Cilt: 12 Sayı: 2, 2025, ss. 462-476
PAMUKKALE ÜNİVERSİTESİ
İŞLETME ARAŞTIRMALARI DERGİSİ (PIAR)
Pamukkale University Journal of Business Research
https://dergipark.org.tr/tr/pub/piar

Predicting Global Health Expenditures Using Machine Learning and Regularized Regression Methods

Makine Öğrenmesi ve Düzenlileştirilmiş Regresyon Yöntemleri ile Küresel Sağlık Harcamalarının Tahmini

Hakan ÖZTÜRK [1] **Elvan HAYAT** [2*]

[1] Aydın Adnan Menderes University, ozturk@adu.edu.tr, ORCID: 0000-0001-8112-4934
[2] Aydın Adnan Menderes University, elvan.hayat@adu.edu.tr, ORCID: 0000-0001-8200-8046
[*] Yazışılan Yazar/Corresponding author

Makale Geliş/Received: 28.09.2025 Makale Kabul/Accepted: 16.10.2025

Araştırma Makalesi / Research Paper DOI: 10.47097/piar.1792425

Abstract

Health expenditures are crucial for countries' economic sustainability and the effectiveness of health policies. Accurately modeling these expenditures is complex and requires methods beyond classical regression. This study aimed to estimate per capita health expenditures using machine learning and regularized regression approaches based on 2022 World Bank data from 190 countries.

Missing values were imputed using the Multiple Imputation by Chained Equations (MICE) method. The dependent variable was per capita health expenditure, while independent variables included socioeconomic and demographic indicators. Six models—Support Vector Regression (SVR), Random Forests (RF), Extreme Gradient Boosting (XGBoost), Elastic Net, Lasso, and Ridge—were compared using RMSE, MAE, and R^2 metrics. SVR achieved the best performance (RMSE = 463 ± 13.3, R^2 = 0.940 ± 0.003). XGBoost yielded the lowest MAE (262 ± 15.5) with high accuracy (R^2 = 0.923 ± 0.007). GDP per capita was the most important predictor, followed by the proportion of elderly population, life expectancy, and urbanization rate. SVR and XGBoost models demonstrated high predictive power, highlighting their potential as decision-support tools for forecasting health expenditures.

Keywords: *Health Expenditure, Machine Learning, XGBoost, Support Vector Regression, Random Forests.*

JEL Kodları: *I10, C1, C45.*

Öz

Sağlık harcamaları, ülkelerin ekonomik sürdürülebilirliği ve sağlık politikalarının etkinliği açısından kritik öneme sahiptir. Bu harcamaların doğru biçimde modellenmesi karmaşık bir süreçtir ve klasik regresyon yöntemlerinin ötesinde yaklaşımlar gerektirir. Bu çalışma, 190 ülkenin 2022 yılına ait Dünya Bankası verileri kullanılarak kişi başına sağlık harcamalarını makine öğrenmesi ve düzenlileştirilmiş regresyon yöntemleriyle tahmin etmeyi amaçlamıştır.

Veri bütünlüğünü sağlamak için eksik değerler Zincirleme Denklemlerle Çoklu Atama (MICE) yöntemiyle tamamlanmıştır. Bağımlı değişken kişi başına sağlık harcaması olup, bağımsız değişkenler sosyoekonomik ve demografik göstergeleri içermektedir. Altı model—Destek Vektör Regresyonu (SVR), Rastgele Ormanlar (RF), Aşırı Gradyan Artırma (XGBoost), Elastic Net, Lasso ve Ridge regresyonu—RMSE, MAE ve R^2 ölçütleri kullanılarak karşılaştırılmıştır. En iyi performans SVR modeliyle elde edilmiştir (RMSE = 463 ± 13.3, R^2 = 0.940 ± 0.003). XGBoost modeli en düşük MAE değerine (262 ± 15.5) ve yüksek doğruluk oranına (R^2 = 0.923 ± 0.007) ulaşmıştır. Kişi başına düşen GSYİH en güçlü yordayıcı olurken, yaşlı nüfus oranı, yaşam beklentisi ve kentleşme oranı ikincil katkılar sağlamıştır. SVR ve XGBoost modelleri yüksek tahmin gücü sergileyerek sağlık harcamalarının öngörülmesinde politika yapıcılar için değerli karar destek araçları olarak öne çıkmaktadır.

Anahtar Kelimeler: *Sağlık Harcaması, Makine Öğrenmesi, XGBoost, Destek Vektör Regresyonu, Rastgele Ormanlar.*

JEL Codes: *I10, C1, C45.*

Atıf için (Cite as): Öztürk, H., & Hayat, E. (2025). Predicting global health expenditures using machine learning and regularized regression methods. *Pamukkale Üniversitesi İşletme Araştırmaları Dergisi*, *12*(2), 462-476. https://doi.org/10.47097/piar.1792425

1. INTRODUCTION

Health expenditures are a fundamental macro-indicator that directly affects the health status of societies, access to health services, and the sustainability of health systems. The level of health expenditure per capita is affected by many demographic, socio-economic, and institutional determinants. Current Organization for Economic Co-operation and Development (OECD) comparisons and World Bank definitions emphasize that the current health expenditure per capita indicator covers the sum of public and private expenditures and that there are significant differences between countries (OECD, 2023a; World Bank, 2023). This diversity exhibits a dynamic that changes over time, along with population aging, urbanization rates, financing arrangements, and differences in service delivery models.

Comprehensive reviews in the literature on the determinants of health expenditure per capita indicate that age structure, health and social expenditure composition, institutional factors, and labor market indicators, particularly per capita income, are prominent. The income-health expenditure relationship exhibits different elasticities in high-income and emerging economies; it has been reported that social expenditures and health expenditures tend to increase together in high-income countries (Martin et al., 2011; Mihaylova et al., 2011; Sinha et al., 2016). These findings imply that health expenditure is a multidimensional and context-sensitive outcome, and therefore linear assumptions in forecasting models may often be inadequate.

