# Combination of machine learning and data envelopment analysis to measure the efficiency of tax service office (#104017)

Second revision

## Guidance from your Editor

Please submit by **18 Nov 2024** for the benefit of the authors .

**Structure and Criteria**
Please read the 'Structure and Criteria' page for guidance.

**Raw data check**
Review the raw data.

**Image check**
Check that figures and images have not been inappropriately manipulated.

If this article is published your review will be made public. You can choose whether to sign your review. If uploading a PDF please remove any identifiable information (if you want to remain anonymous).

### Files

Download and review all files from the materials page.

1 Tracked changes manuscript(s)
1 Rebuttal letter(s)
9 Figure file(s)
2 Latex file(s)
14 Table file(s)
16 Raw data file(s)
1 Other file(s)

# Structure and Criteria

## Structure your review

The review form is divided into 5 sections. Please consider these when composing your review:

1. **BASIC REPORTING**
2. **EXPERIMENTAL DESIGN**
3. **VALIDITY OF THE FINDINGS**
4. General comments
5. Confidential notes to the editor

🗋 You can also annotate this PDF and upload it as part of your review

When ready submit online.

## Editorial Criteria

Use these criteria points to structure your review. The full detailed editorial criteria is on your guidance page.

### BASIC REPORTING

- ✓ Clear, unambiguous, professional English language used throughout.
- ✓ Intro & background to show context. Literature well referenced & relevant.
- ✓ Structure conforms to PeerJ standards, discipline norm, or improved for clarity.
- ✓ Figures are relevant, high quality, well labelled & described.
- ✓ Raw data supplied (see PeerJ policy).

### EXPERIMENTAL DESIGN

- ✓ Original primary research within Scope of the journal.
- ✓ Research question well defined, relevant & meaningful. It is stated how the research fills an identified knowledge gap.
- ✓ Rigorous investigation performed to a high technical & ethical standard.
- ✓ Methods described with sufficient detail & information to replicate.

### VALIDITY OF THE FINDINGS

- ℹ **Impact and novelty is not assessed.** Meaningful replication encouraged where rationale & benefit to literature is clearly stated.
- ✓ All underlying data have been provided; they are robust, statistically sound, & controlled.
- ✓ Conclusions are well stated, linked to original research question & limited to supporting results.

# Standout reviewing tips

The best reviewers use these techniques

| Tip | *Example* |
|---|---|
| **Support criticisms with evidence from the text or from other sources** | *Smith et al (J of Methodology, 2005, V3, pp 123) have shown that the analysis you use in Lines 241-250 is not the most appropriate for this situation. Please explain why you used this method.* |
| **Give specific suggestions on how to improve the manuscript** | *Your introduction needs more detail. I suggest that you improve the description at lines 57- 86 to provide more justification for your study (specifically, you should expand upon the knowledge gap being filled).* |
| **Comment on language and grammar issues** | *The English language should be improved to ensure that an international audience can clearly understand your text. Some examples where the language could be improved include lines 23, 77, 121, 128 – the current phrasing makes comprehension difficult. I suggest you have a colleague who is proficient in English and familiar with the subject matter review your manuscript, or contact a professional editing service.* |
| **Organize by importance of the issues, and number your points** | *1. Your most important issue*<br>*2. The next most important item*<br>*3. …*<br>*4. The least important points* |
| **Please provide constructive criticism, and avoid personal opinions** | *I thank you for providing the raw data, however your supplemental files need more descriptive metadata identifiers to be useful to future readers. Although your results are compelling, the data analysis should be improved in the following ways: AA, BB, CC* |
| **Comment on strengths (as well as weaknesses) of the manuscript** | *I commend the authors for their extensive data set, compiled over many years of detailed fieldwork. In addition, the manuscript is clearly written in professional, unambiguous language. If there is a weakness, it is in the statistical analysis (as I have noted above) which should be improved upon before Acceptance.* |

# Combination of machine learning and data envelopment analysis to measure the efficiency of tax service office

**Shofinurdin Shofinurdin** [1] , **Arif Bramantoro** [Corresp., 2] , **Ahmad A. Alzahrani** [3]

[1] Faculty of Information Technology, Universitas Budi Luhur, Jakarta, Indonesia

[2] School of Computing and Informatics, Universiti Teknologi Brunei, Bandar Seri Begawan, Brunei

[3] Faculty of Computing and Information Technology, King Abdulaziz University, Jeddah, Saudi Arabia

Corresponding Author: Arif Bramantoro
Email address: arif.bramantoro@utb.edu.bn

The Tax Service Office, a division of the Directorate General of Taxes, is responsible for providing taxation services to the public and collecting taxes. Achieving tax targets efficiently, while utilizing available resources, is crucial. To assess the performance efficiency of decision-making units (DMUs), Data Envelopment Analysis (DEA) is commonly employed. However, ensuring homogeneity among the DMUs is often necessary and requires the application of machine learning clustering techniques. In this study, we propose a three-stage approach: Clustering, DEA, and Regression, to measure the efficiency of all tax service office units. Real datasets from Indonesian tax service office units are used, with confidentiality strictly maintained. Unlike previous studies that considered both input and output variables, we focus solely on clustering input variables, as it leads to more objective efficiency values when combining the results from each cluster. The results revealed three clusters with a silhouette score of 0.304 and Davies Bouldin Index of 1.119, demonstrating the effectiveness of Fuzzy C-Means clustering. Out of 352 DMUs, 225 or approximately 64% were identified as efficient using DEA calculations. To measure the efficiency of newly added dynamic data, we propose a regression algorithm as DEA can only handle static data. The optimization of multilayer perceptrons using Genetic algorithms reduced the Mean Squared Error by about 75.75%, from 0.0144 to 0.0035. Based on our findings, the overall performance of tax service offices in Indonesia has reached an efficiency level of 64%. These results show a significant improvement over the previous study, in which only about 18% of offices were considered efficient. The main contribution of this research is the development of a comprehensive framework for evaluating and predicting tax office efficiency, offering valuable insights for performance improvements.

# Combination of Machine Learning and Data Envelopment Analysis to Measure the Efficiency of Tax Service Office

**Shofinurdin**[1], **Arif Bramantoro**[2], **and Ahmad A. Alzahrani**[3]

[1]**Faculty of Information Technology, Universitas Budi Luhur, Jakarta, Indonesia**
[2]**School of Computing and Informatics, Universiti Teknologi Brunei, Bandar Seri Begawan, Brunei Darussalam**
[3]**Faculty of Computing and Information Technology, King Abdulaziz University, Jeddah, Saudi Arabia**

Corresponding author:
Arif Bramantoro[1]

Email address: arif.bramantoro@utb.edu.bn

## ABSTRACT

The Tax Service Office, a division of the Directorate General of Taxes, is responsible for providing taxation services to the public and collecting taxes. Achieving tax targets efficiently while utilizing available resources, is crucial. To assess the performance efficiency of decision-making units (DMUs), Data Envelopment Analysis (DEA) is commonly employed. However, ensuring homogeneity among the DMUs is often necessary and requires the application of machine learning clustering techniques. In this study, we propose a three-stage approach: Clustering, DEA, and Regression, to measure the efficiency of all tax service office units. Real datasets from Indonesian tax service offices were used while maintaining strict confidentiality. Unlike previous studies that considered both input and output variables, we focus solely on clustering input variables, as it leads to more objective efficiency values when combining the results from each cluster. The results revealed three clusters with a silhouette score of 0.304 and Davies Bouldin Index of 1.119, demonstrating the effectiveness of Fuzzy C-Means clustering. Out of 352 DMUs, 225 or approximately 64% were identified as efficient using DEA calculations. To measure the efficiency of newly added dynamic data, we propose a regression algorithm as DEA can only handle static data. The optimization of multilayer perceptrons using Genetic algorithms reduced the Mean Squared Error by about 75.75%, from 0.0144 to 0.0035. Based on our findings, the overall performance of tax service offices in Indonesia has reached an efficiency level of 64%. These results show a significant improvement over the previous study, in which only about 18% of offices were considered efficient. The main contribution of this research is the development of a comprehensive framework for evaluating and predicting tax office efficiency, offering valuable insights for performance improvements.

## INTRODUCTION

The taxation sector in Indonesia plays a crucial role, being the primary contributor to state revenue. In 2022, revenue from taxation amounted to IDR 2,034.54 trillion, accounting for 77.5% of total state revenue (Ministry of Finance of the Republic of Indonesia, 2023). However, the tax ratio remains relatively low at 10.1% of the gross domestic product, which is lower than the average tax ratio of Asia Pacific countries (19%) and the OECD tax ratio (33.5%) (OECD, 2022). To improve the tax ratio, the Indonesian government has consistently pursued policies aimed at increasing tax revenue, as evidenced by the annual increment in tax revenue targets. For instance, the tax target rose from IDR 1,199 trillion in 2020 to IDR 1,718 trillion in 2023 (DGT, 2021). Despite this drive for higher tax revenue, the growth of human resources in the taxation sector has not kept pace. In recent years, the number of tax employees has declined, with figures dropping from 46,607 in 2019 to 45,315 in 2022 (DGT, 2020). This results in a low ratio of employees to taxpayers (1:7,742), far below the average ratio seen in OECD member countries (1:1,657) (DGT, 2020). Addressing this situation requires the Directorate General of Taxes, responsible for tax collection, to function effectively and efficiently with existing resources. Their vision of becoming a trusted partner in national development through efficient, effective, integrity-based, and fair tax administration becomes paramount in achieving the increased revenue target (DGT, 2020).

The commonly employed method for assessing the efficiency of various institutions is Data Envelopment Analysis (DEA). This technique evaluates the efficiency of work units that utilize multiple inputs to achieve desired outcomes. DEA finds extensive application in measuring the performance of diverse entities, including banks, companies, governments, research institutions, and hospitals. It is considered a nonparametric estimation method for assessing the relative efficiency of these entities (Zhang et al., 2022). Originally introduced by Charnes, Cooper, and Rhodes in 1978, DEA has become widely recognized as a modern and valuable tool for efficiency measurement (Rostamzadeh et al., 2021). In recent years, there has been a significant upsurge in publications concerning the theory and application of DEA (Emrouznejad and Yang, 2018). Several studies have employed DEA to evaluate the efficiency of tax agencies in various regions, such as Spain (González and Rubio, 2013), OECD countries (Alm and Duncan, 2014), Brazil (De Carvalho Couy, 2015), Taiwan (Huang et al., 2022), and African countries (ATAF, 2021). In the context of Indonesia, the efficiency of tax service offices has been examined in several instances, including the East Java Regional Office (Triantoro and Subroto, 2016), all tax service offices in 2011 (Suyanto and Saksono, 2013), and 2012 (Fadhila, 2014).

