adult_Mortality',
       'Infant_Deaths', 'Alcohol', 'Percentage_Exp', 'HepatitisB', 'Measles',
       'BMI', 'Under_Five_Deaths', 'Polio', 'Tot_Exp', 'Diphtheria',
       'HIV/AIDS', 'GDP', 'Population', 'thinness_1to19_years',
       'thinness_5to9_years', 'Income_Comp_Of_Resources', 'Schooling'],
      dtype='object')
```

Figure 1: Renamed attributes of the dataset "Life Expectancy"

Figure 1 lists the names of all renamed attributes of the Life Expectancy dataset for each variable was checked the data match according to its data type:

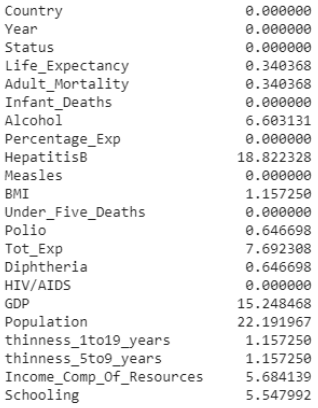


```
 #   Column                    Non-Null Count  Dtype
---  ------                    --------------  -----
 0   Country                   2938 non-null   object
 1   Year                      2938 non-null   int64
 2   Status                    2938 non-null   object
 3   Life_Expectancy           2928 non-null   float64
 4   Adult_Mortality           2928 non-null   float64
 5   Infant_Deaths             2938 non-null   int64
 6   Alcohol                   2744 non-null   float64
 7   Percentage_Exp            2938 non-null   float64
 8   HepatitisB                2385 non-null   float64
 9   Measles                   2938 non-null   int64
 10  BMI                       2904 non-null   float64
 11  Under_Five_Deaths         2938 non-null   int64
 12  Polio                     2919 non-null   float64
 13  Tot_Exp                   2712 non-null   float64
 14  Diphtheria                2919 non-null   float64
 15  HIV/AIDS                  2938 non-null   float64
 16  GDP                       2490 non-null   float64
 17  Population                2286 non-null   float64
 18  thinness_1to19_years      2904 non-null   float64
 19  thinness_5to9_years       2904 non-null   float64
 20  Income_Comp_Of_Resources  2771 non-null   float64
 21  Schooling                 2775 non-null   float64
```

Figure 2: Data types of attributes of the dataset "Life Expectancy"

Figure 2 shows that all data types correspond to their data.

The percentage of zero values in each column was determined:

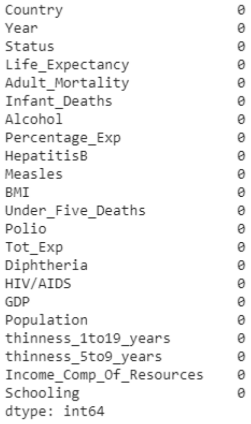


```
Country                     0.000000
Year                        0.000000
Status                      0.000000
Life_Expectancy             0.340368
Adult_Mortality             0.340368
Infant_Deaths               0.000000
Alcohol                     6.603131
Percentage_Exp              0.000000
HepatitisB                 18.822328
Measles                     0.000000
BMI                         1.157250
Under_Five_Deaths           0.000000
Polio                       0.646698
Tot_Exp                     7.692308
Diphtheria                  0.646698
HIV/AIDS                    0.000000
GDP                        15.248468
Population                 22.191967
thinness_1to19_years        1.157250
thinness_5to9_years         1.157250
Income_Comp_Of_Resources    5.684139
Schooling                   5.547992
```

Figure 3: Percentage of zero values in each attribute of the dataset "Life Expectancy"

As shown in Figure 3, zero data is present in the columns of the dataset: Life_Expectancy, Adult_Mortality, Alcohol, HepatitsB, BMI, Polio, Tot_Exp, Diphteria, GDP, Population, Thiness_1to19_years, Thiness_5to9_years, Income_Comces_Of.

Zero values are processed by interpolation, zero values left after interpolation were discarded:

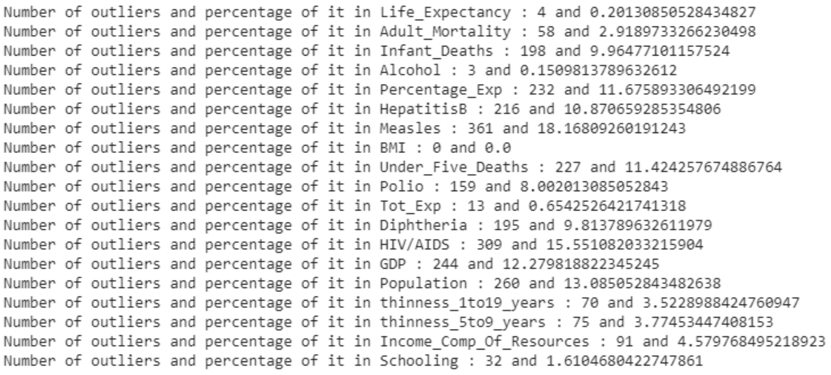


```
Country                     0
Year                        0
Status                      0
Life_Expectancy             0
Adult_Mortality             0
Infant_Deaths               0
Alcohol                     0
Percentage_Exp              0
HepatitisB                  0
Measles                     0
BMI                         0
Under_Five_Deaths           0
Polio                       0
Tot_Exp                     0
Diphtheria                  0
HIV/AIDS                    0
GDP                         0
Population                  0
thinness_1to19_years        0
thinness_5to9_years         0
Income_Comp_Of_Resources    0
Schooling                   0
dtype: int64
```

Figure 4: Percentage of zero values in each attribute of the dataset "Life Expectancy" after interpolation

Figure 4 shows that all zero values are eliminated.

