Classifying ML-DEA: A novel approach integrating Machine Learning classification models with Data Envelopment Analysis to measure productive efficiency

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.~~Seiford & Zhu~~, ~~2002~~1997), healthcare (Olesen et al., 2007), and environmental performance assessment (Zhou et al., 2008), among others.

Comentado [JLZP1]: Me parece una cita más relevante el número especial de EJOR de 1997 (Volumen 98, Issue 2) que es un especial entero sobre banca: https://www.sciencedirect.com/journal/european-journal-of-operational-research/vol/98/issue/2
Se puede citar el editorial: New approaches for analyzing and evaluating the performance of financial institutions, A.N. Berger, P.L. Brockett, W.W. Cooper, J.T. Pastor.
U otro paper que encaje más

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 [JLZP2]: 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 [RG3]: ¿Qué titulo tiene?

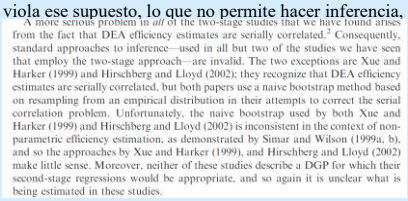
Comentado [RG4R3]: Este?: The estimation of productive efficiency through machine learning techniques: efficiency analysis trees

[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 [JLZP5]: ¿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 [JLZP6]: 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 is employed, revealing trends in EDC performance and identifying areas needing improvement.

Both streams of research have contributed valuable insights and methodologies for integrating ML with DEA. However, despite these 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 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 ~~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)~~as the response variable in the second stage~~. 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

Comentado [JLZP7]: 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 [JLZP8]: 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.

allow us to modify the measurement of the degree of efficiency 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.

Altoghether, 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 techniques that we will apply in this paper (Support Vector Machines and 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:

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

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 $(\boldsymbol{x}_o, \boldsymbol{y}_o)$, its technical efficiency can be calculated through the following program

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

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 the reference technology, with a bigger value indicating a worse degree of efficiency. The radial measure and its associated reference benchmarks on the frontier provides valuable insights into the performance of individual DMUs and can guide decision-makers in identifying opportunities for improvement.

2.2. Two well-known Machine Learning Techniques for Classification

In this subsection, we briefly outline the fundamentals of the two machine learning techniques that will be employed throughout the article: Support Vector Machines (SVM), and Neural Networks (NN), as well as eXplainable Artificial Intelligence (XAI). SVM is a favored supervised

Comentado [JLZP9]: 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.

learning algorithm used for classification and regression tasks. It works by finding the hyperplane that best separates the data points into different classes while maximizing the margin between classes. On the other hand, 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 of SVM and NN, which determine the label and the probability of belonging to that label, we can harness their capabilities to enhance the DEA methodology.

Comentado [RG10]: ¿Con esto bastaría para motivar las técnicas?

2.2.1. Support Vector Machines

Support Vector Machines stand as a stalwart within the machine learning toolbox, praised for their versatility and robust performance, in classification and regression tasks. A classification problem is distinguished from a regression problem by the nature of the target variable. In classification, the target variable is categorical and represents membership in a discrete class or category, whereas in regression, the target variable is continuous and represents a numerical quantity. For example, consider a classification problem where we aim to predict whether an email is spam or not spam based on metadata or various message features such as the frequency of certain keywords, text length, and the presence of hyperlinks. Here, the target variable would be binary: spam or legitimate. Conversely, in a regression problem, we might want to predict the price of a house based on features like size, location, and number of bedrooms, where the target variable would be the price, a continuous amount. In this paper, we will focus our attention on the context of binary classification: efficient units vs inefficient units. In this regard, this section offers a brief exploration of the main elements of SVM when it is used for classification tasks.

At its core, SVM operates on the principle of identifying an optimal hyperplane that effectively separates data points belonging to distinct classes (usually two classes or labels) in the feature space. This hyperplane is strategically positioned to maximize the margin, representing the perpendicular distance between the hyperplane and the closest data points from each class, known as support vectors. The seminal work of Vapnik and Cortes (1995) in the early 1990s laid the theoretical groundwork for SVM, emphasizing the importance of maximizing the margin to enhance generalization performance and to avoid overfitting problems.

A pivotal aspect of SVM lies in its ability to leverage kernels for achieving non-linear transformations in the feature space (the space defined from, for instance, the frequency of certain keywords, text length, and the presence of hyperlinks in the example mentioned above, or the

input-output combinations in production data for efficiency analyses). Kernels serve as a mechanism to map the input data into a higher-dimensional space, where linear separation of the two classes under study becomes feasible. Common kernel functions include the linear kernel, polynomial kernel, radial basis function (RBF) kernel, and sigmoidal kernel. Each kernel induces a specific transformation, altering the shape of the decision boundary and enabling SVM to capture complex relationships within the data. This transformative power of kernels enhances SVM's flexibility and enables it to tackle diverse classification tasks with varying degrees of complexity. However, in practice, the performance of SVM models heavily depends on the selection of hyperparameters, such as the regularization parameter (C), the margin (ε) and the choice of kernel function (which contains several kernel-specific parameters). To optimize model performance and prevent overfitting, cross-validation emerges as a valuable technique. Cross-validation involves partitioning the dataset into multiple subsets, training the SVM model on a subset, and evaluating its performance on the remaining data. By systematically varying hyperparameters and evaluating model performance across different subsets, cross-validation enables the selection of optimal hyperparameters that generalize well to unseen or new data.

