Predicción de la eficiencia de las instituciones de educación superior colombianas con análisis envolvente de datos y minería de datos

Predicting the Efficiency of Colombian Higher Education Institutions with Data Envelopment Analysis and Data Mining

Delimiro Visbal Cadavid
dvisbal@unimagdalena.edu.co

Doctorando en Ingeniería Industrial, Universidad Politécnica de Valencia (España). Magíster en Ingeniería Industrial, Universidad de los Andes (Colombia). Especialista en Gerencia de Producción y Operaciones, Universidad Autónoma del Caribe (Colombia). Ingeniero Químico, Universidad del Atlántico (Colombia). Profesor, Facultad de Ingeniería, Departamento de Ingeniería Industrial, Universidad del Magdalena.

Adel Mendoza Mendoza
adelmendoza@uniatlantico.edu.co.

Magíster en Ingeniería Industrial, Universidad del Norte (Colombia). Especialista en Gerencia de Producción y Operaciones, Universidad Autónoma del Caribe (Colombia). Ingeniero Químico, Universidad del Atlántico (Colombia). Profesor, Facultad de Ingeniería, Departamento de Ingeniería Industrial, Universidad del Atlántico.

Sonia Jacqueline Orjuela Pedraza
sojacor@alumni.uv.es.

Magíster en marketing e investigación de mercados, Universidad de Valencia (España). Especialista en mercadeo, Universidad Jorge Tadeo Lozano (Colombia). Ingeniera Química, Universidad de América (Colombia).

pensamiento y gestión, N° 42
ISSN 1657-6276
http://dx.doi.org/10.14482/pege.41.9704

Resumen

Este trabajo muestra los resultados de una investigación cuyo propósito es evaluar la eficiencia técnica de las instituciones de educación superior en Colombia entre los años 2011-2013, mediante la aplicación del análisis envolvente de datos y técnicas de minería de datos. Con el análisis envolvente de datos se determinó la eficiencia de las instituciones de educación superior esta información es utilizada en la minería de datos. El resultado de la combinación de las dos técnicas facilitó unas reglas de predicción con base a un grupo de indicadores de gestión que pueden ser utilizadas en el diseño de políticas educativas para determinar las razones de algunas ineficiencias de las instituciones de educación superior. Como fuente para los datos, se utilizó la información provista por el Ministerio de Educación Nacional. Se observó que siete de las treinta y dos instituciones consideradas tienen una eficiencia de 100% durante el periodo de estudio.

Palabras clave: *Eficiencia, Análisis envolvente de datos, Minería de datos, Educación superior.*

Abstract

This paper shows the results of a research study whose purpose is to evaluate the technical efficiency of institutions of higher education in Colombia during the years 2011-2013 by applying the data envelopment analysis and data mining techniques. With data envelopment analysis, it is determined the technical efficiency and data mining allows to discover hidden information. The result of the combination of these techniques allows to establish prediction rules based on a group of management indicators that can be used by the designers of educational policies to determine the reasons for inefficiency of institutions of higher education. The information provided by the Ministry of National Education was used as a source for the data. It was observed that seven of the 32 institutions considered had an efficiency of 100% during the period of study.

Keywords: *Efficiency, Data Envelopment Analysis, Data Mining, Higher Education*

Fecha de recepción: 20 de enero de 2017
Fecha de aceptación: 3 de mayo de 2017

1. INTRODUCCIÓN

Hoy día es una exigencia considerar criterios de racionalidad y eficiencia económica en la gestión de las instituciones educativas públicas en sus diferentes niveles, con el fin de mejorar sus procesos mediante la identificación de las variables que puedan afectar dicha gestión de forma significativa. Es por ello, que los diferentes gobiernos y organismos involucrados con la gestión de la educación superior desarrollan planes y estrategias para la mejora de la eficiencia y el funcionamiento de las universidades. La asignación final de los recursos públicos y su uso eficiente están íntimamente correlacionados por lo que los investigadores del campo de la economía de la educación han dedicado tiempo y empeño para evaluar la eficiencia de los sistemas educativos en sus diferentes niveles.

En este sentido, es importante destacar que existen diferencias significativas entre las distintas Instituciones de Educación Superior (IES) en lo referente a los recursos con que cuentan para llevar a cabo sus objetivos misionales (misión estratégica), que de alguna manera inciden en los resultados obtenidos por las mismas. Es factible encontrar universidades que, con menos recursos, pero bien utilizados, muestran mejor desempeño que instituciones con más recursos. Por lo anterior, cabe preguntar por la importancia de las variables al momento de discriminar entre IES eficientes e ineficientes y por las variables sobre las cuales se debe actuar para mejorar el desempeño de las IES. Por esto, la evaluación de la eficiencia es un asunto muy común en muy diversos ámbitos, que brinda herramientas cuantitativas para los responsables del análisis de inversiones y la asignación de recursos (Lan, Chuang y Chen, 2009).

