Predicting Health Spending in Turkey Using the GPR, SVR, and DT Models

Türkiye Sağlık Harcamalarının GPR, SVR ve DT Modelleri ile Tahmini

Didem Güleryüz[1]

ABSTRACT

Rising healthcare costs for countries and the long-term maintainability of this situation are at the center of the political agenda. The steady increase in health spending puts pressure on government budgets, healthcare, and personal patient financing. Policymakers would like to plan reforms to reduce these costs to adapt to problems that may arise. This has led planners to decision support systems and forecasting models. In this paper, three machine learnings algoritms, namely Support Vector Regression (SVR), Decision Tree Regression (DT), and Gaussian Process Regression (GPR) are employed to design a forecasting model for Health Spendings (HS) of Turkey considering various determinants. Gross domestic product per capita, urban population rate, unemployment rate, population ages 65 and above, the life expectancy, the physicians' rate, and the total number of hospital beds are used as inputs. The data set consists of 30 years between 1990-2019, which splits as training and test sets. Developed models were compared considering performance metrics, and the most accurate model was identified. The coefficient of determinations (R^2) for SVR, GPR, and DT models are 0.9929, 0.9989, and 0.9611 in the training phase, 0.9536, 0.8944, and 0.1166 in the testing stage, respectively. Therefore, the SVR model has accurate prediction results with the highest R^2 and the least root mean square error values in the testing phase. The study showed that the proposed SVR model reduced RMSE value by 32.02% and 39.66% compared to the GPR and DT models, respectively. Consequently, the Health Spendings of Turkey can be predicted by employing SVR with high accuracy.

Keywords: Health Spending, Gaussian Process Regression, Decision Tree, Support Vector Regression

ÖZ

Ülkeler için artan sağlık maliyetleri ve bu durumun uzun vadeli sürdürülebilirliği siyasi gündemin merkezinde yer almaktadır. Sağlık harcamalarındaki sürekli artış, hükümet bütçeleri, sağlık hizmetleri ve kişisel hasta finansmanı üzerinde baskı oluşturmaktadır. Politika yapıcılar, ortaya çıkabilecek sorunlara uyum sağlamak ve bu maliyetleri düşürmek için reformlar planlamak isterler. Bu durum, planlayıcıları karar destek sistemlerine ve tahmin modellerine yönlendirmiştir. Bu çalışmada, Türkiye'nin Sağlık Harcaması (HS) için çeşitli belirleyicileri dikkate alan bir tahmin modeli tasarlamak amacıyla Destek Vektör Regresyonu (SVR), Regresyon Ağacı (DT) ve Gauss Süreç Regresyonu (GPR) olmak üzere üç makine öğrenme algoritması kullanılmıştır. Kişi başına gayri safi yurtiçi hasıla, kentsel nüfus oranı, işsizlik oranı, 65 yaş ve üstü nüfus, ortalama yaşam süresi, hekim oranı ve toplam hastane yatak sayısı girdi değişkenleri olarak belirlenmiştir. Veri seti eğitim ve test verisi olarak ayrılmış ve 1990-2019 yılları arası 30 yılı kapsamaktadır. Geliştirilen modeller performans ölçütleri dikkate alınarak karşılaştırılmış ve en iyi model belirlenmiştir. SVR, GPR ve DT modelleri için belirleme katsayısı (R^2) eğitim aşamasında sırasıyla 0.9929, 0.9989 ve 0.9611, test aşamasında sırasıyla 0.9536, 0.8944 ve 0.1166'dır. Ayrıca, SVR modeli, test aşamasında en yüksek R^2 ve en düşük kök ortalama kare hatası değerleri ile en iyi tahmin sonuçlarına sahiptir. Çalışma, önerilen SVR modelinin RMSE değerini diğer GPR ve DT modellerine kıyasla sırasıyla % 32.02 ve % 39.66 azalttığını göstermiştir. Sonuç olarak, Türkiye'nin sağlık harcamaları SVR modeli kullanılarak yüksek doğrulukta tahmin edilebilir.

Anahtar kelimeler: Sağlık Harcaması, Gauss Süreç Regresyonu, Regresyon Ağacı, Destek Vektör Regresyonu

[1](PhD.), Bayburt University, Faculty of Applied Sciences, Department of Management Information Systems Bayburt, Turkey

ORCID: D.G. 0000-0003-4198-9997

Corresponding author:
Didem GÜLERYÜZ
Bayburt University, Faculty of Applied Sciences, Department of Management Information Systems Bayburt, Turkey
E-mail address: dguleryuz@bayburt.edu.tr

1. INTRODUCTION

People are living longer than ever before thanks to developing health technologies and improved living standards. According to Turkstat statistics, Turkey's life expectancy has increased in recent years (Turkstat, 2020). Therefore, healthcare expectations have also increased due to the developments of technological advances. As a corollary result of the increasing expectations for healthcare services, the expenditures of governments in this area will also increase. Determining the reasons for the increase in expenditures is very important for the planning of health systems.

The existing service system must be operated optimally for healthcare services to provide the proper services to the right person at the right time, the problems in the system must be identified, and solutions must be specific to these problems. With developing technology, many methods and technologies can be used in health systems. However, it is not possible to provide all of the technologies. For this reason, the concept of health economics is drawing attention today. Optimizing available resources under constraints on variables such as productivity, patient safety, and clinical efficacy will be an efficient tool for decision-makers.

Health and health services are in a reciprocal relationship with the economy. However, studies defining this relationship are divided into two classes. The first is to investigate the links between health status and income distribution, and the second is to identify the interrelation between health financing policies and economic outcomes (Martín et al., 2011). Health indicators play an important role in determining the development of countries. Some previous studies, such as Bauer et al. (2006) and Dalbokova and Krzyzanowski (2002), claim a positive correlation between countries' health levels and their development. Also, people's welfare, labor productivity, demographic and human capital factors directly impact health spending.

The number of studies that investigate the determinants of healthcare spending and estimate the expenditures according to related determinants is increasing day by day. In the light of previous studies, the determinants affecting health expenditures can be classified as environmental, social, and economic. Gross domestic product, unemployment rate, physicians, urbanization rate, and gas emissions were used frequently as determinants of health expenditure in previous studies. Also, most studies investigated the effects of environmental quality on health expenditures (Martín et al., 2011).

