Determinants of Global Life Expectancy: A Machine Learning Analysis Using Random Forest Models

Xinyan Li

University of California, Irvine, USA

Abstract. In this paper, I adopt a data-driven approach to analyze global life expectancy. Using World Bank data from 174 countries covering the period 2001–2019, I model economic, social, and health factors that may predict a nation's life expectancy. I first performed data preprocessing, which included median imputation for missing values, and then tested several regression models, as well as Random Forest, Boosting, and Decision Tree models. The Random Forest Regressor exhibited the best predictive performance and stability, as it can capture both non-linear and multi-level relationships between income and region. Feature importance analysis showed that undernourishment, communicable diseases, and healthcare expenditure had the most significant impacts on life expectancy, while education expenditure and injury had the least. These results support the view that income alone cannot guarantee population health; instead, investments in health infrastructure and access to basic resources are key to extending overall life expectancy.

Keywords: Life Expectancy; Socioeconomic Determinants; Health Expenditure; Machine Learning; Random Forest Regression.

1. Introduction

Understanding the reasons behind variations in human lifespan has long been a fundamental topic among public health, sociology, and finance researchers. Life expectancy is shaped by a combination of nutrition, education, environment, and healthcare systems. However, despite overall global progress, significant gaps persist across different income groups.

The purpose of this study was twofold: (1) To identify which socioeconomic and health indicators contribute most significantly to life expectancy; (2) To determine which current machine learning methods perform best on our dataset. This was accomplished by fitting and testing various models trained on World Bank life expectancy data from 2001 to 2019. Preliminary results indicate that the Random Forest model outperforms Boosting and Decision Tree models, suggesting that health outcomes are influenced by multiple determinants and that nonlinear modeling must be considered.

2. Literature Review

Associations between national income, healthcare spending, and life expectancy have been of interest to demographers and public health researchers for over a decade. Preston (1975) was the first to demonstrate that GDP per capita is positively associated with life expectancy, with diminishing returns to wealth at higher income levels. The "Preston Curve" established GDP as an upstream determinant of population health: greater wealth correlates with improved access to sanitation, nutrition, and healthcare infrastructure; however, this relationship weakens once certain income thresholds are crossed. Deaton (2003) built on Preston's work, arguing that economic growth enhances health only through mechanisms such as reduced income inequality, improved political quality, and expanded access to social services. Thus, GDP acts as a distal cause whose health effects depend on intermediate factors like health investment and social policy.

Numerous subsequent studies have identified healthcare expenditure as one of the most influential "proxies" linking income to life expectancy. Raeesi et al. (2018) showed that both public and private health spending contribute to longer lifespans, but public spending is more effective because it reaches a broader population and reduces health inequalities. Barber et al. (2017) expanded on this research by introducing a new indicator of health system performance—the HAQ index—which measures the efficiency of health spending across countries. The HAQ index reveals significant

variation in health spending levels among different health systems. Efficiency here refers to factors such as the relative strength of primary care, the proportion of administrative waste, and the emphasis on prevention over tertiary care. Building on this work, Zarulli et al. (2021) noted that many countries could extend life expectancy without increasing spending, provided they reallocate resources more effectively. Taken together, these studies suggest that GDP affects life expectancy not only by increasing total resources allocated to health but also through the manner in which those resources are spent.

Other broader social determinants also strongly mediate the relationship between GDP and life expectancy, with education being the most critical. Raghupathi & Raghupathi (2020) and Balaj et al. (2024) illustrated that education enhances health literacy, enabling individuals to understand medical advice, engage in preventive health behaviors, and adopt healthier lifestyles. They argued that education, as an upstream determinant, may actually be more important than economic growth in influencing morbidity and mortality. Reynolds et al. (2018) found that social spending on education, housing, and welfare yields gains in life expectancy that can be comparable to those from medical spending in high-income countries. These findings indicate that the benefits of income extend beyond healthcare, as they can also be channeled into social spending that builds human capital and resilience.

However, some studies caution against the simplistic assumption that higher income (or expenditure) directly translates to better health. Shkolnikov et al. (2019) replicated the Preston Curve using newer data and found that the marginal gains in life expectancy flatten out rapidly in high-income settings. This reminds us that once most mortality from infectious diseases is eliminated, additional GDP or healthcare spending yields limited returns. Zarulli et al. (2021) highlighted that health systems often waste resources with little tangible benefit, noting that "more can be achieved with existing resources than is currently being done." Deaton (2003) and Reynolds et al. (2018) further argued that inequality and institutional quality moderate the relationship between wealth and health, warning that a narrow focus on spending could blind us to other structural determinants of health.

