Politecnico di Torino

Master's Degree in Computer Engineering

Statistical Analysis on factors influencing Life Expectancy

Course:
Data Spaces
Professor:
Francesco Vaccarino

Students:
Alessandro Bozzella 254764
Rinaldo Clemente 259536

Academic Year 2019 - 2020

Contents

Chapter 1

Introduction

Machine Learning is a cornerstone when it comes to artificial intelligence and big data analysis. It provides powerful algorithms that are capable of recognizing patterns, classifying data, and basically learn by themselves to perform a specific task. This field has incredibly grown in popularity these days, however it still remains unknown for the majority of people and even for most professionals.
The main objective of this work is to apply the notions learned during the course to a dataset using some machine learning algorithms in order to produce a useful result from a scientific point of view: in particular the goal is to find a set of features that affect the feature "Life expectancy", representing our target variable. First of all we started from the data cleaning by detecting and removing null-values and treating outliers; then we moved to the data exploration to understand the relationships among the various features of the dataset. Later we tried the feature extraction to see if it is possible to obtain a subset of features that is the most representative with respect to the initial one and finally we began the data analysis to predict the feature "Life expectancy" applying different classification algorithms.

1.1 Dataset Description

The name of the dataset is Life Expectancy (WHO), accessible on the Kaggle website and related to life expectancy and health factors for 193 countries. It has been collected from the same WHO data repository website and its corresponding economic data was collected from United Nation website; moreover among all categories of health-related factors only those critical were chosen that are more representative. The list of features available in the dataset are the following:

- **Country (nominal)** - the country in which the indicators are from
- **Year (ordinal)** - the calendar year the indicators are from (ranging from 2000 to 2015)
- **Status (nominal)** - whether a country is considered to be 'Developing' or 'Developed' by WHO standards
- **Life Expectancy (ratio)** - the life expectancy of people in years for a particular country and year
- **Adult Mortality (ratio)** - the adult mortality rate per 1000 population (i.e. number of people dying between 15 and 60 years per 1000 population)
- **Infant Deaths (ratio)** - number of infant deaths per a population of 1000 people; similar to above but for infants
- **Alcohol (ratio)** - a country's alcohol consumption rate measured as liters of pure alcohol consumption per capita
- **Percentage Expenditure (ratio)** - expenditure on health as a percentage of Gross Domestic Product per capita
- **Hepatitis B (ratio)** - number of 1 year olds with Hepatitis B immunization over all 1 year olds in population
- **Measles (ratio)** - number of reported measles cases per a population of 1000 people
- **BMI (interval/ordinal)** - average Body Mass Index of a country's total population
- **Under-Five Deaths (ratio)** - number of deaths of people under the age of five per a population of 1000 people
- **Polio (ratio)** - number of 1 year olds with Polio immunization over the number of all 1 year olds in population
- **Total Expenditure (ratio)** - government expenditure on health as a percentage of total government expenditure
- **Diphtheria (ratio)** - Diphtheria Tetanus Toxoid and Pertussis (DTP3) immunization rate of 1 year olds

- **HIV/AIDS (ratio)** - deaths per 1000 live births caused by HIV/AIDS for people under 5; number of people under 5 who die due to HIV/AIDS per 1000 births
- **GDP (ratio)** - Gross Domestic Product per capita
- **Population (ratio)** - population of a country
- **Thinness 1-19 Years (ratio)** - rate of thinness among people aged 10-19
- **Thinness 5-9 Years (ratio)** - rate of thinness among people aged 5-9
- **Income Composition Of Resources (ratio)** - Human Development Index in terms of income composition of resources (index ranging from 0 to 1)
- **Schooling (ratio)** - average number of years of schooling of a population

1.2 Used Tools

For this assignment several tools has been used:

- **LaTeX** - a document preparation system, used for writing this report
- **Python** - an interpreted, high-level, general-purpose programming language widely used in machine learning. In particular, the libraries that have been used are the following:
  - **Pandas** - a library for manipulating data in sequential or tabular format, mainly used for loading standard formats for tabular data, for the simplicity of aggregation of data and execution and display of numerical and statistical operations
  - **NumPy** - a library adding support for large, multi-dimensional arrays and matrices along with a large collection of high-level mathematical functions to operate on these arrays
  - **Matplotlib** - a plotting library for the Python programming language and its numerical mathematics extension NumPy
  - **Seaborn** - a data visualization library based on matplotlib for Python
  - **Graphviz** - a library providing a simple interface for the Graphviz graph-drawing software used for drawing graphs
  - **SciPy** - a library used for scientific and technical computing

  - **Scikit-learn** - a machine learning library, including various classification, regression and clustering algorithms

- **Kaggle** - an online community of data scientists and machine learning practitioners: allows users to find and publish datasets, explore and build models in a web-based data-science environment, work with other data scientists and machine learning engineers and enter competitions to solve data science challenges

- **Notebook** - a type of notebook available on Kaggle: actually it is a Jupyter notebook that provides a cloud computational environment that enables reproducible and collaborative analysis

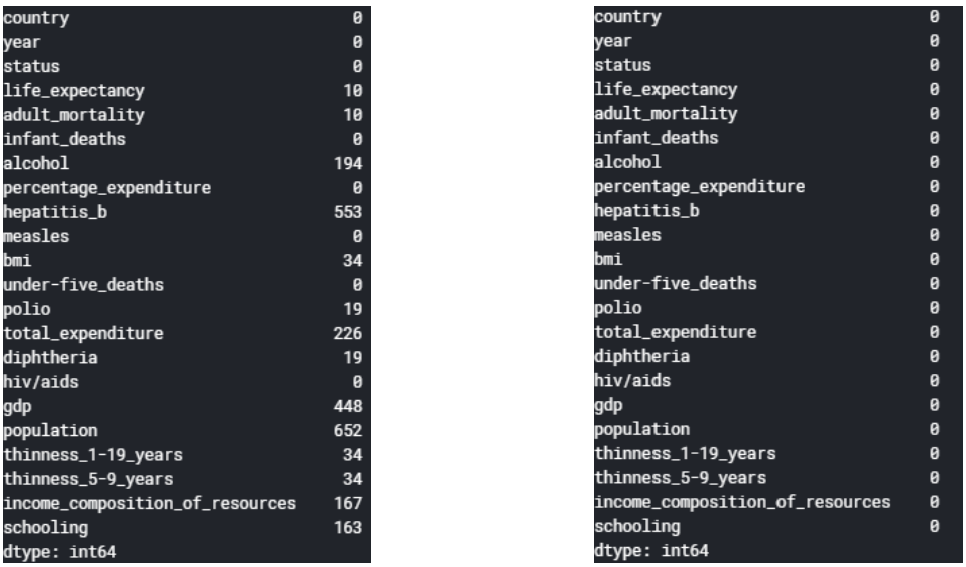
Chapter 2

Data Preprocessing

Data preprocessing is an important step in the data mining process: it is the process of detecting and correcting (or removing) corrupt or inaccurate records from a dataset, or/and focus on identifying incorrect, incomplete, irrelevant parts of the data and then modifying, replacing, or deleting the dirty or coarse data.

