Check for updates

Sustainable Development

WILEY

Sustainable Development

RESEARCH ARTICLE OPEN ACCESS

Predicting Life Expectancy in Upper-Middle-Income Countries Using Machine Learning: A Comparative Analysis

Esra Süzen | Mehmet Kayakuş

Department of Biomedical Engineering, Faculty of Engineering, Akdeniz University, Antalya, Türkiye

Correspondence: Mehmet Kayakuş (mehmetkayakus@akdeniz.edu.tr)

Received: 9 April 2026 | **Revised:** 24 June 2026 | **Accepted:** 7 July 2026

Keywords: global health indicators | life expectancy | machine learning | predictive modeling | random forest | socio-economic determinants | support vector regression

ABSTRACT

This study aims to estimate life expectancy in upper-middle-income countries using socio-economic and health-related variables obtained from the World Bank's World Development Indicators database. To this end, four machine learning algorithms—Support Vector Regression (SVR), Random Forest, Gradient Boosting, and XGBoost—were applied; the models were compared using the 10-fold cross-validation method, employing the R^2, RMSE, and MAE metrics. The analyses revealed that life expectancy is strongly associated with per capita GDP, health expenditure, and the urbanization rate. In the model comparisons, the highest predictive performance was achieved by the SVR model. Despite the limited sample size, SVR successfully captured non-linear relationships thanks to its kernel-based structure, demonstrating both the lowest error values and the most stable residual distribution. While the Random Forest model also performed well, the Gradient Boosting and XGBoost models achieved relatively lower accuracy. The findings demonstrate that machine learning methods provide effective tools for modeling multidimensional health indicators and that the SVR and Random Forest algorithms constitute robust alternatives for such data structures. The study aims to contribute to data-driven decision-making processes in global health and development policies.

1 | Introduction

Life expectancy is regarded as one of the most fundamental indicators that holistically reflects a society's level of health, social welfare, and economic development. The increase in life expectancy recorded worldwide over the last century is closely linked to advances in medical technology, expanded access to healthcare, and improvements in living conditions. However, it is recognized that this increase has not occurred uniformly across countries and income groups and that the factors determining life expectancy, particularly in middle-income countries, exhibit a complex and multidimensional structure (Marmot 2005), which includes variables such as healthcare access, education levels, and economic stability. Consequently, the accurate modeling and forecasting of the socio-economic and health-related determinants of life expectancy are of critical importance for the development of effective public health policies.

According to the World Bank classification, upper-middle-income countries represent a transitional group in terms of economic development and health system capacity (Cabieses and Bird 2014; WB_Group 2024). While these countries possess a certain level of economic size and health infrastructure, they have not yet fully attained the diversity of resources and

This is an open access article under the terms of the Creative Commons Attribution License, which permits use, distribution and reproduction in any medium, provided the original work is properly cited.

© 2026 The Author(s). *Sustainable Development* published by ERP Environment and John Wiley & Sons Ltd.

institutional capacity offered by high-income countries (Cabieses and Bird 2014; Hosen et al. 2021). This situation leads to significant variations in indicators such as per capita income, health expenditure, urbanization rates, and demographic structure and consequently results in life expectancy varying considerably between countries (Aggarwal et al. 2015; Gaziano et al. 2010). Consequently, this heterogeneous structure necessitates the development of customized, high-accuracy forecasting models for upper-middle-income countries.

Traditional statistical and econometric methods have long been used as the primary analytical tools for examining the relationships between life expectancy and income levels, health expenditure, education, and other socio-economic indicators (Lago et al. 2018). However, these methods may be limited in their ability to capture non-linear relationships and multiple interactions between variables, which can lead to an incomplete understanding of how these factors collectively influence life expectancy, particularly in contexts where such interactions are significant, such as in low-income countries or among marginalized populations. Life expectancy, however, is the output of a complex system in which economic, demographic, and health-related variables interact with one another. Consequently, there is a need for more flexible and robust modeling approaches to analyze this multidimensional structure more accurately. Recently, machine learning methods have increasingly been adopted in health research and modeling studies involving socio-economic variables, driven by the proliferation of big data sources and increased computational capacity (Casey et al. 2016; Huang et al. 2018; McDermott et al. 2021; Schwalbe and Wahl 2020). Machine learning algorithms such as SVR, Random Forest, Gradient Boosting, and XGBoost offer significant alternatives to traditional statistical methods due to their ability to capture non-linear patterns, operate effectively on data structures containing many variables, and model complex relationships between variables (Chairunisa et al. 2024; Ghazali et al. 2018). Recent studies have demonstrated that these algorithms yield successful results in predicting multidimensional health indicators such as life expectancy and can deliver high predictive performance (Chairunisa et al. 2024; Lipesa et al. 2023; Vydehi et al. 2020). Recent research has further extended machine learning applications by integrating explainable artificial intelligence (XAI) frameworks to improve model transparency and policy relevance. For example, (Çelik et al. 2026) demonstrated that SHAP and LIME can uncover complex nonlinear structural relationships while enhancing interpretability in governance-related prediction tasks. Similarly, (Yang et al. 2025) applied multiple machine learning algorithms to identify life-expectancy-related risk factors and showed that feature-ranking approaches substantially improved model interpretability. (Tahri et al. 2026) also emphasized that machine learning methods often outperform conventional statistical approaches in multidimensional prediction problems while highlighting the importance of predictor ranking and explainability. However, it is noteworthy that the number of studies in the literature that focus on upper-middle-income countries as a distinct analytical category and comparatively examine the performance of multiple machine learning algorithms is limited (Ogle et al. 2023; Ozyilmaz et al. 2022). This limitation results in the socio-economic, demographic, and health system dynamics specific to the upper-middle-income group not being sufficiently disentangled; consequently, the contextual validity and generalizability of the resulting policy recommendations may also be constrained (Bansal and Jin 2023; Christie et al. 2018). To address this gap, this study estimates life expectancy in upper-middle-income countries using socio-economic and health-related variables obtained from the World Bank's World Development Indicators database. In this context, four different machine learning algorithms—Support Vector Regression (SVR), Random Forest, Gradient Boosting, and XGBoost—were applied, and the models' predictive performance was compared using the coefficient of determination (R^2), Root Mean Square Error (RMSE), and Mean Absolute Error (MAE) metrics. The main contribution of the study is its systematic analysis of the multidimensional determinants of life expectancy from a machine learning perspective and its provision of a data-driven forecasting framework specific to upper-middle-income countries. In this respect, the study not only evaluates model performance but also presents an analytical approach that could contribute to global health and development policies.

2 | Materials and Methods

This section describes the dataset, methods, and machine learning techniques used to estimate life expectancy in upper-middle-income countries. It covers the dataset preparation process, variable selection, data pre-processing, machine learning methods, and methods for evaluating the performance and errors of the models. Figure 1 shows the flowchart of the study.

As shown in Figure 1, the study begins with data collection from the World Bank's World Development Indicators (WDI) database. At this stage, the processes of variable identification, data collection, data preprocessing, and feature selection were carried out. The reliability of the data and the performance of the model were enhanced using the cross-validation method. A 10-fold cross-validation process was performed at this stage. In the machine learning phase, models were created using support vector regression (SVR), random forest, gradient boosting, and XGBoost methods. To determine the models' performance and error, metrics appropriate to the literature—namely, R^2, RMSE, and MAE—were used.

2.1 | Data Set

The dataset used in this study was obtained from the World Bank's World Development Indicators (WDI) database, one of the most widely used and reliable sources of global socio-economic and health-related statistics (World_Bank 2026). By providing comprehensive and standardized data for countries across the globe, the World Bank enables researchers to analyze development indicators and compare them across different income groups. In this study, data relating to upper-middle-income countries were extracted from the World Bank database and used to analyze and estimate life expectancy. The dataset consists of 24 upper-middle-income countries classified according to the World Bank income classification for the year 2024. Each country represents one observation, resulting in a total sample size of 24 observations. Seven predictor variables and one target

10991719, 0, Downloaded from https://onlinelibrary.wiley.com/doi/10.1002/sd.71468 by Mehmet KAYAKUS - Akdeniz University , Wiley Online Library on [17/07/2026]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

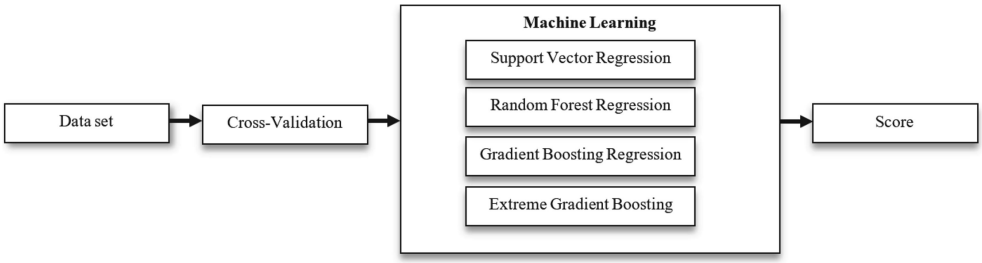


FIGURE 1 | Flowchart of the study.

variable (life expectancy at birth) were included in the analysis. Prior to preprocessing, the dataset contained missing values in a limited number of predictor variables. Missing numerical values accounted for less than 5% of the total dataset and were imputed using mean imputation to preserve sample size and data consistency.

The study's dependent variable is "life expectancy at birth" which represents the average number of years a newborn is expected to live under current mortality conditions. This variable is widely used as an important indicator of public health, living standards, and socio-economic development (Țarcă et al. 2024). In addition to the dependent variable, various demographic, economic, and health-related indicators obtained from the World Bank database have been included in the dataset as independent variables to explain and predict variations in life expectancy (Roffia et al. 2023).

Before proceeding to the machine learning analysis, a series of data preprocessing steps were applied to ensure the quality and consistency of the dataset. First, the dataset was examined to identify missing values and inconsistencies in the variables. To minimize information loss and preserve the structure of the dataset, missing numerical values were handled using the mean imputation method (Liew et al. 2011). Additionally, the data type compatibility of the variables was checked, and only numerical variables suitable for regression-based machine learning models were retained. As many machine learning algorithms are sensitive to differences in variable scales, the independent variables were standardized to ensure that each feature contributes equally to the model training process.

Although the dataset contains a limited number of observations, it is considered sufficient for comparative machine learning modeling when appropriate validation techniques are used. Ten-fold cross-validation was employed in this study to obtain reliable and unbiased model evaluation results. In this approach, the dataset is divided into approximately ten equal subsets. In each iteration, nine subsets are used to train the model, while the remaining subset is used to test the prediction performance. This process is repeated ten times, ensuring that each subset is used as a test set once. The final model performance is obtained by taking the average of the results from all iterations (Malakouti et al. 2023).

The 10-fold cross-validation method is widely recommended in machine learning research because it reduces the risk of overfitting and provides a more robust assessment of model performance compared to a single training-test split (Krstajic et al. 2014). By allowing each observation in the dataset to be used in both the training and testing phases, this method enables a more reliable and stable assessment of the predictive capabilities of the models used in this study.

