Predictive Analytics in Global Health: Modeling Life Expectancy

Toronto
Metropolitan
University

Student Name:- Himit Patel

Student Number :- 501344641

Supervisor’s name :- Dr. Ceni Babaoglu

Date of submission:- 1-Dec-2025

1. Introduction & Research Questions

Life expectancy is one of the most commonly used indicators to assess the overall health, development, and quality of life in a country. Many social, economic, and medical factors contribute to how long people live, and understanding these factors can help governments and policy makers focus on areas that have the most impact. In this project, I used the WHO Life Expectancy dataset to explore which variables are most strongly related to life expectancy and whether these variables can be used to build accurate prediction models.

The dataset includes global records from 2000 to 2015 and contains multiple indicators such as adult mortality, schooling, GDP, alcohol consumption, and infant deaths. The goal of this analysis is not only to identify key predictors but also to compare different machine-learning models and evaluate how well they can predict life expectancy.

Based on the project requirements and the initial exploration, I developed the following research questions:

Research Questions

1. **Which health, economic, and social indicators are most strongly associated with life expectancy across countries?**

2. **How accurately can machine-learning models predict life expectancy using these indicators?**

3. **Which model performs best in terms of prediction accuracy, efficiency, and stability?**

4. **How do the results align with existing research findings on global health outcomes?**

2. Contribution of the Project

This project extends the initial exploratory work into a full predictive modelling study. My contribution includes:

- Cleaning, preparing, and analyzing a global WHO dataset that contains 16 years of country-level health indicators.

- Building and comparing three predictive models: **Linear Regression, Decision Tree, and Random Forest**.

- Evaluating each model using multiple metrics such as R^2, MAE, RMSE, and cross-validation scores.

- Identifying which variables have the strongest impact on life expectancy using feature importance analysis.

- Connecting the results back to the literature to show how the findings support or differ from previous studies.

- Providing a simple, reproducible notebook that includes all steps from data cleaning to final evaluation.

Overall, the project provides a practical machine-learning approach to understanding which factors influence life expectancy and how accurately we can predict it.

3 .Literature Review and Integration

Existing research shows that life expectancy is influenced by a combination of social, economic, and health-related factors. Many studies highlight that countries with higher income, better education, and stronger healthcare systems tend to have higher life expectancy. Education and income often lead to better living conditions and access to medical services, while mortality rates and infectious diseases usually reduce the average lifespan.

Several papers emphasize that **adult mortality** and **infant mortality** are two of the strongest indicators of life expectancy. When mortality levels are high, the average lifespan naturally decreases. Other studies point out that **schooling** has a large effect because educated populations make healthier choices, follow medical guidelines, and have better access to employment and resources. Some research also shows that economic measures like **GDP** are correlated with life expectancy, but the strength of the relationship varies depending on a country's development level.

Integration with My Findings.

The results from my model analysis match many points from the literature. In my Random Forest model, **adult mortality** came out as the most important feature, which agrees with research showing that higher adult mortality rates sharply reduce life expectancy. **Schooling** also appeared as a major predictor, supporting the findings that education has a long-term influence on health outcomes. Some variables like alcohol consumption and BMI played smaller roles, which is similar to studies stating that their effects depend heavily on country-specific behavior and lifestyle patterns.

Overall, the model findings are consistent with what previous research has shown: life expectancy is strongly shaped by mortality patterns and education, while economic factors like GDP play a supportive but less direct role.

4. Methodology and Study Design

To answer the research questions, I followed a structured machine-learning workflow that included data cleaning, exploratory analysis, feature selection, model development, and evaluation. The full workflow was implemented in Google Colab, and every step was made reproducible.

1) Dataset Overview

The dataset used in this project is the WHO Life Expectancy dataset, which contains 2,938 records from 193 countries covering the years 2000–2015. The dataset includes 22 columns with indicators from multiple domains:

- **Health indicators**: adult mortality, infant deaths, hepatitis B, measles, HIV/AIDS, BMI.
- **Economic indicators**: GDP, total health expenditure, percentage expenditure
- **Social indicators**: schooling, population, status (developed/developing)

The target variable for prediction is "Life expectancy", which is a continuous numeric value.

2) Data Preprocessing

Before building models, I cleaned and prepared the dataset to make it suitable for machine-learning.

Handling Missing Values

Several columns contained missing values, especially:

- Population
- Hepatitis B
- GDP
- Total expenditure
- Alcohol
- Income composition of resources
- Schooling

To address this, I used a simple and interpretable imputation strategy:

- **Numeric columns:** filled using the median
- **Categorical columns:** filled using the mode

After cleaning, the dataset had zero missing values, which ensured that all models could be trained without interruption.