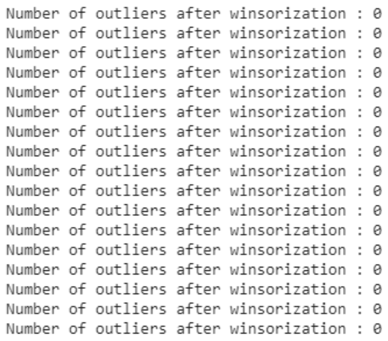
The number and percentage of atypical values for each variable were calculated and deleted using the winsorization technique:



```
Number of outliers and percentage of it in Life_Expectancy : 4 and 0.20130850528434827
Number of outliers and percentage of it in Adult_Mortality : 58 and 2.9189733266230498
Number of outliers and percentage of it in Infant_Deaths : 198 and 9.96477101157524
Number of outliers and percentage of it in Alcohol : 3 and 0.1509813789632612
Number of outliers and percentage of it in Percentage_Exp : 232 and 11.675893306492199
Number of outliers and percentage of it in HepatitisB : 216 and 10.870659285354806
Number of outliers and percentage of it in Measles : 361 and 18.16809260191243
Number of outliers and percentage of it in BMI : 0 and 0.0
Number of outliers and percentage of it in Under_Five_Deaths : 227 and 11.424257674886764
Number of outliers and percentage of it in Polio : 159 and 8.002013085052843
Number of outliers and percentage of it in Tot_Exp : 13 and 0.6542526421741318
Number of outliers and percentage of it in Diphtheria : 195 and 9.813789632611979
Number of outliers and percentage of it in HIV/AIDS : 309 and 15.551082033215904
Number of outliers and percentage of it in GDP : 244 and 12.279818822345245
Number of outliers and percentage of it in Population : 260 and 13.085052843482638
Number of outliers and percentage of it in thinness_1to19_years : 70 and 3.5228988424760947
Number of outliers and percentage of it in thinness_5to9_years : 75 and 3.77453447408153
Number of outliers and percentage of it in Income_Comp_Of_Resources : 91 and 4.579768495218923
Number of outliers and percentage of it in Schooling : 32 and 1.6104680422747861
```

Figure 5: The number and percentage of atypical values for each attribute of the dataset "Life Expectancy"

Figure 5 shows the number and percentage of atypical values in each attribute of the dataset.



```
Number of outliers after winsorization : 0
Number of outliers after winsorization : 0
Number of outliers after winsorization : 0
Number of outliers after winsorization : 0
Number of outliers after winsorization : 0
Number of outliers after winsorization : 0
Number of outliers after winsorization : 0
Number of outliers after winsorization : 0
Number of outliers after winsorization : 0
Number of outliers after winsorization : 0
Number of outliers after winsorization : 0
Number of outliers after winsorization : 0
Number of outliers after winsorization : 0
Number of outliers after winsorization : 0
Number of outliers after winsorization : 0
Number of outliers after winsorization : 0
Number of outliers after winsorization : 0
```

Figure 6: Percentage of atypical values for each attribute of the dataset "Life Expectancy" after winsorization

In Figure 6 is observed that all atypical values are eliminated.

This winsorization technique sets a limit on extreme values in statistics in order to reduce the impact of atypical data that may be erroneous. Data for some variables before and after winsorization using box charts are shown in Figures 7-9

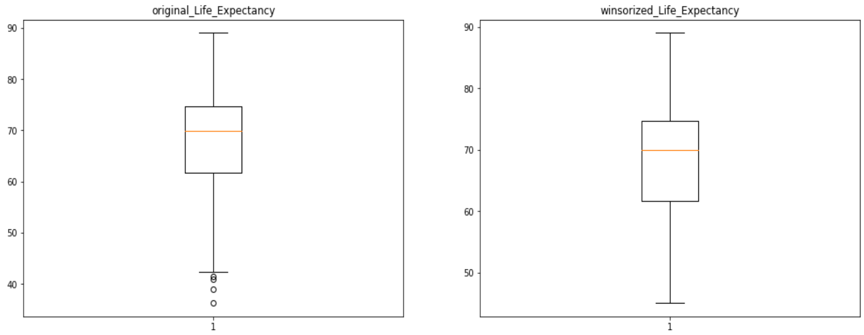


Figure 7: Diagram of the range of values of the attribute Life_Expectancy before and after winsorization

In Figure 7 shows the sample size of the Life_Expectancy attribute before and after winsorization. After winsorization, there are no outliers on the diagram. It is seen that the range of values of the variable has also decreased.

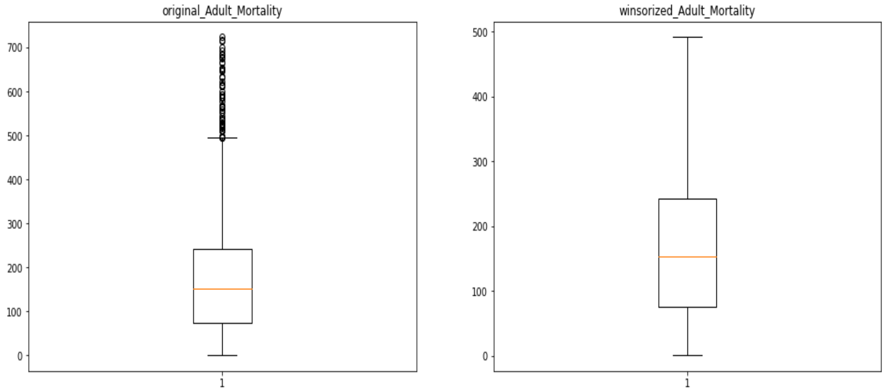


Figure 8: Diagram of the range of values of the attribute Adult_Mortality before and after winsorization

In Figure 8 shows the sample size of the Life_Expectancy attribute before and after winsorization. Before winsorization, the range of sampling values ranged from 0 to 700, after winsorization - from 0 to 500. This was due to the elimination of outliers.

