Inteligencia Artificial Explicable para el análisis de eficiencia probabilístico

Autores: Juan Aparicio, José Luis Zofío, Víctor España y Ricardo González.

XLI Congreso Nacional de Estadística e Investigación Operativa

Proyecto PID2022-136383NB-I00 financiado por MCIN/AEI /10.13039/501100011033 /10.13039/501100011033 y por FEDER, UE

UNIVERSITAS Miguel Hernández RESEARCH INSTITUTE

MINISTERIO DE CIENCIA, INNOVACIÓN Y UNIVERSIDADES

Cofinanciado por la Unión Europea

AGENCIA ESTATAL DE INVESTIGACIÓN

Index

- Introduction
- Methodology
- An empirical application
- Conclusions

Introduction

XAI, DEA y ML

Introduction

- Data Envelopment Analysis (DEA) is one of the main techniques to measure efficiency.
- Traditional DEA approaches may encounter limitations in capturing the intricate patterns and structures inherent in complex datasets.
- Potential overfitting: Dealing with high-dimensional datasets or when the number of DMUs is relatively small compared to the number of inputs and outputs
- Traditional DEA is deterministic in nature.

Introduction

- We propose Machine Learning techniques to enhance the capabilities of DEA.

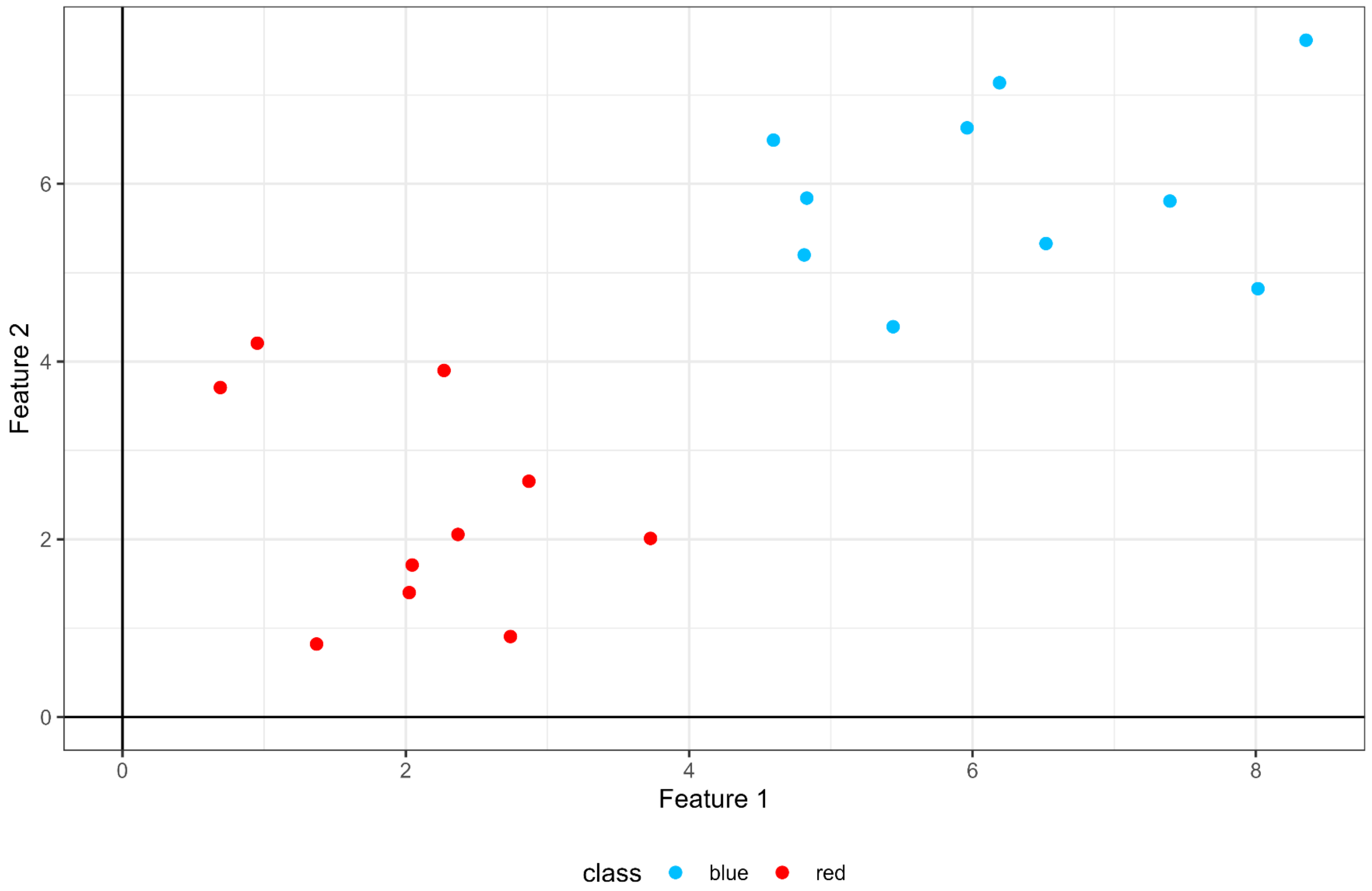
- Two predominant streams of research:

    - Adapting existing ML techniques to satisfy shape constraints
    - A two-stage approach to integrate DEA with ML techniques: 1. Determine efficiency score; 2. Apply a ML technique based on REGRESSION

