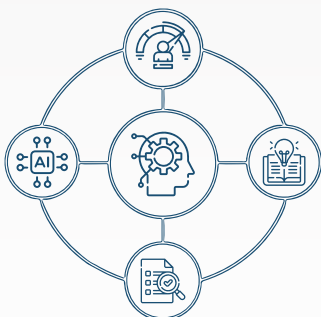




Fase de análisis de modelos de
APRENDIZAJE AUTOMÁTICO
implementados en pruebas académicas estandarizadas

ANALYSIS PHASE OF MACHINE LEARNING MODELS
IMPLEMENTED IN STANDARDIZED ACADEMIC TESTS

SENNOVA

Sistema de Investigación,
Desarrollo Tecnológico e Innovación

Esp. Nathalia Orozco Morales

Corporación Universitaria Minuto de Dios
nathalia.orozco@uniminuto.edu



Esp. Ing. Luis Fernando Rincón Amarillo

Corporación Universitaria Minuto de Dios
luis.rincon-am@uniminuto.edu



Esp. Claudia Liliana Piedrahita Castaño

Corporación Universitaria Minuto de Dios
claudia.piedrahita@uniminuto.edu



Esp. Jeferson Daniel Gallego Agudelo

Corporación Universitaria Minuto de Dios
jeferson.gallego@uniminuto.edu



Esp. Laura Sarita Gallego Martínez

Corporación Universitaria Minuto de Dios
laura.gallego-m@uniminuto.edu



Modelos Predictivos Enfocados a la Ejecución y Mejoras de Resultados Académicos de la Corporación Universitaria Minuto de Dios.

Este proyecto se centra en el desarrollo de modelos predictivos con algoritmos de aprendizaje automático enfocados a la mejora del desempeño académico de los estudiantes de diferentes Instituciones de Educación Superior (IES), tomando como referencia los resultados de las pruebas Saber Pro suministrados por el Instituto Colombiano para la Evaluación de la Calidad de la Educación - ICFES entre los periodos 2021-2 a 2023-1.

Palabras clave: aprendizaje automático, inteligencia artificial, educación, rendimiento académico, aprendizaje.

1.1. Abstract.

The Ministry of National Education (MEN) of Colombia measures the quality of higher education by applying standardized tests, more specifically the Saber Pro tests, with which the cognitive, communicative, mathematical and socioeconomic skills of final year students are evaluated. propaedeutic cycles of programs at this level.

Nowadays, artificial intelligence (AI) algorithms have been developed that focus on machine learning models based on databases in which patterns and relationships between variables are identified or thereby generating a design base of descriptive and predictive analytical models.

In this article, a bibliographic and conceptual review is made about machine learning algorithms and their applications in the academic sector and then present the initial stage (descriptive analysis) and its associated partial results of the project developed by students of the specialization in Artificial Intelligence belonging to the research hotbed Predictive Models Focused on the Execution and Improvement of Academic Results of the Minuto de Dios University Corporation.

1. Resumen

El Ministerio de Educación Nacional (MEN) de Colombia realiza la medición de la calidad de educación superior aplicando pruebas estandarizadas, más concretamente las pruebas Saber Pro, con las cuales se evalúan las habilidades cognitivas, comunicativas, matemáticas y socioeconómicas de los estudiantes de últimos ciclos propedéuticos de programas de este nivel.

Ahora bien, en la actualidad se han desarrollado algoritmos de inteligencia artificial (IA) que se enfocan en modelos de aprendizaje automático partiendo de bases de datos en las cuales se identifican patrones y relaciones entre variables o bien, generando con esto una base de diseño de modelos analíticos descriptivos y predictivos.

En este artículo se hace una revisión bibliográfica y conceptual acerca de algoritmos de aprendizaje automático y sus aplicaciones en el sector académico para luego presentar la etapa inicial (análisis descriptivo) y sus resultados parciales asociados del proyecto desarrollado por estudiantes de la especialización en Inteligencia Artificial pertenecientes al semillero de investigación

The project focuses on the development of predictive models with machine learning algorithms focused on improving the academic performance of students from different Higher Education Institutions (IES), taking as reference the results of the SABER – PRO tests provided by the Institute. Colombian for the Evaluation of the Quality of Education - ICFES between the periods 2021-2 to 2023-1.

Keywords: machine learning, artificial intelligence, education, academic performance, learning.

2. Introducción

El Ministerio de Educación Nacional (MEN) de Colombia, oficializa la medición de la calidad de la educación superior de las Instituciones de Educación Superior (IES) mediante la aplicación de pruebas académicas estandarizadas a sus estudiantes, de carácter oficial y obligatorio, llamadas Pruebas Saber Pro, las cuales se toman como instrumento para la medición de la calidad de la educación formal que los estudiantes que culminan los programas académicos recibieron en su proceso de formación (Ministerio de Educación Nacional – MEN, 2009).

El Instituto Colombiano para la Evaluación de la Educación (ICFES) emplea los resultados de las pruebas Saber Pro como modelos psicométricos para evaluar a los estudiantes de pregrado que se encuentran sobre el 75% de su plan de estudios, con el fin de evidenciar la calidad en sus habilidades académicas adquiridas para emprender su vida laboral y los factores conexos asociados su aprendizaje.

Las pruebas evalúan y miden de forma cuantitativa (entre 0 y 300 puntos) cada competencia genérica y específica del estudiante en diferentes áreas de conocimiento, que se asumen como saberes mínimos propuestos en el pregrado, además de un cuestionario socioeconómico que aporta información adicional del estudiante como características del núcleo familiar y su hogar, estrato, acceso a internet y tiempo dedicado al entretenimiento

deportivo (Guías de orientación examen Saber Pro, 2023), y es de allí de donde se obtiene la información para realizar el análisis y medición del rendimiento académico de los estudiantes y por ende, inferir la calidad del programa académico de la IES a la que pertenece.