In the context of DEA, an important concern is the need for homogeneity among the decision-making units (DMUs) being measured (Omrani et al., 2018), However, existing DMUs often lack this homogeneity (Razavi Hajiagha et al., 2016), necessitating a method to maintain uniformity within the DMU population. To address the issue of DMU heterogeneity, researchers have explored clustering as a solution. Some studies have combined machine learning and DEA methods to evaluate the efficiency of hospitals (Omrani et al., 2018) and banks (Razavi Hajiagha et al., 2016). These studies utilize machine learning algorithms to cluster DMUs into homogeneous groups based on input and output variables.

Another weakness of DEA is its inability to calculate or predict the efficiency value of new data (Zhang et al., 2022). Researchers have sought to overcome this limitation by integrating DEA with machine learning regression algorithms in various domains, including manufacturing companies (Zhu et al., 2021), carbon emissions (Zhang et al., 2022), and bank efficiency in China (Dalvand et al., 2014). In such studies, machine learning algorithms are used to predict the efficiency value of new dynamic data after DEA generates the static efficiency value. The static efficiency score serves as training and testing data for the regression machine learning algorithm.

To address the heterogeneity problem and the limitation of DEA in measuring dynamic efficiency against new data, researchers have undertaken studies that integrate machine learning and DEA methodologies. Some of these studies focused on clustering, measuring, and predicting the efficiency value of poultry farming companies in Iran (Rahimi and Behmanesh, 2012) and the performance of companies in the Tehran stock market (Rezaee et al., 2018). In these investigations, machine learning algorithms were employed to initially cluster the DMUs, then DEA was used to calculate static efficiency values, and regression techniques were applied to predict dynamic efficiency values for new data. However, it is important to note that these studies clustered all input and output variables together.

This paper aims to measure the efficiency value of all tax service offices in Indonesia in 2022 via DEA and machine learning through three stages. The first stage uses machine learning to cluster DMUs to overcome heterogeneity issues. In this stage, clustering has been performed only on input variables to objectively quantify the efficiency value. The second stage is measuring the static efficiency value via the DEA method. The third stage uses machine learning to predict the dynamic efficiency value of new data using regression algorithm. This research is expected to contribute to the objective evaluation of the Tax Service Office's performance. The main contribution of this research is offering a comprehensive and objective evaluation of tax office performance, providing valuable insights into current efficiency and strategies for future improvement, thereby enhancing the overall efficiency of tax services in Indonesia.

## METHODS

This research employs a comprehensive methodology consisting of three approaches to analyze the efficiency of tax offices: clustering techniques, DEA, and regression modeling. The techniques were implemented in Python 3.11, using libraries such as gurobipy for DEA analysis, scikit-learn for modeling, and common libraries like pandas, numpy, and matplotlib.

- **Clustering**: This technique is used to group data into more homogeneous clusters, effectively addressing the issue of heterogeneity within the dataset. We employ various clustering algorithms in machine learning, including Fuzzy C-Means, DBSCAN, K-Medoids, and OPTICS. To evaluate the quality of the clustering results, we utilize the Davies-Bouldin Index (DBI) and Silhouette score, which provide insights into the separation and cohesion of the clusters formed.

- **DEA**: Used to measure the relative efficiency of business units or organizations, enabling the assessment of performance in comparison to other entities. In our analysis, we use DEA-VRS with both input-oriented and output-oriented approaches.

- **Regression**: This technique is employed to predict dynamic efficiency values, which refer to the efficiency derived from new data that is not included in the historical dataset. In this analysis, we utilized Multilayer Perceptron Regressor (MLPR), Support Vector Regressor (SVR), Random Forest Regressor (RFR), and Gradient Boosting Regressor (GBR). To prevent overfitting, we applied K-Fold cross-validation, specifically using 5 folds. Additionally, hyperparameter tuning was conducted using a genetic algorithm. The model's performance was evaluated using the Mean Squared Error (MSE), providing insight into the accuracy of the predictions. Additionally, standard deviation was used as a metric to further evaluate the stability and reliability of the model predictions.

The combination of these three approaches is expected to yield a more comprehensive understanding and accurate outcomes in the analysis. This is based on our previous experience combining several approaches into a unified framework (Murakami et al., 2012).

Machine learning, as a branch of computer science, enables computers to learn from data without explicit programming (Samuel, 2000). It facilitates efficient data processing through the utilization of existing training data, allowing predictions for new data classes that have not been encountered before (Yunianta et al., 2019). By utilizing machine learning, data interpretation becomes more manageable, especially with the large volumes of data available today. Numerous industries have embraced machine learning to extract meaningful information and knowledge relevant to their activities. Machine learning relies on a diverse set of algorithms to solve various data-related problems. Data scientists understand that there is no one-size-fits-all algorithm for problem-solving. The choice of algorithm depends on factors such as the specific problem at hand, the number of variables involved, the most suitable model type, and other relevant considerations. This adaptability allows machine learning to be applied effectively to a wide range of tasks and industries.

Clustering is a fundamental technique in machine learning and data analysis that aims to group a set of objects in such a way that objects in the same group, or cluster, are more similar to each other than to those in other groups. This technique is particularly useful in exploratory data analysis, allowing researchers to discover patterns, structures, and relationships within large datasets. Clustering algorithms can be categorized into several types, including partitioning methods like K-Means, hierarchical methods, density-based methods such as DBSCAN and OPTICS, and soft clustering methods like Fuzzy C-Means. These algorithms facilitate tasks such as customer segmentation, image analysis, and anomaly detection, providing valuable insights across various domains. The effectiveness of clustering often depends on the choice of algorithm, the quality of the data, and the definition of similarity (Jain, 2010).

Fuzzy C-Means (FCM) is a non-hierarchical clustering technique within fuzzy clustering methods. It was initially introduced by Dunn in 1973 and further developed by Bezdek in 1981 (Rezaee et al., 2018). While similar to the K-Means method, FCM incorporates the concept of fuzzy theory to enhance clustering outcomes (Ye and Jin, 2016). In the FCM approach, fuzzy memberships are used, which provide membership degrees for each data point to multiple clusters (Nayak et al., 2015). The process of FCM data clustering begins with an initial estimation of the cluster center. Each data point is then assigned a certain degree of membership to each cluster. The algorithm iteratively updates the cluster centers and reassigns data points to the cluster they are closest to. This iterative process aims to minimize the objective function of the FCM method. The objective function of the FCM method can be represented by the following equation (Bezdek et al., 1984):

$$J_m = \sum_{i=1}^{n} \sum_{j=1}^{c} u_{ij}^{m} \|x_i - v_j\|^2$$

where $J_m$ represents the objective function to be minimized, $u_{ij}$ denotes the degree of membership of the data point $x_i$ in cluster $j$, while $m$ is the fuzziness exponent that controls the fuzziness of the membership values, with $m > 1$, $x_i$ represents the $i$-th data point, and $v_j$ is the centroid of cluster $j$, $\|x_i - v_j\|^2$ represents the squared Euclidean distance between data point $x_i$ and centroid $v_j$. The summations are carried out over all $n$ data points and $c$ clusters.

K-Medoids is a clustering algorithm designed to partition a dataset into a specified number of clusters using medoids as the cluster centers. A medoid is the most representative data point within a cluster, distinguishing it from the K-Means algorithm, which uses the centroid (the average of all points in the cluster). The K-Medoids algorithm begins by randomly selecting a set of medoids and then clusters the data points based on their proximity to these medoids. It optimizes the clustering by minimizing the total dissimilarity between the data points and their corresponding medoids. One of the key advantages of K-Medoids over K-Means is its robustness to outliers; medoids are less influenced by extreme values than centroids. This algorithm is particularly effective for clustering smaller datasets and demonstrates greater resilience to noise in the data. (Kaufman, 1990)

DBSCAN (Density Based Spatial Clustering of Applications with Noise) is ~~presented as~~ a clustering technique that groups data points based on density. The algorithm defines clusters as areas where data points are densely packed, separated by regions of lower density. DBSCAN categorizes points into core points, which have enough neighboring points within a specified distance (Eps); border points, which are close to core points but lack sufficient neighbors to be considered core themselves; and noise points, which do not belong to any cluster. This approach allows DBSCAN to effectively discover clusters of arbitrary shapes, manage noisy data, and eliminate the need to specify the number of clusters in advance, making it highly suitable for large-scale spatial data.(Ester et al., 1996).

OPTICS (Ordering Points To Identify the Clustering Structure) is also a density-based algorithm but is designed to address some of the limitations of DBSCAN. While DBSCAN produces distinct clusters, OPTICS generates an ordering of points that reflects the cluster structure and data density. Using the same parameters as DBSCAN, OPTICS retains information about data density and can differentiate between clusters that have varying densities. This allows OPTICS to build a hierarchy of clusters and perform better in managing data with varying densities, providing users with the flexibility to determine clusters based on different levels of density (Ester et al., 1996).

DEA is nonparametric mathematical programming that is essentially advanced linear programming based on a frontier estimation approach (Coelli, 1996). Nonparametric refers to statistical methods that do not require any parameter assumptions for the population being tested (Wolfowitz, 1949). DEA was first proposed by Charnes, Cooper, and Rhodes in 1978 (Charnes et al., 1978), is an efficiency analysis method used to measure how effectively a business unit or organization utilizes its available inputs to achieve the maximum possible output. By comparing the use of inputs and relative outputs among different units, DEA generates a relative efficiency value, allowing for a comparison of the performance of various business units or organizations. The research steps of the DEA method involve identifying the DMUs or units to be observed, along with their respective inputs and outputs. Efficiency is then calculated for each DMU, providing the input and output targets required to achieve optimal performance (Indrawati, 2009). Initially developed to evaluate non-profit and government organizations, DEA was later applied to assess the performance of service operations in various private companies (Sherman and Zhu, 2013).

The selection of appropriate input and output variables in DEA is critical, as using irrelevant variables can bias the analysis and lead to inaccurate conclusions. In this study, input variables were selected based on a thorough review of previous research on DEA's application in measuring the efficiency of tax service offices. This approach ensures alignment with the operational framework of tax offices in Indonesia. Additional input indicators, such as those proposed by Milosavljević et al. (2023), could be considered for future analyses. DEA models vary in their treatment of variable returns to scale. The two most common models, DEA-CCR (Charnes, Cooper, Rhodes) and DEA-BCC (Banker, Charnes, Cooper), offer different perspectives on efficiency evaluation depending on the specific characteristics of the analyzed units (Charnes et al., 1978) (Banker et al., 1984).

The DEA-CCR model can be customized based on output or input, and the choice of this model depends on the characteristics of DMU in the production frontier. Input-oriented models minimize the inputs for a given level of outputs, whereas output-oriented models maximize the production of outputs for a given level of inputs. Suppose there are n DMUs, and each $DMU_j(j = 1,2...n)$ produces $s$ output $y_rj(r = 1,....s)$ using m inputs $x_{ij}(i = 1,...m)$; then DEA-CCR uses the following equation to evaluate the efficiency of the DMU:

$$\max \theta$$
$$\text{subject to} \sum_{j=1}^{n} \lambda_j y_{rj} \geq \theta y_{r0}, \quad r = 1,2,...,s$$
$$\sum_{j=1}^{n} \lambda_j x_{ij} \leq x_{i0}, \quad i = 1,2,...,m$$
$$\lambda_j \geq 0, \quad j = 1,2,...,n$$

where $\theta$ represents the efficiency score to be maximized. The term $\lambda_j$ refers to the weight assigned to DMU $j$. The $r$-th output for DMU $j$ is denoted by $y_{rj}$, while $x_{ij}$ stands for the $i$-th input of DMU $j$. The values $y_{r0}$ and $x_{i0}$ represent the outputs and inputs of the DMU under evaluation, labeled as DMU 0. The total number of DMUs is $n$, the number of inputs is $m$, and $s$ is the number of outputs.