2.1 Data Cleaning

The first step was the renaming of the features to uniform them to a standard format. After that the data were cleaned: this included detecting and dealing with missing values. Indeed the number of rows with NULL values have been counted and then deleted.



country	0
year	0
status	0
life_expectancy	10
adult_mortality	10
infant_deaths	0
alcohol	194
percentage_expenditure	0
hepatitis_b	553
measles	0
bmi	34
under-five_deaths	0
polio	19
total_expenditure	226
diphtheria	19
hiv/aids	0
gdp	448
population	652
thinness_1-19_years	34
thinness_5-9_years	34
income_composition_of_resources	167
schooling	163
dtype: int64	

country	0
year	0
status	0
life_expectancy	0
adult_mortality	0
infant_deaths	0
alcohol	0
percentage_expenditure	0
hepatitis_b	0
measles	0
bmi	0
under-five_deaths	0
polio	0
total_expenditure	0
diphtheria	0
hiv/aids	0
gdp	0
population	0
thinness_1-19_years	0
thinness_5-9_years	0
income_composition_of_resources	0
schooling	0
dtype: int64	

Figure 2.1: Sum Of NULL Values Before (Left) And After (Right) The Cancellation

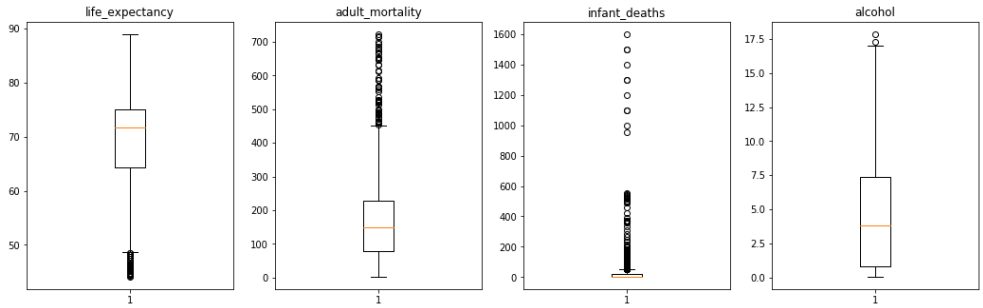
In addition to that the subsequent step was removing the features 'country' and 'year': the reason behind this choice is related to the fact that they are metadata, that is data that provide information about other data and they are not useful to carry on any type of analysis. Once this operation was complete, exploration of the data has been conducted. The last step was the replacing of 'status' column with 0 or 1 respectively for 'Developed' and 'Developing' values, in order to use it as independent variable.

2.2 Data Exploration

In this phase the visualization of the data has been conducted to understand what is in the dataset and the characteristics of the data: these characteristics can include size or amount, completeness, correctness of the data and possible relationships among the several features. All of these activities are aimed at creating a mental model and understanding of the data in the mind of the analyst.

2.2.1 Box Plots

It's a method coming from the descriptive statistics for graphically depicting the distribution of data based on a five number summary ("minimum", first quartile (Q1/25th percentile), median (Q2/50th percentile), third quartile (Q3/75th Percentile) and "maximum"), where: median is the middle value of the dataset, first quartile is the middle number between the smallest number (not the "minimum") and the median of the dataset, third quartile is the middle value between the median and the highest value (not the "maximum") of the dataset, interquartile range (IQR) is 25th to the 75th percentile, maximum is (Q3 + 1.5 $\times$ IQR) and minimum is (Q1 - 1.5 $\times$ IQR). Box plots may also have lines extending from the boxes indicating variability outside the upper and lower quartiles; outliers may be plotted as individual points. The spacings between the different parts of the box indicate the degree of dispersion (spread) and skewness in the data and show outliers.



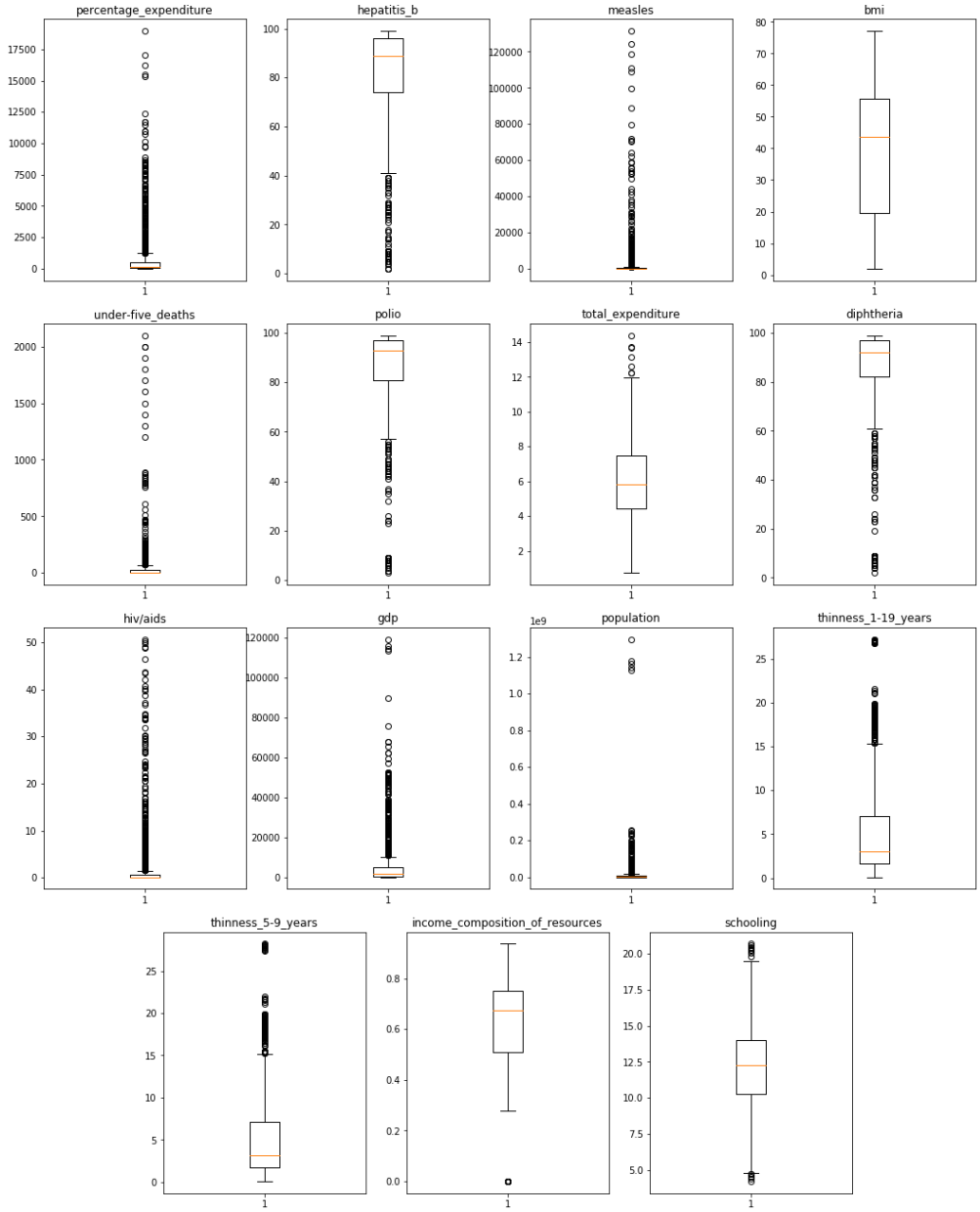


Figure 2.2: Box Plots Before Outliers Removal

Visually, it is plain to see that there are a lot of outliers for all of these numerical features, including the renamed target variable 'life_expectancy': for this reason it has been decided to remove them using the interquartile range (IQR) previously defined. To be more precise, the interquartile range for the data has been calculated, multiplied it by 1.5 (a constant used to discern outliers) and added $(1.5 \times IQR)$ to the third quartile: any number greater than this was a suspected outlier. Equally $(1.5 \times IQR)$ has been subtracted from the first quartile and any number less than this was a suspected outlier.

