eXplainable Artificial Intelligence (XAI) for Probabilistic Efficiency Analysis

Ricardo González-Moyano[1], Juan Aparicio[1,2*], Víctor J. España[1] and José L. Zofío[3,4]

[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

In recent decades, efficiency analysis has made significant progress, particularly in evaluating decision-making units (DMUs) in sectors such as finance, healthcare, education, and manufacturing. Data Envelopment Analysis (DEA) is a non-parametric method used to assess the relative efficiency of DMUs by comparing their input-output relationships. However, traditional DEA approaches face challenges in capturing complex patterns and structures in data, such as overfitting and dealing with nonlinear relationships between inputs and outputs. With the rise of machine learning (ML) techniques, there is an opportunity to enhance DEA's capabilities by leveraging ML's computational power and flexibility. This integration can improve the accuracy, robustness, and interpretability of efficiency assessments, advancing performance analysis. Our paper contributes to this progress by introducing a hybrid methodological framework that uses DEA to label the data and ML classification techniques, specifically Support Vector Machines and Neural Networks, to generate a performance score. We demonstrate the practical implications of this integration through an empirical example using PISA (Programme for International Student Assessment) data. This new synergy between DEA and ML holds promise to further transform efficiency evaluation and enhancing our understanding of complex systems in production.

Keywords: Data Envelopment Analysis, Machine Learning, Classification models, robustness, variable importance.

1. Introduction

In recent decades, the field of efficiency analysis has witnessed significant advancements, particularly in the evaluation of decision-making units (DMUs) 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 relationships. 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 use to be 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 inherent in complex 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). Overfitting in 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 nuanced patterns in the data, leading to biased efficiency estimates. Another significant limitation of traditional DEA is its deterministic nature. Traditional DEA models produce a single efficiency score for each DMU based on the observed input-output data, without accounting for uncertainties or variability inherent 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 leveraging the computational power and flexibility offered by these methods. By integrating machine learning algorithms with DEA, researchers can potentially improve the accuracy, robustness, and interpretability of efficiency assessments, thereby advancing the state-of-the-art in performance analysis. In this context, it becomes a scientific duty to create the necessary bridges between machine learning and other fields, such as Data Envelopment Analysis. Machine learning algorithms can complement DEA by providing advanced techniques for, for example, 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, we have identified certain gaps that we believe our 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, in the literature, there are two predominant streams of research that explore the integration of machine learning with Data Envelopment Analysis[1]. 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 leverage 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. Some of these contributions are the following: Kuosmanen and Johnson (2010) demonstrated the connection between DEA and least-squares regression, introducing Stochastic Non-smooth Envelopment of Data (StoNED). 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

Comentado [JLZP1]: Si, como se dice posteriormente, el paper se enmarca en segundo stream of litearature, parece que se le está restando importancia al considerarlo en Segundo lugar en el lietarature review. Si se ha hecho así la ordenación por cronología yo buscaría algune cirerio que le permitiese dar la Vuelta (p.e., hay más contribunciones en el segundo stream que el primero). Así se le dá más relevancia.

Comentado [RG2]: ¿Qué titulo tiene?

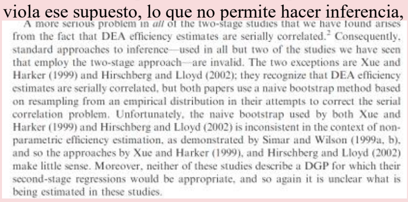
Comentado [RG3R2]: Este?: The estimation of productive efficiency through machine learning techniques: efficiency analysis trees

[1] A third line of research in the literature 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).

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

Comentado [JLZP4]: ¿Esta literature de two-stage no adolece de los problemas econométricos puestos de manifiesto por Simar y Wilson (2007)? https://www.sciencedirect.com/science/article/abs/pii/S0304407605001594
Puede ser relavante para el paper reforzando las conclusiones (2.*Enhanced Interpretability*) por lo que se puede meter una frase enfatizndo esto de acuerdo a lo que pusimom en el paper de LBS-MAFS.

Comentado [JLZP5]: El problema econométrico es que al calcular los índices de eficiencia en la primera etapa, en la segunda no se cumple que laa variables dependiente (los índices) sean independientes (uinos dependen de otros) y se viola ese supuesto, lo que no permite hacer inferencia, etc.



A more serious problem in *all* of the two-stage studies that we have found arises from the fact that DEA efficiency estimates are serially correlated.[2] Consequently, standard approaches to inference—used in all but two of the studies we have seen that employ the two-stage approach—are invalid. The two exceptions are Xue and Harker (1999) and Hirschberg and Lloyd (2002); they recognize that DEA efficiency estimates are serially correlated, but both papers use a naive bootstrap method based on resampling from an empirical distribution in their attempts to correct the serial correlation problem. Unfortunately, the naive bootstrap used by both Xue and Harker (1999) and Hirschberg and Lloyd (2002) is inconsistent in the context of non-parametric efficiency estimation, as demonstrated by Simar and Wilson (1999a, b), and so the approaches by Xue and Harker (1999), and Hirschberg and Lloyd (2002) make little sense. Moreover, neither of these studies describe a DGP for which their second-stage regressions would be appropriate, and so again it is unclear what is being estimated in these studies.

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 CO_2 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) valuated 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 integratingML with DEA. However, despite these developments, there remain certain gaps and limitations that we aim to address in this paper. Specifically, the methodological innovations introduced in this article contribute to both streams of literature. On the one hand the use of ML classifying techniques, like SVM or NN, to label observations as efficient or inefficient represents an alternative method to estimate the production frontier. On the other hand, these techniques offer a second-stage explanation of the efficiency scores that by-pass some of the difficulties of the econometric literature that regresses the DEA scores obtained in the first stage on a set of explanatory variables (e.g., Simar and Wilson, 2007). Despite advances in this field combining bootstrapping and truncated regression techniques, these strategy poses significant challenges in uncertain, indeterminate, and noisy contexts, where distinguishing between 0.9 and 1.0 regarding efficiency score is difficult. Moreover, techniques in this second group use the same DEA efficiency score determined for each DMU in the first stage as the final evaluation for efficiency of the observations. Therefore, the efficiency evaluation of the data sample is not 'improved' by incorporating ML techniques in the second stage and, consequently, the corresponding ranking of DMUs remains the same as the original one. These are the two gaps we identify and aim to address in this paper. In this sense, and for the first time in the literature, we will use a classification model rather than a regression model in the second stage of the approach that combines DEA and ML. In fact, we will employ a standard DEA model in the first stage to identify, through Pareto-dominance efficiency evaluation, a labelling that distinguishes between efficient and inefficient units. And, in the second stage, we will predict this label using all variables of the problem. Additionally, our approach will allow us to modify the measurement of the degree of efficiency

Comentado [JLZP6]: De aquí mi comentario respecto a los problemas econométricos puestos de manifiesto por Simar y Wilson (2007).
Pero si la téncia no está sujeta a estos probelmas pq no hay probelmas econométricos ENTONCES PODEMOS VENDERLA COMO UNA ALTERNATIVA A SIMAR Y WILSON (2007) DESDE EL CAMPO DE ML.

Comentado [JLZP7]: Esto es una alternativa al botstrapping propuesto por Simar y Wilson (1999) para proveer un 'bias adjusted' efficiency score e inferencia estadística respecto a su valor con intervalos de confianza (entre 0.8 y 1 p.e.). De nuevo una alternativa válida al DEA tradicional y los esfuerzos realizados para acomodar el DGP, ruido, etc. I.e. the "uncertainty" of the efficiency scores.

of observations, as the efficiency score will be calculated using an eXplainable Artificial Intelligence (XAI) method based on the use of a counterfactual: technical inefficiency will be defined for an inefficient DMU as the minimum changes required in the observed inputs and outputs (or in a certain direction depending on the model orientation and other factors) to change from the inefficient label to the efficient label. Moreover, in the process we demonstrate that DEA can be viewed as a particular case of a classification model in the sense that the DEA frontier could be interpreted as the separating surface in the input-output space of two classes (labels): technically producible (feasible) units vs. technically non-producible (infeasible) units; with the peculiarity of having all efficient DMUs located on the separating surface (the efficient frontier). This reinterpretation means that traditional efficiency measures for feasible DMUs, which quantify how much the inputs and/or outputs of the evaluated unit would need to change to shift from being labeled as a feasible unit to being labeled as an infeasible unit (projecting the DMU towards the efficient frontier of DEA technology). Therefore, the conceptual foundation motivating the formulation of our counterfactual method aligns with the principles underpinning the conventional approach for quantifying inefficiency in DEA. This entails projecting inefficient units onto the DEA technology frontier until reaching a state where they no longer deviate from the production possibility set (achieving the efficiency status).