3) Exploratory Data Analysis

I created summary statistics and visualizations to better understand relationships between variables.

Key steps included:

- A correlation heatmap, which showed strong negative correlation between life expectancy and adult mortality, and positive correlation with schooling, BMI, and GDP.
- Trend and scatter plots, such as life expectancy by year and GDP vs. life expectancy, to observe general patterns in the data.

These insights helped guide which features should be used in the predictive models.

4) Feature Selection

Based on EDA results and prior literature, I selected a focused set of predictors with strong conceptual and statistical relationships to life expectancy:

- Adult Mortality
- Alcohol
- BMI
- GDP
- Schooling
- Infant deaths

These variables were chosen because they:

1. Showed noticeable correlation with life expectancy
2. Are commonly used in health-related research
3. Helped improve model performance in my initial experiments

This selection also avoids multicollinearity issues that can occur when too many highly correlated variables are included.

5). Model Development

I trained and compared three different models:

1. Multiple Linear Regression (Baseline)

A simple and interpretable model used to establish baseline performance.

Good for understanding the direction and strength of relationships.

2. Decision Tree Regressor

Captures non-linear relationships that linear regression cannot.

Useful for comparing interpretability vs. accuracy.

3. Random Forest Regressor

An ensemble of multiple decision trees.

Usually provides the best accuracy and reduces overfitting through averaging.

6) Train-Test Split

To evaluate model performance fairly, I split the dataset into:

- 80% training data
- 20% testing data

The split was done using a random_state = 42 to ensure reproducibility.

7) Evaluation Metrics

Each model was evaluated using multiple metrics:

- R^2 score: how much variance is explained.
- MAE (Mean Absolute Error): average absolute error in years.
- RMSE (Root Mean Squared Error): penalizes bigger errors more strongly.
- Cross-Validation R^2: checks model stability across different data splits.

These metrics allowed me to assess accuracy, consistency, and robustness.

5. Model Evaluation

In this section, I compared the performance of the three models—Linear Regression, Decision Tree, and Random Forest—using accuracy scores, error metrics, and stability tests. The goal was to determine which model predicts life expectancy most effectively and whether it generalizes well to unseen data.

1) Effectiveness of the Models

To evaluate how well each model predicts life expectancy, I compared the R² scores and error metrics (MAE and RMSE). The results are shown in the table below:

Model	Train R²	Test R²	MAE (years)	RMSE (years)
Linear Regression	0.721	0.731	3.34	4.83
Decision Tree	1.000	0.926	1.50	2.53
Random Forest	0.994	0.958	1.22	1.90

Interpretation

Linear Regression works as a baseline and explains about 73% of the variation in life expectancy. However, the errors (MAE 3.34, RMSE 4.83) are higher compared to the other models.

The Decision Tree model fits the training data perfectly (Train R^2 = 1.0), which means it memorized the dataset. While the test score is much better than linear regression, the perfect training score shows overfitting.

The Random Forest model delivers the best performance with a Test R^2 = 0.958, meaning it explains about 96% of the variation. It also achieves the lowest errors: ~1.2 years MAE and ~1.9 years RMSE.

Conclusion:

The Random Forest is the most effective model for predicting life expectancy in this dataset.

2) Efficiency (How Fast the Models Run)

While exact training times were not measured in milliseconds, the models showed typical behavior:

- Linear Regression trains extremely fast because it solves a simple mathematical equation.
- Decision Trees train slightly slower but still very fast because they split the data into groups.
- Random Forest trains the slowest because it builds 200 trees, but it still finished within a few seconds on Colab.

Even though Random Forest takes longer, the extra time is justified by the large improvement in accuracy.

Conclusion:

All models are computationally feasible, and the Random Forest offers the best accuracy with acceptable training time.

3) Stability (How Consistent the Model Is)

To test stability, I used 5-fold cross-validation on the Random Forest model.

results:

Cross-Validation R² Scores:

[0.916, 0.890, 0.918, 0.922, 0.904]

Mean CV R²:

0.91

Interpretation

The scores across all folds are very close (between 0.89 and 0.92), which means:

- The Random Forest model works consistently across different subsets of the data.
- The performance is not dependent on a single random train-test split.
- The model generalizes well to unseen data.

Conclusion:

Random Forest is the most stable model, with consistent results across all folds.

Overall Model Evaluation Summary

- Linear Regression gives a decent baseline but misses important non-linear relationships.
- Decision Trees capture complex patterns but overfit the data.
- Random Forest balances accuracy and generalization extremely well and is the best-performing model overall.

