**Probability-based Efficiency Analysis through Machine Learning Techniques**

Ricardo González-Moyano1, Juan Aparicio1,2\*, Víctor J. España1 and José L. Zofío3,4

1 Center of Operations Research (CIO). Miguel Hernandez University, Elche, Spain.

2 ValgrAI - Valencian Graduate School and Research Network of Artificial Intelligence, Valencia, Spain.

3 Department of Economics, Universidad Autónoma de Madrid, Madrid, Spain.

4 Erasmus Research Institute of Management, Erasmus University, Rotterdam, The Netherlands.

\* Corresponding author: j.aparicio@umh.es.

**Abstract**

The integration of Data Envelopment Analysis (DEA) and Machine Learning (ML) has paved the way for more robust efficiency assessments of decision-making units (DMUs). Traditional DEA models, while effective, present challenges such as sensitivity to noise, deterministic outputs, and limited ability to account for uncertainty. This paper introduces an innovative classification-based ML approach to enhance DEA by incorporating probabilistic efficiency analysis. Our methodology reformulates DEA as a classification problem, using neural networks to predict efficiency status and estimate the probability of a DMU being classified as efficient. This enables counterfactual-based efficiency benchmarking, where the minimum adjustments required for an inefficient DMU to become efficient can be quantified. A key contribution of this approach is its explainability, achieved through Explainable Artificial Intelligence (XAI) techniques, specifically sensitivity analysis (SA) and counterfactual explanations. We employ variable importance methods to provide targeted efficiency improvement strategies, ensuring that DMUs receive appropriate recommendations based on their specific inefficiencies. Additionally, our framework introduces dynamic peer selection at different probability thresholds, allowing for adaptable benchmarking strategies. The proposed methodology is validated using a real-world dataset from Spain’s food industry, demonstrating its applicability. Results highlight the advantages of this ML-DEA hybrid framework, as a complement of standard DEA, including inferential power, improved discriminatory capacity, and enhanced robustness.

**Keywords:** Data Envelopment Analysis, Machine Learning, Classification models, Neural Networks, variable importance.

# 1. Introduction

In recent decades, the field of efficiency analysis has witnessed significant advancements, particularly in the evaluation of firms, institutions and organizations across various sectors such as finance, healthcare, education, and manufacturing. One prominent methodology that has garnered substantial attention is Data Envelopment Analysis (DEA), initially introduced by Charnes, Cooper, and Rhodes in the late 1970s (Charnes et al., 1978). DEA offers a non-parametric approach to assess the relative efficiency of DMUs by comparing their input-output profiles. The fundamental premise of DEA lies in its ability to evaluate the efficiency of DMUs that operate under multiple inputs and outputs, without imposing restrictive assumptions about functional forms or underlying distributions. This characteristic makes DEA particularly appealing for analyzing complex real-world systems where the relationships between inputs and outputs are likely nonlinear and unknown. Over the years, DEA has been applied to diverse domains, including banking (Berger et al., 1997), healthcare (Olesen et al., 2007), and environmental performance assessment (Zhou et al., 2008), among others.

However, despite its widespread adoption and commendable performance, traditional DEA approaches may encounter limitations in capturing the intricate patterns and structures characterizing involved processes and datasets. One notable challenge lies in the potential for overfitting, wherein the model captures noise or idiosyncratic features in the data rather than true underlying relationships (Esteve et al., 2020). This issue is particularly pronounced in DEA when dealing with high-dimensional datasets or when the number of DMUs is relatively small compared to the number of inputs and outputs, where overfitting is mixed with the curse of dimensionality problem (Charles et al., 2019). Then, DEA can lead to inflated efficiency scores for certain DMUs, thereby distorting the assessment of relative efficiency and potentially misleading decision-makers. Moreover, traditional DEA models rely on linear programming techniques to estimate efficiency scores, which may not adequately capture nonlinear relationships or interactions among inputs and outputs. As a result, the model may overlook certain patterns in the data, leading to biased efficiency estimates. Another significant limitation of traditional DEA is its deterministic nature. Standard DEA models produce a single efficiency score for each DMU based on the observed input-output data, without accounting for the volatile, uncertain, complex, and ambiguous (VUCA) scenarios inherent observed in real-world systems. This deterministic approach fails to acknowledge the stochastic nature of many decision-making processes.

With the advent of machine learning techniques, there exists a compelling opportunity to enhance the capabilities of DEA by exploiting the computational power and flexibility offered by these methods. By integrating machine learning algorithms with DEA, researchers can improve the accuracy, robustness, and interpretability of efficiency assessments, thereby advancing the state-of-the-art in performance analysis. In this context, it is imperative to build the necessary bridges between machine learning and other fields, such as Data Envelopment Analysis. Machine learning algorithms complement DEA by providing advanced techniques for, *inter alia*, data preprocessing (Chen et al., 2014), variable importance measurement (Valero-Carreras et al., 2024), and the treatment of the curse of dimensionality (Esteve et al., 2023), thereby facilitating more accurate and comprehensive efficiency assessments. Moreover, machine learning models can capture nonlinear relationships and interactions among inputs and outputs, addressing one of the key limitations of traditional DEA approaches.

In the literature, several bridges between machine learning (ML) and Data Envelopment Analysis (DEA) have already been established. However, there exist certain gaps that we believe the novel approach introduced in this paper can address. Before mentioning these gaps, we briefly review the main contributions related to ML and DEA. As we are aware, there are two predominant streams of research in the literature that explore the integration of machine learning with Data Envelopment Analysis.[[1]](#footnote-2)

The first stream focuses on adapting existing ML techniques to ensure that the predictive function, typically representing a production function in our context, complies with various shape constraints such as monotonicity or concavity. Researchers in this stream use techniques from ML, such as support vector machines (SVM), neural networks (NN), or decision trees, to develop models that capture the underlying relationships between inputs and outputs by imposing shape constraints on the predictive function. Milestones in this domain are the following: Kuosmanen and Johnson (2010) demonstrated the connection between DEA and least-squares regression, introducing Corrected Concave Nonparametric Least Squares (C2NLS). Parmeter and Racine (2013) proposed innovative smooth constrained nonparametric frontier estimators, incorporating production theory axioms. Daouia et al. (2016) introduced a method using constrained polynomial spline smoothing for data envelopment fitting, enhancing precision and smoothness. Esteve et al. (2020) and Aparicio et al. (2021) developed Efficiency Analysis Trees (EAT), improving production frontier estimation through decision trees. Valero-Carreras et al. (2021) introduced Support Vector Frontiers (SVF), adapting Support Vector Regression for production function estimation. Olesen and Ruggiero (2022) proposed hinging hyperplanes as a nonparametric estimator for production functions. Guerrero et al. (2022) introduced Data Envelopment Analysis-based Machines (DEAM) for estimating polyhedral technologies. Valero-Carreras et al. (2022) adapted SVF for multi-output scenarios, improving efficiency measurement. Guillen et al. (2023a, 2023b, 2023c, 2024) introduced boosting techniques for efficiency estimation in different scenarios. Tsionas et al. (2023) proposed a Bayesian Artificial Neural Network approach for frontier efficiency analysis under shape constraints. Liao et al. (2024) proposed Convex Support Vector Regression (CSVR) to improve predictive accuracy and robustness in nonparametric regression.

The second stream of literature adopts a two-stage approach to integrate DEA with ML techniques. In the first stage, researchers apply a pre-existing DEA model, such as the output-oriented radial model, to compute efficiency scores for each observation in the sample (DMUs). In the second stage, the efficiency scores obtained from DEA are treated as the response variable in a ‘regression’ model based on standard ML techniques (without shape constraints). The original inputs and outputs, along with potentially additional environmental variables, serve as predictor variables in the regression model. By incorporating ML techniques to the performance evaluation framework, researchers aim to develop more robust and accurate predictive models for assessing efficiency. Some of these contributions are the following: Emrouznejad and Shale (2009) explored a novel approach by combining a neural network with Data Envelopment Analysis (DEA) to address the computational challenges posed by large datasets. Liu et al. (2013) compared standard DEA, three-stage DEA, and neural network approaches to measure the technical efficiency of 29 semi-conductor firms in Taiwan. Fallahpour et al. (2016) presented an integrated model for green supplier selection under a fuzzy environment, combining DEA with genetic programming to address the shortcomings of previous DEA models in supplier evaluation. Kwon et al. (2016) explored a novel method of performance measurement and prediction by integrating DEA and neural networks. The study used longitudinal data from Japanese electronics manufacturing firms to show the effectiveness of this combined approach. Aydin and Yurdakul (2020) introduced a three-staged framework utilizing Weighted Stochastic Imprecise Data Envelopment Analysis and ML algorithms to assess the performance of 142 countries against the COVID-19 pandemic. Tayal et al. (2020) presented an integrated framework for identifying sustainable manufacturing layouts using Big Data Analytics, Machine Learning, Hybrid Meta-heuristic and DEA. The paper by Nandy and Singh (2020) presented a hybrid approach utilizing DEA and Machine Learning, specifically the Random Forest (RF) algorithm, to evaluate and predict farm efficiency among paddy producers in rural eastern India. Zhu et al. (2021) proposed a novel approach that combines DEA with ML algorithms to measure and predict the efficiency of Chinese manufacturing companies. Jomthanachai et al. (2021) proposed an integrated method combining Data Envelopment Analysis and Machine Learning for risk management. Boubaker et al. (2023) proposed a novel method for estimating a common set of weights based on regression analysis (such as Tobit, LASSO, and Random Forest regression) for DEA to predict the performance of over 5400 Vietnamese micro, small and medium enterprises. Amirteimoori et al. (2023) introduced a novel modified Fuzzy Undesirable Non-discretionary DEA model combined with artificial intelligence algorithms to analyze environmental efficiency and predict optimal values for inefficient DMUs, focusing on CO2 emissions in forest management systems. Lin and Lu (2024) presented a novel analytical framework utilizing inverse Data Envelopment Analysis and ML algorithms to evaluate and predict suppliers' performance in a sustainable supply chain context. Omrani et al. (2024) evaluated the efficiency of electricity distribution companies (EDCs) from 2011 to 2020 using a combination of DEA, corrected ordinary least squares (COLS), and machine learning techniques. In particular, a three-stage process involving DEA, COLS, support vector regression (SVR), fuzzy triangular numbers, and fuzzy TOPSIS methods are employed, revealing trends in EDC performance and identifying areas needing improvement.

Both streams of research have contributed valuable insights and methodologies for integrating ML with DEA. However, despite these advancements, important questions that offer opportunities for further improvement and innovation remain unaddressed. In this regard, our approach introduces novel functionalities and complementary methodologies to traditional DEA-based analysis by taking advantage of the favorable properties exhibited by classification models in machine learning. Next, we outline the key contributions of our approach, highlighting its methodological innovations, interpretative advantages, and practical implications for efficiency measurement:

* *Methodological Innovation*: As we are aware, for the first time in the literature, we propose a classification-based machine learning approach in the second stage of a DEA-ML hybrid framework, moving beyond the conventional regression-based techniques. In the first stage, we rely on the concept of Pareto-dominance to distinguish between two classes of DMUs: efficient and inefficient. A DMU is Pareto efficient if and only if it is impossible to improve any input or output without worsening some other input or output. We show that this initial classification can be effectively performed with the standard additive DEA model. However, in contrast to regression-based methods that, in the first stage, require for each DMU a precise numerical efficiency score—i.e., a real valued dependent variable—that is subsequently regressed against the set of input and output variables, our approach circumvents the challenges and potential inaccuracies associated with predicting exact numerical outcomes. By adopting a robust binary classification framework as the response variable (efficient vs. inefficient), we eliminate the risk of propagating errors inherent in continuous value predictions, thereby providing a clearer and more reliable distinction between the two classes. In the second stage, we apply classification models to predict the probability of being classified as efficient using the input and output variables.
* *Inferential Power with Probability Estimation*: Indeed, one of the key advantages of our classification-based approach is that it enables the estimation of the probability of a DMU being classified as efficient. In our novel approach this probability constitutes the real valued response variable—i.e., the output of a classification method based on machine learning, that enables efficiency measurement. What is utterly relevant, and unlike traditional DEA models that mainly serve as descriptive tools, our framework incorporates the probabilistic perspective, allowing researchers and practitioners to infer efficiency status based on statistical learning principles. This conceptual shift aligns efficiency measurement with modern inferential analytics, bridging the gap between efficiency measurement and probability-based decision-making.
* *Reinterpreting DEA as a Classification Problem*: Our approach allows recasting traditional DEA as a classification problem. First, the DEA technology differentiates the variable space between two categories or regions: technically feasible and technically infeasible production processes. Second, DEA technical efficiency measures allows classifying DMUs into the efficient and inefficient classes through the identification of the Pareto-efficient production frontier, which can be interpreted as the separating surface between the feasible and unfeasible categories. Under this reinterpretation, technical efficiency measures can be seen as quantifying the minimum required input and/or output modifications necessary for an inefficient unit to transition from the inefficient class to the efficient class, coinciding with the separating Pareto-efficient frontier. In contrast, the new method estimates the probability value at which one specific input-output combination belongs to the efficient boundary of the feasible region. As a result, for any pre-defined level of probability, our method allows classifying DMUs into the efficient or inefficient classes.
* *Algorithm-Agnostic Approach for Robust Efficiency Assessments*: A key advantage of our framework is its flexibility in algorithm selection. Unlike conventional DEA-ML models that rely on a specific regression technique, our method is not tied to a particular classification algorithm. This flexibility might eventually allow us to experiment with a wider range of machine learning models—including decision trees, Support Vector Machines, Neural Networks, and ensemble methods—ensuring that the results remain robust and consistent across multiple techniques. Nevertheless, for the sake of simplicity, in this seminal study we focus on Neural Networks.
* *XAI and Counterfactuals*: Another major contribution of our study is the integration of Explainable Artificial Intelligence (XAI), specifically counterfactual methods, into efficiency analysis. Instead of relying solely on conventional efficiency scores, we define technical inefficiency for an inefficient DMU as the minimum changes required in inputs and/or outputs (depending on the selected orientation) to transition with some probability from the inefficient status to the efficient one. These counterfactual-based adjustments can offer an intuitive and interpretable way to assess inefficiency. For illustrative purposes, and given its increasing popularity in efficiency studies, we choose the Directional Distance Function (Chambers et al., 1998) as the reference measure allowing for efficiency improvements through joint input reductions and output increments.
* *Benchmarking with Variable Importance and Directed Projections*: Our methodology also introduces a novel benchmarking approach by offering a ranking of importance of the inputs and outputs, identified through machine learning models. As highlighted in the literature (e.g., Banker and Morey, 1986; Thanassoulis et al., 2015), understanding the relative importance of variables in efficiency assessments is crucial for strategic decision-making. We propose using this information to assign data-driven weights to inputs and outputs, guiding the projection of inefficient DMUs towards more meaningful and customized improvement paths (directional vectors). This is a significant departure from traditional DEA projections grounded on the Directional Distance Function, which often rely on subjective or arbitrary directional vectors (Wang et al., 2019).
* *Target Setting Through Counterfactual Benchmarking*: The benchmarking framework we propose is further enhanced by incorporating probabilistic efficiency thresholds and applying counterfactual analysis to determine the minimum necessary changes in inputs and outputs that would allow a DMU to be reclassified as efficient. This technique not only generates concrete improvement targets but also allows practitioners to prioritize adjustments based on their impact on efficiency classification.
* *Ranking DMUs and Confidence Thresholds*: Expanding on previous works (e.g., Sexton, 1986; Thanassoulis et al., 2008), we propose a novel ranking system for DMUs based on their probabilistic efficiency scores. By exploiting the information provided by directional projections onto the separating surface of the two classes (efficient vs inefficient), we can rank units according to their likelihood of being classified as efficient, providing an interesting evaluation framework compared to traditional DEA-based ranking methods.
* *Proximity-Based Peer Identification*: Finally, our approach facilitates a more refined peer selection process by identifying, for each DMU and at every efficiency probability threshold, the closest efficient benchmark unit. This selection is performed using proximity-based metrics, such as Euclidean distance, ensuring that benchmark comparisons are contextually relevant and practically achievable.

