Instituto Superior Técnico

Mechanical Engineering

Advanced Automation

Final Project – Report

# Prediction of Life Expectancy

*Autors:*

| | |
|---|---|
| 98482 | Júlio Martinho |
| 91644 | Giovanna Mazzali |
| 101760 | Bruno Oshiro |

April 22, 2025

# Summary

# 1 Introduction

Life expectancy is the approximated number of years a group of individuals is expected to live based on multiple demographic factors. In this report, there will be presented six methods to predict life expectancy of countries, based on data downloaded from organizations API's such as the world bank of data and the World Health Organization (WHO).

The main objective, besides comparing different methods to find which of the analyzed parameters have more influence in the life expectancy value, is to build a database by filtering and merging tables with real world data and put it together in a clean dashboard to visualize the data. In this work, Python, SQL, and powerBI tools were used.

# 2 Database

In this section, it is intended to show which data was selected to compose the database, how it was filtered and the final result, focusing on discussing the decisions that were made regarding it.

## 2.1 Collecting Data from the Internet

Collecting data to analyze is one of the most important parts of the proccess when it comes to building predictors. The first step is to define what data is pertinent to the work. Regarding life expectancy, the defined type of data was: population characteristics, such as cultural aspects (for example, diet); climate and environment elements, such as air pollution; population conditions, such as access to sanitation services; government's issues, such as corruption and $CO_2$ emissions. Another important factor for obtaining data for this object of study is the availability of information per country and years. In order to have the most complete database possible, the tables collected were the ones with information of at least about 100 countries and 20 years.

## 2.2 Filtering Data

This subsection refers to the following jupyter notebook file (located in the project folder): `Projeto_Final_export_tables.ipynb`.

Over 40 tables in **.csv** format were obtained to analysis. From that set, about 25 were indeed treated. The discarted tables were either missing most of the values, had no distinction per country, had a poor range of years or did not have the relevant information expected to the matter.

The 25 remaining ones included adult mortality rate, children overweight, corruption perceptions index, dayli supply of calories, deaths by cause, access to electricity, gross domestic product (GDP), industrialization intensity index, life expectancy at birth, adult literacy rate, mean body mass (both male and female), misery and extreme poverty, under 5 years old deaths, obesity, access to sanitation services, population, rural and urban population percentages, substances related deaths, suicide rate and estimates of homicides, $CO_2$ emissions, air pollution.

Most of these tables came in 4 main formats, which will be presented below. All the tables were treated in the jupyter notebook (Python) in order to follow the pattern defined, that is composed by columns named *Country Name*, *Country Code*, *Year*, *Sex* (when possible) and the parameter name. All the lines with *not a number (nan)* value after the table treatment were excluded. Also, all the spaces in the names of the variables were substituted by ”_”. Only values of years from 1960 were considered because of the lack of data in earlier years in most of the tables.

The first type of table is shown in Table 1. It had the years, the indicator name and indicator code as columns. It was changed to one column for the years, one column named after the parameter in analysis (for the table represented, it was the value for life expectancy), the indicator name and code were both eliminated. The country name and code remained the same.

**Table 1:** Example of format 1 of `.csv` tables obtained - life expectancy.

| Country Name | Country Code | Indicator Name | Indicator Code | 1960 | ⋯ |
|---|---|---|---|---|---|
| Aruba | ABW | Life expectancy at birth, total (years) | SP.DYN.LE00.IN | 65.662 | ⋯ |
| Africa Eastern and Southern | AFE | Life expectancy at birth, total (years) | SP.DYN.LE00.IN | 42.71605 | ⋯ |
| ⋮ | ⋮ | ⋮ | ⋮ | ⋮ | ⋱ |

The second format of table is shown in Table 2. It had a lot of empty columns that had to be eliminated. Besides, the value of the parameter was presented with an interval, which is related to the statistical treatment of the source's organization. For this matter, only the value outside the brackets was considered and transfered to a column for the parameter value. The country name and code were refered as *Location* and *SpatialDimValueCode*, so both were modified to *Country Name* and *Country Code*. The rest of the columns were disconsidered.