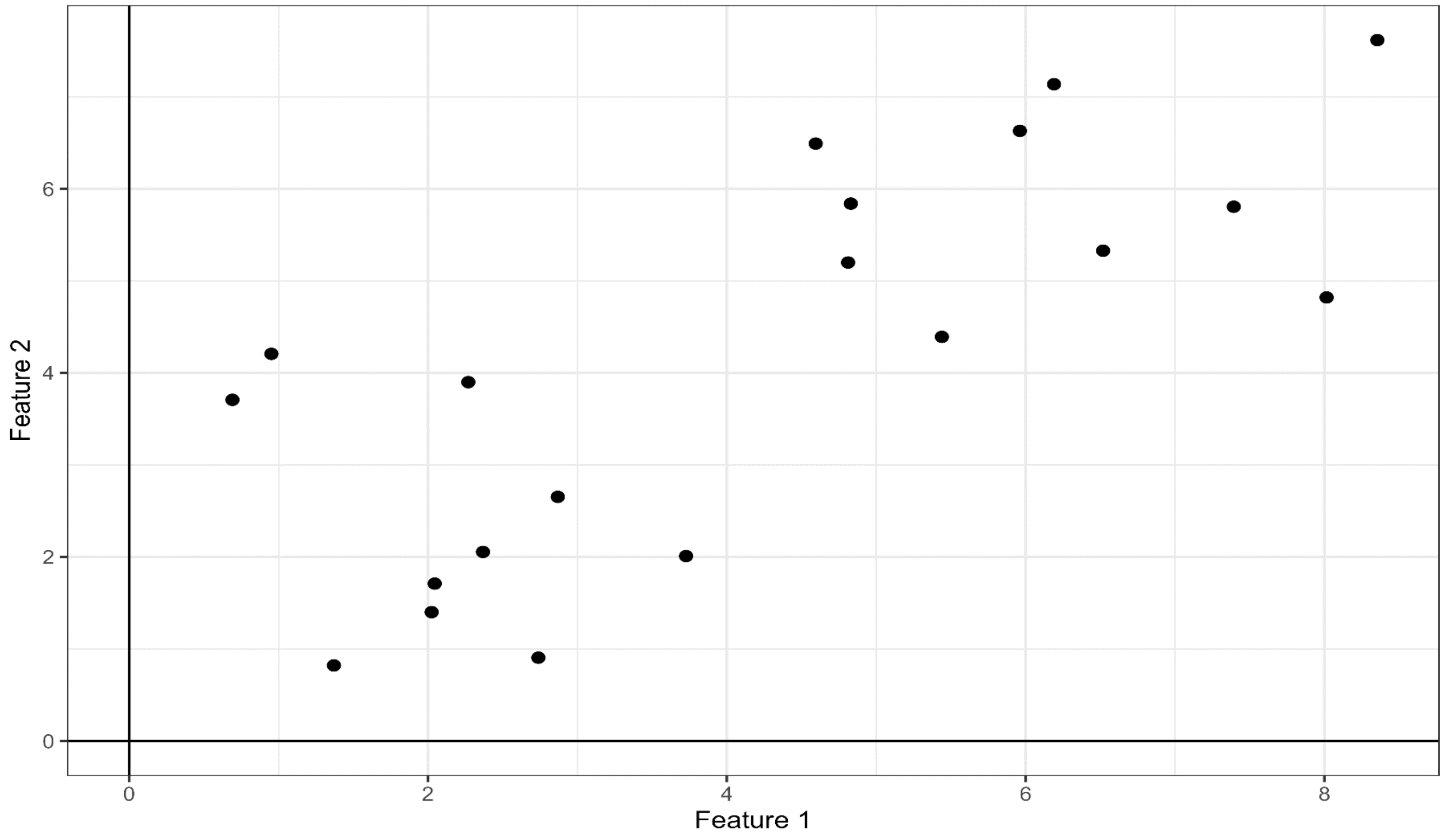
Introduction

- Types of machine learning:

Supervised Learning



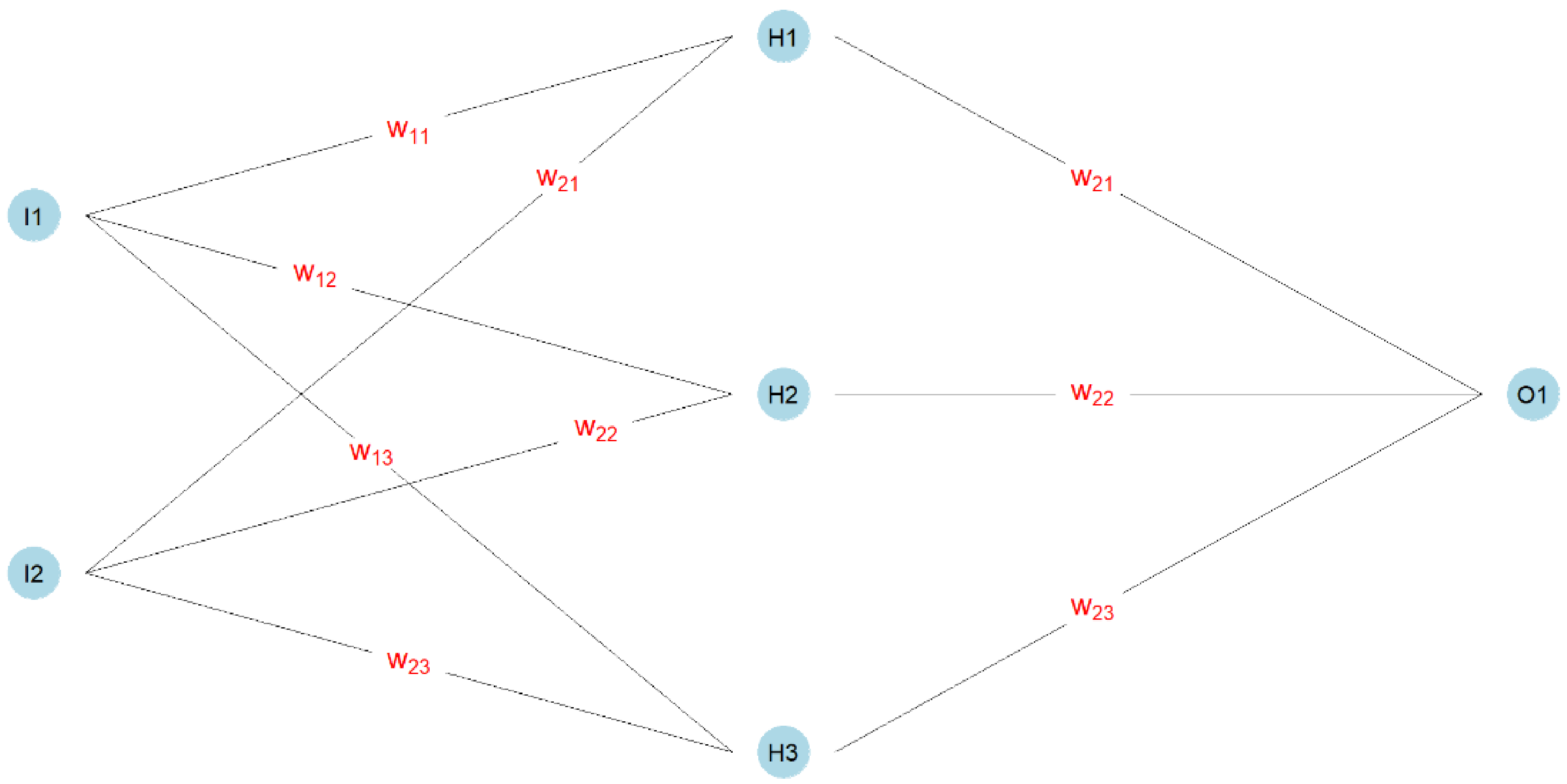
Unsupervised Learning



UNIVERSITAS Miguel Hernández RESEARCH INSTITUTE

Introduction

- Neuronal Network.
  - Iterative process known as backpropagation.
  - Hypermarameters determine network structure.
  - Variables that determine how the network is trained.



Introduction

- 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 inputs and outputs.
- Objective: change from the inefficient label to the efficient label.
- By incorporating advanced machine learning algorithms, we seek to provide more robust and accurate assessments of variable importance.

