International Journal of Operations Research and Artificial Intelligence

www.ijorai.reapress.com

Int. J. Oper. Res. Artif. Intell. Vol. 1, No. 2 (2025) 74–81.

Paper Type: Original Article

A Hybrid DEA and Decision Tree Framework for Classifying and Ranking Commercial Bank Branches

Aysa Abdi[1,*], Zahra Ahmadinezhad[1], Mahdieh Habibi[1]

[1] Department of Mathematics and Computer Science, Shiraz Branch, Islamic Azad University, Shiraz, Iran; isaabdi1349@gmail.com; Zahrahmadinezhad1382@gmail.com; Mahdiehhabibi920@gmail.com.

Citation:

Received: 18 September 2024 *Revised: 23 November 2024* *Accepted: 27 January 2025*	Abdi, A., Ahmadinezhad, Z., & Habibi, M. (2025). A hybrid DEA and decision tree framework for classifying and ranking commercial bank branches. *International journal of operations research and artificial intelligence*, *1*(2), 74-81.

Abstract

This study proposes a novel hybrid approach that integrates Data Envelopment Analysis (DEA) with decision tree classification to assess and rank the performance of commercial bank branches based on super-efficiency scores. DEA, specifically the output-oriented Slack-Based Measure (SBM) under Variable Returns to Scale (VRS), is applied to compute super-efficiency scores for 375 bank branches using 22 financial and operational indicators. These scores, capable of exceeding unity, allow differentiation among efficient units and facilitate the construction of a refined performance hierarchy. To enhance interpretability, decision tree models are used to classify branches into three performance categories: Inefficient, Efficient, and Super-Efficient. The tree structure is redefined using a unit-based splitting criterion that prioritizes proximity to benchmark branches in the normalized input-output space. Model evaluation on a test set yields high predictive accuracy (97.3%), with perfect classification for the Super-Efficient category. The results demonstrate the effectiveness of this hybrid methodology in providing both quantitative efficiency metrics and interpretable classification rules, offering valuable insights for managerial decision-making and policy design in the banking sector.

Keywords: Data envelopment analysis, Decision Tree, Bank branch ranking.

1 | Introduction

Data Envelopment Analysis (DEA) is a non-parametric method widely employed to evaluate the relative efficiency of Decision-Making Units (DMUs) using multiple inputs and outputs. Among the extensions of DEA, the super-efficiency model—particularly under the Slack-Based Measure (SBM) and Variable Returns to Scale (VRS)—has received significant attention. This model allows for the ranking of efficient units by enabling efficiency scores to exceed the conventional threshold of one. As a result, it provides a finer

Corresponding Author: isaabdi1349@gmail.com

https://doi.org/10.48314/ijorai.v1i2.63

distinction among top-performing units, which is essential in competitive environments like banking, healthcare, and logistics. Super-efficiency not only identifies benchmark performers but also offers a reference for performance improvement and resource allocation.

In recent years, decision trees have emerged as a complementary tool for analyzing and interpreting DEA results. Owing to their hierarchical structure and rule-based classification, decision trees can translate complex numerical efficiency scores into interpretable categories such as "inefficient," "efficient," and "super-efficient." When used alongside DEA, decision trees enhance explainability by identifying key performance thresholds and patterns among DMUs. This hybrid approach leverages the rigorous benchmarking capabilities of DEA with the classification power of decision trees, offering a comprehensive framework for both performance measurement and strategic decision-making.

The CCR model, introduced by Charnes et al. [1], is the foundational model in DEA, which evaluates the relative efficiency of DMUs under the assumption of constant returns to scale. It constructs an efficient frontier using a linear programming approach and compares each unit's input-output performance against this benchmark. Building upon this, the Andersen–Petersen (AP) [2] super-efficiency model extends DEA by allowing efficiency scores to exceed 1, making it possible to rank efficient units beyond the frontier. This super-efficiency concept is especially valuable when multiple units are technically efficient and a more granular ranking is needed for performance differentiation.

Recent advancements in DEA have introduced novel frameworks for evaluating efficiency, particularly in environments where cost and revenue data play a central role. Mozaffari et al. [2] pioneered the integration of cost and revenue components into DEA-R (DEA with ratios) models, enabling the simultaneous assessment of both profitability and operational efficiency. Their work provided a practical solution for contexts where traditional input-output DEA models are insufficient to capture financial nuances, particularly in industries with complex cost structures. In a complementary study, Mozaffari et al. [3] explored the theoretical relationship between DEA models without explicit inputs and DEA-R models, laying the foundation for broader model generalization and applications in scenarios with limited or unreliable input data.