The Global Burden of Disease (GBD) 2021 decomposition study (2024) supports this argument by analyzing the causes of increased life expectancy. The authors noted that global gains in life expectancy stem from reduced deaths from neonatal conditions, infectious diseases, and cardiac issues. These are the outcomes driven by public health efforts and new treatment options. This suggests that while GDP serves as a distal driver of resources, the proximate factors (i.e., those reducing the burden of specific causes of death) are the actual drivers of improved longevity. Economic growth can facilitate such progress, but it does not guarantee it in the absence of deliberate health and social policy actions.

3. Data and Methods

The dataset used in this study contains 3,306 observations from 174 countries spanning 2001–2019, with variables including life expectancy, prevalence of undernourishment, CO2 emissions, healthcare and education expenditure as percentages of GDP, unemployment rate, sanitation, and cause-of-death categories (injuries, communicable diseases, cardiovascular diseases, and non-communicable diseases). Prior to analysis, several data preparation steps were performed: the "corruption" column, which contained over 70% missing values, was removed; remaining missing data was imputed using median imputation to preserve the overall data distribution; categorical variables such as "Region" and "IncomeGroup" were encoded via one-hot encoding to ensure compatibility with machine learning models; and the dataset was split into training (80%) and testing (20%) sets with a random state of 42 to ensure result reproducibility.

	Count	Percentage
Life Expectancy World Bank	188	5.686630
Prevelance of Undernourishment	684	20.689655
CO2	152	4.597701
Health Expenditure %	180	5.444646
Education Expenditure %	1090	32.970357
Unemployment	304	9.195402
Sanitation	1247	37.719298

Figure 1. Missing data summary for selected variables. The table shows the number and percentage of missing values for each indicator prior to imputation. Sanitation and education expenditure had the highest missing proportions, while CO2 and health expenditure had relatively complete data. These values were imputed with median values to retain the dataset's original distribution pattern.

Three machine learning models were constructed using the scikit-learn package to predict life expectancy based on the aforementioned economic and social indicators: a Decision Tree Regressor, a non-linear model that recursively splits the input dataset to capture interactions between predictors, with hyperparameters (max_depth = 5 and min_samples_leaf = 10) used to prevent overfitting; a Random Forest Regressor, which aggregates predictions from multiple decision trees trained on random subsets of the data—an ensemble approach that reduces variance and improves generalization, making it well-suited for global datasets with heterogeneous characteristics; and a Gradient Boosting Regressor, which builds trees sequentially, with each subsequent tree correcting residual errors from previous iterations, an iterative boosting method that enhances model accuracy and captures subtle non-linear relationships between social/economic variables and life expectancy. Model performance was evaluated using Mean Squared Error (MSE) and R^2, and all analyses were conducted in Google Colab using Python 3.

4. Results

Across nations, the average life expectancy stood at 69.7 years, with a range of 40.4 to 84.4 years. Correlational analysis identified four variables with the most robust links to life expectancy. The prevalence of undernourishment (r=-0.69) exhibited the strongest negative correlation, indicating that countries facing greater food insecurity tend to have significantly lower life expectancy. Access to sanitation (r=0.68) showed a marked positive association, as improved sanitation reduces the spread of infectious diseases and contributes to longer lifespans. Health expenditure as a percentage of GDP (r=0.33) also correlated positively with longevity, reflecting that increased investment in healthcare supports better population health. In contrast, communicable diseases (r=-0.22) hindered longer lifespans, particularly in regions lacking sufficient public health infrastructure to protect community health.

In terms of predictive performance, the Boosting Model performed well, achieving an R^2 of approximately 0.97 and an MSE of around 2.80. The Decision Tree Regressor had a lower score, with an R^2 of 0.859 and an MSE of 13.03. The Random Forest Regressor delivered the best results, with an R^2 of about 0.971 and the lowest MSE (2.686), demonstrating strong accuracy and generalization capabilities.

