Journal of New Theory
ISSN: 2149-1402

37 (2021) 58-67
Journal of New Theory
https://dergipark.org.tr/en/pub/jnt
Open Access

Efficiency Analysis and Estimation of Factors Affecting the Efficiency with Decision Trees in Imbalanced Data: A Case of Turkey's Environmental Sustainability

Selin Ceren Turan[1], Emre Dünder[2], Mehmet Ali Cengiz[3]

Article History
Received: 08 Dec 2021
Accepted: 28 Dec 2021
Published: 31 Dec 2021
10.53570/jnt.1034246
Research Article

Abstract – Cities have proliferated and experienced increasing environmental issues in the modern world. The concept of environmental sustainability is one of the main problems to solve. Therefore, it is fundamental to establish statistical methods to measure environmental sustainability. The first aim of this study is to measure the environmental sustainability performance of 42 cities in Turkey by Data Envelopment Analysis. The second aim is to solve the imbalance in the efficiency values obtained using Synthetic Minority Oversampling Technique methods. After all, we expose the multiple relationships between input and output variables and efficiency using the Decision trees classifiers approach. As a result of the analyses, three internal factors were found to influence the environmental efficiency levels: residential sales, population intensity, and the number of completed industrial sites. It has been determined that the number of completed industrial sites and the increase in residential sales distorted environmental efficiency.

Keywords – *Classification, decision trees, DEA, environmental sustainability, imbalanced data, SMOTE*

Mathematics Subject Classification (2020) – 62H30, 62P99

1. Introduction

The world has become a situation in which natural resources are depleted rapidly. Besides, it has become difficult to compensate for depleted resources. Factors such as an increase in consumption and population, excessive consumption of natural resources, unplanned urbanization, and industrialization have caused environmental problems to reach serious levels. As a result of all this, an increase in water, soil, and air pollution occurs. Thus, problems arise that threaten the health of living things. This situation, which impacts the lives of all living things, has led people to develop solutions. In this context, the concept of environmental sustainability is tried to be expanded.

In the simplest terms, the concept of sustainability can be defined as developing without harming resources to transfer them to future generations while making use of today's resources. With environmental issues gaining a global dimension, states that have entered the path of developing policies based on the concept of environmental sustainability have signed international protocols and agreements. Besides, the number of established non-governmental organizations has increased, and various studies have been initiated to protect the environment. The United Nations Environment Conference in Stockholm in 1972, the United Nations

[1]scturan1@gmail.com (Corresponding Author); [2]emre.dunder@omu.edu.tr; [3]macengiz@omu.edu.tr
[1,2,3] Department of Statistics, Faculty of Arts and Sciences, Ondokuz Mayıs University, Samsun, Turkey

Environment and Development Conference in 1992, and the Sustainable Development Summit in 2002 are global conferences on environmental issues. In these conferences, political approaches were handled to realize sustainable development, which prevents environmental issues.

Countries and organizations use different methodologies to evaluate the performance of their policies to improve the quality of life and reduce pollution sustainably. Data Envelopment Analysis (DEA) is one of those methods. DEA is a non-parametric statistical method used to define the effectiveness limits of decision-making units (DMUs) with more than input and output. There are some studies about environmental sustainability performance comparisons such as Marshall and Shortle [1], Siong and Hussein [2], Yu and Wen [3], Yoshino et al. [4], Xiaoping et al. [5]. However, all these and similar studies are only on environmental performance measurement.

The imbalance is one of the problems that can be encountered in datasets. It can be defined as the number of observations belonging to classes is not equal. While working with such datasets, the near-perfect accuracy values in the analysis results do not mean that the model is very successful. Therefore, while performing performance evaluation, considering measures such as sensitivity, determination, negative/positive predictive value will also provide a healthier performance evaluation. To improve this gap in performance values, methods such as oversampling, under-sampling, and Synthetic Minority Oversampling Technique (SMOTE) are used to change the sample structure in the dataset. The concept of environmental sustainability has become an important part of scientific studies day by day. However, in the studies in the literature, it has been determined that there is no study that deals with the situation of encountering the imbalance problem in the data sets used. Adebayo et al. [6] investigated coal consumption and environmental sustainability in South Africa by examining the role of financial development and globalization by using a dataset covering the period from 1980 to 2017. Kimhombo et al. [7] investigated by panel analysis whether there is a trade-off between financial globalization, economic growth, and environmental sustainability. Khan et al. [8] investigated how environmental technology contributes to wastewater improvement in 16 selected OECD countries during 2000–2019.

