Exploring Life Expectancy Using World Bank Data: An R Package and Shiny App for Machine Learning Analysis

Team Error 404: Team Not Found

Wenqing (Cathy) Zhang, Guan Gui, Enrui (Ryan) Wang, Zixu (Jerry) Luo

Contents

1 Background

1.1 Research Question

Life expectancy, or the average number of years a person can expect to live, is widely regarded as a key indicator of population health, reflecting the cumulative impact of various social, economic, and environmental factors on human survival [OECD, 2022, Sharma, 2018]. Compared to crude mortality rates, it more accurately summarizes the specific mortality patterns for a certain population group and provides insights into both current health conditions and future demographic trajectories of this population group [Bilas et al., 2014]. Moreover, life expectancy serves as a proxy for broader socioeconomic development, as improvements in education, employment, and living conditions often translate into longer and healthier lives [Bartley and Kelly-Irving, 2016, Dahlgren and Whitehead, 2021]. In fact, increasing life expectancy is also one of the key goals in the United Nations Sustainable Development Goals for 2030 [Nations, 2022].

Yet, average life gains do not flow evenly across all populations, as socioeconomic disparities—rooted in power imbalances, material disadvantages, and differential access to resources—fundamentally shape who lives longer and who does not [Bambra, 2022]. Therefore, by creating predictive models and interactive tools for users to explore life expectancy through socioeconomic indicators, this project seeks to uncover the underlying drivers of long lives. Leveraging the extensive data compiled by the World Bank, it seeks to provide a series of interactive tools that allow an evidence-based understanding of ways to improve life expectancy and deliver actionable insights to policymakers and researchers.

1.2 Objective

For this project, we developed a Shiny app and an R package called `WorldbankAnalysis` that allow users to explore life expectancy. Using data from all countries and across multiple years, these tools enable users to identify key socioeconomic factors influencing life expectancy and leverage machine learning models to predict it. The project aims to provide actionable insights for researchers and policymakers to better understand and address disparities in population health.

1.3 Existing Work

The `wbstats` package [Johnston and Warnes, 2020], available at `https://github.com/gshs-ornl/wbstats`, provides convenient access to a wide range of World Bank indicators, enabling users to quickly retrieve country-level data for various analytical tasks. This package streamlines data retrieval and offers a comprehensive repository of indicators, making it an invaluable tool for researchers seeking easy access to World Bank data. However, the primary focus of `wbstats` is on data acquisition, offering users the ability to download and structure data but leaving tasks such as data cleaning, analysis, and modeling largely up to the user. While this design is suitable for many use cases, it inherently limits the scope of the package to data retrieval rather than providing an integrated framework for deeper analytical tasks.

1.4 Variable Selection

When selecting predictor variables for life expectancy, our primary focus was to identify factors that are both influential and modifiable through policy interventions or strategic planning—unlike genetic factors, which are largely fixed. To achieve this, we grouped potential predictors into broad, actionable categories—demographic, economic, environmental, and social—and chose a subset from each that carried the greatest significance for public health outcomes. By leveraging these categories and the specific indicators within them, policymakers, healthcare planners, and international organizations can more effectively target interventions—such as improving healthcare delivery, fine-tuning economic strategies, and reducing environmental pollutants—to foster better health outcomes and ultimately increase life expectancy.

1.4.1 Demographic Variables

Population: Larger, denser populations can strain healthcare systems and resources, but they can also drive economic growth and infrastructure development, ultimately enhancing healthcare access and longevity [Graves Jr. and Mueller, 1993, Novak and Others, 2016]. For stakeholders, this variable is significant because demographic adjustments in resource distribution, urban planning, and healthcare funding can yield tangible improvements in public health.

Infant Mortality: As a sensitive metric of healthcare quality and socio-economic conditions, reducing infant mortality reflects immediate improvements in maternal-child health services [Wirayuda, 2021]. Stakeholders can use this indicator to prioritize interventions—such as enhanced prenatal care and better nutrition—that have rapid and long-lasting payoffs in extended life expectancy.

1.4.2 Economic Variables

GDP and GDP Per Capita: These measures of economic well-being correlate with better healthcare infrastructure and living standards [Swift, 2011, Țarcă and Others, 2024, Amjad and Khalil, 2014]. Increasing economic prosperity gives stakeholders leverage to invest in healthcare reforms that directly translate to longer, healthier lives.

Unemployment: High unemployment signals reduced income and healthcare access, alerting policymakers to the need for job creation and social safety nets as a strategic route to improving health outcomes [Assari, 2019].

Inflation: As rising prices diminish purchasing power, ensuring affordability of basic needs becomes vital. Managing inflation can help stakeholders maintain access to quality nutrition and healthcare, thereby safeguarding population health [Movsisyan and Others, 2024].

Exports: A robust export sector can fund public health initiatives and infrastructure improvements, offering a clear economic lever for stakeholders to boost life expectancy [Shah and Others, 2021].

1.4.3 Environmental Variables

PM2.5: High levels of air pollution lead to respiratory and cardiovascular issues. By regulating environmental policies and pollution control, stakeholders can directly address a modifiable risk factor that improves overall health and longevity [Qi and Others, 2020].

1.4.4 Social Variables

Education Expenditure: Investment in education enhances health literacy, leads to healthier lifestyles, and promotes socio-economic mobility—all key factors in improving health outcomes and life expectancy [Luy and Others, 2019, Novak and Others, 2016].