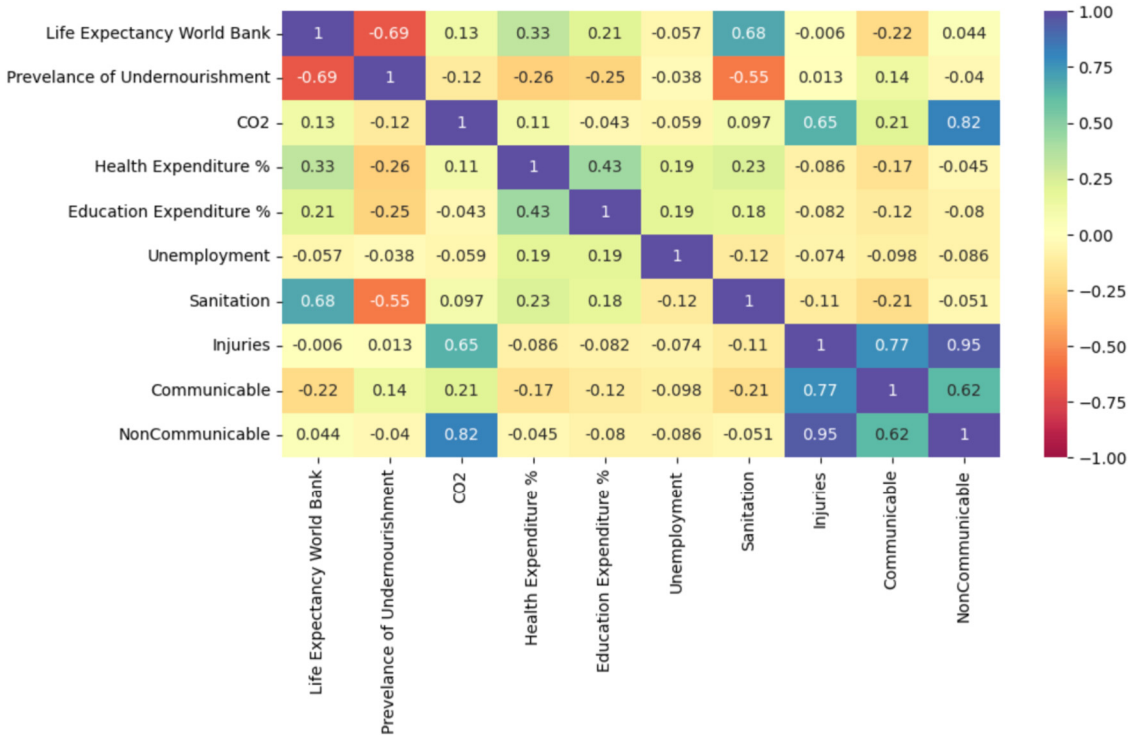

Figure 2. Correlation heatmap illustrating relationships among life expectancy and major socioeconomic and health indicators. Positive correlations are shown in blue, while negative correlations are shown in red, indicating that higher sanitation and health expenditure are associated with longer life expectancy, whereas greater undernourishment and communicable disease prevalence are linked to shorter lifespans.

A further analysis of feature importance shed light on the key drivers of life expectancy. Undernourishment emerged as a primary predictor, highlighting nutrition as the most critical factor in population health. Communicable diseases ranked a close second, underscoring how infections impact a nation's longevity. Health expenditure as a share of GDP was a major positive determinant, as greater spending on healthcare translates to better preventive measures and access to treatment. Access to sanitation was another key factor, encompassing hygiene, public health, and infrastructure needs. CO2 emissions showed a weak yet positive correlation, suggesting that industrialized nations, despite higher pollution levels, tend to have better healthcare and infrastructure. Other factors—such as unemployment, noncommunicable diseases, and injuries—had a weaker impact on mortality, indicating that while labor market conditions and chronic illnesses do affect mortality rates, structural public health factors are the primary drivers of life expectancy.

5. Conclusion

This study applies machine learning techniques to explore a critical question in population health: what factors drive differences in life expectancy across nations? Among the competing models, the Random Forest yielded the strongest and most interpretable results. It reveals that basic needs and investments in robust health systems—including sanitation, nutrition, and public health spending—explain most of the global variation in human lifespan.

As such, policymakers should prioritize allocating resources to strengthen basic healthcare, sanitation, and education rather than focusing solely on GDP growth. Future research could incorporate inequality and government-related indicators to better understand causal relationships, while also extending predictive models to regional or other subnational levels.