2.2 | Machine Learning

Machine learning models, particularly ensemble learning algorithms, have become increasingly important in socio-economic and health-related forecasting studies (Ekundayo 2024). Socio-economic indicators such as life expectancy are influenced by a multitude of interacting factors, including access to healthcare, income levels, education, environmental conditions, and demographic characteristics (Abubakari et al. 2019; Wirayuda and Chan 2021). These variables typically exhibit complex non-linear relationships and interaction effects that are difficult to capture using traditional modeling approaches. Ensemble methods such as Random Forest, Gradient Boosting, and XGBoost combine multiple decision trees to improve predictive performance and reduce model variance. By aggregating the results of multiple learners, these methods can capture complex patterns in the data and provide more stable and reliable predictions (Xiao et al. 2018).

Among machine learning algorithms, gradient-boosting-based models have attracted significant attention due to their strong predictive capabilities. Gradient boosting constructs models sequentially; each new model attempts to correct the errors made by the previous models (Boldini et al. 2023). This iterative learning process enables the algorithm to gradually improve model accuracy and capture complex non-linear relationships between variables (Cai et al. 2022). XGBoost, an optimized implementation of gradient boosting, further enhances this framework by offering regularization techniques, efficient handling of missing values, and parallel computing capabilities (Kyrychek 2025). These features enhance both computational efficiency and predictive performance, making XGBoost one of the most widely used machine learning algorithms in predictive modeling studies across the fields of economics, public health, and the social sciences (Alim et al. 2020).

To evaluate the predictive performance of the models, three commonly used regression evaluation metrics were employed: Root Mean Square Error (RMSE), Mean Absolute Error (MAE), and the coefficient of determination (R^2). RMSE measures the square root of the average of the squared differences between the predicted and observed values and is sensitive to large prediction errors (Hodson 2022). MAE represents the average

10991719, 0, Downloaded from https://onlinelibrary.wiley.com/doi/10.1002/sd.71468 by Mehmet KAYAKUS - Akdeniz University , Wiley Online Library on [17/07/2026]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

absolute difference between predicted and actual values and provides an easily interpretable measure of prediction accuracy (Zhou et al. 2024). R^2 indicates the proportion of variance in the dependent variable that can be explained by the independent variables in the model (Plonsky and Ghanbar 2018). The use of these complementary evaluation metrics provides a comprehensive assessment of model performance and enables a reliable comparison between different machine learning algorithms.

Model hyperparameters were determined based on prior literature, preliminary experimental evaluation, and the characteristics of the dataset. Since the dataset contains a relatively limited number of observations, exhaustive hyperparameter optimization procedures such as Grid Search or Random Search were not applied to avoid excessive model complexity and potential overfitting. Instead, conservative parameter settings commonly recommended in previous studies were adopted to ensure a balance between predictive accuracy and model generalizability.

2.2.1 | Support Vector Regression (SVR)

Support Vector Regression (SVR) is a supervised machine learning method developed to predict continuous numerical outcomes (Montesinos López et al. 2022). It is derived from the principles of Support Vector Machines (SVMs), which were originally designed for classification problems but were later adapted for regression tasks (Valkenborg et al. 2023). SVR is particularly useful when the relationship between dependent and independent variables is not strictly linear, as it can model more complex patterns in the data using kernel functions (Najafzadeh and Niazmardi 2021). Consequently, SVR is widely used in studies involving economic forecasting, health indicators, and multi-dimensional forecasting problems (Qu et al. 2022).

Support Vector Regression (SVR) has been included because life expectancy is influenced by numerous socio-economic and health-related variables, the effects of which do not always follow a simple linear structure. Compared to classical linear regression, SVR offers a more flexible modeling framework by enabling the estimation of non-linear relationships (Klopfenstein and Vaiter 2021). Furthermore, it can perform well even when the dataset is relatively limited, making it suitable for the structure of the present study.

The primary aim of SVR is to find a function that estimates the dependent variable with the smallest possible error while maintaining the simplicity of the model. In general terms, the SVR function can be expressed as follows:

$$f(x) = w^T \emptyset(x) + b \quad (1)$$

where x denotes the input variables, w is the weight vector, b is the bias term, and $\phi(x)$ represents the transformation of the input variables into a higher-dimensional feature space. This transformation enables the model to capture nonlinear relationships that may not be visible in the original input space.

Unlike ordinary regression methods that aim to minimize the total prediction error directly, SVR uses an ε-insensitive loss function. This means that errors falling within a predefined margin, denoted by ε, are ignored. Only errors exceeding this margin are penalized. The optimization problem for SVR can be written as follows:

$$\min \frac{1}{2}\|w\|^2 + C\sum_{i=1}^{n}\left(\xi_i + \xi_i^*\right) \quad (2)$$

tabi olmak

$$y_i - \left(w^T\theta(x_i) + b\right) \le \left(\epsilon + \xi_i^*\right) \quad (3)$$

$$\left(w^T\phi(x_i) + b\right) - \left(y_i \le \epsilon + \xi_i^*\right) \quad (4)$$

$$\xi_i, \xi_i^* \ge 0 \quad (5)$$

In these equations, y_i represents the observed value, ε is the acceptable error margin, C is the penalty parameter, and ξ_i, ξ^*_i are slack variables that measure deviations outside the tolerance interval. The parameter C controls the trade-off between model complexity and prediction error. A larger C places greater emphasis on minimizing training error, whereas a smaller C allows a smoother decision function.

In this study, the SVR model was implemented using the radial basis function (RBF) kernel, one of the most preferred kernel types in non-linear regression problems. The RBF kernel is expressed as follows:

$$K\left(x_i, x_j\right) = \exp\left(-\gamma\left\|x_i - x_j\right\|^2\right) \quad (6)$$

Here, γ determines the effect of individual training observations. The RBF kernel was chosen because it can effectively capture the non-linear relationships between predictors and life expectancy without requiring the functional form of these relationships to be specified in advance.

In the Python implementation, the SVR model was configured with kernel = RBF, C = 100, and $\varepsilon = 0.1$. These parameter values were selected based on prior literature and preliminary experiments to achieve a balance between model flexibility and generalization. The choice of the RBF kernel ensures that the model accounts for non-linear patterns in the data, while the relatively high C value increases the model's sensitivity to predict errors during training. The $\varepsilon = 0.1$ setting defines a narrow error tolerance; this implies that predictions are expected to remain close to the actual observed values. Additionally, the independent variables were standardized prior to model training, as SVR is sensitive to differences in variable scales and performs more reliably when predictors are normalized.

SVR offers a robust alternative to classical regression approaches when dealing with potentially non-linear and multidimensional relationships. Its flexibility, robustness, and suitability for limited datasets make it an appropriate method for estimating life expectancy in upper-middle-income countries.

2.2.2 | Random Forest Regression

Random Forest (RF) is an ensemble machine learning method based on the combination of multiple decision trees (Becker

10991719, 0, Downloaded from https://onlinelibrary.wiley.com/doi/10.1002/sd.71468 by Mehmet KAYAKUS - Akdeniz University , Wiley Online Library on [17/07/2026]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

et al. 2023). Introduced by Leo Breiman, the Random Forest algorithm improves predictive performance by combining the outputs of many individual trees rather than relying on a single model (Breiman 2001). This ensemble structure helps to reduce overfitting and increase the robustness of the model, particularly when dealing with complex datasets involving non-linear relationships and interactions between variables (Zhou et al. 2023).

The Random Forest algorithm is particularly useful in regression problems where the relationship between dependent and independent variables is non-linear or difficult to model using traditional statistical approaches. Unlike classical regression methods, the Random Forest algorithm does not assume a specific functional form between variables (Shanmugasundar et al. 2021). Instead, it learns patterns directly from the data by creating multiple decision trees using different subsets of the training data. This flexibility makes the Random Forest algorithm an effective method for predicting socio-economic and health-related indicators such as life expectancy.

The Random Forest algorithm operates via two fundamental mechanisms: bootstrap sampling and feature randomization (Suthaharan 2016). First, multiple subsets of the training data are created using a technique known as bootstrap sampling; in this technique, samples are drawn from the original dataset at random and with replacement. A decision tree is then trained on each bootstrap sample. Second, at each node of the decision tree, a random subset of predictor variables is selected to determine the best split. This randomization helps to reduce the correlation between individual trees and improve the model's overall predictive accuracy (Salman et al. 2024; Wang et al. 2020).

In regression problems, the final prediction of the Random Forest model is obtained by taking the average predictions from all individual trees in the ensemble. This process can be expressed as follows:

$$\hat{y} = \frac{1}{T}\sum_{t=1}^{T} f_t(x) \tag{7}$$

where: $\hat{y}$ represents the predicted value, T denotes the total number of trees in the forest, $f_t(x)$ represents the prediction generated by the t_{th} decision tree.

Each decision tree in the forest is constructed by recursively partitioning the feature space to minimize the prediction error. The partitioning criterion commonly used in regression trees is the Mean Squared Error (MSE), which measures the variance within each node (Dewi and Chen 2019). The MSE can be expressed as follows:

$$MSE = \frac{1}{n}\sum_{i=1}^{n}\left(y_i - \hat{y}\right)^2 \tag{8}$$

The Random Forest model was implemented using the Random Forest Regressor algorithm available in the Python scikit-learn library. According to the configuration used in the analysis code, the model was trained with 200 decision trees and a maximum tree depth of 10. Increasing the number of trees generally improves prediction stability, while limiting the depth of the trees helps prevent excessive model complexity and reduces the risk of overfitting.

The use of the Random Forest algorithm in this study allows for the comparison of a robust ensemble-based machine learning method with both classical regression approaches and kernel-based methods. Thanks to its ability to model non-linear relationships and capture interactions between variables, Random Forest offers a robust alternative for predicting life expectancy in upper-middle-income countries.

The Random Forest model was trained with 200 trees and a maximum depth of 10, which were selected based on commonly used settings in the literature and preliminary performance evaluation.

2.2.3 | Gradient Boosting Regression

Gradient Boosted Regression (GB) is an ensemble machine learning technique that constructs predictive models by sequentially combining multiple weak learners, typically decision trees (Sahin 2020). Unlike ensemble methods based on bagging, such as Random Forests, where trees are created independently, Gradient Boosting constructs trees sequentially so that each new tree attempts to correct the prediction errors made by the previous trees (Syam and Kaul 2021). This iterative process enables the model to gradually improve its predictive performance and capture complex non-linear relationships between variables (Wahyuningsih et al. 2024).

Gradient boosting is particularly suitable for regression problems involving multi-dimensional datasets where the relationship between the predictor variables and the dependent variable may not follow a simple linear structure (Feng et al. 2018). In socio-economic and health-related studies, variables such as income level, access to healthcare, education, and environmental factors interact with one another in complex ways. Gradient boosting provides a flexible modeling framework capable of capturing these interactions and non-linear patterns in the data (Kalusivalingam et al. 2022).

The main idea behind the gradient boosting method is to minimize the loss function by iteratively adding models that approximately yield the negative gradient of the loss function. The loss function commonly used in regression problems is the Mean Squared Error (Sharma and Juneja 2018). The general form of the gradient boosting model can be written as follows:

$$\mathrm{F_m(x)} = \mathrm{F_{m-1}(x)} + \eta \mathrm{h_m(x)} \tag{9}$$

where $\mathrm{F_m(x)}$ represents the updated model at iteration m, $\mathrm{F_{m\text{-}1}(x)}$ is the model obtained from the previous iteration, $\mathrm{h_m(x)}$ denotes the newly fitted decision tree that models the residual errors, and η represents the learning rate controlling the contribution of each tree to the final model.

