Computer Science and Interdisciplinary Research Journal
Volume 1, Issue 1, ISSN: 3033-1218, DOI: 10.70862/CSIR.2024.0101-04, 2024

# Impact of Global Country Indicators on Life Expectancy

Valentin Georgiev[1], Stanka Hadzhikoleva[2], Emil Hadzhikolev[3]

University of Plovdiv Paisii Hilendarski, 236, Bulgaria Blvd, 4027 Plovdiv, Bulgaria

[1] valentin_georgiev02@abv.bg, [2] stankah@uni-plovdiv.bg, [3] hadjikolev@uni-plovdiv.bg

***Abstract*** – Investigating the factors that influence life expectancy enables government institutions to identify measures to improve public health, pension planning, health services, etc. Understanding these factors would enable the development and implementation of policies to improve the quality of life. This article presents a study that analyzes data from 195 countries covering demographic, economic, health and educational factors to identify those that influence life expectancy. Various models were experimented with, including linear regression, random forest, gradient boosting, SVM, and artificial neural networks. Different machine learning algorithms, feature combinations, and strategies for handling missing data set values were tested in building the models. The best model uses linear regression. Reproductive characteristics, the number of individuals enrolled in higher education, and the number of general practitioners were found to have the highest correlations with life expectancy. An interesting finding is that out-of-pocket health costs were not the leading factor in determining life expectancy.

***Keywords*** – life expectancy, linear regression, random forest, gradient boosting, artificial neural networks, SVM.

**DOI:** https://doi.org/10.70862/CSIR.2024.0101-04

**Corresponding author:** Emil Hadzhikolev,
University of Plovdiv Paisii Hilendarski
**Email:** hadjikolev@uni-plovdiv.bg

*Received: 4 August 2024.*
*Revised: 24 August 2024.*
*Accepted: 25 August 2024.*
*Published: 27 August 2024.*

© 2024 Valentin Georgiev, Stanka Hadzhikoleva & Emil Hadzhikolev. This work is licensed under the Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License.

The article is published with Open Access at http://csirjournal.com/

## 1. Introduction

Life expectancy plays an essential role in the social and economic development of the countries. Life expectancy statistics assist health institutions in creating policies and programs for disease prevention and treatment. This information is needed by governments for demographic planning, including forecasting future needs for health and social services, formulating social security policies, and managing pension systems. High life expectancy is also an indicator of high economic well-being, which provides better living conditions and a functioning health system with access to quality health services. To improve the quality of life, it is necessary to study the factors that influence life expectancy and to create effective policies and laws.

Some of the most significant global organizations conducting research and projects in the field of health are the ***WHO (World Health Organization)*** and the ***IHME (Institute for Health Metrics and Evaluation)***. The WHO leads efforts to improve health globally. It collects and analyzes health data, and develops policies and recommendations for better health [1]. The IHME analyze different health indicators, including the life expectancy [2]. It operates the ***Global Burden of Disease (GBD)*** project, whose goal is to study the health of populations worldwide. This includes quantifying health losses across different locations and over time, to improve health systems and eliminate inequalities. The project is significant and representative, it involves over 12,000 researchers from more than 160 countries [3].

The ***World Bank***, in addition to economic financing, also finances projects and initiatives for improving health. It periodically releases studies and findings on the health status and economic development of the countries [4]. Another global organization concerned with the health and improvement of the quality of life of children worldwide is ***UNICEF***. Its efforts to reduce child

mortality include numerous activities like vaccination, nutrition, and more [5].

The work of ***The Global Fund*** also influences higher life expectancy by financing programs to combat AIDS, tuberculosis, and malaria, having saved 59 million lives over the past 20 years [6]. ***The United Nations Development Programme (UNDP)*** operates a wide range of projects focused on socio-economic development, including health projects that directly impact life expectancy [7]. Current research, statistics, and analyses concerning the health of people in the European Union are published by ***Eurostat*** [8].

Life expectancy is a complex indicator that is influenced by different groups of factors, such as:

- ***healthcare***, which includes access to quality medical services, diagnostics, prevention, vaccination programs, and hereditary diseases;
- ***socio-economic conditions***, such as income levels, education, social protection, working conditions, and workplace safety;
- ***lifestyle factors***, involving dietary habits, physical activity, smoking, alcohol and drug use, and stress levels;
- ***environment***, which covers soil, air, and water pollution, the availability of green spaces, housing conditions, and more.

Numerous studies in the scientific literature explore the relationships between these specific factors and life expectancy across various countries.