6) Findings and Interpretation

The results of the modelling process showed clear patterns about what drives life expectancy and how accurately it can be predicted using machine-learning techniques. After comparing Linear Regression, Decision Tree, and Random Forest, it became clear that the Random Forest model produced the most reliable and accurate predictions.

1) Key Findings from the Analysis

1. Adult Mortality is the strongest predictor of life expectancy.

The feature importance plot shows that adult mortality has the highest importance score by a large margin. Countries with higher adult mortality rates consistently showed much lower life expectancy, which supports the literature linking mortality burden to national health outcomes.

2. Schooling also plays a major role.

Schooling appeared as the second most influential feature. Higher levels of education are usually linked to better awareness of health practices, stable income, and access to medical services — all of which contribute to a longer lifespan.

3. Economic indicators like GDP had a smaller direct effect.

Although GDP is commonly associated with development, its importance was relatively low in the Random Forest model. A possible reason is that adult mortality indirectly captures many economic and healthcare differences between countries.

4. Other variables (BMI, infant deaths, alcohol) had minor contributions.

Their influence was present but not strong enough to change life expectancy predictions significantly. This suggests that lifestyle and isolated health indicators may only affect life expectancy when combined with broader structural factors.

2 Interpretation of Model Results

- Linear Regression provided a basic understanding of how each variable is linearly related to life expectancy. However, it could not capture the non-linear patterns present in the data.
- The Decision Tree model improved accuracy but also showed signs of overfitting, as indicated by the perfect training score (1.0). This means the model learned the training data too specifically.
- The Random Forest model provided the most balanced and accurate results. With a test R^2 score of 0.958 and the lowest error values, it was able to generalize well to unseen data. Cross-validation further confirmed that the model is stable, with consistent R^2 results across different folds.

Together, these results answer the second research question: yes, machine-learning models can predict life expectancy with high accuracy, especially when using ensemble methods like Random Forest.

3) How Findings Address the Research Questions

Research Question	Answer Based on Findings
RQ1: Which indicators influence life expectancy the most?	Adult mortality and schooling are the strongest predictors.
RQ2: Can machine-learning models predict life expectancy accurately?	Yes, especially the Random Forest model (R^2 ~ 0.96).
RQ3: Which model performs the best?	Random Forest performed the best in accuracy, stability, and error reduction.
RQ4: Are the results consistent with past research?	Yes. The findings align with literature emphasizing the importance of mortality rates and education levels.

7. Limitations of the Work

Although the models performed well overall, there are several limitations that may affect the accuracy and generalizability of the results:

1) Missing data handling

Some columns in the dataset contained a significant amount of missing values. Even though I used median and mode imputation, this approach may reduce the true variability in the data and lead to less accurate predictions.

2) Limited feature set

I only selected six features for modeling. The original dataset contains more variables, and some important health or socioeconomic indicators might be missing. Including additional variables could improve accuracy.

3) Country-level aggregation

The dataset represents entire countries, not individual-level data. Country averages can hide important differences within populations, which means the models may not reflect individual health outcomes.

4) Potential multicollinearity

Some features (like GDP, schooling, and expenditure) are correlated. Even though the Random Forest model handles correlation better than Linear Regression, multicollinearity may still reduce interpretability.

5) Model complexity

The Random Forest model performed the best, but it is less interpretable than simpler models. Understanding exactly how each tree makes decisions is difficult.

6) Lack of temporal modeling

The dataset spans 16 years, but the models do not explicitly consider time trends. A time-series model might capture long-term changes more effectively.

Despite these limitations, the models still provide meaningful insights into the key factors shaping life expectancy.

8. Ethical Considerations

Several ethical considerations were taken into account when analyzing the dataset:

1. **Data privacy**

The dataset contains aggregated country-level data, which reduces the risk of identifying individuals. This makes ethical concerns about personal data privacy minimal.

2. **Bias in global health data**

Some countries may have underreported or inconsistent health statistics due to lack of resources or political factors. This can introduce bias into the model's predictions.

3. **Interpretation of results**

Predictive models should not be used to judge the value of populations or countries. Life expectancy is influenced by many structural and historical factors beyond individual control.

4. **Fairness and equity**

The model should not be used to make unfair comparisons between nations. Instead, it should help highlight areas where health investments are needed.

5. **Responsibility in communication**

When presenting results, it is important to avoid implying causal relationships. The models only show associations and predictions, not direct causes.

Overall, the analysis respects ethical guidelines by using non-sensitive data, avoiding discrimination, and focusing on understanding global health patterns.