Según Gómez, 2013 las competencias generales en la educación superior que se establecen desde las diferentes áreas del conocimiento se definen como las habilidades intelectuales que fomentan el aprender a aprender, habilidades sociales que suelen utilizarse para el trabajo en equipo, competencias de aprendizaje o educativas (comunicativas, interpretativas, argumentativas y propositivas), sin embargo, este tipo de evaluaciones estandarizadas presentan desafíos como la subjetividad en la corrección, la falta de retroalimentación y la capacidad de adaptarse a las necesidades individuales de los estudiantes.

Los autores Gil-Vera y Quintero-López, 2021 y Zárate-Valderrama, et al. 2021 demuestran que los resultados de este tipo de pruebas se ven afectados por temas relacionados al contexto personal del estudiante, tales como culturales, sociales, tecnológicos, económicos y de bases académicas que a su vez son importantes para evitar la deserción escolar y el aprendizaje sin fundamento.

Por tal motivo, en este artículo se realiza en primer lugar una revisión bibliográfica y conceptual de los modelos de aprendizaje automático o Machine Learning que pueden ser implementados en las bases de datos generados a partir de los resultados de las pruebas Saber Pro y que sirvan específicamente para realizar un análisis estadístico descriptivo sobre las variables de entrada, hallar una correlación entre las mismas y realizar una evaluación de dicha relación para identificar patrones en los datos que sirvan como base de un modelo predictivo dependiente de entradas como la calificación obtenida en las habilidades del estudiante así como su contexto social y factores de su entorno que puedan llegar a influenciar en la valoración cuantitativa que se obtiene en la prueba.

Luego, se presentará la descripción y resultados parciales de la fase inicial (análisis descriptivo) del proyecto Machine Learning que se está desarrollando dentro del semillero de investigación adscrito a la Corporación Universitaria Minuto de Dios: Modelos Predictivos Enfocados a la Ejecución y Mejoras de Resultados Académicos, en donde se implementan algoritmos de aprendizaje automático, tomando como insumo la base de datos de los resultados de las pruebas Saber Pro de los estudiantes de varias IES que presentaron la prueba en el periodo 2021-2 y 2023-1 y que sirven para el entrenamiento de dichos modelos.

Lo anterior tiene como fin generar un modelo descriptivo y predictivo que permita determinar patrones de relación en las variables de entrada para generar una variable de salida relacionada con el rendimiento académico del estudiante y así poder implementar estrategias más precisas y decisiones organizacionales orientadas a fortalecer su desempeño al momento de la presentación de la prueba y así responder a la pregunta: ¿Es viable implementar un modelo de Machine Learning que permita relacionar variables objetivas y subjetivas de un estudiante que presenta una prueba estandarizada para obtener información que permita determinar su rendimiento?

Al final del artículo se pretende discutir el objetivo general del mismo: Evaluar la viabilidad de un modelo Machine Learning que permita relacionar los resultados cuantitativos con las variables cualitativas relacionadas a un estudiante de IES que presenta una prueba académica estandarizada, para poder obtener información sobre su rendimiento académico.

3. Metodología

Para el presente artículo y sus respectivos resultados y discusión, se presenta un tipo de investigación con enfoque mixto, puesto que para desarrollar modelos analíticos y de aprendizaje automático en la IA, se toma como insumo información proveniente de datos tanto cuantitativos como lo son las calificaciones numéricas de las pruebas Saber Pro como datos cualitativos como los usados

para determinar el contexto social, cultural y económico, entre otros, de cada estudiante que presenta la prueba.

Lo anterior también se argumenta con el hecho de que se parte de un problema bien definido el cual es el de la necesidad de poder realizar un análisis estadístico preliminar sobre una base de datos confiable para luego poder desarrollar un modelo usando algoritmos de Machine Learning o aprendizaje automático que pueda relacionar variables tanto cuantitativas como cualitativas de cada estudiante que presenta una prueba estandarizada de medición de desempeño académico y así definir un modelo que logre realizar una descripción y predicción en los resultados obtenidos por los estudiantes que presentan la prueba o bien, permita identificar patrones para diseñar estrategias de aprendizaje a implementar en los estudiantes para mejorar su rendimiento académico, relacionando directamente las habilidades analíticas de los estudiantes y la inferencia que tiene su contexto económico, social y cultural sobre los resultados obtenidos en la prueba.

La metodología que se aplica en esta fase preliminar del proyecto es de carácter experimental puesto que se parte de un trabajo explicativo haciendo una revisión bibliográfica de lo conceptual para luego aplicar dicho conocimiento sobre un proceso que no tiene casos vistos hasta el momento. Con ello se pretende determinar la relación causa – efecto que se tiene entre las variables independientes o causales y la variable dependiente de ellas (Lerma, 2016).

Para la fase inicial del proyecto se establecieron las siguientes etapas:

- **Etapas de revisión bibliográfica y conceptual:** Esta tiene como base el contexto de la inteligencia artificial aplicada al campo de la educación, y como fin, identificar los algoritmos o modelo de IA más convenientes a implementar en el contexto académico a abordar. Se realizó una búsqueda de los modelos de Machine Learning a nivel global y local, clasificando los algoritmos que son más relevantes para la aplicación en el proyecto. Es importante mencionar que las

pruebas estandarizadas tipo Saber Pro solo se aplican en el contexto local y aunque hay un número variado de estudios a nivel de educación que aportan a los instrumentos del proyecto, esta revisión se centró en la búsqueda de modelos que puedan cumplir el objetivo principal propuesto.

- **Etapas de obtención de base de datos:** En esta fase se utilizó como base de datos la información de los resultados de las competencias genéricas y específicas de la prueba Saber Pro de estudiantes de las diferentes IES del país en el periodo 2021-2 a 2023-1 y suministradas por el ICFES. En total se contaron con 5 bases de datos correspondientes a los periodos de estudio las cuales suman un total de 754.029 registros (estudiantes con sus respectivos datos). Dicha fuente de información es insumo más que suficiente para desarrollar un modelo analítico descriptivo de las variables de estudio.