Methodology

Single input - output example

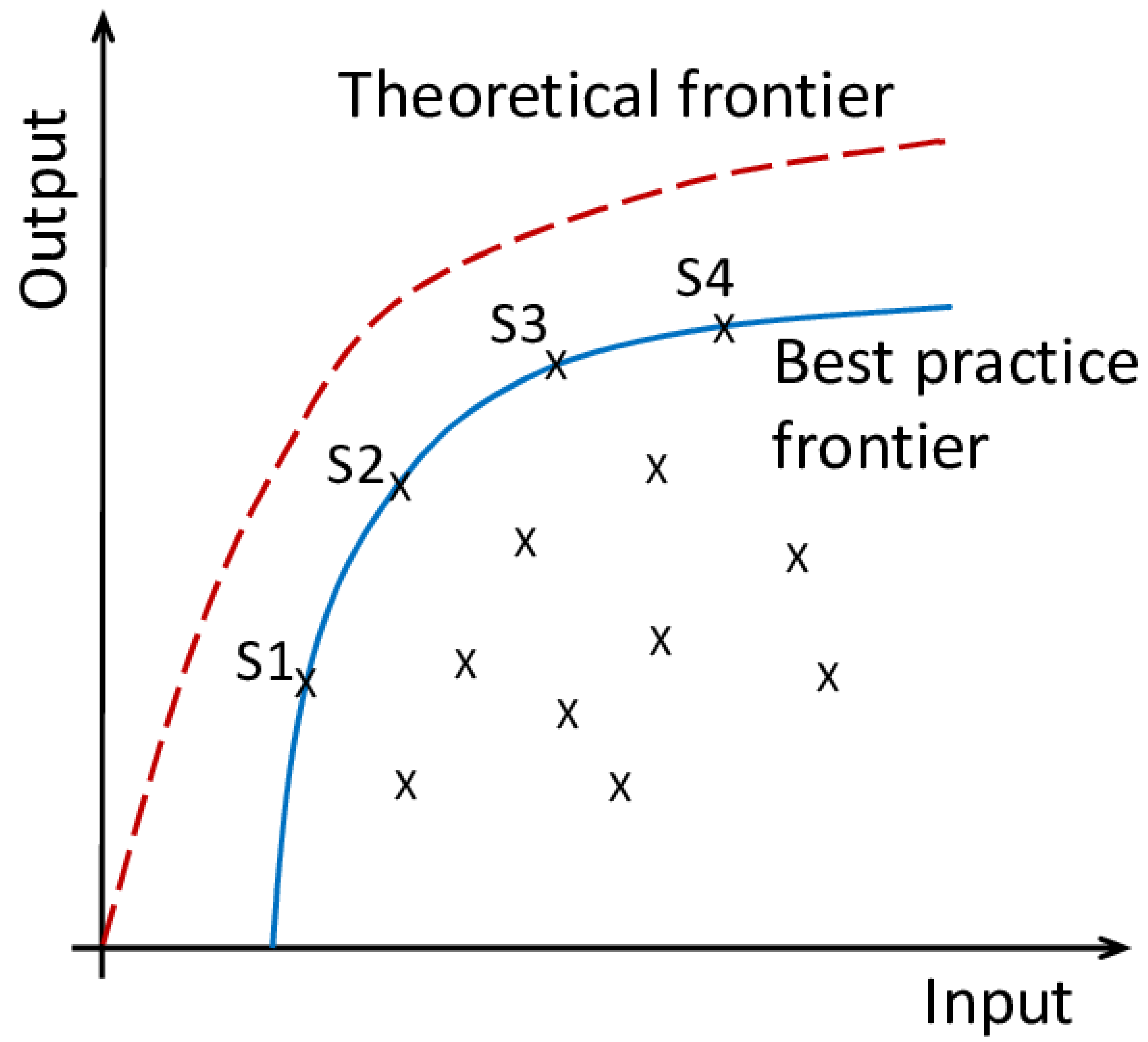
Single input - output example

- Set of Decision Making Units (DMUs) D, where DMU_j consumes $\boldsymbol{x}_j = (x_j^{(1)}, \dots, x_j^{(m)}) \in R_+^m$ to produce $\boldsymbol{y}_j = (y_j^{(1)}, \dots, y_j^{(s)}) \in R_+^s$
- DMUs are generated from some Data Generation Process (DPG) with the form of an unknown non-decreasing function (usually also concave) $f(\boldsymbol{x})\colon R_+^m \to R_+$
- Technical inefficiency occurs $\boldsymbol{y} = f(\boldsymbol{x}) - \boldsymbol{u}, \boldsymbol{u} \geq \boldsymbol{0}$

$$D = \left\{ \left(x_j, y_j \right) \right\}_{j=1}^{40}$$

Single input - output example

- DEA like an expert.
- Estimation of production frontiers
- Technology: $\Psi = \left\{ \left(\boldsymbol{x}, \boldsymbol{y} \right) \in R_{+}^{m+s} : \boldsymbol{x} \text{ can produce } \boldsymbol{y} \right\}$
- Usual Axioms
  - Deterministicness ($f(\boldsymbol{x}_i) \geq \boldsymbol{y}_i$)
  - Free Disposability (non-decreasing production function)
  - Convexity (concave production function)

