Universidad Externado de Colombia

FACULTAD DE ECONOMÍA

IA ECONOMIA

Laboratorio de inteligencia artificial aplicada a Economía

Tu conocimiento, nuestro boletín

Edición
Santiago Mucia

Diseño y diagramación
Diana Medina Ospina

Boletín #1

2025

Tabla
de contenido

Prefacio

Desde la Facultad de Economía de la Universidad Externado de Colombia somos conscientes de la importancia creciente que tiene en el mundo de la cuarta revolución industrial los desarrollos que en los últimos años se han venido gestando en inteligencia artificial aplicados a la economía.

De esa manera, comprometidos con la construcción de puentes entre el sector real, las necesidades de las empresas y el potencial de la academia para proponer soluciones, hemos creado EconomIA, un laboratorio de inteligencia artificial concebido como plataforma de interacción y diálogo constructivo, tendiente a resaltar las principales ideas y proyectos de nuestra comunidad académica de estudiantes y profesores de la facultad quienes, producto de sus investigaciones y actividades de docencia y extensión, desarrollan y materializan soluciones para el sector productivo utilizando la inteligencia artificial desde la economía.

Fruto de este valioso esfuerzo hemos querido recoger una selección de los mejores proyectos de nuestra comunidad para ser presentados al público en general a través de este boletín, con la firme convicción de lograr consolidar un repositorio y carta de presentación del esfuerzo de la comunidad académica de la Facultad de Economía en materia de inteligencia artificial de cara al sector productivo.

Con este documento, no solamente se busca visibilizar este esfuerzo realizado, sino conectar capacidades y talentos con empresas ávidas de poder encontrar soluciones a los retos que impone esta era de permanente cambio tecnológico y continua evolución en la manera de interactuar con sus usuarios.

Así, esperamos incidir decididamente en un propósito totalmente alineado con la idea propuesta recientemente por Ricardo Hausmann, connotado economista director del Centro para el Desarrollo Internacional de la Universidad de Harvard y profesor del Instituto Santa Fe, quien en una conferencia ofrecida el pasado 14 de marzo de este año titulada ¿Por qué seguimos

estancados?[1] ha manifestado como una necesidad apremiante, en respuesta a la pregunta que titula su charla, repensar la lógica de las universidades en su relación con la empresa.

Coincidimos con Hasumann en que los esfuerzos en I+D deben surgir en parte como resultado de la sinergia universidad empresa y como fue manifestado líneas arriba, contar con ese propósito en materia de aplicaciones de la IA a la economía es un paso importante para reafirmar esa sinergia.

Por todo lo anterior, y partiendo de que este boletín es una realidad en su primer número, quisiera celebrar esta iniciativa, y felicitar muy especialmente a nuestro Coordinador Académico de Pregrado en economía, profesor Santiago Murcia Alzate, quien ha sido el artífice tanto del laboratorio EconomIA, como de lo que ha significado darle vida a esta publicación. Su liderazgo proponiendo y materializando este proyecto es cuando menos meritorio. Por supuesto es también la ocasión para felicitar y agradecer a cada uno de los autores de los artículos que resumen los proyectos desarrollados por su esfuerzo y esmero para visibilizar su trabajo y agradecer por supuesto a nuestros lectores de este primer boletín del laboratorio con el especial anhelo de que esta iniciativa académica en este frente sea la primera de muchas que puedan ser posibles en la construcción del futuro de nuestra facultad.

Bienvenido el laboratorio de EconomIA a nuestra Facultad de Economía de la Universidad Externado de Colombia.

Juan Pablo Herrera Saavedra
Decano
Facultad de Economía
Universidad Externado de Colombia

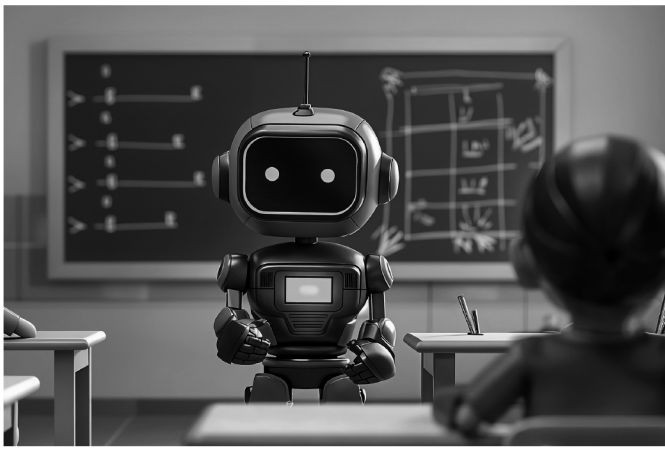
[1] Conferencia: ¿Por qué seguimos estancados? Disponible en: https://politicaspublicas.com.co/porque-seguimos-estancados/

Universidad Externado de Colombia

FACULTAD DE ECONOMÍA

ECONOMIA

Laboratorio de inteligencia artificial aplicada a Economía



Trazado de conocimiento estudiantil:

Una arquitectura basada en LLMs, grafos de conocimiento y razonamiento bayesiano

Autor

Daniel Godoy Ortiz andres.godoy1@uexternado.edu.co

2025

Resumen

El progreso de una nación depende en gran medida de su capacidad para formar una ciudadanía educada con calidad. Sin embargo, en Colombia persiste una debilidad estructural que impide monitorear el aprendizaje estudiantil de forma continua: las únicas mediciones oficiales se realizan al final del ciclo escolar, lo que impide detectar a tiempo las brechas de conocimiento. Esta carencia de datos oportunos afecta la efectividad de las políticas públicas y perpetúa la desigualdad.

A pesar de que cada estudiante produce diariamente evidencia de aprendizaje, el sistema educativo carece de mecanismos para analizarla sistemáticamente.

Este artículo propone una línea de investigación en la que se puede avanzar para aprovechar los avances recientes en inteligencia artificial, en particular los modelos de lenguaje de gran tamaño (LLMs), el trazado bayesiano de conocimiento y los grafos conceptuales, para desarrollar un software que permita transformar tareas y exámenes existentes en diagnósticos dinámicos y personalizados, generando retroalimentación automatizada y apoyo tanto al docente como a la toma de decisiones políticas.

La propuesta busca discutir una solución escalable, sostenible y adaptada al contexto colombiano, que haga posible una política educativa verdaderamente informada por evidencia granular y continua.

Introducción

Para que un país pueda aspirar a un futuro más justo, próspero y competitivo, necesita una población cada vez más educada, no solo en términos de cobertura, sino sobre todo en términos de calidad. En Colombia, esta premisa se enfrenta a una realidad preocupante: el sistema educativo carece de herramientas efectivas para monitorear el proceso de aprendizaje a lo largo del tiempo. A pesar de los esfuerzos por ampliar el acceso a la educación básica y media, sigue sin resolverse un problema estructural: no se cuenta con información suficiente, oportuna y precisa sobre cómo aprenden los estudiantes, qué están entendiendo —y qué no— durante los años cruciales de su formación.

La educación es, por naturaleza, un proceso profundamente individual. Si bien existen patrones comunes de desarrollo cognitivo y estructuras curriculares compartidas, cada estudiante avanza con ritmos, estilos y trayectorias diferentes. En un escenario ideal, cada persona requeriría un acompañamiento personalizado para identificar sus fortalezas, detectar tempranamente sus dificultades y ajustar los contenidos y estrategias pedagógicas en consecuencia. Pero en la práctica, esto es inviable: no hay manera de asignar varios tutores de diferentes áreas del conocimiento a un solo estudiante. Por eso, el sistema adopta soluciones logísticas: agrupa a los estudiantes en aulas y encarga a un docente la tarea de guiar simultáneamente decenas de procesos de aprendizaje distintos, muchas veces con recursos limitados y en contextos de alta vulnerabilidad social.

Ante esa complejidad, se esperaría al menos contar con mecanismos que permitan observar el progreso académico de cada estudiante a lo largo del tiempo, para mejorar las intervenciones de política pública y la focalización de recursos. Sin embargo, en Colombia, la única medición poblacional oficial de resultados de aprendizaje es la prueba Saber 11, que se realiza en el último año de escolaridad, cuando el estudiante está a punto de graduarse. En otras palabras, el país intenta evaluar el resultado de un proceso educativo de más de una década con una sola medición final, sin haber realizado (a nivel poblacional) inspecciones periódicas, sin identificar a tiempo dónde hay debilidades, y sin posibilidad de

hacer ajustes significativos antes de que sea demasiado tarde. Las consecuencias de esta ceguera estructural son profundas: los problemas de aprendizaje se acumulan sin ser detectados, las políticas públicas de calidad de la educación son susceptibles de diseñarse a ciegas, y los esfuerzos de mejora rara vez tienen un sistema que permita evaluar su efectividad.

Además, en un país caracterizado por su diversidad regional, desigualdad social y brechas estructurales, esta falta de información alimenta aún más la inequidad. Quienes se rezagan en silencio —por pobreza, por dificultades cognitivas no diagnosticadas, por falta de acompañamiento familiar— tienen menos probabilidades de recibir ayuda oportuna. Y como el sistema no detecta esas señales de alerta a tiempo, termina reforzando las desigualdades que prometía superar.

Ante esta situación, el desarrollo tecnológico abre una oportunidad transformadora. En particular, los recientes avances en inteligencia artificial, y especialmente en modelos de lenguaje de gran tamaño (LLMs), hacen posible imaginar un nuevo tipo de sistema: uno que permita realizar evaluaciones diagnósticas continuas, adaptativas e interactivas del conocimiento de los estudiantes, sin necesidad de imponer nuevas pruebas estandarizadas ni aumentar la carga docente. El desarrollo de modelos de lenguaje de gran tamaño (LLMs) y sistemas de reconocimiento inteligente de texto o imagen abre la puerta a la evaluación continua sin necesidad de crear nuevos instrumentos de medición. Basta con analizar tareas y exámenes que ya se realizan en el aula, generando diagnósticos precisos de manera casi automática. Esto brindaría a los docentes retroalimentación instantánea y permitiría a los diseñadores de políticas identificar patrones de aprendizaje a gran escala. Lo que marcaría una transformación radical para los hacedores de política pública: pasar de disparar en la oscuridad a contar con un mapa detallado del panorama educativo, en tiempo real y con posibilidades de acción inmediata.

Este artículo explora, en la siguiente sección, los marcos teóricos y hallazgos recientes en torno a la evaluación continua del aprendizaje, incluyendo tanto enfoques pedagógicos como avances tecnológicos en inteligencia artificial aplicados a la educación. A partir de esa base conceptual, en la última sección se propone de forma general una arquitectura técnica para un sistema inteligente que permita monitorear de forma continua y adaptativa el conocimiento de los estudiantes a partir de las evidencias ya presentes en el aula, integrando modelos de lenguaje de gran tamaño (LLMs), grafos de conocimiento y métodos probabilísticos para el trazado de dominios conceptuales. Esta propuesta general busca servir como punto de partida para la investigación en el diseño de soluciones escalables, sostenibles y alineadas con las condiciones del sistema educativo colombiano.

Revisión de Literatura

En Colombia, la evaluación de la calidad educativa ha estado dominada por pruebas estandarizadas de carácter sumativo, con alcance y frecuencia limitados. Históricamente, el ICFES (Instituto Colombiano para la Evaluación de la Educación) instauró exámenes nacionales como el de Estado (hoy Saber 11) para egresados de secundaria, y posteriormente expandió la evaluación a otros niveles con las pruebas Saber de 3°, 5° y 9° grado. Sin embargo, estas últimas se aplican a solo cerca del 2% de los estudiantes del sistema escolar (ICFES, sf).

Así mismo, un informe de la OCDE (2016) señaló que en Colombia la evaluación del aprendizaje es predominantemente sumativa y no se utiliza de forma eficaz para mejorar la enseñanza subsecuente. Los resultados suelen comunicarse como un puntaje numérico aislado, sin brindar a cada alumno información detallada sobre sus fortalezas, dificultades o recomendaciones específicas de mejora.

En la investigación educativa contemporánea, la evaluación continua y personalizada se sustenta en varios marcos teóricos complementarios: la evaluación formativa (assessment for learning), el aprendizaje adaptativo y la instrucción guiada por datos (data-driven instruction). Todos convergen en la idea de que la evaluación debe ser parte integral del proceso de enseñanza-aprendizaje, proporcionando feedback frecuente y ajustes individualizados para optimizar el progreso de cada estudiante.

Desde los trabajos seminales de Black y Wiliam (1998) se reconoce que la evaluación formativa en el aula –es decir, aquella dirigida a obtener evidencia del aprendizaje durante el proceso educativo para retroalimentar a alumnos y docentes– puede elevar significativamente el rendimiento estudiantil (Lui & Andrade, 2025).. Teóricamente, la evaluación formativa se alinea con enfoques constructivistas y de autorregulación: se busca que el estudiante participe activamente de su evaluación, reflexione sobre sus entendidos y desaciertos, y desarrolle metacognición acerca de cómo aprender mejor (Steinert et al., 2024).

Por otro lado, el término aprendizaje adaptativo se refiere a metodologías –a menudo asistidas por tecnología– que ajustan la trayectoria, el nivel de dificultad o el ritmo de enseñanza a las necesidades individuales de cada alumno. En esencia, requiere de evaluación continua: el sistema (sea un docente o un software) debe diagnosticar constantemente el nivel de dominio del estudiante en cada materia para decidir qué contenido o ejercicio presentar a continuación. Este principio tiene raíces en la teoría de la maestría del aprendizaje de Bloom (1968), quien postuló que casi cualquier estudiante puede alcanzar un alto desempeño si se le da el tiempo y apoyo necesarios, identificando y remediando sus lagunas antes de avanzar. En la era digital, el aprendizaje adaptativo se ha operacionalizado mediante plataformas e-learning e tutores inteligentes que monitorizan las respuestas del estudiante en tiempo real. Finalmente, la instrucción basada en datos (data-driven instruction) es precisamente el uso sistemático de información de evaluaciones para planear la enseñanza a nivel micro (aula) y macro (escuela o distrito).

Uno de los desarrollos tempranos más influyentes en IA educativa fue el modelado probabilístico del conocimiento del alumno. Corbett y Anderson (1995) introdujeron el concepto de Knowledge Tracing empleando redes bayesianas para representar el proceso de aprendizaje del estudiante –lo que se conoció como Bayesian Knowledge Tracing (BKT)–. En este enfoque, cada habilidad o concepto se modela como una variable oculta (estado de dominio o no dominio) que se actualiza conforme el alumno responde

preguntas o realiza tareas. El modelo bayesiano incorpora la incertidumbre en las respuestas: por ejemplo, estima la probabilidad de que un acierto sea por adivinanza o que un error sea por descuido, refinando así la estimación real del conocimiento del alumno. Esta técnica, integrada en muchos tutores inteligentes, permite que el sistema "sepa lo que el estudiante sabe" (hasta cierto grado de certeza) en cada momento y por tanto pueda personalizar la secuencia de ejercicios: si BKT infiere que el estudiante aún no domina un tema, continuará proponiendo actividades de práctica en ese tópico; cuando la probabilidad de dominio supera cierto umbral, el tutor pasa al siguiente tema. Durante más de 25 años, BKT y sus extensiones han sido ampliamente utilizados e investigados, en buena medida por su interpretabilidad y fundamento teórico sólido (Šarić-Grgić et al., 2024).

En tiempos recientes han surgido variantes más complejas –como el Deep Knowledge Tracing, que usa redes neuronales recurrentes para lograr predicciones aún más precisas de desempeño futuro– pero incluso estos incorporan ideas de incertidumbre y memoria basadas en BKT (Shen et al., 2024). Para el monitoreo continuo del aprendizaje, los modelos bayesianos ofrecen un marco riguroso: cada nueva respuesta del estudiante actualiza inmediatamente su perfil de dominio con inferencias estadísticas, posibilitando una evaluación fina item a item. Además, estos modelos brindan explicabilidad básica (ej., se puede reportar "el alumno tiene un ~80% de probabilidad de saber resolver ecuaciones lineales de un paso"), lo cual es útil para que docentes y alumnos entiendan el estado del aprendizaje en términos probabilísticos en lugar de solo puntuaciones estáticas.

El advenimiento de LLMs como GPT-3, GPT-4, PaLM, etc., a partir de 2020-2023 ha abierto posibilidades revolucionarias para la educación, especialmente en la evaluación de aprendizajes complejos de forma personalizada. Estos modelos, entrenados con billones de palabras, poseen una capacidad sin precedentes para comprender y generar lenguaje natural, lo cual les permite interactuar con estudiantes y sus producciones de manera muy cercana a como lo haría un humano experto.

En el contexto de la evaluación continua, los LLMs pueden desempeñar varios roles cruciales: *tutor virtual, generador de retroalimentación, evaluador automatizado y creador de contenido adaptativo*. Un avance destacado es el uso de LLMs para analizar respuestas abiertas y proveer retroalimentación instantánea. Por ejemplo, estudios recientes en educación superior muestran que la retroalimentación adaptativa generada por ChatGPT puede mejorar la calidad del desempeño estudiantil. En un experimento con 269 futuros docentes en Alemania, aquellos que recibieron feedback personalizado producido por ChatGPT sobre una tarea escrita de razonamiento pedagógico luego escribieron justificaciones de mejor calidad que aquellos que solo recibieron feedback estático de un experto (Kinder et al, 2025).

A diferencia de los sistemas pre-LLM, que se limitaban a dar hints preprogramados o calificar con rubricas simples, los LLMs ofrecen una capacidad de adaptación dinámica: pueden cambiar la explicación si notan que el estudiante sigue confundido, re-frasear en términos más sencillos, proponer analogías, etc., prácticamente en tiempo real. Esto significa que la frontera entre evaluar y enseñar se difumina positivamente –el mismo acto de evaluar (p. ej. hacer una pregunta y analizar la respuesta) se convierte en una oportunidad de aprendizaje mediante la conversación generada por la IA.

Paralelamente al auge de las redes neuronales y LLMs, la IA educativa reciente enfatiza la integración de representaciones simbólicas del conocimiento (ontologías, grafos conceptuales, reglas lógicas) con los métodos subsimbólicos (aprendizaje profundo) para lograr sistemas más robustos, explicables y pedagógicamente efectivos. Un grafo de conocimiento en educación es básicamente una representación estructurada

de los conceptos de una materia y las relaciones entre ellos (prerequisitos, jerarquías, similitudes, etc.), algo así como un mapa curricular inteligente. Los grafos de conocimiento permiten que un sistema educativo "sepa" la estructura de la disciplina y las rutas óptimas de aprendizaje. Por ejemplo, un grafo puede codificar que para aprender el concepto C es necesario dominar antes A y B; si el alumno falla reiteradamente en C, el sistema puede redirigirlo a reforzar A y B. Actualmente, los grafos de conocimiento son ampliamente utilizados para la planificación personalizada de rutas de aprendizaje en entornos en línea. Estudios han demostrado que algoritmos de recomendación basados en grafos pueden trazar secuencias de contenidos adaptadas que optimizan el progreso del alumno según sus conocimientos previos y objetivos (Hou et al., 2025).

Por ejemplo, Hou et al. (2023) proponen un método de planificación de rutas de aprendizaje personalizadas apoyado en grafos, donde se evalúa la similitud entre conceptos y el dominio del estudiante para decidir el siguiente contenido, logrando así trayectorias más efectivas que un orden fijo de curso.

La ventaja de los grafos es que aportan semántica y estructura a la IA educativa: el sistema puede explicar qué sabe o no el alumno en términos de conceptos ("manejas bien ecuaciones lineales pero aún no sistemas de ecuaciones") y por qué recomienda cierto ejercicio ("porque reforzará un prerrequisito clave"). Esto los hace ideales para aumentar la transparencia y explicabilidad de la evaluación continua, algo crítico para la aceptación por parte de docentes y administradores educativos. Los avances recientes van un paso más allá al combinar grafos de conocimiento con LLMs u otras IA subsimbólicas para aprovechar lo mejor de ambos mundos. Un ejemplo de frontera es el framework FOKE (Forest of Knowledge and Education) propuesto por Hu & Wang (2023), que integra modelos base (foundation models como GPT-4), grafos de conocimiento y prompt engineering en una sola arquitectura (Hu y Wang, 2024).

La convergencia de la necesidad pedagógica de una evaluación continua/personalizada y las capacidades técnicas aportadas por la IA moderna sientan las bases para un nuevo paradigma de evaluación educativa en Colombia. Los hallazgos internacionales indican que un enfoque formativo-adaptativo mejora los aprendizajes, y las tecnologías emergentes –LLMs, grafos, modelos bayesianos– ofrecen medios concretos para materializar ese enfoque a gran escala y de manera sostenible.

Propuesta de línea de investigación para el diseño de sistemas de monitoreo

Una vía viable y estratégica para resolver la falta de evaluación continua del aprendizaje en Colombia, sin recurrir a nuevas pruebas estandarizadas ni a mecanismos costosos de vigilancia externa, es desarrollar un sistema tecnológico que permita extraer información útil a partir de las evidencias que ya produce el sistema educativo: los exámenes, tareas escritas, guías y actividades desarrolladas por los estudiantes en el aula. A diferencia de muchas propuestas tecnocráticas que plantean construir nuevas métricas desde cero, este enfoque parte de una premisa pragmática: los profesores ya evalúan constantemente, los estudiantes ya producen material evaluable. Lo que falta no es evidencia, sino la capacidad de sistematizarla, interpretarla y transformarla en conocimiento accionable para la mejora pedagógica. Aquí es donde los avances recientes en inteligencia artificial pueden marcar la diferencia.

La arquitectura propuesta se fundamenta en cuatro pasos técnicos articulados. El primer paso es construir un sistema de captura digital ágil y descentralizado, que permita a los docentes tomar fotos con sus dispositivos móviles de las actividades escritas de los estudiantes, sin necesidad de modificar sus prácticas actuales. Estas imágenes se envían a una nube donde son procesadas por un motor de reconocimiento óptico de caracteres (OCR) especializado en escritura manuscrita en español; o un modelo fundacional.

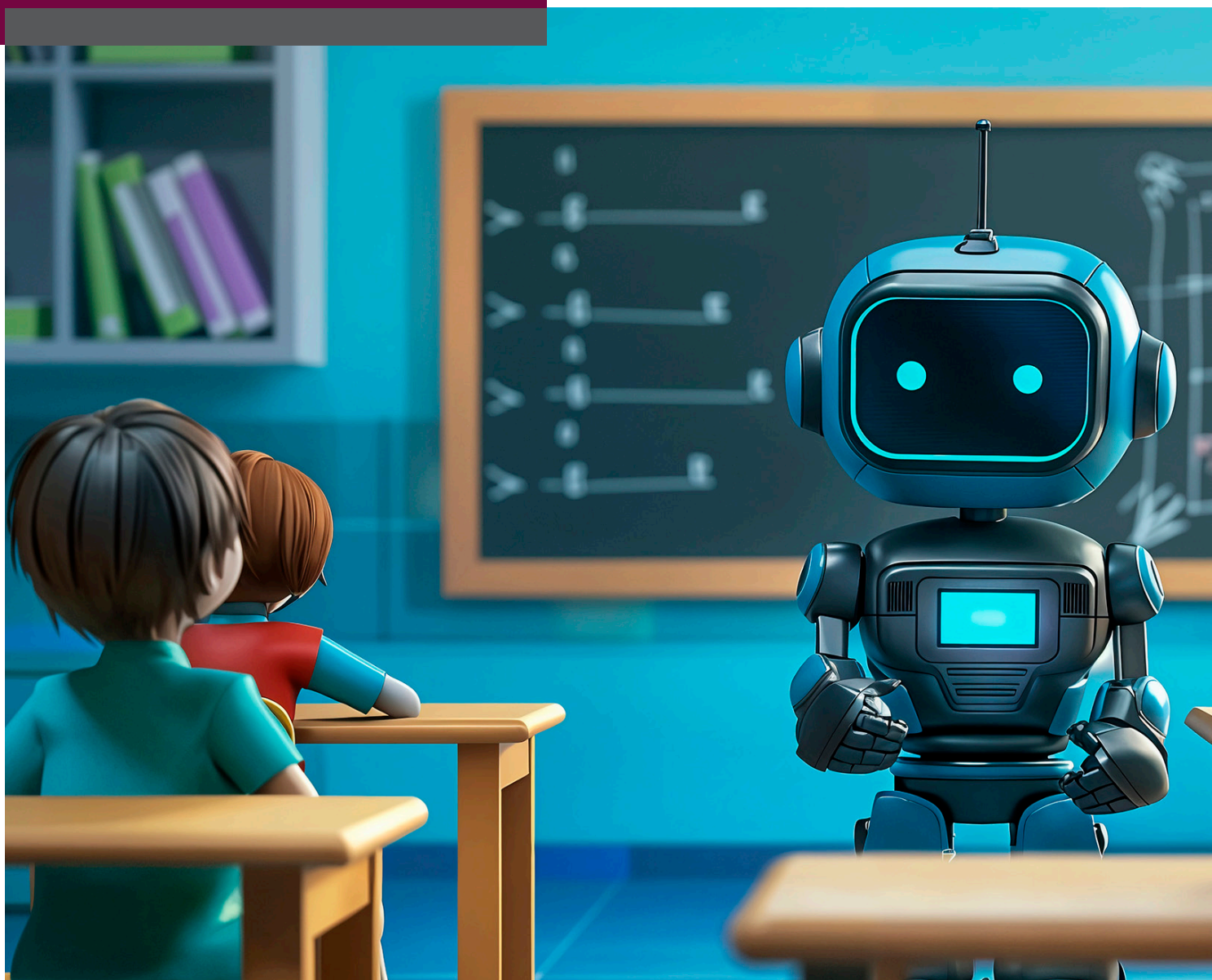
Una vez extraído el texto, se activa el segundo paso: el análisis semántico del contenido mediante un modelo de lenguaje de gran tamaño (LLM) finamente ajustado al dominio educativo. Este modelo tiene como función interpretar lo que escribió el estudiante, identificar el tema o competencia que está siendo evaluada, detectar errores frecuentes, inferir posibles brechas de comprensión y, en los casos necesarios, generar retroalimentación textual. Este análisis no se limita a clasificar respuestas como correctas o incorrectas; el objetivo es comprender las estrategias cognitivas utilizadas, los conceptos involucrados y las pistas que la escritura deja sobre el estado del conocimiento del estudiante. Aquí es donde el potencial de los LLMs, con su capacidad de interpretar lenguaje natural y generar explicaciones contextualizadas, permite que el sistema actúe como un evaluador experto automatizado, que no solo califica, sino que interpreta pedagógicamente cada evidencia.

En el tercer paso, los resultados del análisis semántico se proyectan sobre una base estructurada de conocimiento: un grafo de conocimiento curricular, donde cada nodo representa un concepto o habilidad del currículo nacional, y las aristas codifican relaciones de dependencia, jerarquía o similitud conceptual. Este grafo actúa como una ontología viva del aprendizaje: permite ubicar cada estudiante dentro de una progresión esperada de dominio, identificar los conceptos que ya ha consolidado, aquellos que aún presenta de forma parcial, y los que aún no domina. Con cada nueva evidencia escrita analizada, el perfil del estudiante se actualiza de forma probabilística, utilizando técnicas de trazado de conocimiento como el Bayesian Knowledge Tracing, que permiten modelar el aprendizaje como un proceso dinámico con incertidumbre. Así, el sistema no parte de una fotografía fija del desempeño, sino de una representación en movimiento, que traza una trayectoria de aprendizaje y puede inferir tendencias, estancamientos o avances.

El cuarto paso consiste en activar un motor de razonamiento pedagógico basado en lógica híbrida, que combina reglas simbólicas (por ejemplo, “si el estudiante falla en el concepto X y X depende de Y, revisar Y”) con inferencias subsimbólicas generadas por el LLM (por ejemplo,

“parece haber confusión entre proporción y razón, sugerir ejercicios de aplicación contextual”). Este componente permite generar recomendaciones didácticas específicas, como sugerencias de actividades, rutas de refuerzo personalizadas o agrupamiento de estudiantes con brechas similares. El sistema podría actuar directamente con el estudiante —por ejemplo, a través de un chatbot tutor que lo guíe paso a paso— o entregar informes granulares al docente, para que éste tome decisiones informadas sobre su enseñanza, sin verse reemplazado.

Este sistema puede ser implementado por etapas, comenzando con pilotos en regiones con conectividad básica y disponibilidad docente, y escalando progresivamente a otras zonas del país. Su sostenibilidad radica en que aprovecha infraestructura existente (celulares, tareas escritas, currículo nacional), automatiza procesos actualmente informales, y prioriza el apoyo pedagógico más que el control externo. Es, en esencia, una forma de ampliar la mirada del docente, de dotarlo con una red de inteligencia aumentada capaz de ver lo que, hasta ahora, solo podía intuir. Y para el Estado, representa una oportunidad inédita de contar con datos vivos, finos y continuos sobre el aprendizaje real de millones de estudiantes. En lugar de disparar a ciegas con políticas generales, podríamos, por primera vez, apuntar con precisión quirúrgica hacia donde realmente están las brechas.



CONCLUSIONES

La ausencia de mecanismos efectivos para monitorear el aprendizaje de forma continua ha limitado durante décadas la capacidad del sistema educativo colombiano para actuar con oportunidad, equidad y evidencia. Esta debilidad estructural ha impedido detectar brechas tempranas, orientar los esfuerzos docentes de manera precisa y evaluar con rigor el impacto de las políticas públicas. Sin embargo, los recientes avances en inteligencia artificial —particularmente en modelos de lenguaje de gran tamaño, trazado bayesiano de conocimiento y representación simbólica mediante grafos curriculares— ofrecen una oportunidad inédita para transformar la forma en que se comprende, evalúa y mejora el aprendizaje en el país. Aprovechando las evidencias que ya se producen en el aula, y sin necesidad de nuevos exámenes masivos, es posible imaginar un sistema que acompañe al estudiante a lo largo de su trayectoria escolar, generando diagnósticos dinámicos, retroalimentación automatizada y rutas de mejora personalizadas.

La arquitectura propuesta en este artículo traza una hoja de ruta técnica y pedagógica para desarrollar este tipo de sistemas inteligentes de evaluación continua. No se trata únicamente de una mejora tecnológica, sino de un cambio profundo en la relación entre enseñanza, evaluación y política pública. La posibilidad de contar con información granular, actualizada y accionable sobre el aprendizaje de millones de estudiantes representa un salto cualitativo en la capacidad del Estado y las instituciones para garantizar una educación más justa y efectiva.

Aunque el despliegue de este tipo de soluciones requiere investigación aplicada, pilotos controlados e inversión sostenida, su potencial transformador es claro: pasar de medir a enseñar con inteligencia, de diagnosticar tarde a intervenir a tiempo, y de diseñar políticas educativas en la oscuridad a hacerlo con la luz de los datos.

Referencias

A. T. Corbett and J. R. Anderson, “Knowledge tracing: Modeling the acquisition of procedural knowledge,” User Modelling and User-Adapted Interaction, vol. 4, no. 4, pp. 253–278, 1995.

Bloom, B. S. (1968). Learning for Mastery. Instruction and Curriculum. Regional Education Laboratory for the Carolinas and Virginia, Topical Papers and Reprints, Number 1.

Hou, B., Lin, Y., Li, Y., Fang, C., Li, C., & Wang, X. (2025). KG-PLPPM: A Knowledge Graph-Based Personal Learning Path Planning Method Used in Online Learning. Electronics, 14(2), 255. https://doi.org/10.3390/electronics14020255

Hu, S., & Wang, X. (2024, August). Foke: A personalized and explainable education framework integrating foundation models, knowledge graphs, and prompt engineering. In China National Conference on Big Data and Social Computing (pp. 399-411). Singapore: Springer Nature Singapore.

Instituto Colombiano para la Evaluación de la Educación – ICFES. (s.f.). Pruebas Saber 3°, 5°, 7° y 9°. Tomado de https://www.icfes.gov.co/evaluaciones-icfes/pruebas-saber-3-5-7-y-9/

Kinder, A., Briese, F. J., Jacobs, M., Dern, N., Glodny, N., Jacobs, S., & Leßmann, S. (2025). Effects of adaptive feedback generated by a large language model: A case study in teacher education. Computers and Education: Artificial Intelligence, 8, 100349. https://doi.org/10.1016/j.caeai.2024.100349

Lui AM., Andrade HL. (2022). The Next Black Box of Formative Assessment: A Model of the Internal Mechanisms of Feedback Processing. Front. Educ. 7:751548. doi: 10.3389/feduc.2022.751548

Shen, S., Liu, Q., Huang, Z., Zheng, Y., Yin, M., Wang, M., & Chen, E. (2024). A survey of knowledge tracing: Models, variants, and applications. IEEE Transactions on Learning Technologies, 17, 1858–1879. https://doi.org/10.1109/TLT.2024.3383325

ECONOMIA

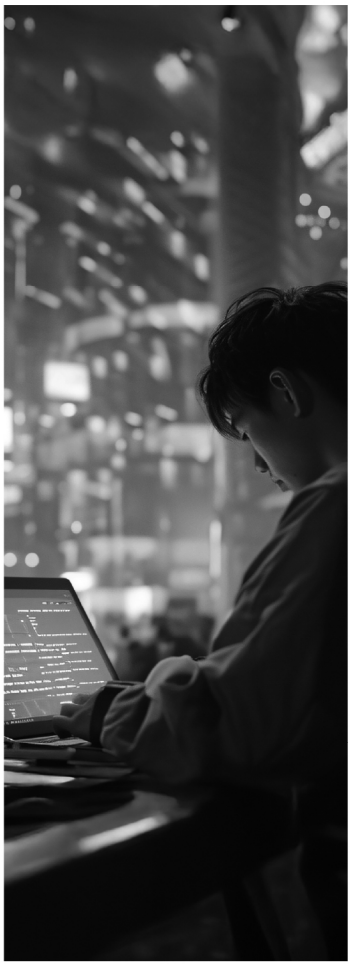
Laboratorio de inteligencia artificial aplicada a Economía

Universidad Externado de Colombia

FACULTAD DE ECONOMÍA

ECONOMÍA

Laboratorio de inteligencia artificial aplicada a Economía



Universidad Externado de Colombia
FACULTAD DE ECONOMÍA

ECONOMIA
Laboratorio de inteligencia artificial aplicada a Economía

Deserción estudiantil en el pregrado de Economía de la Universidad Externado de Colombia:

Un modelo de predicción individualizada de Machine Learning

Autores

Santiago Murcia Álzate santiago.murcia1@uexternado.edu.co
Daniel Fandiño Orjuela daniel.fandino@est.uexternado.edu.co
Santiago Andrés Rodríguez Estrada santiago.rodriguez4@est.uexternado.edu.co
Daniela Valentina Castro Mora daniela.castro4@uexternado.edu.co
Dayan Jasbleidy Pineda Ramírez dayan.pineda@uexternado.edu.co

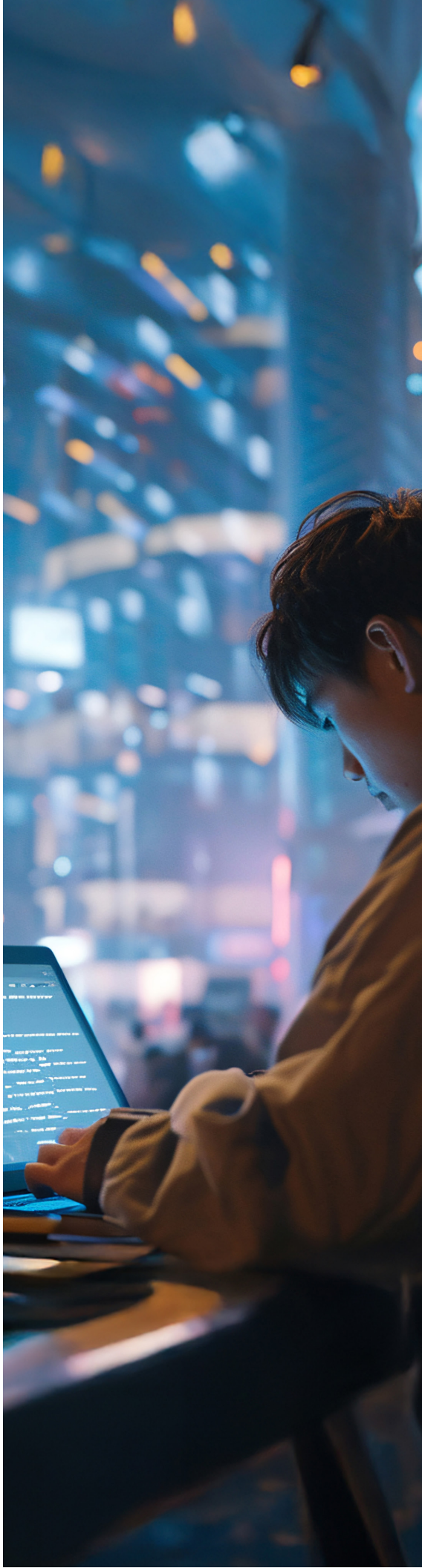
2025

Resumen

El fenómeno de la deserción estudiantil en la educación superior representa uno de los mayores retos para las universidades colombianas y del mundo, especialmente en programas como el pregrado en Economía de la Universidad Externado de Colombia, donde la tasa de deserción ha alcanzado el 29,54% desde 2021. Este artículo analiza el problema desde una perspectiva integral, identificando los factores académicos, socioeconómicos y logísticos que inciden en la decisión de abandonar los estudios, y enfatizando la importancia del bajo rendimiento en áreas clave y la distancia geográfica entre el estudiante y la institución.

Con el propósito de mejorar la identificación temprana de estudiantes en riesgo, se desarrollaron y compararon modelos de machine learning como redes neuronales, árboles de decisión y XGBoost aplicando técnicas avanzadas de análisis predictivo sobre datos históricos institucionales. Los resultados evidencian que la combinación de bajo rendimiento académico y mayor distancia a la universidad constituye el principal determinante de la deserción, y que los modelos utilizados permiten predecir con alta precisión los casos de riesgo, alcanzando un recall del 93% con redes neuronales y una precisión del 88% con XGBoost.

El estudio concluye que la implementación de sistemas predictivos robustos es una herramienta fundamental para diseñar estrategias de intervención focalizadas, facilitando acciones preventivas como tutorías, apoyos psicosociales y logísticos, y promoviendo así la permanencia estudiantil. Esta aproximación no solo contribuye a la gestión institucional, sino que fortalece el papel de la educación superior como motor de movilidad social y desarrollo en el país.



Introducción

Desde el año 2021 la tasa de deserción promedio por cohorte en el pregrado de Economía de la Universidad Externado de Colombia ha sido del 29,54 %, es decir, de los estudiantes que se han matriculado cada semestre para iniciar su proceso de formación 3 de cada 10 de estos estudiantes no logran graduarse. Esto tomando como referencia sólo las cohortes cobijadas bajo la última actualización del plan de estudios a partir del año 2021, puesto que si se toma como referencia los datos de deserción desde el año 2012 el panorama de deserción es mucho más preocupante con una tasa promedio del 37,8 % tal como se puede apreciar en la Figura 1. de movilidad social y desarrollo en el país.

El panorama que enfrenta el pregrado de Economía no es ajeno a la situación general del país en materia de deserción estudiantil en programas de educación superior. De hecho, de acuerdo con datos del Sistema para la Prevención y Análisis de la Deserción en las Instituciones de Educación Superior – SPADIES la tasa de deserción promedio por cohorte desde el año 2017 a nivel nacional es del 26,43 %. Inclusive, de acuerdo con cifras de la OCDE, a nivel internacional se estima que la tasa de deserción promedio es del 20 % en programas de educación superior (Lorenzo-Quiles et al, 2023).

Si bien, tanto en el pregrado en Economía como a nivel nacional, la tendencia de la tasa de

Figura 1. Tasa de deserción estudiantil – Pregrado en Economía Universidad Externado de Colombia

Nota: Datos extraídos del sistema SAREX de la Universidad. La línea de tendencia representa un ajuste lineal de la tasa de deserción entre 2012 y 2023.

deserción es a la baja no deja de ser un problema preocupante que 1 de cada 4 o 1 de cada 3 jóvenes que llegan a una institución de educación superior con un sueño, con un proyecto de vida y con la ilusión de llegar a ser profesionales, no logran graduarse. Esto sumado a la importancia que tiene la educación como motor de desarrollo económico y factor fundamental de movilidad social (Banco Mundial, 2019).