Furthermore, SVM offers a means to assess the importance of predictors in predicting the response variable. In the SVM model, one can gauge the relative influence of different features on the classification outcome. This feature importance analysis provides valuable insights into the underlying data dynamics, guiding feature selection and model interpretation efforts.

To illustrate the practical application of SVM in classification, consider a dataset comprising two classes, depicted by red (southwest) and blue (northeast) points in a two-dimensional feature space. Figure 1 showcases this scenario, where the biggest circles denote the support vectors, the solid line represents the decision boundary, and the dashed lines delineate the margins.

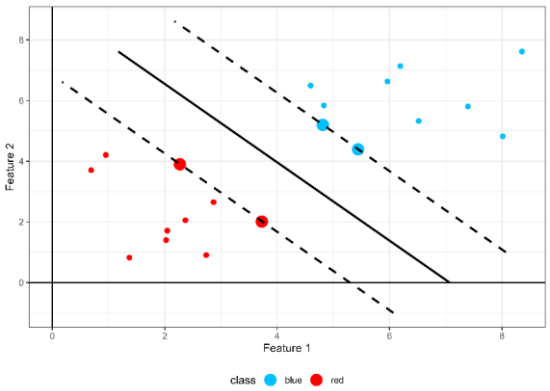


Figure 1. An example of a SVM-based model for classification

2.2.2. Neural Networks

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.

Similar to SVM, 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.

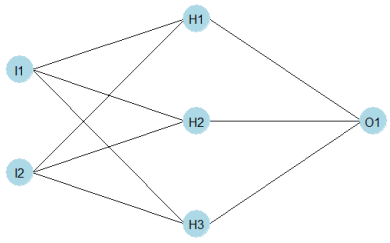


Figure 2. An example of an artificial Neuronal Network

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 (see, for example, 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.

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 to our approach in this paper to measure technical efficiency in the context of machine learning and efficiency analysis (see Section 3).

Comentado [JLZP11]: On esta analogía veo un problema. Resulta que el indicador radial de eficiencia de la ecuación (3) ya nos está proporcionando esta informacón. Entonces, si utilizamos la función radial de outputs, esta pregunta ya estaría contestada con los métodos tradicionales y no hace falta XIA. El lector se va a preguntar esto al leer la última frase de este párrafo. Esto está ilustrado justo en la siguiente sección, pero se podría poner una frase del tipo que las medidas de eficiciencia (como la output oriented radial del modelo (3)) implican contestar a un counterfactual del tipo, ¿qué pesaría si se redujese la ineficiencia en la cuantía determinada por el score de eficencia.?

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 the two previously discussed techniques: support vector machines (SVM) and neural networks (NN).

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 likens 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 [JLZP12]: He reescrito mucho este párrafo con la idea de mejorar la analogia o reinterpretación del DEA como un caso particular de XAI. Espero haberlo conseguido.

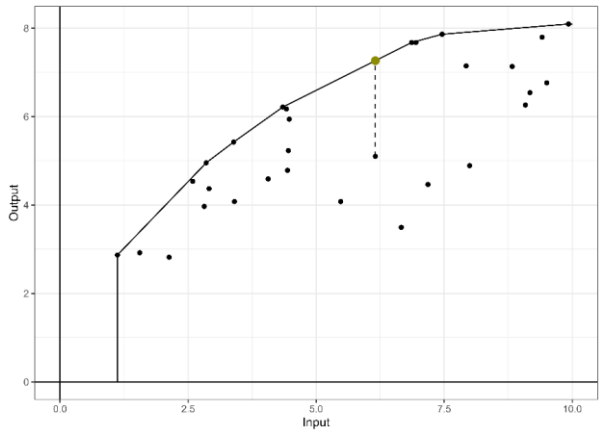


Figure 3. The output-oriented radial measure in DEA

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 a second phase, we employ a classification ML model, wherein the response variable is the efficiency class (efficient vs. inefficient), and the classification features includes 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. For instance, a similar concept can be extended to the efficient units within the framework of DEA. By doing so, we can ascertain a measure indicative of super-efficiency, introduced by Andersen and Petersen (1993), that allows differentiating among the subset of efficient DMUs, which otherwise have the same unitary score. This concept revolves around assessing each observation in relation to all other units within the dataset, wherein the evaluated observation is deliberately omitted from the analysis. Essentially, super-efficiency gauges the efficiency of a DMU by excluding the evaluated observation from the reference technology.