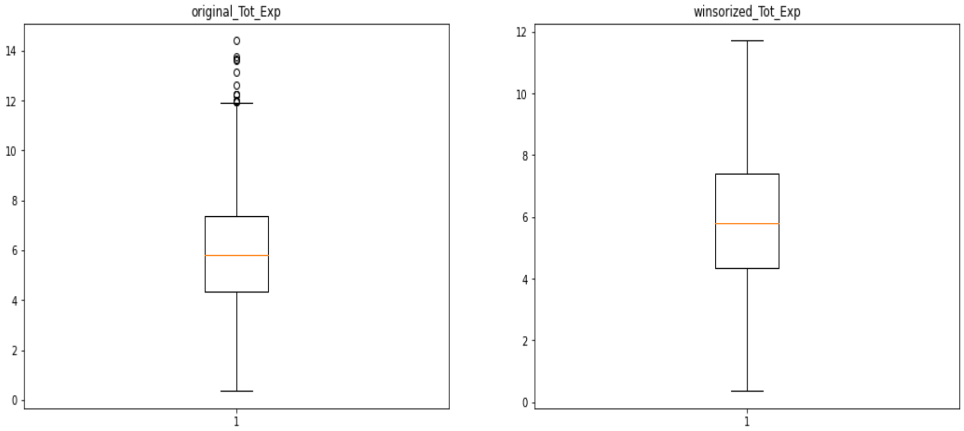


Figure 9: Diagram of the range of values of the attribute original_Tot_Exp before and after winsorization

In Figure 9 shows the sample size of the Life_Expectancy attribute before and after winsorization. Before winsorization, the sample size is in the range from 0 to 14, after - from 0 to 12. Therefore, outliers are eliminated.

As can be seen from the diagrams, atypical data (outliers) were successfully eliminated using the winsorization method. Variables were added to the dataset after winsorization.

Variables winsorized_Life_Expectancy, winsorized_Tot_Exp, winsorized_Schooling are distributed according to the normal distribution (Figures 10-12)

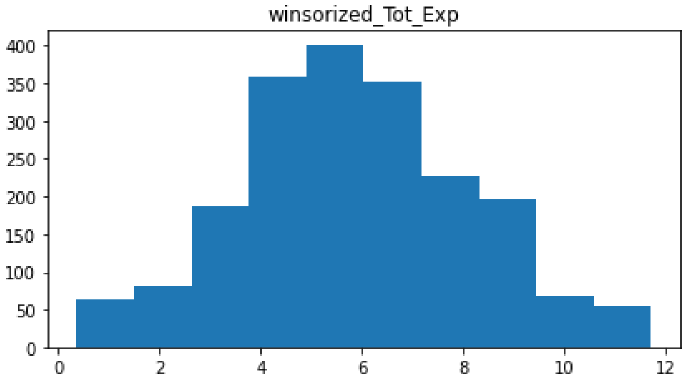


Figure 10: Distribution of the variable winsorized_original_Tot_Exp

Figure 10 shows the distribution of the variable winsorized_Tot_Exp. The diagram shows that this variable is distributed according to the normal distribution.

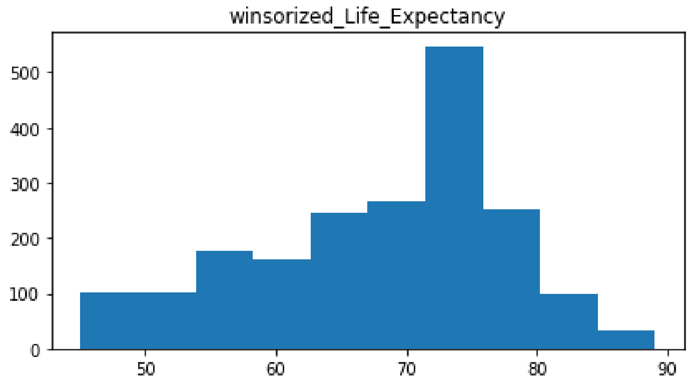


Figure 11: Distribution of the variable winsorized_ Life_Expectancy

Figure 11 shows the distribution of the variable winsorized_Life_Expectancy. The chart shows that this variable is distributed according to the normal distribution.

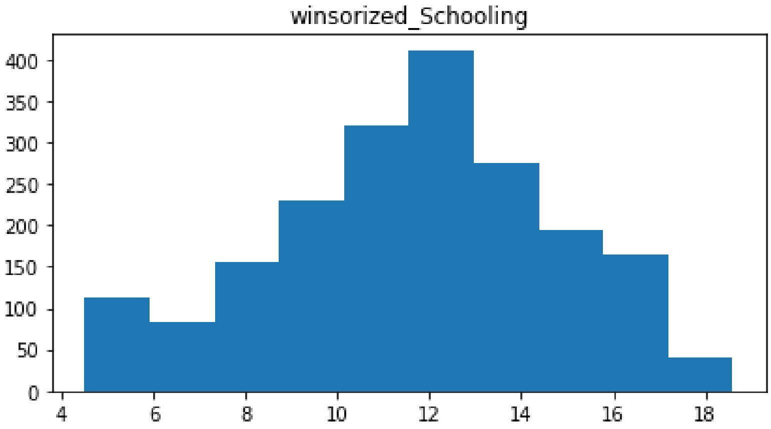


Figure 12: Distribution of the variable winsorized_Schooling

Figure 12 shows the distribution of the variable winsorized_ Schooling. It is observed that this variable obeys the normal distribution.

Analysis of the dependences between the target variable winsorized_Life_Expectancy and other dataset variables shows that there is a direct linear dependence between winsorized_Life_Expectancy and Income_Comp_Of_Resources and Schooling (Figures 13, 14). There is also an inverse linear dependence between winsorized_Life_Expect.

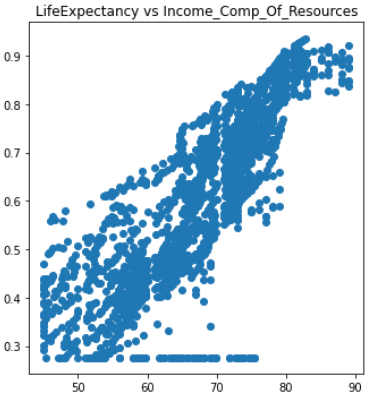


Figure 13: The dependence between the variables LifeExpectancy and Income_Comp_Of_Resources

This chart is traced the linear dependence between the target variable winsorized_Life_Expectancy and Income_Comp_Of_Resources.