The fundamental statistical challenges encountered in modeling health expenditures include strong correlations between variables, potential non-linear relationships, interactions, and a distribution that is excessively skewed to the right. In this context, machine learning regression methods are increasingly being used alongside classical linear approaches due to their ability to capture non-linear patterns and complex interactions. For example, Random Forest (RF) provides robust estimates with high numbers of variables, while Support Vector Regression (SVR) provides a flexible functional approach based on margin optimization using statistical learning theory while gradient-boosted trees achieve competitive predictive success thanks to regularization and efficiency-focused learning algorithms (Breiman, 2001; Friedman, 2001; Chen & Guestrin, 2016).

Country- and period-specific studies on this topic show that machine learning approaches yield promising results in predicting healthcare expenditures. In a study using the 1990–2019 period in Türkiye, SVR and Gaussian Process Regression (GPR) methods were compared; it was reported that SVR provided lower error metrics in the testing phase (Güleryüz, 2021). In another study, Lasso, RF, and SVR methods were tested with different hyperparameters on single-year (2013) data from 214 countries; it was found that the RF method was relatively superior in terms of coefficient of determination (R^2), root mean squared error (RMSE), and mean absolute error (MAE) (Çınaroğlu, 2017). These findings indicate that method and hyperparameter selection significantly affect prediction performance.

This study aims to estimate health expenditure per capita using the World Bank's 2022 indicators covering 190 countries, employing machine learning regression models SVR, RF, and Extreme Gradient Boosting (XGBoost), and regularized regression models Ridge, Lasso, and Elastic Net; to comparatively evaluate the models using RMSE, MAE, and R^2 metrics; and to reveal variable importance. Thus, by providing evidence on the relative performance of

different machine learning and regularized linear approaches on a current and comprehensive set of countries, the study aims to contribute to the use of data-driven forecasting tools in health economics and policy design.

2. METHODS

2.1. Data Source and Sample

The data used in this study were obtained from the World Bank (World Development Indicators) database. The countries included in the study comprise a total of 190 countries with per capita health expenditure (health_exp_pc) data for 2022. Thus, the sample provides a globally representative distribution, excluding countries that were omitted due to data deficiencies. Table 1 explains the definitions, abbreviations, and units of the variables used in this study.

Table 1. Definition of the Variables

Variable	Abbreviation	Unit
Current health expenditure per capita	health_exp_pc	US$
GDP per capita	gdp_pc	US$
GDP growth	gdp_growth	annual %
Population ages 65 and above	age65plus	% of total population
Life expectancy at birth	life_exp	years
Urban population	urban_pop	% of total population
Unemployment, total	unemployment	% of labor force
Inflation, consumer prices	inflation	annual %
Out-of-pocket expenditure	oop_share	% of current health exp.
General government health exp.	gov_share	% of current health exp.
Income Level of Countries	income_level	4-level ordinal
Region of Countries	region	7-category nominal

Source: World Bank 2022

2.2. Handling Missing Data

The dataset contains missing observations only in the unemployment rate and inflation rate variables. To minimize the bias and information loss that missing data may cause, the Multiple Imputation by Chained Equations (MICE) method, one of the multiple imputation methods, was used. This method allows missing values to be estimated multiple times based on statistical relationships with related variables, enabling variance preservation and more reliable results in predictive models (Azur et al., 2011).

2.3. Regression Methods

In this study, per capita health expenditure ($Y = health\_\text{exp}\_pc$) was considered as the dependent variable. The independent variables consist of indicators representing the socioeconomic and demographic characteristics of countries. These indicators are generally expressed by the vector $X = (X_1, X_2, \dots, X_p)$ and include variables related to per capita income, age structure of the population, life expectancy, urbanization rate, unemployment, inflation, and the financing structure of health expenditures.

Six different regression methods were applied in the study to capture linear and non-linear relationships.

Ridge Regression: Ridge regression prevents the coefficients from becoming excessively large by adding an L2 norm penalty term to the least squares method (Hoerl and Kennard, 1970). The objective function in ridge regression is:

$$\min_{\beta} \left\{ \sum_{i=1}^{n} (y_i - \beta_0 - x_i^T \beta)^2 + \lambda \sum_{j=1}^{p} \beta_j^2 \right\} \quad (1)$$

Here, y_i is the i^{th} observation of the dependent variable, x_i is the vector of independent variables for the i^{th} observation, β is the vector of regression coefficients, λ is the regularization (penalty) parameter, and $\sum_{j=1}^{p} \beta_j^2$ represents the squared L2 norm. Ridge regression reduces the coefficients but does not set any of them to zero. It is particularly used to reduce multicollinearity issues.

Lasso Regression: Lasso (Least Absolute Shrinkage and Selection Operator) works with an L1 norm penalty and performs variable selection by directly setting some coefficients to zero (Tibshirani, 1996). The objective function in Lasso regression is:

$$\min_{\beta} \left\{ \sum_{i=1}^{n} (y_i - \beta_0 - x_i^T \beta)^2 + \lambda \sum_{j=1}^{p} |\beta_j| \right\} \quad (2)$$

Here, y_i is the i^{th} observation of the dependent variable, x_i is the vector of independent variables for the i^{th} observation, β is the vector of regression coefficients, λ is the regularization (penalty) parameter, and $\sum_{j=1}^{p} |\beta_j|$ represents the L1 norm. Lasso both regularizes and provides variable selection. However, when there is high correlation among independent variables, the selections may be inconsistent.

