Explaining Life Expectancy

Statistical Learning Project

Federico Bassi – ID: 993443

July 2022

1. Research question

This project moves from a personal curiosity about how Machine Learning techniques can be applied to tackle important social and economic problems. In particular, this project aims at answering the following question: which factors are relevant in explaining life expectancy around the world?

This research question seems to me interesting and important at the same time: finding an answer could provide important information and relevant insights for policymakers and international organizations.

Starting from a dataset containing information about health and economic variables of different States around the world, in this project I developed some Supervised and Unsupervised Learning models that could help explaining the differences between life expectancy in the world.

Given the research question and the data at hand, I focused my attention on Supervised models that could be easily interpreted and potentially exploited for policy analyses: these are linear regression, best subset selection, Ridge Regression, LASSO, Elastic Net and -finally- Regression trees and Random Forests.

Unsupervised Learning algorithms have been used, first of all, to reduce the dimensionality of the dataset, which -as will be shown shortly- contained some strongly correlated features. Moreover, a clustering method has been applied to the set of features at the end of the project.

This report is organized as follows: in Section 2, I describe the dataset and its variables; in Section 3, I describe Linear Regression model and diagnostics; in Section 4, I explain how Principal Component Analysis has been applied to this exercise. Section 5 and Section 6 describe the results of the application of, respectively, Best Subset Selection algorithm and Regularization techniques such as Ridge Regression, LASSO and Elastic Nets. In Section 7, finally, I describe Regression Trees and Random Forests. Section 8 reports a comparison of performances and a conclusion on the Supervised Learning models. The final Section, Section 9, describes K-means clustering algorithm and its application to the dataset.

2. Dataset exploration and descriptive statistics

The dataset upon which this analysis has been carried out was built using data from GHO (Global Health Observatory) and UNESCO (United Nations Educational Scientific and Culture Organization). The original dataset contained information about health and socio-economic indicators for the years 2000-2016 for 183 countries, with the corresponding life expectancy at birth.

The dataset has been restricted to the year 2016, columns with a percentage of missing values higher than 10% have been excluded from the analysis. Moreover, countries for which important data were missing have been excluded from the analysis. Since the dataset contained similar variables coming from different sources (e.g. infant mortality measured by WHO and by UNESCO), I decided to keep only the data from one of the sources.

The resulting dataset contains information about 159 countries. On the columns we have the response variable, `life_exp` along with 15 explanatory variables, whose meaning has been described in Table 1.

Variable	Explanation
life_expect	Life expectancy at birth (measured in years). Response variable.
adult_mortality	Adult Mortality rates for both sexes (probability of dying between 15 and 60 years per 1000 people).
age1.4mort	Death rate between ages 1 and 4.
alcohol	Alcohol, recorded per capita (15+) consumption (in litres of pure alcohol).
bmi	Mean BMI (kg/m^2) (18+) (age-standardized estimate).
age5.19thinness	Prevalence of thinness among children and adolescents, BMI < (median - 2 s.d.) (crude estimate) (in percentage).
age5.19obesity	Prevalence of obesity among children and adolescents, BMI > (median + 2 s.d.) (crude estimate) (in percentage).
hepatitis	Hepatitis B (HepB) immunization coverage among 1-year-olds (in percentage).
measles	Measles-containing-vaccine first-dose (MCV1) immunization coverage among 1-year-olds (in percentage).
polio	Polio (Pol3) immunization coverage among 1-year-olds (in percentage).
diphtheria	Diphtheria tetanus toxoid and pertussis (DTP3) immunization coverage among 1-year-olds (in percentage).

basic_water	Population using at least basic drinking-water services.
gghe.d	Domestic general government health expenditure (GGHE-D) as percentage of gross domestic product (GDP) (in percentage).
che_gdp	Current health expenditure (CHE) as percentage of gross domestic product (GDP) (in percentage).
une_pop	Population (thousands)
une_gni	Gross National Income per capita, Purchasing Power Parity (in current international $).

Table 1: Dataset's variables

3. Linear Regression

3.1 Model

The first model I built was a (simple) linear regression model. Despite its simplicity, a model of this type could be useful given that we are interested in models that could be easily interpreted and used in policymaking.

Before building this model, I checked that the linear regression assumptions were satisfied.

3.2 Linear Regression Diagnostics

Homoskedasticity

In order to check whether the assumption of homoskedasticity was satisfied, I carried out a Breusch-Pagan Test.

```
        studentized Breusch-Pagan test

data:  lin_reg_model
BP = 17.097, df = 15, p-value = 0.3131
```

Figure 1: Breusch-Pagan Test Results

Since the p-value of the test is higher than 0.05, we do not reject the null hypothesis of homoskedasticity. This can be confirmed by looking at the plot of the fitted values vs the residuals (fig.2).

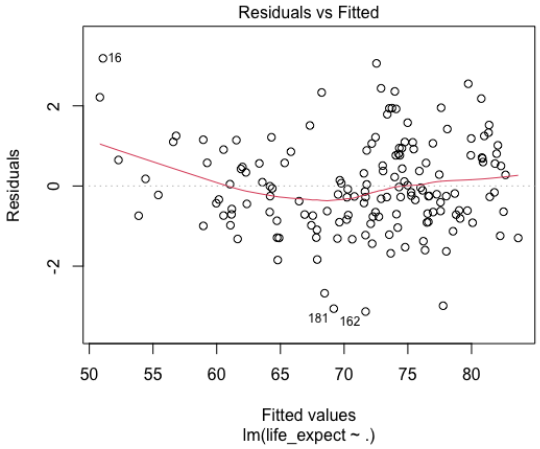

Figure 2: Residuals vs Fitted

Normality of the residuals

To test for the Normality of the residuals, I performed a Shapiro-Wilk test (fig.3).

```
        Shapiro-Wilk normality test

data:  resid(lin_reg_model)
W = 0.9856, p-value = 0.09906
```

Figure 3: Shapiro-Wilk Test Results

Since the p-value is (only slightly) above 0.05, we will not reject the null hypothesis that the residuals were sampled from a normal distribution. This could be confirmed by looking at the corresponding Q-Q plot (fig.4).

