LIFE EXPECTANCY, DOES IMMUNIZATION MATTER?

Contents

Summary

In view of the current pandemic, vaccination does not seem to find the consent of some people in the world. A look at the impact of vaccination on life expectancy could be important to highlight. So, does immunization matter? The dataset related to life expectancy, health factors for 193 countries have been collected as well as its corresponding economic data from 2000-2015. The predicting variables were then divided into several broad categories: Immunization related factors, Mortality factors, Economical factors, and Social factors. The present project assessed the contribution and the relationship of each feature on life expectancy with a special focus on immunization factors and develop a model to predict life expectancy.

- Life expectancy has increased over years in both developed and developing countries
- The mean average of the life expectancy of developed countries is generally higher compared to that of developing countries
- However, the ratio of LE over the decade of 2005 to 2015 showed that life expectancy in developing countries has greatly increased.
- It has been highlighted that immunization has impacted the improvement of life expectancy in a developing country, as well as the reduction in infant deaths.
- Analysis revealed that economic factors played an important role in the system; it is why countries with higher income resources and GDP tend to have high life expectancy even if they are highly populated. In developing countries, an increase in the population tends to impact negatively life expectancy.
- Many(14) regression models have been developed to predict life expectancy, the chosen one is Gradient boost with MAE of 0.202 on train set and 1.431 on the test set, R square is 0.94 on the test set.

1 Introduction

In view of the current pandemic, vaccination does not seem to find the consent of some people in the world. A look at the impact of vaccination on life expectancy could be important to highlight. But also, to identify the other factors which contribute to the improvement of life expectancy in the world. Moreover, it has been observed that in the past 15 years, a huge development in the health sector resulted in improvement of human mortality rates especially in the developing nations in comparison to the past 30 years. The Global Health Observatory (GHO) data repository under World Health Organization (WHO) keeps track of the health status as well as many other related factors for all countries. The dataset related to life expectancy, health factors for 193 countries have been collected as well as its corresponding economic data from 2000-2015. Among all categories of health-related factors, only those critical factors were chosen which are more representative. The final dataset consists of 22 Columns and 2938 rows which means 21 predicting variables. All predicting variables were then divided into several broad categories: Immunization related factors, Mortality factors, Economical factors, and Social factors.

For this project, we are interested in developing a model which predicts life Expectancy based on the factors identified. We could break them down with the following questions:

- What is the trend of life expectancy from 2000 to 2015?
- Are immunization features related to mortality features (such as infant deaths, under five-year-old deaths, and adult deaths)?
- What are the factors that impact the trend of life expectancy?
- Does immunization factors impact the increase of life expectancy compared to other factors?
- What are the factors that affect the life expectancy model?
- What was the level of life expectancy over one decade (2005 -2015)? And what were the top 30 countries with a great increase in life expectancy over this decade?
- What is the 95% confidence interval of life expectancy?
- Is the average mean of life expectancy from developed countries greater than that of developing Countries?

2 Data Processing

The original dataset has 2938 observations and 22 columns. From these 22 variables, 21 are independent variables. Except for the Country, and status, all the variables are numerical. Data processing started with remaining, removing, and replacing symbols on columns name. Next, we analyzed outliers and missing data in the dataset.

2.1 Outliers

Boxplot is used to visualize outliers. The boxplot of all features was done, and it highlights the importance of outliers. The Winsorize method was used to treat outliers. Winsorization is a symmetric process that replaces the k smallest and the k largest data values. We made our upper and lower limits for data our new maximum and minimum points. In this project, we apply the quantile (0.01) and quantile (0.98) as a boundary. We also explore the IQR range method at one standard deviation. This treatment (winsorize) has significant impact in some features such as Life expectancy, but do not improve the spread of others features as Populations.

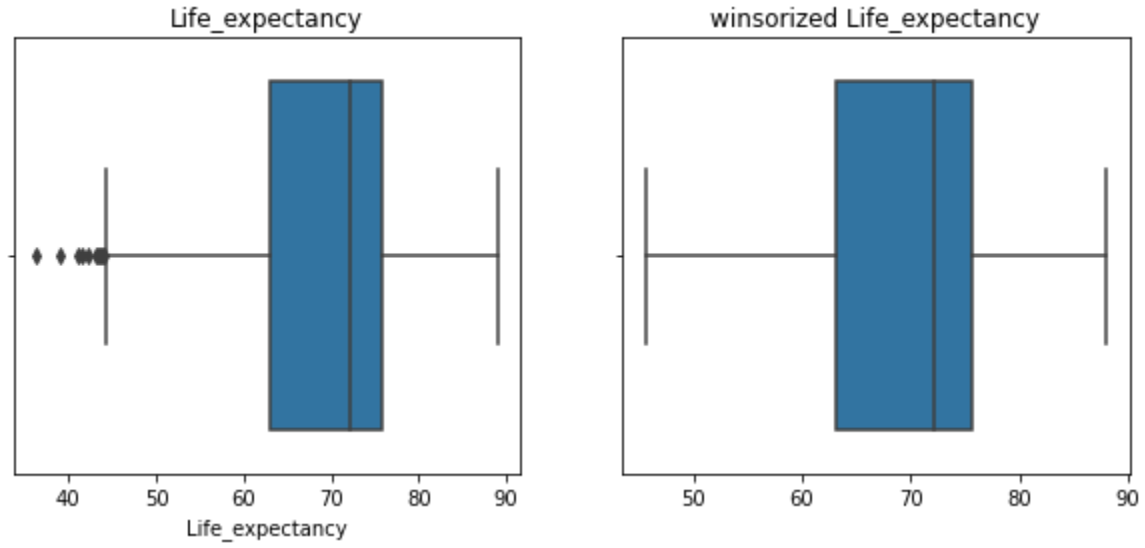


Figure 1. Box plot Life Expectancy feature before and after winsorize

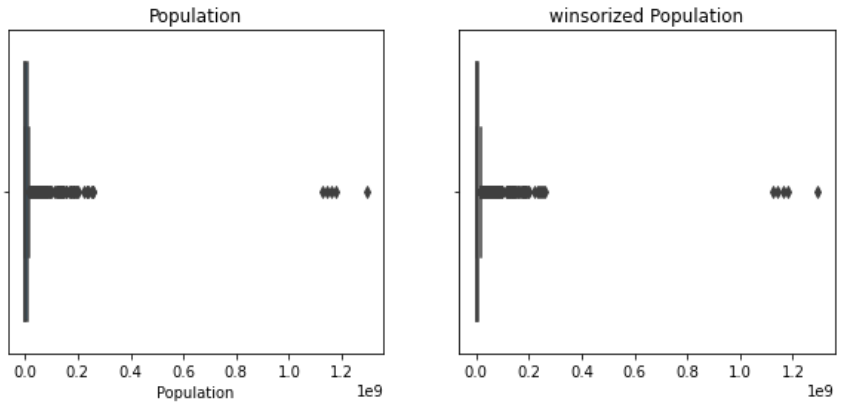


Figure 2. Box plot population feature before and after winsorize.

NB: The data set with outlier treatment was not consider in the dataset where we carryout exploratory data analysis and modeling because we wanted to avoid leakage with this action on dataset.

2.2. Missing values

The original dataset has 14 columns with missing values, where the population has 22% of missing values, followed by Hepatitis_B with 19% and GDP with 15%. The others 11 columns were between 7% and 1%. After the replacement of the uncommon type of missing data (zero was converted to NAN), the number of columns affected by missing data increased (19 columns) as well as the total number of missing values.

