Adıyaman University Journal of Science
https://dergipark.org.tr/en/pub/adyujsci
DergiPark AKADEMİK
ISSN 2147-1630
e-ISSN 2146-586X

# Prediction of Human Development Index with Health Indicators Using Tree-Based Regression Models

Pelin AKIN[1], Tuba KOÇ[2,*]

[1]*Çankırı Karatekin University, Faculty of Science, Department of Statistics,18100, Çankırı, Türkiye*
*pelinakin@karatekin.edu.tr, ORCID: 0000-0003-3798-4827*
[2]*Çankırı Karatekin University, Faculty of Science, Department of Statistics,18100, Çankırı, Türkiye*
*tubakoc@karatekin.edu.tr, ORCID: 0000-0001-5204-0846*

**Received:** 11.03.2021 **Accepted:** 19.11.2021 **Published:** 31.12.2021

**Abstract**

Machine learning is a field of artificial intelligence that allows computers to predict and model future events by making inferences from past information with mathematical and statistical operations. In this study, we used tree-based regression models, one of the machine learning methods, to determine and predict the effect of health indicators of 191 countries on the human development index (HDI) between 2014 and 2018 years. When tree-based regression models were compared according to model performance criteria, it was found that the best model was the gradient boosting model with the highest $R^2 = 0.9962$ and the smallest RMSE = 0.0094. With the gradient boosting model, the three most important variables to HDI are; current health expenditure per capita, physicians and nurses, and midwives, respectively. By selecting the ten countries with the highest HDI values and Turkey, HDI values were estimated for 2018-2019 with a gradient boosting model. The countries for which HDI values are best predicted by the gradient boosting method are Netherlands, Sweden, Norway, Iceland, Denmark, Turkey, Ireland, Germany, Australia, and China.

* Corresponding Author DOI: 10.37094/adyujsci.895084

***Keywords:*** Machine learning algorithms; Tree-based regression model; Gradient boosting method; Human development index; Health indicators.

## Ağaç Tabanlı Regresyon Modelleri Kullanılarak Sağlık Göstergeleri ile İnsani Gelişme Endeksinin Tahmini

**Öz**

Makine öğrenmesi, bilgisayar yardımıyla geçmişteki bilgileri kullanarak matematiksel ve istatistiksel işlemlerle çıkarımlar elde eden ve gelecekteki olaylar hakkında tahmin yürütülmesi modelleme yapılmasına imkân veren bir yapay zekâ alanıdır. Bu çalışmada 191 ülkenin 2014-2018 yıllarında sağlık göstergelerinin insani gelişim endeksi (İGE) üzerindeki etkisini belirlemek ve tahmin yapmak için makine öğrenmesi yöntemlerinden ağaç tabanlı regresyon modelleri kullanılmıştır. Ağaçlı tabanlı regresyon modelleri model performans kriterlerine göre karşılaştırıldığında en iyi modelin en yüksek $R^2 = 0.9962$ ve en küçük RMSE = 0.0094 değeri ile gradyan artırma model olduğu bulunmuştur. Gradyan artırma model ile İGE indeksine en fazla etki eden 3 değişken sırasıyla: kişi başına cari sağlık harcaması, doktorların sayısı ve hemşireler ile ebelerin sayısı olarak bulunmuştur. İGE değeri en yüksek olan 10 ülke ve Türkiye seçilerek gradyan artırma model ile 2018-2019 yılları için İGE değerleri tahmin edilmiştir. Gradyan artırma yöntemi ile İGE değeri en iyi tahmin edilen ülkeler sırasıyla Hollanda, İsveç, Norveç, İzlanda, Danimarka, Türkiye, İrlanda, Almanya, Avustralya ve Çin şeklindedir.

***Anahtar Kelimeler:*** Makine öğrenmesi algoritmaları; Ağaç tabanlı regresyon modelleri; Gardyan arttırma model; İnsani gelişim endeksi; Sağlık değişkenleri.