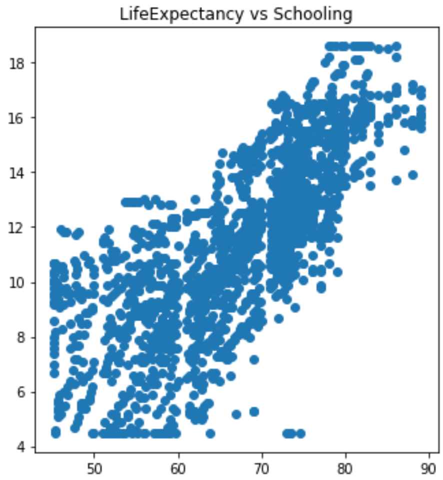


Figure 14: The dependence between the variables LifeExpectancy and Schooling

In this diagram, there is a linear dependence between the target variable winsorized_Life_Expectancy and Schooling.

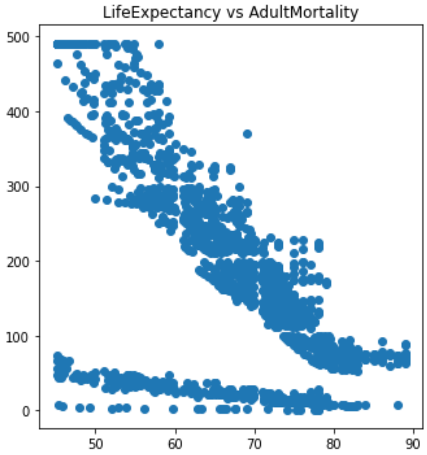


Figure 15: The dependence between the variables LifeExpectancy and AdultMortality

This diagram shows the inverse linear dependence between the target variable winsorized_Life_Expectancy and Schooling

Correlation map of dataset features:

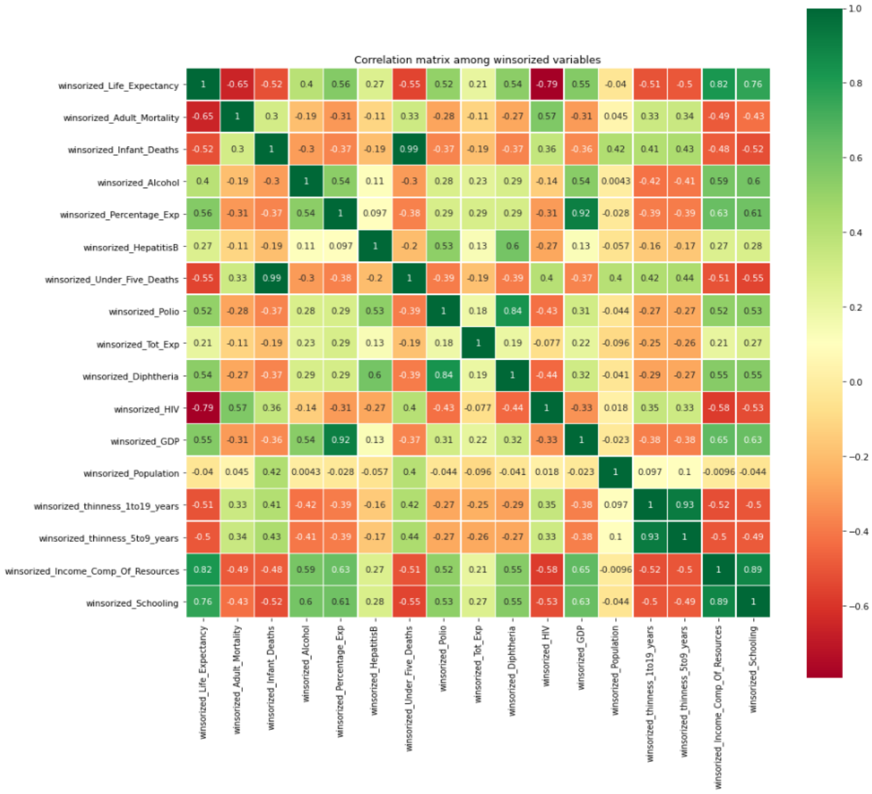


Figure 16: Correlation map of the characteristics of the Life Expectancy datase

From this thermal diagram it is possible to reveal dependences between the following features:

1. There is a dependence between winsorized_Income_Comp_Of_Resources and winsorized_Schooling.
2. There is a dependence between winsorized_thinness_1to19_years and winsorized_thinness_5to9_years
3. There is a dependence between winsorized_Polio and winsorized_Diphtheria
4. There is a dependence between winsorized_Percentage_Exp and winsorized_GDP.
5. There is a dependence between winsorized_Income_Comp_Of_Resources and winsorized_Life_Expectancy.
6. There is a dependence between winsorized_Life_Expectancy and winsorized_Schooling.
7. There is a dependence between winsorized_Infant_Deaths and winsorized_Under_Five_Deaths.
8. There is an inverse dependence between winsorized_HIV and winsorized_Life_Expectancy.
9. There is an inverse dependence between winsorized_Adult_Mortality and winsorized_Life_Expectancy.

The sample has features that correlate with the target variable, which means that the problem of life expectancy (Life Expectancy) can be solved by linear methods.

7. Experiments

Linear regression model

Initially, a linear regression model was created without applying any of the regularizations. This model solves the problem of predicting life expectancy (Life Expectancy) based on 16 other features of the dataset. The scikit-learn library was used for software implementation.

