Research Article/ Araştırma Makalesi

Sağ Aka Derg, 2026; 13(1): 110-121
doi: 10.52880/sagakaderg. 1804617

Sağlık Akademisyenleri Dergisi
Journal of Health Academicians
Journal Home Page: https://dergipark.org.tr/tr/pub/sagakaderg
e-ISSN: 2636-7572

Machine learning models for life expectancy prediction: Vaccination and socioeconomic factors

Yaşam beklentisi tahmini için makine öğrenimi modelleri: Aşılama ve sosyoekonomik faktörler

İsmail Biçer

Dr. Pamukkale University, Denizli/Turkiye, ismailbiceer@gmail.com, 000-0003-1878-0546

ABSTRACT

Introduction and Objective: This study compares machine learning models for predicting life expectancy, focusing on the effects of vaccination rates and socioeconomic factors. The main objective is to identify the most accurate model for life expectancy prediction and to examine the influence of these variables. Materials and Methods: Data preprocessing, statistical analyses, and machine learning modeling were performed using the Python programming language. Model performance was evaluated using RMSE, MAE, and R^2 metrics. The importance of variables was determined through Permutation Importance and SHAP analyses. Results: The XGBoost model achieved the lowest error rates and the highest R^2 value (90.7%), followed by the Random Forest model with an R^2 of 88.8%. Income was identified as the most influential factor affecting life expectancy, followed by health expenditures. Vaccination rates showed lower importance in Linear Regression and Artificial Neural Network models but had higher effects in Random Forest and XGBoost models. Conclusion: The XGBoost model proved to be the most successful method for predicting life expectancy. The findings highlight the crucial role of economic growth and health investments in extending lifespan and suggest that the impact of vaccination rates may vary depending on the model used.

ÖZ

Giriş ve Amaç: Bu çalışma, yaşam beklentisini tahmin etmede makine öğrenimi modellerini karşılaştırmakta ve aşılanma oranı ile sosyoekonomik faktörlerin etkisine odaklanmaktadır. Amaç, yaşam beklentisini en doğru şekilde tahmin eden modeli belirlemek ve bu değişkenlerin etkilerini incelemektir. Gereç ve Yöntem: Veri ön işleme, istatistiksel analizler ve makine öğrenimi modellemeleri Python programlama dili kullanılarak yapılmıştır. Modellerin performansı RMSE, MAE ve R^2 değerleriyle değerlendirilmiş; değişkenlerin önem düzeyi Permütasyon Önemi ve SHAP analizleriyle belirlenmiştir. Bulgular: XGBoost modeli en düşük hata oranlarını ve en yüksek R^2 değerini (%90,7) elde etmiştir. Rastgele Orman modeli %88,8 R^2 ile ikinci sırada yer almıştır. Gelir, yaşam beklentisini en çok etkileyen değişken olarak belirlenmiş; sağlık harcamaları ikinci sırada yer almıştır. Aşılama oranı, doğrusal modellerde düşük, XGBoost ve Rastgele Orman modellerinde daha yüksek etki göstermiştir. Sonuç: XGBoost modeli yaşam beklentisini tahminde en başarılı yöntemdir. Bulgular, ekonomik büyüme ve sağlık yatırımlarının yaşam süresini uzatmadaki önemini ve aşılama oranlarının model türüne bağlı olarak değişen etkisini ortaya koymaktadır.

Key Words:

Life Expectancy, Vaccination, Socioeconomic Factors, Machine Learning

Anahtar Kelimeler:

Yaşam Beklentisi, Aşılama, Sosyoekonomik Faktörler, Makine Öğrenimi

Corresponding Author/Sorumlu Yazar:

Dr. Pamukkale University, Denizli/Turkiye, ismailbiceer@gmail.com, 000-0003-1878-0546

Received Date/Gönderme Tarihi: 15.10.2025
Accepted Date/Kabul Tarihi: 03.03.2026
Published Online/Yayımlanma Tarihi: 31.03.2026

Reference | Atıf : Biçer, İ. (2026). Machine learning models for life expectancy prediction: Vaccination and socioeconomic factors. *Sağlık Akademisyenleri Dergisi*, 13(1), 110-121.

INTRODUCTION

Life expectancy (LE) is a fundamental measure commonly used to assess the overall welfare and health status of a population (Lipesa et al., 2023). LE can be defined as the average number of years a person in a given population is expected to live, based on their year of birth, current age, gender, or geographical location (Ronmi et al., 2023). According to Ritchie et al. (2023), while the average life expectancy of a newborn was 32 years in 1900, this figure more than doubled to reach 71 years by 2021. This substantial increase in life expectancy is attributed to improvements in living standards, economic growth, and poverty reduction, as well as advancements in health-related areas such as nutrition, access to clean water, sanitation, neonatal health, antibiotics, public health interventions, and vaccinations (Roser et al., 2023). Various studies have suggested that LE is influenced by a range of factors, including socioeconomic, environmental, health-related, personal, and immunological determinants.