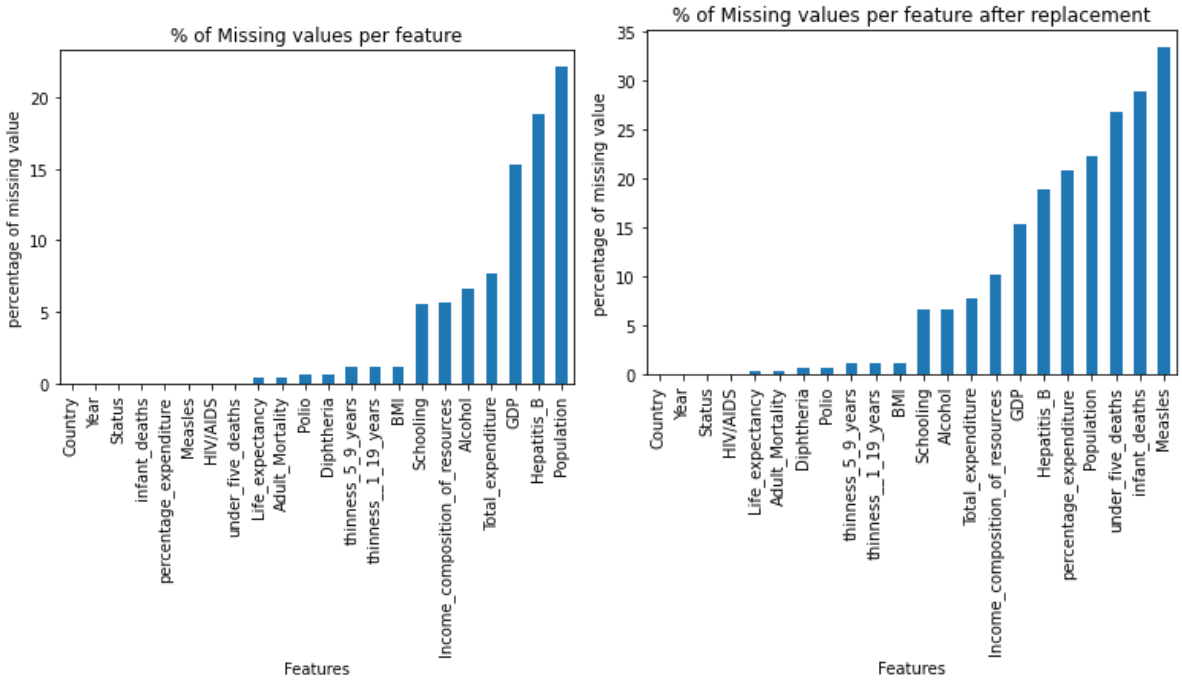


Figure 3. Original state of data with missing value and after the replacement of the uncommon type with NAN

2.3 Treatment of missing data

We would have expected a loss of 69% if we decided to drop all missing values from the dataset representing 911 observations out of 2938. Different imputation technique had been used to address the issues of missing value. Mean, mode, constant and median imputation with SimpleImputer from sklearn, and KNN imputation, MICE imputation using fancyimpute.

The evaluation of the imputation method was done with the simple linear regression, the Ordinary Least Squares (OLS) of stats models. The result showed that mice imputation was the best one. In fact, mice algorithm performs multiple regressions for imputing. An example of plotting was done on population features to see how output of imputation is close to the base line.

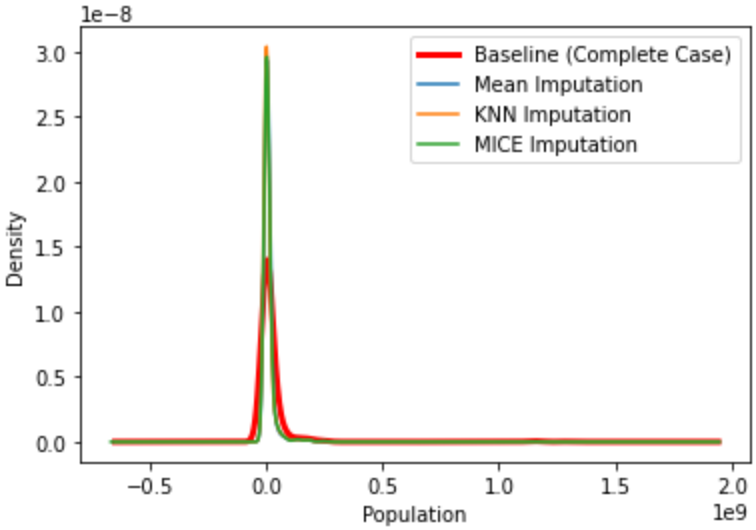


Figure 4. Visualizing Mean, Mice and KNN imputation on Population feature

3 Exploratory Data Analysis

From the exploratory data analysis, we check for duplicates, there was no duplicate found in the dataset. There are 193 countries, where 32 are developed countries (DDC) and 151 are developing countries (DPC).

3.1 Life expectancy versus Categorical features (Countries and status)

Life expectancy increased over years (figure 1) in both countries. The minimum life expectancy in developed countries is 69 whereas in developing countries is 39. The maximum life expectancy in developed countries is 89 and it is 89 in developing countries. The range of life expectancy is huge in developing countries as compared to developed countries. This highlights the fact that we found high difference in life expectancy among countries of developing countries (figure 2).

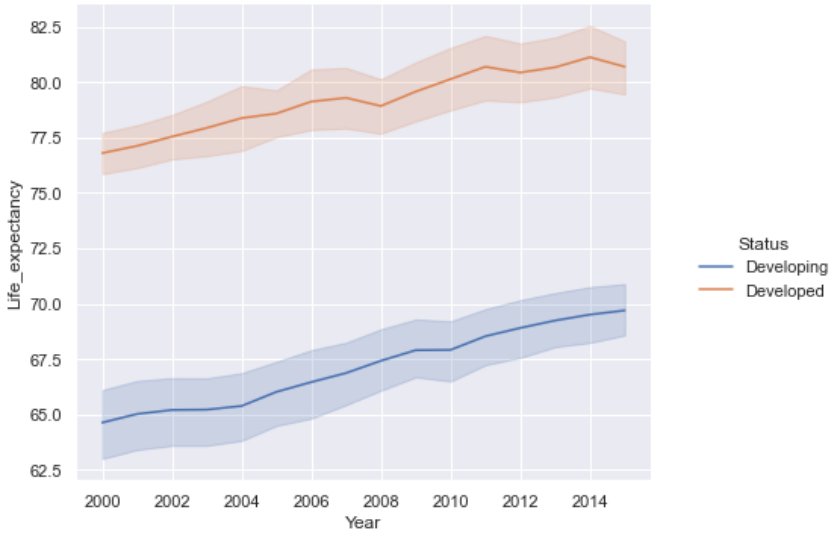


Figure5: Trend of life expectancy overs 2000 to 2015

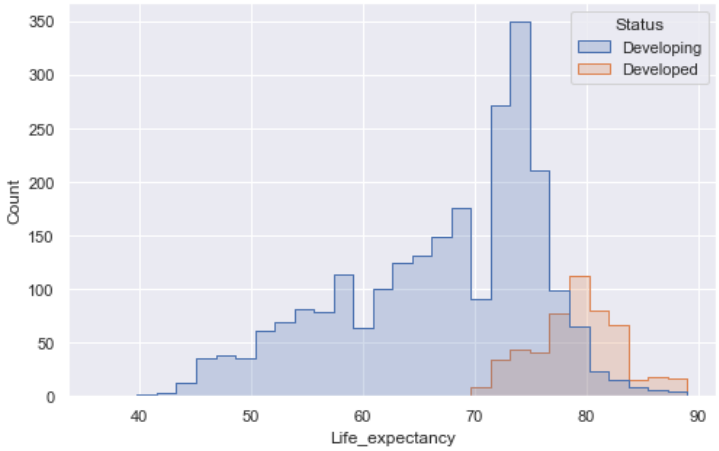


Figure 6. Histogram of the distribution of life expectancy

Top 10 developing Countries with high life_expectancy

	Country	mean	std
0	France	82.21875	3.166958
1	Canada	81.68750	2.240201
2	Israel	81.30000	1.556920
3	Greece	81.21875	3.006264
4	Finland	80.71250	3.354574
5	Republic of Korea	80.48750	3.177184
6	Chile	79.45000	2.073323
7	Costa Rica	78.59375	0.711307
8	Cuba	77.97500	0.745207
9	Qatar	77.03125	0.672031

Top 10 developed Countries with high life_expectancy

	Country	mean	std
0	Lithuania	72.80625	1.848411
1	Bulgaria	72.85000	1.050714
2	Latvia	73.73125	2.312205
3	Hungary	73.82500	1.296405
4	Romania	74.05000	1.995996
5	Slovakia	74.75000	1.120714
6	Poland	75.65000	1.170755
7	Croatia	76.11875	1.144388
8	Czechia	76.76875	1.322986
9	United States of America	78.06250	0.832566

From the statistical analysis, the 95 % confidence interval of life expectancy in developing countries is [64.834, 68.364], and the 95% confidence interval of life expectancy in developed countries is [77.911, 79.450].