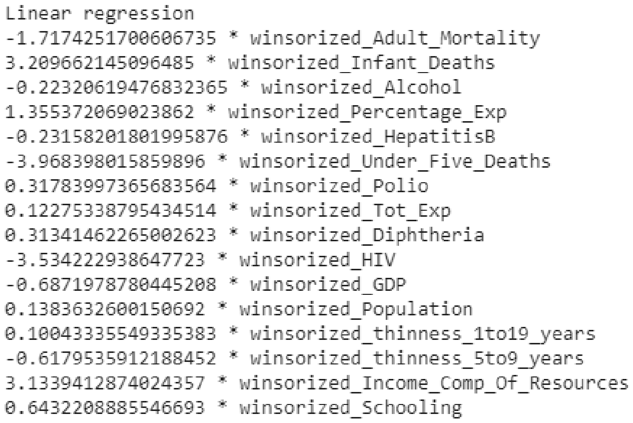
All data were mixed and divided into train and test:

```
train, test = train_test_split(le_shuffled, test_size=0.3)
train_data = scale(train.loc[:, train.columns != "winsorized_Life_Expectancy"])
test_data = scale(test.loc[:, test.columns != "winsorized_Life_Expectancy"])
train_labels = train["winsorized_Life_Expectancy"]
test_labels = test["winsorized_Life_Expectancy"]
```

Creating and learning a linear model:

model = linear_model.LinearRegression().fit(train_data, train_labels)

The obtained weights at the features after learning the model of linear regression (Figure 17):



```
Linear regression
-1.7174251700606735 * winsorized_Adult_Mortality
3.209662145096485 * winsorized_Infant_Deaths
-0.22320619476832365 * winsorized_Alcohol
1.355372069023862 * winsorized_Percentage_Exp
-0.23158201801995876 * winsorized_HepatitisB
-3.968398015859896 * winsorized_Under_Five_Deaths
0.31783997365683564 * winsorized_Polio
0.12275338795434514 * winsorized_Tot_Exp
0.31341462265002623 * winsorized_Diphtheria
-3.534222938647723 * winsorized_HIV
-0.6871978780445208 * winsorized_GDP
0.1383632600150692 * winsorized_Population
0.10043335549335383 * winsorized_thinness_1to19_years
-0.6179535912188452 * winsorized_thinness_5to9_years
3.1339412874024357 * winsorized_Income_Comp_Of_Resources
0.6432208885546693 * winsorized_Schooling
```

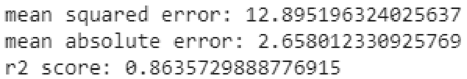
Figure 17: Weights of features of the model of linear regression after training

Absolute value of weights at linearly dependent features are bigger, than at other features. Analytical formula below explain this, it is used to calculate the weights of a linear model in the least squares method:

$$w = (X^T X)^{-1} X^T y \tag{11}$$

If X has collinear (linearly dependent) columns, the matrix X^T becomes degenerate, and the formula ceases to be correct. The more dependent the features, the smaller the determinant of this matrix and the worse the $Xw \approx y$ approximation (the problem of multicollinearity)

Quality metrics of the obtained linear model (Figure 18):



```
mean squared error: 12.895196324025637
mean absolute error: 2.658012330925769
r2 score: 0.8635729888776915
```

Figure 18: Quality metrics of the linear regression model (root mean square error, mean absolute error, coefficient of determination)

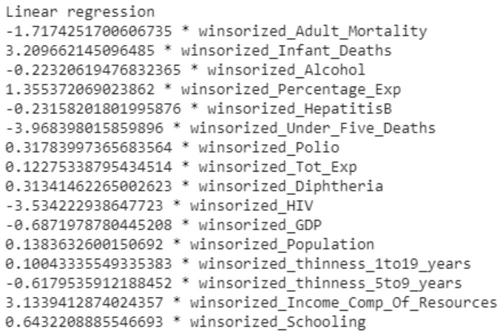
The mean absolute error 2.66, the root mean square error 12.9. The coefficient of determination is 0.86, and therefore is in the range of $0 \le R^2 \le 1$, that indicates that the model works well and explains 86% of the variance in the entire target vector, that is a good characteristic.

The solution of the problem of multicollinearity and overfitting is regularization of the linear model. L1 or L2, or L1 and L2 weight norm multiplied by the regularization coefficient α are added to the optimized functional. In the first case, the method is called Lasso, in the second – Ridge, and the third – Elastic Net.

8. Results

1) Lasso regularization

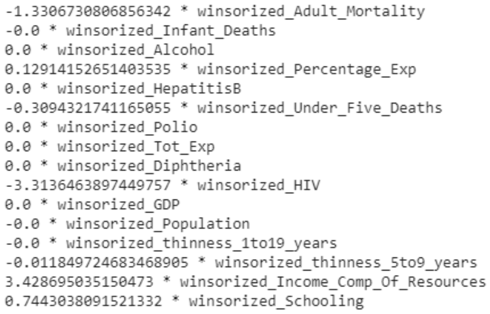
Weights of features of linear regression without regularization (Figure 19):



```
Linear regression
-1.7174251700606735 * winsorized_Adult_Mortality
3.209662145096485 * winsorized_Infant_Deaths
-0.22320619476832365 * winsorized_Alcohol
1.355372069023862 * winsorized_Percentage_Exp
-0.23158201801995876 * winsorized_HepatitisB
-3.968398015859896 * winsorized_Under_Five_Deaths
0.31783997365683564 * winsorized_Polio
0.12275338795434514 * winsorized_Tot_Exp
0.31341462265002623 * winsorized_Diphtheria
-3.534222938647723 * winsorized_HIV
-0.6871978780445208 * winsorized_GDP
0.1383632600150692 * winsorized_Population
0.10043335549335383 * winsorized_thinness_1to19_years
-0.6179535912188452 * winsorized_thinness_5to9_years
3.1339412874024357 * winsorized_Income_Comp_Of_Resources
0.6432208885546693 * winsorized_Schooling
```

Figure 19: Weights of features of the model of linear regression after training

Weights of features of Lasso regression (application of L1-regularization) (Figure 20):



```
-1.3306730806856342 * winsorized_Adult_Mortality
-0.0 * winsorized_Infant_Deaths
0.0 * winsorized_Alcohol
0.12914152651403535 * winsorized_Percentage_Exp
0.0 * winsorized_HepatitisB
-0.3094321741165055 * winsorized_Under_Five_Deaths
0.0 * winsorized_Polio
0.0 * winsorized_Tot_Exp
0.0 * winsorized_Diphtheria
-3.3136463897449757 * winsorized_HIV
0.0 * winsorized_GDP
-0.0 * winsorized_Population
-0.0 * winsorized_thinness_1to19_years
-0.011849724683468905 * winsorized_thinness_5to9_years
3.428695035150473 * winsorized_Income_Comp_Of_Resources
0.7443038091521332 * winsorized_Schooling
```

Figure 20: Weights of features of the linear regression model using L1-regularization after training

In comparison with the weights of the usual linear regression, it is observed that after the use of Lasso regularization the selection of features took place: the weights at non-informative features turned into zero. Weights at other features approached zero.