There are some studies for examining the performance analysis of classifiers on imbalanced datasets in the literature. Akbani et al. [9] investigated how Support Vector Machines (SVM) should be implemented in such datasets. They investigated the cause of the classic SVM classifier's failures and proposed the SMOTE method for high success. It was observed that they achieved higher success with the method they proposed compared to classical methods. In a study conducted by He [10], the most appropriate approaches and performance analysis criteria for imbalanced datasets were examined. They identified what vital areas are encountered with imbalanced datasets. They have determined which vital areas are encountered with imbalanced datasets. They also examined the statistical structures of datasets and mentioned what algorithms the information discovery could take place in such situations.

Inspired by all this information, this study includes an empirical application on DEA related to the environmental sustainability concept. This study consists of two stages. The first stage of this study is the environmental sustainability performance of 42 cities in Turkey is evaluated using the DEA method. For this purpose, input and output variables were selected among the environmental sustainability indicators used in the study conducted by Yu and Wen [3]. Then, it aims to balance the data by using SMOTE and some other types of SMOTE methods, which is one of the resampling methods to solve the imbalance problem in the obtained performance values. The second stage of this study ensures that the relations between the environmental sustainability performance of 42 cities in Turkey and other factors are determined by the Decision Trees classifier. Unlike past studies, this study investigates the solve the imbalance in the efficiency values obtained using SMOTE methods.

The rest of our study includes four more main sections. In the next section, data envelopment analysis, transforming imbalanced data into the balanced dataset, decision trees classifier, and performance criteria are included. While the findings are included in the third section, some comments and studies can be made in the fourth section.

2. Preliminary

2.1. Data Envelopment Analysis

It is of great importance for countries and institutions to know how much production can increase by increasing efficiency due to the rational use of resources. In this context, efficiency measurements of units have become a critical necessity today. DEA can be defined as a linear programming-based method that produces consistent results in the presence of many inputs and outputs. It is used to evaluate their relative effectiveness. Efficiency measurement in this method is carried out as the weighted sum of the outputs divided by the weighted sum of the inputs. In classical DEA models, the unit with the highest efficiency in the observation set is determined by considering the input and output variables of homogeneous DMUs. Then, an efficiency limit is created according to this unit. Efficiency rankings are performed by calculating the distances of other DMUs according to this limit. In this way, it is possible to mathematically interpret how ineffective DMUs can increase or decrease the input/output levels to be effective and which decision points can be used as a reference. The basic assumption in these models is that all DMUs have similar strategic objectives, and in this context, the same type of output is produced using the same kind of input.

One of the most used models of DEA is the model described as the CCR model developed by Charnes et al. [11] based on the assumption of a continuous return to scale (CRS). The second model is the model known as the BCC model, which was expanded by Banker et al. [12] with the assumption of variable returns to scale, see [13]. The model selection for input-oriented or output-oriented is discussed while using CCR and BCC models in DEA studies [1,14]. In the technical efficiency measurement made using the input-oriented efficiency, the minimum input level required to produce the fixed output amount is tried to be obtained. The input oriented CCR model obtained from Charnes et al. [15] is as follows:

$$E_k = \min\theta - \varepsilon\left(\sum_{i=1}^{m} S_i^- + \sum_{r=1}^{s} S_r^+\right)$$

constraints,

$$\sum_{j=1}^{n} x_{ij}\,\lambda_j - \theta x_{ik} + S_i^- = 0, i = 1,2,\dots,m$$

$$\sum_{j=1}^{n} y_{rj}\,\lambda_j - y_{rk} - S_r^+ = 0, r = 1,2,\dots,s$$

$$\lambda\_j, S\_i\text{^}-, S\_r\text{^}+ \geq 0;\ r = 1,2,\dots,s; i = 1,2,\dots,m; j = 1,2,\dots,n \quad (1)$$