In practical applications, the original nonlinear equation of the DEA-BCC model, with infinite optimal solutions, needs to be transformed into a suitable pairwise linear programming model. This conversion ensures that the efficiency evaluation can be effectively implemented. The transformed equation takes the following form (Zhang et al., 2022):

$$\min \theta$$
$$\text{subject to} \sum_{j=1}^{n} \lambda_j x_{ij} \leq \theta x_{ik}$$
$$\sum_{j=1}^{n} \lambda_j y_{rj} \geq y_{rk}$$
$$0 < \theta \leq 1;\ \lambda \geq 0;\ i = 1,2,\ldots,m;\ r = 1,2,\ldots,q;\ j = 1,2,\ldots,n;\ k = 1,2,\ldots,s$$

where $\theta$ represents the efficiency score of the DMU under evaluation and is to be minimized, $\lambda_j$ denotes the non-negative weight assigned to DMU $j$ in the linear combination, $x_{ij}$ refers to the input of DMU $j$ for input category $i$, while $x_{ik}$ is the input of the DMU being evaluated. Similarly, $y_{rj}$ represents the output of DMU $j$ for output category $r$, and $y_{rk}$ is the output of the DMU under evaluation. Here, $k$ represents the DMU being evaluated, and $j$ denotes the other DMUs used for comparison.

The DEA-BCC model is a variant of DEA that assumes variable returns to scale. This means it assumes that the DMU is operating at an optimal scale. However, DEA-BCC also incorporates the notion of variable returns to scale, implying that changes in inputs may not result in a proportional change in outputs (Banker et al., 1984). The DEA-BCC model is particularly suitable for measuring efficiency in the public sector, where operations may not always be at an optimal scale (Kalb, 2010). It allows for a more realistic assessment of efficiency in such contexts. The key distinction between the DEA-BCC and DEA-CCR models lies in the constraints imposed on each weight $\lambda$ in the equation of the DEA-CCR model. These constraints are modified in the DEA-BCC model, resulting in the following equation (Banker et al., 1989):

$$\min \theta$$
$$\text{subject to} \sum_{j=1}^{n} \lambda_j x_{ij} \leq \theta x_{ik}$$
$$\sum_{j=1}^{n} \lambda_j y_{rj} \geq y_{rk}$$
$$\sum_{k}^{n} \lambda_k = 1$$
$$0 < \theta \leq 1;\ \lambda \geq 0;\ i = 1,2,\ldots,m; r = 1,2,\ldots,q;\ j = 1,2,\ldots,n;\ k = 1,2,\ldots,s$$

where $\theta$ represents the efficiency score of the DMU under evaluation and is to be minimized. $\lambda_j$ represents the non-negative weights assigned to each DMU $j$ in the linear combination. These weights determine how much each DMU contributes to the combination.$x_{ij}$ denotes the input $i$ used by DMU $j$, while $x_{ik}$ represents the input $i$ used by the DMU being evaluated (DMU $k$). Similarly, $y_{rj}$ represents the output $r$ produced by DMU $j$, and $y_{rk}$ is the output $r$ produced by the DMU under evaluation. Here, $k$ refers to the DMU being evaluated, and $j$ refers to the other DMUs used for comparison in the linear combination.

In the DEA-BCC model, the efficiency values obtained from the input-oriented and output-oriented approaches are different. Consider the point C, as illustrated in Figure 1. To calculate the input-oriented efficiency value at point C, we divide the distance QC1 by the distance QC. On the contrary, to calculate the output-oriented efficiency value at point C, we divide the distance NC by the distance NC2.

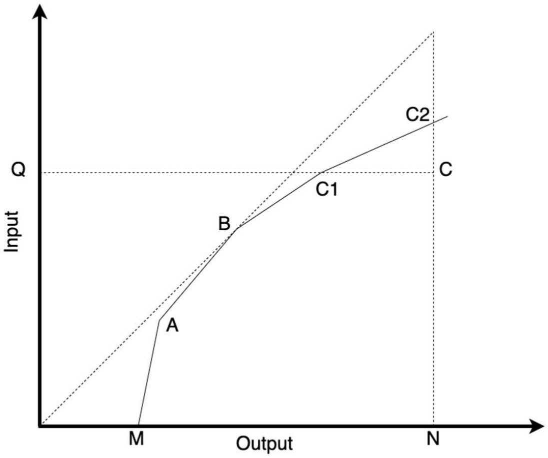

**Figure 1.** Illustration of input-oriented and output-oriented efficiency models in DEA.

The input-oriented approach in DEA-BCC focuses on efficiently using inputs to produce predetermined outputs. In input-oriented DEA-BCC, DMUs are considered units that minimize the use of inputs to produce predetermined outputs. In the input-oriented DEA-BCC model, the efficiency frontier construction technique is employed to evaluate the relative efficiency level of each DMU in utilizing their inputs. DMUs located on the efficiency frontier in input-oriented DEA-BCC are considered efficient in minimizing the usage of inputs to produce the specified outputs. These efficient DMUs serve as benchmarks for other units to strive for in terms of input utilization efficiency. The input-oriented DEA-BCC formula can be represented by the following equation (Banker et al., 1984):

$$\min \theta$$

$$\text{subject to} \sum_{j=1}^{n} \lambda_j x_{jk} \leq \theta \cdot x_{kk}$$

$$\sum_{j=1}^{n} \lambda_j y_{ji} \geq y_{ki}$$

$$\sum_{j=1}^{n} \lambda_j = 1$$

$$0 < \theta \leq 1;\ \lambda \geq 0;\ i = 1,2,\ldots,m; r = 1,2,\ldots,q;\ j = 1,2,\ldots,n;\ k = 1,2,\ldots,s$$

where $\theta$ represents the input-oriented efficiency score of DMU $k$. The weights $\lambda_j$ create a composite DMU from the inputs and outputs of other DMUs. The constraint $\sum_{j=1}^{n} \lambda_j x_{jk} \leq \theta \cdot x_{kk}$ ensures that the total input of the composite DMU does not exceed the scaled input of DMU $k$. The constraint $\sum_{j=1}^{n} \lambda_j y_{ji} \geq y_{ki}$ ensures that the output of the composite DMU is at least as large as the output of DMU $k$. The constraint $\sum_{j=1}^{n} \lambda_j = 1$ ensures that the weights sum to 1, allowing for a proportional adjustment of inputs and outputs. Finally, $\lambda_j \geq 0$ ensures that all weights are non-negative.

The output-oriented DEA-BCC approach focuses on the output produced by DMUs using predetermined inputs. In output-oriented DEA-BCC, DMUs are considered units that use specific inputs to produce the most efficient output possible. Conversely, output-oriented DEA-BCC uses efficiency frontier construction techniques to determine the relative efficiency level of each DMUs in producing their outputs. Decision-making units on the efficiency frontier in output-oriented DEA-BCC are considered efficient in utilizing available inputs to produce the maximum output. The output-oriented DEA-BCC calculation can be represented in the following equation (Banker et al., 1984):

$$\max \theta$$

$$\text{subject to} \sum_{j=1}^{n} \lambda_j y_{ji} \geq \theta \cdot y_{ki}$$

$$\sum_{j=1}^{n} \lambda_j x_{jk} \leq x_{kk}$$

$$\sum_{j=1}^{n} \lambda_j = 1$$

$$0 < \theta \leq 1;\ \lambda \geq 0;\ i = 1,2,\ldots,m; r = 1,2,\ldots,q;\ j = 1,2,\ldots,n;\ k = 1,2,\ldots,s$$

where $\theta$ represents the output-oriented efficiency score of DMU $k$. The weights $\lambda_j$ combine the outputs and inputs from other DMUs to create a virtual DMU for comparison. The constraint $\sum_{j=1}^{n} \lambda_j y_{ji} \geq \theta \cdot y_{ki}$ ensures that the combined output of the virtual DMU is at least equal to the output of DMU $k$, scaled by $\theta$. The constraint $\sum_{j=1}^{n} \lambda_j x_{jk} \leq x_{kk}$ ensures that the input used by the virtual DMU does not exceed the input used by DMU $k$. The constraint $\sum_{j=1}^{n} \lambda_j = 1$ ensures that the weights sum to 1, allowing for scaling adjustments. Finally, $\lambda_j \geq 0$ ensures that the weights are non-negative.

Regression is a statistical technique used to model and analyze the relationship between a dependent variable and one or more independent variables. Its primary purpose is to predict the value of the dependent variable based on the values of the independent variables, enabling insights into how different factors influence outcomes. Regression analysis can take various forms, including linear regression, which assumes a linear relationship, and nonlinear regression, which accommodates more complex relationships. Other advanced techniques such as multiple regression, polynomial regression, and regularized regression (like Lasso and Ridge) further enhance the ability to capture intricate patterns in data. The results of regression analysis provide valuable metrics, such as coefficients indicating the strength and direction of relationships, along with statistical tests for model validity. This technique is widely utilized in fields such as economics, finance, biology, and social sciences to make informed predictions and decisions (Montgomery et al., 2021).

Support Vector Regressor (SVR) is a machine learning algorithm derived from Support Vector Machines (SVM), primarily used for predicting continuous values in regression tasks. It finds a function that deviates from actual observed values by no more than a specified threshold (epsilon), aiming to minimize error within this margin while maintaining generalization for unseen data. SVR utilizes kernel functions to address non-linear relationships and establish complex decision boundaries, with common kernels including linear, polynomial, and radial basis function (RBF). The algorithm is effective in high-dimensional spaces and robust against overfitting, though it can be sensitive to parameter choices, such as regularization and kernel type (Smola and Schölkopf, 2004).

Random Forest Regressor (RFR) is an ensemble learning method that operates by constructing multiple decision trees during training and outputting the average prediction of these trees for regression tasks. It combines the predictions of numerous trees, which helps mitigate the overfitting often seen in individual decision trees and enhances overall model accuracy. The algorithm operates by randomly sampling subsets of data and features, ensuring diversity among the trees, which contributes to its robustness and effectiveness in capturing complex patterns in the data. One of its key advantages is the ability to handle large datasets with high dimensionality while providing insights into feature importance (Breiman, 2001).