Elastic Net Regression: Elastic Net combines Ridge (L2) and Lasso (L1) penalties. Objective function:

$$\min_{\beta} \left\{ \frac{1}{2n} \sum_{i=1}^{n} (y_i - \beta_0 - x_i^T \beta)^2 + \lambda(\alpha \|\beta\|_1 + (1-\alpha)\|\beta\|_2^2) \right\} \quad (3)$$

Here, $\alpha \in [0,1]$ represents the balance parameter between Ridge and Lasso, $\|\beta\|_1$ represents the L1 norm, and $\|\beta\|_2^2$ represents the squared L2 norm. The Elastic Net method is Ridge when $\alpha = 0$, Lasso when $\alpha = 1$, and a mixture of the two methods when $0 < \alpha < 1$. Elastic Net provides more stable results than Lasso when there are highly correlated variables, and more flexible results than Ridge (Zou and Hastie, 2005).

Support Vector Regression (SVR): The objective in SVR is to ensure that the difference between the observed values and the predicted values remains within a certain error tolerance (ε). In this context, the regression function is defined as follows:

$$f(x) = \langle w, x \rangle + b \quad (4)$$

Here, $x \in R^p$ denotes the vector of independent variables, w denotes the weight vector, and b denotes the constant term. The objective in SVR is to find a function $f(x)$ that satisfies the

condition $|y_i - f(x_i)| \leq \varepsilon$. For errors outside this tolerance, penalty parameter C is applied; thus, a balance is achieved between the model's error flexibility and complexity. Nonlinear relationships can be modeled using kernel functions such as the Radial Basis Function (RBF) ($K(x_i, x_j) = exp(-\gamma \|x_i - x_j\|^2)$) (Vapnik, 1995).

Random Forest (RF): RF consists of a large number of decision trees using a combination of bootstrap sampling (bagging) and random variable selection (Breiman, 2001). Each tree is constructed as $T_b(x), b = 1, 2, \ldots, B$. Final prediction:

$$\hat{f}(x) = \frac{1}{B}\sum_{b=1}^{B} T_b(x) \tag{5}$$

Here, B represents the total number of trees created. The best split is made from m randomly selected variables for splitting at each node. This process reduces the risk of overfitting and increases the model's generalizability. RF also stands out for its ability to calculate variable importance. Importance measures are typically based on the decrease in node impurity (Gini importance) or the increase in prediction error.

Extreme Gradient Boosting (XGBoost): XGBoost is a regularized and optimized version of the Gradient Boosted Trees algorithm (Chen and Guestrin, 2016). Objective function:

$$Obj(\theta) = \sum_{i=1}^{n} l(y_i, \hat{y}_i) + \sum_{k=1}^{K} \Omega(f_k) \tag{6}$$

Here, y_i denotes the observed value, $\hat{y}_i$ denotes the predicted value, l denotes the loss function, f_k, denotes the k^{th} decision tree, and $\Omega(f_k)$ denotes the regularization term. In each iteration, a new tree is added using gradient descent to correct the errors of the existing model. XGBoost applies recursive tree pruning and increases computational efficiency with recursive parallelization.

2.4. Model Evaluation Measures

The prediction performance of the models used in the study was evaluated based on three key metrics: RMSE, MAE and R^2.

Root Mean Square Error (RMSE): It is the square root of the average of the squares of the differences between the estimated values ($\hat{y}_i$) and the observed actual values (y_i):

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2} \tag{7}$$

RMSE is a measure that emphasizes high deviations in forecasts, as it assigns greater weight to large errors.

Mean Absolute Error (MAE): It is the average of the absolute values of the differences between the estimated values and the actual values:

$$MAE = \frac{1}{n}\sum_{i=1}^{n}|y_i - \hat{y}_i| \quad (8)$$

MAE reflects the error distribution more evenly and is relatively easy to interpret.

Coefficient of Determination (R^2): It shows the extent to which the model explains the variance of the dependent variable. It is calculated based on the sum of squares total (SST) and the sum of squares residual (SSR):

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2} \quad (9)$$

Here, $\bar{y}$ represents the mean of the dependent variable. The R^2 value ranges from 0 to 1, and as it approaches 1, the explanatory power of the model increases.

When these three criteria are used together, both the error levels (RMSE, MAE) and the explanatory power (R^2) of the models are evaluated. This allows for an objective performance comparison between different regression methods.

2.5. Model Training and Hyperparameter Optimization

Hyperparameter optimization was applied during the training of prediction models to improve performance and prevent overfitting. Before training the models, all continuous variables were scaled using z-transformation (standardization) so that their mean was 0 and their standard deviation was 1. This process was performed to prevent weight imbalances that could arise from different variable scales, particularly in SVR and regularized regression models (Lasso, Ridge, Elastic Net), and to increase model stability. The dataset was randomly split into 80% training and 20% test subsets. A 5-fold cross-validation strategy was applied on the training data to optimize the model hyperparameters. This cross-validation was used to evaluate each model's hyperparameter combinations and select the best parameters. The final model performance was then evaluated on an independent test set using the best hyperparameter combinations determined during this process. For each model, hyperparameter spaces were defined considering the ranges suggested in the literature, and small-scale regular or random grid searches were performed. Hyperparameter ranges and optimization details are summarized in Table 2.