Lelieveld et al. found that ***air pollution*** is a factor that leads to lower life expectancy, notably in cases of existing cardiovascular diseases. They conclude that air pollution is comparable to ***smoking*** when it comes to their impact on life expectancy. The global average loss of life expectancy due to air pollution significantly exceeds that from ***all forms of violence*** combined, with losses of 2.9 years and 0.3 years, respectively. The share of preventable loss of life expectancy from anthropogenic air pollution, often attributed to the use of fossil fuels, is nearly two-thirds globally and up to around 80% in high-income countries [9].

The ***Energy Policy Institute at the University of Chicago***, in their research found that air pollution is the greatest external threat to human life expectancy. According to their findings, consistently reducing global ***PM2.5*** particulate matter pollution to WHO-recommended levels could add 2.3 years to the average human lifespan. The impact of PM2.5 on global life expectancy is more than three times greater than ***alcohol consumption and unsafe water***, more than five times greater than ***transport-related injuries***, and more than seven times greater than ***HIV*** [10].

In this direction is the research of Richard Burnetta and his colleagues. They created a PM2.5-mortality hazard ratio function based only on cohort studies of outdoor air pollution, covering the global exposure range. Their study indicates that PM2.5 exposure may be linked to more causes of death than the five considered by the GBD and that incorporation of risk information from other, ***non-outdoor particle sources*** leads to the underestimation of disease burden, especially at higher concentrations [11].

Jackowska et al. investigated the impact of ***lifestyle*** on life expectancy [12]. They developed a multivariate logistic model to comprehensively assess the influence of lifestyle on mortality among young and middle-aged individuals. The model used predictors such as smoking, diet, physical activity, and body mass index, while also considering the presence of chronic diseases like diabetes, arrhythmia, coronary heart disease, myocardial infarction, stroke, and malignant neoplasms. The study used data from 38,481 participants over the age of 20. The findings highlight that lifestyle is a key factor in determining life expectancy. Premature mortality can be effectively prevented through changes in daily habits, even for healthy, young individuals. The conclusion that lifestyle is both the most controllable and influential factor affecting overall health and disease resilience is supported by other studies [13]. A healthy diet, moderate and regular physical exercise, and sufficient high-quality sleep form the foundation of a healthy lifestyle and have a positive impact on life expectancy.

Fadnes et al. have conducted various studies on how ***diet*** affects life expectancy. They researched how the consistent consumption of different types of food affects longevity. The findings show that maintaining dietary changes can bring significant health benefits to people of all ages, with greater benefits observed when changes are implemented earlier in life. The researchers developed an online calculator, Food4HealthyLife, allowing individuals to explore the effects of their dietary choices on health [14]. In another study, they found that a long-term shift towards healthier eating habits could increase life expectancy by up to 10 years in the United Kingdom. The greatest benefits were achieved by consuming more whole grains, nuts, and fruits, and fewer sugary drinks and processed meats [15]. Another of their study, involving participants from China, France, Germany, Iran, Norway, the United Kingdom, and the United States aimed to evaluate how life expectancy could be extended by switching from typical national dietary patterns to diets optimized for longevity. This research considered changes in dietary habits along with height, weight, and physical activity. Different potential life expectancy values were identified for different countries, with the most significant benefits from dietary changes found in the United States [16].

Roesch et al. conducted a study on ***the differences in life expectancy between black and white populations*** in three American cities. They examined 26 categories of causes of death and found that the primary factors contributing to mortality were circulatory diseases, cancer, injuries, and diabetes, with varying significance and impact across different cities [17]. Roberts et al. also researched this topic, focusing on residents of Washington, D.C. They found that heart disease, homicides, and cancer were the most significant contributors to the life expectancy gap between black and white men. For women, the most critical factors were heart disease, cancer, and unintentional injuries [18].

The GBD study conducted in 2021 found that due to the ***COVID-19 pandemic***, life expectancy decreased by an average of 1.6 years between 2019 and 2021. The negative effects on health were influenced by accompanying non-communicable diseases, high blood sugar, drug use, and obesity.

Despite the challenges posed by current global threats, the authors project that global life expectancy will increase by 4.6 years over the next three decades [19].

Studies like those presented focus on factors directly influencing life expectancy, relying on data collected for numerous individuals. These studies provide insights into average life expectancy based on lifestyle, genetic factors, socio-economic conditions, environmental factors, and more. The primary objective of our research is to analyze the factors affecting life expectancy. Models were created using linear regression, random forest, gradient boosting, SVM, and artificial neural networks [20]. The Orange Data Mining software application was utilized [21], allowing users to create data analysis workflows by connecting various components, called "widgets", in a graphical user interface. The widgets enable visualization of data, statistical analyses, and model building. The Orange application supports multiple machine learning algorithms suitable for classification, regression, clustering, and association tasks.