In the literature, numerous studies have examined the factors determining life expectancy using classical statistical models such as linear regression, generalized linear models, or time series forecasting methods (Bilas et al., 2014; Linden and Ray, 2017). A deep understanding of the factors influencing LE can assist governments and stakeholders in making appropriate health investment and policy decisions. Moreover, in the rapidly expanding field of digital health, comprehending the determinants of LE in detail can aid health professionals and scientists in developing programs and technologies aimed at improving overall well-being and consequently increasing life expectancy. At this point, the emergence of artificial intelligence (AI), machine learning (ML), and data science concepts has made it particularly possible to generate valuable insights from data to achieve this goal (Fosso Wamba and Queiroz, 2021). Additionally, some research has aimed to enhance predictive power by incorporating immunization rates into their models (Agarwal et al., 2019; Lakshmanarao, 2022; Dawoud and Abu-Naser, 2023; Sofi and Yasmin, 2025). Vydehi et al. (2020), in their study utilizing machine learning techniques with various algorithms and datasets to predict life expectancy, found that random forest regression produced the most accurate results for life expectancy prediction. Gill et al. (2023) highlighted the complexity of predicting life expectancy due to multiple influencing factors such as lifestyle, access to healthcare services, and socioeconomic status, employing linear regression (LR) and decision tree classification techniques. Their research demonstrated that linear regression yielded higher accuracy rates compared to decision tree classification. In their study, Ozsahin et al. (2024) utilized ensemble machine learning models such as Random Forest (RF), Light Gradient Boosting Machine (LGBM), Adaptive Boosting (AdaBoost), and eXtreme Gradient Boosting (XGBoost) to predict life expectancy. The results indicated that LGBM demonstrated the best performance among these models. Similarly, recent studies have focused on applying the XGBoost algorithm to predict life expectancy using machine learning approaches. Both Shakeel Ahamad et al. (2025) and Lipesa et al. (2023) employed XGBoost-based models to predict life expectancy across various countries, concluding that these models outperformed other machine learning techniques.

In recent years, machine learning methods have been widely employed for predicting life expectancy, and numerous studies have analyzed the impact of socioeconomic indicators and health expenditures on life expectancy (Alinejad, 2023; Georgiev et al., 2024). Additionally, some research has aimed to enhance predictive power by incorporating immunization rates into their models (Agarwal et al., 2019; Lakshmanarao, 2022; Dawoud and Abu-Naser, 2023; Sofi and Yasmin, 2025). However, although numerous studies have applied ML techniques for LE prediction, few have systematically compared their performances while jointly analyzing socioeconomic and vaccination factors with interpretable ML explainability tools. For these reasons, the primary aim

of this study is to systematically compare the performance of different machine learning models in predicting life expectancy and to detail, on a model-specific basis, the effects of socioeconomic and health indicators such as income, health expenditures, and vaccination coverage rates on life expectancy. Additionally, the study aims to enhance the interpretability of model decision mechanisms through the combined application of Permutation Importance and SHAP analyses. This research fills a gap in the literature by systematically comparing machine learning models for life expectancy prediction and integrating vaccination variables into the analysis. In doing so, it provides both methodological and practical contributions to the fields of health economics, epidemiology, and artificial intelligence applications.

MATERIALS AND METHODS

Data Source and Variables

The data used in this study were obtained from the World Bank (2025) database and cover the period from 2000 to 2022. The database includes a total of 196 countries. Countries with missing data or inaccessible data were excluded from the dataset, resulting in an analysis based on data from 95 countries. Consequently, a total of 2,185 observations were obtained, comprising 23 years of data across these 95 countries.

In this study, life expectancy at birth, total (years) (LEB) was designated as the dependent variable. The independent variables included Immunization, BCG (% of one-year-old children) (BCG); Immunization, DPT (% of children ages 12–23 months) (DPT); Immunization, POL3 (% of one-year-old children) (POL); Immunization, measles (% of children ages 12–23 months) (MEAS); Current health expenditure per capita (current US$) (HEXP); GDP per capita, PPP (constant 2017 international $) (GDP); Inflation, consumer prices (annual %) (INF); and Unemployment, total (% of total labor force) (modeled ILO estimate) (UNEMP). Detailed information regarding the variables used in the study is presented in Table 1.

As the data used in this study were secondary data, ethical approval was not required for this research.

Table 1. Information on the Variables

Variables	Abbreviation	Source
Life expectancy at birth, total (years)	LEB	World Bank (2025)
Immunization, BCG (% of one-year-old children)	BCG	
Immunization, DPT (% of children ages 12-23 months)	DPT	
Immunization, POL3 (% of one-year-old children)	POL	
Immunization, measles (% of children ages 12-23 months)	MEAS	
Current health expenditure per capita (current US$)	HEXP	
GDP per capita, PPP (constant 2017 international $)	GDP	
Inflation, consumer prices (annual %)	INF	
Unemployment, total (% of total labor force) (modeled ILO estimate)	UNEMP	

ANALYSIS

All data preprocessing, statistical analyses, and machine learning modelling in this study were performed using the Python programming language. For data processing and analyses, the libraries pandas, numpy, scikit-learn, xgboost, scipy, and matplotlib were utilized.