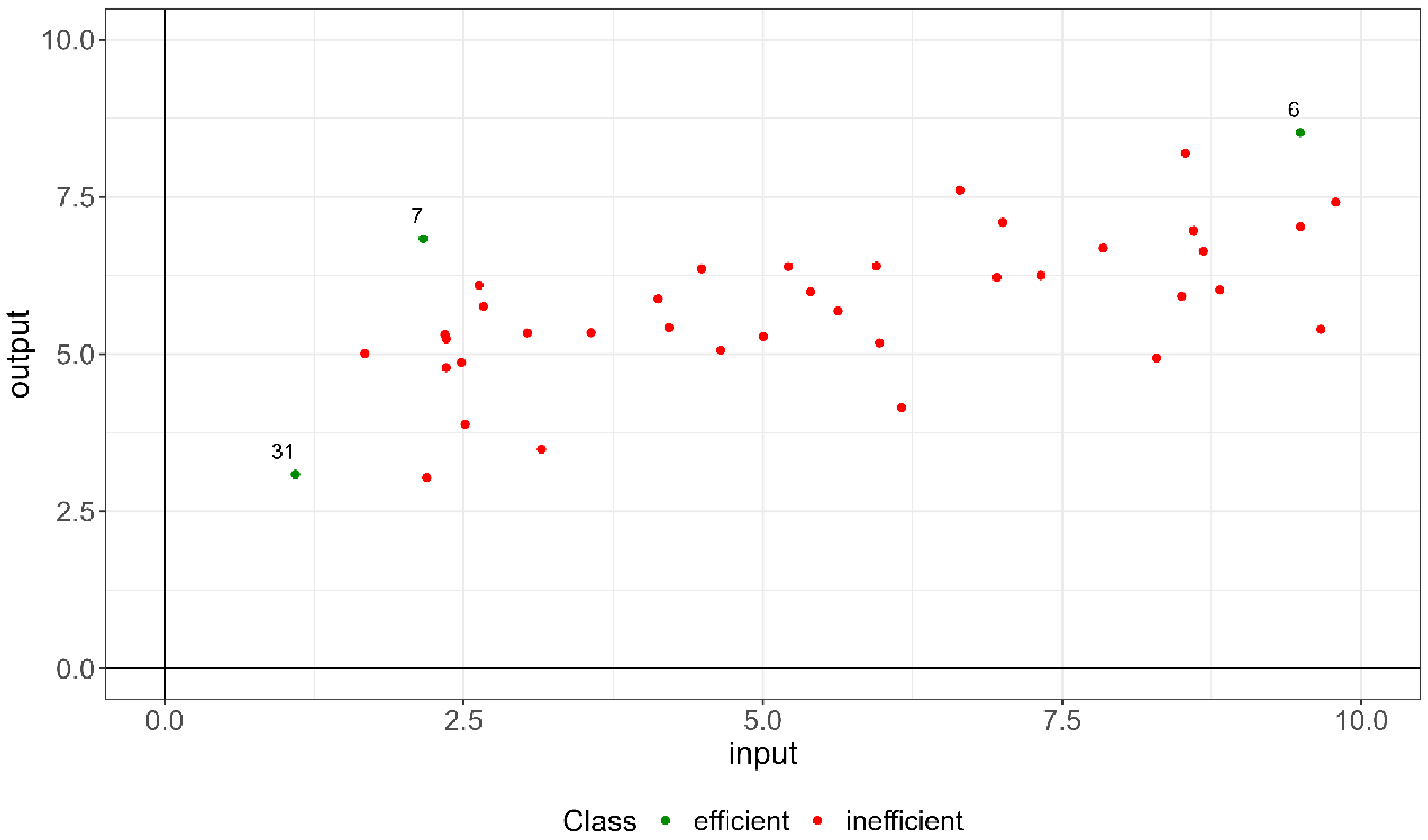


Single input - output example

- Step 1: Data labeling process.

- Utilize the additive DEA model (Charnes et al., 1985) to partition the set of DMUs in two categories.

$$D = E \cup I$$



Single input - output example

- Step 2: Class balancing phase: Synthetic data generation.

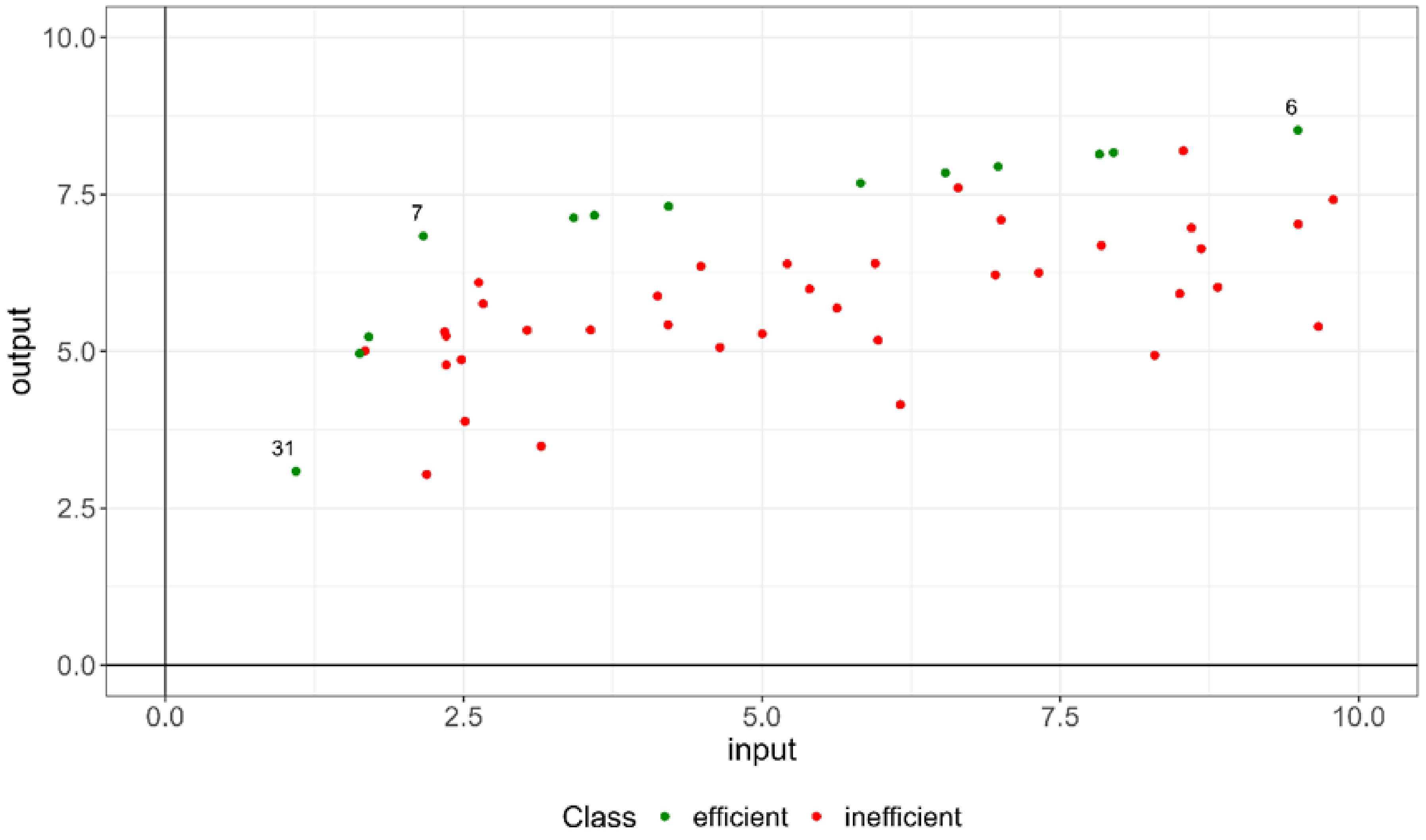
$$\hat{D} = D \cup \hat{E} = E \cup I \cup \hat{E}$$

- Imbalance level desired.

$$\pi_{\min} \in \{0.2, 0.25, 0.3, 0.35, 0.4\}$$

- Best imbalance performance:

$$\pi^*(\gamma^*, \alpha^*)$$

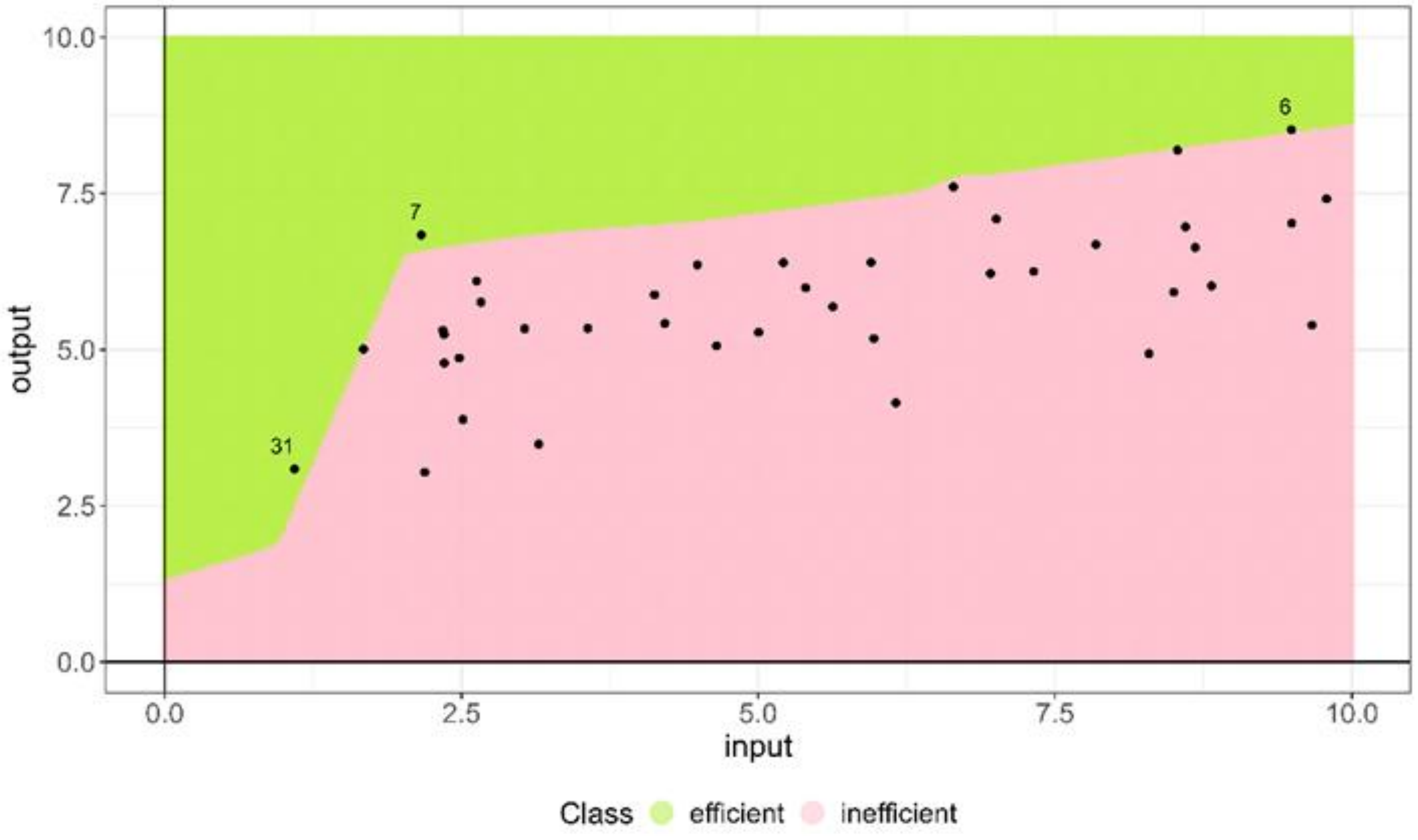


Single input - output example

- Tuning the model. Optimal hyperparameters.

$$\Gamma^*\big(x, y; \pi^*(\gamma^*, \alpha^*)\big)$$

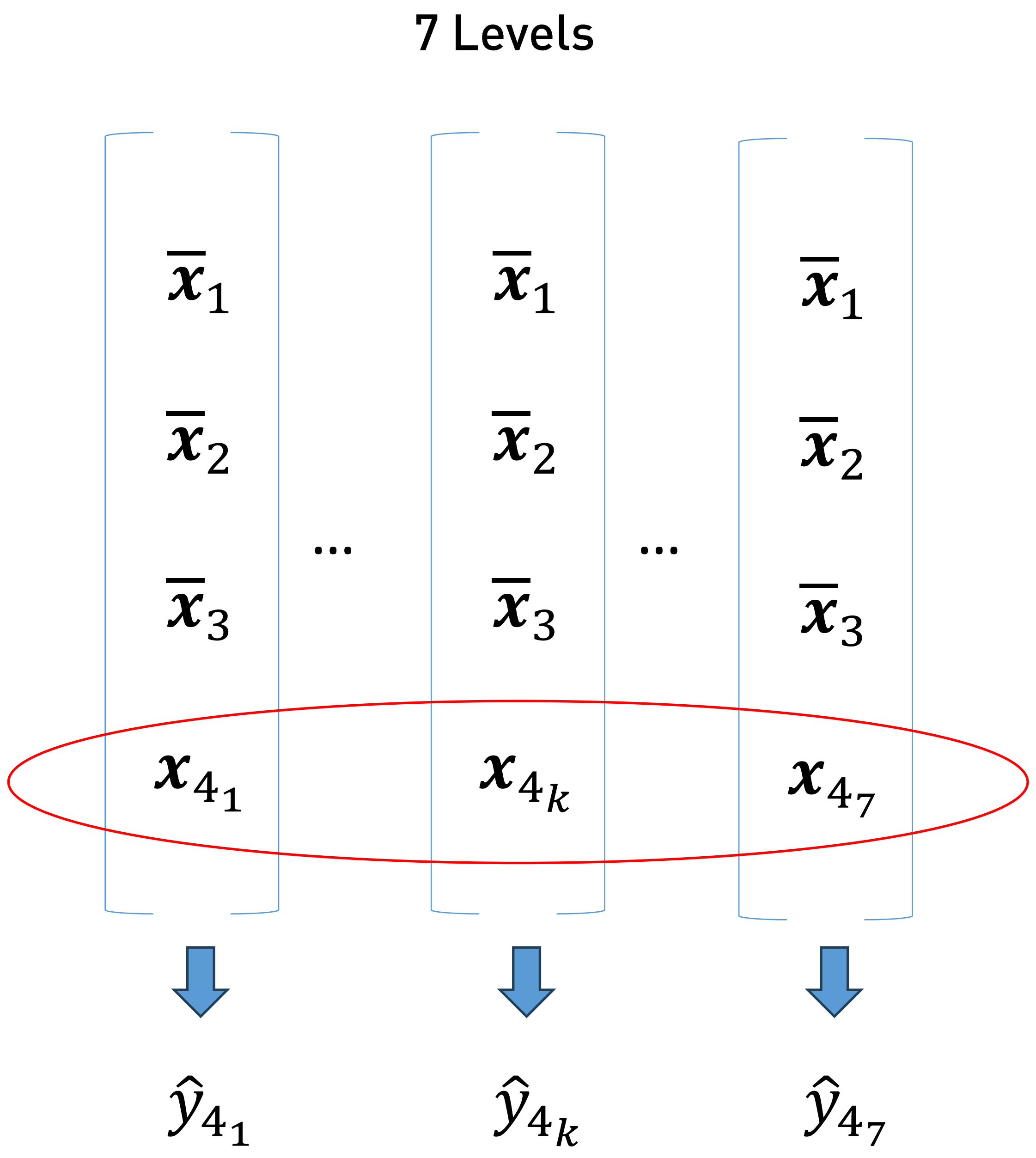
- Final regions are defined.