### 1. Introduction

Machine learning, a new and promising sub-branch of algorithmic data analysis, has rapidly advanced in recent years. Due to the rapid increase in available storage space, processing power, and network connectivity, there has been great progress in data collection, sharing, and processing technologies. Also, given the recent increase in the volume of data from all sources, it is possible to apply learning methods in increasingly complex data that are impossible to analyze with prior technology. Machine learning programs computers use statistical theory to create mathematical models to optimize a performance criterion using sample data or past experiences [1]. Several algorithms are used in machine learning to categorize data sets, analyze their results, and make predictions. Machine learning algorithms are widely used in many sectors such as education, economy, marketing, health, etc. In healthcare, machine learning algorithms are used in many areas such as prediction, diagnosis, disease tracking, and processing of unstructured data [2, 3].

Health services are considered one of the main determinants of economic and social development. Health data belonging to health services guide the development of countries in the field of health. The Human Development Index (HDI) is a concept introduced by the United Nations to measure countries' human development levels. HDI determines the human development levels of societies through three main areas: health, education, and income. Yakut et al. [4] using infant mortality rate, gross national product, high school enrollment rate, growth, foreign direct investment, energy consumption, energy production, inflation, exports, number of internet users, unemployment, imports, mobile phone subscribers, and health expenditures. They made a classification using ordered logistic regression analysis and artificial neural networks on the Human Development Index of 81 countries. Zhang et al. [5] applied the Gradient Boosting Decision Tree Algorithm (GBDT) to analyze health data and make predictions. They concluded that the GBDT algorithm performs better than the traditional least-squares method, ridge regression, lasso regression, ElasticNet, SVR, and KNN algorithm methods. König et al. [6] analyzed the impact of multimorbidity on health care costs in Germany on all sectors of care using an advanced tree-based graphic model. Rençber and Mete [7] classified countries according to the Human Development Index (HDI) using machine learning techniques such as Artificial Neural Network (ANN) and Adaptive Neural Fuzzy Inference System (ANFIS) and compared the results with the HDI. Yakut and Korkmaz [8] created decision trees with C5.0 and Gini algorithms using data from 79 countries from 2010-2017. They determined the HDI factors by the decision trees method and classified the countries as very high, high, medium, and low-level developed countries. They determined that the variables that affect most the HDI are education, employment, and health indicators. Dos Santos et al. [9] used the SMOReg data mining algorithm to predict Latin American countries' human development index and life expectancy. Hu et al. [10] examined the critical factors of high costs for breast cancer patients using the Quantile Regression Forests approach, a flexible tree-based machine learning technique using health data. Saboo et al. [11] compared the ANN and linear regression-based approach in estimating the HDI. Coşar [12] made a classification using Naive Bayes, ANN, and logistic regression methods to determine the effect of healthcare indicators in OECD (Organization for Economic Co-operation and Development) countries on the HDI.

In this study, we used tree-based regression models, one of the machine learning methods, to determine and predict the effect of health indicators of all countries on the human development index. The article is divided as follows: In Section 2, decision tree, random forest, extreme gradient boosting, and gradient boosting from Tree-Based Regression models are defined. In Section 3, the application of Tree-Based Regression models is explained with data for all countries. Finally, a brief discussion is given in Section 4.

## 2. Tree-Based Regression Models

Tree-based regression models use one or more decision trees. We considered four tree-based machine learning methods, decision tree (CART), random forest regression, extreme gradient boosting (XGBoost), and gradient boosting model.

### 2.1. Decision tree

CART was created by [13]. The CART programs construct classification or regression models of a very general structure using a two-step process; the resulting models may be represented as binary trees [14]. The mean squared error is used for the split data in the CART algorithm. Mean squared error (MSE) for a specific node is defined as;

$$\mathrm{MSE_{node}} = \frac{1}{\mathrm{m_{node}}}\sum(\mathrm{y_i} - \bar{y}_{node})^2. \quad (1)$$