We conduct a Z test for 2 independent samples at 95% to test whether the mean average of life expectancy from developed countries is always greater than that of developing countries. We rejected the null hypothesis and found that there is a significant difference between the mean average life expectancy in developed countries and that of developing countries.

We analyzed the increase in life expectancy over one decade (2005 -2015) and got the top 30 countries with the highest rate of life expectancy (figure 3). Except Switzerland (2nd position) and Slovenia (20th position), we found that almost all of the countries (the top 30) which high ratio of life expectancy over a decade of 2005 to 2015 are from developing countries.

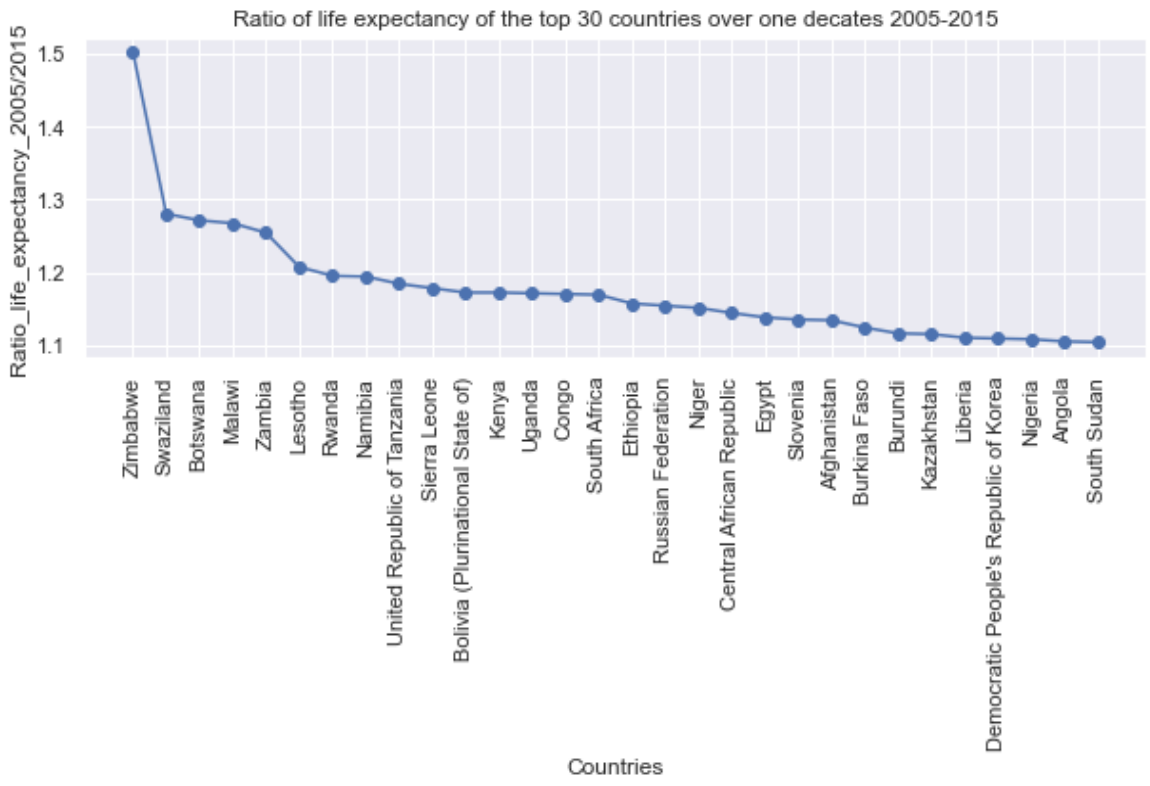


Figure7. life expectancy ratio of the top 30 countries from 2005 to 2015

3.2 Life expectancy versus Numerical features

3.2.1 Immunization factors

Immunization factors identified are :

- Hepatitis B (HepB) immunization coverage among 1-year-olds (%)
- Polio (Pol3) immunization coverage among 1-year-olds (%)
- Diphtheria tetanus toxoid and pertussis (DTP3) immunization coverage among 1-year-olds (%)

The figures below highlight the impact of immunization factors on life expectancy.

Figure 8. correlation life expectancy and immunization

In general, from the figures above, Immunization features has a positive impact on increasing life expectancy in developing countries compared to developed countries where the impact is not important.