The proposed methodology allows us to contribute also to the research focused on the determination of variable (inputs and outputs) importance within DEA models, which has been pivotal in the literature. As highlighted by Banker and Morey (1986), comprehending the contributing factors to relative efficiency empowers organizations to channel efforts towards areas where substantial improvements can be achieved. As suggested by Thanassoulis et al. (2015), identifying the most relevant variables not only facilitates strategic decision-making but also provides valuable insights for optimal resource allocation and the implementation of continuous improvement measures. Hence, the assessment of variable importance in the production process is fundamental for maximizing efficiency and productivity across industries. Our objective is to enhance the new methodological framework for determining variable importance in DEA models. While existing studies have provided valuable insights into the significance of variables (e.g., Pastor et al., 2002), there is still room for refinement and advancement. Specifically, by incorporating advanced machine learning algorithms, we seek to provide more robust and accurate assessments of variable importance, thereby enabling organizations to make informed decisions and drive continuous improvement initiatives effectively.

Altogether, this study introduces a new method that, based on classification models, allows identifying the efficiency status of DMUs and their relative scores. The method exploits existing synergies between DEA and machine learning techniques, elucidating the potential benefits of their integration in the context of efficiency evaluation. Specifically, we discuss various approaches for combining DEA with machine learning within the category of classification models, introducing a new hybrid framework that integrates both techniques. The paper is structured as follows: In Section 2, we provide background information on Data Envelopment Analysis (DEA) and the two machine learning techniques we will utilize, namely Support Vector Machines (SVM) and (Artificial) Neural Networks (NN). Section 3 introduces our novel approach, which integrates DEA with these two classification techniques, aiming to enhance efficiency assessment for DMUs. We demonstrate the practical implications of this integration and its implications for decision-making and policy formulations through an empirical example based on PISA (Programme for International Student Assessment) in Section 4. Section 5 concludes and points out further research lines.

2. Background

This background section provides a concise overview of DEA and the main ML technique that we will apply in this paper (Neural Networks).

2.1. Data Envelopment Analysis

Data Envelopment Analysis (DEA) is a non-parametric method widely used for evaluating the relative efficiency of decision-making units (DMUs) in various fields, including economics, finance, management science and operations research. Introduced by Charnes et al. (1978), DEA offers a powerful framework for assessing the efficiency of DMUs transforming multiple inputs into multiple outputs. DEA operates under the assumption of constant returns to scale (CRS) or variable returns to scale (VRS). VRS is particularly suitable for analyzing real-world production processes, where economies of scale may vary across different units.

In this study we evaluate the performance of *n* observations by measuring their technical efficiency. These observations or DMUs, which could be firms or organizations, utilize *m* various inputs $\mathbf{x}_j = \left(x_{1j}, \ldots, x_{mj}\right) \in R_+^m$, such as resources, to generate *s* various outputs $\mathbf{y}_j = \left(y_{1j}, \ldots, y_{sj}\right) \in R_+^s$, 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’ (also called production possibility set) encompasses all feasible input-output combinations. This concept is typically represented as:

$$T = \left\{ (\mathbf{x}, \mathbf{y}) \in R_{+}^{m+s} : \mathbf{x} \text{ can produce } \mathbf{y} \right\}. \tag{1}$$

Among the non-parametric methodologies utilized to empirically approximate the set T , DEA stands out as one of the most commonly employed approaches in practical applications. Under VRS, Banker et al. (1984) show that that the DEA technology T corresponds to:

$$T_{DEA} = \left\{ (\mathbf{x}, \mathbf{y}) \in R_{+}^{m+s} : y_r \leq \sum_{j=1}^{n} \lambda_j y_{rj}, \forall r, x_i \geq \sum_{j=1}^{n} \lambda_j x_{ij}, \forall i, \sum_{j=1}^{n} \lambda_j = 1, \lambda_j \geq 0, \forall j \right\}. \tag{2}$$

In literature numerous technical efficiency measures are available to calculate the technical efficiency of observations within T_{DEA}—for a general definition of these measures see Pastor et al. (2012). In particular, our focus is directed towards a prevalent measure, namely, the output-oriented radial model. Considering the specific DMU (x_o, y_o), its technical efficiency can be calculated through the following program

$$\begin{array}{llll}
\phi_{DEA}(\mathbf{x}_o, \mathbf{y}_o) = & \max & \phi_o & (3.0) \\
& s.t. & \sum_{j=1}^{n} \lambda_{jo} x_{ij} \leq x_{io}, \quad i = 1, \ldots, m & (3.1) \\
& & \sum_{j=1}^{n} \lambda_{jo} y_{rj} \geq \phi_o y_{ro}, \quad r = 1, \ldots, s & (3.2) \\
& & \sum_{j=1}^{n} \lambda_{jo} = 1, & (3.3) \\
& & \lambda_{jo} \geq 0, \quad j = 1, \ldots, n & (3.4)
\end{array} \tag{3}$$

Under this model, a DMU with a score of one, $\phi_o = 1$, is considered fully efficient, indicating that it operates on the efficient frontier. Conversely, a radial measure greater than one, $\phi_o > 1$, implies inefficiency relative to reference technology, with a bigger value indicating a worse degree of efficiency. The radial measure and its associated reference benchmarks on the frontier provide valuable insights into the performance of individual DMUs and can guide decision-makers in identifying opportunities for improvement.

2.2. Machine Learning Technique for Classification

Comentado [JLZP8]: En esta sección, cuando se decriben SVM y NN se podría poner alguna línea (intuicion) de como se van a utilizar estas tecnicas en el análisis de eficiencia. Como p.e. se hace con XAI con la frase: “contexts (for example, in our production context the question could 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?’)”. Así se “calma” la impaciencia del lector con relición a cómo se van a utilizar junto al DEA.

In this subsection, we briefly outline the fundamentals of the machine learning technique that will be employed throughout the article Neural Networks (NN), as well as eXplainable Artificial Intelligence (XAI). NN are a class of learning algorithms inspired by the structure and function of the human brain. They consist of interconnected layers of neurons that process input data through nonlinear transformations to learn complex patterns and relationships. By understanding the underlying principles NN, which determine the label and the probability of belonging to that label, we can harness their capabilities to enhance the DEA methodology.

Neural Networks represent a cornerstone in the field of machine learning, heralded for their ability to learn complex patterns and relationships from data (LeCun et al., 2015; Goodfellow et al., 2016). In this subsection, we briefly delve into the application of Neural Networks in the context of classification tasks, highlighting their versatility, theoretical foundations, and practical implications.

Neural Networks are inspired by the structure and function of the human brain, comprising interconnected layers of artificial neurons or nodes. The core principle underlying Neural Networks is the process of forward propagation, where input data is sequentially passed through multiple layers of neurons, each layer applying a set of weights and activation functions to produce an output. Through an iterative process known as backpropagation, Neural Networks 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 Neural Networks 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.

The performance of Neural Networks 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. Techniques such as grid search, random search, and Bayesian optimization are commonly employed to systematically explore the hyperparameter space and identify optimal configurations.

Despite their remarkable predictive capabilities, one challenge of Neural Networks lies in their black-box nature, which hinders interpretability and understanding of model decisions. However, techniques such as layer-wise relevance propagation (LRP) and gradient-based attribution methods can provide insights into feature importance and highlight the contribution of input features to model predictions. This feature importance analysis aids in model interpretation and decision-making processes.

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 the output layer, a single neuron would be located to assign the corresponding class to each observation. Between these layers lies the hidden layer, composed of three neurons in this specific case. Figure 2 depicts the structure of this neural network with a configuration of 2-3-1.

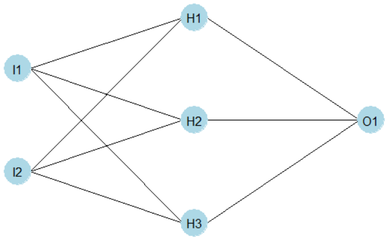


Figure 1~~2~~. An example of an artificial Neural Network

2.3. 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). In this section, 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.

Overall, XAI encompasses a diverse set of methodologies and techniques designed to elucidate the decision-making process of machine learning models. As AI (Artificial Intelligence) systems become increasingly complex and ubiquitous, there is a growing need for transparency and interpretability to foster trust and facilitate human understanding of model behavior. XAI aims to address this need by providing explanations that are understandable, intuitive, and actionable for end-users, stakeholders, and domain experts.

To address this, XAI provides methodologies designed to clarify the decision-making process of ML models. These approaches aim to generate explanations that are understandable, actionable, and intuitive for end-users and stakeholders, enabling better model validation and facilitating the identification of relationships within the data.

2.3.1 Counterfactual Explanations

In particular, 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 could 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?’). 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 will be incorporated into our approach in this paper to measure technical efficiency in the context of machine learning and efficiency analysis (see Section 3).