- **Etapas de procesamiento de la base de datos:** En esta fase se realiza un proceso clave y fundamental en cuanto al manejo y presentación de los datos para modelamiento Machine Learning, que es el procesamiento de los datos, en el cual se estructuran los datos en tablas con el mismo formato, se concatenan dichas tablas, se identifica y hace limpieza de los datos errados, se identifican los datos faltantes y dependiendo de su cantidad se eliminan o imputan, y por último se hace una estandarización de su formato de presentación. Esta base de información procesada es el insumo adecuado para un futuro entrenamiento de los algoritmos Machine Learning a implementar.

- **Etapas de Análisis:** En esta fase se diseña y establece el modelo estadístico descriptivo y se hace una identificación y aproximación de la correlación entre las variables de interés.

4. Resultados y discusión

4.1. Revisión bibliográfica.

En el artículo de García-González et al. 2019 se pretende desarrollar una herramienta que aplique la metodología de descubrimiento

del conocimiento en bases de datos (KDD) y utilice la minería de datos para correlacionar la información del historial académico de los estudiantes con los resultados de la prueba Saber Pro. Se presentan los resultados de un estudio que utilizó redes neuronales para predecir con alta precisión los resultados de la prueba Saber Pro en Colombia.

La investigación se basó en la extracción de conocimiento de una base de datos que incluye el desempeño académico de los estudiantes en áreas relacionadas con la prueba. Se encontró una correlación entre el desempeño académico y los resultados de Saber Pro, lo que sugiere que esta metodología es útil para identificar patrones en los datos y mejorar los resultados de la prueba.

Tras la construcción de la herramienta y la realización de las pruebas en diferentes escenarios (reales, artificiales y uso de datos masivo) se resaltan los siguientes hechos: (1) En todos los casos analizados la herramienta pudo relacionar las asignaturas y pudo calcular el promedio ponderado de cada estudiante; (2) En todos los casos analizados el algoritmo aplicó la correlación y pudo predecir de manera satisfactoria los puntajes de la prueba de razonamiento cuantitativo, comparándolo con el resultado real del estudiante; (3) En algunos casos la predicción no fue exacta, pero los resultados son acordes con el desempeño del estudiante.

En el documento de Contreras, et al. 2023 se proporciona una visión general sobre como el Machine Learning (aprendizaje automático) puede ser usado para identificar condiciones no satisfactorias y predecir el rendimiento académico de los estudiantes en educación superior.

El objetivo es identificar algoritmos de Machine Learning existentes para la evaluación del rendimiento académico. Se concluye que al ingresar técnicas de aprendizaje automático se determinan factores y variables que influyen en el rendimiento de los estudiantes. En última instancia, este enfoque refleja la creciente importancia de la tecnología avanzada en la mejora de la educación y el proceso de aprendizaje.

Froment y de-Besa Gutiérrez, 2022 se centran en cómo las opiniones de los estudiantes sobre el comportamiento de sus profesores en el aula tienen un impacto significativo en su proceso de aprendizaje. Esto afecta aspectos importantes como su motivación académica, que se considera uno de los factores clave para el éxito de los estudiantes. En otras palabras, si los estudiantes tienen una buena percepción de sus profesores, es más probable que estén motivados para aprender y tengan un mejor rendimiento académico. La conclusión principal de este estudio es que existe una relación positiva entre la credibilidad del profesorado, y que esta influye de manera significativa en el compromiso, la satisfacción y la motivación académica en estudiantes universitarios de Educación Infantil y Educación Primaria en la Universidad de Cádiz.

En el artículo Jurado, 2020 la investigación se basa en la creación de un modelo tecnológico que utiliza una red neuronal artificial (NN) y un algoritmo de procesamiento de lenguaje natural. Este modelo busca automatizar la revisión crítica en las evaluaciones virtuales de los estudiantes de pregrado, con el propósito de mejorar el proceso educativo en respuesta a la contingencia que ocasionó la pandemia global del COVID-19. Estas innovaciones se han implementado a través de la plataforma Virtual Moodle, que hoy en día es un recurso esencial para la enseñanza y el aprendizaje en la Universidad Privada Domingo Savio en todas sus carreras.

Igualmente, esta NN desempeña una función adicional al evaluar a los estudiantes para lograr identificar a aquellos con un perfil psicológico adecuado para especializarse en asignaturas de alto nivel. Las redes neuronales artificiales se utilizan como herramienta de clasificación y predicción en la educación virtual, un área de estudio que es aplicable para diversos campos de la educación superior, basándose en el procesamiento, modelado, pruebas y validación estadística, probabilidad y matemáticas puras, dejándonos ver que la tecnología artificial garantiza la calidad de las evaluaciones en

proyectos académicos, alineándose con el concepto del triángulo del conocimiento.

En la comparativa hecha por Mercado Polo, et al. 2015 de las redes neuronales artificiales, se emplearon dos enfoques: el primero es sobre el perceptrón multicapa (MLP) y el segundo sobre las funciones de base radial (RBF), utilizando el algoritmo resilient backpropagation para el MLP y una combinación del algoritmo de los K-Means y el método de la matriz pseudoinversa para la RBF.

La implementación de estas redes se llevó a cabo mediante un sistema de cliente-servidor con vistas a su posible integración en aplicaciones en tiempo real en el futuro. Para la evaluación de estas redes, se utilizaron conjuntos de datos con diversas características y tamaños. Según los resultados obtenidos, se concluye que, para la integración de técnicas de inteligencia computacional en sistemas web, las RBF son preferibles debido a su mejor rendimiento en tiempos de ejecución, aunque es importante destacar que ambos tipos de redes neuronales ofrecen resultados similares en cuanto a la calidad de las respuestas. Los resultados de RBF y MLP son muy similares, respaldados por herramientas estadísticas que indican con un 95% de confianza que no existe una diferencia significativa entre ambos métodos.

