DIALÉCTICA
REVISTA DE INVESTIGACIÓN EDUCATIVA

pp. 65 – 105
Número 25, Vol. 2 (2025)
ISSN en Línea: 2244-7490 / ISSN Impreso: 1316-7243
Depósito Legal: pp 1999802TA777

PORTAFOLIO DE INVESTIGACION

IMPLEMENTACIÓN DE HERRAMIENTAS DE *MACHINE LEARNING* EN EL ANÁLISIS DE LAS INTERRELACIONES ENTRE VARIABLES SOCIODEMOGRÁFICAS Y DE DESEMPEÑO EN ESTUDIANTES DE EDUCACIÓN MEDIA

Manuel Antonio Hoyos García[1]
eucgal07@gmail.com
ORCID: 0009-0009-9408-0470

Eduar Ramiro Burbano Semanate[2]
eduardburbano1234@hotmail.com
ORCID: 0000-0002-3130-0028

Recibido: 15/02/2025 Aprobado: 25/03/2025

RESUMEN

En este artículo se presenta una manera de implementar herramientas *machine learning* en el campo de la educación; así, se muestra la importancia y potencialidad del uso de estas herramientas cuando se trata de analizar grandes volúmenes de datos y más de una variable objetivo. Conforme a lo anterior, el propósito general del trabajo consistió en el diseño de un modelo de predicción de desempeños en competencias básicas mediante técnicas de aprendizaje automático que permitieran establecer relaciones entre variables contextuales (características sociodemográficas) y niveles de rendimiento (competencias básicas) en estudiantes de décimo grado. Seguidamente, para la consecución de tal propósito, se emplea una metodología con paradigma positivista, adoptando métodos mixtos que permiten la integración de elementos cualitativos apoyados en las métricas de fiabilidad en el uso de algoritmos como árboles de decisión dentro del cruce de datos inherentes al contexto de los estudiantes de dos instituciones educativas y sus respectivos desempeños académicos en competencias básicas. Tal desarrollo, apoyado desde una perspectiva compleja, permite concluir que las interrelaciones entre las singularidades de los estudiantes y sus desempeños académicos podrán ser un referente, entre otras cosas, para empezar a adoptar nuevas y mejores herramientas computacionales en el ámbito educativo y replantear las

[1] Insitución Educativa Simón Bolívar, Garzón, Huila, Colombia. Docente de Matemáticas; UNIR, España. Magister en Didáctica de las mátemáticas; USCO, Colombia, Magister en Estudios Interdisciplinarios de la Complejidad; Estudiante de Doctorado en Educación, UPEL, Rubio, Venezuela.

[2] Institución Educativa El Carmelo, La Plata, Huila, Colombia, Coordinador; Uniamazonia, Colombia, Magister en Educación con énfasis en didáctica de la matemática; Estudiante de Doctorado en Educación, UPEL, Rubio, Venezuela

https://doi.org/10.56219/dialctica.v2i25.4065

prácticas docentes para obedecer a la heterogeneidad inevitable de las aulas antes de preponderar la uniformidad forzada en las mismas.

Palabras clave: *Machine learning,* competencias básicas, árbol de decisión, complejidad.

ABSTRACT

This article presents a method for implementing machine learning tools in the field of education, highlighting the importance and potential of these tools when analyzing large volumes of data and multiple target variables. Accordingly, the general objective of this study was to design a predictive model for performance in basic competencies using machine learning techniques, enabling the identification of relationships between contextual variables (sociodemographic characteristics) and achievement levels (basic competencies) among tenth-grade students. To achieve this objective, a positivist paradigm methodology was employed, integrating mixed methods that combine qualitative elements with reliability metrics derived from algorithms such as decision trees. This approach facilitated the analysis of data pertaining to students from two educational institutions and their respective academic performances in basic competencies. Supported by a complex systems perspective, this study concludes that the interrelations between students' individual characteristics and their academic performance can serve as a reference for: adopting new and improved computational tools in education, and rethinking teaching practices to accommodate the inevitable heterogeneity of classrooms rather than enforcing artificial uniformity.

Keywords: Machine learning, basic skills, decision tree, complexity.

https://doi.org/10.56219/dialctica.v2i25.4065

IALÉCTICA
REVISTA DE INVESTIGACIÓN EDUCATIVA
pp. 65 – 105
Número 25, Vol. 2 (2025)
ISSN en Línea: 2244-7490 / ISSN Impreso: 1316-7243
Depósito Legal: pp 1999802TA777
PORTAFOLIO DE INVESTIGACION

INTRODUCCIÓN

El presente artículo es un producto de la tesis de maestría denominada predicción de dificultades en competencias básicas, usando métodos *machine learning*, con estudiantes de décimo grado en dos instituciones del Huila, de la maestría en Estudios Interdisciplinarios de la Complejidad de la Universidad Surcolombiana. En este, se reflexiona que, aun estando en un mundo en constante cambio, los estudiantes, son valorados conforme a criterios similares a pesar de sus singularidades. Omitiendo por ejemplo que, según datos recientes del Departamento Administrativo Nacional de Estadística (DANE), más del 40% de la población colombiana vive en condiciones de pobreza. Pese a esta realidad, estudiantes provenientes tanto de contextos vulnerables como de entornos más favorecidos son evaluados bajo un mismo sistema y criterios estandarizados. En este marco, el presente estudio aborda una problemática centrada en la necesidad de examinar las dificultades que podría presentar un grupo de estudiantes con relación a sus desempeños en competencias básicas, considerando además la posible influencia de sus condiciones sociodemográficas.

En este sentido, se menciona que, en los entornos educativos, los esquemas de evaluación periódica suelen reducir el desempeño estudiantil a escalas numéricas que determinan si se han alcanzado los estándares mínimos de aprendizaje requeridos para la promoción de grado o no. No obstante, como señalan Méndez y López (2019), el rendimiento académico puede estar influenciado por diversos factores de naturaleza no

https://doi.org/10.56219/dialctica.v2i25.4065

necesariamente cuantitativa. Por ello, dichas formas de evaluación podrían resultar insuficientes para captar las circunstancias específicas que inciden en el proceso de aprendizaje, limitando las oportunidades de generar hallazgos que, mediante el análisis de aspectos cualitativos vinculados a los contextos de desarrollo de los estudiantes, orienten mejoras en las prácticas pedagógicas.

De este modo, se vislumbra, en la actualidad virtual que nos envuelve, un potencial relevante al explorar variables sociodemográficas como el lugar de residencia, la estructura familiar o el número de hermanos, y sus posibles vínculos con el desarrollo de competencias básicas. Sin embargo, la identificación de tales relaciones se ve restringida cuando se depende exclusivamente de metodologías convencionales, descriptivas o lineales. Puesto que, aunque los datos sociodemográficos se recopilan desde hace décadas, es improbable que mediante la mera observación o enfoques tradicionales puedan discernirse asociaciones significativas entre dichas características y el desempeño académico de los estudiantes en cuestión.

En consecuencia, este estudio busca identificar posibles interrelaciones entre las características sociodemográficas de los estudiantes y su rendimiento, en un contexto que reconoce la necesidad de reformular las prácticas pedagógicas, evaluativas y curriculares vigentes. La investigación se sustenta en marcos teóricos clásicos como la teoría del desarrollo cognitivo de Piaget (1981) y el concepto de zona de desarrollo

https://doi.org/10.56219/dialctica.v2i25.4065

IALÉCTICA
REVISTA DE INVESTIGACIÓN EDUCATIVA
pp. 65 – 105
Número 25, Vol. 2 (2025)
ISSN en Línea: 2244-7490 / ISSN Impreso: 1316-7243
Depósito Legal: pp 1999802TA777
PORTAFOLIO DE INVESTIGACION

próximo de Vygotsky (1988), los cuales enfatizan el papel formativo de los factores sociales y ambientales en el aprendizaje. Empero, reconocer la existencia de vínculos potenciales entre múltiples variables sociodemográficas y los resultados académicos implica abordar una problemática compleja no solo por la cantidad de factores intervinientes, sino también por los riesgos metodológicos en el tratamiento de dichos datos.

Así, esta investigación plantea un desafío de carácter no predecible: comprender cómo o qué variables podrían influir en las dificultades asociadas al desarrollo de competencias básicas en contextos educativos diferenciados, como es el caso de la Institución Educativa La Merced en el municipio de El Agrado y la Institución Educativa Ramón Alvarado Sánchez en la zona rural de Garzón, Huila. Por lo tanto, se entiende que el objetivo general del presente estudio fue:

Diseñar un modelo de predicción de desempeños en competencias básicas mediante técnicas de aprendizaje automático –*machine learning*-, que permita establecer relaciones entre variables contextuales (características sociodemográficas) y niveles de rendimiento (competencias básicas) en estudiantes de décimo grado, con el fin de respaldar la implementación de enfoques pedagógicos diferenciados en las instituciones educativas Ramón Alvarado Sánchez y La Merced.