El resultado del presente trabajo permitirá a los entes tomadores de decisión del sector educativo superior en Colombia, definir políticas y lineamientos que redunden en la mejora del desempeño de las IES, lo cual se traduce en acciones basadas en la evidencia empírica mostrada por estos resultados, y no en creencias y percepciones de los funcionarios de las instituciones o de directrices normativas.

2. 2. MARCO TEÓRICO

2.1. Análisis envolvente de datos (DEA por su sigla en inglés)

La metodología del Análisis Envolvente de Datos (DEA) fue propuesta por Charnes, Cooper y Rhodes con base en los conceptos planteados por Farrel en 1957 (Charnes, Cooper y Rhodes, 1978) quien evaluó la eficiencia relativa de unidades de producción con múltiples insumos y productos. Así, se generó una frontera de valores eficiente y los índices mismos de eficiencia dentro del grupo de unidades de producción estudiadas (Decision Making Unit, DMU).

El análisis envolvente de datos (DEA) es una de las principales técnicas usadas en el sector público y privado, su uso es tan amplio que podemos citar entre sus aplicaciones las realizadas, por ejemplo, en la evaluación de eficiencias en el sector financiero (Tsolas y Charles, 2015), en el desempeño de fuerzas policiales (Aristovnik, Seljak y Mencinger, 2014); en la asignación de recursos (Fang y Li, 2015), en la evaluación de la eficiencia medioambiental (Woo, Chung, Chun, Seo y Hong, 2015). En el área educativa, el análisis DEA ha sido usado en evaluación del desempeño en instituciones de educación básica (Huguenin, 2015), en evaluación del desempeño de universidades (Thanassoulis, Kortelainen, Johnes, G. y Johnes, J., 2011), en evaluación de programas académicos (Gökşen, Doğan y Özkarabacak, 2015), en la evaluación de centros de investigación (Da Silva y Gonçalves, 2015).

Entre los trabajos que combinan DEA con técnicas de minería de datos para clasificar entre observaciones eficientes e ineficientes, se encuentran los realizados por Emrouznejad y Anouze (2010), quienes utilizan el análisis envolvente de datos con árboles de regresión y clasificación con validación *bootstrapping* para investigar los factores asociados con la eficiencia de bancos en los países árabes del golfo. Por otro lado, Chuang, Chang y Lin (2011) también combinan DEA con árboles de regresión y clasificación para evaluar el desempeño de un conjunto de hospitales en Taiwán.

El análisis envolvente de datos también ha sido bastante utilizado junto con máquina de soporte vectorial. Dentro de estos se pueden destacar

los realizados por: Yeh, Chi y Hsu (2010) para mejorar la precisión en la predicción del riesgo de quiebra empresarial; Kao, Chang, T. y Chang, Y. (2013) para llevar a cabo la estimación de la eficiencia y posterior clasificación de la seguridad de las páginas web de instituciones médicas; Farahmand, Desa y Nilashi (2014) para evaluar y predecir la eficiencia de grandes conjuntos de unidades tomadoras de decisión; Poitier y Cho (2011) para determinar la frontera eficiente del desempeño organizacional; Song y Zhang (2009) para evaluar el desempeño productivo de refinerías petroleras; Jiang, Chen, Zhang y Pan (2013) para evaluación de proveedores y Bazleh, Gholami y Soleymani (2011) para evaluar y realizar un ranking del desempeño de diversos algoritmos de clasificación.

En el análisis envolvente de datos (DEA), el valor de la eficiencia para cada una de las unidades evaluadas se define como la relación entre la ponderación de la suma de las salidas (outputs) y las entradas (inputs). Si $Y_o = (y_{1o}, y_{2o}, y_{3o}, ..., y_{so})$ y $X_0 = (x_{1o}, x_{2o}, x_{3o}, ..., x_{so})$ describe el valor de las entradas y las salidas, de la DMU_0, que es la unidad que se está evaluando, la medida escalar de la eficiencia para esta DMU_0 se obtiene de la solución óptima del modelo mostrado a continuación.

$$Max \ \square = \frac{\sum_{r=1}^{s} u_{ro} y_{ro}}{\sum_{i=1}^{m} v_{io} x_{io}} \qquad (1)$$

sujeto a:

$$\frac{\sum_{r=1}^{s} u_{rj} y_{rj}}{\sum_{i=1}^{m} v_{ij} x_{ij}} \leq 1 \ \ j = 1,2,\ldots n \qquad (2)$$

$$U_{rj}, v_{ij} \geq 0 \ r = 1,..s \quad i = 1,...m$$

2.2. Minería de datos

La minería de datos se puede definir como un proceso donde se descubren nuevas y significativas relaciones, patrones y tendencias al examinar y explorar cantidades de datos muy grandes (López, 2007). Las técnicas de Minería de Datos (MD) se dividen en dos grupos, según el objetivo de

análisis: técnicas de aprendizaje supervisado o métodos predictivos donde hay una variable que debe ser explicada por las otras, es decir, se tiene una clasificación *a priori* de los individuos y técnicas de aprendizaje no supervisado o métodos descriptivos donde no hay una variable preferente, que debe ser explicada por las otras, es decir, cuando los individuos no están previamente clasificados.