4.2. Revisión conceptual.

Algoritmo de Aprendizaje Automático

Este término se refiere al sistema de IA que aprende a través de datos expuestos por un experto y suele mejorar su desempeño en una tarea específica a través del tiempo (Vuralli, et al, 2023). Los modelos de Machine Learning utilizan métodos de clasificación, regresión y agrupamiento para realizar sus predicciones, dentro de estos se pueden encontrar los algoritmos Naive Bayes, Random Forest, Árboles de Decisión, Clusterización y Redes Neuronales Artificiales, cada uno de estos modelos suele ser utilizado según el número de datos y el problema a solucionar (Rojas Rubio & Meneses Villegas, 2022).

Tipos de aprendizaje automático

El aprendizaje automático ofrece muchas ventajas en el contexto educativo comparado con otros métodos tradicionales. Este tipo de algoritmos pueden ser utilizados debido a los avances tecnológicos y a los datos que se reflejan desde los dispositivos electrónicos que se evidencian actualmente (Valero, 2017). Las técnicas utilizadas en la actualidad han permitido a los investigadores modelar problemas de rendimiento académico, análisis de sentimientos y comportamiento social que sirven para automatizar los procesos estudiantiles y analizar los comportamientos sociales (Selwyn, 2019), es así como se explican los tres tipos de AA utilizados:

- Según Pérez del Barrio, et al. 2022 el aprendizaje reforzado es un algoritmo que ejecuta acciones utilizando el proceso de ensayo y error.
- El Aprendizaje automático no supervisado contiene data que no está etiquetada, por lo cual, el algoritmo se centra en encontrar características similares en el conjunto de datos y realizar agrupación.
- En el aprendizaje automático supervisado los datos se encuentran etiquetados, por tal motivo, el algoritmo genera una clasificación o regresión según lo disponga el problema. Según Mamani-Coaquira, et al. 2021 la clasificación se refiere a asignar a los datos de entrada categorías específicas binarias o multiclase, mientras que la regresión se utiliza para hallar la relaciones entre variables.

Desarrollo de competencias digitales

P. Gómez, et al. 2023 indica que la formación de competencias digitales en estudiantes universitarios es muy importante para mejorar las capacidades de la Inteligencia Artificial (IA) en la personalización del aprendizaje. La adaptación de la IA conlleva a que las necesidades individuales de cada estudiante están alineadas con la premisa de que los estudiantes deben adquirir habilidades tecnológicas para mantenerse al día con los avances en la educación.

Esto implica el aumento de la capacidad de los sistemas de tutoría inteligente para proporcionar retroalimentación y recursos personalizados, generando mejoras sustanciales en la eficacia del proceso de aprendizaje. Sin embargo, la implementación de la IA en la educación superior plantea preocupaciones y desafíos, entre ellos la necesidad de adaptar los sistemas a las diversas realidades y necesidades de las instituciones universitarias, así mismo, surge la inquietud por la posible sustitución de la enseñanza humana por sistemas de aprendizaje basados en computadoras.

Inteligencia Artificial (IA)

Ruíz, 2023. La inteligencia artificial (IA) ha experimentado un rápido desarrollo en los últimos años, impulsado por avances en la potencia informática, la disponibilidad de grandes conjuntos de datos y la evolución de nuevos algoritmos. La IA se define como la capacidad de las máquinas para realizar tareas que normalmente requieren inteligencia humana, como el razonamiento, el aprendizaje, la creatividad y la planificación y con ello ha transformado diversos aspectos de nuestra vida cotidiana, desde la forma en que compramos en línea hasta cómo interactuamos con nuestros dispositivos móviles.

A medida que la IA continúa avanzando, es probable que veamos aún más aplicaciones innovadoras en diversas industrias, con un impacto significativo en la eficiencia, la personalización y la calidad de los servicios ofrecidos. Sin embargo, también plantea desafíos éticos y sociales que deben abordarse para garantizar un desarrollo y aplicación responsable de esta tecnología. Algunas aplicaciones clave de la inteligencia artificial se expresan a continuación:

- **Búsquedas en la Web:** Los motores de búsqueda aprovechan la IA para mejorar la relevancia de los resultados, aprenden de la enorme cantidad de datos generados por los usuarios ajustando constantemente sus algoritmos para proporcionar respuestas más precisas y personalizadas.

- **Asistentes Personales Digitales:**

Los Smartphone emplean la IA para ofrecer experiencias personalizadas a los usuarios. Los asistentes virtuales, mediante el procesamiento del lenguaje natural y el aprendizaje automático, responden preguntas, brindan recomendaciones y ayudan a organizar las actividades diarias de los usuarios.

- **Traducciones Automáticas:**

Los programas de traducción de idiomas hacen uso de la IA, tanto para texto escrito como oral. Este enfoque mejora la calidad de las traducciones y permite una comunicación más fluida entre hablantes de diferentes idiomas. Además, la IA también se aplica en tareas como el subtitulado automático, mejorando la accesibilidad y la comprensión en entornos multilingües.

La Inteligencia Artificial en el aprendizaje

El impacto de la IA en los entornos de aprendizaje personal de los estudiantes universitarios es un tema de creciente interés en la educación. En el artículo “Inteligencia artificial y ambientes personales de aprendizaje: atentos al uso adecuado de los recursos tecnológicos por estudiantes universitarios” de Castillejos López, 2022 se hace reflexión en el uso de hacks académicos y herramientas de inteligencia artificial que facilitan prácticas poco éticas en el aula.

El objetivo principal del artículo es dar un alcance frente al impacto de la inteligencia artificial con relación a los entornos de aprendizaje para estudiantes universitarios, específicamente en relación a las herramientas para uso de mejores prácticas, resaltando la ética. La metodología utilizada es principalmente de análisis y reflexión sobre las implicaciones de estas prácticas en el pensamiento crítico y creativo de los estudiantes (Salas & Tello, 2023).

