statistical inference - midterm project

Dan Ben Ami, ID: 316333079
Elad Sofer, ID: 312124662

December 2022

1 Introduction

In this task, we asked to predict Life Expectancy(for data given in test.csv file) using Linear Regression tool from the data provided here(train.csv file). The data is originally from WHO and United Nations website (courtesy: Deeksha Russell and Duan Wang). The features in train subset are as follows: Country, Year, Status, Life expectancy, Adult Mortality, infant deaths, Alcohol,percentage expenditure, Hepatitis B, Measles , BMI, under-five deaths, Polio, Total expenditure, Diphtheria, HIV/AIDS, GDP, Population, thinness 1-19 years, thinness 5-9 years, Income composition of resources, and Schooling.

2 Data Processing

Most of the work to improve the predictive capabilities of the model is done through data processing. We will describe here below all the data processing that we tested as well as which processing we finally chose to perform out of all the different attempts.

2.1 Categorical Features

Within the data there are two categorical features: 'Country', 'Status'.

In the 'Status' feature, there are only 2 options, so we moved it to a binary value where: 1='developing', 0='not developing'.

In the 'Country' feature, we checked and saw that there are 183 different countries. Therefore, first when we wanted to perform a linear regression model, we turned each country into a separate feature, that is, we added 183 binary features to the data, one for each country. After tests and trials, we saw that it is better to use the Kernel Ridge model, in which it is better to reduce the number of features, therefore instead of adding 183 features, we left the "country" as a single feature in which we gave each country a numerical value that is proportional to the "life expectancy" value its.

2.2 Missing or Incomplete Data

First we removed all the data points (rows) in which the "life expectancy" value was missing. After that, since the amount of data points with at least one missing value was very large and constituted a significant part of the total data, we chose not to exclude

the aforementioned data points but to complete it in a smart way. We tried several methods to complete the missing cells in each data point such as: completion By the general average/median of the entire feature, completion by the average/median of the feature but only in the country to which it belongs, completion by a value identical to the next data point where the feature exists (the value of the feature r in the next line where it appears). Finally we saw that the best result was completion by a general median of the feature (over all the data).

2.3 Normalization

We tried two different normalization methods:

1. Standard normalization: For each feature, we calculated its average and standard deviation throughout the data, and from each data point in the feature we subtracted the average and divided by the standard deviation.

2. For each feature, we checked the minimum value and the maximum value in this feature, and for each data point in this feature throughout the data, we subtracted the minimum value and divided by the maximum.

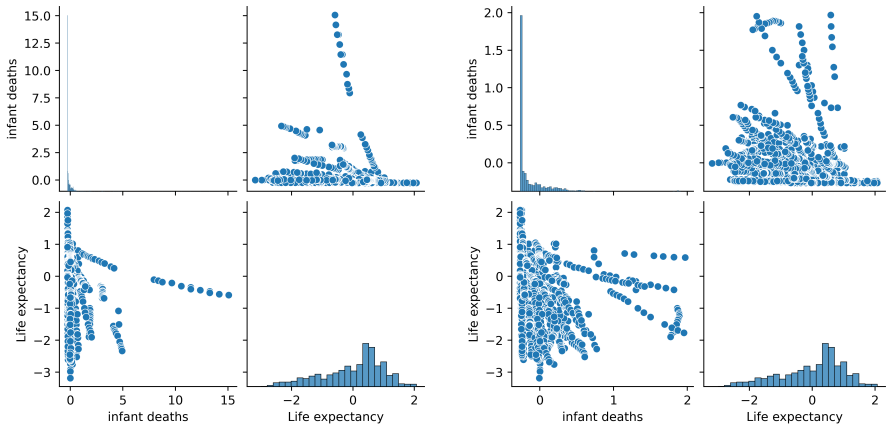
Finally we chose the standard normalization as it produced the best results.

2.4 Feature Selection

In order to reduce the amount of features and find irrelevant features, we performed the Backward Elimination algorithm, in which we actually performed a normal linear regression and then dropped the feature with the largest P-value. Then, we again performed a linear regression (without the removed feature) and again dropped the feature with the largest P-value from the remaining features. This is how we did it time after time again, between each time we checked the performance of the model on the test data, and we stopped dropping features only when the performance of the model on the test data decreased.