If it is assumed that there is a binary split on each node on the tree, it will be divided into left and right. For each division, the error term with

$$\mathrm{MSE_{left}} = \frac{1}{\mathrm{m_{left}}}\sum(\mathrm{y_i} - \bar{\mathrm{y}}_{\mathrm{left}})^2, \quad (2)$$

$$\mathrm{MSE_{right}} = \frac{1}{\mathrm{m_{right}}}\sum(\mathrm{y_i} - \bar{y}_{right})^2. \quad (3)$$

For each attribute j, the following formula is calculated,

$$\min(\mathrm{MSE_{left}} + \mathrm{MSE_{right}}). \quad (4)$$

The smallest of the values is chosen. The dataset splits recursively, which means that the subsets that meet a partition are partitioned until they reach a predetermined expiration criterion [14].

### 2.2. Random Forest regression

The ensemble learning algorithms produce a prediction model by combining the strong points of a group of simpler, a lot of basic models [15]. The most widely used ensemble learning algorithms are bagging and Random Forest algorithms. Breiman's Random Forest classification is an improved version of the bagging technique achieved by adding the randomness feature. The following steps are taken for the Random Forest algorithm: firstly, n bootstrap samples are taken from the original data set. Then CART is created for each bootstrap sampling. A new estimate is made by combining the estimates made by n trees separately. Estimation is made by taking the average of the results made in regression trees [16].

### 2.3. Extreme gradient boosting

Chen and Guestrin [17] proposed a scalable, end-to-end tree strengthening system called XGBoost in 2016. XGBoost, its algorithm is also called a regular gradient boosting technique. The XGBoost algorithm is optimized by using different arrangements of the Gradient Boosting algorithm. XGBoost is fast to interpret, prevents overlearning, and can handle large-sized datasets well.

### 2.4. Gradient boosting model

Friedman [18, 19] established the foundation for the next generation of boosting algorithms. Ridgeway [20] proposed a gradient boosting model using the link between upgrade and optimization. This algorithm can be preferred for regression and classification problems. It creates a model with a combination of its weak models. Increasing the gradient here aims to reach the minimum error values by updating the predictions according to the learning rate.

### 2.5. Evaluation metrics for regression models

The mean absolute error (MAE), the mean squared error (MSE), the root mean squared error (RMSE), and the coefficient of determination ($R^2$) is used for model selection. $R^2$ is used to measure the wellness of the fit by the trained models. MAE, MSE, RMSE are the average error measures [25]. The error measures and $R^2$ are defined as follows

$$\begin{aligned}
&\mathrm{MAE} = \frac{1}{\mathrm{N}}(\sum|\mathrm{y_i} - \hat{\mathrm{y}}|),\\
&\mathrm{MSE} = \frac{1}{\mathrm{N}}\sum(\mathrm{y_i} - \hat{\mathrm{y}})),\\
&\mathrm{RMSE} = \sqrt{\frac{1}{\mathrm{N}}\sum(\mathrm{y_i} - \hat{\mathrm{y}})},\\
&\mathrm{R^2} = 1 - \frac{\sum(\mathrm{y_i} - \hat{\mathrm{y}})^2}{\sum(\mathrm{y_i} - \bar{\mathrm{y}})^2}.
\end{aligned} \tag{5}$$

## 3. Application Part

In this study, the health indicators and HDI of 2014 and 2018 were used for 191 countries. The data obtained are available URL1-2 [21, 22]. The variables used in the study are given in Table 1.

**Table 1:** Description of the variables

| Variable | Description |
|---|---|
| HDI | Human development index |
| $x_1$ | Gross domestic product growth (annual %) (GDP) |
| $x_2$ | Current health expenditure (% of GDP) |
| $x_3$ | Hospital beds (per 1,000 people) |
| $x_4$ | Specialist surgical workforce (per 100,000 population) |
| $x_5$ | Current health expenditure per capita (current US$) |
| $x_6$ | Nurses and midwives (per 1,000 people) |
| $x_7$ | Physicians (per 1,000 people) |

During the preparation of the data set for analysis, the multiple assignments (MICE) method with chained equations was used for the missing value. In the MICE methods, a statistical distribution is obtained over the data set. Then thanks to this distribution, a link is used that fills in the missing value [23]. After the data set was completed, the data for the years 2014-2017 were divided as training and the data for 2018 as test data. CART, random forest regression, XGBoost, and gradient boosting model were applied. Analyzes were performed using version 3.5.2 of the R software.