Asimismo, los objetivos específicos considerados fueron: Caracterizar, los elementos sociodemográficos de los es estudiantes de décimo grado de las instituciones Ramón

https://doi.org/10.56219/dialctica.v2i25.4065

Alvarado Sánchez y La Merced: establecer interrelaciones entre las variables sociodemográficas y de desempeño académico, haciendo uso de herramientas *machine learning* en estudiantes de grado décimo y validar con métricas de incertidumbre, la fiabilidad en las predicciones de los algoritmos para considerar la pertinencia del uso de herramientas ML al momento de predecir dificultades académicas.

Expuesto lo anterior, se clarifica el norte de la investigación, por lo que ahora se procede a hacer una síntesis teórica de las herramientas implementadas en el estudio, de las métricas de incertidumbre necesarias para validarlas, también de la formulación metodológica que orienta el trabajo y que lo hizo posible para, finalmente, mostrar con base a las técnicas e instrumentos usados, un resumen de los resultados y las conclusiones obtenidas tras la culminación del desarrollo investigativo efectuado en el marco de la perspectiva compleja.

https://doi.org/10.56219/dialctica.v2i25.4065

DIALÉCTICA
REVISTA DE INVESTIGACIÓN EDUCATIVA
pp. 65 – 105
Número 25, Vol. 2 (2025)
ISSN en Línea: 2244-7490 / ISSN Impreso: 1316-7243
Depósito Legal: pp 1999802TA777
PORTAFOLIO DE INVESTIGACION

ASPECTOS TEÓRICOS

Relevancia de datos, información y herramientas de aprendizaje automático o *machine learning*

Como proponen Davenport y Prusak (1999), autores citados en el trabajo de Carrion (2017), es fundamental diferenciar entre los conceptos de dato, información y conocimiento, ya que, aunque comparten ciertas características y mantienen relaciones entre sí, no son equivalentes en su naturaleza. En este sentido, los datos pueden concebirse como unidades elementales o fragmentos aislados que representan eventos específicos sin contexto. La información, por otro lado, implica un mayor nivel de organización, pues constituye contenido que adquiere significado y puede ser comprendido por el receptor. Ahora, a partir del momento en que los datos se estructuran y articulan coherentemente, se hace posible la generación de conocimiento.

Justo este último concepto, cabe aclara que, no se limita únicamente a la transformación de datos en información. Este integra dimensiones más amplias, como la experiencia acumulada, los procedimientos sistematizados y las estructuras organizacionales, conformando así un marco complejo de contenido útil, aplicable y con valor educativo y funcional. Por lo tanto, disciplinas como la Minería de Datos (DM-*data mining*-) y el Aprendizaje Automático surgen como enfoques valiosos que aportan claridad a la naturaleza frecuentemente opaca y desorganizada de los datos a gran

https://doi.org/10.56219/dialctica.v2i25.4065

escala, los cuales en su forma cruda pueden parecer confusos, irrelevantes o funcionalmente limitados.

Al reconocer el potencial de los datos y la información en la generación de conocimiento, especialmente en contextos que requieren el análisis y procesamiento de conjuntos de datos extensos, esta sección retoma la DM como herramienta metodológica. Así, basados en el marco propuesto por Ramírez (2021), a continuación, se describen los pasos necesarios para implementar un modelo o algoritmo capaz de revelar relaciones significativas entre las diversas variables que caracterizan a la población estudiantil y que, en términos de educación se tienen en grandes volúmenes si se piensa en instituciones que por ejemplo tienen más de dos mil estudiantes:

Paso 1: Definición de objetivos. Esta fase inicial implica identificar con claridad el tipo de hallazgos o conocimiento que se pretende extraer del conjunto de datos utilizado en la investigación.

Paso 2: Preparación de datos. Esta etapa comprende procedimientos orientados a seleccionar, limpiar, reducir y, en general, transformar el conjunto de datos, así como minimizar el ruido (datos incompletos, errados, etc.), con el fin de garantizar que los datos sean adecuados para el análisis.

https://doi.org/10.56219/dialctica.v2i25.4065

DIALÉCTICA
REVISTA DE INVESTIGACIÓN EDUCATIVA
pp. 65 – 105
Número 25, Vol. 2 (2025)
ISSN en Línea: 2244-7490 / ISSN Impreso: 1316-7243
Depósito Legal: pp 1999802TA777
PORTAFOLIO DE INVESTIGACION

Paso 3: Selección del modelo. Considerada una de las fases más cruciales, este paso implica elegir el algoritmo más apropiado, con el objetivo de abordar exhaustivamente una tarea analítica específica dentro del estudio.

Paso 4: Análisis de resultados. Como en cualquier proceso de investigación formal, es fundamental examinar los resultados obtenidos tras aplicar el algoritmo a los datos procesados, con el fin de evaluar la efectividad y relevancia del modelo en relación con el problema de investigación.

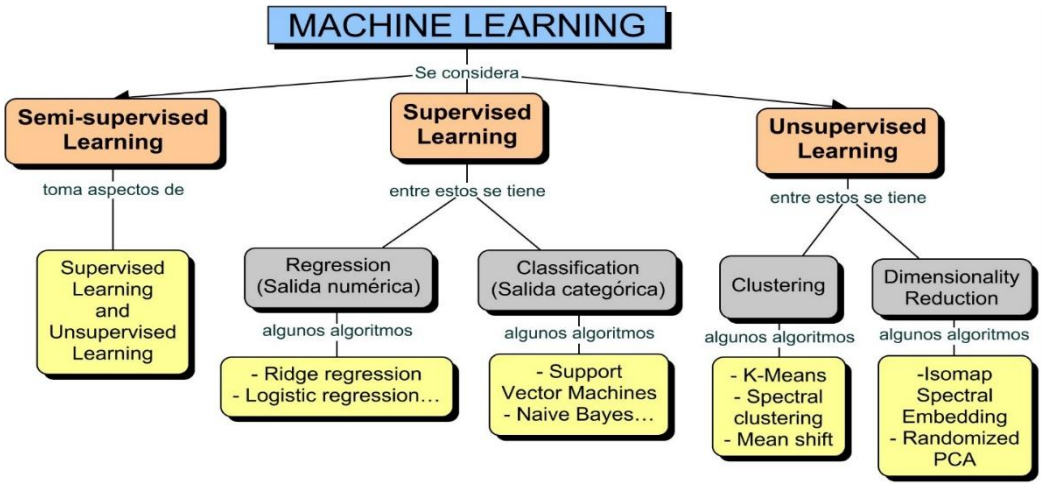
En este contexto, al reconocer el valor de los sistemas inteligentes para el análisis predictivo de datos, el aprendizaje automático o ML se consolida como un enfoque viable para generar resultados automatizados con aplicabilidad transversal en diversos campos del conocimiento, entre los que la educación no escapa a sus alcances por lo ya mencionado con anterioridad. De este modo, se debe destacar que su inherente flexibilidad le permite adaptarse eficazmente a la naturaleza dinámica y heterogénea que caracteriza múltiples procesos de investigación (Méndez y López, 2019).

Considerando entonces el ML, se menciona que, dentro del marco del aprendizaje supervisado (ML supervisado), resulta fundamental ejecutar procesos de monitoreo, entrenamiento y selección de atributos con el fin de optimizar el rendimiento del algoritmo (Flórez & Pulido, 2020). Según la naturaleza del problema de investigación, el aprendizaje supervisado se clasifica típicamente en: 1. regresión, cuando las salidas son

https://doi.org/10.56219/dialctica.v2i25.4065

numéricas; o 2. clasificación, cuando los resultados esperados son categóricos. Considerando esta última característica, se selecciona el ML supervisado como el enfoque más adecuado, dado que permite descubrir patrones relacionales que no son fácilmente identificables mediante el exclusivo análisis humano. De este modo, siempre que los datos estén debidamente procesados y el modelo demuestre un alto rendimiento, las predicciones evaluativas iniciales sobre el desempeño estudiantil pueden considerarse más confiables. La Figura 1 ilustra la taxonomía de herramientas de ML y algunos de los algoritmos correspondientes.

Figura 1. *Taxonomía general de herramientas ML*



Fuente: Basado en los aportes de Beunza, 2020

De acuerdo con los aportes encontrados en Quinlan (1985), se ha reconocido la versatilidad algorítmica de los árboles de decisión como metodología transferible a diversos campos del conocimiento y aplicables, por tanto, en distintas facetas. Se documenta así, su implementación exitosa en sistemas de comercio electrónico,

https://doi.org/10.56219/dialctica.v2i25.4065

IALÉCTICA
REVISTA DE INVESTIGACIÓN EDUCATIVA
pp. 65 – 105
Número 25, Vol. 2 (2025)
ISSN en Línea: 2244-7490 / ISSN Impreso: 1316-7243
Depósito Legal: pp 1999802TA777
PORTAFOLIO DE INVESTIGACION

mecanismos de detección de fraude financiero, modelos predictivos de fallas tecnológicas e, incluso, en protocolos de triage médico basados en sintomatología (Méndez & López, 2019). Conforme a ello, este estudio propone extender su aplicación al ámbito educativo, un sistema complejo que demanda innovaciones capaces de reconfigurar tanto sus procesos administrativos como sus estrategias didácticas, introduciendo entornos educativos virtuales.

