Predicting Life Expectancy Using Global Health Data

Joshua Ewer

2026/01/28

Business Problem

Life expectancy is a widely used indicator of population health and plays a central role in decisions made by governments, non-profit organizations, and global health agencies. Even though a lot of global health data is available, it is many times unclear which factors are most strongly associated with differences in life expectancy across countries (World Health Organization [WHO], 2023). This project aims to model the most influential predictors of life expectancy in order to support data-informed public health planning.

Background / History

Due to family health issues, an interest in life expectancy and population health outcomes motivated this project. For decades, organizations such as the World Health Organization have collected indicators related to mortality, disease prevention, healthcare access, and economic conditions to track global health trends (WHO, 2023). Recent advances in data science and machine learning provide tools to explore these relationships and to evaluate their combined predictive power at scale (Pedregosa et al., 2011).

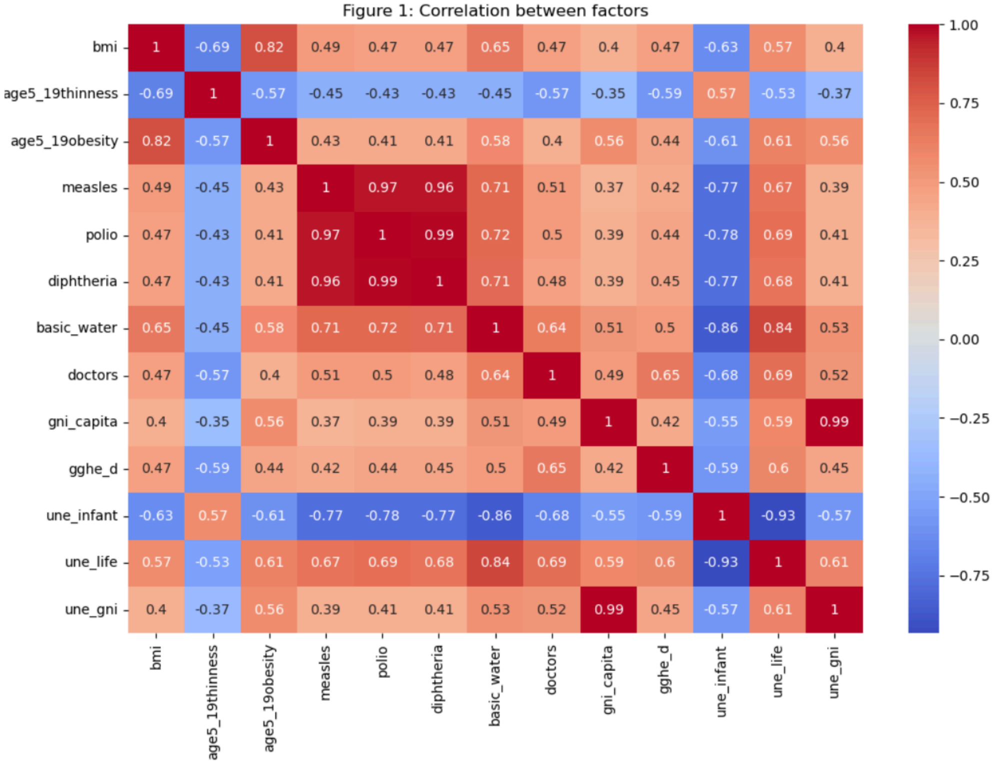
Data Explanation (Data Prep / Data Dictionary)

The dataset used in this project is a publicly available World Health Organization life expectancy dataset obtained from Kaggle. It includes variables such as life expectancy, infant mortality, vaccination coverage, healthcare expenditure, gross domestic product (GDP), alcohol consumption, and poverty-related indicators. All variables are aggregated at the country level and contain no personally identifiable information.

Initial research indicated that the dataset had already undergone basic cleaning, meaning the analysis could focus on exploration and modeling. Missing values were present, particularly among lower-income countries where data collection infrastructure is limited. To address this issue, the data was grouped by country and sorted by year, then forward-filled to replace missing values using prior-year observations. A backward fill was applied to handle missing values at the beginning of each country's time series. After imputation, values were averaged across years so that each country was represented by a single observation, which reduced bias caused by uneven reporting across time.

