Cómo citar este artículo: M. Soto-Acevedo, A. M. Abuchar-Curi, R. A. Zuluaga-Ortiz and E. J. Delahoz-Domínguez, "A Machine Learning Model to Predict Standardized Tests in Engineering Programs in Colombia," in IEEE Revista Iberoamericana de Tecnologias del Aprendizaje, vol. 18, no. 3, pp. 211-218, Aug. 2023, doi: 10.1109/RITA.2023.3301396.

A machine learning model to predict standardized tests in engineering programs in Colombia

Misorly Soto-Acevedo, Alfredo M. Abuchar-Curi, Rohemi A. Zuluaga-Ortiz, Enrique J. Delahoz-Domínguez*

Forecasting of Standardized Test Results for engineering students through Machine Learning

***Abstract*— This research develops a model to predict the results of the national standardized test for Engineering programs in Colombia. The research made it possible to forecast each student's results and thus make decisions on reinforcement strategies to improve student performance. Therefore, a Learning Analytics approach based on three stages was developed: first, analysis and debugging of the database; second, multivariate analysis; and third, the application of machine learning techniques. The results show an association between the performance levels in the Highschool test and the university test results. In addition, the machine learning algorithm that adequately fits the research problem is the Generalized Linear Network Model. For the training stage, the results of the model in Accuracy, AUC, Sensitivity, and Specificity were 0.810, 0.820, 0.813, and 0.827, respectively; in the evaluation stage, the results of the model in Accuracy, AUC, Sensitivity, and Specificity were 0.820, 0.820, 0.827 and 0.813 respectively.**

***Index Terms*— learning Analytics, Machine Learning, Predictive Evaluation, standardized tests.**

I. Introducción

La calidad enfocada como un proceso puede ser estimada objetivamente a través de indicadores de desempeño. Por ejemplo, el análisis longitudinal de los exámenes estandarizados [1] o la interacción entre variables económicas, infraestructura y resultados académicos [2].

Por lo tanto, para lograr la calidad educativa es necesario la implementación de políticas de autoevaluación permanente con el fin de lograr un proceso de mejoramiento continuo [3]. Así, la calidad debe ser evaluada de manera objetiva. A nivel internacional una forma de estimar la calidad en la educación es mediante la certificación Acreditation Board of Engineering and Technology

(ABET), la cual es una entidad no gubernamental, sin ánimo de lucro, conformada por sociedades técnicas y tecnológicas que establecen las políticas del proceso y que acreditan programas en ciencias aplicadas, computación, ingeniería y tecnologías de ingeniería dentro y fuera de los Estados Unidos [4]. Actualmente muchas universidades se encuentran no solamente en procesos de acreditación nacional, sino también en procesos de acreditación internacional, y para ello se deben realizar mediciones en diferentes momentos del proceso aprendizaje y cumplir con los estándares de calidad exigidos.

Por su parte, en Colombia se aplican las pruebas estandarizadas para medir el logro académico de los estudiantes en la educación media, Saber 11, y en la educación superior, Saber Pro. Las pruebas estandarizadas son el principal medio utilizado en Colombia y en el mundo para medir el logro académico [5]. No solamente se aplican las pruebas Saber Pro como indicador de excelencia y calidad para los programas académicos de las universidades, sino que el gobierno nacional, a través del Consejo Nacional de Acreditación -CNA-, ofrece a las universidades y a sus programas académicos la posibilidad de someterse a un proceso de acreditación, ya sea institucional y/o de cada programa académico. Este proceso es voluntario y busca promover el mejoramiento de la calidad, también es una forma para que estas instituciones rindan cuentas ante la sociedad y el Estado sobre el servicio educativo que prestan [6]. En los procesos de acreditación de alta calidad de la universidad o del programa, un aspecto importante es el resultado en las pruebas Saber Pro, razón por la cual es fundamental obtener buenos resultados en estas pruebas para aquellas universidades que aspiren a lograr una acreditación de alta calidad, o inclusive, mantenerla.

Pese a las distintas estrategias del Ministerio de Educación Nacional para el mejoramiento de la calidad de la educación en Colombia, los resultados en las evaluaciones estandarizadas en los niveles medio y básico (Saber 11) demuestran que aún falta mucho por mejorar [7]. El rendimiento académico de los estudiantes es un tema muy complejo y de mucho interés en el ámbito educativo e investigativo y es uno de los mayores retos que actualmente enfrentan las instituciones educativas en la formación básica y universitaria no sólo en Colombia sino a nivel mundial [8], [9].

Desde una perspectiva social las pruebas estandarizadas son una oportunidad para los estudiantes de evidenciar su nivel de conocimiento y además obtener un buen desempeño en estas pruebas permite acceder a becas y descuentos en tasas escolares. Por lo tanto, para un estudiante sería importante tener información objetiva para identificar efectivamente cuáles son las competencias sobre las cuales debería enfocar su proceso de estudio para maximizar el desempeño en el examen estandarizado. Así, la prueba

Manuscrito recibido el día de mes de año; revisado día de mes de año; aceptado día de mes de año.

English versión received Month, day-th, year. Revised Month, day-th, year. Accepted Month, day-th, year.

Misorly Soto-Acevedo, Facultad de Ciencia Básicas, Universidad Tecnológica de Bolívar, Cartagena, Colombia (e-mail: msoto@utb.edu.co)

Rohemi A. Zuluaga-Ortiz, Facultad Ingeniería, Universidad del Sinú, Cartagena, Colombia (e-mail: rohemi.zuluaga@unisinu.edu.co)

Alfredo M. Abuchar-Curi, Facultad de Ingeniería, Universidad Tecnológica de Bolívar, Cartagena, Colombia (e-mail: aabuchar@utb.edu.co)

*Enrique J. Delahoz-Dominguez, Department of Productivity and Innovation, Universidad de la Costa, Barranquilla, Colombia (e-mail: edelahoz13@cuc.edu.co)

SaberPro además se ser un instrumento para evaluar la calidad de la educación superior en Colombia, se convierte en una herramienta de movilidad social para los estudiantes, asociando un alto desempeño en la prueba a un reconocimiento en forma de beneficios económicos, reputacionales y autoconfianza.

Consecuentemente, es pertinente resaltar la importancia de implementar modelos objetivos para lo toma de decisiones en el ámbito educativo, por ejemplo, Timarán, Hidalgo y Caicedo [10] en su investigación analizaron las variables de género, edad, ingreso familiar mensual, tipo de colegio, puntaje obtenido en las pruebas Saber 11 y zona geográfica y su incidencia en el desempeño académico de los estudiantes colombianos que presentaron las pruebas Saber 11 en los años 2015 y 2016. Como conclusión indican que los modelos encontrados, a partir de los datos que se encuentran en las bases de datos del ICFES, son consistentes con la realidad observada.

