heg

Haute école de gestion
Genève

Hes·so // VALAIS WALLIS

ML project

Group DM_ZF

Introduction to Machine Learning

Created by
Fernández Rodríguez Zanya, Marsoni David

Life expectancy dataset

Table of content

1 Context

Our chosen dataset, Life Expectancy (WHO), focuses on analysing life expectancy and related health factors for 193 countries over the period 2000-2015. It integrates data from the WHO Global Health Observatory and the United Nations, covering data such as immunization, mortality, economic, and social factors. You can find the link to this dataset here. (*Life Expectancy (WHO)*, 2018)

2 Methodology

2.1 Problem understanding

For this project, we shall try to answer the following question:

- **In function of the person, what is the life expectancy of a person?**
- **Can we classify correctly a person's life span and predict if she will have a high or low life expectancy?**

To answer this question, we shall understand and predict our target's life expectancy in function of parameters such as their development, education.... This analysis is interesting because it highlights the factors that influence the most in people's life span.

2.2 Data understanding

While missing data from lesser-known countries were excluded, this resulted in a final dataset with 22 columns and 2938 rows, comprising 21 features.

To see the details of all the feature descriptions see annex A.

2.3 Data preparation

The Data preparation part is composed of 3 main parts. First, we need to be able to see and understand the missing or duplicated value of our dataset. Next, we have to do an outliner analysis to see if there are some values that may be errors. Then we need to observe the correlation matrix to avoid keeping value with a too high correlation in our dataset.

2.3.1 Missing and duplicated value

When we checked our values, we quickly realized that while there are no duplicates, there are some features that have a lot of Null values, such as Hepatitis B, which has 553.

To avoid removing more than half of the dataset we have decided to take the mean value of the non-null value to replace all the null values with it.

2.3.2 Outliners

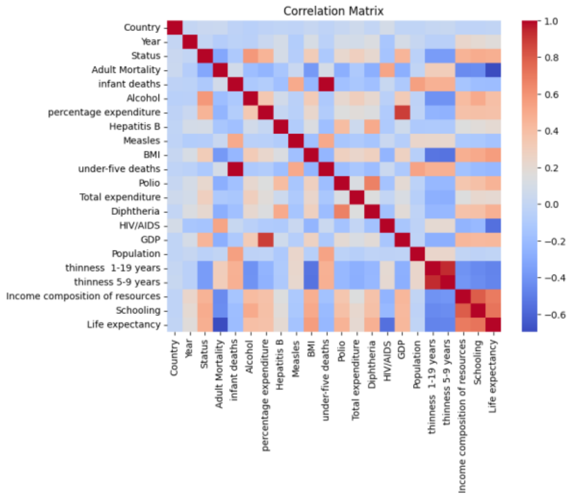
The analysis of the outliner for this dataset is not really relevant as the data are really diversified and vary a lot between the countries. Instead of looking for outliner value, we have focused our effort on aberrant values such as negative values. After our test, we found no negative value.

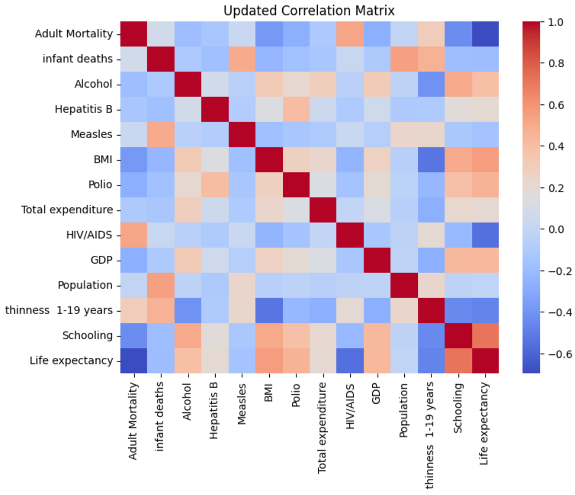
2.3.3 Data correlation

After plotting our correlation matrix, we have found some abnormally high correlation (>80%) between some of our features. To avoid such correlation, we have decided to only keep one of the 2 features correlated. On the correlation graph, we have also seen that our 2 strings have practically no correlation with our target

In the end, we have removed these 7 features

- "Income composition of resources" was highly correlated with "School" and the target feature.
- "Year" and "Country" that have practically no correlation with our target column.
- "under-five deaths" highly correlated with the "infant deaths" feature.
- "Diphtheria was highly correlated with the "Polio" feature.
- "percentage expenditure" was highly correlated with the "GDP" feature.
- "thinness 5-9 years" This feature is contained in the "thinness 1-19 years" column.