Undernourishment: The percentage of the population facing undernourishment directly reflects dietary quality and food security. Reducing undernourishment improves overall health, strengthens immune responses, and increases life expectancy [Kabir, 2008, Djoumessi, 2022].

Health Expenditure: Higher health expenditure as a percentage of GDP typically indicates more robust healthcare infrastructure and widespread access to medical services, both of which are instrumental in reducing mortality rates and extending life spans [Raeesi and Others, 2018].

1.5 Machine-Learning Methods and Models

1.5.1 Principal Component Analysis (PCA)

Principal Component Analysis (PCA) is a widely used dimensionality reduction method in statistical analysis and machine learning. It transforms high-dimensional data into a lower-dimensional space while retaining as much variability (information) as possible.

Suppose we have different independent variables $X_1, X_2, \ldots, X_p$. Because they may correlate with each other, we may encounter the problem of the "curse of dimensionality," which can reduce computational efficiency and introduce redundancy among features. PCA reduces dimensionality by finding principal components (PCs), which are uncorrelated and ordered so that the first few explain most of the variation present in the original variables. Starting with the data matrix $\mathbf{X}$:

$$\mathbf{X} = \begin{bmatrix} x_{11} & x_{12} & \cdots & x_{1p} \\ x_{21} & x_{22} & \cdots & x_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ x_{n1} & x_{n2} & \cdots & x_{np} \end{bmatrix}$$

We estimate the variance-covariance matrix $\mathbf{\Sigma}$:

$$\mathbf{\Sigma} = \frac{1}{n}\mathbf{X}^\top\mathbf{X}$$

PCA replaces the set of initial variables with their linear combinations, creating new variables called principal components. These are expressed as:

$$\mathbf{Z} = \mathbf{XW}$$

where $\mathbf{W}$ is the matrix of eigenvectors of $\mathbf{\Sigma}$, and $\mathbf{Z}$ represents the matrix of principal components [Maćkiewicz and Ratajczak, 1993].

1.5.2 Linear Regression

Multiple linear regression models the relationship between a dependent variable and multiple predictors. For life expectancy, the model is expressed as:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_p X_p + \epsilon$$

where Y is the response variable, β_0 is the intercept, $\beta_1, \ldots, \beta_p$ are the coefficients, and ϵ is the error term. The coefficients are estimated by minimizing the residual sum of squares (RSS):

$$\text{RSS} = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2$$

Linear regression is intuitive and straightforward but is sensitive to multicollinearity, assumes linear relationships, and can be influenced by outliers [Izenman, 2008].

1.5.3 Random Forest

Random Forest is a machine learning algorithm based on decision trees. It creates a large number of decision trees during training and combines their predictions (via averaging for regression) to improve accuracy and robustness. For predicting life expectancy using socioeconomic predictors, Random Forest handles multicollinearity well and captures non-linear relationships that ordinary regression models cannot. Proper tuning ensures robustness to overfitting and allows the estimation of predictor importance [Breiman, 2001].

1.5.4 Support Vector Machines (SVM)

Support Vector Machines (SVM) are supervised learning algorithms effective for regression tasks (Support Vector Regression, SVR). SVMs use kernel functions to map input features into higher-dimensional spaces, enabling the modeling of complex relationships. For life expectancy prediction, SVM handles non-linear relationships effectively with kernel functions such as the radial basis function (RBF). However, SVMs are computationally intensive, sensitive to parameter selection, and can struggle with imbalanced datasets [C. et al., 2020].

1.5.5 Gradient Boosting Machine (GBM)

Gradient Boosting Machines (GBM) combine multiple weak learners (decision trees) iteratively to create a strong predictive model. The goal is to minimize a loss function $L(y, f(x))$:

$$\hat{f}(x) = \arg\min_{f} E[L(y, f(x))|x]$$

For regression tasks, the squared loss function or the Huber loss function is commonly used. GBM effectively captures non-linear dependencies but requires careful tuning due to its high memory consumption and computational demands [Natekin and Knoll, 2013].

1.5.6 Evaluation Metrics

Root Mean Squared Error (RMSE): RMSE measures the difference between predicted values $\hat{y}_i$ and actual values y_i:

$$\text{RMSE} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}$$

A lower RMSE indicates better predictive accuracy.

Coefficient of Determination (R^2): R^2 evaluates the proportion of variance in the dependent variable explained by the model:

$$R^2 = 1 - \frac{\text{RSS}}{\text{TSS}}$$

where RSS is the residual sum of squares, and TSS is the total sum of squares. An R^2 closer to 1 indicates a better fit [James et al., 2013].

1.6 Data Source

For this study, we extracted our data from the World Bank's comprehensive online database [wor, 2024], which compiles global indicators covering a wide range of socio-economic, demographic, environmental, and health-related variables. This database is recognized for its reliability, transparency, and extensive geographic and temporal coverage, enabling comparisons across a large number of countries and regions over multiple years.

2 Methodology

2.1 Programming Paradigms

The analytical workflow implemented in this study is built upon integrating multiple programming paradigms, each contributing to different aspects of the analysis. These paradigms include functional programming, parallel computing, and machine learning, which ensure an efficient, modular, and scalable approach to processing and analyzing World Bank indicator data.