**Table 2:** Example of format 2 of .csv tables obtained - suicide rate.

| Indicator Code | Indicator | ⋯ | Dim1 type | Dim1 | ⋯ | Value |
|---|---|---|---|---|---|---|
| MH_12 | Age-standardized suicide rates (per 100 000 population) | ⋯ | Sex | Male | ⋯ | 0 [0 - 0] |
| MH_12 | Age-standardized suicide rates (per 100 000 population) | ⋯ | Sex | Female | ⋯ | 0.16 [0.11 - 0.22] |
| ⋮ | ⋮ | ⋮ | ⋮ | ⋮ | ⋮ | ⋱ |

The third type of **.csv** file obtained is shown in Table 3. It was the simplest one to treat. The first column was deleted, and the *Country* and *Obesity (%)* ones were renamed to *Country Name* and *Obesity*. Also, the interval data was neglected, and only the outsider value was included.

The fourth format of table obtained is shown in Table 4. It only required renaming the columns to the defined default. In this case, *Entity* was renamed to *Country Name*, *Code* to *Country Code*, and the columns refered to the causes of deaths were changed to the word that indentifies each one best (for example, *Deaths - Meningitis - Sex: Both - Age: All Ages (Number)* became *Meningitis*).

**Table 3:** Example of format 3 of .csv tables obtained - obesity.

| | Country | Year | Obesity (%) | Sex |
|---|---|---|---|---|
| 0 | Afghanistan | 1975 | 0.5 [0.2-1.1] | Both sexes |
| 1 | Afghanistan | 1975 | 0.2 [0.0-0.6] | Male |
| 2 | Afghanistan | 1975 | 0.8 [0.2-2.0] | Female |
| ⋮ | ⋮ | ⋮ | ⋮ | ⋱ |

**Table 4:** Example of format 4 of .csv tables obtained - deaths by cause.

| Entity | Code | Year | Number of executions (Amnesty International) | Deaths - Meningitis - Sex: Both - Age: All Ages (Number) | ⋯ |
|---|---|---|---|---|---|
| Afghanistan | AFG | 2007 | 15 | 2932.559 | ⋯ |
| Afghanistan | AFG | 2008 | 17 | 2730.846 | ⋯ |
| ⋯ | ⋯ | ⋯ | ⋯ | ⋯ | ⋱ |

### 2.3 Exporting to pgAdmin 4 and Final Database

After all the tables were obtained and treated, they were imported to *pdAdmin 4*, more specifically to the *Project* database, to be proccessed and merged in *SQL* in order to compose a final database.

Since most of the data obtained is in the [1990,2016] year range, emerged the need to exclude columns that had *nan* values, because reducing more the dataset would result in a way too poor database. The final table format, object of the prediction methods, is shown in Table 5.

**Table 5:** Final table.

| country_name | country_code | year | life_expectancy | parameter1 | parameter2 | ⋯ |
|---|---|---|---|---|---|---|
| country | NNN | yyyy | value | value1 | value2 | ⋯ |
| ⋯ | ⋯ | ⋯ | ⋯ | ⋯ | ⋯ | ⋱ |

This table allows the study of life expectancy of a country in a determined year based on the parameters in the columns.

## 3 Prediction Methods

In this section, the different prediction methods applied on the database are going to be discussed and compared. It is assumed that the reader is familiar with all the methods. In this approach, training set is composed by all the data of years from 1990 to 2011, and the test set is from 2012 to 2016. The countries of the life expectancy predictions shown in this section are Brazil and South Africa. The accuracy indexes used to compare the results are $R^2$ value and the mean square error (mse).

## 3.1 Linear Regression

By applying the linear regression method, the prediction for both countries are shown in Figures 1 and 2. In this predictions, which had over 40 parameters, the values found for $R^2$ are 0.9691 for Brazil, and 0.4399 for South Africa. It's no surprise that the method turned out this good in Brazil's case, because the life expectancy line for the country has a behaviour very close to a linear function. However, South Africa's line is more similar to a polynomial function, which justifies the lower $R^2$ value.

