UNIVERSIDAD
NACIONAL
DE COLOMBIA

Evaluación de eficiencia en Instituciones Prestadoras de Salud en Colombia: un estudio con Análisis Envolvente de Datos

Efficiency evaluation of Health Promotion Entities (EPS) in Colombia, a case study using data envelopment analysis

Andrés Mauricio Gómez Ardila

Nacional de Colombia
Facultad de Ciencias, Escuela de Estadística
Medellín, Colombia
2021

Evaluación de eficiencia en Instituciones Prestadoras de Salud en Colombia: un estudio con Análisis Envolvente de Datos

Efficiency evaluation of Health Promotion Entities (EPS) in Colombia, a case study using data envelopment analysis

Andrés Mauricio Gómez Ardila

Trabajo Final presentado como requisito para optar al título de:
Magister en Ciencias-Estadística

Director:

René Iral Palomino, Ph.D.
Profesor Asociado Escuela de Estadística
Universidad Nacional de Colombia (Colombia)

Co-director:
Juan G. Villegas, Ph.D.
Profesor Titular Departamento de Ingeniería Industrial
Universidad de Antioquia (Colombia)

Áreas de Investigación:
Estadística Aplicada e Investigación de Operaciones

Universidad Nacional de Colombia
Facultad de Ciencias, Escuela de Estadística
Medellín, Colombia
2021

Agradecimientos

A Dios, por regalarme la salud, la fuerza, sabiduría e inteligencia para poder realizar este trabajo.

A mi asesor René Iral, por haber confiado en mí.

A mis co-asesores Juan Villegas y Andrés Villegas, por el constante apoyo y orientación a lo largo de estos años.

A mi familia, Maria Ardila, Ivan Gómez y Lizeth Gómez por la comprensión y apoyo en todos mis proyectos.

A Ana Gil, por ayudarme, escucharme y alentarme.

Publicaciones:

- Gómez Ardila, A., Villegas Ramirez, A., & Villegas Ramirez, J. (2021). "Método heurístico de selección de variables para el incremento del poder discriminatorio de DEA: caso de aplicación a las EPS Colombianas". *XXX Simposio Internacional de Estadística 2021 - Evento virtual*, 183-190. Disponible en: http://simposioestadistica.unal.edu.co/fileadmin/content/eventos/simposioestadistica/Memorias_2021/Historico_de_memorias_XXX_SIE_2021_Evento_Virtual.pdf

Resumen

El análisis envolvente de datos (DEA) es una técnica no paramétrica para medir la eficiencia relativa de un conjunto de unidades productoras homogéneas (denominadas DMU). Sin embargo, DEA no proporciona pautas claras para la selección de variables de entrada y salida, lo cual es crucial para la precisión y relevancia de los resultados, ya que cuanto mayor sea el número de variables de entradas y salidas en DEA, mayor será la dimensionalidad del espacio de solución y las estimaciones de la eficiencia serán imprecisas. Por consiguiente, mayor será la probabilidad de que algunas unidades ineficientes sean clasificadas como eficientes. En este trabajo presentamos una propuesta heurística para la selección de variables, tanto de entrada como de salida, para mejorar el poder discriminatorio de los modelos de DEA, basada en la eliminación iterativa de variables. La metodología propuesta elige en cada iteración la configuración de variables donde una métrica de impureza (índice Gini o entropía) es máxima. Esta propuesta pretende evitar los criterios ad hoc o juicios particulares en la elección de las variables, proporcionando un conjunto de modelos seleccionados por índices estadísticos de dispersión. Usando datos financieros (Fontalvo Herrera, 2017) para medir la eficiencia de EPS colombianas y datos de estudios previos para este problema bajo otros enfoques (Wong y Beasley, 1990) (Adler y Golany, 2002), se ilustra la metodología propuesta, con la cual se encuentra una reducción en el número de DMUs clasificadas como eficientes, dada una reducción significativa de las variables y garantizando una alta retención de la varianza total de los datos.

Palabras claves: *Análisis envolvente de datos, Selección de variables, Problema de dimensionalidad, Medición del desempeño*

Abstract

Data Envelopment Analysis (DEA) is a non-parametric methodology based on mathematical programming to estimate the performance of a set of Decision Making Unit (DMUs) which use the same set of inputs to produce the same set of outputs. However, DEA does not provide clear guidelines for input and output variable selection, which might prove problematic in the presence of a large number of input and output variables. In such a situation, the solution space is high dimensional resulting in a significant probability that inefficient DMUs are deemed efficient and hindering the discriminatory power of the methodology. In this work, we present a heuristic approach to improve the discriminatory power of DEA models, based on the iterative elimination of variables. The proposed methodology chooses in each iteration the configuration of variables where an impurity metric is maximised. This proposal aims to avoid ad hoc criteria or subjective judgments in the choice of variables, by providing instead a set of models selected using statistical indices of dispersion. We illustrate the proposed methodology using financial data (Fontalvo Herrera, 2017) to measure the efficiency of Colombian Health Promoting Entities (EPS) and datasets from previous studies (Wong y Beasley, 1990) (Adler y Golany, 2002). We find that the number of efficient DMUs is significantly reduced in our method, and this is achieved by selecting fewer variables while the retained variance of the total data is high.

Keywords: *Data envelopment analysis, Variable selection, Curse of dimensionality, Performance measurement*

Índice general

Índice de figuras

Índice de tablas

Capítulo 1

INTRODUCCIÓN

1.1. Medición de la eficiencia, una necesidad

La gestión empresarial está avanzando hacia una evaluación más objetiva del desempeño para soportar la toma de mejores decisiones. Particularmente, el sector salud enfrenta nuevos desafíos todos los días. Las nuevas regulaciones, la extensa gama de tecnologías biomédicas, la creación de nuevas organizaciones y modelos como resultado de políticas públicas , el constante envejecimiento de la población, la reducción de la mortalidad infantil y de enfermedades transmisibles, el surgimiento de nuevas enfermedades y el resurgimiento de otras que se encontraban bajo control, el mayor peso que adquieren las enfermedades crónicas y degenerativas en la condición de salud , entre otros factores, han provocado un aumento considerable en los gastos médicos y la necesidad de valorar el rendimiento de las unidades prestadoras de dichos servicios (Chuang, Chang, y Lin, 2011).

Los administradores de servicios en salud deben responder a estos desafíos con evaluaciones de desempeño que den soporte a planes de trabajo sólidos. Y con ello, una de las tareas más difíciles, es precisar una combinación adecuada de insumos (entradas) y resultados (salidas) inmersos en la prestación de servicios de salud. Desafortunadamente, los puntos de referencia establecidos mediante aproximaciones analíticas antiguas basadas en relaciones múltiples, crearon más dilemas que soluciones (Ozcan, 2014, chap. 1). Paralelamente, una dificultad permanente en la medición de la eficiencia hospitalaria, es la comparación de unidades heterogéneas (O'neill, Liam, 1998), dado que entre hospitales se pueden presentar diferencias, ya sea por su nivel de atención (primero, segundo o tercer nivel, de acuerdo con la presencia de médicos generales, especialistas básicos y de mayor complejidad), por su dedicación u orientación (general, geriátrico, materno-infantil, psiquiátrico, universitario, etc), e incluso por el grado de complejidad en la atención (alta, mediana o baja, de acuerdo a su capacidad resolutiva).

Llegados a este punto, se hace cada vez más relevante una herramienta que transforme

los juicios implícitos en explícitos, y así, los encargados de tomar las decisiones logren anticipar y valorar las posibles consecuencias de las diferentes rutas de trabajo, con criterios sólidos y transparentes respecto al uso de los recursos (Mejía, 2008). Toda organización está interesada en determinar su desempeño con respecto al uso de los recursos, tales como: el talento humano, la materia prima, los equipos, entre otros. Para ello deben compararse con otras unidades organizacionales "pares", con el fin de cuantificar si están siendo eficientes o ineficientes, para finalmente adoptar las mejores prácticas de los líderes o referentes. A este ejercicio se le conoce como benchmarking. La evaluación del desempeño y la evaluación comparativa ayudan a que las operaciones o procesos sean más productivos y eficientes (Zhu, 2014), porque permiten la identificación de buenas prácticas donde, eliminando o perfeccionando las operaciones ineficientes, se reducen costos y se aumenta la productividad.

1.2. Análisis Envolvente de Datos

Durante las últimas décadas han surgido métodos tanto paramétricos como no paramétricos para medir y analizar el desempeño (Ozcan, 2014, chap. 1). Entre los más populares, se encuentra la metodología no paramétrica Análisis Envolvente de Datos (DEA). Esta herramienta ha sido aplicada en diversos sectores tales como: los financieros (banca, seguros y bolsa), servicios de salud, educación, energía, transporte, telecomunicaciones, servicios públicos, entre otros (Liu, Lu, Lu, y Lin, 2013).

DEA intenta identificar y cuantificar las fuentes o factores de ineficiencias contenidas en los insumos o productos generados por las entidades sometidas a evaluación. A dichas entidades se les conoce como "unidades de toma de decisiones" (DMU, Decision Making Unit) (Kohl, Schoenfelder, Fügener, y Brunner, 2018). Bajo esta metodología no se requiere una formulación explícita de las relaciones funcionales subyacentes con formas paramétricas específicas que relacionan las entradas con las salidas del proceso (prestación del servicio), y permite la generación de una medida general (escalar) de eficiencia para cada DMU a partir de los valores observados de sus entradas y salidas, sin requerir el uso de ponderaciones a priori (Charnes, Cooper, y Thrall, 1986).

DEA se ha implementado ampliamente para analizar muchas industrias, utilizado como una herramienta basada en datos para construir un índice compuesto con el fin realizar evaluaciones comparativas justas (Chen, Tsionas, y Zelenyuk, 2021), y popularizado por la incorporación de múltiples entradas y múltiples salidas para medir la eficiencia de unidades productoras. Desafortunadamente, la metodología DEA presenta retos asociados a la ausencia de inferencia estadística formal sobre las medidas de eficiencia (Tsionas y Papadakis, 2010), así como la falta de lineamientos objetivos para la selección de variables, por lo que varios autores obtienen resultados diferentes ante la sensibilidad en el cálculo de eficiencias, para cada selección particular de variables (Fernandez-Palacin, Lopez-Sanchez, y Muñoz-Márquez, 2018).

1.3. Definición del problema

La evaluación del desempeño es una actividad obligatoria en la identificación de deficiencias en los procesos o gestión empresarial, para posteriormente diseñar metas y estrategias de mejoramiento. DEA ha demostrado ser una técnica de utilidad en diversos sectores económicos desde hace varios años para evaluar el desempeño. Sin embargo, la selección de los insumos y los productos ha sido una preocupación constante, ya que a menudo surgen problemas de discriminación entre unidades, y en consecuencia, la evaluación del desempeño puede verse afectada (Charles, Aparicio, y Zhu, 2019).

Aunque se han establecido unas reglas empíricas (Dyson et al., 2001) (Cook, Tone, y Zhu, 2014) para orientar la cantidad de DMUs a ser evaluadas, tales como que el total de DMUs sea como mínimo la multiplicación entre el número de entradas y de salidas, o también que el número de DMUs sea mayor o igual al triple de la cantidad total de variables a usar. Sin embargo, DEA no proporciona pautas objetivas para la selección de variables, lo cual es vital para la precisión y relevancia de los resultados, ya que cuanto mayor sea el número de variables en DEA, mayor será la dimensionalidad del espacio de solución y menos exigente será la estimación de la frontera eficiente. Esta menor exigencia aumenta la probabilidad de que algunas unidades ineficientes sean clasificadas como eficientes. En la literatura, la falta de discriminación a menudo se conoce como la "maldición de la dimensionalidad" (Charles et al., 2019).

Tabla 1.1: Datos - Maldición de dimensionalidad en DEA

	Input 1	Output 1	Output 2	Output 3
EPS 1	1	1	2	3
EPS 2	1	1	3	2
EPS 3	3	3	1	2
EPS 4	3	3	2	1
EPS 5	1	1	1	1

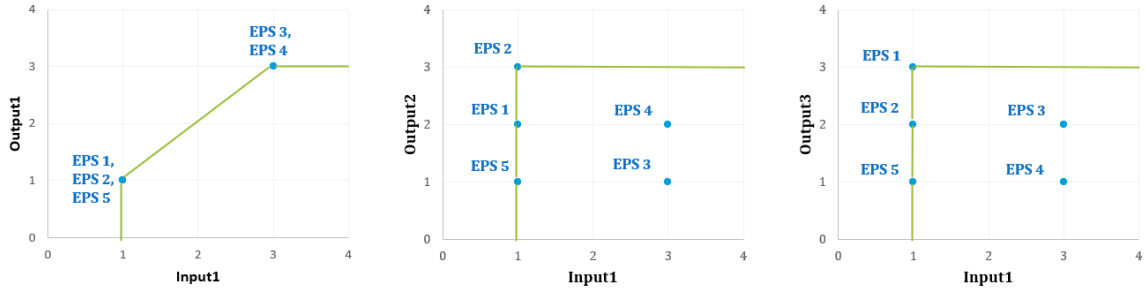

Figura 1.1: Gráficas - Maldición de dimensionalidad en DEA

Ahora bien, para evidenciar la maldición de la dimensionalidad en la Figura 1.1 se grafican los datos (véase la Tabla 1.1) de la variable de entrada con cada una

de las variables de salidas, para observar el aprovechamiento de los insumos en la transformación de productos. Al comparar el *Input1* versus *Output1*, se puede ver en este diagrama de dispersión como todas las unidades evaluadas se ubican en la frontera y ninguna domina a otra, por lo que todas serán clasificadas como eficientes. Sin embargo, en las otras dos gráficas, se puede inferir rápidamente que las EPS4 y EPS3 obtienen los resultados de producción inferiores, dando lugar a las ineficiencias, ya que estas DMUs requieren la misma o una mayor cantidad de insumos (entradas) pero logran una menor cantidad de productos (salidas) con respecto a su pares. Esto resalta la importancia de la selección de las variables en DEA, para eliminar aquellas que generen ruido (note que *Input1* y *Output1* tiene una correlación positiva perfecta) y dejar las que permitan identificar las ineficiencias.

Lo anterior deja en evidencia que cuantas más variables se incluyan en el modelo DEA, mayor es la posibilidad de que algunas unidades ineficientes dominen alguna de las dimensiones y sean clasificadas como eficientes (Charles et al., 2019). Por consiguiente, cobra relevancia la selección de las variables como alternativa para eludir la maldición de la dimensionalidad, una limitación bien conocida de DEA y de prácticamente cualquier estimador no paramétrico, dada la dependencia existente de su precisión ante la dimensión del problema (Chen et al., 2021).

La pregunta que surge en estas circunstancias es: ¿Cómo se puede incrementar el poder discriminatorio de la técnica DEA? En este trabajo se presenta una aproximación heurística en la selección de variables, tanto de entrada como de salida, para mejorar el poder discriminatorio de los modelos de DEA, basada en la eliminación iterativa de variables. La metodología desarrollada elige en cada iteración la configuración de variables donde una métrica de impureza (índice Gini o entropía) es máxima. Esta propuesta pretende evitar los criterios ad hoc o juicios particulares en la elección de las variables, proporcionando un conjunto de modelos seleccionados por índices estadísticos de dispersión. Simultáneamente, se calcula la varianza retenida de los datos para cada configuración de variables, como guía para los tomadores de decisiones en la elección del modelo final.

Estructura del documento

Para abordar el problema de investigación descrito anteriormente, en el Capítulo 2, se presenta la revisión de la literatura, en la cual se aborda la conceptualización de la medición de eficiencia, los fundamentos teóricos de DEA, el problema de dimensionalidad presente en DEA y las aproximaciones tanto clásicas como estadísticas para solucionarlo. Luego, el Capítulo 3 expone el algoritmo y un ejemplo esquemático de la aproximación heurística en la selección de variables para mejorar el poder discriminatorio de los modelos de DEA, basada en la eliminación iterativa de variables y selección de modelos sustentado en la maximización de la métrica de impureza. Posteriormente, el Capítulo 4 presenta los resultados de la metodología propuesta para datos reales del sector salud colombiano y un conjunto de datos de la literatura,

además de la comparación del procedimiento propuesto con los enfoques asociados a: Análisis de Componentes Principales (PCA) y Regresión Lasso. Finalmente, en el Capítulo 5 se consolidan las conclusiones del trabajo.

Capítulo 2

REVISIÓN DE LA LITERATURA

Este capítulo tiene como objetivo hacer un recuento de algunos de los estudios enfocados al perfeccionamiento del rendimiento del análisis de la eficiencia por medio de la metodología DEA, así como a la orientación en la selección de variables, con el propósito de dar mayor contexto y soporte al problema que se plantea en este trabajo. El capítulo comienza con la revisión de la literatura acerca de los fundamentos de DEA, posteriormente se profundiza en el problema de dimensionalidad presente en esta herramienta, y finalmente se exponen los enfoques con los que se ha intentado dar solución, aproximaciones clásicas desde la investigación de operaciones y estadísticas.

2.1. Conceptualización acerca de la medición de la eficiencia

Conceptos de eficiencia, productividad y competitividad

La creciente competitividad sumada a la globalidad de los diversos sectores económicos, han generado un entorno de alta incertidumbre para las empresas en el cual es cada vez más difícil la supervivencia empresarial, de ahí que, la utilización de métodos fiables para evaluación de la eficiencia sea indispensable y cada vez más común. Su importancia radica ya sea en la detección de cambios de un periodo de tiempo a otro, o el simple entendimiento de cómo funcionan las organizaciones en relación con otras del mismo sector productivo (evaluación comparativa o revisión por pares) (Ozcan, 2014, chap. 1).

Con frecuencia en la literatura, y en el lenguaje cotidiano, se utilizan como sinónimos los conceptos eficiencia, productividad y competitividad, porque en el trasfondo todas ellas reflejan que las empresas van por buen camino. De ahí que si las empresas maximizan su beneficio [1], serán consideradas como eficientes. No obstante no todas las

[1] La maximización de beneficios para empresas sin ánimo de lucro, puede interpretarse como la

organizaciones logran dicho objetivo común, dando lugar a situaciones de ineficiencia. Así, la maximización del beneficio requiere que las empresas: *(a)* elijan un nivel máximo de productos [2] con el cual obtienen mayores ganancias, *(b)* seleccionen la combinación de insumos que minimicen los costos pero que garanticen la producción esperada para satisfacer la demanda, y *(c)* empleen la cantidad mínima de insumos, es decir, no generen desperdicios.

Álvarez (2001, chap. 1), define tres tipos de eficiencia:

- **Eficiencia de escala:** se da cuando una empresa maximiza su beneficio produciendo en una escala de tamaño óptima.
- **Eficiencia técnica:** se da cuando una empresa maximiza sus productos (*outputs*) combinando adecuadamente sus recursos (*inputs*).
- **Eficiencia asignativa:** se da cuando una empresa minimiza su gasto de producción combinando adecuadamente sus recursos (*inputs*).

Respecto a la definición de productividad, normalmente se entiende como el cociente entre la cantidad de salidas de un proceso sobre la cantidad de entradas en el mismo. Mientras que por competitividad se hace referencia a la capacidad de sobrevivir en el mercado, comúnmente asociado a la existencia de una ventaja competitiva. En resumen, la creencia de que mejorar en cualquier de estos aspectos es bueno para las empresas, lleva a que se usen indistintamente. Pero no siempre un incremento en la productividad estará ligado a una mejoría en la eficiencia.