2.1.1 Programming Paradigms

Functional programming principles are employed throughout the package to ensure the workflow is modular, reusable, and easy to debug. Key functions, such as `fetch_paginated_data` , `reshape_to_long` , `reshape_to_wide` , and `impute_long_data` , are designed as self-contained, composable units. Each function takes inputs, processes them, and produces outputs without relying on side effects. For instance, `reshape_to_long` cleanly converts wide-format data into long-format while separating variables and years, allowing subsequent functions to work seamlessly with the transformed data. Similarly, `impute_long_data` uses linear interpolation and mean imputation to handle missing values while preserving the integrity of the data structure.

2.1.2 Parallel Computing Paradigms

To handle the computational demands of fetching large-scale indicator data from the World Bank API, the function for extracting data incorporates parallel computing paradigms. The `fetch_and_process_indicators` function uses the parallel package in R to create multiple worker nodes, enabling concurrent data retrieval from different API endpoints. This significantly reduces runtime when fetching data from various indicators. By dynamically exporting necessary functions and libraries to each cluster node, the parallel processing framework ensures smooth execution and compatibility across all cores.

2.1.3 Machine Learning Paradigms

The predictive modeling components of the package are grounded in machine learning paradigms. The `run_analysis` function integrates data preparation, dimensionality reduction using principal component analysis (PCA), and training multiple machine learning models. Supported models include linear regression, random forests, support vector machines (SVM), and gradient boosting machines (GBM). Hyperparameter tuning and cross-validation are incorporated to optimize model performance. The package also provides mechanisms for evaluating models using RMSE and R-squared metrics, with results visualized through scatter plots comparing predicted and actual values. This end-to-end machine learning workflow ensures robust predictions and clear insights into the relationship between indicators and target variables, such as life expectancy.

3 Sample Data Analysis

3.1 Process Method

Using the extracted CSV file, exploratory data analysis (EDA) was conducted to understand and preprocess the global, country-level indicator data. Missing values were quantified for each variable, and any variable with more than 50% missingness was excluded to ensure a sufficient baseline of observed data for imputation. The dataset was transformed from a wide to a long format to facilitate time-aware imputation by splitting year-specific variable columns into separate "variable" and "year" fields. Missing values within each country-variable group were interpolated linearly across years, leveraging temporal trends for more accurate estimates. Remaining gaps were filled using the global mean for each variable as a fallback strategy. The dataset was then reverted to a wide format to ensure ease of downstream analysis.

Descriptive statistics (mean, median, max, min, and standard deviation) were calculated for each variable to summarize central tendencies and variability. A correlation matrix was generated and visualized as a heatmap to explore the relationships between predictor variables and life expectancy. GDP per capita, health expenditure, infant mortality rate, and undernourishment rate showed strong correlations with life expectancy and were selected for further exploration. Scatterplots of these variables against life expectancy confirmed the trends identified in the correlation matrix and allowed for a more detailed analysis of their relationships.

To prevent data leakage, only one column representing life expectancy for a specific year (e.g., 2022) was retained as the target variable, and other life expectancy columns were removed. The dataset was then split into training and testing subsets. PCA transformation was applied to the training predictors, retaining 95% of the variance to reduce dimensionality. Multiple regression and machine learning models—linear regression, random forest, SVM, and gradient boosting—were trained on the PCA-transformed data using five-fold cross-validation. The best-performing models were evaluated on the test set using RMSE and R^2 metrics. Visualization of predicted versus actual values provided insights into model performance. This workflow ensured data quality, extracted meaningful relationships, and identified robust predictive models for life expectancy.

3.2 Analysis Results

3.2.1 Correlation Matrix

The correlation matrix visualized the relationships between the predictor variables and life expectancy. Strong correlations were identified, including:

- GDP per capita ($\rho = 0.49$)
- Health expenditure ($\rho = 0.55$)
- Infant mortality rate ($\rho = -0.86$)
- Undernourishment rate ($\rho = -0.66$)