Gradient Boosting Regressor (GBR) is an ensemble learning technique that enhances predictive performance by sequentially combining multiple weak learners, typically decision trees. The method focuses on correcting the errors made by previous trees, with each new tree added to the ensemble aimed at minimizing the residuals of the combined model from earlier iterations. This optimization is achieved through gradient descent on a specified loss function, enabling the model to capture complex relationships and feature interactions. While Gradient Boosting is effective for various regression tasks, it is sensitive to overfitting, particularly with a high number of trees, requiring careful tuning of parameters like learning rate and tree depth (Friedman, 2001).

Multilayer Perceptron (MLP) networks are among the most popular artificial neural networks used in various scientific fields, particularly in forecasting and prediction. MLP networks consist of an input layer, one or more hidden layers, and an output layer, as illustrated in Figure 2. The input layer receives a vector of data or patterns. The hidden layer spans one or more layers, receives outputs from the previous layer, assigns weights, and passes them through an activation function. In this study, we use the ReLU activation function because it only produces positive values, which aligns well with the nature of DEA that only generates positive outputs. The output layer takes the outputs from the last hidden layer, assigns weights, and potentially passes them through an output activation function to produce a target value. It can be said that a neuron's default activation function is linear. See the following equation:

$$f(x) = x$$

$$x = (-\infty, \cdots + \infty)$$

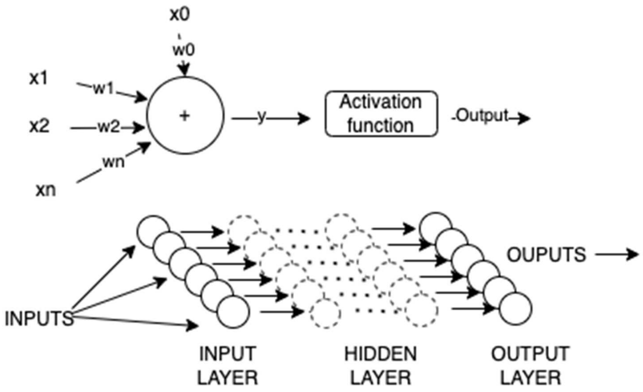

**Figure 2.** Illustration of multilayer perceptron.

K-Fold Cross-Validation is a commonly used method for evaluating machine learning models, dividing the dataset into K equal-sized folds where K-1 folds are used for training and the remaining fold is used for validation. This process is repeated K times, allowing each fold to serve as a validation set once, and the final performance is averaged across all iterations. This approach reduces bias and provides a more generalized estimate of model performance compared to a simple train-test split, as the model is trained and validated across different subsets of the data. K-Fold Cross-Validation also helps prevent overfitting by validating the model on multiple partitions of the data. Variants like Stratified K-Fold are used for handling imbalanced datasets to ensure consistent class distribution across the folds. The method was influenced by the development of resampling techniques, particularly Efron and Tibshirani's work on the bootstrap (Tibshirani and Efron, 1993), and gained wider recognition in machine learning following Kohavi's study on accuracy estimation and model selection (Kohavi, 1995).

Genetic algorithm (GA) belongs to the class of evolutionary algorithms and is inspired by natural selection cycles (Mitchell, 1998). It is a powerful optimization algorithm from the traditional heuristic family, well-suited for handling solutions trapped in local minima. In machine learning optimization, conventional algorithms like gradient descent and grid search may stop at a suboptimal solution due to the risk of getting stuck in a local minimum. However, GA can surpass these local minima and achieve globally better solutions. GA achieves this by employing selection, crossover, and mutation mechanisms to maintain variation within the population and avoid being trapped at a local minimum. It is particularly effective for problems requiring optimization within a countable system (Lambora et al., 2019). In the implementation of GA, a population of candidate solutions evolves iteratively toward a better solution. Each member of the population has a set of characteristics that can change and undergo mutation. The process begins with the initial formation of a population comprising randomly generated individuals. Each iteration, or generation, in the GA, involves calculating the fitness of each member. Fitness usually represents the value of the objective function specific to the problem being solved. Members with higher fitness are then selected from the current population, and their characteristics are combined through crossover to produce offspring with inherited characteristics (Gajić et al., 2020). The GA used in this study can be explained with the following algorithm:

1. **Initialize** the maximum number of generations and population size $n$
2. **Generate** an initial random population of $n$ solutions
3. **While** the number of generations has not reached the maximum:
    - **Evaluate** the fitness function $f(x)$ for each solution in the population
    - **Create** offspring until the desired number is reached:
        - **Select** two parent solutions from the population using the roulette wheel selection method.
        - **Apply** the crossover operator to the selected parents with probability $p$, producing two offspring.
        - **Apply** the mutation operator to the offspring with a probability equal to the mutation rate.
    - **Replace** the current population with the newly generated offspring.
4. **Terminate** when the maximum number of generations is reached or other stopping criteria are met.

There are several methods used to calculate network error, one of which is the Mean Square Error (MSE). MSE measures the average of the squared difference between the predicted value and actual value. The smaller the MSE value, the better the model predicts the data. MSE can be used if there are outliers in the observed data (Chicco et al., 2021). The MSE calculation can be represented in the following equation:

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2$$

where $n$ is number of data point, $y_i$ is actual value, and $\hat{y}_i$ is predicted value.

Additionally, standard deviation is often used to assess the dispersion of prediction errors. A low standard deviation indicates that the prediction errors are tightly clustered around the mean error, reflecting consistent model performance. The standard deviation can be calculated using the following equation (Moore and McCabe, 1989):

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

where STD is the standard deviation, $n$ is the number of data points, $y_i$ is the actual value, and $\hat{y}_i$ is the predicted value.

# RELATED STUDIES

Previous studies on efficiency measurement using DEA and machine learning have been utilized as references in this current study. These studies can be categorized into four main groups based on their focus areas: DEA used in taxation, machine learning for DMU clustering in DEA, machine learning for dynamic efficiency prediction, machine learning in DEA for clustering, and dynamic efficiency prediction.

## DEA on Tax Service Office Field

Previous studies using DEA in the field of taxation are presented in Table 1. None of these studies in the field of taxation use the clustering method. Therefore, heterogeneity is probable. In addition, it has no regression method of measuring the dynamic efficiency value of new data.

**Table 1.** DEA on tax service office field.

| Research Works | Contributions |
|---|---|
| González and Rubio (2013) | Measuring the efficiency value of tax administration performance in Spain using DEA, without clustering DMUs and predicting dynamic efficiency for new data. |
| Alm and Duncan (2014) | Determining the relative efficiency of tax agents in OECD member countries using DEA, without DMU clustering and dynamic efficiency prediction for new data. |
| De Carvalho Couy (2015) | Measuring the efficiency in tax audit performance at the Brazilian tax authority using DEA, without dynamic efficiency prediction for new data and grouping DMUs to cluster. |
| Triantoro and Subroto (2016) | Measuring the efficiency performance of tax service offices using DEA without clustering and predicting dynamic efficiency for new data. |
| Huang et al. (2017) | Measuring efficiency of tax collection and tax management in Taiwan's local tax service offices using DEA without clustering and predicting dynamic efficiency for new data. |
| Suyanto and Saksono (2013) | Analyzing the efficiency of tax service offices in Indonesia using DEA without clustering and predicting dynamic efficiency for new data. |
| ATAF (2021) | Evaluation of tax administration efficiency of African tax administration forum member countries using DEA without clustering and predicting dynamic efficiency for new data. |

### DEA and clustering

Previous studies that use machine learning on DEA to cluster DMUs are presented in Table 2. All studies using machine learning to cluster DMUs were conducted on input and output variables. In addition, these previous studies have not employed any regression method to predict the dynamic efficiency value of new data.

**Table 2.** DEA and clustering machine learning researches.

| Research Works | Contributions |
|---|---|
| Razavi Hajiagha et al. (2016) | Integrating Fuzzy C-Means and DEA to mitigate DMU heterogeneity in Banks. Clustering is performed on the input and output variables, It does not involve regression prediction using machine learning. |
| Omrani et al. (2018) | Integrating Fuzzy Clustering and DEA to find efficiency in hospitals in Iran. The input and output variables are clustered with no regression utilized. |

### DEA and prediction

Previous studies using machine learning on DEA to predict the dynamic efficiency of new data without clustering DMUs are shown in Table 3. All studies that use machine learning to predict the new data do not cluster the DMUs. Hence, DMU heterogeneity is feasible, resulting in efficiency values to be less objective.

### DEA and clustering-prediction

Previous studies that have applied machine learning methods in DEA to cluster DMUs and predict efficiency on new data are shown in Table 4. The results of these studies show that the DMU clustering stage is conducted on the input and output variables. This step may lead to a potential lack of objectivity in the efficiency assessment of the clusters formed.

## PROPOSED APPROACH

This study proposes an integrated framework by combining machine learning and DEA to measure and predict the performance efficiency of tax service offices in Indonesia as the measured DMU. This framework is divided into four processes, namely the data process, the clustering process using machine

**Table 3.** DEA and machine learning prediction researches.

| Research Works | Contributions |
|---|---|
| Dalvand et al. (2014) | Integrating C4.5 classification algorithm and DEA to predict static and dynamic efficiency values for 200 bank branches in Iran. Machine learning is solely used to predict efficiency values for new data, not to cluster DMUs to obtain DMU homogeneity. |
| Appiahene et al. (2020) | Combining DEA with three machine learning approaches to evaluate the efficiency and per formance of banks using 444 bank branches in Ghana. Only efficiency values for new data are predicted using machine learning and DMUs are not clustered to achieve DMU homogeneity. |
| Zhu et al. (2021) | Combining DEA and machine learning to measure and predict the efficiency values of manufacturing companies in China. Instead of clustering DMUs to achieve DMU homogeneity, machine learning is only employed to predict efficiency values for new data. |
| Zhang et al. (2022) | The paper specifically uses the case of China's regional carbon emission performance prediction to demonstrate the effectiveness of the proposed integrated model of DEA and machine learning. No DMUs are clustered to achieve DMU homogeneity; instead, machine learning is utilized to predict efficiency values for new data. |

**Table 4.** DEA and clustering-prediction machine learning researches.

| Research Works | Contributions |
|---|---|
| Rahimi and Behmanesh (2012) | Combining DEA and data mining techniques, such as artificial neural network, and decision tree, to predict the efficiency of poultry companies in Iran by clustering DMUs on input and ouput variables. |
| Rezaee et al. (2018) | Integrating Fuzzy C-Means, DEA, and machine learning to measure the performance of companies in the stock exchange by clustering DMU input and output variable. |

learning algorithms to obtain DMU homogeneity, the process of measuring static efficiency using DEA, and the regression process of measuring dynamic efficiency against new data, as can be observed in Figure 3.

## Data processing

The initial stage comprises data processing, commencing with the input of new historical data and generating output data, primed for further processing in the subsequent stage. The data employed in this study are sourced from the Directorate General of Taxes in Indonesia. Next, the dataset undergoes a data understanding process, aiming to ascertain its suitability for direct consumption or if specific actions are required before further processing. This analysis includes assessing the data's structure, identifying variables with negative values inappropriate for the DEA model, and detecting potential outlier data. Additionally, this stage entails separating the input and output variables. The selection of input variables is based on analyses from previous studies, which have been adjusted to fit the operations of tax service offices in Indonesia and verified by the authorities.