Performance criteria for training and test data are given in Table 2 to compare models.

**Table 2:** Performance measurement for Models

| MODELS | Train Data | | | | Test Data | | | |
|---|---|---|---|---|---|---|---|---|
| | $R^2$ | RMSE | MAE | MSE | $R^2$ | RMSE | MAE | MSE |
| CART | 0.8309 | 0.0620 | 0.0441 | 0.0038 | 0.8704 | 0.0546 | 0.0395 | 0.0029 |
| Random forest | 0.9884 | 0.0170 | 0.0116 | 0.0003 | 0.8865 | 0.0579 | 0.0416 | 0.0034 |
| XGBoost | 0.9884 | 0.0170 | 0.0117 | 0.0002 | 0.7181 | 0.0884 | 0.0664 | 0.0078 |
| Gradient boosting model | 0.9962 | 0.0094 | 0.0073 | 0 | 0.8942 | 0.0539 | 0.0394 | 0.0028 |

The model with the highest $R^2$ value and lower error rates than RMSE, MAE, and MSE shows the best performance. In Table 2, the algorithm that gives the best performance in predicting training data and test data is the gradient boosting model. When the gradient boosting model is applied to the training data, the visible results of the model are as shown in Fig. 1.

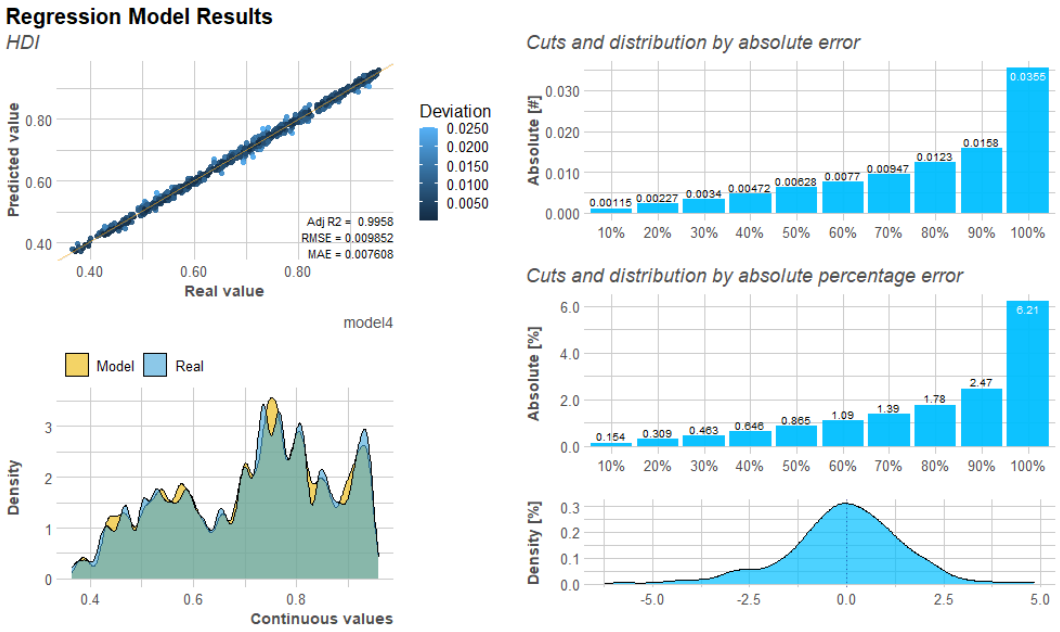


**Figure 1:** Gradient boosting regression model results for train data

Figure 1 shows the regression results plot, errors plot, and distribution plot. When the graphs are examined, it is seen that the gradient boosting model is suitable for the data. Feature importance graph describes which features are relevant.