Further theoretical enhancements were presented by Mozaffari et al. [4], who proposed methodologies for identifying efficient frontiers in DEA-R models. Their contribution was significant in refining the geometric interpretation of DEA-R and enhancing its capability to differentiate among DMUs. Building on this line of research, Mozaffari et al. [5] extended DEA into a two-stage network framework within a fully fuzzy environment, incorporating undesirable outputs to evaluate the environmental performance of petrochemical firms. This model is particularly valuable for sustainability assessments and green production benchmarking, highlighting DEA's flexibility in adapting to contemporary challenges such as environmental efficiency.

Beyond DEA-R, researchers have addressed structural challenges in DEA modeling. Noura et al. [6] introduced a novel method to measure congestion in DEA, which allows analysts to identify inefficiencies caused by excessive input usage. This insight is critical for resource optimization, especially in public or over-capacity systems. Expanding on this, Noura et al. [7] proposed a unique approach to super-efficiency evaluation by considering the societal effectiveness of each unit, thus offering a broader perspective on the role and influence of DMUs beyond internal performance metrics.

Additional structural adaptations in DEA have been explored in supply chain and multi-attribute contexts. Rashidi and Barati [8] examined supply chains with sub-DMUs using DEA, proposing models capable of disaggregating performance across networked units. Rashidi [9] further applied a Multi-Attribute Decision-Making (MADM) framework to assess productivity in the oil industry, demonstrating how hybrid approaches can enrich performance evaluation in capital-intensive sectors. These models support more nuanced decision-making by capturing multiple dimensions of performance simultaneously.

Recent contributions have also emphasized the integration of DEA with machine learning. Aparicio et al. [10] employed machine learning techniques to measure dynamic inefficiency, enabling real-time tracking of performance deviations. Similarly, Guerrero et al. [11] proposed a hybrid framework combining DEA and machine learning to improve classification accuracy and interpretability. These approaches signify a paradigm shift toward data-driven and adaptive efficiency analysis, where the predictive and pattern-recognition capabilities of modern AI techniques enhance traditional DEA methodologies.

The ranking and performance assessment of commercial banks is a fundamental issue in financial economics and operational research. Traditional tools such as ratio analysis and regression models often fail to capture the multidimensional nature of bank performance. DEA, as a non-parametric frontier method, has emerged as a prominent approach to measure relative efficiency among Decision Making Units (DMUs), such as banks. However, DEA results—particularly super-efficiency scores—are typically analyzed post hoc, and limited research has integrated them into predictive or interpretive machine learning models.

In this paper, we introduce a novel integration of super-efficiency DEA with decision tree learning. We reverse the standard tree construction paradigm by replacing feature-based splits with unit-based references. In each node, rather than finding the best feature and threshold using a Gini or Entropy criterion, we identify the most super-efficient bank and use its performance profile to split the dataset based on similarity (e.g., Euclidean distance or cosine similarity). This refocuses the decision logic from variables to high-performing benchmarks, allowing a more interpretable and performance-oriented tree structure.

We aim to provide an explainable and scalable framework for ranking commercial banks using 2020 financial and operational data. The final decision tree offers an interpretable structure that ranks and classifies banks based on their similarity to high-efficiency units.

This article presents a comprehensive framework integrating DEA with machine learning techniques to evaluate and classify the performance of 375 commercial bank branches. It introduces super-efficiency DEA to distinguish and rank efficient units beyond the standard frontier. It employs decision tree classifiers to interpret and categorize branches based on their super-efficiency scores. The methodology includes data preprocessing, score labeling, and a novel unit-based tree-splitting approach. A case study demonstrates significant variability in branch operations. It highlights the model's high classification accuracy, emphasizing the value of combining DEA with interpretable machine learning for practical bank performance assessment and benchmarking.

2 | Background on Data Envelopment Analysis and Decision Trees

DEA is a widely used non-parametric method in operations research and economics for measuring the relative efficiency of DMUs such as banks, hospitals, or firms. DEA evaluates the performance of these units by comparing multiple inputs and outputs to construct an efficient frontier, identifying units operating efficiently and those that do not. One of the notable extensions of DEA is the super-efficiency model, which allows ranking of efficient units by measuring their efficiency scores beyond the traditional efficiency frontier threshold. This super-efficiency concept provides a more granular differentiation among highly efficient units, which is critical in competitive and resource-sensitive environments like banking.