2.3.2 Feature Significance Analysis and Sensitivity Analysis

To complement counterfactual analysis, we incorporate feature significance analysis, focusing on understanding the contribution of input and output variables to the model’s predictions. There are several approaches to feature significance analysis, such as rule extraction methods (see, for example, Tickle et al., 1998; Fogel & Robinson, 2003; Martens et al., 2007), visualization techniques (see, for example, Craven & Shavlik, 1992; Tzeng & Ma, 2005; Cho et al., 2008), Sensitivity Analysis (SA) (Ruck et al., 1990), and more recent methods such as SHAP (Lundberg & Lee, 2017) and LIME (Ribeiro et al., 2016). We decided to use 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 treats the original fitted model, querying it with sensitivity samples and recording the corresponding outputs. Second, visualization techniques are often designed for a specific machine learning method limiting their general applicability. This is a disadvantage compared to SA, which can be applied to any supervised machine learning method. Third, methods like SHAP and LIME are computationally intensive and more challenging to interpret, especially in high-dimensional datasets. In contrast, SA is computationally efficient, simple to implement, and provides clear, actionable insights, making it a practical choice for decision-makers

SA works by perturbing each variable across its range while keeping other variables constant at their baseline values. Sensitivity can be measured in several ways, such as by range, gradient, variance, or the average absolute deviation from the median (AAD), the latter being less sensitive to outliers (see Cortez & Embrechts, 2013, for more details). To obtain probabilities, a One-hot encoding transformation is applied to the class labels, allowing the calculation of the probability for each DMU to belong to a specific class. This process allows us to quantify the absolute relevance of features, providing a ranking of variables based on their impact on the model's predictions. In our context, SA enables the identification of the most influential factors driving inefficiency, offering valuable insights into the interplay between inputs, outputs, and efficiency outcomes. Additionally, SA will be integrated into our approach to measuring technical efficiency in combination with counterfactual analysis (see Section 3).

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. In this case, while other ML classification methods could be considered, we focus here on Neural Networks.

3.1 Classifying DMUs by their (in)efficiency class and measuring technical efficiency

Before introducing our methodology, we aim to elucidate the reinterpretation of DEA, through a graphical toy example (Figure 3), as a classification method that also resorts to counterfactual analysis. DEA can be conceptualized as a classification model wherein the two classes represent feasible and infeasible units of production, with the boundary delineating the separating surface and efficient units positioned precisely onto this surface. Furthermore, within the feasible but inefficient set of DMUs this reinterpretation implies that the 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 radial output-oriented model (3)), signifies a transition within its original class label "feasible, but inefficient" to a new status "feasible and efficient", through its projection onto the efficient frontier (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.

Comentado [JLZP9]: He reescrito mucho este párrafo con la idea de mejorar la analogía o reinterpretación del DEA como un caso particular de XAI. Espero haberlo conseguido.

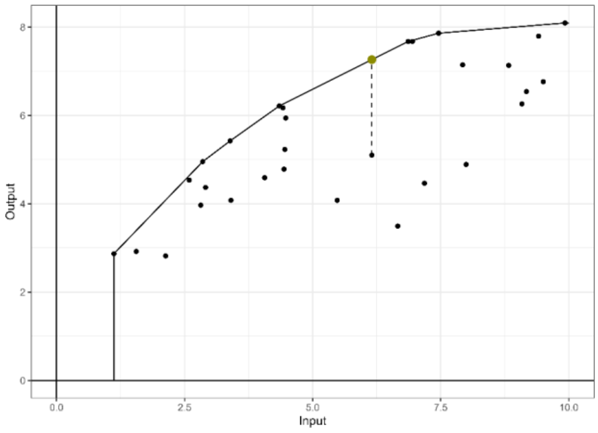


Figure 2. The output-oriented radial measure in DEA

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, particularly from a counterfactual approach, we now proceed to introduce our method. The core concept underlying our model is a multi-stage methodology aimed at enhancing efficiency assessment through the fusion of DEA and ML techniques. Our 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 employ a classification ML model, wherein the response variable is the efficiency class (efficient vs. inefficient), and the classification features include both inputs and outputs. Finally, in the third phase of our approach, we ascertain a robust measure of technical inefficiency through the application of XAI principles. Specifically, given a model measuring technical efficiency (such as the output-oriented radial model), we determine the minimum increase required in the output of each inefficient DMU to transition its class from inefficient to efficient.[2] This structured approach not only facilitates the identification of inefficiencies but also provides actionable insights for decision-makers to enhance performance.

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

Step 1: Utilize the additive DEA model (Charnes et al., 1985), model (4), to partition the set of DMUs into two categories (efficient vs inefficient) based on the optimal value of the optimization program. A value of zero indicates that the evaluated unit is not Pareto-dominated by any technically feasible input-output combination within the standard DEA production possibility set. This condition underscores the efficiency of the evaluated unit, demonstrating that there is no room in the observed sample for enhancing any input and/or output without compromising the feasibility of the unit under assessment.

$$
\begin{aligned}
A_{DEA}\left(\boldsymbol{x}_o,\boldsymbol{y}_o\right)= \quad \max \quad & \sum_{i=1}^{m} s_{io}^{-} + \sum_{r=1}^{s} s_{ro}^{+} & & \quad (4.0)\\
s.t. \quad & \sum_{j=1}^{n}\lambda_{jo}x_{ij} = x_{io}-s_{io}^{-}, & i=1,...,m & \quad (4.1)\\
& \sum_{j=1}^{n}\lambda_{jo}y_{rj} = y_{ro}+s_{ro}^{+}, & r=1,...,s & \quad (4.2)\\
& \sum_{j=1}^{n}\lambda_{jo} = 1, & & \quad (4.3)\\
& \lambda_{jo} \geq 0, & j=1,...,n & \quad (4.4)\\
& s_{io}^{-}, s_{ro}^{+} \geq 0, & \forall i, \forall r & \quad (4.5)
\end{aligned} \tag{4}
$$

[2] We consider the radially oriented output measure (3) for simplicity, but other ‘graph’ measures accounting for both inputs and outputs like the directional distance function or hyperbolic function could be considered.

If, $A_{DEA}\left(\boldsymbol{x}_o,\boldsymbol{y}_o\right)>0$, then DMU $\left(\boldsymbol{x}_o,\boldsymbol{y}_o\right)$ is (technically) inefficient. The set of all inefficient DMUs is denoted as I. Otherwise, that is, if $A_{DEA}\left(\boldsymbol{x}_o,\boldsymbol{y}_o\right)=0$, then DMU $\left(\boldsymbol{x}_o,\boldsymbol{y}_o\right)$ is (technically) efficient. The set of all efficient DMUs is denotes as E.

Step 2: Addressing the challenge of class imbalance (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, 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). This adaptation allows us to tailor the synthetic data generation process to better fit the characteristics of our dataset and context. By doing so, we balance the class distribution, mitigate the bias introduced by the original imbalance, and enhance generalization 'out-of-the-sample' by expanding the decision boundary for the minority class. 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 units). There is no exact proportion that guarantees an ideal balance in the dataset. Weiss and Provost (2003) suggest testing performance with different percentages of minority class examples to identify the optimal class distribution for the training set. They conclude that the proportion of the minority class should ideally fall between 20% and 50%. Like them, we test the performance of 20%, 25%, 30%, 35% and 40% and generate the synthetic units for each scenario.

Step 2a: On the one hand, we can achieve balance by increasing the number of efficient DMUs, which is the most common approach in our production context. Specifically, we generate convex combinations of $m+s$ between the DMUs labeled as efficient in Step 1. The total number of combinations is calculated as $\binom{n}{k}$, where n is the number of efficient DMUs and k is the sum of the number of inputs and outputs, $m+s$. For each convex combination, a synthetic unit is generated by applying the same weights to the DMUs involved in that combination. The weight is defined as $\upsilon = \frac{1}{m+s}$. Once all convex combinations have been created, we use the additive DEA model (4) to identify which of these combinations are Pareto-efficient. If the number of synthetic units remains insufficient, additional random DMUs are generated based on efficient convex combinations. In this process, the weights are randomly selected within the range [0.05, 0.95] to ensure that no weight is equal to zero. 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 balance is achieved, the generation of synthetic units stops.

Step 2b: On the other hand, sometimes it is not uncommon to deal with a dataset with high-dimensional spaces. In this case, the additive model may classify many DMUs as efficient due to the curse of dimensionality (Bellman, 1966). However, machine learning techniques are well-suited to handle such challenges, as they can model complex relationships and identify patterns even in high-dimensional data. By leveraging these techniques, it is possible to better discriminate between efficient and inefficient DMUs, mitigating the issues caused by the sparsity of data in high-dimensional spaces. In such situations, the minority class may correspond to the "inefficient" DMUs. To improve the model's performance, it is necessary to achieve a certain balance too. Additional synthetic inefficient DMUs are generated following a similar process described in the previous paragraph. First, synthetic convex combinations with equal weights are created. Second, the additive DEA model (4) is used to identify which convex combinations are inefficient. Third, a random sample of inefficient convex combinations, 20 times larger than the desired number of inefficient units, is selected. Fourth, the subset is divided into quantiles based on their slack values, and an equal number of units is randomly chosen from each quantile until the desired balance is achieved.