Input variables consist of:

- Vin1: Number of corporate taxpayers
- Vin2: Number of treasury taxpayers
- Vin3: Number of individual taxpayers
- Vin4: Number of non-employee taxpayers
- Vin5: Number of tax auditors
- Vin6: Number of account representatives

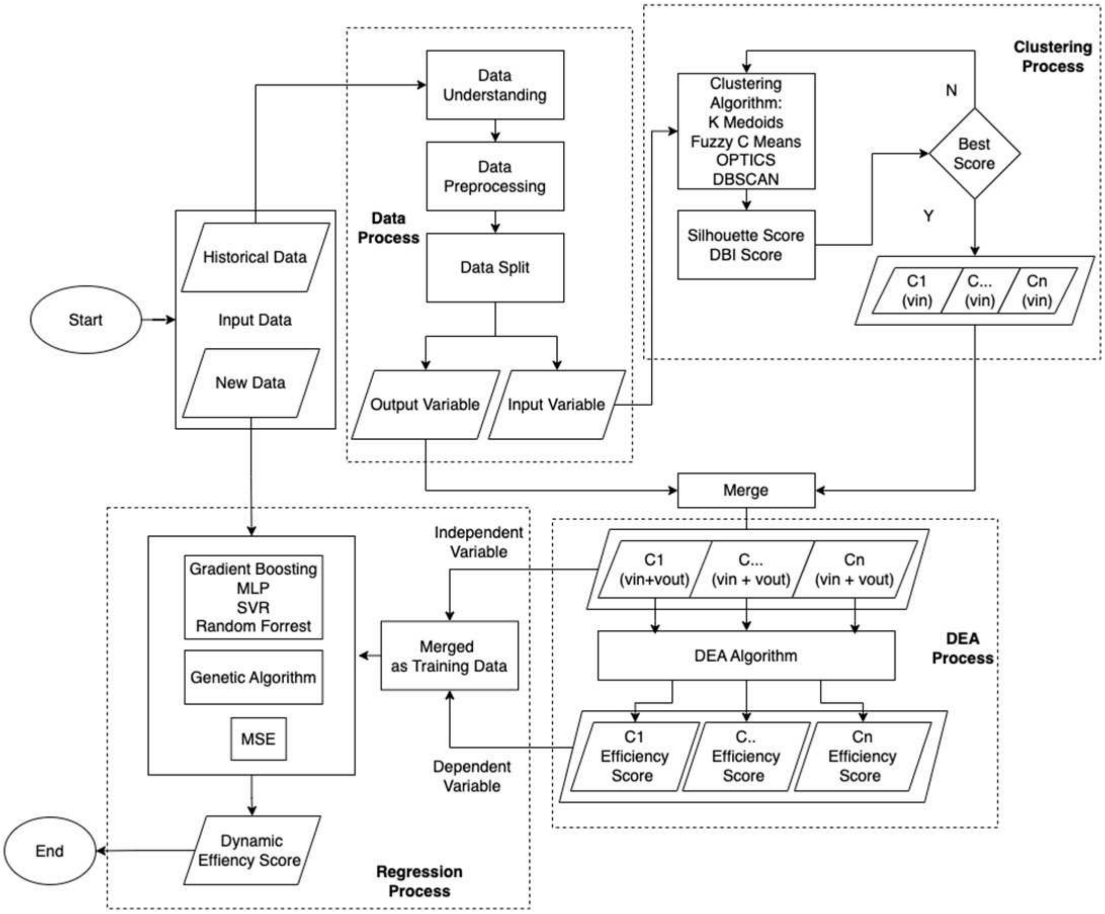

**Figure 3.** Flowchart of proposed methods.

- Vin7: Budget realization amount

Output variables consist of:

- Vout1: Compliance rate of annual tax return submission
- Vout2: Percentage of revenue achievement
- Vout3: Percentage of revenue growth achievement
- Vout4: Number of issued tax advisories
- Vout5: Number of paid tax advisories
- Vout6: Number of completed tax audits

## Clustering

The second stage involves the clustering process to categorize DMUs into several clusters for enhanced homogeneity. Machine learning algorithms like K-Medoids, FCM, DBSCAN, and OPTICS are employed for clustering, and their effectiveness is assessed using the silhouette value and Davies Bouldien Index (DBI). Higher silhouette values indicate more accurate clustering, while lower DBI values signify better cluster quality. The selection of these clustering techniques is based on previous experience using the basic K-Means algorithm for document clustering (Usino et al., 2019).

Based on preliminary experiments with ideal data, specifically data that contains four combinations of low and high input-output variables, as shown in Table A1, it is proposed to cluster the input variables alone. This approach yields a more objective efficiency value compared to clustering both input and output variables. The experiment's results, presented in Table 5, show that clustering inputs and outputs produced four clusters, maximizing efficiency results due to the relative nature of DEA. In contrast, clustering only the input variables resulted in two clusters, with maximum and minimum efficiency values.

This lack of objectivity can be observed in the combination of low input and output variables, leading to the maximum value in the input-output clustering. This situation is less objective compared to the combination of low input and high output variables, both of which produce maximum efficiency values. This situation is different from the clustering of input variables only. The combination of the low input and output variables produces a minimum efficiency value, and the combination of the low input and high output variables produces a maximum efficiency value. This phenomenon is also observed in the combination of high input and low output variables compared with the combination of high input and output variables. The clustering of input variables is more objective than the clustering of input and output variables.

**Table 5.** Clustering simulation and efficiency results.

| **Input and Output Combination** | | **Input and Output Clustering** | | **Input Clustering Only** | |
|---|---|---|---|---|---|
| Input | Ouput | Cluster | Efficiency | Cluster | Efficiency |
| Low | Low | C1 | Max | C1 | Min |
| Low | High | C2 | Max | C1 | Max |
| High | Low | C3 | Max | C2 | Min |
| High | High | C4 | Max | C2 | Max |

### DEA Process

The third stage of the process involves using DEA to calculate static efficiency values for each cluster formed in the previous clustering stage. DEA determines these static efficiency values based on historical data. Efficiency values are represented on a scale from 0 to 1, with 1 indicating maximum efficiency. DEA compares the utilization of exploiting inputs and the production of outputs among the units. Units with higher efficiency values are considered more efficient in resource utilization. The analysis helps identify units with the potential to improve their efficiency by adopting best practices from the most efficient units. At this stage, the results obtained from clustering the input variables are combined with their corresponding output variables, creating clusters consisting of both types of variables. These combined clusters are then subjected to DEA analysis to determine their static efficiency values. The DEA process utilizes the BCC method, assuming that all DMUs in the cluster have not yet reached the optimum performance level. This method employs input-oriented and output-oriented approaches, with a focus on identifying the optimal combination of inputs to produce a given output. The main objective is to measure the relative efficiency of each DMU in achieving optimal results while utilizing available resources. Using the BCC method, the DEA stage of the process can furnish information on the static efficiency level of each pre-formed cluster, taking into account the relevant input and output variables. This approach aids in comprehending the efficiency of resources for each cluster and highlights areas where improvements can be made to achieve higher levels of efficiency.

### Regression Process

The final stage is the regression process, which is the stage to predict the dynamic efficiency value. Dynamic efficiency refers to the efficiency value derived from new data that does not exist in the historical data. This stage can be used to determine the value of each input and output variable when forming a new tax service office that does not yet exist in the historical data in order to obtain the maximum efficiency value. This stage overcomes the DEA method which can only produce static efficiency values, namely the efficiency values of tax service offices that already exist in historical data and have known values of input and output variables. In addition, this regression stage can also be used to evaluate tax service offices that have not maximized their efficiency by adding or reducing the value of each input and output variable.

To achieve this, several machine learning regression algorithms are employed, including Gradient Boosting Regressor (GBR), Multilayer Perceptron Regressor (MLPR) neural network regression algorithm, Support Vector Regressor (SVR), and Random Forest Regressor (RFR). Prior to applying these regression algorithms, K-fold cross-validation is employed with K=5 to assess model performance and mitigate the risk of overfitting. This technique involves dividing the training data into five subsets, training the model on four of these subsets, and validating it on the remaining subset. This process is repeated five times, ensuring that each subset is used for validation once, which helps provide a more robust estimate of model performance. The regression process utilizes the static efficiency results obtained from the DEA stage as training data. The independent variables (x) consist of the input and output variables, while the dependent variable (y) is the static efficiency value generated in the DEA process.

Initially, the regressor models are created, and their performance is then optimized using the GA. The GA is an optimization method inspired by natural evolution principles. It is used in this context to find the best configuration of model parameters for each regressor. These parameters may include the number of trees, the depth of the tree (max depth), the number of neurons in the hidden layer (for MLPR), and the learning rate. The GA will iteratively experiment with various parameter combinations, evaluate each model's performance, and select the best configuration based on the objective function. In this study, the MSE value is used as the objective function for model evaluation. The aim is to minimize the MSE and create the most accurate regression model for predicting dynamic efficiency values. Additionally, standard deviation is monitored to assess the variability of the predictions and ensure the robustness of the final model.

## RESULT AND DISCUSSIONS

This study uses data derived from population data of tax service offices in Indonesia, with samples presented in Table A2. The data is available in two formats: Microsoft Excel and PDF. Before conducting the analysis, it is necessary to merge the data from both sources and adjust the format accordingly. The dataset comprises 352 rows and 14 columns. The DMU column contains three-digit identity codes of the tax service offices in a masked form. The details of 14 columns can be seen in Table A3. The columns comprise one “DMU” column serving as the identity and primary key, seven columns of input variables, and six columns of output variables.

Before proceeding with data processing, duplicate data detection is performed using methods such as edit distance, Jacobson, and cosine similarity. Fortunately, no duplicate data is found in the dataset, eliminating the need for duplicate data removal. Null or empty data detection is also conducted, and it is concluded that there are no null data entries in the dataset. As a result, no further steps are required for null data handling.

The descriptive statistical analysis results indicate that each attribute in the dataset exhibits a wide range, as represented by large standard deviations. Detailed information about this analysis can be found in Table A4. The next step is to perform a more comprehensive analysis to gain deeper insights into the data distribution, skewness, and the presence of outliers.

For the purpose of visualizing the data distribution, the Principal Component Analysis (PCA) method is employed. PCA transforms the data into a two-dimensional representation, allowing for easier visualization and understanding. The results are displayed in Figure 4, providing insights into the patterns and relationships between variables, aiding in comprehending the overall structure and distribution of the data. Notably, several outlier data points are identified, located far away from other data clusters. However, further analysis is required to confirm the presence of outliers at a later stage.

The skewness analysis reveals that most variables in the dataset exhibit positive values, indicating a rightward skew in their data distributions. However, one variable, “vin6,” displays a negative skewness value, suggesting a left-skewed data distribution for this particular variable. A negative skewness value means that the tail of the data distribution tends to be longer on the left side of its center value. This finding highlights that the ”vin6” variable's distribution asymmetry differs from the other variables. Therefore, when analyzing and interpreting the data, special attention should be given to the ”vin6” variable due to its distinct distribution characteristics. Complete results of the skewness values can be found in Table A5.