In (Equation 10), the algorithm follows the steepest descent path of the loss function at each iteration to minimize the errors produced by the current model.

$$\mathrm{h_m(x)} = -\frac{\partial L(\mathrm{y}, \mathrm{F(x)})}{\partial F(\mathrm{x})} \tag{10}$$

10991719, 0, Downloaded from https://onlinelibrary.wiley.com/doi/10.1002/sd.71468 by Mehmet KAYAKUS - Akdeniz University , Wiley Online Library on [17/07/2026]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

where $h_m(x)$ represents the weak learner fitted at iteration m, L(y, F(x)) denotes the loss function measuring the difference between the observed value *y* and the predicted value F(x), and the gradient of the loss function indicates the direction in which the model should be updated to reduce prediction error. This formulation demonstrates that the gradient boosting algorithm operates as an optimization method based on the gradient descent principle. At each iteration, a new model is fitted that approximately represents the negative gradient of the loss function — which represents the residual errors of the current model — leading to a gradual improvement in prediction performance.

In each iteration, the algorithm calculates the residuals from the previous model and applies a new decision tree to these values. Gradient boosting gradually improves prediction accuracy by focusing on the errors generated in previous stages. This sequential learning mechanism enables the model to capture complex interactions between variables that might not be detected by simpler regression techniques (Zhang et al. 2019).

A gradient boosting model was implemented using the Gradient Boosting Regressor algorithm found in the Python scikit-learn library. According to the configuration used in the analysis code, the model was trained with 200 trees, a learning rate of 0.05, and a maximum tree depth of 3. The number of predictors determines how many trees are created sequentially, while the learning rate controls how strongly each tree contributes to the final prediction model.

Gradient boosting was used in this study because it is a powerful ensemble method based on boosting that can model complex non-linear relationships between socio-economic and health-related variables. Its sequential error correction mechanism makes it particularly suitable for predicting life expectancy in upper-middle-income countries.

The hyperparameters (200 trees, learning rate = 0.05, max depth = 3) were selected based on preliminary experimentation and prior studies to reduce overfitting risk while maintaining predictive performance.

2.2.4 | Extreme Gradient Boosting (XGBoost)

Extreme Gradient Boosting (XGBoost) is an advanced implementation of the gradient boosting algorithm, designed to improve predictive performance and computational efficiency (Bentéjac et al. 2021). Developed by Chen and Guestrin (2016), XGBoost has become one of the most widely used machine learning algorithms in predictive modeling due to its scalability, pruning mechanisms, and ability to efficiently process complex datasets (Chen and Guestrin 2016).

Like the standard gradient boosting algorithm, XGBoost constructs predictive models by sequentially combining multiple decision trees; each tree attempts to minimize the prediction errors produced by the previous models (Ali et al. 2023). However, XGBoost offers various methodological improvements that distinguish it from traditional gradient boosting algorithms. These improvements include pruning techniques to control model complexity, efficient handling of missing values, parallel computing capabilities, and optimized tree learning procedures (Bentéjac et al. 2021).

The objective function of the XGBoost algorithm consists of two components: a loss function that measures the difference between predicted and observed values, and a regularization term that penalizes model complexity (Zhang et al. 2022). The objective function can be expressed as follows:

$$\text{Obj} = \sum_{i=1}^{n} L\left(y_i, \hat{y}_i\right) + \sum_{k=1}^{K} \Omega\left(f_k\right) \tag{11}$$

where $L(y_i, \hat{y}_i)$ represents the loss function measuring prediction error, f_k denotes the individual decision trees in the ensemble, and $\Omega(f_k)$ represents the regularization term controlling the complexity of the model. This regularization component helps prevent overfitting and enhances the algorithm's generalization ability.

Another key feature of XGBoost is its use of second-order gradient optimization, which incorporates both the first and second derivatives of the loss function during the training process. This lets the algorithm converge faster and optimize more accurately than traditional gradient descent methods (Wang et al. 2024).

The XGBoost model was implemented using the XGBRegressor algorithm available in the Python XGBoost library. According to the configuration used in the analysis code, the model was trained with 200 trees, a learning rate of 0.05, and a maximum tree depth of 4. These parameters were selected to strike a balance between model complexity and prediction performance while reducing the risk of overfitting.

XGBoost was included in this study as it represents one of the most powerful machine learning algorithms for predictive modeling. Its ability to capture non-linear relationships, incorporate regulatory factors, and process large datasets efficiently makes it particularly suitable for predicting life expectancy based on socio-economic and health-related indicators.

These hyperparameters were selected based on literature-guided tuning and preliminary experimental testing to balance model complexity and predictive stability.

2.2.5 | Performance Criteria

To evaluate the predictive performance of the machine learning models used in this study, three commonly used performance metrics were employed: the coefficient of determination (R^2), root mean square error (RMSE), and mean absolute error (MAE). These metrics measure the difference between observed and predicted values, providing complementary insights into the accuracy and reliability of the prediction models.

The coefficient of determination (R^2) is a statistical measure indicating how well the independent variables explain the variation in the dependent variable (Piepho 2019). In other words, it measures the proportion of the variance in the observed data that is explained by the prediction model (Piepho 2019). The R^2 value ranges from negative infinity to 1. A value close to 1 indicates

10991719, 0, Downloaded from https://onlinelibrary.wiley.com/doi/10.1002/sd.71468 by Mehmet KAYAKUS - Akdeniz University , Wiley Online Library on [17/07/2026]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

that the model explains a large proportion of the variation in the dependent variable, while values close to 0 indicate weak explanatory power (Sharma et al. 2019).

The R^2 metric can be calculated using the following formula:

$$R^2 = 1 - \frac{\sum_{i=1}^{n}\left(y_i - \hat{y}_i\right)^2}{\sum_{i=1}^{n}\left(y_i - \hat{y}\right)^2} \quad (12)$$

where: y_i represents the observed values, $\hat{y}_i$ represents the predicted values, $\hat{y}$ represents the meaning of the observed values, n denotes the total number of observations.

A higher R^2 value indicates that the model performs better and has a stronger predictive ability (Shmueli et al. 2019).

Root Mean Square Error (RMSE) measures the average magnitude of prediction errors by calculating the square root of the average of the squared differences between observed and predicted values (Hodson 2022). RMSE is particularly useful because, due to the squaring operation, it penalizes larger errors more heavily than smaller ones. Lower RMSE values indicate better model accuracy (Hodson 2022).

RMSE is calculated as follows:

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}\left(y_i - \hat{y}_i\right)^2} \quad (13)$$

where: y_i represents the actual observed value, $\hat{y}_i$ represents the predicted value, n is the number of observations.

As RMSE (Root Mean Square Error) is expressed in the same units as the dependent variable, it provides an interpretable measure of prediction error (Hodson 2022).

Mean Absolute Error (MAE) measures the average absolute difference between the observed and predicted values (Frias-Paredes et al. 2018). Unlike RMSE, MAE does not square the errors, making it less sensitive to outliers. Consequently, MAE provides a simple and easily interpretable measure of the average prediction error (Hodson 2022).

The MAE metric is calculated using the following formula:

$$MAE = \frac{1}{n}\sum_{i=1}^{n}\left|y_i - \hat{y}_i\right| \quad (14)$$

where: y_i represents the actual observed value, $\hat{y}_i$ represents the predicted value, n denotes the number of observations.

Lower MAE values indicate that the model exhibits better predictive performance (Kaliappan et al. 2021).

In this study, these three performance criteria were used together to provide a comprehensive assessment of the predictive capabilities of machine learning models. While R^2 measures how well the model explains the variability in the data, RMSE and MAE quantify the magnitude of prediction errors. The combined use of these metrics allows for a more reliable comparison of the predictive performance of models used to estimate life expectancy in upper-middle-income countries.

3 | Results

This section presents the results of the study, which aims to evaluate the predictive performance of machine learning models used to estimate life expectancy in upper-middle-income countries. Firstly, descriptive statistics of the variables in the dataset were examined to provide an overview of the data structure and the variability among the indicators. Subsequently, the predictive performance of four machine learning models (Support Vector Regression (SVR), Random Forest, Gradient Boosting and XGBoost) was compared. The models were evaluated using three widely accepted regression performance metrics: the coefficient of determination (R^2), root mean square error (RMSE), and mean absolute error (MAE). The results provide insights into the effectiveness of different machine learning algorithms in predicting life expectancy and allow for a comparative assessment of their predictive capabilities. Table 1 shows that economic indicators, such as GDP per capita and health expenditure per capita, exhibit greater variability, while health indicators remain relatively stable.

Table 1 presents the descriptive statistics of the variables included in the analysis, providing an overview of the measures of central tendency and the levels of dispersion. In the upper-middle-income countries in the dataset, average life expectancy at birth is 73.88 years, with values ranging from 70.77–76.19 years,

TABLE 1 | Statistical analysis of the variables.

	Average	Minimum	Maximum	SD	SE
Life expectancy at birth, total (years)	73.88	70.77	76.19	1.68	0.34
Current health expenditure (% of GDP)	5.40	5.01	6.08	0.30	0.06
Hospital beds (per 1000 people)	2.80	1.99	4.55	0.78	0.16
GDP per capita (current US$)	5991.27	1829.52	10657.97	2864.58	584.73
Urban population (% of total population)	59.50	48.15	69.27	6.68	1.36
Current health expenditure per capita (current US$)	330.00	94.05	647.67	170.89	34.88
Immunization, measles (% of children ages 12–23 months)	91.19	86.02	94.59	3.03	0.62
Population age 65 and above (%)	8.47	6.44	11.70	1.53	0.31

10991719, 0, Downloaded from https://onlinelibrary.wiley.com/doi/10.1002/sd.71468 by Mehmet KAYAKUS - Akdeniz University , Wiley Online Library on [17/07/2026]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