En cuanto a la necesidad de generar un proyecto enfocado a enseñar la IA en las instituciones con la finalidad de contribuir a la integración de los aspectos humanos en programas de formación dirigidos a los jóvenes, se tiene lo siguiente:

“Debemos renovar este compromiso, ya que nos acercamos a una época en que la inteligencia artificial – la convergencia de tecnologías emergentes – transformará todos los aspectos de nuestras vidas” (La Inteligencia Artificial y su uso en la Educación Inclusiva, 2023).

Pruebas de Estado Saber Pro

Son exámenes estandarizados administrados a nivel estatal o nacional para evaluar el rendimiento educativo a nivel superior. Los estudiantes de educación superior en Colombia deben presentar las pruebas Saber Pro sobre un 75% de su nivel de pregrado y se constituye como un instrumento estandarizado para la evaluación externa de la calidad de la educación superior. El examen se ofrece en modalidad electrónica en un sitio de aplicación definido por el ICFES y para su aplicación, el estudiante debe dirigirse a su Universidad/Facultad para realizar la preinscripción certificada, posteriormente, el ICFES le proporcionará un usuario y clave para darse de alta en la plataforma y realizar la inscripción vía formulario (Guías de orientación examen Saber Pro, 2023).

La Inteligencia Artificial en la Transformación Digital

En gran variedad de estudios se demuestra que la utilización de la IA constituye una herramienta muy importante y de gran relevancia para el éxito de los negocios, ya que evalúa una serie de características que no son tan perceptibles pero que generan valor y conllevan a un aumento en la capacidad para el aprendizaje de los modelos de negocios digitales (Rojas Rubio & Meneses Villegas, 2022).

En la nueva revolución industrial (Industrias 4.0) se retan nuestras prácticas comunes al ser comparadas con la nueva era digital para la integración sistemática de objetos inteligentes (Internet de las Cosas – IoT), sistemas de cómputo con mayores prestaciones, como lo son particularmente los servicios de nube y la utilización de tecnología de la información y las comunicaciones (TIC), entre otras. Lo anterior no plantea un gran

interrogante sobre aquellas buenas o mejores prácticas que podemos relucir para que haya una mayor generación de beneficios con lo que nos ofrecen las nuevas tecnologías emergentes (Rojas Rubio & Meneses Villegas, 2022).

4.3. Obtención de base de datos.

En esta etapa se logró obtener una generosa base de datos provenientes de 5 fuentes de información suministradas por el ICFES y que corresponde a los periodos académicos de estudio (del 2021-2 al 2023-1) y detalladas de la siguiente manera:

1. SaberPro_2021.txt, con un total de registros de 246498.
2. SaberPro_Específicas_20212.txt con un total de registros de 198167.
3. SaberPro_2022-1.TXT con un total de registros de 42734.
4. DATA_ICFES_SABERPROPROFESIONAL_2022-2.TXT con un total de registros de 127342.
5. DATA_ICFES_SABERPROPROFESIONAL_20231.TXT con un total de registros de 139288.

Se consolido un total de 754.029 registros, es decir, ficheros de estudiantes con sus respectivos datos. La presentación de la información proveniente de las fuentes se presenta en formatos TXT como se muestra en la figura 1:

Figura 1.
Presentación y formato de las fuentes de información obtenidas



Nota. ICFES 2021

4.4. Procesamiento de Base de Datos.

Para realizar el procesamiento de la información de la base de datos y en sí, la normalización del formato de la base de datos (.txt) se emplean las herramientas de desarrollo del lenguaje de programación de Python y sus respectivas librerías, todo dentro del entorno de desarrollo Google Colaboratory (servicio en la nube prestado por Google).

En esta etapa, se realizó el procedimiento de concatenación de los 5 archivos de la base de datos obtenida en la fase anterior, es decir, se unieron los 5 archivos en uno solo y se estandarizó la información del nuevo archivo, a tal punto de que los datos de interés tuvieran el mismo formato y fueran compatibles. Luego, con librerías y funciones propias del lenguaje Python se procede a hacer la limpieza o cleaning de la información.

En dicho proceso se identifican los datos erróneos o mal digitados, los datos faltantes o perdidos y se proceden a eliminar puesto que la cantidad de estos datos no es significativa en comparación con la base de datos total y por consiguiente, no alterará el resultado final. A continuación, en la figura 2 se muestra la base de datos limpia que se obtuvo:

Figura 2.
Base de datos obtenida después del proceso de cleaning realizado

Column	ESTU_TIPDOCUMENTO	ESTU_NACIONALIDAD	ESTU_GENERO	ESTU_FECHANACIMIENTO	ESTU_EXTERIOR	PERIODO	ESTU_CONSECUATIVO	ESTU_ESTUDIANTE	ESTU_PAIS_RESIDE	...
0	0	CC	COLOMBIA	F	14/01/2000	NO	2012	EK002120038705	ESTUDIANTE	COLOMBIA
1	1	CC	COLOMBIA	F	09/11/1984	NO	2012	EK002120032222	ESTUDIANTE	COLOMBIA
2	2	CC	COLOMBIA	F	22/12/1987	NO	2012	EK002120070304	ESTUDIANTE	COLOMBIA
3	3	CC	COLOMBIA	F	16/09/1993	NO	2012	EK002120035227	ESTUDIANTE	COLOMBIA
4	4	CC	COLOMBIA	M	28/02/1996	NO	2012	EK002120054325	ESTUDIANTE	COLOMBIA
...	...	...	...	...	...	...	...	...	...	...
452015	130283	CC	COLOMBIA	F	19/10/2001	NO	2031	EK002120236306	ESTUDIANTE	COLOMBIA
452016	130284	CC	COLOMBIA	M	25/01/1999	NO	2031	EK002120239645	ESTUDIANTE	COLOMBIA
452017	130285	CC	COLOMBIA	M	24/01/1996	NO	2031	EK002120115566	ESTUDIANTE	COLOMBIA
452018	130286	CC	COLOMBIA	M	30/03/2001	NO	2031	EK002120111790	ESTUDIANTE	COLOMBIA
452019	130287	CC	COLOMBIA	F	15/07/1999	NO	2031	EK002120174486	ESTUDIANTE	COLOMBIA

Nota. Elaboración propia.