Table 2. Hyperparameter Ranges and Optimization Details of the Models

Model	Hyperparameter(s)	Range / Value	Number of Combinations	Notes
Lasso regression	penalty (λ)	10^{-6} – 1 (log-scaled, 10 levels)	10	mixture (α) = 1
Ridge regression	penalty (λ)	10^{-6} – 1 (log-scaled, 10 levels)	10	mixture (α) = 0
Elastic Net	penalty (λ), mixture (α)	λ: 10^{-6} – 1; α: 0.1 – 0.9	40 (10x4)	Regular grid search

SVR	cost (C), rbf_sigma, margin (ε)	C: random; σ: random; ε = 0.1 (fixed)	12 (random)	RBF kernel; regression mode
Random Forest	mtry, min_n	mtry: 1 – total number of predictors; min_n: default range	12 (random)	trees = 500 (fixed)
XGBoost	trees, tree_depth, learning_rate, loss_reduction, min_n, mtry, sample_size	trees: 200–500; depth: 2–8; learning_rate: 0.001–0.2 (log10); others: default ranges	15 (random)	grid_space_filling approach

The hyperparameter settings for each model were selected based on the combination that yielded the lowest RMSE value on the validation set. The final models were retrained with these parameters, and the RMSE, MAE, and R^2 performance metrics were calculated on the test set. The analysis process was repeated for each completed dataset obtained with MICE; the results were combined in accordance with Rubin's pooling principles.

2.6. Software and Application

The R program (version 4.3) was used for data organization and analysis. The "mice" package was used for multiple imputation of missing values, while the "caret", "glmnet", "randomForest", and "xgboost" packages were preferred for regression analyses.

3. RESULTS

The dataset contains missing observations in the unemployment and inflation rate variables, with values missing for 14 and 19 countries, respectively. Little's MCAR test was applied to determine the missing data mechanism, and the test results showed no statistically significant difference ($p = 0.425$). This finding indicates that the missing data are consistent with the "missing completely at random" (MCAR) mechanism. Therefore, the MICE method was applied to prevent missing data from causing information loss and bias in the analysis results. During the imputation process, the Predictive Mean Matching (PMM) method, which is a suitable approach for continuous variables, was preferred, with the number of multiple assignments set to 20 and the number of iterations for each assignment set to 10. The distribution of missing observations by variable is shown in the heat map presented in Figure 1, revealing that the missingness was limited to only these two variables, with rates of 7% and 10%, respectively. As a result of this process, the missing values were obtained in 20 different completed data sets, and the analyses were performed separately on each set. Subsequently, the results obtained from these analyses were combined using Rubin's pooling rules, and the final findings were reported (Rubin, 1987).

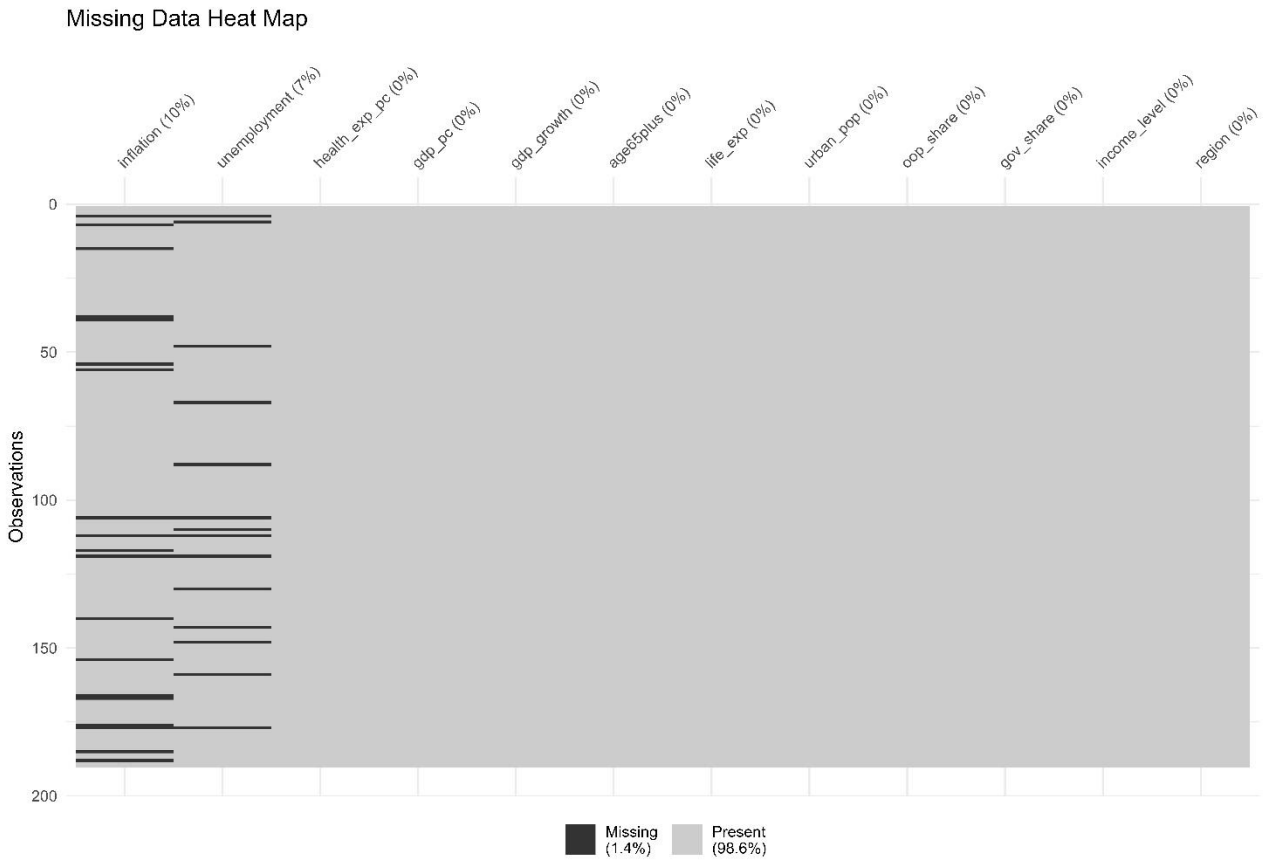

Figure 1. Missing Data Heat Map

Descriptive statistics for the dataset are presented in Table 3. The average per capita health expenditure was 1313 ± 2113 USD, while the per capita GDP was found to be 17358 ± 27459 USD. While 9.4% of the population was aged 65 and over, the average life expectancy at birth was calculated to be 72.3 ± 8.1 years. The unemployment rate is 7.27 ± 5.93%, and the inflation rate is 12.38 ± 19.58%. Looking at the financing structure of health expenditures, out-of-pocket expenditures account for 30.0 ± 18.5%, while public expenditures account for 52.1 ± 22.6%. When examining the distribution of countries according to income level, 30.5% of countries are in the high-income group and 28.4% are in the lower-middle-income group. In terms of regional distribution, Europe and Central Asia have the highest representation at 26.2%, while North America has the lowest representation at 1.1%.