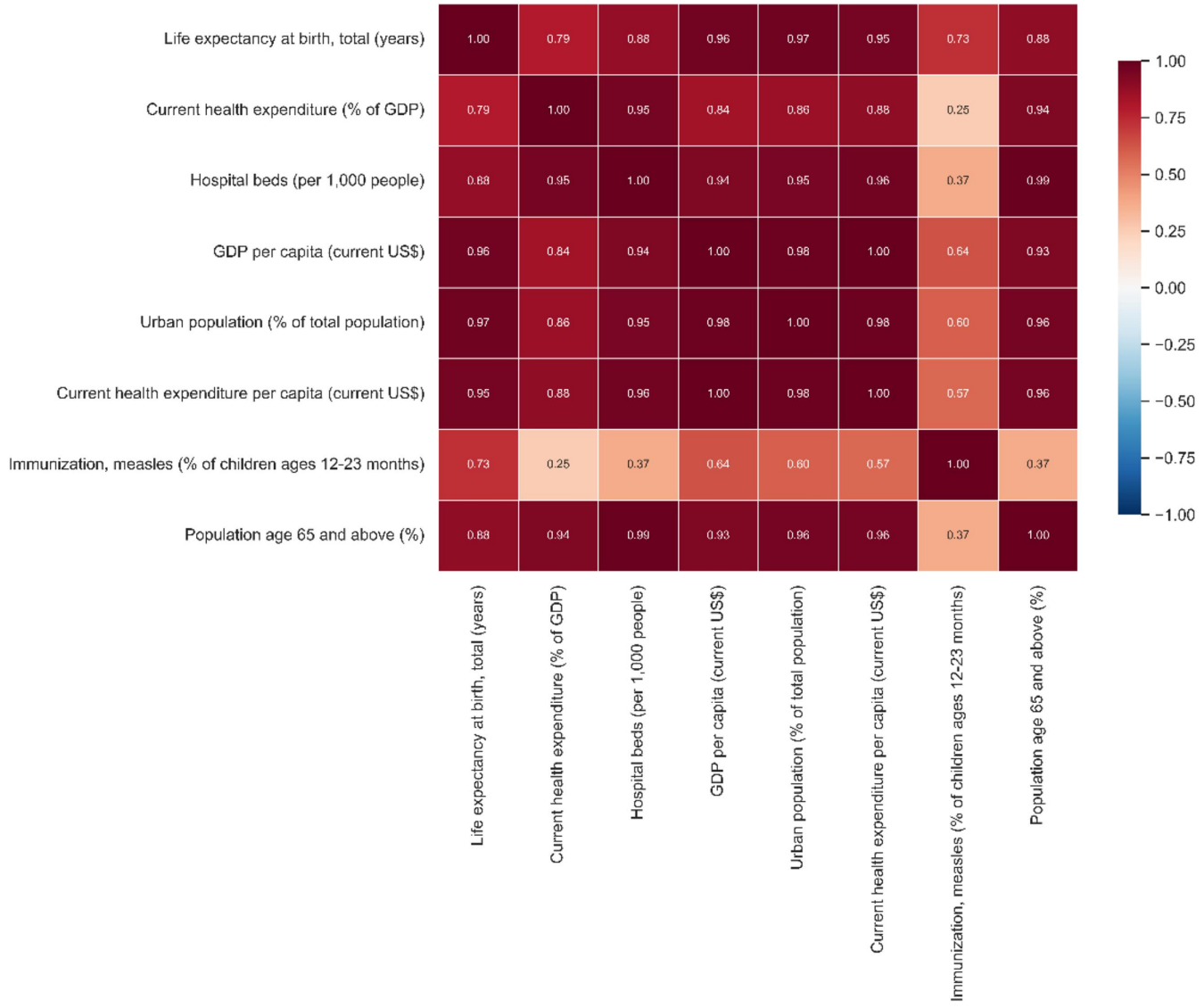


FIGURE 2 | Correlation matrix.

indicating a relatively moderate variation across the sample. Health-related indicators, such as vaccination coverage and health expenditure as a percentage of GDP, exhibit relatively stable distributions and point to similar public health conditions across the analyzed countries. In contrast, economic indicators—particularly GDP per capita and health expenditure per capita—exhibit a higher degree of variability, as reflected in higher standard deviation values. This variability suggests that economic conditions differ significantly among upper-middle-income countries and may play a significant role in explaining differences in life expectancy. Overall, descriptive statistics provide an initial understanding of the dataset and highlight the potential impact of both economic and health-related factors on life expectancy.

Figure 2 shows the correlation matrix illustrating how the study's variables are related to one another.

The correlation matrix in Figure 2 shows that the variables used in this study to estimate life expectancy in upper-middle-income countries are, for the most part, strongly and positively correlated. The very high correlation coefficients ($r \approx 0.95–0.97$) between life expectancy and GDP per capita, current health expenditure per capita, and the urban population indicate that the level of economic prosperity and investment in the health system are crucial for extending life expectancy. The strong positive correlation between the number of hospital beds and the population aged 65 and over, on the one hand, and life expectancy, on the other ($r \approx 0.88$), indicates that improvements in healthcare infrastructure and demographic structure also influence life expectancy. The fact that the measles vaccination rate has lower correlation values compared to other variables is due to the rate already being high in upper-middle-income countries and not varying significantly from one country to another. The correlation analysis confirms the primary objective of this study by demonstrating that variables such as economic development, health expenditure, healthcare infrastructure, and demographic structure are critical determinants of life expectancy and that these factors provide important explanatory information for machine learning models.

10991719, 0, Downloaded from https://onlinelibrary.wiley.com/doi/10.1002/sd.71468 by Mehmet KAYAKUS - Akdeniz University , Wiley Online Library on [17/07/2026]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

The predictive performance of the machine learning models used in the study is presented in Table 2. The models were evaluated using three commonly used regression performance metrics: the coefficient of determination (R^2), root mean square error (RMSE), and mean absolute error (MAE). These metrics provide a comprehensive assessment of how accurately each model predicts life expectancy.

As shown in Table 2, the predictive performance of machine learning models varies in terms of accuracy and error rates. Among the models evaluated, Support Vector Regression (SVR) achieved the best predictive performance with the highest R^2 value (0.949) and the lowest RMSE (0.288) and MAE (0.225) values. These results indicate that the SVR model provides the most accurate predictions of life expectancy in the dataset. The Random Forest model also demonstrated strong predictive capability with an R^2 value of 0.913 and relatively low error metrics. In contrast, the Gradient Boosting and XGBoost models exhibited slightly lower predictive performance, as reflected by their lower R^2 values and higher RMSE and MAE scores. Overall, the results presented in Table 2 indicate that, in this study, SVR outperformed other machine learning models in predicting life expectancy for upper-middle-income countries.

Figure 3 illustrates the comparative performance of machine learning models across three evaluation metrics (R^2, RMSE, and MAE) and presents a visual comparison of the models' accuracy and prediction errors.

TABLE 2 | Analysis of the models.

	R^2	RMSE	MAE
SVR	0.949	0.288	0.225
Random Forest	0.913	0.357	0.274
Gradient Boosting	0.892	0.383	0.305
XGBoost	0.864	0.435	0.376

As shown in Figure 3, a graphical comparison of the evaluation metrics reveals significant differences in model performance. The bar chart shows that the SVR model achieves the highest R^2 value while maintaining the lowest RMSE and MAE values, indicating a strong balance between explanatory power and predictive accuracy. The Random Forest model follows closely behind, with slightly lower explanatory performance and moderately higher error values. The Gradient Boosting and XGBoost models, however, exhibit relatively lower R^2 scores and higher prediction errors; this trend can be observed from the increasing bar heights in the RMSE and MAE metrics. The visual representation in Figure 3 highlights the relative performance differences between the models and facilitates an intuitive comparison of their predictive capabilities.

To further assess whether machine learning models provide meaningful improvements over conventional statistical approaches, an additional benchmark analysis was conducted using Multiple Linear Regression (MLR). The benchmark model achieved substantially lower predictive performance ($R^2 = 0.722$, RMSE = 2.528, MAE = 0.268) compared with all machine learning models, particularly SVR ($R^2 = 0.949$). These findings indicate that machine learning algorithms, especially nonlinear models, capture complex relationships among socio-economic and health-related predictors more effectively than conventional linear regression.

Figure 4 presents line charts comparing actual life **expectancy values with** the predicted values generated by the SVR, Random Forest, Gradient Boosting, and XGBoost models.

As shown in Figure 4, the predicted values generated by the machine learning models generally follow the trend of the actual life expectancy values, demonstrating that the models can capture the overall pattern in the data. The SVR model exhibits the closest fit between actual and predicted values, with the prediction line closely tracking the observed values throughout the dataset. The Random Forest model also shows a strong

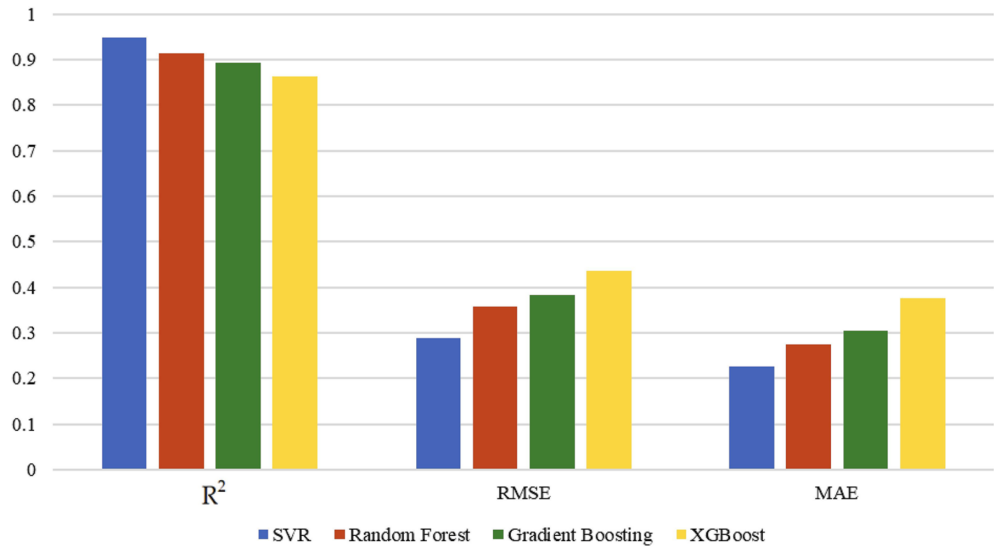


FIGURE 3 | Evaluation of the models.

10991719, 0, Downloaded from https://onlinelibrary.wiley.com/doi/10.1002/sd.71468 by Mehmet KAYAKUS - Akdeniz University , Wiley Online Library on [17/07/2026]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

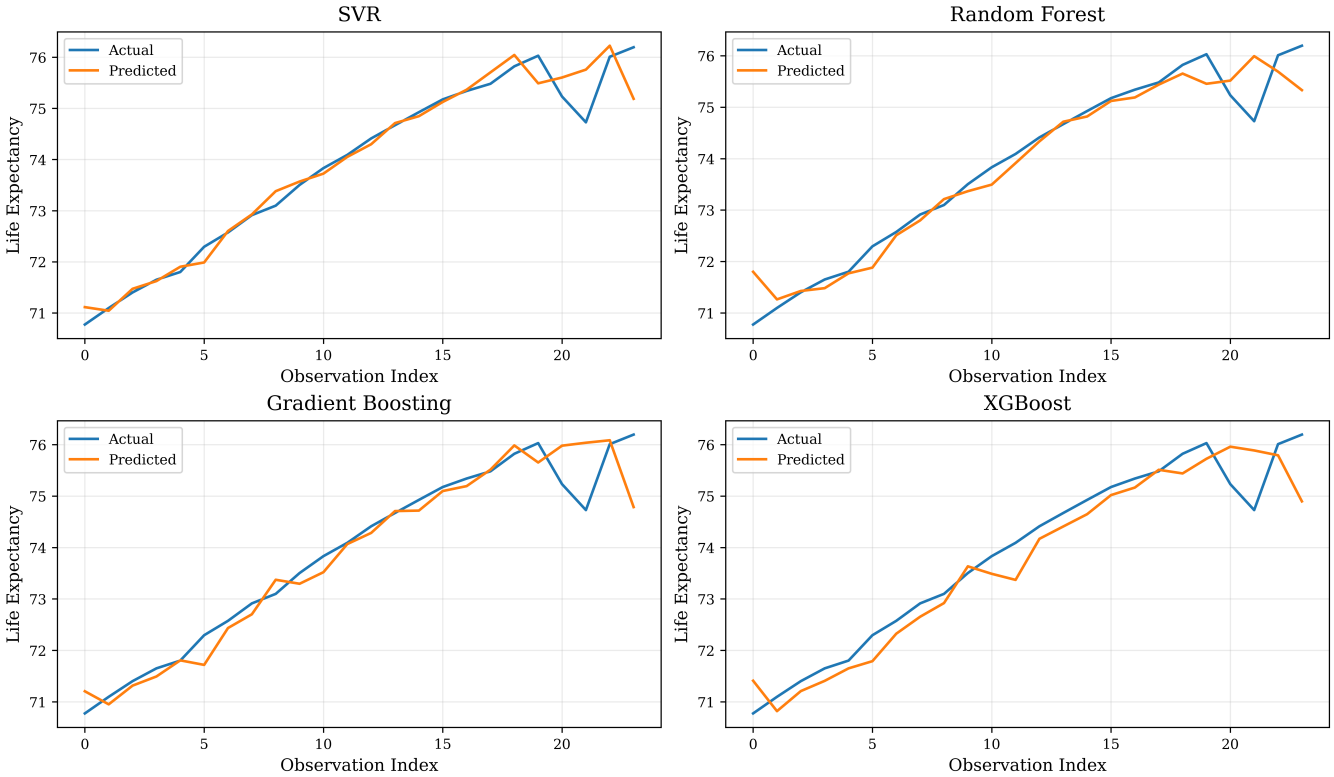


FIGURE 4 | Actual and predicted life expectancy values of the models.

fit, although small deviations can be observed at certain points. In contrast, the Gradient Boosting and XGBoost models exhibit slightly larger fluctuations between actual and predicted values, particularly in later observations. Overall, the graphical comparison in Figure 3 shows that all models successfully capture the general upward trend in life expectancy but that the SVR and Random Forest models provide relatively more consistent predictions across observations.

