**Global Determinants**

**of Life Expectancy**

***A Predictive Analysis Using WHO Global Health Data (2000–2015)***



**Prepared by:**Linda Chimombo,

Haggai Chipeta,

Fungai Mawire,

Peter McKechnie

Pace University

CS 661 – Python Programming

# 1. Abstract

The project explores the key determinants of life expectancy from 2000 to 2015 using the World Health Organization (WHO) Life Expectancy dataset. The aim is to analyze and model how variables such as education, income, healthcare access, environmental factors, and public health indicators influence life expectancy across 193 countries. Methods used included machine learning techniques, particularly linear regression, decision trees and random forest algorithms, to identify the most impactful contributor to longevity.

Through Exploratory Data Analysis (EDA) our key findings highlight that schooling, income composition, and immunization rates have strong positive correlations with life expectancy. The results not only provide a data-driven perspective on global health disparities but reinforce the need for integrated socio-economic and healthcare policies to improve lifespan outcomes. This analysis serves as base tool for health researchers and policymakers aiming to foster equitable and sustainable public health improvements worldwide.

# 2.Introduction

Life expectancy is the main Key Performance Indicator (KPIs) of a country’s overall development and the well-being of its people. It reflects the impact of healthcare systems, social structures, economic opportunities, and environmental conditions. Understanding what influences longer or shorter lifespans is essential for drafting and implementing smarter policies, effective resource allocation, and improving health outcomes both within a country and globally.

In this project we analyse the World Health Organization (WHO) Life Expectancy dataset, for 193 countries from the year 2000 to 2015. Our aim is to explore factors such as education levels, income distribution, access to medical care and environmental quality and conclude on their level of impact in shaping life expectancy trends across the globe.

To identify the existing patterns and relationship within the data, we performed Exploratory Data Analysis (EDA), through application of machine learning techniques, particularly linear regression, decision trees and random forest models, to predict life expectancy and assess the importance of each contributing variable. Beyond just producing predictions, we aimed to turn raw data into meaningful insights that support informed decision-making.

In conclusion this analysis not only contributes to a deeper understanding of the social and economic levers that promote longer, healthier lives. It is a tool for public health professionals and policymakers to build a more equitable and resilient health systems.

# 3.Dataset

The dataset used in this analysis originates from the World Health Organization (WHO) and comprises over 2,900 observations spanning 193 countries. Each record includes numerous features such as:

* Health metrics (e.g., HIV/AIDS, BMI, infant mortality, immunization rates),
* Socio-economic indicators (e.g., schooling, income composition, GDP),
* Demographics (e.g., year, status of development).

Key target variable: Life expectancy.

# 4.Preprocessing

Several preprocessing steps were applied:

* Missing Values: Imputation or removal depending on the nature and sparsity of the column.
* Feature Selection: Based on correlation heatmaps and domain knowledge, key variables like Schooling, log\_GDP, HIV/AIDS, Polio, and BMI were selected.
* Feature Engineering: GDP was log-transformed (log\_GDP) to reduce skewness.
* Train-Test Split: The dataset was split into training and test sets to evaluate model generalization performance.

# 5.Literature Review

For years researchers, governments and global organizations have been fascinated by what determines human longevity. Life expectancy is more than just a number, it reflects the systems of socioeconomics, politics and health for any country. Research has analyzed how the following factors including: income, education, health access, environmental quality and policy impact life expectancy.

**Socioeconomic Drivers: Wealth and Education**

There is a clear link between a country's level of wealth, and its population's length of life. Teh et al. (2025) conducted a multi-country study of 125 countries and concluded that higher GDP per capita and better healthcare systems and educational investment drove improvements in life expectancy. This study supported earlier work by Mimi et al. (2024), who emphasized that while income matters, it is the quality of services provided, and equality of access that influences life expectancy.

**Education: A Lifespan Multiplier**

Scholars have reinforced that education has a greater impact on life expectancy. The notion being more years of schooling is associated with longer lives, healthier behaviors, and better healthcare utilization. Thus, education is regarded as a social vaccine.