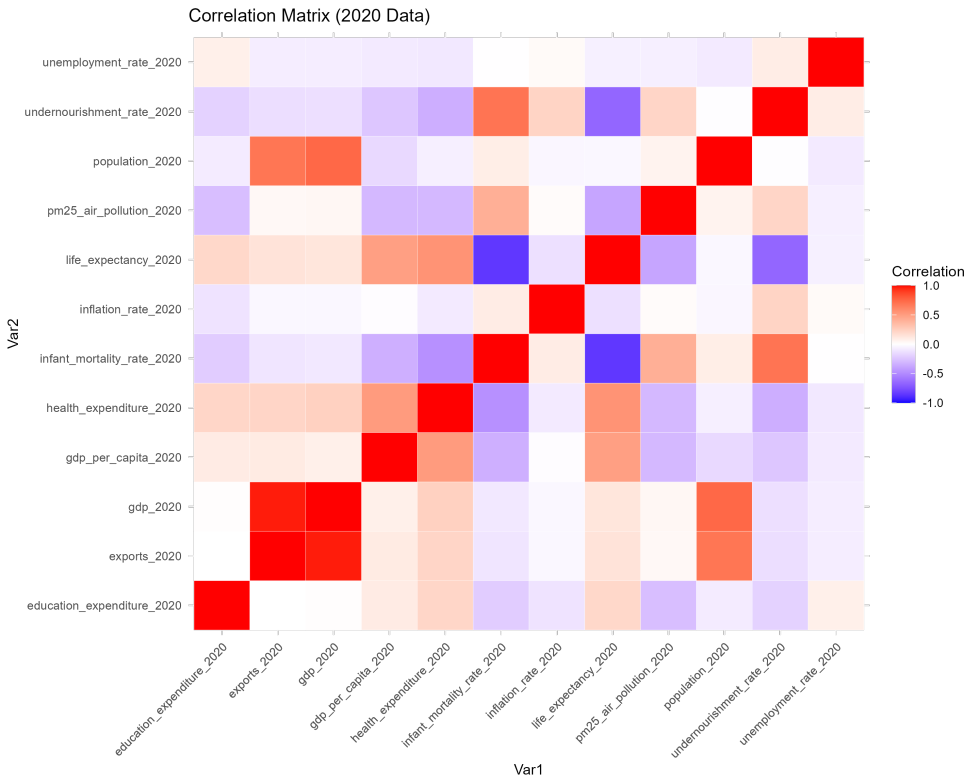

Figure 1: Correlation matrix of predictors and life expectancy (2020, imputed).

3.2.2 Scatterplots of Correlated Variables

Scatterplots were generated to explore the relationships between the four strongly correlated variables and life expectancy. An interactive version of the scatterplots is available online at `https://rpubs.com/ryan_wang19/1255223`.

Figure 2: Relationships between life expectancy and key predictors (2020).

Observations:

- GDP per capita has a strong positive correlation with health expenditure.
- The undernourishment rate has a strong positive correlation with the infant mortality rate.
- Countries with a GDP per capita above the global average generally have a life expectancy above 70. However, further increases in GDP per capita do not consistently lead to higher life expectancy.

3.2.3 Principal Component Analysis (PCA)

The PCA scree plot illustrates the proportion of variance explained by each principal component. The first two principal components (PCs) capture most of the variation in the data, making them critical for dimensionality reduction.

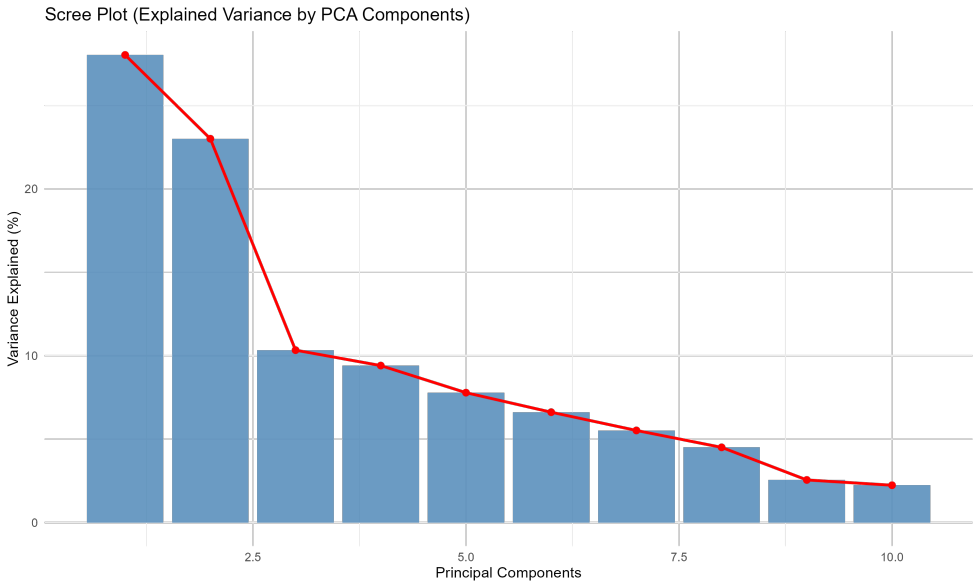

Figure 3: PCA scree plot showing the variance explained by each principal component.

3.2.4 Model Performance

Predicted versus actual values for life expectancy were visualized to evaluate the performance of the trained models. All four models showed good performance, with predicted values lying close to the diagonal line.