In this study, descriptive statistical analyses were first conducted to calculate the mean, standard deviation, minimum, and maximum values of the variables. Subsequently, data preprocessing steps such as logarithmic transformation, Winsorization, and MinMaxScaler normalization were applied to the variables in the dataset to meet model assumptions and improve model performance. Pearson correlation analysis was then performed to determine the linear relationships between variables, and the results were presented using a heatmap. Prior to modelling, the dataset was split into training and test sets, with 80% used for training and 20% for testing. For predicting the relationships between the dependent variable and the selected independent variables, Linear Regression, Random Forest, XGBoost, and Artificial Neural Network (ANN) models were employed. For each model, hyperparameter optimization, cross-validation performance analysis, test set evaluation, Permutation Importance analysis, and SHAP analysis were conducted. Hyperparameter optimization for all models (Random Forest, XGBoost, and ANN) was performed using GridSearchCV with 5-fold cross-validation, ensuring robust parameter selection and preventing overfitting. The Friedman test was used to compare the RMSE values of the models.

MACHINE LEARNING MODELS

Linear Regression

Linear regression is a supervised machine learning method widely used for predicting continuous target variables. This approach aims to make predictions by establishing a linear mathematical relationship between one or more independent variables (X) and a dependent variable (y) (Ngo, 2012).

Random Forest Regression

Random Forest is an ensemble machine learning method composed of multiple decision trees, which combines the results of these trees to improve the accuracy and stability of predictions. This supervised learning algorithm can be used for both classification and regression analyses. In this method, various sub-datasets are created from the original dataset using bootstrap sampling with replacement. Additionally, at each split, a random subset of variables is selected instead of using all variables, thereby reducing correlations among trees. Each decision tree is structured based on the defined splitting criteria (Das et al., 2025).

XGBoost

XGBoost is an advanced version of the Gradient Boosting algorithm, offering high computational efficiency and strong predictive capabilities. This machine learning method incorporates various features such as regularization, computational optimization, and advanced enhancement techniques to prevent overfitting and improve model performance (Mesut et al., 2023).

Artificial Neural Network

Artificial Neural Network (ANN) are supervised machine learning algorithms inspired by the structure and functioning of the human brain (Das et al., 2025). These models are designed to approximately learn complex functional relationships and are widely used in various applications such as classification, regression, and deep learning. ANN architectures consist of layers composed of interconnected neurons, each processing the input it receives to produce an output, which is then passed on to the next layer.

Traditional regression methods rely on strict parametric assumptions and often fail to capture complex nonlinear relationships among variables. In contrast, machine learning models are nonparametric and therefore do not require the assumptions of normality, homoscedasticity, multicollinearity, and autocorrelation that are tested in classical regression models (Umarov, 2025).

RESULTS

Descriptive Statistics

The descriptive statistics of the variables are presented in Table 2. The mean value of the LEB variable was calculated as 67.30 years, with a standard deviation of 8.30. Regarding vaccination rates, the BCG vaccine had a mean coverage of 90.61% with a standard deviation of 11.07, DPT had a mean of 84.78% and a standard deviation of 15.83, polio (Pol3) had a mean of 84.47% and a standard deviation of 15.74, and measles had a mean of 83.64% with a standard deviation of 16.29. For the HEXP variable, the mean per capita health expenditure was calculated as 552.12 USD, with a standard deviation of 692.46 USD. The GDP variable had a mean of 13,684 USD and a standard deviation of 18,696 USD. Among macroeconomic indicators, the INF rate had a mean of 6.52% with a standard deviation of 13.80%, while UNEMP had a mean of 7.45% with a standard deviation of 5.59%.

Table 2. Descriptive Statistics of the Variables

Variables	Mean	Standard Deviation	Minimum	Maximum
LEB	67.30	8.30	14.67	83.60
BCG	90.62	11.08	28.00	99.00
DPT	84.79	15.83	19.00	99.00
POL	84.47	15.75	8.00	99.00
MEAS	83.64	16.30	16.00	99.00
HEXP	552.12	692.46	20.67	6657.93
GDP	13684.27	18696.06	795.77	145591.02
INF	6.52	13.80	-16.86	359.09
UNEMP	7.45	5.59	0.10	34.01

Data Preprocessing

In this study, logarithmic transformation was initially applied to normalize the distributions of the independent variables and to reduce their skewness. This procedure was chosen to strengthen model assumptions and mitigate the effects of extreme outliers (Changyong et al., 2014). All independent variables had positive values; for variables containing zero or negative values, a log(x+1) transformation was applied.

After the logarithmic transformation, Winsorization was applied to reduce the influence of outliers on model parameters. As widely recommended in the literature, the lower 5% and upper 95% limits were used for each variable (Mohamed et al., 2025). This method suppresses the impact of extreme outliers without causing data loss, thereby contributing to the generation of robust estimates. In this way, outlier observations were retained in the dataset with minimal information loss while reducing their undue influence on the data distribution. Winsorization particularly contributes to meeting parametric assumptions and enhancing the stability of analysis results in regression and ANN models (Demir and Sahin, 2023).

In this study, MinMaxScaler normalization was applied to improve the learning performance of models, particularly gradient-based algorithms such as artificial neural networks. With this method, each variable was scaled to the [0,1] range using its own minimum and maximum values. MinMaxScaler normalization standardizes the data, thereby enhancing model performance, and this normalization process positively influences the success and accuracy rates of optimization-based algorithms, especially ANNs (Chepino et al., 2023).

