Comparison of Machine Learning Methods in Prediction Gini Coefficient for OECD Countries

Tuba Koç[1]*, Pelin Akın[2]
[1,2] *Department of Statistic, Faculty of Science, Çankırı Karatekin University, Çankırı, Turkey*

***Abstract*— Income inequality refers to the situation where income distribution is not shared regularly and fairly. Income inequality is among the essential problems of countries in both economic and social terms. The Gini coefficient is widely used to measure income inequality. In this study, random forest, support vector algorithms, and multiple linear regression model, which are among the machine learning algorithms, were applied to estimate the Gini coefficient of Organization for Economic Co-operation and Development (OECD) countries for 2015-2018. When the models were compared according to performance criteria, the random forest model was the highest R^2 = 0.7085 and the smallest RMSE = 0.0264. The random forest model results show that the tax revenue variable has the greatest impact on the Gini coefficient. The country with the highest Gini coefficient is Mexico, and the lowest is the Slovak Republic. Also, it has been observed that the lowest tax income value belongs to Mexico. Consequently, the Gini coefficients of OECD countries can be predicted by random forest algorithm for the future period.**

Keywords: Support vector, Random forest, Gini coefficient, OECD.

I. INTRODUCTION

Inequality of income is an unfair distribution of income among individuals in which an individual's share in total income is less than other individuals. Inequalities of the income distribution are used as an indicator of the general economic inequality of societies. There has been an increase in inequality of income with globalization and rising welfare in the world. The economic growth experienced by low-income countries with the opportunities of globalization has deepened the inequality between them and other developing countries. The increase in income inequality poses a major obstacle to economic development, as it reduces the share of the poor in economic growth.

Inequality of income is one of the important issues discussed and emphasized both economically and socially. Many measures have been developed to understand better the main reasons and effects of inequality of income. The commonly used measurements are the Gini coefficient, Lorenz curve, and Palma ratio [1]. The Gini coefficient is widely used due to its effective analysis of income distribution, easier computing, and graphical tools. The Gini coefficient measures how the income distribution among individuals or households deviates from the whole egalitarian level in the economy. In literature, many studies are using the Gini coefficient. Li et al. [2] found a quadratic relationship between corruption and the Gini coefficient, and the highest Gini coefficient in countries with medium-level corruption concluded that the Gini coefficient has a lower value in countries with low and high corruption rates. Peçe et al. [3], 1977-2013 period Turkey's GDP per person has analyzed the impact on income distribution and used the Gini coefficient as the income distribution criterion in their studies. Yazgan [4] among the provinces of Turkey in the 1999-2017 period public investment has examined the extent to which unequally distributed by computing the Gini Coefficient. Öz [5] investigated the interaction between the varying Gini coefficient and poverty in Turkey before and after the economic breaking point. Demir [6] examined the relationship between the Gini coefficient used to measure income distribution inequality and some selected countries' luxury goods import expenditures. Zaman et al. [7] determined the factors affecting the Gini coefficient with the beta regression method. Basumatary et al. [8] have evaluated per capita electricity consumption inequalities in the northeastern states of India using the Gini Index and the Lorenz Curve.

This study aims to contribute to the literature by using machine learning methods to estimate GINI coefficients of OECD countries. In this study, we used random forest and support vector algorithms, one of the machine learning methods, a multiple regression model to determine and predict the Gini coefficient of OECD countries. The paper is divide as follows: In section 2, support vector and random forest regression from machine learning algorithms and multiple regression models are defined. In section 3 explains the application of the machine learning algorithms with Gini coefficient data. Finally, a brief discussion is given in Section 4.

II. METHODS

A. Support Vector Regression

Support vector machines are machine learning techniques based on statistical learning theory and the principle of structural risk minimization. The first study on support vector machines was presented in 1992 by Vladimir Vapnik, Bernhard Boser, and Isabelle Guyon [9]. Consider the problem of the set of training data $D\{(x^1, y^1), \ldots, (x^l, y^l)\}$ with a linear function,