Medición de la eficiencia

La medición de la eficiencia tiene como base la comparación del comportamiento de una unidad, sea un individuo, un programa, una organización, etc., con respecto a un óptimo. Sin embargo, ante la ausencia de la información exacta y completa de los mercados, procesos, instrumentos, equipos, el entorno y demás, una buena alternativa es referenciarse con otras unidades similares. Esta idea se le atribuye a Farell en 1957, quien determinó empíricamente un estándar de referencia o frontera con la cual comparar y establecer quienes son o no eficientes (Álvarez, 2001, chap. 1).

El concepto de frontera permite la convivencia entre el análisis empírico de la producción y la teoría económica, dado que las funciones de producción, costo y beneficio son funciones de frontera (Álvarez, 2001, chap. 1), y con ello, la posibilidad de interpretar las desviaciones de las unidades versus la frontera como indicadores de ineficiencia.

adecuada utilización de los recursos disponibles, u otras métricas más allá de la rentabilidad como la maximización de la cobertura.

[2] Entendiendo por *"producto"* la fabricación de bienes materiales o la prestación de servicios, es decir productos tangibles o intangibles.

Tipos de frontera

Álvarez (2001, chap. 1) habla de dos tipos de frontera:

- **Frontera determinística**: Se caracteriza por atribuir toda la desviación de la frontera a la ineficiencia, es decir, no contempla situaciones exógenas sobre las unidades valoradas. Una función de producción de frontera determinística puede representarse como $y = f(x) - u$ donde y es el output, x el vector de inputs, f es una función de producción que representa la tecnología, y u es una alteración mayor o igual a cero que simboliza la distancia de la unidad particular a la frontera.
- **Frontera estocástica**: Esta admite la naturaleza estocástica de la producción, que puede escribirse así: $y = f(x) + e$, con $e = v - u$, donde y es el output, x el vector de inputs, f es una función de producción que representa la tecnología, y v es la perturbación aleatoria asociada a situaciones exógenas. El término v se supone idéntico e independientemente distribuido con media 0, y el término u es no negativo e independiente a v, los cuales conforman el término de ineficiencia con choces aleatorios (e).

2.2. Fundamentos de DEA

DEA es una técnica no paramétrica que emplea programación matemática desarrollada por Charnes, Cooper, y Rhodes (1978) para medir la eficiencia relativa de un conjunto de n DMUs que utilizan el mismo conjunto de m entradas para producir el mismo conjunto de salidas s. DEA calcula la eficiencia de cada DMU con base a dos conjuntos de pesos, uno para las entradas y otro para las salidas, elegidos de tal manera que cada DMU alcance la máxima eficiencia factible.

En DEA las n DMU objeto de evaluación, requieren diferentes cantidades de los m insumos para producir s resultados. Entonces la DMU k requiere la cantidad x_{ik} de insumo i y produce la cantidad y_{rk} de la salida r.

Para medir el desempeño de la DMU_k (con $k = 1, 2, ..., n$) se resuelve un problema de optimización, el cual busca maximizar la eficiencia (cociente entre las salidas y las entradas). Como se tienen múltiples entradas y múltiples salidas se construye una salida y una entrada virtual usando ponderaciones u_r y v_i para cada resultado y cada insumo, respectivamente (Restrepo y Villegas, 2014). Todo esto condicionado a que, en la medición del desempeño, ninguna DMU puede presentar una eficiencia mayor al 100 %. Se tiene entonces, el siguiente problema de optimización:

$$
\begin{aligned}
&\underset{u_r, v_i}{\text{Maximizar}} \quad h_k = \left(\sum_{r=1}^{s} u_r \; y_{rk}\right) \left(\sum_{i=1}^{m} v_i \; x_{ik}\right)^{-1} \\
&\text{Sujeto a:} \quad \left(\sum_{r=1}^{s} u_r \; y_{rj}\right) \left(\sum_{i=1}^{m} v_i \; x_{ij}\right)^{-1} \leq 1 \qquad \forall \; j = 1, ..., n \\
&\qquad\qquad u_r \geq 0 \qquad \forall \; r = 1, ..., s \\
&\qquad\qquad v_i \geq 0 \qquad \forall \; i = 1, ..., m
\end{aligned} \tag{2.1}
$$

Este modelo es conocido como CCR, es el primer modelo propuesto en DEA (considera rendimientos constantes a escala) y es conocido como CCR por la inicial de los autores Charnes, Cooper y Rhodes quienes lo desarrollaron (Charnes et al., 1978), donde h_k es la función objetivo (medida de la eficiencia de la k-ésima DMU), y_{rj} es la salida r-ésima de la unidad j-ésima, x_{ij} es el insumo i-ésimo de la unidad j-ésima, y u_r y v_i son las ponderaciones de los insumos y salidas respectivamente (soluciones del programa).

El programa fraccional CCR, puede linealizarse transformándose en un programa lineal, resultando el siguiente modelo:

$$
\begin{aligned}
&\underset{h_k, u_r}{\text{Maximizar}} \quad h_k = \sum_{r=1}^{s} u_r \; y_{rk} \\
&\text{Sujeto a:} \quad \sum_{i=1}^{m} v_i \; x_{ik} = 1 \\
&\qquad\qquad \sum_{r=1}^{s} u_r \; y_{rj} \; - \; \sum_{i=1}^{m} v_i \; x_{ij} \leq 0 \qquad \forall \; j = 1, ..., n \\
&\qquad\qquad u_r \; , \; v_i \geq 0 \quad \forall \; r = 1, ..., s \quad \forall \; i = 1, ..., m
\end{aligned} \tag{2.2}
$$

Vale la pena aclarar que existen otras formulaciones que varían según la orientación del modelo, que puede ser con enfoque hacia las salidas o hacia las entradas, también según el tipo de tecnología que se está evaluando (como los rendimientos constantes o variables a escala), y otras formulaciones, por ejemplo, basadas en holguras (conocidos como aditivos o no orientados).

Habitualmente en los estudios de eficiencia, primero se estima el índice de eficiencia, y después, en una "*segunda etapa*" donde surgen preguntas como ¿por qué hay unidades más eficientes que otras? Kohl et al. (2018) resumen el proceso de estimación de eficiencia con métodos DEA en varios momentos (véase la Figura 2.1). En su fase inicial se realiza la *(a) Selección de los datos*, seguida de la *(b) Especificación del modelo*, para finalmente implementar *(c) Técnicas subsecuentes*, generalmente métodos estadísticos, que ayuden a validar su puntuación contra otros índices, la validación del modelo en sí mismo, o el ajuste de las estimaciones. Entre algunas técnicas subsecuentes (o análisis de segunda etapa) se encuentran el análisis de conglomerados o cluster, el cual

pretende segmentar las unidades de análisis y realizar evaluaciones sobre unidades más homogéneas, y así garantizar las mismas condiciones de arranque; también mediante el análisis de regresión donde se pueden identificar los factores que más contribuyen a la eficiencia y con esto, proponer alternativas de mejoramiento de aquellas unidades que no son tan eficientes.

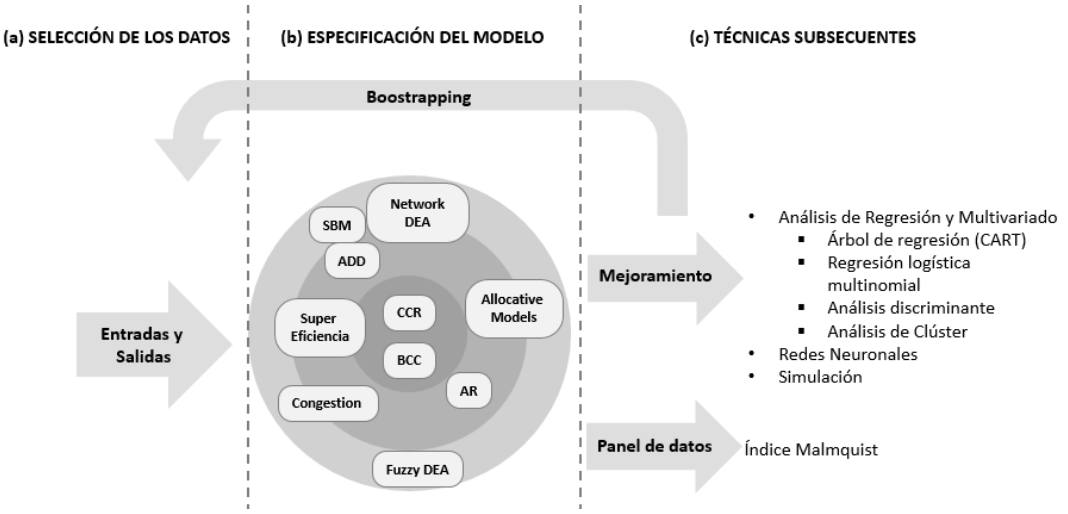

Figura 2.1: Etapas en la estimación de eficiencia con métodos DEA

Por otra parte, aunque DEA generalmente se relaciona al área de la investigación de operaciones con un enfoque de programación lineal no paramétrico para medir la eficiencia productiva, recientemente también se le asocia al área de la estadística por su interpretación como una regresión de mínimos cuadrados no paramétrica con restricciones de convexidad. Esta interpretación se le debe a Kuosmanen (2006) y Kuosmanen y Johnson (2010), quienes demostraron que la ineficiencia estimada por DEA orientada a las salidas bajo el enfoque de rendimientos variables a escala (VRS), a menudo denominado modelo BCC, es equivalente a la eficiencia estimada por SCNLS (sign-constrained convex nonparametric least squares) en un caso de salida única (Chen et al., 2021).

La representación SCNLS viene dada por:

$$
\begin{aligned}
\underset{\alpha,\beta,\varepsilon}{\text{Minimizar}} \quad & \sum_{i=1}^{n} \varepsilon_i^2 \\
\text{Sujeto a:} \quad & y_i = \alpha_i + \sum_{j=1}^{p} \beta_{ij} x_{ij} + \varepsilon_i \quad ;\forall \;\; i = 1, ..., n \\
& \alpha_i + \sum_{j=1}^{p} \beta_{ij} x_{ij} < \alpha_h + \sum_{j=1}^{p} \beta_{hj} x_{ij} \quad ;\forall \;\; i = 1, ..., n \;\; , h = 1, ..., n \\
& \beta_{ij} \geq 0 \quad ;\forall \;\; i = 1, ..., n \;\; , j = 1, ..., p \\
& \varepsilon_i \leq 0 \quad ;\forall \;\; i = 1, ..., n
\end{aligned}
\tag{2.3}
$$

Dónde ε es el término de ineficiencia, representando la desviación de una DMU a la frontera SCNLS estimada, y al igual que en un modelo de regresión, los parámetros α y β son el intercepto y los coeficientes de los predictores, respectivamente. Así como en la regresión estándar de mínimos cuadrados, esta función objetivo minimiza la suma de los residuos cuadrados con la diferencia que hay restricciones adicionales, lo que transforma el problema en un estimador de regresión convexa (Chen et al., 2021).

2.3. Problema de dimensionalidad

Resulta claro que, el análisis envolvente de datos (DEA) es una herramienta popular de los métodos cuantitativos para la toma de decisiones, que mide el desempeño relativo de un conjunto de unidades productoras (DMUs) con múltiples métricas de valoración (entradas y salidas de sus procesos productivos). Sin embargo, el número inicial de variables incluidas en un modelo DEA suele ser muy grande, y la formulación de DEA no proporciona pautas objetivas para la selección de entradas y salidas (Villanueva-Cantillo y Munoz-Marquez, 2021). Por lo cual, los problemas de discriminación entre DMUs eficientes e ineficientes a menudo surgen cuando hay un número relativamente grande de medidas de desempeño (variables) en comparación con el número de DMU; esto puede llevar a que las unidades eficientes se clasifiquen incorrectamente como ineficientes y las unidades ineficientes se clasifiquen erróneamente como eficientes (Charles et al., 2019).

Este problema puede poner en peligro la precisión o incluso la relevancia de los resultados (Chen et al., 2021), y surge debido a los tamaños de muestra pequeños, porque la varianza del estimador aumenta considerablemente a medida que aumenta la dimensionalidad del dominio de la función (Lee y Cai, 2020). En este caso, muchas DMUs serán clasificadas como eficientes debido a una pobre estimación de la frontera de producción.

Liu, Lu, y Lu (2016) identificaron en su estudio sobre los frentes de investigación en DEA, desde el 2000 hasta el 2014, que la selección de variables era una subárea de investigación solida en la literatura de DEA, citando 38 artículos dedicados a este tema.

Lo cual es entendible, dada la sensibilidad de la forma y la ubicación de la frontera eficiente, y consigo la variación en los puntajes de eficiencia, ante la configuración de variables utilizadas en DEA.

En consecuencia, existen estudios que se han enfocado en mejorar el poder discriminatorio de DEA, algunos con la intención de aumentar el número de DMUs y conservar el mismo número de variables, usando datos de corte transversal y series de tiempo (Charles et al., 2019). En cambio, otros han incorporado técnicas estadísticas como el análisis de covarianza parcial (Jenkins y Anderson, 2003), o análisis de componentes principales combinado con DEA (Adler y Golany, 2002), con el objetivo de reducir el número de variables utilizadas y conservando el mismo número de DMUs.

Por otra parte, otros estudios han incorporado enfoques basados en: la inclusión de juicios por un panel de expertos, la limitación del rango de valores de las ponderaciones, modelos de supereficiencia, entre otros (Charles et al., 2019).

2.4. Aproximaciones clásicas

Desde la investigación de operaciones, para los trabajos enmarcados en DEA, se han establecido reglas empíricas (reglas de oro) para evitar que demasiadas DMUs sean calificadas como eficientes. Estos umbrales empíricos relacionan el número de variables con el número de observaciones, entre ellas están:

- el número de DMU debería ser al menos el doble del número de entradas y salidas (Golany y Roll, 1989).
- el número de DMU debería ser al menos tres veces el número de entradas y salidas (Friedman y Sinuany-Stern, 1998).
- el número de DMU debería ser al menos el doble del producto del número de entradas y el número de salidas (Dyson et al., 2001).
- $n \geq max(m \times s, 3(m + s))$, donde n es el número de DMU, m es el número de entradas y s es el número de salidas. (Cooper, Seiford, y Tone, 2007).

A pesar de las reglas antes mencionadas, existen muchos estudios en medición del desempeño que no las cumplen (Adler y Golany, 2001; Sarkis, 2000). Sabiendo que cuando la cantidad de DMUs está por debajo de los niveles de umbral empírico, el poder discriminatorio entre las DMU puede debilitarse (Charles et al., 2019), se han propuesto otros enfoques clásicos en DEA, como la región de aseguramiento, el cual busca restricción en los valores que pueden tomar las ponderaciones de todos los factores de entrada v_i y salida u_r (se requiere la designación de límites superior e inferior para las ponderaciones de los factores).

2.5. Aproximaciones estadísticas

Algunos estudios de investigación en la literatura existente que se dedicaron a mejorar el poder discriminatorio de DEA, generalmente usan la reducción de variables justificada en un análisis de covarianza parcial (Jenkins y Anderson, 2003) o en el análisis de componentes principales combinado con DEA (Adler y Golany, 2001, 2002; Adler y Yazhemsky, 2010), o basado en la regularización de variables (Chen et al., 2021; Lee y Cai, 2020).

Covarianza Parcial

La idea de reducir el número de variables surgió al detectar la alta correlación entre las variables objeto de estudio y considerarlas para su exclusión. Por eso Jenkins y Anderson (2003), sustentados en esa premisa, propusieron un método estadístico sólido para la eliminación de variables, para evitar los criterios ad hoc o juicios particulares del investigador como se observa en la práctica. En su objetivo por reducir la dimensionalidad del espacio de solución del programa lineal y hacer más exigente el análisis, describen un método estadístico sistemático para decidir cuál de las variables correlacionadas pueden omitirse garantizando la menor pérdida de información y cuál deben conservarse en el estudio.

A causa de que la interrelación entre variables que están parcialmente correlacionadas rara vez es obvia (Jenkins y Anderson, 2003), y decidir cuáles pueden eliminarse con la menor pérdida de información, no es posible hacerlo solo con mirar la matriz de correlación. Por esto, usando estadística multivariada se pretende identificar qué variables se pueden omitir con la menor pérdida de varianza explicada, de la siguiente manera:

1. Estandarizar las variables para que tengan una media de 0 y una varianza de 1, y así tratar a todas las variables por igual (denotaremos con $\mathbf{X}$ a la matriz de datos normalizados).
2. Calcular la matriz de varianzas y covarianzas de $\mathbf{X}$ como:

$$var(\mathbf{X}) = \mathbf{V} = \begin{pmatrix} \sigma_{11} & \sigma_{12} & \cdots & \sigma_{1m} \\ \sigma_{21} & \sigma_{22} & \cdots & \sigma_{2m} \\ \vdots & \vdots & \ddots & \vdots \\ \sigma_{m1} & \sigma_{m2} & \cdots & \sigma_{mm} \end{pmatrix}$$

La covarianza entre dos variables $\mathbf{X}_i, \mathbf{X}_{i'}$ esta dado por:

$$\begin{aligned} Cov(\mathbf{X}_i, \mathbf{X}_{i'}) &= E\left\{[\mathbf{X}_i - E(\mathbf{X}_i)]\,[\mathbf{X}_{i'} - E(\mathbf{X}_{i'})]\right\} \\ &= E(\mathbf{X}_i\mathbf{X}_{i'}) - E(\mathbf{X}_i)E(\mathbf{X}_{i'}) \\ &= \sigma_{ii'} \end{aligned}$$

3. Dividir las m variables en dos conjuntos, de forma que $i = 1, ..., p$ sean las variables que serán omitidas y $i = p + 1, ..., m$ sean las variables que se conservaran como representativas con la mayoría de la información. Dicha partición de la matriz de varianzas y covarianzas $\mathbf{V}$ se representa como:

$$\mathbf{V} = \begin{pmatrix} \mathbf{V}_{11} & \mathbf{V}_{12} \\ \mathbf{V}_{21} & \mathbf{V}_{22} \end{pmatrix}$$

dónde $\mathbf{V}_{11}$ representa la matriz de varianzas y covarianzas de las variables $i = 1, ..., p$ y $\mathbf{V}_{22}$ la matriz de varianzas y covarianzas de las variables $i = p+1, ..., m$.

4. Calcular la matriz de varianzas y covarianzas parcial como:

$$\mathbf{V}_{11,2} = \mathbf{V}_{11} - \mathbf{V}_{12}\mathbf{V}_{22}^{-1}\mathbf{V}_{21}$$

Donde $\mathbf{V}_{11,2}$ representa la contribución de las variables $i = 1, ..., p$ después de condicionar con respecto a las $m - p$ variables retenidas $i = p + 1, ..., m$. Cabe recordar que, la covarianza parcial entre dos variables i e i' separando una tercera variable i'' viene dada por

$$\sigma_{ii'.i'} = \sigma_{ii'} - \sigma_{ii''}\sigma_{i'i''}/\sigma_{i''}^2$$

5. Finalmente, de la varianza total se resta el efecto de $(\mathbf{V}_{11,2})$ y se calcula su peso respecto a la misma varianza total, así: $(m - \mathbf{V}_{11,2})/m$. Puesto que la varianza condicional de una variable, denotada $\sigma_{ii.i''}$ se interpreta como la varianza que queda en la variable i cuando se elimina el efecto de la variable i''. Cabe aclarar que la varianza total es igual a m, porque las variables están estandarizadas.