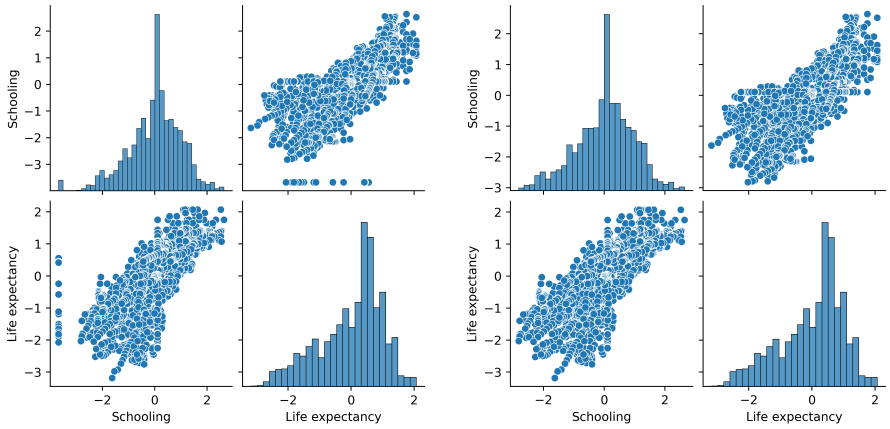
2.5 Outliers Correction

After we finished processing the data and normalizing it, we tried to remove outliers. For this purpose, for each feature we created a graph that describes the "life expectancy" as a function of that feature, in addition to the histogram of the feature. For this purpose, for each feature we created a graph that describes the "life expectancy" as a function of that feature, in addition to the histogram of the feature. After that we examined the graphs and decided which data points and which features should correct the outliers. Because we didn't want to remove the data points of the outliers just because of an extreme value in a single feature. For each feature we identified the corresponding data points, performed a two-dimensional (inverse) linear regression with the remaining data points (those not identified as outliers) of the feature as a function of "life expectancy" and thus assigned a new feature value to the outliers according to their "life expectancy". In figure 1-3, he showed several examples of the graphs we produced and how they looked before and after changing the values of the outliers.



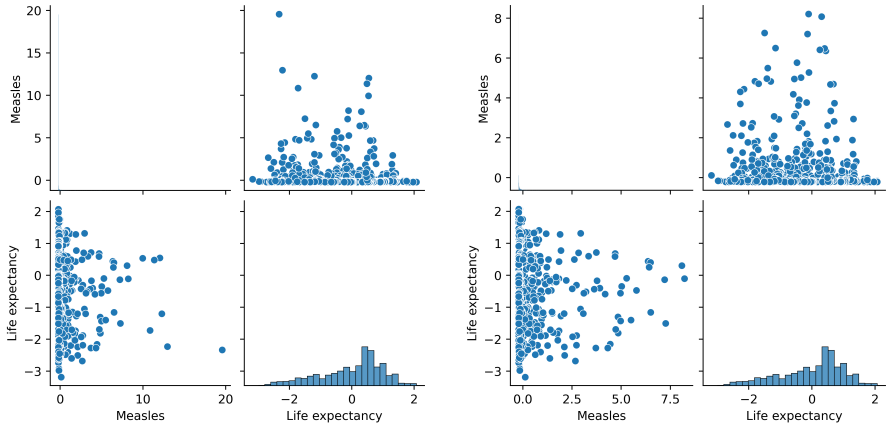
(a) before outliers correction (b) after outliers correction

Figure 1: "infant deaths" feature as a function of "life expectancy" feature.



(a) before outliers correction (b) after outliers correction

Figure 2: "Schooling" feature as a function of "life expectancy" feature.



(a) before outliers correction

(b) after outliers correction

Figure 3: "Measles" feature as a function of "life expectancy" feature.

3 Model Selection

We examined a number of different models: linear regression, ridge regression, lasso regression and kernel ridge. The model that produced the best results was Kernel Ridge, so we decided to continue with it.

4 Conclusions

During the task, we tested and learned about a lot of data processing methods, normalization, checking outliers, removing features, completing missing values, different models, etc. One of the most striking conclusions from the task was that there is no one method that is better than the rest, each model behaves differently with different data processing, and in many cases several types of data processing resulted in different and opposite results with different models. Therefore, we tried to examine as many combinations of different types of data processing and different types of models as possible.