Health spending in Turkey, primarily financed by the government, should be optimized to avoid the extra cost of public spending and it is expected to increase in the coming years. This situation depends on ensuring financial efficiency in health services. Excessive healthcare financed by the Turkish government and the advancement of technology in health services negatively affect health expenditures, threatening the system's financial sustainability. The optimization of the system is possible with properly selected management strategies. For instance, the application of lean service tools in the health sector will increase the efficiency of the system and decrease expenditures (Efe and Efe, 2016). Therefore, decision support systems designed to manage health expenditures should predict health expenditure, determine the indicators of health expenditures and optimize the resources (Ministry of Development, 2014).

As healthcare needs increase, health expenditures are expected to increase to meet these demands. In most industrialized economies, healthcare spending is proliferating more than the growth of income. Figure 1 shows the average health spendings for OECD countries and Turkey's health spendings during the last fifteen years (OECD, 2021). According to this graph, health expenditures increased between 2003 and 2019. Unfortunately, Turkey's per capita health spending level remains below the OECD average. According to OECD data, there was a 15.9% increase in health spendings from 2012 to 2018. Although it is not valid for all OECD countries, some OECD countries with higher income and development levels spend more on health services. Identifying the factors that lead to increased health spending is essential for policymakers and scientists to control and plan the spendings.
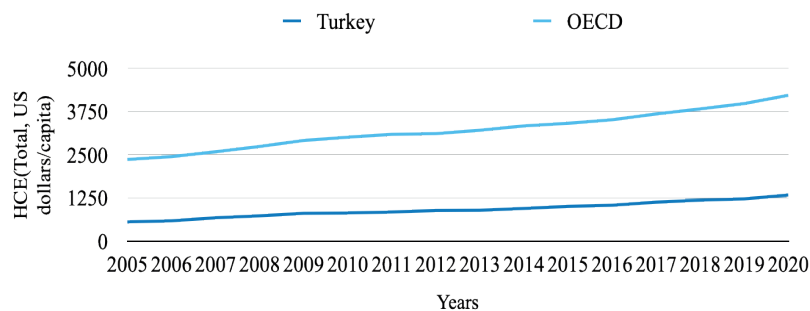
Figure 1. The annual health spendings of Turkey and OECD countries (OECD, 2021)

Improving the health status of citizens will have the potential to increase government spending on healthcare. As seen in Figure 2, the trend in rate of health spendings in GDP has a risen and declined over twenty-nine years in Turkey.
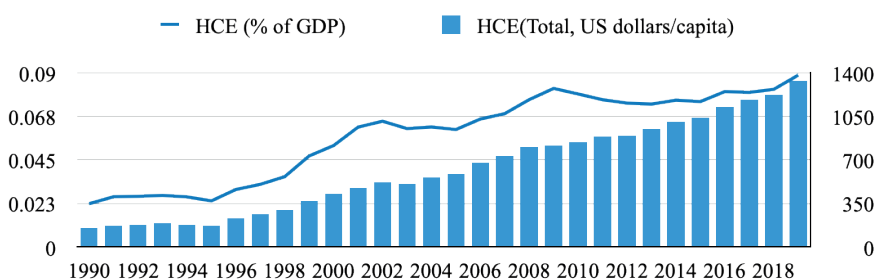

Figure 2. The change of HCE (% GDP) and HCE (Total, US dollars/capita) in Turkey (OECD, 2021)

Moreover, the government's ability to manage health spendings depends on determining the main factors influencing health spendings and selecting a suitable prediction model. With the development of machine learning methods, nonlinear relationships between inputs and outputs can be modeled quickly, and variables that do not explain the target can be included in the model. In previous studies, these relationships were mostly modeled using statistical models. Also, statistical methods widely used in the literature require large data sets. It can show low predictive performance in small-volume data sets. Machine learning models can overcome this problem and have better predictive results with smaller data sets. For this reason, Machine learning models have been widely used recently.

Various methods were employed to predict health spendings in previous studies. These methods can be classified as statistical methods and AI-based methods (Maksimović et al., 2017; Cinaroglu, 2017; Özcan and Tüysüz, 2018). It is impossible to examine all previous studies. For this reason, some selected studies to reflect the literature have been reviewed.

Maksimović et al. (2017) used soft computing methods to estimate gross domestic product (GDP) using health spending determinants. According to the obtained results, the SVR model has better estimation ability than the other employed methods (Maksimović et al., 2017). Mladenović et al. (2016) used a hybrid artificial neural network and fuzzy system (ANFIS) to examine health spending's influence on economic growth. It has been shown that the GDP growth rate is the most useful feature of health expenditures (Mladenović et al., 2016). Cinaroglu (2018) compared ML-based prediction models, namely Lasso Regression, Random Forest, and Support Vector Regression for Turkey. The total health expenditure per capita was selected as the dependent variable. The independent variables are: income, geographical region, total population over 65 years of age, life expectancy at birth, and total population. The different hyperparameters for each model were used in determined intervals, and the results were compared. As a result, the Random Forest model had better estimation results according to different hyperparameter values. Özcan and Tüysüz (2018) used a Grey estimation model to estimate Turkey's per capita healthcare spendings. Since the grey prediction model's performance depends on the model's parameters, these parameters were optimized using the genetic algorithm. Three different mechanisms have been developed to test the predictive ability of the models. As a result, the Grey model's prediction performance has increased with the genetic algorithm (Özcan