Por otro lado, Pentel y Kaiva [11] desarrollan una investigación en Estonia, los autores buscan predecir los resultados de los exámenes de los estudiantes basados en notas previas y datos demográficos, además encontrar las materias y características más importantes que contribuyen al resultado del examen estatal donde tuvieron en cuenta variables como el género, lengua materna y las notas en algunas materias desarrollaron dos clases de modelos, continuos y discretos. Para los modelos continuos usaron regresión lineal, K-vecinos más cercanos y bosques aleatorios, mientras que para los modelos discretos usaron regresión logística, K-vecinos más cercanos, C4.5 y bosques aleatorios. Los autores encontraron varias materias que influyeron en el resultado del examen. Como se esperaba, los predictores más importantes incluyeron las materias en las cuales fue tomado el examen, pero no siempre fue así; hubo asignaturas que tuvieron un efecto sorprendente en algunos resultados de los exámenes estatales y algunas materias tuvieron una fuerte repercusión negativa en estos.

Finalmente, Yang y Li [12] proponen un modelo para seguir el progreso de los estudiantes que forma un componente central de los sistemas de e-Learning para lo cual incluyeron variables como las notas de los estudiantes en las materias y las puntuaciones en las habilidades del aprendizaje utilizando redes neuronales de propagación hacia atrás. Los autores indican dentro de sus conclusiones que los resultados del experimento mostraron que el potencial del progreso previsto puede expresar, intuitivamente, el potencial de los estudiantes para progresar en términos de sus habilidades y desempeño. Además, los resultados del experimento también mostraron que las características estimadas de los estudiantes, el desempeño esperado de los estudiantes y la relación causal basada en estos atributos, son correctos. Especialmente, el rendimiento estimado basado en la agrupación generó resultados más precisos y tomó menos tiempo utilizando una menor cantidad de datos de capacitación, lo que también aprobó que los resultados de la clasificación de los estudiantes sean correctos y significativos.

Para el presente estudio, se cuenta con una base de datos de los resultados de las pruebas Saber 11 y Saber Pro de diferentes estudiantes suministrada por el Instituto Colombiano para la Evaluación de la Calidad de la Educación (ICFES). Usando esta base de datos se establece el objetivo del estudio, el cual es crear un modelo de predicción que permita identificar y clasificar, a partir de los resultados de las pruebas Saber 11, las variables socioeconómicas, la universidad y el programa seleccionado por los estudiantes, los resultados en las pruebas Saber Pro en algunos programas de Ingeniería en Colombia, mediante técnicas de aprendizaje automático.

Finalmente, para el desarrollo de este objetivo el trabajo se divide en 4 secciones: marco teórico, metodología, resultados de la investigación, conclusiones y discusión de los resultados.

II. MARCO TEÓRICO

En este apartado se presentan los modelos de aprendizaje automático utilizados en la investigación. Cabe resaltar que la selección de los modelos se realiza de acuerdo con lo presentado en la revisión de literatura ubicada en la introducción. Finalmente, este capítulo tiene 8 incisos desde la A hasta H. Del inciso A al inciso G se presentan los modelos de aprendizaje automático y el inciso H se muestra las métricas de evaluación utilizadas para comparar los modelos.

A. Modelo de los k-vecinos más cercanos

El algoritmo del k vecinos más cercano (KNN, por sus siglas en inglés) es un modelo de machine learning de corte supervisado para la clasificación o regresión. Este algoritmo clasifica creando patrones de comportamientos para identificar la pertenencia de una observación a una categoría específica, esto anterior se realiza mediante el cálculo de las distancias entre observaciones [13]. Por su parte, debido a este análisis de distancias se le conoce como algoritmo de proximidad, similitud o punto más cercano. En otras palabras, dado un nuevo ejemplo, KNN encuentra sus ejemplos más similares, llamados vecinos más cercanos, según una métrica de distancia como la distancia euclidiana, y predice su valor como una agregación de los valores objetivos asociados con sus vecinos más cercanos [14].

B. Modelo Lineal Generalizado en Red

Un modelo lineal generalizado (GLMNET, por sus siglas en inglés) es una modificación flexible de la regresión lineal debido a que permite la variable de salida sea una función de comportamiento no lineal de la entrada mediante una función de activación [15]. Por lo general, los GLMNET hacen uso de modificaciones utilizadas de la regresión logística y la regresión de Poisson con la regresión lineal. Generalmente, los GLM son usados combinando otras técnicas debido a los altos desempeños que obtiene en diferentes escenarios como el reconocimiento de objetos, el reconocimiento de acciones humanas, el análisis sintáctico y la traducción automática [16].

C. Modelo de Random Forest

Los bosques aleatorios (RF, por sus siglas en inglés) es un modelo de aprendizaje automático supervisado ampliamente usado para la clasificación [17]. Este algoritmo utiliza arboles de decisiones para crear múltiples respuestas y a partir de esto, se clasifica de acuerdo con la respuesta con mayor frecuencia [18]. El modelo de bosques aleatorios tiene dos parámetros fundamentas: el número de árboles (k)

y el número de variables utilizadas para dividir los nodos (m).

D. Modelo de Máquina de Soporte Vectorial

Los modelos de máquina de soporte vectorial (SVM, por sus siglas en inglés) son clasificadores binarios que utilizan la función de kernel y se basan en la teoría de aprendizaje automático [19]. Este algoritmo busca identificar el mejor hiperplano de separación que puede clasificar las distintas observaciones. Este método es ampliamente usado para datos que estén correlacionados y sean no lineales.

E. Modelo Naïve Bayes

El modelo Naïve Bayes (NB) se basa en el teorema de Bayes con supuestos de independencia entre los predictores. Esta teoría permite calcular la probabilidad de P(a|b), a partir de P(a), P(b) y P(b|a). Este clasificador parte del supuesto que el efecto del valor de la variable dependiente b en una clase dada a es independiente de los valores de otras variables dependientes [20]. Esta suposición se denomina independencia condicional de la clase. Por su parte, debido a la simplicidad del modelo es ampliamente usado y sorprende con sus resultados incluso mejores que algoritmos sofisticados.

F. Modelos de Árboles

El modelo de árboles o arboles de decisión (DT, por sus siglas en inglés) es uno de los algoritmos de corte supervisado más utilizados actualmente en el aprendizaje automático. Es un método no paramétrico utilizado para clasificación y regresión [21]. Este algoritmo tiene como objetivo crear un modelo que busque estimar el valor o la categoría de una variable de respuesta a través de reglas de decisión simples que nacen de las características de los datos.