## 2. Materials and methods

In the current study, a dataset containing 195 records with a total of 35 features for all countries in the world for the year 2023 was used [22]. It includes demographics and economic indicators, environmental factors, health and education statistics, and more. The data is collected from multiple sources, including the WHO and the World Bank. The features are presented in *Appendix A. Dataset characteristics*.

Some of the features in the dataset were not used in the models, as they were assessed to have no significant impact on life expectancy or were highly correlated with each other, as determined during the study.

In the initial phase of the study, the correlation between life expectancy and other features was examined. We used both Pearson's and Spearman's correlation coefficients to capture different aspects of the relationships between factors and *Life Expectancy*. While Pearson's correlation coefficient is suitable for detecting only linear relationships, Spearman's correlation coefficient is used to detect relationships that are linear or non-linear.

The ten features with the highest Pearson and Spearman correlations with life expectancy are presented in Tables 1 and 2, respectively. A high negative correlation was identified for factors related to reproduction rates, such as Birth Rate, Fertility Rate, Maternal Mortality Ratio, and Infant Mortality. A relatively high positive correlation was observed for Gross Tertiary Education Enrollment (%) and Physicians per Thousand. This is likely due to the trend in developed countries, where declining birth rates coincide with greater access to education and higher education opportunities. Additionally, a higher standard of living in these countries is associated with a greater availability of doctors, contributing to better overall health. In developing countries, the opposite trends are observed.

**Table 1.** Features with the highest Pearson correlation with Life Expectancy

| Feature | Pearson Correlation coefficient |
|---|---|
| Infant Mortality | -0.915 |
| Birth Rate | -0.866 |
| Fertility Rate | -0.846 |
| Maternal Mortality Ratio | -0.817 |
| Gross Tertiary Education Enrollment (%) | +0.714 |
| Physicians per Thousand | +0.676 |
| Minimum Wage | +0.496 |
| Tax Revenue (%) | +0.340 |
| Out of Pocket Health Expenditure | -0.315 |
| Agricultural Land (%) | -0.241 |

The correlation for *Minimum Wage, Tax Revenue (%), GDP, CO2 Emissions*, and *CPI* is relatively weak, which can again be explained by the differences between developed and developing countries. There is minimal correlation for *Out of Pocket Health Expenditure* (Spearman's correlation is -0.316), leading to the conclusion that higher out-of-pocket payments by patients for healthcare do not necessarily impact life expectancy.

**Table 2.** Features with the highest Spearman Correlation with Life Expectancy

| Feature | Spearman Correlation coefficient |
|---|---|
| Infant Mortality | -0.904 |
| Maternal Mortality Ratio | -0.857 |
| Birth Rate | -0.837 |
| Fertility Rate | -0.837 |
| Physicians per Thousand | +0.784 |
| Gross Tertiary Education Enrollment (%) | +0.767 |
| Minimum Wage | +0.583 |
| GDP | +0.493 |
| Co2-Emissions | +0.442 |
| CPI | -0.417 |

For conducting experiments with artificial intelligence methods, *Life Expectancy* was selected as the target feature. The data preparation process involved several steps:

- Excluding irrelevant features from the initial dataset;
- Addressing the issue of missing values;
- Scaling the data.

Initially, certain features that were not directly related to *Life Expectancy* were removed from the original dataset: *Country, Abbreviation, Armed Forces Size, Calling Code, Capital/Major City, Currency Code, Gasoline Price, Largest City*, and *Official Language*. Geographic location, described by *Latitude* and *Longitude*, was also excluded to prevent the models from overfitting to specific locations. The selection of features in this study was guided by a focus on variables that directly impact life expectancy and are widely recognized in existing literature. Geographic location, while potentially influential due to its connection to climatic conditions, was excluded to prevent the models from overfitting to specific regional characteristics. Including geographic location could have introduced multicollinearity issues, particularly since many of the socioeconomic and healthcare-related features in the dataset are themselves influenced by geographic factors. Furthermore, the inclusion of latitude and longitude might have led to the models capturing specific location-based effects rather than generalizable patterns applicable across diverse contexts.

In the initial stage, high correlations were identified between the following features:

- *Fertility Rate* and *Infant Mortality*, *Birth Rate, Maternal Mortality Ratio*
- *Population* and *Urban Population (%), CO2* Emissions
- *Consumer Price Index (CPI)* and *CPI Change (%)*

Therefore, the final dataset included *Fertility Rate, Population*, and *CPI*.

As a result, in the initial stage, the dataset was organized such that the characteristics could be grouped as follows:

- ***Demographic and Geographic Data***: *Land Area (Km2); Density (P/Km2); Population; Agricultural Land (%)*.
- ***Economic Data***: *GDP; CPI (Consumer Price Index); Minimum Wage; Tax Revenue (%); Total Tax Rate (%); Unemployment Rate (%); Labor Force Participation (%)*.
- ***Healthcare***: *Fertility Rate; Out of Pocket Health Expenditure (%); Physicians per Thousand*.
- ***Education***: *Gross Primary Education Enrollment (%); Gross Tertiary Education Enrollment (%)*.