and Tüysüz, 2018). Dritsakis and Klazoglou (2019) modeled the total health expenditure of the USA using the Box-Jenkins method. The developed ARIMA (2,1,0) model is better than other employed models considering obtained performance criteria (Dritsakis and Klazoglou, 2019). Ram et al. (2019) employed ARIMA to forecast health expenditure in Iran. The ARIMA (1,1,1) model had the most accurate prediction results (Ram et al., 2019). Samah et al. (2020) aimed to observe health spendings and CO_2 emissions during the COVID-19 outbreak in Malaysia. The dynamic panel data system GMM prediction model was used. It was shown that health spendings and the growth in health services, inflation rate, and unemployment rate significantly impact CO_2 emissions (Samah et al., 2020). Shi et al. (2020) investigated the effects of urbanization on healthcare spendings and behavioral changes. As a result of the study, it was found that increasing urbanization affects the total treatment costs. Moreover, it has been observed that health spendings are affected by the urbanization rate and health insurance level (Shi et al., 2020). Ceylan (2020) estimated Turkey's health expenditures using greenhouse gas emission levels as input variables via optimized support vector regression, feedforward backpropagation neural network, and multivariate linear regression. The SVR model with Bayesian optimization has more effective health expenditure estimation results (Ceylan, 2020). Ceylan and Atalan (2020) used artificial intelligence techniques to estimate health expenditures per capita in Turkey. In the study, optimum determinants from the variables were selected using the genetic algorithm method. Random Forest model with genetic algorithm feature selection had the superior performance among all used models (Ceylan and Atalan, 2020). Cinaroglu (2020) presented a study comparing the predictive ability of logistic regression and random forest models for households that will face catastrophic out-of-pocket (OOP) health expenditure. The dataset was obtained from the household budget survey compiled by TURKSTAT for the year 2012. As a result, the developed random forest model had superior performance in showing households facing catastrophic OOP health expenditure. Durmaz and Pabuçcu (2020) used the ANFIS method to investigate the relevance between public health spendings and labor productivity. G20 countries data were used to analyze relations. The health spendings rate, trade gap, and GDP growth rate were used as input variables, and labor productivity as an output variable. It was determined that the relationship between health spendings and labor force productivity has different effects periodically, and there are linear and opposite relationships (Durmaz and Pabuçcu, 2020). Jia et al. (2021) developed a forecasting model using the gray system theory called the New Structure of the Multivariate Grey Prediction Model. The main driving factors affecting China's health sector were used as input variables, and health expenditures were estimated. The input variables in the model were determined considering the economy, population, health service utilization, and policy. The suggested NSGM (1, N) model provided good predictive results.

In the literature review, few studies used the SVR and DT methods for health spendings prediction. Ceylan (2020) and Maksimović et al. (2017) employed an SVR model to estimate health spendings. Akca et al. (2017) developed a decision tree model to identify features that have measurable effects on HS for OECD countries. In this study, GDP per capita (GDP), population ages 65 and above (PO65), urban population (UP), unemployment rate (UR), life expectancy (LE), the physicians' rate (PY), the total number of hospital beds (HB) were used as input variables to predict health spendings per capita (HS) in Turkey. Some or all of these variables have been used as inputs in the HS estimation literature (Häkkinen et al., 2008; Jakovljevic et al., 2020; Mladenović et al., 2016).

This study aims to contribute to the literature by developing forecasting models using machine learning methods. In this study, SVR, GPR, and DT models were used to estimate health spendings. According to the authors' knowledge, GPR has not been used previously to model health spendings. The SVR and GPR models' results were very close, so that GPR might also be used to model health spendings. Furthermore, the SVR and GPR models provided a stable estimation in a short time. The paper includes four sections. The materials and methods, evaluation of the model accuracy, and developed models' structure are placed in Section 2. Section 3 and Section 4 contain the results of the developed models and the conclusion, respectively.

2. MATERIAL AND METHOD

2.1. Data Collection Process

Turkey's health spending per capita was estimated using GDP per capita, population ages 65 and above, life expectancy, unemployment rate, the number of physicians, the total number of hospital beds, and urban population data in this case. The

data set was gained from several data sources covering the period 1990-2019. Table *1* explains the definitions, abbreviations (Abbr.), and units of the variables and shows the sources.

Table 1

Definition of the Variables

Variable	Abbr.	Unit	Source
Health spendings per capita	HS	Total, US dollars/capita	(OECD, 2021)
GDP per capita	GDP	Constant 2010 billion US$	(Worldbank, 2021)
Population ages 65 +	PO65	% of the total population	(Worldbank, 2021)
Urban population	UP	% of the total population	(Worldbank, 2021)
Unemployment rate	UR	% of total labor force	(Worldbank, 2021)
Life expectancy	LE	years	(Worldbank, 2021)
Number of Physicians	PY	Per 1,000 people	(OECD, 2021)
Number of Hospital beds	HB	Unit	(Turkstat, 2020)

GDP, PO65, UP, UR, LE, PY, HB are inputs that affect the HS. Table *2* shows the descriptive statistics of inputs and output.

Table 2

The statistics of the data set

	GDP	PO65	UP	UR	LE	PY	HB	HS
Mean	9991.31	6.55	67.54	9.26	71.71	1.44	171828	627.85
Std	2710.69	1.16	4.96	1.70	4.11	0.31	34337	360.03
Minimum	6709.09	4.61	59.20	6.49	64.26	0.9	120738	151.80
Maximum	15068.98	8.73	75.63	13.67	77.54	1.88	237504	1337.17
Kurtosis	-0.97610	-0.96	-1.27	0.12	-1.17	-1.35	-1.12	-1.15
Skewness	0.60	-0.02	0.001	0.48	-0.28	-0.18	0.26	0.27

The input indicators were determined by examining the studies on health expenditures prediction in the literature. The change in the HS is also related to GDP. Many studies investigate the relationship between HS and GDP. Jakovljevic et al. (2020) and Rana et al. (2020) examined the relevance between HS and GDP. Since the elderly population is increasing, scientific studies on the elderly population have increased. It is inferred that the elderly population will rise in the upcoming period. There are many studies investigating the relationship between HS and PO65. Di Matteo (2005), Häkkinen et al. (2008) investigated the effects of the elderly population on health spendings.