The dataset was divided into training and test sets in an 80%-20% split to evaluate the predictive power and generalizability of the models. The training set contained 1,748 observations, while the test set included 437 observations. This split ratio, widely used in the literature (Khamparia et al., 2021), allows for a robust assessment of the model's learning performance as well as its validation performance on an independent dataset. In this way, the model's generalization ability is ensured, and the risk of overfitting is kept under control.

The boxplot distribution analysis after data transformation and normalization is presented in Figure 1. All variables were scaled to the [0,1] range, with vaccination rate variables showing a concentration near the upper limit and some low-value outliers at the lower end. In contrast, economic and macroeconomic indicators exhibited a more balanced distribution. The LEB variable was also normalized to bring it onto a common scale for analysis, with its median observed at a medium-high level. These results indicate that the applied data transformation procedures effectively reduced the influence of outliers and made the variables suitable for modelling.

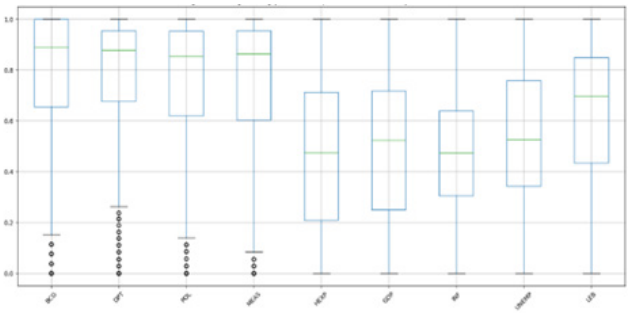


Figure 1. Boxplot Results

Correlation Analysis (Heatmap)

The results of the Pearson correlation analysis performed on the data after Winsorization and logarithmic transformation are presented in Figure 2. According to the analysis, there was a strong positive correlation between LEB and GDP (r=0.75). Similarly, a strong positive correlation was found between LEB and HEXP (r=0.77). The vaccination rates of BCG, DPT, POL, and MEAS showed moderate positive relationships with LEB. A low-level negative correlation was identified between INF and LEB, while UNEMP demonstrated a low-level positive correlation with LEB. The correlation heatmap visualization clearly highlights the strong relationships among LEB, GDP, and HEXP, as well as the vaccination rates. While multicollinearity poses a statistical issue in regression and panel data analyses, it does not adversely affect model performance in nonlinear models since direct statistical significance testing or coefficient interpretation is not conducted in these models.

Figure 2. Heatmap Correlation Matrix

Model Performance

Table 3 presents the 5-fold cross-validation and test set results. As classical linear regression does not require hyperparameter tuning, it was trained directly using sklearn's LinearRegression function. Cross-validation yielded an average RMSE of 4.52, MAE of 3.39, and R^2 of 0.69, indicating that 69% of the variance in the dependent variable was explained. On the test set, RMSE was 4.34, MAE was 3.30, and R^2 increased to 0.74, suggesting stable generalization and no overfitting, with reliable predictive performance on new data.

The Random Forest model was built using sklearn and optimized with GridSearchCV. Cross-validation yielded an average RMSE of 3.02, MAE of 2.04, and R^2 of 0.86, indicating that 86% of the variance was explained in the training data. On the test set, RMSE was 2.84, MAE was 1.81, and R^2 increased to 0.89, suggesting strong generalization performance without overfitting and very high explanatory power on unseen data.

The XGBoost model, based on gradient boosting, was implemented using sklearn and xgboost, with hyperparameters optimized via GridSearchCV. Cross-validation yielded an average RMSE of 3.02, MAE of 2.02, and R^2 of 0.86, explaining 86% of the variance in the training data. On the test set, RMSE decreased to 2.59, MAE to 1.72, and R^2 increased to 0.91, indicating very strong generalization performance without overfitting and high explanatory power on unseen data.

An artificial neural network with an MLP architecture was implemented using sklearn and optimized with GridSearchCV. Cross-validation results showed an average RMSE of 4.08, MAE of 3.06, and R^2 of 0.75, explaining 75% of the variance in the training data. On the test set, RMSE decreased to 3.29, MAE to 2.80, and R^2 increased to 0.80, indicating good generalization performance without overfitting and reasonable explanatory power on unseen data.

According to the findings of the study, the test set and cross-validation performances of the Linear Regression, Random Forest, XGBoost, and Artificial Neural Network models were compared. In terms of RMSE and MAE values, the XGBoost model achieved the lowest error rates, followed by the Random Forest model. The model with the highest R^2 value was XGBoost at 90.7%, with Random Forest ranking second at 88.8%. The performances of the Linear Regression and ANN models were found to be lower compared to these two models. These results indicate that XGBoost outperforms other methods in predicting life expectancy.

The models were ranked based on RMSE to assess predictive accuracy. Since parametric test assumptions were unmet, the Friedman test was used to compare model performances, as recommended for error comparisons on the same dataset (Korkmaz & Onemli, 2011). Results showed significant differences among models (χ^2=14.755, p=0.002). Nemenyi post-hoc analysis

revealed that XGBoost performed significantly better than Linear Regression ($p<0.01$), while the difference between Linear Regression and Random Forest was marginal ($p=0.05$). No other pairwise differences were significant, indicating that XGBoost had superior RMSE performance compared to Linear Regression.