4.5. Análisis Descriptivo.

Al disponer de una base de datos limpia, sólida y confiable (sin errores ni valores perdidos), se procede a realizar una serie de pruebas de funciones estadísticas que poseen las librerías Python que por excelencia manejan modelos estadísticos. Estas librerías son Pandas y NumPy.

Figura 3.
Matriz de correlación general de las variables
de la base de datos obtenida

[illegible]

En esta matriz relaciona una variable con las demás y se asigna un número entre cero y uno que indica el grado de relación entre estas variables. Entre más próximo sea el valor a 1, mayor correlación existirá entre dichas variables.

Figura 4.
Distribución del estrato socioeconómico de los estudiantes que presentaron la prueba Saber Pro entre el periodo 2021-2 y 2023-1

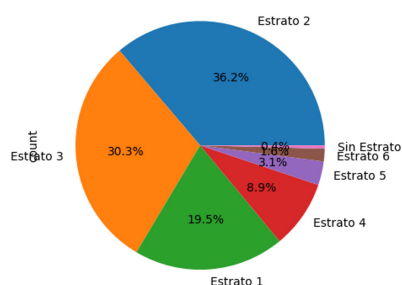


Figura 5.
Distribución del tipo de IES de la que provienen los estudiantes que presentaron la prueba Saber Pro entre el periodo 2021-2 y 2023-1

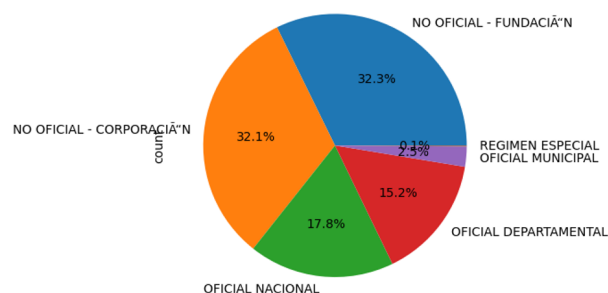
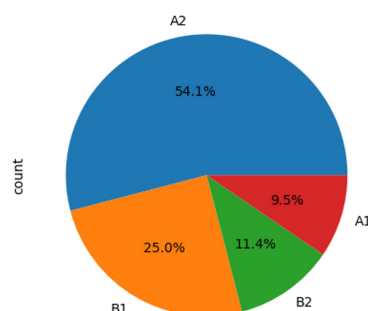


Figura 6.
Distribución del nivel de inglés que poseen los estudiantes que presentaron la prueba Saber Pro entre el periodo 201-2 y 2023-1



4.6. Discusión.

Diferentes investigaciones coinciden en que diferentes factores influyen sobre los resultados de las pruebas Saber Pro ya que el rendimiento académico entre los estudiantes se asocia al estrato social, variables demográficas, la calidad del sueño y el uso de dispositivos electrónicos (Adelantado-Renau, et al. 2019).

Además, Plant, et al. 2005 menciona que la cantidad de tiempo invertido en el estudio es un predictor del éxito académico, por tanto, entre más tiempo se dedique a la retroalimentación de los saberes, mejores serán las calificaciones del estudiante, sin embargo, es importante mencionar que todo estudiante debe fortalecer sus habilidades personales.

En Colombia las pruebas Saber Pro se sustentan en bases teóricas que buscan evaluar habilidades cognitivas, matemáticas, comunicativas entre otras asociadas a cada carrera, sin embargo, es importante revisar un cuestionamiento frente a evaluar la eficacia y adecuación de estas bases teóricas (Amalia & Sanabria, 2013).

En primer lugar, es necesario considerar que las habilidades cognitivas y comunicativas son aspectos fundamentales en la formación de cada estudiante y es así como se pretende que con estas pruebas se pueda evaluar la oportunidad y eficacia, además de que los estudiantes hayan adquirido dichas habilidades que son necesarias para desenvolverse en el ámbito académico y profesional.

En segundo lugar, la inclusión de la evaluación de estas habilidades obtenidas durante el transcurso de la carrera evaluadas en las pruebas Saber Pro refleja una preocupación por formar profesionales con un sólido conocimiento, sin embargo, Corvalán, 2018 indica que existe una serie de variables a tener en cuenta que están afectando esta medición y evaluación y he aquí donde se encuentra que a través de la inteligencia artificial podemos apalancarnos para detectar y determinar aquellos aspectos en los cuales debemos fortalecer o profundizar.

Así, al entrelazar las Tecnologías de la Información y la Comunicación (TIC's) emergentes con procesos y métodos académicos, se logra abordar aspectos relevantes en cuanto al desarrollo de las habilidades académicas e identificación de estilos de aprendizaje en el entorno educativo, donde se puede ofrecer a los estudiantes experiencias de aprendizaje adaptadas a cada población, lo que se traduce en una mayor calidad en la formación, compromiso y retención estudiantil.

La necesidad de identificar estas habilidades que pueden ser innatas o adquiridas por el estudiante dentro del entorno educativo puede ayudar a identificar variables no tan perceptibles por el docente

que pueden llegar a ser un aporte importante para identificar las falencias en cuestión de aprendizaje y por tanto garantizar una educación adecuada para cada individuo del sistema educativo (Giró-Gracia, et al, 2022).

También se tiene en cuenta que los elementos que conforman este estudio se basaron en la premisa de que los modelos IA puede revolucionar la forma de determinar y fortalecer la calidad de la educación superior al proporcionar herramientas analíticas que puedan definir estrategias más personalizadas y efectivas para el aprendizaje, lo que beneficia a los estudiantes y a las IES al garantizar un enfoque claro y medible logrando un mejor resultado del rendimiento estudiantil (Rojas, et al, 2022) y por resorte, mejorar los indicadores de medición de la calidad en la formación.