Table 3. Descriptive Statistics of Dataset

Variables	**Mean±SD and median (25.-75.percentile) / n (%)**
health_exp_pc	1313 ± 2113 418 (91 – 1396)
gdp_pc	17358 ± 27459 6522 (2207 – 20832)
gdp_growth	4.69 ± 6.27 4.25 (2.56 – 6.30)
age65plus	9.40 ± 6.99

	6.43 (3.60 – 15.09)
life_exp	72.3 ±8.1 73.1 (67.1 – 77.5)
urban_pop	60.1 ± 22.9 59.9 (42.4 – 79.2)
unemployment	7.27 ± 5.93 5.27 (3.49 – 9.26)
inflation	12.38 ± 19.58 7.92 (5.35 – 11.90)
oop_share	30.0 ± 18.5 27.7 (13.6 – 40.5)
gov_share	52.1 ± 22.6 54.5 (36.3 – 71.4)
income_level	
Low income	26 (13.7)
Lower middle income	54 (28.4)
Upper middle income	52 (27.4)
High income	58 (30.5)
region	
Sub-Saharan Africa	48 (25.3)
Middle East, North Africa, Afghanistan and Pakistan	21 (11.1)
South Asia	8 (4.2)
Europe and Central Asia	50 (26.2)
East Asia and Pacific	29 (15.3)
Latin America and Caribbean	32 (16.8)
North America	2 (1.1)

SD: Standard Deviation

The completed datasets were randomly split into an 80% training set and a 20% test set. Hyperparameters were optimized for all models using a grid search with 5-fold cross-validation, and the optimal parameter combination was selected based on the lowest RMSE for each imputation. The performance metrics obtained were calculated over 20 imputations and presented as the mean ± standard deviation in Table 4 and Figure 2. According to the findings, SVR stood out in terms of prediction power by achieving the lowest RMSE value (RMSE= 463 ± 13.3); while also being the most successful method with a low error level (MAE= 282 ± 5.66) and the highest determination coefficient (R^2= 0.940 ± 0.003). The XGBoost model showed the best performance in terms of average absolute error (MAE= 262 ± 15.5) and also attracted attention with low error levels (RMSE= 520 ± 25.3) and a high coefficient of determination (R^2= 0.923 ± 0.007). The RF method also showed a similarly strong performance (RMSE= 577 ± 17.4, MAE= 320 ± 6.82, R^2= 0.905 ± 0.007) and provided stable results among tree-based algorithms. In contrast, regularized regression methods showed relatively lower performance. Ridge regression (RMSE= 768 ± 9.81, MAE= 484 ± 5.75, R^2= 0.849 ± 0.003) showed the lowest level of success, while Lasso (RMSE= 718 ± 13.1, MAE= 450 ± 7.36, R^2= 0.868 ± 0.004)

and Elastic Net (RMSE= 721 ± 13.1, MAE= 453 ± 7.24, R^2= 0.867 ± 0.004) methods produced relatively better results, but still lagged behind tree-based methods and SVR.

ANOVA results confirmed that the performance differences between models were statistically significant ($p < 0.001$). Post-hoc Tukey HSD tests showed that SVR achieved significantly lower RMSE and higher R^2 values than all other models ($p < 0.001$ in all pairwise comparisons), while XGBoost achieved the lowest MAE ($p < 0.001$).

Table 4. Performance Results of Regression Models

Model	RMSE	MAE	R^2
Ridge	768±9.81[a]	484±5.75[a]	0.849±0.003[a]
Lasso	718±13.1[b]	450±7.36[b]	0.868±0.004[b]
Elastic Net	721±13.1[b]	453±7.24[b]	0.867±0.004[b]
SVR	463±13.3[c]	282±5.66[c]	0.940±0.003[c]
Random Forest	577±17.4[d]	320±6.82[d]	0.905±0.007[d]
XGBoost	520±25.3[e]	262±15.5[e]	0.923±0.007[e]

Similar superscript letters within the same column indicate no statistically significant difference, whereas different letters denote significant differences (Tukey HSD test, $p < 0.001$).