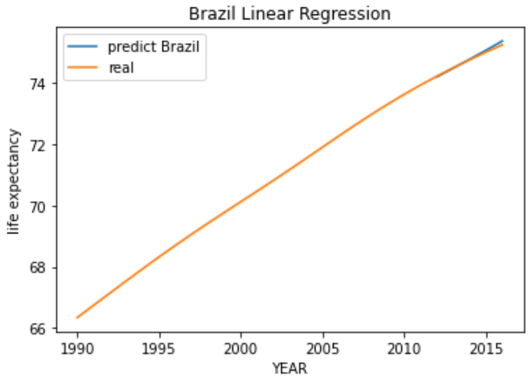


**Figure 1:** Brazil's life expectancy prediction by linear regression.

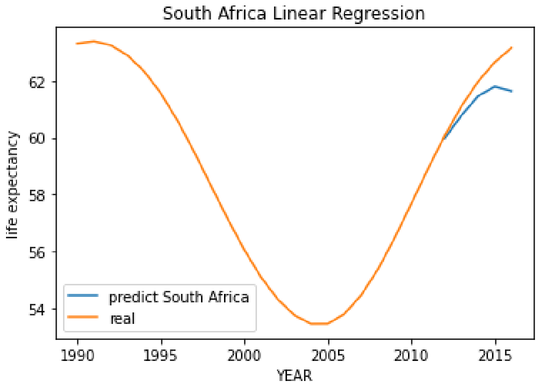


**Figure 2:** South Africa's life expectancy prediction by linear regression.

## 3.2 Ridge Regression

The essential part of ridge regression is determining the value of the tuning parametrer $\alpha$ that reaches the best *bias-variance trade-off* result and leads to the minumum mse value. For both countries analyzed, the relation between the $\alpha$ values and the mse value are shown in Figures 3 and 4. For Brazil's ridge regression prediction, the best $\alpha$ and mse values found were respectively 4328761281 and 0.0006193. For South Africa's study, the best mse found was 0.004103 for $\alpha = 932.60$.

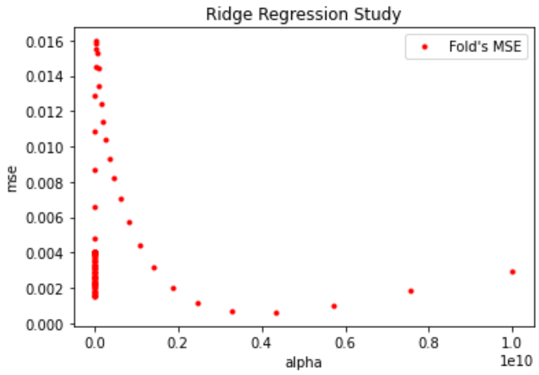


**Figure 3:** Brazil's mse vs $\alpha$.

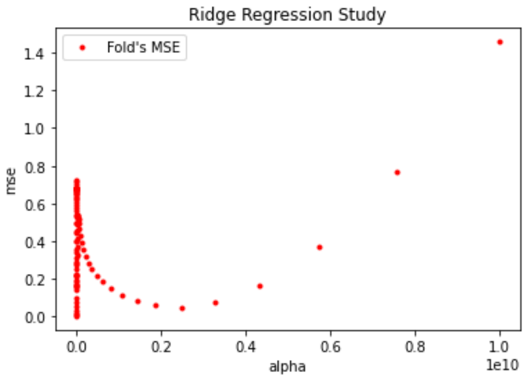


**Figure 4:** South Africa's mse vs $\alpha$.

The plot of the life expectancy prediction of each country by ridge regression is shown in Figures 5 e 6. The $R^2$ values found were, respectively, 0.9952 and 0.9963.