Omitir variables que, por su alta correlación, aportan poca información a la metodología DEA es un enfoque lógico, particularmente cuando esta acción mejorara su rendimiento. Esta propuesta ofrece un procedimiento robusto con un criterio estadístico para garantizar la mínima pérdida de información en la reducción de variables. Lo anterior es necesario para evitar juicios arbitrarios o imposiciones de los investigadores. Sin embargo, Jenkins y Anderson (2003) afirman que sus resultados no son concluyentes, pués al evaluar su procedimiento con diferentes casos ilustrativos, demuestran que los resultados de DEA varían demasiado.

Análisis de Componentes Principales (PCA)

Como un número excesivo de variables en un modelo DEA provoca en los resultados una gran cantidad de unidades eficientes, es preferible mantener baja la relación entre el número de entradas y salidas con respecto al número de DMU, y para tal fin Adler y Golany (2001) usaron el Análisis de Componentes Principales (PCA) con el propósito de agregar entradas o salidas, con una pérdida mínima de información, desarrollando una nueva formulación del modelo DEA, denominada PCA–DEA

PCA describe la estructura de la varianza de una matriz de datos mediante combinaciones lineales de variables, lo que permite por medio de pocas componentes principales, explicar un porcentaje determinado de la varianza en todos los datos. Para ilustrar la técnica, supongamos que $X = [X_1, X_2, ..., X_p]$ es un vector aleatorio (las variables de entradas y salidas) con matriz de varianzas y covarianzas V con valores propios $\eta_1 \geq \eta_2 \geq ... \geq \eta_p \geq 0$ y vectores propios normalizados $l_1, l_2, ..., l_p$. Entonces, las componentes principales (X_{pc_i}) serán combinaciones lineales no correlacionadas de los datos originales y ordenadas descendentemente por la magnitud sus varianzas, así:

$$
\begin{aligned}
&X_{pc_i} = l_{1i}X_1 + l_{2i}X_2 + ... + l_{pi}X_p \\
&Var(X_{pc_i}) = (l_i)^t(V)(l_i) = \eta_i \qquad i = 1, 2, ..., p \\
&Cov(X_{pc_i}, X_{pc_k}) = l_i^t X l_k = 0 \qquad i = 1, 2, ..., p; \quad k = 1, 2, ..., k \quad i \neq k
\end{aligned}
$$

En vista de que los resultados de un PCA pueden tomar valores negativos y, tanto las entradas como las salidas de DEA deben ser estrictamente positivas, es necesario incrementar en un valor $b = 1 - min(X_{PC_i})$ para garantizar que las $X_{PC_i} = X_{PC_i} + b$ sean estrictamente positivas, esto aprovecha la propiedad del modelo DEA BCC de la Ecuación 2.4 (con retornos de escala variable y orientado a la salida), el cual es invariante a traslación en las entradas (Adler y Golany, 2001):

$$
\begin{aligned}
\underset{\lambda,s,\sigma}{\text{Minimizar}} \quad & T - e^t s - e^t \sigma \\
\text{Sujeto a:} \quad & Y\lambda - s = Y^a \\
& -X\lambda - \sigma = -TX^a \\
& e^t \lambda = 1 \\
& \lambda, s, \sigma \geq 0
\end{aligned}
\tag{2.4}
$$

Sin embargo, para la implementación del caso de estudio de este trabajo, se usó el modelo aditivo de DEA combinado con PCA presentado por Adler y Golany (2002) (véase la Ecuación 2.5), porque este aceptan valores negativos para X_{pc} y Y_{pc}. Esta formulación para la estimación de eficiencia de la DMU_a ($z = 1, 2, 3, ..., n$; $n =$ número de DMUs), agrega las variables originales y las variables transformadas por las componentes principales para las entradas ($X = \{X_o, X_{Lx}\}$) y salidas ($X = \{Y_o, Y_{Ly}\}$), donde X_o (Y_o) representa las entradas (salidas) de las variables originales, mientras que X_{Lx} (Y_{Ly}) las variables de entrada (salida) transformadas por PCA. Así mismo, L_x (L_y) representa la matriz de ponderaciones para las entradas (salidas) obtenidos con PCA, con la cual, $X_{pc} = L_x X_{Lx}$ y $X_{pc} = L_y Y_{Ly}$.

$$
\begin{aligned}
\underset{s_o,\sigma_o,s^+_{pc},s^-_{pc},\sigma^+_{pc},\sigma^-_{pc}}{\text{Minimizar}} \quad & -e^T s_o - e^T L'^{-1}(s^+_{pc} - s^-_{pc}) - e^T \sigma_o - e^T L'^{-1}(\sigma^+_{pc} - \sigma^-_{pc}) \\
\text{Sujeto a:} \quad & Y_o\lambda - s_o = Y^a_o \\
& Y_{pc}\lambda - (s^-_{pc} - s^+_{pc}) = Y^a_{pc} \\
& -X_o\lambda - \sigma_o = -X^a_o \\
& -X_{pc}\lambda - (\sigma^-_{pc} - \sigma^+_{pc}) = -X^a_{pc} \\
& L_y^{-1}(s^-_{pc} - s^+_{pc}) \geq 0 \\
& L_x^{-1}(\sigma^-_{pc} - \sigma^+_{pc}) \geq 0 \\
& \lambda, s_o, \sigma_o, s^+_{pc}, s^-_{pc}\sigma^+_{pc}, \sigma^-_{pc} \geq 0
\end{aligned}
\tag{2.5}
$$

Aquí, las holguras para determinar las ineficiencias están dadas para las entradas (σ_o) y salidas (s_o) originales, como para las variables transformadas ($\sigma^+_{pc}, \sigma^-_{pc}, s^+_{pc}, s^-_{pc}$), como variables auxiliares no negativas. Con el propósito de traducir los resultados del modelo adictivo, el cual arroja resultados entre $[0, \infty)$ (donde una holgura igual a cero significa ser eficiente), en un intervalo $[0, 1]$, se usa la Ecuación 2.6 propuesta por Adler y Volta (2019) para la comparabilidad con otros modelos.

$$
0 \leq 1 - \frac{\sum_{i=1}^{m} \frac{h^*_r}{x_{ki}} + \sum_{i=1}^{s} \frac{\sigma^*_r}{y_{kj}}}{m+s} \leq 1 \tag{2.6}
$$

Regresión Lasso

Lasso (least absolute shrinkage and selection operator - Lasso) pretende seleccionar un modelo de menor dimensión, llevando a cero el efecto estimado de algunas variables mediante la incorporación de un coeficiente de regularización o penalización al problema original, usualmente por mínimos cuadrados (Tibshirani, 1996). Esta regularización ayuda a minimizar el error de predicción del modelo, haciéndolo más parsimonioso y fácil de explicar. Este método es popular cuando la dimensión del problema es mayor que el tamaño de la muestra (Chen et al., 2021).

Para ilustrar, suponga un problema de regresión: un conjunto de datos con n observaciones, con variables independientes x_{ij} y una variable dependiente y_i para $i = 1, .., n$ y $j = 1, .., p$, con el cual se quiere estimar los coeficientes del siguiente modelo de regresión lineal:

$$
y_i = \alpha + \sum_{j=1}^{p} x_{ij}\beta_j \qquad i = 1, ..., n,
$$

donde α y β_j son el intercepto y los coeficientes de los regresores.

Las técnicas más comunes para ajustar este modelo son mediante el uso de Mínimos Cuadrados Ordinarios (OLS - Ordinary Least Squares) o por Mínimos Cuadrados Ponderados (WLS - Eeighted Least Squares). Pero cuando hay un gran número de variables explicativas, OLS o WLS pueden no ser muy confiables (Chen et al., 2021). Ante este problema (Tibshirani, 1996) sugiere regularizar el modelo mediante la penalización sobre la suma total de coeficientes, es decir, propone resolver el siguiente problema de regresión penalizado:

$$min_{(\alpha,\beta)} \frac{1}{2} \sum_{i=1}^{n} \left(y_1 - \alpha - \sum_{j=1}^{p} x_{ij}\beta_j \right)^2 + \lambda \sum_{j=1}^{p} |\beta_j| ,$$

donde λ es factor de penalización elegido por el investigador.

Una de las características más distintivas de Lasso, es que, además de reducir los coeficientes para reducir la varianza, este método también permite la selección de variables, reduciendo algunos coeficientes a cero, dependiendo del parámetro de penalización λ. Para valores grandes de λ mayor será el número de predictores que se reducirán a cero (Chen et al., 2021). Esta característica resulta de vital importancia al integrar Lasso y DEA.

Como un camino para eludir la maldición de la dimensionalidad Lee y Cai (2020) propusieron un enfoque de aprendizaje automático basado en el operador de selección y contracción mínima absoluta (least absolute shrinkage and selection operator - Lasso) para la selección de variables, combinado con mínimos cuadrados convexos no paramétricos restringidos por signo (sign-constrained convex nonparametric least squares - SCNLS, un caso particular de DEA), este se denominó Lasso-SCNLS.

En el estudio de simulación, Lee y Cai (2020) propusieron una comparación de diferentes métodos reduciendo la dimensión de DEA para conjuntos de datos pequeños. Concluyeron que, para el escenario de única salida y nueve entradas, la adaptación que integra el análisis de componentes principales (PCA) con DEA presenta un desempeño superior que la integración de Lasso con SCNLS. Mientras que, para escenario de dos salidas y nueve entradas, concluyeron que PCA-DEA y los métodos de selección de variables aleatorios no presentan resultados tan atractivos como si evidencia Lasso-SCNLS y sus variantes.

Lee y Cai (2020) adaptaron SCNLS con Lasso para la selección de variables (en un caso de salida única) de la siguiente manera:

$$
\begin{aligned}
\underset{\alpha,\beta,\varepsilon}{\text{Minimizar}} \quad & \sum_{i=1}^{n} \varepsilon_i^2 + \lambda \sum_{i=1}^{n} \sum_{j=1}^{p} \beta_{ij} \\
\text{Sujeto a:} \quad & y_i = \alpha_i + \sum_{j=1}^{p} \beta_{ij} x_{ij} + \varepsilon_i \qquad ; \forall_i = 1, ..., n \\
& \alpha_i + \sum_{j=1}^{p} \beta_{ij} x_{ij} \leq \alpha_h + \sum_{j=1}^{p} \beta_{hj} x_{ij} \qquad ; \forall i = 1, ..., n \\
& \beta_{ij} \geq 0 \qquad ; \forall i = 1, ..., n, j = 1, ..., n \\
& \varepsilon_i \leq 0 \qquad ; \forall i = 1, ..., n
\end{aligned}
\tag{2.7}
$$

Lasso tiene grandes ventajas para equilibrar de manera óptima el sesgo y la varianza y hacer que el modelo sea parsimonioso con una precisión de predicción alta, pero como con cualquier método, Lasso tiene sus limitaciones, las cuales puede heredar Lasso-SCNLS. Como se ha evidenciado en varios trabajos, una de las principales debilidades de Lasso es que su desempeño es débil ante la presencia de altas correlaciones entre regresores (Chen et al., 2021).

Por lo anterior, Chen et al. (2021) exploraron una versión más avanzada de Lasso, enfoque conocido como Elastic Net (EN), lo adaptaron a DEA y proponen el enfoque EN-DEA. Cuando en el modelo EN-DEA $\delta = 1$, este se convierte en LASSO-SCNLS, mientras que, Cuando $\delta = 0$ el modelo podría llamarse "Ridge-DEA" (o Ridge-SCNLS). Y así mismo, cuando $\beta_{ij} = 0$ para todas observaciones (DMU) en una variable determinada, dicha variable se elimina del análisis. La formalización matemática de EN-DEA, o en otras palabras EN-SCNLS, es la siguiente:

$$
\begin{aligned}
\underset{\alpha,\beta,\varepsilon}{\text{Minimizar}} \quad & \sum_{i=1}^{n} \varepsilon_i^2 + \lambda \left[\delta \sum_{j=1}^{p} \sum_{i=1}^{n} \beta_{ij} + (1-\delta) \sum_{j=1}^{p} \sum_{i=1}^{n} \beta_{ij}^2 \right] \\
\text{Sujeto a:} \quad & y_i = \alpha_i + \sum_{j=1}^{p} \beta_{ij} x_{ij} + \varepsilon_i, \forall_i = 1, ..., n \\
& \alpha_i + \sum_{j=1}^{p} \beta_{ij} x_{ij} \leq \alpha_h + \sum_{j=1}^{p} \beta_{hj} x_{ij}, \forall i = 1, ..., n \\
& \beta_{ij} \geq 0, \forall i = 1, ..., n, j = 1, ..., n \\
& \varepsilon_i \leq 0, \forall i = 1, ..., n
\end{aligned}
\tag{2.8}
$$

Exite un enfoque dos etapas que denominado como "Lasso-DEA", el cual sorprendentemente, resulta ser más prometedor que oitras variaciones de Lasso como la regularización Elastic Net (Chen et al., 2021). Este procedimiento de dos etapas, para solucionar la maldición de la dimensionalidad en DEA, en su primera etapa utiliza Lasso estándar, para seleccionar un número óptimo de predictores

debido a una elección óptima del parámetro de ajuste λ con validación cruzada; para finalmente, en la segunda etapa ejecutar DEA (o SCNLS) solo en las variables seleccionadas como relevantes del paso anterior.

Capítulo 3

APROXIMACIÓN HEURÍSTICA

Este capítulo presenta la metodología que permite la construcción de un conjunto de modelos DEA para la selección de variables desde un punto de vista heurístico, en pro de mejorar el poder discriminatorio de DEA con una gran cantidad de variables. Las siguientes secciones muestran el seudocódigo de la propuesta heurística, una breve explicación de las medidas de impureza o diversidad utilizadas y un ejemplo ilustrativo.

3.1. Medidas de diversidad

El Coeficiente de Concentración de Gini (CG) es uno de los más utilizados en el estudio de la desigualdad, y aunque existen varias formas de derivar la expresión algebraica usada en su cálculo, también es posible deducirlo mediante la curva de Lorenz (Medina, 2001). El índice de Gini se construye al comparar la distribución empírica de los datos observados y la distribución teórica derivada de la curva de Lorenz (línea de igualdad perfecta).

Para calcular el coeficiente de concentración del ingreso, supóngase que se tiene el valor del ingreso de n individuos ordenados en forma ascendente, de la forma

$$y_1 \leq y_2, ..., \leq y_n \ ,$$

y se estiman las distribuciones de frecuencias relativas simples (fn_i) y acumuladas (Fn_i) de la población objeto de análisis, así como de la variable a distribuir (ingreso per cápita) fy_i y Fy_i. Por el ordenamiento de los datos, tenemos que $Fy_i \leq Fy_{i+1}$.

Conforme a lo anterior, el Gini se basa en la suma de las diferencias de las distribuciones acumuladas ($Fn_i - Fy_i$). Con el propósito de estandarizar su recorrido al intervalo $[0, 1]$, la expresión anterior se divide entre $\sum_{i=1}^{n-1} Fn_i$, con lo cual obtenemos una de las

fórmulas utilizadas para calcular el índice *Gini* (Medina, 2001):

$$Gini = \frac{\sum_{i=1}^{n-1}(Fn_i - Fy_i)}{\sum_{i=1}^{n-1} Fn_i}$$

Extrapolando este concepto a la metodología DEA, tendríamos que los individuos serian las DMUs, mientras que los ingresos serian, bajo un enfoque discreto, la categorización de la eficiencia (donde 1 = eficiente y 0 = ineficiente), y para un enfoque continuo, los puntajes de eficiencia estimados. Como se puede observar en la Figura 3.1, las valoraciones de esta formulación Gini tienden a incrementar conforme se pierde la noción de igualdad perfecta, en la cual todas las unidades productoras son eficientes, y a medida que baja la cantidad de DMUs eficientes la brecha entre la distribución teórica y empírica se incrementa.

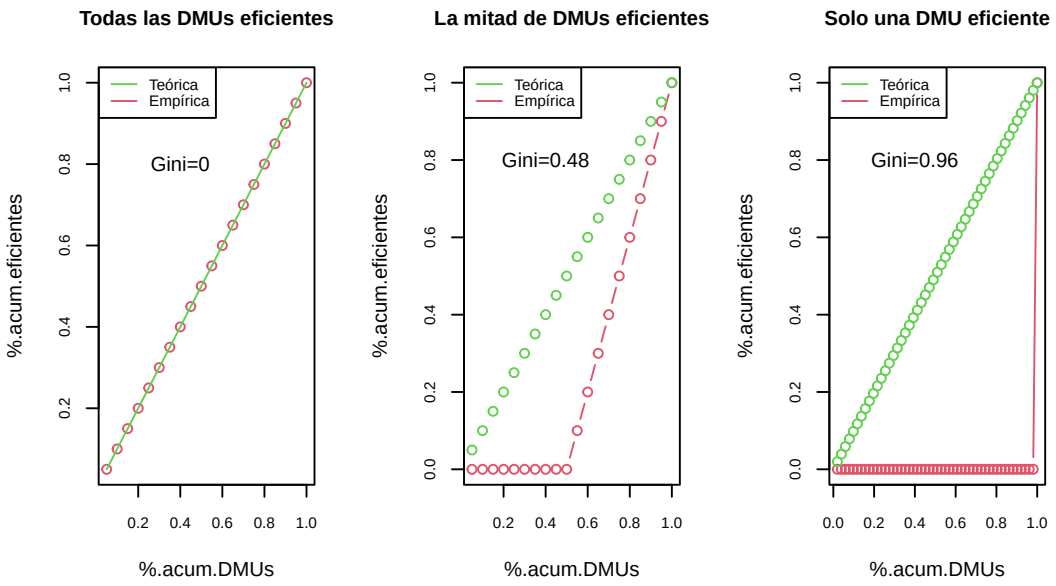

Figura 3.1: Curvas de Lorenz para tres escenarios DEA

Otra alternativa es la Entropía, usualmente utilizada en problemas de clasificación como criterio para realizar divisiones binarias en los nodos (James, Witten, Hastie, y Tibshirani, 2013). La entropía está dada por:

$$E = -\sum_{i=1}^{K}(\hat{p}_i)(log\hat{p}_i),$$

donde p_i representa la proporción de observaciones que pertenecen a la i-ésima categoría (Categorías = K = {eficiente, ineficiente}). En la Figura 3.2 se puede apreciar la

dinámica de esta métrica para las diferentes proporciones de DMUs eficientes, y como esta indica un valor de cero (0) ante la presencia de una alta desigualdad, tanto para una alta proporción de DMUs eficientes como para una alta proporción de DMUs ineficientes. Además, esta logra una valoración de 1 cuando las proporciones entre clases están completamente balanceadas.

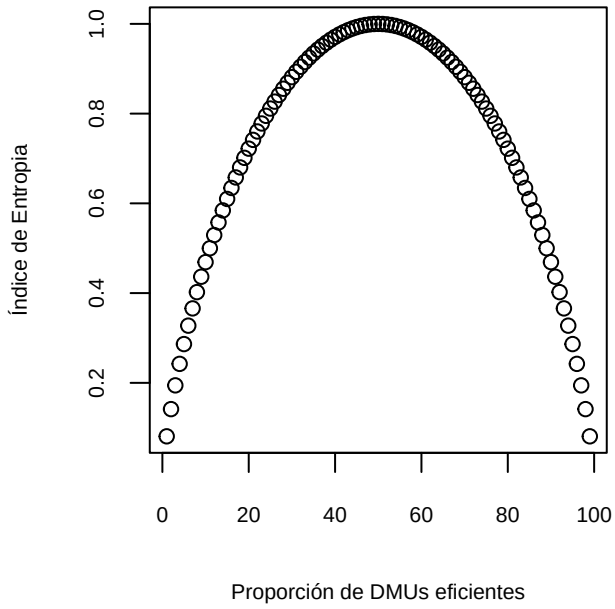

Figura 3.2: Rango del Índice de Entropía

3.2. Esquema general

La estructura básica de esta propuesta heurística se describe en la Tabla 3.2, dicho algoritmo fue implementado en `R` (R Core Team, 2021), y particularmente se resalta el uso de la función `dea` de la librería `Benchmarking` (Bogetoft y Otto, 2020) para la estimación de los puntajes de eficiencia.