The paper is structured as follows: In Section 2, we provide background information on Data Envelopment Analysis (DEA) and the machine learning techniques we will utilize, (Artificial) Neural Networks (NN). Section 3 introduces our novel approach, which integrates DEA with the classification technique, aiming to enhance efficiency assessment for DMUs. We demonstrate the practical implications of this integration and its implications for decision-making through an empirical example based on SABI (Iberian Balance Sheet Analysis System) in Section 4. Section 5 concludes and points out further research lines.

# 1. Background

This background section provides a concise overview of DEA and the classifying ML technique of choice in this study, corresponding to Neural Networks.

## 1.1 Data Envelopment Analysis

As aforementioned, Data Envelopment Analysis (DEA) is a non-parametric method widely used for evaluating the relative efficiency of a set of observations or decision-making units (DMUs). DEA offers a powerful framework for assessing the efficiency of DMUs transforming multiple inputs into multiple outputs. From a technological perspective, DEA offers great flexibility when modeling the production technology, e.g., convex or non-convex, strong or weak disposability of inputs and outputs, constant or variable returns to scale, etc. (Orea and Zofío, 2019). In what follows we rely on the canonical model characterized by convexity, strong disposability and variable returns to scale. The former feature is particularly suited for analyzing real-world production processes where economies of scale may vary across different units.

The process of evaluating the technical efficiency of *n* observations requires comparing the performance of these DMUs in terms of the *m* inputs that they use, , such as labor, capital, and other resources, to generate *s* outputs , like goods or services. In this notation, input and output vectors for a specific observation *j* are presented in bold typeface. In a conceptual framework, the term ‘technology’ encompasses all feasible input-output combinations. This concept is typically represented as:

.

Based on the observed input-output processes, DEA empirically approximates through the following production possibility set, Banker et al. (1984):

.

The DEA technology allows differentiates the input-output variable space (i.e., the non-negative orthant ) in two regions: technically feasible and technically infeasible production processes. Within the feasible region represented by , numerous technical efficiency measures are available to calculate the technical efficiency of observations⎯for a general definition of these measures see Pastor et al. (2012). Here, we consider one of the most prominent measures, the Directional Distance Function (DDF) proposed by Chambers et al. (1998). For a specific DMU , and associated directional vector  specifying the direction for its projection onto the efficient frontier of , its technical efficiency corresponds to the distance between the observation and the frontier, which is be calculated through the following program:



This program allows classifying DMUs as technically efficient and technically inefficient. A DMU with an efficiency score of zero, i.e.,  (with \* indicating optimality), is considered efficient, signaling that it operates on the efficient frontier of ⎯i.e., the boundary separating the feasible and infeasible regions. That is, it is impossible to simultaneously reduce inputs and increase outputs given the technology. Conversely, a value strictly greater than zero, i.e., , implies inefficiency relative to the reference technology , with a bigger value indicating a worse degree of technical efficiency.

## 1.2 Neural Networks

Here we outline the fundamentals of Neural Networks (NNs) as our machine learning technique of choice to undertake classification tasks, highlighting their versatility, theoretical foundations, and practical implications. NNs represent a cornerstone in the field of machine learning, recognized for their ability to learn complex patterns and relationships from data (LeCun et al., 2015; Goodfellow et al., 2016). By understanding the principles of NNs, which will allow us to determine both the label and the probability of belonging to that label, we can leverage their capabilities to enhance efficiency methods. NNs are inspired by the structure and function of the human brain, comprising interconnected layers of artificial neurons or nodes. These neurons that process input data through nonlinear transformations to learn complex patterns and relationships and predict, in this case, the probability of being efficient. The core principle underlying NNs is the process of propagation, where input data is sequentially passed through multiple layers of neurons, each layer applying a set of weights and activation functions to produce a response variable or output. Through an iterative process known as backpropagation, NNs adjust the weights of connections between neurons based on the error between predicted and actual outputs, thereby minimizing a certain loss function and improving predictive accuracy. In this sense, activation functions play a crucial role in NNs by introducing non-linearity into the model, enabling it to capture complex relationships within the data. Common activation functions include the sigmoidal function, hyperbolic tangent (tanh) function, and rectified linear unit (ReLU) function. Each activation function introduces different properties to the model, influencing its ability to learn and generalize from data.

In a binary classification problem as the one we propose to identify efficient and inefficient observations, the neural network is designed to distinguish between two possible classes: a positive class () and a negative class (). The network processes an input  through multiple layers of neurons, applying weighted transformations and nonlinear activation functions. The hidden layers compute intermediate representations:

,

where  and  are the weight matrix and bias vector at layer , and  is a nonlinear activation function. At the final layer, a single neuron outputs a logit , which is mapped to a probability using the sigmoid function:

.

This probability represents the likelihood that the given input belongs to the positive class . Since there are only two possible outcomes, the probability of the negative class  is simply: .

The performance of NNs hinges on the selection of hyperparameters such as the number of layers, the number of neurons per layer, learning rate, and regularization parameters. Hyperparameter tuning is essential to optimize model performance and prevent issues like overfitting or underfitting.

An illustrative example of the configuration of a neural network in the context of a binary classification problem, with two predictor variables, would consist of two neurons in the input layer, reflecting the number of variables involved in the model. In simplest technology these two variables would be the input and output of the production process. In the response layer, a single neuron would be located to assign the corresponding class to each observation: efficient or inefficient with some probability. Between these layers lies, in this case, one hidden layer, composed of a pre-defined number of neurons, three in this case. Figure 1 depicts the structure of this neural network with a configuration of 2-3-1.

Diagrama

Descripción generada automáticamente

Figure 1. An example of a very simple artificial Neural Network.

Figure 2 illustrates the nonlinear decision boundary (separating surface) generated by the neural network, which partitions the feature space into two distinct regions. One region is associated with class  (i.e., efficient class), while the other corresponds to class  (i.e., inefficient class). This separating surface emerges as a result of the network’s learned transformations, effectively capturing complex patterns in the data that a linear classifier would fail to model. The figure provides a visual representation of how the network adapts to the underlying structure of the dataset (technology), demonstrating its ability to perform nonlinear classification.

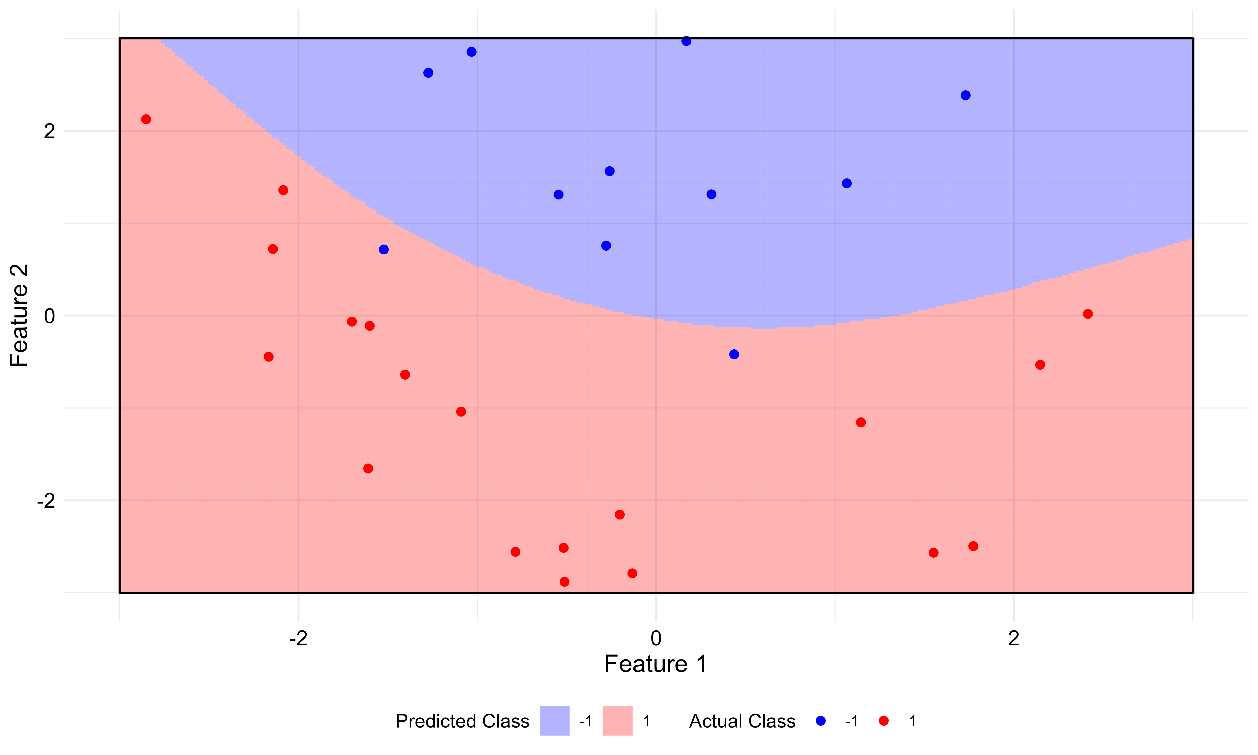


Figure 2. Nonlinear separating surface generated by a Neural Network.

# 2. eXplainable Artificial Intelligence

The so-called eXplainable Artificial Intelligence (XAI) has emerged as a critical area of research aimed at enhancing the transparency, interpretability, and trustworthiness of machine learning models (Wachter et al., 2017). As before, we provide an overview of XAI principles and delve into the concept of counterfactual methods, a subset of XAI techniques that facilitate insightful explanations of model predictions. This will allow us to determine the relative importance of the input and output variables in the prediction of the productively efficient and inefficient classes.

Overall, XAI encompasses a diverse set of methodologies and techniques designed to elucidate the decision-making process of machine learning models, thereby facilitating human understanding and interpretability of model behavior and increasing our trust in the attained results. XAI aims to address this need by not only providing explanations that are understandable, intuitive, and actionable for end-users, stakeholders, and domain experts, but also by enhancing accountability, fairness, and regulatory compliance in AI-driven decision-making. These methodologies play a crucial role in mitigating biases, ensuring ethical AI practices, and enabling more robust model validation. By clarifying the decision-making process of ML models, XAI facilitates the identification of meaningful relationships within data, ultimately leading to more reliable and trustworthy AI systems.

## 2.1 Counterfactual Explanations

Counterfactual methods represent a prominent approach within the realm of XAI, focusing on the generation of alternative scenarios or ‘counterfactuals’ to explain model predictions. The fundamental concept underlying counterfactual methods is the creation of hypothetical instances that are similar to the observed data but differ in one or more attributes. By systematically altering the features of a given instance and observing the corresponding changes in model predictions, counterfactual methods provide valuable insights into the factors driving model decisions and predictions. Moreover, counterfactual explanations offer intuitive and interpretable insights into machine learning models by highlighting the causal relationships between features and model outcomes. These explanations typically take the form of ‘what-if’ scenarios, where adjustments are made to features to generate counterfactual instances that lead to desired outcomes. By identifying the minimal changes required to alter a model prediction, counterfactual explanations shed light on the underlying decision-making process and enable decision-makers to understand the model's behavior in specific contexts. For example, in our production context the question will be ‘What is the minimum amount of adjustment in inputs and/or outputs that a technically inefficient DMU would need to undertake to transition into being considered efficient with a certain probability?’ Thus, the counterfactual method involves projecting an observation from one class onto the separating surface of the two classes, meaning the projection stops just before a change in label occurs. This ‘projection’ strategy is incorporated into our approach to measure technical efficiency in the context of machine learning and efficiency analysis (next Section 3).