G. Modelos Boosting

Los algoritmos de tipo Boosting se centran en predicciones potentes y sofisticadas que se realizan a partir de un único modelo. Estos algoritmos buscan mejorar el poder de predicción entrenando una secuencia de modelos débiles, luego, el modelo final se alimentará de cada lección aprendida por los modelos individuales [22]. Este algoritmo se conoce también como un algoritmo de carácter genérico y no especifico, por lo que es crucial definir el modelo base (por ejemplo, DT, GLMNET, NB, entre otros) y luego se mejorará. Por su parte, en esta investigación se aplicará Boosting al modelo lineal generalizado (GLMBOOSTING).

H. Métricas de evaluación

Los modelos de aprendizaje automático para la clasificación son evaluados mediante las métricas de desempeño que se extraen de una matriz de confusión, esta matriz compara los valores predichos contra los valores reales conocidos (ver Tabla I). Entonces, a partir de la matriz de confusión se extrae [23]: Verdadero Positivo (VP) aparece cuando se predice una observación como positiva y realmente es positiva; Verdadero Negativo (VN) aparece cuando se predice una observación como negativa y realmente es negativa; Falso Positivo (FP) aparece cuando se predice una observación como positiva y realmente es negativa; y finalmente, Falso Negativo (FN) aparece cuando se predice una observación como negativa y realmente es positiva.

TABLA I
MATRIZ DE CONFUSIÓN.

Predicción \ Real	1	0
1	VP	FP
0	FN	VN

De la matriz de confusión se extrae la métrica Accuracy. Esta métrica de desempeño es el porcentaje de predicción correcta para los datos de la prueba. Su resultado se genera en el cociente entre predicciones correctas sobre el total de predicciones (ver Ecuación (1)).

$$Accuracy = \frac{VP + VN}{VP + VN + FP + FV} \quad (1)$$

Otro indicador del desempeño del modelo es el área bajo la curva (AUC) de características operativas del receptor (ROC). La curva ROC presenta la tasa de verdaderos positivos (sensibilidad) en función de la tasa de falsos positivos (especificidad). Por su parte, el valor de ROC maximiza los valores de sensibilidad y especificidad al tiempo. Por su parte, la métrica del área bajo la curva se utiliza para problemas de clasificación binaria y es una de las más utilizadas. El valor AUC de un modelo será aproximadamente igual a la probabilidad de que el modelo clasifique mejor un ejemplo positivo elegido al azar que un ejemplo negativo elegido al azar [24].

Por otro lado, la tasa de verdaderos positivos (sensibilidad) se define como el cociente entre Verdaderos Positivos sobre la suma de los Falsos Negativos y Verdaderos Positivos. Y finalmente, la tasa de verdaderos negativos (especificidad) se define como el cociente entre Verdaderos Negativos sobre la suma de los Falsos Positivos y Verdaderos Negativos.

III. METODOLOGÍA

La presente investigación se enmarca bajo un enfoque de análisis de los datos cuantitativos, en el área conocida como Analítica del Aprendizaje (Learning Analytics), en la que se promueve el uso de los datos generados en los diferentes niveles del proceso, en este caso educativo, con el fin de crear herramientas que soporten la toma de decisiones objetivas [25] para los diferentes grupos de interés del proceso: estudiantes, profesores, gestores educativos, entidades gubernamentales, entidades de acreditación, entre otras.

Por su parte, la investigación se divide en tres etapas: Análisis de Componentes Principales (PCA), entrenamiento de modelos y finalmente, evaluación de modelos.

Por lo anterior, se puede concluir que esta investigación es de tipo aplicativo teniendo en cuenta que busca inferir, por medio de aprendizaje automático, los resultados de los estudiantes en la prueba Saber Pro a partir de los resultados obtenidos en las pruebas Saber 11 y algunos aspectos específicos de la universidad seleccionada.

Es importante señalar que el software utilizado para el análisis de los datos, construcción y evaluación de los modelos es R [26].

A. *Descripción de la base de datos*

Para la investigación se cuenta con una base de datos suministrada por el ICFES correspondiente al 2018 de las pruebas Saber Pro y los correspondientes resultados previos de la Saber 11 [27], que incluye los resultados de 12410 estudiantes de diferentes programas de ingeniería a nivel nacional. En la Tabla II se muestran los programas seleccionados y el número de estudiantes registrados en la base de datos, así como información de si el programa está acreditado o no. La base de datos incluye estudiantes de diferentes universidades a nivel nacional.

Adicional, es importante señalar que de la base de datos diseñada por Delahoz-Dominguez et al. [27], la información utilizada corresponde a las variables de las pruebas Saber 11, las variables socioeconómicas, la universidad y el programa seleccionado por los estudiantes (ver Tabla IV).

Finalmente, previo al desarrollo de las etapas de la investigación, se prepara la base de datos manteniendo la fidelidad de la información con el objetivo de mejorar el rendimiento de los modelos. Este acondicionamiento incluyó la creación de categorías por cada variable con el fin de reducir la variabilidad de la información presentada por variables.

TABLA II
PROGRAMAS A REALIZAR EL MODELO DE PREDICCIÓN

Programa	Número de estudiantes totales	Acreditada
Ingeniería Civil	3320	No
Ingeniería Eléctrica	278	No
Ingeniería Química	1001	No
Ingeniería Electrónica	849	Si
Ingeniería Industrial	5318	Si
Ingeniería Mecánica	1136	Si
Ingeniería Mecatrónica	78	Si

IV. RESULTADOS

En este capítulo se divide en tres etapas de acuerdo con lo presentado en el capítulo de metodología: Análisis de Componentes Principales, entrenamiento de modelos y evaluación de modelos.

A. *Etapa 1: Análisis de Componentes Principales (PCA)*

El PCA se realiza con una rotación ortogonal (rotación de varianza máxima). En la Fig. 1 se presenta el gráfico biplot y se observa una clara diferenciación por niveles del puntaje global en las pruebas Saber Pro, además se observa claramente la relación directa entre los resultados de las diferentes áreas evaluadas en las pruebas Saber Pro y los niveles de rendimiento definido en el puntaje global. De igual forma se puede apreciar en el primer cuadrante la relación fuerte entre los estudiantes que tuvieron buenos resultados en las pruebas Saber 11 y los niveles de desempeño de la prueba Saber Pro (ver Tabla III).