Este método de solución está basado en el concepto de búsqueda iterativa, donde se aumenta el poder discriminatorio de DEA, explorando las configuraciones de variables que maximizan el índice Gini o la entropía de los puntajes de eficiencia. Los parámetros de inicialización son los valores de las variables de entrada (insumos requeridos por las unidades productoras) y de salida (productos generados por las unidades evaluadas), y el criterio de parada, que puede ser:

- el valor mínimo sugerido en las reglas empíricas (como que la cantidad de variables sea por lo menos un $\frac{1}{2}$ o $\frac{1}{3}$ de la cantidad de DMUs) (Dyson et al., 2001) (Cook et al., 2014)

- un valor mínimo en la variación entre la métrica de impureza máxima anterior y la actual.

Cabe aclarar que en la estructura de los datos representan las DMUs en las filas y las variables en las columnas, y también que la selección de variables debe considerar como mínimo una entrada y una salida, ya que dentro de los principios de producción, los recursos deben transformarse en productos.

El algoritmo heurístico propuesto se basa en la eliminación iterativa de variables (Iterative Search Algorithm, ISA-DEA). Como se puede observar en la Tabla 3.2, se parte con la lectura de los datos, inicialización de parámetros, y el cálculo de las métricas que guían el análisis (puntajes de eficiencia, los índices Gini, y la varianza retenida) del modelo con todas las variables. Luego, comienza a iterar calculando las posibles combinaciones de modelos omitiendo una variable, y explora tanto los modelos que consideran la eliminación de una entrada (combinatoria ${}_mC_{m-1}$) como los que eliminan una salida (combinatoria ${}_sC_{s-1}$). Después de estimar las métricas trazadoras, se busca la configuración de variables que maximiza el índice de desigualdad sobre los puntajes de eficiencia estimados, ya que con $\alpha = 1$ tenemos

$$\alpha \times (Gini\_Eficiencia) + \cancelto{0}{(1-\alpha)} \times (Gini\_Ineficiencia) = \alpha \times (Gini\_Eficiencia),$$

ya sea con la depuración de una variable de entrada ($arg\ max_{M'}$) o una de salida ($arg\ max_{S'}$); lo anterior para definir la reducción del conjunto de variables y consigo del espacio de búsqueda que usará la siguiente iteración; siempre y cuando el criterio de parada establecido no se haya cumplido, ya sea el $\Delta Impureza$(Modelos) o las reglas empíricas de DEA, o hasta quedar con tan solo una variable de entrada y una de salida (requerimiento propio en DEA). Finalmente, el algoritmo retorna todos los modelos evaluados y la ruta de los modelos máximos heurísticamente (modelos donde el Gini fue máximo).

Además, en la Tabla 3.3 se expone una extensión en la que se ponderan los índices Gini de la eficiencia y la ineficiencia $(1 - eficiencia)$, de la siguiente forma:

$$\alpha \times (Gini\_Eficiencia) + (1-\alpha) \times (Gini\_Ineficiencia)$$

variando el parámetro α (con $0 \leq \alpha \leq 1$). Esta extensión propone balancear la métrica que orienta la búsqueda en la selección de variables entre las dos categorías del análisis, donde una DMU es eficiente si su puntaje de eficiencia es igual al 100 % (se ubica en la frontera), y por el contrario una DMU es ineficiente si su estimación es menor al 100 %. En la Figura 3.3 se presenta gráficamente el comportamiento de los índices Gini ante diferentes proporciones de DMUs eficientes, por facilidad se ilustra el caso de clasificación binaria ($eficiente = 1$, e $ineficiente = 0$). Esto con el propósito de explorar más el espacio de solución y contar con una métrica que salvaguarde un equilibro entre las dos categorías inmersas en el análisis, ya que no se pretende desaparecer o minimizar por completo alguna de ellas.

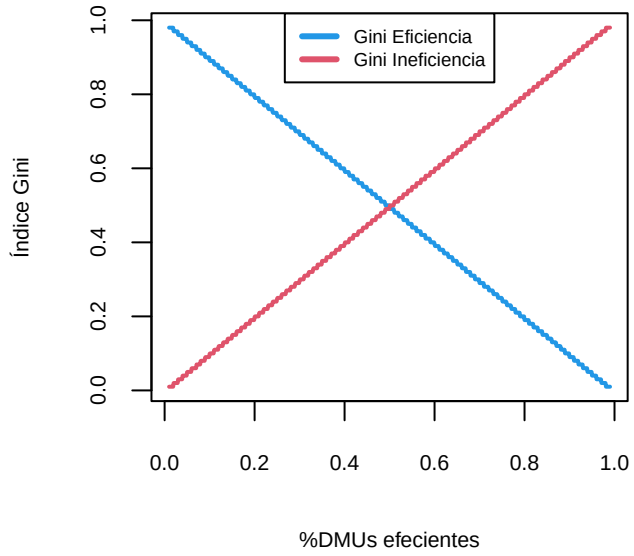

Figura 3.3: Comportamiento de índices según proporción DMUs eficientes

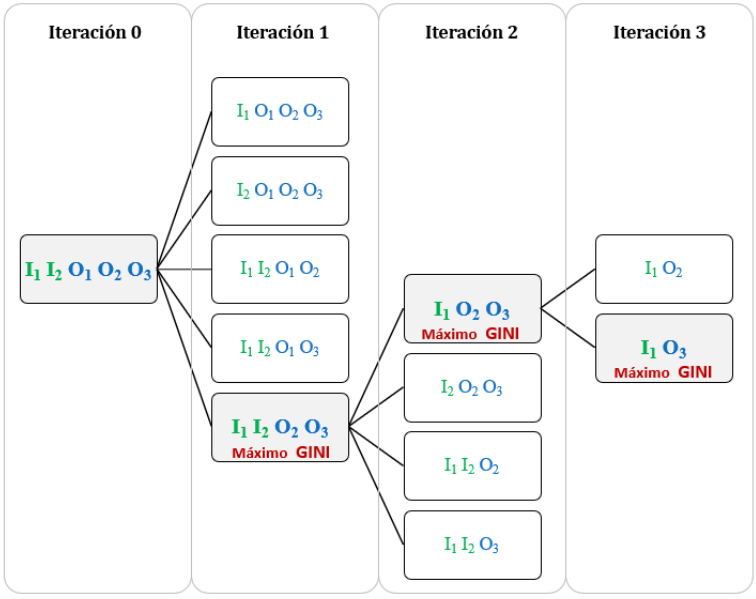

Figura 3.4: Ejemplificación del proceso iterativo

3.3. Ejemplo esquemático

Como se muestra en la Figura 3.4 para el caso de dos entradas (I_1 I_2) y tres salidas (O_1 O_2 O_3), inicialmente se calculan los puntajes de eficiencia para el conjunto total de

variables de entrada y salida (Iteración 0 - I_1 I_2 O_1 O_2 O_3), para una variancia retenida del 100 %, y donde la métrica de impureza se espera que sea baja (ante la presencia de muchas DMUs clasificadas como eficientes).

Luego, la siguiente iteración toma las variables del modelo anterior y crea un conjunto con todas las posibles combinaciones de variables, excluyendo solo una en cada configuración, tanto para las variables de entrada como de salida; a estos modelos candidatos se les calcula sus puntajes de eficiencia, métricas de impureza (índices Gini y Entropía) y varianza explicada. Aquí, se elige el modelo que presente el índice Gini más alto, este va orientando la búsqueda iterativa, garantizando una reducción en el exceso de DMUs clasificadas como eficientes.

Finalmente, se repite el proceso de forma iterativa hasta que se cumpla el criterio de parada, o hasta que la eliminación de variables haya sido exhaustiva, al punto de solo contar con dos variables (una de entrada y una de salida). Siguiendo el ejemplo de la Figura 3.4, en la iteración 1 se elige el modelo con las variables I_1 I_2 O_2 O_3, las cuales serán el insumo de la iteración 2, para la cual se elige el modelo I_1 O_2 O_3, finalmente, el modelo que mejora el poder discriminatorio de DEA en la iteración 3 sería el modelo con las variables I_1 O_3.

Cabe resaltar que, para este pequeño ejemplo de cinco variables, el enfoque heurístico evaluaría 12 modelos hasta quedarse con 1 entrada y 1 salida, en vez de 60 modelos que representaría la exploración del espacio factible de forma exhaustiva. Que, aunque parece factible realizarlo gracias a la capacidad de cómputo en la actualidad, el doblar la cantidad de variables de 5 a 10, eleva la cantidad de modelos de 60 a 1,800,000. En la Tabla 3.1 se presenta la cantidad de posibles modelos a evaluar dependiendo del número de variables, y resulta claro que el espacio de búsqueda crece exponencialmente para una búsqueda exhaustiva, y le da relevancia a un propuesta heurística para realizar la búsqueda de soluciones de una forma práctica y certera.

Tabla 3.1: Cantidad de modelos posibles con una búsqueda exhaustiva

Numero Variables	**Modelos posibles**
2	1
3	3
4	12
5	60
6	360
7	2,520
8	20,160
9	181,440
10	1,814,400
11	20,000,000
12	240,000,000
13	3,110,000,000
14	43,600,000,000
15	654,000,000,000

3.4. Implementación en R

En el Apéndice A se detalla la programación y funciones utilizadas en este trabajo. Allí, en la Sección I se encuentra la función con la que iteración tras iteración se construyen las configuraciones de variables. La Sección II contiene la sistematización de la metodología de Jenkins y Anderson (2003) para el cálculo de la covarianza parcial y para la estimación de proporción de la varianza explicada, dada la selección de variables en los diversos modelos; así mismo, en la Sección III esta la sistematización del coeficiente de concentración de Gini de Medina (2001). Por otra parte, en la Sección IV y Sección V se detalla la programación de la propuesta metodológica de este trabajo. Y finalmente, en la Sección VI y Sección VII se integran las propuestas expuestas en la literatura de Regresión Lasso (Chen et al., 2021; Lee y Cai, 2020) y Análisis de Componentes Principales (Adler y Golany, 2001, 2002; Adler y Yazhemsky, 2010), respectivamente.

Tabla 3.2: Metodología aplicada solo con Gini de la eficiencia

Algoritmo de búsqueda iterativa (con $\alpha = 1$)

Parámetros: input_data, output_data, stop_criterion, épsilon (por defecto 0.0001)
dmu : número de unidades a evaluar
m: entradas
s: salidas
M: ${}_mC_m = \binom{m}{m}$
S: ${}_sC_s = \binom{s}{s}$
$Umbral \leftarrow$ stop_criterion $\in \{$épsilon, $\lfloor dmu/2 \rfloor, \lfloor dmu/3 \rfloor\}$
Calcular métricas: Eficiencia(M-S), Impureza(M-S), Varianza_Retenida(M-S)
$Bandera \leftarrow |m| + |s|$
Mientras que $Bandera > Umbral$ **haga**
  M': ${}_mC_{m-1} = \binom{m}{m-1}$
  S': ${}_sC_{s-1} = \binom{s}{s-1}$
  Si $|m| > 1$ **entonces**
    Para $i = 1$ hasta M' **haga**
      Calcular métricas: Eficiencia(M'_i-S), Impureza(M'_i-S), Varianza_Retenida(M'_i-S)
    $i = i + 1$
    Fin Para
  Fin si
  Si $|s| > 1$ **entonces**
    Para $j = 1$ hasta S' **haga**
      Calcular métricas: Eficiencia(M-S'_j), Impureza(M-S'_j), Varianza_Retenida(M-S'_j)
    $j = j + 1$
  Fin Para
  Fin si
  Guardar y acumular proceso (modelos candidatos)
  Si $arg\ max_{M'}(Impureza(M')) > arg\ max_{S'}(Impureza(S'))$ **entonces**
    $M \leftarrow arg\ max_{M'}(Impureza(M'))$
    $|m| \leftarrow |m| - 1$
  Si no
    $S \leftarrow arg\ max_{S'}(Impureza(S'))$
    $|s| \leftarrow |s| - 1$
  Fin si
  Guardar y acumular ruta (modelos máximos heurísticamente)
Si stop_criterion = épsilon **entonces**
  $Bandera \leftarrow \Delta Impureza(\text{Modelos}_{\text{máximos heurísticamente}})$
Si no
  $Bandera \leftarrow |m| + |s|$
Fin si
Fin Mientras
Retornar el proceso (modelos candidatos) y la ruta (modelos máximos heurísticamente)

Tabla 3.3: Metodología aplicada con Gini de la eficiencia y ineficiencia

Algoritmo de búsqueda iterativa (con $0 \leq \alpha \leq 1$)

Parámetros: input_data, output_data, stop_criterion, épsilon (por defecto 0.0001)
Mientras que $\alpha \leq 1$ **haga**
  dmu : número de unidades a evaluar
  m: entradas
  s: salidas
  M: ${}_mC_m = \binom{m}{m}$
  S: ${}_sC_s = \binom{s}{s}$
  $Umbral \leftarrow$ stop_criterion $\in \{$épsilon, $\lfloor dmu/2 \rfloor, \lfloor dmu/3 \rfloor\}$
  Calcular métricas: Eficiencia(M-S), Impureza(M-S), Varianza_Retenida(M-S)
  $Bandera \leftarrow |m| + |s|$
    Mientras que $Bandera > Umbral$ **haga**
      M': ${}_mC_{m-1} = \binom{m}{m-1}$
      S': ${}_sC_{s-1} = \binom{s}{s-1}$
      Si $|m| > 1$ **entonces**
        Para $i = 1$ hasta M' **haga**
          Calcular métricas: Eficiencia(M'_i-S), Impureza(M'_i-S), Varianza_Retenida(M'_i-S)
        $i = i + 1$
        Fin Para
      Fin si
      Si $|s| > 1$ **entonces**
        Para $j = 1$ hasta S' **haga**
          Calcular métricas: Eficiencia(M-S'_j), Impureza(M-S'_j), Varianza_Retenida(M-S'_j)
        $j = j + 1$
      Fin Para
      Fin si
      Guardar y acumular proceso (modelos candidatos)
      Si $arg\ max_{M'}(Impureza(M')) > arg\ max_{S'}(Impureza(S'))$ **entonces**
        $M \leftarrow arg\ max_{M'}(Impureza(M'))$
        $|m| \leftarrow |m| - 1$
      Si no
        $S \leftarrow arg\ max_{S'}(Impureza(S'))$
        $|s| \leftarrow |s| - 1$
      Fin si
      Guardar y acumular ruta (modelos máximos heurísticamente)
    Si stop_criterion = épsilon **entonces**
      $Bandera \leftarrow \Delta Impureza(\text{Modelos}_{\text{máximos heurísticamente}})$
    Si no
      $Bandera \leftarrow |m| + |s|$
    Fin si
  Fin Mientras
$\alpha = \alpha + 0{,}01$
Fin Mientras
Retornar el proceso (modelos candidatos) y la ruta (modelos máximos heurísticamente)

Capítulo 4

CASO DE ESTUDIO

Este capítulo esta dividido en dos secciones: la primera (Estrategias de solución) tiene como objetivo contrastar las aproximaciones metodológicas en DEA combinadas con técnicas estadísticas para robustecer el análisis de eficiencia, a través de datos con los cuales ya se han validado propuestas en DEA, y en la segunda sección (Evaluación de eficiencia del sector salud colombiano) se aplica el método con mejor rendimiento que aumenta el poder discriminatorio de DEA, y con el se aborda el análisis de eficiencia para el sector salud colombiano.

4.1. Estrategias de solución

Tabla 4.1: Datos DEA – Facultades universitarias

	Entradas (Inputs)			**Salidas (Outputs)**		
DMU	**I1** **Empleados Docentes**	**I2** **Salarios Docentes**	**I3** **Empleados Administrativos**	**O1** **Estudiantes de pregrado**	**O2** **Estudiantes de posgrado**	**O3** **Publicaciones (Artículos)**
DMU1	12	400	20	60	35	17
DMU2	19	750	70	139	41	40
DMU3	42	1500	70	225	68	75
DMU4	15	600	100	90	12	17
DMU5	45	2000	250	253	145	130
DMU6	19	730	50	132	45	45
DMU7	41	2350	600	305	159	97

Con el finde ilustrar el rendimiento de la aproximación heurística propuesta, detallada en el Capítulo 3, usaremos un conjunto de datos existente en la literatura, reportados en los trabajos de Wong y Beasley (1990) y Adler y Golany (2002) para presentar sus resultados. En esta ilustración numérica, se comparan siete DMUs (facultades universitarias) con tres variables de insumos (número de docentes, salarios de los docentes, número de empleados administrativos) y tres variables de salida (número de

estudiantes de pregrado, número de estudiantes de posgrado, y cantidad de artículos publicados), como puede verse en la Tabla 4.1. Cabe resaltar que, al aplicar DEA con todas las variables para las siete unidades productoras, la DMU_4 es la única ineficiente (véase la Tabla 4.2). A continuación, se aplican las técnicas PCA-DEA, Lasso-DEA e ISA-DEA (Iterative Search Algorithm, la metodología propuesta), las cuales buscan mejor la clasificación de eficiencia en DEA.

Tabla 4.2: Puntajes de eficiencia con DEA con todas las variables

Puntajes de eficiencia (DMU eficiente= 1, DMU ineficiente= (0,1))							
DMU	DMU1	DMU2	DMU3	DMU4	DMU5	DMU6	DMU7
Puntaje	1	1	1	0.820	1	1	1

4.1.1. Análisis de Componentes Principales combinado con DEA (PCA-DEA)

Se aplicó la técnica de análisis de componentes principales combinado con DEA, con la formulación de la Ecuación 2.5 para todas las posibles configuraciones de componentes, esto con el objetivo de abarcar todo el espectro en la reducción de dimensionalidad y su impacto en la estimación de eficiencia. Para los nueve posibles modelos evaluados se presentan en la Tabla 4.3 las métricas de impureza (Gini para la eficiencia, la ineficiencia y el ponderado), la cantidad y la proporción de DMUs eficientes, las varianzas retenidas para las entradas y salidas (métricas propias de PCA), y la varianza retenida por el total de variables (usando la Covarianza Parcial de Jenkins y Anderson (2003)).