Here, E_k is k[th] DMU's efficiency value. *i* and *r*, respectively, is the number of inputs and outputs. *j* indicates the number of DMU. θ represents how much the input amount can be reduced without modifying the output and λ_j is the variable used in the determination of the reference set. Decision units with $\lambda_j > 0$ are called effective, and these effective units are obtained as a reference set for ineffective decision units. S_i^- and S_r^+ show the excess in the input variables and the lack of the output variables, respectively. In the case that $\theta = 1$ and $S_i^- = S_r^+ = 0$ for each input and output variable, DMU is considered effective. In this study, an output oriented CCR model was used. The output oriented CCR model aims to maximize the output value without the need for more than the current input values. The output oriented CCR model obtained from Cooper et al. [16] is as follows:

$$E_k = \max\gamma - \varepsilon\left(\sum_{i=1}^{m} S_i^- + \sum_{r=1}^{s} S_r^+\right)$$

constraints,

$$\sum_{j=1}^{n} x_{ij}\,\beta_j - x_{ik} + S_i^- = 0, i = 1,2,\dots,m$$

$$\sum_{j=1}^{n} y_{rj}\,\beta_j - \gamma y_{rk} - S_r^+ = 0, r = 1,2,\dots,s$$

$$\beta\_j, S\_i\text{^}-, S\_r\text{^}+ \geq 0; r = 1,2,\dots,s; i = 1,2,\dots,m;\ j = 1,2,\dots,n \quad (2)$$

Here, γ represents the expansion coefficient that determines how radially the output of decision-makers can be increased. β_j is the variable used in the determination of the reference set.

2.2. Transforming Imbalanced Data into Balanced Datasets

While making the classification, it is tried to be guessed which units will be included in different classes or groups depending on some input variables. Samples of courses in a dataset can have an imbalanced distribution and are frequently encountered with such datasets. Datasets such as fraud detection and delayed invoice estimation are widely known examples of imbalanced datasets.

Applying methods such as increasing the sample rate by collecting more samples, trying different machine learning algorithms, changing class weights, and punishing models help to make the imbalanced datasets more balanced. In addition, there are many suggested methods in the literature to eliminate the effect of imbalanced datasets on classification. Under-sampling and oversampling methods are the most used of these methods. Under-sampling can be defined as approximating the number of majority classes to the number of minority classes in the dataset. In oversampling methods, the number of observations of the minority class increases, or the number of observations of the majority class is reduced. Thus, it is made to achieve balance. The most important disadvantages are that the under-sampling method causes loss of information by decreasing the number of samples in the dataset.

In contrast, the oversampling method may increase the number of samples and cause an overfitting problem. On the other hand, the under-sampling method reduces training time significantly with this sample reduction in the dataset and significantly saves memory. An increase in training time is observed since the oversampling method greatly increases the size of the data set. Also, the memory used occupies a considerable amount of space.

Synthetic Minority Oversampling Technique (SMOTE), which is based on the principle of operation of the oversampling method, is among the recommended methods to eliminate the effect of imbalanced datasets on classification. In this study, SMOTE algorithms are focused on. SMOTE method selects the most recent neighbours by applying the K-nearest neighbours (KNN) algorithm and then combines them. Thus, synthetically replicates the class type where the available imbalanced data distribution. The algorithm calculates distances between vectors using feature vectors and their closest neighbours. These differences are multiplied by the random number between 0 and 1 and added back to the dataset. Thus, the data becomes balanced [17]. The SMOTE algorithm is a pioneering algorithm for algorithms such as ADAptive SYNthetic (ADASYN), Density-Based SMOTE (DBSMOTE), Relocating Safe-level SMOTE (RSLS).