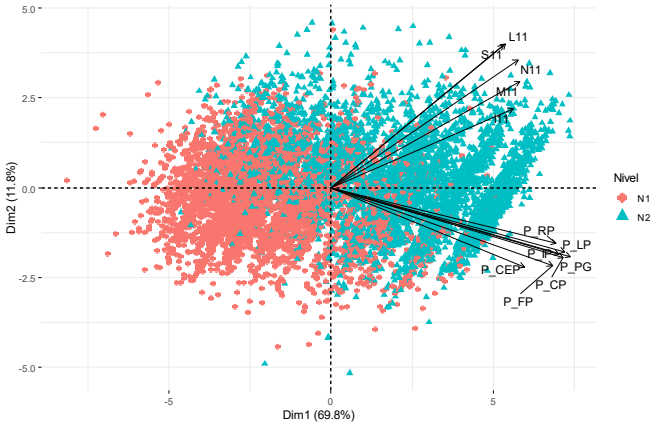

Fig. 1: Plano 1-2 del Análisis PCA

Consecuentemetne, en la Tabla III se presenta la distribución de las observaciones del plano 1-2 del análisis PCA. Se observa que los estudiantes con mayor desempeño académico se encuentran ubicados en mayor proporción en los cuadrantes uno y cuatro, con aproximadamente 74.4% de los estudiantes que están en nivel dos. Por su parte, en los cuadrantes dos y tres se encuentran aproximadamente 83.1% de los estudiantes que están en nivel uno.

TABLA III
DISTRIBUCIÓN DE LAS OBSERVACIONES EN EL PCA PLANO 1-2

Nivel	Cuadrante			
	I	II	III	IV
1	5.1%	36.6%	46.5%	11.9%
2	37.5%	20.2%	5.4%	36.9%

Finalmente, en la primera etapa exploratoria de la base de datos, se presenta la Fig. 2 con la información de la importancia de las variables.

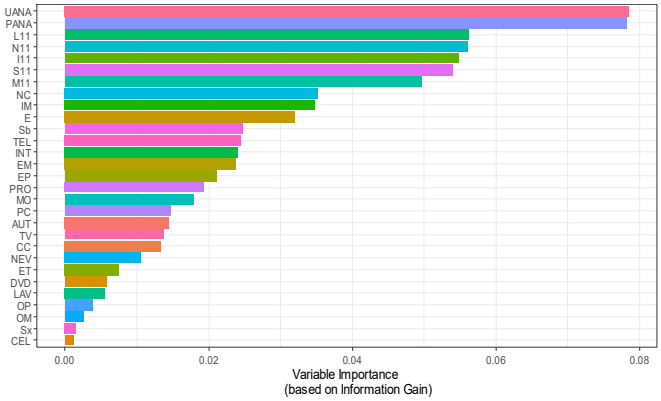

Fig. 2: Importancia de las variables.

Dentro de los resultados presentados en la Fig. 2 se puede evidenciar que para la construcción de un modelo que sea capaz de predecir el resultado de las pruebas Saber PRO es importante tener en cuenta la acreditación de la universidad, del programa académico y los resultados del estudiante en su evaluación de Saber 11.

B. *Etapa 2: Entrenamiento de modelos*

Para el entrenamiento de los modelos se utilizó el método de validación cruzada y se trabajó con una base de entrenamiento correspondiente al 70% de los datos, quedando 3300 en el N1 y 3576 en el N2, y una base para la

TABLA IV
VARIABLES DEL ESTUDIO

Variables	Código	Clase	Categorías
Genero	Sx	Cat	Femenino (F); Masculino (M)
Educación Padre	EP	Cat	Educación profesional completa (epc); Educación profesional incompleta (epi); Ninguno (n); No sabe (ns); Postgrado (pos); Primaria completa (pc); Primaria incompleta (pi); Secundaria (Bachillerato) completa (bc); Secundaria (Bachillerato) incompleta (bi); Técnica o tecnológica completa (tc); Técnica o tecnológica incompleta (ti)
Educación Madre	EM	Cat	
Ocupación Padre	OP	Cat	Empleado con cargo como director o gerente general (G); Empleado de nivel auxiliar o administrativo (A); Empleado de nivel directivo (D); Empleado de nivel técnico o profesional (T); Empleado obrero u operario (EO); Empresario (E); Hogar (H); Otra actividad u ocupación (OA); Pensionado (P); Pequeño empresario (PE); Profesional independiente (I); Trabajador por cuenta propia (CP)
Ocupación Madre	OM	Cat	
Estrato	E	Cat	Estrato 1 (E1); Estrato 2 (E2); Estrato 3 (E3); Estrato 4 (E4); Estrato 5 (E5); Estrato 6 (E6)
Sisbén	Sb	Cat	Está clasificada en otro nivel del SISBEN (SO); Nivel 1 (S1); Nivel 2 (S2); Nivel 3 (S3); No está clasificada por el SISBEN (SN)
Personas casa	Per	Num	Una; Dos; Tres; Cuatro; Cinco; Seis; Siete; Ocho; Nueve; Diez; Once; Doce o más
Tipo de piso	TP	Cat	Cemento, gravilla, ladrillo; Madera burda, tabla, tablón; Madera pulida, baldosa, tableta, mármol, alfombra; Tierra, arena
Familia tiene internet	INT	Cat	Si; No
Familia tiene servicio TV	TV	Cat	Si; No
Familia tiene computador	PC	Cat	Si; No
Familia tiene lavadora	LAV	Cat	Si; No
Familia tiene microondas	MO	Cat	Si; No
Familia tiene automóvil	AUT	Cat	Si; No
Familia tiene DVD	DVD	Cat	Si; No
Familia tiene nevera	NEV	Cat	Si; No
Familia tiene celular	CEL	Cat	Si; No
Familia tiene teléfono	TEL	Cat	Si; No
Ingreso familiar mensual	IM	Cat	Menos de 1 SMLV; Entre 1 y menos de 2 SMLV; Entre 2 y menos de 3 SMLV; Entre 3 y menos de 5 SMLV; Entre 5 y menos de 7 SMLV; Entre 7 y menos de 10 SMLV; 10 o más SMLV
Estudia o trabaja actualmente	ET	Cat	No; Si, 2 horas o más a la semana; Si, menos de 2 horas a la semana
Nombre del colegio	N.C	Cat	
Naturaleza del colegio	NC	Cat	No oficial (NO); Oficial (OF)
Carácter del colegio	CC	Cat	Académico (ACA); No aplica (NA); Técnico (TEC); Técnico/Académico(T/A)
Puntaje Matemáticas Saber11	M11	Num	Rango: 0-100
Puntaje Lectura Crítica Saber11	L11	Num	Rango: 0-100
Puntaje Sociales Ciudadanas_Saber11	S11	Num	Rango: 0-100
Puntaje Ciencias Naturales Saber11	N11	Num	Rango: 0-100
Puntaje Inglés Saber11	I11	Num	Rango: 0-100
Nombre de universidad	UN	Cat	
Puntaje Razonamiento Cuantitativo Saber Pro	RP	Num	Rango: 0-100
Puntaje Lectura Crítica Saber Pro	LP	Num	Rango: 0-100
Puntaje Competencias Ciudadanas Saber Pro	CP	Num	Rango: 0-100
Puntaje Inglés Saber Pro	IP	Num	Rango: 0-100
Puntaje Comunicación Escrita Saber Pro	CEP	Num	Rango: 0-100
Puntaje global Saber Pro	PG	Num	Rango: 0-300
Formulación de Proyectos de Ingeniería Saber Pro	FP	Num	Rango: 0-300
Programa académico	PRO	Cat	Construcciones Civiles (CCI); Ingeniería Aeronáutica (AER); Ingeniería Catastral y Geodesia (CYG); Ingeniería Civil (CIV); Ingeniería de Control (CON); Ingeniería de Producción (PRO); Ingeniería de Productividad y Calidad (PYC); Ingeniería de Transporte y Vías (TYV); Ingeniería Eléctrica (ELE); Ingeniería Electromecánica (ELM); Ingeniería Electrónica (ETR); Ingeniería Electrónica y Telecomunicaciones (ETRT); Ingeniería en Automática Industrial (AUTI); Ingeniería en Automatización (IAU); Ingeniería en Control (CON); Ingeniería en Control y Automatización Industrial (CYA); Ingeniería Industrial (IND); Ingeniería Mecánica (MEC); Ingeniería Mecatrónica (MTR); Ingeniería Química (QUI); Ingeniería Topográfica (TOP)
Universidad acreditada	UANA	Cat	UA (Universidad acreditada); UNA (Universidad no acreditada)
Programa acreditado	PANA	Cat	PA (Programa acreditado); PNA (Programa no acreditado)