Visualization of weight dynamics with increasing regularization parameter α (Figure 21):

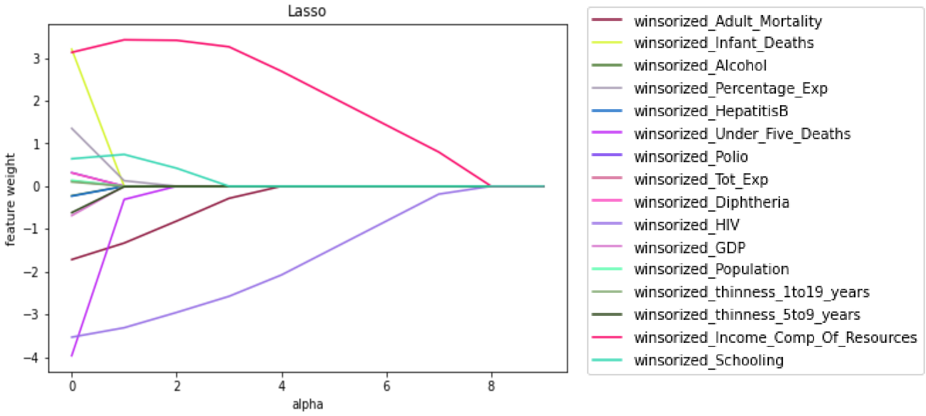
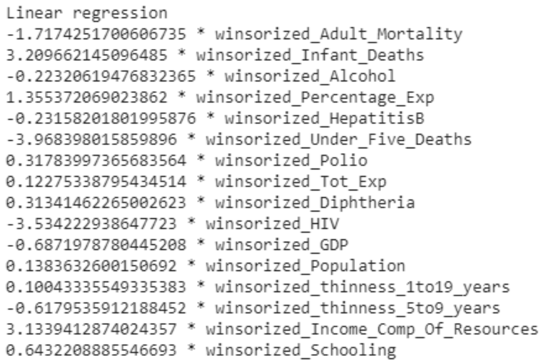


Figure 21: Chart with the dynamics of weights relative to the parameter of regularization α when using Lasso-regression

It is observed that as the parameter α of the L1-regularizer increases, the weights of the features rapidly go to zero, and as the value of α weights increases, more and more features turn to zero and the model becomes simpler.

2) Ridgeregularization

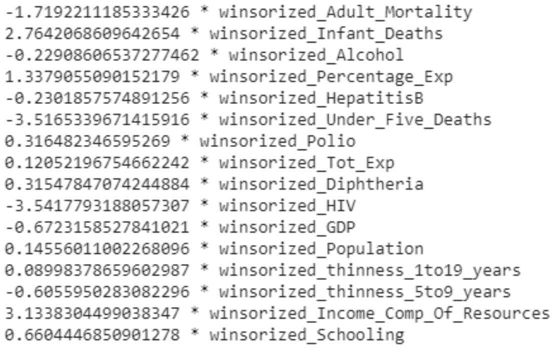
Weights of features of linear regression without regularization:



```
Linear regression
-1.7174251700606735 * winsorized_Adult_Mortality
3.209662145096485 * winsorized_Infant_Deaths
-0.22320619476832365 * winsorized_Alcohol
1.355372069023862 * winsorized_Percentage_Exp
-0.23158201801995876 * winsorized_HepatitisB
-3.968398015859896 * winsorized_Under_Five_Deaths
0.31783997365683564 * winsorized_Polio
0.12275338795434514 * winsorized_Tot_Exp
0.31341462265002623 * winsorized_Diphtheria
-3.534222938647723 * winsorized_HIV
-0.6871978780445208 * winsorized_GDP
0.1383632600150692 * winsorized_Population
0.10043335549335383 * winsorized_thinness_1to19_years
-0.6179535912188452 * winsorized_thinness_5to9_years
3.1339412874024357 * winsorized_Income_Comp_Of_Resources
0.6432208885546693 * winsorized_Schooling
```

Figure 22: Weights of features of the model of linear regression after training

Weights of features of ridge regression (application of L2-regularization) (Figure 23):



```
-1.7192211185333426 * winsorized_Adult_Mortality
2.7642068609642654 * winsorized_Infant_Deaths
-0.22908606537277462 * winsorized_Alcohol
1.3379055090152179 * winsorized_Percentage_Exp
-0.2301857574891256 * winsorized_HepatitisB
-3.5165339671415916 * winsorized_Under_Five_Deaths
0.316482346595269 * winsorized_Polio
0.12052196754662242 * winsorized_Tot_Exp
0.31547847074244884 * winsorized_Diphtheria
-3.5417793188057307 * winsorized_HIV
-0.6723158527841021 * winsorized_GDP
0.14556011002268096 * winsorized_Population
0.08998378659602987 * winsorized_thinness_1to19_years
-0.6055950283082296 * winsorized_thinness_5to9_years
3.1338304499038347 * winsorized_Income_Comp_Of_Resources
0.6604446850901278 * winsorized_Schooling
```

Figure 23: Weights of features of the linear regression model using L2-regularization after training

In comparison with the weights of the usual linear regression, it is observed that after the use of Ridge regularization, the larger weights of the features decreased (approached zero), but did not turn into zero. So the selection of signs did not take place, but we set a penalty for disproportionately large weights and brought them closer to zero.