One of the crucial effects of urbanization is that healthcare services' structure has changed from rural to urban. This situation necessitated the transition from local institutions, where preventive health services are provided in rural areas, to hybrid institutions, where more complex health services are provided in cities. Therefore, health expenditures are affected by the rate of urbanization. Aliyu and Amadu (2017), Biadgilign et al. (2019) have investigated HS and UR relations. Improvements in economic factors generally have positive effects on health services. Increasing household incomes facilitates the financing of healthcare needs. Also, the decrease in family unemployment prevents the loss of employment-based health insurance and provides an opportunity to meet healthcare needs. Several studies were conducted to analyze this relationship, particular examples being Jakovljevic et al. (2020) and Papanicolas et al. (2019). Turkey's life expectancy has increased from year to year, as seen in Figure 3. Therefore, in parallel with this life expectation, the quality expectation in health services will affect the expenditures. Some studies have investigated this relationship, for example, Anderson and Frogner (2008), Crémieux et al. (1999). The number of hospital beds and physicians are the direct determinants of health spendings. Therefore, these determinants are used to predict health spendings. The annual change of the indicators related to the health spendings of Turkey is seen in Figure 3.
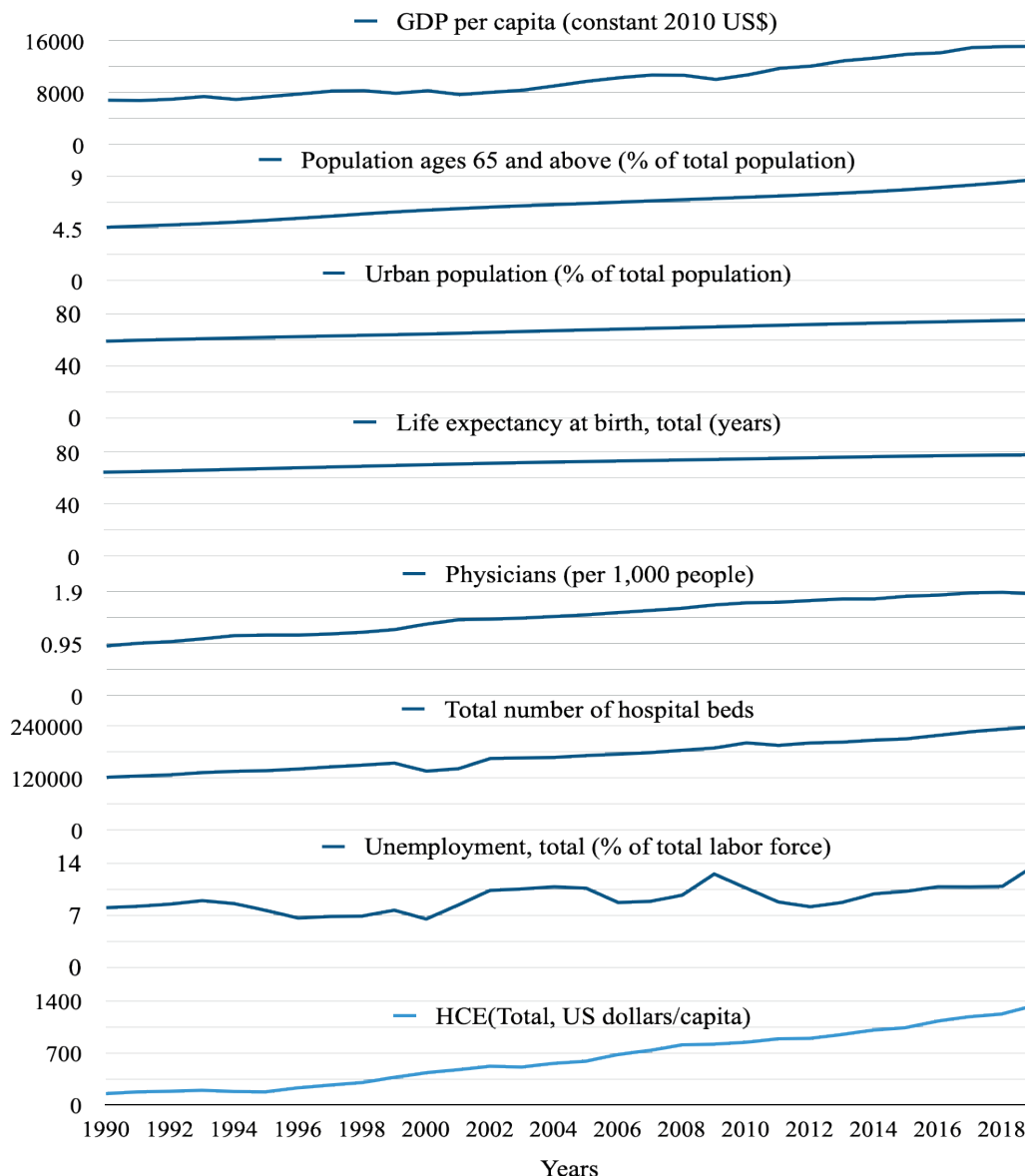
Figure 3. Change of inputs and output between 1990-2019

Preprocessing data for the machine learning algorithms is necessary and time-consuming. Normalizing data aims to transform variable values into a standard scale without disturbing the differences in value ranges. Normalization is only required when features have different ranges. In this case, since the variables' units are different, bringing the data into a standard scale will return more accurate results. Normalizing variables in the range of 0.05- 0.95 will give better results to make more meaningful comparisons in big data applications (Wang et al., 2015). The data set is divided into two, 70% used in the training stage and 30% in the test stage.

2.2. Support Vector Regression (SVR)

Support Vector Machine (SVM) is a kernel-based machine learning tool to address classification and regression problems. Although the support vector algorithm was initially employed for classification, it can be used in regression problems. The ability of this method to be applied for two purposes provides the solution to data science. Support vector regression (SVR) differs from SVM by having a single reference point. SVR does not scan to optimize the distance between two or more reference points to find the optimal hyperplane. The model can be used for linear and nonlinear problems. A linear or nonlinear range can be obtained when applying the SVR model according to the selected kernel function. It tries to find a regression function for the used data set. The SVR architecture is formulated as follows, as seen in Eq. (1) (Quan et al., 2020).