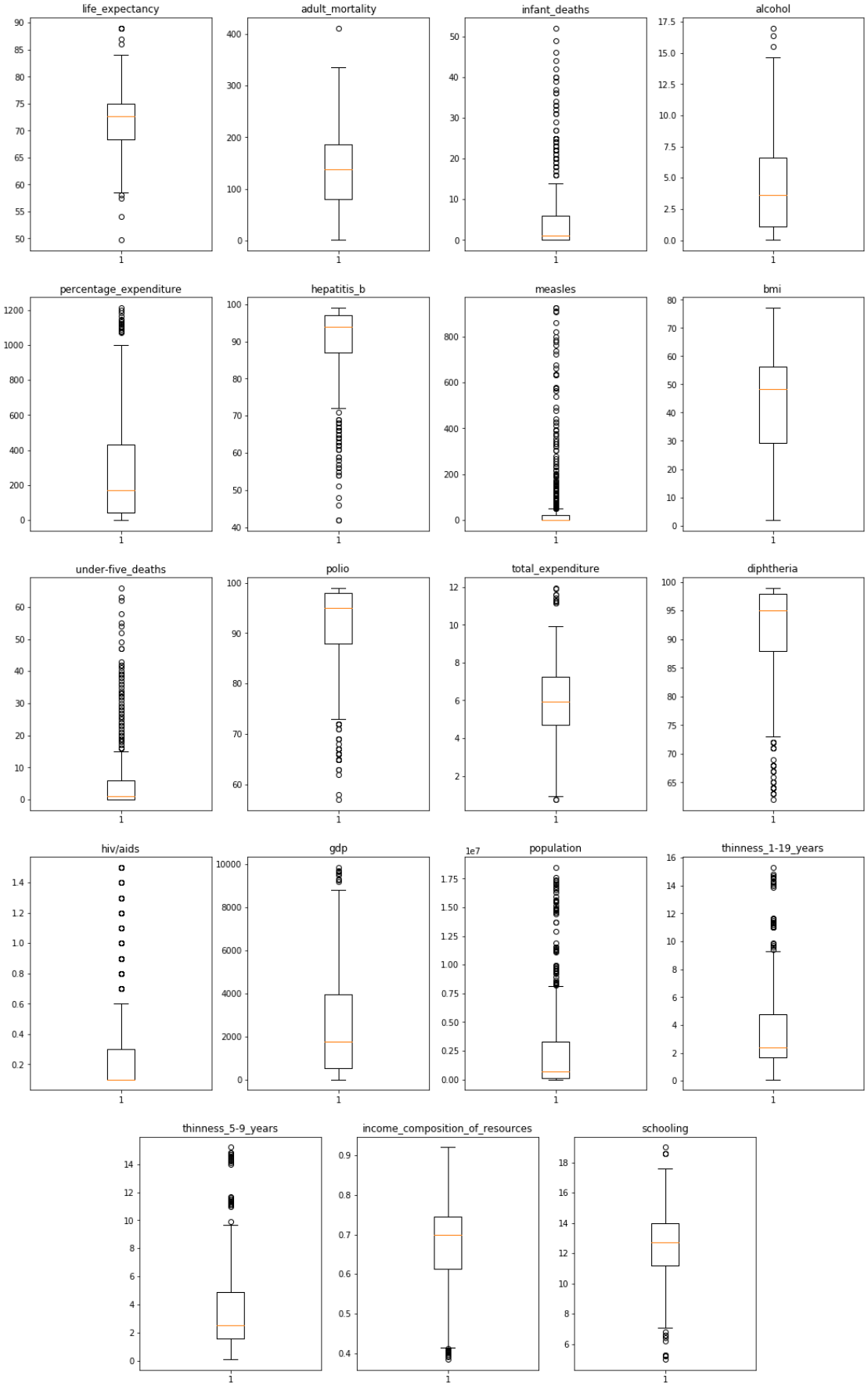


Figure 2.3: Box Plots After Outliers Removal

Although a certain number of outliers has been removed, some of them are still present. This is for the fact that we are analysing the context of different countries around the world and it is highly probable that there aren't so many common factors among themselves: for example, a disease could be totally absent in a country and greatly spread in another one. Accordingly they are not exactly outliers.

2.2.2 Heatmap

An heatmap is a graphical representation of data that uses a system of color-coding to represent different values. The primary purpose is to better visualize the volume of locations/events within a dataset and assist in directing viewers towards areas on data visualizations that matter most. It is innately self-explanatory: the darker the shade, the greater the quantity (the higher the value, the tighter the dispersion, etc.). Said that, the inclusion of it within the elaborate is for showing correlations among the features.