Por los objetivos de análisis y la naturaleza de los datos, este trabajo será analizado bajo las técnicas de aprendizaje supervisado como lo son: Árbol de Clasificación, Vecino más próximo, Random Forest y Máquina de Soporte Vectorial.

Árboles de clasificación (AC)**.** Tratan de predecir una variable respuesta categórica en lugar de una variable respuesta continua. Son útiles porque proporcionan predictores bastante comprensibles en situaciones en las que hay muchas variables que interactúan de forma no lineal, pueden simplificar las complejidades de grandes colecciones de datos, para construir un AC se inicia con todas las observaciones en el nodo raíz del árbol y el algoritmo divide recursivamente los datos en particiones binarias hasta que todas las observaciones en los nodos de la hoja sean de la misma clase (Lau, Salibian-Barrera y Lampe, 2016).

Vecino más cercano. Es un tipo de aprendizaje que no intenta construir un modelo interno general sino simplemente almacena los casos de los datos de entrenamiento. K vecinos más cercanos es un algoritmo simple que almacena todos los casos disponibles y clasifica los nuevos sobre la base de una medida de similitud. El algoritmo de k-vecinos más cercanos es uno de los algoritmos de aprendizaje de máquina más simples. Se basa en la idea de que "los objetos que están" cerca "unos de otros también tendrán características similares. Así, si conoce las características de uno de los objetos, también puede predecirlo para su vecino más cercano". (Khamis, Cheruiyot y Kimani, 2014, p. 122).

Random Forest. Es una de las técnicas de aprendizaje más exitosas que han demostrado ser muy poderosa en el reconocimiento de patrones para la clasificación de alta dimensionalidad y problemas sesgados (Azar, Elshazly, Hassanien y Elkorany, 2014). En los bosques aleatorios, no es ne-

cesario para la validación cruzada una prueba separada configurada para obtener una estimación no sesgada del error de prueba, ya que se estima internamente. Cada árbol se construye utilizando una muestra de arranque diferente de los datos originales. Alrededor de un tercio de los casos se quedan fuera de la muestra de arranque y no se utiliza en la construcción del árbol de orden k. De esta manera, se obtiene una clasificación de prueba para cada caso en alrededor de un tercio de los árboles.

Máquina de Soporte Vectorial. SVM (Support Vector Machines) es una técnica útil para la clasificación de datos. Se puede emplear una máquina de soporte vectorial cuando sus datos tienen exactamente dos clases, como se presenta en este trabajo. Una SVM clasifica los datos mediante la búsqueda del mejor hiperplano que separa todos los puntos de datos de una clase de los de la otra clase. El mejor hiperplano para una SVM significa el que tiene el mayor margen entre las dos clases. Margen significa la anchura máxima de la losa paralelo al hiperplano que no tiene puntos de datos interiores.

3. MATERIALES Y MÉTODOS

La investigación realiza un estudio cuantitativo mediante el Análisis Envolvente de Datos (DEA) y herramientas de clasificación de minería de datos, consistente en un análisis de la eficiencia técnica de las universidades públicas colombianas, durante los años 2011 a 2013, utilizando el modelo CCR-O (modelo CCR enfocado a salidas) y un posterior estudio mediante técnicas de minería de datos con el propósito de determinar las variables que mejor clasifican una IES entre eficiente e ineficiente, y la validación del poder predictivo de cada modelo estudiado. El estudio también permite predecir la clasificación de una nueva observación con base en los modelos formulados.

Cabe resaltar que la mayor parte de las contribuciones en DEA son hechas por dos modelos estándar el CCR y el BCC (Suzuki, Nijkamp, Rietveld y Pels, 2010). El modelo CCR calcula las eficiencias bajo la hipótesis de retornos constantes a escala, es decir compara unidades homogéneas y no homogéneas, mientras que el modelo BBC calcula las eficiencias con retornos variables a escala en donde cada Unidad ineficiente solo se com-

para con una Unidad eficiente, pero de sus mismas características. Por lo anteriormente descrito en este artículo de investigación, la eficiencia de las IES consideradas objeto de estudio será evaluada mediante el modelo DEA CCR-O.

3.1. IES incluidas en el análisis

En el presente estudio de eficiencia de las IES públicas se consideran las 32 universidades públicas colombianas, pertenecientes al Sistema de Universidades Estatales (SUE). En la tabla 1 se muestran las instituciones consideradas en el estudio

Tabla 1. IES consideradas en el estudio

Universidad Colegio Mayor de Cundinamarca	Universidad del Quindío
Universidad de Antioquia	Universidad del Tolima
Universidad de Caldas	Universidad del Valle
Universidad de Cartagena	Universidad Distrital
Universidad de Córdoba	Universidad Francisco de Paula Santander - Cúcuta
Universidad de Cundinamarca	Universidad Francisco de Paula Santander - Ocaña
Universidad de La Amazonía	Universidad Industrial de Santander
Universidad de La Guajira	Universidad Militar Nueva Granada
Universidad de Los Llanos	Universidad Nacional Abierta y a Distancia (UNAD)
Universidad de Nariño	Universidad Nacional de Colombia
Universidad de Pamplona	Universidad Pedagógica Nacional
Universidad de Sucre	Universidad Pedagógica y Tecnológica de Colombia (UPTC)
Universidad del Atlántico	Universidad Popular del Cesar
Universidad del Cauca	Universidad Surcolombiana (Neiva)
Universidad del Magdalena	Universidad Tecnológica de Pereira
Universidad del Pacífico	Universidad Tecnológica del Chocó

Fuente: Elaboración propia.