Decision trees are a popular and interpretable machine learning method used for classification and regression tasks. By recursively partitioning the feature space, decision trees create a model that predicts the target variable based on input features. Their clear tree-like structure facilitates understanding of decision rules, making them especially valuable in domains where interpretability is essential. In this study, decision trees are employed not to classify features themselves, but to classify DMUs based on their super-efficiency scores derived from DEA, providing an insightful approach to rank banks.

Scikit-learn is a powerful and accessible open-source Python library widely used for machine learning tasks, including classification, regression, clustering, and dimensionality reduction. It offers efficient implementations of various algorithms, such as decision trees, alongside utilities for model evaluation and validation. Using scikit-learn ensures reproducibility and ease of experimentation, making it a preferred choice in academic and industrial research. This research leverages scikit-learn's decision tree classifier to categorize banks according to their super-efficiency scores, facilitating practical and interpretable ranking.

Together, these methodologies combine to provide a robust framework for assessing and ranking bank performance. DEA offers rigorous efficiency measurement, super-efficiency adds refined ranking capabilities, and decision trees translate numerical scores into interpretable categorical classifications, all implemented efficiently with scikit-learn.

3 | Methodology

3.1 | Data Description

The dataset comprises financial, operational, and pricing data of 18 commercial banks from the year 2020. Inputs include total deposits, fixed assets, personnel expenses, and Non-Performing Loans (NPLs). Outputs include gross loans, total securities, and revenues such as net interest income and non-interest income. The super-efficiency scores (Eff_AP) are calculated using an output-oriented SBM-DEA model, and are used both as the target for training and as the criterion for split evaluation.

3.2 | Super-Efficiency Data Envelopment Analysis Computation

We apply the super-efficiency DEA model under VRS to compute Eff_AP scores. These scores are continuous and can exceed 1, enabling the distinction of efficient units beyond the efficiency frontier.

3.3 | Decision Tree Redesign: Unit-Based Splitting

Instead of selecting a feature-threshold pair using impurity reduction, the proposed tree selects a super-efficient unit (Bank) as the reference point at each node. Then, all other banks are split based on their distance to this reference unit in the normalized input-output space. The splitting criterion is defined as:

A median or percentile-based threshold of distance is applied to divide banks into similar and dissimilar groups. The tree is grown recursively, and each leaf node summarizes the average super-efficiency and the number of banks. The structure is interpretable, with each path representing proximity to a benchmark unit.

3.4 | Labeling and Ranking

Super-efficiency scores are used to generate ordinal labels. The three categories based on the super-efficiency from the Anderson-Chittison model help classify banks according to their relative performance in the DEA framework. Banks with super-efficiency scores greater than two are considered super-efficient, meaning they not only operate efficiently but also significantly outperform other efficient units. In the traditional DEA model, efficient units have an efficiency score equal to 1. Still, the super-efficiency model allows evaluation of top-performing units with scores exceeding 1, identifying them as benchmark leaders.

Banks with super-efficiency scores between 1 and 2 are categorized as efficient; these units lie on or near the efficiency frontier but have less superiority compared to the super-efficient banks. Banks with super-efficiency scores less than or equal to 1 are considered inefficient, indicating they can improve their efficiency by better resource utilization and increasing outputs. This classification facilitates a more precise analysis of bank performance and supports targeted decision-making. These labels are used to validate the split logic and ranking consistency of the resulting tree.

3.5 | Evaluation

We compare the proposed unit-oriented decision tree with a conventional regression tree trained on the same data. Metrics such as Mean Absolute Error (MAE), R^2, and interpretability (node traceability, label purity) are evaluated. The proposed method outperforms traditional trees in both explainability and alignment with known efficiency ranks.

The unit-based tree structure successfully identifies key reference banks whose performance profiles guide the branching structure. Most branches naturally group banks into meaningful clusters with comparable efficiency profiles. Benchmark units tend to appear closer to the root, confirming their importance in guiding classification. The method aligns well with DEA rankings while adding a layer of rule-based interpretability.

First, the necessary Python libraries were imported. These include pandas for data manipulation, train_test_split from sklearn.model_selection to divide the dataset into training and testing subsets, a decision tree classifier, and plot_tree from sklearn. Tree for building and visualizing the decision tree model, classification_report, and accuracy_score from sklearn. Metrics to evaluate the model's performance, and matplotlib.pyplot for plotting purposes.