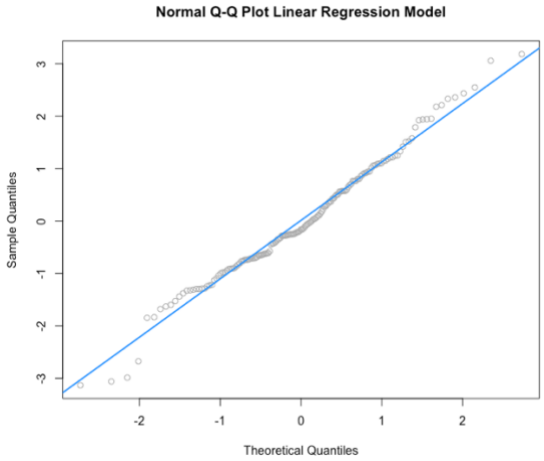

Figure 4: Normal Q-Q plot

Outliers and high-leverage points

Fig. 5 displays the distribution of each variable in the dataset.

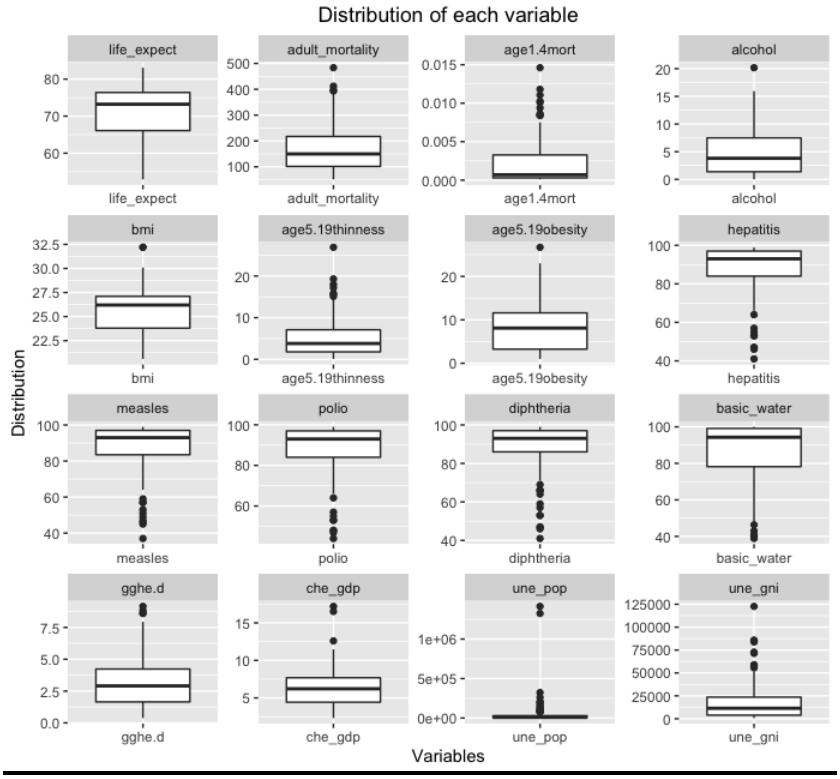

Figure 5: Variables' Distribution

In order to discover whether outliers (i.e. points with an unusual value of the residual) or high-leverage points (i.e. points with an unusual value of some predictors) exist in our dataset, I first of all inspected the Residuals vs Leverage plot (fig.6):

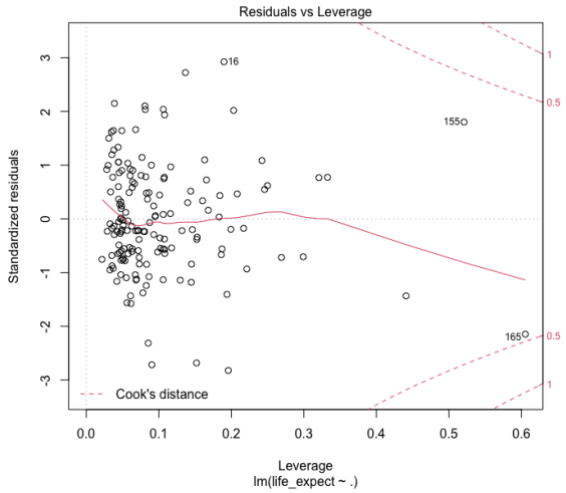

Figure 6: Residuals vs Leverage

From this plot, we can see that both some outliers and high-leverage points are present in the dataset. If the values of the outliers are somehow acceptable (the studentized residuals should ideally stay between -2 and 2), some observation exhibits very high leverage scores (see also fig.7).

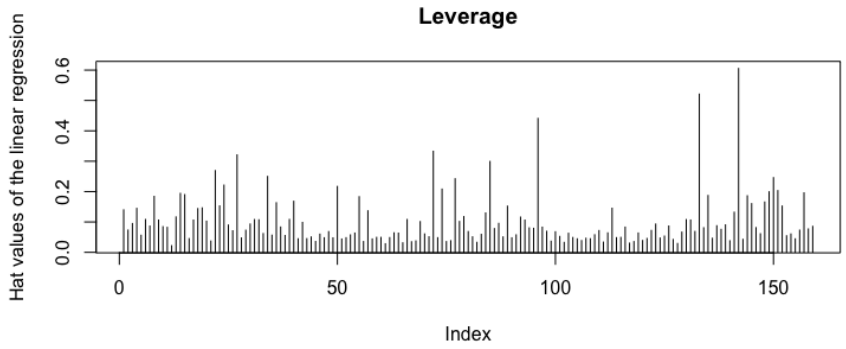

Figure 7: Leverage

According to a rule of thumb, observations should have a leverage score lower than 2p/n. In this case, however, since the problematic observations represent measurements for very important countries (such as India and China), I decided not to exclude them from the following analysis.