$$\text{maximize} \begin{cases} \frac{1}{2}\sum_{i=1}^{l}(\omega_i - \omega_i{}^*)(\omega_i - \omega_i{}^*)K\langle x_i, x_j\rangle \\ -\epsilon\sum_{i=1}^{l}(\omega_i + \omega_i{}^*) + \sum_{i=1}^{l} y_i(\omega_i - \omega_i{}^*) \end{cases} \tag{1}$$

$$\text{s.t} \begin{cases} \sum_{i=1}^{l}(\omega_i - \omega_i{}^*) = 0 \text{ and } \omega_i, \omega_i{}^* \in [0, C] \\ 0 \leq \omega_i, \omega_i{}^* \leq \frac{C}{l} \\ i = 1,2,\dots, l \end{cases}$$

In Eq. (1), ω_i and ωi^* are nonnegative multipliers for each observation. x_i is current data, l is data size, the penalty coefficient is depicted C, ϵ is the penalty dimension, and the kernel function is presented K (x_i, x_j). Adjust the $\omega = [\omega_1, \omega_1{}^*, \dots, \omega_1, \omega_j{}^*]^T$ to get the optimum solution. The regression equation is formulated in Eq. (2) (Quan et al., 2020).

$$f(x) = \sum_{i=1}^{l}(\omega_i - \omega_i^*)\, K(x_i - x_j) + b^* \tag{2}$$

2.3. Gaussian Process Regression (GPR)

Gaussian process regression (GPR) is a probabilistic model based on kernels. The inputs (x_i) and the outputs (y_i) are taken from an unknown distribution. A GPR model searches for a solution for predicting the output variable via the input vector. As seen in Eq. (x), the form is based on linear regression (Zhang et al., 2016).

$$y = x^T\beta + \varepsilon$$

Where $\varepsilon \sim N(0, \sigma^2)$ and the error variance and β are predicted using the training data, considering the Gaussian process, p(f) is zero; a matrix K depicts a kernel function.

$$p(\mathbf{f}) = N(\mathbf{f} \mid \mathbf{0}, K) \tag{3}$$

Let $K_{ij} = K(x_i, x_j)$. Eq. (4) shows the function of y.

$$p(y) = \int p(\mathbf{y} \mid \mathbf{f})p(\mathbf{f})d\mathbf{f} = N(\mathbf{f} \mid \mathbf{0}, K_y) \tag{4}$$

There is a hidden variable f (x_i) calculated for each xi in the GPR model. Let $\mathbf{o_*} = [o(\mathbf{x_*}, \mathbf{x_1}), \dots, o(\mathbf{x_*}, \mathbf{x_M})]^T$ and $o_{**} = k(x_*, x_*)o_{**} = k(x_*, x_*)$. Eq. (5) and Eq. (6) show the mean and variance of $P(y_* \mid y)$.

$$\mu(\mathbf{x_*}) = o_*^T K_y^{-1} \mathbf{y} \tag{5}$$

$$\sigma^2(\mathbf{x_*}) = o_{**} - o_*^T K_y^{-1} o_* + \sigma_n^2 \tag{6}$$

2.4. Decision Tree (DT)

The Decision Tree algorithm is frequently used in machine learning literature and the business environment to provide solutions to both classification and regression problems. The decision tree obtains the predictive value by asking a series of questions to the data until each model reaches a single predictive value. The model determines the order and content of the problem. During the training phase, the model is equipped with a historical data set of the relevant field. Thus, the decision tree model learns the relationship between input data and target value. After the training phase, the decision tree builds a tree and calculates the best questions, and nodes are sequenced to make the most accurate predictions. The developed model can predict by providing the same data format as in the training phase. In the regression model, Let X = X_1, X_2, …. ,X_n be independent variables. Y_1, Y_2, …, Y_n is defined as dependent variables, and n is the number of the observed values. The candidate split can represent $\Upsilon = (f, th_t)$, and node depicts t, feature, and threshold value present f and th, respectively. Eq. (7) presents the dividing data into Υ candidate, and the DT left side node is gained (Pekel et al., 2020).

$$Q_l(\varkappa) = (x,y)|x_f \leq th_t \tag{7}$$

Eq. (7) depicts the dividing data into $\varkappa$ candidate, and the right side of the DT is found.

$$Q_r(\varkappa) = (x,y)|x_f > th_t \tag{8}$$

Also, Eq. (8) can be written as $Q_r(\varkappa) = Q/Q_l(\varkappa)$. The predictive value is estimated based on the historical data.

2.5. Performance evaluation criteria

The model's accuracy was measured using certain metrics - root mean square error (RMSE), the coefficient of determination (R^2), and mean absolute error (MAE). The comparison of these metrics is necessary to make relevant interpretations. The performance criteria equations are given in Eq. (9) to Eq. (11), respectively (Guleryuz and Ozden, 2020).

$$MAE = \frac{1}{n}\sum_{t=1}^{n} |HS_i^{observed} - HS_i^{predicted}| \tag{9}$$

$$RMSE = \sqrt{\frac{1}{n}\sum_{t=1}^{n} (HS_i^{observed} - HS_i^{predicted})^2} \tag{10}$$

$$R^2 = \left(\frac{\sum_{i=1}^{n}(HS_i^{observed} - \overline{HS_i^{observed}})(HS_i^{predicted} - \overline{HS_i^{predicted}})}{\sqrt{\sum_{i=1}^{n}(HS_i^{observed} - \overline{HS_i^{observed}})^2 (HS_i^{predicted} - \overline{HS_i^{predicted}})^2}} \right)^2 \tag{11}$$

n represents the number of observations, $HS_i^{observed}$ is the observed value of HS and $HS_i^{predicted}$ is the estimated value of HS at the time i.

3. RESULTS

3.1. Results of SVR, GPR, and DT

The SVR model was established to define the relationship between seven inputs and the health spendings variable. The SVR model includes hyperparameters, parameters which affect the predictive performance of the model. The developer usually determines these values. There are studies on hyperparameter optimization in the literature. However, since three different models were compared in this study, hyperparameter optimization of each model was not targetted. The developed model was tested with different hyperparameters, and the parameters that led to the best prediction performance were selected. Kernel function, box constraint, kernel scale, and epsilon are hyperparameters of the SVR model. Table 3 includes the parameters of the SVR model.

The Gaussian process is a nonparametric Bayesian method that can model complex relationships by handling uncertainty. Like SVR, a dataset that includes seven inputs and one output is used to develop a regression model. In this case, Matern 3/2 was employed as a kernel function among many kernels since its predictive ability is better than others. Other parameters of the GPR model can be seen in Table 3.