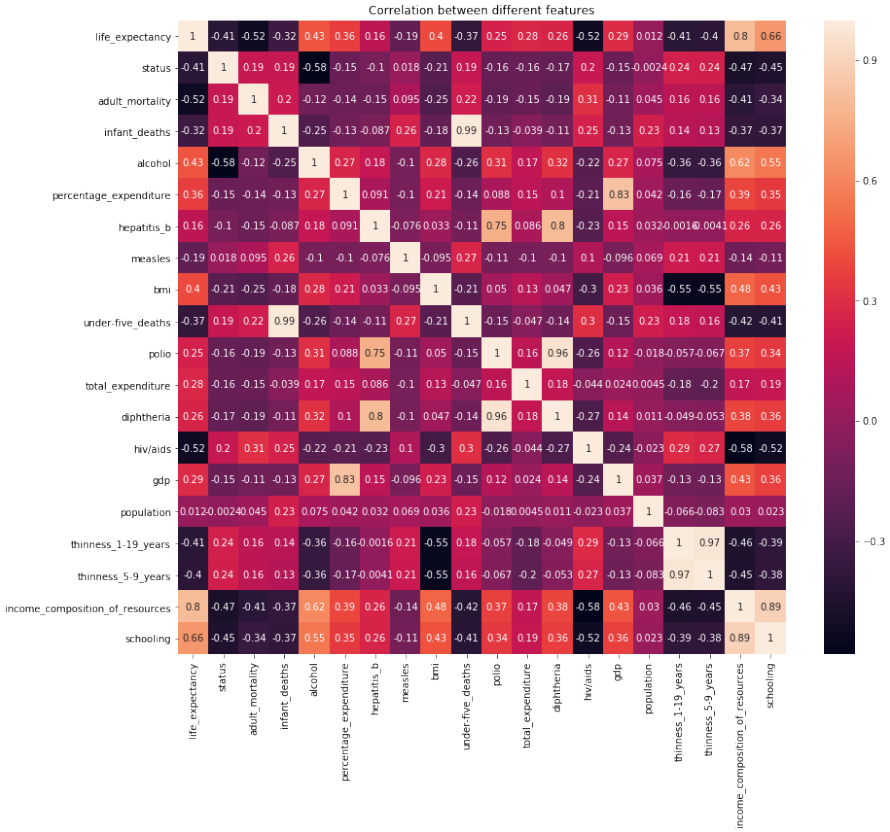


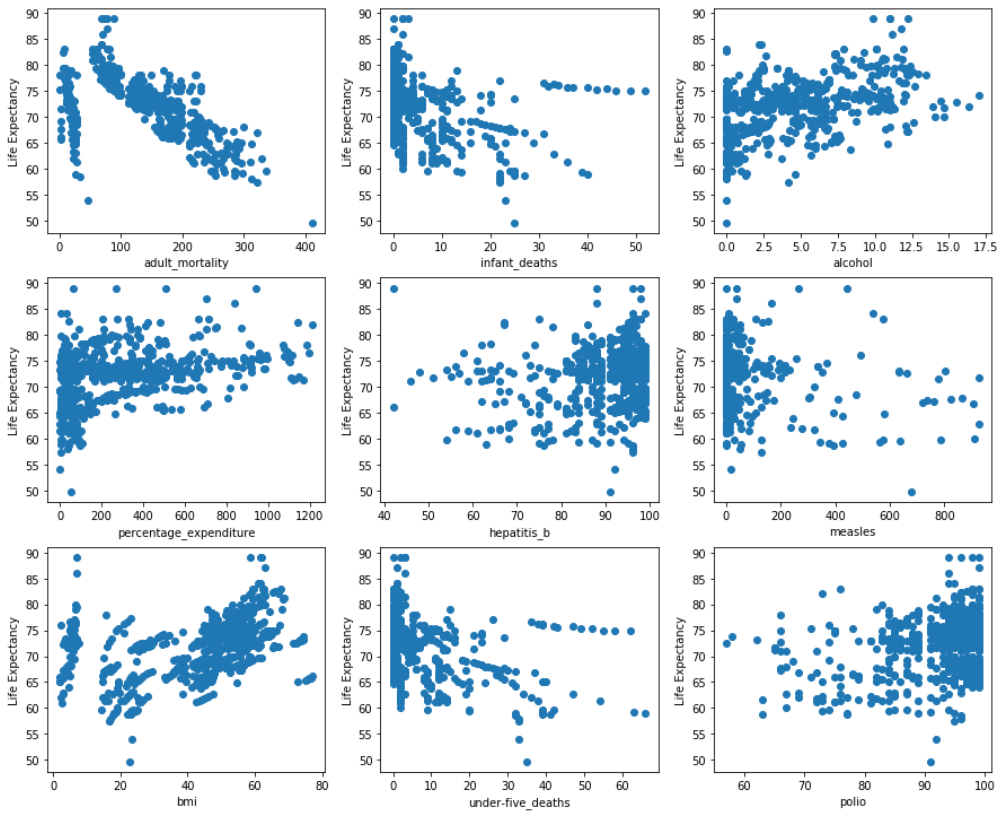
Figure 2.4: Heatmap

As it can be seen above, the correlation matrix has been plotted visualizing it with an heatmap: the legend tells that the lighter colors show high and positive

correlation, while the darker ones high and negative; otherwise the colors closest to the red show low correlation.

2.2.3 Scatter Plots

A scatter plot is a type of plot or mathematical diagram using cartesian coordinates to display values for typically two variables for a set of data. The data are displayed as a collection of points, each having the value of one variable determining the position on the horizontal axis and the value of the other variable determining the position on the vertical axis. A scatter plot can be used either when one continuous variable that is under the control of the experimenter and the other depends on it or when both continuous variables are independent. If exists a parameter that is systematically incremented and/or decremented by the other, it is called the control parameter or independent variable and is customarily plotted along the horizontal axis. The measured or dependent variable is customarily plotted along the vertical axis; if no dependent variable exists, either type of variable can be plotted on either axis and a scatter plot will illustrate only the degree of correlation (not causation) between two variables.



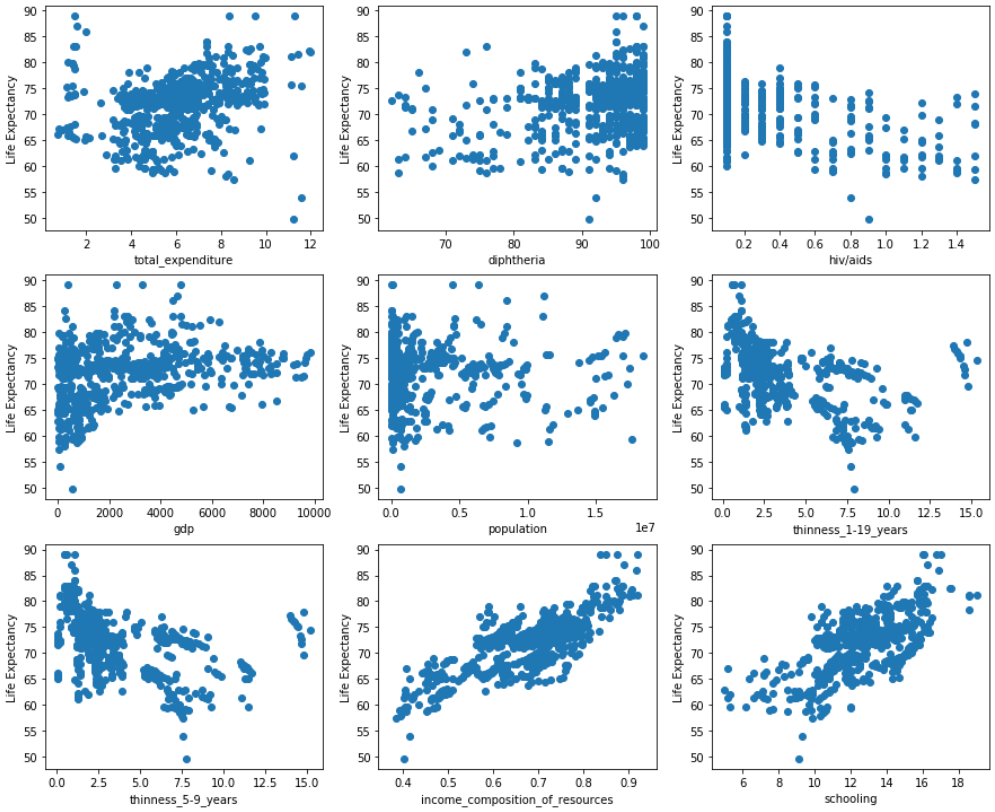


Figure 2.5: Scatter Plots

In the scatter plots the feature 'life_expectancy' was plotted against some other features to see if there is any correlation between them. There seem to be a positive correlation with 'income_composition_of_resources', 'alcohol', 'total_expenditure', 'percentage_expenditure', 'schooling', 'gdp' and 'bmi', while there is a negative one with 'adult_mortality', 'thinness_1-19_years', 'thinness_5-9_years' and 'hiv/aids'; there does not seem to have any correlation with 'hepatitis_b', 'measles', 'infant_deaths', 'polio', 'under-five_deaths', 'diphtheria' and 'population'.

2.3 Feature Extraction

Feature extraction is a process of dimensionality reduction by which an initial dataset of raw data is reduced to a more manageable group for processing. It can be applied to large datasets where there is a large number of features that require a lot of computing resources to process. Feature extraction is the name for methods that select and/or combine features, effectively reducing the amount of data that must be processed, while still accurately and completely describing the original dataset. The process of feature extraction is useful when it is necessary to reduce the number of resources needed for processing without losing important or

relevant information. Feature extraction can also reduce the amount of redundant data for a given analysis. Also, the reduction of the data and the machine's efforts in building variable combinations (features) facilitate the speed of learning and generalization steps in the machine learning process.

2.3.1 Principal Component Analysis

Principal Component Analysis (PCA) is a linear dimensionality reduction procedure and an unsupervised learning technique that can be utilized for extracting information from a high-dimensional space by projecting it into a lower-dimensional sub-space. It tries to preserve the essential parts that have more variation of the data and remove the non-essential parts with fewer variation. It is a statistical procedure that uses an orthogonal transformation to convert a set of observations of possibly correlated features (entities each of which takes on various numerical values) into a set of values of linearly uncorrelated features called principal components.
A preliminary point has been done by normalizing the data since PCA's performance is influenced on the scale of the features of the data. During the normalization step each feature of data is normally distributed such that it will scale the distribution to a mean of zero and a standard deviation of one.
The subsequent step is to perform the eigendecomposition of the covariance matrix $\sum$, that is a matrix where each element represents the covariance between two features. The covariance between two features is calculated as follows:

$$\sigma_{jk} = \frac{1}{n-1} \sum_{i=1}^{N}((x_{ij} - \bar{x}_j)(x_{ik} - \bar{x}_k))$$

The eigendecomposition of the covariance matrix has been performed to find eigenvalues and eigenvectors. The eigenvectors and eigenvalues of a covariance (or correlation) matrix represent the "core" of PCA: the eigenvectors (principal components) determine the directions of the new feature space and the eigenvalues determine their magnitude. In other words, the eigenvalues explain the variance of the data along the new feature axes. In order to decide which eigenvectors can be dropped without losing too much information for the construction of lower dimensional subspace, the corresponding eigenvalues have been inspected: the eigenvectors with the lowest eigenvalues bear the least information about the distribution of the data; those are the ones can be dropped. In order to do so, the common approach is to rank the variance correlated to eigenvalues from highest to lowest and choose the top variances.