Figure 5 presents the prediction error plots for the machine learning models, illustrating the differences between the actual and predicted life expectancy values for each model.

As shown in Figure 5, prediction error plots provide a visual assessment of how closely the predicted values track the actual life expectancy values across the observations. The SVR model exhibits relatively lower prediction errors and more stable prediction performance, as evidenced by the smallest deviations between the two lines. The Random Forest model also exhibits a similar pattern, but some noticeable deviations occur in the middle observations where the prediction errors increase slightly. In contrast, the Gradient Boosting and XGBoost models exhibit greater fluctuations between actual and predicted values, particularly at a few peak points where the predicted values diverge more markedly from the observed values. Overall, the prediction error plots highlight that while all models capture the general pattern of life expectancy, the SVR and Random Forest models provide comparatively more accurate and stable predictions across the dataset.

Figure 6 presents scatter plots comparing actual life expectancy values with predicted values generated by the machine learning models.

As shown in Figure 6, the scatter plots illustrate the relationship between the observed and predicted life expectancy values for each model. Generally, the data points are clustered close to the diagonal reference line; this indicates that the predicted values are largely consistent with the actual observations. The SVR model demonstrates a high level of prediction accuracy, as evidenced by the points being strongly aligned along the diagonal line. The Random Forest model also exhibits a similar pattern, although a few observations deviate slightly from the reference line. For the Gradient Boosting and XGBoost models, the points remain relatively close to the diagonal, though they show slightly greater dispersion, particularly in the higher value range. Overall, the scatter plots confirm that the models can capture the relationship between predictors and life expectancy while also highlighting the differences in prediction accuracy between the algorithms.

Figure 7 presents the cumulative histograms of the machine learning models, illustrating the distribution of prediction errors for each algorithm.

As shown in Figure 7, the residual histograms provide information about the distribution of prediction errors generated by the models. In general, the residuals of all models are concentrated around zero; this pattern indicates that the predicted values are close to the actual observations and that large prediction errors are relatively limited. The SVR model exhibits a relatively symmetrical residual distribution centered close to zero, demonstrating balanced prediction errors and stable model performance. The Random Forest model also exhibits a similar pattern, although a slightly wider spread of residual values can be observed. In contrast, the Gradient Boosting and XGBoost models show a slightly wider distribution of

10991719, 0, Downloaded from https://onlinelibrary.wiley.com/doi/10.1002/sd.71468 by Mehmet KAYAKUS - Akdeniz University , Wiley Online Library on [17/07/2026]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

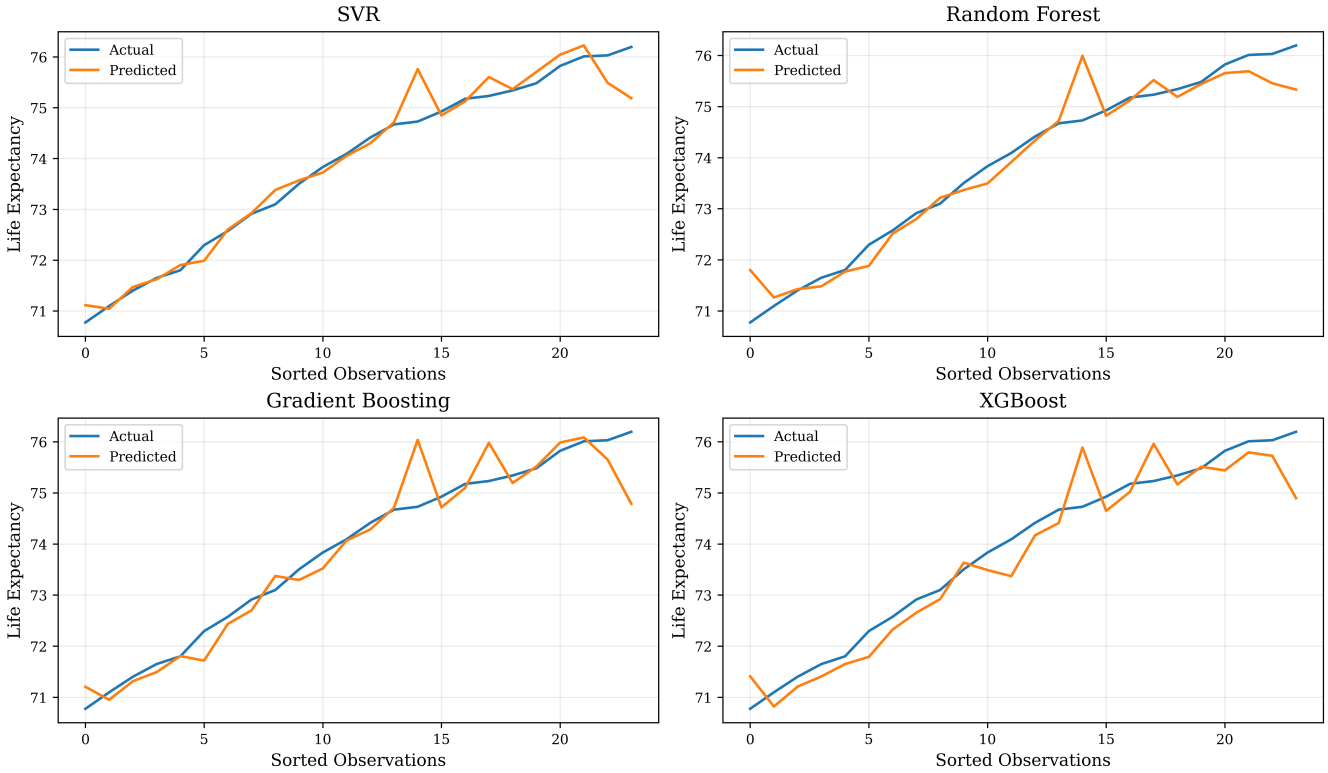


FIGURE 5 | Prediction error plots of the models.

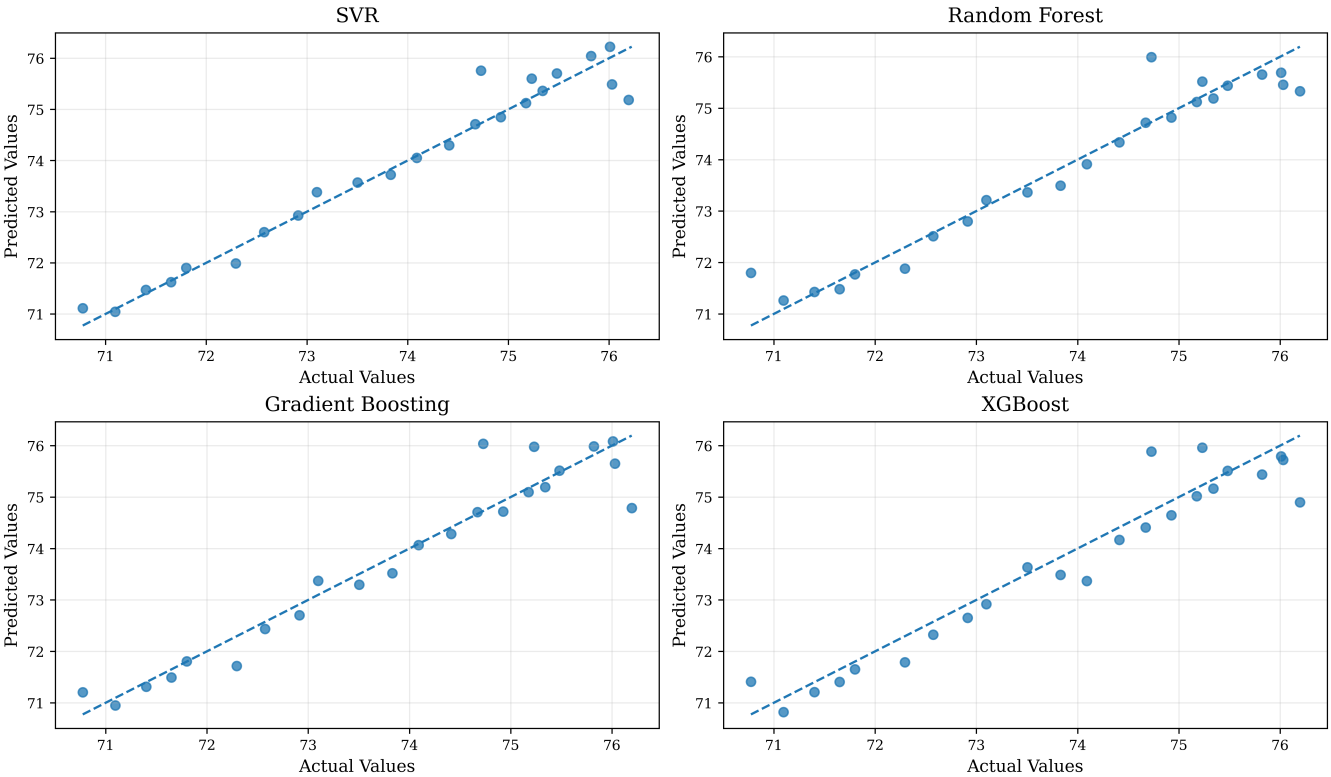


FIGURE 6 | Scatter plots of the models.

residuals, indicating greater variability in prediction errors between some observations. Overall, the residual distributions indicate that, while all models produce reasonably accurate predictions, the SVR and Random Forest models tend to generate more concentrated and stable error patterns compared to the other algorithms.