## 2.2. Feature Significance Analysis and Sensitivity Analysis

To complement counterfactual analysis, we incorporate feature significance analysis, focusing on understanding the contribution of variables to the model’s predictions. In this regard, several approaches exist for feature significance analysis. One approach is rule extraction methods, which aim to derive interpretable decision rules from complex models (Tickle et al., 1998; Fogel & Robinson, 2003; Martens et al., 2007). Another method involves visualization techniques, which provide graphical representations of feature influence (Craven & Shavlik, 1992; Tzeng & Ma, 2005; Cho et al., 2008). Additionally, a widely used approach is Sensitivity Analysis (SA), which assesses the impact of input variations on model predictions (Saltelli et al., 2008; Sobol’, 1993; Hamby, 1994). Lastly, more recent techniques include SHAP (Lundberg & Lee, 2017), based on cooperative game theory, and LIME (Ribeiro et al., 2016), which builds local surrogate models to approximate feature influence.

Given the variety of feature significance methods, we choose SA for several reasons. First, extraction rules typically simplify the model's complexity to produce more understandable rules, which involves discretizing the classifier, leading to information loss, and failing to accurately represent the original model. Instead, SA is a straightforward method that works with the original fitted model by systematically perturbing input features and measuring the corresponding changes in output, without requiring additional model retraining. Second, visualization techniques are often designed for a specific machine learning method, limiting their general applicability. This is a disadvantage compared to SA, which is model-agnostic and can be applied to any supervised machine learning method. Third, methods like SHAP and LIME pose challenges in high-dimensional datasets. SHAP, for instance, relies on Shapley values, which require exponential complexity in the number of features. Meanwhile, LIME constructs local surrogate models to approximate feature influence, but its reliance on sampling around a given instance can lead to high variance in the explanations, especially as dimensionality increases. This makes LIME’s interpretability less stable in complex settings like production processes. In contrast, SA directly analyzes the original model, mitigating these issues while remaining computationally efficient, simple to implement, and providing clear, actionable insights for decision-makers.

Specifically, SA quantifies feature influence by systematically perturbing input variables (either one at a time or through global variance decomposition) while keeping others fixed at their baseline values. This allows for an intuitive assessment of how variations in individual features influence model predictions. Several forms of SA exist, including local sensitivity analysis, which examines small perturbations in feature values to estimate their immediate effect (Hamby, 1994), and global sensitivity analysis, which explores the entire input space to assess broader variable importance (Saltelli et al., 2008). SA has been widely applied not only in the fieldof neural networks (Dimopoulos et al., 1999), but also finance (Saltelli et al., 2004), and engineering (Song et al., 2015) due to its flexibility and efficiency. In Machine Learning, SA has been particularly useful for improving the interpretability of predictive models. Ruck et al. (1990) introduced gradient-based SA for neural networks, estimating feature relevance by computing partial derivatives of the output with respect to each input feature. Later, Dimopoulos et al. (1999) expanded this approach, integrating perturbation-based methods to assess variable importance in neural networks. Gevrey et al. (2003) further compared multiple SA techniques, evaluating their advantages and limitations in complex models. These contributions have established SA as a key tool for feature significance analysis in Machine Learning. Despite its advantages, it is important to note that SA assumes feature independence, meaning that in cases where input variables are highly correlated, results may be misleading (Li et al., 2010). However, this limitation can often be mitigated through careful preprocessing or complementary analytical techniques.

To summarize, we employ local SA to assess feature significance in Machine Learning models, emphasizing computational efficiency and interpretability. By perturbing one variable at a time while keeping others fixed, we obtain a direct and isolated measure of feature influence, avoiding the added complexity of global methods. Furthermore, our approach does not rely on gradient computations, making it applicable to a broader range of models. This design choice ensures a scalable, interpretable, and robust framework for feature importance assessment.

# 3. Integrating ML techniques for classification and Data Envelopment Analysis

In this section, we perform the integration of machine learning techniques for classification tasks with Data Envelopment Analysis (DEA) to enhance the measurement of technical efficiency. By combining the strengths of both methodologies, we aim to provide robust and insightful efficiency assessments of a set of DMUs. As previously mentioned, while other ML classification methods could be considered, we focus on Neural Networks.

## 3.1. Reinterpreting DEA as a Classification Machine Learning Technique and its way to measure technical efficiency as a XAI method.

Before introducing the methodology, we illustrate the reinterpretation of DEA as a classification method that also resorts to counterfactual analysis when measuring technical efficiency. First, following Figure 3, the DEA technology differentiates the input-output space into feasible and infeasible regions. Second, DEA technical efficiency measures can be conceptualized as a classification model within the feasible category wherein the two classes represent efficient and inefficient input-output production processes, with the efficient units positioned precisely onto the boundary of the feasible-unfeasible regions, i.e., the Pareto-efficient frontier. As a classification method within the feasible region, technical efficiency measurement implies that the typical efficiency measures utilized in DEA can be reinterpreted within the realm of eXplainable Artificial Intelligence (XAI) principles, particularly in relation to the notion of counterfactual scenarios. Specifically, the movement of an inefficient DMU, by improving its observed inputs and/or outputs in accordance with the orientation and type of efficiency measure selected (e.g., using the Directional Distance Function defined by ), signifies transition within its original class 'feasible' toward the exact threshold where any further minimal change would result in the unit being classified as 'unfeasible' (through its projection onto the efficient frontier, i.e., the separating surface). This movement resembles a counterfactual that quantifies the level of technical inefficiency within the ‘feasible’ class through DEA, thus highlighting the conceptual linkage between DEA and XAI principles.

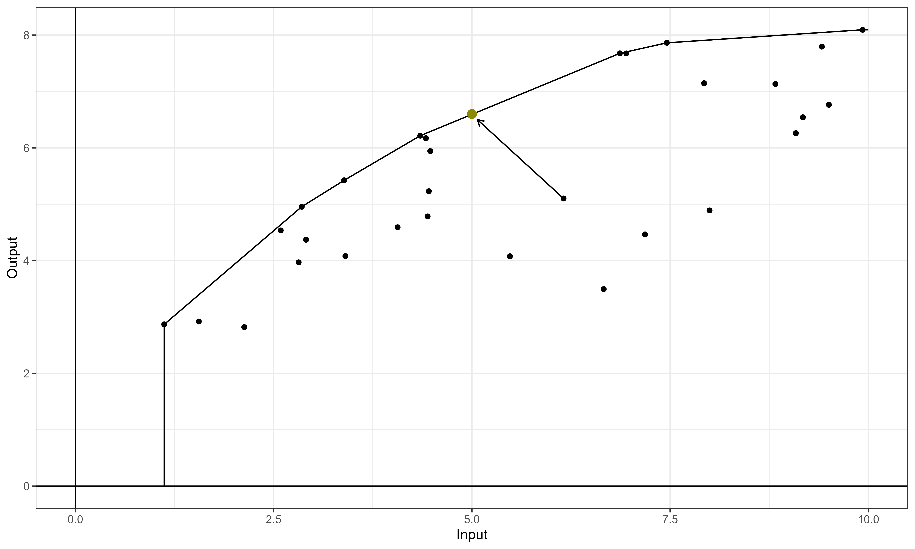


Figure 3. Data Envelopment Analysis and the Directional Distance Function.

After drawing a parallel between standard DEA approaches and classification ML methods, showing that DEA efficiency measures can be considered as a specific case of XAI, we now proceed to introduce the proposed methodology.

## 3.2. Classifying DMUs by their (in)efficiency class and measuring technical efficiency

The core concept underlying our model is a multi-stage methodology aimed at enhancing efficiency assessment through the fusion of DEA and ML techniques. The approach operates in three distinct phases: Firstly, we employ standard DEA to categorize DMUs into efficient and inefficient categories. Subsequently, in the second phase, we address the challenge of class imbalance (efficient vs inefficient). In the third phase, we employ a classification ML model, wherein the response variable is the efficiency status, and the classification features include both inputs and outputs. Finally, in the fourth phase of our approach, we ascertain a robust measure of technical inefficiency through the application of XAI. Specifically, given a model measuring technical efficiency (such as the DDF model ), we determine the required input reductions and output expansions of each inefficient DMU to transition its class from inefficient to efficient.

Next, we introduce our approach in the form of an algorithm with different steps:

**Step 1 [Data labeling process]:** Based on the concept of Pareto-dominance, we resort to the additive DEA model below (Charnes et al., 1985) to partition the set of DMUs into two categories: efficient vs. inefficient. A value of zero in the optimal solution of the linear program indicates that the evaluated DMU is not Pareto-dominated by any technically feasible input-output combination within the DEA technology . This condition underscores that for the DMU under evaluation there is no room for enhancing its input and/or output combination without compromising its technological feasibility (i.e., both input slack reductions and output slack increments are infeasible:  =  = 0, ∀*i*, *r*).



Consequently, if , then DMU  is (technically) inefficient. The set of all inefficient DMUs is denoted as . Otherwise, if , then DMU  is (technically) efficient. The set of all efficient DMUs is denoted as .

**Step 2 [Class balancing phase]**: Addressing the challenge of class imbalance between DMUs (efficient and inefficient) is crucial for prediction by means of ML techniques (see, for example, He & Garcia, 2009). Imbalanced datasets often compromise the performance of standard algorithms, favoring the majority class and neglecting the minority class. In our production context, large datasets typically exhibit a higher proportion of inefficient units, which can skew model outcomes and adversely affect the accuracy of predictions. To address this issue, we adopt a modified version of the Synthetic Minority Oversampling Technique (SMOTE) (Chawla et al., 2002) to generate synthetic examples of the minority class (efficient units).

However, our approach goes beyond simply balancing class proportions. Rather than just addressing class imbalance, we focus on refining the delimitation of the best-practice efficient frontier, allowing the model to learn more effectively while minimizing biases introduced by the balancing process. Depending on the dataset structure, we generate either efficient or inefficient synthetic units until the desired balance threshold is met, but not both simultaneously. If the minority class is composed of efficient units, we distribute the synthetic units along the entire frontier, ensuring that the model has enough references to properly define the efficiency boundary. If additional data are still needed to reach the desired proportion, we generate extra units randomly along the frontier. Conversely, if the minority class consists of inefficient units, we first calculate their inefficiency scores and then distribute the synthetic inefficient units evenly across the corresponding quartiles. This approach prevents an artificial increase in density within a specific region of the feature space⎯technology⎯while ensuring a more accurate representation of the efficiency frontier. Finally, by evaluating the model's performance using a validation dataset that was not used during training, we further mitigate potential biases, as a biased model will exhibit lower performance when confronted with new data. This adaptation allows us to tailor the synthetic data generation process to better fit the characteristics of our dataset and context. Next, we describe the specific implementation process of our adapted approach to generate synthetic units.

First, we determined the necessary number of synthetic units to balance the proportion of units in both classes (efficient vs. inefficient). Since there is no universally optimal ratio, Weiss and Provost (2003) suggest testing different minority proportions to identify the most effective distribution for the training set. Considering the minimization of the classification error, they conclude that the ideal proportion of minority class should ideally fall between 20% and 40%. Following this approach. We treat this proportion as a hyperparameter, denoted by . To achieve this balance we adopt two different strategies depending on the dataset’s observed class distribution and the selected minority class proportion .

*Case 2a (minority class: efficient DMUs):* As the proportion of the originally observed efficient DMUs does not reach the balance level , it is necessary to increase their number. Specifically, we generate convex combinations from sets of  DMUs labeled as efficient in Step 1. The objective of this approach is to populate complete faces of the convex polyhedron. The total number of combinations is calculated as , where  is the number of efficient DMUs labeled as efficient in Step 1 and  is the total number of variables considered in the problem. For each combination of observed efficient DMUs in the dimensions, a synthetic unit is generated by applying the same weights to all DMUs involved in that linear combination. The weight is defined as , chosen arbitrarily to position the synthetic units at the midpoint of the convex combination. Note that the by choosing efficient DMUs in all input and output dimensions, , we ensure that the reference frontier used to create the synthetic DMUs corresponds to full-dimensional facets. Once all convex combinations have been created, we use the additive DEA model to identify which of these combinations are Pareto-efficient. If the number of efficient units remains insufficient, additional random DMUs are generated based on combinations where we found an efficient synthetic point. In this process, the weights for the observed efficient DMUS are randomly selected within the open interval (0, 1). To maintain consistency and ensure that the sum of all weights equals 1, each weight is normalized by dividing it by the total sum of all weights, yielding a new relative weight for each DMU. When the pre-fixed balance level is achieved, the generation of synthetic units stops. If none of the  combinations yield a Pareto-efficient point, we proceed by selecting combinations of  efficient DMUs. If this approach also fails, we reduce the number to  and continue iterating in this manner until a solution is found.

*Case 2b (minority class: inefficient DMUs):* Now it is necessary to create additional synthetic inefficient DMUs. The process consists of four stages. First, as in the previous case, synthetic convex combinations from efficient DMUs are obtained using equal weights. However, rather than identifying and keeping those combinations that are efficient, we select those that are inefficient. Therefore, in the second stage we use the additive DEA model to determine which of these combinations are inefficient. Third, and separately, a large random sample of convex combinations of the initially identified inefficient DMUs is generated—for example, 20 times the desired number of synthetic inefficient units—is selected. Fourth, based on a quantile choice (e.g. quartiles, quintiles, etc.) of the originally observed distribution of inefficiency scores,  > 0, equally numbered subsets of synthetic units by quantiles is randomly selected until the targeted balance proportion is achieved. This systematic approach ensures a well-distributed (populated) set of synthetic inefficient DMUs by maintaining representativeness across different inefficiency levels.