2.3. Decisions Trees

Tree-based learning algorithms are one of the most used supervised learning algorithms. This is because they can generally be adapted to the solution of all classification and regression problems. Decision trees are a basic data mining classification algorithm used for classifying data. It is used in many different datasets due to its ease of creating and interpreting results, transparency, and sensitivity to noisy data. The decision tree is used to break data sets containing many records into smaller sets by applying decision rules. That is, a tree structure is taken as an example in this method. In this structure, while going from the stem to the leaves, there is a

query in each node and the answers given to the queries connected to the nodes. Each query process starts from the body of the decision tree and repeats towards the leaves recursively [18]. There are class labels on the leaves of the tree. Commonly known decision tree classification methods are C4.5, Iterative Dichotomiser 3 (ID3), and Classification and Regression Trees (CART).

2.4. Performance Measurement Criteria

In classification applications, consistency criteria are considered to measure model performances in general. But the consistency criterion may show bias towards the majority class in imbalanced classification problems. For this reason, the F-score, G-mean, and the area under the ROC curve (AUC) criteria, which are more suitable for measuring model performances, were used in this study.

3. Results and Discussion

In the application part, we implemented DEA on 42 different cities existing in Turkey to measure the environmental efficiency levels. Water consumption, electricity consumption, practising nurses, fuel consumption and total environmental public expenditure, are used as input variables, while $(\mathrm{S02})^{-1}$, $(\mathrm{PM10})^{-1}$, $(\mathrm{Water\ waste})^{-1}$, and $(\mathrm{Solid\ Waste})^{-1}$ are used as an output variable for the DEA approach in the study. These input and output variables were selected among the variables used in environmental sustainability analysis in the literature. The data sets were taken from the Turkey Statistical Institute (TURKSTAT) (https://data.tuik.gov.tr/) and the Ministry of Environment and Urbanization (https://csb.gov.tr/).

We obtained the efficiency values in the first phase and then carried out the decision tree algorithm by handling the class-imbalanced problem. We selected one of the most popular CART algorithms. Following this way, we conducted a two-stage approach to identify the potential internal factors on the efficiencies. Statistical analyses in this study were carried out using MaxDEA 8 Basic (available at http://maxdea.com/MaxDEA.htm) and R-Project softwares [19]. We utilized Benchmark [20] and rpart [21] packages existing in R-Project.

Table 1 shows the definitions of the inputs, outputs, and internal factors considered in the analysis part.

Table 1. Definition of the variables

Type	Variables	Explanation
Inputs	i_1	Water Consumption
	i_2	Electricity consumption
	i_3	Fuel consumption
	i_4	Total environmental public expenditure
Outputs	o_1	$(\mathrm{S02})^{-1}$
	o_2	$(\mathrm{PM10})^{-1}$
	o_3	$(\mathrm{Water\ waste})^{-1}$
	o_4	$(\mathrm{Solid\ Waste})^{-1}$
Internal factors	x_1	Residential sales
	x_2	Population in the cities
	x_3	Population density
	x_4	Farming areas
	x_5	Number of completed industrial sites
	x_6	Number of cars (per thousand)
	x_7	Number of motor vehicles by cities
	x_8	GDP per capita ($)

Table 2 denotes the descriptive statistics of the inputs, outputs, and internal factors. The mean, standard deviation, maximum and minimum values are given for each variable.

Table 2. Descriptive statistics of the variables

Type	V	Mean	SD	Min	Max
Inputs	i_1	74840.595	144731.774	4633.000	931885.000
	i_2	2070006.167	4125846.530	69057.000	25304170.000
	i_3	239729.690	445398.485	18360.000	2824675.000
	i_4	135730601.690	384771460.557	6672868.000	2450908325.000
Outputs	o_1	0.070	0.044	0.007	0.167
	o_2	0.016	0.006	0.008	0.033
	o_3	0.002	0.006	0.000	0.033
	o_4	0.000	0.000	0.000	0.000
Internal factors	x_1	9383.476	24208.458	146.000	153897.000
	x_2	1159361.643	2056903.741	74412.000	13255685.000
	x_3	153.931	385.930	19.360	2551.133
	x_4	3233593.714	2447038.310	143688.000	12591457.000
	x_5	1416.595	1065.354	26.000	5147.000
	x_6	81.429	33.517	18.000	147.000
	x_7	236638.262	448382.847	9247.000	2794236.000
	x_8	18975527.071	52964266.728	726575.000	343536128.000

V: Variables, **SD**: Standard deviation, **Min**: Minimum, **Max**: Maximum

As seen from Table 1, we used four inputs and four outputs with the complete data set. Because of the opposite nature of the data, we used the reciprocal of the outputs. The sample size assumption is satisfied since the sample size is relatively large, at least three times higher ($n = 42 > p = 24$).