Nota. Las categorías para la columna Clase son: Numérica (Num) y Categórica (Cat).

evaluación de los modelos correspondiente al 30% de los datos, quedando 1453 en el N1 y 1494 en el N2. Debido a que se observa una distribución muy pareja no se balancean los datos. Adicional, la configuración de los diferentes modelos se presenta en la Tabla V.

Entonces, la parametrización del modelo KNN, el valor k indica el número adecuado de grupos para el conjunto de datos. Para la parametrización del modelo GLMBOOST, el valor Mstop indica el número de iteraciones que realizará el modelo. Para la parametrización del modelo GLMNET, el valor de la penalización de la red elástica está controlado por Alpha, por su parte, el valor lambda controla la fuerza general de la penalización del modelo; por su parte en la Tabla V este modelo presenta un valor de Alpha igual 0.55 y lambda igual 0.0057, esto indica la configuración de un modelo con penalización lasso. Así, para la parametrización del modelo RF, el valor Mtry indica el número de variables para dividir en cada nodo; el valor splitrule indica el modo para estimar la probabilidad en la clasificación, y finalmente, min.node.size indica el tamaño mínimo del nodo. Para la parametrización del modelo SVM, el valor de Sigma actúa como un parámetro de suavización y el valor de costo (C) controla la complejidad del límite entra los vectores de soporte. Para la parametrización del modelo NB, el valor de Laplace indica un suavizado de tipo aditivo para el modelo (si es cero, no hay suavizado) y, además, el uso de kernel (ya sea ajustado o no). Finalmente, para la parametrización del modelo DT se requiere el factor de complejidad y esto disminuye la falta de ajuste global en un factor de acuerdo con el valor determinado.

TABLA V
PARAMETRIZACIÓN DE LOS MODELOS DE LA INVESTIGACIÓN.

Modelo	Parametrización
KNN	k = 43
GMLBOOST	Mstop = 250
GLMNET	Alpha = 0.55; lambda = 0.0057
RF	Mtry = 7; splitrule = extratrees; min.node.size = 1
SVM	Sigma = 0.11; c = 0.25
NB	Laplace = 0; Kernel ajustado
DT	Cp = 0.0018

La Fig. 3 muestra resultados comparativos de los modelos de la investigación, y aquí, el Modelo Lineal Generalizado en Red (GLMNET) y el modelo GLMBOOST son los que entregan mejores resultados.

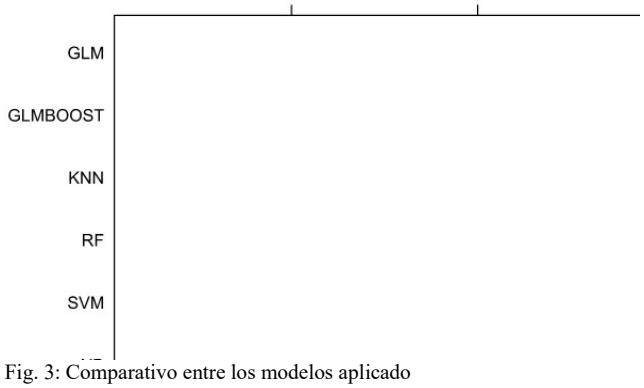

Fig. 3: Comparativo entre los modelos aplicado

Como se puede observar en la Fig. 4 en el modelo GLMNET la distribución de los resultados de la validación cruzada en el diagrama de caja y bigote muestra menor variabilidad frente al modelo GLMBOOST, además es un modelo que tiene menor costo computacional debido a su baja complejidad, en comparación con el modelo GLMBOOST, razón por la cual se considera el mejor modelo para predecir el nivel de desempeño en las pruebas Saber Pro.

En la Tabla VI se presenta los resultados de la etapa de entrenamiento de los modelos. Claramente el algoritmo con desempeño más alto, AUC, es modelo lineal generalizado (GLMNET) y por su parte, el algoritmo con menor desempeño es árbol de decisión (DT). Mientras que los resultados del modelo GLMNET en Accuracy, AUC, Sensitivity y Specificity son 0.810, 0.820, 0.813 y 0.827 respectivamente, los resultados del modelo DT son 0.770, 0.803, 0.837 y 0.770.