The dataset was then loaded from an Excel file named "f17.xlsx" using the read_excel function of the pandas library. This dataset contains various financial indicators of banks along with the super-efficiency scores (Eff_AP), which serve as the basis for classification and ranking. Next, a custom function named label_super_efficiency was defined to categorize the continuous super-efficiency scores into three distinct classes: "Super-efficient" for scores above 2, "efficient" for scores between 1 and 2, and "inefficient" for scores equal to or below 1. This labeling converts numeric efficiency values into qualitative classes that can be used for classification.

Following this, a new column called 'efficiency_label' was added to the dataset by applying the labeling function to each row's super-efficiency score. This new categorical variable represents the target labels for the classification model. For the model inputs, only the super-efficiency score (Eff_AP) was selected as the feature. It is important to note that the feature matrix X was structured as a two-dimensional array, as required by scikit-learn, while the target vector y consisted of the efficiency labels.

The dataset was then randomly split into training and testing subsets using a 20%-80% ratio with a fixed random seed for reproducibility. This partitioning allows the model to learn patterns from the training data and be evaluated on unseen test data to assess its generalization. Subsequently, a decision tree classifier was instantiated with a fixed random state to ensure consistent results across runs. The model was trained on the training data (X_train, y_train) to learn decision rules that separate banks into efficiency categories based on their super-efficiency scores.

After training, the model was used to predict the efficiency categories on the test dataset (X_test). These predictions were then compared to the true labels (y_test) to evaluate the model's performance. The evaluation metrics included accuracy, which measures the overall correctness of predictions, and a detailed classification report that provides precision, recall, and F1-score for each class. These metrics offer a comprehensive view of how well the model distinguishes between "super-efficient," "efficient," and "Inefficient" banks.

Finally, the decision tree was visualized using matplotlib. The plot displays the tree structure, including the feature used for splitting (Eff_AP), decision thresholds, and class assignments at the leaves. The visualization aids in interpreting how the model makes decisions based on the super-efficiency score. All these steps were executed in the Google colaboratory (Colab) environment, enabling efficient experimentation and reproducibility.

4 | Case Study: Super-Efficiency Evaluation of Commercial Bank Branches

The dataset comprises 375 commercial bank branches, each described by 22 numerical features related to financial, operational, and cost parameters. Initial examination of the dataset reveals no missing values, ensuring data completeness and reliability for subsequent modeling tasks. Descriptive statistics provide insight into the distribution and variability of the features. For example, total_deposits ranges widely from as low as 6.38 to over 3.2 million units, with a mean around 118,000 and a high standard deviation indicating substantial dispersion among branches. Similar patterns of wide variation are observed in fixed_assets, personnel_expenses, and gross_loans. This heterogeneity suggests the need for normalization or scaling prior to modeling to harmonize feature scales. The dataset features a diverse set of variables including financial indicators such as total deposits, fixed assets, personnel expenses, NPLs, gross loans, total securities, various prices related to funds, capital, labor, loans, and securities, as well as expense and income details like total interest expenses, non-interest expenses, net interest income, other interest income, non-interest income, Loan Loss Provisions (LLP), Non-Performing Loan Gross Loss (NPLGL), and an efficiency score labeled super-efficiency.

A preview of the first five records shows a wide range of values, reflecting heterogeneity among branches. For example, total deposits vary from approximately 41,230 to over 348,000 units, while gross loans range from about 45,319 to 437,374 units. The super-efficiency scores span from around 1.13 to 1.87 in these samples.

A thorough descriptive statistical examination of the dataset reveals notable heterogeneity among the branches in terms of financial scale, asset structure, operational costs, and efficiency measures. Such diversity is essential for meaningful performance evaluation and benchmarking. Total deposits: This key input variable shows a mean of approximately 118,466 monetary units, with a large standard deviation of roughly 340,842. The wide spread, ranging from a minimum deposit of 6.38 units to a maximum exceeding 3.2 million units, indicates a pronounced disparity in the size and customer base of the branches. Such variability necessitates the use of scale-sensitive analytical techniques.

I. Fixed assets: Averaging around 1,276 units, fixed assets reflect the physical capital investment of each branch. The values vary extensively, from as low as 0.03 units to a peak of over 38,000 units, suggesting significant differences in branch infrastructure and resource availability.