Collinearity

Collinearity might be a serious problem in this dataset: since some of the variables might be different measures of a similar health or economic phenomenon, two different variables might end up being highly correlated.

To understand whether a problem of correlation exists in our dataset, I first plotted the correlation matrix between the explanatory variables (fig.6).

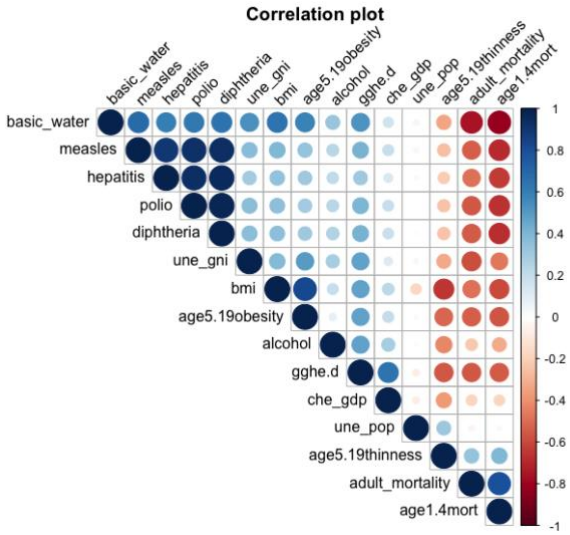

Figure 8: Correlation Plot

As we can see from the plot, some of the variables end up being highly correlated: this fact might cause instability of the regression coefficient estimate. Moreover, I computed and plotted the value of the VIF for each variable. The resulting plot is displayed in fig.9:

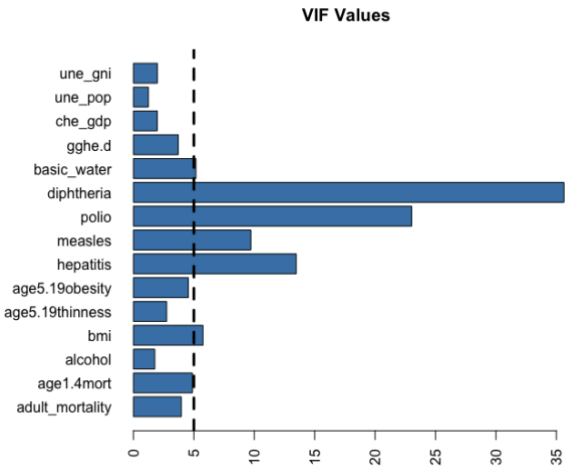

Figure 9: VIF Values

The plot confirms the presence of collinearity in the dataset, since some variables (e.g. “diphtheria”, “polio”, “measles”, “hepatitis”) have VIF > 5.

Given the strong correlation between the variables, in the following section I will built in the following section some algorithms to perform variable selection.

3.3 Results

Linear Regression result for the RMSE computed on the test set are displayed in Table 2, while coefficient estimates on the whole dataset are displayed in Figure 10.

Model	**Test RMSE**
Linear Regression	1.01

Table 2: Linear Regression Performance

```
Coefficients:
                  Estimate Std. Error t value Pr(>|t|)
(Intercept)      8.681e+01  2.563e+00  33.866  < 2e-16 ***
adult_mortality -5.320e-02  2.210e-03 -24.072  < 2e-16 ***
age1.4mort      -5.491e+02  7.491e+01  -7.330 1.56e-11 ***
alcohol          1.182e-01  3.271e-02   3.614 0.000416 ***
bmi             -4.285e-01  1.028e-01  -4.170 5.26e-05 ***
age5.19thinness -1.506e-01  3.728e-02  -4.040 8.68e-05 ***
age5.19obesity   5.841e-02  3.786e-02   1.543 0.125094
hepatitis       -3.058e-02  2.879e-02  -1.062 0.289927
measles          2.223e-02  2.278e-02   0.976 0.330737
polio            2.220e-02  3.814e-02   0.582 0.561342
diphtheria      -1.821e-02  4.757e-02  -0.383 0.702483
basic_water      5.178e-02  1.331e-02   3.890 0.000153 ***
gghe.d           4.423e-01  9.323e-02   4.744 5.02e-06 ***
che_gdp         -5.512e-02  5.393e-02  -1.022 0.308478
une_pop         -3.859e-07  6.797e-07  -0.568 0.571086
une_gni          1.409e-05  6.816e-06   2.067 0.040571 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1.21 on 143 degrees of freedom
Multiple R-squared:  0.9764,    Adjusted R-squared:  0.9739
F-statistic: 394.1 on 15 and 143 DF,  p-value: < 2.2e-16
```

Figure 10: Linear Regression Coefficients Estimate

The regressors which are significant at 99% confidence are: ‘adult_mortality’, ‘age1.4mort’, ‘alcohol’, ‘bmi’ and ‘age5.19thinness’. From the coefficient table, we can also see that -as expected- the variables that measure mortality have a negative relationship with respect to life expectancy, the same for Body-Mass Index and thinness among children. Also the positive relationship between life expectancy and basic water, health expenditure and gross national income seems intuitive. The positive relationship between life expectancy and alcohol consumption, however, seems somehow counter-intuitive.

4. Principal Component Analysis

Given the strong correlation present in our dataset, I performed a Principal Component Analysis on the variables that were more strongly correlated between each other: “diphtheria”, “polio”, “measles”, “hepatitis”.

4.1 Algorithm

PCA is a deterministic algorithm that helps summarizing a set of correlated variables with fewer “principal components”, which explain most of the variability within the original dataset.

Since all the four variables selected for PCA measure the immunization to severe illnesses, the interpretation of the PCA would in this case be facilitated: we could easily interpret the first principal component – which is the direction along which the original data vary the most - as a measure of the immunization to illnesses within each country’s population and, therefore, a proxy of the efficiency of each national health system.