Para entender mejor este fenómeno se debe partir de una definición clara, el Ministerio de Educación Nacional ha definido la "Deserción" como: *"Estado de un estudiante que de manera voluntaria o forzosa no registra matrícula por dos o más períodos académicos consecutivos del programa en el que se matriculó; y no se encuentra como graduado, o retirado por motivos disciplinarios. La deserción es el resultado del efecto de diferentes factores como individuales, académicos, institucionales, y socioeconómicos".*

Partiendo de esta definición, que también es la base para calcular este indicador para el pregrado en Economía, es importante hacer una adecuada revisión de literatura para entender, al menos en los últimos 10 años, cómo se ha estudiado este problema, cuáles son sus principales causas, qué investigaciones se han hecho y qué se puede aprender de sus autores para analizar el caso que nos ocupa.

Teniendo en cuenta la importancia de la educación tanto para el desarrollo humano como para el desarrollo económico, la deserción estudiantil ha sido un fenómeno ampliamente investigado tanto desde un enfoque cuantitativo como cualitativo.

En general se reconoce que la deserción es un problema multifactorial con aristas Macro, Meso y Micro: En la parte macro se tiene en cuenta factores como políticas públicas educativas, acceso al crédito y posibilidades de financiación. Desde lo meso se estudian factores como la calidad de la enseñanza y demás cualidades institucionales, mientras que, desde la lógica micro, se contemplan factores asociados al estudiante como antecedentes académicos, su situación socioeconómica y motivaciones personales (Behr et al, 2020).

La literatura sugiere que las variables más relevantes a la hora de explicar la deserción estudiantil en educación superior se pueden clasificar en 3 categorías: (i) Académicas, (ii) Económicas y (iii) Sociales (Chalela-Naffah et al, 2020). Dentro de lo académico, se ha demostrado que un estudiante que presente un bajo rendimiento académico y/o que llega a la universidad con muy malas bases de su proceso de educación media tienen mayor probabilidad de suspender sus estudios ya sea por reprobar asignaturas, perder motivación o sentir que no cumplen las exigencias académicas.

Desde lo económico, cuando los estudiantes o sus familias presentan dificultades financieras para asumir los costos de la matrícula, manutención o materiales de estudio tiene una mayor probabilidad de desertar. En países en desarrollo, el estrés financiero es particularmente evidente. Un estudio sobre deserción en universidades colombianas encontró que el acceso a la financiación educativa suele limitarse a créditos más que a becas, lo que genera una presión económica sobre el estudiante y su familia (Valencia-Arias et al, 2023). Sin embargo, la pobreza o la falta de recursos no operan solas: suelen venir acompañadas de menos apoyo familiar o capital cultural, lo que agrava el riesgo de abandonar. No obstante, está claro que el alivio de las barreras económicas mediante becas, subsidios de matrícula, ayudas de transporte o alimentación tiende a mejorar la retención.

Finalmente, desde la perspectiva personal y social, el entorno familiar y el apoyo emocional resultan decisivos: estudiantes sin una red de apoyo (familia, amigos o mentores) que los motive a continuar sus estudios pueden sentirse aislados o abrumados, facilitando el abandono (Chalela-Naffah et al, 2020). Lo anterior sumado a una falta de sentido de pertenencia por la institución, dificultades sociales para crear lazos con su entorno y factores de salud mental cierran el conjunto de dificultades sociales que detonan, en parte, las decisiones de deserción.

El enfoque cuantitativo ha dominado el estudio de la deserción, pero la inteligencia artificial está redefiniendo el análisis.

Metodológicamente el problema de la deserción estudiantil se ha abordado tanto cualitativamente como cuantitativamente. Sin embargo, el enfoque cuantitativo ha sido el predominante. Por ejemplo, análisis multivariados como la regresión logística son muy comunes para estimar la probabilidad de deserción en función de variables académicas (promedios, créditos aprobados), económicas (ingresos, estrato socioeconómico) y demográficas (Núñez-Naranjo, 2024).

Estudios recientes han utilizado modelos de regresión logística multinivel y técnicas de *data mining* para lograr predicciones más precisas del abandono, alcanzando grados de predictibilidad del 75 % o más al clasificar estudiantes desertores y persistentes. Asimismo, es habitual el uso de análisis de supervivencia (Kaplan-Meier, modelos de riesgo proporcional) para modelar el tiempo hasta la deserción dentro de una cohorte, o algoritmos de aprendizaje automático (árboles de decisión, *random forest*, redes neuronales) especialmente en contextos de analítica de datos educativos (Núñez-Naranjo, 2024).

Frente a esto, surgen revisiones de metodologías como la de Sánchez et al. (2023), quienes señalaron la necesidad de ampliar la evaluación de variables e incluir otras que podrían influir en la deserción. Frente a esto, los modelos econométricos tradicionales enfrentan limitaciones, ya que muchos asumen relaciones lineales y pueden no capturar la complejidad ni las interacciones entre múltiples factores que influyen en la decisión de abandonar los estudios.

Para superar estas limitaciones, la incorporación de modelos de Machine Learning e Inteligencia Artificial (IA) permite analizar múltiples variables simultáneamente y detectar patrones complejos. Urbina-Nájera et al. (2021) identificaron seis factores clave en la deserción universitaria: asistencia, período, último semestre cursado, porcentaje de materias reprobadas, créditos cursados y programa académico. No obstante, su estudio demostró que las bases de datos utilizadas influyen significativamente en la determinación de las causas de deserción.

Dado que los modelos de Machine Learning (ML), deep learning y árboles de decisión pueden manejar estas relaciones complejas, su validación es un paso crítico para garantizar su rendimiento. Para ello, se utilizan diferentes técnicas como la validación cruzada y la división en conjuntos de entrenamiento; y la validación y prueba, evitando problemas como sobreajuste (overfitting) o subajuste (underfitting). Así mismo, la efectividad de estos modelos se mide mediante métricas como Accuracy, ROC-AUC, MSE y R^2, dependiendo del tipo de problema a resolver.

En este contexto, el uso de modelos de redes neuronales ha demostrado ser una alternativa eficiente. Cruz et al. (2022) y Pelima et al. (2024) analizaron diversos trabajos que emplean algoritmos como Random Forest, Support Vector Machines y redes neuronales, concluyendo que la precisión de predicción puede alcanzar hasta el 90 %. Además, destacaron el uso de sistemas de gestión del aprendizaje (LMS) para extraer datos en tiempo real, lo que permite identificar con mayor precisión a los estudiantes en riesgo de abandonar sus estudios.

Miranda et al. (2024) desarrolló un modelo predictivo basado en aprendizaje automático, señalando que la deserción estudiantil está influenciada por múltiples factores como el entorno familiar, la situación económica, el estado de bienestar, así como variables académicas, financieras, socioeconómicas, psicológicas y sociales. Esto resalta la necesidad de considerar un enfoque más integral en el análisis de la

deserción, donde los modelos de IA pueden capturar la complejidad del fenómeno y generar predicciones más precisas.

Metodología aplicada para la Facultad de Economía de la Universidad Externado de Colombia

En esta ocasión los esfuerzo de la Facultad de Economía, en línea con la tendencia investigativa actual, se han centrado en construir un modelo de machine learning con el propósito de predecir la probabilidad individualizada de deserción de cada estudiante del pregrado en Economía partiendo de variables micro del tipo académico y socioeconómico como:

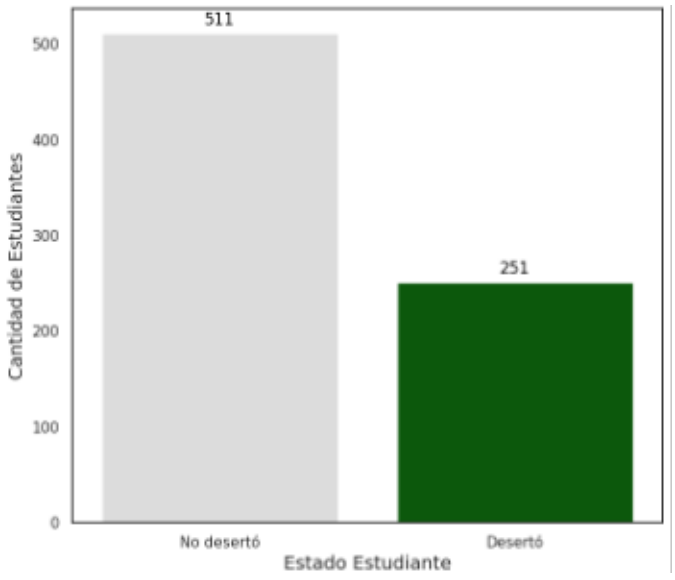
Estado Estudiante:

Esta será la variable explicativa, la cual es dicotómica y muestra si el estudiante desertó o no. La clasificación se da si el estudiante no se sigue registrando después de dos semestres, cuyas proporciones se muestran en la Figura 2. Allí se evidencia que la deserción es alta, ya que representa el 32,93 % de los registros, demostrando que es una problemática para la cual se deben implementar mecanismos urgentes para minimizar esta cifra.

Sexo:

El sexo del estudiante puede llegar a ser un indicador de la deserción, el cual presenta en la

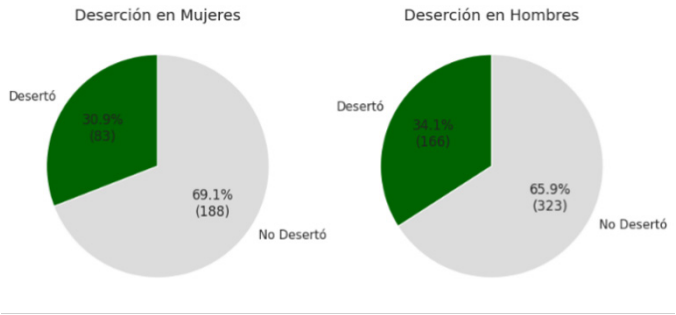
Figura 2. Balance de desertores y no desertores



Nota: Elaboración propia con base a registros estudiantiles de la Facultad de Economía, Universidad Externado de Colombia (2015 – 2024).

Figura 3 que los hombres tienen una mayor tasa de deserción respecto a las mujeres, con una diferencia de cuatro puntos porcentuales.

Figura 3. Comparación de deserción por sexo



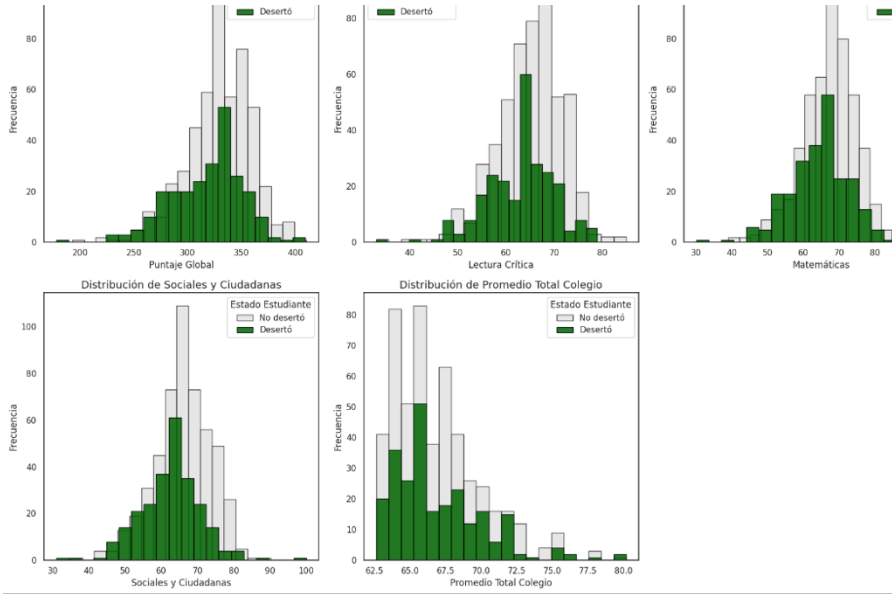
Nota: Elaboración propia con base a registros estudiantiles de la Facultad de Economía, Universidad Externado de Colombia (2015 – 2024).

Preparación previa:

Como acercamiento a la preparación que traían los estudiantes antes del ingreso a la Universidad, se tiene la prueba de estado ICFES, que busca medir la calidad educativa intermedia. Esta prueba evalúa áreas como lo son Lectura Crítica, Matemáticas y Competencias Sociales y Ciudadanas. La prueba da un puntaje global individual y el promedio por colegio.

Algo intuitivo y que se hace evidente en la Figura 4,

Figura 4. Comparación de resultados de prueba ICFES de desertores y no desertores



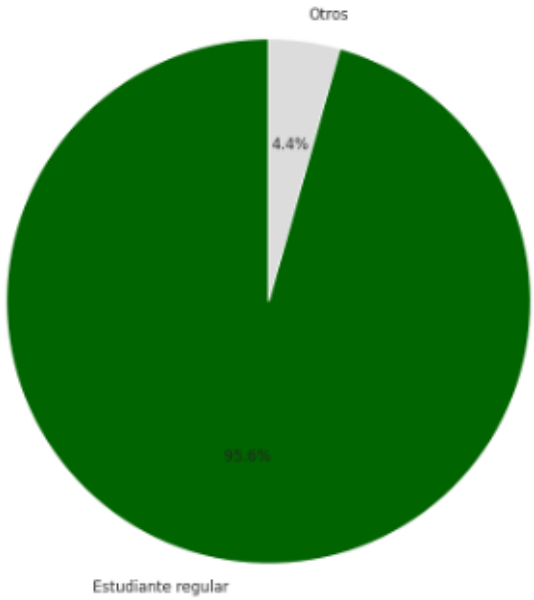
Nota: Elaboración propia con base a registros estudiantiles de la Facultad de Economía, Universidad Externado de Colombia (2015 – 2024).

es que los estudiantes que evidenciaron mayores puntajes en el ICFES (mayor preparación), fueron clasificados en no desertores, mientras que aquellos que han dejado la carrera, han sido aquellos estudiantes con bajo desempeño en la prueba.

Clase oyente:

Esta es una variable que se refiere al tipo de ingreso por el cual el estudiante se vinculó a la Universidad. Incluye diferentes categorías: el estudiante regular, que cursa su programa académico de manera convencional; el estudiante SPP, vinculado a un programa especial de apoyo académico; la comunidad indígena, que accede mediante cupos o programas específicos; la modalidad de inmersión, que incluye estudiantes en programas de intercambio o cursos intensivos; Jóvenes a la U, una iniciativa que facilita el acceso a educación superior para jóvenes en condiciones de vulnerabilidad; y convenios, que agrupa a estudiantes admitidos a través de acuerdos institucionales con otras entidades o programas (Figura 5).

Figura 5. Proporción de estudiantes regulares respecto a otros estudiantes en deserción



Nota: Elaboración propia con base a registros estudiantiles de la Facultad de Economía, Universidad Externado de Colombia (2015 – 2024).

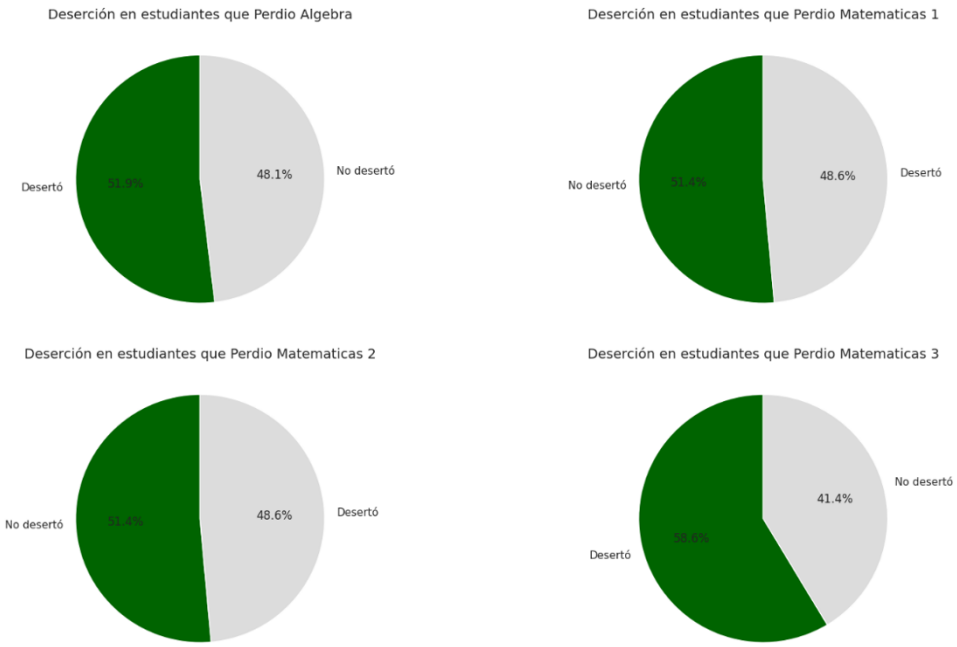
Variables del área de matemáticas:

Se consideró tener en cuenta la pérdida de las materias del área de matemáticas, ya que estas materias suelen ser prerrequisito de otras asignaturas y su reprobación puede generar atrasos en la carrera, lo que podría contribuir a la deserción. Además, la carrera de Economía exige un alto nivel de conocimientos y habilidades matemáticas, por lo que las dificultades en

esta área pueden ser un indicador de riesgo de abandono.

En la Figura 6 se muestran las proporciones de los estudiantes que perdieron alguna de las cuatro materias del área de matemáticas dentro del total de deserción. Se observa que más de la mitad de los estudiantes que pierden estas materias, terminan desertando, especialmente en el caso de Matemáticas 3, donde el 58,6 % de los estudiantes que reprobaron, terminaron desertando. Similar a estas variables, también se consideró el rendimiento presentado en otras asignaturas.

Figura 6. Proporción de la deserción por pérdida de materia del área de matemáticas

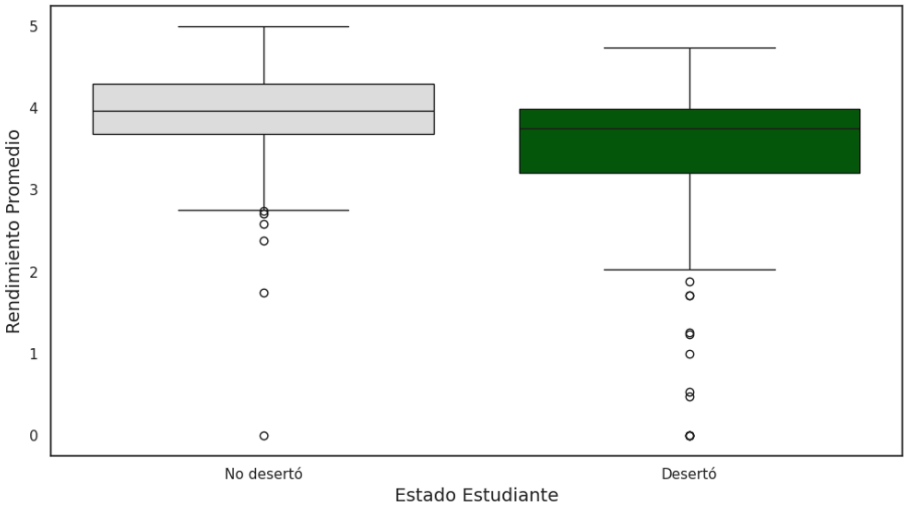


Nota: Elaboración propia con base a registros estudiantiles de la Facultad de Economía, Universidad Externado de Colombia (2015 – 2024).

Introducción a la Economía I:

Introducción a la Economía I: Otra materia clave para identificar posibles estudiantes desertores es la materia introductoria a la carrera. Si un estudiante la percibe como compleja, esto puede incentivarlo a buscar otras opciones académicas. En la Figura 7 se puede observar que los estudiantes que desertaron obtuvieron calificaciones menores en la asignatura, mientras que aquellos que continuaron en la carrera, han alcanzado promedios más altos. Este análisis se basó en el rendimiento académico, el cual refleja la nota final de aquellos que solo vieron la asignatura una sola vez o con el promedio de aquellos que perdieron y la volvieron a cursar.

Figura 7. Comparación de rendimiento en Introducción a la Economía I entre estudiantes desertores y no desertores.



Nota: Elaboración propia con base a registros estudiantiles de la Facultad de Economía, Universidad Externado de Colombia (2015 – 2024).

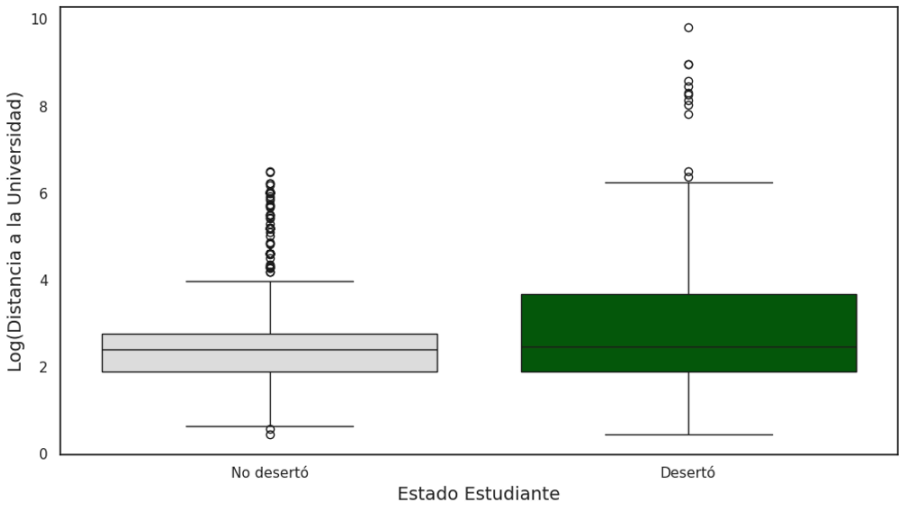
Distancia de la residencia a la Universidad (Km):

La distancia del lugar de residencia al colegio, universidad o trabajo ha sido un tema de debate y estudio a nivel mundial, en donde la evidencia ha mostrado que, a mayor distancia, menor es la productividad.

Respecto a esto, un estudio realizado en el Politécnico de Milán (2024) investigó el impacto del tiempo de desplazamiento en el rendimiento académico de estudiantes de primer año. Los resultados indicaron que un mayor tiempo de viaje se asociaba negativamente con el promedio de calificaciones, sugiriendo que desplazamientos prolongados pueden afectar el desempeño académico.

Por esta razón, se decidió considerar esta variable, la cual se calculó por medio de la geolocalización de la residencia registrada por los estudiantes y calcula la distancia hasta la ubicación de la Universidad. De hecho, en la Figura 8 se muestra que los estudiantes desertores presentan valores y una media más alta en términos de distancia que los estudiantes que no han desertado, permitiendo observar que la variable de distancia puede explicar la deserción.

Figura 8. Comparación de la distancia de estudiantes desertores y no desertores en Kilómetros (logaritmo)

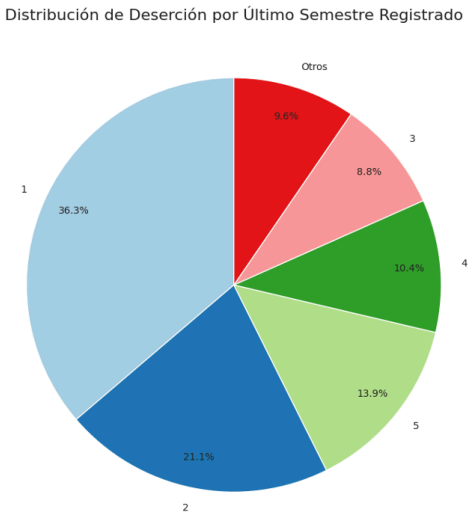


Nota: Elaboración propia con base a registros estudiantiles de la Facultad de Economía, Universidad Externado de Colombia (2015 – 2024).

Deserción por semestres:

En análisis también se observó en cuáles semestres son los que más se presenta deserción, lo que intuitivamente se puede determinar que es más alta la deserción al inicio de carrera que cuando se está concluyendo, lo que se refleja en la Figura 9. Esta variable refleja el último semestre registrado en la base de cada estudiante en el momento de corte cuando se extrajeron los datos.

Figura 9. Deserción por semestre



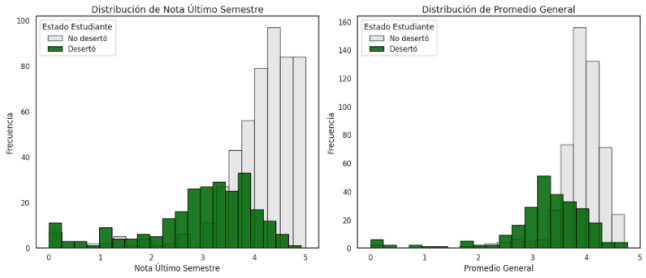
Nota: Elaboración propia con base a registros estudiantiles de la Facultad de Economía, Universidad Externado de Colombia (2015 – 2024).

Último promedio y acumulado:

Uno de los mayores determinantes en la deserción es el rendimiento del estudiante en la carrera, lo cual se ve reflejado en la calificación. Por ello, se consideraron las notas promedio del último semestre registrado, así como el acumulado que puede reflejar el historial del estudiante a nivel académico en la Universidad. La Figura 10 refleja lo que intuitivamente se puede concluir: los estudiantes con menor desempeño académico tienen mayores tasas de deserción.

Otras variables que se consideraron fueron: el Índice de Pobreza Multidimensional (IPM), el Grupo Étnico, la Edad de Inicio de la Universidad y si el estudiante cursó los diferentes niveles de inglés requeridos para graduarse.

Figura 10. Deserción por rendimiento académico (promedio último semestre y acumulado



Nota: Elaboración propia con base a registros estudiantiles de la Facultad de Economía, Universidad Externado de Colombia (2015 – 2024).

Modelo de Redes Neuronales

En el actual paradigma digital, el *Deep Learning* ha revolucionado campos como la visión por computadora, el procesamiento de lenguaje natural, la robótica y muchos otros, gracias a su capacidad para manejar grandes volúmenes de datos y extraer patrones que serían difíciles de identificar con métodos tradicionales. A diferencia de los enfoques clásicos de *Machine Learning*, que requieren un diseño manual de características (*feature engineering*), el *Deep Learning* permite que el modelo aprenda automáticamente las representaciones relevantes a partir de los datos crudos.

En este contexto, estas herramientas se han utilizado en la creación de inteligencias artificiales y modelos económicos, abordando problemas en los que los enfoques tradicionales suelen obtener bajas precisiones o presentar ineficiencias. Por ello, se ha decidido analizar el problema de la deserción estudiantil en la Facultad de Economía de la Universidad Externado de Colombia mediante redes neuronales, un tipo de modelo representativo del *Deep Learning*, utilizado en aplicaciones complejas como *ChatGPT* y *DeepSeek*.

Las redes neuronales artificiales buscan imitar el comportamiento del cerebro, ajustando los pesos de las conexiones entre neuronas durante el proceso de entrenamiento. En este proceso, las variables de entrada pasan a través de varias capas ocultas, cuya cantidad y estructura se definen mediante hiperparámetros antes de la implementación del modelo, hasta llegar a la capa de salida. Allí, la salida predicha se compara con las salidas observadas utilizando una función de pérdida, que mide el error. A través del proceso de *Backpropagation* (retropropagación del error), el modelo ajusta los pesos de las conexiones con el objetivo de minimizar el error. Este ajuste se repite a lo largo de múltiples iteraciones llamadas épocas, lo que permite mejorar la precisión del modelo. Una vez finalizado el entrenamiento, el modelo genera predicciones basadas en la

función de activación especificada en la capa de salida, de acuerdo con la naturaleza del problema abordado.

Previo a la aplicación del modelo, se realizó un rebalanceo de los datos, ya que, como se observa en la Figura 2, la clase mayoritaria (No desertó) duplicaba a la clase minoritaria (Desertó). Esto limitaba el aprendizaje del modelo y afectaba su desempeño. Para corregir este desbalance, se utilizó la técnica *SMOTE*, que genera registros sintéticos para la clase minoritaria basándose en los datos existentes. En este caso, se aplicó un rebalanceo para que la clase minoritaria alcanzara el 75% del tamaño de la clase mayoritaria, manteniendo la coherencia con la tendencia observada de que la deserción es menor que la no deserción

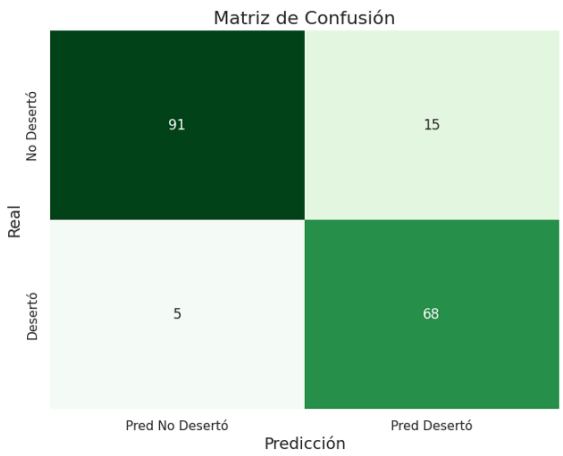
Una vez rebalanceada las clases de la variable y 'Estado Estudiante', se usó una red neuronal que cuenta con dos capas, cada una con 32 y 16 neuronas, respectivamente. A su vez se usa una función de activación de *ReLu*, que permite modelar relaciones lineales y no lineales, con una función de optimización llamada *Adam* para ajustar los pesos. A su vez se usó un *dropout*, lo cual es una técnica de regularización con la que se busca reducir el sobreajuste del modelo mediante el apagado aleatorio de determinado número de neuronas y así evitar la dependencia excesiva de determinadas conexiones, que para este caso fue del 50%. Finalmente, se usó una función de salida sigmoidal, la cual sirve para problemas de clasificación binomial, y su función de pérdida que corresponde a *Binary Crossentropy* (la cual es la más adecuada para problemas de clasificación binomial).

Este modelo se centró en optimizar la métrica *Recall*, que sirve para aquellos problemas de clasificación en donde los falsos positivos no son tan costosos como los falsos negativos, por lo que se busca minimizar esto último mediante la siguiente ecuación.

$$Recall = \frac{Verdaderos\ Positivos}{Verdaderos\ Positivos + Falsos\ Negativos}$$

De lo cual, se obtuvo la matriz de confusión de la Figura 11, que muestra el número de clasificaciones de acuerdo con lo observado y lo predicho.

Figura 11. Matriz de confusión para clasificaciones de deserción y no deserción.

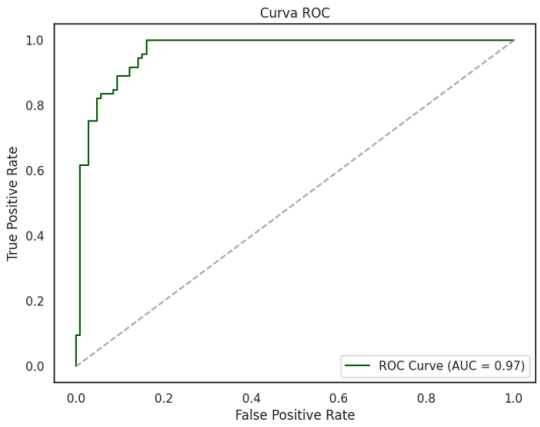


Fuente: elaboración propia

De este modelo se obtuvo un *Recall* que, luego de optimizar con un umbral de decisión de 0,4 (lo que significa que la clasificación de 1 o positivo se va a dar después de este valor), dio un resultado de 0,93, lo que significa que el modelo está clasificando correctamente el 93 % de los estudiantes que realmente desertan.

La razón para escoger un umbral de 0,4 se justifica por medio del Área Bajo la Curva (AUC, por sus siglas en inglés) de la curva ROC, que compara los verdaderos y falsos positivos, con lo que se evalúa la capacidad de discriminación del modelo. Visualmente, entre mayor sea el AUC, es decir, entre más se acerque a la esquina superior izquierda, el modelo clasifica de mejor manera las clases, mientras que, si el AUC es pequeño o se acerca mucho a la línea diagonal, el modelo prácticamente predice las clases aleatoriamente. Para este caso, con un umbral de decisión de 0.4 (con el fin de aumentar el *Recall*, pero que no todo se clasifique como deserción) se tiene un AUC amplio con un coeficiente de 0,97, como se muestra en la Figura 12. Esto último se interpreta de manera que el modelo clasifica correctamente el 97 % de los casos entre 0 y 1.

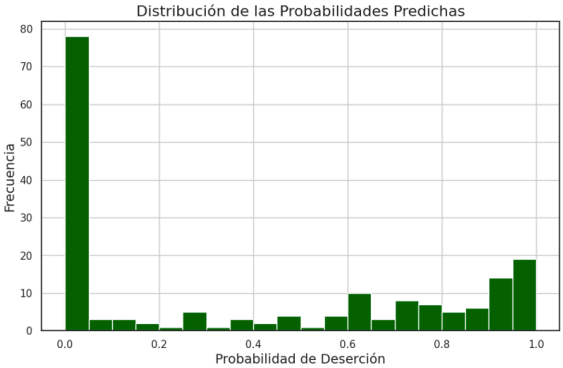
Figura 12. *AUC de la curva ROC*



Fuente: elaboración propia

Por su parte, esta clasificación se hizo teniendo en cuenta que 'Desertan' o 'No desertan', por lo que, con el fin de observar las probabilidades de manera continua, en la Figura 13 se observa en el histograma realizado que en la muestra las probabilidades de desertar se reparten de manera que en su mayoría tienen bajas probabilidades de desertar.

Figura 13. Distribución de probabilidades de deserción



Fuente: elaboración propia

Modelo de Árbol de Decisión

El modelo de Árbol de Decisión es una técnica de aprendizaje supervisado basada en la segmentación recursiva del espacio de características para generar reglas de decisión (Breiman et al., 1984). Se eligió este modelo debido a su facilidad de interpretación y su capacidad para manejar relaciones no lineales en los datos.

En el contexto del problema de deserción estudiantil, esta elección se debe a la necesidad de interpretar de manera clara y concisa las razones detrás de la clasificación de los estudiantes. Un árbol de decisión con una profundidad moderada permite visualizar patrones y factores clave que influyen en la deserción, lo que facilita la toma de decisiones por parte de las instituciones educativas. Sin embargo, debido a que este modelo puede no capturar completamente relaciones complejas en los datos, se complementará con un modelo más avanzado como *XGBoost* para mejorar la precisión y sensibilidad en la detección de estudiantes en riesgo.

El modelo de Árbol de Decisión se configuró con una profundidad máxima de 4 (*max_depth=4*) y una semilla aleatoria (*random_state=42*) para garantizar la reproducibilidad de los resultados. La elección de una profundidad limitada busca evitar el sobreajuste, ya que un árbol muy profundo podría memorizar los datos de entrenamiento en lugar de generalizar correctamente a nuevos casos. Además, al entrenar el modelo con *model_tree.fit(X_train, y_train)*, se permite que el algoritmo divida los datos en función de la mejor ganancia de información en cada nodo, generando una estructura jerárquica de reglas de decisión.

Figura 14. Entrenamiento de Árbol de Decisión

```
from sklearn.tree import DecisionTreeClassifier

# Creación y ajuste del modelo de Árbol de Decisión
model_tree = DecisionTreeClassifier(max_depth=4, random_state=42
model_tree.fit(X_train, y_train)

# Predicción sobre el conjunto de prueba
y_pred_tree = model_tree.predict(X_test)

# Evaluación del modelo
print(classification_report(y_test, y_pred_tree))
print("Precisión:", accuracy_score(y_test, y_pred_tree))
```

Fuente: elaboración propia.

Los resultados obtenidos para el modelo de Árbol de Decisión indican una precisión general del **80%**, lo que muestra un desempeño aceptable en la clasificación de los estudiantes en riesgo de deserción. A continuación, se presenta el reporte de clasificación detallado (Figura 15):

Figura 15. Resultados de Árbol de Decisión

```
=== Evaluación de Árbol de Decisión ===
Precisión: 0.80

Reporte de clasificación:
              precision    recall  f1-score   support

           0       0.83      0.89      0.86       101
           1       0.75      0.63      0.69        52

    accuracy                           0.80       153
   macro avg       0.79      0.76      0.77       153
weighted avg       0.80      0.80      0.80       153
```

Fuente: elaboración propia.

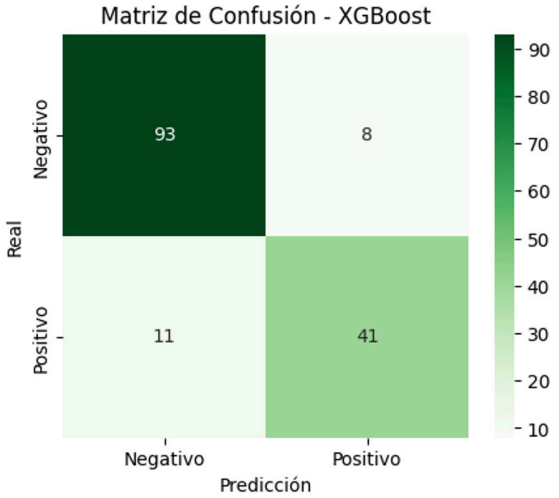
Evaluación del XGBoost

Dado que el modelo de Árbol de Decisión mostró limitaciones en la capacidad de identificación de estudiantes en riesgo de deserción, se optó por implementar *XGBoost* (*eXtreme Gradient Boosting*). Este algoritmo de aprendizaje supervisado se basa en el ensamblado de múltiples árboles de decisión débiles mediante *boosting*, lo que le permite mejorar la capacidad predictiva y reducir errores de clasificación (Chen & Guestrin, 2016).

XGBoost fue elegido debido a su capacidad para manejar conjuntos de datos desbalanceados, capturar relaciones no lineales entre variables y optimizar el desempeño en tareas de clasificación a través de técnicas avanzadas como la regularización y la minimización del error en cada iteración. Su eficiencia computacional y precisión en comparación con otros métodos tradicionales hacen que sea una opción adecuada para abordar el problema de predicción de la deserción estudiantil.

XGBoost mostró una mejora significativa en la precisión del modelo, se realizó un análisis detallado de métricas como la matriz de confusión, la tasa de falsos positivos y falsos negativos, y el equilibrio entre precisión y *recall* (Figura 16). Esto permitió evaluar con mayor profundidad la capacidad del modelo para identificar correctamente a los estudiantes en riesgo y minimizar los errores de clasificación.

Figura 16. Matriz de confusión para clasificaciones de deserción y no deserción

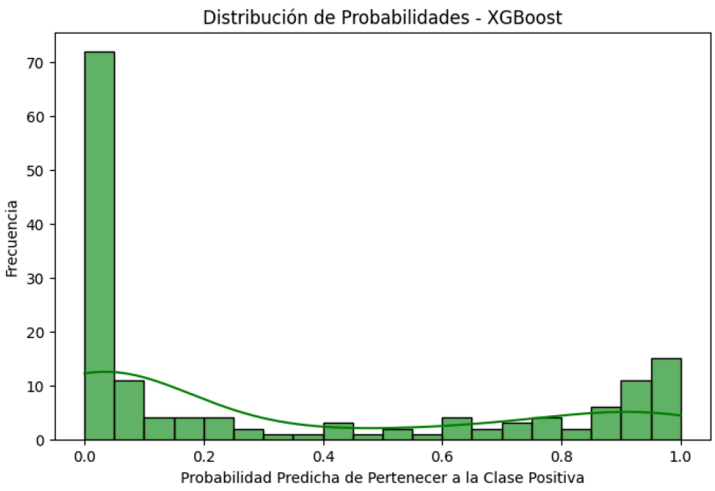


El modelo *XGBoost* demostró un excelente desempeño en la clasificación, logrando una precisión del 88%, lo que indica una alta capacidad para identificar correctamente las clases en los datos de prueba. Su rendimiento se refleja en la matriz de confusión, donde se observa que el modelo clasificó de manera acertada 93 casos negativos y 41 casos positivos, lo que evidencia su fiabilidad y robustez en la identificación de patrones.

Además, el modelo presentó un *f1-score* alto en ambas clases, con valores de 91% para la clase negativa y 81% para la clase positiva, lo que confirma su equilibrio entre precisión y *recall*. Si bien se registraron 8 falsos positivos y 11 falsos negativos, la cantidad de aciertos sigue siendo significativamente mayor, lo que resalta la efectividad del modelo en la toma de decisiones.