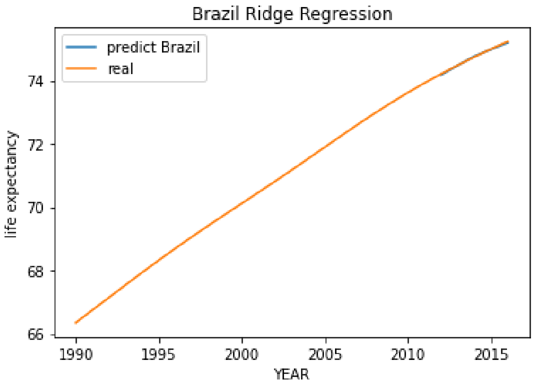


**Figure 5:** Brazil's ridge regression prediction.

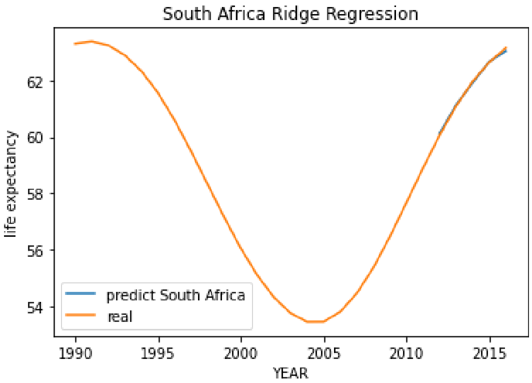


**Figure 6:** South Africa's ridge regression prediction.

## 3.3 Lasso Regression

The third method used to predict life expectancy is lasso regression. Similarly to ridge regression, lasso regression also estimates the best tuning parameter $\alpha$ in order to lower the mse value. The main difference is that the $\alpha$ value here results in such a penalty that performs a selection of parameters, so irrelevant parameters may have no influence at all in the prediction. The study of $\alpha$ values for each prediction are shown in Figures 7 and 8.

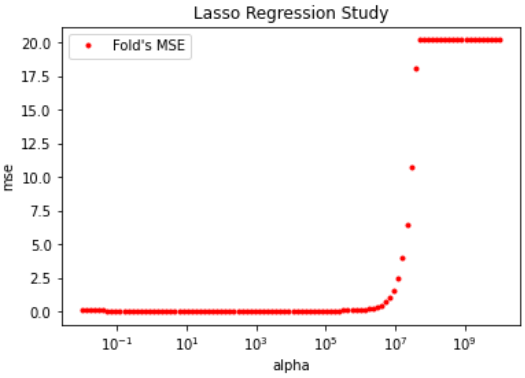


**Figure 7:** Brazil's mse vs $\alpha$.

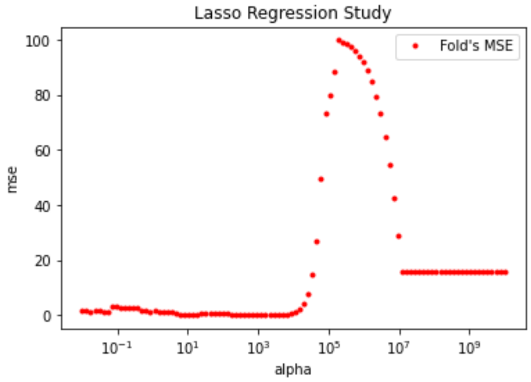


**Figure 8:** South Africa's mse vs $\alpha$.

The best fit for Brazil resulted in $R^2 = 0.9971$, mse $= 0.0003752$ and $\alpha = 14.17$, while South Africa resulted in $R^2 = 0.9612$, mse $= 0.04724$ and $\alpha = 3764.93$. Even though Brazil's error indicators are better than South Africa's, both predictions resulted in satisfying behaviours, as shown in Figures 9 and 10.