In the second stage, a significant amount of missing data was identified for *Minimum Wage* (45) and *Tax Revenue* (%) (23). The importance of these features was assessed using Univariate Regression and RReliefF, both of which yielded relatively high values for these characteristics. To determine whether the absence of these data points would affect the models, experiments were conducted with input datasets that both included and excluded these two characteristics.

Considering the diversity of characteristics and their significance, the following approach was adopted to handle missing data: two separate models were created. In one model, records with missing values were removed, while in the other, the missing values were imputed with the mean value of the corresponding characteristic.

In the third stage of preparing the experimental dataset, the values were scaled to the range of [-1, 1] using Min-Max normalization. The main advantages of scaling include improving model performance, enhancing the efficiency of algorithms with similar scales, preventing any single characteristic from dominating, and facilitating the interpretation of results [23].

A part of the data preprocessing is illustrated in Figure 1.

To test the SVM method, an additional categorical feature – *Life Expectancy (categorical)* – was created by dividing *Life Expectancy* into four categories. Given that *Life Expectancy* ranges from $[52.8, 85.4]$, the following categories were defined:

- short: $x \in [52.8, 60)$;
- medium_l: $x \in [60, 70)$;
- medium_h: $x \in [70, 80)$,
- long: $x \in [80, 85.4]$.