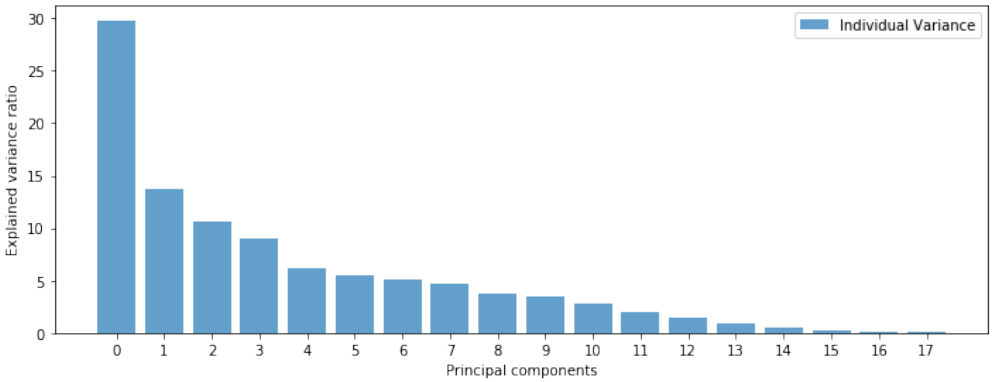


Figure 2.6: Principal Components Variances

The plot above clearly shows that maximum variance (around 30%) can be explained only by the first component. The second and third also are important (around 13% and 10%); the fourth component gives less information (8%) than the first three but more than the other ones. Fifth to tenth components share almost equal amount of information as compared to the rest but it cannot be ignored that they contributed almost with about 26% of the data. Finally the components from eleventh to follow could have been discarded as they contain less than 10% of the variance.

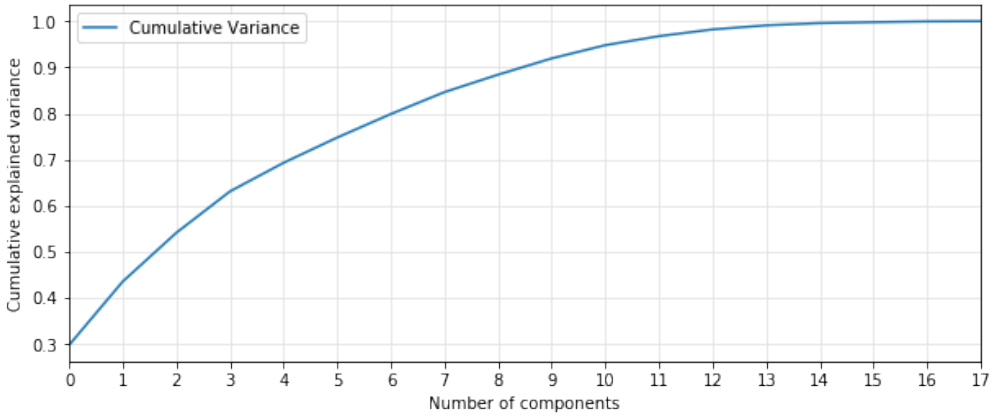


Figure 2.7: Cumulative Components Variances

The last plot shows almost 90% of variance by the first ten components. Those features are as follows:

1. adult_mortality
2. infant_deaths
3. alcohol

4. percentage_expenditure
5. hepatitis_b
6. measles
7. bmi
8. under-five_deaths
9. polio
10. total_expenditure

At the end of this subsection it has been decided to continue the analysis of the dataset on the entire set of features without removing any of them. The motivation related to this choice is correlated to the variance in the principal components. Specifically almost the 90% of variance is reached by the first ten components on a total of eighteen: this implies that the adopted reduction technique doesn't reach an optimal trade-off between the number of principal components obtained and the associated total variance. One ideal case for what concerns the number of principal components could have been less than five with almost the 90% of variance: indeed this could have conducted to a significant reduction in the complexity of the dataset structure avoiding at the same time a considerable loss of information.

Chapter 3

Data Analysis

Data analysis is the process of inspecting, cleaning, transforming and modeling data with the goal of discovering useful information and supporting decision-making. In today's business world, data analysis plays a role in making decisions more scientific and helping businesses to operate more effectively. With the purpose of predicting the target variable, classification algorithms have been applied for establishing the one that best fits in terms of performance to the dataset under consideration. Before applying each of these algorithms, the cleaned dataset has been properly normalized and divided in two subsets according to fixed proportions: the training set used for building and training the model and the test set for assessing the model performance.

3.1 Linear Regression

Linear regression is a supervised learning and linear approach used for modeling the relationship between a scalar response (or dependent variable) and one or more explanatory variables (or independent variables). The case of one explanatory variable is called simple linear regression. Simple linear regression models are used to show or predict the relationship between two variables. The variable that is being predicted (the one that solves the equation) is called the dependent variable; the one used to predict the value of the dependent variable is called the independent variable: in simpler terms, it is the 'line that best fits to the points'. It has been considered the following model function:

$$Y = \beta_0 + \beta_1 X$$

which describes a line with regression intercept β_0 and regression slope β_1.

In general this relationship could not be applied in the right way for a large number of unobserved values: these unobserved deviations can be quantified introducing the errors and the equation above, supposing the observation of n points, can be rewritten as:

$$Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i$$

where Y_i is the Y value for the i-th observation, β_0 is the regression intercept, β_1 is the regression slope, X_i is the value of X for the i-th observation and ε_i is the random deviation or random error term in the model.
The intent is to find estimated values $\hat{\beta}_0$ and $\hat{\beta}_1$ for the parameters β_0 and β_1 which would provide the "best" fit in some sense for the data points; this purpose is realized mathematically through the principle of the least squares that states that the top choice of this linear relationship is the one that minimizes the square in the vertical distance from the y values in the data and the y values on the regression line.
For sure, the best parameter which minimizes the random error term (ε_i), also according the heatmap (Figure 2.4) with the most correlated feature with the target variable ('life_expectancy'), is the feature 'income_composition_of_resources', that has been selected to perform the linear regression. In order to do that, just these two features of the dataset have been taken and the other ones have been discarded. Then the dataset is splitted into train (70%) and test (30%) set.

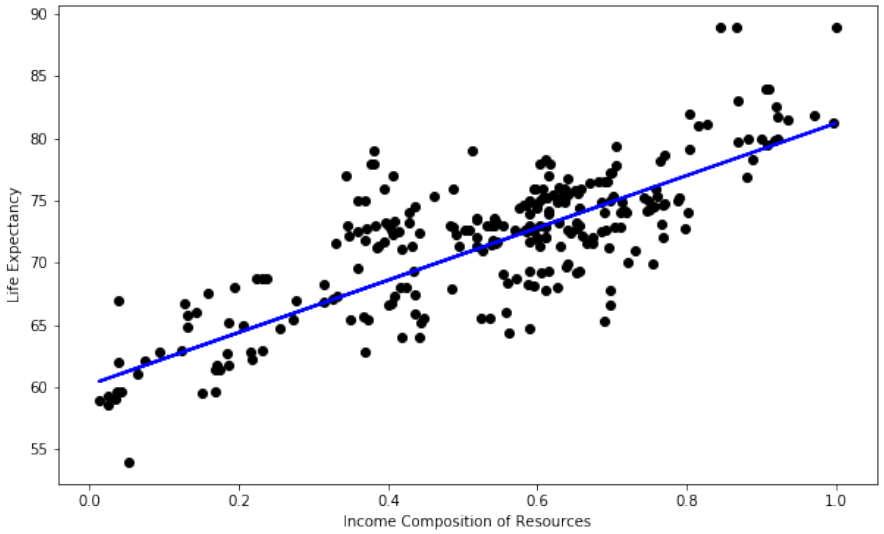


Figure 3.1: Linear Regression

From the scatter plot there is an evidence of aggregated data and a high-significance linear regression. As a matter of fact a straight line can be seen in the plot, show-

ing how linear regression attempts to draw a straight line that will minimize the residual sum of squares between the observed responses in the dataset and the responses predicted by the linear approximation.
Turning the explanation into more technical things, to evaluate the accuracy of the linear model the following metrics have been calculated: the coefficients of the line, the mean squared error and the coefficient of determination (R^2).