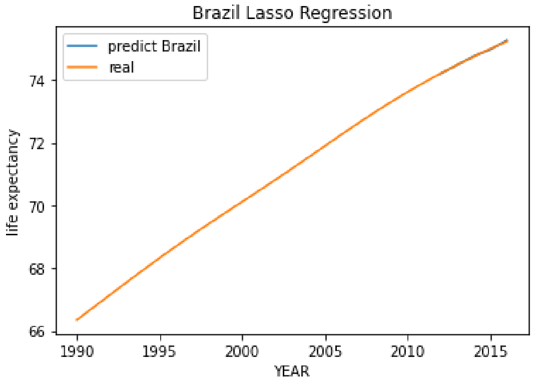


**Figure 9:** Brazil's lasso regression prediction.

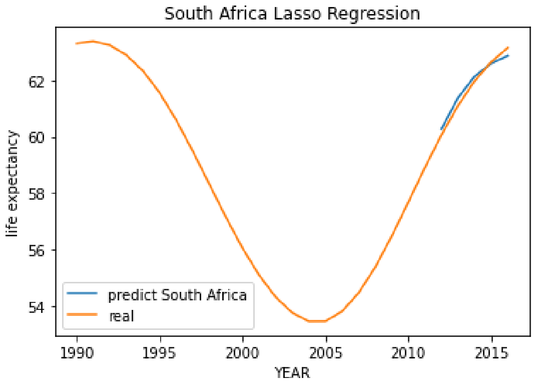


**Figure 10:** South Africa's lasso regression prediction.

## 3.4 Decision Tree

This appproach relies on a tree-based method for regression. The decision tree splitting rules to segment the predictor space can be summarized in a tree. The Figures 11 and 12 show the study of the depth (splits of the tree) versus the mse value. The more splitted the tree is, the more complex the model is. In Brazil's prediction, the best values found are $max\_depth = 6$ and $R^2$ = -32.40. For South Africa's one, the values are $max\_depth = 29$ and $R^2$ = -0.04143. Negative values for $R^2$ indicator suggest that the model cannot be trusted. The prediction for both countries is shown in Figures 13 and 14. It's clear from the error indicator and the figures that the prediction did not perform well for both countries.

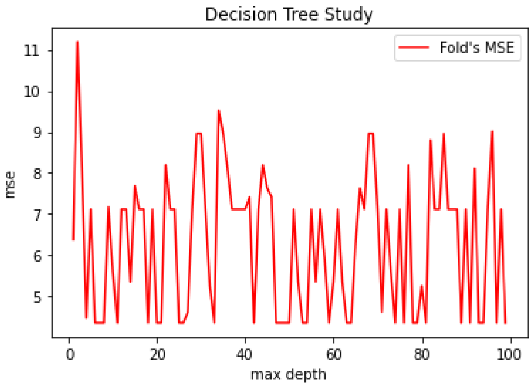


**Figure 11:** Brazil's mse vs depth.

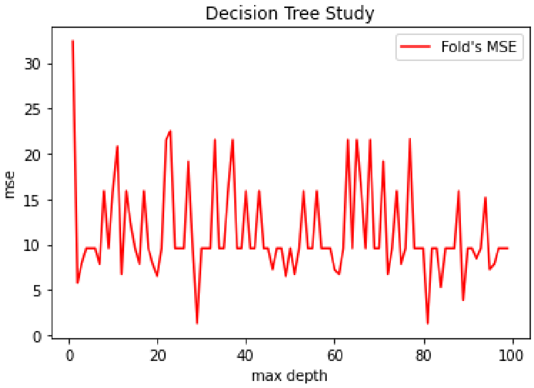


**Figure 12:** South Africa's mse vs depth.

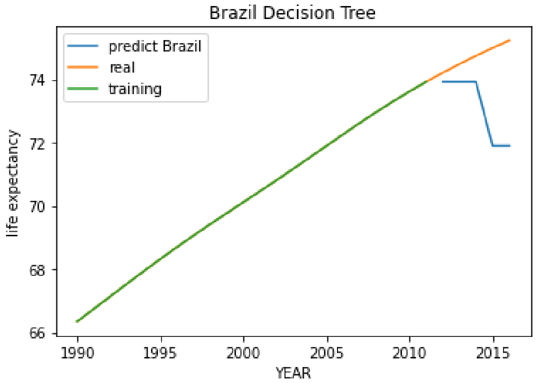


**Figure 13:** Brazil's decision tree prediction.

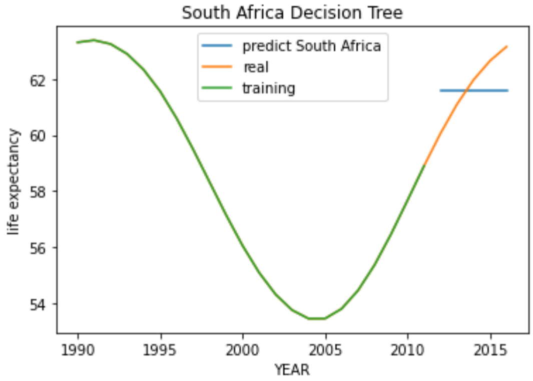


**Figure 14:** South Africa's decision tree prediction.

## 3.5 Random Forest

Random forest is a alternative method to decision tree regression. In simple words, it differs by having a proccess that decorrelates the trees, thereby making the average of the resulting trees less variable and hence more reliable. In this predictor, the best values found in terms of maximum number of features and mse (Figures 15 and 16) are, respectively, 21 and 1.4245 for Brazil, and 20 and 11.2509 for Africa.