TABLA VI
MÉTRICA DEL DESEMPEÑO DE LOS MODELOS EN EL ENTRENAMIENTO

	Accuracy		AUC		Sensitivity		Specificity	
Model	Mean	sd	Mean	sd	Mean	sd	Mean	sd
KNN	0.810	0.015	0.810	0.014	0.813	0.015	0.807	0.016
GLMNET	0.810	0.015	0.820	0.014	0.813	0.016	0.827	0.017
RF	0.810	0.015	0.813	0.014	0.815	0.020	0.812	0.019
SVM	0.810	0.014	0.816	0.014	0.825	0.017	0.807	0.017
NB	0.790	0.015	0.790	0.015	0.742	0.014	0.837	0.014
DT	0.770	0.015	0.803	0.014	0.837	0.031	0.770	0.028
GLMBOOST	0.810	0.014	0.818	0.014	0.808	0.014	0.829	0.016

C. Etapa 2: Fase de evaluación

Finalmente, la última fase de la metodología consiste en evaluar los modelos entrenados con el fin de observar su desempeño para predecir nuevas observaciones. Consecuentemente, en la Tabla VII se presenta el desempeño de los modelos en la etapa de evaluación. Se observa que en esta etapa nuevamente el Modelo Lineal Generalizado en Red (GLMNET) es el que tiene mejor capacidad de predicción, por su parte, el modelo de Naïve Bayes (NB) obtuvo en menor desempeño en esta etapa. Mientras que los resultados del modelo GLMNET en Accuracy, AUC, Sensitivity y Specificity son 0.820, 0.820, 0.827 y 0.813 respectivamente, los resultados del modelo NB son 0.790, 0.792, 0.937 y 0.742.

TABLA VII
MÉTRICAS DE DESEMPEÑO EN LA ETAPA DE EVALUACIÓN DE LOS MODELOS

	Accuracy		AUC		Sensitivity		Specificity	
Model	Mean	sd	Mean	sd	Mean	sd	Mean	sd
KNN	0.810	0.015	0.810	0.014	0.807	0.013	0.813	0.014
GLMNET	0.820	0.014	0.820	0.014	0.827	0.012	0.813	0.013
RF	0.813	0.015	0.813	0.014	0.812	0.013	0.815	0.170
SVM	0.816	0.014	0.817	0.014	0.807	0.012	0.825	0.013
NB	0.790	0.015	0.792	0.015	0.837	0.014	0.742	0.014
DT	0.803	0.015	0.805	0.014	0.770	0.013	0.837	0.013
GLMBOOST	0.818	0.014	0.818	0.014	0.829	0.012	0.801	0.013

Finalmente, nótese que en la Fig. 4 y Fig. 5 se ilustra el desempeño de los modelos mediante la curva ROC, en estas figuras se observa que el Modelo Lineal Generalizado en Red (GLMNET) tiene mayor área y esto, se traduce en mayor capacidad predictiva.

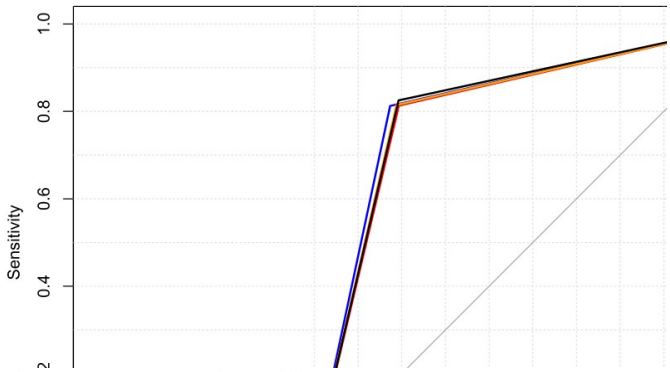

Fig. 4: Curva ROC para los modelos GLMNET, KNN, RF y SVM

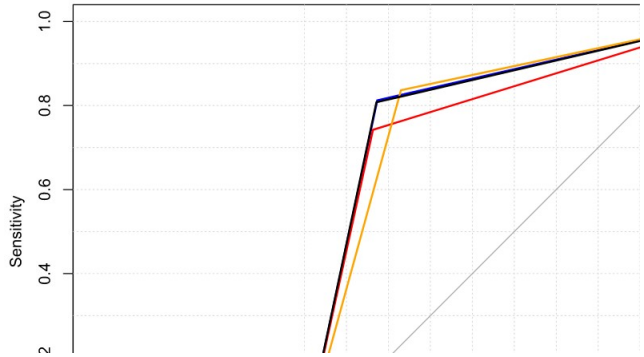

Fig. 5: Curva ROC para los modelos GLMNET, NB, DT y GLMBOOST

V. Discusión

En esta sección se presenta una interpretación comparativa de los principales hallazgos asociados al desarrollo de un modelo predictivo de las pruebas estandarizadas SaberPro en Colombia. Los resultados, tienen implicaciones más allá del ajuste de los datos a un modelo de aprendizaje supervisado y servirían como una herramienta para la gestión de los recursos universitarios. Así, identificar tempranamente el desempeño académico futro de los estudiantes permitiría gestionar eficientemente los recursos educativos.

En primer lugar, el análisis descriptivo multivariable a través del PCA evidencia diferencias en los resultados de matemáticas en los alumnos con un buen y mal desempeño en las pruebas SaberPro. Estas diferencias se pueden expliar por diferentes razones. i) Diferencias por ubicación geográfica, Estudios previos muestran que la heterogeneidad regional es una fuente importante de desigualdad regional en los países emergentes [28] . ii) Debido a la brecha entre las universidades privadas y públicas [29].

En segundo lugar, el análisis de la importancia de las variables para generar la clasificación del modelo de machine learning indica que la acreditación de la universidad es el factor más importante, estos resultados son congruentes con la investigación de , donde utilizan un modelo de regresión logística para clasificar los programas de ingeniería en Colombia usando un modelo de regresión lineal [30].

En tercer lugar, desde el punto de vista de la capacidad de los modelos de machine learning para predecir los resultados del desempeño en los resultados utilizados en nuestra investigación, los resultados de la precisión y AUC fueron del 82%. En comparación con el trabajo desarrollado por Kaur et al. [31], donde construyen un modelo de Redes Neuronales para estimar el desempeño académico de los estudiantes utilizando variables económicas, familiares y educativas, obteniendo una precisión de 75%. Consecuentemente, Jishan et al. [32] desarrollan un modelo con los algoritmos Naïve Bayes y Redes Neuronales utilizando variables netamente académicas, obteniendo un valor AUC del 81%. Por su parte, Lau et al. [33] desarrollaron un modelo de Redes Neuronales utilizando 11 variables de entrada, dos capas de neuronas ocultas y una capa de salida, para esto, emplean el algoritmo de Levenberg-Marquardt como regla de entrenamiento de retro-propagación, alcanzando una precisión de 84.8%. Por lo tanto, los resultados de nuestra investigación asociados a la capacidad predictiva de los modelos de machine learning implementados son competitivos y ajustados a resultados de investigaciones con enfoques similares en educación. Sin embargo, la comparación de los modelos va mucho más allá de las cifras de las métricas de desempeño, desde el punto de vista metodológico es importante señalar que las investigaciones mencionadas solo utilizan variables del contexto académico, generando un sesgo por la asociación del desempeño estudiantil a un único contexto. Por el contrario, en nuestra investigación se tuvo en cuenta variables socioeconómicas, asimilando el proceso de aprendizaje como la interacción de múltiples variables del entorno del estudiante.