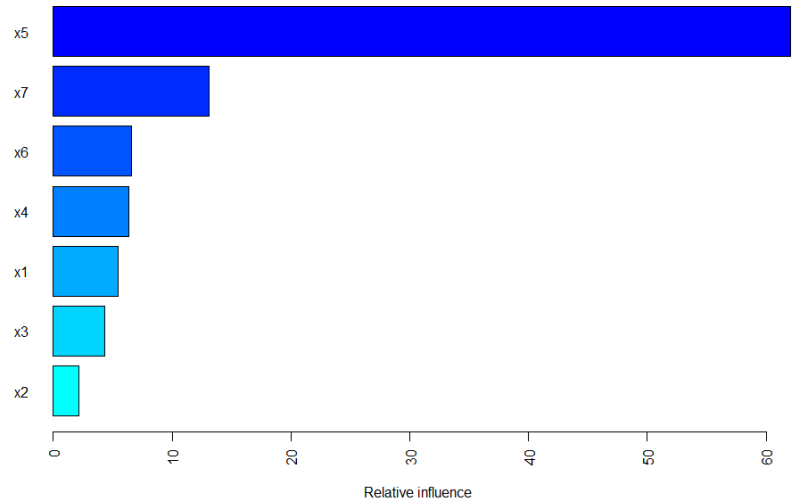


**Figure 2:** Gradient boosting model feature importance bar chart

In Fig. 2, it is seen that the three most important variables for the gradient boosting model used to determine the effect of health indicators on HDI are x5, x7, and x6, respectively.

When the gradient boosting model is applied to the data of the ten countries with the highest HDI index and Turkey, the predicted values for 2018 and 2019 are as follows.

**Table 3:** HDI index prediction with gradient boosting model

| **Countries** | **HDI$_{2018}$** | **Predicted$_{2018}$** | **Predicted$_{2019}$** |
|---|---|---|---|
| Norway | 0.947 | 0.956 | 0.95304 |
| Switzerland | 0.953 | 0.955 | 0.90801 |
| Ireland | 0.922 | 0.951 | 0.93866 |
| Germany | 0.908 | 0.946 | 0.91917 |
| China | 0.879 | 0.946 | 0.92261 |
| Iceland | 0.958 | 0.946 | 0.95179 |
| Australia | 0.893 | 0.943 | 0.92787 |
| Sweden | 0.951 | 0.943 | 0.90990 |
| Netherlands | 0.943 | 0.942 | 0.93263 |
| Denmark | 0.952 | 0.939 | 0.93321 |
| Turkey | 0.804 | 0.817 | 0.85975 |

When the 2018 actual values (HDI) of the countries selected from Table 3 and the predicted values for 2018 and 2019 are examined, the country with the best-predicted HDI value with the gradient boosting method is found to be the Netherlands, and the worst predicted country China. HDI forecast values change charts of selected countries for 2018 and 2019 are given in Fig. 3.