Step 3: In this phase, a classification ML model is implemented where the dependent variable denotes the efficiency status (efficient [class +1] vs. inefficient [class -1]), while the independent variables (features) comprise the input and output vectors. In this step, the parameters of the ML model will also be fine-tuned through cross-validation, ensuring the determination of an optimal parameter configuration, an ideal balance rate and a final classification model $\Gamma(\boldsymbol{x}, \boldsymbol{y})$. The best balance rate is selected by comparing the model's performance. If the original dataset is large, we propose creating training, testing, and validation partitions to evaluate how the model interacts with data not used during the fitting phase. If dividing our data into three partitions is not the best option, we propose testing the model's performance on the original dataset and selecting the balance level that provides the best results. In case multiple models exhibit equal performance, the balance level will be determined based on the model's performance with its respective balanced dataset. If the equality persists, the smallest balance level will be selected, following the principle of parsimony. Finally, the best $\Gamma(\boldsymbol{x}, \boldsymbol{y})$ predicts the probability of belonging for each class and classifies the input-output bundle $(\boldsymbol{x}, \boldsymbol{y})$ as (technically) efficient (+1) or inefficient (-1) for certain level of confidence, by default 0.5.

Step 4: Select a standard technical efficiency measure (for example, a directional vector approach to simultaneously capture improvements across multiple dimensions). The objective of this measure is to reduce inputs while increasing outputs. Although we focus on the directional vector approach, the commonly used radial input- or output-oriented models are also viable alternatives, representing extreme cases of this approach determined by SA. Our directional vector is defined as $(-g_x, g_y)$ where g_x represents the relative importance of each input multiplied by its respective observed mean. Similarly, g_y represents the relative importance of each output, taking into account the output variables.

To find the optimal projection, it is necessary to establish a confidence level to define what will be considered efficient. By default, this confidence level is set at 0.5; however, we propose using a higher confidence threshold. Next, the unknown parameter β is selected using contrafactual analysis. They are tested by generating intervals within a given range and observing the corresponding efficiency results. The resulting projections are calculated as (5) where $\boldsymbol{x}_i$ and $\boldsymbol{y}_i$ are the observed inputs and outputs for DMU_i, respectively:

$$(\boldsymbol{x}_i - \beta \boldsymbol{g}_x, \boldsymbol{y}_i + \beta \boldsymbol{g}_y) \tag{5}$$

If the target probability lies between two consecutive β values, we refine the search by iteratively testing new β values within that range and recalculating their associated probabilities. When the algorithm converges to the probability that meets the desired confidence threshold, the corresponding β value is recorded as the minimum adjustment required for the DMU to achieve efficiency.

If the efficiency probability of a DMU observed exceeds the determined threshold, the β value will be set to 0, and the projection will coincide with the observed DMU. Moreover, there are cases where the projection does not achieve the threshold because we restrict projections. The algorithm does not consider projections with resource values lower than the observed ones. As a result, some DMUs may never reach the established threshold. In such cases, we assign the most efficient feasible projection and record its β value.

3.3. An illustrative example.

Next, we will illustrate our method through a numerical example, complemented by several figures. For the classification ML model, we employ Neural Networks (NN).

In this example we create a data set made of 40 DMUs (D) that uses a single input to produce a single output. Following the algorithm, step 1 labels the available data according to the additive model through standard DEA. In this example, 3 DMUs are efficient with all their optimal slacks in model (4) equal to 0, labeling them as such. The remaining 37 are marked as 'inefficient'. The efficient DMUs are: 6, 7 and 31 (see Figure 4). In this case, there is an imbalance in the labels, with 7.5% of the units being efficient and 92.5% being inefficient.

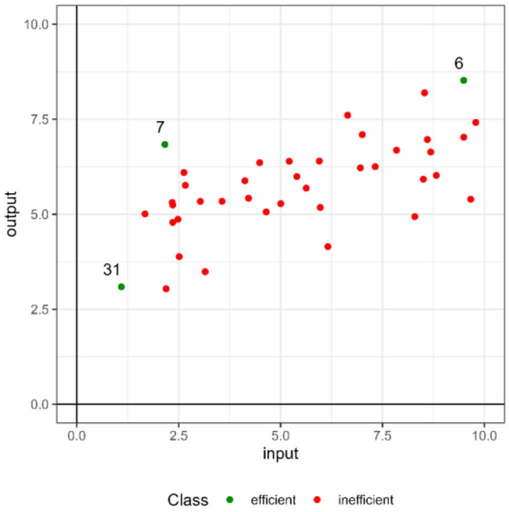
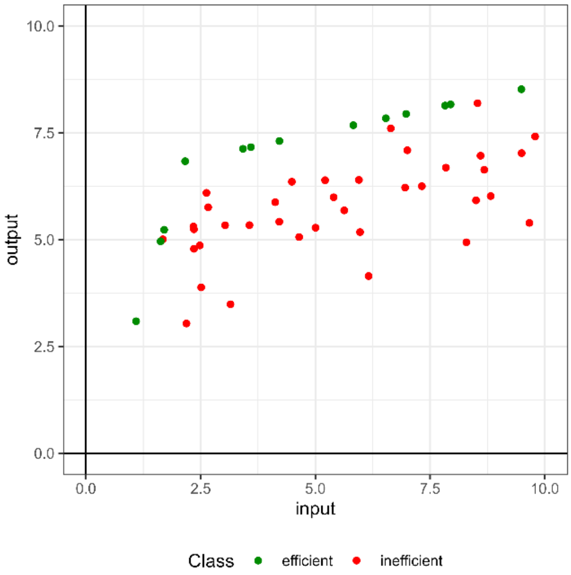


Figure 4. Labeling through the standard DEA additive model

Step 2 of the method balances the dataset by generating synthetic units when necessary. In this example, the minority class consists of efficient DMUs, which are augmented to achieve balance. The procedure for creating new efficient synthetic units (set $\hat{E}$) is creating convex combinations described in step 2a.

Figure 5 illustrates the augmented dataset with a 25% minority class level. Initially, there were only 3 observations labeled as 'efficient', which increased to 13 after the creation of synthetic efficient units. Once the data imbalance has been addressed, the dataset consists of $\widehat{D} = D \cup \hat{E}$ with 50 units, with an approximately 1:4 ratio between units labeled as 'efficient' and 'inefficient'.



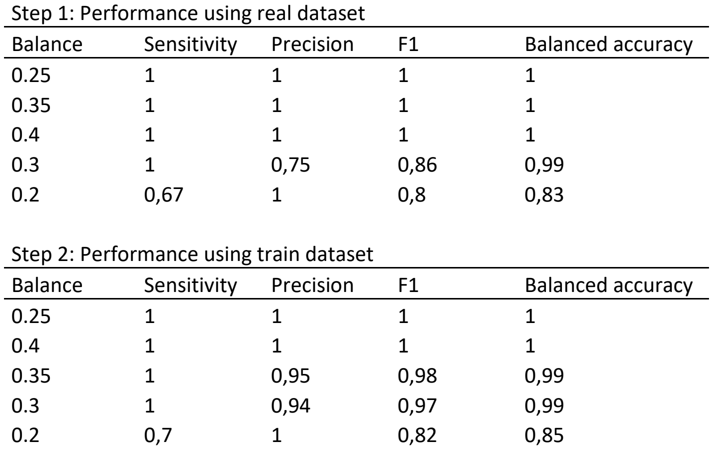
Con formato: Centrado

Figure 5. The top left section displays the original data and shows all units labeled as efficient after label balancing (step 2a), the top right section displays the original DMUs and shows all units labeled as inefficient after worsening the original inefficient DMUs (step 2b) and below is the labeled dataset that will be used for model training.

The third step involves training the NN machine learning model. For this purpose, we use the R package caret (Kuhn, 2008) to facilitate model training, specifically leveraging the NN implementation from the nnet package (Venables & Ripley, 2002). A grid is defined with selected hyperparameters for model fitting size (1, 5, 10, 20 and 30) and decay (0, 0.1, 0.01, 0.001, 0.0001). To determine these hyperparameters, a 5-fold cross-validation was implemented. After adjusting the model, the optimal hyperparameters for this dataset were: size = 20 and decay = 0.

Table 1 presents the performance of the fitted models for their respective balanced datasets. We use standard metrics commonly applied in ML problems. Our focus is on metrics related to the 'efficient' class, such as sensitivity (the proportion of actual positives correctly identified, or true positive rate), precision (the proportion of positive predictions that are actually correct, or positive predictive value), 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, ensuring equal weight is given to all classes regardless of imbalance). However, other metrics may be more appropriate depending on the specific case.

Due to the limited number of DMUs, a validation partition was not created. Performance was evaluated using the observed data, which remains consistent across all models. Since there is a tie for balance levels 0.25, 0.35, and 0.4, we consider the performance using the entire dataset, which differs by the balance level established. After this evaluation, 0.25 and 0.4 remained tied. Following the principle of parsimony, we selected the 0.25 imbalance dataset.