Otro aspecto clave es que el modelo tiene un buen desempeño en la detección de la clase mayoritaria y logra una generalización efectiva sobre los datos de prueba, minimizando los errores de predicción. Además, su robustez frente a datos complejos demuestra que es una herramienta potente para la clasificación en este contexto.

Figura 17. Distribución de probabilidades de deserción.



El gráfico de distribución de probabilidades predichas por el modelo *XGBoost* revela un comportamiento altamente diferenciador entre las clases. Se observa una marcada concentración de probabilidades cerca de 0,0, lo que indica que el modelo clasifica con alta confianza muchas instancias como pertenecientes a la clase negativa.

Así mismo, aunque en menor medida, también hay una agrupación de probabilidades cerca de 1,0, lo que sugiere que el modelo identifica correctamente varios casos de la clase positiva.

La presencia de una curva de densidad bien definida respalda la capacidad del modelo para asignar probabilidades de manera estructurada, evitando distribuciones excesivamente dispersas o uniformes. Este comportamiento es indicativo de un modelo bien ajustado, capaz de distinguir entre ambas clases sin una tendencia significativa a la incertidumbre. Estos resultados refuerzan la solidez de *XGBoost* como una herramienta eficaz para la clasificación en este contexto, demostrando su capacidad para capturar patrones subyacentes en los datos con un alto nivel de confianza.

Análisis de Probabilidad de Deserción Estudiantil en XGBoost

A continuación, se presenta una tabla que estima la probabilidad de deserción para cada uno de los 52 estudiantes de primer semestre que llegan a la universidad. Esta estimación se ha realizado a partir del análisis de variables clave que históricamente han demostrado influir en la continuidad académica, concretamente el "Promedio Total Colegio" y la "Distancia a Universidad (km)". Combinando estos indicadores, se asignó heurísticamente un porcentaje de riesgo de deserción a cada observación, lo que nos permite identificar de manera preliminar a aquellos estudiantes que podrían necesitar un mayor apoyo o intervención temprana. Este análisis no solo busca cuantificar el riesgo, sino también sentar las bases para el desarrollo de modelos predictivos más robustos y personalizados en futuras evaluaciones. Aquí se evidencia cómo factores académicos y logísticos se integran para ofrecer una visión completa del perfil de riesgo de los estudiantes recién ingresados (Tabla 1).

Tabla 1 - Probabilidades de deserción para primer semestre

Nº de Estudiante	Prob. de Deserción (%)	Justificación
41	70%	Muy bajo rendimiento y máxima distancia: situación de riesgo altísimo.
6	68%	Rendimiento muy bajo y alta distancia: situación de alto riesgo.
46	67%	Rendimiento muy deficiente y alta distancia; riesgo muy elevado.
1	65%	Rendimiento excelente; casi sin fricción, pero gran distancia.
30	65%	Rendimiento deficiente junto a máxima distancia, riesgo muy alto.
51	64%	Muy bajo rendimiento y alta distancia, indicando un riesgo muy elevado.

Nº de Estudiante	Prob. de Deserción (%)	Justificación
10	63%	Rendimiento deficiente y gran distancia se reflejan en un alto riesgo.
35	62%	Muy bajo rendimiento y alta distancia se asocian a un riesgo muy alto.
29	60%	Bajo rendimiento y alta distancia se reflejan en un riesgo elevado.
16	55%	Rendimiento deficiente y distancia algo elevada incrementan el riesgo.
45	55%	Rendimiento bajo y distancia media-alta hacen que el riesgo sea considerable.
7	54%	Promedio preocupante junto a una distancia extensa; riesgo elevado.
40	52%	Rendimiento bajo y alta distancia; riesgo notablemente elevado.
15	50%	Promedio bajo y distancia moderada; combinación de factores de riesgo.
50	50%	Promedio bajo y considerable distancia combinan para un riesgo medio.
9	49%	Promedio justo y distancia moderada; riesgo casi a la mitad.
20	48%	Rendimiento justo y gran distancia; riesgo cercano al 50 %.
34	46%	Rendimiento bajo y distancia considerable; riesgo elevado.
4	45%	Bajo rendimiento y amplia distancia combinan para un riesgo notable.
25	44%	Rendimiento bajo y distancia considerable dan un riesgo notable.
19	38%	Promedio regular y distancia considerable; riesgo moderado.
8	37%	Rendimiento sólido, pero una distancia media genera cierta dificultad.
39	37%	Promedio moderado junto a una distancia considerable elevan el riesgo intermedio.
28	36%	Rendimiento regular y distancia media indican riesgo moderado.
14	35%	Rendimiento justo con distancia media; riesgo intermedio.
49	35%	Rendimiento regular y distancia moderada: riesgo intermedio.

Nº de Estudiante	Prob. de Deserción (%)	Justificación
33	34%	Promedio justo y distancia media; riesgo intermedio.
24	33%	Rendimiento moderado y distancia razonable generan riesgo intermedio.
3	32%	Promedio regular y distancia considerable elevan el riesgo moderado.
27	30%	Rendimiento aceptable, pero la distancia sustancial eleva el riesgo.
44	30%	Promedio justo y distancia intermedia; riesgo moderado-bajo.
2	28%	Rendimiento aceptable, pero la distancia moderada añade cierto obstáculo.
48	28%	Rendimiento aceptable y distancia moderada; riesgo leve a moderado.
32	26%	Rendimiento aceptable y distancia razonable generan un riesgo moderado bajo.
13	25%	Rendimiento aceptable y poca distancia, lo que da un riesgo moderado bajo.
21	24%	Rendimiento aceptable y buena proximidad; riesgo moderado bajo.
38	23%	Buen rendimiento y baja distancia, lo que sugiere un riesgo moderado.
42	23%	Rendimiento aceptable y proximidad favorable; riesgo ligeramente bajo.
18	22%	Rendimiento sólido y poca distancia aportan estabilidad.
5	20%	Buen rendimiento y distancia razonable reducen el riesgo.
31	19%	Buen rendimiento y distancia moderada; riesgo relativamente bajo.
12	18%	Buen rendimiento y muy corta distancia; riesgo reducido.
52	18%	Rendimiento sobresaliente y mínima distancia; casi sin riesgo.
37	17%	Excelente rendimiento y corta distancia, manteniendo el riesgo muy bajo.
17	16%	Alto rendimiento y distancia corta; riesgo muy reducido.
43	16%	Buen rendimiento con corta distancia, lo que disminuye el riesgo.

Nº de Estudiante	Prob. de Deserción (%)	Justificación
26	15%	Alto rendimiento y corta distancia; muy pocos impedimentos.
22	14%	Rendimiento sobresaliente y proximidad total reducen el riesgo.
47	14%	Rendimiento sobresaliente y mínima distancia; casi sin riesgo.
23	13%	Rendimiento excepcional con mínima distancia; riesgo casi nulo.
11	12%	Rendimiento sobresaliente y distancia mínima; barreras casi inexistentes.
36	12%	Rendimiento sobresaliente y proximidad extrema; casi sin riesgo.

CONCLUSIONES

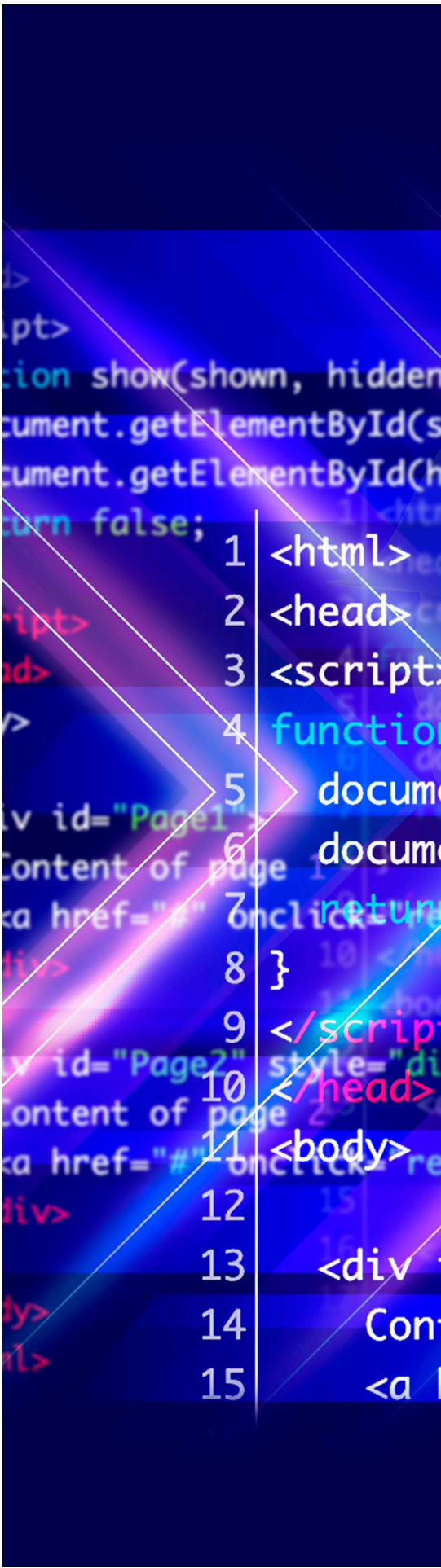
La tasa de deserción en el pregrado de Economía de la Universidad Externado (29,54% desde 2021) supera la media nacional e internacional, lo que refuerza la urgencia de aplicar modelos predictivos que permitan una intervención temprana. Diferentes estudios han demostrado que la deserción responde a factores multifactoriales que deben ser considerados simultáneamente, lo que justifica el uso de modelos más complejos y precisos como los de inteligencia artificial y *machine learning*.

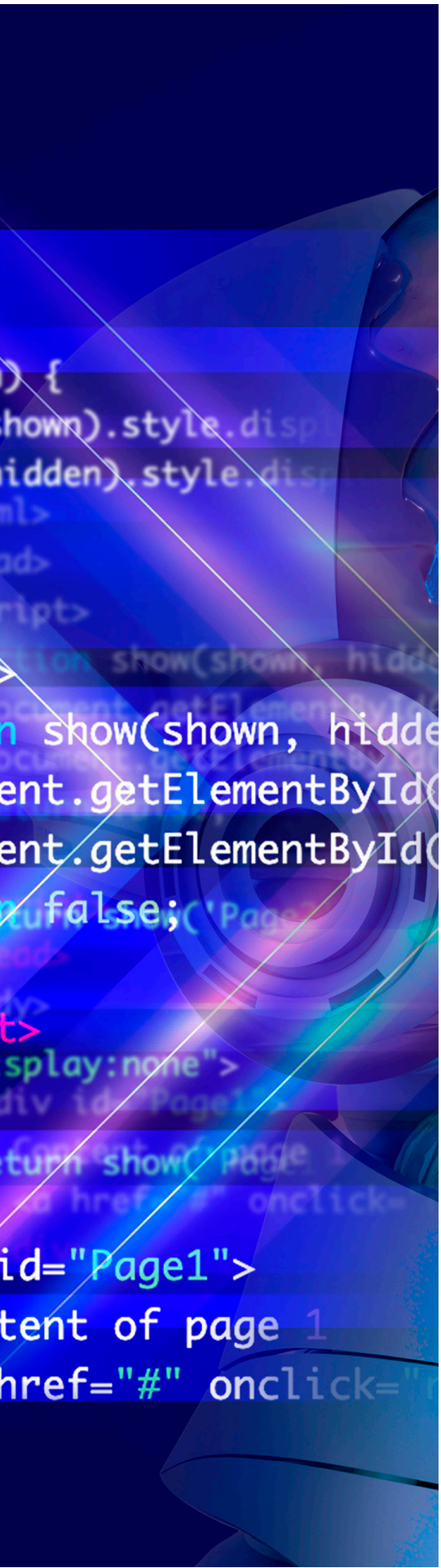
Por esta razón, medir con precisión la probabilidad de deserción permite transformar la educación en una verdadera herramienta de movilidad social, al evitar que los estudiantes más vulnerables abandonen sus estudios. En ese sentido, la literatura ha evidenciado que los modelos tradicionales de análisis no capturan completamente la interacción entre variables académicas, económicas y personales, lo que limita la eficacia de las estrategias de retención. En ese sentido, se buscó con este estudio comparar los resultados de tres modelos de aprendizaje automático: redes neuronales, árboles de decisión y *XGBoost*.

En estos modelos se utilizaron variables como el sexo del estudiante, el puntaje del ICFES, el tipo de ingreso, el rendimiento en asignaturas clave como Matemáticas e Introducción a la Economía y la distancia a la Universidad, las cuales resultaron ser variables significativas para predecir la deserción. Por tanto, su integración en modelos predictivos sigue siendo fundamental.

En este contexto, trabajar con bases de datos históricas de la Universidad, se pueden generar perfiles de riesgo desde el primer semestre, facilitando acciones de intervención preventiva por parte de la institución. Así mismo, comprender las características individuales de los estudiantes permite diseñar apoyos focalizados, como tutorías, becas o acompañamiento psicosocial para mejorar la retención.

Los resultados del modelo de redes neuronales desarrollado para la Universidad alcanzaron un *Recall* del 93%, lo que demuestra





su capacidad para identificar correctamente a la mayoría de los estudiantes en riesgo real de deserción. Se llegó a este resultado gracias a la utilización técnicas como *SMOTE* para corregir el desbalance en los datos. El modelo logró una mayor precisión sin sesgarse hacia la clase mayoritaria, lo cual es clave en contextos como la educación superior. Así mismo, métricas como *AUC* y la optimización del umbral de decisión permitieron ajustar el modelo para priorizar la minimización de falsos negativos, asegurando que los estudiantes en riesgo no pasen desapercibidos.

Por su parte, los resultados del modelo de árbol de decisión, aunque en este caso mostró una menor capacidad predictiva (80%), ofrece una lectura clara sobre qué variables inciden más en la deserción. Por otro lado, cuando se realizó la evaluación del modelo *XGBoost* aumentó la precisión a 88% y *f1-scores* altos, *XGBoost* se posiciona como una herramienta eficaz para la clasificación de estudiantes en riesgo de abandono académico, combinando interpretación y rendimiento. Según la matriz de confusión, se observa que el modelo clasificó de manera acertada 93 casos negativos y 41 casos positivos, lo que evidencia su fiabilidad y robustez en la identificación de patrones.

Basado en este modelo se hizo un análisis de probabilidades de deserción de los estudiantes de primer semestre y se encontró una constante clara: la combinación de bajo rendimiento académico y una alta distancia entre el lugar de residencia y la universidad es el patrón más recurrente entre los perfiles de alto riesgo. En todos los casos analizados, estas dos variables actúan de forma conjunta, elevando la probabilidad de abandono. Esta tendencia valida la capacidad del modelo para identificar factores críticos de manera precisa, confirmando que tanto el desempeño académico como las barreras logísticas deben ser atendidos desde una perspectiva institucional.
En particular, los estudiantes identificados con probabilidades de deserción iguales o superiores al 65% se encuentran en una situación de riesgo muy alto. Aquellos que presentan un rendimiento muy bajo o deficiente y, al mismo tiempo, residen

a gran distancia del campus universitario, enfrentan obstáculos tanto cognitivos como físicos para continuar sus estudios. Incluso el caso del estudiante 1, que presenta un rendimiento excelente, demuestra que la distancia por sí sola puede ser un factor crítico, al alcanzar también un 65% de probabilidad de deserción. Esto refleja que el modelo no depende exclusivamente del rendimiento académico, sino que evalúa de manera integral cómo interactúan las variables.

En niveles ligeramente inferiores, con probabilidades entre el 52% y el 64%, se agrupan estudiantes con características como rendimiento bajo o promedio preocupante y distancias grandes. Estos perfiles son igualmente relevantes para la acción institucional, ya que, si bien no están en el nivel más crítico, podrían cruzar el umbral de riesgo si no se toman medidas de apoyo. En ese sentido, la sensibilidad del modelo *XGBoost* para captar estos matices sugirió que es posible anticiparse a la deserción incluso antes de que se presenten signos evidentes, ofreciendo así una ventana de oportunidad para la intervención oportuna.

Basado en todos los resultados, se puede evidenciar cómo estos modelos pueden abrir una puerta a la intervención temprana y al diseño de estrategias que se pueden realizar al interior de la Facultad de Economía de la Universidad Externado para hacer una política de priorización y acompañamiento a aquellos estudiantes con mayor probabilidad de deserción.

Gracias a la distribución de probabilidades se pueden focalizar los programas según los determinantes de deserción y así diseñar diferentes intervenciones institucionales según la gravedad del caso. Algunos ejemplos son los programas de tutoría intensiva con Departamento de Matemáticas de la Universidad, focalización de monitores y acompañamiento personalizado, acompañamiento psicosocial con Bienestar Universitario y apoyo logístico, especialmente en temas de movilidad o alojamiento con aliados estratégicos que se puedan consolidar en la Universidad de la mano con el apoyo institucional de Bienestar.

Referencias

Banco Mundial. (2019). *Informe sobre el desarrollo mundial 2019: La naturaleza cambiante del trabajo*. Banco Mundial.

Behr, A., Giese, M., Teguim Kamdjou, H. D., & Theune, K. (2020). *Dropping out of university: a literature review*. Review of Education, 8(2), 614-652. https://doi.org/10.1002/rev3.3202

Burzacchi, A., Rossi, L., Agasisti, T., Paganoni, A. M., & Vantini, S. (2024). *Urban mobility and learning: Analyzing the influence of commuting time on students' GPA at Politecnico di Milano. Studies in Higher Education*, 1–26. https://doi.org/10.1080/03075079.2024.2374005

Chalela-Naffah, S., Valencia-Arias, A., Ruiz-Rojas, G. A., & Cadavid-Orrego, M. (2020). Factores psicosociales y familiares que influyen en la deserción en estudiantes universitarios en el contexto de los países en desarrollo. *Revista Lasallista de Investigación*, 17(1), 103-115. https://doi.org/10.22507/rli.v17n1a9

Chen, T., & Guestrin, C. (2016). XGBoost: A scalable tree boosting system. Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 785-794.

Hastie, T., Tibshirani, R., & Friedman, J. (2009). The elements of statistical learning: Data mining, inference, and prediction (2nd ed.). Springer. https://doi.org/10.1007/978-0-387-84858-7 https://doi.org/10.1145/2939672.2939785

Lorenzo-Quiles, O., Galdón-López, S., & Lendínez-Turón, A. (2023). *Factors contributing to university dropout: a review. Frontiers in Education*, 8, Article 1159864. https://doi.org/10.3389/feduc.2023.1159864

McKinney, W. (2017). Python for Data Analysis: Data wrangling with Pandas, NumPy, and Jupyter (2nd ed.). O'Reilly Media.

Núñez-Naranjo, A. F. (2024). Analysis of the determinant factors in university dropout: a case study of Ecuador. *Frontiers in Education*, 9, Article 1444534. https://doi.org/10.3389/feduc.2024.1444534

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., ... & Duchesnay, E. (2011). Scikit-learn: Machine learning in Python. Journal of Machine Learning Research, 12, 2825-2830. https://jmlr.org/papers/v12/pedregosa11a.html

Valencia-Arias, A., Chalela, S., Cadavid-Orrego, M., Gallegos, A., Benjumea-Arias, M., & Rodríguez-Salazar, D. Y. (2023). University Dropout Model for Developing Countries: A Colombian Context Approach. Behavioral Sciences, 13(5), 382. https://doi.org/10.3390/bs13050382

Universidad Externado de Colombia

FACULTAD DE ECONOMÍA

ECONOMIA

Laboratorio de inteligencia artificial aplicada a Economía

Universidad Externado de Colombia
FACULTAD DE ECONOMÍA

ECONOMIA
Laboratorio de inteligencia artificial aplicada a Economía

El sueño de un hogar propio:

Predicciones para la adquisición de vivienda en Bogotá

Autores

Laura Nathaly Camacho Cepeda laura.camacho05@est.uexternado.edu.co
Ricardo Andrés Sinning Sanabria ricardo.sinning@est.uexternado.edu.co
Oscar Fabian Rodríguez Sarmiento oscar.rodriguez08@est.uexternado.edu.co

2025

Resumen

El acceso a vivienda propia en Bogotá es un desafío para muchas familias debido a factores como los precios del mercado, la disponibilidad de financiamiento y las condiciones socioeconómicas de los hogares. Este estudio analiza los determinantes de la adquisición de vivienda a partir de un modelo de Bosque Aleatorio, utilizando datos de la Encuesta Multipropósito 2021 del DANE. Se identificaron las principales variables que influyen en la tenencia de vivienda propia completamente paga, en proceso de pago o en arriendo.

Los resultados muestran que el pago de arriendo o leasing, la razón por la cual no se adquiere vivienda, el ingreso per cápita y la informalidad laboral son los factores más relevantes en la decisión de compra de vivienda. Asimismo, el mapa de calor de los precios de vivienda por UPZ evidencia que los hogares con vivienda propia tienden a ubicarse en las periferias de la ciudad, donde los precios del suelo son más bajos, mientras que en zonas centrales y de mayor costo predominan los hogares en arriendo.

A pesar de la existencia de programas como Subsidio de Concurrencia y Generación FNA, que buscan facilitar el acceso a la vivienda, los hallazgos sugieren que es necesario fortalecer las políticas públicas enfocadas en la reducción del monto de la cuota de pago y en la estabilidad laboral, con el fin de mejorar la inclusión financiera y ampliar las oportunidades de acceso a vivienda propia.

Este estudio demuestra la utilidad de los modelos de Inteligencia Artificial en la predicción de la tenencia de vivienda, permitiendo una mejor comprensión de los factores que influyen en esta decisión y proporcionando herramientas para el diseño de estrategias que promuevan un acceso más equitativo a la vivienda en Bogotá.

Palabras clave: vivienda propia, arriendo, acceso a vivienda, Bosque Aleatorio, inteligencia artificial, subsidios de vivienda, inclusión financiera.

Introducción

La adquisición de vivienda propia es el sueño de muchas familias. Sin embargo, lograrlo se dificulta para la mayoría de los habitantes de las grandes ciudades. Por diversos factores, como los precios, la situación del mercado inmobiliario, las condiciones macroeconómicas y las condiciones internas de los hogares, estos se enfrentan a la disyuntiva de vivir en arriendo, muchas veces bajo condiciones precarias; o a comprar vivienda propia, asumiendo riesgos o dificultades en el pago de éstas. Así sucede en Bogotá, donde sus habitantes deben decidir si seguir pagando arriendo, o tratar de conseguir vivienda a bajo costo, pero en las periferias de la ciudad o entornos hostiles.

Muestra de estas dificultades es que, según Camacol, en el primer semestre de 2023 la comercialización de vivienda en Bogotá cayó un 49%, debido en gran parte por la coyuntura que el país atravesaba. No obstante, en las declaraciones no se ahonda en las condiciones socioeconómicas del grueso de hogares de la ciudad, lo que puede aportar a la discusión de los factores que inciden en la decisión de tener vivienda propia. Además, según el Dane (2022) en 2021, el 39,4% de los hogares del país vivía en una vivienda propia (totalmente pagada o en proceso de pago) y el 38,6% vivía en arriendo o subarriendo.

Es por eso que, en este artículo se pretende crear un programa que identifique si un hogar tendría vivienda propia o viviría en arriendo, dadas sus condiciones socioeconómicas y del entorno en el que se desenvuelven, incluyendo un análisis espacial.

En particular, la pregunta de investigación que se pretende responder es: ¿cuál sería el estatus de vivienda de una persona en Bogotá, entendido como tener una vivienda propia completamente paga, propia en proceso de pago o vivir en arriendo, teniendo en cuenta condiciones socioeconómicas y demográficas, e incluyendo factores espaciales, con datos de 2021?

Para el desarrollo del modelo se utilizó la Encuesta Multipropósito para Bogotá y Cundinamarca elaborada por el DANE en el año 2021 (EMB 2021).

Análisis y metodología IA

Se construyó una variable objetivo multidimensional, que muestra la cantidad de hogares en Bogotá que ocupan una vivienda:

1. Propia completamente pagada, 2. Propia en proceso de pago y, por último, 3. en arriendo/sin propiedad. Para ello se realizó una transformación a la variable NHCCP1 de la EMB 2021, que nos dice si la vivienda habitada por el hogar es:

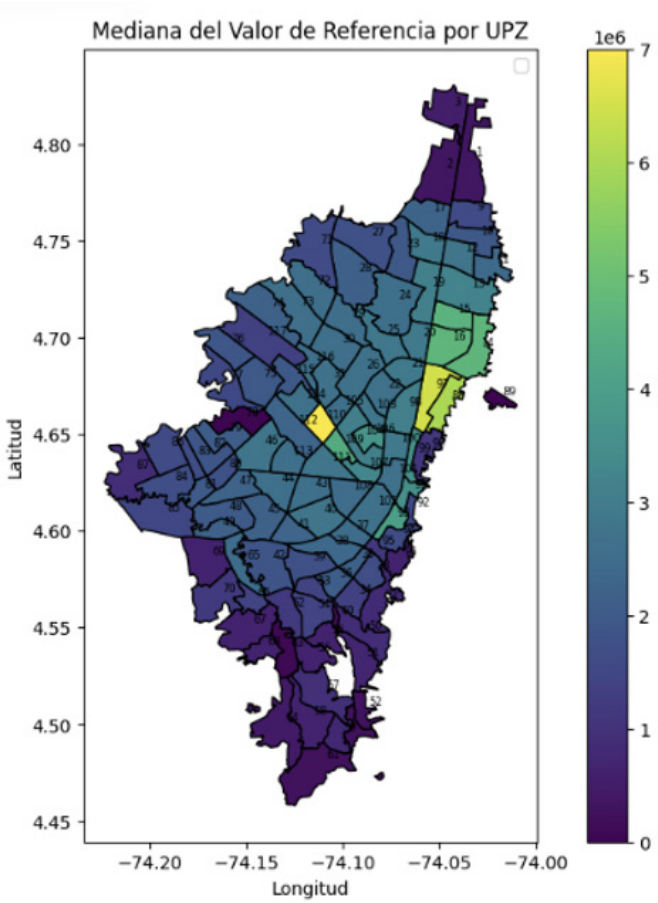
1. Propia, totalmente pagada. 2. Propia, la están pagando. 3. En arriendo, subarriendo. 4. Leasing 5. En usufructo. 6. Otra forma de tenencia (posesión sin título, ocupante de hecho, propiedad colectiva, entre otras).

De esta manera, en la categoría 1 se ubicaron los hogares que tienen vivienda totalmente pagada (1); en la categoría 2 están los hogares que son propietarios de la vivienda, pero aún la están pagando (2); y en la categoría 3 irán todos los hogares que viven en arriendo o no tienen propiedad sobre la vivienda (3, 4, 5, 6).

Las variables explicativas se escogieron y construyeron después de un proceso de revisión de literatura e intuición económica. Se enlistan a continuación: sexo, valor del arriendo de la vivienda, razón para no adquirir vivienda, ingreso por persona en el hogar, ingreso total del hogar, ocupación informal, la mediana del valor del suelo por UPZ, estrato, tiempo en minutos que le toma llegar a Transmilenio, factores de seguridad y salubridad alrededor de su lugar de vivienda, entre otras.

Es importante notar que para hacer una aproximación geoespacial se decidió crear una variable nueva para implementarla en el modelo, la cual es mediana del valor del suelo por UPZ en Bogotá (V_ref_UPZ). Para su construcción se tomaron datos de Mapas Bogotá de la mediana del valor de terreno por manzanas, y de Laboratorio Bogotá de las UPZ de la ciudad. De esta manera obtuvimos el mapa de calor del Gráfico 1, dónde un color más claro indica un valor del terreno más alto.

Gráfico 1. Mapa de calor del valor de la vivienda por UPZ

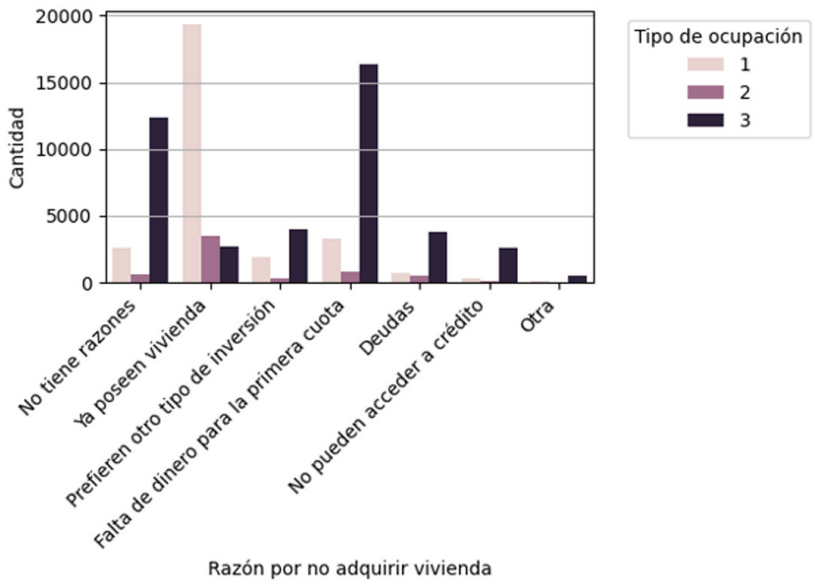


Fuente: Elaboración propia con datos de Mapas Bogotá y Laboratorio Bogotá

De este modo se puede evidenciar que, hacia las periferias de la ciudad, dado un valor más bajo por el suelo, es más probable tener vivienda propia, mientras que, hacia el oriente, en la zona de Chapinero se encuentran los valores más altos. Con esto los habitantes de la ciudad se enfrentan al trade-off de tener vivienda propia en lugares alejados de la ciudad o vivir en zonas céntricas, pero con una probabilidad mayor de vivir en arriendo, teniendo en cuenta únicamente la variable del valor de la vivienda en este caso, es decir, por el lado de la oferta.

Ahora bien, por el lado de la demanda se tienen entre otras variables, la razón por la cual las personas no adquirirían vivienda y el tipo de ocupación que tienen, siendo los tipos de ocupación 1, 2 y 3 la vivienda propia totalmente paga, vivienda propia en proceso de pago y en arriendo respectivamente. En el Gráfico 2 se observa que la mayoría de las personas que viven en arriendo no adquieren vivienda por falta de dinero para la primera cuota, aunque no es despreciable la cantidad de personas de este mismo tipo de ocupación que en realidad no tienen razones para no adquirir vivienda, mostrando así su preferencia y sus gustos personales. Los resultados para las personas que tienen vivienda propia paga son consistentes con la razón de "ya poseen vivienda".

Gráfico 2. Relación entre la razón para no adquirir vivienda y tipo de ocupación



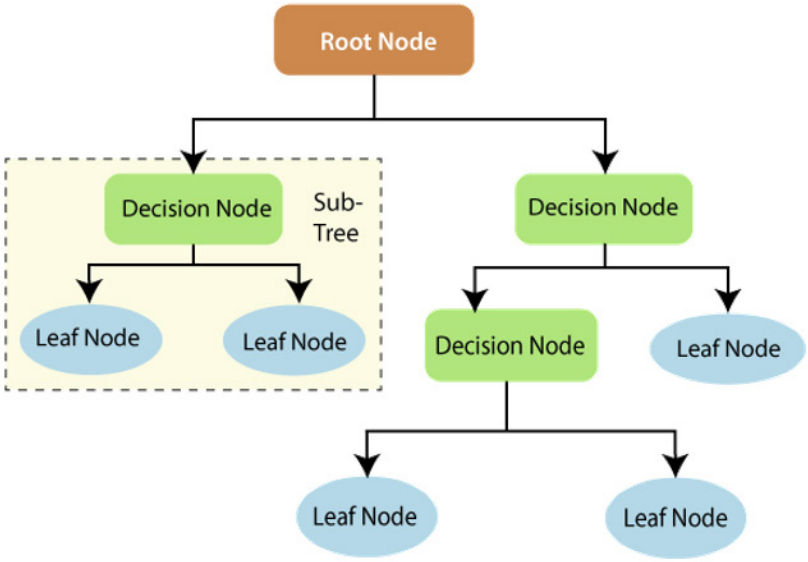
Fuente: Elaboración propia

Metodología

Con esta información se creó un modelo de 'Bosque Aleatorio' el cual se entrenó usando el 80% de los datos con el fin de probar su precisión en el conjunto restante de los datos. El modelo funciona dividiendo los datos en conjuntos aleatorios los cuales se les asignan también aleatoriamente características, cada uno de estos conjuntos se denomina un árbol de decisión.

Cada árbol genera su propia predicción a la disyuntiva entre tener casa propia y vivir en arriendo, luego en su conjunto, el bosque aleatorio determina cuales fueron los árboles que dado su conjunto de variables obtuvieron una mayor precisión en la predicción, una de las ventajas de este modelo es esta variedad entre los árboles lo cual robustece el modelo evitando el riesgo de sobre-ajustarse a los datos disponibles. La forma en la que funciona este modelo se puede entender mejor con la Figura 1.

Figura 1. Funcionamiento del modelo Random Fores



Fuente: Dynamic Ticket Price Prediction Models for Spanish Renfe Railways: A Comparative Analysis of Machine Learning Algorithms (2024).

Aunque los modelos en Machine-Learning son difícilmente interpretables, una ventaja de los bosques aleatorios es poder visualizar cuales variables el modelo encontró más importantes para la predicción y así contribuir a la generación de conclusiones o incluso nuevos modelos.

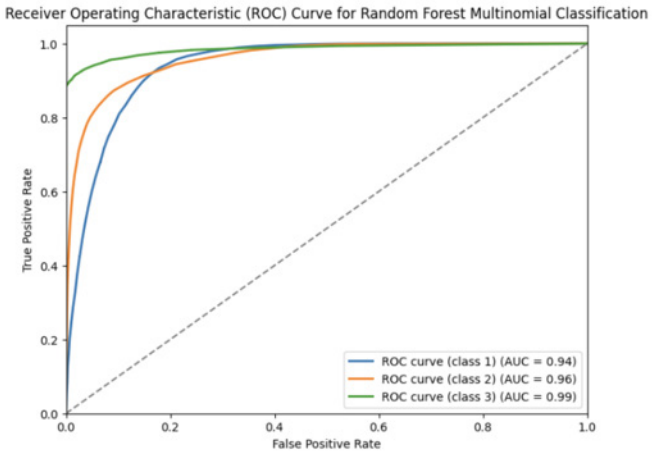
Resultados

Para hablar de los resultados del modelo, es primero necesario hablar de la precisión del modelo. El modelo tiene una precisión del 87%, lo que quiere decir que del 100% de las predicciones que realiza el modelo, 87% son verdaderas. De igual manera, se tiene un puntaje F1 (que combina la precisión y la sensibilidad del modelo) de un 87%. La predicción general del modelo es que un hogar en Bogotá tendría una vivienda propia en proceso de pago.

Por otro lado, se tiene la curva ROC (Receiver Operating Characteristic), en el que el eje X (False Positive Rate - FPR) representa la tasa de falsos positivos, es decir, el porcentaje de veces que el modelo predice incorrectamente una clase positiva cuando en realidad es negativa mientras que el eje Y (True Positive Rate - TPR) representa la tasa de verdaderos positivos, que es el porcentaje de veces que el modelo predice correctamente una clase positiva.

Como se observa en el Gráfico 3, donde se evalúa el rendimiento del modelo para clasificar tres clases, cada curva representa un tipo de ocupación y muestra cómo el modelo distingue entre las clases positivas y negativas, con los valores de AUC (Área bajo la curva) indicando la calidad del modelo. Los AUC para las tres clases son muy altos: 0.94 para la clase 1 (hogares con vivienda propia paga), 0.96 para la clase 2 (hogares con vivienda propia en proceso de pago) y 0.99 para la clase 3 (hogares que viven en arriendo), lo que refleja un excelente desempeño del modelo, especialmente en la clase 3.

Gráfico 3. Curva ROC

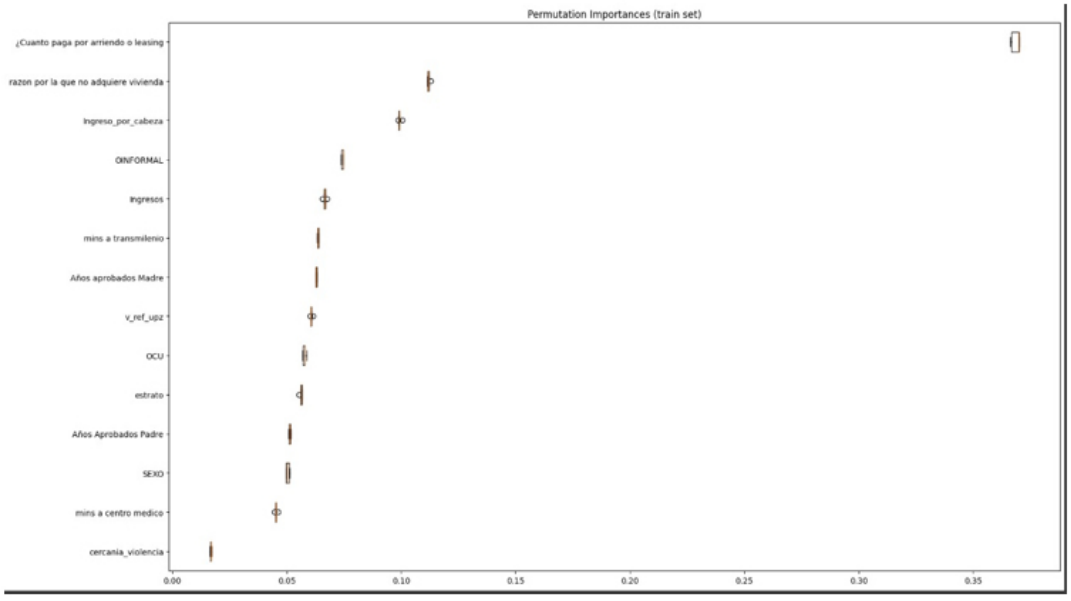


Fuente: Elaboración propia

Ya que se pudo comprobar que el modelo tuvo un buen rendimiento, se puede observar la importancia que el modelo le da a las características para asignar un tipo de ocupación o clase a cada hogar.

Ahora, en el Gráfico 4, se representa la importancia de las variables por permutación en el conjunto de entrenamiento. Esta es una técnica utilizada para evaluar qué tan influyente es cada variable en la predicción del modelo.

Gráfico 4: Importancia de las variables por permutación en el conjuto de entrenamiento



Fuente: Elaboración propia

Para interpretar la gráfica, se observan las variables que consideramos podrían ser relevantes para definir los resultados del modelo (se usaron más, pero no se encontró relevante mostradas). En el eje X se observa la importancia de cada variable medida a través de la pérdida en el desempeño del modelo cuando esa variable es permutada aleatoriamente. Es decir, cuanto mayor es el valor, más relevante es la variable para la predicción.

De esta manera, tenemos que existen varias variables importantes, por ejemplo, La variable "¿Cuánto paga por arriendo o leasing?" tiene la mayor importancia en la predicción. Esto sugiere que el pago de arriendo o leasing es un factor clave en la determinación de la variable objetivo (posiblemente si la persona tiene vivienda propia o no). También son relevantes las variables: "Razón por la que no adquiere vivienda", "Ingreso por cabeza" e "Ingresos" dada su importancia a la hora de poder cumplir con una cuota, y por último se encuentra la variable "OINFORMAL" que se refiere a si la cabeza del hogar tiene empleo informal. Esto también puede determinar si una persona tendría un tipo de ocupación determinada.

Por otro lado, las variables "Cercanía a violencia" y "Minutos a centro médico" tienen la menor importancia en la predicción del modelo. Esto indica que su permutación no afecta significativamente la precisión del modelo, lo que sugiere que no son determinantes en la adquisición de vivienda o en la variable objetivo.

CONCLUSIONES

Los resultados obtenidos en este estudio permiten identificar los principales factores que influyen en la adquisición de vivienda propia en Bogotá. A partir del modelo desarrollado, se observó que el "pago de arriendo o leasing", "la razón por la que no se adquiere vivienda", "el ingreso per cápita" y la "informalidad laboral" son variables determinantes en la clasificación de los hogares en las categorías de vivienda propia, en proceso de pago o en arriendo.