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)(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

Con formato: Revisar la ortografía y la gramática

[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.

there is no room in the observed sample for enhancing any input and/or output without compromising the feasibility of the unit under assessment.[3]

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

If $A_{DEA}(\boldsymbol{x}_o,\boldsymbol{y}_o) > 0$, then DMU $(\boldsymbol{x}_o,\boldsymbol{y}_o)$ is (technically) inefficient. The set of all inefficient DMUs is denoted as I. Otherwise, if $A_{DEA}(\boldsymbol{x}_o,\boldsymbol{y}_o) = 0$, then DMU $(\boldsymbol{x}_o,\boldsymbol{y}_o)$ 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). In particular, 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 overcome this hurdle, we propose balancing the sample of data. This step involves adjusting the class distribution to achieve parity between efficient and inefficient units. The selected technique for achieving this balance is synthetic data generation. In practice, this method is primarily applied to augment the representation of efficient units, which are often less prevalent in real datasets. This enrichment of the dataset contributes to more effective generalization 'out-of-the-sample' by mitigating the bias introduced by the original class imbalance. Next, we talk about the process that we implement in practice to generate the synthetic units.

Step 2a: First, we determined the necessary number of synthetic units to balance the proportion of units in both classes (efficient vs. inefficient units). To achieve this equilibrium, we project~~ed~~ the inefficient DMUs onto the DEA frontier using a radial model and incorporate~~d~~ them into the training set. However, we performed a conditioned selection of the synthetic units to cover

[3] As opposed to radial measures like (3), the use of additive measures in this first stage prevents the appearance of 'slacks' in the measurement of technical inefficiency, i.e., efficient DMUs belong to the so-called 'strongly' efficient frontier, and therefore are 'Pareto-efficient' and cannot be dominated in individual input or output dimensions, see Pastor et al. (2022; Chap. 2) for an introduction to technical efficiency measurement with DEA.

Comentado [JLZP13]: He metido una nota explicando que las medidas aditivas excluye la posibilidad que aparezcan slacks como con las radiales. Tenemos que hablar esto porque esto hace que el uso de (3) posteriomente pueda ser criticado, si es que puede haber slacks con el nuevo método.

Con formato: Fuente: Sin Negrita, Cursiva

Con formato: Sangría: Izquierda: 0,32 cm

Comentado [JLZP14]: Aquí tenenos el problema de los 'slacks' y la inconsistencia entre (4) y (3), aplicada a este step. O se utiliza siempre (4) en todo el paper, o no decimos nada como hasta ahora (borranado la nota), y cruzamos los dedos para que los evaluadores no se den cuenta.

as much of the frontier as possible within the region of observed inputs and outputs (bounded by the minimum and maximum observed values in the data).

Step 2b: Second, for each non-synthetic DMU, we assessed whether both inputs and outputs were situated in the first quartile. If a unit was found to be in the first quartile in at least half of the dimensions, the synthetic unit was generated through an input-oriented projection. This procedure ensured that we increased data density on the standard DEA frontier using the input-oriented radial model. Similarly, additional units needed to balance the classes were projected using an output-oriented radial model. This approach increased data density in the remaining area of the frontier. Subsequently, all produced synthetic units were classified as efficient and included in the dataset.

Step 2c: ~~Furthermore~~Third, to provide additional information to the ML model, we generated new inefficient synthetic units following the same methodology. Our investigations indicated that model predictions improved with this last~~step~~ addition, especially in cases with 50 DMUs or fewer. In this process, we considered the original DMUs and worsened them (in terms of more input and less output), resulting in new synthetic units. ~~After this step~~The goal is to obtain ~~, the~~ a proportion of efficient to inefficient DMUs of at least ~~was approximately~~ 1:2, which is deemed acceptable in the literature (He & Garcia, 2009).

Step 3: Implement a classification ML model in this phase, either Support Vector Machines (SVM) or Neural Networks (NN) as discussed in Section 2.2, 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 and a final classification model $\Gamma(\boldsymbol{x},\boldsymbol{y})$. $\Gamma(\boldsymbol{x},\boldsymbol{y})$ predicts the classification of input-output bundle $(\boldsymbol{x},\boldsymbol{y})$ as (technically) efficient (+1) or inefficient (-1).

Step 4: Select a standard technical efficiency measure (for example, the output-oriented radial model). Then, calculate the minimum changes required in inputs and outputs (following the projection strategy marked by the chosen efficiency measure) of each inefficient DMU to transition its classification from inefficient to efficient. In this way, we are applying ~~a~~the previously discussed ~~–~~counterfactual analysis. The optimization program to be solved is the following one in the case of resorting to the output-oriented radial model for evaluating unit $(\boldsymbol{x}_o,\boldsymbol{y}_o)\in I$:

Comentado [JLZP15]: Esto no se entiende bien. ¿Cual es el criterio para quedarse con algunas de las proyectas y otras no?

Comentado [RG16R15]: Esto se explica en el siguiente párrafo. Propongo que vayan juntos estos dos párrafos el 2a y el 2b

Comentado [JLZP17]: Outputs: top 1st quartile and inputs lowest 1st quartile (in value). Right? Para generar la synthetic con las mejores. Aclarar

Comentado [RG18R17]: No, la idea es que las DMUs muy ineficientes en alguna de las variables, se proyecten con orientación input. Asi poblamos forzamos a poblar la frontera para valores pequeños, existen mas ejemplos y el modelo intenta fallar menos en esta region.