9. Future Work and Recommendations

There are several ways this project could be expanded or improved:

1.Using more features

Adding additional socioeconomic variables (e.g., income inequality, sanitation access, vaccination coverage) may increase predictive accuracy.

2.Testing more models

Advanced models like XGBoost, Gradient Boosting, or Neural Networks might outperform Random Forest and provide new insights.

3.Hyperparameter tuning

The models in this project used mostly default settings. Tuning parameters such as the number of trees, depth limits, or learning rates could significantly improve performance.

4.Time-series modeling

Since the dataset spans multiple years, using models that incorporate time (e.g., ARIMA, LSTM) could capture trends more effectively.

5.Regional analysis

Breaking the dataset into regions (e.g., Africa, Europe, Asia) could highlight regional patterns that global averages may hide.

6.Improved data preprocessing

More sophisticated imputation methods—such as KNN imputation—might preserve data relationships better than median/mode.

These improvements could help build a more accurate and comprehensive model for predicting life expectancy.

10. Conclusion

This project explored the key factors influencing life expectancy and evaluated machine-learning models to predict it using global WHO data. The findings showed that adult mortality and schooling are the strongest predictors, while economic indicators like GDP played a smaller role. Among the models tested, the Random Forest model delivered the highest accuracy, with a test R^2 of 0.96 and strong stability across cross-validation folds.

The results were consistent with previous research, confirming that mortality rates, education, and national health structures play a major role in shaping life expectancy. Although the project had limitations, such as missing data and limited features, the overall outcomes were meaningful and demonstrated how machine-learning can support global health analysis.

The work provides a solid starting point for future improvements, including adding more features, testing advanced models, and exploring time-based patterns. Overall, the project successfully answered the research questions and showed that machine-learning can be a powerful tool for understanding life expectancy trends.

11.Code and Reproducibility Section

All code used in this project is available in my GitHub repository at the link below:

GitHub Repository: https://github.com/Himitpatel1/WHO-Life-Expectancy-Project

The repository includes the following:

- The full Google Colab notebook ("Patel_Himit_InitialResults.ipynb") containing all preprocessing, EDA, modeling, and evaluation steps.
- A PDF export of the notebook for easier viewing (`Patel_Himit_InitialResults.pdf`).
- A README.md file with instructions for running the notebook.
- The dataset (if required) or a link to the WHO data source.

To reproduce the results, open the notebook in Google Colab, upload the dataset when prompted, and run all cells from top to bottom. All metrics, figures, and model outputs will be generated automatically. The results are fully reproducible on any machine with Python, scikit-learn, pandas, and matplotlib installed.

12.References Page

- Barocas, S., Hardt, M., & Narayanan, A. (2023). *Fairness and machine learning: Limitations and opportunities. MIT Press. https://fairmlbook.org

- Bloom, D. E., & Canning, D. (2019). The health-wealth gap. Finance & Development, 56*(2), 10–13. https://www.imf.org/external/pubs/ft/fandd/2019/06/health-and-wealth.htm

- Breiman, L. (2001). Random forests. *Machine Learning, 45*(1), 5–32. https://doi.org/10.1023/A:1010933404324

- Hunter, J. D. (2007). Matplotlib: A 2D graphics environment. Computing in Science & Engineering, 9*(3), 90–95.

- Kaggle. (n.d.). Life Expectancy (WHO) Dataset. Retrieved from https://www.kaggle.com/datasets/kumarajarshi/life-expectancy-who

- Kuhn, M., & Johnson, K. (2013). Applied predictive modeling.* Springer. https://doi.org/10.1007/978-1-4614-6849-3

- Lutz, W., & Kebede, E. (2018). Education and life expectancy: A global perspective. Demographic Research, 38*(9), 247–272. https://doi.org/10.4054/DemRes.2018.38.9

- McKinney, W. (2017). *Python for data analysis. O'Reilly Media.

- Molnar, C. (2022). Interpretable machine learning: A guide for making black-box models explainable. https://christophm.github.io/interpretable-ml-book/

- Pedregosa, F., et al. (2011). Scikit-learn: Machine Learning in Python.Journal of Machine Learning Research, 12*, 2825–2830.

- Pandas Development Team. (2024). pandas documentation. https://pandas.pydata.org

- Wang, H., Abajobir, A. A., & Abate, K. H. (2017). Global, regional, and national under-5 mortality, adult mortality, and life expectancy. The Lancet, 390*(10100), 1084–1150. https://doi.org/10.1016/S0140-6736(17)31833-0

- World Health Organization. (2015). *WHO Life Expectancy Data.* WHO Global Health Observatory. https://www.who.int/data/gho