10991719, 0, Downloaded from https://onlinelibrary.wiley.com/doi/10.1002/sd.71468 by Mehmet KAYAKUS - Akdeniz University , Wiley Online Library on [17/07/2026]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

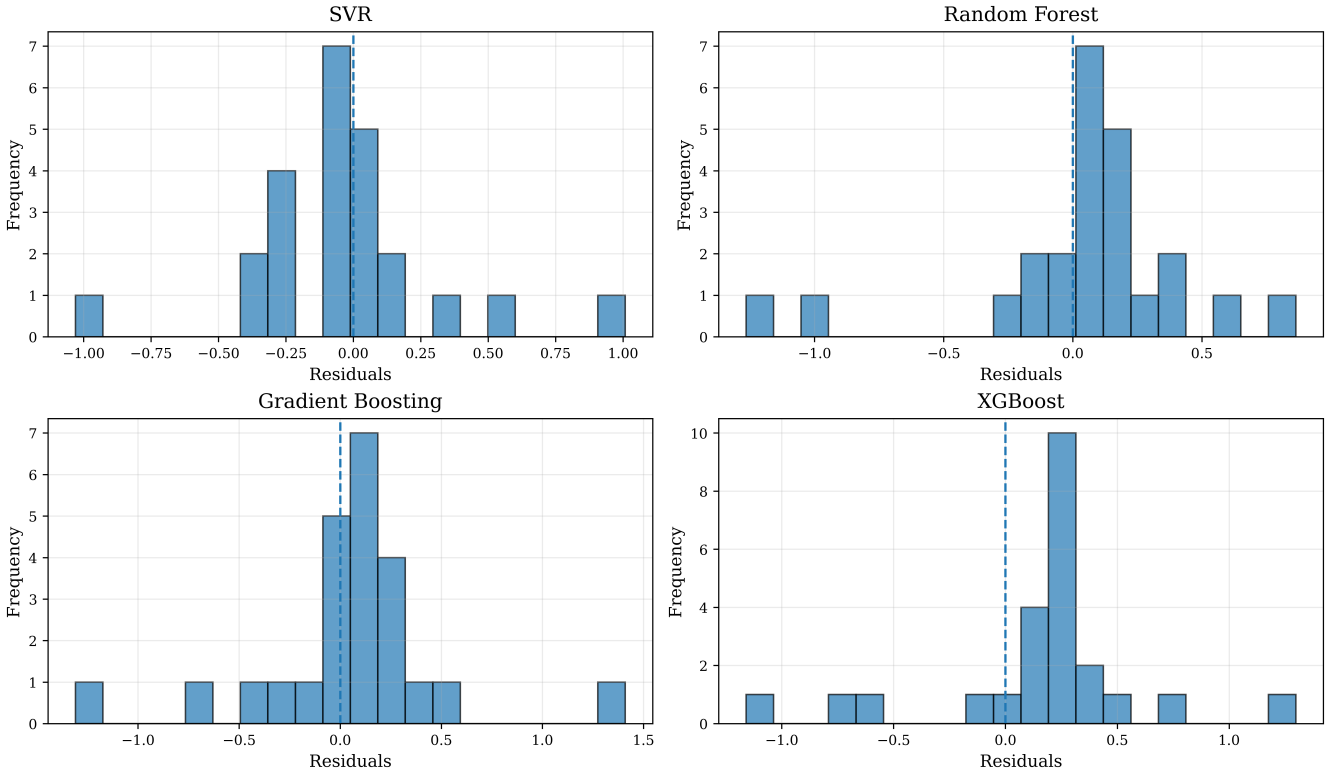


FIGURE 7 | Residual distribution of the models.

Figure 8 presents the residual plots of the machine learning models, illustrating the distribution of prediction errors relative to the predicted life expectancy values.

As shown in Figure 8, residual plots provide a visual assessment of the error patterns associated with each machine learning model. Ideally, the residuals should be randomly distributed around the zero line; this distribution indicates that the model predictions are unbiased and that there is no systematic pattern in the errors. The SVR model typically exhibits residuals distributed around zero with a relatively limited spread, indicating a stable prediction structure. The Random Forest model also shows a largely random distribution of residuals, although there are slightly larger deviations in a few observations. In contrast, the Gradient Boosting and XGBoost models exhibit a slightly wider residual distribution, particularly within certain predicted value ranges. Despite these differences, the residual plots show that all models produce errors centered predominantly around zero and capture the underlying relationship in the data without a strong systematic bias.

4 | Discussion

In this study, the performance of four different machine learning algorithms in predicting life expectancy in upper-middle-income countries was evaluated comparatively. The findings indicate that socio-economic and health-related indicators provide a high level of explanatory power in predicting life expectancy, and that machine learning approaches offer effective tools for modeling this multidimensional structure. This finding is consistent with previous studies reporting that life expectancy can be successfully modeled using machine learning methods based on various health, education, and economic indicators (Ghazali et al. 2018; Lipesa et al. 2023).

The fact that the highest predictive performance was achieved with the SVR model suggests that this algorithm constitutes a robust alternative, particularly for data structures with limited sample sizes and where relationships between variables may not be linear. The superiority of SVR in this study should be evaluated not only in terms of the algorithm's theoretical properties but also in relation to the dataset's size, the distribution of variables, the scaling structure, and the alignment of model parameters with the data. Furthermore, as SVR's kernel-based structure can flexibly capture non-linear patterns, it has provided a significant advantage in modeling a multi-dimensional and complex output such as life expectancy (Ghazali et al. 2018). Its ability to offer higher explanatory power compared to classical linear methods supports the view that this approach could be an effective option for small and medium-sized datasets. Although the predictive performance of the SVR model was exceptionally high, this result should be interpreted cautiously due to the relatively small sample size. To reduce overfitting risk, all models were evaluated using 10-fold cross-validation within a pipeline structure, ensuring that scaling and preprocessing steps were learned exclusively from training folds. This procedure minimized data leakage and improved the robustness of performance estimates. Nevertheless, the possibility of optimistic performance due to sample size cannot be completely excluded.

The fact that the Random Forest model also demonstrates high performance indicates that ensemble learning-based approaches can be effective in modeling multivariate outcomes such as life

10991719, 0, Downloaded from https://onlinelibrary.wiley.com/doi/10.1002/sd.71468 by Mehmet KAYAKUS - Akdeniz University , Wiley Online Library on [17/07/2026]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

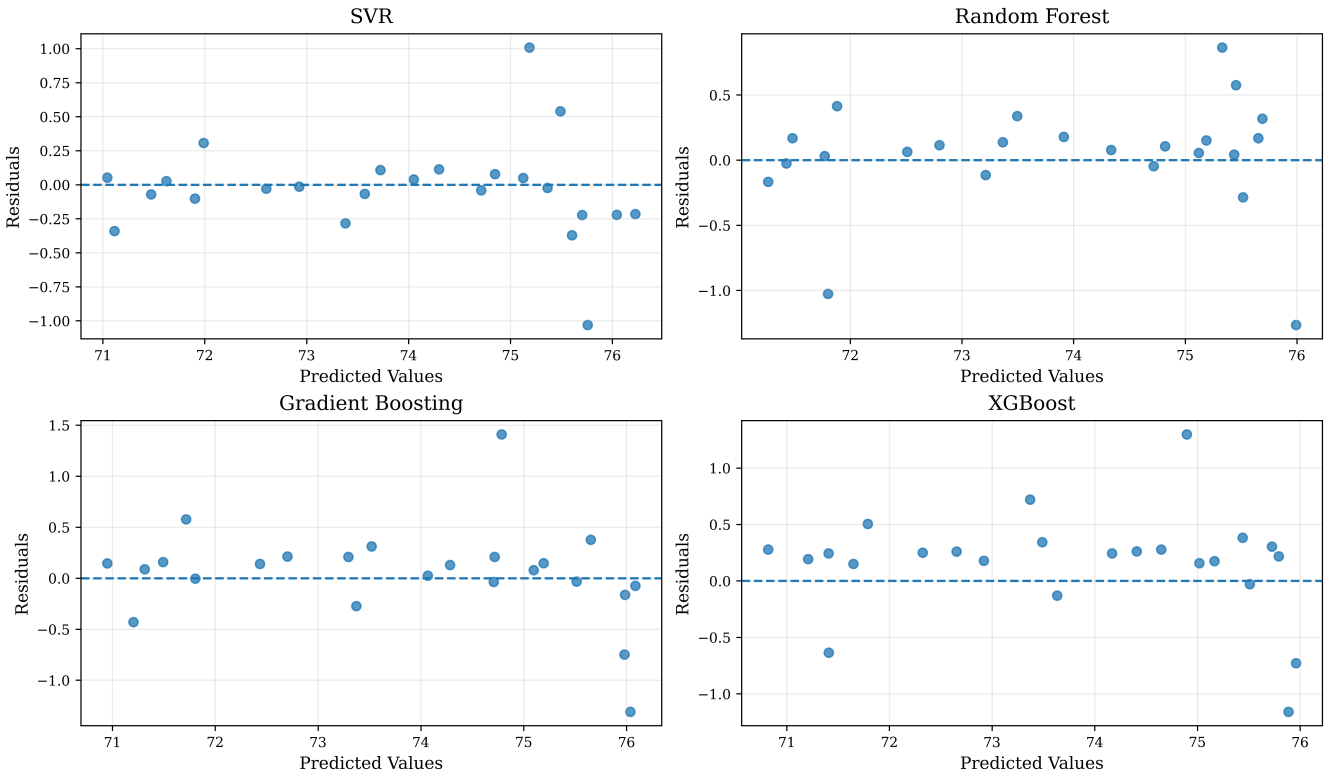


FIGURE 8 | Residual plot of the models.

expectancy. The fact that the Random Forest algorithm operates via multiple decision trees and can relatively reduce the risk of overfitting may provide an advantage, particularly in data structures containing distinct differences in economic indicators. This finding is consistent with recent studies reporting that tree-based approaches such as Random Forest, Gradient Boosting, and XGBoost can be effective in life expectancy prediction (Chairunisa et al. 2024). However, the high performance of the Random Forest does not indicate that the relevant variables are causally determinative on their own; rather, it suggests that these indicators successfully capture the structural patterns that vary alongside life expectancy.

The fact that Gradient Boosting and XGBoost models performed less well than SVR and Random Forest in this study may seem unexpected at first glance, as these algorithms are among the most robust methods for many regression problems (Chairunisa et al. 2024). However, it is well known that the performance of machine learning models is sensitive to the data context. Boosting-based algorithms can be significantly influenced by factors such as sample size, hyperparameter optimization, variable structure, and signal-to-noise ratio. Therefore, the relatively lower performance of these models suggests not that the algorithms are generally inadequate, but rather that the structure of the dataset and the limited sample size have shaped the model's success. Indeed, it has been reported that model performance in various life expectancy estimation studies can vary depending on the data processing approach, missing data handling method, and sample structure (Bahadure et al. 2025; Lipesa et al. 2023).

The findings of the correlation analysis indicate very strong positive relationships between life expectancy and GDP per capita, health expenditure per capita, and the urbanization rate. This result is consistent with extensive literature showing that economic development, health investment, and structural transformation processes are closely linked to life expectancy (Bansal and Jin 2023). However, the high level of correlation between these variables also suggests that the structural processes influencing life expectancy do not operate independently of one another. In other words, economic development, health expenditure, and urbanization are likely to be interrelated and partially overlapping processes in their association with life expectancy. Therefore, the possibility of multiple connections and structural overlap must be considered when interpreting the individual effects of these variables. The fact that the vaccination rate shows a lower correlation compared to other variables may stem from the more homogeneous distribution of this indicator in upper-middle-income countries; this suggests that, rather than the variable being insignificant, its contribution to explaining cross-country variation may be more limited.

Although formal feature importance analysis was beyond the primary scope of this study, the correlation matrix and model outputs consistently indicate that GDP per capita, health expenditure per capita, and urbanization are among the most influential predictors of life expectancy. These variables exhibited the strongest associations with life expectancy and appear to contribute substantially to the predictive performance of the machine learning models. This finding further supports the importance of socio-economic development and

10991719, 0, Downloaded from https://onlinelibrary.wiley.com/doi/10.1002/sd.71468 by Mehmet KAYAKUS - Akdeniz University , Wiley Online Library on [17/07/2026]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

healthcare investment in explaining cross-country differences in life expectancy.