Step 1: Performance using real dataset

Balance	Sensitivity	Precision	F1	Balanced accuracy
0.25	1	1	1	1
0.35	1	1	1	1
0.4	1	1	1	1
0.3	1	0,75	0,86	0,99
0.2	0,67	1	0,8	0,83

Step 2: Performance using train dataset

Balance	Sensitivity	Precision	F1	Balanced accuracy
0.25	1	1	1	1
0.4	1	1	1	1
0.35	1	0,95	0,98	0,99
0.3	1	0,94	0,97	0,99
0.2	0,7	1	0,82	0,85

Table. 1 Performance results from every fitted model.

As shown in Table 1, the classification efficiency models demonstrate very high performance, primarily because the most critical factor in classification models is the dataset's complexity (He & Garcia, 2009). In our context, the datasets consist of only two well-defined classes (efficient and inefficient) with no overlap or rare cases, which directly influences the complexity level. While the relative imbalance amplifies this complexity, the issue has been effectively addressed by incorporating synthetic units.

Comentado [RG10]: He ampliado los puntos de la leyenda como has indicado.

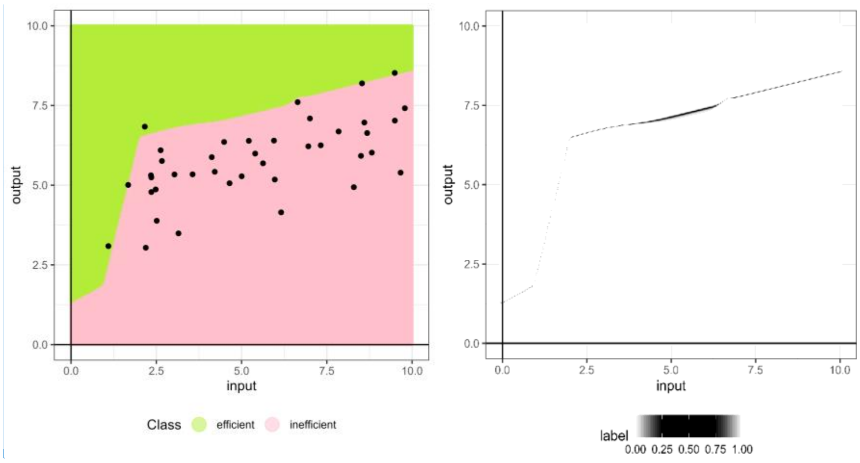


Figure 6. On the left, predicted regions by the new approach and the original DMUs unlabeled. On the right, the uncertainty shaded in black, as predicted by the fitted model (certainty region shown in white).

Figure 6 (on the left) displays the separating hyperplane generated by the trained model. Because it is not possible to visualize the hyperplane directly, we create a grid ranging from 0 to 10 in both dimensions and predict the class for each point. Green points represent those with a probability higher than 0.5, classified as ‘efficient,’ while the red points correspond to those with a probability of 0.5 or lower, classified as ‘inefficient. DMUs (black points) are classified using this method. In this two-dimensional example, where complexity is minimal (as previously discussed and confirmed by the performance metrics in Table 1), the fitted model shows high confidence in its predictions.

Figure 6 (right) represents the uncertainty through a gradient based on the levels predicted by the model. Areas with a predicted probability between 0.25 and 0.75 are shaded in black, indicating maximum uncertainty, while regions with probabilities closer to 0 or 1 are displayed in white, reflecting greater certainty. We can observe that the area of greatest uncertainty, near 5 on the input value in Figure 6 (right), corresponds to a region in Figure 5 where the units are more widely dispersed throughout the space.

The ability to obtain the probability of being efficient for all DMUs allows us to construct a ranking based on these probabilities. This ranking helps distinguish between efficient units, prioritize improvement efforts, and communicate results more effectively to stakeholders. Greater complexity and uncertainty result in higher discriminatory power observed within the ranking. Additionally, probabilities allow us to perform alternative analyses, such as peer selection. By setting a probability threshold to classify a DMU as 'efficient,' we can calculate the Euclidean distance (or weighted Euclidean distance) between all DMUs and those classified as 'efficient' at the determined threshold. The peer for an inefficient DMU is identified as the closest 'efficient' DMU based on the chosen distance metric. Examples of both analyses will be presented in the next section.

Seeking to improve the efficiency of the DMUs, we look for projections that reach the established probability level. First, we perform a SA analysis using the Rminer library (Cortez et al. 2004) and we find out that the model considers the output variable to be twice as important as the input variable when classifying a DMU as efficient or inefficient. The SA result for the input is 0.333 and for the output 0.667. Second, after SA analysis result, we can define the director vector $\left(-g_x, g_y\right)$ as $(-0.333, 0.667)$. Third, we determine the β value that, in the projection, achieves the established probability level for each DMU.

Figure 7 illustrates the projection of DMU 22, which is classified as inefficient with an input value of 4.49 and an output value of 6.36. Using the director vector, we calculate the minimum distance required to reach the specified efficiency confidence level, such as 0.75, through counterfactual analysis. The resulting projection reduces the input to 4.17 and increases the output to 7.03, corresponding to a β value of 0.18.

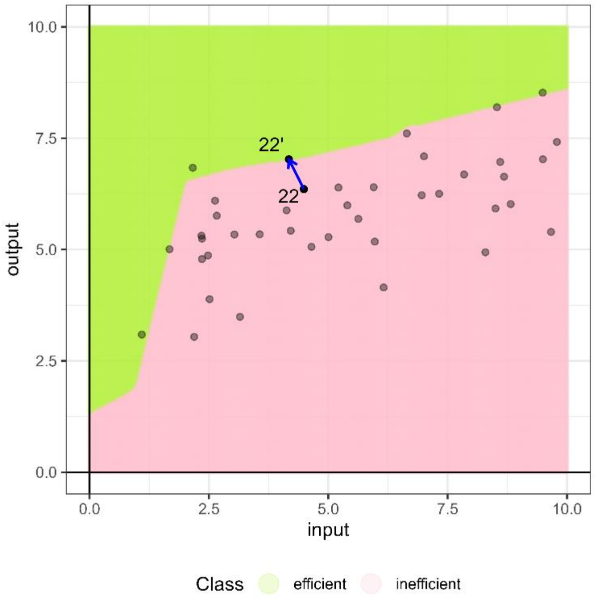


Figure 3. Projection of DMU 22 considering an efficiency level of 0.75.

Following the same process, we can determine the necessary resources for all DMUs to achieve the required efficiency level. However, in some cases, it may not be possible to reach the desired confidence level in the projections. This limitation arises because DMUs are not allowed to have resource levels below those observed in any variable. In such cases, we present the best result that the DMU can achieve under these constraints.

Counterfactual analysis enables the evaluation of statistical metrics such as mean, median, or standard deviation of the projections, providing a way to quantify the effort required to reach the desired efficiency confidence level.

In the following section, we demonstrate the merits of our method through its application to an empirical example based on data from the Programme for International Student Assessment (PISA) report. This empirical application will serve to showcase the practical effectiveness and utility of our approach in real-world scenarios, particularly in the context of educational performance evaluation and policy formulation.

4. An empirical application: the efficiency assessment of the Spanish educational sector

In this section, we exemplify the application of our novel algorithm to a dataset sourced from a public service. To illustrate the new methodology, we use data obtained from the Programme for International Student Assessment (PISA), administered by the Organization for Economic Co-operation and Development (OECD). PISA evaluates the competencies of students nearing the end of compulsory education, assessing their aptitude in essential academic skills necessary for effective participation in contemporary societies. Our empirical investigation focuses on analyzing schools as the fundamental unit, consistent with prevailing practices in educational efficiency evaluations (Johnes, 2015; Witte and López-Torres, 2017). This selection ensures alignment with prior research and relevance to ongoing discussions concerning educational institutions and their operational effectiveness. The dataset utilized encompasses data from the year 2018, comprising anonymized records from 999 Spanish schools randomly selected by the OECD.

Spain's educational system is decentralized, organized into 17 autonomous communities, each with distinct educational policies and practices. This decentralized structure adds complexity to our analysis, as variations across regions can significantly influence overall educational performance in PISA assessments. Understanding these regional nuances is essential for accurate interpretation and targeted interventions within Spain's diverse educational landscape. Additionally, assessing efficiency in the education sector involves examining input variables such as educational resource quality (EDUQUAL), reflecting available physical resources; the socioeconomic status index of students (ESCS), and the teacher-student ratio (TSRATIO), representing human resources within each school. Output variables considered are standardized test scores in mathematics (PVMATH), reading (PVREAD), and science (PVSCIE). We also consider two contextual variables: region (autonomous community) and type of school (SCHLTYPE) (public, private or charter school).