To ensure comparability during data analysis and modification, the normalization stage is performed using various methods, including z-score scaler, min-max scaler, and log transformation. Tables A6, A7, and A8 present the results of the normalization process. Among these methods, only the min-max scaler results in all positive values. Since DEA requires all data to be positive to achieve more robust efficiency results, the min-max scaler method is chosen for further processing (Wei and Wang, 2017).

Additionally, in the data processing stage, outlier detection is carried out to identify any outlier data. The boxplot method is used for this purpose, and the results indicate the presence of outliers in each input variable. Outliers represent significant extreme values within the data, and their identification is crucial

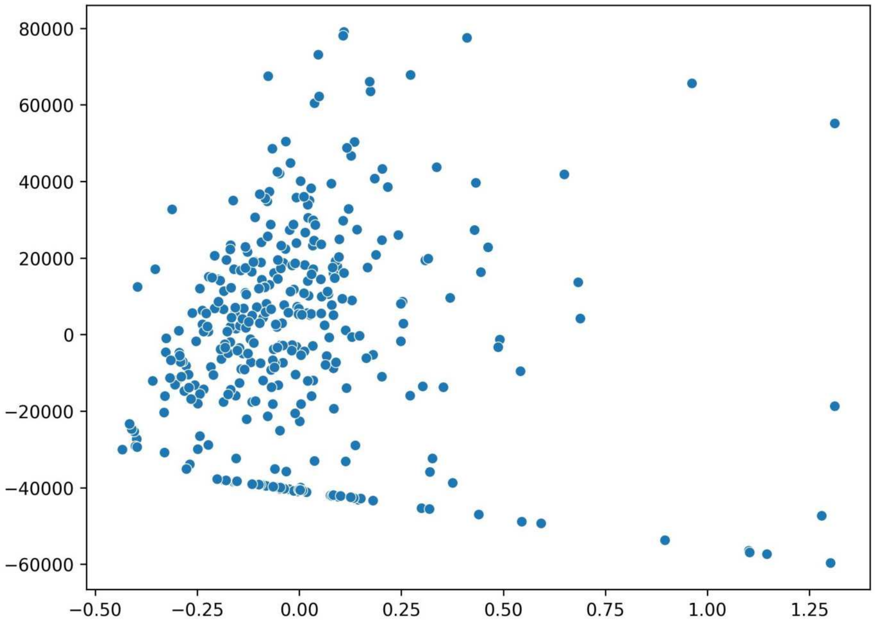

**Figure 4.** Data distribution visualization using a two-dimensional scatter plot in PCA.

as they can influence the selection of clustering and regression algorithms. The findings of the outlier detection process are depicted in Figure 5.

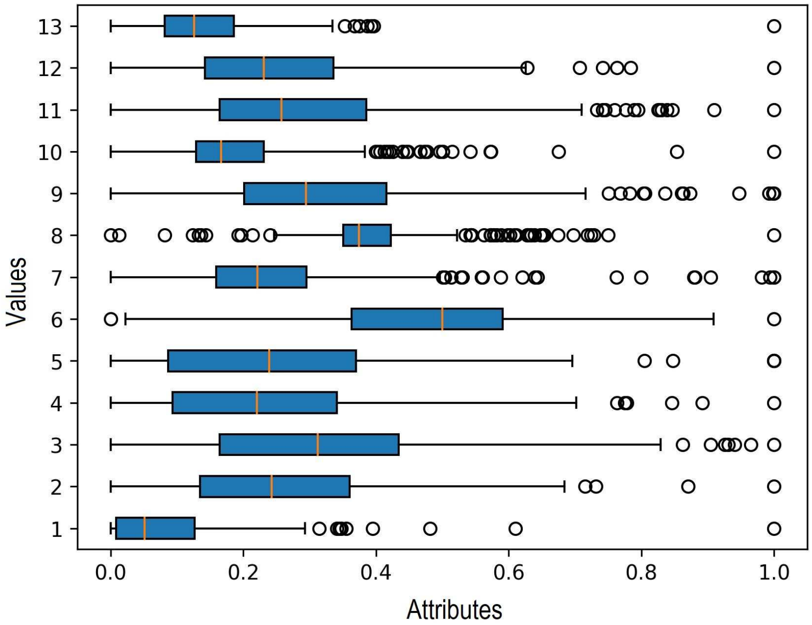

**Figure 5.** Outlier detection in boxplot chart.

Various clustering methods are used to group data into similar or homogeneous clusters. Because there were outliers in each input variable to be clustered, we used four clustering methods resistant to data outliers: K-Medoids, Fuzzy C-Means, DBSCAN, and OPTICS. Based on the experiment with five clusters from these four methods, K-Medoids and FCM provided the best silhouette score. The results of the silhouette score calculation for all the methods are shown in Table 6 below.

Based on the initial five clusters' results, the K-Medoids and FCM algorithms were repeatedly tested to obtain the optimal number of clusters. The best results obtained in K-Medoids are two clusters with a

| | K-MEDOIDS | OPTICS | FCM | DBSCAN |
|---|---|---|---|---|
| Silhouette Score | 0.197174 | 0.094126 | 0.235802 | 0.061974 |

**Table 6.** Silhouette score for five clusters.

silhouette value of 0.265 and DBI 1.388, whereas the best number of clusters obtained in FCM is three clusters with a silhouette value of 0.304 and DBI 1.119. The test results of the number of clusters and silhouette value can be observed in Table 7 and Figure 6

| Number of Cluster | Silhouette Score | | Davies Bouldin Index | |
|---|---|---|---|---|
| | FCM | K-Medoids | FCM | K-Medoids |
| 2 | 0.270 | **0.265** | 1.382 | **1.388** |
| 3 | **0.304** | 0.132 | **1.119** | 1.897 |
| 4 | 0.249 | 0.091 | 1.308 | 1.852 |
| 5 | 0.236 | 0.197 | 1.341 | 1.471 |
| 6 | 0.107 | 0.156 | 2.335 | 1.570 |
| 7 | 0.074 | 0.164 | 2.134 | 1.531 |

**Table 7.** Silhouette and DBI score for K-medoids and FCM.

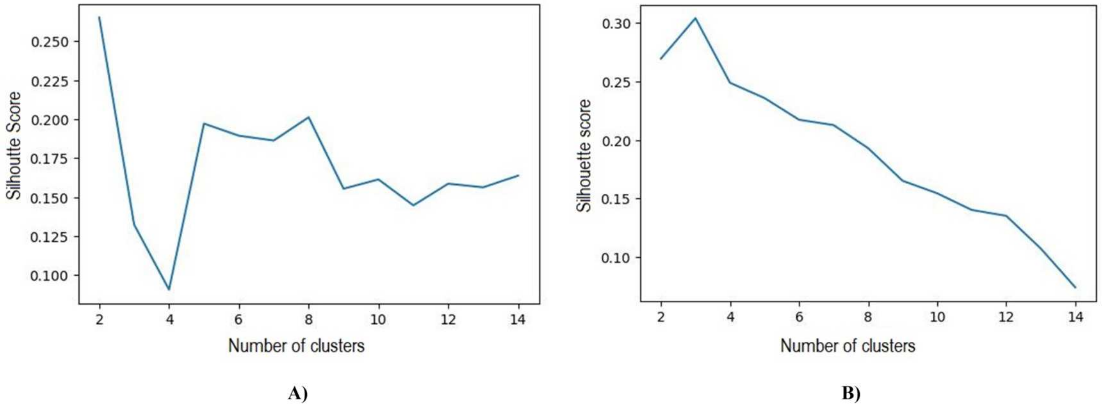

**Figure 6.** (A) Best silhouette score for K-Medoids clustering. (B) Best silhouette score for FCM clustering.

The clustering outcomes, featuring the most favorable silhouette scores from both the K-Medoids and FCM techniques, can be effectively visualized via a scatter plot, offering valuable insights into cluster memberships. Figure 7 exhibits this scatter plot, presenting the clustering pattern in two dimensions through PCA. PCA serves to diminish the data's high dimensionality, thereby simplifying the analysis and comprehension of intricate data. The plot illustrates a well-defined division of cluster members, with no instances of cluster overlap. The analysis leads to the conclusion that the FCM clustering algorithm, employing three clusters with a silhouette score of 0.304 and DBI of 1.119, outperforms the K-Medoids algorithm, which employs two clusters and achieves a silhouette score of 1.265 and DBI of 1.388. A comprehensive comparison of these results is available in Table 7. Based on these findings, the FCM algorithm with three clusters was chosen for further investigation. The respective cluster and centroid data are visually presented in the two-dimensional scatter plot graph in Figure 8. Additional information, including the cluster membership and centroid details, can be found in Table A9. Furthermore, the specific cluster results are detailed in Table A10. The clustering results can serve as recommendations for stakeholders in classifying offices into categories such as small, medium, and large.

After identifying the optimal clusters, the subsequent stage involves conducting DEA modeling. The primary objective of DEA is to determine static efficiency values for each DMU within each cluster. Each member of the cluster undergoes a separate DEA analysis. The efficiency values are computed using both input-oriented DEA-BCC and output-oriented DEA-BCC methods. These methods help assess

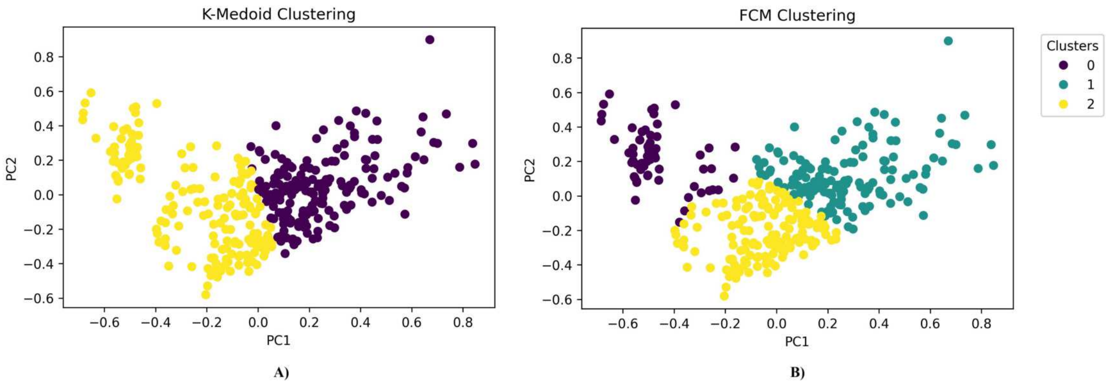

**Figure 7.** (A) Scatterplot results of clustering with the best silhouette score value for K Medoids clustering. (B) Scatterplot results of clustering with the best silhouette score value for FCM clustering.