**Environmental and Health System Influences**

Environmental and health system factors also impact life expectancy. Clean air, access to vaccines, and functioning health systems are essential. Davey and the GBD Collaborators (2024) stated a global shift from infectious to chronic diseases as a main health burden.

**Historical Trends and WHO Global Patterns**

De Souza and Rêgo (2018) analyzed WHO data from 2000 to 2015 and reported improvements in life expectancy. Mathers et al. (2018) then confirmed that non-communicable diseases are now the leading causes of death worldwide.

**The Combined Impact: An Interdisciplinary Perspective**

There is no single factor that drives life expectancy. Education, income equality, clean environments, and health access together produce longer and healthier lives. Tosun and Yilmaz (2024) argue for an integrated policy approach to tackle life expectancy disparities.

# 6.Architecture / Methodology

The methodology outlines the step-by-step process we followed in analyzing the WHO Life Expectancy dataset. It includes data collection, preprocessing, exploratory analysis, feature selection, model implementation, evaluation, and interpretation using machine learning techniques.

|  |  |
| --- | --- |
| **Step** | **Description** |
| * 1. Data Acquisition | * Dataset sourced from WHO covering 193 countries from 2000–2015 with 22 features. |
| * 1. Data Preprocessing | * Removed whitespace in column names, dropped the 'Status' column, and removed rows with missing target values. * The numerical missing values were filled using median/mode value. * The categorical values were filled using mode. * To ensure consistency we standardized numerical features using StandardScaler. |
| * 1. Exploratory Data Analysis (EDA) | * Visualized data distributions and checked correlations through summary statistics, histograms, and correlation heatmaps. * We identified outliers, trends and identified relationships between life expectancy and features like schooling, income composition, and immunization rates. |
| * 1. Feature Selection | * Applied SelectKBest with f\_regression to extract the top 10 most relevant features that influenced life expectancy. * Visualized the resulting scores with a bar chart to illustrate the variance and contribution of each feature. |
| * 1. Model Selection | * Implemented three machine learning models: Linear Regression for the baseline model. * The Random Forest Regressor was the main model which handled non-linear relationships. |
| * 1. Model Evaluation | * Used R², Mean Absolute Error (MAE), and Mean Squared Error (MSE) to assess model performance. * The result: Random Forest outperformed Decision Trees and Linear Regression by achieving higher accuracy and lower error rates. |
| 6.7. Interpretation | * The variables that demonstrated the strongest predictive power in estimating life expectancy were the following:  1. Schooling, 2. Income composition, 3. Health expenditure, 4. Immunization, and 5. Adult Mortality. |
| 6.8 Tools and Libraries | * The project was implemented in Python using the following tools and libraries:  1. **Data Manipulation –** pandas and numpy. 2. **Data Visualization-** matplotlib.pyplot and seaborn. 3. **Machine Learning Models-** Linear Regression, Logistic Regression, Random Forest Regressor, Decision Trees. 4. **Model Evaluation -** mean squared error, r2 score, mean absolute error, accuracy score, precision\_score, recall\_score, f1\_score. 5. **Preprocessing and Data Splitting-** train test split and StandardScaler. |

# 7.Results

The correlation heatmap and model results revealed several significant predictors of life expectancy. Below are the top factors that Boost Life Expectancy (Strong Positive Relationships):

**Education** (Schooling): The correlation heatmap shows a strong positive correlation (0.74) between years of schooling and life expectancy. This indicates that the more an individual is educated the longer he/she tends to live, and this could be attributed to better health knowledge and job opportunities.

**Basic Needs** (Income Composition of Resources): A high correlation (0.72) with life expectancy highlights the importance of access to basic needs like food, shelter, and health services, leading to a longer lifespan.

**Wealth** (log\_GDP): Wealthier countries tend to have better healthcare systems, and the correlation heatmap confirms this with a positive correlation of 0.56 between log\_GDP and life expectancy.

**Nutrition** (BMI): A healthy BMI range is associated with longer lives. The heatmap reveals a positive correlation of 0.56, suggesting better nutrition supports a longer lifespan.