Las IES en la etapa de análisis se nombraron de forma aleatoria de IES-1 a IES-32, esto con el fin de garantizar la confidencialidad de los resultados de la evaluación.

3.2. Variables consideradas en el estudio

La información referente a las variables consideradas en la realización de este estudio y su respectiva descripción se obtuvo del Ministerio de Educación Nacional de Colombia (2015). El análisis envolvente de datos requiere la identificación de las variables de entradas, las cuales se corresponden con los recursos utilizados para llevar a cabo la función misional, y de las variables de salida, las cuales se identifican como los productos u objetivos misionales de cada IES.

Variables de entrada (recursos):

- **Docentes tiempo completo equivalente (DTCE):** Número de docentes en tiempos completos equivalentes, incluyendo catedráticos y ocasionales.
- **Gastos en personal administrativo (GPA):** Gasto para el pago del personal no docente (COP).
- **Recursos financieros (Recfin):** Recursos financieros provenientes del Estado y generados por la universidad en desarrollo de su actividad (no incluye ingresos generados por extensión e investigación) (COP).
- **Recursos físicos (M2):** Área de los espacios físicos construidos disponibles para las actividades universitarias misionales y de apoyo administrativo. (m^2)
- Variables de salida (productos):
- **Matrícula de pregrado (Matpreg):** Número ponderado de matriculados por niveles de formación y metodologías de enseñanza en posgrado.
- **Matrícula de posgrado (Matpost):** Número ponderado de matriculados por niveles de formación y metodologías de enseñanza en posgrado.

- **Saber PRO (Saberpro):** Número ponderado de estudiantes de la universidad que obtuvieron un puntaje mayor al quintil superior en las pruebas saber pro.
- **Revistas indexadas (Revindex):** Número ponderado de revistas indexadas de la institución de acuerdo la legislación vigente (Colciencias).
- **Artículos en revistas indexadas (Artículos):** Número ponderado de artículos publicados en revistas indexadas.
- **Movilidad de Docentes (Movdoc):** Número de docentes vinculados a procesos de movilidad promovidos desde la IES a la que pertenecen.

4. RESULTADOS

4.1. Análisis Envolvente de Datos (DEA)

Los resultados de eficiencia del modelo CCR-O de los años 2011 a 2013 se muestran en la tabla 2. En este modelo, una DMU es considerada eficiente si el valor obtenido de la eficiencia es igual a uno ($\theta = 1$) y no se presentan holguras (es decir la holgura en cada una de las variables es cero).

Tabla 2. Eficiencia de las IES

IES	Eficiencia 2011	Eficiencia 2012	Eficiencia 2013
IES 1	Eficiente	Ineficiente	Eficiente
EIS 2	Ineficiente	Ineficiente	Ineficiente
IES 3	Ineficiente	Ineficiente	Eficiente
IES 4	Ineficiente	Ineficiente	Ineficiente
IES 5	Eficiente	Eficiente	Eficiente
IES 6	Eficiente	Ineficiente	Ineficiente
IES 7	Ineficiente	Ineficiente	Ineficiente
IES 8	Ineficiente	Eficiente	Eficiente
IES 9	Ineficiente	Eficiente	Ineficiente
IES 10	Ineficiente	Ineficiente	Eficiente
IES 11	Eficiente	Eficiente	Eficiente

Continúa...

IES	Eficiencia	Eficiencia	Eficiencia
	2011	2012	2013
IES 12	Eficiente	Eficiente	Eficiente
IES 13	Eficiente	Ineficiente	Ineficiente
IES 14	Ineficiente	Ineficiente	Ineficiente
IES 15	Eficiente	Eficiente	Ineficiente
IES16	Eficiente	Eficiente	Eficiente
IES 17	Ineficiente	Ineficiente	Ineficiente
IES 18	Ineficiente	Ineficiente	Ineficiente
IES 19	Eficiente	Eficiente	Eficiente
IES 20	Eficiente	Ineficiente	Eficiente
IES 21	Eficiente	Eficiente	Ineficiente
IES 22	Eficiente	Eficiente	Eficiente
IES 23	Eficiente	Eficiente	Ineficiente
IES 24	Ineficiente	Ineficiente	Ineficiente
IES 25	Eficiente	Ineficiente	Ineficiente
IES 26	Ineficiente	Ineficiente	Ineficiente
IES 27	Ineficiente	Ineficiente	Ineficiente
IES 28	Ineficiente	Ineficiente	Ineficiente
IES 29	Ineficiente	Ineficiente	Ineficiente
IES 30	Ineficiente	Ineficiente	Ineficiente
IES 31	Eficiente	Ineficiente	Ineficiente
IES 32	Eficiente	Eficiente	Eficiente

Fuente: Elaboración propia.