Comentado [JLZP19]: See my previous comments about slacks and contradiction among measures.

$$\min\left\{\tau_o : (\boldsymbol{x}_o, \tau_o \boldsymbol{y}_o) \in E, \tau_o \geq 1\right\} = \min\left\{\tau_o : \Gamma(\boldsymbol{x}_o, \tau_o \boldsymbol{y}_o) = +1, \tau_o \geq 1\right\}. \quad (5)$$

In particular, to solve model (5)~~(5)~~, we will employ an approximate strategy outlined as follows (inspired ~~ated~~ by the line search algorithm without using derivatives considered by Bazaraa et al., 2006):

Step 4a: If the model initially classifies a unit as inefficient (-1), we increase the outputs proportionally by 1% (τ_o=1.01) to ~~enhance~~improve its efficiency level and determine $\Gamma(\boldsymbol{x}_o, \tau_o \boldsymbol{y}_o)$. We repeat this process until the ML model classifies the input-output point as efficient (+1). The required increase in outputs to reach the decision surface is then determined as the midpoint between the last two calculated increments. The search is terminated when the difference between the τ_o variable positions in two consecutive iterations is less than a predefined tolerance. The last value determined for τ_o is considered to be the efficiency score of the output-oriented radial model for unit $(\boldsymbol{x}_o, \boldsymbol{y}_o)$.

Step 4b: Additionally, it is possible to extend ~~the~~ Step 4a above to efficient units to measure super-efficiency, thereby distinguishing among the subset of Pareto-efficient DMUs in the data sample. To do that, we must solve the following optimization program for each observation $(\boldsymbol{x}_o, \boldsymbol{y}_o) \in E$:

$$\max\left\{\tau_o : (\boldsymbol{x}_o, \tau_o \boldsymbol{y}_o) \in I, \tau_o < 1\right\} = \max\left\{\tau_o : \Gamma(\boldsymbol{x}_o, \tau_o \boldsymbol{y}_o) = -1, \tau_o < 1\right\}. \quad (6)$$

In comparison to model (5)~~(5)~~, ~~in model (6),~~ we ~~have~~ replace~~d~~ 'min' with 'max' in model (6)~~(6)~~. This adjustment is made because, in this scenario, we aim to identify the first value of τ_o, with $\tau_o < 1$, for which the output-oriented radial projection of $(\boldsymbol{x}_o, \boldsymbol{y}_o)$, representing an efficient unit, transitions to being considered inefficient according to the classification model, that is, the first value of τ_o such as $(\boldsymbol{x}_o, \tau_o \boldsymbol{y}_o) \in I \Leftrightarrow \Gamma(\boldsymbol{x}_o, \tau_o \boldsymbol{y}_o) = -1$.

3.2. Feature significance analysis: The drivers of input and output inefficiency

~~Furthermore, W~~we also use our chosen classification ML techniques, ~~specifically~~ Support Vector Machine (SVM) and Neural Networks (NN), to elucidate the significance of variables within our model. ML methods offer a robust framework for feature importance analysis, allowing us to

Con formato: Revisar la ortografía y la gramática

Con formato: Fuente: Sin Negrita, Cursiva

Comentado [JLZP20]: Esta es la contradicción. No se puede utilizar la medidad radial para identificar 'Pareto-efficiency'. ¿O si? Es decir, ?pueden los modelos (5) y (6) generar slacks? Si no hay que explicar y vender esto porque es super importante. Es decir, la técnica permite acabar con el probelma de como el DEA adolece del problema de los slacks.

Con formato: Revisar la ortografía y la gramática

Con formato: Fuente: Sin Negrita, Cursiva

Con formato: Fuente: Sin Negrita, Cursiva

Con formato: Fuente: Sin Negrita, Cursiva

Con formato: Sangría: Izquierda: 1,27 cm, Sin viñetas ni numeración

discern the most influential factors driving the efficiency classification of DMUs. For SVM models, variable importance is typically inferred through examining the weights assigned to support vectors, where larger weights correspond to greater importance in separating different classes or categories. Additionally, techniques such as Recursive Feature Elimination (RFE) can be employed to iteratively identify and remove less relevant variables, thereby emphasizing the ones contributing most significantly to model performance. On the other hand, NN employ diverse strategies for assessing variable importance, including sensitivity analysis, gradient-based methods, and layer-wise relevance propagation. Sensitivity analysis involves perturbing individual input variables and observing the resulting changes in model output, providing insights into their relative impact. Gradient-based methods leverage the gradients of loss functions with respect to input variables to quantify their contribution to model predictions. Layer-wise relevance propagation decomposes prediction scores across network layers, attributing relevance to input features based on their influence on subsequent layers' activations. By harnessing these sophisticated techniques within our SVM and NN frameworks, we aim to unravel the nuanced interplay between input-output variables and efficiency outcomes, thus enhancing the interpretability and utility of our DEA-ML integration approach.