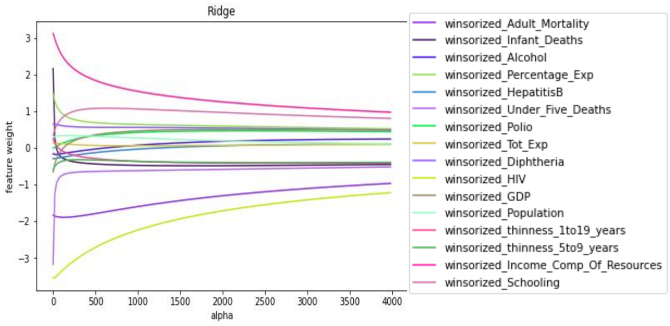


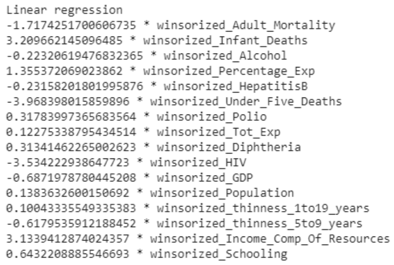
Figure 24: Chart of the dynamics of weights relative to the parameter of regularization α when using Ridge-regression

Figure 24 Legend for the chart of the dynamics of weights relative to the parameter of regularization α when using Ridge regression

It is observed that with increasing L2-regularization parameter, the weights gradually go to zero, but do not turn to zero. At α = 0 the weights of the features are directed to minimize the error function.

3) Elastic Net regularization

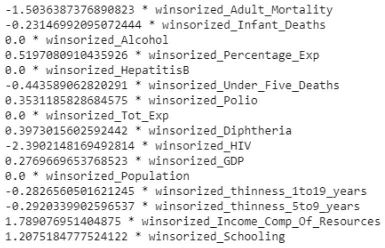
Weights of signs of linear regression without regularization (Figure 25):



```
Linear regression
-1.7174251700606735 * winsorized_Adult_Mortality
3.209662145096485 * winsorized_Infant_Deaths
-0.22320619476832365 * winsorized_Alcohol
1.355372069023862 * winsorized_Percentage_Exp
-0.23158201801995876 * winsorized_HepatitisB
-3.968398015859896 * winsorized_Under_Five_Deaths
0.31783997365683564 * winsorized_Polio
0.12275338795434514 * winsorized_Tot_Exp
0.31341462265002623 * winsorized_Diphtheria
-3.534222938647723 * winsorized_HIV
-0.6871978780445208 * winsorized_GDP
0.1383632600150692 * winsorized_Population
0.10043335549335383 * winsorized_thinness_1to19_years
-0.6179535912188452 * winsorized_thinness_5to9_years
3.1339412874024357 * winsorized_Income_Comp_Of_Resources
0.6432208885546693 * winsorized_Schooling
```

Figure 25: Weights of features of the model of linear regression after training

Weight of features of elastic net regression (application of L1-L2-regularization) (Figure 26):



```
-1.5036387376890823 * winsorized_Adult_Mortality
-0.23146992095072444 * winsorized_Infant_Deaths
0.0 * winsorized_Alcohol
0.5197080910435926 * winsorized_Percentage_Exp
0.0 * winsorized_HepatitisB
-0.443589062820291 * winsorized_Under_Five_Deaths
0.3531185828684575 * winsorized_Polio
0.0 * winsorized_Tot_Exp
0.3973015602592442 * winsorized_Diphtheria
-2.3902148169492814 * winsorized_HIV
0.2769669653768523 * winsorized_GDP
0.0 * winsorized_Population
-0.2826560501621245 * winsorized_thinness_1to19_years
-0.2920339902596537 * winsorized_thinness_5to9_years
1.789076951404875 * winsorized_Income_Comp_Of_Resources
1.2075184777524122 * winsorized_Schooling
```

Figure 26: Weights of features of the linear regression model using ElasticNet-regularization after training

In comparison with the weights of simple linear regression, it is observed that after the use of ElasticNet regularization, some weightsofnon-informative features turned to zero, and other disproportionately large weights approached zero. This was achieved through the use of two penalties L1 and L2 regularization.

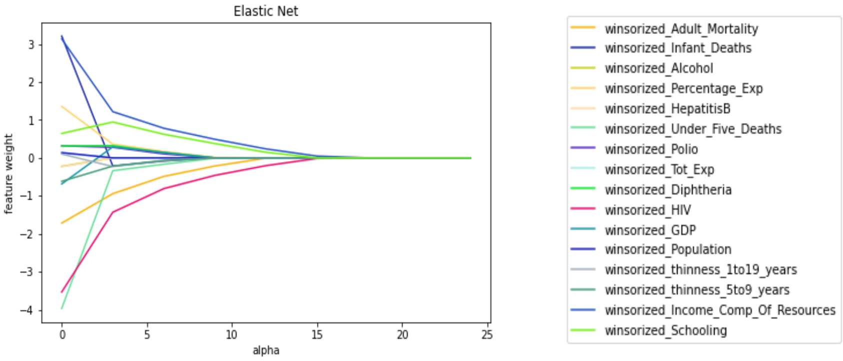


Figure 27: Chart of the dynamics of weights relative to the parameter of regularization α when using ElasticNet-regression

Figure 27 Legend for the chart of the dynamics of weights relative to the parameter of regularization α when using ElasticNet-regression.

It is observed that with increasing ElasticNet-regularization parameter, the weights at the features go to zero, but not as fast as it happens in L1-regularization and not as slowly as observed when using L2-regularization.

9. Conclusion

L1-, L2-, ElasticNet-regularizations of classification and regression were investigated in the course of work.

First of all, the WHO Statistics on Life Expectancy dataset was analyzed, prepared and cleaned. It was checkedif the data types match the attributes of dataset. Zero values were eliminated by interpolation, and atypical values of each attribute of the dataset were eliminated by the method of winsorization. The distribution and scope of the values of each variable are investigated and demonstrated using scale diagrams and bar charts, respectively. Linear dependences between the target variable and the rest of the dataset variables are determined. A correlation map of attributes was constructed and on the basis of it was determined that the set task of life expectancy prediction can be realized by linear methods.