The observed variability in input and output variables across regions underscores significant disparities in educational resources and outcomes, emphasizing the need to investigate regional differences comprehensively. Given that the PISA dataset represents only a subset of the total population, our objective is not to calculate precise technical efficiencies of observed schools. Instead, we aim to leverage the estimated education production function to predict outcomes for schools beyond the observed sample. Consequently, a compelling scenario for educational decision-makers involves optimizing the allocation of educational and human resources to enable schools to attain or surpass certain thresholds in mathematics, reading, and science scores. Notably, modifying the socioeconomic status of students (ESCS), primarily determined by school location, may not be readily feasible for this purpose.

Building upon this production framework, we will employ the technique described in this paper, which combines ML techniques for classification and DEA, to determine a robust technical efficiency analysis. This approach allows us to capture the complex intricacies and idiosyncrasies of the educational sector in Spain, providing a more accurate and contextualized perspective efficiency.

Table 1 shows descriptive statistics for the sample: mean, standard deviation, number of DMUs per region, and the number of schools per type. Public schools represent 63.76% of the total, while charter schools account for 28.93% and private schools for 7.3%. Out of the 999 DMUs, the additive model identifies 38 as efficient, representing 3.8% of the total units evaluated. After identifying the efficient units, balancing the dataset, and increasing the number of inefficient units, the dataset used to train the model consists of 2921 units (961 efficient (32.90%) and 1960 inefficient (67.10%))

Con formato: Justificado, Interlineado: 1,5 líneas

		OUTPUTS						INPUTS							Type of school		
	Region	PVSCIE		PVMATH		PVREAD		ESCS		TSRATIO		EDUQUAL		Samples	Private	Charter	Public
1	Andalusia	469,45	(33,27)	466,20	(30,17)	464,93	(39,21)	2,53	(0,52)	10,16	(13,28)	3,43	(1,12)	49	0	10	39
2	Aragon	493,11	(29,1)	496,02	(29,85)	489,20	(34)	2,90	(0,41)	10,90	(13,25)	3,87	(1,1)	50	4	14	32
3	Asturias	496,11	(29,85)	490,71	(31,53)	494,58	(33,12)	2,84	(0,52)	13,06	(15,42)	3,80	(1,07)	54	1	16	37
4	Balearic Islands	481,14	(28,85)	481,90	(29,88)	478,25	(31,11)	2,76	(0,49)	15,27	(18,53)	3,69	(1)	50	5	11	34
5	Canary Islands	468,79	(32,75)	459,63	(33,48)	471,35	(35,28)	2,50	(0,5)	9,18	(3,93)	3,52	(1,11)	51	6	6	39
6	Cantabria	494,08	(29,38)	497,45	(32,8)	482,56	(31,43)	2,89	(0,42)	9,91	(2,72)	4,16	(0,76)	52	1	16	35
7	Castile and Leon	499,80	(30,33)	501,24	(30,55)	494,67	(33,74)	2,85	(0,41)	11,41	(11,91)	3,93	(1,06)	56	2	18	36
8	Castile-La Mancha	485,29	(26,66)	479,60	(27,87)	478,63	(30,77)	2,66	(0,48)	8,87	(2,32)	2,97	(1,2)	51	2	8	41
9	Catalonia	487,10	(37,09)	488,47	(35,9)	483,12	(39,7)	2,99	(0,5)	15,15	(19,33)	3,97	(1,12)	50	4	14	32
10	Extremadura	472,58	(32,85)	468,39	(31,9)	463,31	(35,65)	2,52	(0,43)	11,19	(3,46)	3,85	(1)	52	0	11	41
11	Galicia	510,17	(22,46)	497,23	(24,58)	492,40	(28,14)	2,82	(0,45)	10,38	(3,56)	4,07	(1,03)	54	4	11	39
12	La Rioja	481,83	(35,81)	492,65	(37,69)	461,53	(43,33)	2,71	(0,44)	8,56	(2,8)	3,92	(0,97)	42	0	20	22
13	Community of Madrid	495,18	(36,9)	495,74	(39,96)	482,40	(46,47)	3,23	(0,62)	11,71	(16,6)	4,08	(0,95)	130	39	29	62
14	Region of Murcia	480,64	(35,74)	474,97	(35,65)	482,96	(38,23)	2,52	(0,5)	8,89	(4,92)	3,63	(1,05)	52	0	15	37
15	Navarre	492,91	(35,72)	502,96	(35,03)	472,72	(43,18)	2,91	(0,46)	11,44	(10,25)	4,11	(0,99)	47	0	18	29
16	Basque Country	481,96	(35,08)	491,50	(41,26)	469,31	(39,94)	2,89	(0,52)	11,39	(11,76)	3,99	(1,02)	108	0	58	50
17	Valencian Community	479,33	(28,66)	474,87	(29,52)	474,18	(36,1)	2,73	(0,48)	11,85	(13,25)	3,82	(1,09)	51	5	14	32

Table 1. Descriptive statistics for the PISA dataset, Spanish schools, 2018.

Two ML techniques have been employed to classify the schools: Support Vector Machines (SVM) with a polynomial kernel (Karatzoglou et al. 2004) and neural networks (NN) with a hidden layer (Venables and Ripley, 2002). A grid is defined with selected hyperparameters for SVM model tunning: $degree\ (1, 2, 3, 4\ and\ 5)$, $data\ scaling\ (0.01, 0.1, 1, 10\ and\ 100)$ and $cost\ (0.001, 0.1, 1, 10\ and\ 100)$. For the neural network, a grid with selected hyperparameters is also defined for model fitting: $size\ (1, 5, 10\ and\ 20)$ and $decay\ (0, 0.1, 0.01, 0.001, 0.0001)$. The best models after tunning were: SVM with a polynomial kernel ($degree = 2$, $scale = 0.1$ y $C = 1$) with a cut off of 0.69 and neural network ($size = 5,\ decay = 0.1$) with a 24-5-1 structure with a cut off of 0.67.

Comentado [JLZP11]: Lo mismo, apenas se dice nada en la sección 3.2 de neural networks para seguir la metodología y el significado de los hyperparámetros.

Subsequently, the efficiency score was determined, also considering the case of calculating super efficiency. In the case of the scores estimated by the SVM model, it was not possible to calculate the efficiency score for 8 out of 999 units, since we got results related to infeasibilities. The ~~Spearmen~~ Pearson correlation between SVM and NN scores calculated according to our methodology is 0.961, showing the compatibility and robustness of both ML classification methods. It is important to note that direct comparison of DEA efficiency scores with those obtained using our novel method is not feasible due to fundamental differences in their underlying principles. Traditional DEA constructs an enveloping surface that envelops the observed data from above, representing the production possibility frontier. Efficiency scores in DEA are then calculated based on the distance of each DMU to this frontier, indicating how much outputs can be proportionally increased for the DMU to become efficient. Conversely, our novel method employs a classification model to determine a separating surface between efficient and inefficient units. This separating surface does not function as an enveloping frontier but rather as a boundary that discriminates between the two classes of DMUs. Efficiency scores in our method are derived from the distance of each DMU to this separating surface, reflecting the minimal changes required for an inefficient unit to be reclassified as efficient. Thus, while DEA efficiency scores measure the degree of deviation from an optimal production frontier, our method's scores quantify the classification margin relative to the separating boundary. However, although the scores themselves are inherently different and thus incomparable, the relative ranking of the units can still provide valuable insights. To evaluate the consistency in ranking between DEA and our novel method, we can use Spearman's rank correlation coefficient. This statistical measure assesses the degree to which the rankings of the DMUs are preserved across the two methods, offering a means to compare the ordering of efficiency even if the absolute scores differ. By examining Spearman's rank correlation, we can ascertain the alignment in relative efficiency rankings and gain a better understanding of the concordance between the two approaches in evaluating DMU performance. The Spearman's rank correlation between SVM's scores and traditional DEA is 0.962 and

Comentado [JLZP12]: ?

Comentado [JLZP13]: Yo creo que se podría calcular igual índice de Spearman y ver que pasa ¿no?. Estoy seguro que esto le pedirán los revisores. Eso si, manteniendo todos estos caveats.

between NN's scores and traditional DEA is 0.967. Both correlations show that the relationship between them is very high.

Comentado [JLZP14]: Añadir DEA como tercera línea dado que está en la gráfica

	Min.	1st Quartil	Median	Mean	3rd Quartil	Max.
DEA radial model	1.00	1.058	1.096	1.101	1.136	1.348
DEA super efficiency	0.899	1.060	1.097	1.100	1.137	1.348
SVM	0.925	1.035	1.075	1.079	1.115	1.305
Neuronal Network	0.795	1.035	1.075	1.078	1.105	1.325

Comentado [RG15R14]: He añadido el BCC y superefficiency

Table 2. Descriptive statistics of the scores for SVM and NN

In Table 2, we compare the results obtained by applying the selected ML models using our methodology. The median and the first quartile of the SVM and NN scores are identical. The significant difference is observed in the minimum value. This is illustrated in Figure 7, where the kernel density for SVM and the neural network overlap and are nearly identical. The results show that the DEA production frontier is 'further away' from the original observations as it is skewed to the right when compared to the disttrbutions of the ML classification methods. Also note that the efficiency scores can be smaller than one for DEA, correspoding to the superefficiency calculations for the efficient DMUs.