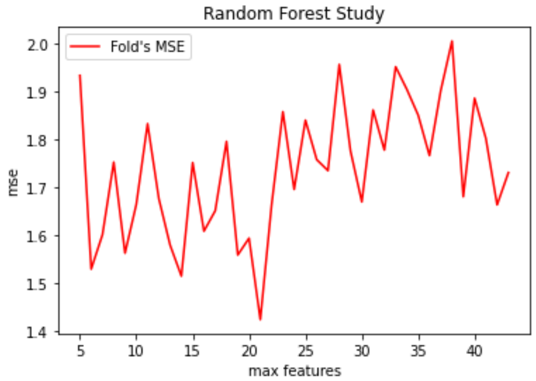


**Figure 15:** Brazil's mse vs maximum number of features.

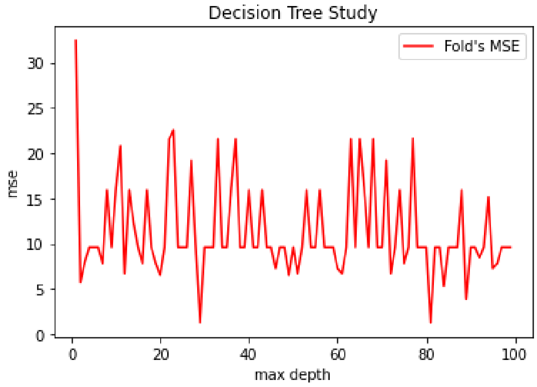


**Figure 16:** South Africa's mse vs maximum number of features.

In Figure 17 the Brazil's life expectancy prediction is shown. Even though it's sligthly better than the previous one, it's still a bad prediction. Africa's prediction got a very similar result to the decision tree one, as Figure 18 shows.

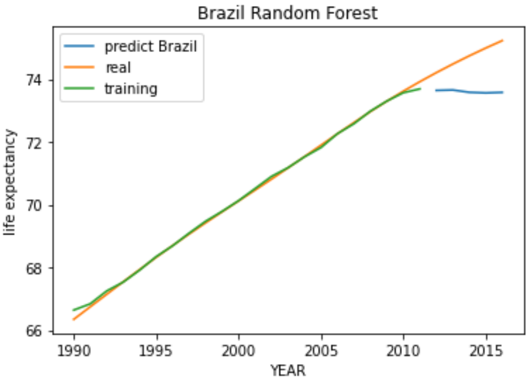


**Figure 17:** Brazil's random forest prediction.

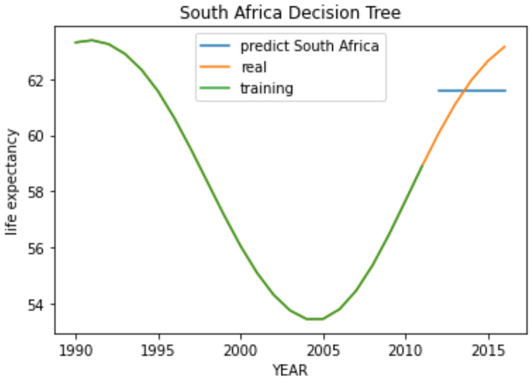


**Figure 18:** South Africa's random forest prediction.

## 3.6 Gradient Boosting

In gradient boosting, the tree is fit using the residuals, rather than the outcome. In Figures 19 and 20, the mse versus learning rate value study is shown. For Brazil's prediction, the learning rate found is 1.6, for South Africa's case, it is 0.05.