Además, es necesario analizar la efectividad de la utilización de la inteligencia artificial sin perder el factor humano pero sí aprovechando estos nuevos mecanismos para fortalecer aquellas posibles variables que no son tan perceptibles, pero que sí generan un gran valor al momento de la aplicabilidad en las pruebas Saber Pro, lo anterior, ya que estas pruebas como herramienta para evaluar la calidad de la educación superior pueden proporcionar información sobre el desempeño individual de los estudiantes, los programas educativos y las instituciones.

Ahora bien, dentro de la inteligencia artificial, el desarrollo de modelos Machine Learning puede llevar al uso de algoritmos robustos y adaptables, y su implementación en el contexto de las pruebas estandarizadas Saber Pro necesita ser cuidadosamente evaluado, ya que se requiere una discusión profunda sobre cómo estos modelos pueden complementar o mejorar las bases teóricas establecidas, sin comprometer la equidad, la validez de las pruebas y el factor humano.

5. Conclusiones

Con la digitalización de los entornos educativos, incluyendo los métodos de enseñanza aprendizaje el número de datos

es cada vez mayor y por tanto los procesos que se pueden realizar con estos, por tanto, es una puerta para el análisis y la predicción de los datos.

Los resultados de la búsqueda bibliográfica permiten identificar modelos de IA utilizados en la educación, así como el impacto de estos algoritmos en la búsqueda de patrones y tendencias que podrían reflejarse en estudios futuros.

El análisis descriptivo realizado en las etapas iniciales del proyecto permitió identificar variables clave, como el estrato socioeconómico, el tipo de IES de procedencia del estudiante y su nivel de inglés. Estas variables resultan fundamentales para entender mejor los factores que influyen en el rendimiento académico y deberían ser consideradas en futuros análisis más exhaustivos para profundizar en las causas y correlaciones observadas.

De acuerdo con los resultados obtenidos, más precisamente, la base de datos obtenida y su respectiva organización, es totalmente viable la aplicación de un modelo estadístico y descriptivo sobre las variables cualitativas y cuantitativas de la misma base, para posteriormente desarrollar un modelo de entrenamiento Machine Learning e implementar los algoritmos adecuados en la obtención de una relación entre ellas e identificar patrones que permitan idealizar estrategias enfocadas al mejoramiento del rendimiento académico de los estudiantes de las Instituciones de Educación Superior.

6. Referencias Bibliográficas

Ministerio de Educación Nacional - MEN. Decreto Numero 3963: Examen de Estado de Calidad de la Educación Superior. En: Ley 1324. República de Colombia (2009).

Adelantado-Renau, M., Díez-Fernandez, A., Beltran-Valls, M. R., Soriano-Maldonado, A., & Moliner-Urdiales, D. (2019). The effect of sleep quality on academic performance is mediated by Internet use time: DADOS study. *Jornal de Pediatria*, 95(4), 410–418. <https://doi.org/10.1016/j.jped.2018.03.006>

Amalia, C., & Sanabria, C. (2013). Una Mirada A Los Resultados De Aprendizaje. *Edu.Co*. https://www.colombiaaprende.edu.co/sites/default/files/files_public/2022-07/Una%20mirada%20a%20los%20resultados%20de%20aprendizaje

Aznarte, J. L. (2020). Consideraciones éticas en torno al uso de tecnologías basadas en datos masivos en la UNED. *RIED Revista Iberoamericana de Educación a Distancia*, 23(2), 237. <https://doi.org/10.5944/ried.23.2.26590>

Bonami, B., Piazzentini, L., & Dala-Possa, A. (2020). Education, Big Data and Artificial Intelligence: Mixed methods in digital platforms. *Comunicar*, 28(65), 43–52. <https://doi.org/10.3916/c65-2020-04>

Castillejos López, B. (2022). Inteligencia artificial y los entornos personales de aprendizaje: atentos al uso adecuado de los recursos tecnológicos de los estudiantes universitarios. *Educación*, 31(60), 9–24. <https://doi.org/10.18800/educacion.202201.001>

Contreras, L. E., Tarazona Bermúdez, G. M., & Alemán Cardona, A. P. (2023). Machine Learning aplicado al rendimiento académico en educación superior: factores, variables y herramientas. Editorial Universidad Distrital Francisco José de Caldas.

Corvalán, J. G. (2018). Inteligencia artificial: retos, desafíos y oportunidades – Prometea: la primera inteligencia artificial de Latinoamérica al servicio de la Justicia. *Revista de Investigações Constitucionais*, 5(1), 295. <https://doi.org/10.5380/rinc.v5i1.55334>