Descriptive statistics reveal that, in the upper-middle-income countries under examination, health indicators exhibit a relatively more homogeneous distribution, whereas economic indicators display a more variable distribution. This finding demonstrates that upper-middle-income countries, despite belonging to the same income group, exhibit significant structural heterogeneity within their own ranks. Consequently, differences in life expectancy between countries must be assessed not only in terms of health system indicators but also in conjunction with the distribution of economic resources, patterns of urbanization, and the varying manifestations of development levels. Studies emphasizing that life expectancy in developing countries is shaped by multi-layered economic and social determinants also support this interpretation (Bansal and Jin 2023).

The residuals and the distribution pattern between observed and predicted values indicate that all models generally operate without producing a significant systematic bias; however, the fact that the SVR model's residuals exhibit a narrower and more balanced distribution suggests that this model produces more stable predictions. The closer alignment of observed and predicted values along the diagonal axis also visually supports this result. This suggests that SVR stands out not only in terms of performance metrics but also in terms of model reliability. This finding is particularly noteworthy in decision-support contexts such as health policies, where not only high accuracy but also a stable error structure is important.

From a public health policy perspective, the findings suggest that improving life expectancy in upper-middle-income countries requires integrated interventions extending beyond healthcare delivery alone. Policies aimed at increasing healthcare accessibility, strengthening preventive care, reducing regional inequalities, and supporting balanced urban development may produce substantial long-term gains in population health. In addition, targeted investments in economically disadvantaged regions may help reduce health inequalities and improve overall life expectancy outcomes.

The findings of this study should be interpreted within certain limitations. The inclusion of only upper-middle-income countries in the analysis limits the generalizability of the results to other income groups. Furthermore, the relatively small sample size and the variable set based on selected indicators from international databases should be considered as key methodological factors that may influence model performance and the scope of the findings. It should be noted that models developed in different clinical and social contexts may not perform similarly; therefore, interpreting model results in a context-sensitive manner is important (Christie et al. 2018). Nevertheless, it is considered that studying makes significant contributions. Firstly, its treatment of upper-middle-income countries as a specific unit of analysis is valuable in highlighting the internal heterogeneity of this income group. Furthermore, the comparison of multiple machine learning algorithms on the same data structure demonstrates the need for context-sensitive method selection. The findings suggest that kernel-based methods such as SVR could be robust alternatives, particularly under conditions of limited sample size and multidimensional data. From a policy perspective, the joint assessment of economic and health-related indicators could contribute to a more precise monitoring of differences in life expectancy. In future studies, the use of larger samples, the comparative analysis of different income groups, the expansion of the variable set to include social and environmental determinants, and the incorporation of methods that enhance model interpretability could further deepen the body of knowledge in this field.

4.1 | Data and Code Availability

The Python codes used for data preprocessing, model training, cross-validation, and visualization are available in the Supporting Information and can also be provided by the corresponding author upon reasonable request to ensure reproducibility and transparency.

5 | Conclusion

The findings of this study demonstrate that machine learning methods provide a robust and effective analytical framework for predicting life expectancy in upper-middle-income countries. Among the evaluated models, Support Vector Regression (SVR) achieved the highest predictive performance, highlighting its suitability for modeling complex and non-linear relationships, particularly in datasets with limited sample sizes. The Random Forest model also exhibited strong predictive capability, whereas Gradient Boosting and XGBoost showed comparatively lower performance within the context of the dataset.

Beyond methodological contributions, the results reveal that life expectancy is strongly associated with key socio-economic and structural indicators, including GDP per capita, health expenditures, and urbanization. These results show that longer life expectancy is not just a result of how well the health system works but is also a part of larger development trends. In this respect, life expectancy can be considered not only a health indicator but also a critical proxy for sustainable development outcomes.

From a sustainability perspective, the study highlights that investments in healthcare systems, equitable economic growth, and balanced urban development are essential components of long-term human well-being. These dimensions align closely with several of the United Nations Sustainable Development Goals, particularly Goal 3 (Good Health and Well-being) and Goal 11 (Sustainable Cities and Communities). The strong relationships identified between life expectancy and socio-economic variables suggest that integrated and data-driven policy approaches are necessary to ensure sustainable improvements in population health.

Furthermore, the use of machine learning models offers significant potential for supporting evidence-based policymaking within the framework of sustainable development. By enabling more accurate predictions of health outcomes, these models can assist policymakers in identifying priority areas, optimizing resource allocation, and designing targeted interventions that promote both health equity and long-term sustainability.

10991719, 0, Downloaded from https://onlinelibrary.wiley.com/doi/10.1002/sd.71468 by Mehmet KAYAKUS - Akdeniz University , Wiley Online Library on [17/07/2026]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

However, the findings should be interpreted considering certain limitations, including the relatively small sample size and the focus on upper-middle-income countries, which may restrict the generality of the results. Future research could expand the scope by incorporating larger datasets, additional socio-environmental variables such as environmental quality and climate indicators, and comparisons across different income groups. Such extensions would further strengthen the integration of machine learning approaches into sustainable development research.

In conclusion, this study makes both methodological and practical contributions. Methodologically, it demonstrates that Support Vector Regression can outperform widely used ensemble learning methods in predicting life expectancy under limited-sample and non-linear data conditions. Practically, the findings highlight that economic development, healthcare investment, and urbanization are key strategic levers for improving population health. These results provide policymakers with evidence-based insights for prioritizing interventions aimed at increasing life expectancy and reducing health inequalities in upper-middle-income countries.

Author Contributions

Esra Süzen: conceptualization, writing – original draft, writing – review and editing, funding acquisition, project administration, data curation, supervision, validation, formal analysis. **Mehmet Kayakuş:** methodology, software, data curation, writing – review and editing, writing – original draft, visualization, validation, investigation, formal analysis.

Conflicts of Interest

The authors declare no conflicts of interest.

References

Abubakari, M., N. S. Owoob, and E. Nketiah-Amponsah. 2019. "Socio-Economic Determinants of Life Expectancy in Sub-Saharan Africa." *Ghanaian Journal of Economics* 7, no. 1: 156–177.

Aggarwal, A., K. Unger-Saldaña, G. Lewison, and R. Sullivan. 2015. "The Challenge of Cancer in Middle-Income Countries With an Ageing Population: Mexico as a Case Study." *Ecancermedicalscience* 9: 536.

Ali, Z. A., Z. H. Abduljabbar, H. A. Taher, A. B. Sallow, and S. M. Almufti. 2023. "Exploring the Power of eXtreme Gradient Boosting Algorithm in Machine Learning: A Review." *Academic Journal of Nawroz University (AJNU)* 12, no. 2: 320–334.

Alim, M., G.-H. Ye, P. Guan, D.-S. Huang, B.-S. Zhou, and W. Wu. 2020. "Comparison of ARIMA Model and XGBoost Model for Prediction of Human Brucellosis in Mainland China: A Time-Series Study." *BMJ Open* 10, no. 12: e039676.

Bahadure, N. B., R. Khomane, D. Raut, et al. 2025. "Comparative Analysis of Life Expectancy Prediction Using Regression Algorithms." *Intelligent Decision Technologies* 19, no. 1: 429–442.

Bansal, S., and Y. Jin. 2023. "Heterogeneous Effects of Obesity on Life Expectancy: A Global Perspective." *Annual Review of Resource Economics* 15, no. 1: 433–554.

Becker, T., A.-J. Rousseau, M. Geubbelmans, T. Burzykowski, and D. Valkenborg. 2023. "Decision Trees and Random Forests." *American Journal of Orthodontics and Dentofacial Orthopedics* 164, no. 6: 894–897.

Bentéjac, C., A. Csörgő, and G. Martínez-Muñoz. 2021. "A Comparative Analysis of Gradient Boosting Algorithms." *Artificial Intelligence Review* 54, no. 3: 1937–1967.

Boldini, D., F. Grisoni, D. Kuhn, L. Friedrich, and S. A. Sieber. 2023. "Practical Guidelines for the Use of Gradient Boosting for Molecular Property Prediction." *Journal of Cheminformatics* 15, no. 1: 73.

Breiman, L. 2001. "Random Forests." *Machine Learning* 45, no. 1: 5–32.

Cabieses, B., and P. Bird. 2014. "Glossary of Access to Health Care and Related Concepts for Low-and Middle-Income Countries (LMICs): A Critical Review of International Literature." *International Journal of Health Services* 44, no. 4: 845–861.

Cai, W., R. Wei, L. Xu, and X. Ding. 2022. "A Method for Modelling Greenhouse Temperature Using Gradient Boost Decision Tree." *Information Processing in Agriculture* 9, no. 3: 343–354.

Casey, J. A., B. S. Schwartz, W. F. Stewart, and N. E. Adler. 2016. "Using Electronic Health Records for Population Health Research: A Review of Methods and Applications." *Annual Review of Public Health* 37, no. 1: 61–81.

Çelik, S., S. Yavan, and G. Çimen. 2026. "Unveiling Structural Patterns of Global Budget Transparency Using Advanced Machine Learning and Explainable Artificial Intelligence Frameworks." *Economics of Governance* 27, no. 1: 24.

Chairunisa, G., M. K. Najib, S. Nurdiati, et al. 2024. "Life Expectancy Prediction Using Decision Tree, Random Forest, Gradient Boosting, and XGBoost Regressions." *Jurnal Sintak* 2, no. 2: 71–82.

Chen, T., and C. Guestrin. 2016. Xgboost: A Scalable Tree Boosting System Proceedings of the 22nd ACM sigkdd International Conference on Knowledge Discovery and Data Mining.

Christie, S. A., A. E. Hubbard, R. A. Callcut, et al. 2018. "Machine Learning Without Borders? An Adaptable Tool to Optimize Mortality Prediction in Diverse Clinical Settings." *Journal of Trauma and Acute Care Surgery* 85, no. 5: 921–927.

Dewi, C., and R.-C. Chen. 2019. "Random Forest and Support Vector Machine on Features Selection for Regression Analysis." *International Journal of Innovative Computing, Information and Control* 15, no. 6: 2027–2037.

Ekundayo, F. 2024. "Using Machine Learning to Predict Disease Outbreaks and Enhance Public Health Surveillance." *World Journal of Advanced Research and Reviews* 24, no. 3: 794–811.

Feng, J., Y. Yu, and Z.-H. Zhou. 2018. "Multi-Layered Gradient Boosting Decision Trees." *Advances in Neural Information Processing Systems* 31: 1–11.

Frias-Paredes, L., F. Mallor, M. Gaston-Romeo, and T. León. 2018. "Dynamic Mean Absolute Error as New Measure for Assessing Forecasting Errors." *Energy Conversion and Management* 162: 176–188.

Gaziano, T. A., A. Bitton, S. Anand, S. Abrahams-Gessel, and A. Murphy. 2010. "Growing Epidemic of Coronary Heart Disease in Low-and Middle-Income Countries." *Current Problems in Cardiology* 35, no. 2: 72–115.

Ghazali, M., I. Fitriati, and R. Purnamasari. 2018. "Support Vector Regression for Modeling Effect of Education Rate on Life Expectancy Rate in Indonesia." *IJECA (International Journal of Education and Curriculum Application)* 1, no. 1: 64–69.