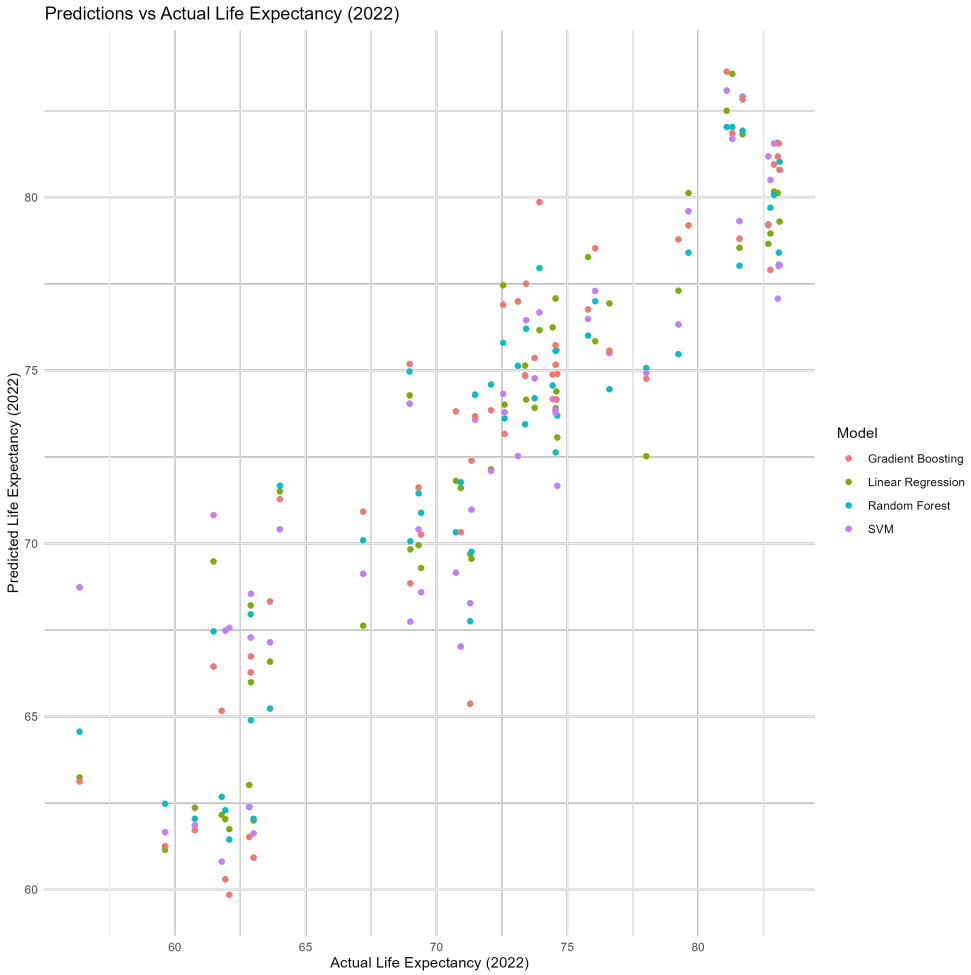

Figure 4: Predicted versus actual life expectancy values across models.

4 Dashboard - Shiny App

4.1 Overview

An interactive dashboard was developed using `flexdashboard` and `shiny` to provide a dynamic interface for model exploration and performance assessment. This dashboard directly leverages the imputed, cleaned, and integrated dataset produced in earlier steps, as well as insights gained from the exploratory data analysis. The primary goal of the dashboard is to enable users to visually and interactively examine how changes in modeling parameters, such as selected predictors, train-test splits, and model types, affect performance, interpretability, and stability. Additionally, the dashboard allows users to explore relationships between different predictor variables and life expectancy, providing theoretical insights into the key drivers of health disparities.

4.2 Dashboard Design and Functionality

The dashboard's layout was designed to allow intuitive parameter selection and immediate feedback. Users can select a target year (from 2015 to 2022) and choose which predictors to include in the model. Predictors are sourced from the previously imputed dataset, ensuring that missing values have already been addressed and that the chosen variables have sufficient coverage. A slider is provided to adjust the train/test split ratio, allowing the user to allocate different proportions of the data for model training and validation. Additionally, a checkbox enables optional application of Principal Component Analysis (PCA) to reduce dimensionality and mitigate multicollinearity by retaining 95% of the variance in the predictor set.

Once the user inputs their desired parameters, the dashboard automatically subsets the dataset to include only the selected predictors and target variable for the specified year. Rows with any remaining missing values in the chosen variables are excluded to ensure that each model is trained on a fully observed subset. If PCA is chosen, the selected predictors in the training set are standardized and transformed into principal components. The same PCA transformation is then applied to the test set, ensuring that the model evaluation occurs on the same transformed scale.

4.3 Model Selection and Training

To facilitate robust comparisons, users can select multiple modeling techniques simultaneously, including linear regression, random forest, support vector machines (SVM), and gradient boosting. Allowing users to select different statistical models in the dashboard enhances its flexibility and usefulness by enabling them to compare the performance and suitability of various approaches for the data. Different models have unique strengths—some prioritize accuracy, while others emphasize interpretability. By offering this choice, the dashboard supports diverse user needs, such as exploring complex relationships, understanding trade-offs, and selecting the most appropriate model for specific goals.

Each selected model is trained on the training portion of the data using a standardized approach. Five-fold cross-validation is employed to ensure that hyperparameters are tuned, and that the models' generalizability is estimated before final evaluation on the test set. For models requiring preprocessing (e.g., centering and scaling), these steps are integrated into the training workflow, either prior to PCA or as part of the modeling pipeline.

4.4 Performance Evaluation and Visualization

After the models are trained, their performance is automatically evaluated on the test set. The dashboard computes the Root Mean Squared Error (RMSE) and R^2 for each model and displays these metrics in a comparative table. Additionally, visual plots show the relationship between predicted and actual life expectancy values, enabling quick assessment of model bias, variance, and overall predictive accuracy. A correlation matrix of the selected predictors provides immediate insight into potential redundancy or relationships among variables, and can help the user decide when PCA might be beneficial.

For PCA-based analyses, the dashboard also offers a variable importance visualization. This step involves extracting PCA loadings and mapping the model's feature importance (in principal component space) back to the original variables, illustrating which features contribute most strongly to the predictive signals uncovered by the model. Such interpretability aids in understanding the underlying data structure and informs subsequent feature engineering or variable selection decisions.

4.5 Dynamic Experimentation

The dashboard environment allows iterative experimentation. Users may adjust year, variables, model choices, and preprocessing steps and instantly observe changes in correlation patterns, model performance, and feature importance. This dynamic approach extends beyond the static exploratory data analysis (EDA) and modeling steps described earlier, offering a richer, more flexible means of exploring how methodological choices influence outcomes. In doing so, the dashboard provides a platform for both method selection and result interpretation, ultimately enhancing the transparency, usability, and rigor of the analysis workflow.