**Step 3 [Fitting the ML model]:** In this phase, a classification-based ML model⎯NN in this study⎯ is implemented, where the dependent variable denotes the efficiency status (efficient [class +1] vs. inefficient [class -1]), while the independent variables (features) comprise all inputs and outputs. Additionally, model parameters are fine-tuned at this stage. Specifically, these parameters are determined by the minority class proportion  and the hyperparameters of the selected methodology .

The training phase consists of two stages. In the first stage, we use *k*-fold cross-validation to determine the optimal parameter configuration for the NN model at each  level . In the second stage, we compare the performance of the best-trained models to identify the ideal balance proportion between efficient and inefficient DMUs. Finally, the best classification model  is obtained.

To measure the performance of a specific configuration , we can use standard metrics commonly used in ML classification problems. Our primary focus is on metrics related to the 'efficient' (minority) class, such as sensitivity (i.e., the proportion of actual positives correctly identified), precision (the proportion of positive predictions that are actually correct), F1 score (the harmonic mean of precision and sensitivity, balancing detection accuracy and reliability), and balanced accuracy (the average of true predictions for each class). Depending on the specific context and the metrics of interest, we evaluate model performance by ranking the results according to our chosen criteria.

If the original dataset is sufficiently large, we propose creating training and validation partitions to evaluate the model’s performance with data not used during the fitting phase. This approach helps mitigate overfitting and provides a reliable estimate of the model’s real-world performance when applied to new data. However, if the dataset is too small to allow for such partitioning, we suggest testing the models trained with different balance levels on the original dataset. This ensures that all models are evaluated on the same dataset, allowing for a fair performance comparison. If multiple models achieve the same performance on the original dataset, only the tied models will be further evaluated based on their performance when applied to their respective balanced datasets, whose sizes vary depending on the balance level. This restriction prevents the selection of models that may overfit to highly augmented datasets while performing poorly on real data. If equality persists, the smallest  level is selected, following the principle of parsimony (as fewer synthetic observations are required to reach ).

Finally, the best classification model  predicts the probability of being technically efficient and classifies the input-output bundle  as technically efficient (+1) or inefficient (-1) through the following rule: if , then  is classified as efficient; otherwise,  is classified as inefficient.

**Step 4 [Measuring technical inefficiency]:** The final step is to measure technical efficiency using a standard technical efficiency measure. For example, the popular DDF presented in model , which simultaneously captures improvements across input-output dimensions by reducing inputs and increasing outputs. The directional vector is defined as  and can be interpreted as the relative weight assigned to inputs and outputs when measuring inefficiency. Each element of  represents the weight capturing the relative importance of input *i* multiplied by its respective observed mean, and each element of  represents the relative importance of output *r*, multiplied by its respective observed mean. The relative importance of each variable is calculated through the SA method (see Section 2.3.2). As we are aware, the selection of this type of directional vector is also original (Wang et al., 2019). This choice of directional vector is particularly noteworthy for several reasons. First, incorporating the relative importance of input and output variables—derived from their contribution to predicting efficiency status—introduces a novel and objective criterion for defining the direction of improvement. Unlike traditional approaches that often rely on subjective or uniform weightings, this method exploits data-driven insights obtained through sensitivity analysis. This could result eventually in inputs or outputs being dropped from the directional vector evaluating inefficiency if they do not contribute to the predictive status of the observations. Second, using the same directional vector to evaluate inefficiency across all DMUs ensures that the inefficiency values⎯optimal ’s⎯remain comparable across units. This uniformity enhances the interpretability of inefficiency scores, as differences in inefficiency are not influenced by unit-specific directional choices but rather reflect genuine performance gaps relative to a common yardstick. Third, by constructing the directional vector based on the mean values of inputs and outputs, we ensure that the measure is inherently dependent on the original units of measurement. This property makes the inefficiency measure units’ invariant, meaning that results remain consistent regardless of the scale or units in which the inputs and outputs are expressed. This avoids potential distortions that could arise from differences in variable magnitudes or measurement scales.

Next, the value of technical inefficiency is determined using counterfactual analysis. In particular, we answer the question: “What is the minimum modification required for the evaluated DMU to be classified as efficient with at least probability *p*?”. The corresponding projection (input-output targets) is calculated through where  and  are the observed inputs and outputs for , respectively, and  takes values within a predefined grid:

.

If the target probability *p* lies between two consecutive  values, we refine the search by iteratively testing new  values within that range and recalculating their associated probabilities . When the algorithm converges to the probability that meets the desired confidence threshold, the corresponding  value is recorded as the minimum adjustment required for the DMU to achieve technical efficiency. Given a predefined threshold *p*, the benchmark for each evaluated DMU is identified as the closest unit⎯e.g., based on the Euclidean distance—classified as efficient at level *p*. This selection ensures that the chosen peer represents the most comparable efficient unit, providing a meaningful benchmark for performance evaluation. By resorting to proximity within the input-output space, the approach facilitates a more intuitive interpretation of efficiency adjustments required for the assessed DMU to reach the predefined efficiency level. Note that peers can change for each DMU depending on *p*. Formally, for DMU , and given probability threshold *p*, its corresponding peer is determined as . In case of a tie, we report the group of peers and randomly select one of the possible solutions.

Additionally, if in the evaluation process, the efficiency probability of an observed DMU exceeds the pre-fixed threshold *p*, its is set to 0 as it is ‘superefficient’ at the selected probability level. This implies that we do not calculate negative ’s increasing inputs and reducing outputs to project the DMUs ‘backwards’ towards lower level probability frontiers⎯although it would be possible following Anderse and Petersen (1993). Finally, when searching for the potential benchmarks our algorithm does not consider projections with input values smaller than the minimum observed across the sample, as it could result in zero valued inputs, which is unrealistic. As a result, some DMUs could never reach the established threshold *p*. In such cases, we consider  the biggest feasible  in the considered grid.

## 3.3. An illustrative example

Next, we illustrate our method using NNs as ML classification technique of choice through a numerical example, complemented by several figures.

In this example, we consider a simulated dataset  comprising 40 DMUs, where each  employs a single input  to produce a single output . In accordance with the proposed algorithm, Step 1 assigns efficiency labels by solving the additive model via standard DEA. Specifically, a  is classified as efficient if and only if all its (optimal) slacks in model are zero. In our simulated dataset, only 3 DMUs are Pareto-efficient, namely DMUs 6, 7 and 31 (see Figure 4), while the remaining DMUs are inefficient. Consequently, the label distribution is high imbalanced, with approximately 7.5% of the DMUs classified as efficient and 92.5% as inefficient.

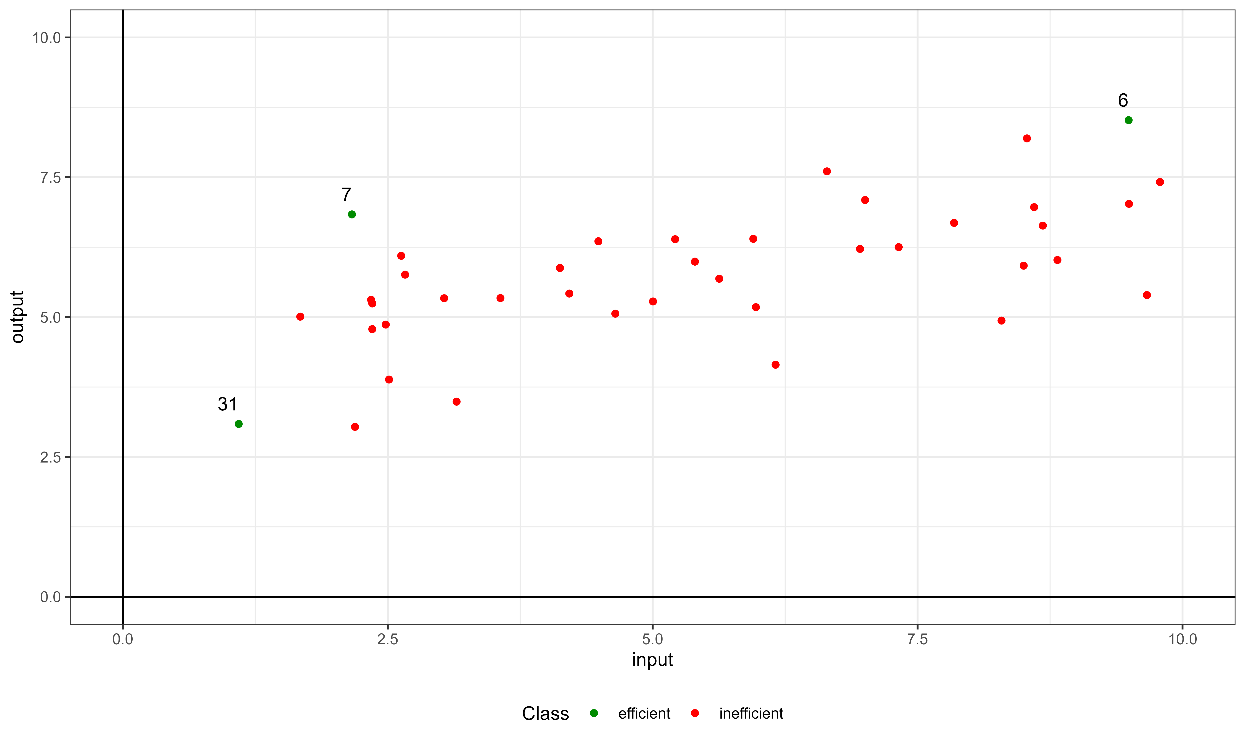


Figure 4. Labeling through the standard DEA additive model.

Step 2 addresses the dataset imbalance by generating synthetic observations for the efficient minority class in this example. Synthetic efficient units, denoted by , are generated as linear combination of the 3 efficient DMUs until the total number of efficient units reaches 13. This augmentation results in a new dataset, , where the efficient DMUs represent approximately 25% of the observations. This particular balance percentage is merely illustrative of the outcomes obtained through the balancing process. In practice, we use the neural network to determine the optimal proportion of efficient units, typically ensuring that the minority class constitutes between 20% and 40% of the dataset.

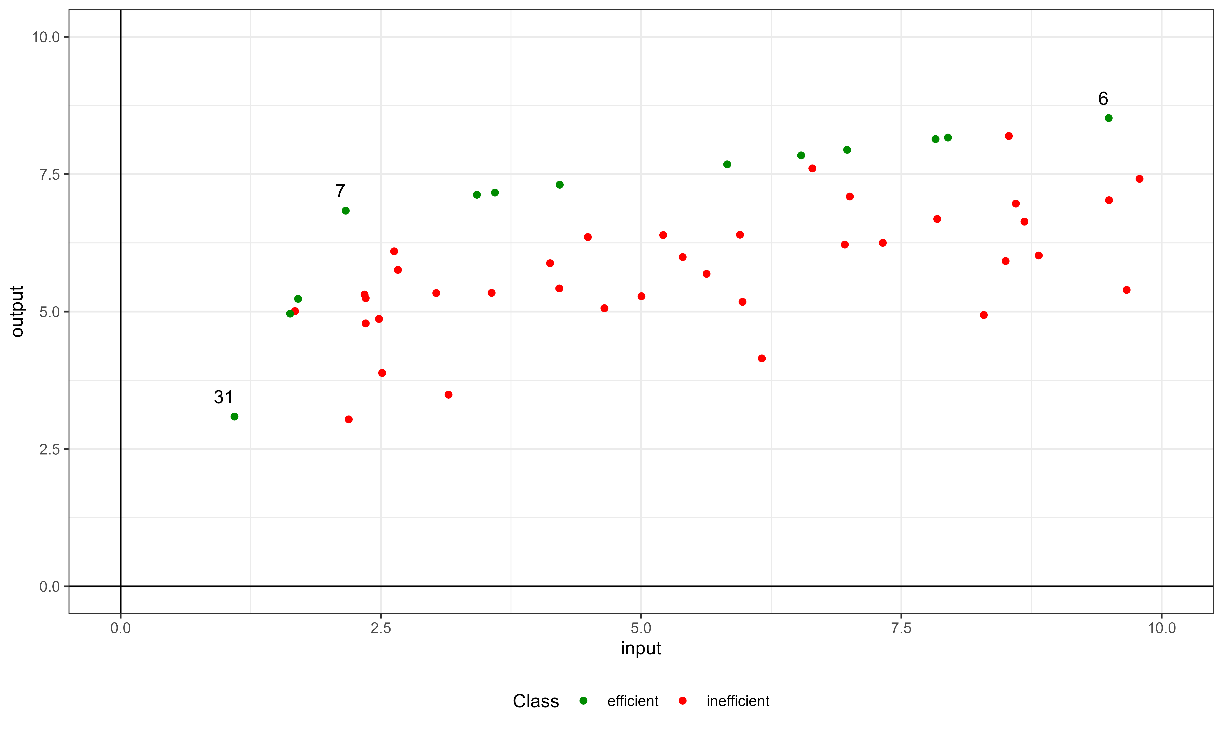


Figure 5. The labeled dataset that will be used for model training.

The third step involves training the machine learning model. Seeking simplicity, we employ a NN with an initial (input) layer containing as many nodes as input and output variables (represented in in Figure 1 by I1 (input) and I2 (output)), a single hidden layer (H1), and a final exit (output) layer consisting of a single neuron (O1). In practice, we use the *R* package ‘*caret*’(Kuhn, 2008) to facilitate model training, specifically using the NN implementation from the ‘*nnet*’ package (Venables & Ripley, 2002).