The decision tree is an iterative algorithm that uses a tree structure that starts using a single node and splits branches to find the current value. The input values move in a specific way considering the nodes. In this case, the decision tree starts with a single node, then the fitness function is found, and the split is chosen with the best fitness function value. This step is repeated until it reaches the stopping criteria. Table 3 shows the parameters of the developed DT model. Also, 3-fold cross-validation was employed in the training phase to impede overfitting.

Linear, Polynomial, Gaussian, and Sigmoid kernel functions were used in the developed SVR models. SVR with Gaussian kernel function had the best accurate prediction result. Also, in the GPR model, various kernel functions such as Linear, Polynomial, Squared Exponential, Rational Quadratic, Matern 3/2, and Matern 5/2 were used. The model with the lowest prediction error among the developed GPR models was GPR with Matern 3/2 kernel function. The parameters of these models are given in Table 3.

Table 3

The parameters of the developed models

SVR		GPR		DT	
Kernel function	: Gaussian	**Kernel function**	: Matern3/2	**Minimum leaf size**	: 1
Box constraints (*C*)	: 750.8097	**Sigma**	: 0.010018	**Prune criterion**	: MSE
Epsilon, *ε*	: 0.00025521	**Kernel scale**	: 0.15866	**Minimum parent**	: 10
Kernel scale	: 29.917	**Prediction speed (obs/ sec)**	: 2700	**Number of nodes**	: 7
Prediction speed (obs/ sec)	: 2900	**Training time (sec)**	: 67.691	**Prediction speed (obs/ sec)**	: 2900
Training time (sec)	: 85.414			**Training time (sec)**	: 26.236

3.2. Accuracy evaluations of the developed models

In this paper, the HS was a target value, and seven indicators related to HS were inputs for all developed models. The regression functions that establish the relationship between inputs and output were obtained using the training data set. The developed SVR, GPR, and DT models were tested using the test dataset to measure the predictability. The estimation results' performances were compared according to three metrics: MAE, RMSE, and R^2.

The performance metrics values of all models are given in Table 4. When the dataset is small, there is much debate over whether the coefficient of determination is a valid performance measure for tree-based models. In this way, developed models are ranked according to their RMSE values. However, even if the models are ranked according to R^2 values, the developed SVR model has the best predictive performance during the test phase.

Table 4

Performance metrics of the developed models

	Models	MAE	R2	RMSE	Rank
Training Phase	SVR	0.0097	0.9929	0.0151	2
	GPR	0.0049	0.9989	0.0060	1
	DT	0.0292	0.9611	0.0350	3
Testing Phase	SVR	**0.1452**	**0.9536**	**0.1520**	**1**
	GPR	0.1802	0.8944	0.2236	2
	DT	0.2256	0.1166	0.2519	3

All models successfully estimated health spendings using determined indicators. The R^2 values of these three models changed between 0.9611 and 0.9929 during the training phase. This situation confirms that the developed models had good predictive performance. The predictive performance of the models depends on the relationships between inputs and outputs during the training phase. If this relationship is modeled correctly, the machine learning algorithm will give satisfactory results during the test phase.

For this reason, 30% of the data was reserved as the test data. Table 4 includes the prediction ability of the developed models in the testing phase. The R^2 value of SVR, GPR, and DT were 0.9929, 0.9989, and 0.9611 in the training phase, with 0.9536, 0.8944, and 0.1166 in the testing phase. Machine learning models establish the relationship between inputs and outputs in the training phase. Therefore, since this established relationship was evaluated during the test phase, the testing phase's performance metrics should be compared. Table 4 shows that the developed SVR model has stronger prediction ability than the DT and GPR models. The developed SVR model found the R^2 and RMSE values to be 0.9536 and 0.1520, respectively, in the testing phase. The computational results indicate that the SVR can use to predict HS, while meeting the least estimation error. Figure 4 shows the comparison of developed models based on MAE, RMSE, and R^2.
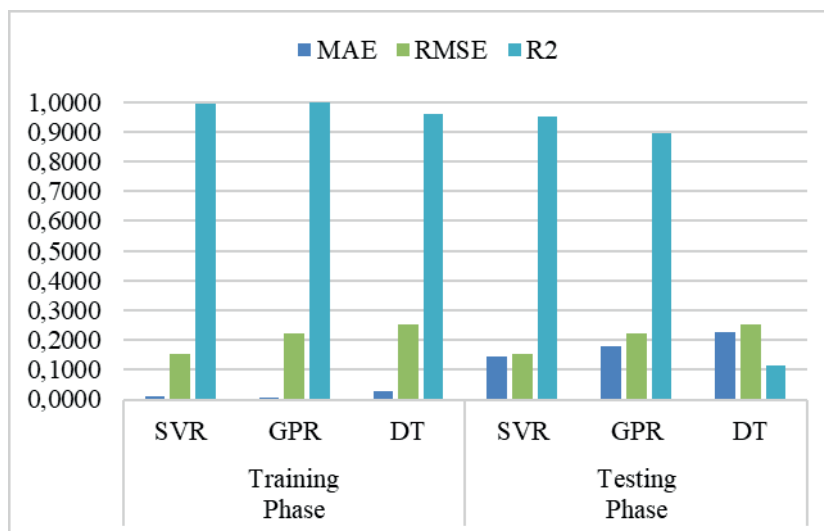
Figure 4. The comparison of developed models

4. DISCUSSION

Prosperous countries have a healthier population, mainly because there is a positive relationship between health services and economic power (Boyce and Brown, 2019). Since the problems related to countries' health systems differ according to the country, the main reasons that affect the health systems of many countries, such as the aging population and chronic diseases, should be monitored carefully. Thus, problems that may arise in the health system can be easily identified. Data science and machine learning techniques play an active role in solving these problems and can support decision science.