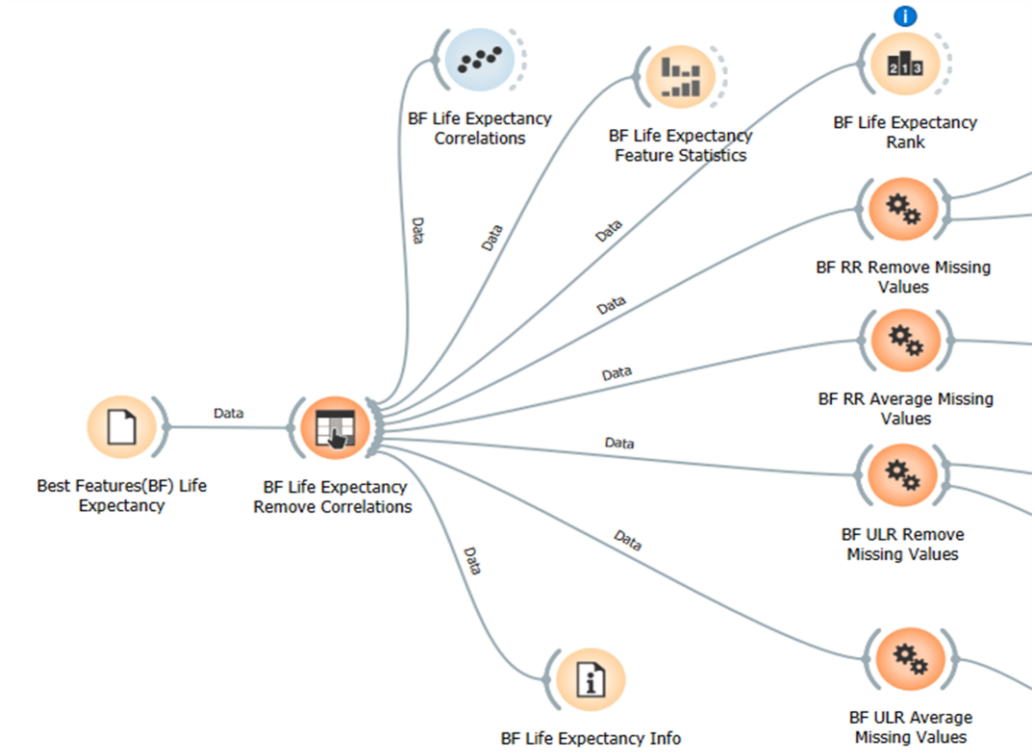



**Figure 1**. Data preprocessing of selected features in Orange

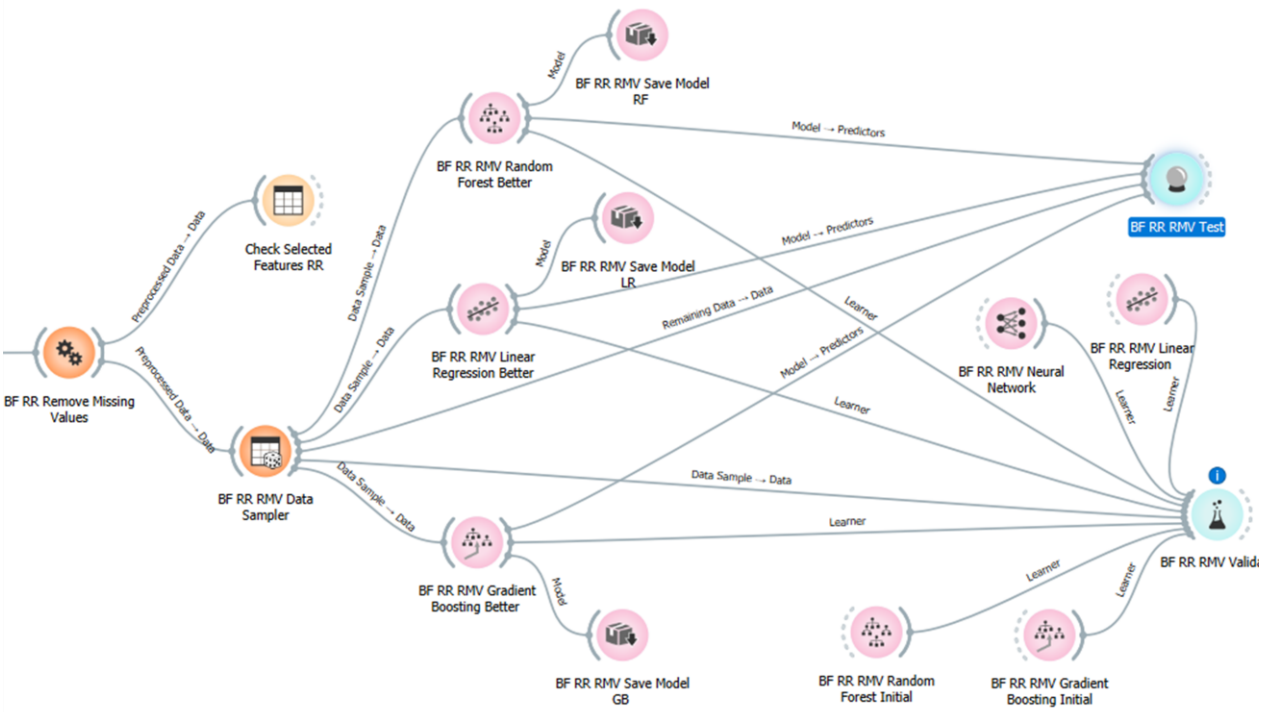



**Figure 2**. Diagram of the forecasting process with parts of the created models

To train the models, the data was split into two parts: training and validation (80%) and testing data (20%). We used cross-validation to select appropriate models, dividing the validation data into five non-overlapping subsets. The models were trained on all combinations of four subsets, and the predictions were then compared with the fifth subset. The accuracy of the created models was evaluated using the following metrics:

- MSE $= \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2$ (Mean Squared Error);
- RMSE $= \sqrt{MSE}$ (Root Mean Squared Error);
- MAE $= \frac{1}{n}\sum_{i=1}^{n}|y_i - \hat{y}_i|$ (Mean Absolute Error);
- MAPE $= \frac{1}{n}\sum_{i=1}^{n}\left|\frac{y_i - \hat{y}_i}{y_i}\right|$ (Mean Absolute Percentage Error);
- $R^2 = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2}$ (Coefficient of Determination),

where n is the total number of samples, $y_i$ is the actual value, $\hat{y}_i$ is the predicted value, and $\bar{y}$ is the mean of the target variable.

## 3. Experiments

Initial tests showed that Linear Regression, Random Forest, and Gradient Boosting outperformed SVM and Artificial Neural Networks (ANNs), which may be due to the small number of records in the dataset. Although SVM and ANN are powerful tools, they often require larger datasets to function effectively [24]. On the other hand, Linear Regression, Random Forest, and Gradient Boosting tend to perform better with smaller datasets. Due to the poor performance of SVM and ANNs, they were excluded from further experiments.

Using the three selected methods – Linear Regression, Random Forest, and Gradient Boosting – 24 models were developed with various features and approaches to handling missing data. Depending on the chosen features, the models were as follows:

- ***Models using the most relevant features according to RReliefF***: Minimum Wage, Labor Force Participation, Unemployment Rate (%), Agricultural Land (%), Tax Revenue (%), Gross Tertiary Education Enrollment (%), Fertility Rate, Forested Area (%), Out of Pocket Health Expenditure;
- ***Models using the most relevant features according to Univariate Linear Regression***: Fertility Rate, Gross Tertiary Education Enrollment (%), Physicians per Thousand, Minimum Wage, Tax Revenue (%), Out of Pocket Health Expenditure, Agricultural Land (%), Total Tax Rate (%), CPI;
- ***Models using all features***;
- ***Models using all features except Minimum Wage and Tax Revenue (%)***, which had many missing values in the original dataset.