Uno de los hallazgos más relevantes es la fuerte relación entre el nivel de ingresos y la capacidad de acceso a vivienda propia. La falta de recursos para cubrir la cuota inicial sigue siendo una de las principales barreras para la compra de vivienda, lo que resalta la importancia de mecanismos de financiamiento accesibles y políticas públicas que faciliten el acceso a vivienda digna.

El mapa de calor de precios de vivienda por UPZ confirma que la ubicación es un factor clave en la decisión de compra de vivienda. En Bogotá, los valores más altos del suelo se concentran en el oriente de la ciudad, en zonas como Chapinero y el centro financiero, lo que hace que la compra de vivienda en estas áreas sea poco accesible para muchos hogares. En contraste, en las periferias de la ciudad, donde el valor del suelo es significativamente menor, la probabilidad de que los hogares tengan vivienda propia aumenta. Esto indica que muchas familias se ven obligadas a elegir entre acceder a vivienda propia en zonas más alejadas o vivir en arriendo en zonas más centrales y con mejores servicios.

En este sentido, programas como el Subsidio de Concurrencia —que permite la articulación entre subsidios de cajas de compensación familiar y el programa Mi Casa Ya— han demostrado ser herramientas valiosas para ampliar el acceso a la vivienda propia, especialmente para hogares de ingresos bajos y medios. Asimismo, la iniciativa Generación FNA, que ofrece condiciones preferenciales a jóvenes entre 18 y 28 años, incluyendo una reducción de la cuota inicial al 10% y tasas de interés más bajas, representa un avance en la democratización del acceso a vivienda propia.

Sin embargo, aunque existen estos mecanismos de apoyo, el modelo indica que aún hay desafíos estructurales en el acceso a la vivienda. Por ello, es necesario fortalecer y ampliar las políticas públicas orientadas a facilitar la compra de vivienda, especialmente aquellas que aborden el factor del monto de la cuota de pago, ya que su reducción podría aumentar significativamente la posibilidad de que más hogares accedan a una vivienda propia.

Además, la estabilidad laboral y la formalización del empleo emergen como elementos que se deben tener en cuenta en este análisis. La informalidad laboral, que afecta a una gran proporción de la población, limita el acceso al crédito hipotecario y dificulta la planificación financiera a largo plazo. En este sentido, sería recomendable explorar políticas que promuevan la inclusión financiera y la estabilidad económica de los trabajadores informales.

Finalmente, la implementación de modelos de inteligencia artificial como el desarrollado en este estudio permite comprender mejor los patrones y barreras en la adquisición de vivienda, facilitando la toma de decisiones informadas tanto para los formuladores de políticas como para las entidades financieras. La combinación de datos socioeconómicos con herramientas analíticas avanzadas puede contribuir a diseñar estrategias más efectivas para mejorar el acceso a la vivienda en Bogotá y en otras ciudades del país.

Referencias

Dane. (2022). Comunicado de prensa ECV 2021. Recuperado de: https://www.dane.gov.co/files/investigaciones/condiciones_vida/calidad_vida/2021/comunicado_ECV_2021.pdf

Encuesta Multipropósito Bogotá - Cundinamarca - EM – 2021. Dane. Consultado en: https://microdatos.dane.gov.co/index.php/catalog/743

Mediana del Valor de Referencia de Terreno por Manzana. Unidad A. Especial de Catastro Distrital (2022): https://www.ideca.gov.co/recursos/mapas/mediana-del-valor-de-referencia-de-terreno-por-manzana

UPZ Bogotá. Laboratorio urbano Bogotá (2022). Consultado en: https://bogota-laburbano.opendatasoft.com/explore/dataset/upz-bogota/table/

Universidad Externado de Colombia

FACULTAD DE ECONOMÍA

ECONOMÍA

Laboratorio de inteligencia artificial aplicada a Economía

Universidad Externado de Colombia
FACULTAD DE ECONOMÍA

ECONOMIA
Laboratorio de inteligencia artificial aplicada a Economía

C-FIRE:

Desarrollo y Evaluación de un Modelo de Machine Learning para la Gestión Preventiva de Incendios en Colombia

Autores

Santiago Andrés Rodríguez Estrada santiago.rodriguez4@est.uexternado.edu.co
Laura Sofía Romero Suárez laura.romero6@est.uexternado.edu.co

2025

Resumen

El modelo C-FIRE (Colombian Fire Incidence and Risk Estimator) es una herramienta de Machine Learning desarrollada para predecir incendios en Colombia a partir de variables asociadas a la actividad humana, como el uso del suelo, la infraestructura vial y la minería ilegal. Su enfoque busca identificar patrones que contribuyen a la ocurrencia de incendios, proporcionando información clave para una gestión preventiva más eficiente. Aunque no incorpora variables climáticas, su aplicación permite orientar estrategias de mitigación y concientización para reducir prácticas de riesgo y proteger los ecosistemas.

Introducción

El calentamiento global ha empeorado debido al aumento de gases de efecto invernadero, principalmente por actividades humanas, lo que ha generado cambios climáticos extremos como altas temperaturas, olas de calor y sequías severas. Esto ha favorecido la proliferación de incendios forestales incontrolables en regiones como el Amazonas, Australia y California (Grottola, 2021; WWF, 2020; Damasio, 2021; CNN, 2025).

Aunque Colombia no enfrenta esta crisis con la misma intensidad, ha sufrido temporadas de altas temperaturas, como el Fenómeno del Niño, que aumentan el riesgo de incendios forestales, en 2023, los cerros orientales de Bogotá y algunos páramos del país fueron afectados por incendios provocados en su mayoría por desechos arrojados en la zona, que, combinados con la sequía y el calor, generan llamas que se expanden rápidamente (UNGRD, 2023; El Espectador, 2024).

Ante este contexto, el presente proyecto busca predecir la ocurrencia de incendios forestales a nivel sectorial en Colombia (que se refiere a una unidad territorial utilizada para delimitar áreas dentro de barrios, comunas, corregimientos o veredas, facilitando la planificación y gestión de servicios.), considerando variables especialmente centradas en la actividad humana. En particular, se analizarán factores como el tipo de vías, que pueden facilitar el acceso y propagación de incendios, el uso de la tierra, que influye en la disponibilidad de material combustible; y la presencia de minería ilegal, una actividad que ha sido vinculada con la generación de focos de calor y deforestación. Identificar patrones a partir de estas variables permitirá mejorar la prevención y gestión del riesgo de incendios en el país.

Contexto económico

En economía, la prevención de riesgos es clave, especialmente en contextos vulnerables como los incendios forestales. La literatura destaca que los costos de prevención son significativamente menores que los de mitigación, optimizando recursos y evitando gastos posteriores. Kunreuther y Michel-Kerjan (2012) y Freeman et al. (2014) sostienen que cada dólar invertido en prevención puede ahorrar entre cuatro y siete dólares en daños evitados. Además, Shafran (2008) demostró que el uso de inteligencia artificial en la prevención de incendios en California en 2007 redujo en un 30% las pérdidas frente a estrategias reactivas, resaltando la eficiencia de una asignación presupuestaria adecuada.

Colombia ha sido afectada por incendios forestales, agravados por el Fenómeno del Niño, el calentamiento global y la expansión agrícola. Según el IDEAM, entre 2015 y 2020 se quemaron más de 200.000 hectáreas de bosque, causando pérdidas anuales superiores a 3.000 millones de pesos. Para mitigar estos desastres, la ley 1523 de 2012 establece la Política Nacional de Gestión del Riesgo y Desastres. Además, estudios como el de Velásquez y Rosales (2019) proponen el uso de inteligencia artificial para mejorar la prevención, optimizar recursos y fortalecer la capacidad de respuesta ante emergencias.

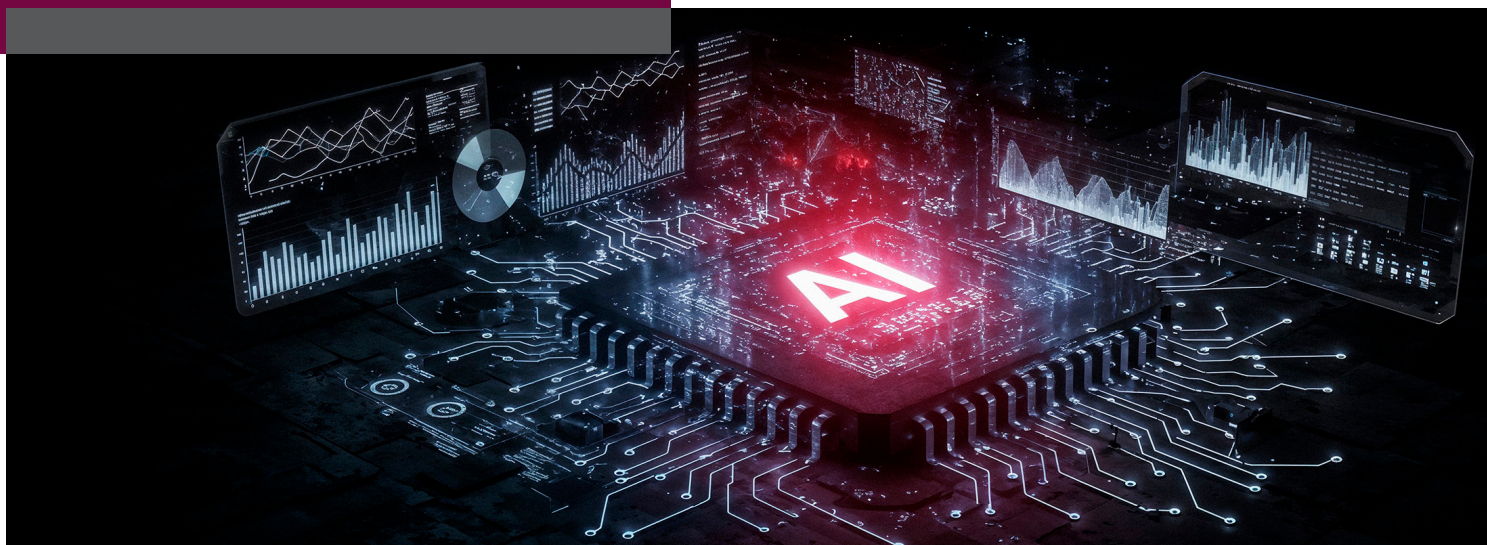
Machine Learning en la gestión de riesgos naturales

En la gestión de riesgos, el Machine Learning se ha adoptado como una herramienta con gran potencia para predecir y prevenir eventuales desastres y lo que eso conlleva. Los modelos elaborados a partir de esta metodología permiten procesar grandes bases de datos, reconocer patrones complejos y realizar predicciones bastante acertadas. Esto es algo trascendental en la toma de decisiones eficientes en la prevención y, de ser el caso, en la reacción frente a situaciones críticas como lo podría ser un incendio forestal. Como mencionan Goodfellow, Bengio y Courville (2016), los algoritmos en el modelaje por medio de Machine Learning evaden uno de los supuestos más limitantes de la econometría como lo es la relación lineal entre variables, por lo que, al establecer relaciones complejas, la precisión de estos modelos se incrementa significativamente, algo crucial en la predicción de desastres naturales.

Para el uso específico de modelos en esta área, se usan datos topográficos, climáticos, sobre la vegetación de las zonas, actividades humanas y otros factores que suelen ser determinantes en la generación de incendios. Con base a estos datos se generan modelos supervisados (como lo son Árboles de Decisión, Bosques Aleatorios, Máquinas de Soporte Vectorial) que permiten clasificar áreas de riesgo y predecir incendios, mientras que modelos no supervisados (como lo es clustering) permiten segmentar regiones en función las características similares que las hagan vulnerables al fuego.

Dentro del estado del arte se encuentran trabajos que implementan un uso más complejo y riguroso de la aplicación de Machine Learning, como lo es el uso de Redes Neuronales, en donde su aplicación ha arrojado grandes resultados como lo fue la investigación de Jain et al. (2020), la cual implemento esta herramienta de Deep Learning junto con imágenes satélites para obtener un modelo con una precisión de predicción de incendios del 85%. Además de esta investigación, la realizada por Zhao et al. (2018) muestra un enfoque crucial en la prevención de riesgos naturales como lo son incendios forestales, buscando predecir la severidad de los incendios en Australia.

En el caso de Colombia, la adopción de tecnologías de Machine Learning en la gestión de incendios representa una oportunidad invaluable para enfrentar los desafíos asociados con el cambio climático y las dinámicas locales. Los ecosistemas diversos del país, sumados a su variabilidad climática, generan un entorno donde la implementación de estas herramientas no solo es pertinente, sino urgente. Aprender de los casos internacionales y adaptar modelos existentes a las características propias de los ecosistemas colombianos puede ser un paso crucial hacia la prevención y mitigación de incendios forestales.



Resultados

Para la elaboración del modelo C-FIRE se crearon variables en base a fuentes que se consideraron pertinentes para la explicación en la generación de incendios, especialmente aquellas que pudieran reflejar la influencia humana. Para este fin, se tomaron principalmente datos del sistema FIRMS y datos del marco geoestadístico del DANE.

El Sistema de Información sobre Incendios para la Gestión de Recursos (FIRMS, por sus siglas en inglés), es una herramienta desarrollada por la NASA que proporciona acceso a datos casi en tiempo real sobre la actividad de incendios a nivel global. Su objetivo principal es ofrecer a los gestores de recursos naturales y al público en general información actualizada sobre la ubicación, extensión e intensidad de los incendios forestales. FIRMS ofrece datos globales disponibles dentro de las 3 horas posteriores a la observación satelital. Para Estados Unidos y Canadá, algunas detecciones de incendios activos están disponibles en tiempo real. El sistema utiliza imágenes satelitales de instrumentos como MODIS y VIIRS, que detectan incendios activos y puntos calientes en la superficie terrestre.

Datos geoespaciales, como las vías principales, secundarias y terciarias, y el mapa de Colombia dividido en secciones fueron adquiridos del geoportal del DANE, el cual es un conjunto de datos geográficos y territoriales que se utiliza para la recolección, procesamiento y análisis de información estadística en Colombia. Este marco tiene como propósito proporcionar un sistema de referencia geográfica consistente, que permita identificar y localizar las áreas geográficas relevantes para la producción de estadísticas y censos nacionales, como los límites de municipios, corregimientos, veredas, y otras divisiones administrativas.

El marco geoespacial del DANE incluye información detallada sobre las fronteras y las características físicas y administrativas del territorio colombiano. Además, facilita la integración de datos estadísticos con mapas y otras herramientas geográficas, mejorando la precisión en los análisis y la planificación en diversas áreas como la educación, la salud, la infraestructura, y el desarrollo económico.

Los datos originales, que más luego serían transformados y generadores de otras variables fueron:

Registros de incendios:

A partir del sistema FIRMS se obtuvieron registros de apariciones de puntos de calor detectados como incendios por los satélites de la NASA. Esta base (para Colombia) fue el punto de partida para la obtención de datos y posterior generación de variables para el desarrollo del modelo C-FIRE. Con el fin de obtener una amplia cantidad de registros, se tomó el periodo comprendido desde enero de 2020 (2020-01) hasta septiembre de 2024 (2024-09), fecha en la que empezó el desarrollo de C-FIRE.

División sectorial de Colombia:

Para que las predicciones del modelo C-FIRE fueran lo más precisas posible, era fundamental definir una división territorial adecuada. Usualmente, las divisiones administrativas tradicionales, como departamentos o municipios, abarcan áreas demasiado extensas, lo que podría introducir sesgos y pérdida de detalle espacial en el análisis de incendios. Dado que los incendios son

fenómenos altamente localizados, influenciados por factores como cobertura vegetal, topografía y clima, se requería una segmentación más pequeña del territorio. Por esta razón, se optó por utilizar una división sectorial.

Vías principales, secundarias y terciarias:

La infraestructura vial juega un papel importante en la ocurrencia y propagación de incendios, ya que las vías pueden actuar como barreras o como conductores de actividad humana que incremente el riesgo de que ocurran estos siniestros. Además, es importante tener en cuenta que la proximidad de incendios a vías facilita el acceso de las entidades encargadas de la mitigación y control del fuego, por un lado, mientras que por el otro puede significar la presencia de tránsito humano, cuya actividad puede ser el iniciador de incendios.

Uso de la Tierra:

El uso del suelo es un factor determinante, ya que ciertas actividades económicas y agrícolas pueden incrementar la probabilidad de incendios. Las zonas destinadas a la agricultura, ganadería, explotación forestal e industrias suelen producir residuos o realizar quemas controladas que pueden extenderse e influir en la generación de grandes incendios forestales.

Puntos de Minería Ilegal:

Se incluyeron los puntos donde se identificó actividad minera ilegal debido a que estas zonas suelen ser altamente inflamables. La minería ilegal suele involucrar deforestación y el uso de materiales combustibles, lo que incrementa el riesgo de incendios.

Variables

Las variables del modelo fueron creadas a partir de datos originales para capturar patrones en la ocurrencia de incendios. Se incluyeron indicadores de actividad humana, infraestructura y dinámica espacial del fuego.

La variable principal del modelo es la presencia de incendios por mes y sector, ya que actúa como la variable objetivo a predecir. Se trata de una variable dicotómica que indica si ocurrió un incendio en un sector durante un mes determinado, construida a partir de los registros del sistema FIRMS.

A partir de esta, se incluyeron variables explicativas clave. La frecuencia de incendios por sector y la distancia promedio entre incendios permiten identificar tendencias y posibles patrones de propagación. Además, se consideraron variables relacionadas con la infraestructura vial, como el conteo de vías, la distancia a vías y la moda del tipo de vía, así como el conteo de minería ilegal, debido a su impacto ambiental. También se incluyó la moda del uso de la tierra, que refleja las actividades predominantes en cada sector. Finalmente, las variables de sector, año y mes permiten contextualizar los datos en tiempo y espacio.

Resultados

El modelo de Arboles de decisión es el más adecuado para el problema que se decidió tratar. Además, al observar el reporte de clasificación del modelo, se puede ver que alcanzó una precisión general de 0,97, lo que refleja un buen desempeño en términos de clasificación global. Los resultados detallados en la Tabla 1 muestran que el modelo es muy efectivo al predecir la clase 0 (no hubo incendio), con una precisión de 0.96 y un recall de 0,98. Para la clase 1 (hubo incendio), que podría ser más difícil de predecir debido a su menor representación en el conjunto de datos, se obtuvo una precisión de 0,87 y un recall de 0,95, lo que sugiere que el modelo tiene un buen equilibrio entre la predicción de ambas clases.

El F1-Score, que combina precision y recall, es 0,97 para la clase 0 y 0,96 para la clase 1, lo que indica un buen balance entre las métricas de clasificación y un desempeño robusto en ambos casos. En general, los resultados sugieren que el modelo tiene un rendimiento confiable y adecuado para la tarea, destacándose por su capacidad para manejar el desbalance entre clases y obtener predicciones precisas y consistentes.

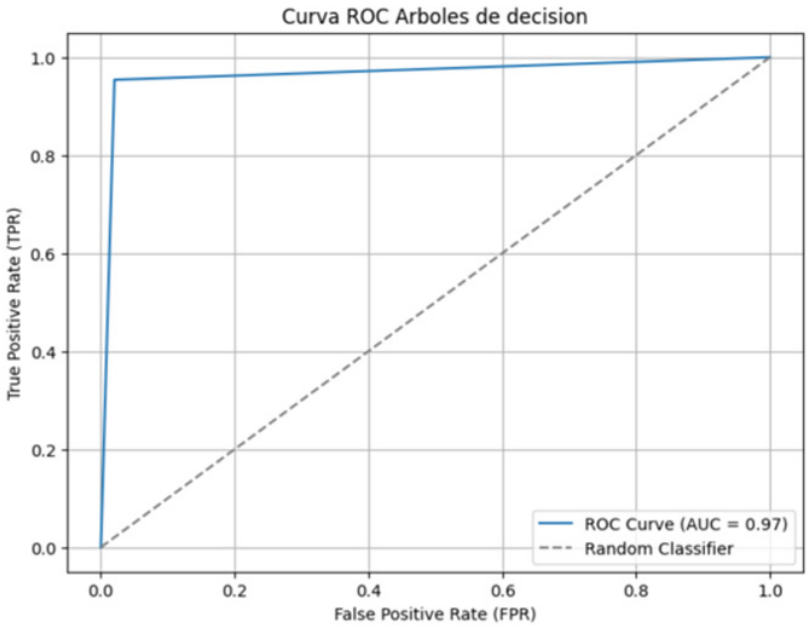
Tabla 1 Matriz de métricas de clasificación del modelo Árboles de Decisión

	Precision	Recall	F1-Score	SVM
0	0,96	0,98	0,97	15673
1	0,97	0,95	0,96	12538
Accuracy			0,97	28211

Fuente: Elaboración propia

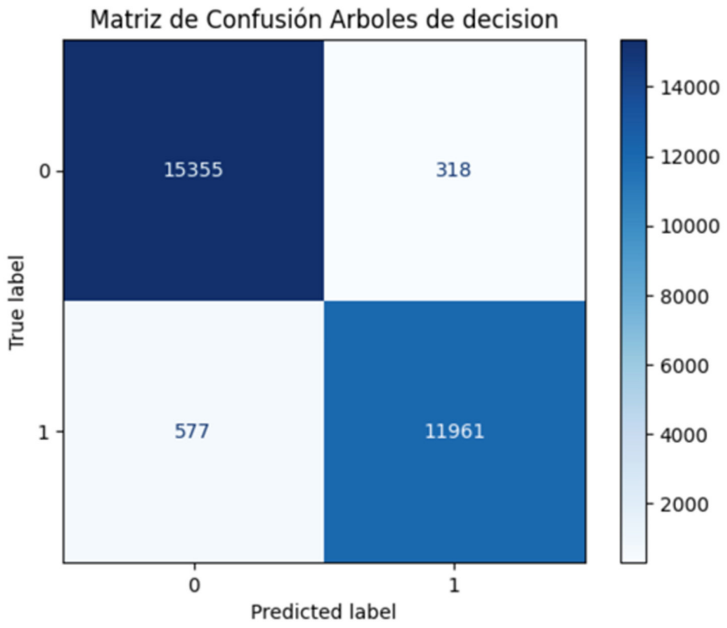
Por su parte, el resultado de la AUC se observa en la Figura 1, en donde la curva ROC de Arboles de decisión muestra un excelente desempeño con un AUC de 0.97, indicando una alta capacidad para distinguir entre clases. La curva está muy por encima del clasificador aleatorio, con una baja tasa de falsos positivos y alta de verdaderos positivos. En conclusión, el modelo es altamente efectivo para la tarea de clasificación.

Figura 1 Curva ROC para el modelo Arboles de decisión



Fuente: Elaboración propia

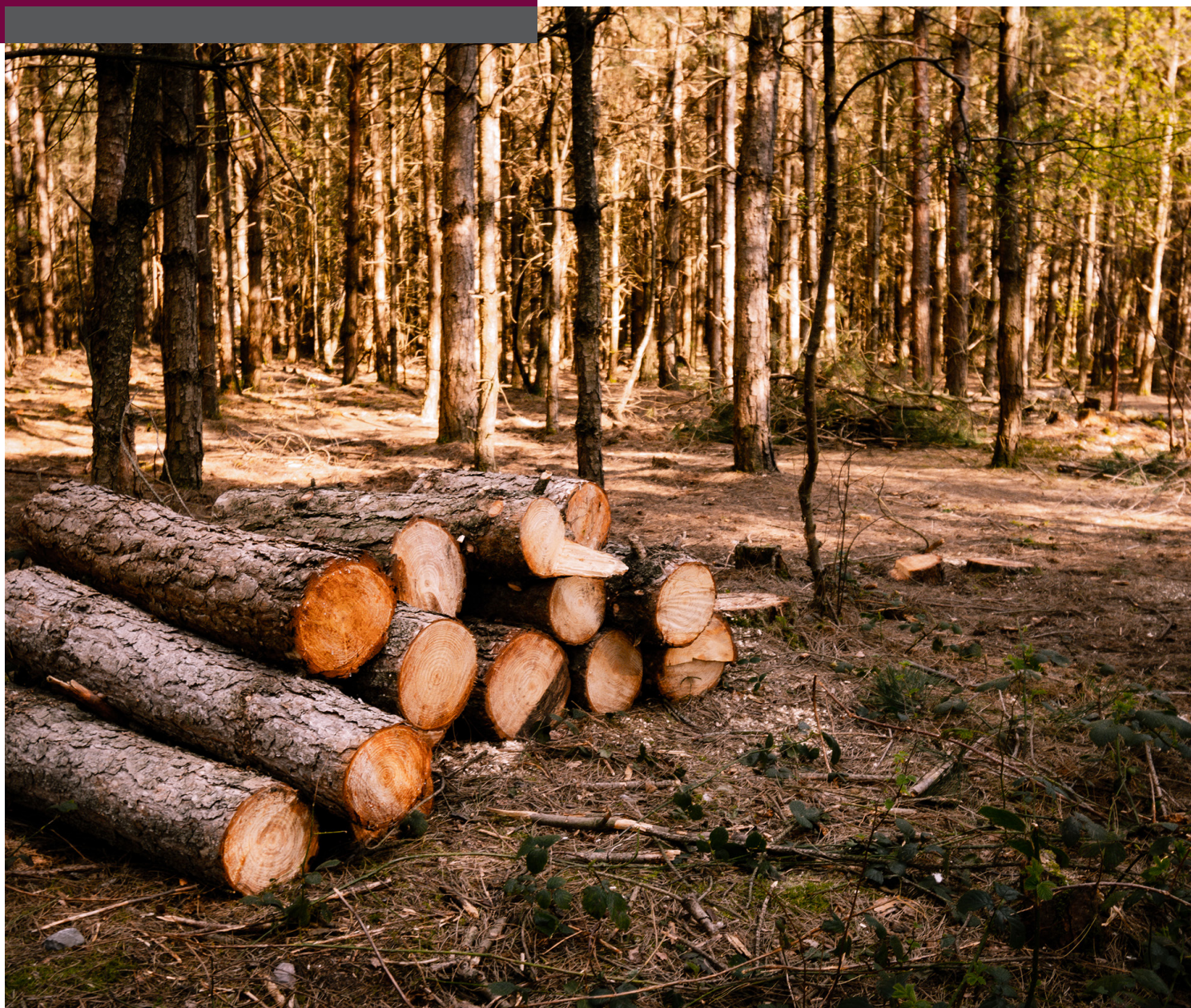
Figura 2 Matriz de confusión para modelo de Arboles de decisión



Fuente: Elaboración propia

Variables

El modelo de Árboles de Decisión demostró el mejor desempeño predictivo y una alta capacidad para clasificar correctamente los datos, alcanzando una precisión del 97%. Los mejores hiperparámetros identificados incluyen una profundidad máxima de 6, un mínimo de 2 muestras por hoja y por división de nodo, lo que equilibra complejidad y generalización. El reporte de clasificación muestra valores de F1-score cercanos a 0.97, indicando un buen balance entre precisión y recall para ambas clases. La matriz de confusión revela que el modelo predice correctamente la mayoría de los casos, con 15,355 verdaderos positivos en la clase 0 y 11,661 en la clase 1, aunque aún existen algunos errores, como 577 falsos negativos y 118 falsos positivos. En general, estos resultados sugieren que el modelo es robusto y confiable para la predicción de incendios en Colombia.



CONCLUSIONES

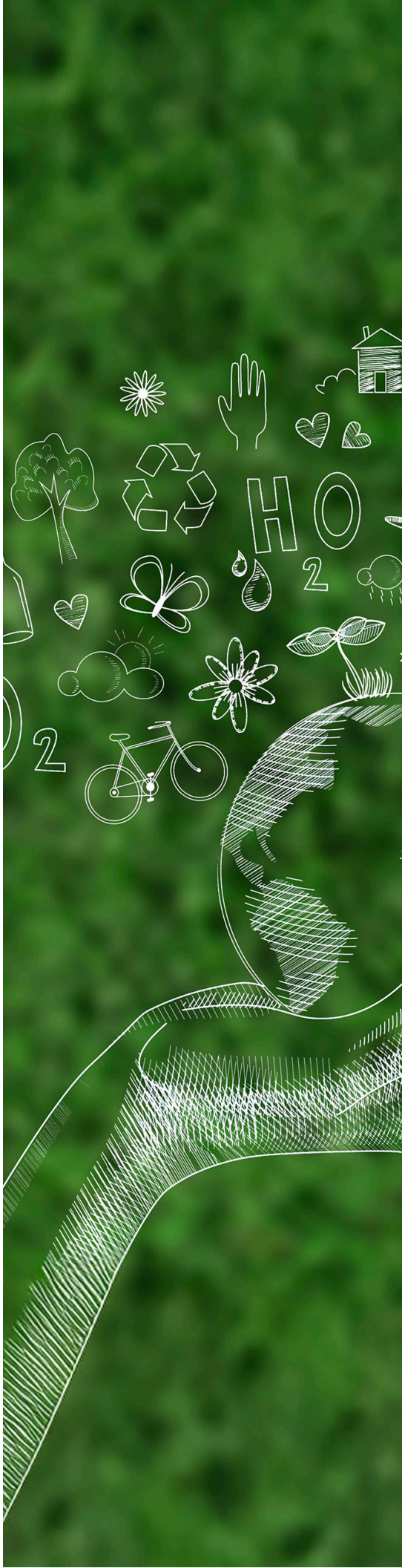
El desarrollo del modelo C-FIRE representa un avance significativo en la gestión preventiva de incendios en Colombia. A través del uso de técnicas de Machine Learning, se logró crear una herramienta robusta y precisa, con una precisión general del 97%, capaz de predecir la ocurrencia de incendios basándose en variables relacionadas con el uso humano del suelo, la infraestructura vial y la actividad minera ilegal. Los resultados demuestran que el modelo es altamente efectivo para identificar áreas de riesgo, lo que permite una toma de decisiones más informada y oportuna para la prevención y mitigación de incendios.

Sin embargo, es importante destacar que una de las principales limitaciones del modelo fue la falta de acceso a variables climáticas, como temperatura, humedad y precipitaciones, las cuales son factores clave en la propagación de incendios. Aunque el modelo se basó en variables antropogénicas, que resultaron ser altamente predictivas, la inclusión de datos climáticos en futuras iteraciones podría mejorar aún más su capacidad predictiva, especialmente en regiones donde las condiciones meteorológicas son determinantes.

Dado que el análisis reveló que las actividades humanas son una de las principales causas de incendios en Colombia, es fundamental implementar campañas de concientización dirigidas a la población. Estas campañas deberían enfocarse en:

• Educación sobre prácticas seguras:
Informar a las comunidades sobre los riesgos asociados con actividades como la quema de basura, la quema controlada de terrenos agrícolas y el uso de materiales inflamables en áreas propensas a incendios. Es crucial que las personas comprendan cómo sus acciones pueden desencadenar incendios y cómo evitarlos.

• Promoción de alternativas sostenibles:
Fomentar prácticas agrícolas y ganaderas más sostenibles que reduzcan la necesidad de quemas controladas. Esto podría incluir



la promoción de técnicas de cultivo que no dependan del fuego y la implementación de sistemas de manejo de residuos más seguros.

• Conciencia sobre el impacto ambiental:
Sensibilizar a la población sobre las consecuencias ambientales y económicas de los incendios forestales, incluyendo la pérdida de biodiversidad, la degradación del suelo y el impacto en la calidad del aire. Esto podría lograrse a través de talleres comunitarios, material educativo y campañas en medios de comunicación.

• Participación comunitaria:
Involucrar a las comunidades locales en la prevención de incendios, fomentando la creación de brigadas comunitarias y programas de vigilancia. La participación de la población puede ser clave para identificar y mitigar riesgos de manera temprana.

• Colaboración interinstitucional:
Establecer alianzas entre entidades gubernamentales, organizaciones no gubernamentales y el sector privado para fortalecer las estrategias de prevención y respuesta. Esto incluye la integración de tecnologías avanzadas, como el modelo C-FIRE, en los planes de gestión de riesgos.

El modelo C-FIRE es una herramienta valiosa para la prevención de incendios en Colombia, pero su efectividad puede ser potenciada mediante la integración de datos climáticos y la implementación de campañas de concientización que aborden las causas humanas de los incendios. La combinación de tecnología avanzada y educación comunitaria puede ser la clave para reducir significativamente la incidencia de incendios forestales y proteger los ecosistemas colombianos.

Referencias

Kunreuther, H., & Michel-Kerjan, E. (2012). Managing Catastrophes through Insurance: Challenges and Opportunities for Reducing Future Risks. Journal of Risk and Uncertainty.

Freeman, P. K., Keen, M., & Mani, M. (2014). Dealing with Increased Risk of Natural Disasters: Challenges and Costs. IMF Working Paper.

Shafran, A. P. (2008). Risk Externalities and the Problem of Wildfire Risk. Journal of Urban Economics.

Instituto de Hidrología, Meteorología y Estudios Ambientales (IDEAM). Informes sobre incendios forestales en Colombia, 2015-2020.

Velásquez, C., & Rosales, J. (2019). Políticas de gestión del riesgo en Colombia: desafíos y oportunidades en un contexto de cambio climático. Revista Colombiana de Ciencias Sociales.

Goodfellow, I., Bengio, Y., & Courville, A. (2016). Deep Learning. MIT Press.

Jain, P., Coogan, S. C., Subramanian, S. G., Crowley, M., Taylor, S., & Flannigan, M. D. (2020). A review of machine learning applications in wildfire science and management. Environmental Reviews.

Zhao, F., Wang, J., Yang, J., & Zheng, J. (2018). Using Deep Learning for Predicting Fire Spread in Forests. Forest Ecology and Management.

Departamento de Bosques y Protección contra Incendios (CAL FIRE). Informes técnicos sobre sistemas predictivos.

Instituto Nacional de Técnica Aeroespacial (INTA). Proyecto FUEGO: Informe de resultados 2020

ECONOMIA

Laboratorio de inteligencia artificial aplicada a Economía

Universidad Externado de Colombia

FACULTAD DE ECONOMÍA

ECONOMIA

Laboratorio de inteligencia artificial aplicada a Economía

Predicción de las hectáreas de cultivos de coca:

Un modelo de Machine Learning por municipios en Colombia

Autores

Roberto Carlos Chapman Diaz roberto.chapman@est.uexternado.edu.co
Laura Juliana Bolívar Roa laura.bolivar02@est.uexternado.edu.co

2025

Resumen

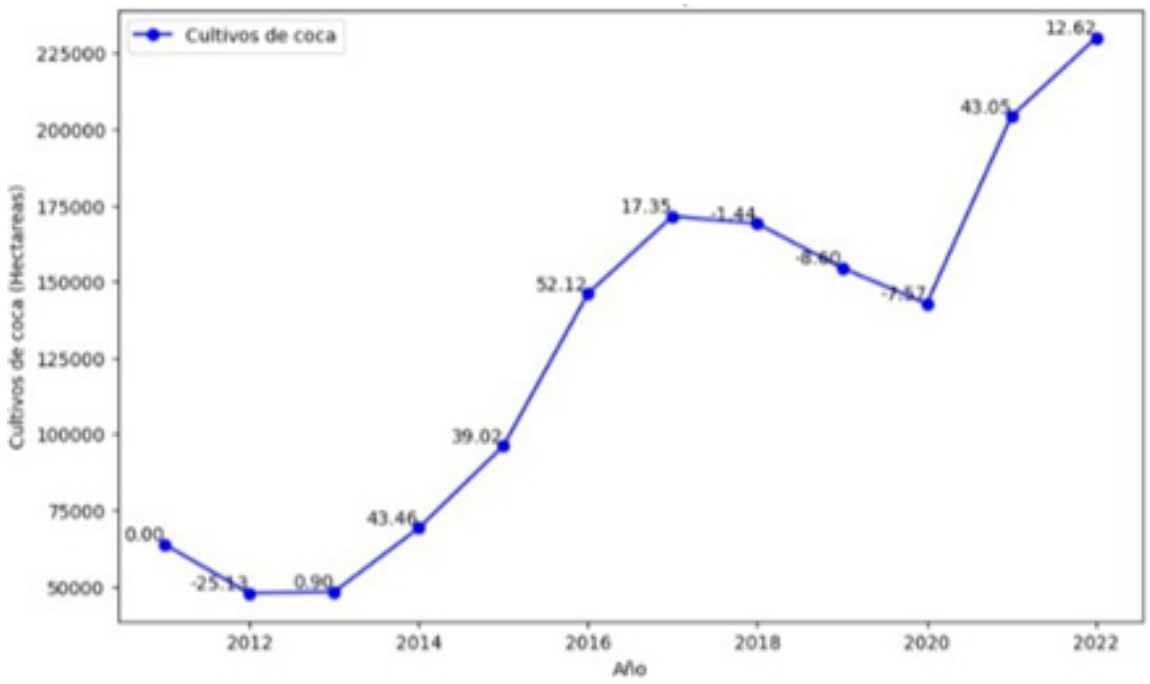
La persistencia y expansión de los cultivos ilegales de hoja de coca en Colombia, impulsada por el lucrativo negocio internacional de las drogas, ha sido un problema significativo que afecta la institucionalidad del país debido a su vinculación con la violencia, el conflicto y la presión internacional. Aunque autores de diversas disciplinas han investigado las causas y consecuencias del crecimiento de estos cultivos, la falta de información actualizada sobre la extensión de las hectáreas sembradas en distintos municipios ha dificultado la implementación de medidas efectivas, resultando en un aumento continuo de estas áreas en los últimos años. Con datos actualizados solo hasta 2022 por el Sistema Integrado de Monitoreo de Cultivos Ilícitos, Colombia se ha consolidado como el mayor productor de cocaína, con un récord de 230,000 hectáreas cultivadas en 2022. Es por esto que nuestro principal objetivo es solucionar la falta de datos actualizados sobre la extensión de los cultivos de coca.

Este estudio propone desarrollar un modelo de machine learning con alta capacidad predictiva que permita determinar cuántas hectáreas de cultivo de coca tiene un municipio en Colombia. Este modelo será una herramienta fundamental para comprender la distribución y la evolución de los cultivos de coca en el país, proporcionando información precisa y actualizada que respalde la toma de decisiones y la implementación de políticas efectivas para combatir el narcotráfico en el país

Introducción

La producción y el tráfico de coca en Colombia representan un fenómeno complejo y multifacético que tiene profundas implicaciones socioeconómicas, ambientales y políticas. Estas actividades dependen de una variedad de factores los cuales algunos de ellos se van a abarcar más adelante. En este contexto, los cultivos de coca han experimentado un crecimiento notable a lo largo de los años, pasando de tener cerca de 60,000 hectáreas en 2011 a 230,000 hectáreas en 2022, como se puede apreciar en el Gráfico 1. Este aumento no solo refleja la expansión de esta actividad ilegal, sino que también subraya la complejidad del desafío que representa para las autoridades y la sociedad en su conjunto.

Gráficos 1 Hectáreas de Cultivos de Coca por municipios y evolución por años



Fuente: Cálculos propios a partir de: Datos Abiertos Colombia. (s.f.).

Estas actividades tienen un significativo valor agregado en la economía de ciertos municipios, en departamentos como lo son Nariño, Norte de Santander y Putumayo, que representan conjuntamente cerca del 65% del total de los cultivos, siendo una fuente importante de ingresos para diversas comunidades, especialmente en zonas rurales. Los cultivos de coca, aunque ilegales, representan una parte significativa de la actividad agrícola en más de 300 municipios.

Análisis y Metodología

Se empleó el modelo Random Forest para predecir los cultivos de coca en Colombia, seleccionado por su capacidad para manejar grandes conjuntos de datos con múltiples variables predictoras y capturar relaciones complejas entre ellas. Se recopilaron datos históricos sobre cultivos de coca y variables socioeconómicas, geográficas y ambientales relevantes para entrenar el modelo. Tras dividir el conjunto de datos en entrenamiento y prueba, se realizó una validación cruzada para evaluar su rendimiento y capacidad predictiva. Se ajustaron los parámetros del modelo para optimizar su rendimiento y minimizar el error de predicción.

Para añadir una explicación matemática de cómo funciona el modelo, habría que empezar desde lo más sencillo, que es el árbol de decisión, el componente básico de un bosque aleatorio. Cada árbol se construye dividiendo recursivamente los datos en diferentes ramas basadas en reglas de decisión. Estas decisiones se toman optimizando ciertas métricas de impureza, en este caso utilizamos la impureza de Gini:

$$G(S) = 1 - \sum_{i=1}^{k} p_i^2$$

donde *pi* es la proporción de instancias de la clase i en el conjunto de datos.