Table 3. Analysis Results of ML Models

Model	Test Set Error			Cross-Validation Error		
	RMSE	MAE	R2	RMSE	MAE	R2
Linear Regression	4.3407	3.2988	0.7385	4.5209	3.3904	0.6949
Random Forest	2.8357	1.8061	0.8884	3.0161	2.0388	0.8645
XGBoost	2.5906	1.7168	0.9068	3.0157	2.016	0.8645
Artificial Neural Network	3.2891	2.7978	0.7965	4.0799	3.0605	0.7525

The models were ranked based on RMSE to assess predictive accuracy. Since parametric test assumptions were unmet, the Friedman test was used to compare model performances, as recommended for error comparisons on the same dataset (Korkmaz & Onemli, 2011). Results showed significant differences among models ($\chi^2=14.755$, $p=0.002$). Nemenyi post-hoc analysis revealed that XGBoost performed significantly better than Linear Regression ($p<0.01$), while the difference between Linear Regression and Random Forest was marginal ($p=0.05$). No other pairwise differences were significant, indicating that XGBoost had superior RMSE performance compared to Linear Regression.

Table 4. Friedman Test Results

	Linear Regression	Random Forest	XGBoost	ANN
Linear Regression	1			
Random Forest	0.05*	1		
XGBoost	0.00*	0.76	1	
ANN	0.61	0.53	0.09	1

Friedman test statistic = 14.755, p=0.002

Importance Analyses of the Models

Table 5 presents the results of the Permutation Importance and SHAP analyses, evaluating the variable importance and their effects on LEB across four different models (LN, RF, XGBoost, ANN).

In the linear regression model, the Permutation Importance analysis indicated that GDP and HEXP were the most influential variables. The SHAP summary plot confirmed these findings, showing strong positive effects of GDP and HEXP on LEB. In contrast, the contributions of DPT and UNEMP were relatively limited, while variables such as POL, MEAS, BCG, and INF showed minimal impact.

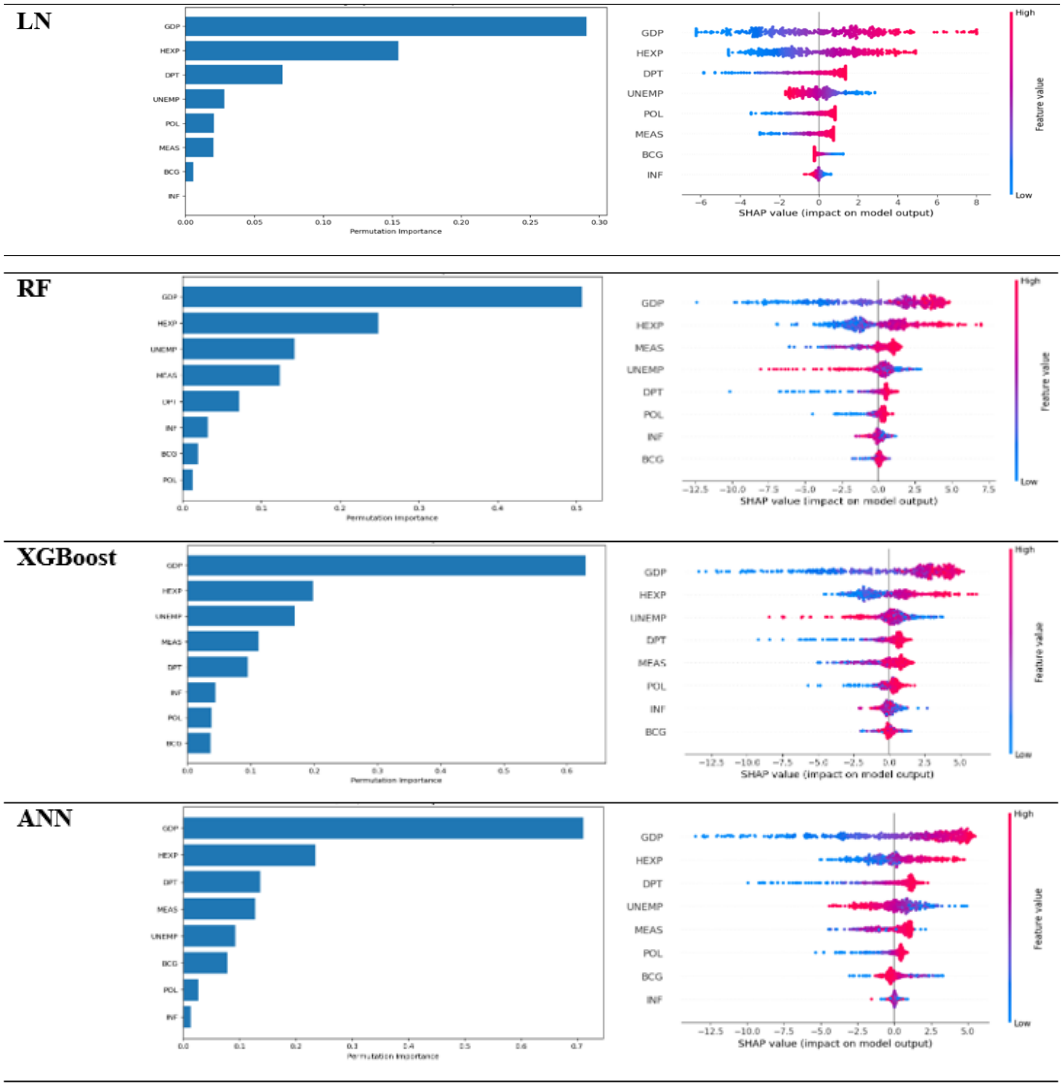
For the Random Forest model, GDP again emerged as the most important variable, followed by HEXP, MEAS, and UNEMP. The SHAP analysis supported these results, demonstrating strong positive effects of GDP and HEXP, with MEAS and UNEMP having moderate impacts. Conversely, DPT, POL, INF, and BCG showed very limited contributions.