UNIVERSITAS Miguel Hernández RESEARCH INSTITUTE

Single input - output example

- Considering a given baseline vector $\boldsymbol{b}$.
  - Typically, $\boldsymbol{b}$ contains the mean or median.
  - But any other vector can be used.
- Then, it cycles through all variables: $\{\boldsymbol{x}_a : a \in \{1, \dots, m+s\}\}$.
- For each variable, L varaibles examples are built using all $\boldsymbol{b}$ values except $\{\boldsymbol{x}_{a_k} : k \in \{1, \dots, L\}\}$.



Single input - output example

- Sensitivity measure of variable importance
- Average Absolute Desviation (AAD)

$$\varsigma_a = \sum_{j=1}^{L} \left| \widehat{\boldsymbol{y}}_{a_j} - \widetilde{\boldsymbol{y}}_a \right| \Big/ L$$

- Relative importance

$$SA_a = \varsigma_a \Big/ \sum_{i=1}^{M} \varsigma_i$$

SA results:

- $SA_{x1} = 0.333$
- $SA_y = 0.667$

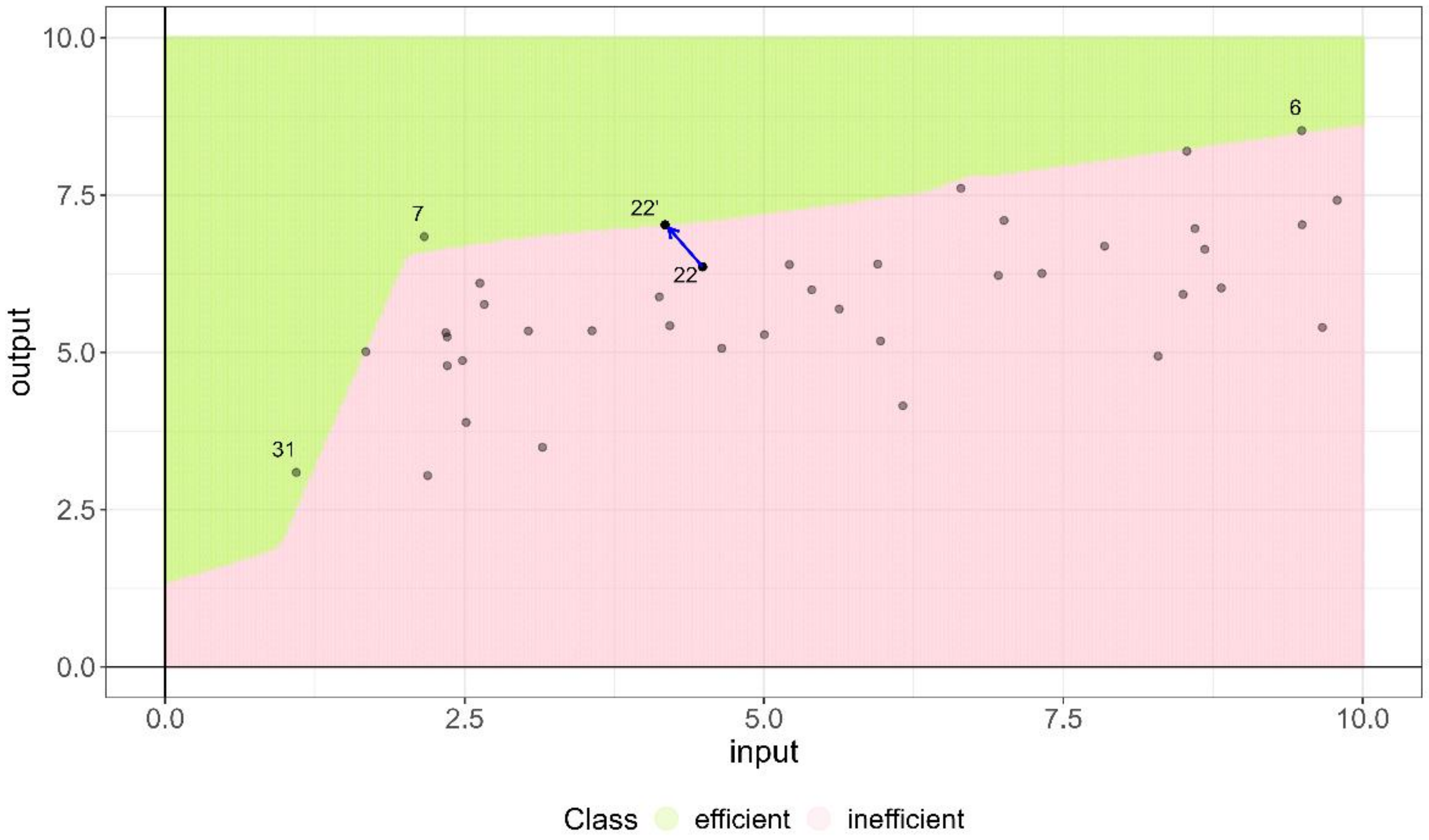
Single input - output example

- Measure technical inefficiency using counterfactual analysis.

- Define a directional vector:
  - $(\boldsymbol{g}_x, \boldsymbol{g}_y) = (SA_x \cdot \bar{x}, SA_y \cdot \bar{y})$

- β takes values within a predefined grid.

$$\Gamma^*(\boldsymbol{x}_\circ - \beta \boldsymbol{g}^x, \boldsymbol{y}_\circ + \beta \boldsymbol{g}^y \,; \pi^*(\gamma^*, \alpha^*))$$