2.4 Modelling

All our models will use cross-validation with 5-fold as it helps to evaluate the true performance of a model by testing it on multiple training/testing sets. It aims to reduce the overfitting by training the model on different subsets of data. We have chosen 5-fold as it does not split the dataset into too small sections and helps to see the real result of the model train.

All the following information is based on both the 62-62 Data Exploitation course and the introduction to machine learning. As a complement, we have searched some explanations on the internet for the models and here are the 2 main websites where we have retrieved information. (*Machine Learning Algorithms | Microsoft Azure*, n.d.) and (Ibm, 2024)

2.4.1 Classification

To be able to conduct some classification tasks on this dataset we needed to create a new target binary target to create this target feature we have used the median of the dataset as a threshold to be able to create 2 categories for the life expectancy. The median choice allows us to have a balanced dataset with the same number of low and high life expectancies. This newly created feature will be used as a reference to train all our models.

Decision Tree

The decision tree model is a supervised learning algorithm used for classification tasks. This algorithm uses a tree-like structure where each node is a condition that splits the data. This model predicts a case by navigating the tree and executing their conditions.

This model doesn't need any normalization or standardization of the data provided as it cares about the data values to create its nodes, not the data magnitude.

KNN

K-Nearest Neighbors (kNN) algorithm classifies a data point based on the most-resembling class of its k closest neighbours. It calculates the Euclidean distance to identify the closest neighbours.

As this algorithm is based on the distance between points the normalization and standardization of the data is essential to not introducing biases with data that have different magnitudes. Standardization is used to be able to extract valuable data from the string values like the "Status" feature.

To have the best performance with this model we have conducted some tests to find the best k based on the F1 score.

Logistic regression

The logistic regression model is a statistical model used for binary classification that predicts the probability of an outcome belonging to one of two categories. This model uses a logistic function, such as the sigmoid curve, to output a probability. This probability is then converted into one of the two classes using a threshold, usually set at 0.5.

The model needs to have data that are normalize and standardized to be able to improve the convergence during the training phase and avoid misinterpretation of data that have different scales.

Random Forest

The random forest algorithm builds multiple decision trees during training and merges their result to improve accuracy and prevent overfitting. This model as the name suggests takes a random subset of the feature to create a test on it. This algorithm differs from cross-validation as it aggregates the result of all the trees not just evaluating the result values of trees.

As the decision tree algorithm, this model doesn't need to be normalised or standardised because it uses multiple decision trees that are not affected by the magnitude of the data.

Naïve Bayes

The Naïve Bayes probabilistic algorithm is based on the Bayes' theorem, which assumes independence between features. This model will calculate the probability of each class given the feature input and assign the classes with the highest probability.

The data input in this model doesn't need to be normalised and standardised because it uses probability distribution that inherently handles the different scales of the features.

2.4.2 Regression

For the regression task the target feature "Life expectancy" will be used as a target to train the different models.

Linear / Polynomial regression

The linear regression algorithm is a statistical model that tries to find the best straight line to predict the target variable using all the features provided. In addition, the polynomial regression algorithm is an extension to capture more complex nonlinear trends for the target.

These 2 algorithms need to have data that he normalizes and standardizes to be able to converge more efficiently and avoid giving more weight to data with bigger scales.

Decision tree regression

The decision Tree Regression algorithm predicts a continuous target by splitting data into subsets based on feature thresholds, forming a tree structure. Each split tries to minimize the variance, and the leaf of the tree provides the predictions.

This algorithm didn't need to have data that are standardize or normalize as it creates nodes with thresholds that handle different scales of data.