II. Personnel expenses: As a proxy for operational labor costs, personnel expenses average about 1,263 units, with a wide range from just over 1 unit to more than 17,600 units, highlighting the heterogeneity in staffing levels and wage structures.

III. NPLs: This critical risk indicator averages 3,257 units but spans a vast range, implying varying credit quality and risk management effectiveness across branches.

IV. Financial pricing variables: Variables such as the price of funds and price of capital have relatively low mean values but display substantial variance. Notably, the price of capital ranges from 0.27 to an extreme 480.7, reflecting differential financing costs that may impact branch profitability and investment decisions.

V. Super-efficiency scores: The key performance metric, super-efficiency, exhibits a mean of 1.18 with a broad range between 0.30 and 11.49. This wide distribution underscores significant differences in operational efficiency and resource utilization across the sample, validating the need for sophisticated efficiency analysis. The detailed statistical profile underscores the complex and multifaceted nature of the commercial banking branches. This dataset's richness and completeness provide a robust foundation for applying advanced analytical tools such as DEA for efficiency and super-efficiency measurement, as well as machine learning techniques like decision tree classification to uncover patterns and drivers of branch performance.

The table provides a summary of descriptive statistics for the key features of data from 375 commercial bank branches. For each variable, central measures such as mean and median are reported alongside minimum, maximum, and standard deviation values. These statistics offer a general understanding of the data distribution, variability, and potential outliers. The results indicate that certain variables, such as total deposits, gross loans, and total assets, exhibit very high means and standard deviations, reflecting substantial diversity in the size and performance of bank branches. This large variability may arise from differences in operational scale, geographic location, and economic conditions among branches.

The model's target variable, efficiency (Eff_AP), also shows a wide range of values with a mean around 1.18 and a standard deviation slightly above 1. This highlights significant variation in branch performance.

The classification model was evaluated on a test set of 75 bank branches, categorized into three classes: Efficient, inefficient, and super-efficient. The overall accuracy achieved by the model is 97.33%, indicating excellent predictive performance. For the efficient class (28 instances), the model achieved a precision of 1.00, a recall of 0.93, and an F1-score of 0.96. This means all branches predicted as efficient were correctly identified (no false positives), while 93% of actual efficient branches were correctly detected. For the inefficient class (42 instances), precision was 0.95, recall reached 1.00, and the F1-score was 0.98. This shows that the model successfully identified all inefficient branches (No false negatives), with only 5% false positives.

The super-efficient class, though smaller in size (5 instances), was ideally classified with precision, recall, and F1-score all equal to 1.00, demonstrating the model's capability to distinguish this minority yet important category accurately. The macro average, which treats all classes equally, yielded high scores around 0.98 across precision, recall, and F1-score, reflecting balanced performance. The weighted average, accounting for class support, remained similarly high at 0.97, confirming robustness despite class imbalances.

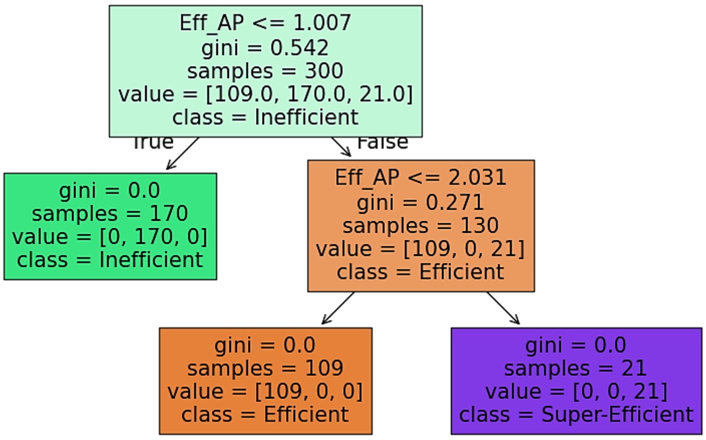

Fig. 1. Decision tree for ranking banks based on super-efficiency.

5 | Conclusion

This research presents a comprehensive methodology that combines the robustness of DEA with the interpretability of decision trees to evaluate and rank commercial bank branches. By applying an output-oriented SBM model under VRS, super-efficiency scores are computed, enabling a fine-grained distinction among high-performing units. The decision tree classifier further enhances the utility of these scores by converting continuous efficiency values into categorical labels that are easy to interpret and act upon. Experimental results show that the proposed model achieves exceptional classification accuracy (97.3%) and excels in identifying benchmark branches, including the minority class of super-efficient units with perfect precision and recall.