Algoritmos de clasificación y medición de incertidumbre

Como se ha establecido previamente, el Aprendizaje Automático, engloba diversos paradigmas de aprendizaje que varían según la presencia o ausencia de supervisión, o la integración de ambos enfoques. Con base en esta clasificación y respaldados por distintos trabajos en el ámbito de los sistemas inteligentes [Jones-Ortiz & Guzmán-Seraquive (2022); Bausili (2021); Joaquín (2020); Méndez & López (2019); Correa et al. (2008); Sucar & Tonantzintla (2006)], esta sección describe los algoritmos de clasificación más relevantes para la presente investigación:

Árboles de Decisión (DT): Utilizan métricas de teoría de la información (entropía de Shannon, índice de Gini)

Mecanismo: Evalúan la pureza de los nodos mediante particiones recursivas del espacio de características

Regresión Logística (LR): En términos generales, es un tipo de modelo lineal generalizado

https://doi.org/10.56219/dialctica.v2i25.4065

Mecanismo: Clasificación binaria mediante transformación sigmoide de predictores lineales

Redes Bayesianas: Su estructura se basa en el uso de Grafos Acíclicos Dirigidos (GAD)

Mecanismo: Representan dependencias condicionales mediante relaciones probabilísticas

Redes Neuronales Artificiales (NN): Las cuales en su arquitectura trabajan con perceptrones multicapa de inspiración biológica (de ahí su nombre)

Mecanismo: Procesamiento distribuido de información mediante interconexiones ponderadas

Máquinas de Vectores Soporte (SVM): La implementación de estos, es la construcción de hiperplanos óptimos de separación en espacios multidimensionales

Mecanismo: Clasificación de margen máximo

Competencias desde la perspectiva educativa

Antes de abordar la clasificación de las competencias según su alcance, dimensiones o ámbitos de aplicación, resulta imprescindible realizar una reflexión detallada sobre el significado del término "competencia". Al respecto, es necesario considerar las diversas perspectivas que han contribuido a su desarrollo conceptual. Precisamente, desde una perspectiva fundacional, cabe resaltar los aportes realizados en décadas anteriores durante el período en que el enfoque basado en competencias

https://doi.org/10.56219/dialctica.v2i25.4065

DIALÉCTICA
REVISTA DE INVESTIGACIÓN EDUCATIVA
pp. 65 – 105
Número 25, Vol. 2 (2025)
ISSN en Línea: 2244-7490 / ISSN Impreso: 1316-7243
Depósito Legal: pp 1999802TA777
PORTAFOLIO DE INVESTIGACION

cobró relevancia en educación, con autores como Ouellet (2000), Bogoya (2000) y Bunk (1994), quienes plantearon que las competencias implican la capacidad de movilizar conocimientos para abordar situaciones específicas. Desde este enfoque, la noción remite a la aplicación contextualizada de saberes que, una vez interiorizados, permiten actuar de manera creativa, efectiva y pertinente en diversos contextos.

Ahora, desde una perspectiva crítica y ampliación conceptual, Tobón (2003) propone una ampliación crítica de esta perspectiva. Si bien reconoce el valor de las conceptualizaciones previas, cuestiona su tendencia a centrarse en una lógica funcionalista y economicista, frecuentemente dejando de lado elementos esenciales como la responsabilidad ética, la convivencia social y la solidaridad. Desde la teoría del pensamiento complejo, se postula por el autor que las competencias deben entenderse como procesos integradores y dinámicos que fusionan el saber, el hacer y el ser. El autor las define como :

> procesos complejos que las personas activan en la terna acción-desempeño-creación para realizar actividades sistémicas y resolver problemas relacionados con el trabajo y la vida cotidiana, con el objetivo de avanzar en la autorrealización personal, vivir la vida con autenticidad y contribuir al bienestar humano (p. 10).

De manera complementaria, desde el marco internacional y con una perspectiva más holística, cabe destacar el rol de la Organización de las Naciones Unidas para la Educación, la Ciencia y la Cultura (UNESCO). En 2015, la UNESCO impulsó iniciativas

https://doi.org/10.56219/dialctica.v2i25.4065

orientadas a transformar la educación a nivel global. A través del informe Repensar la educación: ¿Hacia un bien común mundial? (UNESCO, 2015), convocó a revisar los modelos educativos, enfatizando la necesidad de replantear la enseñanza y el aprendizaje mediante un enfoque por competencias más integral. En sintonía con los objetivos del presente estudio, el organismo subraya la importancia de diseñar políticas educativas que respondan a la complejidad del mundo contemporáneo y reconozcan la interconexión entre los diversos agentes involucrados en los procesos educativos.

MÉTODOS

Paradigma y orientación del trabajo investigativo

Siguiendo referentes como Hernández-Sampieri (2018), esta investigación se construye desde el paradigma positivista, pero, adopta un diseño de métodos mixtos que integra variables cualitativas (características sociodemográficas del estudiantado) y datos cuantitativos (rendimiento en competencias básicas e índices de confiabilidad de predicciones generadas por algoritmos de machine learning). El estudio se enfoca en estudiantes de las instituciones educativas La Merced y Ramón Alvarado Sánchez, utilizando un enfoque cuantitativo no experimental de tipo longitudinal, donde los datos se categorizaron sin manipulación experimental para identificar patrones, incluyendo fases iterativas de prueba de modelos ML que justifican su temporalidad extendida.

https://doi.org/10.56219/dialctica.v2i25.4065

DIALÉCTICA
REVISTA DE INVESTIGACIÓN EDUCATIVA
pp. 65 – 105
Número 25, Vol. 2 (2025)
ISSN en Línea: 2244-7490 / ISSN Impreso: 1316-7243
Depósito Legal: pp 1999802TA777
PORTAFOLIO DE INVESTIGACION

Respecto al alcance, la investigación posee un carácter correlacional, buscando examinar asociaciones entre el desempeño académico en competencias básicas y factores sociodemográficos, mediante la aplicación de algoritmos de aprendizaje automático que evalúen su potencial predictivo sobre dificultades en el desarrollo competencial. Este abordaje permite analizar sistemáticamente estas interrelaciones para generar insumos que contribuyan a la mejora educativa en los contextos estudiados y se apoya en la complejidad que, desde la naturaleza del problema sugiere la importancia de usar métodos, herramientas y técnicas no lineales para considerar la problemática desde su totalidad y no de manera seccionada.

Delimitación del estudio

Esta investigación ha sido diseñada para implementarse en contextos educativos tanto rurales como urbanos, específicamente en los municipios de Garzón y El Agrado (en el departamento del Huila), por lo que la población objeto de estudio corresponde a estudiantes de décimo grado de las instituciones educativas seleccionadas, grado elegido por coincidir con la edad previa en que se aplican evaluaciones estandarizadas que deberían reflejar el grado de desarrollo de sus competencias básicas; la Tabla 1 presenta una caracterización general de la muestra estudiada, destacando variables como género y procedencia residencial de los alumnos, con el fin de aportar una comprensión más detallada de sus particularidades contextuales.

https://doi.org/10.56219/dialctica.v2i25.4065

Tabla 1. Particularidades contextuales de los estudiantes

	Zona de Residencia		Edad			Reprobación previa		
	Rural	Urbana	14	15	> = 16	Sí	No	
Femenino	25	27	2	28	22	8	44	52
Masculino	27	13	0	23	17	12	28	40[3]
Total	52	40	2	51	39	41	51	92

Fuente: Elaboración propia

Fases del Estudio

En coherencia con los objetivos propuestos y la idea general de abordar el problema para considerar la interrelación entre variables sociodemográficas y de desempeño académico, en la tabla 2, se establecen las fases seguidas para el desarrollo de la investigación, en ella, no solo se precisan los objetivos que son la orientación del trabajo, sino que, además, se especifican las acciones genéricas que permitirán el éxito de cada fase, en consecuencia, se deja trazada una guía resumida que permite introducirse en el cómo se usarán las herramientas ML para que se obtengan resultados con fundamento científico.