Coefficient: [20.87836766]
Mean squared error: 11.54
R^2 score: 0.57

The first one indicates the estimated coefficients for the linear regression problem; the second one measures the average of the squares of the errors (that is, the average squared difference between the estimated values and the actual value). The last one is a statistical measure of how close the data are to the fitted regression line.
In summary the variation of the target variable can be explained with a great grade of approximation by the predictor variable: not by chance R^2 suggests that the linear regression model explained much of the variability in the outcome.

3.2 Multiple Linear Regression

Multiple Linear Regression, as the Single one, is a supervised learning regression based on this following math model:

$$\hat{y} = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + ... + +\beta_n x_n$$

Where:

- $\hat{y}$ is the result of the model;
- β_0 is the intercept, it means the $\hat{y}$ value when all x_i are equals to 0;
- β_1 is the coefficient of x_1 (first feature);
- β_n is the coefficient of x_n (n-th feature);
- $x_1, x_2, ..., x_n$ are the independent variables of the model.

The equation shows the relationship between one dependent variable ($\hat{y}$) and two or more independent variables ($x_1, x_2, ..., x_n$). β_n are real numbers and they are called regression coefficients estimated by the model.

The object is to determine the best hyperplane covered by the data. To do that, is needed to estimate β_n values that best predict target values. This purpose is realized mathematically through the principle of the least squares that states that the top choice of this linear relationship is the one that minimizes the square in the vertical distance from the y values in the data and the y values on the regression line.
Multiple Linear Regression analysis is an algorithm uses to identify the effect of independent variables on the dependent one. The objective is to understand how much dependent variable changes when independent variables change. This algorithm could permits to predict effects or impacts of changes in real situations.
The evaluation of Multiple Linear Regression accuracy have been calculated like Linear Regression: the coefficients of the lines, the mean squared error and the coefficient of determination (R^2).

Coefficients: [-1.10123467 -6.05913045 12.56178554 -1.69728779 5.24253066
0.5545919 -2.24012194 -0.09972823 -13.49562558 -1.50760409 3.90769657
-0.67585599 -1.49658085 -5.38186767 -0.0663229 1.15573658 -1.01818128
22.2303531 -9.43518131]
Mean squared error: 9.07
R^2 score: 0.76

According to the scores, it is easy to understand that the linear model which best perform accuracy of prediction is the Multiple Linear Regression, due to the fact that in the fist algorithm we perform it only on the most correlated feature with target feature.

3.3 Logistic Regression

Logistic Regression is a supervised learning algorithm, which uses statistical methods, to show a result which represents a probability that an input value belongs to a specific class. In logistic regression problems, the probability that the output belongs to a class is P, to the other class is 1-P (where P is a number between 0 and 1).

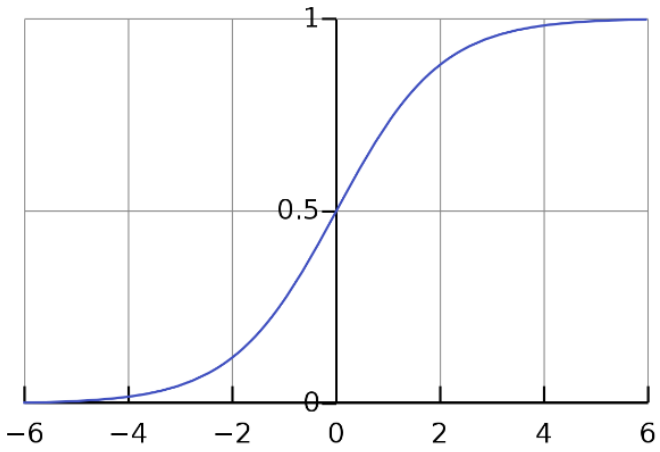


Figure 3.2: Logistic Regression

As all the regression algorithm, logistic regression is a predictive analysis uses to show the relation between dependent variable and one or more independent variables, estimating probabilities with a logistic function with the following formula:

$$y = \frac{e^{-(b_0+b_1 x)}}{1+e^{-(b_0+b_1 x)}}$$

Where:

- X are the input values;
- b_0 and b_1 are the coefficients of input values;
- Y is the output value to be predicted.

This model permits to search values for coefficients which minimize the error in terms of probability. First of all, the probability of an input (X) belongs to a specific class (Y=1) is defined as follows:

$$P(X) = P(Y = 1|X)$$

Using the logistic function, it can be expressed in following way:

$$P(X) = \frac{e^{-(b_0+b_1 X)}}{1+e^{-(b_0+b_1 X)}}$$

Binomial logistic regression permits to classify observations estimating the probability of an observation falls into a particular category (0 or 1). To do that, the dataset, more precisely the target variable, 'life_expectancy' is replaced with boolean values. The average 'life_expectancy' was calculated and it was the treshold of the separation's condition. Major values were replaced with 1, 0 otherwise.

Then the logistic regression was performed and the confusion matrix was calculated.

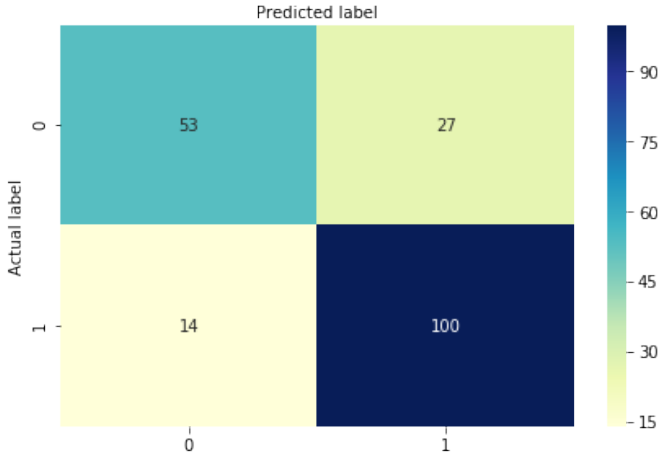


Figure 3.3: Logistic Regression

According to the above image, its easy to understand that the 79% of observation are classified well and there is a 21% of misclassification. This can be expressed also in the next table, with accuracy on train and test score.

	Train Score	Test Score
LogisticRegression	0.83	0.79

Table 3.1: Comparison Between Train And Test Score

3.4 Decision Tree

Decision tree algorithm belongs to the family of supervised learning algorithms: unlike the other ones it can be used for solving regression and classification problems too. The goal of using it is to create a training model that can be used to predict the value of the target variable by learning simple decision rules inferred from prior data (training data); in fact it is able to take a dataset of unknown data and extract a set of rules through which an individual can understand the problem and the results. This type of algorithm has the following advantages: low computation time, easy understanding of the results; the disadvantages instead are: could treat irrelevant data and it is prone to overfitting. Just like the word suggests the algorithm exploits a tree as data structure.