4.2. Aplicación de la Minería de datos para estimación de eficiencia de la IES

Los modelos de clasificación utilizados son Árbol de Clasificación, Vecino más próximo, Random Forest y Máquina de Soporte Vectorial. Cuando se construyen estos modelos de clasificación se puede predecir cuál será su índice de error al aplicarlos a datos sin clasificar (García, Navelonga y García Peñalvo, 2007).

4.2.1. Árbol de clasificación

El árbol de clasificación nos muestra como primer criterio de clasificación a la variable Matpreg (cantidad de alumnos matriculados en programas de grado), seguida por Revindex (número de revistas indexadas) y por último a la variable Recfin (recursos financieros).

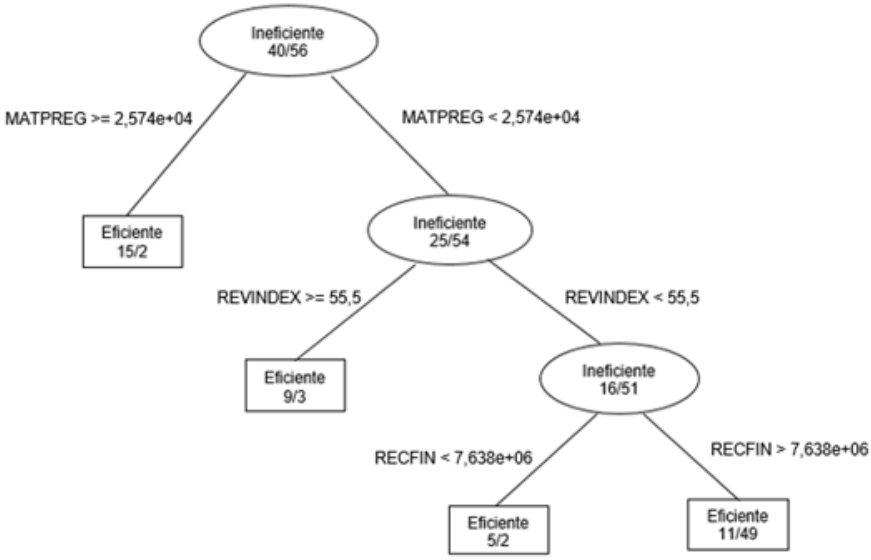

Figura 1. Árbol de clasificación entrenamiento

Reglas de clasificación:

- Nodos internos: categoría
- Arcos: valores del nodo de origen
- Hojas: valor de clasificación “eficiente” e “ineficiente”

Disyunción de reglas proposicionales

- Si MATPREG>=2,574e+04 entonces IES = "Eficiente"
- Si MATPREG<2,574e+04 y REVINDEX>=55,5 entonces IES = "Eficiente"
- Si MATPREG<2,574e+04 y REVINDEX<55,5 y RECFIN<7,638e+06 entonces IES = "Eficiente"
- Si MATPREG<2,574e+04 y REVINDEX<55,5 y RECFIN>=7,638e+06 entonces IES = "Ineficiente"

Por tanto, de un total de 96 IES; categorizadas como 56 eficientes y 40 ineficientes, podemos observar que mediante el modelo de árbol de clasificación de 56 IES ineficiente, 49 clasificadas correctamente, y de 40 IES eficientes, 29 son clasificadas correctamente, por lo que la tasa global de correcta clasificación es de 81,25%.

4.2.2. Vecino más próximo (KNN)

El entrenamiento lo realizamos con una muestra de tamaño aleatoria de 72 observaciones sin reemplazo. Se determina la tasa de clasificación correcta utilizando las 24 observaciones de prueba y también toda la base de datos. La matriz de confusión con las observaciones no utilizadas para el entrenamiento (observaciones de prueba).

Tabla 3. Matriz de confusión (Validación KNN)

	Eficiente	**Ineficiente**
Eficiente	6	6
Ineficiente	4	8
Total	10	14

Fuente: Elaboración propia.

La tasa global de correcta clasificación es del 58,33%. La tasa de verdaderos positivos es 60% y la tasa de verdaderos negativos es de 57,14%

La matriz de confusión con todas las observaciones.

Tabla 4. Matriz de Confusión (Entrenamiento KNN)

	Eficiente	Ineficiente
Eficiente	15	25
Ineficiente	17	39
Suma	32	64

Fuente: Elaboración propia.

La tasa global de correcta clasificación es del 56,25%. La tasa de verdaderos positivos es 46,87% y la tasa de verdaderos negativos es de 60,94%.