Tabla 4.3: Métricas de impureza y varianza retenida para los modelos con PCA-DEA

Modelo	**Componentes Principales**		**Gini**	**Gini**	**Gini**	**DMUs**	**% DMUs**	**Varianza Retenida**		
	Entradas (CPi)	**Salidas (CPo)**	**Eficiencia**	**Ineficiencia**	**Ponderado**	**Eficientes**	**Eficientes**	**Entradas**	**Salidas**	**Total**
$Modelo_1$	3	3	0.0412	0.7500	0.3956	6	85.7 %	1.0000	1.0000	1.0000
$Modelo_2$	3	2	0.0398	0.7500	0.3949	6	85.7 %	1.0000	0.9755	1.0000
$Modelo_3$	3	1	0.0363	0.5546	0.2955	3	42.9 %	1.0000	0.9456	1.0000
$Modelo_4$	2	3	0.1036	0.5663	0.3349	4	57.1 %	0.9995	1.0000	1.0000
$Modelo_5$	2	2	0.1010	0.5638	0.3324	4	57.1 %	0.9995	0.9755	0.9728
$Modelo_6$	2	1	0.0673	0.2255	0.1464	1	14.3 %	0.9995	0.9456	0.9122
$Modelo_7$	1	3	0.1253	0.5431	0.3342	4	57.1 %	0.8679	1.0000	0.9839
$Modelo_8$	1	2	0.1204	0.4907	0.3056	2	28.6 %	0.8679	0.9755	0.8893
$Modelo_9$	1	1	0.0803	0.2167	0.1485	1	14.3 %	0.8679	0.9456	0.7882

Resulta claro que al utilizar menos componentes principales, la varianza retenida disminuye, y se evidencia en la Tabla 4.3 como esto tiene un impacto en la reducción de DMUs estimadas como eficientes, pasando de un 85.7 % a 14.3 % de DMUs eficientes entre el $Modelo_1$ y el $Modelo_9$, modelos con la máxima y mínima retención de varianza respectivamente. Además, se puede apreciar como la DMU6 es la única eficiente en los nueve modelos, bajo todas las configuraciones de reducción de dimensionalidad (véase la Tabla 4.4), convirtiéndola en un referente clave del grupo evaluado, resultado que es consistente con los resultados de las tres configuraciones de componentes principales presentados por Adler y Golany (2002).

Tabla 4.4: Puntajes de eficiencia para los modelos con PCA-DEA

Modelo	Puntajes de eficiencia (DMU eficiente= 1, DMU ineficiente= (0,1))						
	DMU1	**DMU2**	**DMU3**	**DMU4**	**DMU5**	**DMU6**	**DMU7**
Modelo_1	1.0000	1.0000	1.0000	0.6352	1.0000	1.0000	1.0000
Modelo_2	1.0000	1.0000	1.0000	0.6471	1.0000	1.0000	1.0000
Modelo_3	1.0000	0.9438	0.9582	0.7663	0.9015	1.0000	1.0000
Modelo_4	1.0000	1.0000	0.7721	0.4849	1.0000	1.0000	0.6609
Modelo_5	1.0000	1.0000	0.7790	0.5074	1.0000	1.0000	0.6502
Modelo_6	0.8186	0.7643	0.7890	0.6466	0.7104	1.0000	0.6625
Modelo_7	1.0000	1.0000	0.6248	0.4185	1.0000	1.0000	0.6448
Modelo_8	0.9636	0.9405	0.6330	0.4560	1.0000	1.0000	0.6278
Modelo_9	0.6664	0.8784	0.6457	0.6177	0.6569	1.0000	0.6419

Tradicionalmente, la selección de modelos bajo el enfoque de componentes principales se hace bajo la fijación de un bajo número de componentes que retengan una alta proporción de varianza, este umbral de retención de varianza es fijado por quien realiza el análisis (por ejemplo, mediante el uso del gráfico de sedimentación). No obstante, con el propósito de guiar la selección de variables de los modelos para mejorar el poder discriminatorio de DEA, complementamos los resultados calculando las métricas de impureza (véase la Tabla 4.3), donde se identifica que los modelos con el máximo índice "Gini de Eficiencia", son el Modelo_7 y el Modelo_8 (con 0.1253 y 0.1204 respectivamente) para los cuales se observa una disminución del 33 % y 67 % en la cantidad de DMUs clasificadas como eficientes respecto al modelo con el total de variables disponibles.

Por otra parte, basado en las reglas empíricas para la técnica DEA, desde la investigación de operaciones, donde relacionan el número máximo de variables (de entrada y de salida) y la cantidad de DMUs evaluadas (Dyson et al., 2001) (Cook et al., 2014), cabe resaltar que el Modelo_8 es el único cuya división entre la cantidad de componentes usadas (1 de entrada y 2 de salidas) y el número de DMUs (7 unidades productoras) se encuentra en el rango $[\frac{1}{3}, \frac{1}{2}]$, por lo que se consideraría un modelo apto para aplicar DEA.

Finalmente, en la Figura 4.1 por medio de un gráfico burbuja, también se puede detectar rápidamente como la cantidad de DMUs eficientes, determinada por el tamaño del circulo, disminuye conforme la varianza explicada en variables de entrada y salida es menor. Esta herramienta visual permite identificar los modelos que se encuentran en umbrales aceptables, tanto para el sacrificio de varianza explicada como en el exceso de unidades clasificadas como eficientes. En color rojo se resalta el Modelo_8 como la mejor opción a elegir, ya que es el segundo modelo con el índice Gini para la eficiencia más alto, y además, se encuentra en el intervalo $[\frac{1}{3}, \frac{1}{2}]$, una vez extrapolamos las reglas empíricas usadas en DEA. Este modelo resaltado con la etiqueta "CPi:1−CPo:2" indica que para las entradas (inputs) se tome la primera componente principal (CPi:1) mientras que para las salidas se tomen las dos primeras (CPo:2).

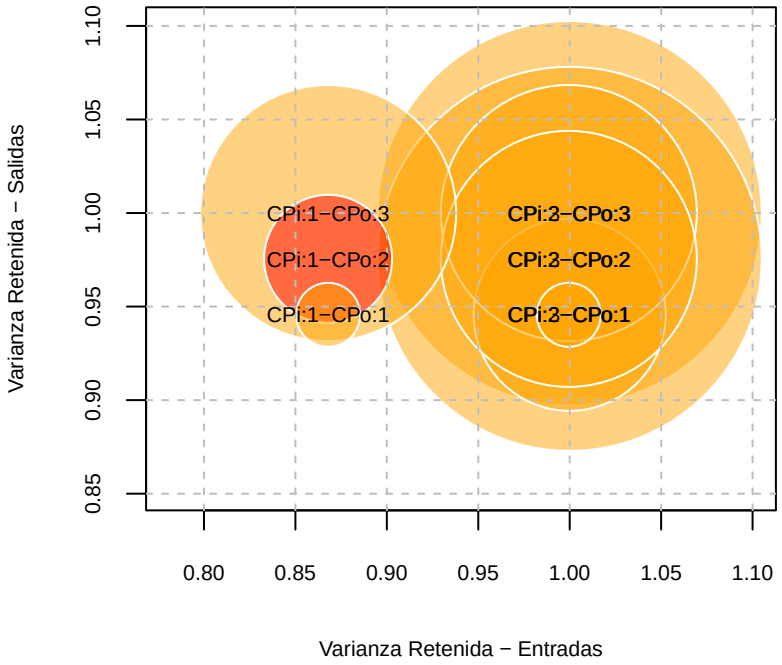

Figura 4.1: Cantidad de DMUs eficientes dada las varianzas retenidas (PCA-DEA)

4.1.2. Regresión Lasso combinado con DEA (Lasso-DEA)

Para la selección de variables en DEA, usualmente se emplea la regresión Lasso en la búsqueda de aquellas variables que mejor explican la eficiencia. Aquí aplicamos la penalización Lasso sobre todas las variables disponibles (entradas y salidas), y al igual que en el ejercicio anterior, calculamos las métricas de impureza y la covarianza parcial para cada uno de los modelos, con el propósito de homologar esta aproximación con la propuesta metodológica expuesta en el presente trabajo.

Tabla 4.5: Métricas de impureza y varianza retenida para los modelos con Lasso-DEA

Modelo	λ	Varibles						Gini	Gini	Gini	DMUs	% DMUs	Varianza Retenida		
		I1	I2	I3	O1	O2	O3	Eficiencia	Ineficiencia	Ponderado	Eficientes	Eficientes	Entradas	Salidas	Total
Modelo_1	0	✓	✓	✓	✓	✓	✓	0.0198	0.7500	0.3849	6	85.7 %	1.0000	1.0000	1.0000
Modelo_2	0.0012	✓	✓	✓	✓	✓		0.0198	0.7500	0.3849	6	85.7 %	1.0000	0.9602	1.0000
Modelo_3	0.0021	✓	✓		✓	✓	✓	0.0334	0.6274	0.3304	5	71.4 %	0.9958	1.0000	1.0000
Modelo_4	0.0037		✓	✓	✓	✓	✓	0.0334	0.6368	0.3351	5	71.4 %	0.9983	1.0000	1.0000
Modelo_5	0.0095			✓	✓	✓	✓	0.2400	0.4161	0.3281	2	28.6 %	0.6955	1.0000	0.9920
Modelo_6	0.0115			✓	✓		✓	0.2426	0.3898	0.3162	1	14.3 %	0.6955	0.9631	0.9879
Modelo_7	0.0429			✓	✓			0.2629	0.3773	0.3201	1	14.3 %	0.6955	0.8913	0.9435
Modelo_8	0.0518	✓		✓	✓			0.0435	0.5863	0.3149	4	57.1 %	0.9992	0.8913	0.9649
Modelo_9	0.0754	✓			✓			0.0719	0.3742	0.2230	1	14.3 %	0.7515	0.8913	0.9222

En la Tabla 4.5 se resumen los modelos para cada λ de penalización para el cual se dio una cambio en la configuración de variables, es decir, la inclusión u omisión de alguna variable. Por ejemplo, el Modelo_1 tiene un $\lambda = 0$ para el cual se marca que se usaron

todas las variables. Entre las nueve configuraciones de variables, se encuentran que los índices Gini para la eficiencia más altos son 0.2400, 0.2426 y 0.2629, y corresponden al Modelo_5, Modelo_6 y Modelo_7, respectivamente, en los cuales se logra obtener una reduccción en el porcentaje de DMUs eficientes por debajo al 28.5 %. Por el contrario, con la recomendación de la regla empírica el Modelo_6 y Modelo_8, presentan una relación apta entre el número de variables y la cantidad de DMUs.

Adicionalmente, cabe resaltar que con esta aproximación metodológica, la DMU clasificada como eficiente en la mayoría de modelos es la DMU3 (véase la Tabla 4.6), diferente a lo evidenciado con PCA-DEA donde era la DMU6, esto refuerza la importancia de la selección de variables en DEA con criterios objetivos, dada la sensibilidad de sus estimaciones. Conforme a lo anterior, en la Figura 4.2 se puede observar el diagrama de dispersión de acuerdo a las varianzas explicadas de los diferentes modelos, etiquetados con el λ usado en la penalización; aquí se resaltan en rojo el Modelo_5, Modelo_6 y Modelo_7 como los mejores candidatos, según los criterios anteriormente mencionados.

Tabla 4.6: Puntajes de eficiencia para los modelos con Lasso-DEA

Modelo	**Puntajes de eficiencia (DMU eficiente= 1, DMU ineficiente= (0,1))**						
	DMU1	**DMU2**	**DMU3**	**DMU4**	**DMU5**	**DMU6**	**DMU7**
Modelo_1	1.0000	1.0000	1.0000	0.8197	1.0000	1.0000	1.0000
Modelo_2	1.0000	1.0000	1.0000	0.8197	1.0000	1.0000	1.0000
Modelo_3	1.0000	1.0000	0.8267	0.8197	1.0000	1.0000	1.0000
Modelo_4	1.0000	1.0000	1.0000	0.8094	1.0000	1.0000	0.8421
Modelo_5	1.0000	0.6178	1.0000	0.2800	0.5082	0.8577	0.1759
Modelo_6	0.9333	0.6178	1.0000	0.2800	0.4853	0.8400	0.1581
Modelo_7	0.9333	0.6178	1.0000	0.2800	0.3148	0.8213	0.1581
Modelo_8	0.9333	1.0000	1.0000	0.8164	0.7663	1.0000	1.0000
Modelo_9	0.6721	0.9834	0.7201	0.8066	0.7558	0.9339	1.0000

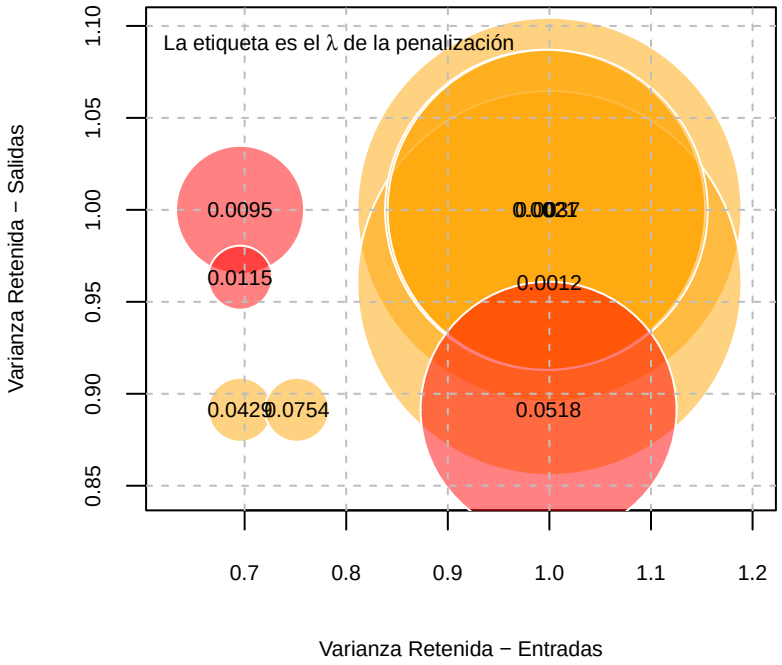

Figura 4.2: Cantidad de DMUs eficientes dada las varianzas retenidas (Lasso-DEA)

4.1.3. Metodología propuesta: Iterative Search Algorithm with DEA (ISA-DEA)

Tabla 4.7: Métricas de impureza y varianza retenida para los modelos con ISA-DEA

Modelo	α	Variables						Gini	Gini	Gini	DMUs	% DMUs	Varianza Retenida		
		I1	I2	I3	O1	O2	O3	Eficiencia	Ineficiencia	Ponderado	Eficientes	Eficientes	Entradas	Salidas	Total
$Modelo_1$	0	✓		✓	✓	✓	✓	0.0202	0.7500	0.7500	6	85.7 %	0.9992	1.0000	1.0000
$Modelo_2$	0.02	✓		✓	✓	✓		0.0202	0.7500	0.7354	6	85.7 %	0.9992	0.9602	0.9954
$Modelo_3$	0.46	✓		✓		✓		0.1636	0.5062	0.3486	3	42.9 %	0.9992	0.8998	0.9903
$Modelo_4$	0.49			✓		✓	✓	0.2661	0.4108	0.3399	2	28.6 %	0.6955	0.9565	0.9660
$Modelo_5$	0.62			✓			✓	0.2662	0.3470	0.2969	1	14.3 %	0.6955	0.8957	0.9553
$Modelo_6$	0.68			✓		✓		0.3201	0.2339	0.2925	1	14.3 %	0.6955	0.8998	0.9129
$Modelo_7$	0.93	✓				✓		0.1945	0.3240	0.2036	1	14.3 %	0.7515	0.8998	0.9359

En cuanto a los resultados de la propuesta metodológica presentada en el Capítulo 3, tras realizar la búsqueda iterativa de modelos orientada por la maximización del "Gini ponderado" ($Gini\ Ponderado = \alpha(Gini\ Eficiencia) + (1-\alpha)(Gini\ Ineficiencia)$), esto para diferentes ponderaciones, con valores de α entre $[0, 1]$, se identifican siete modelos con métricas únicas después de omitir los resultados para aquellos α donde no se presentó un cambio en los resultados. En la Tabla 4.7 se presentan los modelos sugeridos por este enfoque, para los cuales se logra apreciar que cuando el α es cercano a cero la cantidad DMUs eficientes es alta, mientras que para valores cercanos a uno, la discriminación se hace tan estricta que solo se tiene una DMU eficiente (tipo ranking,

lo cual no es foco de este trabajo). Las puntuaciones de eficiencia para cada modelo y cada DMU se exponen en la Tabla 4.8

Tabla 4.8: Puntajes de eficiencia para los modelos con ISA-DEA

Modelo	Puntajes de eficiencia (DMU eficiente= 1, DMU ineficiente= (0,1))						
	DMU1	DMU2	DMU3	DMU4	DMU5	DMU6	DMU7
$Modelo_1$	1.0000	1.0000	1.0000	0.8164	1.0000	1.0000	1.0000
$Modelo_2$	1.0000	1.0000	1.0000	0.8164	1.0000	1.0000	1.0000
$Modelo_3$	1.0000	0.7017	0.5551	0.2422	1.0000	0.7915	1.0000
$Modelo_4$	1.0000	0.5476	1.0000	0.1587	0.5082	0.8577	0.1759
$Modelo_5$	0.7933	0.5333	1.0000	0.1587	0.4853	0.8400	0.1509
$Modelo_6$	1.0000	0.3347	0.5551	0.0686	0.3314	0.5143	0.1514
$Modelo_7$	0.7521	0.5564	0.4175	0.2063	0.8309	0.6107	1.0000

Es necesario recalcar que el $Modelo_3$ y $Modelo_4$ con ponderaciones cercanas al promedio ($\alpha = 0{,}5$), logran reducir el exceso de DMUs eficientes a menos de la mitad, particularmente 3 y 2 DMUs eficientes respectivamente (de un total de 7 DMUs evaluadas). Además, la correspondencia entre las variables usadas y el total de DMUs analizadas para estos dos modelos es aceptable, según las reglas empíricas existentes y por ello, su configuración de variables, sugiere estos como los mejores candidatos a ser seleccionados para realizar estudios de eficiencia (se resaltan en rojo en la Figura 4.3).

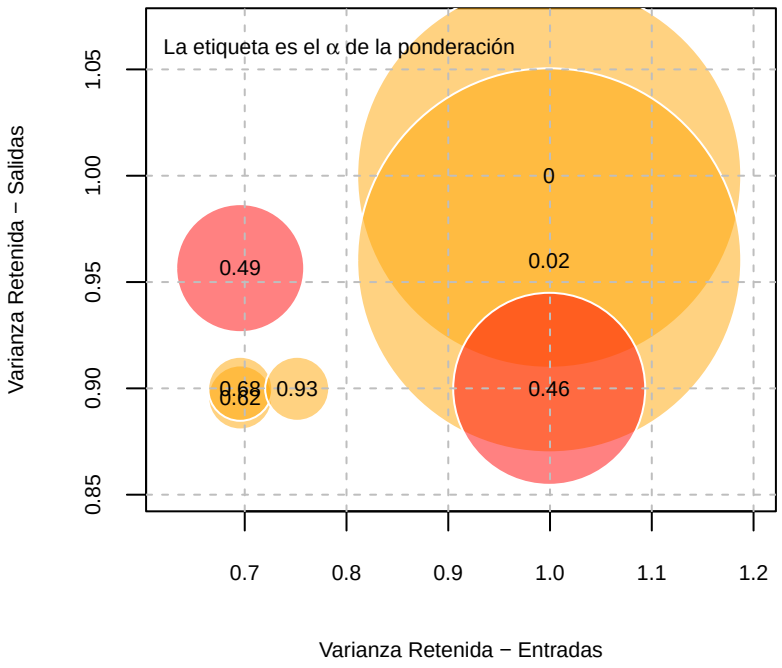

Figura 4.3: Cantidad de DMUs eficientes dada las varianzas retenidas (ISA-DEA)

4.1.4. Análisis comparativo

Resulta claro que cada metodología tiene sus bondades y desventajas, como por ejemplo el PCA que aprovecha todas las variables y las sintetiza por medio de combinaciones lineales no correlacionadas, de ahí que este enfoque combinado con DEA resulte una alternativa muy beneficiosa para la solución de la maldición de la dimensionalidad. Sin embargo, con respecto a las otras aproximaciones como Lasso-DEA e ISA-DEA, con PCA-DEA no se puede identificar de forma directa cuáles variables están impactando el modelo. La Tabla 4.10 y Tabla 4.11 en sus primeras columnas indican las variables usadas y omitidas en cada modelo, lo cual le permite al analista de información ver exploratoriamente que variables son más relevantes y contrastarlo con su experticia y conocimiento a cerca del sector estudiado. Por ejemplo, para estos modelos sugeridos, entre las seis variables utilizadas, las tres con mayor frecuencia de uso son I3, O1 y O2.