Table 3 reports the environmental efficiency values of all the cities.

Table 3. The efficiency values of the cities

City	Efficiency	Decision
Adana	0.182	Not efficient
Adıyaman	0.418	Not efficient
Afyonkarahisar	0.146	Not efficient
Antalya	0.037	Not efficient
Balıkesir	0.162	Not efficient
Bayburt	1.000	**Efficient**
Bilecik	1.000	**Efficient**
Bitlis	0.704	Not efficient
Burdur	0.661	Not efficient
Bursa	0.075	Not efficient
Çanakkale	0.382	Not efficient
Çorum	0.201	Not efficient
Diyarbakır	0.152	Not efficient
Edirne	0.261	Not efficient
Elazığ	0.207	Not efficient
Erzincan	0.419	Not efficient
Eskişehir	0.616	Not efficient
Gaziantep	0.066	Not efficient
Giresun	0.653	Not efficient
Hatay	0.215	Not efficient
Isparta	0.275	Not efficient
İstanbul	0.021	Not efficient
İzmir	0.033	Not efficient

Table 4. (Continued) The efficiency values of the cities

City	Efficiency	Decision
Kars	1.000	**Efficient**
Kırklareli	1.000	**Efficient**
Kütahya	0.263	Not efficient
Malatya	0.174	Not efficient
Manisa	0.247	Not efficient
Ordu	0.280	Not efficient
Osmaniye	1.000	**Efficient**
Rize	0.553	Not efficient
Samsun	0.214	Not efficient
Siirt	0.369	Not efficient
Sivas	0.663	Not efficient
Şanlıurfa	0.198	Not efficient
Tekirdağ	0.089	Not efficient
Tokat	0.354	Not efficient
Trabzon	0.224	Not efficient
Uşak	0.515	Not efficient
Van	0.276	Not efficient
Yalova	1.000	**Efficient**
Yozgat	0.462	Not efficient

According to table 3, only a few cities (seven cities) are efficient in terms of environmental sustainability, and the ratio of the efficient units is 14.3%.

We labelled the units as efficient and not efficient and implemented the decision tree algorithms using eight internal factors in the second stage. We have two purposes of conducting decision trees: 1) To determine the important variables that affect environmental efficiency 2) To construct reasonable rules for proposing further suggestions.

However, the data mining algorithms are rather sensitive to the distribution of the response variable, especially in the presence of a class imbalanced problem. Our DEA results clearly point out the class-imbalanced problem because of the ratio of the efficient units. To overcome this difficulty, we applied four different oversampling techniques such as SMOTE, DBSMOTE, RSLS, and ADASYN. We checked three performance metrics: AUC, F-score, and G-mean of the decision tree results.

Table 4 presents the results of the AUC, F-score and G-mean for each oversampling method and when there is no oversampling.

Table 5. The results of the performance metrics for oversampling methods

Method	AUC	G-mean	F-score
No oversampling	0.500	0.000	0.000
SMOTE	**0.990**	0.039	0.028
DBSMOTE	0.975	**0.060**	0.031
RSLS	0.938	0.051	**0.036**
ADASYN	**0.990**	0.039	0.028

According to the metrics, the oversampling methods obviously improve the performance of the decision tree algorithms. When compared with the raw efficiency data, the over-sampled data sets overcome the class-imbalanced problem by increasing AUC, G-mean, and F-scores. However, there is no absolute discrimination among oversampling methods in achievement. Because of that reason, we interpret the decision tree results and struggle to make inferences on general findings.

Table 5 shows the variable selection results for each oversampling method.