Manuscript received July 05, 2021; accepted September 06, 2021.
*Corresponding author: pelinakin@karatekin.edu.tr

$$y = \langle w, x \rangle + b \quad (1)$$

The optimum regression function is provided from the minimum of the function,

Minimize

$$\frac{1}{2}\|w\|^2 + c\sum(\xi_i + \xi_i^*) \quad (2)$$

Constraints

$$\begin{aligned} y_i - (wx_i) - b &\leq \varepsilon + \xi_i \\ wx_i + b - y_i &\leq \varepsilon + \xi_i^* \\ \xi_i, \xi_i^* &\geq \end{aligned} \quad (3)$$

where (*) symbolizes both the vector with and without asterisks. ξ_i, ξ_i^* slack variable and c > 0 is a penalty parameter [10]. The constrained optimization problem is then reworded as a dual problem using Lagrange multipliers $a_i\ a_i^*$ for each constraint. Lagrange multipliers are determined by solving the issue with Quadratic Programming (QP). After $a_i\ a_i^*$ are determined, the optimal weights w and the base b can be calculated, and the final predictor is given in the equation [11]

$$y = \sum_i^m (a_i^* - a_i)(x_i - x) + b \quad (4)$$

B. Random Forest Regression

The Random Forests algorithm is one of the ensemble learning algorithms. The ensemble learning algorithms produce a prediction model by combining the strong points of a group of simpler, a lot of basic models [12]. 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 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 a regression tree (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[13] .

C. Multiple Linear Regression

Multiple linear regression is the most common form of linear regression analysis. Multiple linear regression explains the relationship between one continuous dependent variable and two or more independent variables. The multiple regression is given by

$$Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \cdots + \varepsilon_i \quad (5)$$

where y is the dependent variable, β_i the regression parameters (i=0,1,2,…, p), X_i the independent variables (i=0,1,2,…, p) and ε_i the random error term. The least-squares method is used for parameter estimation in the multiple regression model. Regression coefficients are estimated as follows.

$$\hat{\beta} = (X'X)^{-1}(X'Y) \quad (5)$$

Once regression coefficients are obtained, a prediction equation can then be used to predict the dependent value as a linear function of one or more independent inputs. The reason for the popularity of the regression models is the interpretability of model parameters and ease of use.

D. Evaluation Metrics for Regression models

Commonly used metrics to evaluate forecast accuracy are the mean absolute error (MAE), the mean squared error (MSE), the root mean squared error (RMSE), and the Coefficient of determination (R^2) [14]. R^2 is used to measure the wellness of the fit by the trained models. A high R^2 value indicates that the prediction relationship is good. MAE, MSE, RMSE are the average error measure, so low values indicate good performance. The error measures are defined as follows

$$\begin{aligned} \text{MAE} &= \frac{1}{N}(\sum|y_i - \hat{y}| \\ \text{MSE} &= \frac{1}{N}\sum(y_i - \hat{y}) \\ \text{RMSE} &= \sqrt{\frac{1}{N}\sum(y_i - \hat{y})} \\ R^2 &= 1 - \frac{\sum(y_i - \hat{y})^2}{\sum(y_i - \bar{y})^2} \end{aligned} \quad (6)$$

III. Application

In this study, Gini coefficient data were used for OECD countries for the years 2015-2018. The data obtained is available URL1-2. A set of 6 continuous variables are used in this paper and described as Gini coefficient, tax revenue, gross domestic product(us dollar/capita), unemployment rate(% of the labor force), inflation (Annual growth rate %), current health expenditure (% of GDP).

The data for 2015-2017 were divided as training and the data for 2018 as test data. Support vector regression, random forest regression, and multiple linear regression were applied. Analyzes were performed using version 3.5.2 of the R software. The performance criteria values obtained from the test are given.

Table 1. Performances measurements for models

Performance measurements	Support Vector Regression	Random Forest Regression	Multiple Linear Regression
MAE	0.0232	0.0199	0.0303
MSE	0.0010	0.0006	0.0014
RMSE	0.0326	0.0264	0.0374
R-squared	0.6611	0.7085	0.4283

Table 1, random forest regression algorithm is the best model with the highest R^2 and the smallest MAE value. Feature importance graph describes which features are relevant. The order of importance of variables is as given in Fig. 1.