Por último, desde un enfoque aplicado nuestra investigación contribuye al espectro de conocimiento de los modelos predictivos que son útiles para la gestión de las instituciones de educación superior; sin embargo, esto no limita la posibilidad de extrapolar la metodología a otros niveles académicos y áreas de conocimiento. Teniendo en cuenta lo anterior, es vital señalar que la educación siendo el principal motor de desarrollo para las sociedades debe contar con herramientas que sean capaces de identificar las falencias en el proceso de aprendizaje de los educandos. La herramienta propuesta no solo será de apoyo para la predicción de los niveles del desempeño académico de los estudiantes, sino también, que al identificar estos niveles se podrán identificar las variables críticas que aportan la mayor cantidad de información para explicar el desempeño académico y así, será posible intervenir sobre los distintos aspectos que son oportunidad de mejora (condiciones socioeconómicas, falencias académicas, entre otras variables que se consideren dentro del estudio).

VI. Conclusión

En la presente investigación se utilizó una base de datos de los resultados de las pruebas Saber 11 y Saber Pro de diferentes estudiantes, suministrada por el ICFES por medio de la cual se creó un modelo de aprendizaje automático para la predicción del desempeño académico de los estudiantes que ingresan a una Universidad.

Es así como la investigación permite realizar un pronóstico de los resultados de cada estudiante y de esta manera poder tomar decisiones de estrategias de refuerzo para mejorar los resultados en las pruebas Saber Pro. En definitiva, el modelo propuesto permite identificar grupos de estudiantes que podrían tener bajos resultados en las pruebas Saber Pro con el fin de crear estrategias durante todo el proceso de enseñanza-aprendizaje para estos estudiantes y de esta manera mejorar su formación académica, mejorar los

resultados de las pruebas Saber Pro y potencializar su desempeño profesional y laboral.

Finalmente, la principal limitación del estudio es el posible sesgo presentado al tener solo programas de ingeniería, sin embargo, esta metodología se puede ser aplicada a otras áreas de estudio. En este orden de idea, dentro de las investigaciones futuras se puede considerar la aplicación a otros campos de estudio y un análisis de entradas difusas de las variables para observar los cambios en la variable de respuesta ante un espectro más amplio y dinámico de las entradas.

AGRADECIMIENTOS

Agradecimiento especiales al Instituto Colombiano para la evaluación de la calidad de la educación por facilitar los datos utilizados en esta investigación.

REFERENCIAS

[1] J. Aparicio, S. Perelman, y D. Santín, «Comparing the evolution of productivity and performance gaps in education systems through DEA: an application to Latin American countries», *Oper. Res.*, jun. 2020, doi: 10.1007/s12351-020-00578-2.

[2] D. Visbal-Cadavid, M. Martínez-Gómez, y F. Guijarro, «Assessing the efficiency of public universities through DEA. A case study», *Sustainability*, vol. 9, n.º 8, p. 1416, 2017.

[3] M. Campo, «Capital humano para el avance colombiano, Editorial en Educación superior 20». p. 1, 2012.

[4] L. Valencia, H. Trefftz, y I. Delgado-González, «Acreditación Internacional de Carreras de Ingeniería», *Educ. En Ing.*, vol. 15, n.º 29, pp. 28-33, 2020.

[5] R. Hoyos Martínez, M. Borja Maturana, R. Gómez Lorduy, y G. Casadiegos Aponte, «Calidad en la escuela vs. prácticas pedagógicas: los relatos como medio para la reflexión y la emancipación de los maestros en tiempos de la eficiencia», *Esfera*, vol. 5, n.º 2, p. 16, 2015.

[6] J. Guerrero, «La acreditación de alta calidad en Colombia», 2018.

[7] L. A. Sanabria James, M. C. Pérez Almagro, y L. E. Riascos Hinestroza, «Pruebas de evaluación Saber y PISA en la Educación Obligatoria de Colombia», *Educ. Siglo XXI*, vol. 38, n.º 3 Nov-Feb, pp. 231-254, 2020, doi: 10.6018/educatio.452891.

[8] L. A. Melo-Becerra, J. E. Ramos-Forero, y P. O. Hernández-Santamaría, «La educación superior en Colombia: situación actual y análisis de eficiencia», *Desarro. Soc.*, vol. 2017, n.º 78, pp. 59-111, 2017, doi: 10.13043/DYS.78.2.

[9] Y. Bernal y C. Rodríguez, «Factores que Inciden en el Rendimiento Escolar de los Estudiantes de la Educación Básica Secundaria», Universidad Cooperativa de Colombia, 2017.

[10] R. Timarán-Pereira, J. Caicedo-Zambrano, y A. Hidalgo-Troya, «Árboles de decisión para predecir factores asociados al desempeño académico de estudiantes de bachillerato en las pruebas Saber 11°», *Rev. Investig. Desarro. E Innov.*, vol. 9, n.º 2, pp. 363-378, 2019, doi: 10.19053/20278306.v9.n2.2019.9184.

[11] A. Pentel y L. L. Kaiva, «Predicting Students' State Examination Results based on Previous Grades and Demographics», *11th Int. Conf. Inf. Intell. Syst. Appl. IISA 2020*, 2020, doi: 10.1109/IISA50023.2020.9284401.

[12] F. Yang y F. W. B. Li, «Study on student performance estimation, student progress analysis, and student potential prediction based on data mining», *Comput. Educ.*, vol. 123, n.º April, pp. 97-108, 2018, doi: 10.1016/j.compedu.2018.04.006.

[13] S. Zhang, X. Li, M. Zong, X. Zhu, y R. Wang, «Efficient kNN Classification With Different Numbers of Nearest Neighbors», *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 29, n.º 5, pp. 1774-1785, may 2018, doi: 10.1109/TNNLS.2017.2673241.

[14] A. Moldagulova y R. Bte. Sulaiman, «Using KNN algorithm for classification of textual documents», en *2017 8th International Conference on Information Technology (ICIT)*, Amman, Jordan, may 2017, pp. 665-671. doi: 10.1109/ICITECH.2017.8079924.

[15] P. K. Dunn y G. K. Smyth, «Chapter 5: Generalized Linear Models: Structure», en *Generalized Linear Models With Examples in R*, P. K. Dunn y G. K. Smyth, Eds. New York, NY: Springer, 2018, pp. 211-241. doi: 10.1007/978-1-4419-0118-7_5.

[16] D. Zhang, «A Coefficient of Determination for Generalized Linear Models», *Am. Stat.*, vol. 71, n.º 4, pp. 310-316, oct. 2017, doi: 10.1080/00031305.2016.1256839.