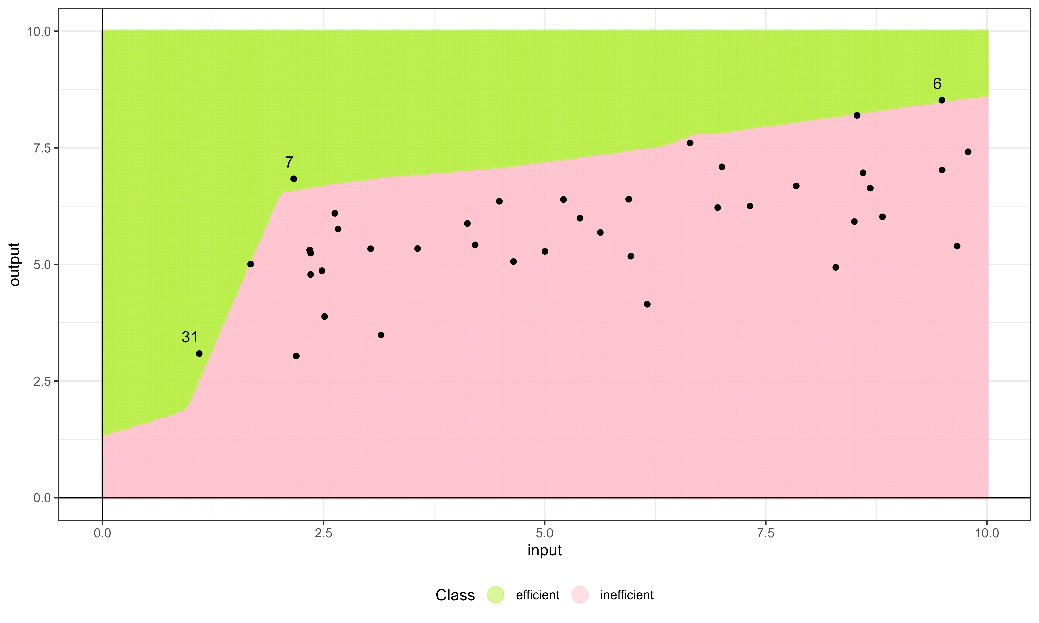
1. First, we define the different levels of the minority class as . Then, for each value in , we construct a grid of possible values for two hyperparameters: (i) model fitting size , representing the number of neurons in the hidden layer and can take the following values , and a decay parameter , acting as a regularization constrained, which can take values . To determine these hyperparameters, we perform 5-fold cross-validation. After that, we evaluate the performance of each hyperparameter configuration  for a given minority class proportion using the full dataset . Due to the limited sample size of , we do not use in this example a separate validation set. Instead, performance is assessed directly on the observed data, ensuring consistency across all models.

Table 1 presents the performance of the fitted models for different values of , using standard metrics commonly applied in ML problems as defined in Step 3. Although the choice of metrics for evaluating the performance of an ML model is arbitrary and depends on the specific characteristics of the problem, we propose the following approach for our specific efficiency analysis. First, we select the model with the highest accuracy rate while considering both classes, using balanced accuracy. If multiple models achieve the same balanced accuracy, we then evaluate them based on metrics that prioritize the minority 'efficient' class, which is our primary interest. Specifically, we use the F1-score, as it balances detection and confidence through the harmonic mean. If a tie persists, we further assess models based on confidence using precision, and finally, on detection using sensitivity. Since performance for all metrics is tied for minority class distributions of 25%, 35%, and 40% (see upper panel of Table 1. Stage 1: Performance using real dataset), we further evaluate performance using the augmented dataset  for these balance levels, which included both real and synthetic units⎯i.e. used as train dataset for each balance level. After this evaluation, configurations with minority class proportion of 25% and 40% remained tied (see lower panel of Table 1. Stage 2: Performance using train dataset). Hence, following the principle of parsimony, we select the dataset with a minority class proportion of 0.25.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Stage 1: Performance using real dataset (*D*) | | | | |
|  | Balance accuracy | F1 | Precision | Sensitivity |
| 0.25 | 1 | 1 | 1 | 1 |
| 0.35 | 1 | 1 | 1 | 1 |
| 0.4 | 1 | 1 | 1 | 1 |
| 0.3 | 0.99 | 0.86 | 0.75 | 1 |
| 0.2 | 0.83 | 0.8 | 1 | 0.67 |
|  |  |  |  |  |
| Stage 2: Performance using train dataset  (only tied balance levels) | | | | |
|  | Balanced accuracy | F1 | Precision | Sensitivity |
| 0.25 | 1 | 1 | 1 | 1 |
| 0.4 | 1 | 1 | 1 | 1 |
| 0.35 | 0.99 | 0.98 | 0.95 | 1 |

Table. 1 Performance results for each fitted model across different class balancing levels, ranked from highest to lowest performance.

1. Once the optimal set of hyperparameters has been selected following the above method (i.e., ,  and ), the NN model is retrained on the augmented dataset  without cross-validation. Figure 6 (on the left) presents the results of this final adjustment. The observed DMUs are represented as black points. Additionally, two distinct regions are identified choosing 0.5 as the probability threshold to differentiate among the efficient and inefficient classes. The probability level *p* can be modified at will when representing the areas corresponding to the two classes. In this case we identify the green region with input/output production processes to which the model assigns a probability greater than 0.5, classifying these units as “efficient”, and the pink region, where the probability is 0.5 or lower, classifying units as “inefficient”. Moreover, Figure 6 (on the right) represents uncertainty through a gradient based on the probabilities predicted by the model. Areas representing units with a predicted probability between 0.25 and 0.75 are shaded in black, indicating high uncertainty about their efficient or inefficient status, while regions with probabilities closer to 0 or 1 are displayed in white, reflecting greater certainty about their class.



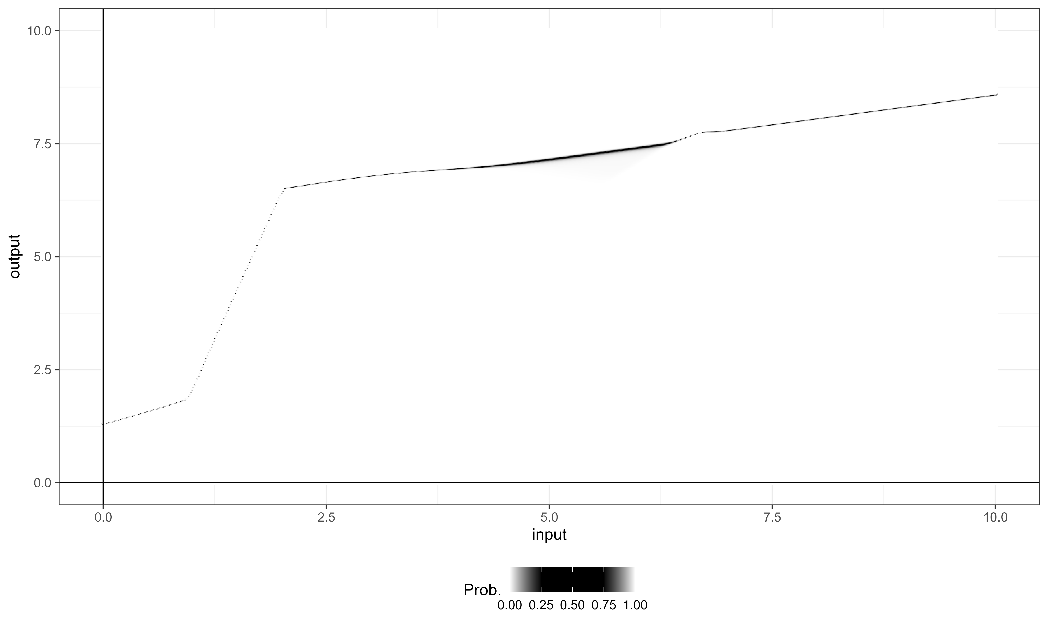


Figure 6. At the top, the predicted (efficient/inefficient) regions generated by the new approach are displayed alongside the original unlabeled DMUs. At the bottom, the uncertainty regions are shaded in black as predicted by the fitted model, with certainty regions shown in white.

Once the final model is established, we use it to measure technical efficiency in step 4. First, we perform a SA analysis (see Section 2.3.2) using the ‘*Rminer*’ library (Cortez, 2010) to determine the directional projection vector for the DDF model defined in . Specifically, the SA results are  for inputs and  for outputs. This analysis reveals that the model assigns twice as much importance to the output variable as to the input variables when classifying a DMU as efficient or inefficient. We then define the directional vector as , where  and  represent the mean value of input  and output , respectively. Lastly, we determine  (the efficiency score derived from the DDF model), along with the input and output targets and the peers for each DMU.

Next, we illustrate the measurement of inefficiency at a 0.75 confidence level. Figure 7 displays the separating hyperplane, highlighting DMUs 6, 7 and 31⎯the three DEA efficient DMUs⎯ with a probability of being efficient greater than 0.75, and located in the green area (*p*>0.75). Additionally, the DDF projection of DMU 22, classified as inefficient with an input value of 4.49 and an output value of 6.36, is shown. Using the previously defined directional vector, we calculate the value of  required for DMU 22 to reach the specified efficiency confidence level (*p=*0.75). As a result, its projection reduces the input to 4.17 (its input target) and increases the output to 7.03 (its output target), corresponding to a . From a benchmarking perspective, DMU 22 has DMU 7 as its peer at a 0.75 confidence level. This selection is based on the Euclidean distance, where DMU 7 is the closest among the originally observed efficient units (6, 7, and 31) to the projection of DMU 22 on the efficient frontier.

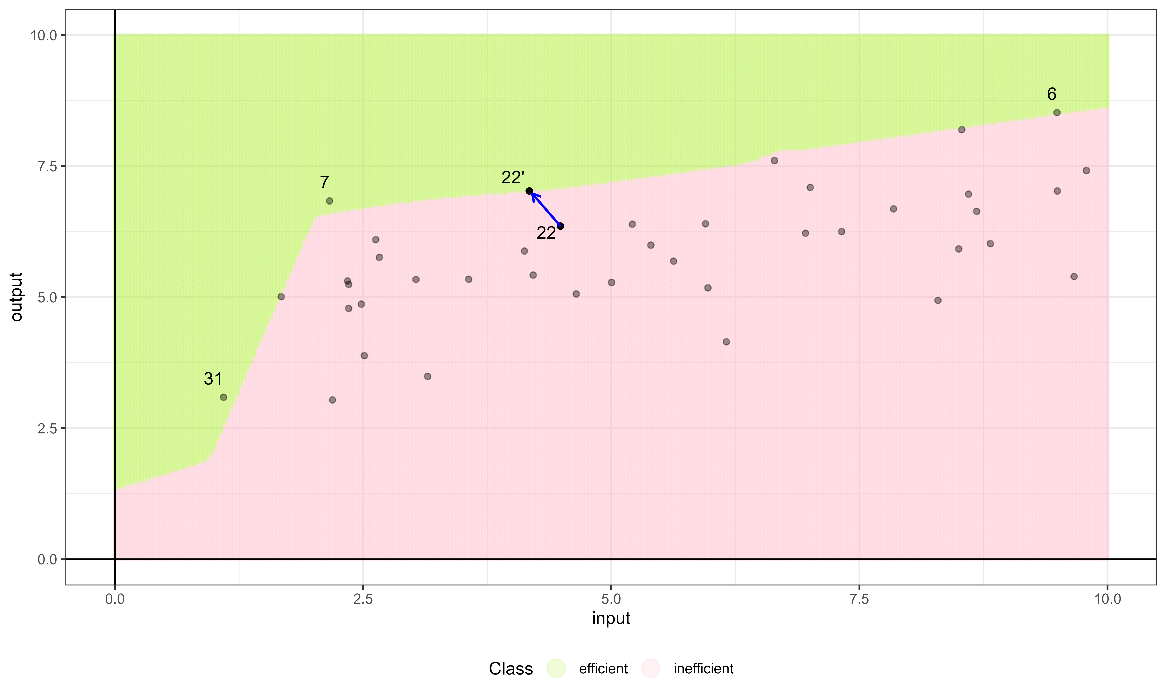


Figure 7. Projection of DMU 22 at an efficiency confidence level of p=0.75.

# An empirical application: Efficiency Assessment of the Valencian Food Industry

In this section we show that the new approach can be empirically implemented using real-world data of firms operating in the Spanish food industry, a sector that plays a crucial role in the country's economy. The food industry in Spain is both economically significant and culturally rich, seamlessly combining traditional practices with modern technological innovations. Its scope covers the entire food value chain, transforming raw agricultural products into a wide variety of food items consumed domestically and exported internationally. This industry is supported by a diverse ecosystem, ranging from small-scale farmers dedicated to preserving heritage techniques to large companies adopting advanced industrial production systems. It is also a vital source of employment, spanning agriculture, processing, distribution, and retail. Numerous studies worldwide have analyzed efficiency in food industries, with examples including India (Kumar and Basu, 2008), Mexico (Flegl et al., 2022), Taiwan (Dadura and Lee, 2011), and Indonesia (Machmud et al., 2019). Such analyses provide valuable insights into the operational dynamics of food sectors across different regions.

In Spain, the institutional and economic environment is shaped by its 17 autonomous communities, each characterized by distinct policies and market conditions. This regional diversity introduces significant complexity into any analysis, as variations in regulations and management frameworks influence business operations at the community level. The Valencian Community, selected as the focus of this study, exemplifies such diversity. Known for its strong agricultural exports and medium-sized enterprises, this region provides a representative case for evaluating efficiency in the Spanish food industry.