División de un Nodo:

En cada nodo *t*, el conjunto de datos *Dt* se divide en dos subconjuntos *Dizquierda* y *Dderecha* basados en una condición que involucra una de las características. La condición seleccionada es aquella que maximiza la reducción en la medida de impureza o maximiza la ganancia de información.

Ganancia de Información:

$$IG(t) = I(t) - \left(\frac{|D_{\text{izquierda}}|}{|D_t|} I(D_{\text{izquierda}}) + \frac{|D_{\text{derecha}}|}{|D_t|} I(D_{\text{derecha}}) \right)$$

donde *I(t)* es la impureza del nodo *t* antes de la división, y *I(Dizquierda)* e *I(Dderecha)* son las impurezas de los datos después de la división.

Combinando Árboles en un Random Forest:

Un Random Forest combina varios árboles de decisión, cada uno construido de la manera descrita anteriormente, para formar un modelo más robusto.

Número de Árboles *B:*

El Random Forest construye *B* árboles independientes. La predicción del conjunto se hace por mayoría de votos (clasificación) o promediando las predicciones (regresión) de todos los árboles.

Predicción de Clase en Random Forest:

$$C(x) = \text{moda}\{C_1(x), C_2(x), \ldots, C_B(x)\}$$

donde *Ci(x)* es la predicción del i-ésimo árbol.

Predicción de Regresión en Random Forest:

$$R(x) = \frac{1}{B} \sum_{i=1}^{B} R_i(x)$$

donde Ri(x) es la predicción del i-ésimo árbol.

Ganancia de Información:

La combinación de varios árboles reduce la varianza de las predicciones, haciendo al modelo menos propenso a sobreajuste en comparación con un solo árbol de decisión.

Este enfoque de ensamble aprovecha el hecho de que un conjunto de modelos (árboles) que trabajan juntos pueden alcanzar mejor desempeño que cualquiera de los modelos individuales operando por sí solos.

Datos

Tabla 1 Conjunto de variables utilizadas en el modelo

	DATOS	
Municipios	Zona de Frontera	Zonas más Afectadas por el Conflicto Armado
Valor Agregado	Kilómetros de Vías Primarias y Secundarias	Zonas Húmedas
Cultivos de Coca	Área Total del Municipio	Volumen de Gasolina Demandada
Población	Altura Media del Municipio	Número de Resguardos Indígenas
Población Rural	Incautaciones de Coca	TRM
Valor Primario	Zonas más Afectadas por el Conflicto Armado	Precio de la Hoja de Coca

Fuente: Elaboración propia

Resultados

Para nuestro estudio empleamos el modelo Random Forest para predecir con un alto nivel de precisión la cantidad de cultivos de coca por municipios en Colombia, seleccionando cuidadosamente variables socioeconómicas y geográficas como predictores del modelo.

Los resultados revelaron una capacidad explicativa del 89.5% en los datos de prueba y del 97.8% en los de entrenamiento, con un bajo error medio absoluto (MAE) de 0.043 y 0.023 respectivamente.

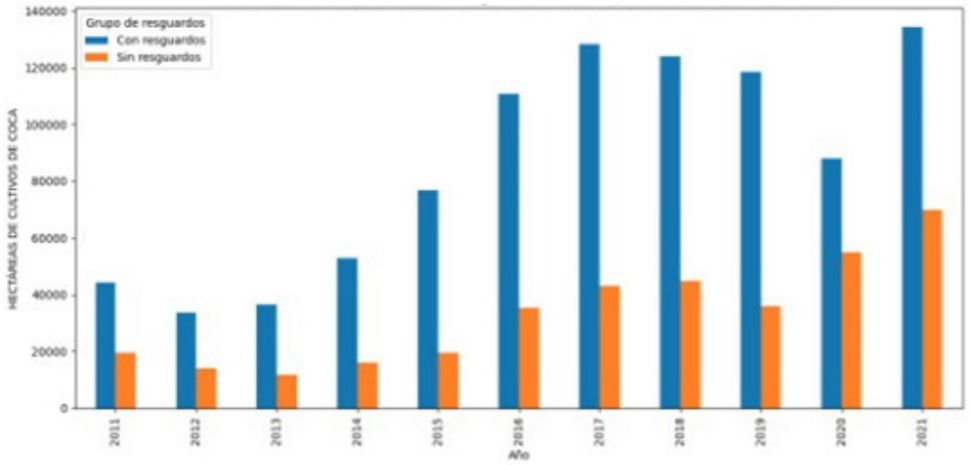
Estos indicadores demuestran una sólida comprensión de los factores que influyen en los cultivos de coca, lo que hace que el modelo sea altamente efectivo y confiable para la predicción precisa de la cantidad de cultivos en los diferentes municipios.

Esta alta capacidad predictiva lo convierte en una herramienta invaluable para la planificación pública, ya que aborda el desafío de la falta de datos actualizados, permitiendo tomar decisiones informadas y diseñar estrategias eficaces para combatir los cultivos de coca y sus implicaciones socioeconómicas.

Una primera revisión de nuestros datos

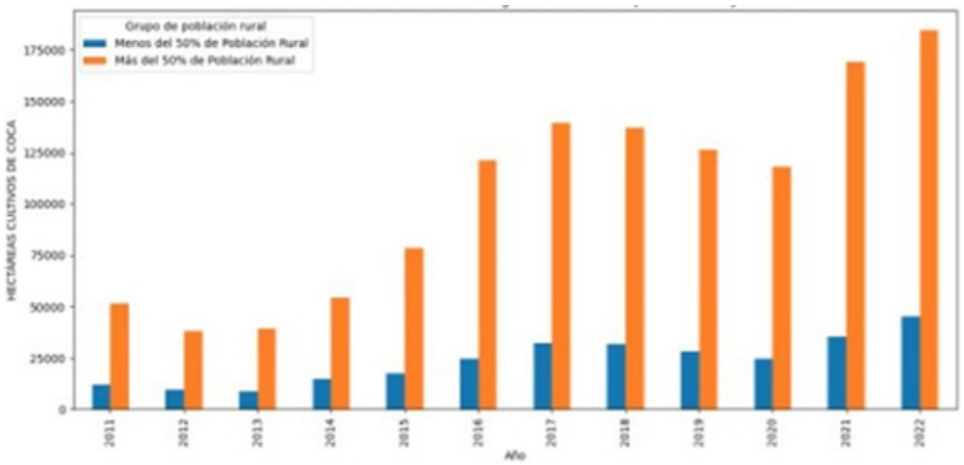
Los cultivos de coca son una parte importante de la agricultura en el país, ligados a la población rural y a los resguardos indígenas. El crecimiento alarmante de los cultivos de coca en estos territorios, alcanzando el 11% del total nacional en 2014, refleja una transformación acelerada debido a actividades perjudiciales como monocultivos y minería. A pesar de protecciones legales, la autonomía de los resguardos se ve amenazada por actores externos. La ubicación remota y la baja presencia estatal facilitan la expansión de los cultivos ilícitos, afectando negativamente a las comunidades indígenas en aspectos sociales, de salud y ambientales. La legislación que exige una Consulta Previa para la erradicación de cultivos en territorios indígenas complica aún más la intervención estatal en esta problemática

Gráfico 2 Cultivos de Coca en municipios con resguardos indígenas



Fuente: Cálculos propios a partir de: Departamento Nacional de Planeación. (s.f.). Terridata.

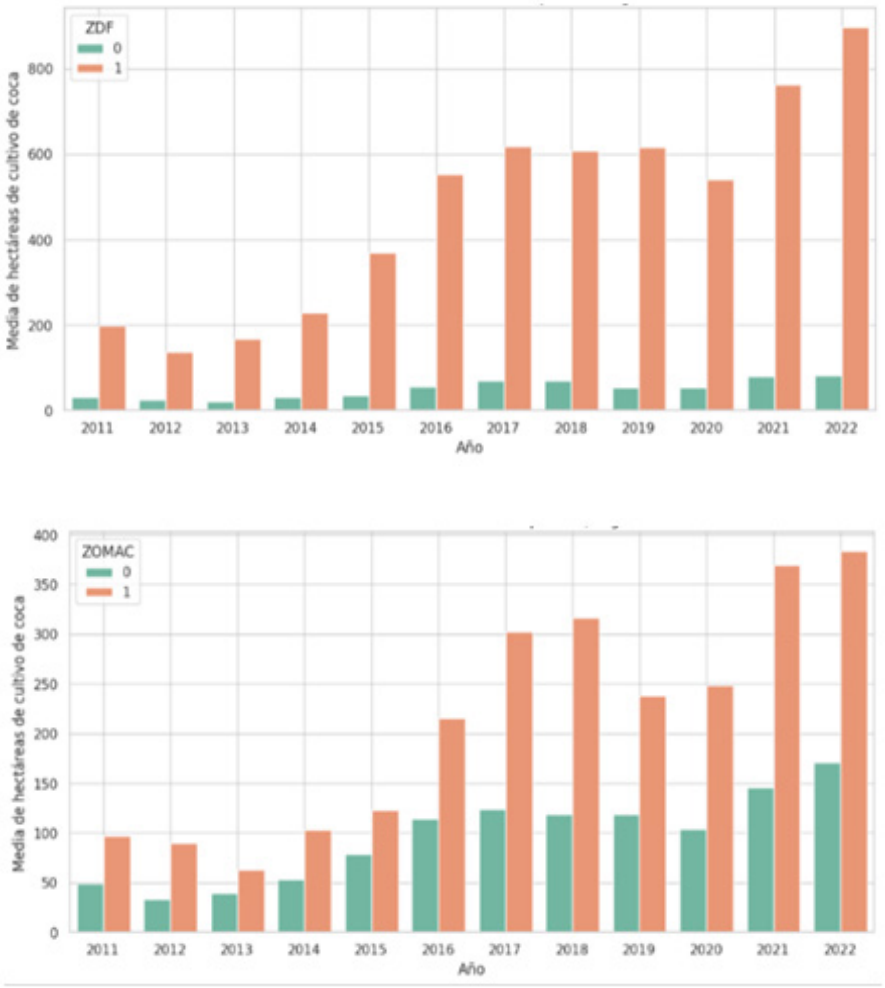
Gráfico 3 Cultivos de Coca en municipios por grupo de población rural



Fuente: Cálculos propios a partir de: Departamento Nacional de Planeación. (s.f.). Terridata

Los cultivos de coca suelen estar controlados por grupos armados ilegales en áreas conflictivas y zonas de frontera, donde la presencia del Estado es limitada. Estas áreas proporcionan condiciones propicias para el cultivo, producción y tráfico de cocaína, así como para la expansión del conflicto armado y la violencia. A pesar del acuerdo de paz, las nuevas estructuras criminales han extendido su influencia en el país, financiándose en gran parte a través del narcotráfico, lo que perpetúa la violencia y afecta negativamente a los cultivadores de coca.

Gráfico 4 Relación de los cultivos de Coca con Las Zonas Más Afectadas por el Conflicto Armado Y Relación de los Cultivos de Coca con las Zonas de Frontera

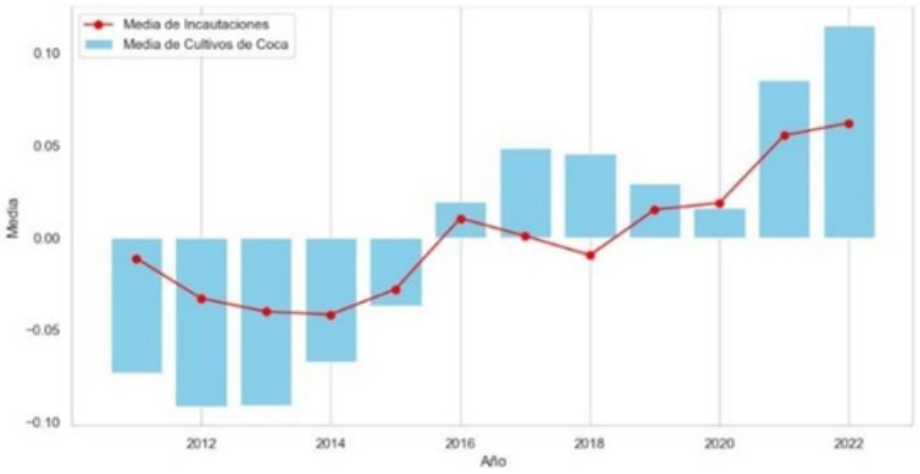


Fuente: Cálculos propios a partir de: Datos Abiertos Colombia. (s.f.).

El conflicto armado y la falta de recursos dificultan el control de los cultivos de coca. En municipios costeros como Buenaventura, se incautaron más de 22,730 kilogramos de cocaína en 2022, junto con más de 1,000 hectáreas de cultivos de coca. Esta ciudad, afectada por la violencia y su alta humedad, destaca la importancia de considerar no solo las zonas fronterizas, sino también las áreas costeras.

Los puertos como Buenaventura no solo son propensos a la presencia de cultivos de coca, sino que también se convierten en puntos clave para la exportación de drogas ilícitas.

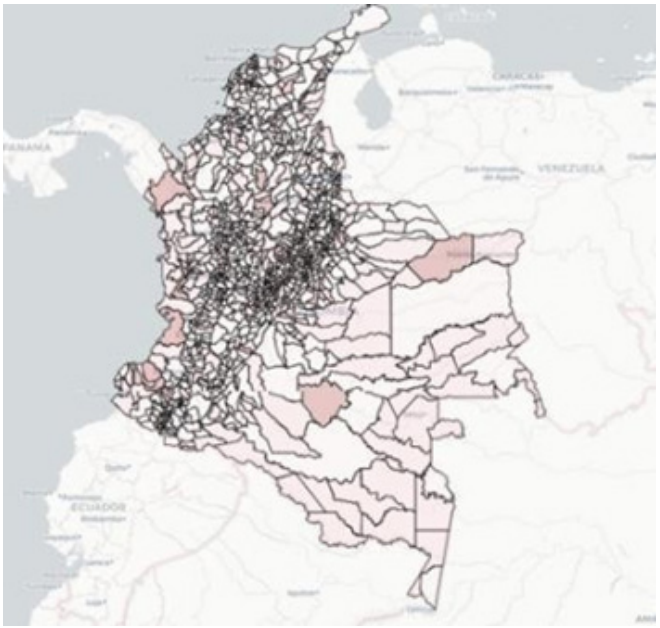
Gráfico 5 Comparación de Medias de Incautaciones y Cultivos de Coca a lo largo del tiempo



Fuente: Cálculos propios a partir de: Departamento Nacional de Planeación. (s.f.). Terridata

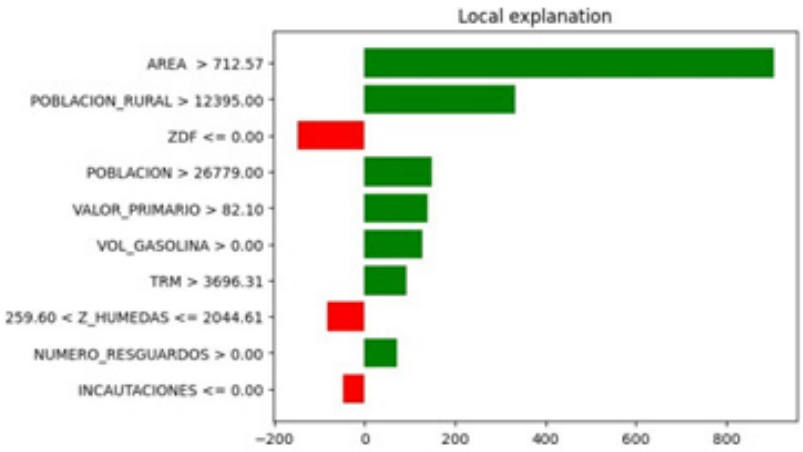
La gasolina desempeña un papel esencial en la producción de los cultivos, siendo un elemento presente en diversas etapas del proceso. En primer lugar, se emplea en la extracción de la pasta básica de cocaína a partir de las hojas de coca, facilitando la separación de los alcaloides de la planta. Esta actividad suele llevarse a cabo en áreas remotas y de difícil acceso, donde la gasolina es el combustible preferido debido a su disponibilidad y portabilidad. Lo que sugiere un posible desvío de combustible hacia la producción de cocaína.

Gráfico 6 Mapa de municipios según la gasolina per cápita



Fuente: Cálculos propios a partir de: Datos Abiertos Colombia. (s.f.).

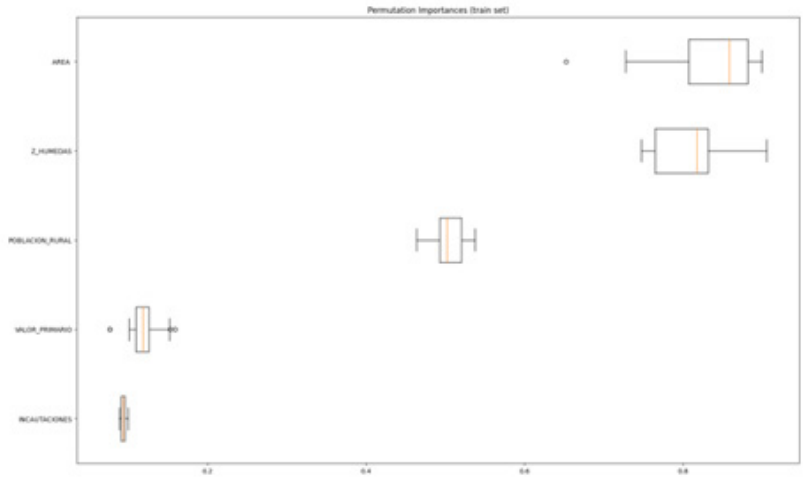
Gráfico 7 Local Interpretable Model-Agnostic Explanations



Fuente: Elaboración propia

LIME es un método para interpretar modelos de aprendizaje automático a nivel local, perturbando ligeramente los datos de entrada y construyendo un modelo sencillo para comprender el comportamiento del modelo complejo. En una gráfica de LIME, las barras verdes indican características que aumentan la predicción (como `AREA` y `POBLACION_RURAL`), mientras que las rojas indican características que la disminuyen (como `ZDF` y `Z_HUMEDAS`). Esto ayuda a entender qué características son más importantes para una predicción específica y en qué medida influyen

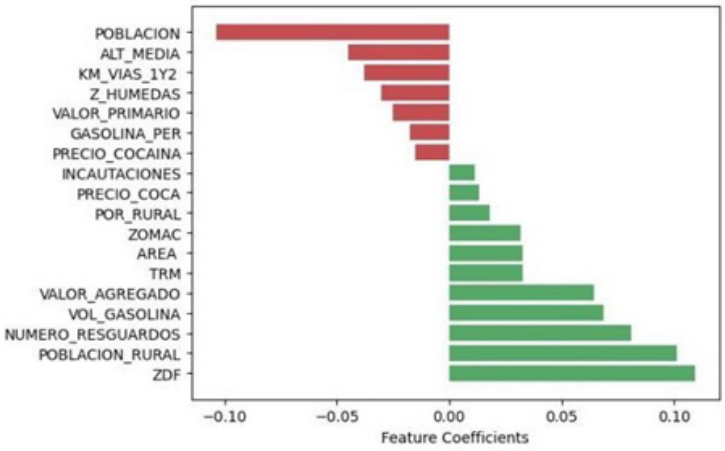
Gráfico 8 Permutation Feature Importance



Fuente: Elaboración propia

El método de Permutation Feature Importance evalúa la importancia de las características al permutar aleatoriamente sus valores y medir el aumento en el error del modelo. Las características que causan un mayor aumento en el error se consideran importantes. Esto es útil ya que no depende de la estructura del modelo y se puede aplicar a cualquier modelo de aprendizaje automático. En la gráfica, `AREA` y `Z_HUMEDAS` son las características más importantes, con valores alrededor de 0.8 y 0.6 respectivamente, lo que indica su crucial papel en la precisión del modelo para predecir hectáreas de cultivo de coca. Las cajas de bigotes muestran la consistente influencia de estas características en la predicción del modelo, mientras que características como `INCAUTACIONES` tienen una importancia mucho menor.

Gráfico 9 Global Surrogate Model



Fuente: Elaboración propia

Un Global Surrogate Model es un modelo interpretable que se ajusta para aproximar un modelo complejo en su totalidad, permitiendo entender cómo toma decisiones el modelo complejo al examinar un modelo más simple.

En la gráfica, los coeficientes de las características indican su influencia en la variable objetivo, las hectáreas de cultivo de coca. `ZDF` y `POBLACION_RURAL` tienen coeficientes positivos altos, aproximadamente 0.10 y 0.08 respectivamente, sugiriendo una fuerte asociación con un aumento en las hectáreas de cultivo de coca. `POBLACION` y `ALT_MEDIA` tienen coeficientes negativos significativos, aproximadamente -0.10 y -0.08, indicando una asociación con una disminución en las hectáreas de cultivo de coca. Estos resultados son útiles para identificar qué características gestionar o intervenir para controlar el cultivo de coca en diferentes regiones.

CONCLUSIONES

El uso de este modelo predictivo de Machine Learning resulta importante para abordar el desafío persistente de la ausencia de datos actualizados sobre cultivos de hoja de coca en el país. Al predecir este tipo de cultivo ilícito con una precisión del 89.5%, se proporciona una fuente confiable de información, que puede ser fundamental para la planificación y ejecución de estrategias más eficaces. Esta capacidad para llenar el vacío de datos no solo beneficia a los encargados de formular políticas, sino que también puede empoderar a las comunidades locales, permitiéndoles tomar decisiones más informadas y proactivas en la gestión de sus tierras y recursos.

Estas predicciones tendrían un impacto crucial en Colombia. Primero, serviría para guiar la asignación precisa de recursos hacia el desarrollo rural y la erradicación de cultivos, focalizando esfuerzos en las áreas más necesitadas y proporcionando alternativas económicas viables para las comunidades rurales. Segundo, refuerzan la capacidad del gobierno para combatir el narcotráfico y el conflicto armado al dirigir los recursos de seguridad hacia las zonas identificadas como centros de producción de coca.

Asimismo, el monitoreo continuo de la evolución de los cultivos en las diferentes regiones permite evaluar la efectividad de las políticas y programas implementados y realizar ajustes según sea necesario. Por lo tanto, la precisión en la predicción de los cultivos de coca en los municipios de Colombia brinda la capacidad de adaptación y aprendizaje continuo, lo cual es fundamental para mejorar las estrategias y asegurar resultados sostenibles a largo plazo.

Referencias

Garzon, J., & LLorente, M. (2014). ¿Por qué siguen aumentando los cultivos de coca en Colombia? *Los desafios para el proximo gobierno*.

Bogotá DC: Fundacion Ideas Para la Paz. Obtenido de http://cdn. ideaspaz. org/media/website/document/5b33d29448b3b. pdf.

Díaz, A. M., & Sánchez, F. (2004). Geografía de los cultivos ilícitos y conflicto armado en Colombia (No. 2766). Universidad de los Andes, Facultad de Economía, CEDE.

Cárdenas-Torres, M. A. (2006). Estimación de la deforestación por cultivos ilícitos en la zona de reserva forestal del Río Magdalena. Colombia forestal, 9(19), 136-154.

Montes Jaramillo, E. (2016). Efectos de la presencia de cultivos ilícitos en resguardos indígenas sobre la autonomía indígena y la conservación.

Gómez, C. Y., Sastoque, T. G., & Mantilla, S. C. (2019). Los estudios sobre el fenómeno de los cultivos ilícitos de coca en Colombia: una revisión desde los enfoques de la geografía. Análisis político, 32(97), 24-44.

Datos Abiertos Colombia. (s.f.). https://www.datos.gov.co/

Departamento Nacional de Planeación. (s.f.). Terridata.

Banco de la República. (2023). Tasa Representativa del Mercado (TRM - Peso por dólar) del 1 de enero de 2023.

Ministerio de Justicia y del Derecho. (2021). Boletín sobre precios de las drogas ilícitas año 2021.

ECONOMIA

Laboratorio de inteligencia artificial aplicada a Economía

Universidad Externado de Colombia

FACULTAD DE ECONOMÍA

ECONOMIA

Laboratorio de inteligencia artificial aplicada a Economía

De Datos a Destinos:

Un Enfoque de Machine Learning para Estimar el Turismo Extranjero en Colombia

Autor

Pablo Alejandro Reyes Granados pablo.reyes4@est.uexternado.edu.co

2025

Resumen

En el contexto de la predicción de flujos turísticos, los métodos econométricos tradicionales han mostrado limitaciones para capturar dinámicas complejas y variables interrelacionadas en datos panel.

En esta investigación, se evaluaron diversas técnicas de machine learning para predecir el número de turistas extranjeros mensuales en las 81 principales ciudades de Colombia, identificando- se a XGBoost como el modelo con mejor desempeño. Además, no solo se quedó con un modelo “caja negra”, sino que se aplicó un análisis LIME para interpretar la influencia de las variables.

Los resultados no solo evidencian una mejora significativa en la precisión predictiva respecto a las técnicas econométricas clásicas, sino que también proporcionan una visión más detallada de los factores determinantes del flujo turístico.

Este estudio destaca el potencial de los enfoques basados en machine learning para optimizar la toma de decisiones en el sector turístico y aportar una nueva herramienta para planificar estrategias por parte de las ciudades colombianas.

Introducción

Es innegable que el turismo extranjero es un motor económico para todas las ciudades colombianas. Concretamente, según el Ministerio de Comercio, Industria y Turismo, en ciudades principales como Bogotá o Medellín, el turismo representa del 3 al 4.5 por ciento del PIB y se han llegado a cifras máximas del 10 por ciento; mientras que, en ciudades con distritos netamente turísticos como Cartagena o Santa Marta, la participación del turismo en el PIB no baja del 7 por ciento. Es por esto por lo que, naturalmente, surge la necesidad de que los economistas traten de estimar los factores que permiten a las ciudades atraer un flujo turístico significativo o, en su defecto, lograr una estimación estadísticamente representativa del futuro turístico de las ciudades.

Tradicionalmente, se han aplicado métodos econométricos clásicos, como las regresiones lineales, para desentrañar estos patrones y formular conclusiones sobre el fenómeno turístico. Aunque en teoría estos modelos pueden resultar adecuados, en la práctica los datos relacionados con el turismo a menudo exhiben comportamientos no lineales y complejos. Este escenario plantea un reto para las técnicas econométricas, pues no logran capturar de manera eficaz la complejidad inherente a la dinámica turística. En este contexto, el machine learning (ML) surge como una aproximación prometedora y robusta para abordar el problema de la predicción del turismo. Algoritmos avanzados, entre los que destacan los métodos de boosting, permiten modelar relaciones no lineales y detectar patrones sutiles en los datos que las metodologías convencionales podrían pasar por alto.

Esta investigación se centra en evaluar y comparar diferentes modelos de machine learning, demostrando que, más allá de lograr una mayor precisión en la predicción del número de turistas extranjeros, estos nuevos enfoques permiten superar las limitaciones de los métodos econométricos clásicos, ofreciendo herramientas más eficaces y precisas para la toma de decisiones en el ámbito del turismo.

Aproximación al problema y metodología

En esta sección se describe la aproximación adoptada para abordar el análisis del flujo de turistas extranjeros. Se especifican las fuentes de datos y el proceso de recolección utilizado. Se define detalladamente la variable dependiente y las variables independientes. Se explica cómo se construyeron las variables más significativas y se realizó su preprocesamiento. Todo ello sienta las bases metodológicas para el posterior modelado del fenómeno turístico.

Recolección de los datos

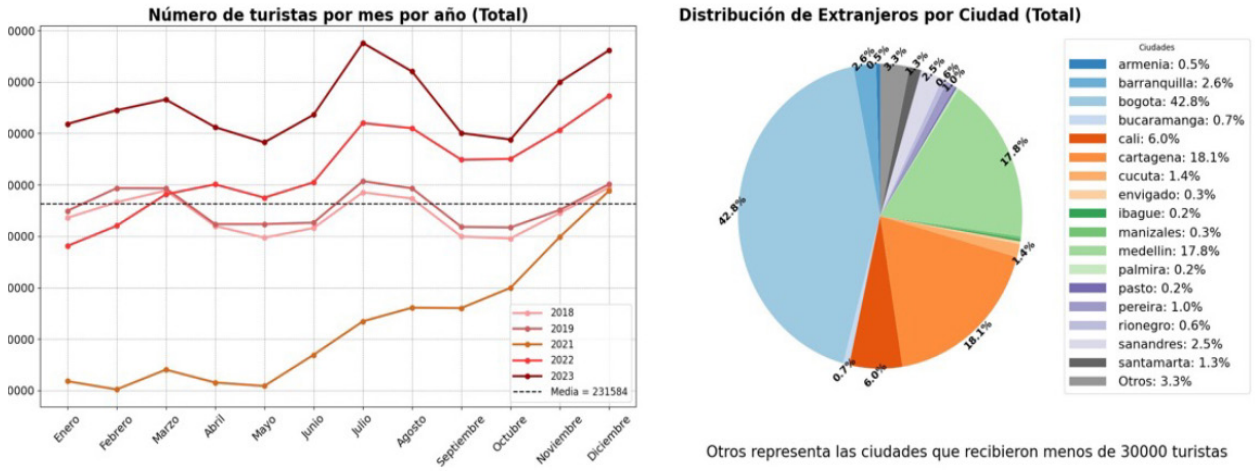
Esta es, sin duda alguna, la parte más importante de toda investigación enfocada en Machine Learning, puesto que la capacidad de predicción de un modelo depende en gran medida de la calidad de los datos empleados.

Con el fin de identificar variables capaces de predecir el turismo extranjero en Colombia, se construyó un conjunto de datos de tipo panel, en donde cada observación representa una ciudad en un mes, en el periodo de tiempo desde 2018 hasta 2023, exceptuando el año 2020, debido a las alteraciones significativas en la dinámica turística ocasionadas por la pandemia. La selección de las ciudades se realizó en función del tamaño poblacional, optando por aquellas con un mínimo de 20 mil habitantes, lo que resultó en un total de 81 ciudades a lo largo del país. Esta metodología garantiza una muestra representativa, facilitando un análisis robusto de los factores que influyen en el flujo turístico.

Variable dependiente

Considerando que nuestro modelo es de tipo regresión, la variable dependiente corresponde al destino final de todos los visitantes no residentes que llegan a Colombia por los diferentes puntos migratorios. Esta información es publicada por el Ministerio de Industria, Comercio y Turismo mensualmente. Esta variable es continua y se expresa en números absolutos de turistas por ciudad, lo cual permite comprender la magnitud de los flujos y facilita la realización de comparaciones temporales.

Gráfico 1: Scores de las variables y número de extranjeros



Fuente: Elaboración Propia

Como se puede ver en el gráfico 1, el año en el que se recibió mayor cantidad de turistas fue 2023 y el de menor afluencia, 2021, debido a los rezagos de la pandemia. Los picos de turismo se registran en junio, julio y diciembre.

Por otro lado, el promedio mensual en la serie de tiempo es de 231 mil visitantes extranjeros. En la figura de la derecha se observa que la mayoría de estos visitantes acuden a Bogotá (42,8 por ciento), seguida de Cartagena y Medellín; en general, se evidencia que el grueso del turismo se concentra en solo tres ciudades, dejando a los 78 restantes con únicamente el 20 por ciento de los visitantes.

Variables independientes

Una de las mayores limitaciones de la econometría clásica es la cantidad de supuestos que se requieren para que las estimaciones del modelo sean validas. En particular, establecer una forma funcional clara que capture las complejas relaciones entre las variables es un desafío considerable, ya que en muchos casos las interacciones en el mundo real no son lineales, polinomiales o multifacéticas. Esta primera necesidad, representa una gran restricción al análisis del turismo desde un enfoque clásico de econometría, pues muchas de las variables en este campo no se ajustan a una especificación sencilla. Por otro lado, la multicolinealidad impide trabajar con variables que muestran un alto grado de correlación entre sí, lo que representa otro gran problema, ya que se deben omitir variables que se esperaría que tengan un efecto significativo de forma combinada, solo por estar correlacionadas.

En contraste, al pasar al paradigma del Machine Learning, no se imponen supuestos preestablecidos sobre la forma funcional de las relaciones entre variables; es el propio modelo el que descubre la estructura subyacente a partir de los datos. Esto libera al investigador de tener que definir de antemano cómo se relacionan las variables, permitiendo que incluso los modelos más básicos identifiquen patrones no lineales y complejos de manera autónoma.

Teniendo esto claro podemos empezar a describir cuales son y como se construyeron las variables que se usaron en el modelo. Para esta investigación se crearon más de 31 variables independientes. Dada la gran cantidad de variables, se optó por clasificarlas en grupos temáticos amplios, lo que facilita su análisis.

Variables espaciales

En esta temática encontramos todas las variables que incorporan un componente geográfico. Entre ellas se incluyen: la distancia ponderada de cada ciudad a los diferentes puntos de acceso internacional del país, la distancia ponderada a los principales puntos turísticos, el área urbana, el área rural de cada ciudad y el número de vías principales cercanas.

Variables económicas

Aquí se incluyen las variables económicas que se consideraron más significativas para la predicción, tales como: el precio del dólar, el PIB ponderado por ciudad, un proxy de pobreza de cada ciudad y la inflación.

Variables de seguridad

Dado que la percepción de seguridad es fundamental al elegir un destino turístico, se incluyeron indicadores de criminalidad. Específicamente: el número de homicidios, hurtos y delitos sexuales de cada ciudad en cada mes.

Variables de infraestructura turística

En esta sección se agrupan variables que reflejan la capacidad y oferta turística de cada ciudad, como: el número de establecimientos de turismo de cada ciudad, el número de habitaciones, el número de camas y el número de eventos representativos que ocurren en cada mes en las ciudades.

Variables climáticas

El clima es otro factor significativo al momento de escoger un destino turístico, por lo que incluimos la temperatura promedio de las ciudades y el área de agua que tiene cada una de ellas.

Variables de de gastos turísticos

Se analizaron variables que capturan el comportamiento del gasto por viaje en cada ciudad, en base a variables como: Gasto Promedio Viaje, Gasto Aloja- miento Viaje, Gasto Transporte Viaje, Gasto Alimentos Viaje y Otros Gastos.

Descripciones variables esenciales

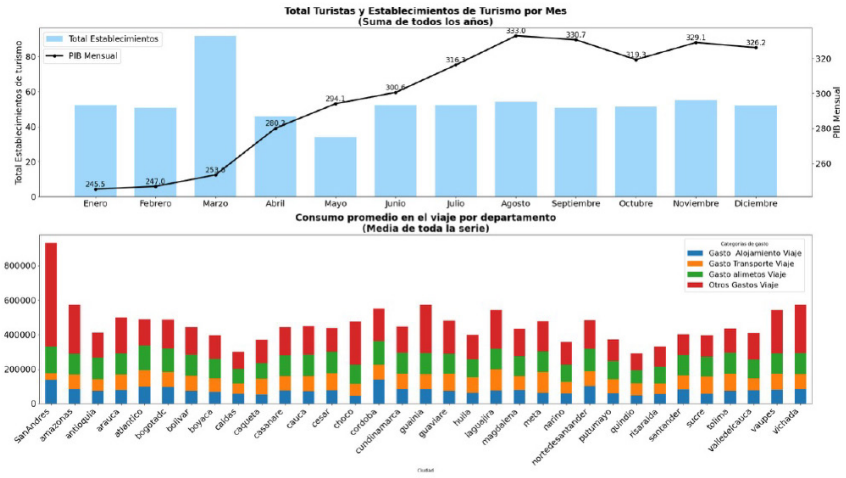
Debido a la gran cantidad de variables que se crearon para el modelo, explicar individualmente el origen y el proceso de extracción de cada una resultaría poco práctico. Por ello, nos centraremos en detallar aquellas variables que consideramos esenciales y que requirieron un mayor nivel de procesamiento o conocimiento técnico para su cálculo.

Descomposición y desagregación de series temporales mediante LOESS

Un problema que encontramos al extraer variables fue que varias presentaban una periodicidad anual, como el PIB por ciudad, el número de establecimientos de turismo, camas, habitaciones y todas las variables de gastos turísticos. Aunque una opción habría sido asignar el mismo valor a lo largo del año, confiando en que el modelo aprendería de esta constante, optamos por aplicar una descomposición LOESS a las series temporales.

La descomposición LOESS es una técnica de suavizamiento local que ajusta regresiones en ventanas móviles de la serie temporal, permitiendo capturar tendencias y patrones estacionales sin asumir una forma funcional global. El principal desafío de esta técnica radica en definir un patrón mensual adecuado para la descomposición. En nuestro caso, para las variables de PIB y costos de viaje, utilizamos la inflación mensual de cada año, asumiendo que el PIB se ve impulsado por la tendencia inflacionaria y que los costos guardan una relación directa con la misma. Por otro lado, para las variables de establecimientos se adoptó la tendencia mensual publicada por el Ministerio de Industria y Comercio. De esta forma, logramos resultados como los presentados en la figura 2, en donde podemos observar los establecimientos de turismo y una desagregación mensual de los costos del viaje por departamento.

Gráfico 2: Establecimientos de turismo por mes y costos del viaje por departamento



Fuente: Elaboración Propia

Variables de área

Las variables de área urbana, área rural y área de agua de cada ciudad se calcularon utilizando las imágenes tomadas por el satélite Sentinel-2. Para ello, se accedió a los servidores del satélite y a la serie histórica de imágenes en las ´áreas de cada ciudad desde 2018 hasta 2023. Tras aplicar un filtro de nubes, se calcularon los índices de: NDVI, GNDVI, EVI, NDWI, MNDWI, NDBI y AWEIsh, los

cuales permiten diferenciar entre zonas urbanas, rurales y acuáticas. Gracias al método Otsu, se determinaron umbrales óptimos para cuantificar en km² cada una de estas áreas y analizar su evolución a lo largo del tiempo. En la Table 1 se presentan los resultados de las áreas promedio por ciudad cabecera, mediana y pequeña.

Tabla 1: Media de las variables de área por el tipo de ciudad

Tipo Ciudad	Nmero Extranjeros	Á rea Urbana	Á rea Rural	Á rea Agua
Ciudad Pequeña	915.63	8.63	2.59	0.29
Ciudad Media	8990.02	56.23	16.08	0.77
Cabecera	74579.17	153.44	36.48	0.67

Fuente: Elaboración Propia

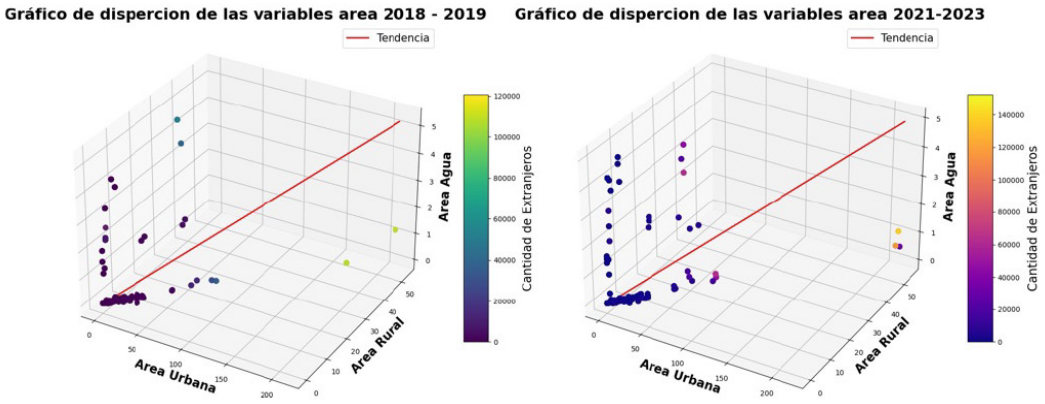
La integración de estas variables facilitó la generación de estadísticas que reflejan con precisión las distintas coberturas del suelo. Cada imagen fue procesada en lotes, garantizando la consistencia temporal y espacial en la cuantificación de las áreas, lo que permitió identificar tendencias en el crecimiento urbano y rural. Además, la validación de los resultados se realizó mediante la comparación con datos oficiales de referencia, confirmando la robustez del enfoque.