Table 6. The selected variables by decision tree algorithm

Oversampling method	x_1	x_2	x_3	x_4	x_5	x_6	x_7	x_8
No oversampling	NS	NS	NS	NS	NS	NS	NS	NS
SMOTE	+	NS	+	NS	+	NS	NS	NS
DBSMOTE	+	NS	NS	NS	+	NS	NS	NS
RSLS	+	NS	NS	NS	NS	NS	NS	NS
ADASYN	+	NS	+	NS	+	NS	NS	NS

NS: Not selected

The "+" symbol in Table 5 indicates that the relevant variables were selected. When the class-imbalanced problem was not solved, i.e., the decision tree algorithm did not select any variables when the oversampling was not applied. The variable of residential sales (x_1) was commonly selected by all the algorithms. Besides, the variables of the population intensity (x_3) and the number of completed industrial sites (x_5) was selected by the other algorithms. We can observe that the residential sales, population intensity, and completed industrial sites significantly affect the environmental efficiency levels.

Table 6 reports the extracted rules and the ratios from the decision tree algorithms with the oversampling methods.

Table 7. The extracted rules from decision trees

Tree	**Rules**	**Inference**	**Ratio**
SMOTE	1	If x_5>=713 then inefficient	42
+	2	If x_5<713 and x_3 ≥83 then efficient	31
CART	3	If x_5<713 and x_3<83 and x_1<883 then efficient	18
	4	If x_5<713 and x_3<83 and x_1 ≥883 then efficient	10
DBSMOTE +	1	If x_5 ≥713 then inefficient	45
CART	2	If x_5<713 and x_1<1241 then efficient	42
	3	If x_5<713 and x_1 ≥1241 then inefficient	13
RSLS + CART	1	If x_1<972 then efficient	65
	2	If x_1 ≥972 then inefficient	35
ADASYN	1	If x_5 ≥713 then inefficient	41
+	2	If x_5<713 and x_3 ≥84 then efficient	32
CART	3	If x_5<713 and x_3<84 and x_1<782 then efficient	18
	4	If x_5<713 and x_3<84 and x_1 ≥782 then efficient	10

According to the rules, higher residential sales leads to environmental inefficiency for all the cases. Mostly, the rules point out the environmental inefficiency when the number of completed industrial sites becomes high, singly. The environmental efficiency is observed simultaneously when residential sales are relatively lower, and the population intensity is more elevated. Even though the threshold values differ among the trees, they are rather similar. Also, the rules containing the number of completed industrial sites have the highest frequency in general.

4. Conclusion

It is crucial to provide efficient management in environmental issues in worldwide. Due to the importance of this topic, we should understand the main determinants and possible effects of these determinants that may affect environmental efficiency. Within this purpose, we attempted to identify the related factors on the environmental efficiency levels of Turkish cities. We followed a three-way approach using DEA, oversampling methods, and decision tree algorithms to reach our purpose. First, we obtained the efficiencies of the cities and then applied one of the decision trees, the CART algorithm, with oversampling procedures.

We utilized the capacity of data mining with the CART algorithm for selecting the most relevant factors and making further inferences with the extracted rule sets. The imbalance of the distribution of the efficiency levels led us to use oversampling methods. Also, the use of oversampling methods proved the rightness of our way since the raw efficiency data did not produce any rules. One of the most benefits of this paper is to demonstrate how reasonable results can be obtained by handling the class-imbalanced problem in second stage DEA.

According to the results, three internal factors were found to influence the environmental efficiency levels: residential sales, population intensity, and the number of completed industrial sites. Mainly, the increment of the residential sales and the number of completed industrial sites distort environmental efficiency. However, environmental efficiency can be ensured, even in high population intensity, when the residential sales are relatively low.

Our findings give insight into the improvement of the environmental efficiency process. We propose to decrease the residential and industrial buildings, together with the population intensity

Author Contributions

S.C.T: designing the study, providing data and statistical analyses, and writing the manuscript. E.D: statistical analyses and interpreting the results. M.A.C: literature research and discussed the results.

Conflict of Interest

The authors declare no conflict of interest.

References

[1] E. Marshall, J. Shortle, *Using DEA and VEA to evaluate quality of life in the Mid-Atlantic States*, Agricultural and Resource Economics Review 34(2) (2005) 185–203.

[2] H. C. Siong, M. Z. S. M Hussein, *Modeling Urban Quality of Life with Data Envelopment Analysis Methods*, Research Result Report, *Universiti Teknologi Malaysia*, VOT78513, 2008.