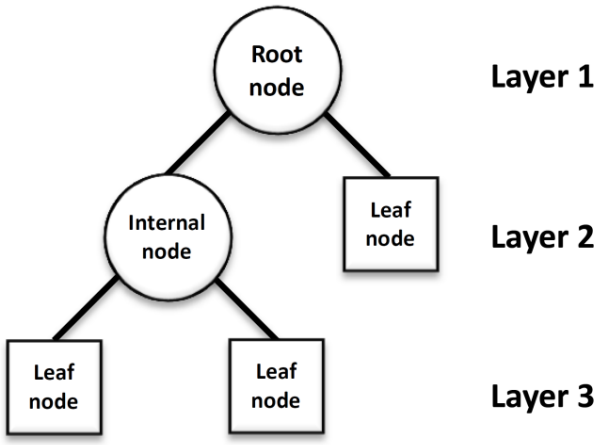


Figure 3.4: Decision Tree

A tree is built by splitting the source set, constituting the root node of the tree, into subsets - which constitute the successor children. The splitting is based on a set of splitting rules based on classification features. This process is repeated on each derived subset in a recursive manner called recursive partitioning. The recursion is completed when the subset at a node has all the same values of the target variable, or when splitting no longer adds value to the predictions. Different algorithms implementing the decision tree use different criteria for measuring the "top" subset: these generally measure the homogeneity of the target variable within the subsets. These metrics are applied to each candidate subset and the resulting values are combined (for instance averaged) to provide a measure of the quality of the split. The previous figure shows an example of decision tree in which each internal node represents a "test" on an feature, each branch represents the outcome of the test and each leaf node represents a class label (decision taken after computing all the attributes). The paths from the root to the leaf represent classification rules.
In Decision Trees some division point are tried and tested using a regression cost function. The division with less cost is select. In the first division, all the features are considered and training data are divided into group according to that division. For example, if there are 3 features, there will be 3 kind of divisions. For each one the accuracy of division is calculated and selected the best one.
For what concerne problems of regression, the cost function is the Sum Squared Error (SSE or Residual Sum Squared, RSS) between all data in the division:

$$RSS = \sum (y - prediction)^2$$

Where y is the real observed value and prediction is the predicted value.
In this work the algorithm has been used for regression purposes, attempting to predict a quantitative response for the target variable. Additionally it has been

trained with several values of depth (the deeper the tree, the more complex the decision rules and the fitter the model).

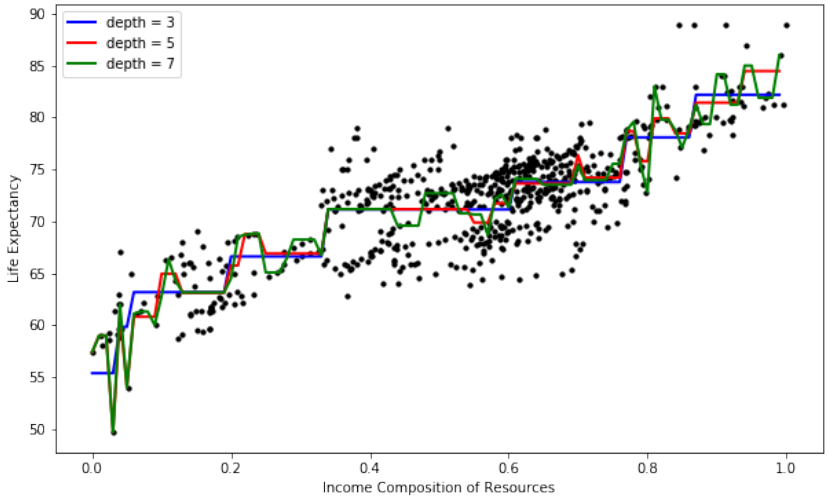


Figure 3.5: Scatter Plot For Different Decision Trees With Different Depths

Here the study focused on choosing a decision tree with the best compromise between complexity and result (plotting the outcome of the algorithm for the target variable and the most correlated one), placing particular attention in the reduction of the number of visited nodes (limited depth of the tree) in favor of a reasonable execution time of the algorithm.

DecisionTreeRegressor	Train Score	Test Score
max_depth = 3	0.74	0.73
max_depth = 5	0.87	0.78
max_depth = 7	0.93	0.72

Table 3.2: Comparisons Among Different Depths

With reference to Table 3.2 the results were computed with three different levels of depth: increasing the depth the classifier tends to the overfitting (train score increases while test score decreases). Selecting the depth equal to three what has been deduced is an appropriate result on the basis of a good score in the training phase, with a difference only of about the 1% from the test score.

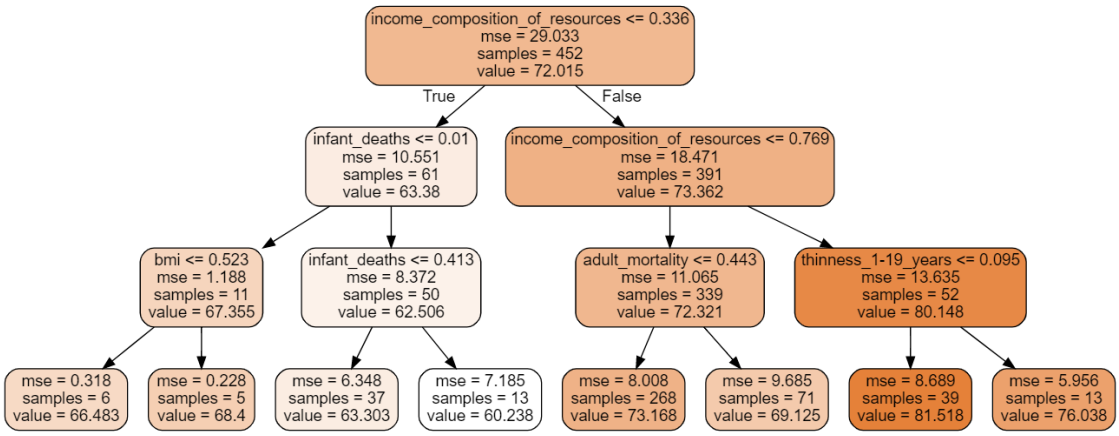


Figure 3.6: Life Expectancy Decision Tree

Some parameters inside every decision block are metrics used to measure the performance of the algorithm: *mse* is the average squared difference between the estimated values and what is actually estimated, *samples* means that the node 'contains' n samples. Since it is the root node this means that the tree was trained on n samples; *value* tells how many samples at the given node fall into each category. Looking at the last figure, the feature which divides substantially the dataset is 'income_composition_of_resources', which is the feature in the first split: in the beginning there were 452 samples (instances). The first partition of the samples into 2 groups was made by comparing the value of the feature 'income_composition_of_resources' with 0.336. The same predictor variable can be used to split many nodes: for example, node number 2 (the one to the right in figure) is further divided using it; on the contrary the other node (on the left) is better split with feature 'thinness_1-19_years'. The process continues up to depth equal to 3: in this case the more the samples of a class, the closer the color to the previous node.

3.5 Random Forest

Random forest is a supervised learning algorithm that consists of a large number of individual decision trees that operate using an ensemble method (Bootstrap Aggregation, Bagging briefly). Ensemble methods use multiple learning algorithms to obtain better predictive performance than could be obtained from any of the constituent learning algorithms alone. Each individual tree in the random forest produces a class prediction; the algorithm combines hundreds or thousands of decision trees, training each of them on a slightly different set of observations, splitting nodes in each tree and considering a limited number of the features for each

one. The final prediction is made by averaging the predictions of each individual tree to get a more accurate and stable prediction. The low correlation among models is the key: uncorrelated models can produce ensemble predictions that are more accurate than any of the individual predictions. While some trees may be wrong, many others will be right, so as a group the trees are able to move in the correct direction.