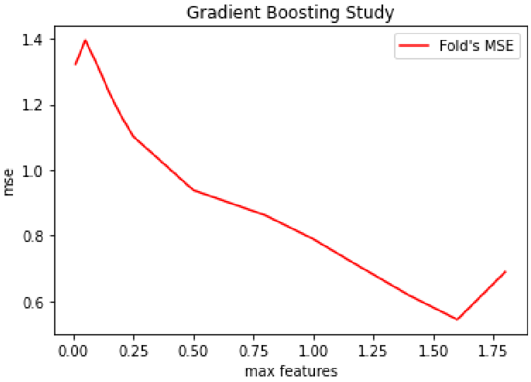


**Figure 19:** Brazil's mse vs maximum number of features.

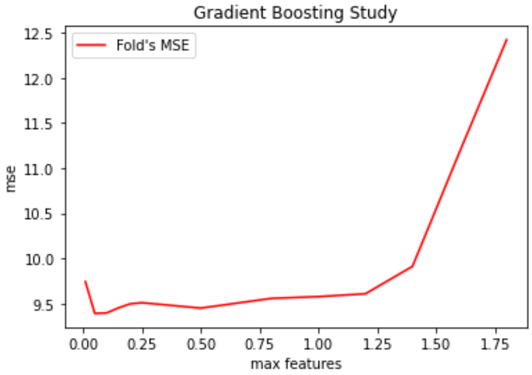


**Figure 20:** South Africa's mse vs maximum number of features.

In Figures 21 and 22, it's shown the plot for the predictions. With $R^2$ scores of -3.1610 and -6.6998, both Brazil and South Africa approximations cannot be trusted.

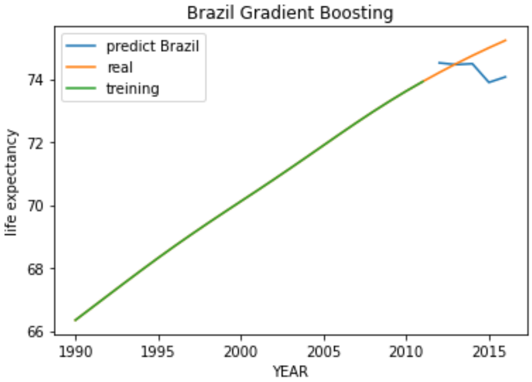


**Figure 21:** Brazil's gradient boosting prediction.

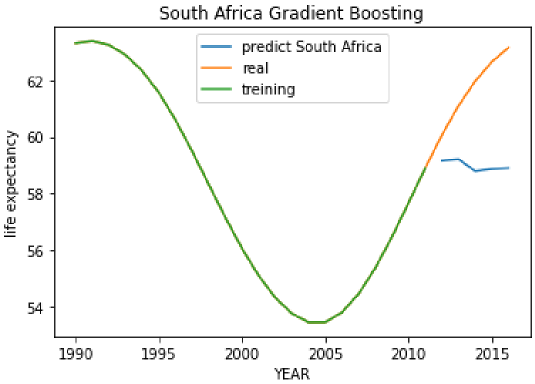


**Figure 22:** South Africa's gradient boosting prediction.

### 3.7 Discussion of Results

It's reasonable to agree that the linear, ridge and lasso predictors resulted in good models for life expectancy prediction. On the other hand, decision tree, random forest and gradient boosting predictors resulted in not trustable life expectancy predictions.

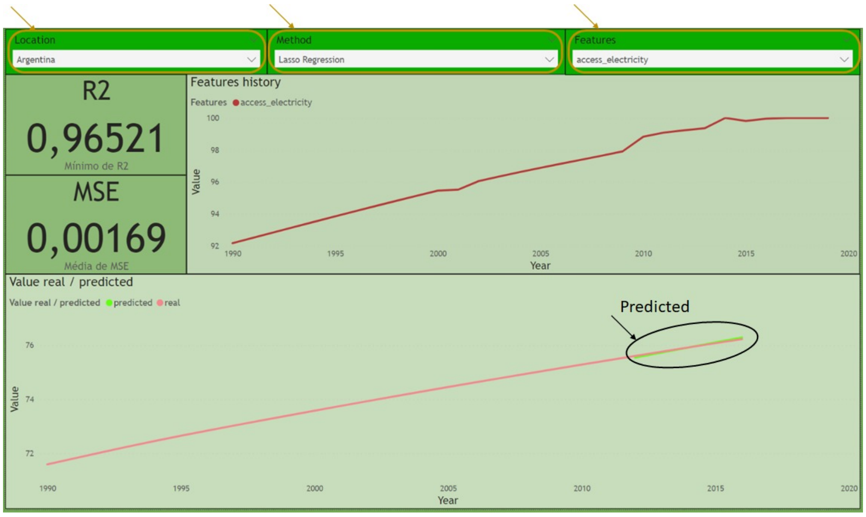
Comparing the first three predictors, the results of ridge and lasso predictions are slightly better than the linear regression result. The main difference between these predictors is that ridge and lasso apply a penalty on unimportant parameters, resulting in more reliable models.