Random Forest Regression

Random Forest Regression model that combines multiple decision trees to improve prediction accuracy. Each of the subtree trees is trained on a random subset of data and features, and then the prediction is calculated with the average of each tree. It reduces overfitting compared to a single decision tree by averaging the results, making it more robust.

This algorithm didn't need to have data that are standardize or normalize as it create a forest or tree with different nodes with thresholds that handle different scales of data.

2.5 Evaluation

2.5.1 Classification

Methods

To evaluate the best our regression model we have decided to compute the following score:

- **AUC** (Area Under the Curve): Measures the area under the ROC curve, indicating the model's ability to distinguish between classes.
- **Accuracy**: Proportion of correctly predicted instances out of total instances.
- **Precision**: Proportion of true positive predictions out of all positive predictions made.
- **Recall**: Proportion of true positive predictions out of all actual positive instances.
- **F1 Score**: Harmonic means of precision and recall, balancing the trade-off between them.
- **CV Score** (Cross-Validation Score): Average performance metric (e.g., accuracy, precision) from cross-validation, assessing the model's generalization.

Choice of the best model

To choose our best model we have first sort it by the AUC and then we have compared this score with the CV_Score. This approach allows us to have a model that have a good ability to distinguish the low and high life expectancy.

Results

Model	AUC	Accuracy	Precision	Recall	F1	CV_Score
Random Forest	0.993	0.951	0.952	0.946	0.949	0.906
Decision Tree	0.959	0.917	0.896	0.936	0.916	0.886
k-Nearest Neighbors	0.968	0.893	0.868	0.917	0.892	0.856
Logistic Regression	0.957	0.882	0.847	0.920	0.882	0.860
Naïve Bayes	0.929	0.82	0.758	0.917	0.830	0.817

Here our best-performing model is the Random Forest. This model has the best overall metrics in all the categories. Another model tester performs also well on this dataset. To see the detailed result see annex B.

2.5.2 Regression

Methods

To evaluate the best our regression model we have decided to compute the following score:

- **MSE** (Mean Squared Error): Average squared difference between actual and predicted values, penalizing large errors.
- **MAE** (Mean Absolute Error): Average absolute difference between actual and predicted values, less sensitive to outliers.
- $\boldsymbol{R^2}$(Coefficient of Determination): Proportion of variance in the target explained by the model, ranges from 0 to 1.
- **RSS** (Residual Sum of Squares): Total squared differences between actual and predicted values, measures model error.
- **CV Score** (Cross-Validation Score): Average performance metric (e.g., MSE, MAE) from cross-validation, indicating model robustness.

Choice of the best model

To choose our best model we have first sort it by the R^2 sore and then compare it to the cross-validation score to see if the model was not overfitting. The other score allow use to have a deeper analysis of each model.

Results

Model	MSE	MAE	$\boldsymbol{R^2}$	RSS	CV_Score
Random forest regressor	5.667	1.534	0.939	4997.982	0.885
Decision Tree Regressor	7.486	1.774	0.918	6744.619	0.846
k-Nearest Neighbors Regressor	7.995	1.982	0.914	7051.722	0.803
Linear Regression	19.01	3.179	0.797	16766.384	0.767
Polynomial Regression	82.792	6.689	0.114	...	-0.088

Here our best-performing model to predict the continuous life expectancy variable is the Random Forest regressor which has the best metrics. The other algorithms are also performing well except for the Polynomial Regression which clearly doesn't explain the model. That means that our model is better explained by a linear function than a curve. To see the detailed result of our best model see the annex C.

2.6 Conclusions

This dataset was really interesting to use because it forced us to really follow carefully the steps of the CRISP-DM methodology to first clean the data for all the missing and error values then remove the highly correlated column. (IBM, 2018)

The modelling part was interesting because this dataset was suitable for Either classification or regression task that have open us a variety of possibilities.

In addition, our result is better that we were expecting and that make us happy because the model train are practically suitable for real cases.

