Integrating Machine Learning and DEA: Technical Efficiency Assessment through Counterfactual Analysis and Explainability

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

V Congreso Anual Internacional de Estudiantes de Doctorado

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

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.

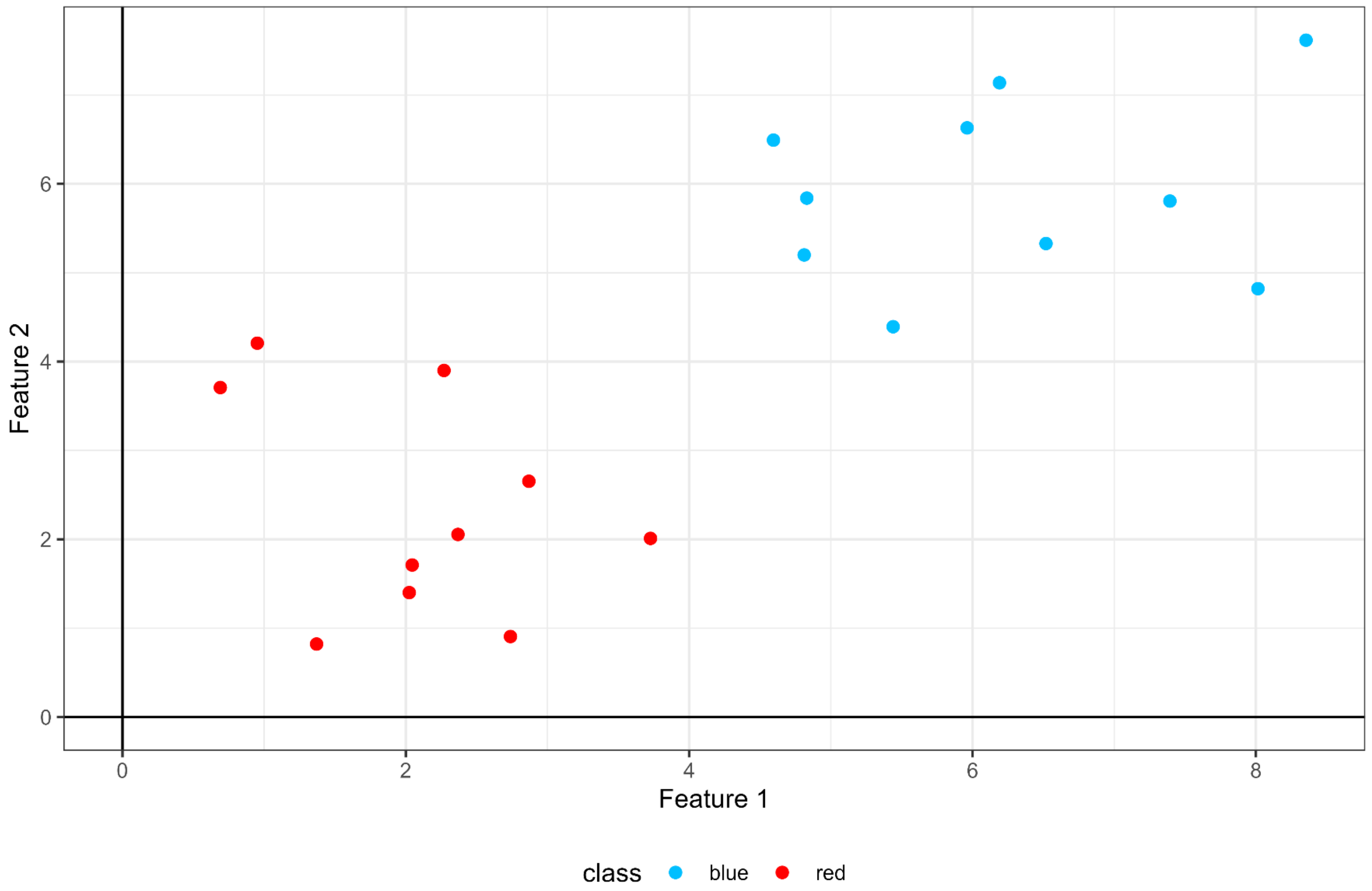
- Two predominant streams of research:

  - 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

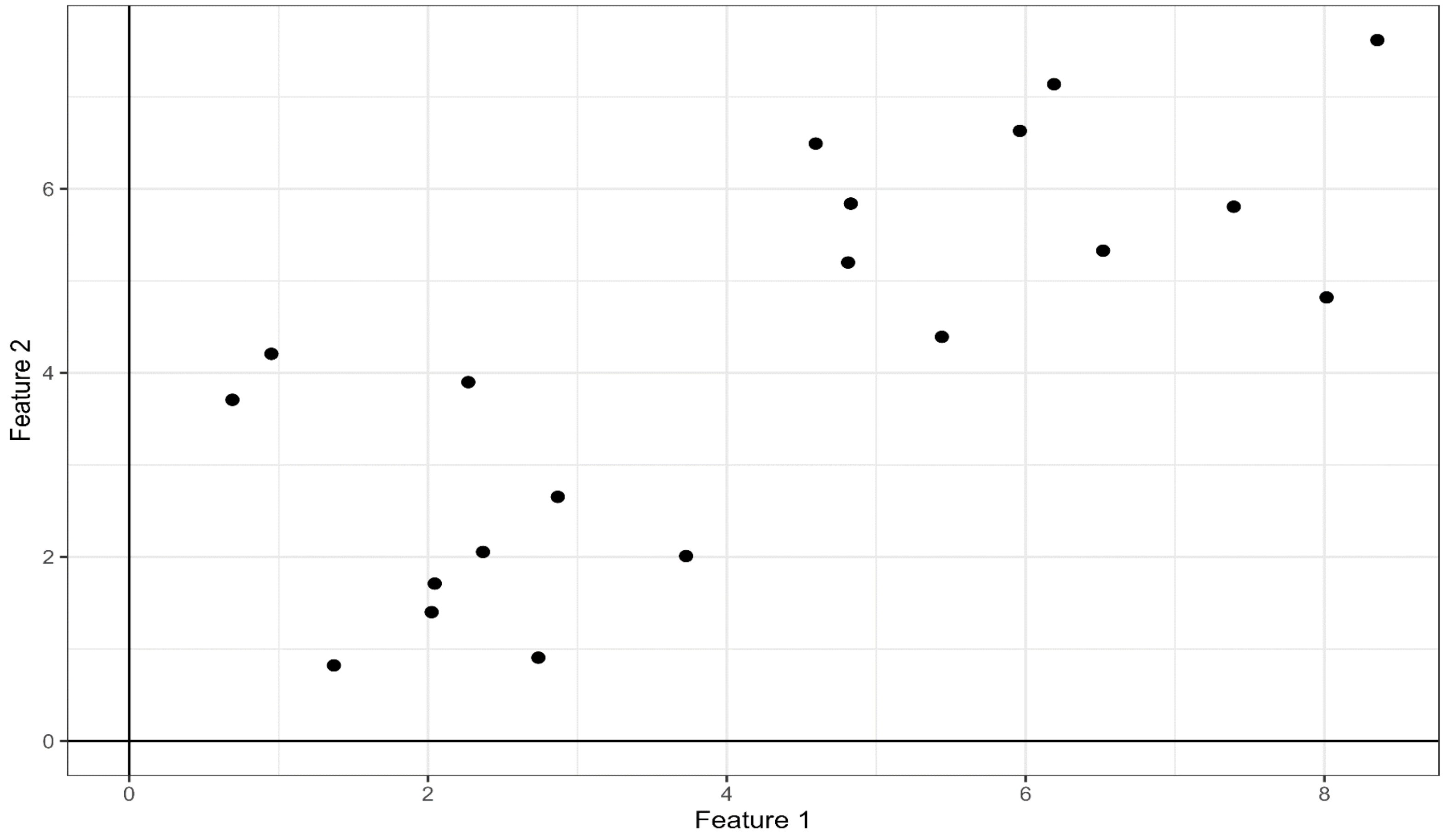
Introduction

- Types of machine learning:

Supervised Learning



Unsupervised Learning



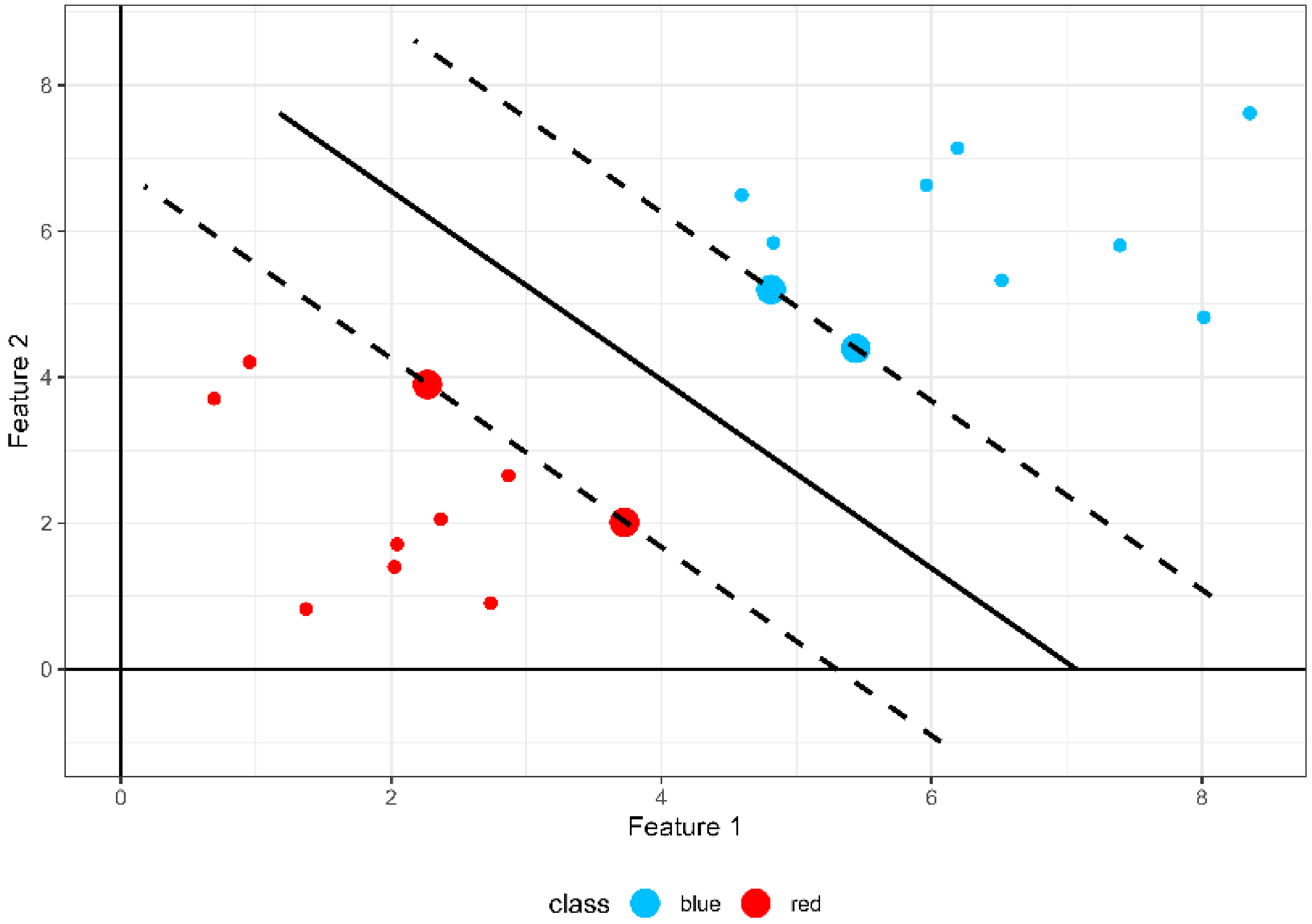
UNIVERSITAS Miguel Hernández RESEARCH INSTITUTE