For countries that want to provide sustainable health services, it is essential to measure and compare the performance of health processes reliably. Thus, it is crucial to obtain data on time to carry out the measurement and comparison process effectively. This can help to provide high-quality services at optimum costs continuously and responsively (WHO, 2018). This whole process requires more comprehensive analysis with new data. This situation opens the door to the use of machine learning and deep learning models. In order to manage expenditures effectively in health systems, the indicators affecting the system should be determined. Especially if a prediction model is developed to support the decision-maker, the variables and the relationships between these variables affect the model considerably. Linear and nonlinear relationships between data can be modeled with machine learning methods. In addition, higher accuracy prediction results can be obtained by using less data than statistical methods with machine learning methods.

This study was conducted to model the effects of selected indicators, such as population ages 65 and above, life expectancy, unemployment rate, the number of physicians, the total number of hospital beds, and urban population on health spending via ML-based models. Three different ML methods were developed, and the developed SVR model outperformed GPR and DT. Various models in the literature estimate health expenditures using various combinations of selected indicators with different models. High accuracy was obtained via developed models in this study, and this result is in accordance with Jakovljevic et al. (2020), Cinaroglu (2018), Aliyu and Amadu (2017), Häkkinen et al. (2008), and Anderson and Frogner (2008).

5. CONCLUSIONS

The Health economy has become an essential field in recent years. In modern economies, health spendings occupy a central place due to social policy. Health problems are considerably different for different countries. Thus, would be complicated to tell countries how much they will spend on health services without knowing the problems faced by each of them. Also, the indicators affecting the health expenditures of the countries should be appropriately determined. However, many different administrative and logistic activities and sustainability indicators can change health spending efficiency. The availability,

advantages, and disadvantages of each continuous improvement method differ by country. Health spendings are related to certain social, economic, and environmental indicators. In this study, prediction models have been developed with machine learning methods by selecting some of the indicators which are related to health expenditures.

Since GPR has not been used previously to model health spendings, this study contributes to the literature by developing forecasting models. Machine learning-based prediction models were designed with determined indicators via Support Vector Regression, Gaussian Process Regression, and Decision Tree architectures. The developed models were tested with test data, and their prediction performances were compared. The developed SVR model's prediction ability was better than GPR and DT. The MAE, RMSE, and R^2 values of the SVR model were 0.1452, 0.1520, and 0.9536, respectively, in the testing phase. Policymakers can benefit from these results to make healthcare spending plans. The study showed that the proposed SVR model reduced the RMSE value by 32.02% and 39.66% compared to the GPR and DT models, respectively.

The choice of determinants as input is important for accurate estimation results. Although this is the main limitation of the employed methods, good predictive performance has been achieved with this study's selected inputs. This limitation could be eliminated in future studies by combining the employed models with feature selection algorithms such as the genetic algorithm, grey wolf optimization, and particle swarm optimization. Also, the hyperparameters of the models could be optimized by heuristic or mathematical optimization methods.

References/Kaynaklar

Akca, N., Sonmez, S., Yilmaz, A., 2017. Determinants of health expenditure in OECD countries: A decision tree model. Pakistan J. Med. Sci. 33, 1490–1494. https://doi.org/10.12669/pjms.336.13300

Aliyu, A.A., Amadu, L., 2017. Urbanization, cities, and health: The challenges to Nigeria - A review. Ann. Afr. Med. 16, 149–158. https://doi.org/10.4103/aam.aam_1_17

Anderson, G.F., Frogner, B.K., 2008. Health spending in OECD countries: Obtaining value per dollar. Health Aff. 27, 1718–1727. https://doi.org/10.1377/hlthaff.27.6.1718

Bauer, G., Davies, J.K., Pelikan, J., Consortium, on behalf of the E.T.W.G. and T.E., 2006. The EUHPID Health Development Model for the classification of public health indicators. Health Promot. Int. 21, 153–159. https://doi.org/10.1093/heapro/dak002

Biadgilign, S., Ayenew, H.Y., Shumetie, A., Chitekwe, S., Tolla, A., Haile, D., Gebreyesus, S.H., Deribew, A., Gebre, B., 2019. Good governance, public health expenditures, urbanization and child undernutrition Nexus in Ethiopia: An ecological analysis. BMC Health Serv. Res. 19, 1–10. https://doi.org/10.1186/s12913-018-3822-2

Boyce, T., Brown, C., 2019. Economic and social impacts and benefits of health systems. World Heal. Organ. 56.

Ceylan, Z., 2020. Investigation the insights between health expenditures and air quality. Int. J. Glob. Warm. 20, 203–215. https://doi.org/10.1504/IJGW.2020.106594

Ceylan, Z., Atalan, A., 2020. Estimation of healthcare expenditure per capita of Turkey using artificial intelligence techniques with genetic algorithm-based feature selection. J. Forecast. 279–290. https://doi.org/10.1002/for.2747

Cinaroglu, S., 2020. Modelling unbalanced catastrophic health expenditure data by using machine-learning methods. Intell. Syst. Accounting, Financ. Manag. 27, 168–181. https://doi.org/10.1002/isaf.1483

Cinaroglu, Songül, 2017. Sağlık Harcamasının Tahmininde Makine Öğrenmesi Regresyon Yöntemlerinin Karşılaştırılması. Uludağ Univ. J. Fac. Eng. 22, 179–200. https://doi.org/10.17482/uumfd.338805

Cinaroglu, Songul, 2017. International Journal of Economics and Financial Issues A Fresh Look at Out-of-Pocket Health Expenditures after More than a Decade Health Reform Experience in Turkey: A Data Mining Application. Int. J. Econ. Financ. Issues 7, 33–40.

Crémieux, P.Y., Ouellette, P., Pilon, C., 1999. Health care spending as determinants of health outcomes. Health Econ. 8, 627–639. https://doi.org/10.1002/(SICI)1099-1050(199911)8:7<627::AID-HEC474>3.0.CO;2-8

Dalbokova, D., Krzyzanowski, M., 2002. Environmental health indicators: Development of a methodology for the WHO European Region. Stat. J. UN. Econ. Comm. Eur. 19, 93–103. https://doi.org/10.3233/SJU-2002-191-209

Di Matteo, L., 2005. The macro determinants of health expenditure in the United States and Canada: Assessing the impact of income, age distribution and time. Health Policy (New. York). 71, 23–42. https://doi.org/10.1016/j.healthpol.2004.05.007

Dritsakis, N., Klazoglou, P., 2019. Time Series Analysis using ARIMA Models: An Approach to Forecasting Health Expenditure in USA. Econ. Internazionale / Int. Econ. 72, 77–106.