Con el análisis del Gráfico 3 y la Tabla 1, podemos concluir que existe una relación directa entre la cantidad de área y el número de extranjeros recibidos en cada ciudad, siendo las ciudades con mayor área urbana las que reciben más extranjeros a lo largo de los años; sin embargo, otro factor importante es que las ciudades con mayor área acuática, aunque sean pequeñas en comparación, reciben más extranjeros que sus pares.

Índices de distancia

Por otro lado, las variables de distancia a los puntos de acceso se ponderaron en función de la cantidad de turistas que llegan a cada uno, fundamentándose en la idea de que resulta más relevante estar cerca del Aeropuerto El Dorado que de, por ejemplo, la frontera vial Ecuador–Colombia.

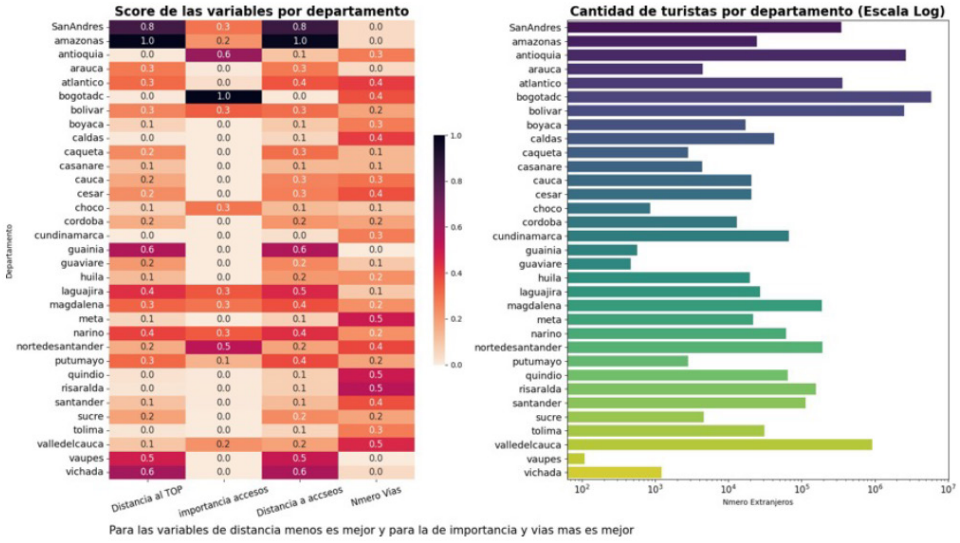
Gráfico 3: Evolución variables de área y su relación con el turismo



Elaboración Propia

Del mismo modo, la variable de distancia a los puntos turísticos se construyó considerando los 50 primeros lugares que aparecen en TripAdvisor al buscar 'Colombia' y se ponderó por el número de estrellas y de reviews de cada lugar. Finalmente, para el número de vías se utilizaron los datos de la red vial de INVIAS y se calculó la cantidad de vías principales existentes en un radio de 10 km de cada ciudad. Con estas variables obtuvimos estos resultados:

Gráfico 4: Scores de las variables y número de extranjeros



Fuente: Elaboración Propia

En el Gráfico 4, es fácil ver que existe una relación entre la distancia de una ciudad respecto a los accesos y lugares reconocidos y el flujo de turistas. Los departamentos más alejados son los que reciben menos turismo. Mientras tanto, ciudades como Bogotá, que cuentan con la menor distancia a lugares turísticos y a las entradas internacionales, reciben el mayor flujo de turistas del país.

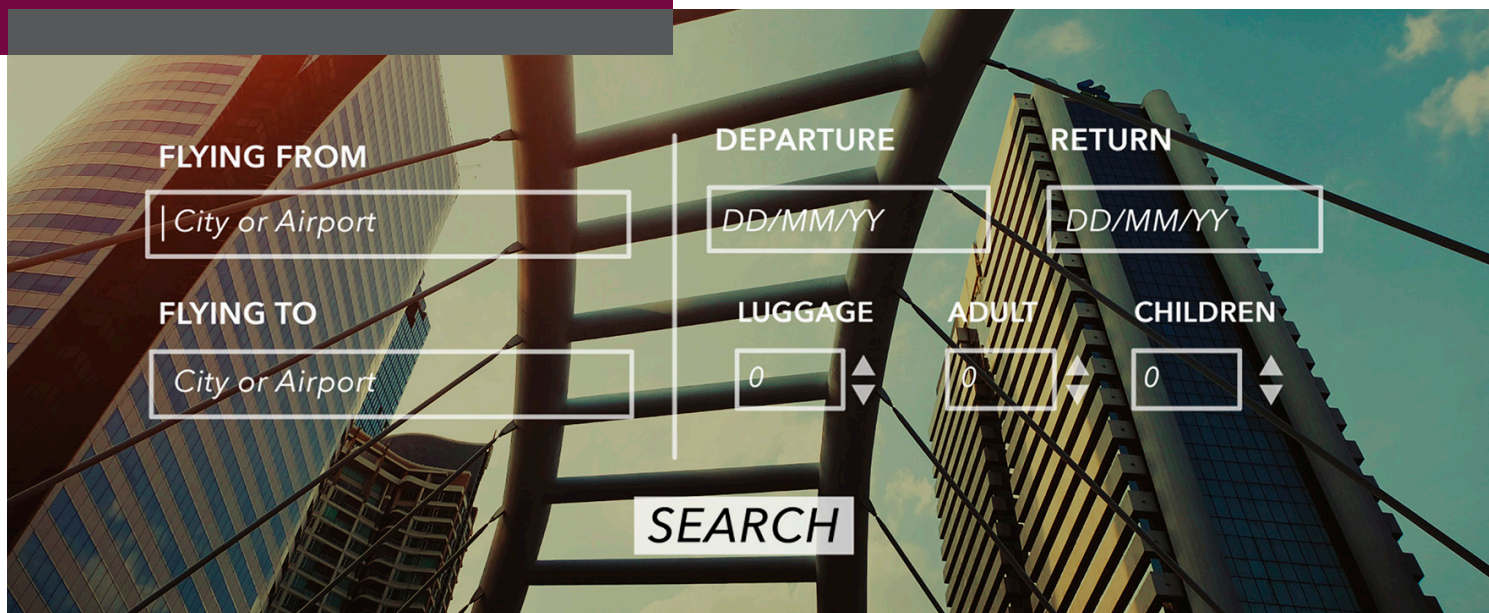
Preprocesamiento de los datos

Una vez entendidas las variables en bruto, es pertinente detallar el preprocesamiento aplicado a los datos antes de ser introducidos en el modelo. En esta etapa, el manejo de datos faltantes resultó crucial, ya que identificamos ausencias en las variables de temperatura, establecimientos de turismo y costos. En ciencia de datos, existen diversas técnicas para tratar datos faltantes, siendo una de las más valoradas la imputación mediante K-Nearest-Neighbor; sin embargo, otro método de ML, que no es tan conocido, es el Kriging. Mientras que KNN identifica a los K vecinos más cercanos basándose en similitudes entre características, el Kriging realiza una interpolación espacial, lo que lo hace especialmente útil para variables con una relación geoespacial. Por ello, se optó por emplear Kriging en las variables de temperatura y costos, y KNN para la variable de establecimientos de turismo.

Por último, se realizó un debido análisis de valores atípicos y se normalizaron los datos para que todos los modelos trabajen correctamente con los datos.

Modelaje y evaluación de desempeño

Al momento de modelar, se optó por realizar un amplio ejercicio comparativo entre diversos modelos de Machine Learning, empezando desde los modelos más básicos (Regresiones Lineales Regularizadas), hasta llegar a métodos de ensamble (Random Forest), métodos de boosting (XGBoost, GradientBoos- ting y CatBoost) y terminar con SVM. En total se corrieron 12 modelos diferentes y, para cada uno, se hizo una búsqueda exhaustiva (grid search) de hiperparámetros, optimizando su desempeño. Con esta metodología se logró tener un amplio abanico de opciones, lo que permitió seleccionar no solo el modelo con mejores métricas de desempeño, sino también aquel que minimizara el uso de recursos computacionales. Para validar los modelos, se dividió la muestra en conjuntos de entrenamiento (70 por ciento), validación (20 por ciento) y prueba (10 por ciento), seleccionándose aquel que maximizó el coeficiente R2 en entrenamiento y validación, con el objetivo de evitar un modelo sobreajustado a los datos de entrenamiento.



Métricas de desempeño de los modelos

Tabla 2: Métricas de desempeño de los modelos en el conjunto de testeo

	R^2	MSE	MAE	$RMSE$
Regresión Lineal	0.7326	0.376	0.356	0.519
Regresión Lasso	0.8272	0.230	0.374	0.412
Regresión Ridge	0.7910	0.213	0.291	0.491
Elastic Net	0.9337	0.017	0.059	0.132
KNN	0.9221	0.023	0.011	0.105
Arboles de Decisión	0.9344	0.051	0.046	0.226
Random Forest	0.9518	0.029	0.021	0.143
GradientBoosting	0.9533	0.032	0.026	0.161
XGBoost	**0.9816**	**0.037**	**0.015**	**0.122**
CatBoost	0.9412	0.091	0.100	0.197
SVM	0.9001	0.162	0.141	0.212
SVM con Kernel	0.9310	0.043	0.091	0.161

En la Tabla 2 se pueden evidenciar las métricas de desempeño de todos los modelos que se corrieron, posteriormente a realizar la optimización de hiperparámetros. Se observa que los métodos basados en técnicas de boosting y ensemble, especialmente XGBoost, destacan por alcanzar un R2 casi perfecto (0.9816) y errores bajos en MSE, MAE y RMSE. En contraste, modelos tradicionales como la regresión lineal presentan un rendimiento inferior, lo que respalda la ventaja de los enfoques modernos para capturar la complejidad de los datos turísticos.

Aunque modelos como SVM superan significativamente a los enfoques de regresión clásica en términos de precisión, es importante destacar que su tiempo de entrenamiento fue de aproximadamente 4 horas, lo que limita su implementación en un escenario real. En contraste, XGBoost no solo maximizó las métricas de desempeño, sino que completó su entrenamiento en menos de un minuto. Esta combinación de alta precisión y eficiencia fue la principal razón para seleccionarlo como el modelo óptimo.

Comparación econométrica

Aunque en la sección anterior se observó que las regresiones lineales alcanzaban desempeños muy inferiores a los modelos de Machine Learning, cabe señalar que dichas regresiones no son “puramente econométricas”, ya que se emplearon variables altamente correlacionadas. Esto se traduce en coeficientes beta con mayor varianza, sesgados y con dificultades para aislar los efectos independientes de cada variable, lo que complica la interpretación precisa de sus resultados. Por ello, se desarrollaron dos regresiones que cumplen con los estándares econométricos, con el fin de contrastarlas de manera más adecuada con los modelos de ML. La primera regresión que se planea estimar aprovecha el factor del tiempo del panel, y su fórmula se presenta en la ecuación (1):

$$NExtranjerosit = \alpha i + \beta 0 + \beta 1 \cdot NExtranjerosi,t-1 + \beta 2 \cdot NExtranjerosi,t-2 + \beta 3 \cdot PIBi,t-1 + \epsilon it$$

Al estimar esta regresión se obtuvieron estos resultados:

Tabla 3 Coeficientes de la regresión

Variable	Coeficiente	Error Est.	*t*-valor	*p*-valor
Constante	2257.5641	202.793	11.132	0.000
Nmero Extranjeros$_{t-1}$	-0.1539	0.032	-4.877	0.000
Nmero Extranjeros$_{t-2}$	0.1548	0.014	10.941	0.000
Pib Ponderado$_{t-1}$	0.6657	0.135	4.919	0.000

Fuente: Elaboración propia

Como se ve en la Tabla 3, aunque todos los coeficientes beta son estadísticamente significativos, tanto el R2 como el R2 Ajustado no superan 0.1; lo cual representa un modelo peor que el azar en términos predictivos y una reducción del 90 por ciento con respecto a XGBoost. Por ello, esta aproximación econométrica resulta en un fracaso total.

Alejándonos de los datos netamente temporales y estimando una regresión para datos panel con efectos aleatorios, la cual podemos describir con la ecuación (2):

$$NExtranjerosit = \beta 0 + \beta 1 \cdot Homicidiosit + \beta 2 \cdot Gasto\ Promedio\ Viajeit + \beta 3 \cdot Establecimientos\ de\ Turismoit$$

$$+ \beta 4 \cdot D´olarit + \beta 5 \cdot Temperaturait + \beta 6 \cdot PIBit + \beta 7 \cdot Eventosit + ui + \epsilon it \quad (2)$$

Este enfoque permite aprovechar la variabilidad tanto entre las diferentes ciudades (Efectos “bet- ween”) como a lo largo del tiempo (Efectos “within”), lo que se traduce en estimaciones más robustas y en la reducción del sesgo por heterogeneidad. Los resultados se presentan en la Tabla 4.

Tabla 4: Coeficientes de la regresión Random Effects

Variable	Coeficiente	Error Est.	*t*-valor	*p*-valor
Constante	1861.6000	699.2300	2.6623	0.0078
Homicidios	3.6310	12.1120	0.2998	0.7644
Gasto Promedio Viaje	-0.0001	0.0007	-0.2092	0.8343
Establecimientos de turismo	10.3510	0.5159	20.0640	0.0000
Dólar	-0.6097	0.1362	-4.4764	0.0000
Temperatura	15.5170	21.1510	0.7336	0.4632
Pib Ponderado	2.4648	0.0677	36.4200	0.0000
Eventos	199.2500	54.1680	3.6784	0.0002

Fuente: Elaboración propia

La regresión de Efectos Aleatorios representa una mejora notable respecto al modelo previo, elevando el R2 de 0.08 a 0.73. Este incremento en la capacidad explicativa se debe a que el enfoque de datos panel captura tanto las variaciones entre ciudades como las variaciones a lo largo del tiempo. Se observa que variables como Establecimientos de Turismo, Dólar, PIB Ponderado y Eventos resultan altamente significativas, subrayando su impacto en el flujo de turistas extranjeros. Por otro lado, la disparidad entre el R2 "Within"(0.2718) y el "Between"(0.8091) indica que el modelo explica de manera más efectiva las diferencias estructurales entre ciudades que las fluctuaciones internas a lo largo del tiempo.

No obstante, es importante recordar que XGBoost logró un R2 de 0.98, lo que enfatiza la clara diferencia en desempeño entre una regresión econométrica bien planteada y un enfoque de Machine Learning capaz de capturar patrones complejos con mayor eficacia.

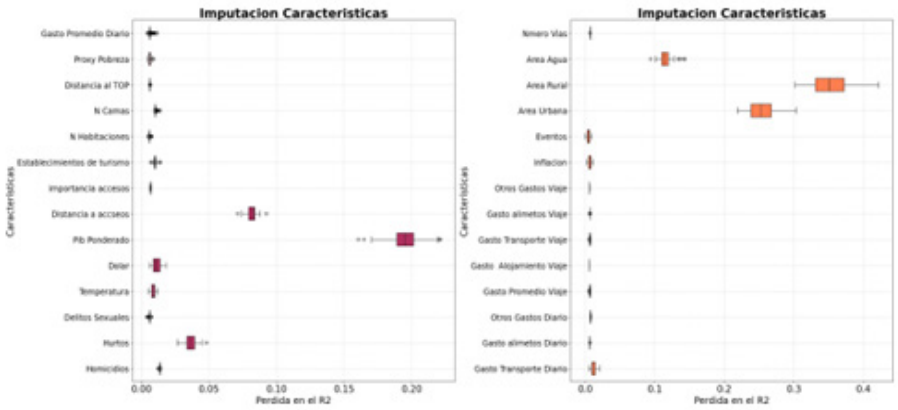
Resultados de *XGBoost*

Una de las principales ventajas de los modelos econométricos es la facilidad para interpretarlos, lo que permite establecer relaciones causales y dependencias significativas entre las variables. En con- traste, cuando se utilizan modelos avanzados de Machine Learning, la interpretación directa de los parámetros se vuelve complicada debido a la naturaleza de "caja negra" de estos métodos. Sin embargo, aunque los parámetros en sí carezcan de una interpretación directa, existen diversas técnicas que permiten identificar dependencias y relaciones entre las variables predictoras y la variable objetivo. Entre las técnicas más famosas y usadas se encuentran la imputación de características, LIME y Partial Dependence Plots (PDP).

Imputación de características

La imputación de características es una técnica que evalúa el impacto de cada variable en el desempeño del modelo. Al modificar o sustituir sistemáticamente los valores de una característica y observar los cambios en las métricas del modelo, se pueden identificar cuáles son las variables más significativas para el modelo.

Gráfico 5: Imputación de características XGBoost.



Fuente: Elaboración Propia

El Gráfico 5 revela que solo tres variables: Área Urbana, Área Rural y Área de Agua son cruciales para el modelo, ya que su eliminación implicaría una pérdida del 12 al 37 por ciento de la capacidad para predecir la variable dependiente, siendo el Área Rural la más influyente.

Por otro lado, podemos ver que el PIB y la distancia a los accesos internacionales también resultan importantes, ya que su omisión hace que el modelo pierda entre el 9 y el 20 por ciento de su capacidad predictiva.

Por último, otros indicadores, como los establecimientos de turismo o los costos de viaje, tienen un impacto marginal, reduciendo solo alrededor del 1 por ciento. Esta gráfica evidencia claramente que XGBoost se apoya principalmente en las variables relacionadas con las áreas, el PIB y las distancias para identificar patrones en el flujo turístico, y son estas variables las que, según el modelo, guardan una relación más intensa con el flujo turístico de cada ciudad.

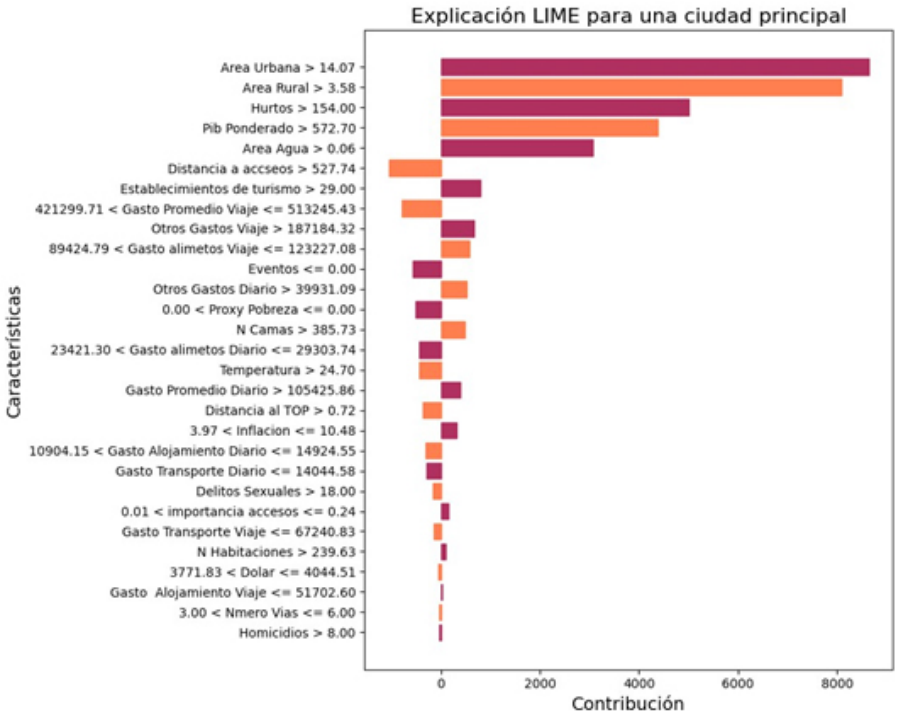
LIME

LIME (Local Interpretable Model-Agnostic Explanations) es una técnica que genera interpretaciones locales de las predicciones de modelos complejos. Se basa en aproximar el comportamiento del modelo en un vecindario alrededor de una instancia específica mediante un

modelo de regresión lineal, lo que permite identificar la contribución individual de cada variable. De esta forma, LIME facilita la comprensión y validación de las decisiones del modelo, ofreciendo una visión similar a la interpretación de una regresión, pero aplicada de manera local.

Dado que en LIME se debe especificar en base a cuál registro (ciudad) se hará la interpretación, vamos a realizar dos análisis LIME: uno en una de las grandes ciudades, observando cuáles son los efectos marginales (betas) que resultan, y otro en una ciudad pequeña. El propósito de esto es identificar cuáles son las variables que mías influencian el turismo en las grandes y pequeñas ciudades, o si se trata de las mismas.

Gráfica 6 Lime para una ciudad principal



Fuente: elaboración propia

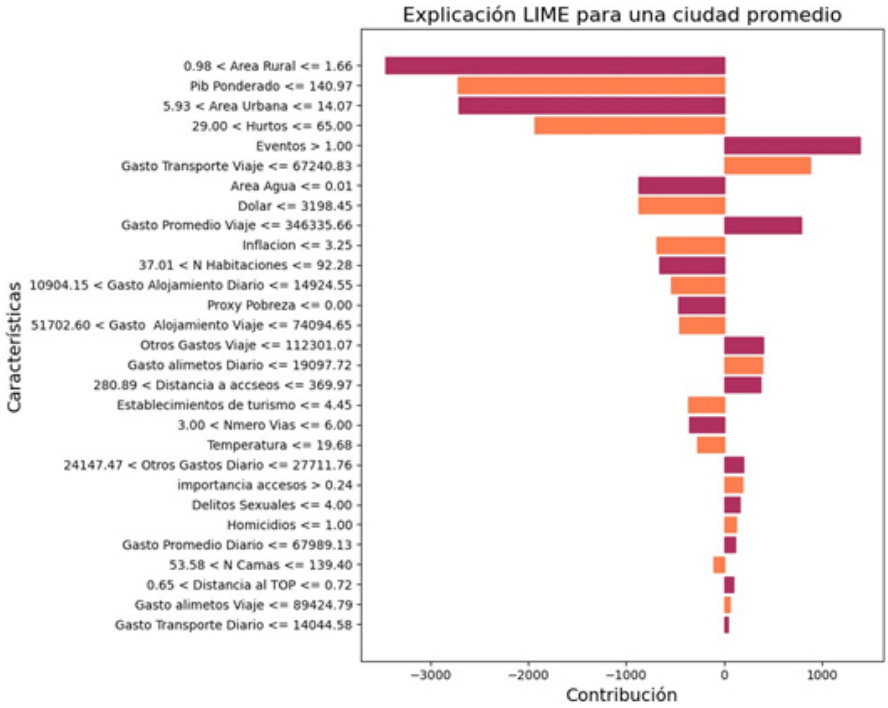
Según el Gráfico 6 y utilizando LIME, se observa que, para las grandes ciudades, contar con un área mayor a 14 km² y un área rural mayor a 4 km² aumenta el número de turistas en más de 8.000, lo que deja claro que estas son las características más significativas para una ciudad grande. Además, los hurtos superiores a 155 también tienen una influencia positiva en la predicción; esto se debe a que esta variable está funcionando como un proxy del tamaño o la actividad económica, pero no establece una causalidad directa.

Asimismo, un PIB mayor a 572 billones por mes y más de 30 establecimientos de turismo aumentan el flujo de turistas entre 2.000 y 4.000, lo que refuerza la idea de que el desarrollo económico es crucial para sostener grandes flujos turísticos, evidenciando como factores macroeconómicos y de infraestructura se combinan para atraer visitantes.

Por otro lado, LIME permite destacar que otras características, como el área de agua, logran atraer un gran flujo de turistas. En contraste, contar con altos niveles de pobreza, altos gastos promedio del viaje, una baja relevancia de puntos de acceso internacionales (como aeropuertos o fronteras) y tener un clima frio reducen significativamente

el número de turistas. Finalmente, las variables que menos destacan en las ciudades grandes son: la inflación, los delitos sexuales, el gasto en transporte y el número de vías cercanas. Una posible explicación es que, en las ciudades principales, estos factores se ven opacados por características de mayor impacto mediático y estructural, como la infraestructura turística, la actividad económica y la misma popularidad de la ciudad.

Gráfico 7 Lime para una ciudad promedio



Fuente: elaboración propia

En la interpretación de LIME para una ciudad promedio, se evidencia una dinámica distinta en comparación con las grandes urbes. En este contexto, el tamaño del área urbana, rural y de agua, especialmente cuando se sitúa entre 0.98 y 14.16 km², es el factor que más contribuye negativamente a la predicción del número de turistas, lo que indica que las ciudades con áreas más reducidas tienen una capacidad limitada para atraer flujos turísticos significativos. Sumado a esto, variables como un PIB bajo, grandes distancias a los principales puntos de acceso internacionales y el precio del dólar reducen aún más la capacidad de la ciudad para atraer turistas.

Sin embargo, la cantidad de eventos en la ciudad emerge como el principal impulsor positivo, subrayando que en estas ciudades las actividades culturales y recreativas pueden compensar, en parte, la carencia de una infraestructura de gran escala. También se puede ver que unos costos de viaje reducidos potencian de gran manera el flujo de turistas; aunque, entre estos, los costos de transporte son los más importantes, lo que indica que los turistas prefieren viajar a ciudades donde les es más barato llegar.

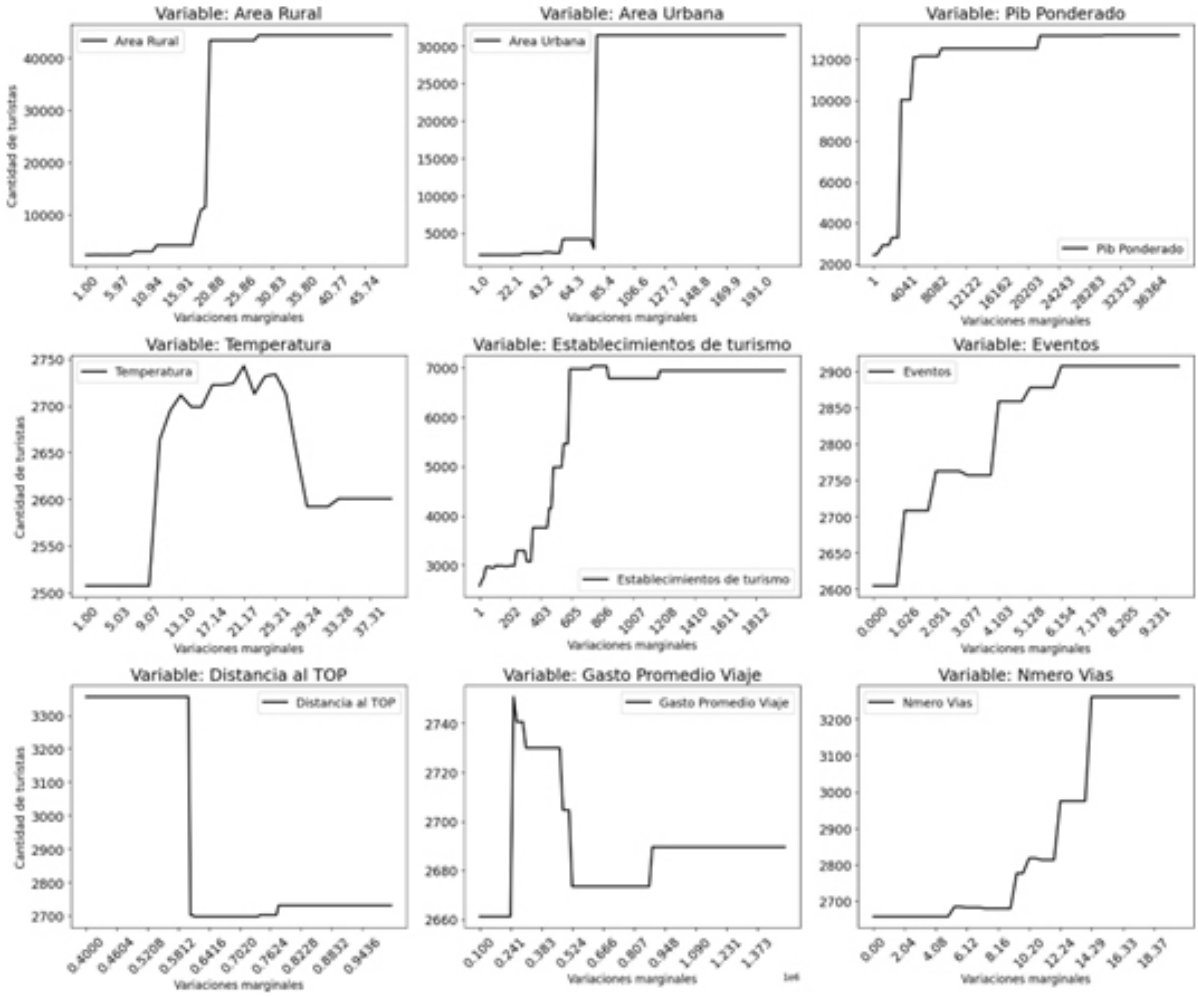
Por último, se puede ver que todas las variables de infraestructura aportan negativamente; esto refuerza la importancia de invertir en infraestructura turística local. En conclusión, en grandes ciudades como Bogotá, el tamaño del área urbana es el

principal determinante del turismo, apoyado por una robusta oferta de infraestructura y servicios; en cambio, en las ciudades de menor escala, la realización de eventos atractivos se presenta como la estrategia clave para contrarrestar las limitaciones inherentes a un área urbana reducida.

Partial Dependence Plots

Las Partial Dependence Plots son gráficos interpretativos que muestran el efecto marginal de una variable sobre la predicción del modelo. Estos gráficos permiten visualizar de manera clara cómo varían las predicciones al modificar el valor de una característica, ayudando a identificar patrones no lineales y relaciones de interacción.

Figura 14. Entrenamiento de Árbol de Decisión



Fuente: elaboración propia.

En el Gráfico 8 se presentan las Partial Dependence Plots (PDP) para nueve variables que, a lo largo de la investigación, se han identificado como las más influyentes en la predicción del flujo turístico. En la primera fila se evidencia que el flujo de turistas tiene un crecimiento exponencial cuando el área urbana supera los 20 km² y el área rural supera los 60 km², incrementándose la cantidad de turistas esperados en un rango de 30 a 40 mil. En complemento, el PIB presenta un aumento de 12 mil turistas cuando se superan los 4041 millones de pesos.

En la segunda fila, se destaca que las temperaturas extremas (muy frías o altas) reducen el potencial turístico, siendo el rango óptimo entre 17 y 25 °C. Sin embargo, el efecto positivo de aumentar el número de establecimientos de turismo y la realización de eventos culturales es notable y termina de validar los resultados de LIME, ya que contar con más de 600 establecimientos incrementa el flujo en aproximadamente 7 mil turistas, y la celebración de más de tres eventos aporta un aumento de 2800 turistas.

Por último, la tercera fila enfatiza que tanto una mayor distancia a los puntos de interés más reconocidos como unos costos turísticos elevados en la ciudad tienen efectos negativos, mientras que disponer de un mayor número de vías principales de acceso favorece la llegada de turistas extranjeros. Estos resultados subrayan la importancia de la infraestructura, la actividad cultural y la accesibilidad para impulsar el turismo en las diferentes ciudades.

CONCLUSIONES

A lo largo de la investigación, hemos evidenciado que los métodos de Machine Learning ofrecen ventajas significativas sobre los enfoques econométricos tradicionales en la predicción de fenómenos complejos, como el turismo. Mientras que en las regresiones econométricas se debe estipular una forma funcional clara para las variables y cumplir con diversos supuestos —por ejemplo, la linealidad de los parámetros y la ausencia de multicolinealidad—, los modelos de ML, como XGBoost, capturan relaciones no lineales y patrones complejos sin estar limitados a una representación en forma de ecuación, lo que les permite lograr precisiones predictivas excepcionales.

Por otro lado, la investigación evidencia que el turismo en Colombia se ve determinado por una compleja interacción de factores, cuyos efectos varían según el tamaño y las características de cada ciudad. En grandes urbes como Bogotá, un área urbana extensa (mayor a 70 km²) y un PIB elevado (superior a 9000 mil millones anuales) se consolidan como los principales impulsores del flujo turístico; estas dos variables no son más que un complemento de una robusta infraestructura y una amplia oferta de servicios. No obstante, estos beneficios se ven contrarrestados por barreras como altos costos de viaje, elevados gastos en alimentación, deficiencias en la conectividad internacional y altas tasas de crímenes y pobreza, que limitan el potencial turístico pese a las ventajas estructurales.

En contraste, en ciudades pequeñas y promedio, donde las limitaciones de infraestructura y tamaño son evidentes, los eventos locales y el crecimiento en el número de establecimientos turísticos emergen como estrategias clave para estimular el turismo. Concretamente, la realización de más de dos eventos culturales y el superar umbrales críticos en el número de establecimientos permite compensar la falta de área construida, mientras que la conectividad mediante vías principales y la proximidad a puntos de interés facilitan la llegada de visitantes. Factores transversales como la diversidad territorial, las

condiciones climáticas moderadas y la seguridad también juegan un papel crucial en la atracción turística.

Por último, es fundamental implementar estrategias diferenciadas que potencien las fortalezas de cada región. En las grandes ciudades se debe mejorar la conectividad internacional y reducir las barreras de costos, mientras que en las ciudades más pequeñas es prioritario promover eventos culturales y desarrollar la infraestructura turística local. Los análisis de LIME y las Partial Dependence Plots revelan que, en las urbes de mayor escala, ampliar el área urbana y optimizar la infraestructura de transporte, así como mejorar la percepción de seguridad general, contribuyen significativamente a incrementar el flujo de turistas; por ello, se recomienda invertir en la modernización de los servicios de movilidad, en proyectos de infraestructura urbana correctamente planificada y en programas de justicia y seguridad.

En contraste, para las ciudades pequeñas, donde la limitación de espacio es un desafío, la realización de eventos culturales y el fomento de la identidad local emergen como estrategias clave para compensar la menor infraestructura, de modo que el impulso de festivales, ferias y actividades recreativas, junto con alianzas estratégicas con operadores turísticos locales, puede transformar estas limitaciones en ventajas competitivas y atraer a un público extranjero diverso.

Referencias

Banco Mundial. (2019). *Informe sobre el desarrollo mundial 2019: La naturaleza cambiante del trabajo*. Banco Mundial.

Behr, A., Giese, M., Teguim Kamdjou, H. D., & Theune, K. (2020). *Dropping out of university: a literature review*. Review of Education, 8(2), 614-652. https://doi.org/10.1002/rev3.3202

Burzacchi, A., Rossi, L., Agasisti, T., Paganoni, A. M., & Vantini, S. (2024). *Urban mobility and learning: Analyzing the influence of commuting time on students' GPA at Politecnico di Milano. Studies in Higher Education*, 1–26. https://doi.org/10.1080/03075079.2024.2374005

Chalela-Naffah, S., Valencia-Arias, A., Ruiz-Rojas, G. A., & Cadavid-Orrego, M. (2020). Factores psicosociales y familiares que influyen en la deserción en estudiantes universitarios en el contexto de los países en desarrollo. *Revista Lasallista de Investigación*, 17(1), 103-115. https://doi.org/10.22507/rli.v17n1a9

Chen, T., & Guestrin, C. (2016). XGBoost: A scalable tree boosting system. Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 785-794.

Hastie, T., Tibshirani, R., & Friedman, J. (2009). The elements of statistical learning: Data mining, inference, and prediction (2nd ed.). Springer. https://doi.org/10.1007/978-0-387-84858-7 https://doi.org/10.1145/2939672.2939785

Lorenzo-Quiles, O., Galdón-López, S., & Lendínez-Turón, A. (2023). *Factors contributing to university dropout: a review. Frontiers in Education*, 8, Article 1159864. https://doi.org/10.3389/feduc.2023.1159864

McKinney, W. (2017). Python for Data Analysis: Data wrangling with Pandas, NumPy, and Jupyter (2nd ed.). O'Reilly Media.

Núñez-Naranjo, A. F. (2024). Analysis of the determinant factors in university dropout: a case study of Ecuador. *Frontiers in Education*, 9, Article 1444534. https://doi.org/10.3389/feduc.2024.1444534

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., ... & Duchesnay, E. (2011). Scikit-learn: Machine learning in Python. Journal of Machine Learning Research, 12, 2825-2830. https://jmlr.org/papers/v12/pedregosa11a.html

Valencia-Arias, A., Chalela, S., Cadavid-Orrego, M., Gallegos, A., Benjumea-Arias, M., & Rodríguez-Salazar, D. Y. (2023). University Dropout Model for Developing Countries: A Colombian Context Approach. Behavioral Sciences, 13(5), 382. https://doi.org/10.3390/bs13050382

Universidad Externado de Colombia

FACULTAD DE ECONOMÍA

ECONOMIA

Laboratorio de inteligencia artificial aplicada a Economía

Inclusión Financiera en Zonas Rurales por medio Machine Learning

Autores

Daniela Andrea Orduz Macías daniela.orduz1@est.uexternado.edu.co nicol.
Nicoll Reina Moreno reina@est.uexternado.edu.co
Oscar Fabian Rodríguez Sarmiento oscar.rodriguez08@est.uexternado.edu.co

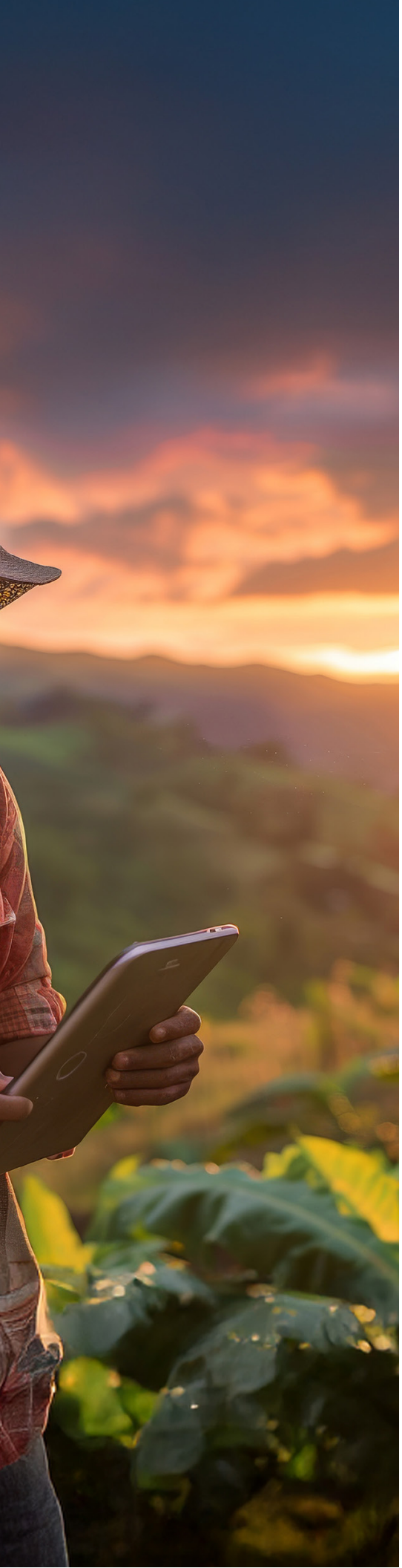
2025

Resumen

Históricamente la población rural colombiana se ha enfrentado a profundas desigualdades en diferentes aspectos, esto se puede evidenciar claramente en la incidencia de pobreza monetaria y pobreza monetaria extrema donde son los centros poblados y rurales dispersos los que poseen un porcentaje de incidencia mayor; para 2023 el 30.6% de la población de las cabeceras se enfrentaba a pobreza monetaria frente a un 41.2% en las zonas rurales (DNP, 2024). De esta manera es necesario que además de los diferentes servicios básicos como la educación, salud, vivienda, entre otros, estos individuos sean soportados por la existencia de un respaldo financiero robusto que les permita acceder a mayores oportunidades crediticias.

Ante esto, el presente documento presenta la implementación de una aplicación de análisis de comportamiento financiero de los campesinos con el fin de diseñar instrumentos financieros personalizados a las necesidades de cada campesino según su enfoque laboral (agrícola, ganadero, apicultor, entre otros), sus características personales y las características de su entorno. Esto con el objetivo de ofrecer mayor acceso a créditos e inversiones según las necesidades de los usuarios reduciendo la burocracia y trabas a las que suelen enfrentarse por ser una población con un escaso historial crediticio y poca educación financiera.

Palabras Clave: Redes Neuronales, IBF, desigualdad, Clusterización.