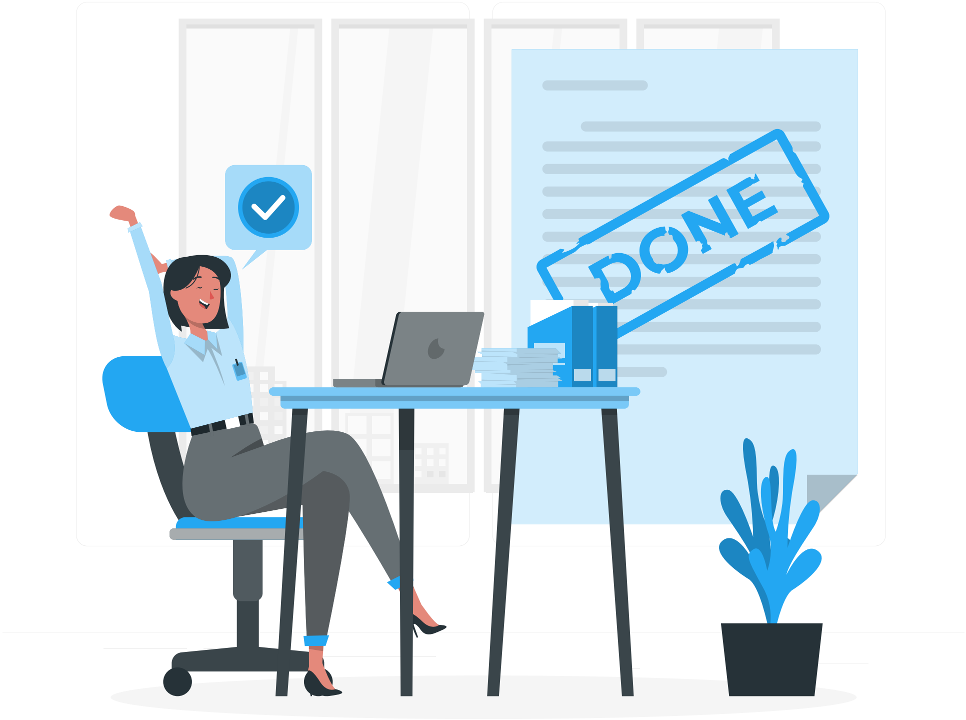
Finally, to answer formally to our questions:

In function of the person, what is the life expectancy of a person?

For this question with the random forest regressor we are able to achieve pretty good score for predicting the life expectancy of a person. If you want more detail about the result of our best model see the Annex C.

Can we classify correctly a person's life span and predict if she will have a high or low life expectancy?

This question was good resolve by our random forest classifier that have achieve an accuracy score of 94,7%. If you want more detail about the result of our best model see the Annex B.



3 References

Life expectancy (WHO). (2018, February 10). Kaggle. https://www.kaggle.com/datasets/kumarajarshi/life-expectancy-who

IBM. (2024, December 19). Machine Learning Algorithm. *Machine Learning Algorithm*. https://www.ibm.com/think/topics/machine-learning-algorithms

IBM. (2018). IBM SPSS Modeler CRISP-DM Guide. In *IBM SPSS Modeler CRISP-DM Guide*. https://www.ibm.com/docs/it/SS3RA7_18.3.0/pdf/ModelerCRISPDM.pdf

Machine Learning Algorithms | Microsoft Azure. (n.d.). https://azure.microsoft.com/en-us/resources/cloud-computing-dictionary/what-are-machine-learning-algorithms

Storyset. (n.d). *Illustration*. https://storyset.com/illustration/

4 Annexes

4.1 Annex A

Features

- **Country**: Full name of the country being referred to.
- **Year**: The year when the data was recorded.
- **Status**: Economic classification of the country ‘Developed' or 'Developing'.
- **Adult Mortality**: Adult Mortality Rates of both sexes (probability of dying between 15 and 60 years per 1000 population).
- **Infant deaths**: Number of deaths of infants under per 1,000 population.
- **Alcohol**: Recorded per capita (15+) alcohol consumption in litters of pure alcohol.
- **Percentage expenditure**: Expenditure on health as a percentage of Gross Domestic Product per capita (%).
- **Hepatitis B**: Percentage of infants who received three doses of the Hepatitis B vaccine.
- **Measles**: Number of reported cases of measles per 1000 population.
- **BMI**: Average Body Mass Index (BMI) of the population.
- **Under-five deaths**: Number of deaths of children under five years per 1,000 live births.
- **Polio**: Percentage of children under one year who received three doses of the Polio vaccine.
- **Total expenditure**: Total government health expenditure as a percentage of GDP.
- **Diphtheria**: Percentage of children under one year who received three doses of the DPT (Diphtheria, Pertussis, and Tetanus) vaccine.
- **HIV/AIDS**: Deaths per 1 000 live births HIV/AIDS (0-4 years).
- **GDP**: Gross Domestic Product per capita, expressed in current US dollars.
- **Population**: Total population of the country in the given year.
- **Thinness 1-19 years**: Percentage of individuals aged 1–19 who are underweight.
- **Thinness 5-9 years**: Percentage of individuals aged 5–9 who are underweight.
- **Income composition of resources**: A human development index measurement of income sustainability and availability of resources.
- **Schooling**: Average number of years of schooling attained by individuals in the country.