4.2.3. Random Forest (RF)

La importancia de cada una de las variables utilizadas para clasificar una IES como eficiente o ineficiente, se muestra en la tabla 5 en donde se observa que las variables más importantes según el Mean Decrease Dini son Matpreg, Matpost, Saberpro, según el Mean Decrease Accuracy, las variables más importantes son Matpreg, Movdoc, Matpost y Saberpro, indicándonos que Matpreg, Matpost y Saberpro son las variables de mayor importancia al momento de clasificar a una IES como eficiente o ineficiente.

Tabla 5. Importancia de las variables predictoras según RF

	Eficiente	Ineficiente	Mean Decrease Accuracy	Mean Decrease Gini
MATPREG	8,6632807	13,8690230	15,39271	6,67331
MATPOST	6,9190638	12,0835930	13,51881	5,59438
SABERPRO	-0,8016486	15,4357860	12,65028	5,10904
GPA	6,3029290	10,4781970	11,58101	4,78363
M2	7,8450084	8,7910740	11,25235	4,76469
RECFIN	5,4457624	6,6912960	8,39539	4,61298
MOVDOC	4,3607542	15,2369000	14,25431	4,26152
REVINDEX	7,1559971	11,5097820	12,22567	3,95734
DTCE	3,4253285	9,7260280	9,54733	3,74111
ARTICULOS	-0,5155326	5,2837390	3,96518	2,66988

Fuente: Elaboración propia.

4.24. Máquina de Soporte Vectorial (Support Vector Machine, SVM)

El modelamiento de las SVM se llevó a cabo utilizando los paquetes "*kernlab*" y "*e1071*" del software estadístico R. Las clasificaciones realizadas por cada uno de los paquetes se muestran en las tablas 6 y 7.

Tabla 6. Matriz de confusión (SVM-K)

pred	Eficiente	Ineficiente
Eficiente	27	0
Ineficiente	13	56
Suma	40	56

Fuente: Elaboración propia.

Tabla 7. Matriz de confusión (SVM-e)

pred	Eficiente	Ineficiente
Eficiente	28	0
Ineficiente	12	56
Suma	40	56

Fuente: Elaboración propia.

4.3. Validación de los modelos de minería de datos

Para la validación de los modelos se utiliza el remuestreo (*bootstrapping*) y validación simple (*holdout*).

El remuestreo o *bootstrapping* como método para validación del poder predictivo de los modelos de clasificación se construyen (se entrenan) con las 96 observaciones disponibles y se valida realizando el remuestreo también de tamaño 96 pero con reemplazo, exceptuando el modelamiento de Vecino más próximo, el cual se entrena eligiendo al azar 72 observaciones sin reemplazo, y se aplica validación cruzada con las 24 observaciones restantes. Se trata de un procedimiento relativamente simple, pero repetidos tantas veces, entonces la técnica Bootstrap dependen en gran medida de cálculos por ordenador. Al analizar los datos, a menudo necesitan compa-

rar la eficacia de varios modelos, y una forma de medir la efectividad de la predicción es calculando el área bajo la curva (AUC). (Izrael, Battaglia, A., Hoaglin y Battaglia, M., 2002).

La validación simple o técnica Holdout consiste en utilizar una porción de registros como conjunto de entrenamiento, eligiendo al azar sin reemplazo muestras con un tamaño del 75% de las observaciones disponibles (72 observaciones) para construir el modelo y el resto, 25% (24 observaciones) como conjunto de prueba.

Para la validación por *bootstrapping* se realizan 100 muestreos de 96 observaciones con repetición. La tabla 8, que se muestra a continuación, indica que el mejor método de clasificación es SVP (SVM librería R kernlab), seguido por SVM (SVM librería R e1071).

Tabla 8. Cuartiles AUC Bootstrapping

	2.5%	50%	97.5%
AUC.Tree	0,1143390	0,1929547	0,2654014
AUC.Vecino	0,5040793	0,5870288	0,6695527
AUC.SVM	0,9418346	0,9725641	0,9897518
AUC.SVP	***0,9486383***	***0,9794344***	***0,9931274***

Fuente: Elaboración propia.

Los resultados de la validación por Holdout se muestran a continuación:

Para la técnica árbol de clasificación se observan la matriz de confusión y las tasas de clasificación de los datos. En la matriz de confusión podemos notar que de las 24 observaciones de validación 6 de 12 observaciones eficientes fueron correctamente clasificadas, correspondiendo esto a una tasa de verdaderos positivos (eficientes) del 50%, mientras que 9 de 12 observaciones ineficientes fueron correctamente clasificadas, esto corresponde a un 75% de tasa de verdaderos negativos (ineficientes). Globalmente el modelo clasifica correctamente 15 de 24 observaciones, ofreciendo una tasa de correcta clasificación (TCC) del 62,50%.

Tabla 9. Matriz de confusión. Árbol clasificación

pred	Eficiente	Ineficiente
Eficiente	6	3
Ineficiente	6	9

Fuente: Elaboración propia.

Tabla 10. Tasa de clasificación. Árbol de clasificación

TVF	TVP	TCC
0,750	0,500	0,625

Fuente: Elaboración propia.