[3] Se aclara que, el número obedece a la suma de los hombres, y no a los subtotales de las categorías

https://doi.org/10.56219/dialctica.v2i25.4065

DIALÉCTICA
REVISTA DE INVESTIGACIÓN EDUCATIVA
pp. 65 – 105
Número 25, Vol. 2 (2025)
ISSN en Línea: 2244-7490 / ISSN Impreso: 1316-7243
Depósito Legal: pp 1999802TA777
PORTAFOLIO DE INVESTIGACION

Tabla 2. Resumen descriptivo de las fases de la investigación

Objetivo	Fase	Descripción
Caracterizar, los elementos sociodemográficos de los es estudiantes de décimo grado de las instituciones Ramón Alvarado Sánchez y La Merced	1. Caracterización	En primera instancia se realiza una caracterización de la población objeto de estudio, para ello se usa un cuestionario Google que servirá como medio de recolección de datos, luego se hará un estudio de las singularidades sociodemográficas de cada estudiante.
Establecer interrelaciones entre las variables sociodemográficas y de desempeño académico, haciendo uso de herramientas machine learning en estudiantes de grado décimo.	2. Comprensión de los datos	Previo a la recolección de datos sociodemográficos, es esencial definir las categorías de las variables y los criterios analíticos para abordar el problema de investigación.
	3. Limpieza de los datos	Definidas las variables de estudio, es crucial depurar los datos mediante: eliminación de registros incompletos, unificación en una sola base y estandarización de formatos para el análisis algorítmico, considerando posibles ajustes adicionales durante el proceso.
	4. Prueba del modelo ML	Tras completar la depuración de datos, se ejecutan los algoritmos

https://doi.org/10.56219/dialctica.v2i25.4065

		seleccionados para clasificar los estudiantes según su desempeño en competencias básicas, utilizando como variables predictoras sus características sociodemográficas. El modelo óptimo resultante deberá integrarse en una plataforma web que funcione como entorno virtual de apoyo docente, facilitando prospectivamente la mejora de prácticas pedagógicas.
Validar con métricas de incertidumbre, la fiabilidad en las predicciones de los algoritmos para considerar la pertinencia del uso de herramientas ML al momento de predecir dificultades académicas.	5. Evaluación	La implementación de sistemas basados en Big Data y aprendizaje automático para predecir dificultades académicas requiere un riguroso análisis de la confiabilidad del modelo, evaluando métricas clave como: exactitud (accuracy), precisión (precision), exhaustividad (recall) y puntuación F1 (F1-score). Estas métricas, previamente definidas en el marco teórico, permiten cuantificar la robustez algorítmica y los niveles de incertidumbre en los datos. Complementariamente, el análisis de la entropía informacional emerge

https://doi.org/10.56219/dialctica.v2i25.4065

DIALÉCTICA
REVISTA DE INVESTIGACIÓN EDUCATIVA
pp. 65 – 105
Número 25, Vol. 2 (2025)
ISSN en Línea: 2244-7490 / ISSN Impreso: 1316-7243
Depósito Legal: pp 1999802TA777
PORTAFOLIO DE INVESTIGACION

		como criterio técnico fundamental para seleccionar el modelo óptimo, garantizando así la pertinencia predictiva del sistema implementado.

Fuente: Elaboración propia

https://doi.org/10.56219/dialctica.v2i25.4065

RESULTADOS

En la primera fase (Caracterización) del desarrollo de la investigación (ver tabla 2), se definieron algunas de las variables que referencian las características inherentes a la singularidad de los estudiantes de las dos instituciones educativas objeto de estudio y que fueron los predictores respecto a los desempeños en competencias básicas. La tabla 3 resume dichas variables las cuales se convirtieron en el insumo para la generación de la base de datos inicial para aplicación de algoritmos.

Tabla 3. Variables de entrada, convenciones y numeración de sub categorías

No.	Variable	Convención	Sub Categoría	Equivalencia numérica
1	Residencia	TP_RES	Rural	0
			Urbano	1
2	Acceso a internet	INT	Sí	1
			No	0
3	Seguridad de la zona	SEG_ZO	Baja	0
			Media	1
			Alta	2
4	Lejanía a la institución	LEJA	Menos de 15 min	0
			15 a 30 min	1
			30 a 60 min	2
			Más de 1h	3
5		ACT_COMP	Ninguna	0
			1 actividad	1

https://doi.org/10.56219/dialctica.v2i25.4065

DIALÉCTICA
REVISTA DE INVESTIGACIÓN EDUCATIVA
pp. 65 – 105
Número 25, Vol. 2 (2025)
ISSN en Línea: 2244-7490 / ISSN Impreso: 1316-7243
Depósito Legal: pp 1999802TA777
PORTAFOLIO DE INVESTIGACION

	Realización de actividades complementarias		2 actividades	2
			3 actividades	3
6	Riesgo a drogas o alcohol en la zona	RIES_DR_ALC	Baja	0
			Media	1
			Alta	2
7	Tipo de familia	TP_FAM	elemental	0
			consanguínea	1
			monoparental	2
			padres separados	3
			vive con otro familiar	4
			homoparental	5
8	Integrantes de la familia	NU_INT	valor numérico	0,1,2,3,…
9	Usted ha sido educado de manera	TP_EDU	autoritaria	0
			democrática	1
			permisiva	2
			indiferente	3
10	Ingresos familiares	INGR	1 salario mínimo	1
			1 y 2 salarios	2
			más de 2 salarios	3
11	Escolaridad de la madre	ESC_M	Primaria	0
			Básica secundaria	1
			Bachillerato	2
			Técnico o tecnólogo	3
			Universitario	4

https://doi.org/10.56219/dialctica.v2i25.4065

			Ns/Nr	5
12	Escolaridad del padre	ESC_P	Primaria	0
			Básica secundaria	1
			Bachillerato	2
			Técnico o tecnólogo	3
			Universitario	4
			Ns/Nr	5
13	Relación con la madre	REL_M	Mala	0
			Regular	1
			Buena	2
			excelente	3
			Ns/Nr	4
14	Relación con el padre	REL_P	Mala	0
			Regular	1
			Buena	2
			excelente	3
			Ns/Nr	4
15	Manejo de los conflictos	CONFL	Diálogo	0
			discusiones	1
16	Consideración de la infraestructura de la Institución para el aprendizaje	INFR	Mala	0
			Regular	1
			Buena	2
			excelente	3
17	Relación con sus compañeros	REL_C	Mala	0
			Regular	1
			Buena	2
			excelente	3
18	Relación con sus docentes	REL_D	Mala	0
			Regular	1
			Buena	2
			excelente	3

https://doi.org/10.56219/dialctica.v2i25.4065

DIALÉCTICA
REVISTA DE INVESTIGACIÓN EDUCATIVA
pp. 65 – 105
Número 25, Vol. 2 (2025)
ISSN en Línea: 2244-7490 / ISSN Impreso: 1316-7243
Depósito Legal: pp 1999802TA777

19	Tiempo habitual que dedica a la lectura por entretenimiento	LEE	no lee	0
			<= 30 minutos	1
			(30 y 60]	2
			(1 y 2] h	3
			más de 2 horas	4
20	Siente agrado y considera que la institución lo apoya en su proceso de formación	APOY	muy poco	0
			Regular	1
			a gusto	2
			muy a gusto	3
21	Reprobación en años anteriores	REPR	entrada numérica	0,1,2,3,…
22	Edad	ED	entrada numérica	0,1,2,3,…
23	Género Biológico	GEN_B	Femenino	0
			Masculino	1
Fuente: Elaboración propia				

Se debe recalcar que, las equivalencias numéricas en la anterior tabla, son estrictamente necesarias para su análisis en cualquier algoritmo de clasificación, es decir, aunque se están haciendo preguntas de tipo cualitativo como su relación con padres o percepción del apoyo de la institución, las máquinas operan con números, por lo que, se hace dicha conversión. Ahora, el desempeño académico de los estudiantes encuestados en áreas relacionadas con las competencias básicas, específicamente Matemáticas, Ciencias Naturales, Ciencias Sociales, Inglés y Lectura Crítica, fue registrado de manera continua. Estas áreas representan las variables dependientes o de

https://doi.org/10.56219/dialctica.v2i25.4065

salida en el modelo propuesto. En consecuencia, la base de datos fue enriquecida con estas variables y sus respectivas categorías.

Es importante señalar que los niveles de desempeño en cada una de las asignaturas asociadas a las cinco competencias básicas evaluadas por el sistema educativo colombiano se clasifican en cuatro categorías: Bajo, Básico, Alto y Superior. Sin embargo, para mejorar el potencial de desempeño de los algoritmos predictivos, estas categorías fueron reclasificadas en dos grupos: Reprobado (0) y Aprobado (1). Este resultado binomial se considera suficiente para las necesidades interpretativas del estudio y se alinea con el objetivo general de predecir dificultades académicas. Subsecuente, durante la fase de comprensión de los datos, se observó que la mayoría de las variables son de naturaleza categórica. Por lo tanto, con base en el marco teórico sobre la importancia e influencia de las variables sociodemográficas, se aplicó un filtro de selección, resumido en la Tabla 3. Adicionalmente, las variables de salida se definieron como asignaturas que promueven las competencias básicas y fueron representadas como salidas binarias para fines de modelado.

Se debe referenciar que, la elección de los estudiantes de grado décimo se da debido a su posición previa a la presentación de exámenes estandarizados y porque, de encontrarse un modelo favorable de predicción, podrían establecerse alternativas para mejoramiento de competencias básicas en el siguiente año. De manera continua, dentro del tratamiento de los datos, se procede a una fase de limpieza para eliminar en mayor

https://doi.org/10.56219/dialctica.v2i25.4065

DIALÉCTICA
REVISTA DE INVESTIGACIÓN EDUCATIVA
pp. 65 – 105
Número 25, Vol. 2 (2025)
ISSN en Línea: 2244-7490 / ISSN Impreso: 1316-7243
Depósito Legal: pp 1999802TA777
PORTAFOLIO DE INVESTIGACION

medida, el ruido de la base de datos generado por, entre otras cosas, datos incompletos, datos errados, datos repetidos, etc. Con lo anterior, se logra, además, evitar la redundancia de dimensionalidad (exceso de atributos).