The dataset used for this analysis consists of 97 food industry companies located in the Valencian Community, each employing more than 50 workers, and collected from the SABI database for year 2023.[[2]](#footnote-3) The dataset includes several variables that comprehensively reflect the operational and financial profiles of the companies. The output variable, operating income (in millions of Euros), captures revenue generated from core business activities. Input variables include total assets (in millions of Euros), representing the resources utilized; the number of employees, indicating workforce size; tangible fixed assets (in millions of Euros), such as buildings and machinery essential for production; and personnel expenses (in millions of Euros), encompassing costs like salaries, benefits, and training. Together, these variables enable a detailed examination of resource allocation, labor engagement, and financial investments, forming the basis for a robust analysis of operational efficiency within the Valencian food industry (see Table 2). To better understand the characteristics of the dataset and the challenges it presents for analysis, Table 2 presents the descriptive statistics for the sample. Examining the data, we observe that the dataset includes both very small and very large companies. The maximum and minimum values are significantly distant from the mean and median, highlighting the wide dispersion in the data. This dispersion affects the central tendency measures, resulting in a notable difference between the mean and median.

Building upon this production framework, we employ the technique described in this paper, which combines ML techniques for classification and DEA, to determine a robust technical efficiency analysis. This approach should allow us to capture the complex intricacies and idiosyncrasies of the food industry in the Valencian Community, providing a more accurate and contextualized perspective on efficiency.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | Inputs | | | | Output |
|  | Total assets | Employees | Fixed assets | Personnel expenses | Operating income |
| Min. | 1.537 | 50 | 0.142 | 1.037 | 2.382 |
| 1st Qu. | 8.989 | 75 | 2.680 | 2.167 | 12.994 |
| Median | 24.555 | 98 | 6.258 | 3.059 | 29.138 |
| Mean | 41.030 | 201 | 15.280 | 6.757 | 62.307 |
| 3rd Qu. | 52.409 | 240 | 20.096 | 8.244 | 72.688 |
| Max. | 258.825 | 1076 | 140.689 | 36.789 | 460.578 |

Table. 2 Summary statistics of input and output variables.

## 4.1. Balancing the classes in the dataset.

In this dataset on firms , the additive model identifies 15 out of the 97 firms as efficient. After classifying the data, the next step involves balancing the dataset  and fine-tuning the NN. Regarding the balance, since 15.43% of the firms are labeled as efficient, which is a value close to the minimum recommended percentage (20%), we compare the performance of this initial model without balancing to scenarios where the efficient class is balanced between 20% and 40% at increments of 5%. In each scenario, we fine-tune a NN with a unique hidden layer. A grid of selected hyperparameters is defined for model fitting, including model fitting size , decay parameter  and the proposed proportions for the minority class . We follow the same process for selecting the best-performing ML models as described in Section 3.3, where models are first evaluated based on balanced accuracy first, followed by the F1-score, precision, and sensitivity. Table 3 presents the performance of each parameter configuration, with the best results obtained by ,  and . Notably, the second-best performance in terms of balanced accuracy is observed when the dataset is not balanced, with a minority class distribution of 15.43%. However, this model has the worst precision while achieving the highest sensitivity, indicating that it tends to overclassify firms as efficient, resulting in less reliable predictions.

For this dataset, balancing the classes improves precision at the expense of sensitivity when comparing the model trained on the original dataset  to other models trained on balanced datasets , as observed in the datasets with 20% and 40% balance, where precision reaches 1. Regarding sensitivity, the unbalanced model tends to overclassify units as efficient, as we previously mentioned. However, as we introduce balancing, this issue diminishes, achieving a very good performance across all metrics at 20% balance. At this balance level, balanced accuracy remains nearly the same, but the maximization of precision makes the model's predictions more reliable. Nevertheless, further increasing the minority class proportion seems to introduce confusion into the model, making classification more challenging. At 25%, 30%, and 35% balance levels, the performance declines compared to the 20% balance dataset. However, when the dataset reaches a minority class percentage of 40%, the detection of true positives (sensitivity) increases again to 0.93. Additionally, precision reaches 1, meaning that all observations detected by the model are correctly classified in the validation set, with no false positives. The NN trained on the dataset with 40% balance achieves the best performance across all metrics, making it the most consistent model regardless of the evaluation criteria, including Balanced Accuracy, F1-Score, Precision, and Sensitivity.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Performance using real dataset | | | | |
|  | Balanced accuracy | F1 | Precision | Sensitivity |
| 0.40 | 0.97 | 0.97 | 1 | 0.93 |
| 0.15 | 0.94 | 0.75 | 0.60 | 1 |
| 0.20 | 0.93 | 0.93 | 1 | 0.87 |
| 0.30 | 0.85 | 0.79 | 0.85 | 0.73 |
| 0.35 | 0.85 | 0.79 | 0.85 | 0.73 |
| 0.25 | 0.79 | 0.69 | 0.82 | 0.60 |

Table. 3 Performance results of different models depending on balancing levels, ranked by their performance.

## 4.2. Efficiency probabilities and sensitivity analysis

After selecting the best-performing model, we predict the probability of being efficient for each firm. Table 4 ranks the firms in the sample based on their probabilities of being efficient estimated through the fitted NN. We report the top 25 firms, with the last firm showing already a very low probability value, which indicates that the remaining 69 firms are classified as inefficient with certainty. Moreover, A total of 14 DMUs are predicted to have a probability of efficiency exceeding 0.5, one less than initially labelled in the first step.

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Ranking | Firm | Probability of being efficient | *p* = 0.75 | | *p* = 0.85 | | *p* = 0.95 | |
|  | Peer |  | Peer |  | Peer |
| 1 | 2 | 0.9999 | 0 | 2 | 0 | 2 | 0 | 2 |
| 2 | 18 | 0.9998 | 0 | 18 | 0 | 18 | 0 | 18 |
| 3 | 3 | 0.9996 | 0 | 3 | 0 | 3 | 0 | 3 |
| 4 | 17 | 0.9983 | 0 | 17 | 0 | 17 | 0 | 17 |
| 5 | 20 | 0.9962 | 0 | 20 | 0 | 20 | 0 | 20 |
| 6 | 36 | 0.9960 | 0 | 36 | 0 | 36 | 0 | 36 |
| 7 | 46 | 0.9894 | 0 | 46 | 0 | 46 | 0 | 46 |
| 8 | 1 | 0.9868 | 0 | 1 | 0 | 1 | 0 | 1 |
| 9 | 56 | 0.9705 | 0 | 56 | 0 | 56 | 0 | 56 |
| 10 | 62 | 0.9486 | 0 | 62 | 0 | 62 | 0 | 46 |
| 11 | 93 | 0.9441 | 0 | 93 | 0 | 93 | 0 | 46 |
| 12 | 92 | 0.9335 | 0 | 92 | 0 | 92 | 0 | 46 |
| 13 | 9 | 0.9288 | 0 | 9 | 0 | 9 | 0.87 | 3 |
| 14 | 97 | 0.9176 | 0 | 97 | 0 | 97 | 0 | 46 |
| 15 | 25 | 0.4981 | 0.3033 | 17 | 0.3650 | 17 | 0.4815 | 17 |
| 16 | 26 | 0.4909 | 0.4131 | 17 | 0.5025 | 17 | 0.5025 | 17 |
| 17 | 91 | 0.0735 | 0.1005 | 93 | 0.1005 | 93 | 0.1005 | 56 |
| 18 | 22 | 0.0560 | 0.2665 | 18 | 0.2942 | 18 | 0.3480 | 18 |
| 19 | 43 | 0.0549 | 0.1005 | 46 | 0.1005 | 46 | 0.105 | 46 |
| 20 | 85 | 0.0490 | 0.0994 | 93 | 0.1005 | 93 | 0.105 | 46 |
| 21 | 95 | 0.0338 | 0.1986 | 56 | 0.2145 | 56 | 0.2495 | 56 |
| 22 | 83 | 0.0335 | 0 | 92 | 0 | 92 | 0 | 46 |
| 23 | 44 | 0.0183 | 0.0710 | 56 | 0.0824 | 56 | 0.1091 | 56 |
| 24 | 75 | 0.0167 | 0 | 62 | 0 | 62 | 0 | 46 |
| 25 | 94 | 0.0099 | 0.1005 | 93 | 0.1005 | 93 | 0.1005 | 46 |

Table 4. Top 25 DMUs ranked by the probability of being efficient thresholds and their corresponding peers at different thresholds.

Next, we perform SA on the selected model. The relative importance of the variables used by the model to assign efficiency probabilities is presented in Figure 8. Operating income accounts for 50.8% of the total importance, highlighting its dominant role in the model’s decision-making process. Among the input variables, the relative importance is distributed as follows: employees (25.4%), total assets (12.5%), fixed assets (11.2%), and personnel expenses (0.1%). These relative importance results are subsequently used to define the directional vector as  and calculate the  value for each DMU through counterfactual analysis, as explained in Section 3.

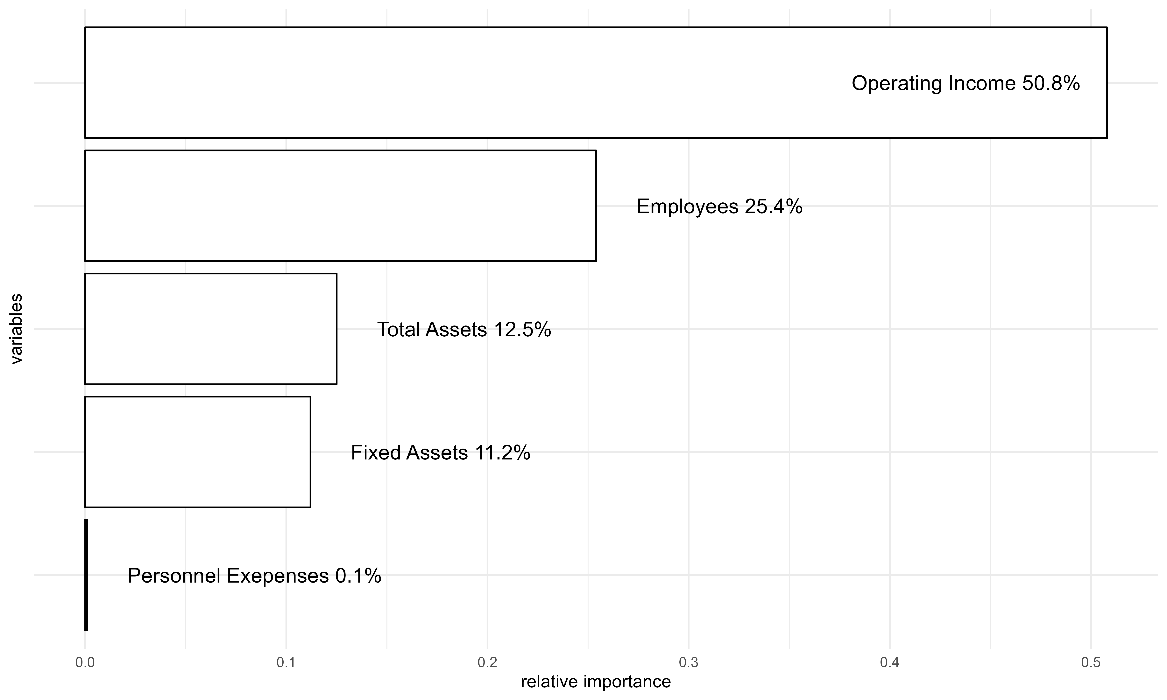


Figure 8. Relative importance of variables, ordered by significance.

## 4.3. Technical efficiency and benchmark peers

We report in Table 4 the inefficiency values and benchmark peers for each firm using three thresholds corresponding to *p* = 0.75, *p* = 0.85 and *p* = 0.95. For example, the top 9 firms are efficient at the 95% level and therefore their ’s are equal to zero. Decreasing the asked probability threshold results in more firms being classified as efficient. The first 14 firms exhibit probabilities very close to 1, being efficient at *p* = 0.90. Meanwhile, firms ranked 15th and 16th can be labeled inefficient because their efficiency probabilities do not reach the threshold *p*=0.5 by less than 1 percentage point. The remaining firms show probabilities close to 0. Notably, firm #26 (ranked 16th), originally labelled as efficient by the additive DEA model , is now classified as inefficient, demonstrating the ability of NNs to overcome the deterministic nature of the traditional approach by leveraging the probabilistic paradigm associated to ML classifying models. In the peer columns, we indicate the reference benchmark for each firm at each probability threshold. If a firm is efficient at a given threshold, it is its own peer. Considering the highest *p* = 0.95 scenario, all inefficient firms identify their reference peer in the top 9, all of which have probabilities exceeding 0.95. Across the entire dataset, all peers belong to the 14 firms that are efficient with a probability greater than 0.5.

Additionally, we compute the main statistics for the efficient projections in each scenario to analyze the inputs and outputs adjustments required to reach the frontier. For the thresholds , a total of 25, 31, and 37 firms, respectively, fail to achieve the established probability level due to the constraints discussed in Section 3—specifically, the restriction that prevents reducing any input below the minimum observed value in each dimension. To further analyze these results, Tables 5, 6, and 7 summarize the mean, median, and standard deviation for these scenarios, highlighting the impact of increasing probability thresholds on the projections (in particular, on the input and output targets). The percentage increments are shown in brackets for better interpretation. According to Table 5, if all firms were to adjust the processes following our directional vector to be considered efficient, total assets would, on average, decrease by 13% (for *p=*0.75) to 15% (for *p=*0.95), employees by 26% to 30%, and fixed assets by 12% to 13%. Operating income would increase by 53% to 61%, depending on the scenario, while personnel expenses remain unchanged, reflecting their negligible role in the adjustments. On average, the probability of being efficient in the observed dataset is 0.15, rising to 0.6, 0.67, and 0.74 in the predicted scenarios. These results depict realistic situations where firms cannot reach the efficient frontier for high thresholds, leading the sector to converge toward a specific probability level below 1. Additionally,  values progressively increase to 1.03, 1.14, and 1.20, reflecting the growing effort required to achieve higher confidence thresholds.