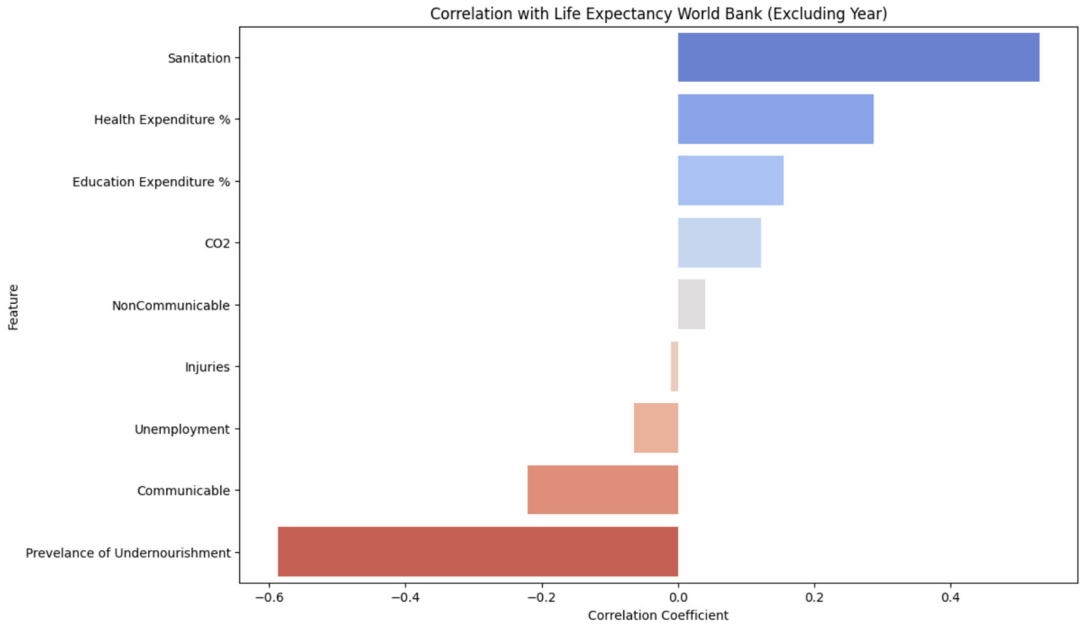

Figure 3. Correlation coefficients between socioeconomic and health indicators and life expectancy. This bar chart visualizes the correlation of each variable with life expectancy. Sanitation, health expenditure, and education expenditure show strong positive associations with longevity, whereas the prevalence of undernourishment exhibits the strongest negative correlation.

References

[1] Barber, R. M., et al. (2017). Healthcare access and quality index based on mortality from causes amenable to personal health care in 195 countries and territories, 1990–2015. The Lancet, 390(10091), 231–266. https://doi.org/10.1016/S0140-6736(17)30818-8.

[2] Balaj, M., et al. (2024). Education and life expectancy: Cross-national evidence. Social Science & Medicine, 347, 116833. https://doi.org/10.1016/j.socscimed.2024.116833.

[3] Deaton, A. (2003). Health, inequality, and economic development. Journal of Economic Literature, 41(1), 113–158. https://doi.org/10.1257/jel.41.1.113.

[4] Preston, S. H. (1975). The changing relation between mortality and level of economic development. Population Studies, 29(2), 231–248. https://doi.org/10.1080/00324728.1975.10410201.

[5] Zarulli, V., et al. (2021). Health system efficiency and the reallocation of resources. European Journal of Public Health, 31(4), 617–623. https://doi.org/10.1093/eurpub/ckab026.

[6] Morina, F., Komoni, A., Kilaj, D., Selmonaj, D., Grima, S. (2022). The effect of health expenditure on life expectancy. International Journal of Sustainable Development and Planning, Vol. 17, No. 5, pp. 1389-1401. https://doi.org/10.18280/ijsdp.170502.

[7] Reynolds, M., et al. (2018). Social spending and life expectancy in high-income countries. Journal of Epidemiology & Community Health, 72(3), 240-246. 10.1016/j.amepre.2017.09.001.

[8] Shkolnikov, V. M., Andreev, E. M., Tursun-Zade, R., & Leon, D. A. (2019). Patterns in the relationship between life expectancy and gross domestic product in Russia in 2005-15: a cross-sectional analysis. The Lancet. Public health, 4(4), e181–e188. https://doi.org/10.1016/S2468-2667(19)30036-2.

[9] Global Burden of Disease Study Collaborators. (2024). Decomposition of global life expectancy gains. The Lancet, 403(10440), 2100–2132. https://doi.org/10.1016/S0140-6736(24)00367-2.

[10] Chavan, S. S. (n.d.). Life expectancy & socio-economic (World Bank) [Data set]. Kaggle. Retrieved October 29, 2025, from https://www.kaggle.com/datasets/mjshri23/life-expectancy-and-socio-economic-world-bank.