Como se ha mencionado y expuesto en todo el documento, la importancia de la selección de variables es crucial en DEA, y más bajo criterios objetivos, por ello para las tres metodologías en la Tabla 4.9, Tabla 4.10 y Tabla 4.11 se presentan solo aquellos modelos donde la métrica de impureza es máxima y los que son aptos bajo la mirada de las reglas empíricas. Para la última regla, donde se busca que la relación entre las variables y las unidades de evaluación este entre en el rango $[\frac{1}{3}, \frac{1}{2}]$, se incluyó la condición de que los modelos presenten como mínimo dos DMUs eficientes, porque el propósito es contar con varias unidades productoras como referentes más allá de hacer un ranking (aquí si interesa identificar la mejor del DMU entre el conjunto evaluado).

Tabla 4.9: Modelos sugeridos con PCA-DEA

Modelo	Componentes Principales		Gini	Gini	Gini	DMUs	% DMUs	Varianza Retenida		
	Entradas	Salidas	Eficiencia	Ineficiencia	Ponderado	Eficientes	Eficientes	Entradas	Salidas	Total
$Modelo_7$	1	3	0.1253	0.5431	0.3342	4	57.1 %	0.8679	1.0000	0.9839
$Modelo_8$	1	2	0.1204	0.4907	0.3056	2	28.6 %	0.8679	0.9755	0.8893

Tabla 4.10: Modelos sugeridos con Lasso-DEA

Modelo	λ	Varibles						Gini	Gini	Gini	DMUs	% DMUs	Varianza Retenida		
		I1	I2	I3	O1	O2	O3	Eficiencia	Ineficiencia	Ponderado	Eficientes	Eficientes	Entradas	Salidas	Total
$Modelo_7$	0.0429			✓	✓			0.2629	0.3773	0.3201	1	14.3 %	0.6955	0.8913	0.9435
$Modelo_8$	0.0518	✓		✓	✓			0.0435	0.5863	0.3149	4	57.1 %	0.9992	0.8913	0.9649

Tabla 4.11: Modelos sugeridos con ISA-DEA

Modelo	α	Variables						Gini	Gini	Gini	DMUs	% DMUs	Varianza Retenida		
		I1	I2	I3	O1	O2	O3	Eficiencia	Ineficiencia	Ponderado	Eficientes	Eficientes	Entradas	Salidas	Total
$Modelo_3$	0.46	✓		✓		✓		0.1636	0.5062 8	0.3486	3	42.9 %	0.9992	0.8998	0.9903
$Modelo_4$	0.49			✓		✓	✓	0.2661	0.4108	0.3399	2	28.6 %	0.6955	0.9565	0.9660
$Modelo_6$	0.68			✓		✓		0.3201	0.2339	0.2925	1	14.3 %	0.6955	0.8998	0.9129

Como resultado general, se evidencia que los modelos sugeridos por la metodología

propuesta (ISA-DEA) presentan las mejores valoraciones, considerando las métricas trasversales anteriormente mencionadas, e independientemente de las particularidades metodológicas de cada técnica, ya sea la sustitución de los datos por nuevas variables no correlacionadas o la identificación de unas cuantas variables con la mayor relación ante una variable independiente (en este caso la estimación de eficiencia con DEA). El Modelo_3 y Modelo_4 del enfoque ISA-DEA para valores de α de 0.46 y 0.49 respectivamente, obtienen los valores más altos para el Gini Ponderado y de varianza explicada (véase la Tabla 4.11), con los cuales se logra disminuir considerablemente el porcentaje de DMUs clasificadas como eficientes, usando la mitad de las variables iniciales.

En definitiva, el Modelo_3 con la aproximación ISA-DEA es el de mayor rendimiento, ya que reduce del 85.7 % al 42.9 % el número de DMUs clasificadas como eficientes, con tan solo la selección del 50 % de las variables iniciales. Este modelo usa dos insumos (I1: Empleados Docentes, e I3: Empleados Administrativos) y un producto (O2: Estudiantes de posgrado) en la estimación de eficiencia de las siete facultades universitarias (DMUs), variables con las cuales se explica el 99.03 % de la varianza total de los datos. Adicionalmente, este modelo con $\alpha = 0{,}46$ es el que reporta el Gini Ponderado más elevado con el menor número de DMUs eficientes, igual a tres, porque para los modelos de la Tabla 4.9 y Tabla 4.10 donde se maximiza esta métrica, la cantidad de unidades productoras eficientes es más alto (de cuatro), y con mayor perdida en la varianza retenida.

4.2. Evaluación de eficiencia del sector salud colombiano

Después de ratificar el rendimiento de la metodología propuesta al comprarse con otros enfoques comúnmente utilizados en la literatura, ahora se expone su aplicación con datos reales del sector salud colombiano. Se presentan los resultados expuestos en el Simposio Internacional de Estadística (Gómez Ardila, Villegas Ramírez, y Villegas Ramírez, 2021), como hallazgos parciales de este trabajo, donde se usó el parámetro $\alpha = 1$, y finalmente, se muestran y discuten los resultados con valores de α entre $[0, 1]$.

4.2.1. Descripción del caso de estudio

El sector salud, al igual que muchos otros, presenta múltiples variables con las cuales podría realizarse análisis de eficiencia, tales como variables financieras, operativas, o incluso de resultados clínicos. Con el propósito de evitar los sesgos generados por la maldición de la dimensionalidad o la elección arbitraria de variables, aplicamos la aproximación ISA-DEA a los mismos datos usados por Fontalvo Herrera (2017), quien analizó la eficiencia del sector salud Colombiano. El conjunto de datos está compuesto por indicadores financieros del 2011, reportados ante la Superintendencia de Salud para 17 entidades promotoras de salud (EPS) que hacen parte del régimen contributivo.

Aquí se aplicó, para la estimación de la frontera eficiente con DEA, la formulación "CCR-O", es decir, rendimientos constantes a escala y orientado a las salidas, donde la eficiencia técnica $= 1/\eta^*$, con η como variable objetivo del modelo de optimización y η^* como valor óptimo (véase la Ecuación 4.1). Para esta y otras formulaciones de DEA puede consultarse Cooper et al. (2007):

$$\begin{aligned} \underset{\eta,\mu}{\text{Maximizar}} \quad & \eta \\ \text{Sujeto a:} \quad & x_o \geq X\mu \\ & \eta y_o \leq Y\mu \\ & \mu \geq 0 \end{aligned} \tag{4.1}$$

En la Tabla 4.12 se resumen el valor de las variables de entradas y salidas. Vale aclarar que el estudio realizado por Fontalvo Herrera (2017) no cumple las reglas empíricas sugeridas en DEA, las cuales establecen, por ejemplo, que la cantidad de DMUs debe ser tres veces la cantidad total de las variables. Para este caso en particular se debieron haber evaluado como mínimo 21 DMUs dadas las 7 variables que fueron utilizadas en el modelo. Dichas variables están compuestas por cuatro variables de entradas: los activos totales (I1), las inversiones (I2), las cuentas por cobrar FOSYGA (I3), y los activos fijos (I4) como la propiedad, planta y equipo con los que cuentan las EPS; y tres variables de salidas, que son: los ingresos por la unidad de pago por captación de usuarios (O1), los ingresos por recobros al FOSYGA (O2), es decir, ingresos de cuentas por conceptos no incluidos por el plan básico de salud (PBS) pero autorizados por el comité técnico o por fallos de tutelas, y por último los ingresos operacionales (O3).

Tabla 4.12: EPS colombianas: Entradas y salidas de las 17 DMUs (cifras COP)

	DMU	**I 1**	**I 2**	**I 3**	**I 4**	**O 1**	**O 2**	**O 3**
DMU1	Aliansalud	74,659,289	19,712,425	43,573,089	506,726	196,384,549	81,430,596	193,387,216
DMU2	Comfenalco Valle	73,192,236	810,002	36,472,629	15,382,635	154,675,269	39,994,397	138,684,343
DMU3	Compensar	147,183,633	15,127,388	75,399,989	4,243,481	408,946,118	108,062,571	425,911,666
DMU4	Coomeva	552,127,504	12,757,631	403,192,739	12,549,111	1,412,872,026	384,858,018	1,464,245,760
DMU5	EPM	11,497,792	8,997,307	0	0	11,851,157	0	8,712,201
DMU6	Comfenalco Antioquia	55,003,905	4,500,000	43,468,103	1,489,166	177,047,201	50,863,177	177,934,783
DMU7	Sura	247,430,381	78,407,570	101,337,602	17,131,377	696,932,272	198,384,339	713,300,668
DMU8	Famisanar	254,676,457	21,000	176,767,704	8,508,024	669,227,215	124,616,077	666,839,628
DMU9	Fondo de Pasivo Ferrocarriles	39,370,561	16,297,183	3,739,882	229,932	54,289,171	0	104,784,504
DMU10	Golden Group	8,873,885	0	1,003,682	686,940	24,376,955	522,175	25,567,417
DMU11	Nueva EPS	1,058,955,212	3,191,139	725,870,151	2,942,152	1,903,254,182	788,690,698	1,811,061,005
DMU12	Colpatria	29,626,063	18,584,149	982,376	6,142	44,524,895	6,509,171	43,522,988
DMU13	Salud Total	365,396,570	198,238,511	75,315,458	35,242,132	787,461,401	109,833,981	817,409,281
DMU14	Saludvida	68,974,504	4,785,992	7,520,665	7,904,296	42,480,528	6,890,123	29,513,895
DMU15	Sanitas	379,950,832	25,376,938	242,987,341	3,000,112	589,479,187	227,092,530	582,135,254
DMU16	S.O.S.	118,171,938	1,319,225	86,320,782	3,688,966	376,769,004	89,411,705	359,358,087
DMU17	Solsalud	106,869,758	2,613,716	46,573,124	3,457,365	87,208,246	17,644,963	87,656,995

4.2.2. Resultados y Análisis

En la Tabla 4.13 se resumen los resultados de la aplicación de la metodología propuesta al caso de estudio, para $\alpha = 1$, lo que significa que el algoritmo guía su búsqueda exclusivamente con el Gini de la eficiencia. Al implementar el heurístico propuesto se observa que el primer modelo tiene un 70.59 % de las DMUs clasificadas como eficientes, usando el total de variables propuestas por Fontalvo Herrera (2017), modelo para el cual se obtienen sus mismos resultados (12 EPS eficientes). Para los modelos restantes, que fueron seleccionados por la metodología por tener el mayor índice Gini sobre los puntajes de eficiencia en cada iteración, se observa como la cantidad de variables de entrada o salida disminuyen al igual que la cantidad de DMUs clasificadas como eficientes. Cabe resaltar tras la ejecución del algoritmo ISA-DEA, que el número de DMUs ubicadas en la frontera se reducen del 70.6 % al 5.9 %, del Modelo_1 al Modelo_5 respectivamente, pasando de usar todas las variables, a tan solo usar el 34 % de las variables iniciales y reteniendo el 95.09 % de la varianza de los datos.

Tabla 4.13: Modelos seleccionados por el heurístico ISA-DEA con $\alpha = 1$

	I_1	I_2	I_3	I_4	O_1	O_2	O_3	Indice Gini	DMUs Eficientes	% DMUs Eficientes	Varianza Retenida		
											Entradas	Salidas	Total
Modelo_1	✓	✓	✓	✓	✓	✓	✓	0.0778	12	70.6 %	1.0000	1.0000	1.0000
Modelo_2	✓	✓		✓	✓	✓	✓	0.1284	9	52.9 %	0.9986	1.0000	0.9997
Modelo_3		✓		✓	✓	✓	✓	0.4452	5	29.4 %	0.5249	1.0000	0.9946
Modelo_4				✓	✓	✓	✓	0.7732	1	5.9 %	0.4380	1.0000	0.9543
Modelo_5				✓	✓	✓		0.7732	1	5.9 %	0.4380	0.9995	0.9509

Sin embargo, el Modelo_5 presenta una sola DMU eficiente, lo que se traduce en un solo referente entre los candidatos evaluados; este tipo de resultaddos no es el foco de este análisis comparativo, y además la varianza retenida de las entradas es muy baja (0.4380). En contraste, el Modelo_2 obtiene un 52.9 % de las DMUs eficientes y explica un 99.86 % y 100 % de la varianza por las entradas y salidas respectivamente (en conjunto un 99.97 %), y además, es el único cuya relación entre el número de variables y cantidad de DMUs es aceptable (bajo regla empírica), haciendo su configuración de variables la mejor selección bajo esta parametrización.

Tabla 4.14: Modelos seleccionados por el heurístico ISA-DEA con $0 \leq \alpha \leq 1$

Modelo	α	**Varibles**							**Gini**	**Gini**	**Gini**	**DMUs**	**% DMUs**	**Varianza Retenida**		
		I1	**I2**	**I3**	**I4**	**O1**	**O2**	**O3**	**Eficiencia**	**Ineficiencia**	**Ponderado**	**Eficientes**	**Eficientes**	**Entradas**	**Salidas**	**Total**
Modelo_1	0	✓	✓	✓	✓		✓	✓	0.0803	0.7788	0.7788	12	70.6 %	1.0000	0.9996	0.9999
Modelo_2	0.02	✓	✓	✓			✓	✓	0.0836	0.7633	0.7497	12	70.6 %	0.9279	0.9996	0.9773
Modelo_3	0.42	✓	✓	✓	✓		✓		0.2075	0.6749	0.4786	8	47.1 %	1.0000	0.9245	0.9939
Modelo_4	0.45	✓	✓	✓			✓		0.2075	0.6749	0.4645	8	47.1 %	0.9279	0.9245	0.9501
Modelo_5	0.59				✓		✓		0.7732	0.0815	0.4896	1	5.9 %	0.4380	0.9245	0.9350
Modelo_6	0.62		✓		✓	✓			0.4625	0.3465	0.4184	5	29.4 %	0.5249	0.9659	0.9803
Modelo_7	0.69				✓	✓	✓		0.7732	0.0815	0.5588	1	5.9 %	0.4380	0.9995	0.9509

En segunda instancia, se decidió ejecutar el algoritmo de búsqueda iterativa, orientado por la ponderación entre los puntajes de eficiencia e ineficiencia, mediante la

modificación del parámetro α entre cero y uno, esto con el objetivo de suavizar la métrica y proporcionar mayor balance entre estas dos categorías. Realizando incrementos del 0.01 en el α, se tendrían 101 parámetros (α={0, 0.01, 0.02, … , 0.98, 0.99, 1}), esto se traduce en 101 ejecuciones del algoritmo ISA-DEA. El último modelo sugerido en las rutas exploradas para cada valor de α (el cual reune el mayor grado de discriminación) se consolida en la Tabla 4.14 (aquí se ocultan los modelos duplicados).

Basta con observar la cantidad de variables usadas en los siete modelos de la Tabla 4.14, para evidenciar que la regularización del Gini Ponderado permite una exploración más diversificada, ya que se alcanzan modelos con dos variables, que en los resultados para $\alpha = 1$ no se generaron debido a la rápida convergencia. De igual forma, los modelos aquí sugeridos indican que el algoritmo tomo otras rutas en el espacio de solución porque hay modelos con nuevas cantidades de DMUs eficientes, lo que enriquece el conjunto de candidatos para la selección del modelo final. Los puntajes de eficiencia de los siete modelos para cada una de las DMUs se presentan en la Tabla 4.15.

Tabla 4.15: Puntajes de eficiencia para los modelos con ISA-DEA y $0 \leq \alpha \leq 1$

	Puntajes de eficiencia (DMU eficiente= 1, DMU ineficiente= (0,1))						
	Modelo$_1$	**Modelo$_2$**	**Modelo$_3$**	**Modelo$_4$**	**Modelo$_5$**	**Modelo$_6$**	**Modelo$_7$**
DMU 1	1.000	1.000	1.000	1.000	0.152	0.531	0.152
DMU 2	1.000	1.000	0.981	0.981	0.003	0.096	0.003
DMU 3	1.000	1.000	1.000	1.000	0.024	0.148	0.024
DMU 4	0.975	0.894	0.882	0.882	0.029	0.184	0.029
DMU 5	1.000	1.000	0.000	0.000	0.000	1.000	0.000
DMU 6	1.000	1.000	1.000	1.000	0.032	0.183	0.032
DMU 7	1.000	1.000	1.000	1.000	0.011	0.062	0.011
DMU 8	1.000	1.000	1.000	1.000	0.014	1.000	0.014
DMU 9	1.000	1.000	0.000	0.000	0.000	0.295	0.000
DMU 10	1.000	1.000	1.000	1.000	0.001	1.000	0.001
DMU 11	1.000	1.000	1.000	1.000	0.253	1.000	0.253
DMU 12	1.000	1.000	1.000	1.000	1.000	1.000	1.000
DMU 13	0.848	0.848	0.609	0.609	0.003	0.034	0.003
DMU 14	0.502	0.502	0.477	0.477	0.001	0.014	0.001
DMU 15	0.741	0.741	0.741	0.741	0.071	0.296	0.071
DMU 16	1.000	1.000	0.991	0.991	0.023	0.385	0.023
DMU 17	0.346	0.338	0.322	0.322	0.005	0.053	0.005

Finalmente, como en los anteriores análisis, se sugiere el Modelo$_6$ (con $\alpha = 0{,}62$) por el criterio extrapolado de las reglas empíricas. Este utiliza solo el 43 % de las variables (O1, I2, I4) identificando como eficientes al 29.4 % de las DMUS evaluadas y reteniendo el 98.03 % de la varianza total de los datos. Este se observa marcado con rojo en la Figura 4.4 donde se puede comparar con los demás modelos según la proporción de varianza explicada de sus entradas y salidas, etiquetados con el valor del α, y la cantidad de unidades eficientes representada en el tamaño del circulo.

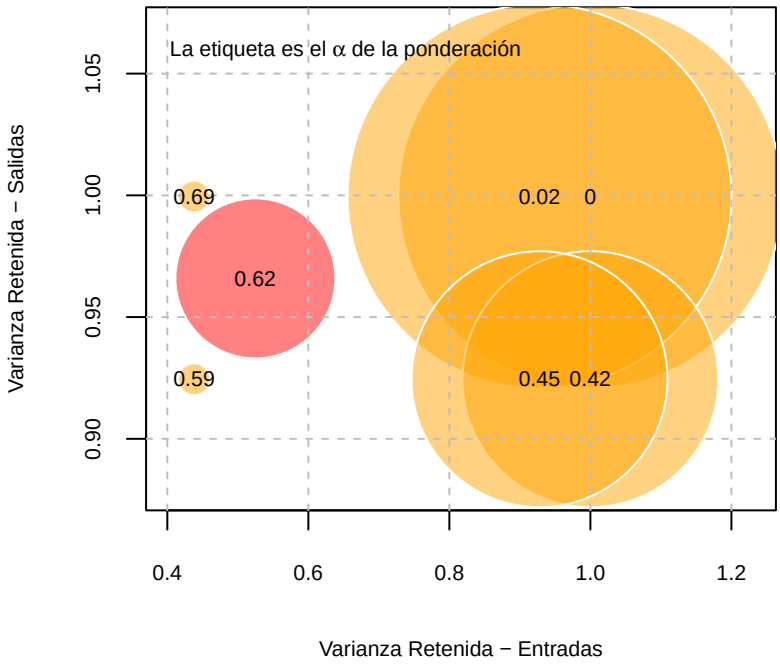

Figura 4.4: Modelos ISA-DEA ponderado: Varianza retenida versus DMUs eficinetes

Capítulo 5

CONCLUSIONES

En este trabajo se ilustra el problema de dimensionalidad presente en DEA ante la presencia de una gran cantidad de variables de entrada y salida, y un bajo número DMUs, donde es alta la probabilidad de clasificar una DMU como eficiente cuando no lo es. El método heurístico de selección de variables propuesto para el incremento del poder discriminatorio de DEA permite generar y analizar diversas configuraciones de variables, en las cuales tanto el número de DMUs eficientes como la varianza retenida por las variables usadas son importantes. De este modo, se obtienen mejores soluciones que las que se alcanzarían cuando no se tiene en cuenta la dimensionalidad.