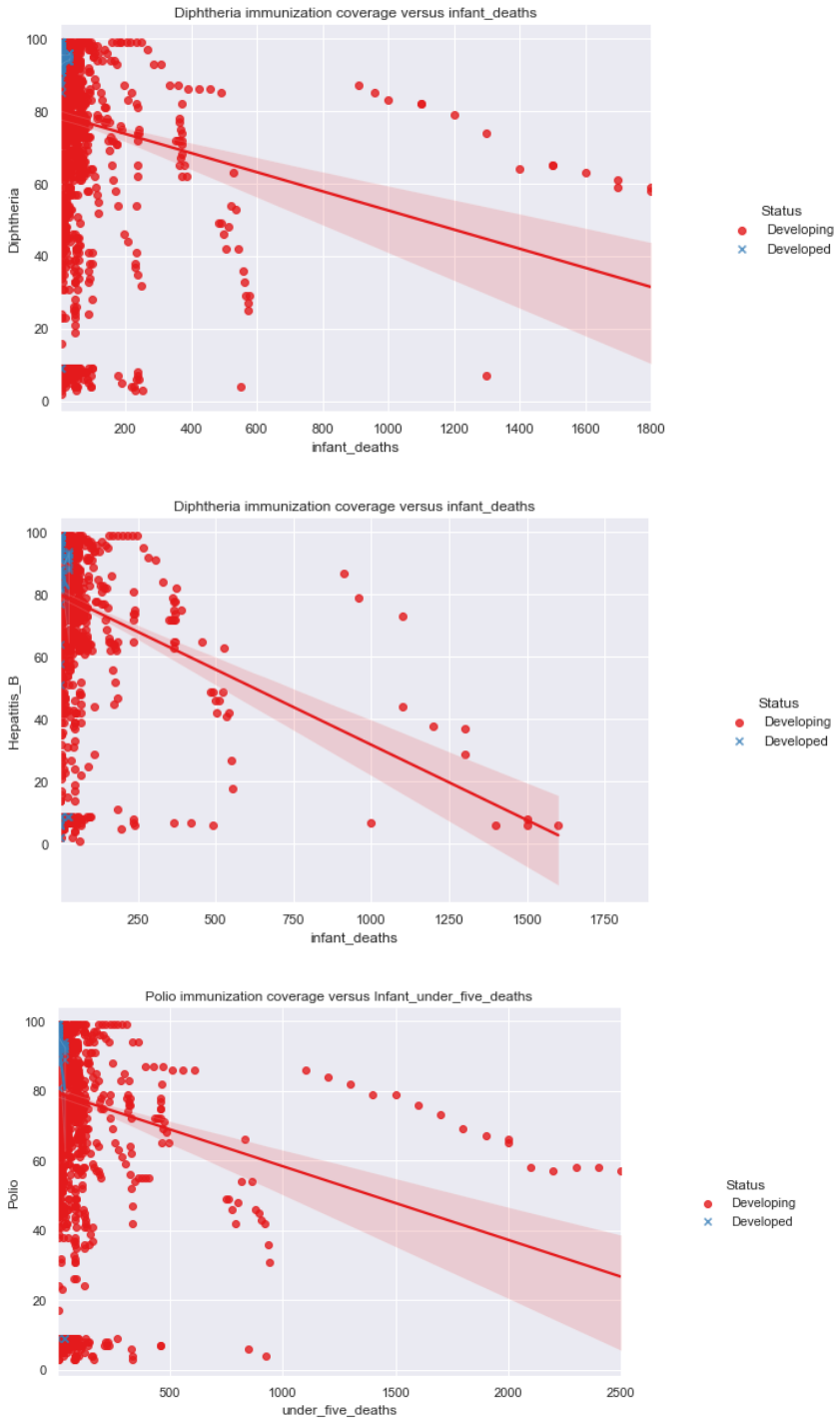


Figure 9. correlation immunization and death factors

From the figures above, we observed that immunization features are related to mortality features such as infant deaths, under five-year-old deaths, and adult deaths. The mortality decreases with the increase of immunization.

3.2.2 Mortality factors

Mortality factors are :

- Adult Mortality Rates of both sexes (probability of dying between 15 and 60 years per 1000 population)
- Number of Infant Deaths per 1000 population
- Number of under-five deaths per 1000 population
- Deaths per 1 000 live births HIV/AIDS (0-4 years)
- Measles - number of reported cases per 1000 population

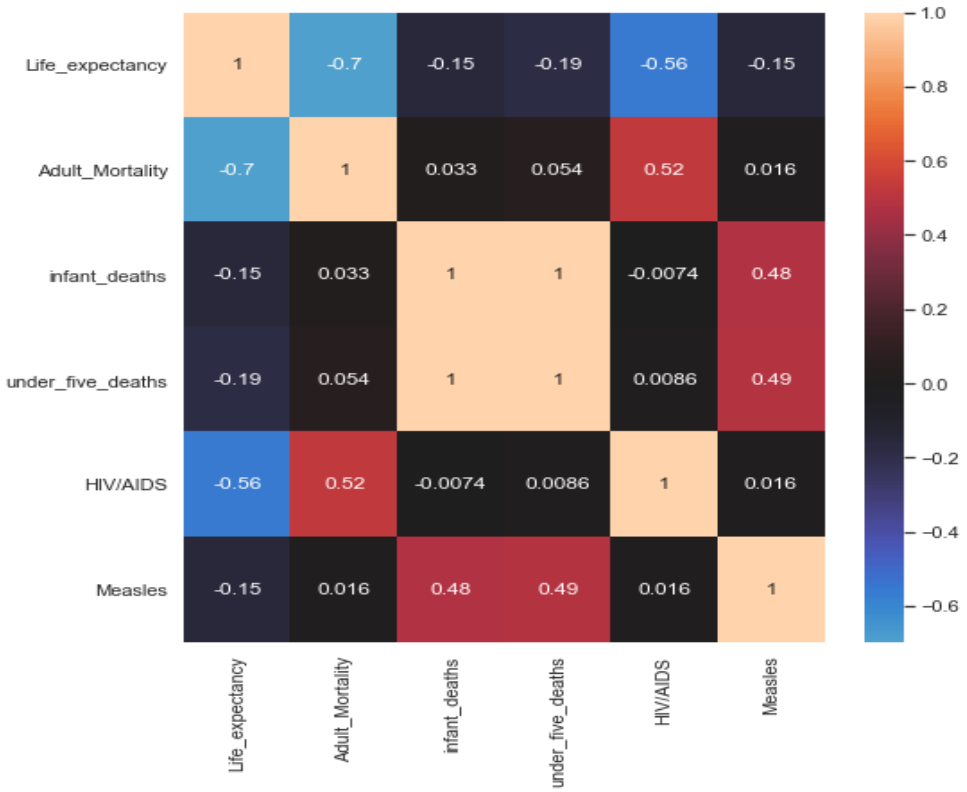


Figure 10. heatmap mortality factors

The main output from this correlation matrix is that mortality factors are all negatively correlated to life expectancy. Adult mortality has the highest correlation score of -0.7 followed by HIV/AIDS.

3.2.3. Social factors

Social factors are:

- Alcohol, recorded per capita (15+) consumption
- Number of years of Schooling(years)
- Average Body Mass Index of entire population
- Prevalence of thinness among children and adolescents for Age 10 to 19 (%)
- Prevalence of thinness among children for Age 5 to 9(%)

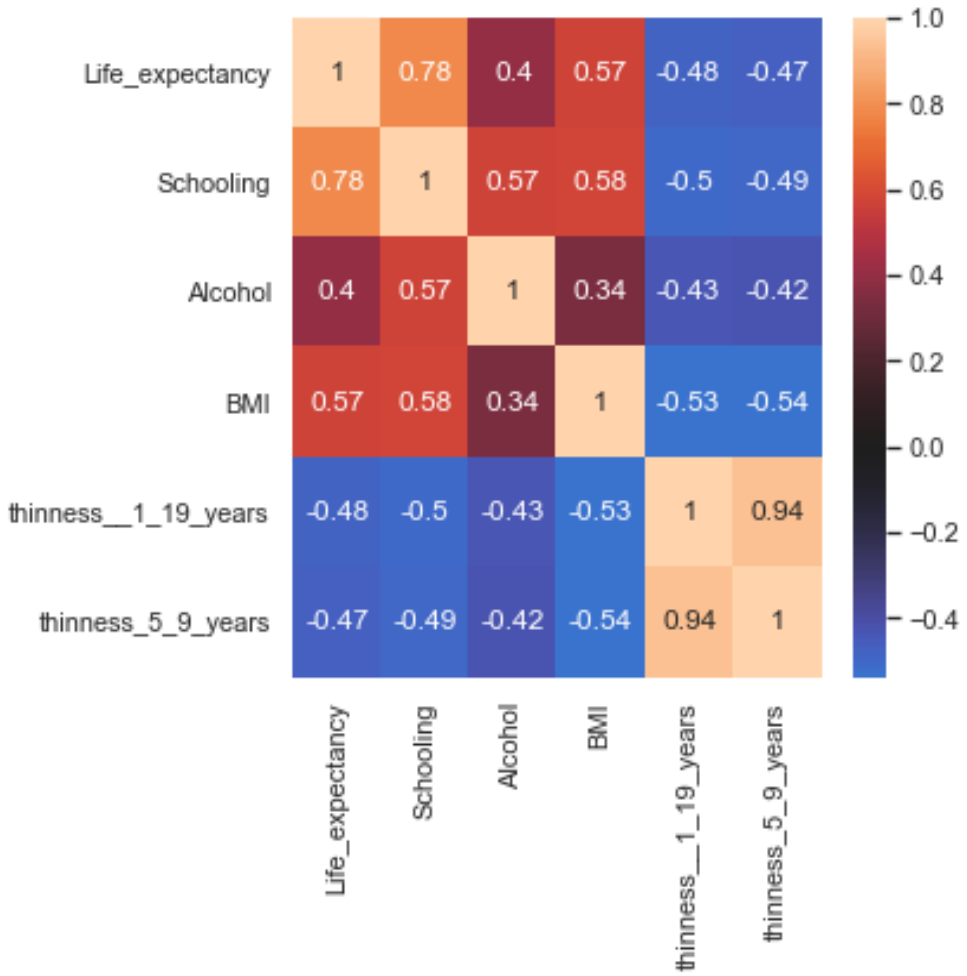


Figure 11. heatmap social factors

From correlation matrix of social factors, schooling has the highest positive correlation score of 0.78 with life expectancy followed by BMI and Alcohol.