In the XGBoost model, Permutation Importance analysis showed that GDP had the highest contribution, followed by HEXP and UNEMP. The SHAP summary plot confirmed the prominent positive effects of GDP and HEXP on life expectancy, while UNEMP, DPT, and MEAS exhibited moderate effects. POL, INF, and BCG contributed negligibly.

For the artificial neural network model, Permutation Importance indicated that GDP and HEXP were again the most influential variables. The SHAP summary plot corroborated these results, revealing strong positive impacts of GDP and HEXP on LEB, while DPT, UNEMP, and MEAS had relatively limited contributions. BCG, POL, and INF showed minimal influence.

Overall, GDP and HEXP were consistently identified as the most important variables with strong positive effects on life expectancy across all models. Variables such as UNEMP, DPT, and MEAS provided moderate contributions in some models, whereas POL, INF, and BCG had limited effects. These analyses enhance the understanding of the models' decision-making processes and enable reliable interpretation of variable importance.

Table 5. Permutation Importance and SHAP Analysis Results of the Models



DISCUSSION

According to the findings obtained in this study, the XGBoost model had the lowest error rates in terms of RMSE and MAE values and the highest R^2 value. This model was followed by Random Forest, Artificial Neural Network, and Linear Regression models, respectively. These results indicate that the XGBoost model outperformed other methods in predicting life expectancy. In the literature, studies using machine learning methods to predict life expectancy have also demonstrated that XGBoost outperforms other machine learning techniques (Pandey and Chhikara, 2020; Lipesa et al., 2023; Dangety et al., 2024; Shakeel Ahamad et al., 2025). In this context, the findings of the present study are consistent with those reported in the literature. Although the XGBoost model demonstrated the highest accuracy rates and lowest error values in the study findings, the Random Forest model also exhibited notable performance. Due to its decision tree-based structure, the Random Forest model offers advantages such as rapid computation time and high accuracy in datasets with high variability and complex interactions, and it also provides ease of interpretation, which makes it widely preferred in studies (Vydehi et al., 2020; Das et al., 2025). Artificial Neural Networks (ANNs), on the other hand, can achieve superior performance in modelling large datasets with complex nonlinear relationships and offer flexible modelling capabilities without requiring parametric assumptions (Sharma et al., 2017). However, in the present study, the ANN model was outperformed by ensemble tree-based models such as XGBoost and Random Forest. This performance gap can be attributed to the nature of the dataset; for tabular data with a relatively limited number of observations (n=2185), gradient boosting algorithms often demonstrate superior efficiency compared to deep learning architectures, which typically require much larger datasets to realize their full predictive potential. Therefore, although the XGBoost model generally showed the best performance, Random Forest and ANN models should also be considered as alternative approaches depending on data structures, variable characteristics, and application objectives. The ability of the XGBoost algorithm to better model inter-variable relationships in complex datasets makes it stand out. While Random Forest and ANN also achieved high accuracy rates, the lower error values obtained by XGBoost further strengthen its potential for use in critical fields such as healthcare. These findings indicate the importance of using gradient boosting-based models in predicting multifactorial and complex target variables such as life expectancy.

According to other important findings from this research, GDP and HEXP were identified as the strongest determinants of life expectancy across all models. Additionally, in the linear regression model, the most important determinants after GDP and HEXP were DPT and UNEMP; in the Random Forest model, MEAS and UNEMP; in the XGBoost model, UNEMP and DPT; and in the ANN model, DPT and UNEMP. A study conducted by Dangety et al. (2024) also identified income and health expenditures as the most influential factors in determining life expectancy. Other studies have similarly found that determinants such as vaccination rates (Lakshmanarao, 2022; Sofi and Yasmin, 2025) and unemployment (Sofi and Yasmin, 2025) have significant effects on life expectancy. In this regard, the findings of the present study are seen to support the existing literature.