También, dentro de la comprensión y limpieza de los datos, se tiene que, en la etapa de codificación e inicio de la implementación de los algoritmos de clasificación, los datos mostrados en la Figura 2 se transformaron en un conjunto de datos más adecuado y con menos ruido, lo que permitió un procesamiento más efectivo mediante los lenguajes de programación seleccionados. Para lograrlo, se establecieron convenciones específicas con el fin de agilizar el procedimiento, se aclara que, todo este desarrollo conforme a los pasos expuestos en los aspectos teóricos, favorecen, en términos de manejos computacionales, la potencialidad de las herramientas de aprendizaje automático.

Figura 2. *Muestra de atributos considerados en el dataset*

Nombre		Número de i	3. Lugar de residencia	En su lugar de residen	Seguridad de la zona	Lejanía a la instituci	Modo de desplazamie	Clase social (e	Grupo social especial:	Actividades comple	Riesgo a drogas
sanchez	1	123456789	Urbano	Acueducto, Alcantarill	Media	Menos de 15 min	Caminando	2	No pertenezco a ningu	Ninguna	Alta
herrera palma	2	12279381	Rural	Acueducto, Alcantarill	Media	Menos de 15 min	Moto	1	No pertenezco a ningu	Deportes	Media
bustos	3	1.023E+09	Urbano	Electricidad, Gas dom	Media	Menos de 15 min	Moto	1	No pertenezco a ningu	Deportes	Baja
Carvajal Morales	4	1.079E+09	Urbano	Acueducto, Electricida	Media	15 a 30 min	Moto	1	Victima de conflicto	Ninguna	Baja
orjuela mancera	5	1.079E+09	Rural	Electricidad, Televisior	Media	30 a 60 min	Transporte escolar	1	No pertenezco a ningu	Ninguna	Baja
sandoval cruz	6	1.074E+10	Rural	Electricidad, internet,	Media	15 a 30 min	Moto	2	Victima de conflicto	Ninguna	Baja
navia	7	1.06E+09	Urbano	Electricidad, Computa	Baja	15 a 30 min	Caminando	1	Desplazado	Deportes	Baja
sofia leal rojas	8	1.079E+09	Urbano	Acueducto, Electricida	Alta	Menos de 15 min	Caminando	1	No pertenezco a ningu	Pintura, Deportes,	Baja
Brinyth	9	1.079E+09	Urbano	Acueducto, Alcantarill	Media	15 a 30 min	Caminando	1	No pertenezco a ningu	Canto y banda	Baja
herrera cortes	10	1.082E+09	Urbano	Moto	Baja	15 a 30 min	Moto	1	No pertenezco a ningu	galda	Baja
herrera	11	1.082E+09	Urbano	Acueducto, Alcantarill	Media	15 a 30 min	Moto	1	No pertenezco a ningu	Danzas, Pintura, D	Media
Camila dussan	12	1.082E+09	Urbano	Acueducto, Alcantarill	Media	Menos de 15 min	Moto	1	No pertenezco a ningu	Danzas, Teatro, De	Media
almario losada	13	1.078E+09	Urbano	Acueducto, Alcantarill	Media	15 a 30 min	Caminando	1	No pertenezco a ningu	Ninguna	Media
perdomo fierro	14	1.082E+09	Urbano	Acueducto, Alcantarill	Media	15 a 30 min	Caminando	1	No pertenezco a ningu	Ninguna	Media
Naranjo	15	26451127	Urbano	Acueducto, Alcantarill	Media	15 a 30 min	Moto	1	No pertenezco a ningu	Danzas	Media
perdigon ruiz	16	1.107E+09	Urbano	Acueducto, Alcantarill	Baja	Menos de 15 min	Caminando	1	No pertenezco a ningu	Deportes	Baja
Nicolle Leiva	17	1.078E+09	Urbano	Acueducto, Alcantarill	Media	Menos de 15 min	Caminando	1	No pertenezco a ningu	Ninguna	Baja
Jurado Chavarro	18	1.079E+09	Urbano	Acueducto, Alcantarill	Media	Menos de 15 min	Moto	1	Afrodescendiente	Danzas, Deportes	Media
Quintero joven	19	1.082E+09	Urbano	Acueducto, Alcantarill	Alta	15 a 30 min	Moto	1	No pertenezco a ningu	Ninguna	Baja
Hoyos Almario	20	1.079E+09	Urbano	Acueducto, Alcantarill	Media	Menos de 15 min	Moto	1	No pertenezco a ningu	Deportes	Alta
Parra Vargas	21	1.079E+09	Urbano	Acueducto, Alcantarill	Media	15 a 30 min	Caminando	1	No pertenezco a ningu	Ninguna	Media
trujillo Escalante	22	1.079E+09	Rural	Acueducto	Media	15 a 30 min	Caminando	0	No pertenezco a ningu	Danzas	Baja
Ramírez	23	1.075E+09	Urbano	Electricidad, Computa	Media	Menos de 15 min	Moto	1	No pertenezco a ningu	Danzas	Baja
Adames Olaya	24	1.082E+09	Rural	Electricidad, internet,	Baja	30 a 60 min	Moto	0	No pertenezco a ningu	Ninguna	Baja

Fuente: Elaboración propia

https://doi.org/10.56219/dialctica.v2i25.4065

La Figura 3 por su parte, muestra una parte representativa del conjunto de datos final, el cual, ya tratado conforme a los procesos de limpieza y uso de convenciones en los datos, se utilizó para probar los algoritmos dentro de la plataforma Anaconda inicialmente y usando el interfaz Spyder para Python, con ello se evidencia una clara transformación que es necesaria para que los algoritmos operen en su lenguaje natural.

Figura 3. *Dataset preliminar: 92 muestras*

TP_RES	INT	SEG_ZO	LEJA	ACT_COMP	RIES_DR_ALC	TP_FAM	UN_INT	TP_EDU	INGR	ESC_M	ESC_P	REL_M	REL_P	CONFL	INFR
1	1	1	0	0	2	0	5	1	1	2	2	2	2	0	1
0	0	1	0	1	1	0	5	0	2	1	2	4	3	0	2
1	0	1	0	1	0	0	4	2	1	4	1	3	3	0	2
0	0	1	2	0	0	3	2	0	1	2	0	3	0	0	1
0	1	1	1	0	0	0	4	2	1	0	1	3	3	0	2
1	0	1	1	2	0	2	3	1	1	0	0	2	3	0	1
1	0	0	1	1	0	0	4	1	1	2	0	2	2	0	1
1	1	1	1	3	1	4	3	0	1	2	0	1	2	0	1
1	1	1	0	3	1	0	3	0	1	2	4	3	3	0	1
1	0	1	1	0	1	4	4	1	2	1	1	3	3	0	1
1	1	1	1	1	1	0	5	1	2	3	1	2	3	0	2
1	1	0	0	1	0	0	5	1	1	0	2	3	3	0	1
1	1	1	0	0	0	2	2	2	1	0	0	2	2	0	2
1	1	1	0	2	1	1	3	1	1	4	4	2	4	0	1
1	1	2	1	0	0	0	4	0	1	4	4	2	2	0	2
1	1	1	0	1	2	2	2	1	1	3	2	3	1	0	1
0	0	1	1	1	0	0	4	0	3	1	1	3	3	0	3
1	1	1	0	1	0	0	3	1	1	2	3	2	2	0	1
1	1	1	1	1	1	1	6	0	1	2	0	3	4	0	2
1	0	1	1	1	1	2	6	1	1	1	1	1	2	0	2
1	1	1	0	0	2	0	6	1	1	4	5	3	3	0	3

Fuente: Elaboración propia

Una vez finalizada la fase de limpieza de datos, todos los componentes necesarios estuvieron listos para iniciar la etapa de experimentación. En este contexto, el primer paso consistió en identificar el tipo de marco de aprendizaje adecuado para el problema en cuestión. Dado que la variable objetivo es categórica (aprobado o reprobado en cada una de las asignaturas seleccionadas) y que el objetivo es entrenar algoritmos para luego

https://doi.org/10.56219/dialctica.v2i25.4065

DIALÉCTICA
REVISTA DE INVESTIGACIÓN EDUCATIVA
pp. 65 – 105
Número 25, Vol. 2 (2025)
ISSN en Línea: 2244-7490 / ISSN Impreso: 1316-7243
Depósito Legal: pp 1999802TA777
PORTAFOLIO DE INVESTIGACION

evaluar su desempeño predictivo, el estudio se enmarca en el aprendizaje supervisado. Además, dado que la variable de salida no es continua ni numérica, los algoritmos de regresión se consideraron inadecuados, optándose en su lugar por algoritmos de clasificación. Así, entre la amplia gama de modelos disponibles y con base en los referentes teóricos consultados, se consideraron aquellos algoritmos con eficacia comprobada en aplicaciones comerciales, pero con potencial de adaptabilidad al contexto educativo.