Figure 12. correlation life expectancy and schooling

3.2.4 Economical factors

Economic factors are:

- Gross Domestic Product per capita (in USD)
- General government expenditure on health as a percentage of total government expenditure (%)
- Expenditure on health as a percentage of Gross Domestic Product per capita (%)
- Human Development Index in terms of income composition of resources (index ranging from 0 to 1)

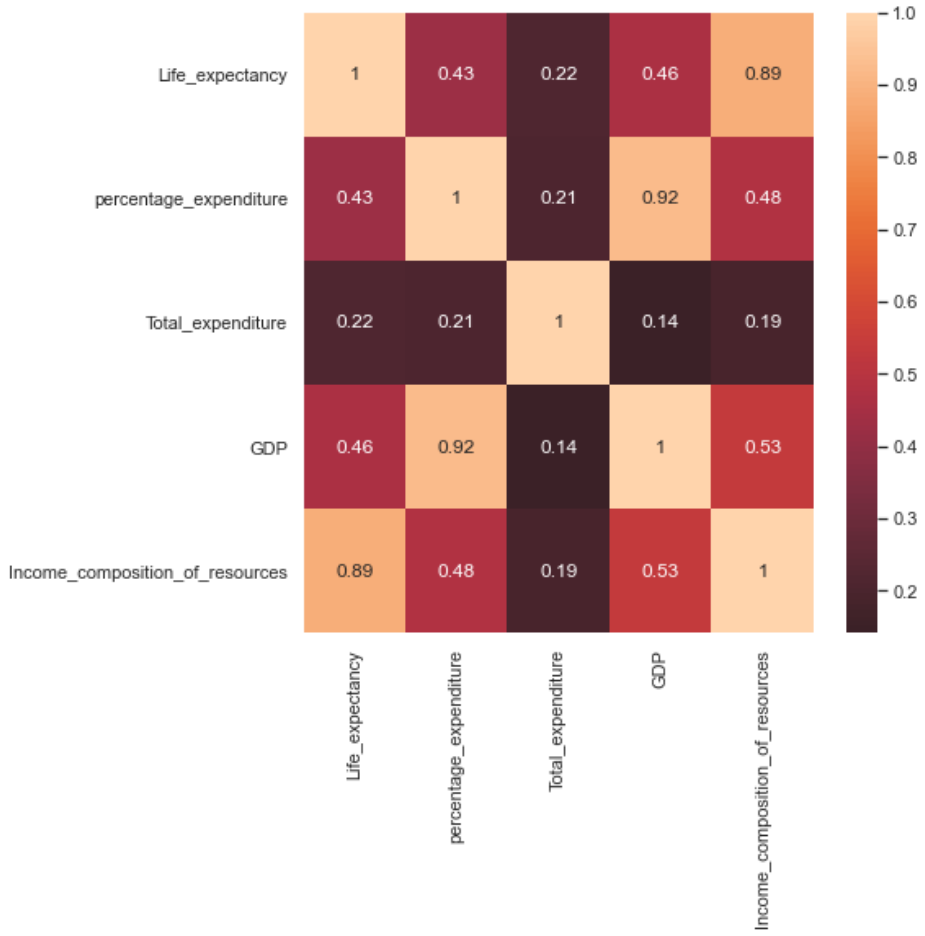


Figure 13. heatmap Economic factors

All the economic factors have positive correlation with life expectancy. The highest factor is income composition of resources followed by GDP and percentage expenditure.

3.2.5 Population and life expectancy

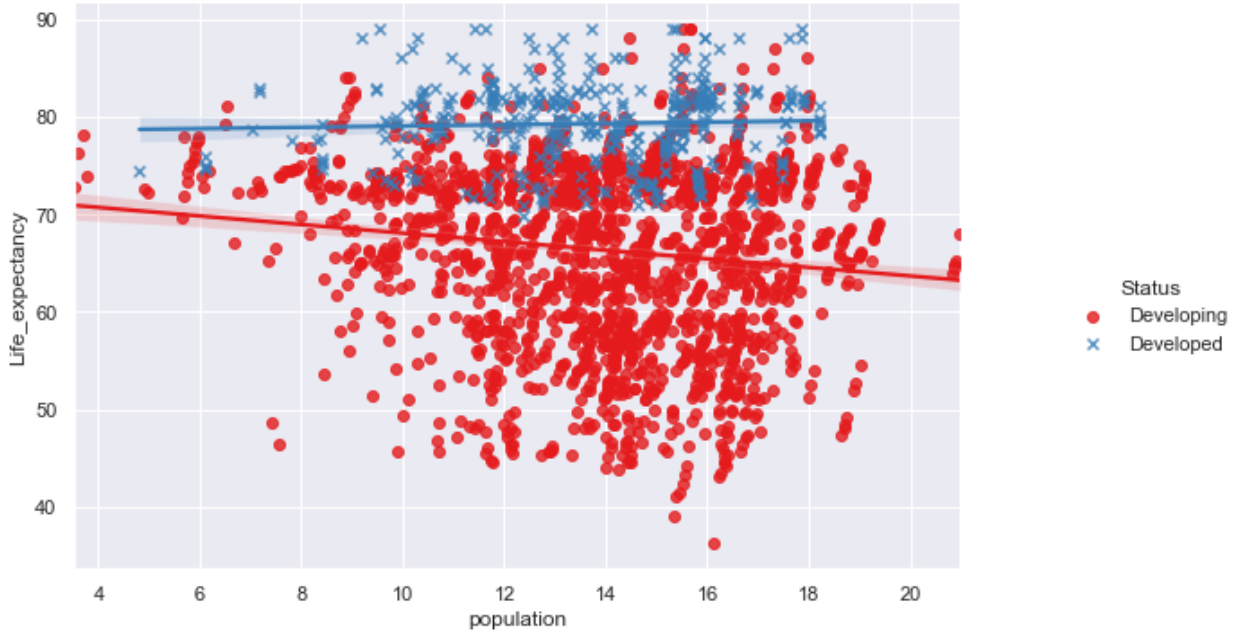


Figure 14. Correlation life expectancy and population

It was important to highlight the importance of population on life expectancy. The figure shows the facts that, the increase of population in developing countries impact negatively the level of life expectancy. This is not the case in developed countries where we observed a slightly increase in life expectancy with the increase of population.

These discrepancies could be explained by the level of economic factors in both developed and developing countries such as GDP and income composition of resources which are lower in developing countries compared to that of developed countries. In fact economic factors have a great impact in the implementation of heath politics.