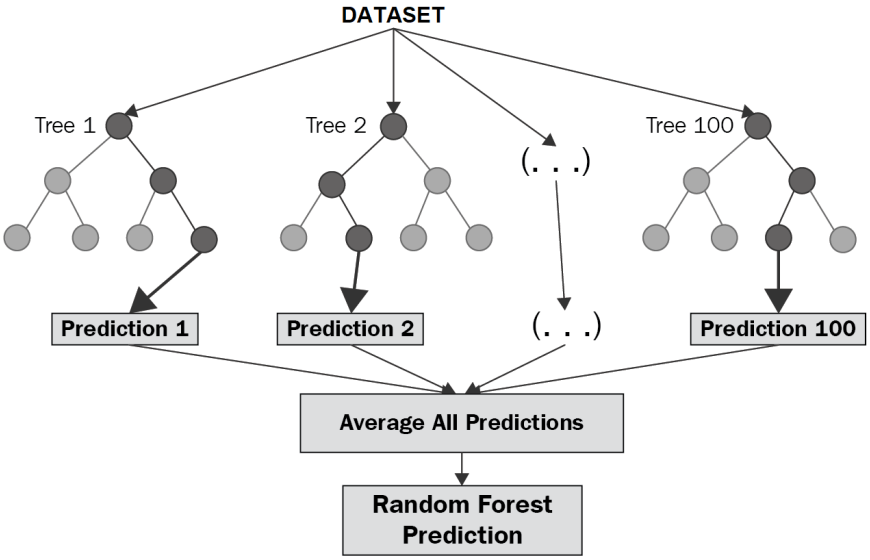


Figure 3.7: Random Forest

In Figure 3.7 can be observed that each decision tree has voted or predicted a specific class $P_n(c|f)$. The final output or class predicted by the random forest will be calculated using the next formula:

$$P(c|f) = \sum_{i=1}^{n} P_n(c|f)$$

During training, every single tree in a Random Forest learn from a casual subset of data. Those subset are designed with substitution (Bootstrap), it means that some subset will be analyzed more times in a single tree. Usually, this subset is set to $\sqrt{n\_features}$, but in regression Random Forest could be trained also considering all the features.
Briefly, the Random Forest mixes hundreds of decision trees, trains each one on a different subset of observations, subdividing nodes in each tree considering less features. On test set, prediction were done calculating the average of predictions of each decision tree.

This means that if there are such as 18 features, random forest will only use a certain number of those in each model: in this way in each tree can be utilized n random features. If there are many trees in the 'forest', eventually many or all of the present features will have been included. This inclusion of many features will help limit the error due to the bias and the variance. If the features were not chosen randomly, trees may have been able to become highly correlated. This is because few features could be particularly predictive and hence the same ones would be chosen in many of the base trees. If many of these trees include the same features there's no need of combating error due to variance. With that said, random forests are a strong modeling technique and much more robust than a single decision tree: they aggregate many decision trees to limit overfitting as well as error due to bias and therefore yield useful results.

	Train Score	Test Score
RandomForestRegressor	0.96	0.90

Table 3.3: Comparison Between Train And Test Score

In the conducted work the algorithm was exploited for a regression problem and as already said, the accuracy of 90% on the test set confirms that is better than the accuracy of individual models due to the fact that the algorithm combines the results from the individual models and provides a final outcome. A fundamental characteristic and of big utility is the feature importance through which it can be easily measured the effect associated to every feature. The scikit-learn python's library that has been adopted finds the importance of a single feature by simply computing how much the nodes of the tree that use that feature reduce the impurity on the mean of all the trees. Strictly speaking, it's a weighted mean where each weight of a node is equal to the number of training data related to the node itself. This technique is realized after the training step and the importance values connected to the features are such that their sum is equal to 1.

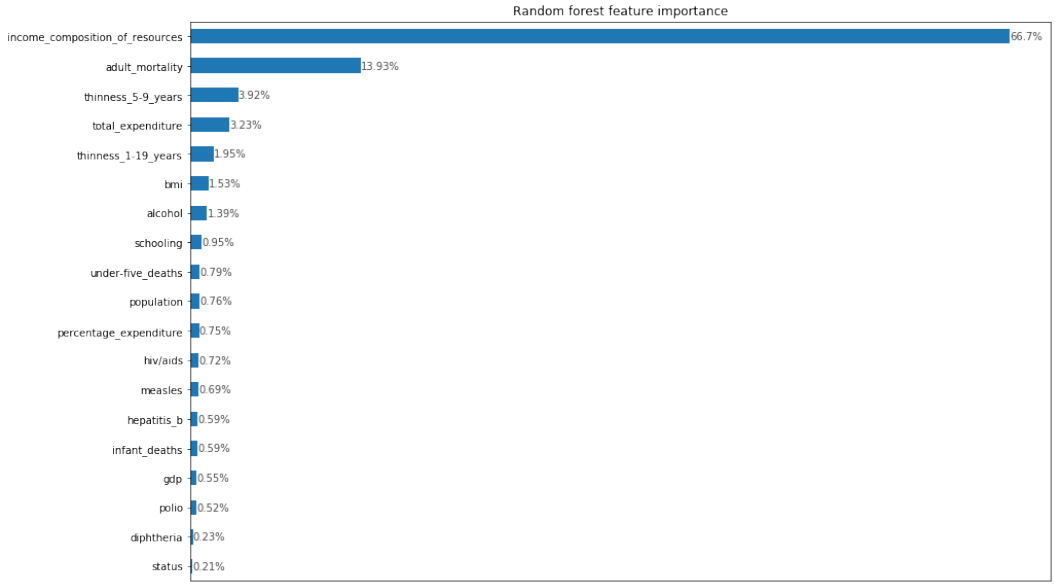


Figure 3.8: Feature Importance Using Random Forest

Let's take a brief look at the feature importance according this analysis: the top 3 most important features are:

- 'income_composition_of_resources'
- 'adult_mortality'
- 'thinness_5-9_years'

It is worth to mention also 'total_expenditure' on health and the 'alcohol' consumption rate in line with Figure 3.8: they are important for predicting 'life_expectancy'. Intuitively the other ones should be less important for predictive purposes.
The essential limitation of this algorithm is that it's not the best when there are real-time data because it requires an higher computation time than the decision tree and it's necessary that some variables must be inserted manually to increase the performances.

Conclusions

The dataset started with 21 uncleaned features (including the target variable) and has been pared down to 18 features to describe the target variable 'life_expectancy'. The first step was to give a general description about the dataset, the features and the used tools so that a better understanding could be gathered.
The second one was to clean the data: the features were renamed to uniform them to a standard format; the categorical features 'year', 'status', and 'country' were removed as metadata. Then the cleaning included detecting and dealing with both missing values and outliers; once missing values were deleted, the next step was detecting and dealing with outliers. Extreme value detection was done by using box plots with a standard IQR threshold of 1.5: using this technique, each feature was investigated on a one by one basis to eliminate outliers while limiting the loss of data. Once this step was complete, exploration of the data has been conducted. The now cleaned dataset was explored using several techniques: the primary method used in this assignment was the correlation matrix in conjunction with the heatmap in order to detect variables that were highly correlated to one another; later from those pairs keep the features which were highly correlated with the target. This was the main foundation for feature selection: but before moving on, some categorical variables were compared to the target variable using scatter plot. Then the dimensionality reduction method of PCA was used: the number of variables could be dropped from 18 down to 10 features but didn't seem to garner very useful results.
Finally data analysis has been performed, which represent the core of this work. The dataset was subject to five different techniques: linear regression, multiple linear regression, logistic regression, decision tree and random forest. First, the linear regression was executed using the most correlated feature ('income_composition_of_resources') with the target on the basis of the heatmap and it was found an accuracy of 57%, then with the Multiple Linear Regression was found a test score of 76%, that means that using all the features instead of only the most correlated one is better for linear regression. Moreover, to perform logistic regression we divided 'life_expectancy' into boolean values using the average value as treshold, and it was obtain an accu-

racy of 79%. Using all the available features the decision tree gave an accuracy of 73% with best depth (3); last but not least the accuracy reached by random forest is better than the previous ones being more than 90%.
In general for the analysed dataset the used techniques are good enough: indeed they are able to hold for the models a discrete accuracy (leaving out linear regression) above the 70% and the models will be able to predict in the same way the target variable for new records that will be added to the dataset in a second time as long as they will have the same features.
In conclusion the elaborate has been showed an investigation aimed at choosing the relevant factors that affect the life expectancy with the attempt of generalising the result although the different context present in the various countries; above all it gave a concrete opportunity for deepening some topics explained during the course.