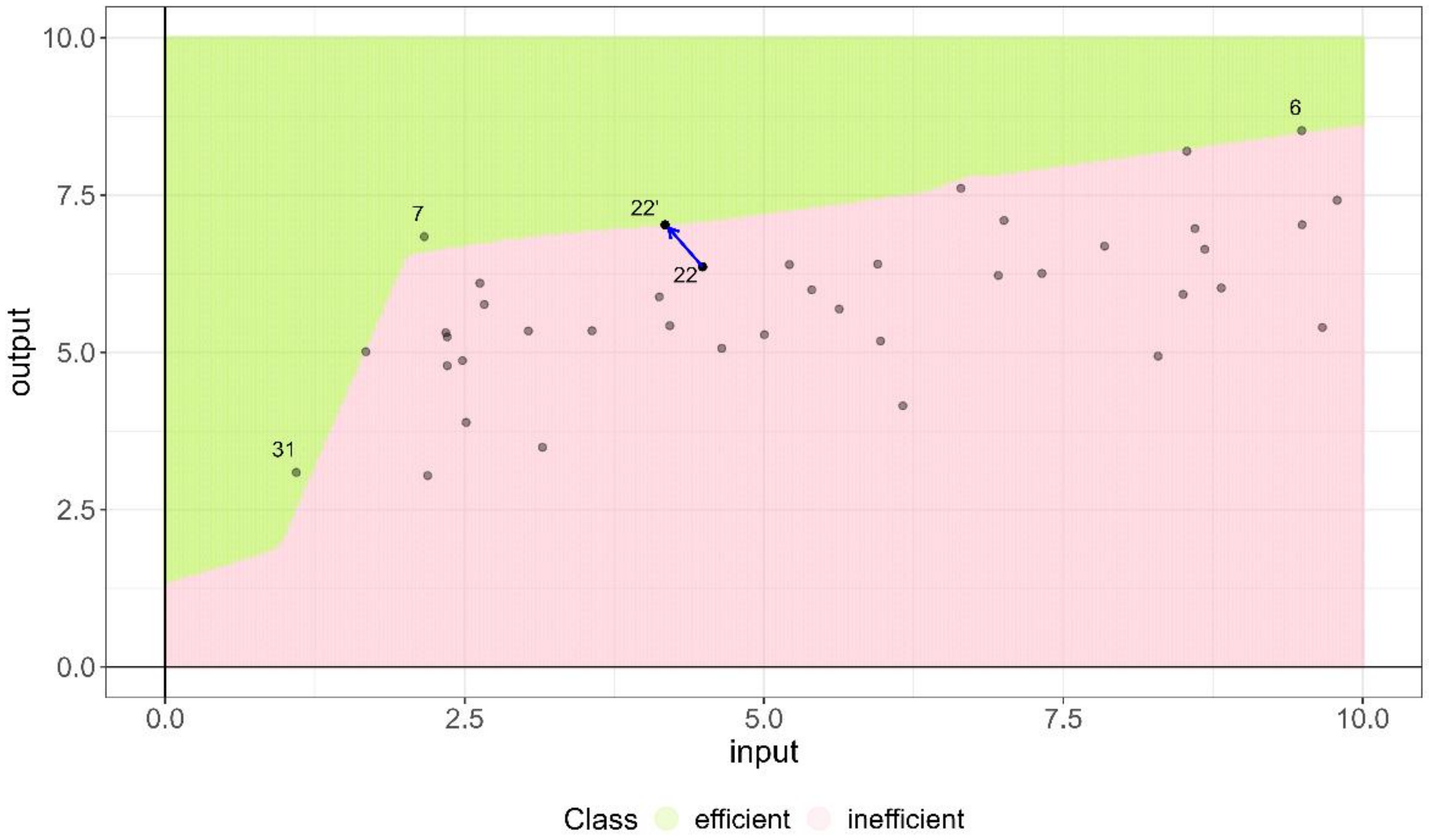


UNIVERSITAS Miguel Hernández RESEARCH INSTITUTE

Single input - output example

- Measuretechnical inefficiency using counterfactual analysis.
- Efficiency confidence level ($\bar{p} = 0.75$).

β	x1	y
0	4.49	6.36
0.17	4.17	7.03



UNIVERSITAS Miguel Hernández RESEARCH INSTITUTE

An empirical application

Efficiency Assessment of the Valencian Food Industry

The Efficiency Assessment of the Valencian Food Industry

- A dataset obtained from Iberian Balance Sheet Analysis System (SABI).
- The dataset utilized encompasses data from the year 2023, comprising records from 97 food industry companies located in the Valencian Community, each with more than 50 employers.
- Input variables: Total assets, number of employees, tangible fixed assets and personnel expenses.
- Output variable: Operating income.

The Efficiency Assessment of the Valencian Food Industry

- Step 1: Data labeling process.
  - 15 DMUs are efficient (15,46%)

- Step 2: Class balancing phase.

- Step 3: 6 NNs are fine-tuned

$$\begin{cases} \pi_{\min} \in \{0.20,\ 0.25,\ 0.30,\ 0.35,\ 0.40\} \\ \alpha \in \{0,\ 0.1,\ 0.01,\ 0.001,\ 0.0001,\ 0.00001\} \\ \gamma \in \{1,\ 5,\ 10,\ 20,\ 30\} \end{cases}$$

π_{min}

The Efficiency Assessment of the Valencian Food Industry

Performance using observed dataset.

π_{min}	Balanced accuracy	F1-Score	Precision	Sensitivity
0.40	0.97	0.97	1	0.93
0.1543*	0.94	0.75	0.60	1
0.20	0.93	0.93	1	0.87
0.30	0.85	0.79	0.85	0.73
0.35	0.85	0.79	0.85	0.73
0.25	0.79	0.69	0.82	0.60

The Efficiency Assessment of the Valencian Food Industry

- Hyperparameters

$$\pi_{\min} \in \{0.20,\ 0.25,\ 0.30,\ 0.35,\ 0.40\} \longrightarrow \pi^{*}_{\min} = 0.4$$

$$\alpha \in \{0,\ 0.1,\ 0.01,\ 0.001,\ 0.0001,\ 0.00001\} \longrightarrow \alpha^{*} = 0.01$$

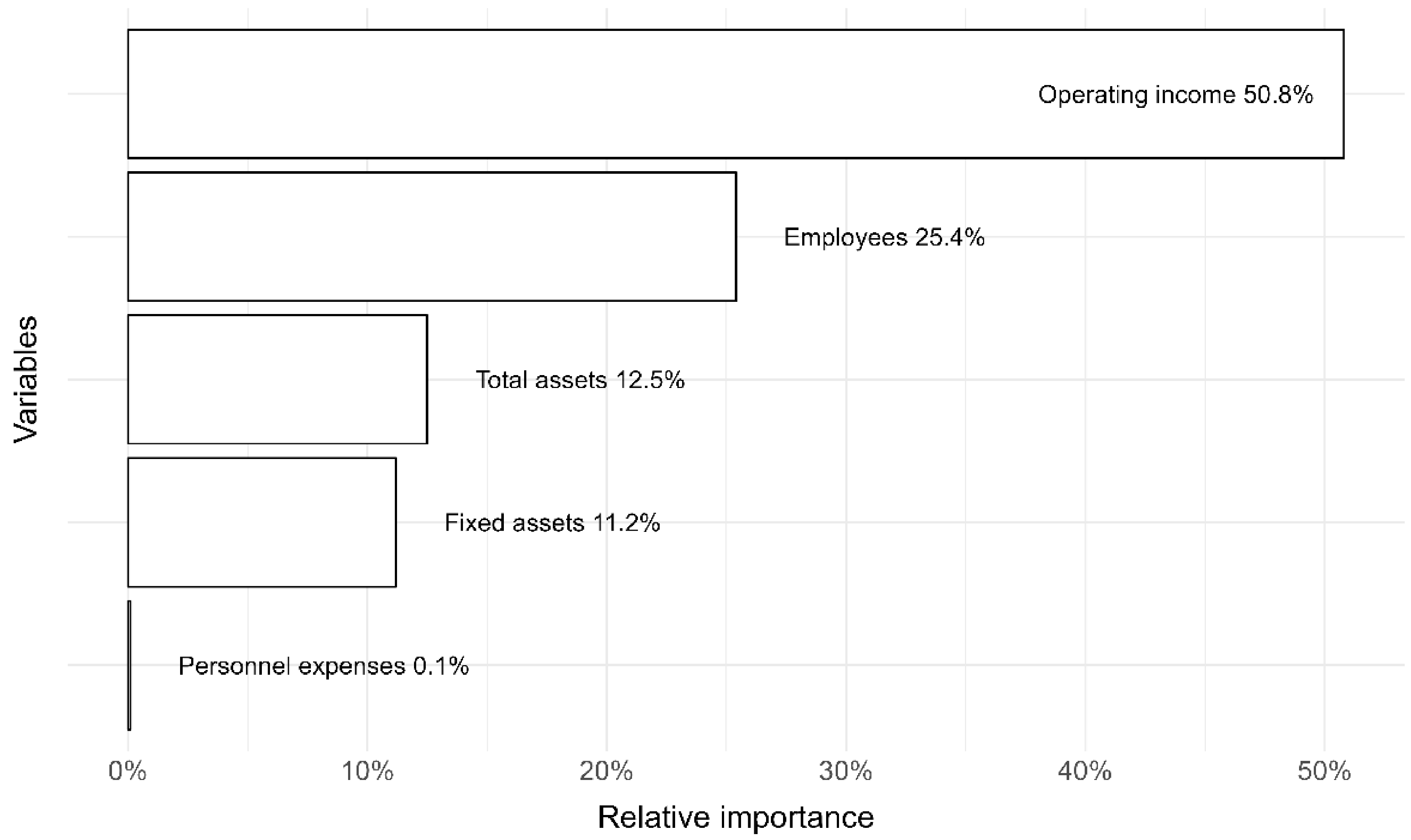
$$\gamma \in \{1,\ 5,\ 10,\ 20,\ 30\} \longrightarrow \gamma^{*} = 20$$