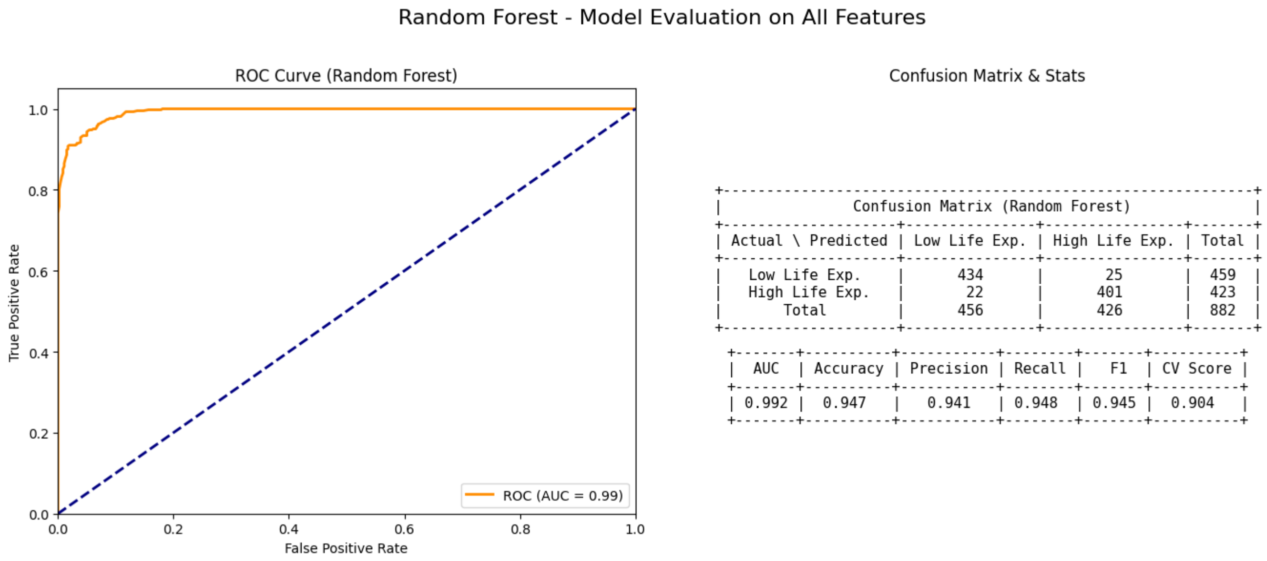
Target

- **Life expectancy**: Average number of years a newborn is expected to live if current mortality rates remain constant.

Life expectancy (WHO). (2018, February 10). Kaggle.
https://www.kaggle.com/datasets/kumarajarshi/life-expectancy-who

5 Annexe B

Here is the detailed result of our best classification model (Random Forest):



```
+-----------------------------------------------------------------+
|                  Confusion Matrix (Random Forest)               |
+------------------+----------------+-----------------+-------+
| Actual \ Predicted | Low Life Exp. | High Life Exp. | Total |
+------------------+----------------+-----------------+-------+
|   Low Life Exp.   |      434       |       25        |  459  |
|   High Life Exp.  |       22       |      401        |  423  |
|       Total       |      456       |      426        |  882  |
+------------------+----------------+-----------------+-------+

+-------+----------+-----------+--------+-------+----------+
|  AUC  | Accuracy | Precision | Recall |  F1   | CV Score |
+-------+----------+-----------+--------+-------+----------+
| 0.992 |  0.947   |   0.941   | 0.948  | 0.945 |  0.904   |
+-------+----------+-----------+--------+-------+----------+
```

The Random Forest model demonstrates great performance in predicting the life expectancy binary class with a high accuracy of 94,7%. The AUC score of 99,2% shows good discrimination of the model between the high and low life expectancy classes.