[17] E. De La Hoz, R. Zuluaga, y A. Mendoza, «Assessing and Classification of Academic Efficiency in Engineering Teaching Programs», *J. Effic. Responsib. Educ. Sci.*, vol. 14, n.º 1, Art. n.º 1, mar. 2021, doi: 10.7160/eriesj.2021.140104.

[18] G. Louppe, «Understanding Random Forests: From Theory to Practice», *ArXiv14077502 Stat*, jul. 2014, Accedido: 23 de julio de 2019. [En línea]. Disponible en: http://arxiv.org/abs/1407.7502

[19] S. Suthaharan, «Support Vector Machine», en *Machine Learning Models and Algorithms for Big Data Classification: Thinking with Examples for Effective Learning*, S. Suthaharan, Ed. Boston, MA: Springer US, 2016, pp. 207-235. doi: 10.1007/978-1-4899-7641-3_9.

[20] D. Buzic y J. Dobsa, «Lyrics classification using Naive Bayes», en *2018 41st International Convention on Information and Communication Technology, Electronics and Microelectronics (MIPRO)*, Opatija, may 2018, pp. 1011-1015. doi: 10.23919/MIPRO.2018.8400185.

[21] K. David Kolo, S. Adepoju, y J. Kolo Alhassan, «A Decision Tree Approach for Predicting Students Academic Performance», *Int. J. Educ. Manag. Eng.*, vol. 5, n.º 5, pp. 12-19, oct. 2015, doi: 10.5815/ijeme.2015.05.02.

[22] T. Chen y C. Guestrin, «XGBoost: A Scalable Tree Boosting System», en *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining - KDD '16*, San Francisco, California, USA, 2016, pp. 785-794. doi: 10.1145/2939672.2939785.

[23] A. J. Hung, J. Chen, y I. S. Gill, «Automated Performance Metrics and Machine Learning Algorithms to Measure Surgeon Performance and Anticipate Clinical Outcomes in Robotic Surgery», *JAMA Surg.*, vol. 153, n.º 8, p. 770, ago. 2018, doi: 10.1001/jamasurg.2018.1512.

[24] Z. H. Hoo, J. Candlish, y D. Teare, «What is an ROC curve?», *Emerg. Med. J.*, vol. 34, n.º 6, pp. 357-359, jun. 2017, doi: 10.1136/emermed-2017-206735.

[25] D. Gašević, V. Kovanović, y S. Joksimović, «Piecing the learning analytics puzzle: a consolidated model of a field of research and practice», *Learn. Res. Pract.*, vol. 3, n.º 1, pp. 63-78, 2017, doi: 10.1080/23735082.2017.1286142.

[26] R Core Team, *R: A language and environment for statistical computing, R Foundation for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing, 2013. Accedido: 15 de abril de 2020. [En línea]. Disponible en: http://www.polsci.wvu.edu/duval/PS603/Notes/R/fullrefman.pdf

[27] E. Delahoz-Dominguez, R. Zuluaga, y T. Fontalvo-Herrera, «Dataset of academic performance evolution for engineering students», *Data Brief*, vol. 30, p. 105537, jun. 2020, doi: 10.1016/j.dib.2020.105537.

[28] P. Herrera-Idárraga, E. López-Bazo, y E. Motellón, «Regional Wage Gaps, Education and Informality in an Emerging Country: The Case of Colombia», *Spat. Econ. Anal.*, vol. 11, n.º 4, pp. 432-456, oct. 2016, doi: 10.1080/17421772.2016.1190462.

[29] J. Moreno-Gómez, J. Calleja-Blanco, y G. Moreno-Gómez, «Measuring the efficiency of the Colombian higher education system: a two-stage approach», *Int. J. Educ. Manag.*, vol. 34, n.º 4, pp. 794-804, ene. 2020, doi: 10.1108/IJEM-07-2019-0236.

[30] E. J. Delahoz-Dominguez, S. Guillen-Ibarra, T. Fontalvo-Herrera, E. J. Delahoz-Dominguez, S. Guillen-Ibarra, y T. Fontalvo-Herrera, «Análisis de la acreditación de calidad en programas de ingeniería industrial y los resultados en las pruebas nacionales estandarizadas, en Colombia», *Form. Univ.*, vol. 13, n.º 1, pp. 127-134, feb. 2020, doi: 10.4067/S0718-50062020000100127.

[31] P. Kaur, M. Singh, y G. S. Josan, «Classification and Prediction Based Data Mining Algorithms to Predict Slow Learners in Education Sector», *Procedia Comput. Sci.*, vol. 57, pp. 500-508, 2015, doi: 10.1016/j.procs.2015.07.372.

[32] S. T. Jishan, R. I. Rashu, N. Haque, y R. M. Rahman, «Improving accuracy of students' final grade prediction model using optimal equal width binning and synthetic minority over-sampling technique», *Decis. Anal.*, vol. 2, n.º 1, p. 1, dic. 2015, doi: 10.1186/s40165-014-0010-2.

[33] E. T. Lau, L. Sun, y Q. Yang, «Modelling, prediction and classification of student academic performance using artificial neural networks», *SN Appl. Sci.*, vol. 1, n.º 9, p. 982, ago. 2019, doi: 10.1007/s42452-019-0884-7.

Misorly Soto Acevedo es ingeniera de Sistemas, Politécnico Gran Colombiano. Especialista en Estadística Aplicada, Universidad del Norte. Magister en Estadística Aplicada, Universidad Tecnológica de Bolívar. Profesora catedrática en la Facultad de Ciencias Básicas de la Universidad Tecnológica de Bolívar.

Alfredo Miguel Abuchar Curi es Ingeniero mecánico, Universidad Tecnológica de Bolívar (UTB). Magister en

Ingeniería mecánica, universidad de los Andes. Magister en Estadística Aplicada, Universidad Tecnológica de Bolívar. Profesor de Tiempo Completo de la UTB con más de 25 años de experiencia. Actualmente secretario de la Facultad de Ingeniería de la UTB. Su área de investigación es mecánica de fluidos.

Rohemi Alfredo Zuluaga Ortiz es magister en ingeniería de la Universidad Tecnológica de Bolívar (UTB). Actualmente es profesor de la Universidad del Sinú. Sus áreas de investigación son eficiencia y Learning Analytics.

Enrique De La Hoz es máster en Investigación de Operaciones por la Universitat de Barcelona y la UPC. Actualmente es profesor de la Universidad de la Costa en el departamento de productividad e Innovacion. Sus áreas de investigación son Learning Analytics, sistemas de recomendación y minería de datos a gran escala.