Para la selección de variables se implementó un algoritmo de búsqueda iterativo, donde se van eliminando variables, tanto de entrada como salida, explorando soluciones que maximicen la diversidad entre los puntajes de eficiencia. Este algoritmo puede ejecutarse maximizando solo el Gini de eficiencia, cuando el parámetro de $\alpha = 1$, o también variando dicho parámetro entre cero y uno ponderando las categorías del análisis (eficiencia e ineficiencia); bajo ambos escenarios se obtienen buenas aproximaciones de la frontera, sin exceso de unidades clasificadas como eficientes.

Al compararse el desempeño de la aproximación propuesta (ISA-DEA) usando datos de la literatura para ilustrar el rendimiento frente a otros enfoques, se observan mejores resultados con esta que con las otras aproximaciones que combinan DEA con técnicas estadísticas, como el análisis de componentes principales (PCA-DEA) y la regresión Lasso (Lasso-DEA). Se debe agregar que ISA-DEA permite evidenciar de forma directa las variables utilizadas en cada modelo, las métricas de desigualdad para las estimaciones de eficiencia y la proporción de varianza explicada para las entradas, las salidas y el conjunto de datos total, lo cual resulta especialmente útil para que los tomadores de decisiones eviten la aplicación de criterios ad hoc o juicios particulares en la elección de las variables, proporcionando un conjunto de modelos seleccionados sistematicamente por índices estadísticos (conocidos como ínidices de impureza o desigualdad).

Por otra parte, este método se aplicó al sector de salud, usando datos financieros para medir la eficiencia de EPS colombianas, y se evidencian reducciones considerables en en el número de DMUs clasificadas como eficientes, empleando menos del 50 % de las variables iniciales y alcanzando retenciones superiores al 89 % de la varianzas total de los datos en varios de los modelos sugeridos.

Este trabajo deja abierta diferentes opciones para realizar extensiones, entre ellas, que la metodología presentada pueda robustecerse mediante la aleatorización en la selección de los modelos, que serán los puntos de partida en cada iteración, ya que permite diversificar el espectro de solución. Esta noción es muy común en las propuestas heurísticas y se ha evidenciado que tiene buenos resultados en la práctica. Adicionalmente, puede ponerse a prueba el desempeño de esta metodología, usando datos simulados frente a diferentes tecnologías de producción y enfoques para el análisis de fronteras.

Bibliografía

Adler, N., y Golany, B. (2001). Evaluation of deregulated airline networks using data envelopment analysis combined with principal component analysis with an application to western europe. *European Journal of Operational Research*, *132*(2), 260–273.

Adler, N., y Golany, B. (2002). Including principal component weights to improve discrimination in data envelopment analysis. *Journal of the Operational Research Society*, *53*(9), 985–991.

Adler, N., y Volta, N. (2019). Ranking methods within data envelopment analysis. En *The palgrave handbook of economic performance analysis* (pp. 189–224). Springer.

Adler, N., y Yazhemsky, E. (2010). Improving discrimination in data envelopment analysis: Pca–dea or variable reduction. *European Journal of Operational Research*, *202*(1), 273–284.

Álvarez, A. (2001). *La medición de la eficiencia y la productividad*. Ediciones Pirámide.

Bogetoft, P., y Otto, L. (2020). Benchmarking with dea and sfa [Manual de software informático]. (R package version 0.29)

Charles, V., Aparicio, J., y 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., y Rhodes, E. (1978). Measuring the efficiency of decision making units. *European journal of operational research*, *2*(6), 429–444.

Charnes, A., Cooper, W. W., y Thrall, R. M. (1986). Classifying and characterizing efficiencies and inefficiencies in data development analysis. *Operations Research Letters*, *5*(3), 105–110.

Chen, Y., Tsionas, M., y Zelenyuk, V. (2021). Lasso+ dea for small and big wide data. *Omega*, 102419.

Chuang, C.-L., Chang, P.-C., y Lin, R.-H. (2011). An efficiency data envelopment analysis model reinforced by classification and regression tree for hospital performance evaluation. *Journal of medical systems*, *35*(5), 1075–1083.

Cook, W. D., Tone, K., y Zhu, J. (2014). Data envelopment analysis: Prior to choosing a model. *Omega*, *44*, 1–4.

Cooper, W. W., Seiford, L. M., y Tone, K. (2007). *Data envelopment analysis: a comprehensive text with models, applications, references and dea-solver software*

(Vol. 2). Springer.

Dyson, R. G., Allen, R., Camanho, A. S., Podinovski, V. V., Sarrico, C. S., y Shale, E. A. (2001). Pitfalls and protocols in dea. *European Journal of operational research*, *132*(2), 245–259.

Fernandez-Palacin, F., Lopez-Sanchez, M. A., y Muñoz-Márquez, M. (2018). Stepwise selection of variables in dea using contribution loads. *Pesquisa Operacional*, *38*(1), 31–52.

Fontalvo Herrera, T. J. (2017). Eficiencia de las entidades prestadoras de salud (eps) en colombia por medio de análisis envolvente de datos. *Ingeniare. Revista chilena de ingeniería*, *25*(4), 681–692.

Friedman, L., y Sinuany-Stern, Z. (1998). Combining ranking scales and selecting variables in the dea context: The case of industrial branches. *Computers & Operations Research*, *25*(9), 781–791.

Gómez Ardila, A., Villegas Ramírez, A., y Villegas Ramírez, J. (2021). Método heurístico de selección de variables para el incremento del poder discriminatorio de dea: caso de aplicación a las eps colombianas. *XXX Simposio Internacional de Estadística 2021 - Evento virtual*, 183–190.

Golany, B., y Roll, Y. (1989). An application procedure for dea. *Omega*, *17*(3), 237–250.

James, G., Witten, D., Hastie, T., y Tibshirani, R. (2013). *An introduction to statistical learning* (Vol. 112). Springer.

Jenkins, L., y Anderson, M. (2003). A multivariate statistical approach to reducing the number of variables in data envelopment analysis. *European Journal of Operational Research*, *147*(1), 51–61.

Kohl, S., Schoenfelder, J., Fügener, A., y Brunner, J. O. (2018). The use of Data Envelopment Analysis (DEA) in healthcare with a focus on hospitals. *Health care management science*, 1–42.

Kuosmanen, T. (2006). Stochastic nonparametric envelopment of data: combining virtues of sfa and dea in a unified framework.

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

Lee, C.-Y., y Cai, J.-Y. (2020). Lasso variable selection in data envelopment analysis with small datasets. *Omega*, *91*, 102019.

Liu, J. S., Lu, L. Y., y Lu, W.-M. (2016). Research fronts in data envelopment analysis. *Omega*, *58*, 33–45.

Liu, J. S., Lu, L. Y., Lu, W.-M., y Lin, B. J. (2013). A survey of dea applications. *Omega*, *41*(5), 893–902.

Medina, F. (2001). *Consideraciones sobre el índice de gini para medir la concentración del ingreso*. Cepal.

Mejía, A. M. (2008). Evaluación económica de programas y servicios de salud. *Revista Gerencia y Políticas de Salud*, *7*(15), 91–113.

O'neill, Liam. (1998). Multifactor efficiency in data envelopment analysis with an application to urban hospitals. *Health Care Management Science*, *1*(1), 19–27.

Ozcan, Y. A. (2014). Evaluation of Performance in Health Care. En *Health Care Benchmarking and Performance Evaluation: An Assessment using Data Envelopment Analysis (DEA)* (pp. 3–14). Boston, MA: Springer US.

R Core Team. (2021). R: A language and environment for statistical computing [Manual de software informático]. Vienna, Austria. Descargado de https://www.R-project.org/

Restrepo, M. I., y Villegas, J. G. (2014). Clasificación de grupos de investigación colombianos aplicando análisis envolvente de datos. *Revista Facultad de Ingeniería*(42), 105–119.

Sarkis, J. (2000). A comparative analysis of dea as a discrete alternative multiple criteria decision tool. *European journal of operational research*, *123*(3), 543–557.

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society: Series B (Methodological)*, *58*(1), 267–288.

Tsionas, E. G., y Papadakis, E. N. (2010). A Bayesian approach to statistical inference in stochastic DEA. *Omega*, *38*(5), 309–314.

Villanueva-Cantillo, J., y Munoz-Marquez, M. (2021). Methodology for calculating critical values of relevance measures in variable selection methods in data envelopment analysis. *European Journal of Operational Research*, *290*(2), 657–670.

Wong, Y.-H., y Beasley, J. (1990). Restricting weight flexibility in data envelopment analysis. *Journal of the Operational Research Society*, *41*(9), 829–835.

Zhu, J. (2014). *Quantitative Models for Performance Evaluation and Benchmarkings*. Springer.

Apéndice A

Implementación en R

I. Programa en R: Combinación de variables

```
combi_1vm <- function(n){
  library(gtools)
  g <- n-1
  y <- c(1:n)
  y_com <- combinations(n, g, y)
  return(y_com)
}
```

II. Programa en R: Covarianza parcial

```
parcor<-function(data, var_include){
  x<-data

  vt<-dim(x)[2]
  vars<-seq(1:vt)
  (p_include<- var_include)
  (p_exclude<- vars[-p_include])

  x_reorder<-x_standardized<-x[,c(p_exclude,p_include)]
  x_reorder
  for (i in 1:dim(x)[2]){
    x_standardized[,i] <- (x_reorder[,i] - mean(x_reorder[,i]))/sd(x_reorder[,i])}
  x_standardized
  varcov<-cor(x_standardized)

  le<-length(p_exclude)
  v11<-varcov[1:le,1:le]
  v12<-varcov[1:le,(le+1):vt]
  v21<-varcov[(le+1):vt,1:le]
  v22<-varcov[(le+1):vt,(le+1):vt]

  vexp<-vt-sum(diag(v11-(v12 %*% solve(v22) %*% v21)))
  vexp
  var_exp<-vexp/vt
  var_exp
```

```
  return(variance_explained=var_exp)
}
```

III. Programa en R: Índice Gini para puntajes de eficiencia

```
gini_eff<-function(vector){
  if(sum(vector)==0){gini_s<-round(0,7)}else{
    vector<-vector[order(vector)]
    aux<-rep(1,1,length(vector))
    num<-sum((cumsum(aux)/max(cumsum(aux)))-(cumsum(vector)/max(cumsum(vector))))
    deno<-sum(cumsum(aux)/max(cumsum(aux)))
    gini_s<-round(num/deno,7)}
  return(gini_s)
}
```

IV. Programa en R: ISA-DEA (Iterative Search Algorithm)

```
isa_dea_gini_pond<-function(input_data, output_data, performance_metric, stop_criterion,
                            epsilon=0.00001, alpha=0){
  library(Benchmarking)
  # library(glmnet)
  library(dplyr)
  inputs<-input_data
  outputs<-output_data
  criterio<- performance_metric
  param_rts="crs"
  param_orient="in"
  alfa<-alpha
  # umbral<-stop_criterion

  # inputs<-x
  # outputs<-y
  # criterio<-6
  # stop_criterion<-1
  # epsilon=0.0001
  # alfa<-0.62

  colnames(inputs)<-paste(rep("I",dim(inputs)[2]),seq(1,dim(inputs)[2]), sep = "")
  rownames(inputs)<-paste(rep("DMU",dim(inputs)[1]),seq(1,dim(inputs)[1]), sep = "")

  colnames(outputs)<-paste(rep("O",dim(outputs)[2]),seq(1,dim(outputs)[2]), sep = "")
  rownames(outputs)<-paste(rep("DMU",dim(outputs)[1]),seq(1,dim(outputs)[1]), sep = "")

  input_data=inputs
  output_data=outputs

  inputnames<-colnames(inputs)
  outputnames<-colnames(outputs)

  namesDMU<-paste(rep("DMU",dim(inputs)[1]),seq(1,dim(inputs)[1]), sep = "")
  dmu<-dim(inputs)[1]
  m<-dim(inputs)[2]
  s<-dim(outputs)[2]
```

```
sol<-NULL
map<-NULL
tabla<-NULL
valor<-m+s
if(stop_criterion==1){umbral<-epsilon}else{
  if(stop_criterion==2){umbral<-2}else{
    if(stop_criterion==3){umbral<-trunc(dmu/3)}
  }
}
# criterio<-4 #1: entropy_bi; 2: gini_bi; 3: entropy_eff; 4: gini_eff;
ite<-1


aux<-round(Benchmarking::dea(X=inputs,Y=outputs, RTS=param_rts,
                             ORIENTATION=param_orient)$eff, digits = 4)
eff<-length(which(aux==1));neff<-length(which(aux!=1));n<-length(aux)
vo<-ifelse(criterio==1,-(eff/n)*log2(eff/n) -(neff/n)*log2(neff/n),
ifelse(criterio==2,(eff/n)*(1-(eff/n)) + (neff/n)*(1-(neff/n)),
ifelse(criterio==3,round(sum((-1/length(aux))*log2(aux)),7),
ifelse(criterio==4,gini_eff(aux),
ifelse(criterio==5,(gini_eff(aux)+gini_eff(1-aux))/2,
ifelse(criterio==6,alfa*gini_eff(aux)+(1-alfa)*gini_eff(1-aux),NA))))))

var_exp_m<-1
var_exp_s<-1
var_exp_total<-1
ge<-gini_eff(aux)
g1e<-gini_eff(1-aux)
ruta_var<-c(colnames(outputs),colnames(inputs),vo,0,ge,g1e,eff,
            var_exp_m,var_exp_s,var_exp_total,aux)
vartotal<-m+s
if(valor > umbral){
  while(valor > umbral) {
    if(m==1){dimcombi_m=0;comb_int<-m}else{dimcombi_m=dim(combi_1vm(m))[1]
    comb_int<-combi_1vm(m)}
    if(s==1){dimcombi_s=0;comb_out<-s}else{dimcombi_s=dim(combi_1vm(s))[1]
    comb_out<-combi_1vm(s)}
    ncols=12+dmu
    rownames(input_data)
    var<-matrix(rep(0,ncols*(dimcombi_m+dimcombi_s)), ncol=ncols)
    colnames(var)<-c("m","s","dmu_eff","entropy_bi","gini_bi","entropy_eff",
                     "gini_eff","gini_eff_prom","gini_eff_pond","gini_1-eff",
                     "Iteracion","max_local",namesDMU)

    #para las entradas
    if(m>1){
      for(i in 1:dim(combi_1vm(m))[1]){
        # i=1
        aux<-round(Benchmarking::dea(X=inputs[,combi_1vm(m)[i,]],Y=outputs,
              RTS=param_rts, ORIENTATION=param_orient)$eff, digits = 4)

        if(length(combi_1vm(m)[i,])==1){var[i,1]<-mm<-1}else{
            var[i,1]<-mm<-dim(inputs[,combi_1vm(m)[i,]])[2]}
        if(s==1){var[i,2]<-ms<-1}else{var[i,2]<-ms<-dim(outputs)[2]}
        var[i,3]<-eff<-length(which(aux==1))
        neff<-length(which(aux!=1))
        n<-length(aux)
        var[i,4]<-entropy<- -(eff/n)*log2(eff/n) -(neff/n)*log2(neff/n)
        var[i,5]<-gini<- (eff/n)*(1-(eff/n)) + (neff/n)*(1-(neff/n))
        var[i,6]<-round(sum((-1/length(aux))*log2(aux)),7)
        var[i,7]<-gini_eff(aux)
```

```
      var[i,8]<-(gini_eff(aux)+gini_eff(1-aux))/2
      var[i,9]<-alfa*gini_eff(aux)+(1-alfa)*gini_eff(1-aux)
      var[i,10]<-gini_eff(1-aux)
      var[i,11]<-ite
      var[i,13:ncols]<-aux
      aux
    }
  }
  var
  #para las salidas
  if(s>1){
    for(i in 1:dim(combi_1vm(s))[1]){
      # i=1
      aux<-round(Benchmarking::dea(X=inputs,Y=outputs[,combi_1vm(s)[i,]],
                 RTS=param_rts, ORIENTATION=param_orient)$eff, digits = 4)
      if(m==1){var[dimcombi_m+i,1]<-sm<-1}else{var[dimcombi_m+i,1]<-sm<-dim(inputs)[2]}
      if(length(combi_1vm(s)[i,])==1){var[dimcombi_m+i,2]<-ss<-1}else{
        var[dimcombi_m+i,2]<-ss<-dim(outputs[,combi_1vm(s)[i,]])[2]}
      var[dimcombi_m+i,3]<-eff<-length(which(aux==1))
      neff<-length(which(aux!=1))
      n<-length(aux)
      var[dimcombi_m+i,4]<-entropy<- -(eff/n)*log2(eff/n) -(neff/n)*log2(neff/n)
      var[dimcombi_m+i,5]<-gini<- (eff/n)*(1-(eff/n)) + (neff/n)*(1-(neff/n))
      var[dimcombi_m+i,6]<-round(sum((-1/length(aux))*log2(aux)),7)
      var[dimcombi_m+i,7]<-gini_eff(aux)
      var[dimcombi_m+i,8]<-(gini_eff(aux)+gini_eff(1-aux))/2
      var[dimcombi_m+i,9]<-alfa*gini_eff(aux)+(1-alfa)*gini_eff(1-aux)
      var[dimcombi_m+i,10]<-gini_eff(1-aux)
      var[dimcombi_m+i,11]<-ite
      var[dimcombi_m+i,13:ncols]<-aux
      # var[dimcombi_m+i,10]<-parcor(data=outputs, var_include=combi_1vm(s)[i,])
    }
  }
  var
  pos<-3+criterio

  # if(identical(which.max(var[,pos]), integer(0))){ban_inp=1}else{ban_inp=0}
  # if(identical(which.max(var[,pos]), integer(0))){ban_out=1}else{ban_out=0}


  if(identical(which.max(var[,pos]), integer(0))){valor<-epsilon}else{

    if(which.max(var[,pos])<=dimcombi_m){
      aux<-which.max(var[,pos])
      var[which.max(var[,pos]),]
      vincf<-NULL; for(i in 1:length(inputnames[combi_1vm(m)[aux,]])){
        vinc<-which(colnames(input_data)==inputnames[combi_1vm(m)[aux,]][i])
        vincf<-cbind(vincf,vinc)}
      var_exp_m<-round(parcor(data=input_data, var_include=vincf),7)
      inputs<-inputs[,combi_1vm(m)[aux,]]
      if(is.null(dim(inputs)[2])){m<-1;var_in<-inputnames[aux]}else{m<-dim(inputs)[2]}
    }else{aux<-which.max(var[,pos])-dimcombi_m
    var[which.max(var[,pos]),]
    vincf<-NULL; for(i in 1:length(outputnames[combi_1vm(s)[aux,]])){
      vinc<-which(colnames(output_data)==outputnames[combi_1vm(s)[aux,]][i])
      vincf<-cbind(vincf,vinc)}
    var_exp_s<-round(parcor(data=output_data, var_include=vincf),7)
    outputs<-outputs[,combi_1vm(s)[aux,]]
    if(is.null(dim(outputs)[2])){s<-1;var_out<-outputnames[aux]}else{s<-dim(outputs)[2]}
    }
```

```
      if(ite==1){
        delta<-round((var[which.max(var[,pos]),pos]/vo)-1,7)
        anterior<-var[which.max(var[,pos]),pos]
        if(m==1){var_in<-inputnames[1]}
        if(s==1){var_out<-inputnames[1]}

        }else{
          delta<-round((var[which.max(var[,pos]),pos]/anterior)-1,7)
          anterior<-var[which.max(var[,pos]),pos]}


      if(m==1){inputnames<- var_in}else{inputnames<- colnames(inputs)}
      if(s==1){outputnames<- var_out}else{outputnames<-colnames(outputs)}

      sol<-list(sol,Solucion=ite, Variables=c(outputnames,inputnames))
      var[which.max(var[,pos]),12]<-1
      if(ite==1){tabla<-var}else{tabla<-rbind(tabla,var)}



      aux_pyf<-NULL; for(i in 1:length(outputnames)){
        aux_py<-which(colnames(output_data)==outputnames[i])
        aux_pyf<-cbind(aux_pyf,aux_py)}
      aux_pxf<-NULL; for(i in 1:length(inputnames)){
        aux_px<-which(colnames(input_data)==inputnames[i])
        aux_pxf<-cbind(aux_pxf,aux_px)}
      aux_data_exclu<-cbind(input_data[,-aux_pxf], output_data[,-aux_pyf])
      aux_data<-cbind(aux_data_exclu,cbind(input_data[,aux_pxf], output_data[,aux_pyf]))
      var_exp_total<-round(parcor(data=aux_data,
                  var_include=(dim(aux_data_exclu)[2]+1):dim(aux_data)[2]),7)

      ruta_var<-rbind(ruta_var,
                      c(outputnames,
                        inputnames,
                        rep("",vartotal-length(outputnames)-length(inputnames)),
                        var[which.max(var[,pos]),pos],
                        delta,
                        var[which.max(var[,pos]),7], #gini_eff
                        var[which.max(var[,pos]),10], #gini_1-eff
                        var[which.max(var[,pos]),3],
                        var_exp_m,
                        var_exp_s,
                        var_exp_total,
                        var[which.max(var[,pos]),13:ncols]))
      # t(ruta_var)

      sol
      map<-list(map,Iteracion=ite, mapa=var)#paste("Iteracion",ite, sep=" ")
      map
      if(m>=1 | s>=1){valor<-m+s}else{valor<-epsilon}
      if(stop_criterion==1){if(delta<epsilon){valor<-epsilon}else{
        if(m+s<umbral){valor<-epsilon}}}
      if(m==1 & s==1){valor<-umbral}
      ite<-ite+1
    }

  }
  sol
  map
  tabla
  colnames(ruta_var)[vartotal+1]<-"Metrica"
```

```
    colnames(ruta_var)[vartotal+2]<-"Delta"
    colnames(ruta_var)[vartotal+3]<-"Gini_Eff"
    colnames(ruta_var)[vartotal+4]<-"Gini_1-Eff"
    colnames(ruta_var)[vartotal+5]<-"DMUs_Eff"
    colnames(ruta_var)[vartotal+6]<-"VarExp_In"
    colnames(ruta_var)[vartotal+7]<-"VarExp_Out"
    colnames(ruta_var)[vartotal+8]<-"VarExp_Model"
    # t(ruta_var)
    ruta_var

    return(list(Proceso=tabla,Ruta=ruta_var))
  }else{print(paste("El número de variables (",valor,") es inferior al umbral
                    (",umbral,") especificado en el criterio de parada"))}
}
```