- Escandón Barbosa, D. M., & Ayala, A. H. (2017).** El compromiso exportador en Colombia: un análisis de redes neuronales. 24, 362–388. <https://www.proquest.com/scholarly-journals/el-compromiso-exportador-en-colombia-un-analisis/docview/2013604511/se-2>
- Froment, F., & de-Besa Gutiérrez, M. (2022).** La predicción de la credibilidad docente sobre la motivación de los estudiantes: el compromiso y la satisfacción académica como variables mediadoras. *Revista de psicodidáctica*, 27(2), 149–157. <https://doi.org/10.1016/j.psicod.2022.04.003>
- García-González, J. R., Sánchez-Sánchez, P. A., Orozco, M., & Obredor, S. (2019).** Extracción de Conocimiento para la Predicción y Análisis de los Resultados de la Prueba de Calidad de la Educación Superior en Colombia. *Formación Universitaria*, 12(4), 55–62. <https://doi.org/10.4067/s0718-50062019000400055>
- Gil-Vera, V. D., & Quintero-López, C. (2021).** Predicción del rendimiento académico estudiantil con redes neuronales artificiales. *CIT Informacion Tecnologica*, 32(6), 221–228. <https://doi.org/10.4067/s0718-07642021000600221>
- Giró-Gracia, X., Universidad de Barcelona, Sancho-Gil, J. M., & Universidad de Barcelona. (2022).** La Inteligencia Artificial en la educación: Big data, cajas negras y solucionismo tecnológico. *Revista latinoamericana de tecnología educativa*, 21(1), 129–145. <https://doi.org/10.17398/1695-288x.21.1.129>
- Gómez, B. R. (2013).** Fundamentos teóricos de la evaluación por competencias: trazabilidad histórica del concepto/. 13(2), 14–23. <https://www.proquest.com/scholarly-journals/fundamentos-teoricos-de-la-evaluacion-por/docview/1495407097/se-2?accountid=48797>
- Gómez, P., Martínez, D., Delgado, R., Lopez, R., & Freire, P. (2023).** Desarrollo de Competencias Digitales Docentes en el Ecuador. *RISTI - Revista Ibérica de Sistemas e Tecnologías de Informação*, 62, 157–165. <https://www.proquest.com/docview/2880950505?parentSessionId=J2RZw33b8KX2I2sL5I8%2BRHrSPjN%2B7XiaHCL6uegACM8%3D&pq-origsite=summon&accountid=48797>
- Guías de orientación examen Saber Pro. (2023).** ICFES. <https://www.icfes.gov.co/gu%C3%ADas-de-orientaci%C3%B3n-examen-saber-pro>
- Jurado, H. F. M. (2020).** Modelos De Redes Neuronales Artificiales, Como Sustento Evaluativo Al Crecimiento Pedagógico Virtual En Educación Superior. *Revista Científica de Publicación del Centro Psicopedagógico y de Investigación en Educación Superior*, 7(2), 11–11. <https://ojs.cepies.umsa.bo/index.php/RCV/article/view/61>
- La Inteligencia Artificial y su uso en la Educación Inclusiva. (2023, noviembre 22).** Argentina.gob.ar. <https://www.argentina.gob.ar/andis/la-inteligencia-artificial-y-su-uso-en-la-educacion-inclusiva>
- Mamani-Coaquira, Y., Ibarra, M. J., Vilca, E. M., Ramos, E. O., & Cruz, M. A. (2021).** Identificar Sentimientos en Cuarentena por la Covid-19 mediante Clasificador Léxico y Aprendizaje Supervisado. *RISTI - Revista Ibérica de Sistemas e Tecnologías de Informação*, 19, 618–631. <https://www.proquest.com/scholarly-journals/identificar-sentimientos-en-cuarentena-por-la/docview/2493869956/se-2>
- Mercado Polo, D., Pedraza Caballero, L., & Martínez Gómez, E. (2015).** Comparación de Redes Neuronales aplicadas a la predicción de Series de Tiempo. *Prospectiva*, 13(2), 88. <https://doi.org/10.15665/rp.v13i2.491>

- Pérez Del Barrio, A., Menéndez Fernández-Miranda, P., Sanz Bellón, P., Lloret Iglesias, L., & Rodríguez González, D. (2022).** Artificial Intelligence in Radiology: an introduction to the most important concepts. *Radiología (English Edition)*, 64(3), 228–236. <https://doi.org/10.1016/j.rxeng.2022.03.005>
- Pérez del Barrio, A., Menéndez Fernández-Miranda, P., Sanz Bellón, P., Lloret Iglesias, L., & Rodríguez González, D. (2022).** Inteligencia artificial en Radiología: introducción a los conceptos más importantes. *Radiología*, 64(3), 228–236. <https://doi.org/10.1016/j.rx.2022.03.003>
- Plant, E. A., Ericsson, K. A., Hill, L., & Asberg, K. (2005).** Why study time does not predict grade point average across college students: Implications of deliberate practice for academic performance. *Contemporary Educational Psychology*, 30(1), 96–116. <https://doi.org/10.1016/j.cedpsych.2004.06.001>
- Ruíz Ramírez, E. (2023).** Avances y desafíos de la inteligencia artificial en el mundo laboral. Una visión desde el sector público y privado. *Dialnet*. <https://dialnet.unirioja.es/servlet/articulo?codigo=9015608>
- Raschka, S., & Mirjalili, V. (2023).** Python machine learning - (2a ed.). Packt Publishing.
- Rojas Rubio, L., & Meneses Villegas, C. (2022).** Una comparación empírica de algoritmos de aprendizaje automático versus aprendizaje profundo para la detección de noticias falsas en redes sociales. *Ingeniare. Revista Chilena de Ingeniería*, 30(2), 403–415. <https://doi.org/10.4067/s0718-33052022000200403>
- Rojas, S., González, J., Suárez, M., & González, J. (2022).** Fortalecimiento de canales de comunicación para personas con discapacidad fono-auditiva en ambientes de educación básica mediante captura de movimientos usando lenguajes de señas. *Ingeniería y Competitividad*, 25(1). <https://doi.org/10.25100/iyv.v25i1.12066>
- Salas, N. D., & Tello, M. B. (2023).** La Inteligencia Artificial en la Educación: Herramientas de Aprendizaje y Evaluación de la Escritura (en L1). *Revista de logopedia, foniatría y audiología (Internet)*, 43(4), 100328. <https://doi.org/10.1016/j.rlfa.2023.100328>
- Selwyn, N. (2019).** ¿Deberían los robots sustituir al profesorado? La IA y el futuro de la educación. Morata.
- Valero, C. S. (2017).** Aplicación de métodos de aprendizaje automático en el análisis y la predicción de resultados deportivos (Application of automated learning methods for analyzing and predicting sports outcomes). *Retos digital*, 34, 377–382. <https://doi.org/10.47197/retos.v0i34.58506>
- Vuralli, D., Polat, S., Isök, E. I., Öksüzler, M., & Göker, A. S. Ö. & (2023).** Crista Galli Morphometry and Morphology: An Anatomical, Radiological, and Machine Learning Application Study.
- Zárate-Valderrama, J., Bedregal-Alpaca, N., & Cornejo-Aparicio, V. (2021).** Modelos de clasificación para reconocer patrones de deserción en estudiantes universitarios. *Ingeniare. Revista Chilena de Ingeniería*, 29(1), 168–177. <https://doi.org/10.4067/s0718-33052021000100168>
- Fecha de recepción: 27/11/2024**
Fecha de aceptación: 11/03/2025