Introduction

- Support Vector Machines.
  - Tries to find the best separating hyperplane.
  - Depends on:
    - selection of hyperparameters

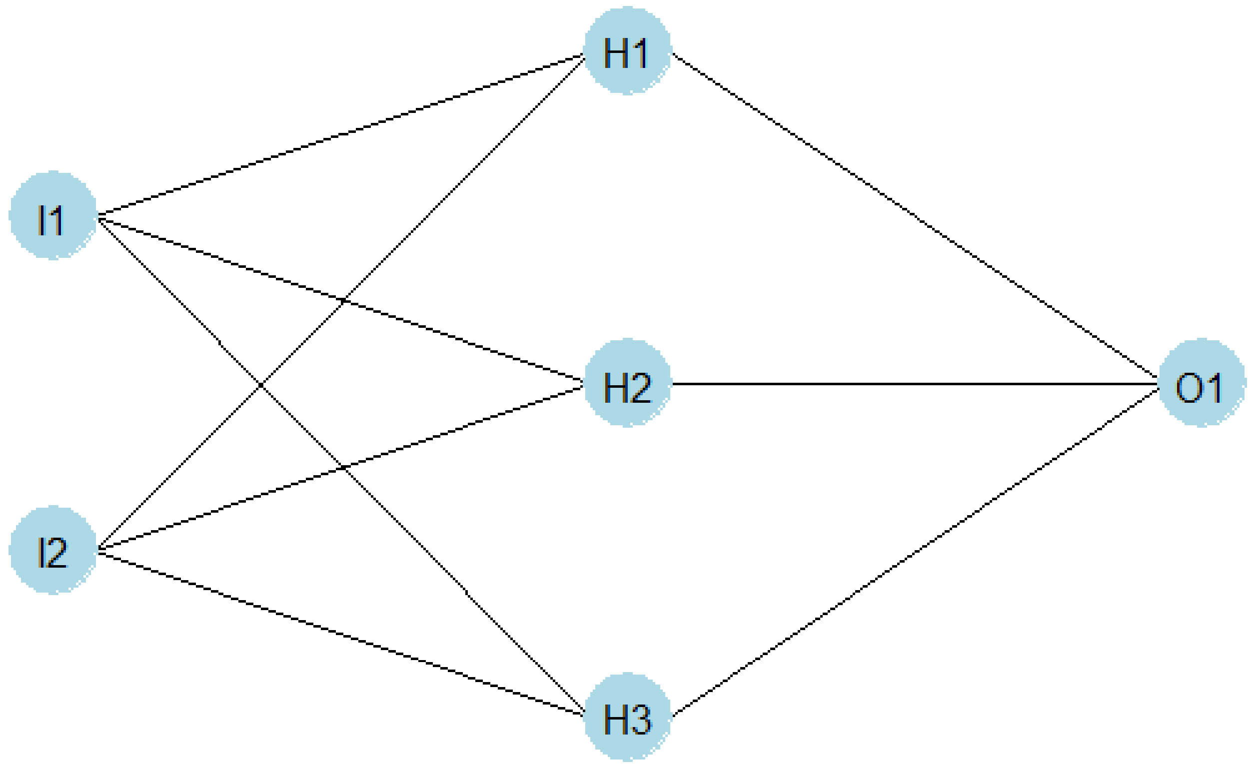
      regularization parameter (C)