**Vaccinations** (Diphtheria, Polio, Hepatitis B): The heatmap shows positive correlations around 0.55-0.57 for immunizations and life expectancy, indicating that vaccinated populations live longer.

Below are the Strong Negative Relationships:

**HIV/AIDS:** The correlation heatmap reveals a very strong negative correlation (-0.80) between HIV/AIDS and life expectancy, emphasizing the critical role of disease control in public health. This means that HIV/AIDS impacts negatively on life expectancy, thus reducing survival years for those infected.

**Adult Deaths** (Adult Mortality): A strong negative correlation (-0.69) highlights the direct impact of adult mortality rates on reducing average life expectancy.

**Child Deaths** (Under-Five Deaths, Infant Deaths): High child mortality rates, as shown by negative correlations of -0.60 and -0.57 respectively, significantly lower a country's average life span.

**Measles**: The heatmap indicates a negative correlation of -0.34, suggesting that weak vaccination coverage and poor healthcare contributes to lower life expectancy.

Other Observations from the Heatmap include:

**Alcohol**: A moderate correlation (0.39) suggests higher alcohol consumption is somewhat associated with higher life expectancy. This does not mean alcohol causes people to live longer, but there could be other confounding factors e.g., wealthier countries both drink more and have better healthcare and this could be dynamic at play.

**Population Size**: The heatmap shows a near-neutral correlation (-0.09) between log Population and life expectancy, indicating that population size does not directly affect the lifespan of individuals.

# 8.Conclusion

The correlation heatmap provides visual confirmation of the key factors influencing life expectancy. Education, basic needs, wealth, nutrition, and vaccinations are positively correlated with longer lives, while diseases, high mortality rates, and poor healthcare are negatively correlated.

Based on the model performance comparison, the Random Forest Regressor demonstrated the best predictive capability. It achieved the highest R² score, indicating that it explains the largest proportion of variance in life expectancy, while simultaneously producing the lowest RMSE and MAE values, reflecting minimal prediction error. In contrast, the Linear Regression and Decision Tree models showed comparatively lower R² scores and higher error metrics. Therefore, Random Forest is the most suitable model for accurately predicting life expectancy in this study.

# 9.References

**Dataset**

K Jarshi, “*Life Expectancy (WHO)*,” Available at [https://www.kaggle.com/datasets/kumarajarshi/life-expectancy-who] (Accessed April – May 2025)

**Sources**

 Davey, G., & GBD 2021 Causes of Death Collaborators. (2024). *Global burden of 288 causes of death and life expectancy decomposition in 204 countries and territories and 811 subnational locations, 1990–2021: a systematic analysis for the Global Burden of Disease Study 2021*.

 De Souza, F. C., & Rêgo, L. C. (2018). Life expectancy and healthy life expectancy changes between 2000 and 2015: an analysis of 183 World Health Organization member states. *Journal of Public Health, 26*, 261–269.

 Mathers, C., et al. (2018). *Global and regional causes of death: patterns and trends, 2000–15*.

 Mimi, M. B., Kibria, M. G., & Selim, M. M. I. (2024). How do economic, health, environmental and demographic factors affect life expectancy? A novel attempt for developed and developing economies. *International Journal of Sustainable Development & World Ecology, 31(7)*, 776–791.

 Teh, W. S., Ling, G. L. H. T., Rusli, N., Teh, H. H., Tan, K. P., Faisal, S., … Abdullh Mohamed, M. S. (2025). Factors Influencing Life Expectancy: Global Evidence from 125 Countries. *International Journal of Built Environment and Sustainability, 12(2)*, 17–28. <https://doi.org/10.11113/ijbes.v12.n2.1294>

 Tosun, M. S., & Yilmaz, S. (2024). Impact of education and income inequalities on life expectancy in the EU. *Frontiers in Public Health*. <https://www.frontiersin.org/articles/10.3389/fpubh.2024.1397585>

 Van de Poel, E., et al. (2020). The influence of education on health: an empirical assessment. *Archives of Public Health*. <https://archpublichealth.biomedcentral.com/articles/10.1186/s13690-020-00402-5>

# 10.GitHub Repository for Code