To address the issue of missing data, two approaches were chosen: removing records or substituting with the mean value of the respective feature.

A part of the experimental model for "BF RR remove missing values" is shown in Figure 2.

Figure 3 presents the results of the evaluation of three models – *Linear Regression*, *Random Forest*, and *Gradient Boosting* – after removing the features *Minimum Wage* and *Tax Revenue* (%), as well as records with missing values. The x-axis shows the actual *Life Expectancy*, and the y-axis represents the error of each model in predicting this value. It is observed that Linear Regression performs the best in this case, with the smallest extreme values of error.

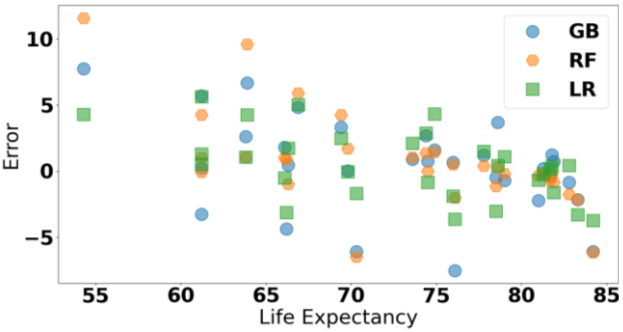



**Figure 3**. Results of tests on Gradient Boosting (GB), Random Forest (RF), and Linear Regression (LR)

Table 3 presents the results from ten of the most effective models.

**Table 3.** Selected effective models for predicting Life Expectancy

| № | Model | Features | Missing values | MSE | RMSE | MAE | MAPE | $R^2$ |
|---|---|---|---|---|---|---|---|---|
| 1 | LR | R2F | RMV | 6.718 | 2.592 | 2.061 | 0.029 | 0.895 |
| 2 | RF | RR | RMV | 8.242 | 2.871 | 2,283 | 0.033 | 0.877 |
| 3 | LR | RR | RMV | 8.764 | 2.960 | 2.358 | 0.034 | 0.870 |
| 4 | GB | RR | RMV | 9.446 | 3.073 | 2.403 | 0.036 | 0.859 |
| 5 | LR | ULR | RMV | 9.827 | 3.135 | 2.505 | 0.037 | 0.858 |
| 6 | RF | ULR | RMV | 10.018 | 3.165 | 2.435 | 0.037 | 0.855 |
| 7 | RF | ULR | AMV | 8.284 | 2.878 | 2.289 | 0.032 | 0.780 |
| 8 | LR | ALL | RMV | 15.981 | 3.998 | 2.991 | 0.045 | 0.724 |
| 9 | RF | RR | AMV | 12.216 | 3.495 | 2.799 | 0.039 | 0.676 |
| 10 | GB | RR | AMV | 12.436 | 3.526 | 2.718 | 0.037 | 0.670 |

The following abbreviations are used

- ***Model*:** LR – Linear Regression; RF – Random Forest; GB – Gradient Boosting.
- ***Features***: RR – Top 10 features selected with RReliefF; ULR – Top 10 features selected with Univariate Linear Regression; ALL – All features used; R2F – All features used except Minimum Wage and Tax Revenue (%);
- ***Missing Values***: RMV – Records with missing values removed; AMV – Missing values filled with mean values of the respective feature.

The parameters of the created models are presented in Tables 4, 5, and 6.

**Table 4.** Parameters of the Linear Regression Models

| Model № | L1 (Lasso) α |
|---|---|
| 1 | 0.03 |
| 3 | 0.3 |
| 5 | 0.3 |
| 8 | 0.1 |

**Table 5.** Parameters of the Random Forest Models

| Model № | Number of trees | Number of attributes at each split | Maximum depth of the trees | Minimum number of instances in leaves |
|---|---|---|---|---|
| 2 | 500 | 5 | 7 | 5 |
| 6 | 425 | 3 | 10 | 5 |
| 7 | 225 | 3 | 11 | 2 |
| 9 | 150 | 5 | 9 | 5 |

**Table 6.** Parameters of the Gradient Boosting Models

| Model № | Number of trees | Learning rate | Maximum depth of the trees | Minimum number of samples per leaf | Fraction of training instances |
|---|---|---|---|---|---|
| 4 | 450 | 0.075 | 2 | 3 | 0.75 |
| 10 | 450 | 0.075 | 2 | 4 | 0.75 |

Models 1 through 6 are the best according to the coefficient of determination $R^2$, with three of them being linear regression models. It is notable that the most effective models work with datasets from which records with missing values for *Minimum Wage* and *Tax Revenue (%)* have been removed. Clearly, substituting missing values with mean values is not a good idea. This can be explained by the inherent complexity of the features, which makes the mean values unreliable substitutes. For Model 8, where records with missing values have been removed, the poor performance is likely due to the loss of data (almost 25% of the original dataset). Its alternative version (Model 1), which includes all records except for the two problematic features, is the best-performing model.