In order to perform PCA, I transformed all the variables involved to have mean 0 (this is indeed done automatically by the R function “prcomp”), while I did not scale the variables to have a standard deviation of 1. This is because all the variables involved are in this case measured with the same unit, i.e. percentage of total population immune to the disease.

4.2 Proportion of variance explained

A first measure of the effectiveness of PCA can be seen by inspecting the scree plot (fig. 11&12):

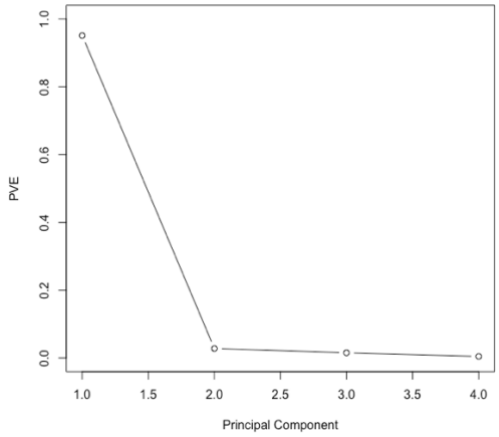

Figure 11: Scree Plot

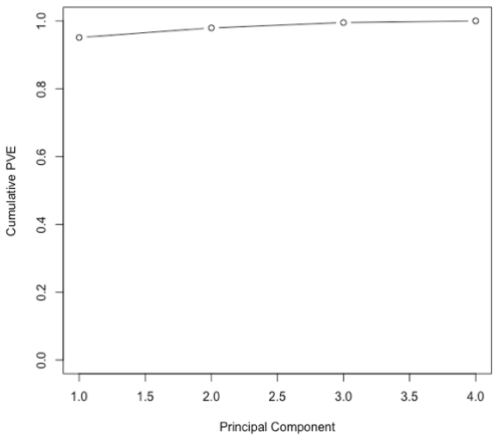

Figure 12: Scree Plot (cumulative PVE)

The variance explained by the principal component is 95% of the total variance: therefore, we can confidently say that the four original variables are well represented by the first principal component, which explains well-enough the variability of disease variables.

4.3 Biplot

Further insights from the results of PCA can be gained by inspecting the biplot (fig.13, countries' names have been substituted by numbers to improve the visualization), which displays both the principal components scores and the loading vectors.

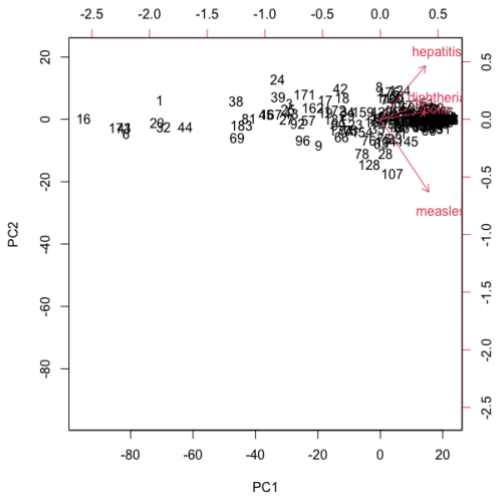

Figure 13: Biplot

From the plot, we can see that the first component places approximately equal weight on hepatitis, diphtheria, polio and measles, while the second component places a positive weight on hepatitis, a negative weight on measles, and approximately zero weight on the other two variables.

I was not able to find a satisfying interpretation of the second principal component, but I personally do not think this is a big issue, since -as we have seen in the previous paragraph- the second component explains a tiny proportion of the variance.

4.4 Linear regression after PCA

After PCA has been performed on the dataset, I ran again a Linear Regression on life expectancy considering as a regressor the first principal component of the four disease variables (named 'disease_PC'). Results of the Test RMSE and coefficient estimate on the whole dataset are displayed in Table 3 and Figure 14. Unfortunately, Test RMSE does not seem to have improved after PCA. Moreover, the coefficient estimate of the variable 'disease_PC' is not statistically significant at the 95% confidence level and it has a negative sign. This is quite counter-intuitive: as immunization to disease rises, life expectancy seems to go down.

Model	Test RMSE
Linear Regression after PCA	1.02

Table 3: Linear Regression after PCA Performance

```
Coefficients:
                  Estimate Std. Error t value Pr(>|t|)
(Intercept)      8.586e+01  2.369e+00  36.244  < 2e-16 ***
adult_mortality -5.297e-02  2.149e-03 -24.646  < 2e-16 ***
age1.4mort      -5.528e+02  7.463e+01  -7.407 9.54e-12 ***
alcohol          1.085e-01  3.213e-02   3.375 0.000944 ***
bmi             -4.174e-01  1.022e-01  -4.085 7.23e-05 ***
age5.19thinness -1.527e-01  3.715e-02  -4.111 6.54e-05 ***
age5.19obesity   4.722e-02  3.692e-02   1.279 0.202980
basic_water      5.461e-02  1.278e-02   4.272 3.48e-05 ***
gghe.d           4.625e-01  9.008e-02   5.135 8.88e-07 ***
che_gdp         -5.060e-02  5.374e-02  -0.942 0.347978
une_pop         -3.288e-07  6.750e-07  -0.487 0.626912
une_gni          1.585e-05  6.705e-06   2.364 0.019378 *
disease_PC      -3.090e-03  5.774e-03  -0.535 0.593292
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 1.209 on 146 degrees of freedom
Multiple R-squared:  0.9759,    Adjusted R-squared:  0.9739
F-statistic: 493.1 on 12 and 146 DF,  p-value: < 2.2e-16
```

Figure 14: Linear Regression after PCA Coefficients Estimate

5. Best subset selection

5.1 Algorithm

Another approach to tackle the problem of collinearity is to select the best (where "best" will be defined shortly) subset of variables contained in the full model through the "best subset selection" algorithm[1].