Para la técnica el vecino más próximo en la matriz de confusión, podemos observar que de las 24 observaciones de validación 4 de 7 observaciones eficientes fueron correctamente clasificadas, correspondiendo esto a una tasa de verdaderos positivos (eficientes) del 57,14%, mientras que 9 de 17 observaciones ineficientes fueron correctamente clasificadas, esto corresponde a un 52,94% de tasa de verdaderos negativos (ineficientes). Globalmente el modelo clasifica correctamente 13 de 24 observaciones, ofreciéndonos una tasa de correcta clasificación del 54,17%.

Tabla 11. Matriz de confusión. Vecino más próximo

pred	Eficiente	Ineficiente
Eficiente	4	8
Ineficiente	3	9

Fuente: Elaboración propia.

Tabla 12. Tasa de clasificación. Vecino más próximo

TVF	TVP	TCC
0,5294	0,5714	0,5417

Fuente: Elaboración propia.

En las tablas 13 y 14 se muestran la matriz de confusión y las tasas de clasificación, respectivamente, para SVM con la librería *kernlab*. La matriz de confusión podemos observar que de las 24 observaciones de validación 9 de 12 observaciones eficientes fueron corectamente clasificadas, correspondiendo esto a una tasa de verdaderos positivos (eficientes) del 75%, mientras que 11 de 12 observaciones ineficientes fueron correctamente clasificadas, esto corresponde a un 91,67% de tasa de verdaderos negativos (ineficientes). Globalmente el modelo clasifica correctamente 15 de 24 observaciones, ofreciendonos una tasa de correcta clasificación (TCC) del 83,33%.

Tabla 13. Matriz de confusión - SVM kernlab

pred	Eficiente	Ineficiente
Eficiente	9	1
Ineficiente	3	11

Fuente: Elaboración propia.

Tabla 14. Tasa de correcta clasificación SVM kernlab

TVF	TVP	TCC
0,9167	0,7500	0,8333

Fuente: Elaboración propia.

En las tablas 15 y 16 se observa que el modelamiento de SVM con el paquete e1071 presenta una tasa de correcta clasificación (TCC) del 75%. Por su lado la tasa de verdaderos positivos es del 58,33% y la tasa de verdaderos negativos del 91,67%.

Tabla 15. Matriz de confusión- SVM e1071

pred	Eficiente	Ineficiente
Eficiente	7	1
Ineficiente	5	11

Fuente: Elaboración propia.

Tabla 16. Tasa de correcta clasificación SVM e1071

TVF	TVP	TCC
0,5833	0,9167	0,750

Fuente: Elaboración propia.

5. CONCLUSIONES

La revisión de la literatura respecto al uso del análisis envolvente de datos conjuntamente con herramientas de minería de datos indica que se logra una mejora sustancial en el análisis de eficiencia y clasificación.

En la construcción y validación del árbol de decisión por Holdout se encontró que el modelo clasifica correctamente 15 de 24 observaciones de validación, ofreciéndonos una tasa de correcta clasificación (TCC) del 62,50%.

Los resultados de RF muestran que las variables más importantes al momento de predecir la eficiencia de una IES son Número de alumnos matriculados en pregrado, Número de alumnos matriculados en postgrado y número de estudiantes con resultados en las Pruebas Saber Pro en el quintil superior. Por otro lado solo 16,67% de las observaciones que no participan en el entrenamiento son mal clasificadas por Random Forest.

En la construcción y validación del modelo de vecino más próximo por Holdout se encontró que 4 de 7 observaciones eficientes fueron corectamente clasificadas, correspondiendo esto a una tasa de verdaderos positivos (eficientes) del 57,14%, mientras que 9 de 17 observaciones ineficientes fueron correctamente clasificadas, esto corresponde a un 52,94% de tasa de verdaderos negativos (ineficientes). Globalmente el modelo clasifica correctamente 13 de 24 observaciones, ofreciéndonos una tasa de correcta clasificación del 54,17%.

El modelo de clasificación SVM con la librería *kernlab* indica que 9 de 12 (75%) observaciones eficientes fueron correctamente clasificadas, mientras que 11 de 12 (91,67%) observaciones ineficientes fueron correcta-

mente clasificadas. Globalmente el modelo clasifica correctamente 15 de 24 (83,33%) observaciones.

El modelamiento de SVM con el paquete *e1071* presenta una tasa de correcta clasificación (TCC) del 75%. Por su lado la tasa de verdaderos positivos es del 58,33% y la tasa de verdaderos negativos del 91,67%.

Los resultados de la validación indican que el ranking de los modelos considerados en este estudio de mejor a peor poder predictivo es SVM > SVP > > Árbol de clasificación > Vecino más próximo.

REFERENCIAS

Aristovnik, A., Seljak, J. and Mencinger, J. (2014). Performance measurement of police forces at the local level: A non-parametric mathematical programming approach. *Expert Systems with Applications*, *41*(4), 1647-1653.

Azar, A. T., Elshazly, H. I., Hassanien, A. E. and Elkorany, A. M. (2014). A random forest classifier for lymph diseases. *Computer Methods and Programs in Biomedicine*, *113*(2), 465-473.