The Efficiency Assessment of the Valencian Food Industry

Results:

- 14 DMUs are predicted to have a probability of efficiency exceeding 0.5, one less than initially labelled in the first step by standard DEA.

- Operating income is the most important variable and personnel expenses is not important.



The Efficiency Assessment of the Valencian Food Industry

Efficiency profiles of firms 22, 26, 83, and 36 for $\bar{p} = 0.85$.

	Firm #22			Firm #26		
	Peer (#18)			Peer (#17)		
	Observed	Predicted		Observed	Predicted	
Total assets	24.71	23.20	(-6.11%)	27.93	25.36	(-9.23%)
Employees	212	196.97	(-7.09%)	80	54.33	(-32.09%)
Fixed assets	11.46	10.96	(-4.39%)	13	12.14	(-6.62%)
Personnel expenses	8	8.00	(-0.02%)	2.17	2.16	(-0.16%)
Operating income	80.89	90.21	(11.51%)	72.57	88.47	(21.92%)
Probability, p	0.056	$\bar{p}$ =0.85		0.49	p^{max} = 0.81	
β^*	-	0.29		-	0.50	
	Firm 83			Firm 36		
	Peer (#92)			Peer (#36)		
	Observed	Projection		Observed	Predicted	
Total assets	7.24	7.24	(0%)	67.59	67.59	(0%)
Employees	51	51	(0%)	53	53	(0%)
Fixed assets	0.66	0.66	(0%)	5.21	5.21	(0%)
Personnel expenses	2.04	2.04	(0%)	3.91	3.91	(0%)
Operating income	8.89	8.89	(0%)	54.91	54.91	(0%)
Probability, p	0.03	p^{max} = 0.03		0.99	$\bar{p}$ =0.85	
β^*	-	0		-	0	

UNIVERSITAS Miguel Hernández RESEARCH INSTITUTE

The Efficiency Assessment of the Valencian Food Industry

Ranking	Firm	Probability of being efficient	$\bar{p} = 0.75$			$\bar{p} = 0.85$			$\bar{p} = 0.95$		
			β^*	Reached Probability	Peer	β^*	Reached Probability	Peer	β^*	Reached Probability	Peer
1	2	0.9999	0	0.75	2	0	0.85	2	0	0.95	2
2	18	0.9998	0	0.75	18	0	0.85	18	0	0.95	18
3	3	0.9996	0	0.75	3	0	0.85	3	0	0.95	3
4	17	0.9983	0	0.75	17	0	0.85	17	0	0.95	17
5	20	0.9962	0	0.75	20	0	0.85	20	0	0.95	20
6	36	0.9960	0	0.75	36	0	0.85	36	0	0.95	36
7	46	0.9894	0	0.75	46	0	0.85	46	0	0.95	46
8	1	0.9868	0	0.75	1	0	0.85	1	0	0.95	1
9	56	0.9705	0	0.75	56	0	0.85	56	0	0.95	56
10	62	0.9486	0	0.75	62	0	0.85	62	0	0.9486	46
11	93	0.9441	0	0.75	93	0	0.85	93	0	0.9441	46
12	92	0.9335	0	0.75	92	0	0.85	92	0	0.9335	46
13	9	0.9288	0	0.75	9	0	0.85	9	0.8734	0.95	3
14	97	0.9176	0	0.75	97	0	0.85	97	0	0.9176	46
15	25	0.4981	0.3033	0.75	17	0.3650	0.85	17	0.4815	0.95	17
16	26	0.4909	0.4131	0.75	17	0.5025	0.8078	17	0.5025	0.8078	17
17	91	0.0735	0.1005	0.5385	93	0.1005	0.5385	93	0.1005	0.5385	56
18	22	0.0560	0.2665	0.7500	18	0.2942	0.85	18	0.3480	0.95	18
19	43	0.0549	0.1005	0.4121	46	0.1005	0.4121	46	0.1005	0.4121	46
20	85	0.0490	0.0994	0.75	93	0.1005	0.7596	93	0.1005	0.7596	46
21	95	0.0338	0.1986	0.7500	56	0.2145	0.85	56	0.2495	0.95	56
22	83	0.0335	0	0.0335	92	0	0.0335	92	0	0.0335	46
23	44	0.0183	0.0710	0.75	56	0.0824	0.85	56	0.1091	0.95	56
24	75	0.0167	0	0.0167	62	0	0.0167	62	0	0.0167	46
25	94	0.0099	0	0.0099	93	0	0.0099	93	0	0.0099	46

The Efficiency Assessment of the Valencian Food Industry

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

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

UNIVERSITAS Miguel Hernández RESEARCH INSTITUTE

Conclusions

…and future work

Conclusions and future work

- Improved Accuracy and Robustness.
- Enhanced Interpretability.
- Flexibility and Customization.
- Exploration of other machine learning techniques.
- The application of our integrated ML-DEA model to other domains.
- Development of more sophisticated counterfactual methods within the ML-DEA framework.

Thanks for your attention!

XLI Congreso Nacional de Estadística e Investigación Operativa

Proyecto PID2022-136383NB-I00 financiado por MCIN/AEI /10.13039/501100011033 /10.13039/501100011033 y por FEDER, UE

UNIVERSITAS
Miguel Hernández
RESEARCH INSTITUTE

Center of Operations Research University Institute - Miguel Hernández University of Elche
Avda. de la Universidad s/n – Edificio Torretamarit
03202 Elche (Alicante) - Spain
Phone: (+34) **96 665 85 72** - Fax: (+34) 96 665 87 15

cio@umh.es - cio.umh.es

MINISTERIO DE CIENCIA, INNOVACIÓN Y UNIVERSIDADES
Cofinanciado por la Unión Europea
AGENCIA ESTATAL DE INVESTIGACIÓN