      Kernel function



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

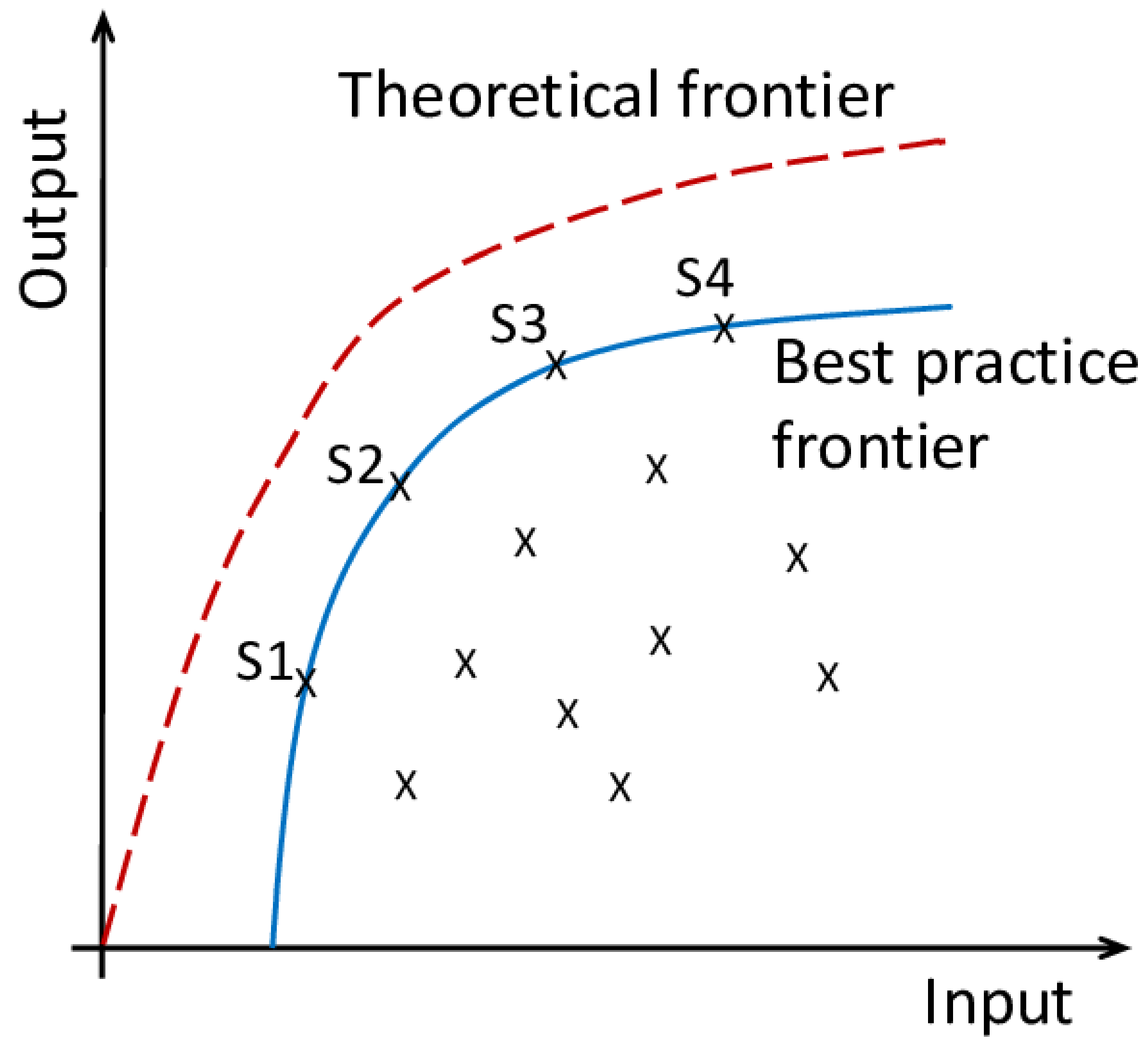
- Set of Decision Making Units (DMUs), where DMU_k consumes $\boldsymbol{x}_k = (x_k^{(1)}, \dots, x_k^{(m)}) \in R_+^m$ to produce $\boldsymbol{y}_k = (y_k^{(1)}, \dots, y_k^{(s)}) \in R_+^s$

- 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 $\boldsymbol{y} = f(\boldsymbol{x}) - \boldsymbol{u}, \boldsymbol{u} \geq \boldsymbol{0}$