[3] Y. Yu, Z. Wen, *Evaluating China's Urban Environmental Sustainability with Data Envelopment Analysis*, Ecological Economics (69) (2010) 1748–1755.

[4] D. Yoshino, A. Fujiwara, J. Zhang, *Environmental Efficiency Model Based on Data Envelopment Analysis and Its Application to Environmentally Sustainable Transport Policies*, Transportation Research Record 2163(1) (2010) 112–123.

[5] Z. Xiaoping, L. Yuanfang, W. Wenjia, *Evaluation of Urban Resource and Environmental Efficiency in China Based on The DEA Model*, Journal of Resources and Ecology 5(1) (2014) 11–19.

[6] T. S. Adebayo, D. Kirikkaleli, I. Adeshola, D. Oluwajana, G. D. Akinsola, O. S. Osemeahon, *Coal Consumption and Environmental Sustainability in South Africa: The Role of Financial Development and Globalization*, International Journal of Renewable Energy Development 10(3) (2021) 527–536.

[7] S. Kihombo, A. I. Vaseer, Z. Ahmed, S. Chen, D. Kirikkaleli, T. S. Adebayo, *Is There a Trade-Off Between Financial Globalization, Economic Growth, and Environmental Sustainability? An Advanced panel Analysis*, Environmental Science and Pollution Research (2021) 1–11.

[8] S. A. R. Khan, P. Ponce, Z. Yu, H. Golpîra, M. Mathew, *Environmental Technology and Wastewater Treatment: Strategies to Achieve Environmental Sustainability*, Chemosphere 286 (2022) 131532.

[9] R. Akbani, S. Kwek, N. Japkowicz, *Applying Support Vector Machines to Imbalanced Datasets*, Machine Learning: ECML of the series Lecture Notes in Computer Science 3201 (2004) 39–50.

[10] H. He, *Learning from Imbalanced Data*, IEEE Transactions on Knowledge and Data Engineering 21(9) (2009) 1263–1284.

[11] A. Charnes, W. W. Cooper, E. Rhodes, *Measuring the Efficiency of Decision-Making Units*, European Journal of Operations Research 2(6) (1978) 429–444.

[12] R. D. Banker, A. Charnes, W. Cooper, *Some Models for Estimating Technical and Scale Inefficiencies in Data Envelopment Analysis*, Management Science 30(9) (1984) 1078–1092.

[13] T. Coelli, *A Guide to DEAP Version 2.1: A Data Envelopment Analysis (Computer) Program*, Centre for Efficiency and Productivity Analysis, University of New England, Australia 96(08) (1996) 1–49.

[14] R. Fare, S. Grosskopf, *Modeling Undesirable Factors in Efficiency Evaluation: Comment*, European Journal of Operational Research 157(1) (2004) 242–245.

[15] A. Charnes, W. Cooper, A. Y. Lewin, L. M. Seiford, *Data Envelopment Analysis Theory, Methodology and Applications*, Journal of the Operational Research Society 48(3) (1997) 332–333.

[16] W. W. Cooper, L. M. Seiford, K. Tone, *Data Envelopment Analysis: A Comprehensive Text with Models, Applications, References and DEA-Solver Software*, Journal of the Operational Research Society 52(12) (2001) 1408–1409.

[17] S. Kotsiantis, D. Kanellopoulos, P. Pintelas, *Handling Imbalanced Datasets: A Review*, GESTS International Transactions on Computer Science and Engineering 30(1) (2006) 25–36.

[18] R. S. Mitchell, J. G. Michalski, T. M. Carbonell, *An Artificial Intelligence Approach*, Berlin, Germany, Springer, 2013.

[19] R Core Team (2021). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria, https://www.R-project.org/

[20] P. Bogetoft, L. Otto, Benchmarking with Dea, Sfa, and r (Vol. 157). Springer Science & Business Media, 2010.

[21] T. Therneau, B. Atkinson, (2019). rpart: Recursive Partitioning and Regression Trees. R package version 4.1-15. https://CRAN.R-project.org/package=rpart