Comentado [JLZP16]: Esto es así ¿os es que el Kernel "rellena" valores menores que uno ?

Comentado [RG17R16]: Por utilizar super efficiency DEA. Nos pareció una comparación mas justa que con DEA normal

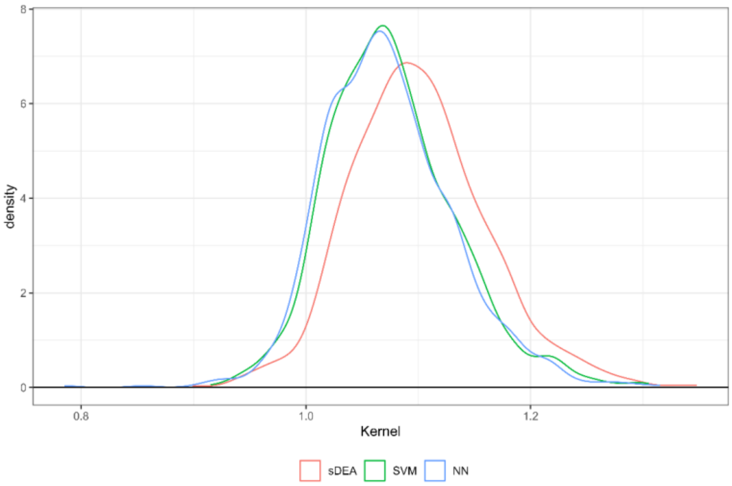


Figure 7. Kernel density estimation of the scores.

Comentado [JLZP18]: ¿Que representa la "s" delante de DEA?

Comentado [RG19R18]: Como nuestro metodo está preparado para detectar supereficiencia, para compararlo, utilizamos super efficiency DEA. La "s" es de ahí.

Another characteristic of estimating the efficiency score using a machine learning technique is the ability to discriminate Pareto-efficient DMUs. DEA models consider all Pareto-efficient DMUs as equally efficient. In contrast, our methodology calculates the distance between each DMU and the separating frontier and is capable of identifying some DMUs that are classidied as Pareto-efficient by DEA as inefficient. This is one of the advantages of applying machine learning techniques: they unveil measurement errors of the deterministic DEA based on a single sample thereby offering a better separating frontier, which is more flexible and aims to be more generalizable. In Table 3, we present 40 Pareto-efficient DMUs detected by the additive model (with an unitary effeincy score) and the scores achieved with our methodology. Many of these DMUs have scores below 1 based on ML classification methods. SVM identifies 9 DMUs as inefficient while NN identifies 2. The maximum score estimated by SVM is 1.065, while for NN it is 1.025. There are 5 DMUs that are infeasible for SVM, but NN can determine their scores. The minimum score estimated by NN is 0.785, and for SVM, it is 0.915. For this dataset, NN is able to estimate the score for all the DMUs, whereas SVM tends to classify more DMUs as inefficient, with slighter higher scores than NN as shown in Figure 2.

Comentado [JLZP20]: Una vez más el problema de la Pareto-eficiencia con las medida radial de output.

Comentado [JLZP21]: Hay que darle una vuelta a esto, Si una observación es Pareto-eficiente y no está dominada, no puede ser ineficiente. Esto es una defínción que se aplicaría a cualuier técnica. ¿Realmente las técnicas pueden identificar las DMUs que son Pareto-eficientes según DEA? Es decir, ¿todas las DMUs clasificadas como eficientes son Pareto-Eficientes? Es lo que parece intuirse del texto. Si no es asi no sñe si el concepto de P-E tiene mucho senbtidocon ML. Esto se une a todo lo dicho antes con el uso de medidas de eficiencia fuertes (aditivas) y débiles (radial).

Comentado [JLZP22]: He cambiado esta frase que creo refleja mejor lo que se pretende seguir.

Comentado [JLZP23]: ¿Tienen todos un índice 1? Lo he puesto porque no se sabe si se está trabajando con supereficiencia como en el gráfico 2.

Comentado [RG24R23]: El aditivo tradicional, sin super eficiencia.

	DMU	SVM	NN
1	18	0.945	0.945
2	67	0.945	0.975
3	85	1.045	0.995
4	117	0.975	0.925
5	145	0.945	0.965
6	149	0.995	0.975
7	241	-	0.915
8	250	1.055	0.985
9	268	0.965	0.975
10	273	0.985	0.995
11	316	0.955	0.975
12	318	-	0.925
13	335	1.035	0.935
14	391	0.965	0.955
15	442	0.975	0.985
16	462	0.965	0.975
17	480	-	0.785
18	520	0.985	0.975
19	557	0.965	0.965
20	588	0.975	0.995
21	698	0.935	0.855
22	700	0.935	0.975
23	706	0.965	0.965
24	745	0.915	0.965

25	759	1.035	1.025
26	776	1.045	1.005
27	787	0.975	0.975
28	801	-	0.995
29	803	1.065	0.945
30	804	0.975	0.965
31	863	0.965	0.965
32	874	0.955	0.925
33	878	1.015	0.995
34	882	0.975	0.955
35	906	1.005	0.985
36	910	-	0.905
37	986	1.005	0.965
38	992	0.945	0.965

Table 3. Pareto-efficient DMUs in the Pisa 2018 dataset, as identified through the additive model, along with their scores calculated using SVM and neural networks.

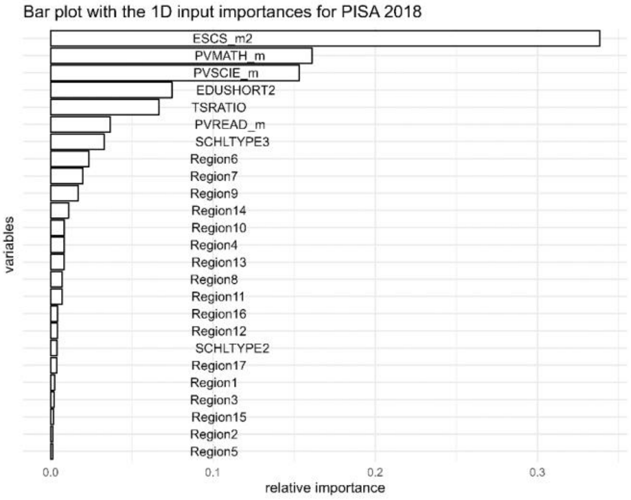
Tabla con formato

Comentado [RG25]: Hay que actualizar la tabla a los 40

The sensitivity analysis conducted on the SVM-calculated model reveals the following order of importance: the input ESCS (0.431) is the most important variable. It is followed by two outputs: PVMATH (0.193), PVSCIE (0.161), the remaining inputs: EDUQUAL (0.102), TSRATIO (0.04), SCHLTYPE (0.03), the last output, PVREAD (0.029) and one context variable: Region (0.015). The same analysis applied to the model using NN, results in the following variable importance ranking: ESCS (0.418), PVMATH (0.32), PVSCIE (0.09), SCHLTYPE (0.066), EDUQUAL (0.057), Region (0.027), TSRATIO (0.015) and PVREAD (0.007). Both results highlight the importance of the ESCS input in model training, assigning it similar significance. However, the SVM model's analysis distributes the remaining importance among more variables, such as PVMATH and PVSCIE, while the NN model focuses it on the second variable, PVMATH. In both models, the variables Region and SCHLTYPE are not very important in the presence of the other predictor variables, although the importance attributed by the NN is twice that of the SVM.

Comentado [JLZP26]: De nuevo no se sabe que representan estos valores porque habría que decirlo en la sección 3.2. Decir cuales son output y cuales inputs

Comentado [JLZP27]: Yo haría una tabla con dos columnas, Las variables por filas y las columnas SVM y NN



Finally, it is worth mentioning that our integration of Machine Learning with Data Envelopment Analysis may be also used to extrapolate efficiency assessments to unseen data, such as schools not included in the initial PISA sample. This capability is particularly valuable in educational policy making, where decision-makers need to predict and evaluate the efficiency of organizations that were not part of the (random) data sample that was used in the original study. In particular, our method utilizes classification models trained on known PISA data to establish a predictive framework that can assess whether an unseen school would likely operate efficiently or not based on its inputs, outputs and context variables. In cases where a school is predicted to be inefficient, our model not only quantifies the level of inefficiency but also provides specific output targets that the school needs to achieve to be considered efficient through the application of the XAI method. Moreover, this predictive ability enhances the practical utility of standard DEA by extending its applicability beyond the traditional analysis of existing units to include even potential future or hypothetical units. By enabling the evaluation of schools outside the observed dataset, our approach offers a robust tool for continuous improvement and strategic planning in education systems.