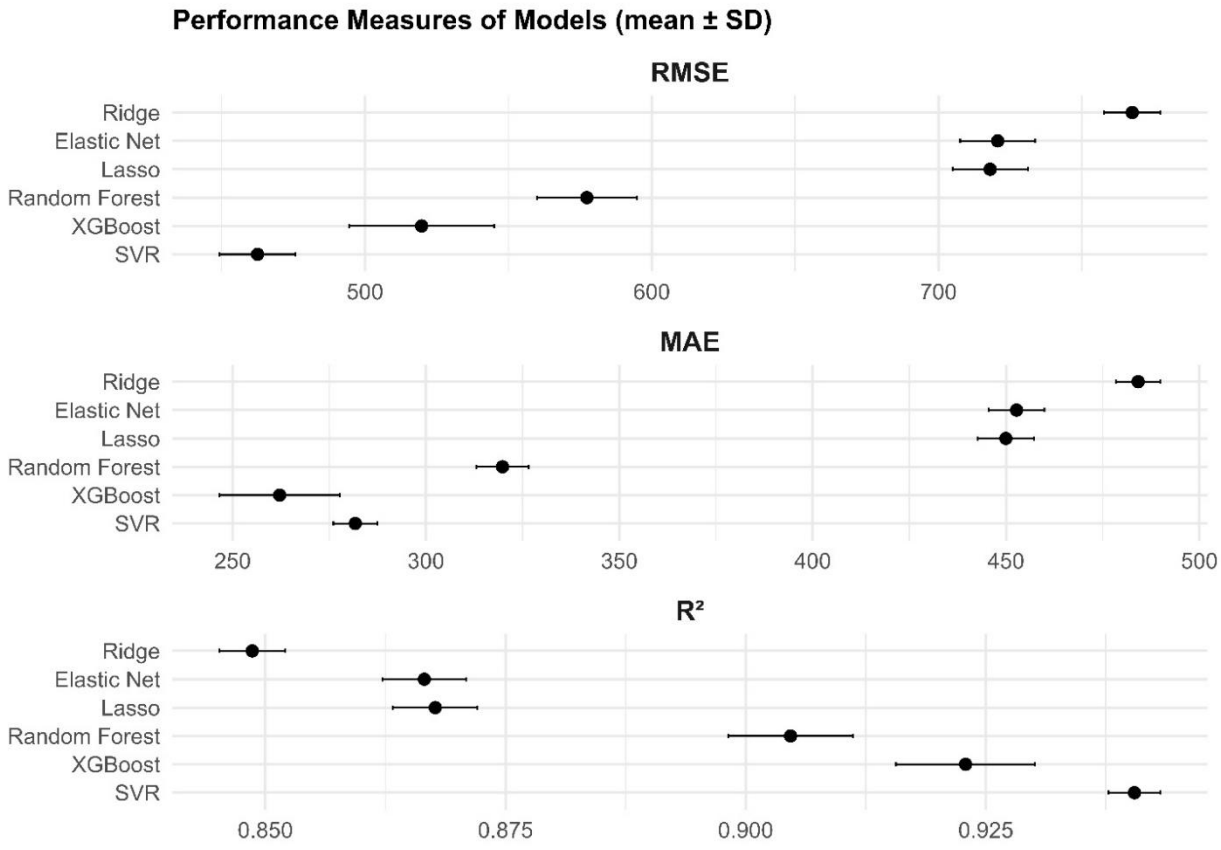

Figure 2. Performance Measures of Regression Models

Variable importance rankings in regression models are presented in Figure 3. In all methods, gross domestic product per capita (gdp_pc) emerged as the strongest predictor. This variable

was followed by the proportion of the population aged 65 and over (age65plus), life expectancy at birth (life_exp), and the proportion of the population living in urban areas (urban_pop). Tree-based methods (RF and XGBoost) also attributed significant importance to the income level variable, whereas the effect of this variable was limited in Ridge, Lasso, and Elastic Net models. Contributions from variables representing the financing structure of healthcare expenditures (oop_share and gov_share) and macroeconomic indicators (inflation and unemployment) were low across all models.