Introducción

La inclusión financiera es un concepto amplio el cual puede cambiar dependiendo el contexto y la población objetivo, en las grandes ciudades los esfuerzos se centran en la generación de créditos o el ofrecer servicios financieros digitales como Nequi o Daviplata, los cuales funcionan como billeteras virtuales incluso otorgando créditos de bajo monto también llamados nano-créditos.

Para la población agrícola es difícil acogerse a este concepto de inclusión por la falta de información del sistema financiero sobre estas poblaciones, que dificulta y/o encarece el acceso al crédito, por lo que se usan servicios de crédito informal 'Gota a Gota'. Por otro lado, en el caso de la población campesina de menores ingresos tiene un problema con la naturaleza de su negocio, dado que es altamente volátil a cambios climáticos o políticos, como fue el resultado de la guerra en Ucrania, entre otros. Esto genera incertidumbre en los ingresos de la población y deriva de nuevo en el uso de servicios de crédito informal.

Por lo tanto, se considerará como inclusión financiera el permitir el acceso al crédito para la población rural con el fin de generar resiliencia financiera apoyando así el crecimiento de los negocios, reduciendo la incertidumbre y abaratando el acceso al crédito para la población.

Como uno de los problemas para la generación de estos créditos es la falta de información comparada con poblaciones de grandes ciudades, se buscará construir una aplicación que genere predicciones sobre potenciales usuarios de estos créditos agrícolas usando modelos de redes neuronales e inteligencia artificial.

Se considera el modelo de redes neuronales útil para este fin dada su construcción en capas la cual, permite que se generen patrones con poca información y así se genere un proceso de entrenamiento donde el modelo pueda pronosticar e identificar con mayor eficiencia el comportamiento de los usuarios.

Análisis y metodología IA

Para la construcción del modelo, se recopilaron datos de diversas fuentes oficiales como el DANE, FINAGRO y Banca de Oportunidades, complementados con datos sintéticos generados mediante técnicas de simulación. Las principales variables utilizadas incluyen datos demográficos como edad, nivel de escolaridad y ubicación geográfica; datos económicos como ingreso mensual, gastos en producción y mantenimiento; acceso financiero medido a través de la cantidad de cuentas de ahorro, corresponsales bancarios y distancia a puntos financieros; resiliencia financiera reflejada en la existencia de fondos de emergencia, diversificación de ingresos y participación en cooperativas; y progreso financiero evaluado mediante la inversión en activos, crecimiento de producción y uso de tecnología. Una vez obtenidos los datos, se realizaron procesos de normalización utilizando la técnica de MinMaxScaler para estandarizar las variables numéricas. Además, las variables categóricas fueron transformadas a valores binarios para su uso en los modelos de IA.

El bienestar financiero de cada campesino fue cuantificado mediante el desarrollo de un Índice de Bienestar Financiero (IBF), calculado a partir de cuatro dimensiones clave. La seguridad financiera se evaluó a través de la relación entre ingresos, disponibilidad de efectivo y gastos. La resiliencia financiera se determinó con base en la capacidad del campesino para enfrentar imprevistos económicos. El progreso financiero se midió mediante indicadores de crecimiento económico y adopción de tecnología, mientras que el acceso financiero se analizó considerando la disponibilidad de servicios bancarios y la capacidad de uso. Cada una de estas dimensiones fue calculada mediante combinaciones ponderadas de variables clave, proporcionando una representación integral de la situación financiera del usuario.

Para la clasificación y análisis del bienestar financiero, se utilizaron dos enfoques principales. Primero, se aplicó el algoritmo K-Means Clustering para segmentar a los campesinos según patrones de acceso y uso financiero. Se determinaron varios clústeres que agrupan a los usuarios en función de su nivel de inclusión financiera, permitiendo identificar perfiles de alto, medio y bajo acceso a servicios financieros. Posteriormente, se empleó un modelo Random Forest Classifier para evaluar la relación entre las variables socioeconómicas y el bienestar financiero. Este modelo fue entrenado con la base de datos generada y logró una precisión significativa en la identificación de los factores más influyentes en la estabilidad financiera de los campesinos. Entre estos factores, se destacan la educación financiera, la diversificación de ingresos y la disponibilidad de efectivo mensual como elementos clave en la predicción del bienestar económico.

Adicionalmente, se desarrolló un modelo de predicción de riesgo crediticio, diseñado para estimar la capacidad de pago y el nivel de riesgo de cada solicitante. Utilizando técnicas de aprendizaje supervisado, este modelo clasifica a los campesinos en diferentes niveles de riesgo con base en su historial financiero simulado y su clasificación dentro del IBF. Los resultados de este análisis permiten al neobanco determinar no solo la elegibilidad del usuario para un crédito, sino también el monto óptimo que se le puede otorgar sin comprometer su estabilidad financiera.

Los modelos desarrollados serán integrados en la aplicación del neobanco, permitiendo evaluar solicitudes de crédito de manera automatizada y eficiente. A través de la app, los campesinos podrán ingresar sus datos financieros y recibir una evaluación en tiempo real basada en el IBF. La aplicación recomendará montos de financiamiento personalizados y estrategias de ahorro adaptadas a la situación particular de cada usuario. Además, el neobanco podrá utilizar los modelos de segmentación para ofrecer productos financieros específicos según el perfil

del usuario, como seguros agrícolas, cuentas de ahorro incentivadas y líneas de crédito con tasas diferenciadas. Con esta infraestructura basada en IA, se busca facilitar el acceso a créditos justos y promover el crecimiento económico sostenible de la población campesina colombiana.

La evaluación del desempeño del modelo de predicción de riesgo crediticio se realizó utilizando métricas como precisión, recall y F1-score, asegurando su robustez mediante técnicas de validación cruzada y división en conjuntos de entrenamiento y prueba. La selección de los modelos de IA responde a la necesidad de contar con herramientas explicables y eficientes. K-Means fue elegido por su capacidad para segmentar grandes volúmenes de datos sin necesidad de etiquetas previas, permitiendo la identificación de perfiles financieros diferenciados. Random Forest se utilizó debido a su alta precisión y capacidad para manejar relaciones no lineales entre las variables, superando métodos tradicionales como la regresión logística, que tienden a ser menos efectivos en la captura de patrones complejos. La aplicación del neobanco integrará estos modelos para evaluar solicitudes de crédito en tiempo real, ajustando tasas de interés y montos según el nivel de riesgo determinado por el IBF. Además, se contempla la inclusión de programas de educación financiera dentro de la plataforma para mejorar la gestión de recursos de los campesinos.

Sin embargo, la metodología presenta ciertas limitaciones, principalmente en el uso de datos sintéticos que, aunque alineados con información real, requieren validación con datos bancarios efectivos. Para mejorar la precisión del sistema, en futuras versiones se podrían integrar datos provenientes de entidades financieras y explorar modelos más avanzados como XGBoost o redes neuronales, permitiendo una evaluación de riesgo aún más refinada y ajustada a las particularidades del sector agrícola colombiano.

Diseño de la aplicación

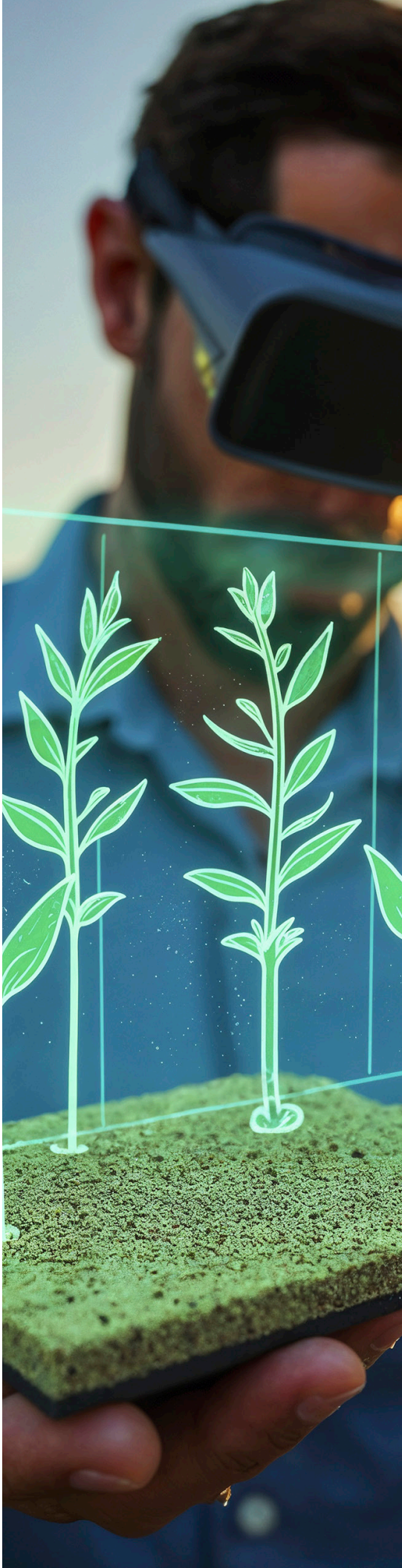
Una vez establecido el objetivo y el funcionamiento de la aplicación de inclusión financiera se ve necesaria la implementación de una interfaz sencilla de manejar por lo cual el usuario solo debe ingresar sus datos y diferentes transacciones, ingresos y deudas que tenga con el fin de que esta información sea recibida por una API (Application Programming Interface) interna de Backend donde nuestro modelo no solo interpretará la información sino que posteriormente enviará la información y los resultados a un BlockChain guardando la información relevante de las transacciones de los usuarios de manera inmutable y permitiéndolos desarrollar un historial financiero cada vez más robusto. Este proceso puede ser igualmente desarrollado por medio de machine learning pues por medio del aprendizaje del programa a la hora de identificar las transacciones y recomendar un instrumento financiero, es posible que estas mismas recomendaciones sean clasificadas y generen el puntaje crediticio del usuario dándole no solo acceso a nuestros servicios sino también la posibilidad de diversificar su portafolio.

CONCLUSIONES

Se identificó la necesidad de diseñar una herramienta financiera que permita una inclusión genuina y sencilla de los individuos rurales, pues, si bien existen instituciones financieras como el Banco Agrario, no suelen tener en consideración las diferentes necesidades de crédito que tiene cada individuo.

En el sector rural, los criterios de financiamiento suelen ser demasiado generales, ignorando que un agricultor, por ejemplo, enfrenta mayores riesgos debido a factores climáticos y de cosecha en comparación con un ganadero, cuyo flujo de ingresos puede ser más estable.

Además, los requisitos exigidos por las entidades financieras, como un historial crediticio sólido, garantías formales y una gran cantidad de documentación dificultan el acceso al crédito para pequeños productores, limitando su crecimiento y capacidad de inversión.



Referencias

DNP (2024). Pobreza Monetaria y Desigualdad. Dirección Nacional de Planeación. https://colaboracion.dnp.gov.co/CDT/PublishingImages/Planeacion-y-desarrollo/2024/Agosto/pdf/pobreza-monetaria.pdf

ECONOMIA

Laboratorio de inteligencia artificial aplicada a Economía

Universidad Externado de Colombia

FACULTAD DE ECONOMÍA

ECONOMIA

Laboratorio de inteligencia artificial aplicada a Economía

Construcción y análisis para Colombia de un índice de bienestar financiero por clústeres mediante Machine Learning no supervisado

Autores

Laura Nathaly Camacho Cepeda laura.camacho05@est.uexternado.edu.co
Ricardo Sinning Sanabria ricardo.sinning@est.uexternado.edu.co
Juan Miguel Rodríguez Trujillo juan.rodriguez68est.uexternado.edu.co

2025

Resumen

El bienestar financiero es un concepto multidimensional que va más allá del acceso a productos financieros, abarcando la capacidad de los individuos para administrar sus recursos, responder a imprevistos y planificar su futuro económico. En este estudio se propone la construcción de un Índice de Bienestar Financiero (IBF) para la población en Colombia, utilizando datos de la Encuesta de Carga Financiera y Educación Financiera de los Hogares (IEFIC) - 2018 realizado por el DANE. La metodología empleada se basa en un enfoque de Machine Learning No Supervisado, específicamente el algoritmo K-Means, con el objetivo de segmentar a la población en grupos homogéneos según su nivel de bienestar financiero.

El análisis de los clústeres revela diferencias significativas entre los grupos identificados. Los individuos con menores ingresos presentan mayor endeudamiento relativo, menor acceso a productos financieros formales y una alta dependencia de fuentes de crédito informales. Además, registran niveles más bajos de educación financiera, lo que podría influir en la limitada planificación y gestión de sus finanzas. Estos resultados evidencian que el bienestar financiero no depende únicamente del nivel de ingresos, sino también de factores como el acceso y uso de instrumentos financieros, los hábitos de ahorro e inversión y el conocimiento financiero. La segmentación basada en aprendizaje no supervisado permite una mejor comprensión de los patrones financieros dentro de la población, lo que podría contribuir al diseño de políticas públicas más eficaces en términos de inclusión y bienestar financiero.

Introducción

El bienestar financiero es un concepto ampliamente discutido tanto en el ámbito académico como en el sector financiero. Más allá de incrementar el número de clientes en el sistema financiero, el objetivo debe ser garantizar que los usuarios accedan a servicios de calidad y disfruten de una experiencia positiva. Según la UNSGSA (2021), el bienestar financiero se define como la capacidad de las personas y familias para gestionar eficazmente sus obligaciones económicas actuales y sentirse seguras respecto a su futuro financiero. Este concepto abarca cuatro dimensiones fundamentales:

1. Gestión diaria

Capacidad para manejar las finanzas cotidianas y cumplir con las necesidades de consumo.

2. Resiliencia

Habilidad para absorber y recuperarse de imprevistos financieros.

3. Metas a futuro

Progreso hacia la consecución de objetivos financieros a largo plazo.

4. Confianza

Sensación de seguridad y control sobre la situación financiera personal.

En Colombia, la población que percibe hasta dos salarios mínimos mensuales legales vigentes (SMMLV) enfrenta desafíos particulares en términos de bienestar financiero. Según La República (2024), el 81% de los adultos colombianos ganan menos de dos SMMLV. Esta cifra pone en evidencia la vulnerabilidad económica de una gran proporción de la población y la necesidad de evaluar su bienestar financiero más allá del simple acceso a servicios bancarios.

El Reporte de Inclusión Financiera 2023 señala que el 94,6% de los adultos en el país posee al menos un producto financiero, lo que indica un alto nivel de acceso al sistema financiero. Sin embargo, solo el 35,3% de la población adulta tenía acceso a algún producto de crédito para 2023 (Banca de las Oportunidades, 2023). Esta disparidad sugiere que, aunque el acceso general a servicios financieros es elevado, la utilización de productos crediticios sigue siendo limitada, especialmente entre los segmentos de menores ingresos.

La construcción de un índice de bienestar financiero específico para la población que gana hasta dos SMMLV permitiría evaluar de manera más precisa su situación económica y diseñar políticas públicas más efectivas.

De esta manera, teniendo guía en el Grupo de Trabajo de Salud Financiera de UNSGSA (2020) y el índice de Bienestar Financiero (BF) hecho por el Consumer Financial Protection Bureau (CFPB) de Estados Unidos, se ha construido un índice de Bienestar Financiero de acuerdo con 4 dimensiones:

Gráfico 1: Dimensiones que componen el Índice de Bienestar Financiero

Fuente: Elaboración Propia

Análisis y metodología IA

Para la construcción del índice se tomaron datos de la Encuesta de Carga Financiera y Educación Financiera de los Hogares (IEFIC) - 2018 realizada por el DANE, cuyos encuestados fueron la población civil de la zona urbana de las ciudades de Bogotá, Medellín y Cali. En esta encuesta se encuentran diversas preguntas que permiten establecer en qué dimensiones los colombianos tienen fortalezas a nivel financiero, y en qué aspectos se puede mejorar.

Para la construcción del Índice de Control de las Finanzas (ICF) se usaron las secciones B y F de la encuesta, estas se llaman "consumo" y "deuda no hipotecaria" respectivamente. En general, las preguntas realizadas permiten indagar sobre si los ingresos cubren completamente los gastos, los motivos de ahorro, y el uso de crédito para cubrir déficits, como el endeudamiento para cubrir gastos excedentes. También se incluyen preguntas sobre los métodos utilizados para cubrir gastos en meses donde estos exceden los ingresos y el tipo de activos o servicios que poseen.

En el caso del Índice de Educación Financiera (IEF) se utilizó toda la sección E, de educación financiera. En esta sección se hacen preguntas sobre datos o procesos financieros en los que solo hay una respuesta correcta; si la persona responde de manera correcta, esto indicará que la persona conoce instrumentos financieros o ha tenido algún tipo de acercamiento a las finanzas, que es un buen indicador. El cálculo de este índice se realizó así:

$$\mathrm{IEF} = \left(\frac{\sum_{i=1} C_i}{n}\right)$$

Donde Imagen 1178720940, Picture es una variable binaria que toma el valor de 1 si la respuesta a la pregunta i es correcta, y 0 si es incorrecta; n es el número total de preguntas sobre educación financiera.

Para la construcción del Índice de Libertad Financiera (ILF) se utilizó la sección B, de consumo, en la que se analiza la Relación de Ingresos – Gastos, y si los ingresos son mayores a los gastos, se analiza en qué usan las personas el dinero restante.

Finalmente, en la construcción del Índice de acceso financiero (IAF) se usaron las siguientes variables dicótomas: seguro, tarjeta de crédito, crédito con prestamistas, crédito con amigos, CDT's, acciones y aceptación de pagos electrónicos.

Todos estos cálculos fueron normalizados para mantener una escala de 0 a 1, de manera que se pudiera obtener un índice de cada uno. El resultado de este proceso es el Índice de Bienestar Financiero (IBF), el cual es un número que va de 0 a 1, donde un valor más cercano a 1 indica un mejor nivel de bienestar financiero. Dado que en nuestro análisis el grupo de interés son las personas que ganan hasta 2 SMMLV, comparamos las estadísticas descriptivas de los componentes del IBF entre este grupo y los individuos que ganan más de este límite.

Los resultados se pueden ver en la Tabla 1.

Tabla 1. Comparación IBF por grupos de interés.

Medida	clas_ing ≤ 2	clas_ing > 2
Mean	0.129406	0.242207
Std	0.150384	0.206138
Min	0.000000	0.000000
25%	0.000000	0.088889
50%	0.088889	0.177778
75%	0.177778	0.366667
Max	1.000000	1.000000

	Gr. 1 ICF	Gr. 2 ICF
mean	0.530001	0.592650
std	0.342378	0.362164
25%	0.330033	0.330033
75%	0.990099	0.995050

Estadística	Gr. 1 IEF	Gr. 1 ILF	Gr. 2 IEF	Gr. 2 ILF
Media	0.338	0.023	0.402	0.049
DS	0.232	0.073	0.234	0.105
Per 25%	0.200	0.000	0.200	0.000
Per 75%	0.500	0.000	0.600	0.000

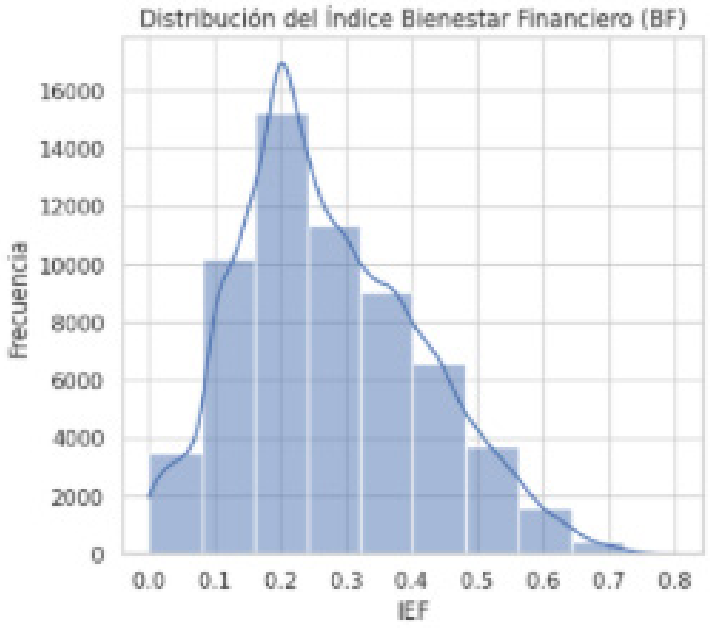
Fuente: Elaboración Propia

Como resultado, se observa que la media de los índices que componen el índice de Bienestar financiero es siempre menor para aquellos colombianos que ganan hasta 2 SMMLV. A su vez, se observa de manera general que el índice en el que mejor les va a los colombianos es en el de Control de las Finanzas mientras que en que peor les va es en el de Acceso a servicios Financieros.

De igual manera,es interesante ver que en los dos grupos de los cuatro índices existen personas con in IBF iguales o muy cercanos a 1, lo que indica que el nivel de ingreso, en este caso, no es el único determinante del nivel de bienestar financiero.

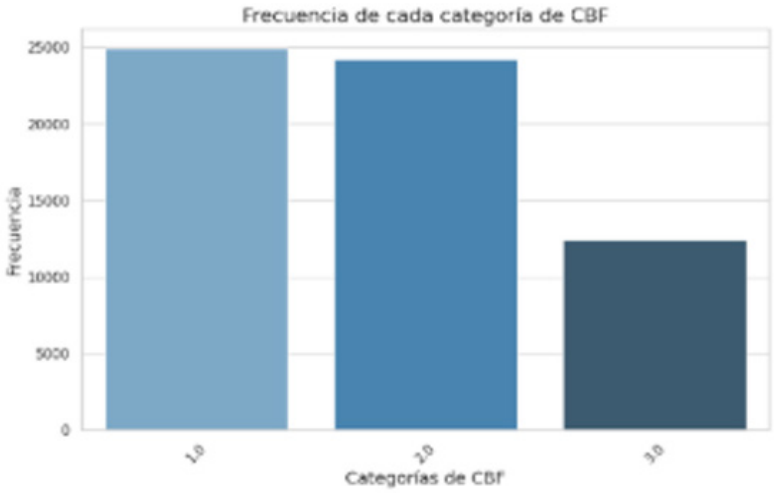
Para poder implementar el modelo de análisis se decidió crear la variable “Categoría de Bienestar Financiero” (CBF), la cual categoriza el IBF según los deciles (gráfica 1), demostrando que la mayoría de la población se encuentra en los deciles más bajos de Bienestar Financiero. Luego, el CBF (gráfica 2) tiene las clases 1, 2 y 3, donde en 1 se agrupan las personas con el IBF bajo, el 2 los individuos con un resultado medio y en 3 las personas de la muestra con un mayor nivel de bienestar financiero correspondiente con el índice creado.

Gráfica 2: Distribución del Bienestar Financiero por deciles



Fuente: Elaboración Propia

Gráfica 3: Categoría del Bienestar Financiero



Fuente: Elaboración Propia

Se observa que el Bienestar Financiero según los cálculos de este artículo, es algo que tiene la minoría de la población, por esto es relevante saber qué características comparten las personas que tienen un mayor bienestar financiero.

Modelo

Para el análisis del Índice de Bienestar Financiero, se utilizó un modelo K-Means, un algoritmo de Machine Learning No Supervisado basado en la agrupación de observaciones en diferentes grupos o clústeres según su similitud. A diferencia de los modelos supervisados, que requieren una variable objetivo para entrenar el modelo, los métodos no supervisados identifican patrones en los datos sin una etiqueta predefinida.

El algoritmo K-Means opera bajo un enfoque basado en distancias, agrupando observaciones en K grupos distintos. Cada grupo es representado por un centroide, el cual se actualiza dinámicamente en cada iteración para minimizar la variabilidad dentro del clúster. El funcionamiento interno de K-Means se desarrolla en los siguientes pasos:

1. Inicialización

Se seleccionan aleatoriamente K centroides en el espacio de las variables.

2. Asignación de clústeres

Cada observación es asignada al clúster cuyo centroide se encuentre a la menor distancia euclidiana.

3. Actualización de centroides

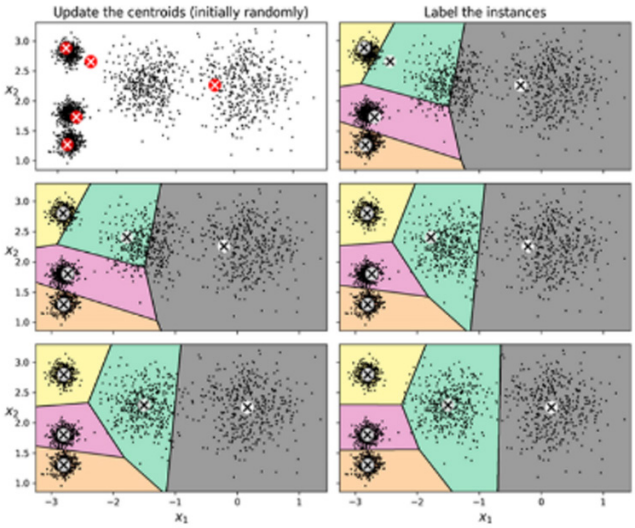
Se recalculan los centroides como el promedio de las observaciones asignadas a cada grupo.

4. Iteración

Se repiten los pasos 2 y 3 hasta que los centroides convergen y los clústeres dejen de cambiar significativamente.

Este proceso se puede observar en el Gráfico 4.

Gráfico 4. El algoritmo de K-Means

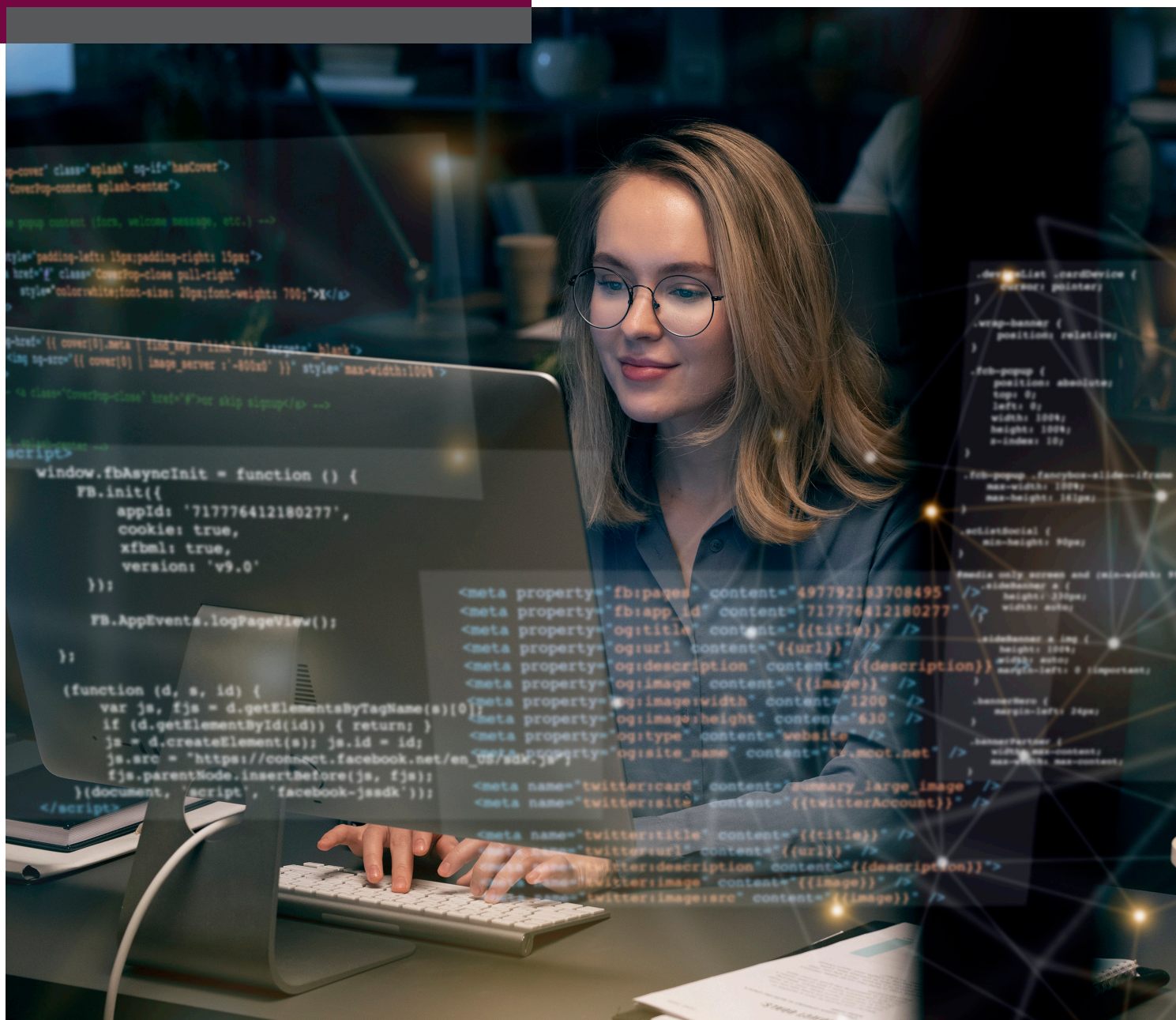


Tomado de: Geron, A. (2023). Hands-On Machine Learning with Scikit-Learn, Keras, and TensorFlow (3rd edition). O'Reilly Media.

Antes de aplicar K-Means, se escalaron las variables para garantizar que todas contribuyan de manera equitativa en la asignación de clústeres. Esto se debe a que K-Means se basa en distancias euclidianas, y sin escalado, variables con valores numéricamente mayores pueden tener mayor peso en el proceso de agrupación, distorsionando los resultados. Para evitar este problema, las variables fueron normalizadas mediante escalado estándar (media 0 y desviación estándar 1).

Para seleccionar el número óptimo de clústeres (K), se utilizó el método del codo, el cual se basa en la métrica de inercia. La inercia mide la suma de las distancias cuadradas entre cada observación y el centroide de su clúster. A medida que aumenta el número de clústeres, la inercia disminuye, pero en algún punto la reducción de la inercia se vuelve marginal. Este punto de inflexión en la gráfica de inercia se conoce como “el codo” y es donde se establece el valor óptimo de K.

Con esta metodología, se logró una segmentación de los individuos en distintos grupos homogéneos según su bienestar financiero, permitiendo una mejor interpretación de los patrones dentro de la población analizada.



Resultados

El análisis de los clústeres muestra diferencias marcadas en el perfil financiero de los individuos agrupados. En particular, aquellos con menores ingresos presentan un mayor nivel de endeudamiento, tanto absoluto como relativo, lo que sugiere una mayor dependencia del crédito. Además, este grupo tiene un acceso más limitado a productos financieros formales, evidenciado por una menor proporción de individuos con seguros, tarjetas de crédito y cuentas de ahorro. Cabe destacar que es el único grupo que accede a créditos con prestamistas o “gota a gota”, lo que indica una posible exclusión del sistema financiero tradicional y una exposición a condiciones crediticias más desfavorables.

En términos de hábitos financieros, los individuos con menores ingresos no presentan comportamiento de ahorro ni inversión, lo que limita su capacidad de acumulación de recursos a largo plazo. Asimismo, registran los niveles más bajos de educación financiera, lo que podría estar relacionado con su menor acceso a productos financieros formales y su preferencia por opciones de financiamiento informales. Estos resultados sugieren la necesidad de implementar estrategias de educación e inclusión financiera para mejorar las condiciones de este grupo.

Conclusiones

Los bajos niveles de bienestar financiero observados en la población que gana hasta dos salarios mínimos en Colombia tienen raíces estructurales en la economía del país. Un factor particularmente preocupante es la alta dependencia de los créditos informales, como los préstamos con “gota a gota”, lo que refleja la falta de acceso a servicios financieros formales y la persistente informalidad del mercado laboral. Además, el endeudamiento no puede ser evaluado de manera aislada como un indicador positivo o negativo, sino que debe entenderse en función del contexto económico y las oportunidades disponibles para esta población.

Para mejorar la situación financiera de estos hogares, es clave promover el acceso a servicios financieros formales mediante la creación de productos de microfinanzas y microcréditos con tasas de interés que se acoplen a su nivel de riesgo y les permitan entrar al sistema, ya que en los modelos de endeudamiento informal las tasas son desproporcionadamente mayores a las que actualmente ofrece el sistema financiero. En este caso, se deben tener en cuenta posibles barreras a tener tasas de interés un poco más altas para aquellos perfiles con mayor riesgo como, por ejemplo, la tasa de usura. Si bien la idea es que las personas puedan acudir al mercado con tasas de interés bajas, también es necesario tener en cuenta que el tener un tope muy bajo en estas puede estar excluyendo del sistema financiero a personas que terminan acudiendo a la financiación no formal, que según un estudio de la ANIF llamado “Análisis de cambios metodológicos de la tasa de usura y su impacto en la inclusión financiera: un enfoque para el desarrollo económico sostenible en Colombia” tienen tasas hasta del 700%, mientras que la tasa máxima legal ronda alrededor del 24% en estos momentos.

Además, la implementación de programas de inclusión y educación financiera permitiría fortalecer la capacidad de estas personas para administrar sus recursos y evitar el sobreendeudamiento en el sistema informal. Estas medidas, en conjunto, podrían contribuir significativamente a la reducción de la vulnerabilidad financiera de este segmento de la población y a su integración en el sistema financiero formal.

Referencias

ANIF. (2025). Análisis de cambios metodológicos de la tasa de usura y su impacto en la inclusión financiera: un enfoque para el desarrollo económico sostenible en Colombia

Banca de las Oportunidades. (2023). Reporte de inclusión financiera 2023: Nuevos avances y retos en Colombia.

Geron, A. (2023). Hands-On Machine Learning with Scikit-Learn, Keras, and TensorFlow (3rd edition). O'Reilly Media.

La República. (2024). En el país, 3.300.000 personas ganan el mínimo, lo que equivale al 15% de la población. Recuperado de https://www.larepublica.co

UNSGSA. (2021). La medición de la salud financiera: Una perspectiva global sobre cómo evaluar el bienestar financiero. Recuperado de La Medición_02.pdf

ECONOMIA

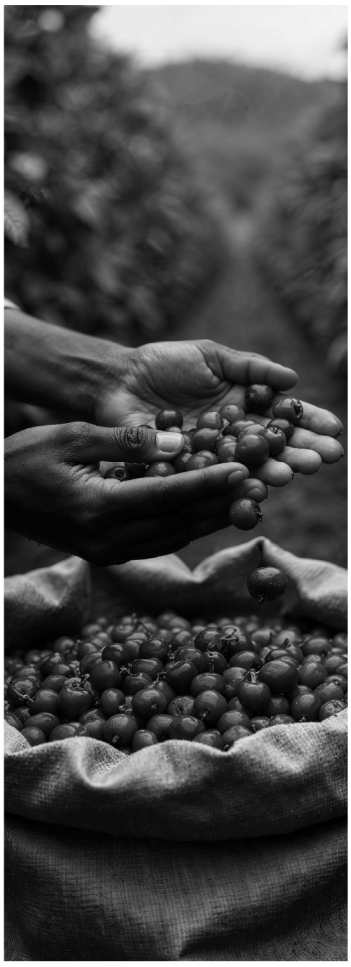
Laboratorio de inteligencia artificial aplicada a Economía

Universidad Externado de Colombia

FACULTAD DE ECONOMÍA

ECONOMIA

Laboratorio de inteligencia artificial aplicada a Economía



La revolución cafetera en Huila:

Desafíos y perspectivas empresariales para el 2025

Autor

Juan Carlos Urbano Rodríguez juan.urbano44@est.uexternado.edu.co

2025

Resumen

El presente trabajo investigativo busca comprender y determinar las posibles problemáticas económicas, sociales y ambientales que enfrenta el sector cafetero en el Departamento del Huila, además de indagar sobre los proyectos productivos que se han implementado en los últimos años teniendo en cuenta el poder comercial de la producción cafetera en este departamento.

Por lo tanto, más allá de identificar las dificultades estructurales de los productores de café, es relevante proponer ideas innovadoras que tengan un gran impacto en la sociedad cafetera con el fin de consolidar en la economía nacional un nuevo sector altamente productivo.

Introducción

En los últimos años el departamento del Huila se ha convertido en el mayor productor del café a nivel nacional. Este hito económico ha consolidado una interdependencia entre los caficultores y los proveedores de café, además las bonanzas cafeteras en el Huila junto con los proyectos productivos gestionados por la Gobernación Departamental repercuten sobre los niveles de producción. En ese sentido, el sector se divide en dos subsectores: cafés especiales (café tostado y café verde) y derivados de café que pueden aportarles valor agregado a los ingresos del café. A partir de lo mencionado anteriormente es pertinente preguntarse ¿Qué estrategias comerciales se deberían implementar para aumentar la rentabilidad y comercialización del café en el Departamento del Huila? Para poder resolver esa pregunta se va a implementar un análisis temporal desde el año 2018 hasta 2024 e indagar acerca de las dinámicas de producción durante ese periodo de tiempo.

En 2024, el departamento del Huila representó el 19,08% de la producción total del café en Colombia según la Federación Nacional de Cafeteros (FNC), sin embargo, aún no hay avances significativos en la transformación productiva del sector cafetero al momento de implementar productos que maximicen las ganancias de los productores con el fin de generar valor agregado sobre la productividad del café en tiempos de bonanza y de crisis cuando se presentan fluctuaciones en su precio. Por lo tanto, el sector cafetero debería diseñar estrategias comerciales que fortalezcan y modifiquen la cadena de valor del café para alcanzar mayores ingresos en el mediano y largo plazo.

Cabe mencionar que durante el 2025 la carga del café pergamino seco ha alcanzado precios récord superiores a los 3 millones de pesos, lo cual repercute sobre las ganancias que puede generar el sector cafetero en el largo y el nivel salarial de los jornaleros de la región del Huila. Por lo tanto, es esencial que los subsectores como el sector de cafés especiales y derivados de café se prioricen en un plan estratégico para el desarrollo productivo del Huila y el país.

Importancia del sector cafetero en Colombia y el Huila

El café del Huila se cultiva en el sur de la Región Andina por comunidades campesinas en 35 municipios, los cuales albergan más de 84.000 familias que cultivan 145.741 hectáreas de café arábico de las variedades Castillo, Colombia, Caturra, Típica, Borbón y Tabí, además el 74% de la población del Huila se dedica a la caficultura.

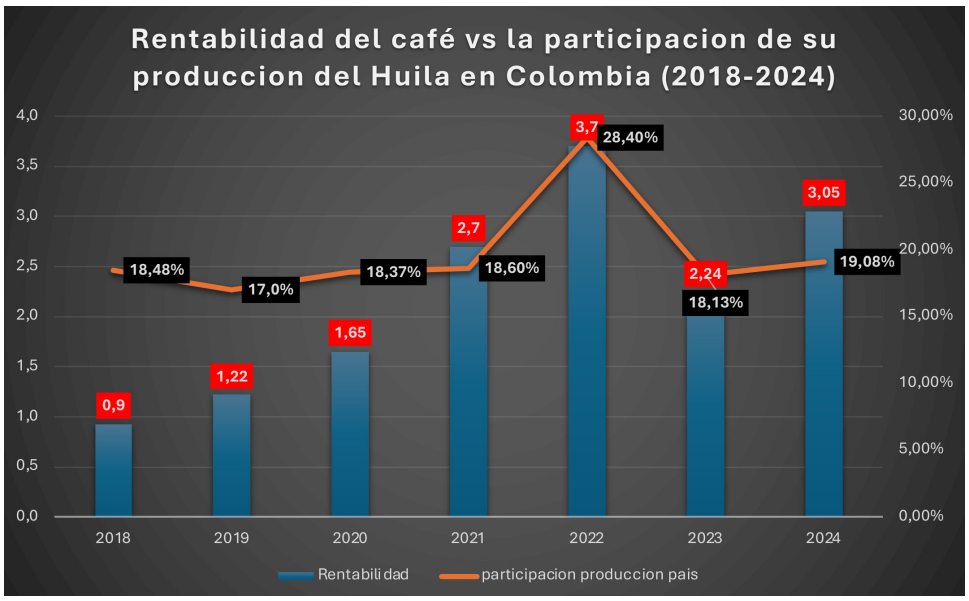
En promedio, el café representa el 8,8% del Producto Interno Bruto departamental, el 58,84% del Producto Interno Bruto agropecuario, a su vez genera el 53% de las exportaciones del departamento, genera más de 101.000 de empleos directos y alrededor de 27.000.000 de jornales al año. (Federación Nacional de Cafeteros del Huila, 2025)

La Gobernación del Huila en los últimos años ha implementado proyectos productivos que se enfocan en mejorar la calidad, optimizar los procesos de eficiencia productivos del café e incrementar la rentabilidad. Para lograr esos propósitos se implementó en el 2023 una estrategia llamada "Más Agronomía, Más Productividad, Más Calidad para una mejor rentabilidad", este programa de asistencia productiva ha alcanzado una cobertura de 94,6% ha beneficiado a alrededor de 199.387 caficultores según cifras de la gobernación del Huila.