The algorithm can be described as follows:

1. Start from the null model;
2. For k = 1, … , p:
   a. Fit with a linear regression all the possible models containing k predictors;
   b. Among the model with k predictors, pick the best one (the one with the lowest RSS)
3. Pick the best models within the one selected in point 2.b, where "best" here means with the lowest cross-validated RMSE[2].

Running the algorithm on the training set (10 folds CV) results in choosing an algorithm with 10 variables (fig.15).

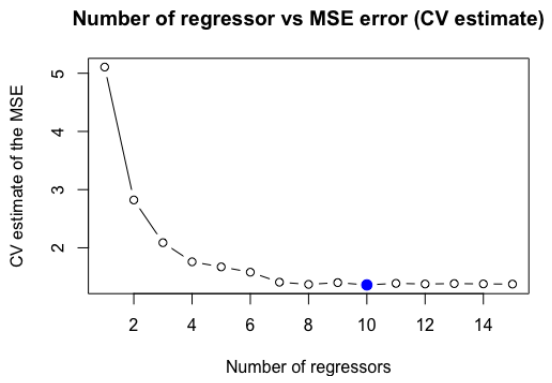

Figure 15: CV estimate for MSE vs Number of Regressors

[1] I did not take into consideration similar algorithms such as Forward or Backward Stepwise selection since, in this case, performing Best Subset Selection was not computationally infeasible, given the relatively small size of the dataset. Since Forward or Backward Stepwise selection might end up not selecting the best sub-model, Best Subset Selection has been preferred in this case.

[2] Even if other measures that account for the complexity of the model, such as AIC, BIC or Adjusted R^2 can be computed, cross-validated RMSE has been preferred in this case in order to obtained a valid comparison with the other models.

5.2 Results

Given that the model with 10 regressors has been chosen, I then ran the algorithm on the whole training set, selecting the best among the models with 10 variables, and computed the RMSE on the test set. The estimate of the test RMSE in this case is 1.05.

Model	Test RMSE
Best subset selection	1.05

Table 4: Best Subset Selection Perfomance

Running again the algorithm, this time on the whole dataset, results in this best 10-variable-model (fig. 16):

```
(Intercept) adult_mortality      age1.4mort         alcohol             bmi
8.650395e+01   -5.294540e-02   -5.689430e+02    1.080825e-01   -4.118719e-01
age5.19thinness  age5.19obesity       hepatitis     basic_water          gghe.d
-1.558309e-01    4.452410e-02   -1.090724e-02    5.545672e-02    4.049359e-01
      une_gni
1.708429e-05
```

Figure 16: Best 10-variables model

It is interesting to notice that the variables “polio”, “diphtheria”, “measles” -along with “che_gdp", whose measure is similar to the variable “gghe.d”, and “une_pop”- have not been included. Therefore, these results are in line with the exercise done in the previous paragraph.

6. Ridge, Lasso, Elastic Net

6.1 Models

Another possible approach to address the problem of multi-collinear predictors, alternative to the one explored in the previous sections, is the one of *regularizing* the coefficient estimate, i.e. shrinking them towards zero.

In this section, I explore three similar approaches for regularization: Ridge Regression, Lasso and Elastic Nets. These three approaches estimate the regression coefficients by finding the minimum of a slightly different function with respect to OLS. In particular, while OLS minimizes the RSS of the regression, Ridge Regression, Lasso and Elastic Nets minimize the following functions:

Ridge:

$$RSS + \lambda \sum_{j=1}^{p} \beta_j^2$$

Lasso:

$$RSS + \lambda \sum_{j=1}^{p} |\beta_j|$$

Elastic Net:

$$\text{RSS} + \lambda \sum_{j=1}^{p} \left(\frac{1-\alpha}{2} \beta_j^2 + \alpha |\beta_j| \right)$$

Therefore, the three approaches differ in the *type* of penalization and, in particular, on which norm of the vector of coefficients -whether a norm-1, norm-2 or a mixture of them- is computed. It is important to notice that, with λ sufficiently large, LASSO may also perform feature selection by shrinking some of the coefficients to 0.

6.2 Results

The performances of the three models are displayed in table 5, which reports estimates of the test RMSE for each model. The values for the penalization parameter λ and for the mixture parameter α – in the case of Elastic Net – have been chosen via cross-validation on the training set (results of hyperparameter tuning are displayed in table 6).

Model	Test RMSE
Ridge	1.15
Lasso	1.10
Elastic Net	1.11

Table 5: Regularization methods performance

	Penalty	**Mixture**
Ridge	0.0000000001	(0)
Lasso	0.0596	(1)
Elastic Net	0.153	0.388

Table 6: Hyperparameter Tuning Results

7. Tree-based methods

Prediction trees are simple models which can be easily interpreted and plotted. In this section, I build a regression tree considering the "reduced" dataset obtained after performing PCA and visualize the tree obtained.

Since simple regression trees are usually not competitive with respect to other supervised models, in this section I also explore ensemble methods such as Random Forest.

7.1 Simple Regression Tree

A regression tree is an algorithm for partitioning the predictor space into distinct regions, for each of which we make a prediction by simply taking the mean value of the training observation falling within the region.

To avoid overfitting, trees are usually grown and then *pruned* adding to the minimization function a penalty (cost-complexity parameter) that prevents the tree from growing too much. In what follows, I have tuned the values for the hyperparameter "cost-complexity", "three depth" and "min_n" [3] on the training set and evaluated the performance of the tree on the test set.

The results of applying a single regression tree to our dataset are displayed in table 7. Results of the hyperparameter tuning are displayed in table 8.

Model	**Test RMSE**
Simple Regression Tree	1.81

Table 7: Simple Regression Tree Perfomance

	Hyperparameter value
Cost Complexity	0.000562
Tree depth	8
Min_n	2

Table 8: Hyperparameters Tuning Results

[3] The hyperparameter "min_n" corresponds to the minimum number of data points in a node required for the node to be splitted further.