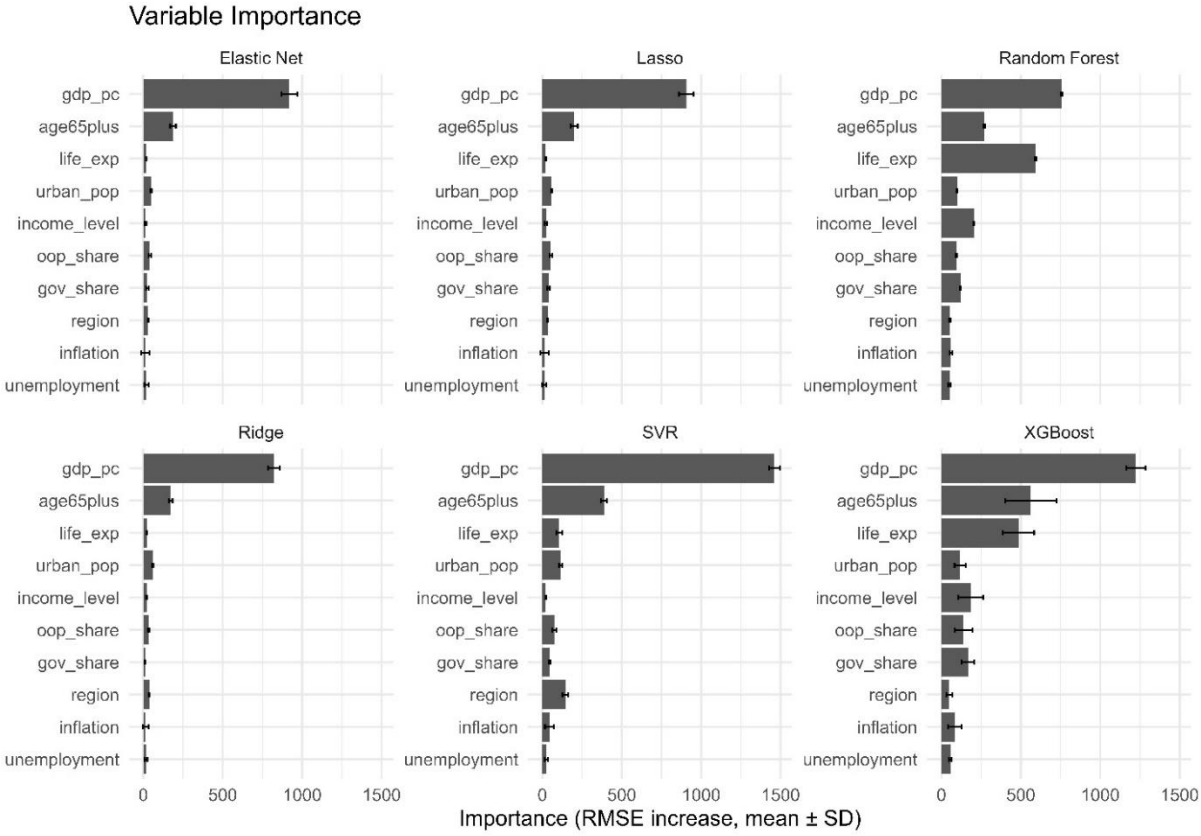

Figure 3. Variable Importance

4. DISCUSSION

This study aimed to estimate health expenditures for 190 countries in 2022 using World Bank data through machine learning methods and regularized regression methods. The findings showed that SVR demonstrated the most successful performance with the lowest RMSE (463 ± 13.3) and highest R^2 (0.940 ± 0.003) values. The XGBoost model stood out by achieving the lowest MAE (262 ± 15.5) value and also demonstrated strong performance with a high explanatory coefficient (R^2= 0.923 ± 0.007). The RF method also yielded similarly stable results (R^2= 0.905 ± 0.007). In contrast, regularized regression methods (Ridge, Lasso, Elastic Net) showed relatively lower performance. This difference may stem from the presence of nonlinear relationships and interactions between socio-economic and demographic variables affecting healthcare expenditure; While SVR and tree-based methods have the capacity to capture such complex structures, regularized linear regressions are limited by linear assumptions (Breiman, 2001; Vapnik, 1995; Friedman, 2001).

Methodologically, the superior performance of the SVR model stems from its ability to capture nonlinear relationships between variables, thanks to its kernel-based flexible structure. Additionally, the regularization term in SVR reduces the risk of overfitting while maintaining the bias–variance balance, thereby increasing generalizability. These features enable SVR to produce more stable and reliable predictions in multidimensional and heterogeneous data structures. Tree-based ensemble methods (XGBoost and RF) can also effectively model nonlinear relationships, but they can be more sensitive to hyperparameter tuning and data variation compared to SVR.

These findings are largely consistent with previous studies in the literature. A study covering the period 1990–2019 in the Turkish sample reported that SVR outperformed GPR and decision trees (Güleryüz, 2021). Similarly, SVR provided the highest predictive performance in this study. XGBoost's low MAE value and overall success are consistent with strong findings in the machine learning literature regarding the predictive performance of gradient boosting algorithms (Chen & Guestrin, 2016). Furthermore, in a study comparing Lasso, RF, and SVR on 2013 World Bank data, RF was reported to be relatively superior (Çınaroğlu, 2017). Although RF lagged behind SVR and XGBoost in the current study, it demonstrated a noteworthy performance with a high R^2 value (0.905). From this perspective, the findings reveal that nonlinear, kernel-based, and community learning approaches are more successful than classical linear regression methods in modeling complex socioeconomic processes such as healthcare expenditures.

The higher performance of SVR, XGBoost, and RF compared to regularized regression methods can be attributed to the nonlinear relationships and interactions between variables among the socioeconomic and demographic factors that determine healthcare expenditure. While tree-based algorithms and kernel-based SVR have the capacity to capture such complex patterns, linear approaches such as Ridge, Lasso, and Elastic Net are inherently limited due to their assumptions. This situation supports the view that machine learning methods are more advantageous than classical linear regression in multidimensional and interaction-sensitive fields such as health economics (Breiman, 2001; Vapnik, 1995).

The findings of the study indicate that the strongest determinant of health expenditures is per capita income level, with the proportion of elderly population and life expectancy also making significant contributions. These results are consistent with findings in the international literature pointing to the central role of income level in explaining health expenditures and the impact of demographic changes (particularly aging) on spending pressures (Baltagi & Moscone, 2010; OECD, 2023a). Regarding the financing mix, increasing the share of public spending and reducing dependence on out-of-pocket payments is considered critical for access and financial protection; global monitoring reports emphasize that high OOP burdens can impoverish households and hinder progress toward UHC goals (World Bank & WHO, 2023; WHO, 2023a).

However, this study has some limitations. Since the analysis is based on cross-sectional data, the relationships between variables should not be interpreted as causal. Furthermore, differences in data collection methods, reporting accuracy, and definitions of health expenditures across countries may lead to measurement errors or unobserved biases. Some structural determinants not included in the model (e.g., health system efficiency, service

accessibility, or policy variables) may limit the explanatory power of the model. Finally, the direct generalizability of this globally developed model to a single country context is limited, as countries exhibit significant differences in socioeconomic structure and health financing systems. Future studies utilizing longitudinal (panel) data and developing country-specific models will strengthen the validity of the results and policy implications.

5. CONCLUSION

In this study, machine learning and regularized regression methods were used to estimate health expenditures per capita using data from 190 countries for the year 2022. The findings revealed that SVR was the most successful method, achieving the lowest RMSE and highest R^2 values, while XGBoost obtained the lowest MAE value and RF demonstrated stable performance. In contrast, regularized linear regression methods such as Ridge, Lasso, and Elastic Net showed relatively lower success. Variable importance analyses showed that per capita GDP was the strongest predictor, while the elderly population ratio, life expectancy, and urbanization rate provided additional contributions.

The findings indicate that machine learning algorithms, particularly SVR and XGBoost, can successfully capture the nonlinear and interactive structure of macro-level socioeconomic determinants of health expenditures. These models can provide valuable analytical tools for policymakers and international organizations (e.g., WHO, OECD, World Bank) to forecast health expenditure trends, identify key factors influencing expenditures, and optimize budget allocation strategies. Integrating such predictive models into health budget planning processes can contribute to enhancing fiscal sustainability and strengthening comparability between countries.