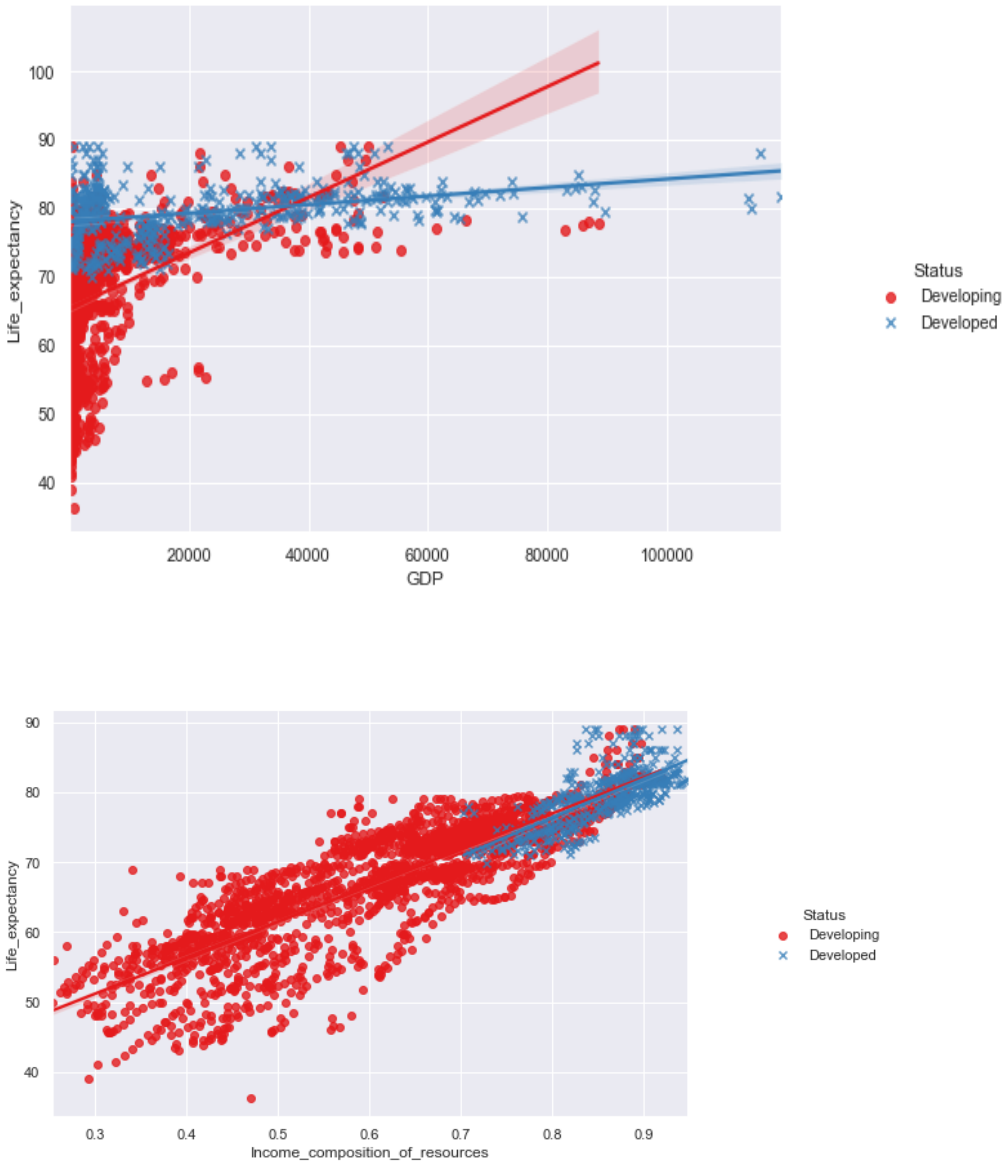


Figure 15. Correlation life expectancy income composition resources and GDP

4. Modeling

4.1 Data Preprocessing

We performed the principal component analysis which reveals that the first five components seem to account for over 75% of the variance, while the first 10 components account for 92% of the variance.

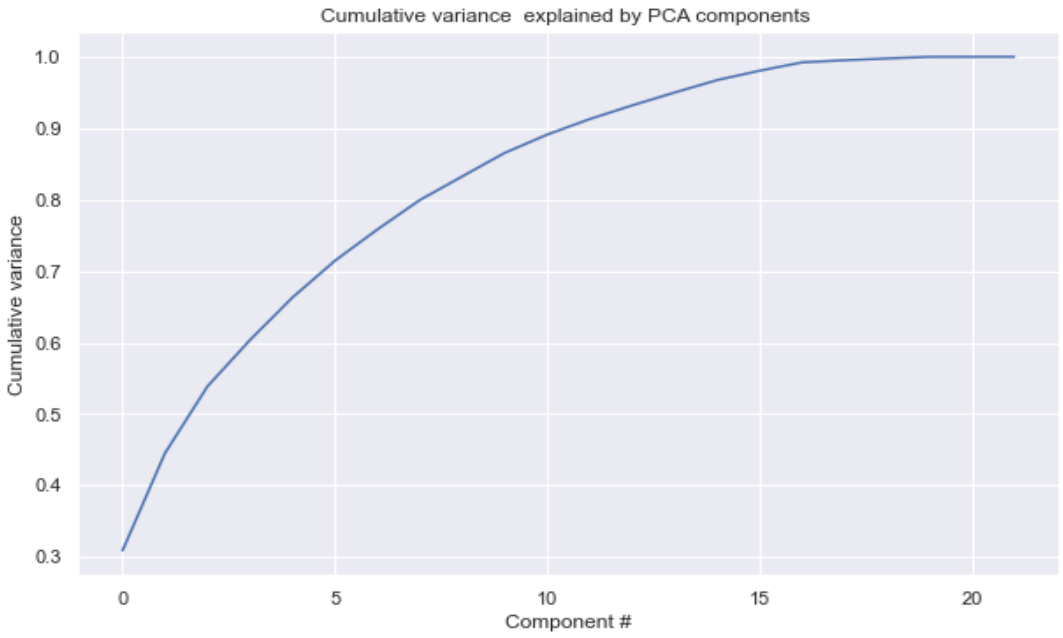


Figure 16. PC features and cumulative variances

The select best K features with linear regression suggested that is k=20. As present in the figure, the variances get stable at 20.

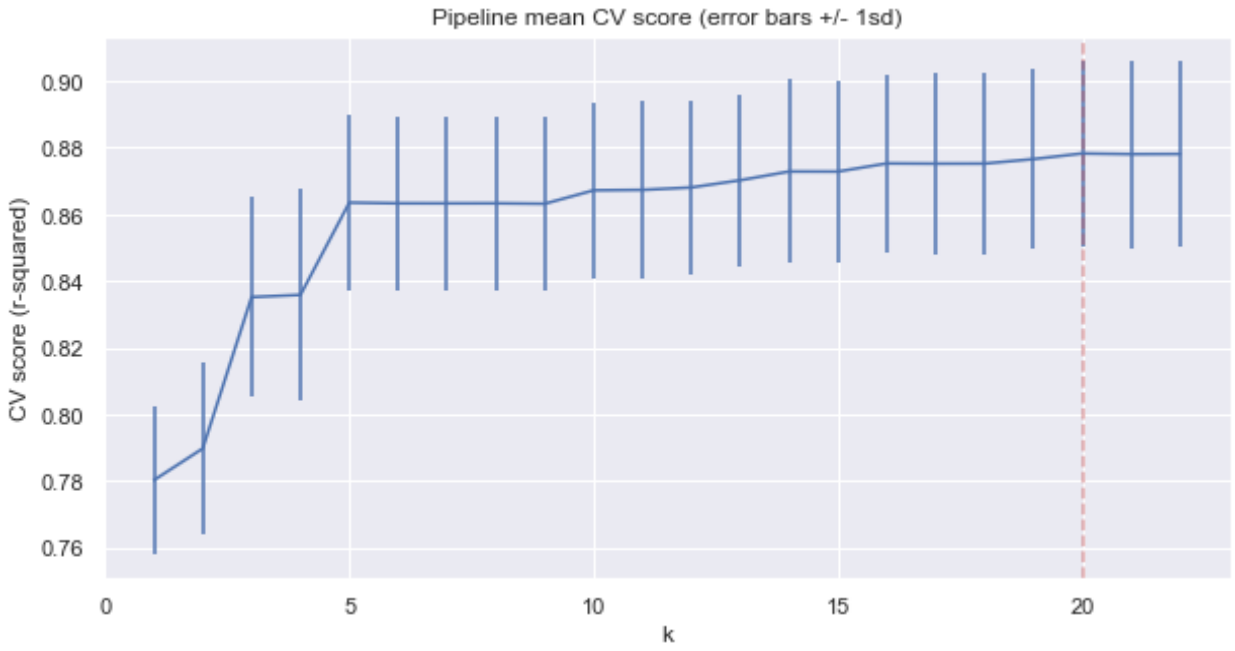


Figure 17. The best K features

The random forest reveals the most important features as presented in the figure below.

Income composition of resources, HIV/AIDS, adult mortality, Schooling is the most important features for the model. However, immunization features (Polio (8^{th}), Diphtheria (13^{th}) and hepatitis B (15^{th}) appear in the list even if it doesn't have high impact.