De acuerdo a lo anterior, se seleccionaron para su prueba los siguientes algoritmos, por su nombre en español (siglas originarias del nombre inglés): Regresión Logística (LR); Árbol de Decisión (DT); K-Vecinos Más Cercanos (KNN); Red Neuronal (NN) y Máquina de Vectores de Soporte (SVM). Estos serán validados posteriormente mediante métricas de desempeño apropiadas.

En la figura 4, después de varias pruebas, se muestran de manera resumida los rendimientos de los algoritmos de predicción mencionadaos en el anterior párrafo, para cada una de las salidas que representan las asignaturas de Matemáticas (MAT), ciencias sociales (CS), ciencias naturales (CN), inglés (ING) y Castellano (LEC). Tales pruebas se dieron mediante codificación en *Python*, haciendo uso de la ya antes mencionada plataforma Anaconda y el interfaz *Spyder.*

Debe mencionarse, además, que, tras la obtención de una base de datos desbalanceada, surgieron algunos errores en las métricas de rendimiento iniciales, por

https://doi.org/10.56219/dialctica.v2i25.4065

lo que, se mejoró esta y se aplicaron nuevamente los algoritmos de donde se pudo visualizar lo que se muestra a continuación, en la figura 4. En esta se deja evidencia de la necesidad de contar con una base de datos cada vez más amplia y lo más balanceada posible con el fin de que, las predicciones mejoren sus índices de fiabilidad. No obstante, se aclara que, para una exploración inicial, se destacan elementos interesantes que permiten la elección posterior de los árboles de decisión para el trabajo de interrelación de variables dentro de la especificidad del estudio.

Figura 4. *Rendimiento de los algoritmos de clasificación*

TIPO		MAT				CS				CN				ING				LEC			
	ALGORITMO	PRECISION	ACCURACY	RECALL	F1-SCORE	PRECISION	ACCURACY	RECALL	F1-SCORE	PRECISION	ACCURACY	RECALL	F1-SCORE	PRECISION	ACCURACY	RECALL	F1-SCORE	PRECISION	ACCURACY	RECALL	F1-SCORE
CLASIFICADOR	LR	0.66	0.44	0.64	0.65	0.62	0.56	0.65	0.63	0.76	0.68	0.62	0.69	0.58	0.58	0.66	0.63	0.86	0.57	0.63	0.6
	DT	0.65	0.55	0.55	0.57	0.72	0.68	0.68	0.69	0.75	0.65	0.65	0.68	0.73	0.65	0.65	0.68	0.68	0.65	0.65	0.66
	KNN	0.65	0.68	0.68	0.64	0.87	0.84	0.84	0.82	0.75	0.68	0.68	0.7	0.71	0.72	0.72	0.71	0.41	0.48	0.48	0.44
	SVM	0.52	0.52	0.52	0.52	0.42	0.52	0.52	0.47	0.6	0.6	0.6	0.6	0.58	0.56	0.56	0.57	0.52	0.52	0.52	0.52
	NN	0.6	0.64	0.64	0.61	0.59	0.4	0.4	0.42	0.61	0.44	0.44	0.48	0.61	0.36	0.36	0.4	0.6	0.36	0.36	0.38

Fuente: Elaboración propia

De acuerdo a los propósitos orientados al campo educativo del trabajo, de manera paralela a la prueba de algoritmos, y, en prospectiva, se realizó una fase de diseño de un entorno virtual en el que, se integró el algoritmo seleccionado posterior a la fase de evaluación. Tal aplicativo se codificó haciendo uso de PHP e integrando módulos de *Python* para probar las herramientas ML; el uso de PHP como lenguaje de programación, de acuerdo con Heuter (2016), Ortiz y Rojas (2018), para el desarrollo de la aplicación, se basó en su amplio reconocimiento en el ámbito del desarrollo web. Específicamente,

https://doi.org/10.56219/dialctica.v2i25.4065

DIALÉCTICA
REVISTA DE INVESTIGACIÓN EDUCATIVA
pp. 65 – 105
Número 25, Vol. 2 (2025)
ISSN en Línea: 2244-7490 / ISSN Impreso: 1316-7243
Depósito Legal: pp 1999802TA777
PORTAFOLIO DE INVESTIGACION

se eligió un lenguaje del lado del servidor (backend) debido a sus características, que lo hacen adecuado y propicio para su implementación en servidores de aplicaciones. Además, se adoptó PHP como lenguaje central para la estructura del sitio web, principalmente porque cuenta con soporte en una amplia gama de plataformas y servidores web. Este respaldo elimina la necesidad de configuraciones adicionales, ya que suele venir preinstalado por defecto en estos sistemas.

Por otro lado, la decisión de utilizar el algoritmo de Árbol de Decisión (*Decision Tree*) se fundamentó en su desempeño consistentemente sólido en todas las variables objetivo, ubicándose apenas por debajo de KNN en términos de precisión predictiva, como se puede apreciar en la figura anterior. En la segunda fase de pruebas, el algoritmo reveló relaciones significativas dentro de cada variable objetivo. En adición, su selección estuvo respaldada por las siguientes tres razones:

En primera instancia, el algoritmo destaca explícitamente las variables sociodemográficas más influyentes en cada nodo del árbol.

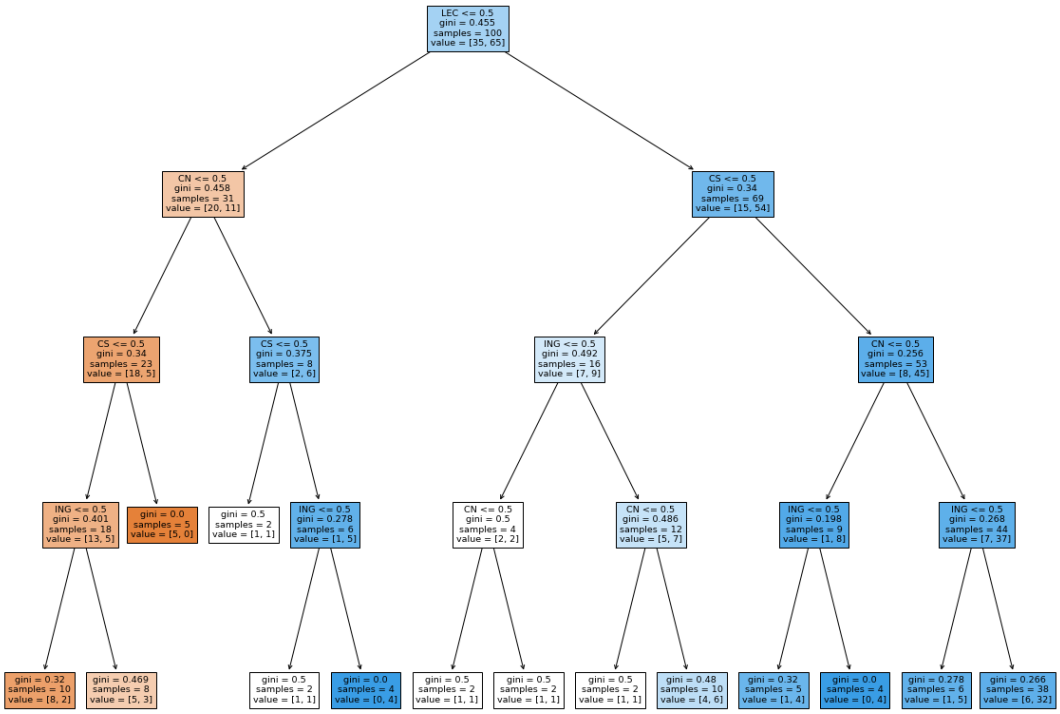
Segundo, genera salidas visuales de fácil interpretación, permitiendo que los educadores independientemente de su familiaridad con herramientas computacionales puedan comprender y utilizar los resultados efectivamente. Además, porque a pesar de su simplicidad, es importante que se pueda detallar el funcionamiento interno del *DecisionTreeClassifier*. Para este propósito, se adoptó como referencia principal el

https://doi.org/10.56219/dialctica.v2i25.4065

algoritmo CART (*Classification And Regression Trees*), implementado mediante la biblioteca Scikit-learn de Python.

A continuación, en la figura 5, se ilustra un árbol de decisión referido a la salida Matemáticas, arrojado directamente por la interfaz trabajada en la plataforma Anaconda y que, como se puede apreciar, contiene distintos parámetros que deben interpretarse para considerar concluir las posibles predicciones. En términos generales, para tratar de interpretarlo apropiadamente, sin un gran fundamento en términos de computación o programación pude decirse que, si se cumple con lo mostrado en cada nodo (recuadros) nos vamos por las ramas de la izquierda y si no, vamos por las ramas derechas del árbol.