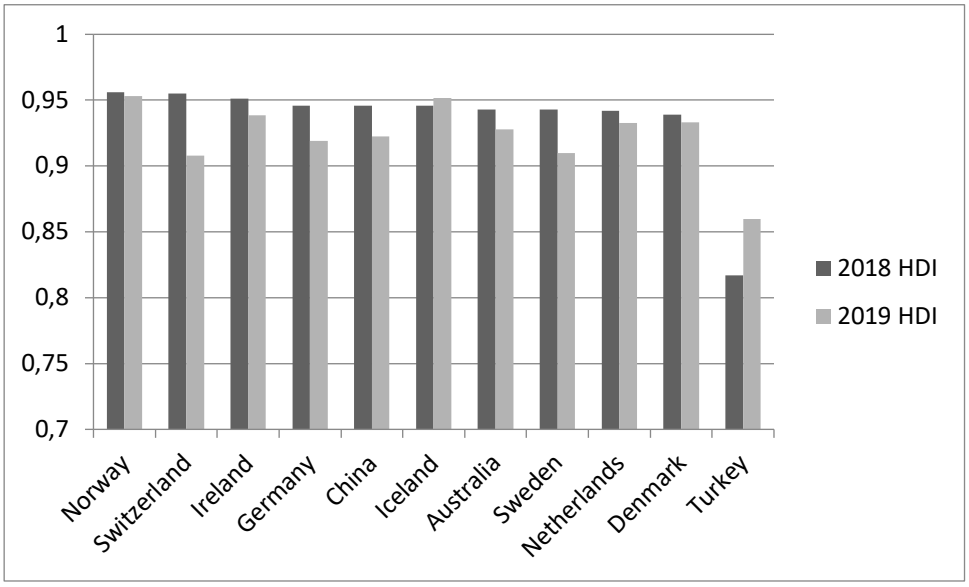


**Figure 3:** HDI forecast values comparison graph of countries

Figure 3 shows HDI predictions with health indicators, and it is seen that HDI values of Iceland and Turkey increase, while HDI values of Norway, Switzerland, Ireland, Germany, China, Australia, Sweden, Netherlands, and Denmark countries decrease.

## 4. Conclusion and Discussion

Human development, which is a concept that aims to raise the living standards of societies to the living standards of the modern world, is defined as the process that enables individuals to live their lives as they value and to exercise their basic human rights [25]. Health is one of the three most important components of the HDI.

In this study, the health indicators and HDI of 2014 and 2018 were used for 191 countries. Health indicators; GDP, current health expenditure, hospital beds specialist surgical workforce, current health expenditure per capita, nurses and midwives, and physicians have been selected from several indicators that may potentially impact the HDI. First, tree-based decision tree, random forest, extreme gradient boosting, and gradient boosting model methods, which are machine learning algorithms, were applied to the training data to determine the effect of health indicators on HDI of all countries and to make predictions. When the model performance criteria were examined, it was found that the best model was the gradient boosting model with the highest $R^2 = 0.9962$ and the smallest RMSE = 0.0094. According to the gradient boosting model results, the three variables that have the greatest effect on the HDI index are current health expenditure per capita, physicians, nurses, and midwives. Then ten countries with the highest HDI values and Turkey were chosen, and HDI values were estimated for 2018-2019 with the gradient boosting model. The countries with the best HDI value estimated by gradient boosting method are Netherlands, Sweden, Norway, Iceland, Denmark, Turkey, Ireland, Germany, Australia, and China, respectively. The gradient boosting method has estimated HDI values with the same mistake for Turkey and Denmark. However, Turkey has increased the HDI in the years, lagged behind many developed countries. Considering the analysis results, it is seen that the effect of the level of health expenditure on the HDI is quite high. Therefore, countries should allocate more resources to the field of health. Also, given the impact of numbers of physicians, nurses, and midwives on human development, it is obvious that more healthcare staff investment will increase human development at higher rates. These results are very consistent with the literature [8]. The limitation of this study is that the HDI index of 2020 and 2021 could not be predicted. HDI predictive values can be calculated when current data on health variables of countries are available.

**References**

[1] Alpaydın, E., *Yapay öğrenme*, Boğaziçi Üniversitesi Yayınevi. 2011.

[2] Alonso, D.H., Wernick, M.N., Yang, Y.Y., Germano, G., Berman, D.S., Slomka, P., P*rediction of cardiac death after adenosine myocardial perfusion SPECT based on machine learning*, Journal of Nuclear Cardiology, 26(5), 1746-1754, 2019.

[3] Kavakiotis, I., Tsave, O., Salifoglou, A., Maglaveras, N., Vlahavas, I., and Chouvarda, I., *Machine learning and data mining methods in diabetes research*, Computational and Structural Biotechnology Journal, 15, 104-116, 2017.

[4] Yakut, E., Gündüz, M., and Demirci, A., *Comparison of classification success of human development index by using ordered logistic regression analysis and artificial neural network methods*, Journal of Applied Quantitative Methods, 10(3), 15-34, 2015.