To minimize the risk of data leakage, variables that were overly correlated with life expectancy were removed from the dataset. Correlation heatmaps were used to identify and guide the removal of these features.

Figure 1: Correlation between factors



Methods

In this project, exploratory data analysis (EDA) was used to explore variable distributions, identify outliers, and better understand relationships between predictors. This is a best practice in data-driven research (Wickham & Grolemund, 2017).

Figure 2: Correlation between Infant Mortality Rates and poverty

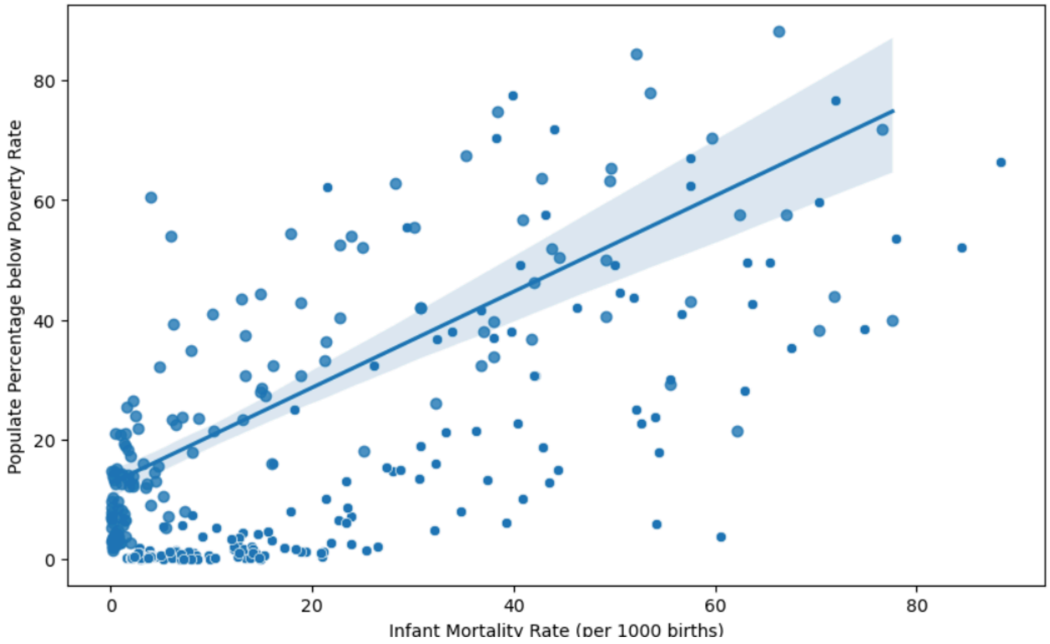
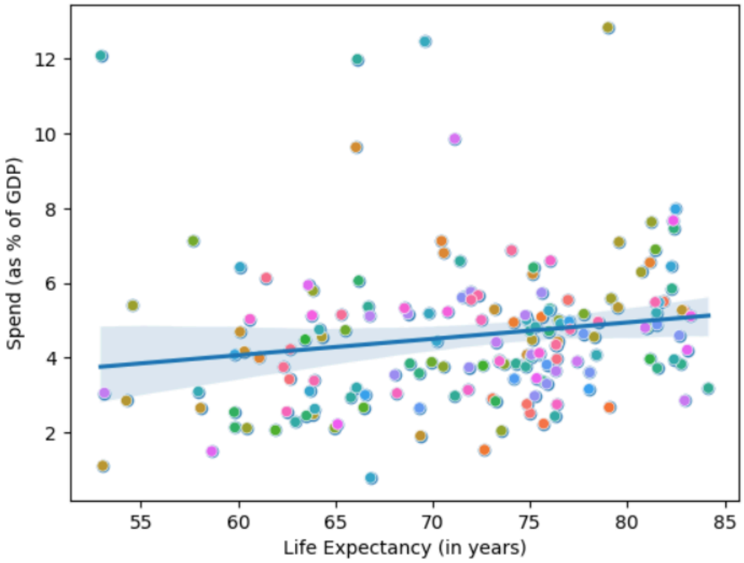


Figure 3: Health Spending Trends



Correlation analysis showed key associations, especially between vaccination rates, poverty, and infant mortality.