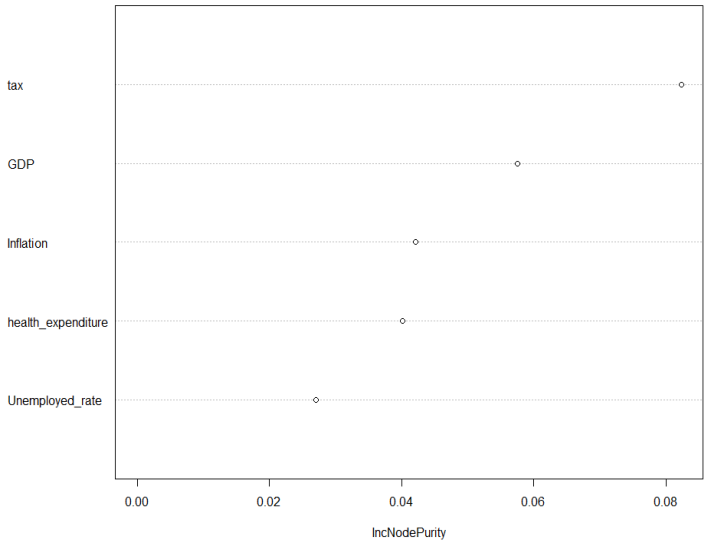

Fig.1. Random Forest features importance

In Fig. 1, it is seen that the most important variable used to determine the effect of the Gini coefficient is tax revenue for the random forest model. The tax revenue variable and Gini coefficient for 2018 are given in Fig. 2.

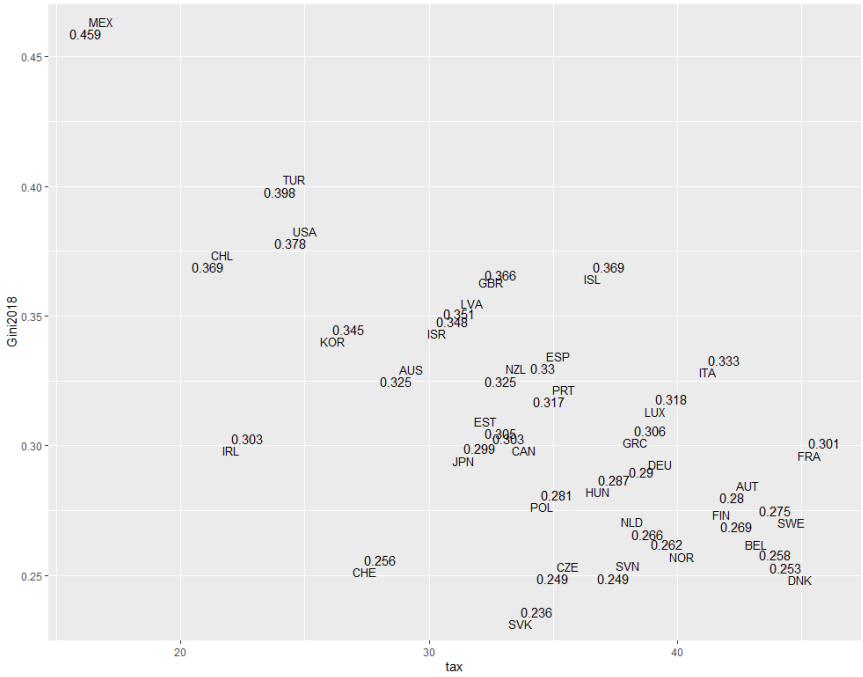

Fig.2. The tax revenue and Gini coefficient value comparison of OECD counties

When the Gini coefficient and tax revenue actual values for 2018 are examined in Fig. 2, the highest Gini coefficient is Mexico, and the lowest country is the Slovak Republic. In addition, the lowest tax revenue value belongs to Mexico. The tax revenue variable and Gini coefficient predicted value is given as follow graphs.