Single input - output example

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

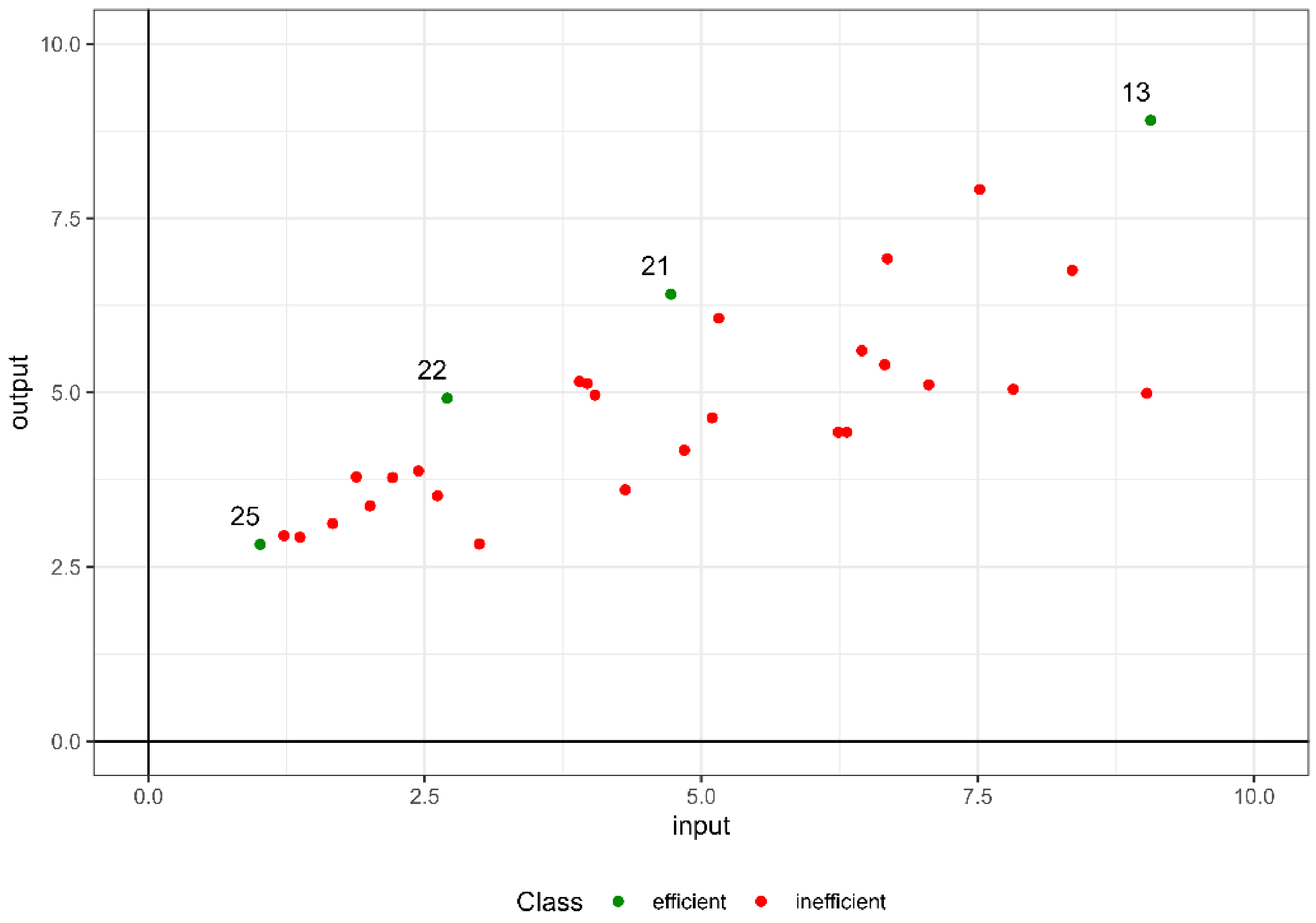


UNIVERSITAS Miguel Hernández RESEARCH INSTITUTE

Single input - output example

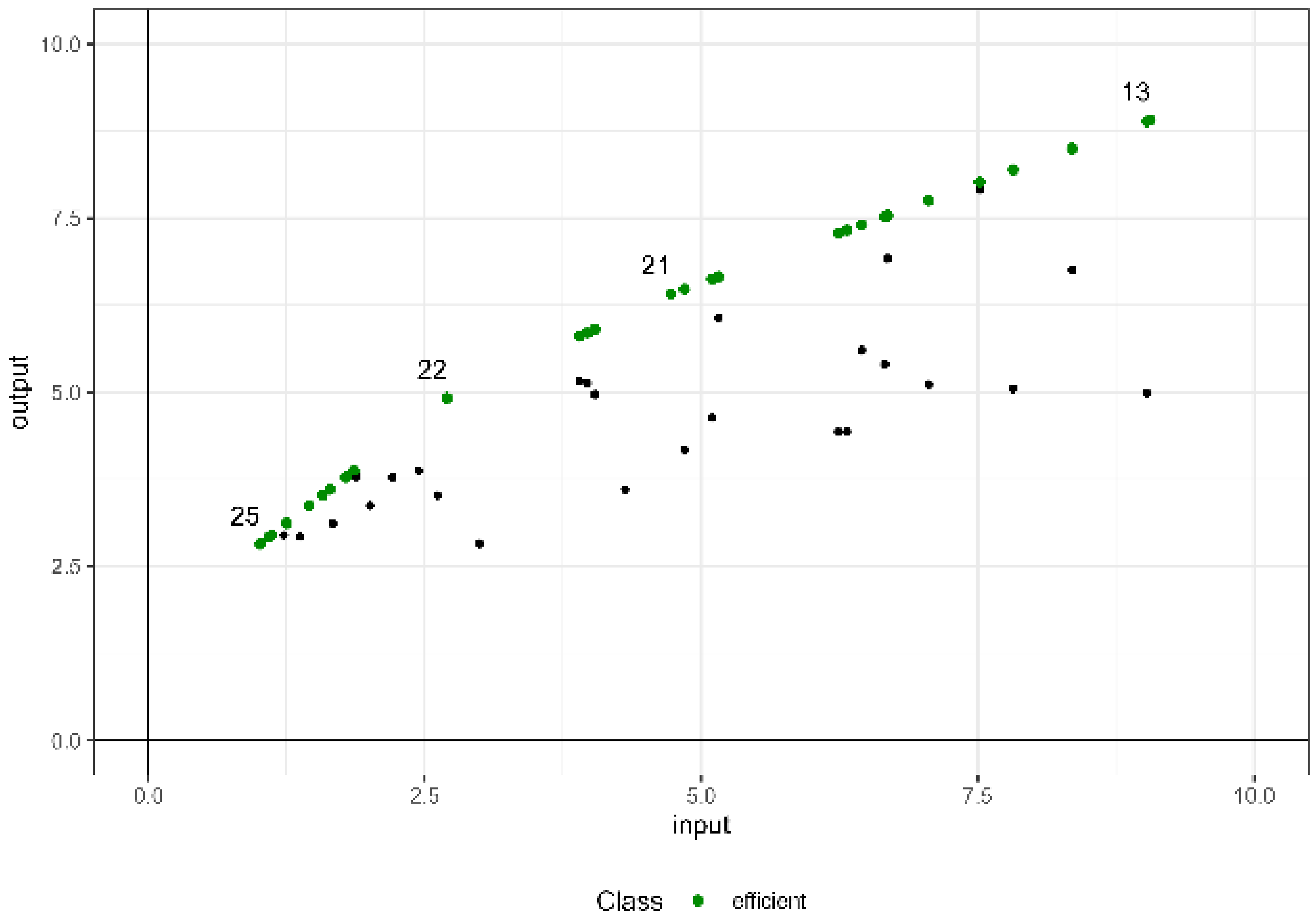
- Step 1: Utilize the additive DEA model (Charnes et al., 1985) to partition the set of DMUs in two categories.