[5] Zhang, B., Ren, J., Cheng, Y., Wang, B., Wei, Z., *Health data driven on continuous blood pressure prediction based on gradient boosting decision tree algorithm*, IEEE Access, 7, 32423-32433, 2019.

[6] Konig, H.H., Leicht, H., Bickel, H., Fuchs, A., Gensichen, J., Maier, W., Mergenthal, K., Riedel-Heller, S., Schafer, I., Schon, G., Weyerer, S., Wiese, B., van den Bussche, H., Scherer, M., Eckardt, M., Grp, M.S., *Effects of multiple chronic conditions on health care costs: an analysis based on an advanced tree-based regression model,* BMC Health Services Research 13(1), 1-13, 2013.

[7] Rençber, Ö.F., Sinan, M., *Reclassification Of Countries According To Human Development Index: An Application With Ann And Anfis Methods,* Business & Management Studies: An International Journal, 6(3), 228-252, 2018.

[8] Yakut, E., Korkmaz, A., *İnsani Gelişmişlik Endeksinin Karar Ağacı Algoritmaları ile Modellenmesi: BM'de Bir Uygulama 2010-2017 Dönemi,* Anadolu Üniversitesi Sosyal Bilimler Dergisi, 20(2), 65-84, 2020.

[9] dos Santos, C.B., Pilatti, L.A., Pedroso, B., Carvalho, D.R., Guimaraes, A.M., *Forecasting the human development index and life expectancy in Latin American countries using data mining techniques,* Ciência & Saúde Coletiva, 23(11), 3745-3757, 2018.

[10] Hu, L., Li, L., Ji, J., Sanderson, M., *Identifying and understanding determinants of high healthcare costs for breast cancer: a quantile regression machine learning approach,* BMC Health Services Research, 20(1), 1066, 2020.

[11] Saboo, A., Parakh, R., Trivedi, P., Potdar, M., *A Comparative Study of Artificial Neural Networks and Multiple Linear Regression by Predicting Human Development Index,* International Journal of Scientific & Engineering Research, 7(9), 424-428, 2016.

[12] Çoşar, K., *OECD sağlık verilerinin veri madenciliği yöntemleri ile analizi*, Marmara University, İstanbul, 2020.

[13] Breiman, L., Friedman, J., Olshen, R., Stone, C., *Classification and regression trees,* Chapman and Hall/CRC, 1998.

[14] Therneau, T.M., Atkinson, E.J., *An introduction to recursive partitioning using the RPART routines*, Technical report Mayo Foundation, 1997.

[15] Friedman, C., Sandow, S., *Utility-based learning from data*, Machine learning & pattern recognition series, Boca Raton: Chapman & Hall/CRC, 397 pages, 2011.

[16] Liaw, A., Wiener, M., *Classification and regression by random Forest,* R News, 2(3), 18-22, 2002.

[17] Chen, T., Guestrin, C., *Xgboost: A scalable tree boosting system*, in Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 2016.

[18] Friedman, J. H., *Greedy function approximation: a gradient boosting machine*, Annals of Statistics, 1189-1232, 2001.

[19] Friedman, J.H., *Stochastic gradient boosting*, Computational Statistics & Data Analysis, 38(4), 367-378, 2002.

[20] Ridgeway, G., *Generalized boosted models: A guide to the gbm package,* Update, 1(1), 2007, 2007.

[21] URL-1, https://data.worldbank.org/ Erişim tarihi 25.12.2020.

[22] URL-2, http://hdr.undp.org/en/content/human-development-index-hdi Erişim tarihi 25.12.2020

[23] Şeker, Ş.E., Eşmekaya, E., *Eksik verilerin tamamlanması (imputation),* YBS Ansiklopedi, 4, 10-17, 2017.

[24] Uğuz, S., Makine Öğrenmesi Teorik Yönleri ve Python Uygulamaları ile Bir Yapay Zeka Ekolü, Nobel, Ankara, 2019.

[25] Uygur, S., Yıldırım, F., *Cinsiyete bağlı insani gelişme endeks yaklaşımları: Türkiye örneği,* Tisk Akademi, 30-59, 2011.