Comentado [JLZP21]: Esto no tiene nada que ver con los inputs y outputs del modelo ¿verdad? Entonces mejor cambiar input por 'decision' o algo así para no rear confusion.

Comentado [RG22R21]: Exacto, no son los inputs de DEA. Aquí se refiere a las entradas, al vector que se perturba para ver cómo varia la predicción del modelo.

Comentado [JLZP23]: De nuevo

Comentado [JLZP24]: De nuevo

Comentado [JLZP25]: Ahora si que se refiere a los inputs y outputs del modelo

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 Support Vector Machines (SVM).

In this example~~,~~ we create a data set made of 30 DMUs (D) that use a single input to produce a single output. Following the algorithm, step 1 ~~is to~~ labels the available data according to the additive model through standard DEA. In this example, 4 DMUs ~~are identified~~ are efficient with all their optimal slacks in model (5) equal to 0, ~~considering them efficient and~~ labeling them as such. The remaining 26 are marked as 'inefficient'. The efficient DMUs are: 13, 21, 22 and 25 (see Figure 4). In this case, there is an imbalance in the labels, with 13.33% of the units being efficient and 86.66% being inefficient.

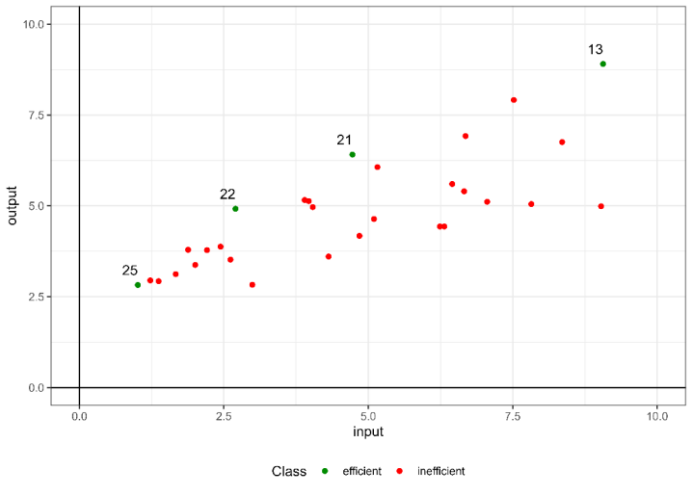


Figure 4. Labeling through the standard DEA additive model

~~The second s~~Step 2 of the method involves the creation of both efficient and inefficient synthetic units. The procedure for creating new synthetic efficient units (set $\hat{E}$) selects those projections of inefficient ones (step 2a) that covers ~~depends on the~~ all the observed regions ~~region~~ of the input-output space ~~where the unit is located~~. Following step 2b, ~~I~~if a unit falls within the first quartile in at least half of the variables, an additional ~~the~~ synthetic unit is created using an input-oriented projection of the radial model. The remaining synthetic units needed to balance the proportion between the two classes are generated using an output-oriented projection of the radial model (step 2c). For the creation of synthetic inefficient units (set $\hat{I}$), an equal number of units are randomly worsened (increased inputs and decreased outputs) as there are original inefficient units.

Figure 5 illustrates the evolution of the dataset. Initially, there were only 4 observations labeled as 'efficient', which increased to 26 after the creation of synthetic efficient units. Additionally, the number of inefficient DMUs increased from the original 26 to 56 after incorporating the synthetic inefficient units. Once the data imbalance has been addressed, the dataset consists of $\hat{D} = D \cup \hat{E} \cup \hat{I}$ with 82 units, with an approximatelyly~~ly~~ 1:2 ratio between units labeled as 'efficient' and 'inefficient'.

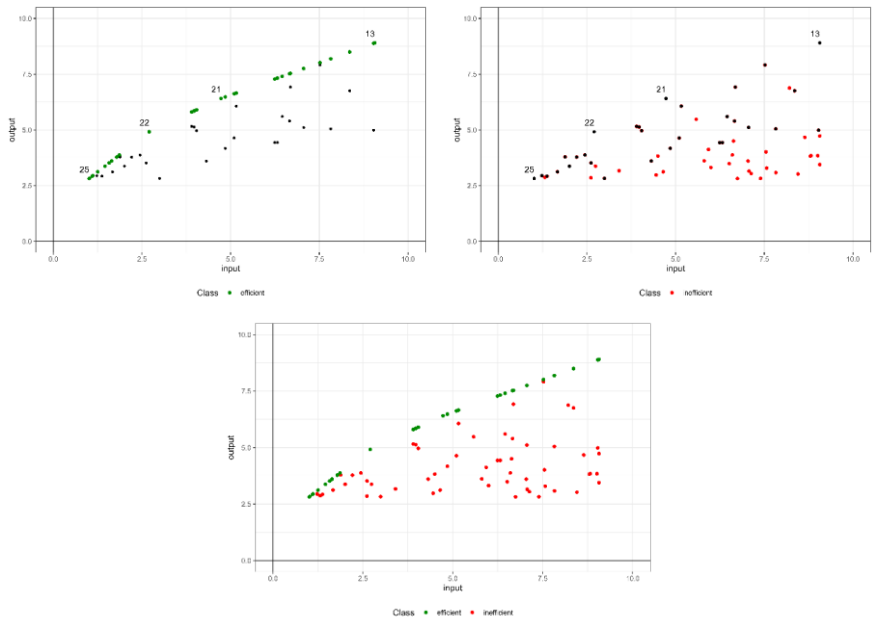


Figure 5. The top left section displays the original data and shows all units labeled as efficient after label balancing (step 2a and 2b), the top right section displays the original DMUs and shows all units labeled as inefficient after worsening the original inefficient DMUs (step 2c) -and below is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