- Efficient vs inefficient units



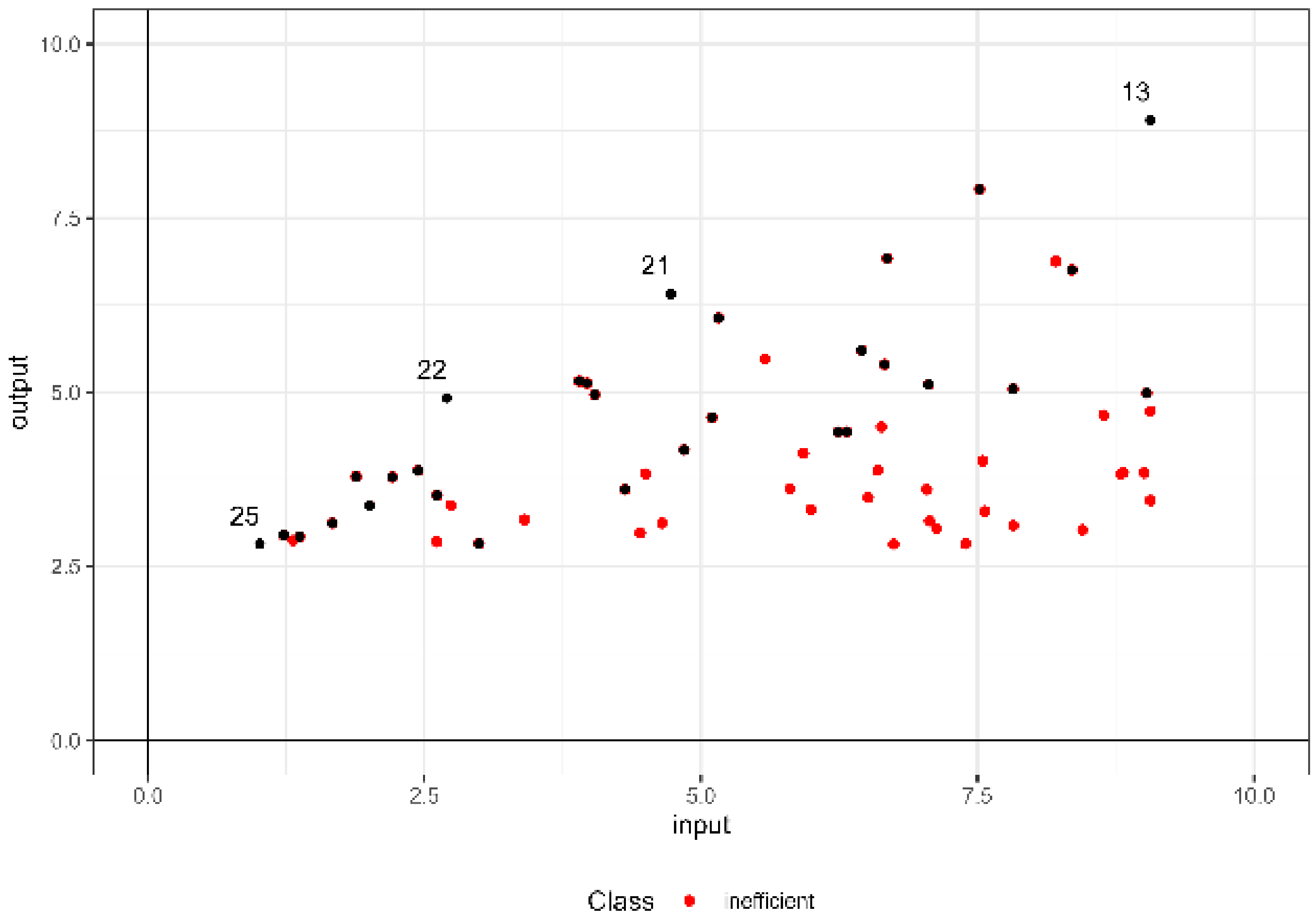
Single input - output example

- Step 2: Balancing the sample of data.
- Synthetic data generation.
- Determinate number of efficient DMUs to achieve the same proportion.



Single input - output example

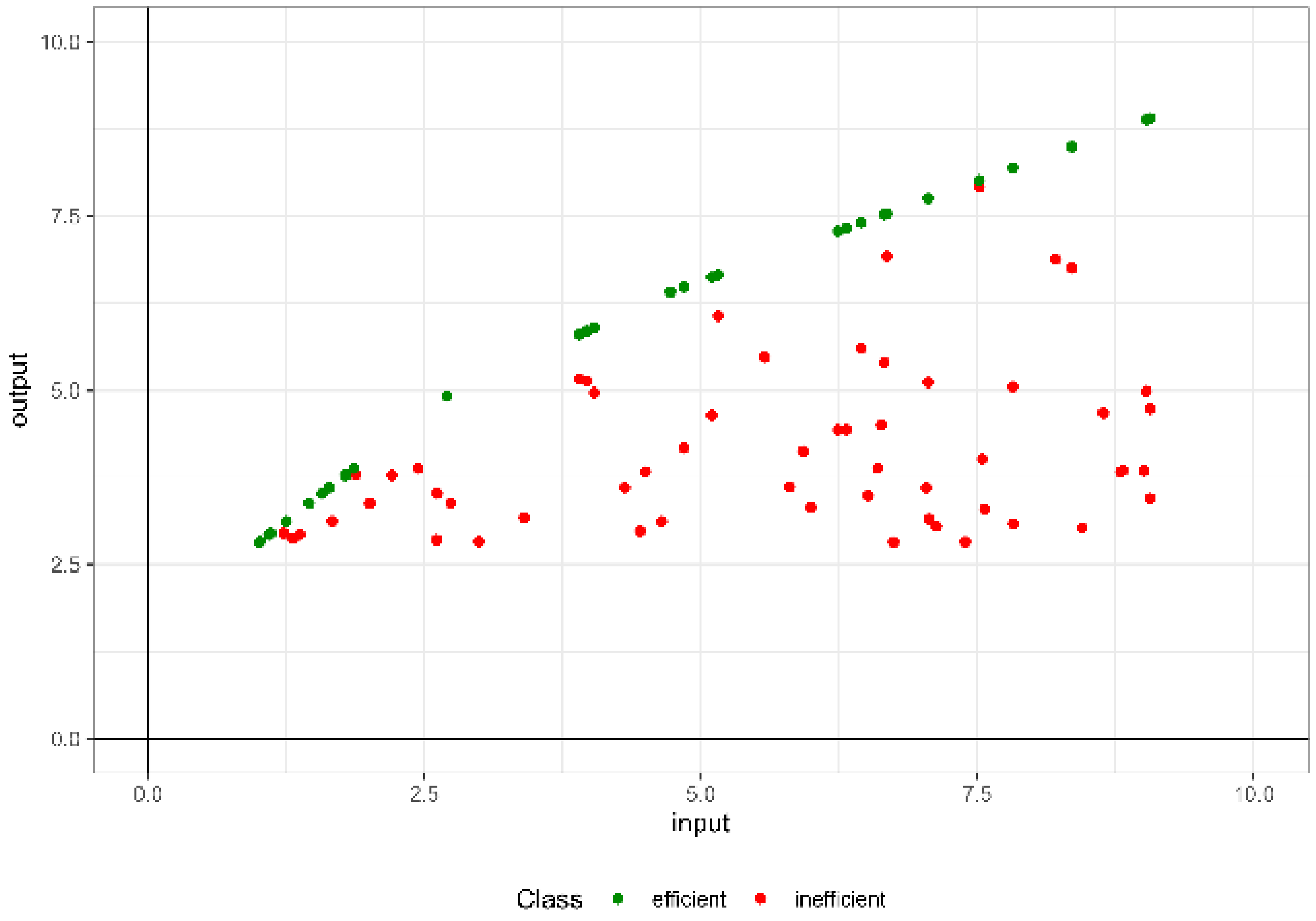
- Step 3: Balancing the sample of data.
- Synthetic data generation.
- Determinate number of efficient DMUs to achieve the same proportion.