V. Programa en R: ISA-DEA con métrica ponderada

```
res_alpha<-NULL
valpha<-seq(0,1,0.01)
for(i in 1:length(valpha)){
  # i=43
  print(valpha[i])
  res<-isa_dea_gini_pond(input_data=x, output_data=y, performance_metric=6, stop_criterion=1, epsilon=0.00001, alpha=valp
  nmod<-dim(res[[2]])[1]
  res_alpha<-rbind(res_alpha,c(vlr_alpha=valpha[i],res[[2]][nmod,]))
}
```

VI. Programa en R: Lasso-DEA (Regression Lasso)

```
Lasso_DEA<-function(input_data, output_data, orientation=c("in","out"), method=1){
  library(Benchmarking)
  library(glmnet)
  library(dplyr)

  colnames(input_data)<-paste(rep("I",dim(input_data)[2]),seq(1,dim(input_data)[2]), sep="")
  rownames(input_data)<-paste(rep("DMU",dim(input_data)[1]),seq(1,dim(input_data)[1]), sep="")

  colnames(output_data)<-paste(rep("O",dim(output_data)[2]),seq(1,dim(output_data)[2]), sep="")
  rownames(output_data)<-paste(rep("DMU",dim(output_data)[1]),seq(1,dim(output_data)[1]), sep="")
  datos<-cbind(input_data,output_data)

  dea_full<-Benchmarking::dea(X=input_data,Y=output_data, RTS="crs", ORIENTATION="in")
  scores<-round(dea_full$eff, digits = 4)
  scores

  xx<- datos
  yy<- scores

  # Iterar con diferentes lambdas
  lambdas_to_try <- c(0,10^seq(-3, 5, length.out = 99))
  round(lambdas_to_try,0)

  map<-NULL
  for(i in 1:length(lambdas_to_try)){

    lambdas_to_try[i]
```

```
lasso_lambdas <- glmnet(x=xx, y=yy, alpha = 1,
        lambda=lambdas_to_try[i], standardize = T, intercept = F)
n<-dim(xx)[2]+1

coef<-as.array(coef(lasso_lambdas))[-1]
nx<-dim(input_data)[2]
ny<-dim(output_data)[2]

pos_coef<-which(coef!=0)
pos_xcoef<-which(coef[1:nx]!=0)
pos_ycoef<-which(coef[(nx+1):n]!=0)
i;pos_coef
if(length(pos_coef)==0 | length(pos_xcoef)==0 | length(pos_ycoef)==0){
  res<-c(itera=i,
         lambda=round(lambdas_to_try[i],4),
         inp_sel=NULL,
         out_sel=NULL,
         var_exp_inp=NULL,
         var_exp_out=NULL,
         var_exp_model=NULL,
         dmu_eff=NULL,
         entropy_bi=NULL,
         gini_bi=NULL,
         entropy_eff=NULL,
         gini_eff=NULL,
         gini_eff_prom=NULL,
         NULL,
         NULL,
         NULL,
         NULL
  )
}else{


  if(length(pos_xcoef)==1){
    x_lasso<-matrix(input_data[,pos_xcoef], ncol=1)
    colnames(x_lasso) <- colnames(input_data)[pos_xcoef]}else{
      x_lasso<-input_data[,pos_xcoef]}
  if(length(pos_ycoef)==1){
    y_lasso<-matrix(output_data[,pos_ycoef], ncol=1)
    colnames(y_lasso) <- colnames(output_data)[pos_ycoef]}else{
      y_lasso<-output_data[,pos_ycoef]}


  data_lasso<-cbind(x_lasso,y_lasso)
  dea_lasso<-Benchmarking::dea(X=x_lasso,Y=y_lasso, RTS="crs", ORIENTATION="in")
  scores_lasso<-round(dea_lasso$eff, digits = 6)

  eff<-length(which(scores_lasso==1))
  eff
  neff<-length(which(round(scores_lasso,6)!=1))
  neff
  n_dmu<-length(scores_lasso)

  criterio1<- -(eff/n_dmu)*log2(eff/n_dmu) -(neff/n_dmu)*log2(neff/n_dmu) #ok
  criterio2<- (eff/n_dmu)*(1-(eff/n_dmu)) + (neff/n_dmu)*(1-(neff/n_dmu))
  criterio3<- round(sum((-1/n_dmu)*log2(round(scores_lasso,6))),7)
  criterio4<- gini_eff(scores_lasso) # ok
  criterio5<- (gini_eff(scores_lasso)+gini_eff(1-scores_lasso))/2
  metrica_aux<- gini_eff(1-scores_lasso)
  varnames<-c(colnames(x_lasso), colnames(y_lasso),
```

```
                rep("",dim(xx)[2]-dim(x_lasso)[2]-dim(y_lasso)[2]))
      i;varnames
      # n_v_inp<- length(colnames(x_lasso))

      var_exp_x<-parcor(data=input_data, var_include=pos_xcoef)
      var_exp_y<-parcor(data=output_data, var_include=pos_ycoef)
      if(nx+ny==length(pos_coef)){var_exp_total<-1}else{
        var_exp_total<-parcor(data=xx, var_include=pos_coef)}


      res<-c(itera=i,
             lambda=round(lambdas_to_try[i],4),
             inp_sel=length(pos_xcoef),
             out_sel=length(pos_ycoef),
             var_exp_inp=round(var_exp_x,4),
             var_exp_out=round(var_exp_y,4),
             var_exp_model=round(var_exp_total,4),
             dmu_eff=eff,
             entropy_bi=round(criterio1,4),
             gini_bi=round(criterio2,4),
             entropy_eff=round(criterio3,4),
             gini_eff=round(criterio4,4),
             gini_eff_prom=round(criterio5,4),
             Gini_1mEff=round(metrica_aux,4),
             configuracion=paste(varnames,collapse = ""),
             scores_lasso,
             varnames
      )
      res
      (map<-rbind(map,res))
    }
  }
  map<- as.data.frame(map)
  map
  rownames(map) <- NULL

  map2<- as.data.frame(map[,(3:dim(map)[2])])
  map2<- map[!duplicated(map2), ]
  map2 <- as.data.frame(map2[order(map2$lambda),])

  return(list(map, map2))
}
```

VII. Programa en R: PCA-DEA (Principal Component Analysis)

```
DEA_PCA_addc_smb<-function(input_data, output_data, exp_var=1){
  library(magrittr)
  library(dplyr)
  library(tidyr)
  library(tidyverse)
  library(ROI)
  library(ROI.plugin.glpk)
  library(ompr)
  library(ompr.roi)
  inputs<-input_data
  outputs<-output_data
```

```
inputs<-x
outputs<-y
exp_var=1
#
acp_inputs<-prcomp(inputs, center=TRUE, scale=TRUE)
summary(acp_inputs)
impor_xcp<-round(summary(acp_inputs)$importance[2,],4)
nxt<-max(which(cumsum(impor_xcp)<=exp_var),1)

acp_outputs<-prcomp(outputs, center=TRUE, scale=TRUE)
summary(acp_outputs)
impor_ycp<-round(summary(acp_outputs)$importance[2,],4)
nyt<-max(which(cumsum(impor_ycp)<=exp_var),1)
ite<-0
map<-NULL
for(i in nxt:1){
  for(j in nyt:1){
    nx<-i
    ny<-j
    Lx<-acp_inputs$rotation
    Lx
    Ly<-acp_outputs$rotation
    Ly
    Lx_inv<-solve(Lx,I=diag(1,nrow=dim(Lx)[1]))
    Lx_inv
    round(Lx%*%Lx_inv,3)
    Ly_inv<-solve(Ly,I=diag(1,nrow=dim(Ly)[1]))
    Ly_inv
    round(Ly%*%Ly_inv,3)

    xo<-inputs
    yo<-outputs

    if(nx==1){Lx<-matrix(Lx[,(1:nx)], ncol=1)}else{Lx<-matrix(Lx[,(1:nx)], ncol=nx)}
    Lx; dim(Lx)
    xpc<-xo%*%Lx
    if(ny==1){Ly<-matrix(Ly[,(1:ny)], ncol=1)}else{Ly<-matrix(Ly[,(1:ny)], ncol=ny)}
    Ly; dim(Ly)
    ypc<-yo%*%Ly
    ypc
    xopc<-xo%*%acp_inputs$rotation
    yopc<-yo%*%acp_outputs$rotation
    vinfx=1:(nxt-(nxt-nx))
    vinfy=1:(nyt-(nyt-ny))
    aux_data_exclu <- cbind(xopc[,-vinfx],yopc[,-vinfy])
    aux_data<-cbind(aux_data_exclu,cbind(xopc[,vinfx],yopc[,vinfy]))
    round(aux_data,2)
    # det(aux_data)
    # diag(aux_data)
    if((nxt-nx)==0){var_exp_total<-1}else{var_exp_total<-round(parcor(data=aux_data, var_include=(dim(aux_data_exclu)[2
    xo<-inputs-inputs
    yo<-outputs-outputs

    (ei<-rep(1,dim(xpc)[2]))
    (eo<-rep(1,dim(ypc)[2]))
    # (inp<-seq(1,dim(xpc)[2],1))

    (inp_pc<-as.integer(seq(1,dim(xpc)[2],1)))
    (out_pc<-as.integer(seq(1,dim(ypc)[2],1)))

    (inp<-as.integer(seq(1,dim(xo)[2],1)))
```

```
(out<-as.integer(seq(1,dim(yo)[2],1)))
(if(nx==1){inp2<-NULL}else{inp2<-as.integer(seq(1,dim(xpc)[2]-1,1))})
(if(ny==1){out2<-NULL}else{out2<-as.integer(seq(1,dim(ypc)[2]-1,1))})
(dmu<-as.integer(seq(1,dim(ypc)[1],1)))
class(inp)
class(out)
class(inp_pc)
class(out_pc)
class(dmu)

score<-rep(0, length(dmu))
remetrica<-rep(0, length(dmu))
for(k in 1:length(dmu)){
  dmua<-k
  m <- MIPModel() %>%
    # m <- MIPModel()%>%
    add_variable(so[o], o=out, type="continuous",lb=0,ub=Inf) %>%
    add_variable(spcu[o], o=out, type="continuous",lb=0,ub=Inf) %>%
    add_variable(spcl[o], o=out, type="continuous",lb=0,ub=Inf) %>%
    add_variable(zi[i], i=inp, type="continuous",lb=0,ub=Inf) %>%
    add_variable(zpcu[i], i=inp, type="continuous",lb=0,ub=Inf) %>%
    add_variable(zpcl[i], i=inp, type="continuous",lb=0,ub=Inf) %>%
    add_variable(lambda[d], d=dmu, type="continuous",lb=0,ub=Inf) %>%

    add_constraint((sum_expr(lambda[d] * yo[d,o], d=dmu)- so[o]) == yo[dmua,o], o=out) %>%
    add_constraint((-sum_expr(lambda[d] * xo[d,i], d=dmu)- zi[i])==-xo[dmua,i], i=inp) %>%

    add_constraint((sum_expr(lambda[d] * ypc[d,o], d=dmu)- (spcu[o]-spcl[o]))==ypc[dmua,o], o=out_pc) %>%
    add_constraint((-sum_expr(lambda[d] * xpc[d,i], d=dmu)- (zpcu[i]-zpcl[i]))==-xpc[dmua,i], i=inp_pc) %>%

    add_constraint(sum_expr(Ly_inv[h,o]*(spcu[h]-spcl[h]), h=out) >=0, o=out) %>%
    add_constraint(sum_expr(Lx_inv[h,i]*(zpcu[h]-zpcl[h]), h=inp) >=0, i=inp) %>%

    add_constraint(sum_expr(Ly_inv[h,o]*(spcu[h]-spcl[h]), h=out) <= outputs[dmua,o], o=out) %>%
    add_constraint(sum_expr(Lx_inv[h,i]*(zpcu[h]-zpcl[h]), h=inp) <= inputs[dmua,i], i=inp) %>%

    set_objective(sum_expr(so[o] + sum_expr(Ly_inv[h,o]*(spcu[h]-spcl[h]), h=out), o=out)
                  +sum_expr(zi[i] + sum_expr(Lx_inv[h,i]*(zpcu[h]-zpcl[h]), h=inp), i=inp)
                  ,"max") %>%

    solve_model(with_ROI(solver = "glpk"))


  get_solution(m,lambda[d])
  get_solution(m,so[o]);get_solution(m,zi[i])
  get_solution(m,zpcu[i]);get_solution(m,zpcl[i]);get_solution(m,spcu[o]);get_solution(m,spcl[o])
  solver_status(m);objective_value(m)

  spc=get_solution(m,spcu[o])$value-get_solution(m,spcl[o])$value
  zpc=get_solution(m,zpcu[i])$value-get_solution(m,zpcl[i])$value

  hx=zpc%*%Lx_inv
  hy=spc%*%Ly_inv
  round(hx,2)
  round(hy,2)

  vx=inputs[dmua,]
  vy=outputs[dmua,]

  eff_sbm=1-(sum(hy/vy)+sum(hx/vx))/(length(hx)+length(hy))
  eff_sbm_no=(1-sum(hx/vx)/length(hx))/(1+sum(hy/vy)/length(hy))
```

```
        eff_sbm_io=1-sum(hx/vx)/length(hx)
        eff_sbm_oo=1+sum(hy/vy)/length(hy)
        (1-sum(hx/vx)/length(hx))
        (1+sum(hy/vy)/length(hy))
        (1-sum(hx/vx)/length(hx))/(1+sum(hy/vy)/length(hy))
        1/(1-sum(hx/vx)/length(hx))

        DMU_i<-dmua
        Status_Mod<-solver_status(m)
        objective_Mod<-round(objective_value(m),9)
        cbind(DMU_i,Status_Mod,objective_Mod)
        if(objective_Mod==0){eff_sbm=1}
        remetrica[k]<-eff_sbm
        score[k]<-objective_value(m)
      }
      round(score,4)
      round(remetrica,4)

      eff<-length(which(round(remetrica,6)==1))
      neff<-length(which(round(remetrica,6)!=1))
      n_dmu<-length(remetrica)
      # 1: entropy_bi
      # 2: gini_bi
      # 3: entropy_eff
      # 4: gini_eff
      criterio1<- -(eff/n_dmu)*log2(eff/n_dmu) -(neff/n_dmu)*log2(neff/n_dmu)
      criterio2<- (eff/n_dmu)*(1-(eff/n_dmu)) + (neff/n_dmu)*(1-(neff/n_dmu))
      criterio3<- round(sum((-1/n_dmu)*log2(round(remetrica,6))),7)
      criterio4<- gini_eff(remetrica)
      criterio5<- (gini_eff(remetrica)+gini_eff(1-remetrica))/2
      metrica_aux<- gini_eff(1-remetrica)
      ite<-ite+1
      res<-c(itera=ite,
             npcx=nx,
             npcy=ny,
             var_inp=cumsum(impor_xcp)[nx],
             var_out=cumsum(impor_ycp)[ny],
             dmu_eff=eff,
             entropy_bi=criterio1,
             gini_bi=criterio2,
             entropy_eff=criterio3,
             gini_eff=criterio4,
             gini_eff_prom=criterio5,
             gini_1meff=metrica_aux,
             var_exp_model=var_exp_total,
             score_dmu=round(score,4),
             remetrica_dmu=round(remetrica,4)
      )
      res
      t(res)
      # as.data.frame(res)
      (map<-rbind(map,res))
      # map<-as.data.frame(map)

    }
  }
  return(map)
}
```