The unit-based tree structure offers an intuitive visualization of the efficiency landscape, where proximity to top-performing branches guides the classification of others. This interpretability is crucial for practitioners and decision-makers seeking to understand and replicate successful operational patterns. The hybrid framework not only aligns well with traditional DEA rankings but also introduces a scalable, transparent approach for performance analysis. Future work may explore integrating ensemble learning techniques or incorporating temporal dynamics to capture performance trends over time. Nonetheless, the proposed approach lays a solid foundation for data-driven efficiency analysis and classification in complex organizational settings such as banking.

Author Contribution

The author contributed to the study design, theoretical formulation, computational coding, testing of the algorithm, performance evaluation, and preparation of the manuscript.

Funding

This study did not receive financial support from any governmental, industrial, or non-profit organizations.

Data Availability

The datasets used and analyzed in this study are fully presented within the article.

Conflicts of Interest

The author reports no potential conflict of interest.

References

[1] Charnes, A., Cooper, W. W., & Rhodes, E. (1978). Measuring the efficiency of decision making units. *European journal of operational research*, *2*(6), 429–444. https://doi.org/10.1016/0377-2217(78)90138-8

[2] Andersen, P., & Petersen, N. C. (1993). A procedure for ranking efficient units in data envelopment analysis. *Management science*, *39*(10), 1261–1264. https://doi.org/10.1287/mnsc.39.10.1261

[3] Mozaffari, M. R., Kamyab, P., Jablonsky, J., & Gerami, J. (2014). Cost and revenue efficiency in DEA-R models. *Computers and industrial engineering*, *78*, 188–194. https://doi.org/10.1016/j.cie.2014.10.001

[4] Mozaffari, M. R., Gerami, J., & Jablonsky, J. (2014). Relationship between DEA models without explicit inputs and DEA-R models. *Central european journal of operations research*, *22*(1), 1–12. https://doi.org/10.1007/s10100-012-0273-4

[5] Mozaffari, M. R., Dadkhah, F., Jablonsky, J., & Wanke, P. F. (2020). Finding efficient surfaces in DEA-R models. *Applied mathematics and computation*, *386*, 125497. https://doi.org/10.1016/j.amc.2020.125497

[6] Mozaffari, M. R., Mohammadi, S., Wanke, P. F., & Correa, H. L. (2021). Towards greener petrochemical production: Two-stage network data envelopment analysis in a fully fuzzy environment in the presence of undesirable outputs. *Expert systems with applications*, *164*, 113903. https://doi.org/10.1016/j.eswa.2020.113903

[7] Noura, A. A., Lotfi, F. H., Jahanshahloo, G. R., Rashidi, S. F., & Parker, B. R. (2010). A new method for measuring congestion in data envelopment analysis. *Socio-economic planning sciences*, *44*(4), 240–246. https://doi.org/10.1016/j.seps.2010.06.003

[8] Noura, A. A., Hosseinzadeh Lotfi, F., Jahanshahloo, G. R., & Fanati Rashidi, S. (2011). Super-efficiency in DEA by effectiveness of each unit in society. *Applied mathematics letters*, *24*(5), 623–626. https://doi.org/10.1016/j.aml.2010.11.025

[9] Rashidi, S. F., & Barati, R. (2014). On the comparison of supply chain with SUB-DMUs in DEA. *Advances in environmental biology*, *8*(7), 2387–2390. https://go.gale.com/ps/i.do?id=GALE%7CA385069491&sid=googleScholar&v=2.1&it=r&linkaccess=abs&issn=19950756&p=AONE&sw=w&cookieConsent=true&analyticsOptout=false&userGroupName=anon~3f96368d&aty=open-web-entry

[10] Rashidi, S. F. (2015). Evaluation of productivity indicators in the oil industry by using multi-attribute decision making approach (MADM). *International journal of advanced and applied sciences*, *2*(6), 25–31. https://B2n.ir/kn4618

[11] Aparicio, J., Esteve, M., & Kapelko, M. (2023). Measuring dynamic inefficiency through machine learning techniques. *Expert systems with applications*, *228*, 120417. https://doi.org/10.1016/j.eswa.2023.120417

[12] Guerrero, N. M., Aparicio, J., & Valero-Carreras, D. (2022). Combining data envelopment analysis and machine learning. *Mathematics*, *10*(6), 909. https://doi.org/10.3390/math10060909