Selection of the best features using RReliefF leads to reliable predictive models that rank from 2nd to 4th place. Selection using Univariate Linear Regression yields slightly weaker results but still has high accuracy.

It should be noted that even the least effective models have an average error of 5%, which is interpreted as less than 3 years, while the best models have an average error of approximately 3.5%, which is 2.2 years.

## 4. Conclusion

The conducted study aimed to identify the factors affecting life expectancy. It was found that reproductive characteristics, the number of people with higher education, and the number of general practitioners have the highest correlation with life expectancy. A surprising discovery was that out-of-pocket health expenses do not significantly impact life expectancy.

Various models were experimented with, including linear regression, random forest, gradient boosting, SVM, and ANNs. Better results were achieved with linear regression, random forest, and gradient boosting. Due to the relatively small dataset, SVM and ANNs did not achieve sufficiently good accuracy. Different machine learning algorithms, feature combinations, and strategies for handling missing data were tested when building the models. The best model used linear regression on a dataset where records with missing values were removed. Selection of the best features, determined by RReliefF and Univariate Linear Regression, also led to satisfactory results. Substituting missing values with mean values resulted in less effective models. The most effective models have an average error of approximately 3.5%, which corresponds to 2.2 years in life expectancy.

## Appendix A. Dataset characteristics

- ***Country***: Name of the country.
- ***Density (P/Km2)***: Population density measured in persons per square kilometer.
- ***Abbreviation***: Abbreviation or code representing the country.
- ***Agricultural Land (%)***: Percentage of land area used for agricultural purposes.
- ***Land Area (Km2)***: Total land area of the country in square kilometers.
- ***Armed Forces Size***: Size of the armed forces in the country.
- ***Birth Rate***: Number of births per 1,000 population per year.
- ***Calling Code***: International calling code for the country.
- ***Capital/Major City***: Name of the capital or major city.
- ***CO2 Emissions***: Carbon dioxide emissions in tons.
- ***CPI***: Consumer Price Index, a measure of inflation and purchasing power.
- ***CPI Change (%)***: Percentage change in the Consumer Price Index compared to the previous year.
- ***Currency Code***: Currency code used in the country.
- ***Fertility Rate***: Average number of children born to a woman during her lifetime.
- ***Forested Area (%)***: Percentage of land area covered by forests.
- ***Gasoline Price***: Price of gasoline per liter in local currency.

- ***GDP***: Gross Domestic Product, the total value of goods and services produced in the country.
- ***Gross Primary Education Enrollment (%)***: Gross enrollment ratio for primary education.
- ***Gross Tertiary Education Enrollment (%)***: Gross enrollment ratio for tertiary education.
- ***Infant Mortality***: Number of deaths per 1,000 live births before reaching one year of age.
- ***Largest City***: Name of the country's largest city.
- ***Life Expectancy***: Average number of years a newborn is expected to live.
- ***Maternal Mortality Ratio***: Number of maternal deaths per 100,000 live births.
- ***Minimum Wage***: Minimum wage level in local currency.
- ***Official Language***: Official language(s) spoken in the country.
- ***Out of Pocket Health Expenditure (%)***: Percentage of total health expenditure paid out-of-pocket by individuals.
- ***Physicians per Thousand***: Number of physicians per thousand people.
- ***Population***: Total population of the country.
- ***Population: Labor Force Participation (%)***: Percentage of the population that is part of the labor force.
- ***Tax Revenue (%)***: Tax revenue as a percentage of GDP.
- ***Total Tax Rate (%)***: Overall tax burden as a percentage of commercial profits.
- ***Unemployment Rate (%)***: Percentage of the labor force that is unemployed.
- ***Urban Population (%)***: Percentage of the population living in urban areas.
- ***Latitude***: Latitude coordinate of the country's location.
- ***Longitude***: Longitude coordinate of the country's location.

## Acknowledgements

The work is funded by the SP23-FMI-008 and MUPD23-FMI-021 projects at the Research Fund of the University of Plovdiv "Paisii Hilendarski".

## References

[1] World Health Organization (WHO), https://www.who.int/.

[2] Institute for Health Metrics and Evaluation (IHME). "Global Burden of Disease 2021: Findings from the GBD 2021 Study". Seattle, WA: IHME, 2024. https://www.healthdata.org/research-analysis/library/global-burden-disease-2021-findings-gbd-2021-study.