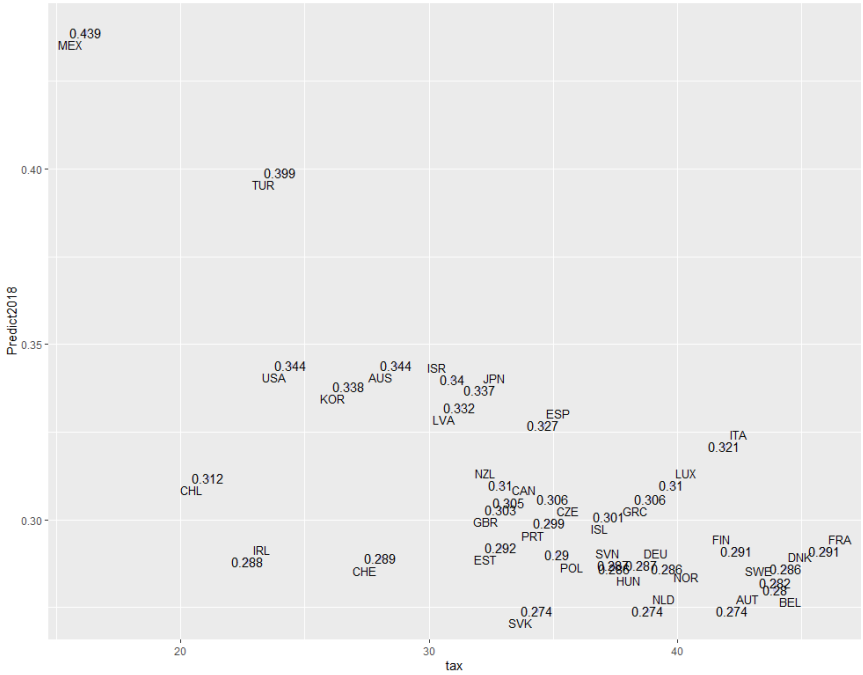

Fig.3. The tax revenue and Gini coefficient value Comparison of OECD counties

In Fig. 3, it is seen that the results of the estimates that are very close to the actual values are obtained with the random forest model. Fig. 4 shows the forecast values for 2019.

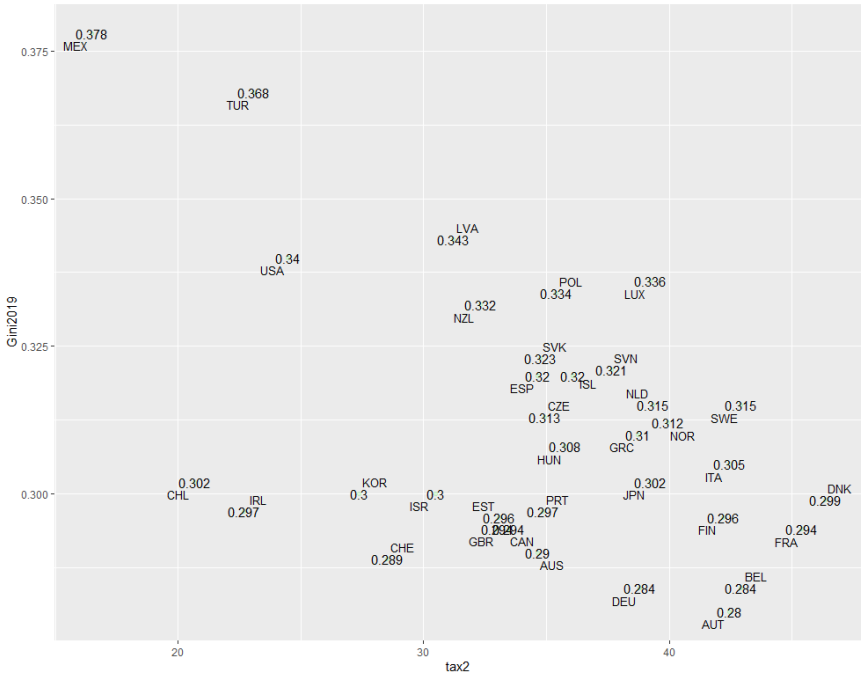

Fig.4. The tax revenue and Gini coefficient predicted values comparison of OECD counties

When Fig. 4 is examined, the highest Gini coefficient was estimated as Mexico with 0.459, and Austria with the lowest Gini coefficient of 0.28. In addition, the lowest tax revenue value was found to be similar to 2018 in Mexico.

IV. Results and Discussion

Income inequality is one of the most important macroeconomic indicators of economic development in developed and developing countries. Increasing income

inequality hinders the healthy and stable growth of the economy. The most popular tool for measuring income inequality is the Gini coefficient.

This study used the Gini coefficient and the factors affecting the Gini coefficient indicators for OECD countries between 2014 and 2018. The variables of tax revenue, gross domestic product, unemployment rate, inflation, and current health expenditure have been selected from several indicators that may potentially impact the Gini coefficient.

First, multiple linear regression, support vector, and random forest were applied to the training data to determine the effect of variables on the Gini coefficient of OECD countries and make predictions.

When the model performance criteria were examined, it was found that the best model was the random forest model with the highest R^2 =0.7085 and the smallest RMSE = 0.0264. The random forest model results show that the tax revenue variable has the greatest impact on the Gini coefficient.