The third step involves training the SVM machine learning model. The R package Caret (Kuhn, 2008) is used for model training—for brevity we dispense with Neural Networks for this example. The selected kernel is polynomial, as the resulting hyperplane shape fits the type of data being studied appropriately. For this purpose, the polynomial kernel model from Caret is utilized, which internally employs the R library Kernlab (Karatzoglou et al. 2004). A grid is defined with selected hyperparameters for model fitting : $degree\ (1, 2, and\ 3)$, $data\ scaling\ (0.1, 1, and\ 10)$, and $cost\ (0.1, 1, and\ 10)$. To determine these hyperparameters, a 5-fold cross-validation was implemented.

After adjusting the model, the optimal hyperparameters for this dataset were: $degree = 3, scale = 1, and\ C = 1$. To classify an observation as efficient, it is proposed that the model's label prediction be greater than 0.82.

Comentado [JLZP26]: While the hyperparameter degree and data scaling are easy to understand, the cost one is not intuitive. Is it related to the loss function? If so, what do the number represent? I would add a footnote explaining to what do the hyperparameters refer, so the reader does not have to go to the R package.

Comentado [RG27R26]: El coste es el parametro de regularizacion que se comenta, nosotros lo indicamos como C, pero la libreira de R lo llama cost.

Comentado [JLZP28]: Explain what the label prediction is. Is it a percentage of the times it is efficient among all the experiments?

Comentado [RG29R28]: Es el umbral que menos errores hemos detectado. Solo las observaciones con un valor superior al umbral se clasifican como eficientes

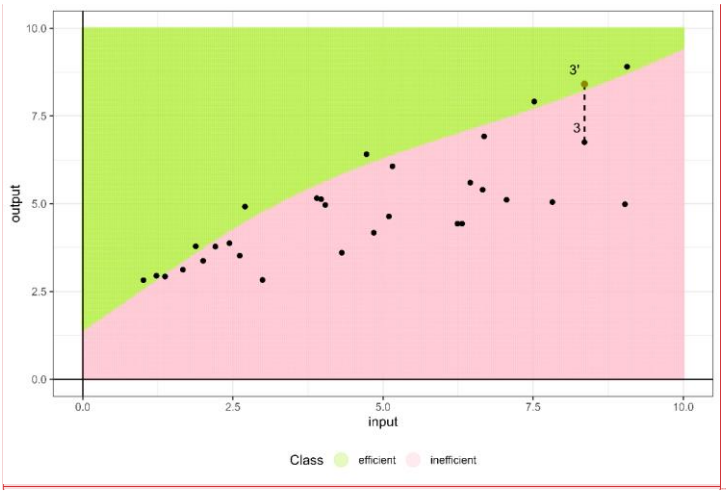


Figure 6. Predicted regions by the new approach

Figure 6 displays the class predictions for a grid of points between 0 and 10 in both dimensions. It is possible to observe the resulting separating hyperplane from the trained model. The original DMUs located in the efficient region (green region) are identified as efficient, with scores of 1 or lower if super-efficiency is applied. Those DMUs situated in the inefficient region (pink region) are identified as inefficient, and the score will be the average of the last two calculated increases. For example, in Figure 6, DMU 3, which is predicted as inefficient by the fitted model, has a resulting score of 1.245. This score suggests that if DMU 3 increases its outputs by at least 24.5%, it will be considered efficient by the model.

To assess the importance of variables in the trained model, we conducted a sensitivity analysis using the Rminer library (Cortez et al. 2004) in R. ~~This analysis relied on the mean absolute deviation over the median as the sensitivity measure~~ We used the Average Absolute Deviation (AAD) from the median as the sensitivity measure, which allowed us to quantify the relevance of each variable by measuring how much the prediction changes in response to alterations in a specific variable (Cortez, P & Embrechts, M. J., 2013). We chose to use this library instead of continuing to utilize the function provided by the Caret package, as the latter does not allow for determining the importance of variables for SVM. According to the analysis results, the most important variable for this dataset is the output, representing 0.521 of the total importance, while for the input, its importance accounts for the remaining 0.479.

Comentado [RG30]: He ampliado los puntos de la leyenda como has indicado.

Comentado [JLZP31]: Esto tampoco se entiende

Comentado [RG32R31]: Me referia que, como no sabemos por donde pasa el hiperplano, lo que hacemos es llenar la region de puntos, de 0 a 10 en ambas variables, y a continuación, predecimos para cada punto. De esta forma, podemos detectar por donde pasa el hiperplano.