Durmaz, A., Pabuçcu, H., 2020. Kamu Sağlık Harcamaları - İşgücü Verimliliği İlişkisi: Sinirsel Bulanık Yaklaşım 8, 1375–1382.

Efe, B., Efe, O.F., 2016. An Application of Value Analysis for Lean Healthcare Management in an Emergency Department. Int. J. Comput. Intell. Syst. 9, 689–697. https://doi.org/https://doi.org/10.1080/18756891.2016.1204117

Guleryuz, D., Ozden, E., 2020. The Prediction of Brent Crude Oil Trend Using LSTM and Facebook Prophet. Eur. J. Sci. Technol. 1–9. https://doi.org/10.31590/ejosat.759302

Häkkinen, U., Martikainen, P., Noro, A., Nihtilä, E., Peltola, M., 2008. Aging, health expenditure, proximity to death, and income in Finland. Heal. Econ. Policy Law 3, 165–195. https://doi.org/10.1017/S174413310800443X

Jakovljevic, M., Timofeyev, Y., Ranabhat, C.L., Fernandes, P.O., Teixeira, J.P., Rancic, N., Reshetnikov, V., 2020b. Real GDP growth rates and healthcare spending - Comparison between the G7 and the EM7 countries. Global. Health 16, 1–13. https://doi.org/10.1186/s12992-020-00590-3

Jia, H., Jiang, H., Yu, J., Zhang, J., Cao, P., Yu, X., 2021. Total Health Expenditure and Its Driving Factors in China: A Gray Theory Analysis. Healthcare 9. https://doi.org/10.3390/healthcare9020207

Maksimović, G., Jović, S., Jovanović, R., Aničić, O., 2017. Management of health care expenditure by soft computing methodology. Phys. A Stat. Mech. its Appl. 465, 370–373. https://doi.org/10.1016/j.physa.2016.08.035

Martín, J.J.M., del Amo González, M.P.L., García, M.D.C., 2011. Review of the literature on the determinants of healthcare expenditure. Appl. Econ. 43, 19–46. https://doi.org/10.1080/00036841003689754

Ministry of Development, 2014. Onuncu Kalkınma Planı 2014-2018 Sağlık Hizmetlerinin Etkinliğinin Artırılması ve Mali Sürdürülebilirlik Özel İhtisas Komisyonu Raporu 2023 (In Turkish)

Mladenović, I., Milovančević, M., Sokolov Mladenović, S., Marjanović, V., Petković, B., 2016. Analyzing and management of health care expenditure and gross domestic product (GDP) growth rate by adaptive neuro-fuzzy technique. Comput. Human Behav. 64, 524–530. https://doi.org/10.1016/j.chb.2016.07.052

OECD, 2021. Health Resources (indicators) [WWW Document]. URL https://data.oecd.org/health.htm (accessed 2.1.21).

Özcan, T., Tüysüz, F., 2018. Healthcare expenditure prediction in Turkey by using genetic algorithm based grey forecasting models, International Series in Operations Research and Management Science. https://doi.org/10.1007/978-3-319-65455-3_7

Papanicolas, I., Woskie, L.R., Orlander, D., John Orav, E., Jha, A.K., 2019. The relationship between health spending and social spending in high-income countries: How does the US compare? Health Aff. 38, 1–9. https://doi.org/10.1377/hlthaff.2018.05187

Pekel, E., Akkoyunlu, M.C., Akkoyunlu, M.T., Pusat, S., 2020. Decision tree regression model to predict low-rank coal moisture content during convective drying process. Int. J. Coal Prep. Util. 40, 505–512. https://doi.org/10.1080/19392699.2020.1737527

Quan, Q., Hao, Z., Xifeng, H., Jingchun, L., 2020. Research on water temperature prediction based on improved support vector regression. Neural Comput. Appl. 4. https://doi.org/10.1007/s00521-020-04836-4

Ram, M., Akbar Haghdoost, A., Hossein Mehrolhassani, M., Abolhallaje, M., Dehnavieh, R., Najafi, B., Akbar Fazaeli, A., 2019. Forecastting health expennditures in Iran using the ARIMA.

Rana, R.H., Alam, K., Gow, J., 2020. Health expenditure and gross domestic product: causality analysis by income level. Int. J. Heal. Econ. Manag. 20, 55–77. https://doi.org/10.1007/s10754-019-09270-1

Samah, I.H.A., Rashid, I.M.A., Husain, WAFW, Ibrahim, S., Hamzah, H., Amlus, M.H., 2020. The impact of healthcare expenditure and healthcare sector growth on CO2 emission using dynamic panel data system GMM estimation model during COVID 19 crisis. Int. J. Energy Econ. Policy 10, 235–241. https://doi.org/10.32479/ijeep.9769

Shi, L., Smit, E., Luck, J., 2020. Panel Survey Estimation of the Impact of Urbanization in China: Does Level of Urbanization Affect Healthcare Expenditure, Utilization or Healthcare Seeking Behavior? Chinese Econ. 0, 1–12. https://doi.org/10.1080/10971475.2020.1848472

TURKSTAT, 2020. Turkish Statistical Institute [WWW Document]. URL http://www.tuik.gov.tr/Start.do (accessed 6.13.20).

Wang, Y., Li, J., Gu, J., Zhou, Z., Wang, Z., 2015. Artificial neural networks for infectious diarrhea prediction using meteorological factors in Shanghai (China). Appl. Soft Comput. J. 35, 280–290. https://doi.org/10.1016/j.asoc.2015.05.047

Worldbank, 2021. World Bank Open Data [WWW Document]. URL https://data.worldbank.org/

WHO, 2018. Delivering quality health services, World Health Organization, World Bank Group, OECD.

Zhang, C., Wei, H., Zhao, X., Liu, T., Zhang, K., 2016. A Gaussian process regression based hybrid approach for short-term wind speed prediction. Energy Convers. Manag. 126, 1084–1092. https://doi.org/10.1016/j.enconman.2016.08.086