5 R Package

5.1 Overview

The `WorldbankAnalysis` package is a comprehensive tool designed to streamline the workflow for analyzing World Bank indicator data. It integrates data retrieval, preprocessing, imputation, and predictive modeling into a seamless framework, allowing users to focus on extracting insights rather than managing tedious technical details. By leveraging functional programming, parallel computing, and machine learning paradigms, this package ensures scalability, reproducibility, and efficiency.

5.2 Process Usage of the Package

The workflow begins with extracting World Bank data using the `fetch_merge_indicators` and `fetch_and_process_indicators` functions. These steps ensure that data from multiple indicators is collected and merged into a single dataset suitable for analysis. Next, missing values in the dataset are handled using the `impute_data` function, which applies both linear interpolation and global mean imputation. Finally, the entire analytical pipeline, including data preparation, dimensionality reduction, model training, evaluation, and visualization, is executed using the `run_analysis` function. Each of these functions is modular, making the workflow adaptable to different research objectives.

5.3 Function Details

The `WorldbankAnalysis` package consists of three major functions: `fetch_merge_indicators`, `impute_data`, and `run_analysis`. These functions serve as the backbone of the package and are supported by utility functions that handle specific tasks. Each major function is designed to perform a critical role in the workflow, ensuring a streamlined and efficient analysis process. The package's design emphasizes scalability, reproducibility, and user-friendliness, making it an invaluable tool for handling large-scale datasets.

The `fetch_merge_indicators` function is the entry point for data retrieval and merging. It allows users to fetch data for multiple World Bank indicators by providing a list of API URLs. This function integrates parallel computing capabilities through the `fetch_and_process_indicators` utility to fetch indicator data concurrently, significantly reducing runtime when handling large datasets. Additionally, the `fetch_paginated_data` utility retrieves paginated data from World Bank API endpoints, ensuring that all available data is fetched regardless of dataset size. Once the data is retrieved, `fetch_merge_indicators` renames columns dynamically to reflect the specific indicators and merges the datasets into a single, wide-format structure. This consolidated dataset includes all the indicators for the specified years, preparing it for further preprocessing and analysis.

The `impute_data` function addresses the common issue of missing values in large datasets. It begins by filtering out variables with excessive missingness and then reshapes the dataset into a long format using the `reshape_to_long` utility. This transformation facilitates the application of imputation techniques. The `impute_long_data` utility performs two types of imputation: linear interpolation, which fills gaps within a time series for each country and variable, and global mean imputation, which replaces any remaining missing values with the global mean of the respective variable. After imputation, the `reshape_to_wide`

utility reshapes the data back into a wide format, preserving its compatibility with downstream analysis. This process ensures that the dataset is complete and retains its integrity, enabling robust and reliable modeling.

The `run_analysis` function provides a comprehensive analytical workflow that integrates data preparation, dimensionality reduction, model training, evaluation, and visualization. It begins by preparing the data with the `prepare_data` utility, which selects relevant predictors and the target variable while ensuring that only numeric variables are included. Dimensionality reduction is achieved through the `perform_pca` utility, which applies Principal Component Analysis (PCA) to reduce the number of variables while retaining a specified proportion of variance. Once the data is prepared, multiple machine-learning models are trained using the `train_models` utility. These models include linear regression, random forests, support vector machines, and gradient boosting machines, with hyperparameter tuning incorporated for GBM to optimize performance. The `evaluate_models` utility then evaluates the trained models using metrics such as RMSE and R-squared, providing both evaluation results and predictions. Finally, the `plot_results` utility generates scatter plots to visualize the relationship between actual and predicted values, allowing users to assess the accuracy of each model visually. This function encapsulates the entire workflow, making it a powerful tool for predictive modeling and analysis.

6 Discussion

6.1 Usability and Documentation

The presented dashboard is designed to be user-friendly. Specifically, it allows users to freely select variables, years, and modeling approaches through well-labeled inputs such as dropdown menus and slider controls. Moreover, visual outputs—including correlation matrices, predicted vs. actual plots, and feature importance charts—are presented in a coherent format that aids quick interpretation. By integrating comprehensible labels, intuitive controls, and easily interpretable plots, this dashboard offers both high usability and adequate documentation that can guide decision-makers, researchers, and other stakeholders in exploring the data and extracting meaningful insights from the results.

In addition, the `WorldbankAnalysis` package has been designed with a strong emphasis on usability and comprehensive documentation, ensuring that it is accessible to both novice and advanced users in data analysis. One of the key features that enhance usability is the detailed vignette, accessible online at `https://ggui6809.github.io/WorldbankAnalysis/articles/introduction.html`. This vignette provides a step-by-step guide that demonstrates the package's functionality, from data retrieval and preprocessing to missing value imputation and predictive modeling.

The functions within the package are thoughtfully designed with intuitive inputs and outputs, minimizing the learning curve for users. For example, the `fetch_merge_indicators` function requires only a list of World Bank indicator URLs, the number of processing cores, and a year range to efficiently retrieve and merge datasets. Similarly, the `impute_data` function simplifies missing data handling by automatically performing linear interpolation and global mean imputation, which are commonly used methods in time-series data analysis. The outputs of all functions are returned in consistent and well-documented formats, ensuring compatibility with standard R workflows.

The package documentation is extensive, with each function accompanied by clear descriptions of its purpose, parameters, and return values. Example code snippets are included to illustrate usage, enabling users to quickly understand how to implement the functions in their analyses. Additionally, the inclusion of a vignette as part of the package ensures that users have a comprehensive reference for the end-to-end analysis workflow. This focus on documentation and ease of use not only makes the package suitable for individual researchers but also for integration into larger, collaborative projects.