Comentado [JLZP33]: Creo que habría que ampliar algo más el punto 3.2 metiendo algo de metodología porque esto se queda muy corto par entender por qué el output o input son más importantes.

Comentado [RG34R33]: Se podría ampliar indicando que utilizamos un método desarrollado por Cortez Monte-Carlo SA, que en vez de coger un vector con los datos que modificar, lo que hace es coger muestras "except that this method uses several training samples instead of the baseline vector ". La justificación para utilizar el Monte-Carlo SA es que ofrece unos porcentajes razonables, con los otros métodos disponibles no. Por eso no lo he añadido, peor podemos meterlo.

In the following section, we ~~will~~ 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 ~~will~~ exemplify the application of our novel algorithm to a dataset sourced from a public service. ~~In particular,~~ To ~~to~~ illustrate ~~our~~ the new methodology, we ~~will utilize~~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: ~~the~~ 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%)).

		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) ~~(Venables and Ripley, 2002)~~ 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.

Subsequently, the efficiency score was determined, also considering the case of ~~detecting~~ 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 ~~neuronal networks~~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 ~~encapsulates~~ 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

Comentado [JLZP35]: Lo mismo, apenas se dice nada en la sección 3.2 de neural networks para seguir la metodología y el significado de los hyperparámetros.

Comentado [JLZP36]: ?

Con formato: Tachado

Comentado [JLZP37]: 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.

SVM's scores and traditional DEA is 0.962 and between NN's scores and traditional DEA is 0.967. Both correlations show that the relationship between them is very high.

	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

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 ~~neural network~~ 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.

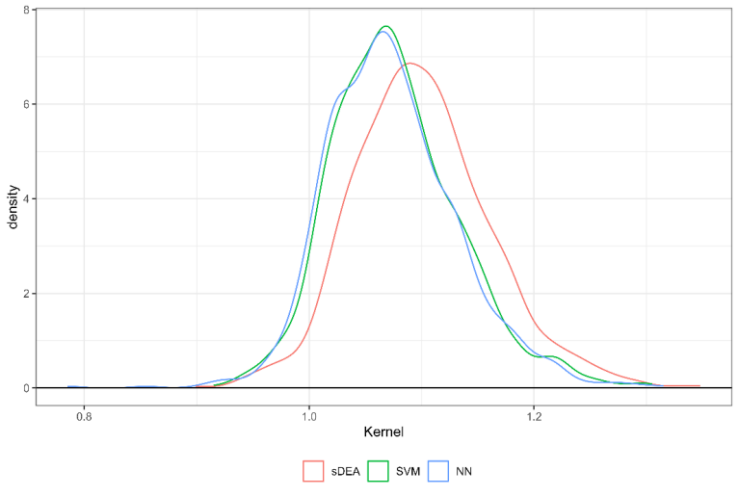


Figure 7. Kernel density estimation of the scores.

Comentado [JLZP38]: Añadir DEA como tercera línea dado que está en la gráfica

Comentado [RG39R38]: He añadido el BCC y superefficiency

Comentado [JLZP40]: Esto es así ¿os es que el Kernel "rellena" valores menores que uno ?

Comentado [RG41R40]: Por utilizar super efficiency DEA. Nos pareció una comparación mas justa que con DEA normal

Comentado [JLZP42]: ¿Que representa la "s" delante de DEA?

Comentado [RG43R42]: 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 ~~Pareto-efficient~~ DMUs that are classidied as Pareto-efficient by DEA as ~~as~~ inefficient. This is one of the advantages of applying machine learning techniques: they unveil ~~allow for~~ measurement errors of the deterministic DEA based on a single sample thereby offering ~~errors in the pursuit of achieving~~ a better separating frontier, which is more flexible and aims to be more generalizable. In Table 3, we present 38 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. ~~but~~ SVM identifie~~d~~s 9 DMUs as inefficient ~~and~~while ~~N~~NN identifie~~d~~s 2 ~~DMUs as inefficient~~. 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 ~~often assigning them~~ slighter higher scores than NN as shown in Figure 2.

Comentado [JLZP44]: Una vez más el problema de la Pareto-eficiencia con las medida radial de output.

Comentado [JLZP45]: 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 [JLZP46]: He cambiado esta frase que creo refleja mejor lo que se pretende seguir.

Comentado [JLZP47]: ¿Tienen todos un índice 1? Lo he puesto porque no se sabe si se está trabajando con supereficiencia como en el gráfico 2.