Bazleh, A., Gholami, P. and Soleymani, F. (2011). A new approach using data envelopment analysis for ranking classification algorithms. *Journal of Mathematics and Statistics*, 7(4), 282-288.

Charnes, A., Cooper, W. and Rhodes, E. (1978). Measuring the efficiency of decision making units. *European Journal of Operational Research*, *2*(6), 429-444.

Chuang, C. L., Chang, P. C. and 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.

Da Silva, G. and Gonçalves, E. (2015). Management of agricultural research centers in Brazil: A DEA application using a dynamic GMM approach. *European Journal of Operational Research*, (240), 819-824.

Emrouznejad, A. and Anouze, A. L. (2010). Data envelopment analysis with classification and regression tree-a case of banking efficiency. *Expert Systems*, *27*(4), 231-246.

Fang, L. and Li, H. (2015). Centralized resource allocation based on the cost-revenue analysis. *Computers & Industrial Engineering*, (85), 395-401.

Farahmand, M., Desa, M. I. and Nilashi, M. (2014). A combined data envelopment analysis and support vector regression for efficiency evaluation of large decision-making units. *International journal of engineering and technology (IJET),* *6*(5), 2310-2321.

García, M., Navelonga, M. y García Peñalvo, F. J. (2007). *Modelos de estimación del software basados en técnicas de aprendizaje automático*. La Coruña, España. Netbiblo.

Gökşen, Y., Doğan, O. and Özkarabacak, B. (2015). A data envelopment analysis application for measuring efficiency of university departments. *Procedia Economics and Finance*, (19), 226-237.

Huguenin, J. M. (2015). Determinants of school efficiency: The case of primary schools in the State of Geneva, Switzerland. *International Journal of Educational Management*, *29*(5), 539-562.

Izrael, D., Battaglia, A. A., Hoaglin, D. C. and Battaglia, M. P. (2002, April). Use of the ROC curve and the bootstrap in comparing weighted logistic regression models. En *Proceedings of twenty-seventh annual SAS users group international conference* (pp. 1-6).

Jiang, B., Chen, W., Zhang, H. and Pan, W. (2013). Supplier's efficiency and performance evaluation using DEA-SVM approach. *Journal of Software*, *8*(1), 25-30.

Kao, H. Y., Chang, T. K. and Chang, Y. C. (2013). Classification of hospital web security efficiency using data envelopment analysis and support vector machine. *Mathematical Problems in Engineering*, *2013,* Article ID 542314.

Khamis, H. S., Cheruiyot, K. W. and Kimani, S. (2014). Application of k-nearest neighbour classification in medical data mining. *International Journal of Information*, *4*(4), 121-128.

Lan, C. H., Chuang, L. L. and Chen, Y. F. (2009). Performance efficiency and resource allocation strategy for fire department with the stochastic consideration. *International Journal of Technology, Policy and Management*, *9*(3), 296-315.

Lau, K., Salibian-Barrera, M. and Lampe, L. (2016). Modulation recognition in the 868MHz band using classification trees and random forests. *AEU-International Journal of Electronics and Communications*, *70*(9), 1321-1328.

López, C. P. (2007). *Minería de datos: técnicas y herramientas*. Madrid, España. Editorial Paraninfo. QUITAR Ministerio de Educación Nacional.

Ministerio de Educación Nacional de Colombia (2015).

Poitier, K. and Cho, S. (2011). Estimation of true efficient frontier of organizational performance using data envelopment analysis and support vector machine learning. *International Journal of Information and Decision Sciences*, *3*(2), 148-172.

Song, J. and Zhang, Z. (2009, January). Oil refining enterprise performance evaluation based on DEA and SVM. En IEEE (2009), *Knowledge discovery and data mining* (pp. 401-404). WKDD 2009. Second International Workshop.

Suzuki, S., Nijkamp, P., Rietveld, P. and Pels, E. (2010). A distance friction minimization approach in data envelopment analysis: a comparative study on airport efficiency. *European Journal of Operational Research*, *207*(2), 1104-1115.

Thanassoulis, E., Kortelainen, M., Johnes, G. and Johnes, J. (2011). Costs and efficiency of higher education institutions in England: a DEA analysis. *Journal of the Operational Research Society*, *62*(7), 1282-1297.

Tsolas, I. E. and Charles, V. (2015). Incorporating risk into bank efficiency: A satisficing DEA approach to assess the Greek banking crisis. *Expert Systems with Applications*, *42*(7), 3491-3500.

Woo, C., Chung, Y., Chun, D., Seo, H. and Hong, S. (2015). The static and dynamic environmental efficiency of renewable energy: A Malmquist index analysis of OECD countries. *Renewable and Sustainable Energy Reviews*, (47), 367-376.

Yeh, C. C., Chi, D. J. and Hsu, M. F. (2010). A hybrid approach of DEA, rough set and support vector machines for business failure prediction. *Expert Systems with Applications*, *37*(2), 1535-1541.