Figura 5. *Relación entre desempeños – matemáticas*



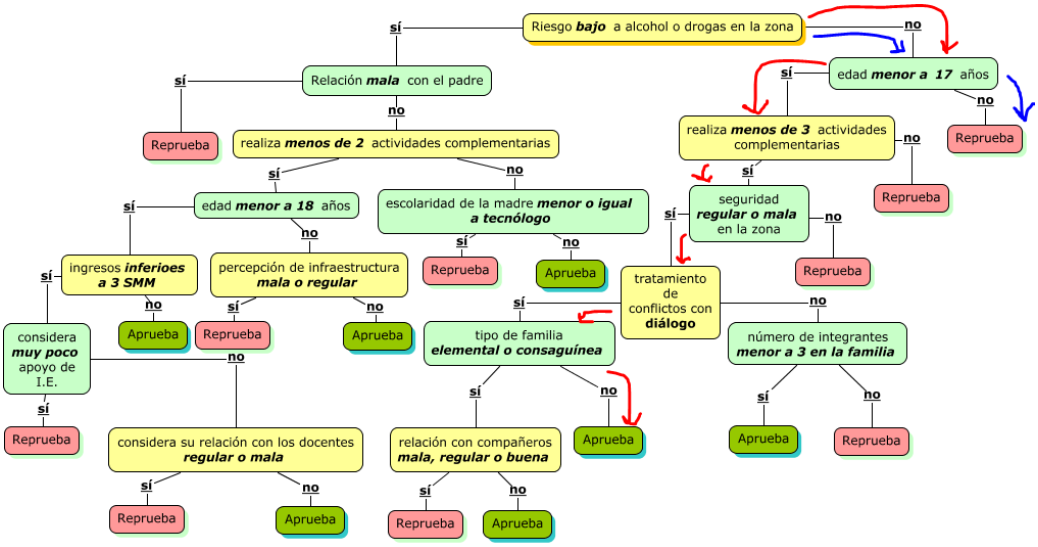
Fuente: Elaboración propia

https://doi.org/10.56219/dialctica.v2i25.4065

DIALÉCTICA
REVISTA DE INVESTIGACIÓN EDUCATIVA
pp. 65 – 105
Número 25, Vol. 2 (2025)
ISSN en Línea: 2244-7490 / ISSN Impreso: 1316-7243
Depósito Legal: pp 1999802TA777

De acuerdo con la prospectiva pedagógica del trabajo investigativo mostrado en el presente artículo, se considera necesario que, los árboles de decisión sean de fácil comprensión para una buena interpretación desde cualquier ente de las instituciones, por tal motivo, en la figura 6. Se ilustra el mismo árbol, expuesto en la figura 5, pero en una versión más amigable con el usuario y en el que se clasifican dos estudiantes nuevos (línea roja y azul) a modo de ejemplo, para considerar si podría o no, presentar dificultades en la competencia básica ya mencionada. De este modo, la línea roja predice que, si el estudiante no presenta riesgo bajo a alcohol o drogas, se va a la derecha al predictor edad, donde si es menor de 17 años, realiza menos de 3 actividades complementarias por semana, considera regular o mala la seguridad en su zona, percibe que el tratamiento de conflictos en su hogar es mediante el diálogo y no tiene familia elemental o consanguínea, entonces aprobaría el área de matemáticas.

Figura 6. *Árbol de decisión para predecir dificultades en la salida MAT*



https://doi.org/10.56219/dialctica.v2i25.4065

Fuente: Elaboración propia

Por último, un elemento adicional que ofrece este algoritmo, radica en su capacidad para evaluar cuantitativamente la importancia de los atributos involucrados en el proceso de clasificación. Esto permite identificar en qué medida las variables de entrada influyen sobre las variables de salida. La Figura 7 ilustra, de manera jerárquica, el nivel de influencia que cada variable de entrada ejerce en relación con cada variable objetivo.

Figura 7. *Jerarquía de los predictores respecto a cada variable objetivo*

Matemáticas		Lectura		Ciencias Nat.		Ciencias Soc.		Inglés	
Predic.	Import.	Predic.	Import.	Predic.	Import.	Predic.	Import.	Predic.	Import.
RIES_DR_ALC	0.203987	ESC_M	0.173797	ACT_COMP	0.166066	REL_M	0.163132	ESC_M	0.420174
ED	0.180532	ESC_P	0.151308	REL_D	0.151774	ESC_M	0.126984	LEE	0.159928
REL_C	0.142265	ED	0.140505	INFR	0.145457	GEN_B	0.115372	REL_D	0.110868
APOY	0.09382	NU_INT	0.138236	REL_P	0.127392	LEJA	0.079275	SEG_ZO	0.079912
ESC_M	0.092956	APOY	0.108134	REL_M	0.122238	NU_INT	0.073409	REPR	0.065082
TP_EDU	0.081355	RIES_DR_ALC	0.080834	GEN_B	0.082966	INFR	0.061175	NU_INT	0.058385
REL_P	0.066726	INFR	0.078498	SEG_ZO	0.056487	ESC_P	0.05793	ACT_COMP	0.050044
LEJA	0.056398	LEJA	0.070169	REL_C	0.051601	LEE	0.04941	INGR	0.044484
GEN_B	0.043403	ACT_COMP	0.058518	REPR	0.042809	ACT_COMP	0.049231	REL_M	0.011121
LEE	0.038558			LEJA	0.038048	REL_P	0.048348		
				RIES_DR_ALC	0.015162	ED	0.04798		
						APOY	0.045881		
						INT	0.038929		
						RIES_DR_ALC	0.024591		
						TP_EDU	0.018352		

Fuente: Elaboración propia

https://doi.org/10.56219/dialctica.v2i25.4065

DIALÉCTICA
REVISTA DE INVESTIGACIÓN EDUCATIVA
pp. 65 – 105
Número 25, Vol. 2 (2025)
ISSN en Línea: 2244-7490 / ISSN Impreso: 1316-7243
Depósito Legal: pp 1999802TA777
PORTAFOLIO DE INVESTIGACION

Las anteriores relaciones, permitieron reflexionar en las instituciones, sobre qué aspectos pueden resultar favorables para estudiarse en pro de prevenir posibles dificultades en cada una de las competencias básicas. Además, qué predictores eran comunes a varias salidas.

https://doi.org/10.56219/dialctica.v2i25.4065

Conclusiones

De acuerdo con el primero de los objetivos planteados en el presente artículo, se recalca que, una caracterización adecuada de los aspectos sociodemográficos de los estudiantes puede ir más allá de la recopilación de datos carente de trascendencia, para convertirse en un insumo que, organizado y bien orientado puede permitir análisis inferenciales y predictivos que ayudarían a mejorar prácticas pedagógicas. En este sentido y con base en la revisión de estudios previos y el marco teórico, puede inferirse que las herramientas de machine learning están adquiriendo una relevancia creciente dentro del sistema educativo debido a los beneficios que ofrecen para procesar grandes volúmenes de información.

Las herramientas de ML no solo permiten describir y representar diversos conglomerados de datos, sino que también contribuyen significativamente mediante prácticas consistentes de programación, a los procesos de toma de decisiones en contextos académicos y sociales dentro de las instituciones educativas. Esto se debe en gran medida a que se apoyan en modelos probabilísticos que pueden guiar de manera más efectiva las acciones orientadas a mejorar los resultados educativos en términos de formación y resultados académicos.

En este contexto, las herramientas computacionales antes mencionadas, asociadas al machine learning, ayudan a esclarecer el intrincado panorama del Big Data,

https://doi.org/10.56219/dialctica.v2i25.4065

DIALÉCTICA
REVISTA DE INVESTIGACIÓN EDUCATIVA
pp. 65 – 105
Número 25, Vol. 2 (2025)
ISSN en Línea: 2244-7490 / ISSN Impreso: 1316-7243
Depósito Legal: pp 1999802TA777
PORTAFOLIO DE INVESTIGACION

generando conocimiento que clarifica cómo abordar fenómenos educativos complejos. Así, para los propósitos específicos de este estudio, su aplicación proporcionó elementos valiosos para repensar las prácticas pedagógicas, promover la enseñanza basada en competencias y reconocer la individualidad del estudiante al identificar posibles relaciones entre diversas variables inherentes a su singularidad.

En vista de lo anterior, y considerando el segundo de los objetivos planteados, la presente investigación demuestra que sí es posible establecer relaciones significativas entre las características sociodemográficas de los estudiantes analizados y su desempeño académico en competencias básicas. No obstante, es importante enfatizar que la interpretación de los resultados derivados de herramientas computacionales debe abordarse con discreción y siempre alinearse con la naturaleza del problema de investigación y los datos utilizados, independientemente de que las métricas de desempeño cumplan con las expectativas o no.

Por otro lado, más allá de identificar relaciones, las técnicas de machine learning también permiten cuantificar el grado de influencia entre las variables involucradas. En este estudio, la implementación del *DecisionTreeClassifier* facilitó establecer dichas relaciones, junto con la posibilidad de jerarquizar su importancia relativa para cada una de las variables de salida analizadas. Con ello, el análisis combinado de variables sociodemográficas y desempeño académico, así como las relaciones observadas entre

https://doi.org/10.56219/dialctica.v2i25.4065

las propias variables de desempeño, llevó a los docentes a reflexionar sobre: 1. El valor de las herramientas computacionales, 2. La necesidad de abordar problemáticas educativas desde una perspectiva de sistemas complejos, y 3. El impacto potencial que estos hallazgos podrían tener en las prácticas pedagógicas.