The slightly below score for the Cross-validation score reflects that the model is slightly overfitting but keeps a good score.

For 100 people, this model is able to predict 95 people correctly, with 5 people wrongly classified.

Among these 5 errors:

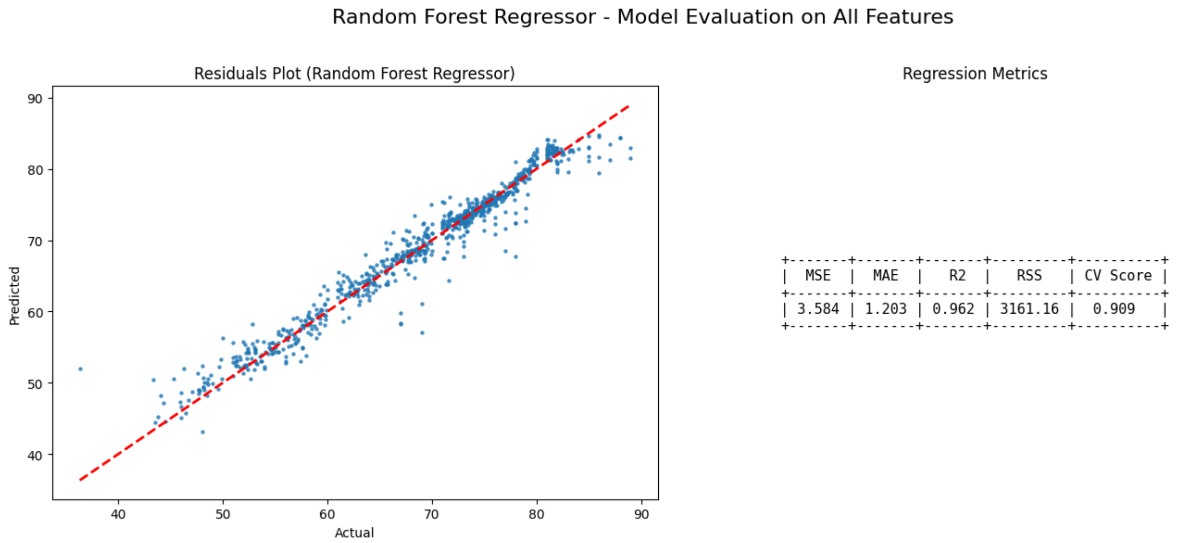
- 3 are classified as false positives, meaning the model incorrectly predicts that a person has a high life expectancy when they actually have a low life expectancy.
- And 2 are false negatives, meaning the model incorrectly predicts that a person has a low life expectancy when they actually have a high life expectancy.

This demonstrates that the model slightly overestimates high life expectancy cases but maintains a strong overall accuracy.

The problem with false positives is that they might cause people with low life expectancy to mistakenly believe they are in good health or have a higher life expectancy than they do

6 Annexe C

Here is the detailed result of our best regression model (Random Forest Regressor):



```
+-------+-------+-------+---------+----------+
|  MSE  |  MAE  |   R2  |   RSS   | CV Score |
+-------+-------+-------+---------+----------+
| 3.584 | 1.203 | 0.962 | 3161.16 |  0.909   |
+-------+-------+-------+---------+----------+
```

The random forest performs quite well on this dataset. The R^2 value of 0.962 indicates that the model is correct 96.2% of the time. This score compared to the CV score is slightly higher that mean that the model is a little overfit.

The low MSE and MAE indicate that the model is performing very well on the dataset. In particular the MAE of 1.203 indicates that the error of the prediction is in a range of 1.2 years. For example, a people that have a known life expectancy of 70 could have a predicted life expectancy of 68.8 year to 71.2 year.