Overall, the `WorldbankAnalysis` package achieves a balance between functionality and user-friendliness. By providing clear documentation and practical examples, it empowers users to effectively analyze World Bank data without requiring extensive technical expertise.

6.2 Originality and Complexity

This project addresses a complex statistical and programming challenge by going beyond data aggregation to create a fully integrated framework for analyzing life expectancy. Unlike the `wbstats` package, which focuses on data acquisition, our framework combines data retrieval, preprocessing, exploratory analysis, predictive modeling, and visualization into a seamless pipeline. This approach not only simplifies workflows but also provides actionable

insights for stakeholders.

By automating traditionally manual processes such as data cleaning and interpretation, and integrating advanced techniques like machine learning and Principal Component Analysis (PCA), the project ensures scalability and usability. The framework's originality lies in its ability to transform raw socioeconomic data into evidence-based insights that inform policy and address health disparities, bridging the gap between data retrieval and decision-making. This combination of creativity and technical rigor demonstrates the project's non-trivial and innovative nature.

7 Limitation

7.1 Limitations of the Dashboard

One limitation of the current dashboard is the inability to simulate "what-if" scenarios by modifying predictor values to explore their impact on life expectancy outcomes. For instance, users cannot adjust key predictors, such as healthcare expenditure, to 20% to observe how these changes might influence predicted life expectancy. This restricts the tool's utility for decision-makers who may want to evaluate potential policy interventions or resource allocation strategies. Without the ability to simulate outcomes under hypothetical conditions, the dashboard primarily serves as a tool for retrospective analysis and model performance assessment rather than a proactive decision-support system for planning and policy formulation. Adding this capability would significantly enhance the dashboard's relevance for evidence-based decision-making.

7.2 Limitations of the R Package

There are two limitations for `WorldbankAnalysis` package. First, the package relies heavily on the availability and quality of data from the World Bank API. Inconsistent data coverage or missing values in certain indicators and regions can limit the scope of analyses, particularly when working with time series or global comparisons. Although the package includes imputation methods to address missing data, the accuracy of imputed values is inherently dependent on the data's underlying structure and assumptions, which may not always reflect the true patterns in real-world datasets.

Second, the package's machine-learning functionality is constrained by the default selection of models and hyperparameters. Although users can experiment with a variety of models, such as linear regression, random forest, support vector machines, and gradient boosting, advanced users might find the lack of full customization options—for example, specifying custom model architectures or hyperparameter grids—limiting for specialized applications.

8 Conclusion

In this project, we developed an integrated framework for analyzing life expectancy using World Bank data. This framework includes the `WorldbankAnalysis` R package and an interactive Shiny dashboard, both designed to streamline the workflow from data acquisition to actionable insights. By leveraging a combination of robust statistical techniques, machine learning models, and dynamic visualization tools, we provide researchers and policymakers with the ability to uncover the socioeconomic factors influencing life expectancy and to make informed decisions aimed at improving public health outcomes.

The `WorldbankAnalysis` package offers a cohesive set of tools for retrieving, preprocessing, and modeling large-scale datasets. Its modular design, along with comprehensive documentation and examples, ensures accessibility for a wide range of users. Meanwhile, the Shiny dashboard enables users to explore the relationships between predictors and life expectancy interactively, evaluate model performance, and compare different machine learning methods in real-time.

Our framework goes beyond traditional data aggregation by incorporating analytical techniques that transform raw data into meaningful insights. These tools highlight disparities in life expectancy across countries and provide evidence-based recommendations for addressing these inequalities. By integrating functional programming, parallel computing, and machine learning paradigms, we ensure that the methodology is scalable, efficient, and adaptable to diverse research objectives.

Despite its contributions, the project has limitations, such as the dependency on the quality and coverage of World Bank data and the lack of advanced customization for machine learning models. Future improvements could address these issues by incorporating additional data sources, enabling more flexible modeling options, and adding "what-if" scenario simulation capabilities to the dashboard.

Overall, this project provides a novel and comprehensive approach to studying life expectancy, empowering users to make data-driven decisions that promote health equity and advance public health research. We hope this framework will serve as a valuable resource for stakeholders seeking to understand and improve global health outcomes.

Acknowledgments

We would like to express our deepest gratitude to Professor Stephanie Hicks for her invaluable guidance and support throughout this project. Her expertise and insights greatly enriched our understanding and helped shape the direction of our work.

We also extend our sincere thanks to the teaching assistants, Joe Sartini and Wenxuan Lu, for their dedication in providing detailed feedback on our assignments and presentations, as well as their continuous support throughout the course.

This project was developed as part of the **Statistical Programming Paradigms and Workflows** course, where we learned foundational and advanced skills that directly contributed to the success of our work. Through this course, we gained proficiency in configuring statistical programming environments, writing advanced R and Python code, and leveraging tools such as SQL and `tidyverse` for data analysis. Additionally, the course provided us with the knowledge to build and organize software packages with thorough documentation, develop code using functional programming paradigms, access and process data from APIs, and design interactive web applications using `Shiny`. These skills were essential in enabling us to develop the `WorldbankAnalysis` R package and the accompanying `Shiny` dashboard, which serve as the backbone of our analysis framework.

We are deeply grateful for the opportunity to apply these skills to a meaningful project that combines statistical programming, machine learning, and interactive data exploration.