Como resultado de las interrelaciones identificadas entre áreas disciplinares, se podrían proponer mejoras mediante trabajos interdisciplinarios y transversales, particularmente a través de metodologías como el aprendizaje basado en problemas y el aprendizaje basado en proyectos, estrategias que, potenciadas con nuevas tecnologías y situaciones que resulten significativas a los estudiantes, podrían promover el interés, la motivación y el desarrollo formativo buscado desde las instituciones educativas.

Con respecto al tercero de los objetivos, es de anotar que, la fiabilidad de las predicciones independientemente del algoritmo que se use, está sujeto a varios elementos imperativos para el rendimiento de las herramientas, así, se establece que, la validación y fiabilidad se relaciona estrictamente con bases de datos lo suficientemente grandes (y entre más grandes mejor) como para que el entrenamiento de la herramienta sea óptimo y con el balanceo de dichas bases de datos, es decir con *datasets* que contengan similar número de muestras para las categorías de las salidas, en este caso, por ejemplo, el no contar con suficientes estudiantes que aprueba y reprueban cada una de las cinco variables objetivo, conlleva inconvenientes en el entrenamiento del algoritmo para predicciones posteriores pertinentes.

https://doi.org/10.56219/dialctica.v2i25.4065

IALÉCTICA
REVISTA DE INVESTIGACIÓN EDUCATIVA
pp. 65 – 105
Número 25, Vol. 2 (2025)
ISSN en Línea: 2244-7490 / ISSN Impreso: 1316-7243
Depósito Legal: pp 1999802TA777
PORTAFOLIO DE INVESTIGACION

Otro elemento importante para destacar es que, los árboles de decisión generados o, en sentido genérico, las predicciones hechas, son de acuerdo al grupo que conforma la base de datos que sirve como insumo, es decir, las predicciones hechas en una base de datos pueden (y quizá deberían) variar en un grupo poblacional distinto, puesto que, las características de singularidad son inevitablemente diferentes, de este modo, se manifiesta la importancia de enriquecer las bases de datos con cada vez más y más muestras de estudiantes, con ello, se entrenan mejor los algoritmos y se hacen cada vez más eficientes las predicciones.

De acuerdo con la presentación de la jerarquía de predictores que permite ver qué variables de entrada se mostraron más relevantes para las distintas variables de salida (o desempeños académicos), en prospectiva, esta investigación resalta el potencial para: Extender el análisis de relaciones entre áreas disciplinares (designando cada una como variable objetivo –o predictor–) a todos los campos del conocimiento impartidos en instituciones educativas. También, se podría extrapolar a examinar las ocupaciones de los egresados en relación con sus perfiles sociodemográficos y su desempeño académico registrado previamente, con dos objetivos principales: 1. Ofrecer una orientación profesional más dirigida a los estudiantes actuales, y 2. Explorar la viabilidad de predecir decisiones profesionales futuras mediante un enfoque analítico similar.

https://doi.org/10.56219/dialctica.v2i25.4065

REFERENCIAS

Ariza-López F. J., Rodríguez-Avi, J., Alba-Fernández, V. (2018). *Control estricto de matrices de confusión por medio de distribuciones multinomiales*, GeoFocus (Artículos), nº 21, p. 215-226. ISSN: 1578-5157. http://dx.doi.org/10.21138/GF.591

Bausili Llamas, E. (2021). Machine Learning para el diagnóstico de COVID-19 [Trabajo de Grado]. Madrid, España. https://repositorio.comillas.edu/xmlui/handle/11531/53241

Carrion, J. (2017). Diferencia entre dato información y conocimiento. http://148.202.167.116:8080/jspui/handle/123456789/869

Castro, R. I. (2020). Aplicación de técnicas de Machine Learning para el estudio de deserción temprana y egreso oportuno en estudiantes de Ingeniería de la Facultad de Ciencias Físicas y Matemáticas [Trabajo de Grado]. Santiago de Chile, Chile. https://repositorio.uchile.cl/handle/2250/178598

Correa, M., Bielza, C., Pamies-Texeira, J., y Alique López, J. R. (2008). Redes Bayesianas vs redes neuronales en modelos para la predicción del acabado superficial. Madrid, España. https://digital.csic.es/handle/10261/13826

Joaquín, R. (2020). Gradient Boosting con Python by, available under a Attribution 4.0 International (CC BY 4.0) https://www.cienciadedatos.net/documentos/py09_gradient_boosting_python.html

https://doi.org/10.56219/dialctica.v2i25.4065

pp. 65 – 105
Número 25, Vol. 2 (2025)
ISSN en Línea: 2244-7490 / ISSN Impreso: 1316-7243
Depósito Legal: pp 1999802TA777

Jones- Ortiz, C. V. ., & Guzmán – Seraquive, J. E. . (2022). Análisis de las técnicas de machine learning aplicadas en la detección de fraudes bancarios. *Revista Científica Ciencia Y Tecnología*, 22(33). https://doi.org/10.47189/rcct.v22i33.516

Beunza, J. J., y Puertas, E. (2020). Tipos de algoritmos. Manual práctico de inteligencia artificial en entornos sanitarios, 35. https://books.google.com/books?hl=en&lr=&id=88nSDwAAQBAJ&oi=fnd&pg=PA35&dq=machine+learning,+tipos+de+aprendizaje&ots=6Q8dGGZ6T_&sig=wDjkify1m62oZMRtuCuJZus4kho

Bogoya, D. (2000). Resultados de la evaluación de competencias básica en lenguaje, matemáticas y ciencias. *Bogotá*, *Secretaría de educación distrital*.

Bunk, G. (1994). La transmisión de las competencias en la formación y perfeccionamiento profesionales de la RFA. *Revista europea de formación profesional*, (1), 8-14. https://dialnet.unirioja.es/descarga/articulo/131116.pdf

Florez, C. y Pulido, K. (2020). Estrategia Didáctica mediada por Técnicas de Machine Learning para Potenciar la Habilidad de Pensamiento Crítico Interdisciplinar en Procesos Académicos y Actitudinales. https://repositoriotesiscomplejidad.blogspot.com/

Hernández-Sampieri, R., Fernández Collado, C., & Baptista Lucio, P. (2018). *Metodología de la investigación* (Vol. 4, pp. 310-386). México: McGraw-Hill Interamericana. https://dspace.scz.ucb.edu.bo/dspace/bitstream/123456789/21401/1/11699.pdf

https://doi.org/10.56219/dialctica.v2i25.4065

Méndez, A. (2007). Terminología pedagógica específica al enfoque por competencias: el concepto de competencia. Innovación educativa. https://redined.educacion.gob.es/xmlui/handle/11162/75735

Méndez, O. y López, J. (2019). Técnicas de Machine Learning para la predicción de desempeño académico en el Desarrollo del espacio proyectivo del Pensamiento Espacial. [Tesis de maestría, Universidad Pedagógica Nacional]. Bogotá, Colombia. https://docs.google.com/document/d/11-pYeDpzAg3Y9wRbczjyEENjtQXo2cJH/edit#

Piaget, J. (1981). La teoría de Piaget. Infancia y aprendizaje, 4(sup2), pp. 13-54. https://www.tandfonline.com/doi/pdf/10.1080/02103702.1981.10821902

Ramírez, D. (2021). Análisis del Machine Learning como estrategia didáctica para el mejoramiento de Competencias Básicas en Educación Secundaría en Entornos Virtuales de Aprendizaje. https://gredos.usal.es/handle/10366/146077

Sucar, L. E., y Tonantzintla, M. (2006). Redes bayesianas. Aprendizaje Automático: conceptos básicos y avanzados, 77, 100. https://ccc.inaoep.mx/~esucar/Clases-mgp/caprb.pdf

Tobón, S. (2003). Las competencias en el sistema educativo: de la simplicidad a la complejidad. Bogotá: CIFE. http://files.nathalyeismenia.webnode.com.ve/200000134-bff07c1e3d/LAS%20COMPETENCIAS%20Y%20EL%20PENSAMIENTO%20COMPLEJO.doc

UNESCO (2015). Replantear la educación ¿Hacia un bien común mundial? París: UNESCO. 93 pp. ISBN-978-92-3-300018.6. *Journal of Supranational Policies of Education*. https://repositorio.uam.es/bitstream/handle/10486/671277/-JOSPOE_4_15.pdf?s

https://doi.org/10.56219/dialctica.v2i25.4065

DIALÉCTICA
REVISTA DE INVESTIGACIÓN EDUCATIVA
pp. 65 – 105
Número 25, Vol. 2 (2025)
ISSN en Línea: 2244-7490 / ISSN Impreso: 1316-7243
Depósito Legal: pp 1999802TA777
PORTAFOLIO DE INVESTIGACION

Vigotsky, L. (1988). El desarrollo de los procesos psicológicos superiores. México: Editorial Crítica, Grupo Editorial Grijalbo. https://www.bibliopsi.org/docs/carreras/obligatorias/CFP-/educacional/erausquin/Unidad%202/Vigotsky%2%20el%20desarrollo%20de%20los%20p.%20cap%204.pdf

https://doi.org/10.56219/dialctica.v2i25.4065