[3] Global Burden of Disease, https://www.healthdata.org/research-analysis/gbd.

[4] World Bank Group, https://www.worldbank.org/.

[5] UNICEF, https://www.unicef.org/.

[6] The Global Fund, https://www.theglobalfund.org/.

[7] United Nations Development Programme, https://www.undp.org/.

[8] Eurostat, https://ec.europa.eu/eurostat/statistics-explained/index.php?title=Category:Health.

[9] Lelieveld, J., A. Pozzer, U. Pöschl, M. Fnais, A. Haines and T. Münzel, "Loss of life expectancy from air pollution compared to other risk factors: a worldwide perspective", Cardiovascular research, vol. 116(11), pp. 1910-1917, 2020. https://doi.org/10.1093/cvr/cvaa025

[10] Energy Policy Institute at the University of Chicago, https://global.uchicago.edu/news/air-quality-life-index-2023-update-finds-air-pollution-reduces-average-persons-life-expectancy-23.

[11] R. Burnett, H. Chen, M. Szyszkowicz, N. Fann, et al. "Global estimates of mortality associated with long-term exposure to outdoor fine particulate matter". Proceedings of the National Academy of Sciences, vol. 115(38), pp. 9592-9597, 2018.

[12] B. Jackowska, P. Wiśniewski, T. Noiński and P. Bandosz, "Effects of lifestyle-related risk factors on life expectancy: A comprehensive model for use in early prevention of premature mortality from noncommunicable diseases", Plos one, vol. 19(3), 2024. https://doi.org/10.1371/journal.pone.0298696

[13] H. Oster and I. Chaves, "Effects of Healthy Lifestyles on Chronic Diseases: Diet, Sleep and Exercise", Nutrients, vol. 15(21), 4627, 2023. https://doi.org/10.3390/nu15214627

[14] L. T. Fadnes, J. M. Økland, Ø. A. Haaland and K. A. Johansson, "Estimating impact of food choices on life expectancy: A modeling study". PLoS medicine, vol. 19(2), 2022. https://doi.org/10.1371/journal.pmed.1003889

[15] L. T. Fadnes, C. Celis-Morales, J. M. Økland, S. Parra-Soto, et al. "Life expectancy can increase by up to 10 years following sustained shifts towards healthier diets in the United Kingdom", Nature Food, vol. 4(11), pp. 961-965, 2023. https://doi.org/10.1038/s43016-023-00868-w

[16] L. T. Fadnes, E. J. Arjmand, J. M. Økland, Celis-Morales, et al. "Life expectancy gains from dietary modifications: a comparative modeling study in 7 countries", The American Journal of Clinical Nutrition, vol. 150(1), pp. 170-177, 2024. https://doi.org/10.1016/j.ajcnut.2024.04.028

[17] P. T. Roesch, N. S. Saiyed, E. Laflamme, F. G. De Maio and M. R. Benjamins, "Life Expectancy Gaps Among Black and White Persons and Contributing Causes of Death in 3 Large US Cities, 2018-2019". JAMA Network Open, vol. 6(3), 2023, e233146. https://doi.org/10.1001/jamanetworkopen.2023.3146

[18] M. Roberts, E. N. Reither and S. Lim, "Contributors to the black-white life expectancy gap in Washington

DC", Scientific reports, vol. 10(1), 13416, 2020. https://doi.org/10.1038/s41598-020-70046-6

[19] Global Burden of Disease 2021: Findings from the GBD 2021 Study, https://www.healthdata.org/research-analysis/library/global-burden-disease-2021-findings-gbd-2021-study.

[20] J. Wang, Encyclopedia of Data Science and Machine Learning, 2022. https://doi.org/10.4018/978-1-7998-9220-5

[21] Orange Datamining, https://orangedatamining.com/.

[22] N. Elgiriyewithana, Countries of the World 2023 [Dataset]. Kaggle. https://www.kaggle.com/datasets/nelgiriyewithana/countries-of-the-world-2023/data.

[23] K. Cabello-Solorzano, I. Ortigosa de Araujo, M. Peña, L. Correia, L and A. Tallón-Ballesteros, "The impact of data normalization on the accuracy of machine learning algorithms: a comparative analysis". In International Conference on Soft Computing Models in Industrial and Environmental Applications. Cham: Springer Nature Switzerland, pp. 344-353, 2023. https://doi.org/10.1007/978-3-031-42536-3

[24] A. R. Ajiboye, R. Abdullah-Arshah, H. Qin and H. Isah-Kebbe, "Evaluating the effect of dataset size on predictive model using supervised learning technique", International Journal of Software Engineering & Computer Sciences, vol. 1(1), pp. 75-84, 2015. https://doi.org/10.15282/ijsecs.1.2015.6.0006