UNIVERSITAS Miguel Hernández RESEARCH INSTITUTE

Single input - output example

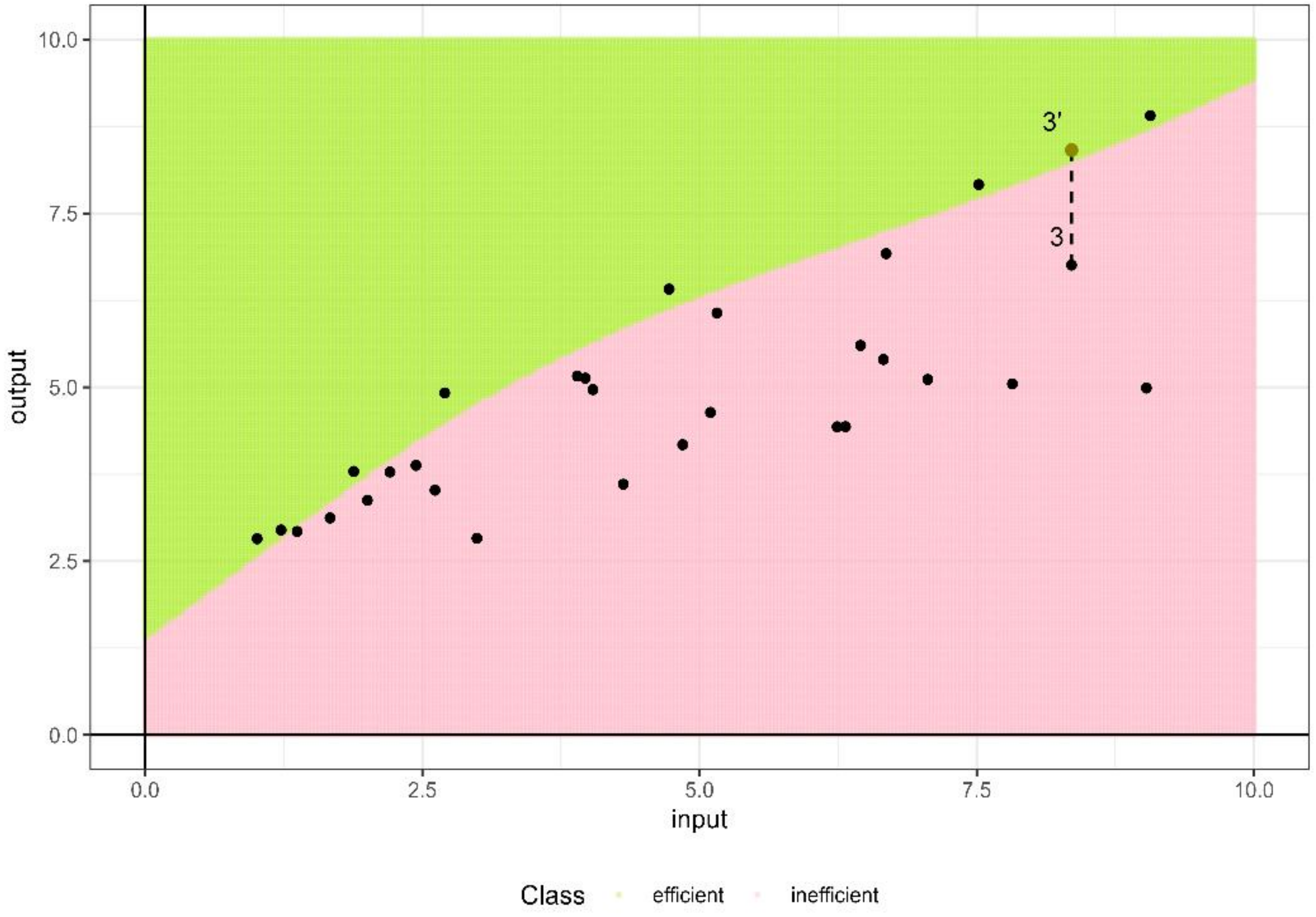
- Final dataset.
- 30 DMUs to 82 DMUs
- 26 efficient vs 56 inefficient



UNIVERSITAS Miguel Hernández RESEARCH INSTITUTE

Single input - output example

- Tuning the model. Optimal hyperparameters.
- Final regions are defined.
- To classify an observation as efficient, it is proposed that the model's label prediction be greater than 0.82.



UNIVERSITAS
Miguel Hernández
RESEARCH INSTITUTE

An empirical application

The efficiency assessment of the Spanish educational sector

The efficiency assessment of the Spanish educational sector

- A dataset obtained from the Programme for International Student Assessment (PISA).
- The dataset utilized encompasses data from the year 2018, comprising anonymized records from 999 Spanish schools randomly selected by the OECD.
- Input variables: EDUQUAL, ESCS and TSRATIO.
- Output variables: PVMATH, PVREAD and PVSCIE.
- Contextual variables: REGION and SCHLTYPE.