Hodson, T. O. 2022. "Root Mean Square Error (RMSE) or Mean Absolute Error (MAE): When to Use Them or Not." *Geoscientific Model Development Discussions* 15: 1–10.

Hosen, M. J., S. Anwar, J. Taslem Mourosi, S. Chakraborty, M. F. Miah, and O. M. Vanakker. 2021. "Genetic Counseling in the Context of Bangladesh: Current Scenario, Challenges, and a Framework for Genetic Service Implementation." *Orphanet Journal of Rare Diseases* 16, no. 1: 168.

Huang, H., A. Wang, R. Morello-Frosch, et al. 2018. "Cumulative Risk and Impact Modeling on Environmental Chemical and Social Stressors." *Current Environmental Health Reports* 5, no. 1: 88–99.

10991719, 0, Downloaded from https://onlinelibrary.wiley.com/doi/10.1002/sd.71468 by Mehmet KAYAKUS - Akdeniz University , Wiley Online Library on [17/07/2026]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

Kaliappan, J., K. Srinivasan, S. Mian Qaisar, K. Sundararajan, C.-Y. Chang. 2021. "Performance Evaluation of Regression Models for the Prediction of the COVID-19 Reproduction Rate." *Frontiers in Public Health* 9: 729795.

Kalusivalingam, A. K., A. Sharma, N. Patel, and V. Singh. 2022. "Leveraging Random Forests and Gradient Boosting for Enhanced Predictive Analytics in Operational Efficiency." *International Journal of AI and ML* 3, no. 9: 1–10.

Klopfenstein, Q., and S. Vaiter. 2021. "Linear Support Vector Regression With Linear Constraints." *Machine Learning* 110, no. 7: 1939–1974.

Krstajic, D., L. J. Buturovic, D. E. Leahy, and S. Thomas. 2014. "Cross-Validation Pitfalls When Selecting and Assessing Regression and Classification Models." *Journal of Cheminformatics* 6, no. 1: 10.

Kyrychek, M. 2025. "Gradient Boosting as a Tool for Solving Classification Problems in Data-Constrained Environments." *Technologies and Engineering* 2, no. 26: 37–47.

Lago, S., D. Cantarero, B. Rivera, et al. 2018. "Socioeconomic Status, Health Inequalities and Non-Communicable Diseases: A Systematic Review." *Journal of Public Health* 26, no. 1: 1–14.

Liew, A. W.-C., N.-F. Law, and H. Yan. 2011. "Missing Value Imputation for Gene Expression Data: Computational Techniques to Recover Missing Data From Available Information." *Briefings in Bioinformatics* 12, no. 5: 498–513.

Lipesa, B. A., E. Okango, B. O. Omolo, and E. O. Omondi. 2023. "An Application of a Supervised Machine Learning Model for Predicting Life Expectancy." *SN Applied Sciences* 5, no. 7: 189.

Malakouti, S. M., M. B. Menhaj, and A. A. Suratgar. 2023. "The Usage of 10-Fold Cross-Validation and Grid Search to Enhance ML Methods Performance in Solar Farm Power Generation Prediction." *Cleaner Engineering and Technology* 15: 100664.

Marmot, M. 2005. "Social Determinants of Health Inequalities." *Lancet* 365, no. 9464: 1099–1104.

McDermott, M. B., S. Wang, N. Marinsek, R. Ranganath, L. Foschini, and M. Ghassemi. 2021. "Reproducibility in Machine Learning for Health Research: Still a Ways to Go." *Science Translational Medicine* 13, no. 586: eabb1655.

Montesinos López, O. A., A. Montesinos López, and J. Crossa. 2022. "Support Vector Machines and Support Vector Regression." In *Multivariate Statistical Machine Learning Methods for Genomic Prediction*, 337–378. Springer.

Najafzadeh, M., and S. Niazmardi. 2021. "A Novel Multiple-Kernel Support Vector Regression Algorithm for Estimation of Water Quality Parameters." *Natural Resources Research* 30, no. 5: 3761–3775.

Ogle, G. D., G. A. Gregory, F. Wang, et al. 2023. "The T1D Index: Implications of Initial Results, Data Limitations, and Future Development." *Current Diabetes Reports* 23, no. 10: 277–291.

Ozyilmaz, A., Y. Bayraktar, E. Isik, et al. 2022. "The Relationship Between Health Expenditures and Economic Growth in EU Countries: Empirical Evidence Using Panel Fourier Toda–Yamamoto Causality Test and Regression Models." *International Journal of Environmental Research and Public Health* 19, no. 22: 15091.

Piepho, H. P. 2019. "A Coefficient of Determination (R2) for Generalized Linear Mixed Models." *Biometrical Journal* 61, no. 4: 860–872.

Plonsky, L., and H. Ghanbar. 2018. "Multiple Regression in L2 Research: A Methodological Synthesis and Guide to Interpreting R2 Values." *Modern Language Journal* 102, no. 4: 713–731.

Qu, F., Y.-T. Wang, W.-H. Hou, et al. 2022. "Forecasting of Automobile Sales Based on Support Vector Regression Optimized by the Grey Wolf Optimizer Algorithm." *Mathematics* 10, no. 13: 2234.

Roffia, P., A. Bucciol, and S. Hashlamoun. 2023. "Determinants of Life Expectancy at Birth: A Longitudinal Study on OECD Countries." *International Journal of Health Economics and Management* 23, no. 2: 189–212.

Sahin, E. K. 2020. "Assessing the Predictive Capability of Ensemble Tree Methods for Landslide Susceptibility Mapping Using XGBoost, Gradient Boosting Machine, and Random Forest." *SN Applied Sciences* 2, no. 7: 1308.

Salman, H. A., A. Kalakech, and A. Steiti. 2024. "Random Forest Algorithm Overview." *Babylonian Journal of Machine Learning* 2024: 69–79.

Schwalbe, N., and B. Wahl. 2020. "Artificial Intelligence and the Future of Global Health." *Lancet* 395, no. 10236: 1579–1586.

Shanmugasundar, G., M. Vanitha, R. Čep, V. Kumar, K. Kalita, and M. Ramachandran. 2021. "A Comparative Study of Linear, Random Forest and Adaboost Regressions for Modeling Non-Traditional Machining." *Pro* 9, no. 11: 2015.

Sharma, N. A., and A. Juneja. 2018. "Extreme Gradient Boosting With Squared Logistic Loss Function." In *Machine Intelligence and Signal Analysis*, 313–322. Springer.

Sharma, P., M. Sarstedt, G. Shmueli, K. H. Kim, and K. O. Thiele. 2019. "PLS-Based Model Selection: The Role of Alternative Explanations in Information Systems Research." *Journal of the Association for Information Systems* 20, no. 4: 4.

Shmueli, G., M. Sarstedt, J. F. Hair, et al. 2019. "Predictive Model Assessment in PLS-SEM: Guidelines for Using PLSpredict." *European Journal of Marketing* 53, no. 11: 2322–2347.

Suthaharan, S. 2016. "Random Forest Learning." In *Machine Learning Models and Algorithms for Big Data Classification: Thinking With Examples for Effective Learning*, 273–288. Springer.

Syam, N., and R. Kaul. 2021. "Random Forest, Bagging, and Boosting of Decision Trees." In *Machine Learning and Artificial Intelligence in Marketing and Sales*, 139–182. Emerald Publishing Limited.

Tahri, D., J. Liu, and M. Aziku. 2026. "Machine-Learning Applications in Predicting Students' Non-Cognitive Skills: Evidence From PISA 2022." *Fudan Journal of the Humanities and Social Sciences* 19: 1–31.

Țarcă, V., E. Țarcă, and M. Moscalu. 2024. "Social and Economic Determinants of Life Expectancy at Birth in Eastern Europe." *Health* 12: 1148.

Valkenborg, D., A.-J. Rousseau, M. Geubbelmans, and T. Burzykowski. 2023. "Support Vector Machines." *American Journal of Orthodontics and Dentofacial Orthopedics* 164, no. 5: 754–757.

Vydehi, K., K. Manchikanti, T. S. Kumari, and S. A. Shah. 2020. "Machine Learning Techniques for Life Expectancy Prediction." *International Journal of Advanced Trends in Computer Science and Engineering* 9, no. 4: 4503–4507.

Wahyuningsih, T., A. Iriani, H. D. Purnomo, and I. Sembiring. 2024. "Predicting Students' Success Level in an Examination Using Advanced Linear Regression and Extreme Gradient Boosting." *Computer Science and Information Technologies* 5, no. 1: 29–37.

Wang, D., H. Guo, Y. Sun, H. Liang, A. Li, and Y. Guo. 2024. "Prediction of Oil–Water Two-Phase Flow Patterns Based on Bayesian Optimisation of the Xgboost Algorithm." *Pro* 12, no. 8: 1660.

Wang, Q., Y. Zhou, W. Ding, Z. Zhang, K. Muhammad, and Z. Cao. 2020. "Random Forest With Self-Paced Bootstrap Learning in Lung Cancer Prognosis." *ACM Transactions on Multimedia Computing, Communications, and Applications (TOMM)* 16, no. 1s: 1–12.

WB_Group. 2024. World Bank Country and Lending Groups.

Wirayuda, A. A. B., and M. F. Chan. 2021. "A Systematic Review of Sociodemographic, Macroeconomic, and Health Resources Factors on Life Expectancy." *Asia-Pacific Journal of Public Health* 33, no. 4: 335–356.

10991719, 0, Downloaded from https://onlinelibrary.wiley.com/doi/10.1002/sd.71468 by Mehmet KAYAKUS - Akdeniz University , Wiley Online Library on [17/07/2026]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

World_Bank. 2026. "Upper Middle Income." https://data.worldbank.org/country/upper-middle-income.

Xiao, Y., J. Wu, Z. Lin, and X. Zhao. 2018. "A Deep Learning-Based Multi-Model Ensemble Method for Cancer Prediction." *Computer Methods and Programs in Biomedicine* 153: 1–9.

Yang, J., Z. Li, M. Wu, Y. Zhang, B. Zhang, and H. Shi. 2025. "Identifying Subjective Life Expectancy Risk Factors in Physically Active and Inactive Middle-Aged and Older Adults Using Machine Learning Models." *BMC Public Health* 25, no. 1: 3506.

Zhang, P., Y. Jia, and Y. Shang. 2022. "Research and Application of XGBoost in Imbalanced Data." *International Journal of Distributed Sensor Networks* 18, no. 6: 15501329221106935.

Zhang, Z., Y. Zhao, A. Canes, D. Steinberg, and O. Lyashevska. 2019. "Predictive Analytics With Gradient Boosting in Clinical Medicine." *Annals of Translational Medicine* 7, no. 7: 152.

Zhou, W., Z. Yan, and L. Zhang. 2024. "A Comparative Study of 11 Non-Linear Regression Models Highlighting Autoencoder, DBN, and SVR, Enhanced by SHAP Importance Analysis in Soybean Branching Prediction." *Scientific Reports* 14, no. 1: 5905.

Zhou, Z., C. Qiu, and Y. Zhang. 2023. "A Comparative Analysis of Linear Regression, Neural Networks and Random Forest Regression for Predicting Air Ozone Employing Soft Sensor Models." *Scientific Reports* 13, no. 1: 22420.

10991719, 0, Downloaded from https://onlinelibrary.wiley.com/doi/10.1002/sd.71468 by Mehmet KAYAKUS - Akdeniz University , Wiley Online Library on [17/07/2026]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License