CONCLUSION

The results obtained from this study demonstrate that socioeconomic factors, health expenditures, and vaccination programs are strong predictors of public health outcomes and longevity. This research highlighted the potential of machine learning methods in enhancing the accuracy and reliability of life expectancy predictions, providing valuable insights for

policymakers, healthcare professionals, and academics seeking to understand and improve population health. Future studies that incorporate factors such as education, environmental sustainability, and health infrastructure as variables in the models will expand the scope of life expectancy predictions and contribute to offering more comprehensive recommendations for policymakers. In line with the findings of this study, it is crucial for health administrators and policymakers to prioritize strategies that enhance economic prosperity, ensure the effective utilization of health expenditures, and strengthen vaccination programs in their planning and implementation efforts aimed at increasing life expectancy. The statistical link between per capita income and health expenditures on determining life expectancy indicates that health policies should be integrated not only with treatment services but also with economic and social development initiatives. Furthermore, the high predictive power of vaccination rates, such as for diphtheria, highlights the critical importance of expanding immunization programs within the scope of preventive healthcare services. Considering the observed negative correlation of unemployment on life expectancy, it would be beneficial for health administrators to collaborate with employment-enhancing social policies and to develop health literacy and psychosocial support programs targeted at unemployed individuals. In conclusion, it is recommended that health administrators and policymakers incorporate data-driven decision-making mechanisms into their strategic plans. For example, implementing national digital health dashboards can enable real-time monitoring and informed resource allocation. Additionally, the integration of machine learning-based predictive models and multisectoral collaborations will help optimize resource utilization and maximize contributions to public health.

This study has some limitations regarding data scope and variable selection. First, due to missing values in several indicators, the final analysis was conducted on 95 countries. This reduction may lead to a 'survivorship bias', as the excluded nations are predominantly lower-income countries with less robust data reporting systems. Consequently, the generalizability of the results to these specific regions may be limited. Furthermore, while the model focuses on vaccination and socioeconomic factors, certain variables such as educational attainment were not included in the analysis due to data unavailability across the entire timeframe and country set. Since education is a significant predictor of life expectancy, its absence is a notable constraint. Future research should aim to incorporate these additional dimensions as more comprehensive datasets become available. Finally, this study indicates statistical associations rather than definitive causal relationships.

REFERENCES

Agarwal, P., Shetty, N., Jhajharia, K., Aggarwal, G., & Sharma, N. V. (2019). Machine learning for prognosis of life expectancy and diseases. International Journal of Innovative Technology and Exploring Engineering, 8(10), 1765-1771. http://doi.org/10.35940/ijitee.J9156.0881019

Alinejad, M. (2023). Analyzing the Impact of Health, Economic, and Demographic Factors on Life Expectancy: A Comparative Study of Developed and Developing Countries. https://tspace.library.utoronto.ca/

Bilas, V., Franc, S., & Bošnjak, M. (2014). Determinant factors of life expectancy at birth in the European Union countries. Collegium Antropologicum, 38(1), 1-9. https://hrcak.srce.hr/120775

Changyong, F. E. N. G., Hongyue, W. A. N. G., Naiji, L. U., Tian, C. H. E. N., Hua, H. E., Ying, L. U., & Xin, M. T. (2014). Log-transformation and its implications for data analysis. Shanghai Archives of Psychiatry, 26(2), 105. http://doi.org/10.3969/j.issn.1002-0829.2014.02.009

Chepino, B. G., Yacoub, R. R., Aula, A., Saleh, M., & Sanjaya, B. W. (2023). Effect of MinMax normalization on ORB data for improved ANN accuracy. Journal of Electrical Engineering, Energy, and Information Technology (J3EIT), 11(2), 29-35. https://jurnal.untan.ac.id/index.php/j3eituntan/article/view/68689/0

Dangety, S., Kasulu, K. V., Swetha, G., Begum, Z., Bhashyam, K. M., & Lakshmanarao, A. (2024). Exploring socioeconomic influences on life expectancy through machine learning ensemble regression techniques. In 2023 4th International Conference on Intelligent Technologies (CONIT) (pp. 1-5). IEEE. http://doi.org/10.1109/CONIT61985.2024.10626849

Das, A., Uddin, M. M., & Karim, M. R. (2025). Predicting life expectancy using machine learning techniques. International Journal of Statistical Sciences, 25(1), 55-70. https://doi.org/10.3329/ijss.v25i1.81045

Dawoud, A. M., & Abu-Naser, S. S. (2023). Predicting life expectancy in diverse countries using neural networks: Insights and implications. Journal of Computational Health, Advance online publication. https://doi.org/10.1016/j.jochealth.2023.101234

Demir, S., & Sahin, E. K. (2023). Application of state-of-the-art machine learning algorithms for slope stability prediction by handling outliers of the dataset. Earth Science Informatics, 16(3), 2497-2509. https://doi.org/10.1007/s12145-023-01059-8

Fosso Wamba, S., & Queiroz, M. M. (2023). Responsible artificial intelligence as a secret ingredient for digital health: Bibliometric analysis, insights, and research directions. Information Systems Frontiers, 25(6), 2123-2138. https://doi.org/10.1007/s10796-021-10142-8

Georgiev, V., Hadzhikoleva, S., & Hadzhikolev, E. (2024). Impact of global country indicators on life expectancy. Computer Science and Interdisciplinary Research Journal, 1(1). https://doi.org/10.70862/CSIR.2024.0101-04

Gill, K. S., Anand, V., Chauhan, R., & Sharma, M. (2023). Predicting life expectancy using machine learning approach through linear regression and decision tree classification techniques. In 2023 3rd International Conference on Smart Generation Computing, Communication and Networking (SMART GENCON) (pp. 1-6). IEEE. https://doi.org/10.1109/SMARTGENCON60755.2023.10441837