According to the study's findings, it is seen that estimation results very close to the actual values are obtained with the random forest model for 2018. The highest Gini coefficient is the country Mexico and the lowest country is the Slovak Republic. In addition, the lowest tax revenue value belongs to Mexico. The predicted value for 2019's year is examined, the highest Gini coefficient was estimated as Mexico with 0.459, the lowest as Austria with 0.28, and the lowest tax revenue value was found to be similar to 2018's year in Mexico. As a result, it has been observed that countries with high tax revenue have low Gini coefficients. The decrease in the Gini index also decreases the inequality of income. These results are very consistent with the literature [15, 16]. The tax system that countries should establish can develop tax systems that do not prevent productive wealth accumulation as required by the economic structure but taxing the unproductive accumulation of wealth and efficient investments that increase employment and technology direct foreign capital investments that require technology transfer. [17].

As a result, the Gini coefficients of OECD countries can be estimated with the random forest algorithm for the next period. In addition, countries will have information to improve the Gini coefficient.

ACKNOWLEDGMENT

Part of this work was presented orally at the IV. International Conference on Data Science and Applications 2021 (ICONDATA'21).

REFERENCES

[1] Niyimbanira F. Analysis of the impact of economic growth on income inequality and poverty in South Africa: the case of Mpumalanga Province. International Journal of Economics and Financial Issues, 2017.

[2] Li H, Xu L C, Zou H f. Corruption, income distribution, and growth. Economics & Politics, 2000. 12(2): p. 155-182.

[3] Peçe M A, Ceyhan M S, Akpolat A. Türkiye'de gelir dağılımının ekonomik büyümeye etkisi üzerine ekonometrik bir analiz. Uluslararası Kültürel Ve Sosyal Araştırmalar Dergisi (Uksad), 2016. 2(Special Issue 1): p. 135-148.

[4] Yazgan Ş. Kamu yatırımları dağılımının gini katsayısı ile ölçülmesi: türkiye üzerine bir uygulama (1999-2017). Uluslararası Ekonomi Siyaset İnsan ve Toplum Bilimleri Dergisi, 2018. 1(1): p. 35-44.

[5] Öz S. Gelir dağılımında gını katsayısı ve p80/p20 oranı arasındaki ilişkiler: 2000-2016 dönemi Türkiye örneği. 2019.

[6] Demir M A. Gelir dağılımı eşitsizliği ve lüks mal ithalatı arasında panel nedensellik analizi. Akademik Araştırmalar ve Çalışmalar Dergisi (AKAD), 2020. 12(23): p. 471-483.

[7] Zaman T, Dünder E, Aydın S. Gini katsayısını etkileyen faktörlerin beta regresyon yöntemi yardımı ile belirlenmesi. Erzincan Üniversitesi Fen Bilimleri Enstitüsü Dergisi, 2019. 12(1): p. 235-240.

[8] Basumatary S, Devi M, Basumatary K. Assessing the Disparities of Per-capita Electricity Consumption in North-Eastern States of India Using Gini Index and Lorenz Curve. Journal of Humanities and Social Sciences Studies, 2021. 3(1): p. 103-107.

[9] Vapnik V. Principles of risk minimization for learning theory. in Advances in neural information processing systems. 1992.

[10] Gunn S R. Support vector machines for classification and regression. ISIS technical report, 1998. 14(1): p. 5-16.

[11] Shokry A, et al., Modeling and simulation of complex nonlinear dynamic processes using data based models: Application to photo-Fenton process, in Computer Aided Chemical Engineering. 2015, Elsevier. p. 191-196.

[12] Friedman J H. Greedy function approximation: a gradient boosting machine. Annals of statistics, 2001: p. 1189-1232.

[13] Liaw A Wiener M. Classification and regression by randomForest. R news, 2002. 2(3): p. 18-22.

[14] Uğuz S, Makine Öğrenmesi Teorik Yönleri ve Python Uygulamaları ile Bir Yapay Zeka Ekolü. 2019, Ankara: Nobel. 312.

[15] Ekici M S. Vergi gelirlerini etkileyen ekonomik ve sosyal faktörler. Elektronik Sosyal Bilimler Dergisi, 2009. 8(30): p. 200-223.

[16] .Balcı B İ, Alyu E. Yoksulluk ve Gelir Dağılımı Eşitsizliği: OECD ve AB Ülkeleri Panel Veri Analizi. Gaziantep Üniversitesi Sosyal Bilimler Dergisi, 2018. 17(3): p. 988-996.

[17] Yanpar A. Gelişmekte olan ülkelerde büyüme yönelimli vergi politikası. Ankara Üniversitesi, SBE-Yüksek Lisans Tezi, 2007.