From the perspective of policymakers, it should be noted that increases in income levels directly affect health expenditures and that income elasticity in health expenditures is significant (Acemoglu et al., 2013; Gerdtham & Jönsson, 2000). In this context, health financing strategies parallel to income growth should be developed in budget planning to maintain financial sustainability. Furthermore, long-term infrastructure and care models should be created to meet the healthcare demands of an aging population; investments should be made to increase healthcare infrastructure in response to the increased burden of healthcare services caused by urbanization. On the other hand, the high share of out-of-pocket (OOP) health expenditures can expose households to financial risk. Therefore, increasing the share of public spending, reducing OOP expenditures, and strengthening pooling mechanisms in health systems are critical for financial protection (WHO, 2023b; OECD, 2023b).

In future studies, the inclusion of time series analysis, a broader examination of structural indicators specific to healthcare systems, and the evaluation of alternative modeling strategies such as deep learning or Bayesian methods will contribute to a more comprehensive and reliable estimation of healthcare expenditures.

DECLARATION OF THE AUTHORS

Approval of ethical committee: All procedures performed in studies comply with the ethical standards of comparable institutional and/or national research committees.

Declaration of Contribution Rate: The authors have equal contributions.

Declaration of Support and Thanksgiving: No support is taken from any institution or organization.

Declaration of Conflicts of Interest: The authors declare no conflict of interest.

REFERENCES

Acemoglu, D., Finkelstein, A., & Notowidigdo, M. J. (2013). Income and health spending: Evidence from oil price shocks. *Review of Economics and Statistics*, *95*(4), 1079-1095.

Azur, M. J., Stuart, E. A., Frangakis, C., & Leaf, P. J. (2011). Multiple imputation by chained equations: What is it and how does it work?. *International Journal of Methods in Psychiatric Research*, *20*(1), 40-49.

Baltagi, B. H., & Moscone, F. (2010). Health care expenditure and income in the OECD reconsidered: Evidence from panel data. *Economic Modelling*, *27*(4), 804–811.

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

Chen, T., & Guestrin, C. (2016). XGBoost: A scalable tree boosting system. *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 785–794. https://doi.org/10.1145/2939672.2939785

Çınaroğlu, S. (2017). Sağlık harcamasının tahmininde makine öğrenmesi regresyon yöntemlerinin karşılaştırılması. *Uludağ Üniversitesi Mühendislik Fakültesi Dergisi*, *22*(2), 179–197. https://doi.org/10.17482/uumfd.338805

Friedman, J. H. (2001). Greedy function approximation: A gradient boosting machine. *Annals of Statistics*, *29*(5), 1189–1232. https://doi.org/10.1214/aos/1013203451

Gerdtham, U. G., & Jönsson, B. (2000). International comparisons of health expenditure: Theory, data and econometric analysis. *In Handbook of health economics*, (1), 11-53.

Güleryüz, D. (2021). Predicting health spending in Turkey using the GPR, SVR, and DT models. *Acta Infologica*, *5*(1), 155–166. https://doi.org/10.26650/acin.885940

Hoerl, A. E., & Kennard, R. W. (1970). Ridge regression: Biased estimation for nonorthogonal problems. *Technometrics, 12*(1), 55–67. https://doi.org/10.1080/00401706.1970.10488634

Martin, A. B., Hartman, M., Benson, J., Catlin, A., & National Health Expenditure Accounts Team. (2011). National health spending in 2011: Overall growth remains low, but some payers and services show signs of acceleration. *Health Affairs*, *32*(1), 87–99. https://doi.org/10.1377/hlthaff.2012.1206

Mihaylova, B., Briggs, A., O'Hagan, A., & Thompson, S. G. (2011). Review of statistical methods for analysing healthcare resources and costs. *Health Economics*, *20*(8), 897–916. https://doi.org/10.1002/hec.1653

OECD. (2023a). *Health at a Glance 2023: OECD Indicators*, OECD Publishing, https://doi.org/10.1787/7a7afb35-en

OECD. (2023b). *Health spending (indicator)*. OECD Data. https://doi.org/10.1787/8643de7e-en

Rubin, D. B. (1987). *Multiple imputation for nonresponse in surveys*. John Wiley & Sons. https://doi.org/10.1002/9780470316696

Sinha, R., Khandelwal, S., & Deshmukh, P. R. (2016). Determinants of out-of-pocket health expenditure: A systematic review. *Journal of Health Management*, *18*(2), 213–242. https://doi.org/10.1177/0972063416637700

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society: Series B (Methodological)*, *58*(1), 267–288. https://doi.org/10.1111/j.2517-6161.1996.tb02080.x

Vapnik, V. N. (1995). *The nature of statistical learning theory*. Springer. https://doi.org/10.1007/978-1-4757-2440-0

World Bank. (2023). *World Development Indicators: Health expenditure per capita (current US$).* The World Bank. https://data.worldbank.org/indicator/SH.XPD.CHEX.PC.CD

World Bank, & World Health Organization. (2023). *Tracking universal health coverage: 2023 global monitoring report.* World Bank /World Health Organization. https://openknowledge.worldbank.org/entities/publication/1ced1b12-896e-49f1-ab6f-f1a95325f39b

World Health Organization. (2023a). *Global spending on health: Report summary.* World Health Organization. https://iris.who.int/bitstream/handle/10665/379750/9789240104495-eng.pdf?isAllowed=y&sequence=1

World Health Organization. (2023b). *Financial protection.* World Health Organization. https://www.who.int/data/gho/indicator-metadata-registry/imr-details/4950

Zou, H., & Hastie, T. (2005). Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, *67*(2), 301–320. https://doi.org/10.1111/j.1467-9868.2005.00503.x