While Table 5 highlights average adjustments, Table 6 focuses on median values, which offer a more representative view of the input and output changes required for the typical firm to reach the efficiency thresholds. For the median firm, total assets would decrease by 22% to 23%, employees by 26% to 31%, and fixed assets by 4% to 5%. Operating income would increase by 52% to 61%, consistent with the mean results. As before, personnel expenses remain unaffected. The larger reductions in total assets observed in the median case suggest that smaller or more typical firms require relatively greater adjustments to achieve efficiency compared to larger DMUs that influence the sector average. Additionally, Table 6 reveals a significant increase in the median probability of efficiency, progressing from 0 in the observed dataset to 0.75, 0.85, and 0.95, underscoring the adaptability of the model. Finally, Table 7 reflects the variability in the inputs reductions and output increases necessary to reach the different efficiency thresholds with respect to the observed values.

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
|  | | | | | | |  |
| Scenario | Observed | *p*=0.75 | | *p*=0.85 | | *p*=0.95 | |
| Total assets | 41.03 | 35.72 | (-13%) | 35.18 | (-14%) | 34.86 | (-15%) |
| Employees | 201.00 | 148.29 | (-26%) | 142.90 | (-29%) | 139.70 | (-30%) |
| Fixed assets | 15.28 | 13.51 | (-12%) | 13.33 | (-13%) | 13.22 | (-13%) |
| Personnel expenses | 6.76 | 6.75 | (0%) | 6.75 | (0%) | 6.75 | (0%) |
| Operating income | 62.31 | 95.05 | (53%) | 98.39 | (58%) | 100.37 | (61%) |
| Probability *p* | 0.15 | 0.60 | (287%) | 0.67 | (333%) | 0.74 | (376%) |
| Beta | 0.00 | 1.03 |  | 1.14 |  | 1.20 |  |

Table 5. Mean values of observed data and projections at different confidence levels, with percentage changes from observed values shown in parentheses.

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
|  | | | | |  |  |  |
| Scenario | Observed | *p*=0.75 | | *p*=0.85 | | *p*=0.95 | |
| Total assets | 24.56 | 19.12 | (-22%) | 19.05 | (-22%) | 18.88 | (-23%) |
| Employees | 98.00 | 72.66 | (-26%) | 69.87 | (-29%) | 68.06 | (-31%) |
| Fixed assets | 6.26 | 6.03 | (-4%) | 6.01 | (-4%) | 5.97 | (-5%) |
| Personnel expenses | 3.06 | 3.05 | (0%) | 3.05 | (0%) | 3.05 | (0%) |
| Operating income | 29.14 | 44.35 | (52%) | 45.47 | (56%) | 46.90 | (61%) |
| Probability *p* | 0.00 | 0.75 | (93079%) | 0.85 | (105503%) | 0.95 | (117927%) |
| Beta | 0.00 | 0.38 |  | 0.40 |  | 0.49 |  |

Table 6. Median values of observed data and projections at different confidence levels, with percentage changes from observed values shown in parentheses.

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
|  | | | | |  |  |  |
| Scenario | Observed | *p*=0.75 | | *p*=0.85 | | *p*=0.95 | |
| Total assets | 49.75 | 45.01 | (-10%) | 44.25 | (-11%) | 44.14 | (-11%) |
| Employees | 216.03 | 162.32 | (-25%) | 157.61 | (-27%) | 156.09 | (-28%) |
| Fixed assets | 23.26 | 21.03 | (-10%) | 20.63 | (-11%) | 20.61 | (-11%) |
| Personnel expenses | 7.69 | 7.68 | (0%) | 7.68 | (0%) | 7.68 | (0%) |
| Operating income | 84.17 | 115.49 | (37%) | 121.50 | (44%) | 122.34 | (45%) |
| Probability, *p* | 0.34 | 0.28 | (-17%) | 0.32 | (-6%) | 0.36 | (5%) |
| Beta |  | 1.65 |  | 1.90 |  | 1.91 |  |

Table 7. Standard deviation of observed data and projections at different confidence levels, with percentage changes from observed values shown in parentheses.

Overall, it is observed that as the probability threshold (confidence level) of belonging to the efficient class increases, the required adjustments in inputs and outputs for the evaluated firms to be classified as efficient also increase. Therefore, as expected, a higher probability threshold demands a greater magnitude of change, meaning that the evaluated firms must undergo more significant input and output adjustments to achieve efficiency. This result highlights the increasing effort required to meet stricter efficiency classification criteria.

While the aggregated results in Tables 5, 6, and 7 provide a comprehensive overview, examining specific cases can reveal additional patterns. In Table 8, we present an example of two inefficient firms and their projections for a scenario with a confidence level of *p=*0.85. These firms are selected as representative examples to illustrate the varying adjustments required. Firm #37, not ranked among the 25 top firms in Table 4 with a probability efficiency of approximately 0, fails to meet the threshold *p=*0.85 and its best achievable performance, is recorded. The adjustments required for firm #22 are notably smaller than those for DMU 37 as shown by the percentage changes in brackets, because its efficiency probability is much higher, 0.056. This difference is further highlighted by the value , with DMU 37's  being 5.55 times larger than DMU 22's, reflecting the significantly greater effort needed for DMU 37 to approach efficiency.

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
|  | DMU 22 | | | DMU 37 | | |
|  | Target (#18) | | | Target (#56) | | |
|  | Observed | Predicted |  | Observed | Predicted |  |
| Total assets | 24.71 | 23.20 | (-6%) | 48.90 | 40.65 | (-17%) |
| Employees | 212.00 | 196.97 | (-7%) | 134.00 | 51.84 | (-61%) |
| Fixed assets | 11.46 | 10.96 | (-4%) | 20.10 | 17.34 | (-14%) |
| Personnel expenses | 8.00 | 8.00 | (0%) | 6.39 | 6.37 | (0%) |
| Operating income | 80.89 | 90.21 | (12%) | 52.77 | 103.67 | (96%) |
| Probability, *p* | 0.056 | 0.85 |  | 0.001 | 0.80 |  |
|  | - | 0.29 |  | - | 1.61 |  |

Table 8. Observed values and projections for DMUs 22 and 37, with percentage changes from observed values shown in parentheses.

# 5. Conclusions and future work

A growing body of literature explores the integration of Machine Learning (ML) with Data Envelopment Analysis (DEA) to enhance efficiency analysis across various sectors. While many studies have focused on improving traditional DEA methodologies through ML techniques based on regression, our research extends this synergy by introducing classification models to predict efficiency probabilities. Our findings show that the integration of ML classifiers with DEA not only predicts the efficiency status of Decision-Making Units (DMUs) but also provides a richer framework for assessing efficiency through probabilistic measures and counterfactual analysis. The advantages of our integrated approach extend beyond just analytical improvements. They also offer practical benefits in terms of scalability and adaptability. The model's ability to handle large datasets efficiently makes it especially relevant in the era of big data, where organizations across sectors are looking to exploit vast amounts of information for enhanced decision-making (Zhu, 2022). Additionally, the flexibility of the ML-DEA framework means it can be tailored to specific sector needs. This novel approach is illustrated through an empirical analysis of SABI (Iberian Balance Sheet Analysis System) firm data, emphasizing its practical utility.

As a summary, let us highlight that the new approach introduces several key methodological, interpretative, and practical contributions to efficiency analysis by integrating machine learning techniques within a DEA framework. First, we propose a novel classification-based machine learning approach in the second stage of a DEA-ML hybrid framework, moving beyond traditional regression-based techniques. In the first stage, we employ a standard DEA model to generate a binary efficiency label, which is then predicted in the second stage using classification models. Second, our framework enhances inferential power by estimating the probability of a DMU being classified as efficient, shifting DEA from a purely descriptive tool to a probabilistic efficiency assessment. This aligns efficiency analysis with modern inferential analytics and decision-making frameworks. Third, we reinterpret DEA as a classification problem, where the efficiency frontier is understood as a separating surface between technically feasible and infeasible input-output profiles, allowing efficiency measures to be framed in terms of the minimal modifications required for reclassification. Fourth, our approach is algorithm-agnostic, enabling robust efficiency assessments across various classification models, including decision trees, SVMs, neural networks, and ensemble methods. Fifth, we integrate Explainable AI (XAI) techniques, particularly counterfactual analysis, to define inefficiency in terms of the minimum changes required for an inefficient DMU to become efficient, offering an interpretable and actionable efficiency assessment. Sixth, we introduce a benchmarking approach that exploits the importance ranking of inputs and outputs obtained with machine learning models of sensitivity analysis that allow assigning data-driven weights to directional projections, thereby improving the interpretability and strategic value of efficiency assessments. Seventh, we enhance benchmarking by incorporating probabilistic efficiency thresholds, allowing for target setting through counterfactual benchmarking, which provides improvement strategies based on minimum necessary input-output modifications. Eighth, we propose a new ranking system for DMUs based on their probabilistic efficiency scores, offering an alternative to traditional DEA ranking methods. Finally, our method facilitates a refined proximity-based benchmark identification strategy, ensuring that each DMU is compared against the closest efficient benchmark at any given efficiency probability threshold, strengthening the practical applicability of DEA for dynamic and adaptive benchmarking. These contributions collectively advance efficiency analysis by bridging the gap between DEA, statistical learning, and explainable AI, offering a more flexible and interpretable approach to performance assessment.

Looking forward, several research avenues appear promising. First. the exploration of other machine learning techniques, such as ensemble methods (e.g., Random Forest or Boosting). could provide further improvements in the robustness and accuracy of efficiency predictions. Indeed, when faced with a real empirical case, we could implement multiple machine learning techniques (and not only NN) in parallel to assess the consistency and robustness of the results. By comparing the outcomes across different models, we could evaluate the stability of efficiency classifications and ensure that our findings are not overly dependent on a specific algorithm. Secondly, the application of our integrated ML-DEA model to other domains, such as environmental sustainability, public sector performance, or benchmarking through composite indicators could be highly beneficial. These areas, where efficiency and resource optimization are critical, may significantly benefit from the enhanced analytical capabilities that our model offers. Additionally, extending our model to handle real-time data could transform operational efficiency monitoring, allowing organizations to make immediate adjustments based on current performance metrics. Lastly, further research should also focus on the development of more sophisticated counterfactual methods within the ML-DEA framework. These methods would not only enhance the interpretability of the model outcomes but also allow decision-makers to perform scenario analysis and policy testing effectively.

# Acknowledgments

The authors thank the grant PID2022-136383NB-I00 funded by MCIN/AEI/ 10.13039/501100011033 and by ERDF A way of making Europe. R. Gonzalez-Moyano thanks the grant FPI/PREP2022-000117 funded by MICIU/AEI/10.13039/501100011033 and by ESF+. V. España thanks the PhD scholarship ACIF/2021/135 supported by the Conselleria d'Educació, Universitats i Ocupació (Generalitat Valenciana). Additionally, J. Aparicio thanks the grant PROMETEO/2021/063 funded by the Valencian Community (Spain).

# References

Amirteimoori. A. Allahviranloo. T. Zadmirzaei. M. & Hasanzadeh. F. (2023). On the environmental performance analysis: a combined fuzzy data envelopment analysis and artificial intelligence algorithms. *Expert Systems with Applications*. *224*. 119953.

Aparicio, J., Esteve, M., Rodriguez-Sala, J. J., & Zofio, J. L. (2021). The estimation of productive efficiency through machine learning techniques: efficiency analysis trees. In *Data-enabled analytics: DEA for big data* (pp. 51-92). Cham: Springer International Publishing.

Aydin. N. & Yurdakul. G. (2020). Assessing countries’ performances against COVID-19 via WSIDEA and machine learning algorithms. *Applied Soft Computing*. *97*. 106792.

Banker. R. D. & Morey. R. C. (1986). Efficiency analysis for exogenously fixed inputs and outputs. Operations Research. 34(4). 513-521.

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

Berger. A. N. Brockett. P. L. Cooper. W. W. & Pastor. J. T. (1997). New approaches for analyzing and evaluating the performance of financial institutions. *European Journal of Operational Research*. *98*(2). 170-174.

Boubaker. S. Le. T. D. Ngo. T. & Manita. R. (2023). Predicting the performance of MSMEs: A hybrid DEA-machine learning approach. *Annals of Operations Research*. 1-23.

Chambers, R. G., Chung, Y., & Färe, R. (1998). Profit, directional distance functions, and Nerlovian efficiency. Journal of optimization theory and applications, 98, 351-364.

Charles. V. Aparicio. J. & Zhu. J. (2019). The curse of dimensionality of decision-making units: A simple approach to increase the discriminatory power of data envelopment analysis. European Journal of Operational Research. 279(3). 929-940.

Charnes. A. Cooper. W. W. & Rhodes. E. (1978). Measuring the efficiency of decision making units. European Journal of Operational Research. 2(6). 429-444.

Charnes. A. Cooper. W. W. Golany. B. Seiford. L. & Stutz. J. (1985). Foundations of data envelopment analysis for Pareto-Koopmans efficient empirical production functions. *Journal of econometrics*. *30*(1-2). 91-107.

Chawla. N. V. Bowyer. K. W. Hall. L. O. & Kegelmeyer. W. P. (2002). SMOTE: synthetic minority over-sampling technique. Journal of artificial intelligence research. 16. 321-357.