Khamparia, A., Singh, P. K., Rani, P., Samanta, D., Khanna, A., & Bhushan, B. (2021). An internet of health things□driven deep learning framework for detection and classification of skin cancer using transfer learning. Transactions on Emerging Telecommunications Technologies, 32(7), e3963. https://doi.org/10.1002/ett.3963

Korkmaz, A., & Onemli, M. B. (2011). Model selection by Friedman statistics. Pakistan Journal of Statistics and Operation Research. https://doi.org/10.18187/pjsor.v7i2-Sp.285

Lakshmanarao, A. (2022). Life expectancy prediction through analysis of immunization and HDI factors using machine learning regression algorithms. International Journal of Online & Biomedical Engineering, 18(13). https://doi.org/10.3991/ijoe.v18i13.33315

Linden, M., & Ray, D. (2017). Life expectancy effects of public and private health expenditures in OECD countries 1970–2012: Panel time series approach. Economic Analysis and Policy, 56, 101-113. https://doi.org/10.1016/j.eap.2017.06.005

Lipesa, B. A., Okango, E., Omolo, B. O., & Omondi, E. O. (2023). An application of a supervised machine learning model for predicting life expectancy. SN Applied Sciences, 5(7), 189. https://doi.org/10.1007/s42452-023-05404-w

Mesut, B., Başkor, A., & Aksu, N. B. (2023). Role of artificial intelligence in quality profiling and optimization of drug products. In Y. Zhao (Ed.), A handbook of artificial intelligence in drug delivery (pp. 35-54). London, UK: Academic Press. https://doi.org/10.1016/B978-0-323-89925-3.00003-4

Mohamed, S. D., Ismail, M. T., & Ali, M. K. B. M. (2025). Improving detectability of the indicator saturation approach through winsorization: An empirical study in the cryptocurrency market. Statistics in Transition New Series, 26(1), 155-181. https://www.ceeol.com/search/article-detail?id=1326166

Ngo, T. H. D., & La Puente, C. A. (2012). The steps to follow in a multiple regression analysis. In Proceedings of the SAS Global forum (pp. 22-25). Princeton, NJ, USA: Citeseer. https://citeseerx.ist.psu.edu/document?repid=rep1&type=pdf&doi=24dc25fdd1921d3b5986690e1659b27271272e96

Ozsahin, D. U., Emegano, D. I., David, L. R., Hussain, A. J., Uzun, B., & Ozsahin, I. (2024). Global life expectancy prediction using machine learning ensemble techniques. In 2024 17th International Conference on Development in eSystem Engineering (DeSE) (pp. 423-427). IEEE. https://doi.org/10.1109/DeSE63988.2024.10912031

Pandey, A., & Chhikara, R. (2020). Analysis of life expectancy using various regression techniques. In 2020 2nd International Conference on Advances in Computing, Communication Control and Networking (ICACCCN) (pp. 209-213). IEEE. https://doi.org/10.1109/ICACCCN51052.2020.9362914

Ritchie, H., Roser, M., & Ortiz-Ospina, E. (2023). Life expectancy. Our World in Data. https://ourworldindata.org/life-expectancy

Ronmi, A. E., Prasad, R., & Raphael, B. A. (2023). How can artificial intelligence and data science algorithms predict life expectancy—An empirical investigation spanning 193 countries. International Journal of Information Management Data Insights, 3(1), 100168. https://doi.org/10.1016/j.jjimei.2023.100168

Roser, M., Rohenkohl, B., Arriagada, P., Hasell, J., Ritchie, H., & Ortiz-Ospina, E. (2023). Economic growth. Our World in Data. https://ourworldindata.org/economic-growth

Shakeel Ahamad, S., Kumar, K. P., Ganesh, S. S., Alharbi, F., Alharby, S. A., Dendukuri, V. S., & Vani, K. S. (2025). Machine learning-based life expectancy prediction in developed and developing regions. IEEE Access. https://doi.org/10.1109/ACCESS.2025.3560890

Sharma, S., Sharma, S., & Athaiya, A. (2017). Activation functions in neural networks. Towards Data Sci, 6(12), 310-316. https://www.ijeast.com/papers/310-316,Tesma412,IJEAST.pdf

Sofi, S. A., & Yasmin, E. (2025). Determinants of life expectancy: Evidence from World Bank income groups using a panel dummy interaction approach. International Journal of Health Care Quality Assurance. https://doi.org/10.1108/IJHCQA-03-2025-0026

Umarov, A. (2025). Application of nonparametric methods in economics and business: A review. SSRN. https://doi.org/10.2139/ssrn.5272342

Vydehi, K., Manchikanti, K., Kumari, T. S., & Shah, S. A. (2020). Machine learning techniques for life expectancy prediction. International Journal of Advanced Trends in Computer Science and Engineering, 9(4), 4503-4507. http://www.warse.org/IJATCSE/static/pdf/file/ijatcse45942020.pdf

World Bank. (2025). World Bank DataBank: Databases [Online database]. Retrieved July 7, 2025, from https://databank.worldbank.org/databases