Project Title: Life Expectancy Analysis

Author: John Christo

Date: August 4, 2025

1. EXECUTIVE SUMMARY

This report presents the development and evaluation of a machine learning model to accurately predict the life expectancy of a country's population. Using a comprehensive dataset from the World Health Organization (WHO), this project aimed to identify key health, economic, and social indicators that influence longevity and build a robust predictive tool.

The methodology involved a thorough process of data cleaning, including mean imputation for missing values, followed by in-depth exploratory data analysis to uncover key relationships. A **Random Forest Regressor** model was trained on the prepared data and rigorously evaluated.

The model achieved exceptional performance, demonstrating a very strong predictive capability. The key findings are:

- An **R-squared (R2) score of 0.97**, indicating the model explains 97% of the variance in life expectancy.
- A **Mean Absolute Error (MAE) of only 1.05 years**, signifying high precision in the predictions.

Given the outstanding accuracy of the Random Forest model, it is recommended as a highly effective tool for analytical and predictive tasks related to public health and global development.

2. INTRODUCTION

Understanding the factors that contribute to life expectancy is a cornerstone of public health and global development policy. By identifying the most influential health and socioeconomic indicators, governments and organizations can better allocate resources and design effective interventions.

The objective of this project was to leverage machine learning to build a highly accurate regression model for predicting life expectancy. The analysis was conducted using the "Life Expectancy (WHO)" dataset from Kaggle, which contains 15 years of data across 193 countries.

This report details the end-to-end process, from data preparation and exploratory analysis to model training and evaluation, concluding with the practical implications of the model's strong predictive performance.

3. METHODOLOGY

A systematic data science workflow was employed to ensure the development of a robust and reliable model.

- **Data Preparation**: The project began with loading the dataset and cleaning the column names. A critical step was handling missing values; to preserve the dataset's integrity, **mean imputation** was used to fill any missing numerical data points, ensuring a complete dataset for analysis.
- **Exploratory Data Analysis (EDA)**: Comprehensive EDA was performed to understand the data's underlying structure. Key visualizations included a histogram of the Life expectancy target variable, a correlation heatmap to identify relationships between all variables, and scatter plots to examine the connection between top predictors (like 'Schooling' and 'Income composition of resources') and life expectancy.
- **Feature Engineering & Preprocessing**:
    - The categorical Status column ('Developed'/'Developing') was converted into a numerical format using one-hot encoding to make it machine-readable.
    - The final feature set was prepared, and the data was split into an 80% training set and a 20% testing set using train_test_split for model evaluation.
- **Model Selection**: A **Random Forest Regressor** was selected for this regression task. This model is a powerful ensemble method known for its high accuracy, robustness, and ability to handle complex interactions between features.
- **Evaluation Metrics**: The model's performance was measured using two standard regression metrics:
    - **Mean Absolute Error (MAE)**: The average absolute difference between the predicted and actual values.
    - **R-squared (R2)**: The proportion of the variance in the dependent variable that is predictable from the independent variables.

4. RESULTS AND ANALYSIS

The Random Forest Regressor model yielded exceptionally strong results when evaluated on the unseen test data.

Quantitative Performance

The model's performance metrics underscore its high accuracy and reliability:

- **R-squared (R2) Score: 0.97**. This outstanding result indicates that the model successfully explains 97% of the variability in life expectancy based on the provided features. An R2 score this close to 1.0 signifies an extremely strong model fit.
- **Mean Absolute Error (MAE): 1.05 years**. This result is equally impressive, showing that, on average, the model's predictions are off by only 1.05 years. This level of precision is remarkable for a complex, real-world prediction task.

Feature Importance

A key benefit of the Random Forest model is its ability to determine the importance of each feature in making predictions. The analysis confirms that socio-economic factors like **'Income composition of resources'** and **'Schooling'**, along with health factors like **'HIV/AIDS'** and **'Adult Mortality'**, were among the most influential predictors in the model.

5. PRACTICAL IMPLICATIONS

A model with this degree of accuracy has significant practical applications for various organizations:

- **Public Health Policy**: Governments can use the model's insights to prioritize interventions. For instance, the high importance of 'Schooling' reinforces the link between education and long-term health outcomes.
- **Resource Allocation**: Non-governmental organizations (NGOs) and international bodies like the WHO can better allocate funds and aid to countries by identifying the factors that could provide the most significant boost to life expectancy.
- **Global Development Insights**: The model serves as a powerful analytical tool to study the complex interplay between a nation's economy, education system, and health infrastructure on the well-being of its population.

6. CONCLUSION

This project successfully developed a high-performing machine learning model capable of predicting life expectancy with an exceptional R-squared score of 0.97. The thorough process

of data cleaning, exploratory analysis, and the application of a robust Random Forest Regressor model culminated in a tool with significant predictive power and practical utility. The model is complete and has achieved a level of performance that exceeds typical project expectations.