Chen. Y. Li. Y. Xie. Q. An. Q. & Liang. L. (2014). Data envelopment analysis with missing data: a multiple imputation approach. International Journal of Information and Decision Sciences. 6(4). 315-337.

Cho. B. H. Yu. H. Lee. J. Chee. Y. J. Kim. I. Y. & Kim. S. I. (2008). Nonlinear support vector machine visualization for risk factor analysis using nomograms and localized radial basis function kernels. *IEEE Transactions on Information Technology in Biomedicine*. *12*(2). 247-256.

Cortez. P. (2010). Data mining with neural networks and support vector machines using the R/rminer tool. In *Industrial conference on data mining* (pp. 572-583). Berlin. Heidelberg: Springer Berlin Heidelberg.

Craven. M. W. & Shavlik. J. W. (1992). Visualizing learning and computation in artificial neural networks. *International journal on artificial intelligence tools*. *1*(03). 399-425.

Dadura, A. M., & Lee, T. R. (2011). Measuring the innovation ability of Taiwan's food industry using DEA. *Innovation: The European Journal of Social Science Research*, *24*(1-2), 151-172.

Daouia. A. Noh. H. & Park. B. U. (2016). Data envelope fitting with constrained polynomial splines. *Journal of the Royal Statistical Society Series B: Statistical Methodology*. *78*(1). 3-30.

Dimopoulos, I., Chronopoulos, J., Chronopoulou-Sereli, A., & Lek, S. (1995). Neural network models to study relationships between lead concentration in grasses and permanent urban descriptors. *Ecological Modelling, 120*(2-3), 261-270.

Dimopoulos, I., Chronopoulos, J., Chronopoulou-Sereli, A., & Lek, S. (1995). Neural network models to study relationships between lead concentration in grasses and permanent urban descriptors. Ecological Modelling, 120(2-3), 261-270.

Emrouznejad. A. & Shale. E. (2009). A combined neural network and DEA for measuring efficiency of large scale datasets. *Computers & Industrial Engineering*. *56*(1). 249-254.

Esteve. M. Aparicio. J. Rabasa. A. & Rodriguez-Sala. J. J. (2020). Efficiency analysis trees: A new methodology for estimating production frontiers through decision trees. Expert Systems with Applications. 162. 113783.

Esteve. M. Aparicio. J. Rodriguez-Sala. J. J. & Zhu. J. (2023). Random Forests and the measurement of super-efficiency in the context of Free Disposal Hull. European Journal of Operational Research. 304(2). 729-744.

Fallahpour. A. Olugu. E. U. Musa. S. N. Khezrimotlagh. D. & Wong. K. Y. (2016). An integrated model for green supplier selection under fuzzy environment: application of data envelopment analysis and genetic programming approach. *Neural Computing and Applications*. *27*. 707-725.

Flegl, M., Jiménez-Bandala, C. A., Sánchez-Juárez, I., & Matus, E. (2022). Analysis of production and investment efficiency in the Mexican food industry: Application of two-stage DEA. *Czech Journal of Food Sciences*, *40*(2).

Fogel. D. B. & Robinson. C. J. (2003). Techniques for extracting classification and regression rules from artificial neural networks.

Gevrey, M., Dimopoulos, I., & Lek, S. (2003). Review and comparison of methods to study the contribution of variables in artificial neural network models. *Ecological modelling*, *160*(3), 249-264.

Goodfellow. I. Bengio. Y. & Courville. A. (2016). Deep Learning. MIT Press.

Guerrero. N. M. Aparicio. J. & Valero-Carreras. D. (2022). Combining Data Envelopment Analysis and Machine Learning. Mathematics 2022. 10. 909.

Guillen. M. D. Aparicio. J. & España. V. J. (2023a). boostingDEA: A boosting approach to Data Envelopment Analysis in R. *SoftwareX*. *24*. 101549.

Guillen. M. D. Aparicio. J. & Esteve. M. (2023b). Gradient tree boosting and the estimation of production frontiers. *Expert Systems with Applications*. *214*. 119134.

Guillen. M. D. Aparicio. J. & Esteve. M. (2023c). Performance Evaluation of Decision-Making Units Through Boosting Methods in the Context of Free Disposal Hull: Some Exact and Heuristic Algorithms. *International Journal of Information Technology & Decision Making*. 1-30.

Guillen. M. D. Aparicio. J. Zofío. J. L. & España. V. J. (2024). Improving the predictive accuracy of production frontier models for efficiency measurement using machine learning: The LSB-MAFS method. Computers & Operations Research. 171. 106793.

Hamby, D. M. (1994). A review of techniques for parameter sensitivity analysis of environmental models. Environmental Monitoring and Assessment, 32(2), 135-154.

He. H. & Garcia. E. A. (2009). Learning from imbalanced data. IEEE Transactions on knowledge and data engineering. 21(9). 1263-1284.

Jin. Q. Kerstens. K. & Van de Woestyne. I. (2024). Convex and nonconvex nonparametric frontier-based classification methods for anomaly detection. OR Spectrum. 1-27.

Jomthanachai. S. Wong. W. P. & Lim. C. P. (2021). An application of data envelopment analysis and machine learning approach to risk management. *Ieee Access*. *9*. 85978-85994.

Kuhn. Max (2008). “Building Predictive Models in R Using the caret Package.” *Journal of Statistical Software*. **28**(5). 1–26.

Kumar, M., & Basu, P. (2008). Perspectives of productivity growth in Indian food industry: a data envelopment analysis. *International Journal of Productivity and Performance Management*, *57*(7), 503-522.

Kuosmanen. T. & Johnson. A. L. (2010). Data envelopment analysis as nonparametric least-squares regression. *Operations Research*. *58*(1). 149-160.

Kwon. H. B. Lee. J. & Roh. J. J. (2016). Best performance modeling using complementary DEA-ANN approach: Application to Japanese electronics manufacturing firms. *Benchmarking: An International Journal*. *23*(3). 704-721.

LeCun. Y. Bengio. Y. & Hinton. G. (2015). Deep learning. Nature. 521(7553). 436-444.

Li, G., Rabitz, H., Yelvington, P. E., Oluwole, O. O., Bacon, F., Kolb, C. E., & Schoendorf, J. (2010). Global sensitivity analysis for systems with independent and/or correlated inputs. *The journal of physical chemistry A*, *114*(19), 6022-6032.

Liao. Z. Dai. S. & Kuosmanen. T. (2024). Convex support vector regression. *European Journal of Operational Research*. *313*(3). 858-870.

Lin. S. W. & Lu. W. M. (2024). Using inverse DEA and machine learning algorithms to evaluate and predict suppliers’ performance in the apple supply chain. *International Journal of Production Economics*. 109203.

Liu. H. H. Chen. T. Y. Chiu. Y. H. & Kuo. F. H. (2013). A comparison of three-stage DEA and artificial neural network on the operational efficiency of semi-conductor firms in Taiwan. *Modern Economy*. *4*(01). 20-31.

Lundberg. S. & Lee, S.-I. (2017). A unified approach to interpreting model predictions. *arXiv preprint arXiv:1705.07874*.

Martens. D. Baesens. B. Van Gestel. T. & Vanthienen. J. (2007). Comprehensible credit scoring models using rule extraction from support vector machines. *European journal of operational research*. *183*(3). 1466-1476.

Nandy. A. & Singh. P. K. (2020). Farm efficiency estimation using a hybrid approach of machine-learning and data envelopment analysis: Evidence from rural eastern India. *Journal of Cleaner Production*. *267*. 122106.

Olesen. O. B. & Ruggiero. J. (2022). The hinging hyperplanes: An alternative nonparametric representation of a production function. *European Journal of Operational Research*. *296*(1). 254-266.

Olesen. O. B. Petersen. N. C. & Podinovski. V. V. (2007). Staff assessment and productivity measurement in public administration: an application of data envelopment analysis. Omega. 35(3). 297-307.

Omrani. H. Emrouznejad. A. Teplova. T. & Amini. M. (2024). Efficiency evaluation of electricity distribution companies: Integrating data envelopment analysis and machine learning for a holistic analysis. Engineering Applications of Artificial Intelligence. 133. 108636.

Orea, L. & Zofío, J.L. (2019). Common methodological choices in non-parametric and parametric analyses of firms’ performance, in ten Raa, T. & Greene, W. H. (eds.) *Palgrave Handbook of Economic Performance Analysis*, Palgrave-MacMillan: Cham, 419-484.

Parmeter. C. F. & Racine. J. S. (2013). Smooth constrained frontier analysis. *Recent Advances and Future Directions in Causality. Prediction. and Specification Analysis: Essays in Honor of Halbert L. White Jr*. 463-488.

Pastor. J. T. Lovell. C. K. & Aparicio. J. (2012). Families of linear efficiency programs based on Debreu’s loss function. *Journal of Productivity Analysis*. *38*. 109-120.

Ribeiro. M. T. Singh. S. & Guestrin. C. (2016. August). " Why should i trust you?" Explaining the predictions of any classifier. In *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining* (pp. 1135-1144).

Ruck. D. W. Rogers. S. K. & Kabrisky. M. (1990). Feature selection using a multilayer perceptron. Journal of neural network computing. 2(2). 40-48.

Saltelli, A., Ratto, M., Andres, T., Campolongo, F., Cariboni, J., Gatelli, D., Saisana, M., & Tarantola, S. (2008). Global Sensitivity Analysis: The Primer. Wiley.

Saltelli, A., Tarantola, S., Campolongo, F., & Ratto, M. (2004). *Sensitivity analysis in practice: a guide to assessing scientific models* (Vol. 1). New York: Wiley.

Sexton, T. R. (1986). The methodology of data envelopment analysis. New directions for program evaluation, 32, 7-29.

Soboĺ, I. M. (1993). Sensitivity estimates for nonlinear mathematical models. *Math. Model. Comput. Exp.*, *1*, 407.

Song, X., Zhang, J., Zhan, C., Xuan, Y., Ye, M., & Xu, C. (2015). Global sensitivity analysis in hydrological modeling: Review of concepts, methods, theoretical framework, and applications. *Journal of hydrology*, *523*, 739-757.

Tayal. A. Solanki. A. & Singh. S. P. (2020). Integrated framework for identifying sustainable manufacturing layouts based on big data. machine learning. meta-heuristic and data envelopment analysis. *Sustainable Cities and Society*. *62*. 102383.

Thanassoulis, E., Portela, M. C., & Despic, O. (2008). Data envelopment analysis: the mathematical programming approach to efficiency analysis. The measurement of productive efficiency and productivity growth, 251-420.

Thanassoulis. E. Boussofiane. A. & Dyson. R. G. (2015). Applied data envelopment analysis. Springer.

Tickle. A. B. Andrews. R. Golea. M. & Diederich. J. (1998). The truth will come to light: Directions and challenges in extracting the knowledge embedded within trained artificial neural networks. IEEE Transactions on Neural Networks. 9(6). 1057-1068.

Tsionas. M. Parmeter. C. F. & Zelenyuk. V. (2023). Bayesian artificial neural networks for frontier efficiency analysis. *Journal of Econometrics*. *236*(2). 105491.

Tzeng. F. Y. & Ma. K. L. (2005). *Opening the black box-data driven visualization of neural networks* (pp. 383-390). IEEE.

Valero-Carreras. D. Aparicio. J. & Guerrero. N. M. (2021). Support vector frontiers: A new approach for estimating production functions through support vector machines. *Omega*. *104*. 102490.

Valero-Carreras. D. Aparicio. J. & Guerrero. N. M. (2022). Multi-output support vector frontiers. *Computers & Operations Research*. *143*. 105765.

Valero-Carreras. D. Moragues. R. Aparicio. J. & Guerrero. N. M. (2024). Evaluating different methods for ranking inputs in the context of the performance assessment of decision making units: A machine learning approach. Computers & Operations Research. 163. 106485.

Venables W.N. & Ripley B.D. (2002). *Modern Applied Statistics with S*. Fourth edition. Springer. New York. ISBN 0-387-95457-0.

Wachter. S. Mittelstadt. B. & Russell. C. (2017). ‘Counterfactual Explanations Without Opening the Black Box: Automated Decisions and the GDPR.’ Harvard Journal of Law & Technology. 31(2). 841-887.

Wang, K., Xian, Y., Lee, C. Y., Wei, Y. M., & Huang, Z. (2019). On selecting directions for directional distance functions in a non-parametric framework: a review. *Annals of Operations Research*, *278*, 43-76.

Weiss. G. M. & Provost. F. (2003). Learning when training data are costly: The effect of class distribution on tree induction. Journal of artificial intelligence research. 19. 315-354.

Zhou. P. Ang. B. W. & Poh. K. L. (2008). A survey of data envelopment analysis in energy and environmental studies. European Journal of Operational Research. 189(1). 1-18.

Zhu. J. (2022). DEA under big data: Data enabled analytics and network data envelopment analysis. Annals of Operations Research. 309(2). 7

Zhu. N. Zhu. C. & 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.

1. A third line of research in the literature, unrelated to this study, employs Data Envelopment Analysis (DEA) as an alternative method to conventional Machine Learning (ML) classification techniques such as Support Vector Machines (SVM), decision trees, and neural networks. In that line, DEA is utilized to classify observations based on their features instead of measuring technical efficiency. For example, it is applied to identify individuals as carriers of a rare genetic disorder from age and several blood measurements. A recent example of this type of contributions is Jin et al. (2024). [↑](#footnote-ref-2)
2. The SABI (Iberian Balance Sheet Analysis System) database is a subset of the ORBIS product commercialized by Moody’, which offers comparable data on private companies: <https://sabi.bvdinfo.com>. [↑](#footnote-ref-3)