Comentado [RG48R47]: 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

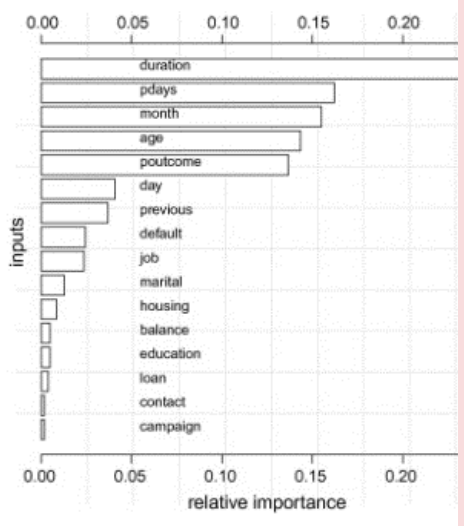
24	745	0.915	0.965
25	759	1.035	1.025
26	776	1.045	1.005
27	787	0.975	0.975
28	801	-	0.995
29	803	1.065	0.945
30	804	0.975	0.965
31	863	0.965	0.965
32	874	0.955	0.925
33	878	1.015	0.995
34	882	0.975	0.955
35	906	1.005	0.985
36	910	-	0.905
37	986	1.005	0.965
38	992	0.945	0.965

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.

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:~~follows,~~ 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.

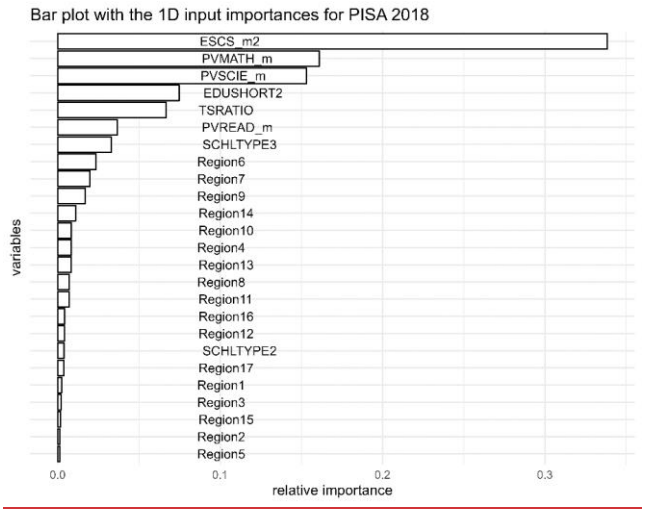
Comentado [JLZP49]: De nuevo no se sabe que representan estos valores porque habría que decirlo en la sección 3.2. Decir cuales son output y cuales inputs

Comentado [RG50R49]: Cortez propone este grafico para visualiar la importancia. ¿Lo meto?



Comentado [RG51R49]: Las variables, se comentan al principio, tambien lo recuerdo aquí entonces?

Comentado [JLZP52]: Yo haría una tabla con dos columnas, Las variables por filas y las columnas SVM y NN



Con formato: Centrado

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 ~~will present~~ evaluate three hypothetical public schools of the same type, each with different characteristics, located in the Valencian Community. We assign ~~T~~the first school ~~has the 25th percentile as the~~ a value of ~~for the~~ ESCS equal to 25th percentile, with the values of the remaining variables set at the average ~~for~~ 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 ~~as~~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 faces 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 ~~Despite having a lower~~ PVMATH ~~score.~~. 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 ~~in overcoming potential academic performance deficits~~. 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~~reinforcing the trend that higher ESCS is associated with higher efficiency~~ 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).

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.

5. Conclusions and future work

~~After examining the existing literature, it is clear that a~~A growing literature ~~number of researchers are~~ is focusing on the combined use of ML-DEA methodologies to predict organizational efficiency across diverse ~~various sectors~~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 ~~Decision Making Units~~ DMUs (or even unseen data) but also in refining the evaluation process of

Comentado [JLZP53]: 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 [RG54R53]: Exacto, a mayor input y menos output, mas ineficiente. La correccion que has puesto JL me parece bien

Comentado [JLZP55]: 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 [RG56R55]: Sí

Con formato: Superíndice

Comentado [JLZP57]: ¿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 [RG58R57]: Parece que me quivoque en la interpretacion. A mas recursos, MAS ineficiencia.
Por mi, eliminamos este parrafo.

Con formato: Color de fuente: Rojo

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 offer~~sing~~ 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, ~~and~~ model overfitting and lack of discriminatory power—but also leverages the computational prowess of ML to ~~uncover~~ 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.

Con formato: Fuente: Cursiva

2. *Enhanced Interpretability*: By employing explainable AI techniques, particularly the use of counterfactual explanations within the ML-DEA framework, our method not only quantifies

Con formato: Fuente: Cursiva

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.

Con formato: Fuente: Cursiva

In conclusion, the new integration of ML with DEA models ~~could~~ 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 ~~environmental~~firms, environmental sustainability and ~~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

~~indispensable~~central tool in strategic planning and resource management, especially in sectors where efficiency gains translate directly into improved outcomes for stakeholders~~ and the environment~~.

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, 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., 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.

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.

~~Seiford, L. M., & Zhu, J. (2002). Modeling undesirable factors in efficiency evaluation. European Journal of Operational Research, 142(1), 16-20.~~

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.

Con formato: Fuente: Cursiva

Con formato: Fuente: Cursiva

Con formato: Fuente: Cursiva

Con formato: Fuente: Cursiva

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.