To conclude, we evaluate three hypothetical public schools of the same type, each with different characteristics, located in the Valencian Community. We assign the first school a value of ESCS equal to 25th percentile, with the values of the remaining variables set at the average of this region. The scores generated by the SVM model and the NN model are 1.055 and 1.065, respectively. The second school has the 90th percentile of ESCS, PVMATH has the 25th percentile, and the values of the remaining variables set at the average. The resulting scores are 1.185 from the SVM model and 1.155 from the NN model. The third school has the 75th percentile as ESCS, with the values of the remaining variables set at the average. The scores estimated by the SVM and NN models are 1.125 and 1.115, respectively. To interpret these results, it is crucial to consider the impact of ESCS on the efficiency scores of the schools. The ESCS is a relevant variable in our model, indicating that higher socio-economic status generally correlates with greater efficiency. For the first school, which has an ESCS in the 25th percentile, the scores generated by the SVM and NN models are 1.055 and 1.065, respectively. This lower ESCS suggests that students attending this school face more challenges in achieving efficiency compared to schools with higher ESCS values. The second school, with an ESCS in the 90th percentile and PVMATH in the 25th percentile, shows scores of 1.185 from the SVM model and 1.155 from the NN model. This school is more inefficient because despite students having a high ESCS, this does not materialize in a high score for PVMATH. We see that according to the sensitivity analysis, the high ESCS contributes to higher overall inefficiency scores, demonstrating the significant influence of socio-economic status. The third school has an ESCS in the 75th percentile and average values for the remaining variables. The SVM and NN models estimate the scores as 1.125 and 1.115, respectively. This school, with a moderately high ESCS, shows efficiency scores that fall between the first and second schools, because the output scores are on average (like the first school), but the socioeconomic status of the students is higher (yet below the 90th percentile of the second school).

Comentado [JLZP28]: The explanation was counterintuitive. I have changed it. The second school has much more ESCS (an input) and therefore it makes sense that it is less efficient if the output scores like PVMATH are below the average and, in particular, below that of the first school.

Comentado [RG29R28]: Exacto, a mayor input y menos output, mas ineficiente. La correccion que has puesto JL me parece bien

Comentado [JLZP30]: I do not understand this reasoning. The higher the efficiency score the worse. Right? In this case, the third school has a moderately high ESCS pero sus outputs están en la media, por lo que la ineficencia cae entre medias del 1 y 2.

Comentado [RG31R30]: Sí

Overall, these results highlight the pivotal role of socio-economic status in determining school efficiency. Rather interestingly Schools with higher ESCS values tend to achieve better efficiency scores, even when other variables such as academic performance (PVMATH) are lower.

Comentado [JLZP32]: ¿Es esto válido para toda la muestra? ¿has calculado la correlación entre el score de eficiencia y ESCS? Porque para las simulaciones de los tres colegios esta conclusion creo que no es válida. Es al revés porque a mayor ESCS no le corresponde mayor output scores y por tanto .

Comentado [RG33R32]: Parece que me quivoque en la interpretacion. A mas recursos, MAS ineficiencia.
Por mi, eliminamos este parrafo.

5. Conclusions and future work

A growing literature is focusing on the combined use of ML-DEA methodologies to predict organizational efficiency across diverse sectors. Although many of these studies focus on utilizing these methodologies to explore the interplay between machine learning enhancements and traditional DEA approaches, our research introduces a new dimension by integrating classification models with DEA. This fusion is not merely theoretical but also practically applicable, as demonstrated through our empirical study using PISA data. Our findings underscore that integrating ML classifiers with DEA not only helps in predicting the efficiency status of DMUs (or even unseen data) but also in refining the evaluation process of observations by introducing new judgment elements into the nature of traditional DEA assessments.

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 leverage 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, whether it be healthcare, education, or finance, providing customized efficiency assessments that are both insightful and actionable.

The integration of Machine Learning models with Data Envelopment Analysis represents a compelling advancement in the realm of efficiency analysis and offers a more nuanced understanding and interpretability of the results through variable importance ranking. This synthesis not only enhances traditional DEA by addressing its limitations—such as handling nonlinearity, model overfitting and lack of discriminatory power—but also leverages the computational prowess of ML to cut through intricate patterns and relationships within data that are otherwise not discernible. By employing ML techniques, particularly classification models, alongside DEA, we can effectively rank inputs, outputs, and contextual variables in terms of their impact on efficiency scores. This ranking is crucial for decision-makers as it identifies key performance drivers, enabling targeted improvements and resource allocation. The incorporation of ML thus empowers organizations to not only measure efficiency but also to understand the underlying factors contributing to inefficiency, facilitating strategic interventions that are both precise and impactful.

Compared to other methods, the integrated ML-DEA approach brings several distinct advantages:

1. *Improved Accuracy and Robustness*: The integration of ML algorithms enhances the robustness of the DEA model by enabling it to handle noise effectively through the cross-validation procedure that creates folds of the observed data into training and test sets.

2. *Enhanced Interpretability*: By employing explainable AI techniques, particularly the use of counterfactual explanations within the ML-DEA framework, our method not only quantifies efficiency but also explains it, constituting a valid alternative to second-stage methods that regress efficiency scores on contextual variables.

3. *Flexibility and Customization*: The modular nature of our approach allows for the integration of any classification ML technique into the algorithm—beyond SVM and NN, depending on the specific characteristics of the dataset and analytical needs. This adaptability ensures that the model remains relevant across different applications and evolves alongside advancements in

machine learning. The adoption of other classification methods and number of labels, e.g. graduating inefficiency score into groups, also constitutes a promising venue of future research.

In conclusion, the new integration of ML with DEA models represents a significant advancement in the field of efficiency analysis that enhances classical methods. Its ability to provide detailed, reliable, and actionable efficiency assessments could make it a valuable tool for researchers and practitioners alike. Ultimately, the true value and relevance of our contribution in the field of efficiency evaluation will be determined by its future application across diverse datasets and contexts, which will validate or challenge the robustness and adaptability of our approach.

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. These techniques, known for their effectiveness in capturing nonlinear relationships and high-dimensional data interactions, could be tailored to complement DEA's framework, potentially leading to more nuanced and detailed efficiency analyses. Secondly, the application of our integrated ML-DEA model to other domains, such as for market oriented organizations like firms, environmental sustainability and other public sector performance, 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. Such developments could make ML-DEA an central tool in strategic planning and resource management, especially in sectors where efficiency gains translate directly into improved outcomes for stakeholders.

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.

Andersen, P., & Petersen, N. C. (1993). A procedure for ranking efficient units in data envelopment analysis. *Management science*, *39*(10), 1261-1264.

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.

Bazaraa, M. S., Sherali, H. D., & Shetty, C. M. (2006). *Nonlinear programming: theory and algorithms*. John wiley & sons.

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.

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.

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.

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.

Cortez, P., & Embrechts, M. J. (2013). Using sensitivity analysis and visualization techniques to open black box data mining models. Information Sciences, 225, 1-17.

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.

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.

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. (2023). boostingDEA: A boosting approach to Data Envelopment Analysis in R. *SoftwareX*, *24*, 101549.

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

Guillen, M. D., Aparicio, J., & Esteve, M. (2023). 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.

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.

Johnes, J. (2015). Operational research in education. *European journal of operational research*, *243*(3), 683-696.

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.

Karatzoglou A, Smola A, Hornik K, Zeileis A (2004). "kernlab – An S4 Package for Kernel Methods in R." *Journal of Statistical Software*, **11**(9), 1–20.

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

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.

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

Lin, H. T., Lin, C. J., & Weng, R. C. (2007). A note on Platt's probabilistic outputs for support vector machines. *Machine learning*, *68*, 267-276.

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.

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.

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.

Pastor, J. T., Ruiz, J. L., & Sirvent, I. (2002). A statistical test for nested radial DEA models. Operations Research, 50(4), 728-735.

Pastor, J.T., Aparicio, J., Zofío, J.L., 2022. *Benchmarking Economic Efficiency*: *Technical and Allocative Fundamentals*. In: International Series In Operations Research & Management Science, ISOR, vol. 315. Springer Verlag.

Platt, J. (1999). Probabilistic outputs for support vector machines and comparisons to regularized likelihood methods. *Advances in large margin classifiers*, *10*(3), 61-74.

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

Simar, L., & Wilson, P. W. (2007). Estimation and inference in two-stage, semi-parametric models of production processes. Journal of econometrics, 136(1), 31-64.

Tayal, A., Solanki, A., & Singh, S. P. (2020). Integrated frame work 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., Boussofiane, A., & Dyson, R. G. (2015). Applied data envelopment analysis. Springer.

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

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.

Vapnik, V., & Cortes, C. (1995). Support-vector networks. Machine learning, 20(3), 273-297.

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.

Witte, K. D., & López-Torres, L. (2017). Efficiency in education: A review of literature and a way forward. *Journal of the operational research society*, *68*, 339-363.

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.

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.

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.

Bellman, R. (1966). Dynamic programming. *science*, *153*(3731), 34-37.

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.

Con formato: Inglés (Reino Unido)

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

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.

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.

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

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.

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

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).