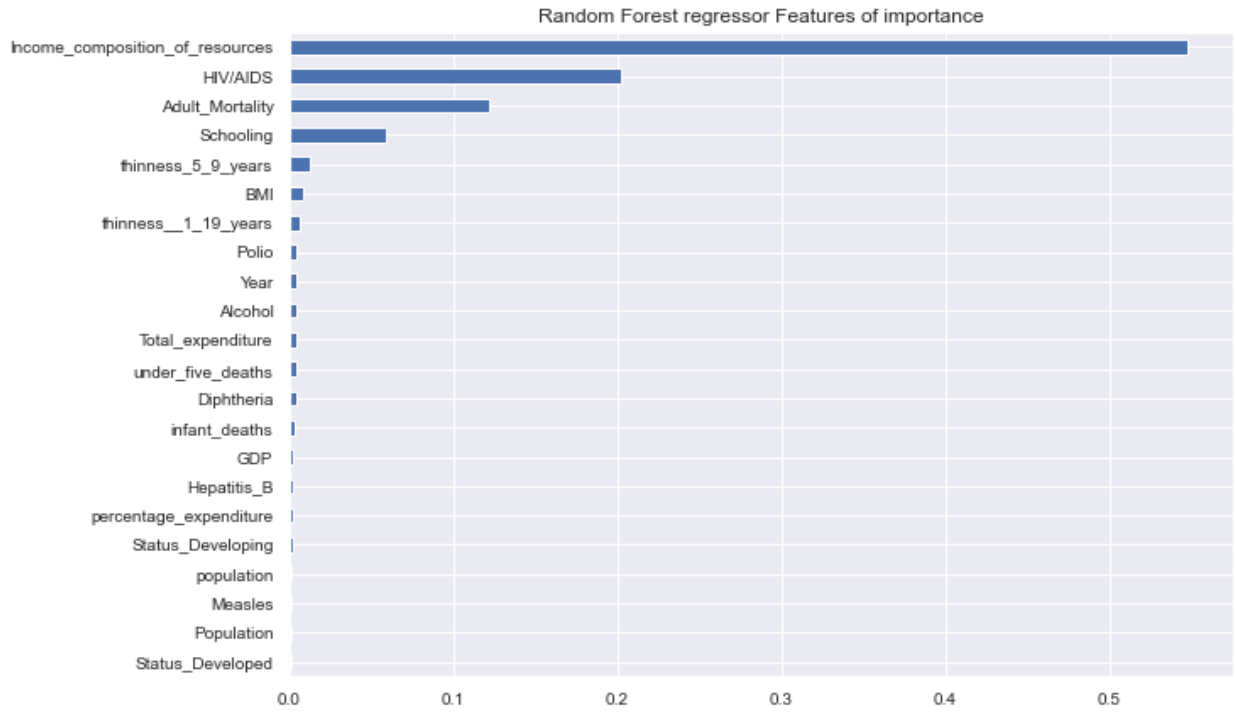


Figure 18. Features of importance with random forest

This result help to understand and explain the fact that increase of population in developing country impact negatively life expectancy. In fact, Income composition resources in the main feature, its correlation with life expectancy reveals that countries with low income tend to have lower life expectancy compared to countries with high income.

4.2 Data preparation or Pipeline definition

Before driving into modeling, we labeled encode the categorical feature (countries status), then the dataset was divided in test sets and training sets (30% and 70%). The pipeline was build following these steps: 1) Imputing missing data, 2) Data scaling, 3) defining the model with corresponding parameters and hyperparameters.

Then we used grid search and random search for hyperparameters tuning, got the best estimators and performed the cross validation technique where we compute the accuracy. From the best estimators, we predict on test set and compute the evaluation on train and test set.

4.3 Modeling

A bunch of models has been developed:

- Regression model: simple Linear regression, linear regression with principal component analysis (PCA), ridge regression, Elastic net regression,
- Tree base model: Decision tree,
- Tree base model - Ensemble model: random forest, Gradient Boosting , XGBoost and voting.

The next table presents the parameters and hyperparameters content of each model.

Table: Model definition

model	**Model_definition with hyperparamters**
linear_reg	Pipeline(steps=[('iterativeimputer', IterativeImputer()), ('standardscaler', StandardScaler()), ('selectkbest', SelectKBest(k=19, score_func=<function f_regression at 0x0000025E60390C10>)), ('linearregression', LinearRegression())])
linear_reg2	Pipeline(steps=[('iterativeimputer', IterativeImputer()), ('standardscaler', StandardScaler()), ('pca', PCA(n_components=19)), ('linearregression', LinearRegression())])
ridge_reg	Pipeline(steps=[('iterativeimputer', IterativeImputer()), ('standardscaler', StandardScaler()), ('ridge', Ridge(alpha=0.5))])
Elastic_net	ElasticNet(alpha=0.0001, l1_ratio=0.4)
decision_tree	DecisionTreeRegressor(max_depth=4, max_features=0.2, min_samples_leaf=0.1, random_state=1)

Table: Model definition

model	model_definition with hyper parameters
random_forest_reg1	RandomForestRegressor(max_depth=4, max_features=0.4, n_estimators=200, n_jobs=-1, random_state=1)
random_forest_reg2	RandomForestRegressor(max_depth=10, max_features=0.4, n_estimators=200, n_jobs=-1, random_state=1)
random_forest_reg3	RandomForestRegressor(max_depth=7, max_features=0.3, n_jobs=-1, random_state=1)
random_forest_reg4	RandomForestRegressor(max_depth=8, max_features=0.6, n_estimators=200, random_state=1)
gradien_boost_1	GradientBoostingRegressor(n_estimators=150, random_state=1)
gradien_boost_2	GradientBoostingRegressor(learning_rate=0.08249999999999999, max_depth=10, max_features=0.6000000000000001, min_samples_leaf=8, min_samples_split=10, n_estimators=118)
gradien_boost_3	GradientBoostingRegressor(learning_rate=0.08249999999999999, max_depth=7, max_features=0.8, min_samples_leaf=4, min_samples_split=12, n_estimators=150, random_state=1)
XGBRegressor	XGBRegressor(base_score=0.5, booster='gbtree', colsample_bylevel=1, colsample_bynode=1, colsample_bytree=1, enable_categorical=False, gamma=0, gpu_id=-1, importance_type=None, interaction_constraints='', learning_rate=0.04, max_delta_step=0, max_depth=5, min_child_weight=1, missing=nan, monotone_constraints='()', n_estimators=200, n_jobs=-1, num_parallel_tree=1, predictor='auto', random_state=0, reg_alpha=0, reg_lambda=1, scale_pos_weight=1, subsample=1, tree_method='exact', validate_parameters=1, verbosity=None)
voting	VotingRegressor(estimators=[('gb', GradientBoostingRegressor(random_state=47)), ('rf', RandomForestRegressor(random_state=47)), ('lr', LinearRegression())])

4.4 Model Evaluation and selection

Models have been evaluated on test and train dataset. Mean absolute error (MAE), Mean squared error (MSE), Root Mean Squared Error (RMSE) and R-squared (R2) are metrics that were used to evaluate the performance of the models developed.

Before jumping into the model selection, few things to keep in mind are: The dataset has 22 independent features, the exploratory data analysis showed that the dataset has outliers.

From the metrics used to evaluate the model, these are some observations that we have to aware about:

- MAE is a metric that summarizes the difference between the predicted and the actual values. It is the mean of absolute error values. It is very robust to outliers. In fact, if one wants to ignore the outlier values to a certain degree, MAE is the choice since it reduces the penalty of the outliers significantly with the removal of the square terms.
- MSE is an absolute measure of the goodness of lit. It is very sensitive to outliers and will show a very high error value even if a few outliers are present.
- RMSE is the root of MSE and is beneficial because it helps to bring down the scale of the errors closer to the actual values, making it more interpretable.
- R-squared helps to identify the proportion of variance of the target variable that can be captured with the help of the independent variables or predictors. R2 increases with any new feature addition. If there are too many independent variables, the model can overfit, performing well on train data and will perform poorly on test data.