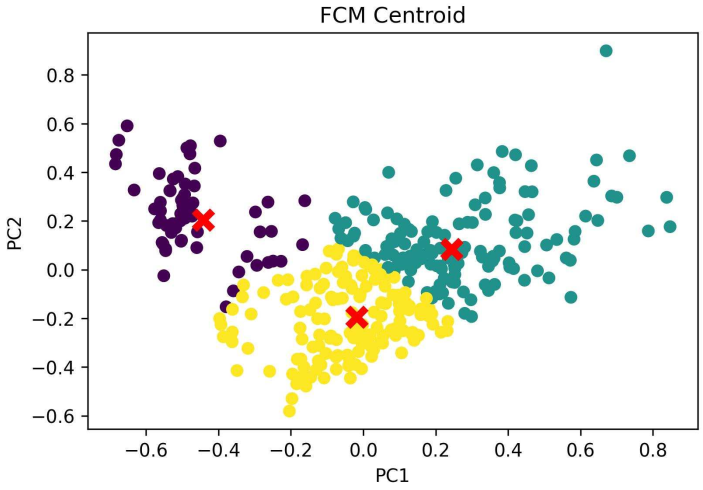

**Figure 8.** FCM cluster and centroid visualization.

the relative efficiency of each DMU within its respective cluster, taking into account input and output measures to evaluate their performance.

The detailed outcomes of the input-oriented DEA-BCC method can be observed in Table A11. From the results, it can be deduced that within cluster C0, 48 DMUs have attained efficiency, while the remaining 17 DMUs have not. In this cluster, the efficiency level reaches 74% of the total 65 DMU members. For cluster C1, 89 DMUs have achieved efficiency, accounting for 60% of the total DMUs, while the remaining 60 DMUs are not efficient. In cluster C2, 89 DMUs have already attained efficiency, constituting 65% of the total DMU members, leaving 49 DMUs yet to achieve efficiency.

The detailed results of the output-oriented DEA-BCC method can be found in Table A12. Within cluster C0, 54 DMUs have reached efficiency, and the remaining 11 DMUs have not yet achieved efficiency. In this cluster, the efficiency level reaches 83% of the total 65 DMU members. For cluster C1, 89 DMUs have attained efficiency, accounting for 60% of the total DMUs, while the remaining 60 DMUs are not efficient. In cluster C2, 88 DMUs have already attained efficiency, representing 63% of the total DMU members, with 50 DMUs yet to achieve efficiency.

The results of input-oriented and output-oriented calculations are then summarized in Table 8. For cluster C0, 48 DMUs, or 74% of the total 65 DMU cluster members, are efficient in both input-oriented and output-oriented approaches. In cluster C1, 89 DMUs, equivalent to 60% of the total cluster members, are efficient in both input-oriented and output-oriented analyses. In cluster C2, 88 DMUs, representing 64% of the total C2 cluster members, are efficient in both input-oriented and output-oriented evaluations. Overall, 225 out of 352 DMUs, or approximately 64%, demonstrated efficiency in both input- and output-oriented assessments. This information is summarized in Table A13. Consequently, it can be concluded that the overall performance of tax service offices in Indonesia has attained efficiency, with an efficiency level of 64% of the total number of offices. These findings indicate a substantial improvement over the previous study by Suyanto and Saksono (2013), which classified only 61 out of 331 tax service offices, or approximately 18%, as efficient.

**Table 8.** Summary of DEA process result.

| Cluster | Number of cluster members | Efficiency | Input Oriented | Output Oriented | Input and Output Oriented | Percentage |
|---|---|---|---|---|---|---|
| C0 | 65 | Efficient | 48 | 54 | 48 | 74% |
| | | Not Efficient | 17 | 11 | 17 | 26% |
| C1 | 149 | Efficient | 89 | 89 | 89 | 60% |
| | | Not Efficient | 60 | 60 | 60 | 40% |
| C2 | 138 | Efficient | 89 | 88 | 88 | 64% |
| | | Not Efficient | 49 | 50 | 50 | 36% |
| **TOTAL** | **352** | **Efficient** | **226** | **231** | **225** | **64%** |
| | | **Not Efficient** | **126** | **121** | **127** | **36%** |
| | | **Sum** | **352** | **352** | **352** | **100%** |

The results of the model optimization using the genetic algorithm (GA) are shown in Table 9. The Multilayer Perceptron Regression algorithm optimized with GA (GA-MLPR) achieved the smallest objective function value of 0.0035, with an execution time of 8 minutes and 13 seconds, and it converged on the 13th iteration. The best parameter configuration consists of five hidden layers, 73 units, a ReLU activation function, and a learning rate of 0.006. The second best-performing algorithm is the Genetic Algorithm SVR (GA-SVR) with an objective function value of 0.0037, followed by the Genetic Algorithm RFR (GA-RFR) with an objective function value of 0.0051, and the Genetic Algorithm Gradient Boosting Regressor (GA-GBR) with an objective function value of 0.0052. These results show a significant decrease in the MSE value for GA-MLPR, which decreased from 0.0144 to 0.0035, reflecting a substantial improvement of 75.75%, as seen in Table 10. We chose GA-MLPR as the best model based on its lowest MSE (0.0035), which indicates the highest predictive accuracy among all tested models. The multilayer perceptron's ability to capture non-linear relationships combined with genetic algorithms, offers good flexibility and adaptability to data variations while controlling the risk of overfitting. Although the standard deviation (0.0821) is slightly higher than some other models, it still indicates adequate stability, making GA-MLPR a robust and reliable solution for regression needs. The visualization of objective function values for each iteration can be seen in Figure 9.

**Table 9.** Model optimization results with genetic algorithm.

| Algorithm | Objective Function | Best solution | Executed time | Iteration |
|---|---|---|---|---|
| GA-MLPR | 0.0035 | layer = 5<br>unit = 73<br>activation function = ReLU<br>learning rate = 0.006 | 8m 13s | 13 |
| GA-SVR | 0.0037 | C = 0.549<br>epsilon = 0.0159 | 14.2s | 5 |
| GA-RFR | 0.0051 | n_estimators = 71<br>max_depth=10<br>max_features = 6 | 14m 17s | 12 |
| GA-GBR | 0.0052 | n_estimator = 187.950<br>max_depth = 2.246<br>learning_rate = 0.09 | 3m 30s | 12 |

**Table 10.** Results of MSE and standard deviation value reduction before and after the algorithm is optimized using genetic algorithm.

| Optimized Algorithm | MSE | Decrease Percentage | Standard Deviation |
|---|---|---|---|
| MLPR<br>GA-MLPR | 0.0144<br>0.0035 | 75.75% | 0.1016<br>0.0821 |
| SVR<br>GA-SVR | 0.0057<br>0.0037 | 34.28% | 0.0714<br>0.0696 |
| RFR<br>GA-RFR | 0.0059<br>0.0051 | 13.78% | 0.0728<br>0.0724 |
| GBR<br>GA-GBR | 0.0057<br>0.0052 | 8.92% | 0.0713<br>0.0712 |

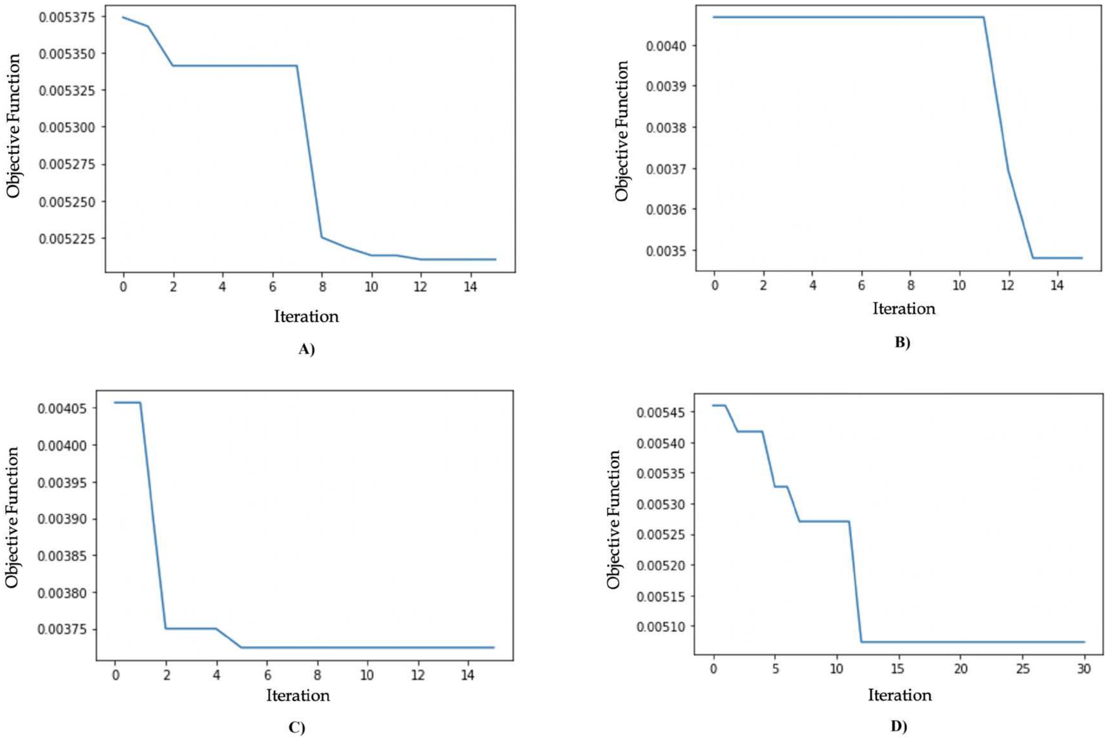

**Figure 9.** (A) Genetic algorithm chart optimization for GA-GBR. (B) Genetic algorithm chart optimization for GA-MLPR. (C) Genetic algorithm chart optimization for GA-SVR. (D) Genetic algorithm chart optimization for GA-RFR.

## CONCLUSIONS

This paper presents an experimental approach to assess the efficiency of tax service offices in Indonesia using real dataset through three stages: clustering with K-Medoids, OPTICs, DBScan, and FCM algorithms to group tax service offices as DMUs; static efficiency measurement using input-oriented and output-oriented DEA-BCC; and dynamic efficiency prediction using machine learning regression algorithms (Gradient Boosting Regressor (GBR), MLPR, SVR, and RFR) optimized with GA. The FCM algorithm, with a silhouette value of 1.304 and a DBI value of 1.119, outperformed other algorithms and produced three clusters of tax service offices.

In the DEA measurement, using the input-oriented DEA-BCC method, 226 tax service offices were found to be efficient DMUs, while using the output-oriented DEA-BCC method, there were 231 efficient DMUs. Overall, 225 out of 352 DMUs demonstrated efficiency in both input- and output-oriented calculations, representing 64% efficient DMUs of the Tax Service Office. These findings show a significant improvement over the previous study, in which only 61 out of 331 tax service offices, or about 18%, were classified as efficient.