Table 7 reveals not such a good performance of this model with respect to the previous ones. A graphical representation of the tree has been displayed in fig. 17:

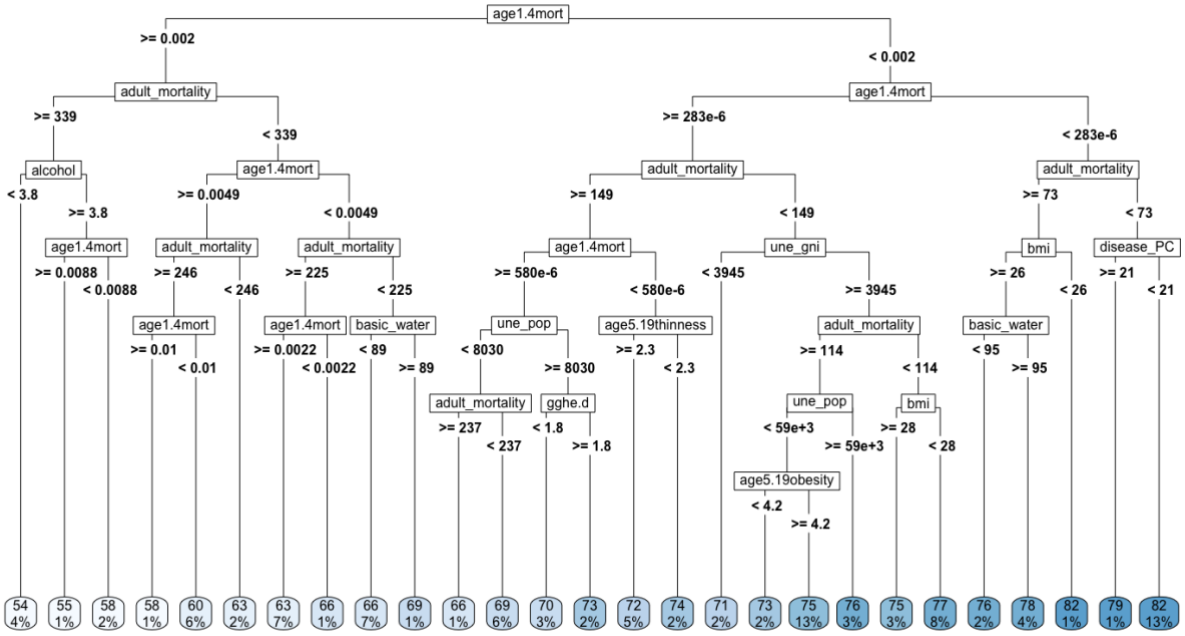

Figure 17: Tree Representation

This figure reports the predicted value for each leaf of the pruned tree and the percentage of data points contained in that leave. We can give to the picture the following interpretation: life expectancy is lower in countries where the variable "age1.4mort" is greater or equal to 0.002 and, within these countries, it is lower where "adult_mortality" is higher than 339. Following the same logic, we can interpret all the possible paths of the tree.

7.2 Random Forest

Random Forest is an algorithm that allows to average the predictions made by different trees on different datasets obtained by bootstrapping the original dataset and, at the same time, allowing the algorithm to see -for each bootstrapped sample- only a fraction of the variables (3 in our case).

Random Forest can help reducing the variance of the trees while also de-correlating the different trees that are ensembled (a good advantage with respect to Bagging).

Unfortunately, since we are building many different tree predictors, we will not be able to visualize a single tree as before, but we will only be able to assess each variable's importance by computing the frequency by which it has been chosen by the different trees.

Results of the hyperparameter tuning on the training set, test set performances and importance of each variable in the Random Forest are reported respectively in table 9 table 10 and figure 18.