The last three predictors, all tree-based methods, resulted in $R^2$ negative values for the prediction. It means that the chosen model does not follow the trend of the data, so fits worse than a horizontal line. It may be explained it taking in account that for tree-based methods it's likely to overfit the data, leading to poor test set performance.

## 4 Dashboard

In order to visualize the results of the different methods studied for each of the countries that compose the database, a dashboard was created. The interface is shown in Figure 23. It has two graphs: the first shows the data of the feature selected; the second shows the life expectancy prediction of the country in analysis. It's also displayed the values for $R^2$ and mse of the prediction.

In the dashboard, it's possible to choose the location (country), the predictor method (from the six presented in this report) and the feature. Figure 24 ilustrates how the interative component of it works.



Location: Argentina

Method: Lasso Regression

Features: access_electricity

R2

0,96521

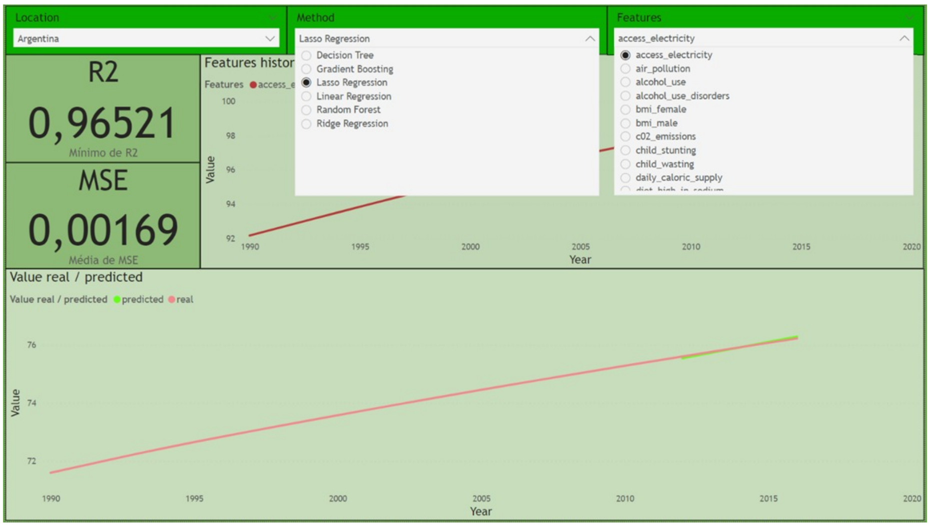
Mínimo de R2

MSE

0,00169

Média de MSE

**Figure 23:** Dashboard interface.



Location: Argentina

Method: Lasso Regression

Features: access_electricity

R2 0,96521 Mínimo de R2

MSE 0,00169 Média de MSE

**Figure 24:** Dashboard 2

## 5 Conclusion

When it comes to making predictions, the most important step is to find good data to rely on. The decisions made when filtering and proccessing it are decisive to whether the prediction reaches good results. Besides that, it's very important to know the data, because it implies on the methods that are used and in what to expect.

From the results obtained, it is possible to conclude that, for different countries, different methods will work best. The main purpose was to analyze the results and verify if it makes sense or not based on the data available.

## References

[1] James Gareth et al. *An Introduction to Statistical Learning: With Applications in R*. 2nd ed. Springer, 2013. ISBN: 978-1-4614-7138-7. DOI: 10.1007/978-1-4614-7138-7.