The efficiency assessment of the Spanish educational sector

- **For SVM polynomial kernel:**

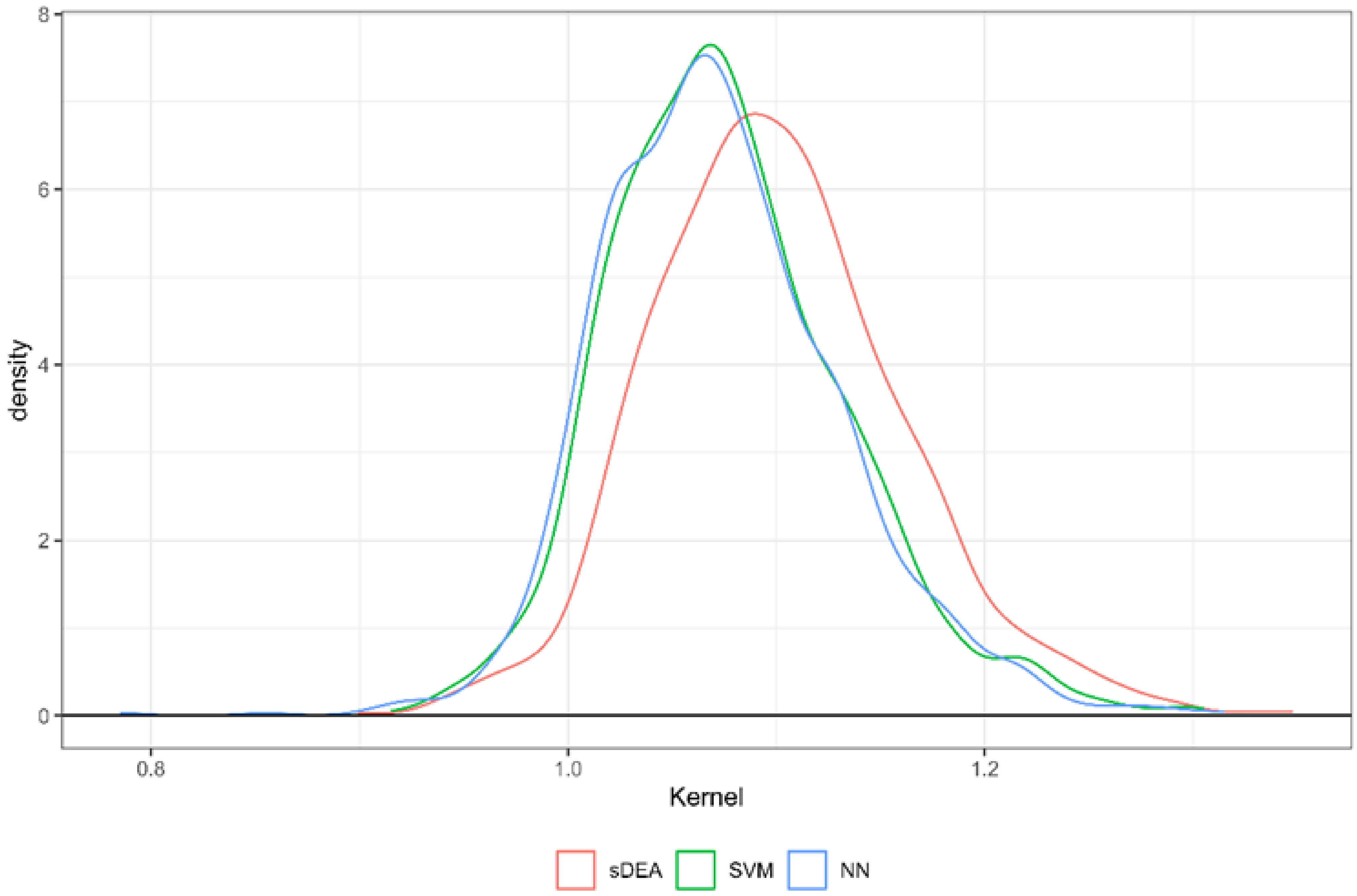
$degree$ (1, 2, 3, 4 and 5),
$data\ scaling$ (0.01, 0.1, 1, 10 and 100) and
$cost$ (0.001, 0.1, 1, 10 and 100).

- *cut off of 0.69*

- **For NN:**

$size$ (1, 5, 10 and 20) and $decay$ (0, 0.1, 0.01, 0.001, 0.0001).

- *cut off of 0.67*
- *24-5-1*



The efficiency assessment of the Spanish educational sector

- For SVM polynomial kernel:

$degree\ (1, 2, 3, 4\ and\ 5)$,
$data\ scaling\ (0.01, 0.1, 1, 10\ and\ 100)$ and
$cost\ (0.001, 0.1, 1, 10\ and\ 100)$.

- *cut off of 0.69*

- For NN:

$size\ (1, 5, 10\ and\ 20)$ and $decay\ (0, 0.1, 0.01, 0.001, 0.0001)$.

- *cut off of 0.67*
- *24-5-1*

	Min.	1st Quartil	Median	Mean	3rd Quartil	Max.
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

The efficiency assessment of the Spanish educational sector

- Sensitivity analysis reveals the following variable importance list:

- SVM model
  - ESCS (0.431)
  - PVMATH (0.193)
  - PVSCIE (0.161)
  - EDUQUAL (0.102)
  - TSRATIO (0.04)
  - SCHLTYPE (0.03)
  - PVREAD (0.029)
  - REGION (0.015)

- NN model
  - ESCS (0.418)
  - PVMATH (0.32)
  - PVSCIE (0.09)
  - SCHLTYPE (0.066)
  - EDUQUAL (0.057)
  - REGION (0.027)
  - PVREAD (0.007)

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!

V Congreso Anual Internacional de Estudiantes de Doctorado

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