References

World bank data, 2024. URL https://data.worldbank.org/. Available at https://data.worldbank.org/.

Ahmed Amjad and Rehman Khalil. Economic prosperity and life expectancy, 2014. URL https://mpra.ub.uni-muenchen.de/70871/.

Shervin Assari. Unemployment and health outcomes. *Journal of Health Disparities*, 2019. URL https://pmc.ncbi.nlm.nih.gov/articles/PMC6392452/.

Clare Bambra. Health inequalities in a global context: Exploring socioeconomic disparities. *Health and Place*, 75:102761, 2022. doi: 10.1016/j.healthplace.2022.102761.

Mel Bartley and Michelle Kelly-Irving. *Health Inequality: An Introduction to Concepts, Theories and Methods*. Polity Press, 2016.

Vlatka Bilas, Sanja Franc, and Marina Bošnjak. Life expectancy and mortality rates in population health. *PubMed*, 2014. URL https://pubmed.ncbi.nlm.nih.gov/24851591/.

Leo Breiman. Random forests. *Machine Learning*, 45:5–32, 2001.

Jair C., Farid G., Lisbeth R., and Asdrubal L. A comprehensive survey on support vector machine classification: Applications, challenges and trends. *Neurocomputing*, 408:189–215, 2020.

Göran Dahlgren and Margaret Whitehead. The dahlgren-whitehead model of health inequalities: A retrospective on public health. *Public Health*, 199:14–21, 2021. doi: 10.1016/j.puhe.2021.08.009.

Evelyne Djoumessi. Undernourishment and mortality. *Nutrition Studies*, 2022. URL https://www.sciencedirect.com/science/article/abs/pii/S0899900722001733.

James Graves Jr. and Susan Mueller. Population density and health care. *Springer Journal*, 1993. URL https://link.springer.com/article/10.1007/BF01435991.

Alan J. Izenman. *Modern Multivariate Statistical Techniques: Regression, Classification, and Manifold Learning*. Springer, 2008.

Gareth James, Daniela Witten, Trevor Hastie, and Robert Tibshirani. *An Introduction to Statistical Learning: with Applications in R*. Springer, 2013.

Christopher H. Johnston and Gregory M. Warnes. wbstats: A package to access world bank data. GitHub repository, 2020. URL https://github.com/gshs-ornl/wbstats. Available at https://github.com/gshs-ornl/wbstats.

Nahid Kabir. Undernourishment and health. *Journal of Health Studies*, 2008. URL https://www.jstor.org/stable/40376184.

Marc Luy and Others. Education expenditure and health outcomes. *Genus*, 2019. URL https://genus.springeropen.com/articles/10.1186/s41118-019-0055-0.

Andrzej Maćkiewicz and Waldemar Ratajczak. Principal components analysis (pca). *Computers & Geosciences*, 19(3):303–342, 1993.

Ani Movsisyan and Others. Inflation and public health. *Public Health Policy*, 2024. URL https://www.sciencedirect.com/science/article/pii/S2214109X24001335.

Alexey Natekin and Alois Knoll. Gradient boosting machines, a tutorial. *Frontiers in Neurorobotics*, 7, 2013. doi: 10.3389/fnbot.2013.00021.

United Nations. Ensure healthy lives and promote well-being for all at all ages, 2022. URL https://www.un.org/sustainabledevelopment/health/.

Dan Novak and Others. Population impacts on health. *International Journal of Innovation*, 2016. URL https://www.inderscienceonline.com/doi/abs/10.1504/IJIL.2016.076673.

OECD. Life expectancy at birth, 2022. URL https://www.oecd.org/en/data/indicators/life-expectancy-at-birth.html.

Yuan Qi and Others. Pm2.5 and life expectancy. *PLOS Medicine*, 2020. URL https://journals.plos.org/plosmedicine/article?id=10.1371/journal.pmed.1003027.

Hamed Raeesi and Others. Health expenditure and mortality reduction. *Global Public Health*, 2018. URL https://pmc.ncbi.nlm.nih.gov/articles/PMC6108280/.

Mirza Shah and Others. Exports and health investments. *Health Economics*, 2021. URL https://pmc.ncbi.nlm.nih.gov/articles/PMC8079350/.

Rakesh Sharma. Life expectancy and its determinants: Evidence from plos one. *PLOS One*, 2018. URL https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0204940.

Richard Swift. Gdp and health outcomes. *Health Economics*, 2011. doi: 10.1002/hec.1590.

Prakoso Wirayuda. Infant mortality and healthcare quality. *Asia-Pacific Journal of Public Health*, 2021. URL https://journals.sagepub.com/doi/10.1177/1010539520983671.

Ioana Țarcă and Others. Gdp per capita and life expectancy. *Health Economics Review*, 2024. URL https://pmc.ncbi.nlm.nih.gov/articles/PMC11171643/.

A Appendix: Project Resources

Team GitHub Repository

The full project repository, including all files, scripts, and resources for the `WorldbankAnalysis` R package, sample analyses, and Shiny dashboard, is hosted on our team GitHub. Access it at:

`https://github.com/jhu-statprogramming-fall-2024/project4-error-404-team-not-found`

Package GitHub Repository

The `WorldbankAnalysis` R package is available as a standalone repository. It includes all functionalities for fetching, preprocessing, and analyzing World Bank data. Access it here:

`https://github.com/ggui6809/WorldbankAnalysis`

For installation and usage instructions, refer to the `README.md` file in the repository.