The Multilayer Perceptron Regression algorithm optimized with Genetic Algorithm (GA-MLPR) obtained optimal results with parameter combination of 73 units, five hidden layers, a ReLU activation function, and a learning rate of 0.006. It achieved an objective function value of 0.0035 during the 13th iteration, significantly reducing the MSE value by approximately 75.75% from 0.0144 to 0.0035.

The findings of this study can serve as a reference for stakeholders to categorize tax offices into small, medium, and large categories based on the clustering results. The DEA process that identifies efficient offices can serve as a benchmark for the efficiency levels that other offices should aim to achieve. Additionally, stakeholders can propose these efficient offices for incentives as a form of reward, which is expected to motivate performance improvement across the tax service sector.

For future research, it is recommended to use data from multiple years and incorporate more variables to enhance the comprehensiveness of the DEA analysis. Employing additional regression algorithms and optimization models from other heuristic algorithms is also suggested to further improve the objective function value.

## REFERENCES

Alm, J. and Duncan, D. (2014). Estimating tax agency efficiency. *Public Budgeting & Finance*, 34(3):92–110.

Appiahene, P., Missah, Y. M., and Najim, U. (2020). Predicting bank operational efficiency using machine learning algorithm: comparative study of decision tree, random forest, and neural networks. *Advances in fuzzy systems*, 2020(1):8581202.

ATAF (2021). Evaluation of tax administration efficiency: Data envelope analysis (dea). *SSRN*.

Banker, R. D., Charnes, A., and Cooper, W. W. (1984). Some models for estimating technical and scale inefficiencies in data envelopment analysis. *Management science*, 30(9):1078–1092.

Banker, R. D., Charnes, A., Cooper, W. W., Swarts, J., and Thomas, D. (1989). An introduction to data envelopment analysis with some of its models and their uses. *Research in governmental and nonprofit accounting*, 5(1):125–163.

Bezdek, J. C., Ehrlich, R., and Full, W. (1984). Fcm: The fuzzy c-means clustering algorithm. *Computers & geosciences*, 10(2-3):191–203.

Breiman, L. (2001). Random forests. *Machine learning*, 45:5–32.

Charnes, A., Cooper, W. W., and Rhodes, E. (1978). Measuring the efficiency of decision making units. *European journal of operational research*, 2(6):429–444.

Chicco, D., Warrens, M. J., and Jurman, G. (2021). The coefficient of determination r-squared is more informative than smape, mae, mape, mse and rmse in regression analysis evaluation. *Peerj computer science*, 7:e623.

Coelli, T. (1996). 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):1–49.

Dalvand, B., Jahanshahloo, G., Lotfi, F. H., and Rostami, M. (2014). Using c4.5 algorithm for predicting efficiency score of dmus in dea. *Advances in Environmental Biology*, 8(22):473–477.

De Carvalho Couy, J. P. (2015). *Asessing Tax Inspection Performance: A Data Evelopment Analysis on Brazilian Federal Tax Offices*. PhD thesis, KDI School of Public Policy and Management.

DGT (2020). *Directorate General of Taxes - Annual Report (Laporan Tahunan DJP)-2020*. DGT.

DGT (2021). *Directorate General of Taxes - Annual Report (Laporan Tahunan DJP)-2021*. DGT.

Emrouznejad, A. and Yang, G.-l. (2018). A survey and analysis of the first 40 years of scholarly literature in dea: 1978–2016. *Socio-economic planning sciences*, 61:4–8.

Ester, M., Kriegel, H.-P., Sander, J., Xu, X., et al. (1996). A density-based algorithm for discovering clusters in large spatial databases with noise. In *kdd*, volume 96, pages 226–231.

Fadhila, D. (2014). Analysis of tax efficiency at the directorate general of taxes for the period 2006-2012 (analisis efisiensi pajak pada direktorat jenderal pajak periode 2006-2012). *Fakultas Ekonomi dan Bisnis UIN Syarif Hidayatullah*.

Friedman, J. H. (2001). Greedy function approximation: a gradient boosting machine. *Annals of statistics*, pages 1189–1232.

Gajić, L., Cvetnić, D., Bezdan, T., Živković, M., and Bačanin, N. (2020). Multi-layer perceptron training by genetic algorithms. In *Sinteza 2020 - International Scientific Conference on Information Technology and Data Related Research*, pages 301–306.

González, P. E. B. and Rubio, E. V. (2013). The efficiency of the regional management centres of the tax administration in spain. *Journal of US-China Public Administration*, 10(1):49–56.

Huang, S.-H., Yu, M.-M., and Huang, Y.-L. (2022). Evaluation of the efficiency of the local tax administration in taiwan: Application of a dynamic network data envelopment analysis. *Socio-Economic Planning Sciences*, 83(C).

Huang, S.-H., Yu, M.-M., Hwang, M.-S., Wei, Y.-S., and Chen, M.-H. (2017). Efficiency of tax collection and tax management in taiwan's local tax offices. *Pacific Economic Review*, 22(4):620–648.

Indrawati, Y. (2009). Analysis of the efficiency of commercial banks in indonesia for the period 2004-2007: Application of the dea method (analisis efisiensi bank umum di indonesia periode 2004-2007: Aplikasi metode dea). *Fakultas Ekonomi Universitas Indonesia*.

Jain, A. K. (2010). Data clustering: 50 years beyond k-means. *Pattern recognition letters*, 31(8):651–666.

Kalb, A. (2010). *Public sector efficiency: applications to local governments in Germany*. Springer Science & Business Media.

Kaufman, L. (1990). Partitioning around medoids (program pam). *Finding groups in data*, 344:68–125.

Kohavi, R. (1995). A study of cross-validation and bootstrap for accuracy estimation and model selection. *Morgan Kaufman Publishing*.

Lambora, A., Gupta, K., and Chopra, K. (2019). Genetic algorithm-a literature review. In *2019 international conference on machine learning, big data, cloud and parallel computing (COMITCon)*, pages 380–384. IEEE.

Milosavljević, M., Radovanović, S., and Delibašić, B. (2023). What drives the performance of tax administrations? evidence from selected european countries. *Economic Modelling*, 121:106217.

Ministry of Finance of the Republic of Indonesia (2023). *Our National Budget (APBN Kita)*. Ministry of Finance of the Republic of Indonesia.

Mitchell, M. (1998). *An introduction to genetic algorithms*. MIT press.

Montgomery, D. C., Peck, E. A., and Vining, G. G. (2021). *Introduction to linear regression analysis*. John Wiley & Sons.

Moore, D. S. and McCabe, G. P. (1989). *Introduction to the practice of statistics.* WH Freeman/Times Books/Henry Holt & Co.

Murakami, Y., Tanaka, M., Bramantoro, A., and Zettsu, K. (2012). Data-centered service composition for information analysis. In *2012 IEEE Ninth International Conference on Services Computing*, pages 602–608. IEEE.

Nayak, J., Naik, B., and Behera, H. (2015). Fuzzy c-means (fcm) clustering algorithm: a decade review from 2000 to 2014. In *Computational Intelligence in Data Mining-Volume 2: Proceedings of the International Conference on CIDM, 20-21 December 2014*, pages 133–149. Springer.

OECD (2022). *Revenue Statistics in Asia and the Pacific 2022 - Indonesia*. OECD.

Omrani, H., Shafaat, K., and Emrouznejad, A. (2018). An integrated fuzzy clustering cooperative game data envelopment analysis model with application in hospital efficiency. *Expert Systems with Applications*, 114:615–628.

Rahimi, I. and Behmanesh, R. (2012). Improve poultry farm efficiency in iran: using combination neural networks, decision trees, and data envelopment analysis (dea). *International Journal of Applied Operational Research*, 2(3):69–84.

Razavi Hajiagha, S. H., Hashemi, S. S., and Amoozad Mahdiraji, H. (2016). Fuzzy c-means based data envelopment analysis for mitigating the impact of units' heterogeneity. *Kybernetes*, 45(3):536–551.

Rezaee, M. J., Jozmaleki, M., and Valipour, M. (2018). Integrating dynamic fuzzy c-means, data envelopment analysis and artificial neural network to online prediction performance of companies in stock exchange. *Physica A: Statistical Mechanics and its Applications*, 489:78–93.

Rostamzadeh, R., Akbarian, O., Banaitis, A., and Soltani, Z. (2021). Application of dea in benchmarking: a systematic literature review from 2003–2020. *Technological and Economic Development of Economy*, 27(1):175–222.

Samuel, A. L. (2000). Some studies in machine learning using the game of checkers. *IBM Journal of research and development*, 44(1.2):206–226.

Sherman, H. D. and Zhu, J. (2013). Analyzing performance in service organizations. *MIT Sloan Management Review*, 54(4):37–42.

Smola, A. J. and Schölkopf, B. (2004). A tutorial on support vector regression. *Statistics and computing*, 14:199–222.

Suyanto, B. and Saksono, R. A. (2013). Analysis of the efficiency of tax service offices using the data envelopment analysis (dea) method (analisis efisiensi kantor pelayanan pajak dengan metode data envelopment analysis (dea)). *Good Governance*, 9(1):1–30.

Tibshirani, R. J. and Efron, B. (1993). An introduction to the bootstrap. *Monographs on statistics and applied probability*, 57(1):1–436.

Triantoro, H. and Subroto, B. (2016). Efficiency performance of primary tax service offices: A data envelopment analysis (dea) approach (kinerja efisiensi kantor pelayanan pajak pratama: Pendekatan data envelopment analysis (dea)). *Jurnal Akuntansi Aktual*, 3(3):215–225.

Usino, W., Prabuwono, A. S., Allehaibi, K. H. S., Bramantoro, A., Hasniaty, A., and Amaldi, W. (2019). Document similarity detection using k-means and cosine distance. *International Journal of Advanced Computer Science and Applications*, 10(2):165–170.

Wei, G. and Wang, J. (2017). A comparative study of robust efficiency analysis and data envelopment analysis with imprecise data. *Expert systems with applications*, 81:28–38.

Wolfowitz, J. (1949). Non-parametric statistical inference. In *Proceedings of the [First] Berkeley Symposium on Mathematical Statistics and Probability*, volume 1, pages 93–114. University of California Press.

Ye, A.-X. and Jin, Y.-X. (2016). A fuzzy c-means clustering algorithm based on improved quantum genetic algorithm. *International Journal of Database Theory and Application*, 9(1):227–236.

Yunianta, A., Basori, A. H., Prabuwono, A. S., Bramantoro, A., Syamsuddin, I., Yusof, N., Almagrabi, A. O., and Alsubhi, K. (2019). Ontodi: The methodology for ontology development on data integration. *International Journal of Advanced Computer Science and Applications*, 10(1):160–168.

Zhang, Z., Xiao, Y., and Niu, H. (2022). Dea and machine learning for performance prediction. *Mathematics*, 10(10):1776.

Zhu, N., Zhu, C., and Emrouznejad, A. (2021). A combined machine learning algorithms and dea method for measuring and predicting the efficiency of chinese manufacturing listed companies. *Journal of Management Science and Engineering*, 6(4):435–448.