A linear regression model was created using the scikit-learn library. After her training, the weights of the model features were obtained and it was observed that the weights at strongly correlated features were greater than the rest. Thus, the problem of multicollinearity was identified. The quality metrics of the linear regression model were calculated, namely: root mean square error, mean absolute error and coefficient of determination. The root mean square error indicated that the model was wrong in 12.9% of cases, the mean absolute error – in 2.7% of cases. The coefficient of determination is ≈ 0.86, which indicates that the trained model describes 86% of the variance and is reasonable because the coefficient of determination is in the range from zero to one.

To eliminate the problem of multicollinearity, 3 regularization methods were applied and compared.

The Lasso regression model was created and after its training the weights of the features were obtained. It was observed that this type of regularization carried out the selection of features and turned the weights at non-informative features to zero. The dynamics of Lasso regression weights with increasing α regularization parameter was monitored. It was found that with increasing α, the weights rapidly approach zero, and with sufficiently large α all weights turn into zero. As α increases, the model becomes simpler.

During the Ridge regression, it was observed that the large weights approached zero, but none of them turned into zero. The dynamics of Ridge regression weights was observed and it was found that even at a sufficiently large α the weights do not turn into zero, but slowly asymptotically approach zero.

After implementing ElasticNet regression, some of the weights turned to zero and some approached zero. This is due to the application of penalties of both L1 and L2 regularizations in this method. The change in weights with increasing α parameter was observed. It was observed that with increasing α the weight of the features tends to zero, but not as rapidly as it occurs when using L1-regularization. But in contrast to L2-regularization, at a sufficiently large α all weights are converted to zero.

Therefore, L1-regularization is better used in cases where it is known that some of the attributes are unimportant, because when using this regularization the selection of features will be conducted that will turn the weight of non-informative features to zero. L2-regularization is better to use when it is known that all variables of the dataset are important in predicting the target variable, because when using this regularization disproportionately large weights will approach zero, but the selection of features will not occur. ElasticNet regularization is universal. It is suitable for the two cases described above, and especially for cases where it is not known which variables are important and which are not, or when the dataset has a very large number of variables.

10. References

[1] A.K Tung, J. Hou, J. Han, Spatial clustering in the presence of obstacles, in: The 17th Intern. conf. on data engineering (ICDE’01), Heidelberg, 2001, pp. 359–367.

[2] C. Boehm, K. Kailing, H. Kriegel, P. Kroeger, Density connected clustering with local subspace preferences, in: Proc. of the 4th IEEE Intern. conf. on data mining, IEEE Computer Society, Los Alamitos, 2004, pp. 27–34.

[3] D. Guo, D.J. Peuquet, M. Gahegan, ICEAGE: Interactive clustering and exploration of large and high-dimensional geodata, vol. 3, N. 7, Geoinfor-matica, 2003, pp. 229–253.

[4] D. Harel, Y. Koren, Clustering spatial data using random walks, in: Proc. of the 7th ACM SIGKDD Intern. conf. on knowledge discovery and data mining, San Francisco, California, 200, pp. 281–286.

[5] N. Boyko, M. Kuba, L. Mochurad, S. Montenegro, Fractal Distribution of Medical Data in Neural Network, in: The 2 nd International Workshop on Informatics & Data-Driven Medicine (IDDM 2019), Volume 1. Lviv, Ukraine, November 11-13, 2019, pp. 307-318.

[6] D.J. Peuquet, “Representations of space and time”, N. Y.: Guilford Press, 2002.

[7] H.-Y. Kang, B.-J. Lim, K.-J. Li, P2P Spatial query processing by Delaunay triangulation, Lecture notes in computer science, vol. 3428, Springer/Heidelberg, 2005, pp. 136–150.

[8] M. Ankerst, M. Ester, H.-P. Kriegel, Towards an effective cooperation of the user and the computer for classification, in: Proc. of the 6th ACM SIGKDD Intern. conf. on knowledge discovery and data mining, Boston, Massachusetts, USA, 2000, pp. 179–188.

[9] N. Shakhovska, N. Boyko, P. Pukach, The information model of cloud data warehouses, in: Advances in intelligent systems and computing III. Selected papers from the International conference on computer science and information technologies, CSIT 2018, September 11-14, Vol. 871, Lviv, Ukraine, pp. 182–191.

[10] C. Zhang, Y. Murayama, Testing local spatial autocorrelation using, in: Intern. J. of Geogr. Inform. Science, vol. 14, 2000, pp. 681–692.

[11] R. Agrawal, J. Gehrke, D. Gunopulos, P. Raghavan, Automatic sub-space clustering of high dimensional data, in: Data mining knowledge discovery, vol. 11(1), 2005, pp. 5–33.

[12] V. Estivill-Castro, I. Lee, Amoeba: Hierarchical clustering based on spatial proximity using Delaunay diagram, in: 9th Intern. Symp. on spatial data handling, Beijing, China, 2000, pp. 26–41.

[13] N. Boyko, V. Korkishko, B. Dohnyak, O. Vovk, Use of Neural Networks in Q-Learning Algorithm, in: 32nd International Symposium on Computer and Information Sciences, ISCIS 2018: Computer and Information Sciences, Poznan, Poland, September 20-21, 2018, pp. 188-195.

[14] I. Turton, S. Openshaw, C. Brunsdon, Testing spacetime and more complex hyperspace geographical analysis tools, in: Innovations in GIS 7, London: Taylor & Francis, 2000, pp. 87–100.

[15] N. Boiko, The issue of access sharing to data when building enterprise information model, in: IX International Scientific and Technical conference, Computer science and information technologies (CSIT 2014), Lviv, Ukraine, 2014, pp. 23-24.

[16] C. Aggarwal, P. Yu, Finding generalized projected clusters in high dimensional spaces, in: Intern. conf. on management of data, ACM SIGMOD, 2000, pp. 70–81.

[17] C.M. Procopiuc, M. Jones, P.K. Agarwal, T.M. Murali, A Monte Carlo algorithm for fast projective clustering, in: Intern. conf. on management of data, ACM SIGMOD, Madison, Wisconsin, USA, 2002, pp. 418–427.