The tables below summarize the output of model's evaluations.

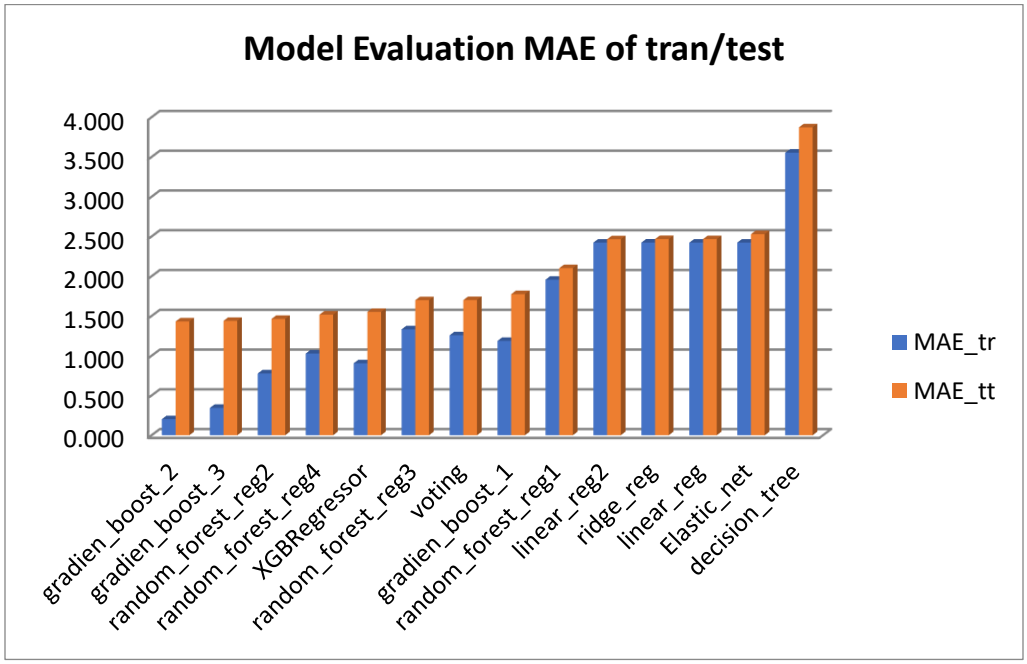
Table 1: Model evaluation on train set

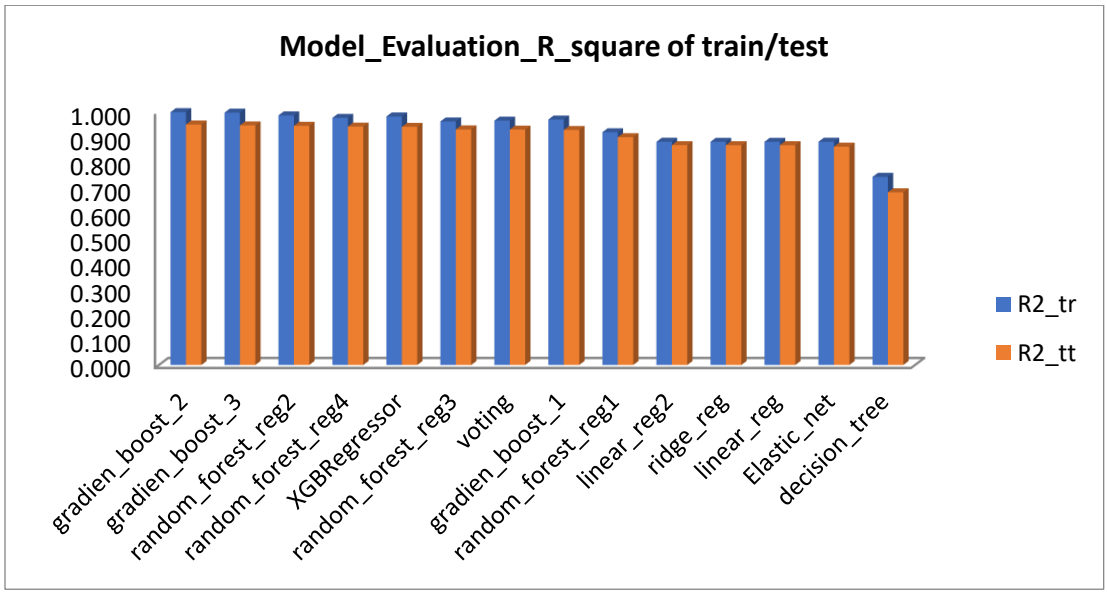
model	MAE_tr	MSE_tr	RMSE_tr	R2_tr
gradien_boost_2	0.202	0.115	0.340	0.999
gradien_boost_3	0.344	0.279	0.528	0.997
random_forest_reg2	0.777	1.267	1.126	0.986
random_forest_reg4	1.027	2.167	1.472	0.976
XGBRegressor	0.905	1.669	1.292	0.982
random_forest_reg3	1.332	3.436	1.854	0.962
Voting	1.257	3.140	1.772	0.965
gradien_boost_1	1.186	2.750	1.658	0.970
random_forest_reg1	1.955	7.328	2.707	0.919
linear_reg2	2.422	10.749	3.279	0.882
ridge_reg	2.423	10.750	3.279	0.882
linear_reg	2.422	10.751	3.279	0.882
Elastic_net	2.422	10.749	3.279	0.882
decision_tree	3.554	23.350	4.832	0.743

Table 1: Model evaluation on test set

model	MAE_tt	MSE_tr	RMSE_tt	R2_tt
gradien_boost_2	1.431	0.115	2.116	0.950
gradien_boost_3	1.438	0.279	2.188	0.947
random_forest_reg2	1.462	1.267	2.220	0.945
random_forest_reg4	1.517	2.167	2.282	0.942
XGBRegressor	1.551	1.669	2.303	0.941
random_forest_reg3	1.698	3.436	2.491	0.931
voting	1.699	3.140	2.503	0.930
gradien_boost_1	1.774	2.750	2.528	0.929
random_forest_reg1	2.101	7.328	2.993	0.900
linear_reg2	2.465	10.749	3.432	0.868
ridge_reg	2.467	10.750	3.434	0.868
linear_reg	2.466	10.751	3.434	0.868
Elastic_net	2.528	10.749	3.507	0.863
decision_tree	3.871	23.350	5.332	0.682

We wanted to compare models and got the one that best predict life expectancy. The dataset is influence by outliers and according to the strength and weakness of each metric to model evaluation and selection; MAE was the best metric that helped to identify the best model. And for explanatory purpose on how the independent variables explained the variability on dependent variable, R square was the second metric chosen. From this perspective, the figure below heighted the best model with a nice visualization.





The 2 figures showed the histogram of model evaluation score on test set and train set.

In fact, the purpose of this project is a model that performs well on unseen data, the ordering of metrics highlighted Gadrien_boost_2 model as the best model with MAE of 0.202 on train set and 1.431 on test set, R square is 0.94 on test set.

5. Conclusion

- Life expectancy has increased over years in both developed and developing countries
- The mean average of the life expectancy of developed countries is generally higher compared to that of developing countries
- However, the ratio of LE over the decade of 2005 to 2015 showed that life expectancy in developing countries has greatly increased.
- It has been highlighted that immunization has impacted the improvement of life expectancy in a developing country, as well as the reduction in infant deaths.
- The analysis revealed that economic factors play an important role in the system, it is why countries with higher income resources and GDP tend to have high life expectancy even if the population is big. In developing countries, an increase in the population tends to impact negatively life expectancy.
- Many (14) regression models have been developed to predict life expectancy, the chosen one is Gradient boost with MAE of 0.202 on train set and 1.431 on the test set, R square is 0.94 on the test set.