Because life expectancy is a continuous target variable, regression-based modeling techniques were applied. Linear regression was used as a baseline model because it is easily interpreted and is often used in health and social science research (James et al., 2021). More flexible models, including Random Forest and XGBoost, were implemented to capture potential non-linear relationships among predictors (Breiman, 2001; Chen & Guestrin, 2016). Gridsearch with cross-validation was used to tune hyperparameters, and model performance was evaluated using the coefficient of determination (R^2).

Analysis

Exploratory analysis revealed several intuitive and meaningful patterns. Vaccination coverage showed a positive association with life expectancy, while poverty rates were positively correlated with infant mortality and negatively correlated with life expectancy. These findings are consistent with prior research on health factors (Marmot et al., 2008). Government healthcare expenditure demonstrated a weaker but still positive relationship with life expectancy.

All three models demonstrated strong predictive performance. Linear regression achieved an R^2 of approximately 0.90, suggesting that much of the variance in life expectancy could be explained by the selected predictors. Random Forest slightly outperformed the other models with an R^2 of approximately 0.903, while XGBoost produced comparable results. Given its slightly stronger performance, likely due to resistance to overfitting, I selected the Random Forest model as the final model (Breiman, 2001).

Conclusion

This project demonstrates that life expectancy can be effectively modeled using a combination of health, economic, and social indicators derived from global health data. Many findings aligned with common expectations, such as the negative impact of poverty and infant mortality on life expectancy. Additional insights, such as the limited value of separating individual vaccination variables, helped refine feature selection and model interpretation.

Assumptions

Several assumptions constrain this analysis. Country-level averages are assumed to reasonably represent national health conditions. Imputed values are assumed to be appropriate approximations where data is missing. The models also assume that relationships between predictors and life expectancy are broadly stable across countries and time periods, despite differences in culture, policy, and data quality.

Limitations

This analysis is limited by the quality and completeness of the available data. Countries with fewer resources often have more missing or less reliable data, which can introduce bias even after imputation. Aggregating data at the country level obscures regional and demographic differences within countries. Additionally, the models identify correlations rather than causal relationships.

Challenges

Key challenges included handling missing data, addressing multicollinearity among predictors, and avoiding data leakage from variables closely tied to the target outcome. Balancing predictive performance with interpretability was also a challenge when comparing simpler linear models with more complex ensemble methods.

Future Uses / Additional Applications

Future work could extend this analysis by including additional lifestyle-related variables, such as diet, physical activity, or substance use. Splitting the data by gender or age group could reveal more patterns in life expectancy. The same modeling framework could also be adapted for regional or intra-national if more granular data is available.

Recommendations

Future analysis should focus on improving data collection and reporting in lower-income countries to reduce systemic bias. Policymakers should interpret model results as exploratory evidence and not definitive conclusions. Combining predictive modeling with public health expertise could lead to more responsible and effective decision-making.

Implementation Plan

In practice, this model could be implemented as a decision-support tool for public health organizations. Updated WHO data could be periodically refreshed and processed using the same data pipelines, allowing the model to generate updated predictions and feature importance summaries. These outputs could inform many policy decisions, including resource allocation and further research initiatives.

Ethical Assessment

Working with global health data raises important ethical considerations. Incomplete or uneven data coverage can reinforce existing global inequalities if not handled carefully (WHO, 2023). There is also a risk of misinterpreting associations as causal relationships, particularly in politically charged climates. For these reasons, results should be treated as exploratory, with an emphasis on transparency. Applying these findings to real world problems would still require expert judgment.

References

Breiman, L. (2001). *Random forests*. Machine Learning.

Chen, T., & Guestrin, C. (2016). *XGBoost: A scalable tree boosting system*. Proceedings of the ACM SIGKDD Conference.

Kaggle. *WHO National Life Expectancy Dataset*.

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

World Health Organization. (2023). *Global Health Observatory data repository*.