	Hyperparameter value
Trees	1500
Min_n	2

Table 9: Hyperparameters Tuning Results

Model	**Test RMSE**
Random Forest	1.39

Table 10: Random Forest Performance

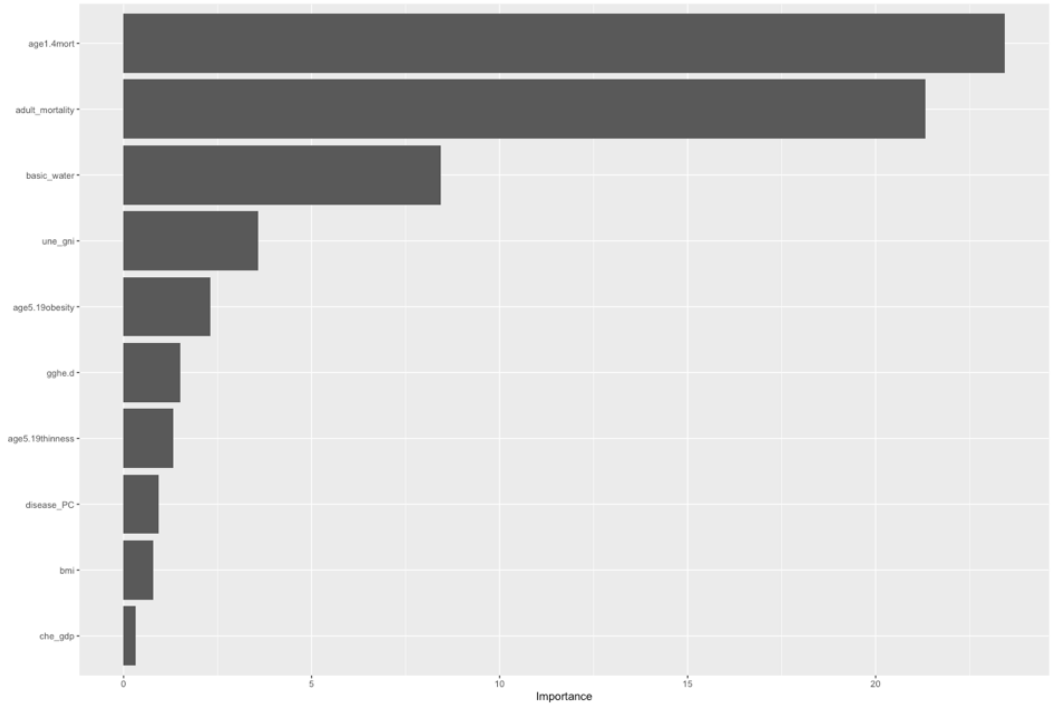

Figure 18: Random Forest's Variables Importance

We can observe that the Random Forest select as important the same variables that are in the upper roots of the Single-Tree Predictor, i.e. ‘age1.4mortality’ and ‘adult_mortality’. ‘Basic water’ appears here as the third variable for importance. The test error has slightly improved with respect to Single Tree Predictor, but it still is higher than the previous models.

8. Supervised models: performance comparison and conclusions

Model	Test RMSE
Linear Regression	1.01
Linear Regression after PCA	1.02
Best subset selection	1.05
Ridge	1.15
Lasso	1.10
Elastic Net	1.11
Simple Regression Tree	1.81
Random Forest	1.39

Table 11: Models Performance Comparison

Table 11 summarizes the performance of all the models. From this comparison, we can see the somehow surprisingly good result of the simple Linear Regression Model, which outperforms all the other. This is not completely surprising since, as we have seen from the diagnostic in Section 3, the dataset exhibits a good amount of linearity.

Since the performance of the model after having performed PCA is similar, the regression coefficients estimated by the "Linear Regression after PCA" model are probably more reliable, since they reduce the amount of collinearity among the features.

9. K-means clustering

To complete the project, I applied to the features of the dataset an unsupervised algorithm – K-means clustering- in order to discover whether it was possible to cluster countries in two different groups (having in mind the well-known distinction between "developed" and "developing" countries) on the basis of the health and socio-economic indicators present in the dataset[4].

9.1 Algorithm

K-means clustering is a clustering technique which is very useful when the number of clusters we want to obtain is known in advance. The idea behind K-means is partitioning the observation in clusters such that the total within-cluster variation (usually defined in terms of squared Euclidean distance) is minimized.

The local minima of the objective function can be found applying iteratively the following algorithm:

1. Assign each observation randomly to one of the K clusters;
2. Repeat until no observation changes its cluster:
    a. Compute the *centroid* of each cluster, i.e.: the vector containing the mean of the features for the observations in each cluster;
    b. Re-assign each observation to the cluster whose centroid is closest (usually in terms of Euclidean distance).

Since the algorithm only finds local optimum of the objective, the result of the procedure will depend on the initial assignment of the observations. Therefore, it is important to repeat the initial assignment procedure more than one time (in this case, it has been repeated 50 times) and to select the result such that the objective is smallest.

9.2 Results

The results of the K-means clustering have been displayed in fig. 19:

[4] I excluded from the original dataset the variable "une_pop" which is the country's population. This choice was motivated by the fact that I wanted to cluster the countries only on the basis of health and socio-economic factors.

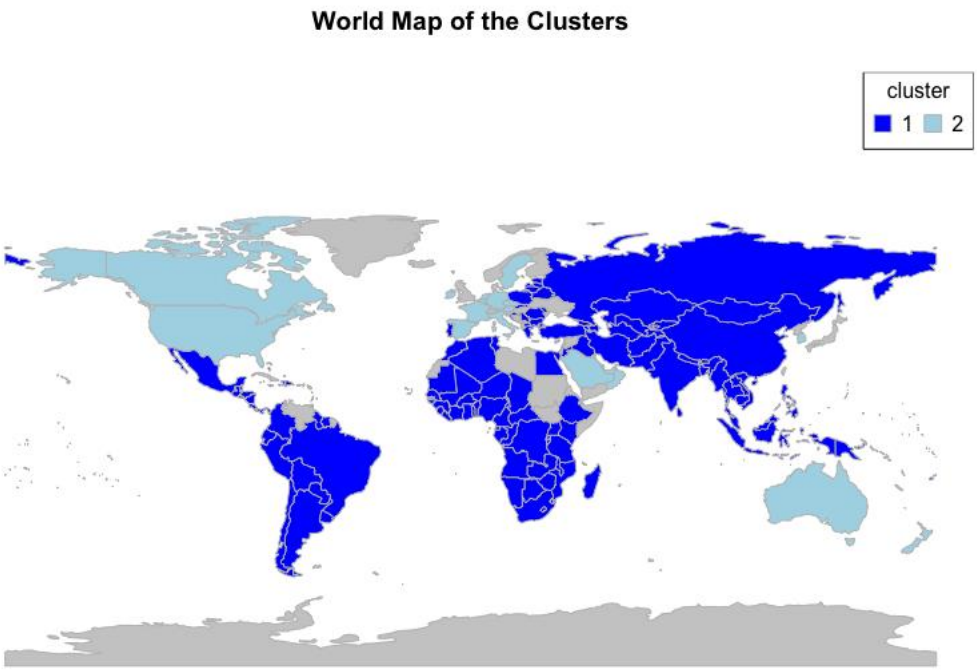

Figure 19: World Map of the clusters computed by K-means

As a comparison, I report in fig. 20 a clustering performed by the International Monetary Fund and the United Nations:

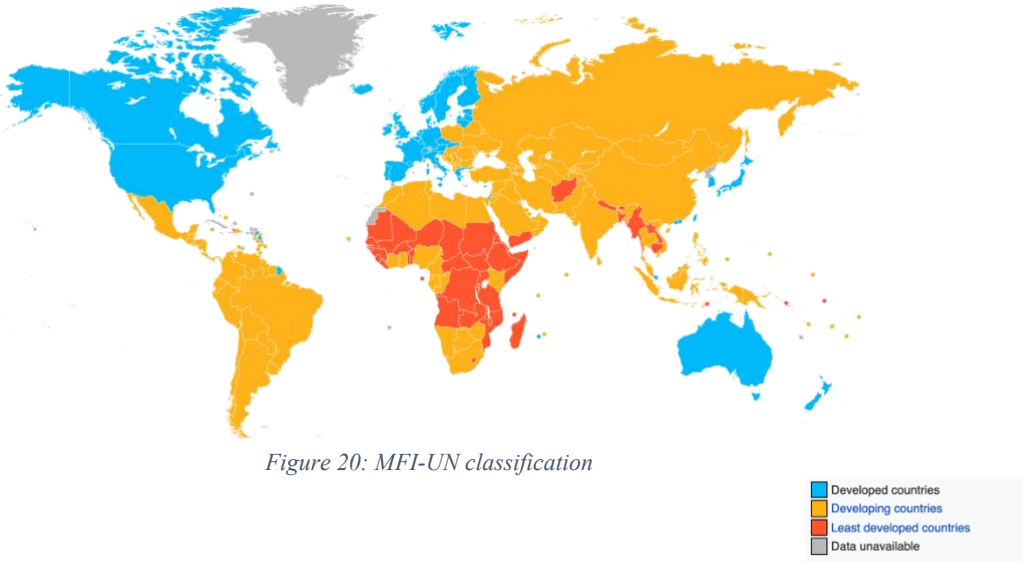

Figure 20: MFI-UN classification

As we can see, the two graphs are comparable, with some differences.

To have further insights into how each cluster behaves, fig. 21 reports the distribution of each variable within each cluster:

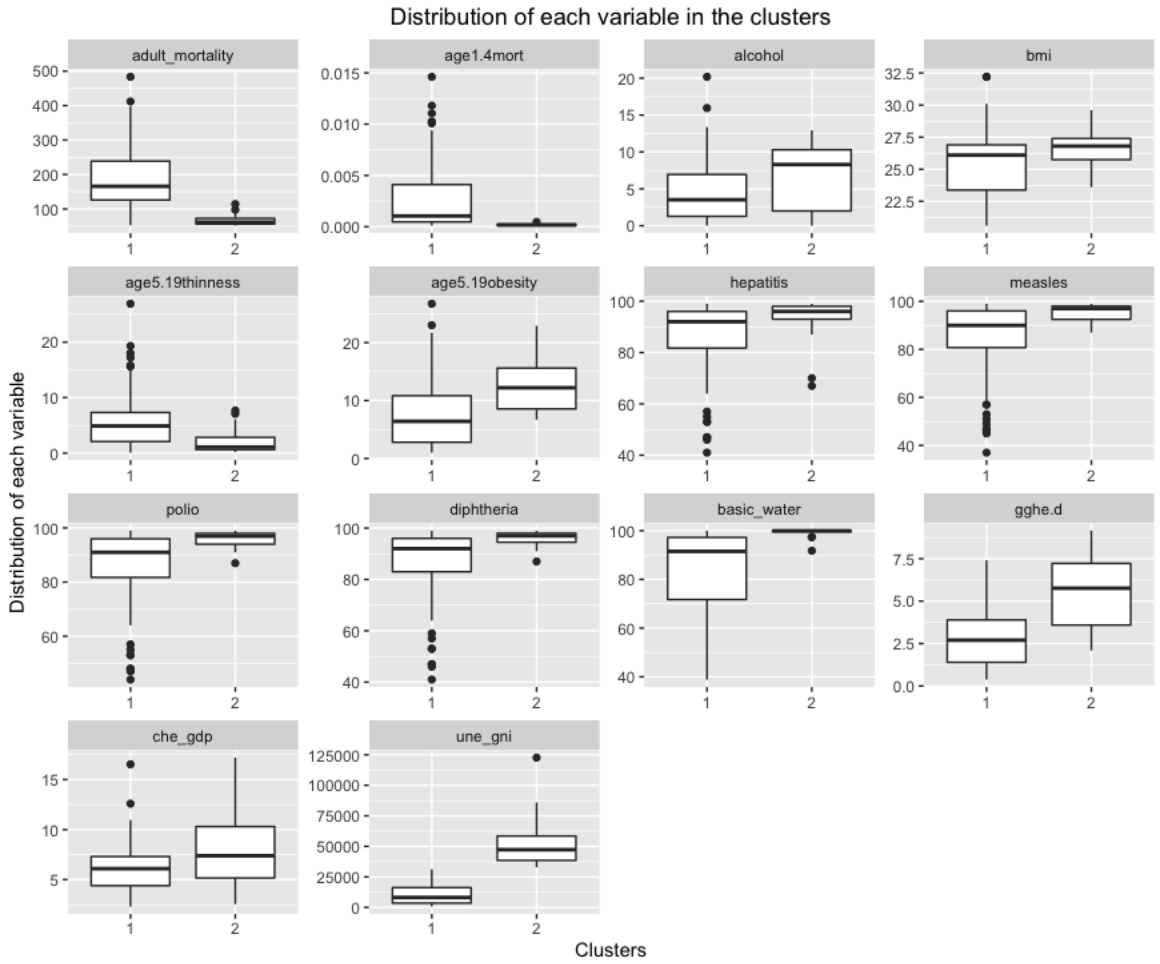

Figure 21: Distribution of each variable in the clusters

From this figure, we can see that the countries clustered in the first group have -with some heterogeneity- characteristics such as high adult and infant mortality, lower immunization to infectious diseases, lower access to drinking water, relatively lower gross domestic product and health expenses with respect to countries in cluster 2.

As I expected by looking at this plot, the countries belonging to cluster 1 belong at the same time to the group classified as “Developing Countries” by the United Nation, while cluster 2 coincides (with some differences) to the “Developed Countries” group.