En ese sentido, este proyecto social se basa en visitas a las fincas para dar pedagogía acerca de métodos productivos para la tecnificación del campo y la implementación de factores calidad que aumentan los rendimientos de la venta de café.

Análisis y metodología IA

Grafico 1. Rentabilidad del café vs participación de su producción del Huila en Colombia (2018-2019)



Fuente: Elaboración propia, datos suministrados de los informes de gestión de la Federación Nacional de Cafeteros del Huila.

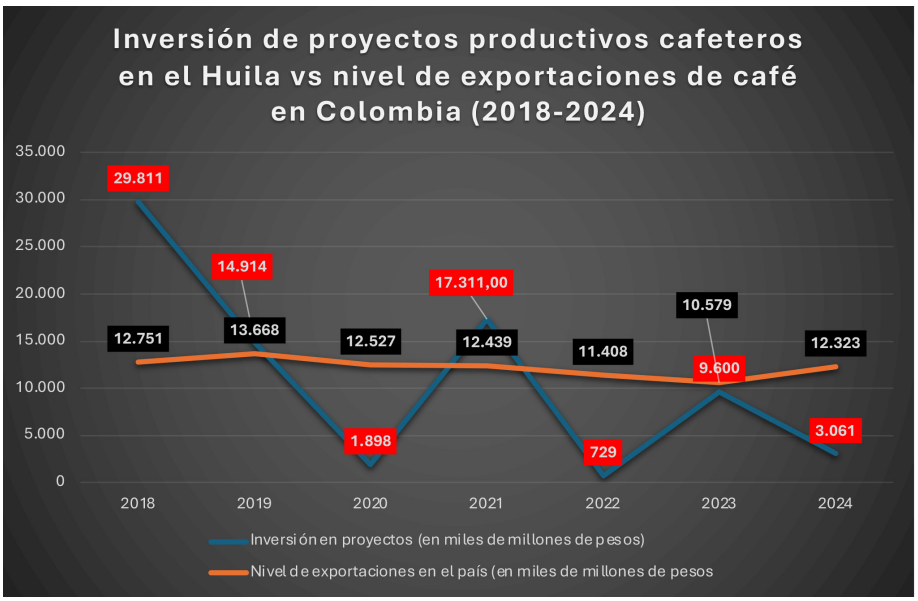
En el gráfico 1. Podemos observar que entre 2018 y 2024, la rentabilidad del café en el Huila mostró una tendencia creciente, pasando de 0,9 en 2018 a un pico de 3,7 en 2022, seguido de una caída a 2,24 en 2023 y una recuperación a 3,05 en 2024, evidenciando alta volatilidad. Mientras tanto, la participación del Huila en la producción nacional fluctuó entre 17,0% y 28,4%, con su punto más alto en 2022, coincidiendo con la mayor rentabilidad, y cerrando en 19,08% en 2024, lo que sugiere una relación no lineal entre ambas variables.

Tabla 1. Aspectos productivos del sector cafetero en el Huila

Año	Ingresos (en billones de peso)	Participación de la producción en el país	Producción (en millones de sacos de café)	Nivel de exportaciones en el país (en miles de millones de pesos
2018	0.928	18,40%	1.29	12.751
2019	1.22	17,00%	1.31	13.668
2020	1.65	18,37%	2.52	12.527
2021	2.7	18,30%	2.5	12.439
2022	3.7	28,40%	2.42	11.408
2023	2.24	18,13%	2.17	10.579
2024	3.05	19,08%	2.48	12.323

Fuente: elaboración propia, datos suministrados de la Federación Nacional de Cafeteros del Huila

Gráfico 2. Inversión de proyectos productivos cafeteros en el Huila vs nivel de exportaciones de café en Colombia (2018-2024).



Fuente: elaboración propia, datos suministrados de la Federación Nacional de Cafeteros del Huila.

Tabla 2. Precios del café y tipo de cambio a lo largo del tiempo

Año	Precio internacional (centavos de dólar por libra)	Precio interno (precio promedio de la carga de café)	Tipo de cambio (promedio en pesos colombianos)
2018	137.15	727.000	3285.51
2019	133.19	815.000	3277.14
2020	158.25	1.010.000	3432.5
2021	218.15	2.150.867	3981.16
2022	279.12	2.500.000	4810.2
2023	209.12	1.788.844	3822.05
2024	254.71	2.480.000	4409.15

Fuente: elaboración propia, datos suministrados de la Federación Nacional de Cafeteros del Huila.

Tabla 3. Inversión social destinada al sector cafetero en el Departamento del Huila

Año	Precio internacional (centavos de dólar por libra)	Precio interno (precio promedio de la carga de café)	Tipo de cambio (promedio en pesos colombianos)
2018	137.15	727.000	3285.51
2019	133.19	815.000	3277.14
2020	158.25	1.010.000	3432.5
2021	218.15	2.150.867	3981.16
2022	279.12	2.500.000	4810.2
2023	209.12	1.788.844	3822.05
2024	254.71	2.480.000	4409.15

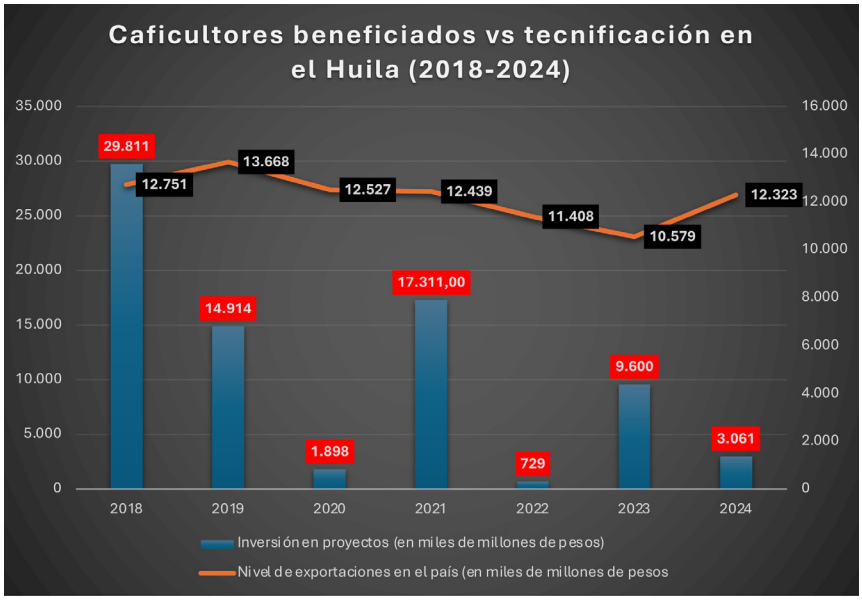
Fuente: elaboración propia, datos suministrados de la Federación Nacional de Cafeteros del Huila.

Al realizar un análisis conjunto de las tablas y gráficos anteriormente presentados podemos inferir que el sector cafetero del Huila ha experimentado un crecimiento significativo en ingresos, pero con volatilidad en la producción y las exportaciones y el aumento en los precios internacionales y la devaluación del peso beneficiaron el sector entre 2020 y 2022, pero la caída de precios en 2023 impactó la rentabilidad, sin embargo la inversión social y la tecnificación han disminuido, lo que va a generar problemas de productividad en el futuro en caso de que no se implementen medidas necesarias.

Cabe mencionar que, para la eficacia y transparencia de la FNC en la gestión y ejecución

de proyectos de desarrollo, la inversión a través de cooperantes nacionales e internacionales se han consolidado como un aliado estratégico, para proyectos que impactan en la cadena de valor del sector cafetero. Esa situación ha permitido que, en 2018, se empezará a desarrollar la Fábrica de Proyectos en la que convergen la innovación y tecnificación del campo y que fortalecen la rentabilidad en el largo plazo de los ingresos de los caficultores y les puede dar un valor agregado a sus cosechas cafeteras.

Gráfico 3. Caficultores beneficiados vs tecnificación en el Huila (2018-2024)



Fuente: elaboración propia, datos suministrados de la Federación Nacional de Cafeteros del Huila

Variables empleadas

Se construyó una base de datos a partir de datos recolectados de la Federación Nacional de Cafeteros del Huila y sus informes de gestión reportados, junto con la información presentada por la Gobernación del Huila basada en los siguientes criterios:

Variables de entrada

• **Eje económico**: participación producción país, producción, exportaciones, Café con factor de rendimiento (café estándar y especial), productores beneficiados, hectáreas cultivadas, valorización lotes, tecnificación, afectaciones productivas, Prima de calidad.

• **Precios:** precio internacional e interno, tipo de cambio.

• **Eje social:** fincas beneficiadas, cantidad de proyectos, inversión proyectos.

• **Eje ambiental:** inversión en reforestaciones, inversión ambiental.

• **Gestión de proyectos:** inversión social, iniciativas cafeteras.

• **Costos de producción:** maquinaria, fertilizantes.

• Opiniones y percepciones sobre el café y sus derivados.

Variable objetivo

Rentabilidad del café (ganancias netas).

Metodología IA

Para realizar esta investigación se utilizaron dos modelos de Machine Learning asociado a redes neuronales y de regresión y clasificación con el fin de construir un análisis económico integral sobre las variables que impactan en las estrategias comerciales cafeteras y por ende en sus niveles de rentabilidad y comercialización.

Modelo de Red Neuronal

El modelo de redes neuronales está enfocado en entender patrones de los proyectos productivos y de inversión en el sector cafetero para optimizar la eficiencia de los factores producción y la calidad de los productos derivados del café. Los datos fueron recolectados a partir de los informes de gestión anuales del comité de Cafeteros del Huila y la Federación Nacional de cafeteros.

Desarrollo de la metodología empleada

Modelo 1 - Red Neuronal para Regresión

• **Arquitectura:**
Capa densa de 64 neuronas con ReLU y input_shape=(X_train.shape[1],)
Capa oculta de 16 neuronas con ReLU
Capa oculta de 8 neuronas con ReLU
Capa de salida de 1 neurona (sin activación, apropiada para regresión)

• **Compilación:**
Optimizador: Adam
Función de pérdida: MSE (Error Cuadrático Medio)
Métrica: MAE (Error Absoluto Medio)

• **Entrenamiento:**
epochs=100
batch_size=8
Validación con validation_data=(X_test, y_test)

Modelo 2 - Red Neuronal para Regresión con Regularización

• **Diferencias clave:**
Uso de Dropout(0.2) después de las capas densas para reducir el sobreajuste.

Kernel regularizador L2(0.001) en las capas densas.
Se ajusta con validation_split=0.2.

Modelo 3 - Red Neuronal para Clasificación Binaria

• **Arquitectura:**
Capa densa de 10 neuronas con ReLU
Capa oculta de 8 neuronas con ReLU
Capa de salida de 1 neurona con sigmoid (para salida binaria)

• **Compilación:**
Optimizador: Adam
Función de pérdida: binary_crossentropy
Métrica: accuracy

• **Entrenamiento:**
epochs=10
batch_size=32
Validación con validation_data=(X_test, y_test)

Evaluación

• Para regresión: R2 y MAE.

• Para clasificación: accuracy, AUC-ROC (Área bajo la curva ROC), Matriz de confusión.

• Se utiliza model.fit() para el entrenamiento y validación.

Incluye diferentes modelos para regresión y clasificación, con técnicas de regularización como Dropout y L2 para prevenir sobreajuste. Se emplea Adam como optimizador y se usa StandardScaler para normalizar los datos antes del entrenamiento.

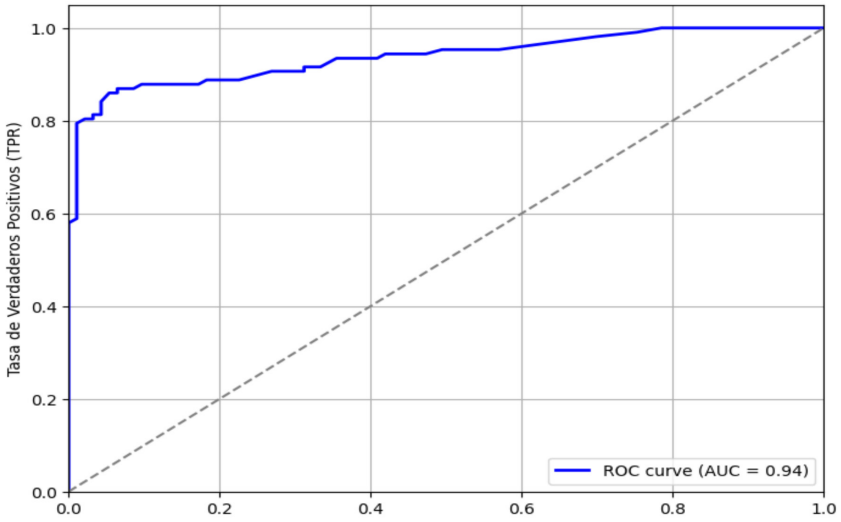
Resultados

Métricas de evaluación.

Métricas relacionadas con problemas de clasificación .

Curva ROC.

Curva ROC



Tasa de falsos positivos (FPR)

A partir del resultado de este gráfico podemos decir los siguiente:

AUC

Indica que el modelo tiene un desempeño excelente, ya que el AUC varía entre 0.5 (modelo aleatorio) y 1.0 (modelo perfecto).

Un valor de 0.94 significa que hay un 94% de probabilidad de que el modelo clasifique correctamente una instancia positiva sobre una negativa.

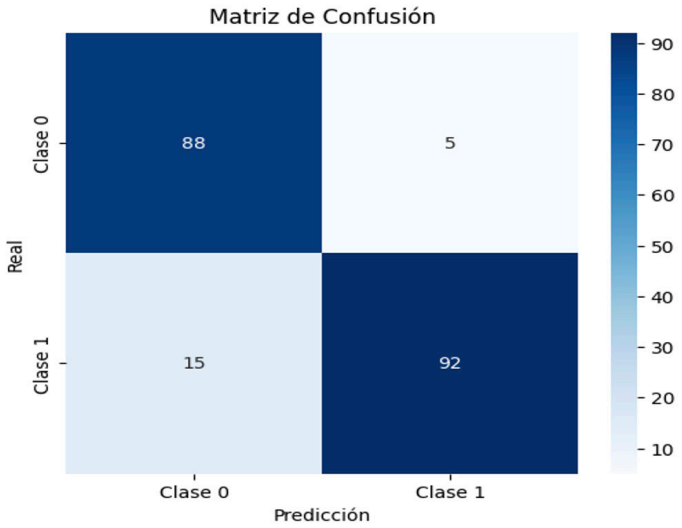
Forma de la curva

Al inicio, el TPR aumenta rápidamente con un FPR bajo, lo que sugiere que el modelo clasifica bien los positivos sin muchos errores.

En valores altos de FPR, la curva se aplana, lo que indica que a medida que se aumentan los verdaderos positivos, también crecen los falsos positivos.

El modelo es altamente efectivo para distinguir entre clases. Con un AUC de 0.94, es probable que tenga una excelente precisión y sensibilidad, con pocos falsos positivos y falsos negativo.

Matriz de confusión



Reporte de clasificación:

	precision	recall	f1-score	support
0	0.85	0.95	0.90	93
1	0.95	0.86	0.90	107
accuracy			0.90	200
macro avg	0.90	0.90	0.90	200
weighted avg	0.90	0.90	0.90	200

A partir del resultado de este grafico podemos decir los siguiente:

Estructura de la matriz:

• Contiene cuatro valores clave:
Verdaderos Positivos (TP = 92): Casos positivos correctamente clasificados.

Verdaderos Negativos (TN = 88): Casos negativos correctamente clasificados.

Falsos Positivos (FP = 5): Casos negativos incorrectamente clasificados como positivos.

Falsos Negativos (FN = 15): Casos positivos incorrectamente clasificados como negativos.

Métricas clave:

• **Precisión (Accuracy):** 90% de las predicciones son correctas.

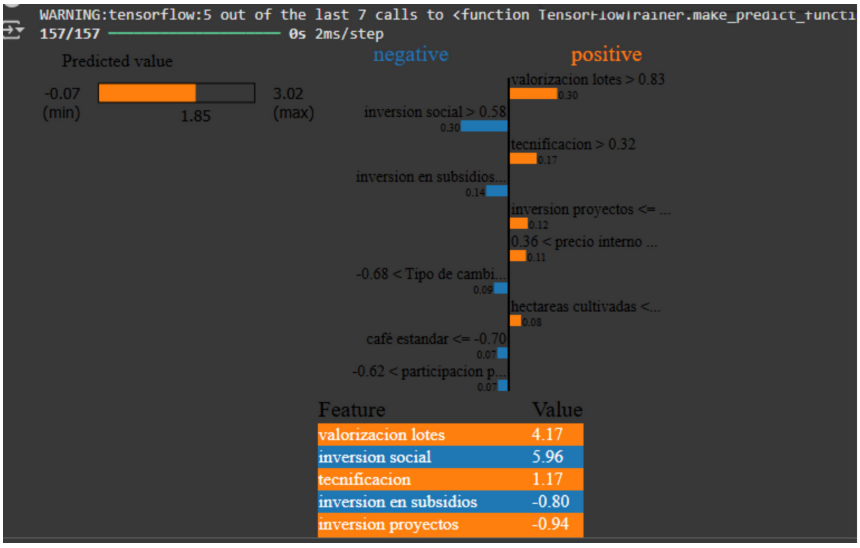
• **Precisión de la Clase 1:** 94.8% de las predicciones positivas son correctas.

• **Sensibilidad (Recall):** 86% de los casos positivos fueron identificados correctamente.

• **Especificidad:** 94.6% de los casos negativos fueron clasificados correctamente.

Teniendo en cuenta lo anterior, se puede decir que el modelo tiene un buen desempeño general. Los falsos negativos (15 casos) pueden ser un problema si la Clase 1 es crítica.

Valores explicativos



A partir de este grafico podemos inferir lo siguiente:

El modelo muestra que la valorización de los lotes y la inversión social son los factores más influyentes en el resultado positivo.

Se debe analizar por qué la inversión en subsidios y proyectos está reduciendo la predicción y si hay formas de optimizar estos factores.

La tecnificación tiene un impacto positivo, aunque menor en comparación con la inversión social.

Métricas relacionadas con problemas de regresión

R2: 0,89

Error medio absoluto: 0,222

El modelo tiene un buen rendimiento, ya que explica el 89% de la variabilidad de los datos y su error medio absoluto es relativamente bajo, lo que indica que las predicciones están cercanas a los valores reales.

CONCLUSIONES

Aunque el departamento del Huila se caracteriza por elevar sus niveles de producción café cada año, esto no es suficiente para aumentar los niveles de rentabilidad. En ese sentido, las estrategias comerciales no solo deberían enfocarse en alcanzar altos niveles de producción sino en optimizar y transformar el café en bienes de consumo que tengan un gran impacto sobre la economía nacional e internacional con el fin de aumentar y diversificar las exportaciones del sector cafetero.

En primera medida, un café se considera especial cuando es percibido y valorado por los consumidores por alguna característica que lo diferencia de los cafés convencionales (sabor, aroma, lugar de origen, suavidad, calidad de la infusión), por lo cual están dispuestos a pagar un precio superior. En segunda instancia, entre los derivados de café se encuentran aquellos productos en los que se procesa el grano para obtener un producto nuevo, como es el caso de los dulces, snacks y demás, que le dan un valor agregado a la eficiencia productiva del café.

El fortalecimiento de la valoración de los lotes de producción cafetera de los pequeños y medianos productores de café puede marcar la diferencia en términos salariales y de inversión estructural en las diferentes fincas cafeteras del departamento del Huila. Por lo tanto, una iniciativa cafetera debería centrarse en promover ferias de café que atraigan la inversión extranjera que en el largo plazo es una inversión que se va a enfocar en elevar los niveles de productividad y calidad del café, y brinda un intercambio de conocimientos que permite afianzar nuevos factores de producción dentro de la tecnificación cafetera.

La inversión departamental y municipal en proyectos productivos debe seguir implementándose principalmente en la inclusión financiera para pequeños y nuevos productores de café que tengan los mecanismos para entrar al mercado y mediante iniciativas cafeteras demostrar que hay una drástica reducción de barreras de entrada, ya que año tras año en el Huila los proyectos dan la posibilidad de reducir los costos de producción

y los medios para comercializar y vender el café a precios justos. Cabe recalcar que la inclusión financiera desde el punto de vista crediticio brinda prestamos a tasas preferenciales, lo cual permite que los caficultores mejoren el factor de rendimiento del café el cual tiene una gran incidencia sobre el precio del café, de esta manera se busca equilibrar el mejoramiento de los ingresos de los caficultores y la calidad del café.

Referencias

Federación Nacional de Cafeteros del Huila (2025). Comité de Cafeteros del Huila https://huila.federaciondecafeteros.org/cafe-de-huila/

Federación Nacional de Cafeteros del Huila (2018). Informe de Gestión 2018 https://federaciondecafeteros.org/static/files/Informe_Gestion_2018.pdf

Federación Nacional de Cafeteros. (2021). Informe de gestión 2020. https://federaciondecafeteros.org/wp/listado-publicaciones/informe-de-gestion-2020/

Federación Nacional de Cafeteros del Huila (2021). Informe de gestión 2021. https://huila.federaciondecafeteros.org/app/uploads/sites/4/2022/05/FNC-Informe-de-Gestion-2021-Comite-Huila_compressed-2-comprimido.pdf

https://huila.federaciondecafeteros.org/tipos/informes/

Federación Nacional de Cafeteros del Huila (2022). Informe de gestión 2022. https://huila.federaciondecafeteros.org/app/uploads/sites/4/2022/05/FNC-Informe-de-Gestion-2021-Comite-Huila_compressed-2-comprimido.pdf

https://huila.federaciondecafeteros.org/tipos/informes/

Federación Nacional de Cafeteros del Huila (2023). Informe de gestión 2023. https://huila.federaciondecafeteros.org/app/uploads/sites/4/2022/05/FNC-Informe-de-Gestion-2021-Comite-Huila_compressed-2-comprimido.pdf

https://huila.federaciondecafeteros.org/tipos/informes/

Gobernación del Huila. Datos abiertos. https://www.datos.gov.co/Econom-a-y-Finanzas/Tasa-de-Cambio-Representativa-del-Mercado-Historic/mcec-87by/data_preview

https://ungc-production.s3.us-west-2.amazonaws.com/attachments/cop_2021/500910/original/Informe de Gestin Federacin Nacional de Cafeteros 2020.pdf?1627405810

Federación Nacional de Cafeteros. (2025). Anuncian apoyo a pequeños productores por 8.572 millones para renovación de cafetales.

https://federaciondecafeteros.org/wp/listado-noticias/anuncian-apoyo-a-pequenos-productores-por-8-572-millones-para-renovacion-de-cafetales/

https://federaciondecafeteros.org/wp/listado-noticias/anuncian-apoyo-a-pequenos-productores-por-8-572-millones-para-renovacion-de-cafetales/ Gobernación del Huila. (2021). Caficultura huilense sigue creciendo.

https://www.huila.gov.co/publicaciones/10606/caficultura-huilense-sigue-creciendo/#:~:text=Durante%20el%202020%20el%20Huila,productor%20del%20grano%20en%20Colombia.

Sociedad de Agricultores de Colombia (SAC). (2019). Cafeteros huilenses proyectan buena cosecha para 2019.
https://sac.org.co/cafeteros-huilenses-proyectan-buena-cosecha-para-2019/

Federación Nacional de Cafeteros de Colombia. (2023). Resumen Ejecutivo IG 2022.
https://federaciondecafeteros.org/app/uploads/2023/05/Resumen-Ejecutivo-IG-2022.pdf

Universidad Externado de Colombia

FACULTAD DE ECONOMÍA

ECONOMÍA

Laboratorio de inteligencia artificial aplicada a Economía

Universidad Externado de Colombia
FACULTAD DE ECONOMÍA

ECONOMIA
Laboratorio de inteligencia artificial aplicada a Economía

Exportaexpress:

Respuestas inmediatas para oportunidades globales

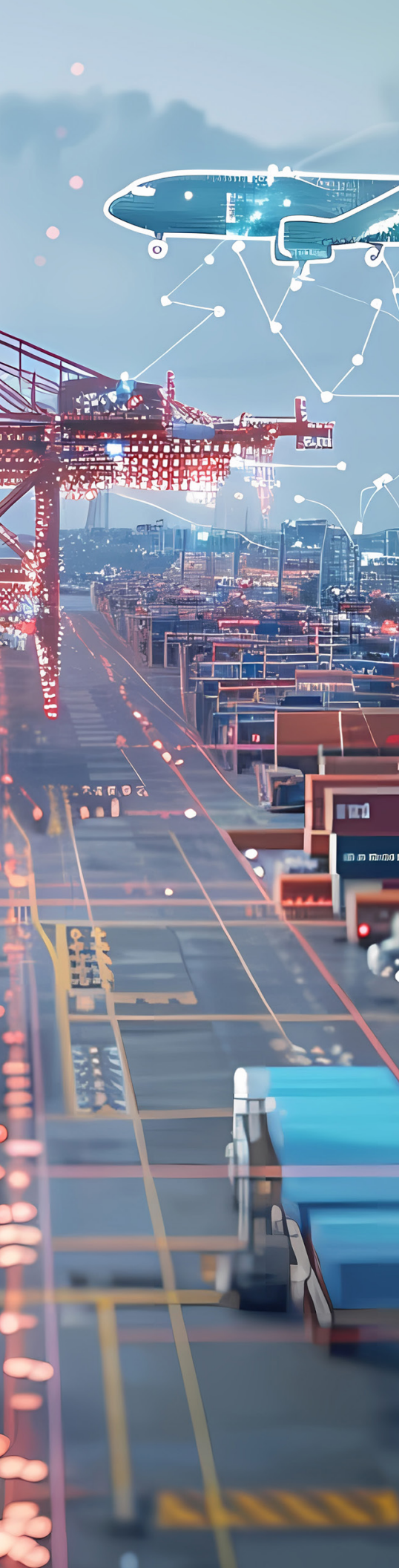
Autores

Santiago Bernal Giraldo santiago.bernal3@est.uexternado.edu.co
María Teresa Camacho Ríos maria.camacho03@est.uexternado.edu.co

2025

Resumen

El presente artículo contiene el desarrollo, la implementación y la evaluación de un asistente basado en inteligencia artificial generativa que atiende la necesidad de la Vicepresidencia de Exportaciones relacionada con la atención oportuna al usuario externo mediante la automatización de la atención de los correos electrónicos recibidos en el buzón exportaciones@procolombia.co, reduciendo los tiempos de atención y la precisión de las respuestas generadas.



Introducción

En el comercio internacional, la rapidez y precisión en la respuesta a consultas son cruciales. ProColombia, encargado de promover las exportaciones no minero energéticas de Colombia, enfrenta el desafío de gestionar eficientemente miles de correos electrónicos anuales en la Vicepresidencia de Exportaciones, lo que genera tiempos de respuesta variables y una alta carga de trabajo.

Mediante el proyecto capstone se desarrolló un asistente basado en inteligencia artificial generativa para automatizar la respuesta a consultas, reducir tiempos de respuesta, mejorar la precisión y aliviar la carga de trabajo del personal, optimizando así la operatividad de ProColombia y mejorando el servicio al cliente.

Este artículo detalla el desarrollo, implementación y evaluación de esta solución, así como conclusiones y recomendaciones buscando mejorar la atención al usuario.

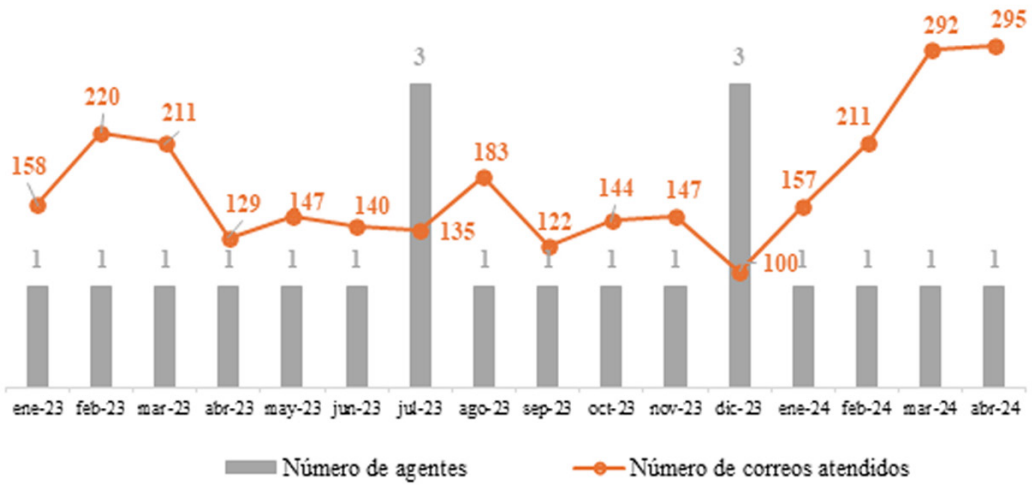
Contexto y Problema

ProColombia es un fideicomiso, del Ministerio de Comercio, Industria y Turismo, encargado de promover el Turismo, la Inversión Extranjera en Colombia, las Exportaciones no minero energéticas y la imagen del país. A través de su red nacional e internacional de oficinas, ofrece apoyo y asesoría integral a los clientes, mediante servicios o instrumentos dirigidos a facilitar el diseño y ejecución de su estrategia de internacionalización, que busca la generación, desarrollo y cierre de oportunidades de negocios.

Exportar no es una tarea fácil para los empresarios colombianos, por lo que la guía de ProColombia es fundamental para promover las exportaciones del país. Por lo tanto, dentro de la Vicepresidencia de Exportaciones se cuenta con el servicio de atención al usuario externo, el cual consiste en dar respuesta a la comunidad empresarial respecto a temas de exportación, de aliados o relacionados con la entidad en general. Para dar respuesta a estas inquietudes, se dispuso del buzón de correo electrónico: exportaciones@procolombia.co a través del cual se recibieron 1.836 correos durante el año 2023.

Este buzón es atendido principalmente por 1 agente, y se anexaron estudiantes en práctica para apoyar la labor en los meses de julio y diciembre de 2023. De acuerdo con la Ilustración No. 1 el número de correos recibidos durante el año 2023 presentó un comportamiento variable donde los meses de mayor cantidad de correos atendidos fueron febrero y marzo con 220 y 211 correos. A partir de enero de 2024 los correos recibidos por la Vicepresidencia de Exportaciones reflejaron una tendencia alcista y los valores mensuales superan el valor mensual promedio del año 2023 que estuvo en 153 correos.

Ilustración No. 1. Estadísticas de correos atendidos



Fuente: Vicepresidencia de Exportaciones ProColombia

El modelo definido para la atención al usuario diseñado en la Vicepresidencia de Exportaciones, inicia con la implementación del documento Guía de Respuestas a Preguntas Frecuentes, el cual contiene respuestas sectorizadas relacionadas con el proceso de exportación con información básica y que cuenta también con información adicional tanto de aliados (por ejemplo, Ministerio de Comercio, Industria y Turismo, DIAN, DANE, entre otros), como de información que puede ser consultada por el usuario relacionada con temas de interés general como participación en ferias o eventos comerciales o de otras áreas de ProColombia. El manejo de esta guía representa una cantidad de tiempo considerable por su gran tamaño y por la información que contiene la cual debe ser revisada de manera juiciosa para que la respuesta sea la indicada para la pregunta y/o preguntas que pueda tener el usuario.

Para llevar a cabo esta labor, se dispone de un agente o persona a cargo de dar respuesta dentro de la Vicepresidencia, (quien además tiene otras funciones), luego de recibir el correo con la consulta, elabora un documento con la información correspondiente que posteriormente envía al empresario. La labor antes descrita hace que los agentes que brindan este servicio extiendan el tiempo de atención con cada usuario por la complejidad del documento y por la necesidad de entregar una respuesta efectiva.

Para ProColombia el servicio de atención al usuario es muy importante, no obstante, tiene oportunidades de mejora debido a la alta rotación del equipo, la curva de aprendizaje lenta y la necesidad de mejorar la precisión en las respuestas que se brindan. Además, de la necesidad de implementar nuevas estrategias tecnológicas que permitan la reducción en tiempos de respuesta.

Inteligencia artificial generativa

De acuerdo con Kulkarni et al (2023) y Corchado et al (2023) la inteligencia artificial generativa es una rama de la inteligencia artificial que se encarga de crear modelos y algoritmos capaces de generar contenido nuevo y original como imágenes, texto, música e incluso videos. Estos modelos difieren de los tradicionales porque pueden ser entrenados para asumir tareas específicas que tienen por objetivo aprender de patrones y comportamientos para generar resultados.

Respecto al procesamiento de texto o procesamiento de lenguaje natural, los modelos generativos pueden usarse para traducir de un idioma a otro, para resumir textos o para crear agentes conversacionales que luego de ser entrenados puedan producir respuestas similares a las brindadas por los humanos.

Modelos GPT

Modelos trasformadores generativos pre-entrenados creados por OpenAI que se basaron en la arquitectura de los transformes introducidos por Vaswani et al (2017) y cuyo modelo GPT-2 fue lanzado en el año 2019 demostrando una gran capacidad para generar texto consistente y realista (Corchado et al 2023). Un año más tarde OpenAI lanzó GPT-3 que mejoró sus capacidades y popularizó las aplicaciones como chat GPT (Abdullah et al 2022).

Modelo GPT-3.5 Turbo

Ofrece la capacidad de ajustar el modelo para mejorar su rendimiento en tareas específicas, como la generación de resúmenes y la extracción de información precisa de documentos. La capacidad de ajuste fino (fine-tuning) permite a los desarrolladores mejorar la precisión y el formato de las respuestas del modelo, lo que es crucial para aplicaciones que requieren un formato de respuesta específico, como la finalización de códigos o la composición de las llamadas API[1]. (Peng et al 2023)

Modelo GPT-3.5 Turbo

GPT-4o ("o" para "omni") es el modelo más avanzado, es multimodal porque acepta entrada de texto o imágenes y genera texto, es mucho más eficiente que GPT-4 Turbo porque genera 2 veces más rápido y es 50% más económico. Este modelo se encuentra disponible para los clientes que cuentan con la suscripción paga. (OpenAI 2024).

Asistentes de la API de OpenAI

Son sistemas de inteligencia artificial desarrollados por OpenAI que utilizan modelos de lenguaje avanzados para ayudar en diversas tareas como la generación de texto, la traducción o la programación, en otras tareas. Estos asistentes pueden interactuar con los usuarios a través de texto o voz para proporcionar respuestas, completar tareas o generar texto basado en instrucciones (OpenAI 2024). Estos sistemas son de gran ayuda para los programadores ya que les permite tener acceso a modelos avanzados como GPT-3,5 o GPT-4o con los que pueden interpretar código, realizar búsquedas en archivos y hacer llamadas a funciones específicas.

Prompts estructurados

Un prompt es una instrucción o conjunto de instrucciones proporcionada a un modelo de inteligencia artificial con la cual se busca la ejecución de una tarea. Estas instrucciones pueden ser preguntas, declaraciones, fragmentos de texto o indicaciones que contextualizan el modelo para la generación de una respuesta.

Una alucinación en el contexto de LLM[2] (Large Language Models) se define como la generación de contenido no esencial y proveniente de una fuente no confiable. (Amatriain, 2024)

Para minimizar la alucinación y para obtener respuestas lo suficientemente estructuradas para interactuar con otros sistemas se utiliza la ingeniería de prompts (Castro, 2024). Esta ingeniería consiste en definir las instrucciones de entrada al modelo con parámetros como el formato a la respuesta esperada, la identificación

[1] Interfaz de programación de aplicaciones.
[2] Modelos de lenguaje a gran escala que son entrenados con enormes cantidades de datos textuales. Pueden generar y comprender texto con un alto nivel de precisión, utilizados en aplicaciones como asistentes virtuales o generación automática de texto.

del query[3] o la tarea, los filtros de seguridad y la citación de las fuentes utilizadas, así como ejemplos de prompts y respuestas esperadas que permiten el entrenamiento del modelo para el paso a producción.

La identificación del query consiste en entender el contexto o la intención del usuario con esto se busca llegar a respuestas precisas. Los filtros de seguridad aseguran que el modelo no responda contenido falso, ofensivo, discriminatorio o que comprometa la privacidad del usuario. A su vez aseguran que no se puedan modificar las instrucciones establecidas. La citación de las fuentes proporciona credibilidad en tanto que el usuario puede verificar la información mediante la referencia y obliga al modelo a usar la fuente suministrada. El formato a la respuesta esperada establece la estructura y la longitud de la respuesta, esto permite extraer fragmentos o secciones específicas del output para su respectivo uso.

Métricas para medir la calidad de generación de texto

Para evaluar la calidad de un modelo de generación de texto se utilizan comúnmente las siguientes métricas:

• **BLEU (Bilingual Evaluation Understudy)**

Esta métrica propuesta por Papineni et al (2002) compara los n-gramas[4] del texto de referencia con los n-gramas del texto generado. Su puntaje se mide en una escala de 0 a 1 donde 1 representa una coincidencia perfecta. Valores de BLEU entre 0.3 y 0.6 se consideran razonablemente buenos en las tareas de generación de texto, y valores superiores al 0.7 indican muy buena coincidencia.

• **ROUGE (Recall- Oriented Understudy for Gisting Evaluation)**

Esta métrica propuesta por Lin (2004) evalúa la calidad de los resúmenes generados al compararlos con textos de referencia creados por humanos. Existen varias variantes de esta métrica que difieren por la medición de la similitud.

• **ROUGE-N:**
Evalúa la superposición de n-gramas ente el texto de referencia y el generado.

• **ROUGE-L:**

Evalúa la longitud de la subsecuencia más larga común entre el texto de referencia y el generado. La L puede tomar el valor de 1 cuando mide la coincidencia de unigramas y de 2 cuando mide la coincidencia de bigramas.
Su puntaje se mide en una escala de 0 a 1, donde 1 representa similitud exacta. Valores de ROUGE entre 0.5 y 0.7 se consideran de una buena calidad.

[3] Consulta o instrucción que un usuario hace a un sistema o base de datos para obtener una respuesta o ejecutar una acción específica.
[4] Secuencia continua de n elementos donde los elementos pueden ser palabras, caracteres o fonemas según el contexto.

Metodología

Para brindar atención oportuna al creciente número de consultas que recibe la Vicepresidencia de Exportaciones de ProColombia, se propuso como solución la creación de un asistente llamado Exporta Express: respuestas inmediatas para oportunidades globales que opera bajo un modelo de generación de respuestas con base en el documento de preguntas frecuentes mediante el uso de modelos de inteligencia artificial generativa para reducir los tiempos de atención al usuario brindando una atención personalizada.

Para el desarrollo de este proyecto capstone se establecieron 4 fases basadas en la metodología CRISP-DM[5]:

Ilustración No. 2. Fases de metodología aplicada

Fuente: Elaboración propia

Entendimiento del negocio

Se realizaron reuniones con la agente de la Vicepresidencia de Exportaciones para identificar los aspectos relevantes a considerar en el diseño y posterior implementación de la solución.

Se identificó en primer lugar que el área recibió 1.836 correos durante el año 2023 y en los primeros 4 meses del año 2024 se recibieron 955 correos, cifra que presentó un incremento del 25% respecto a los correos que se recibieron durante esos mismos meses en el año 2023; y en segundo que el manual de preguntas frecuentes tiene un manejo complejo debido a la variedad de temáticas que maneja.

[5] CRISP-DM (Cross Industry Standard Process for Data Mining): Metodología utilizada para estructurar proyectos de minería de datos, que incluye fases como la comprensión del negocio, análisis de datos, y modelado.

Análisis de la información disponible

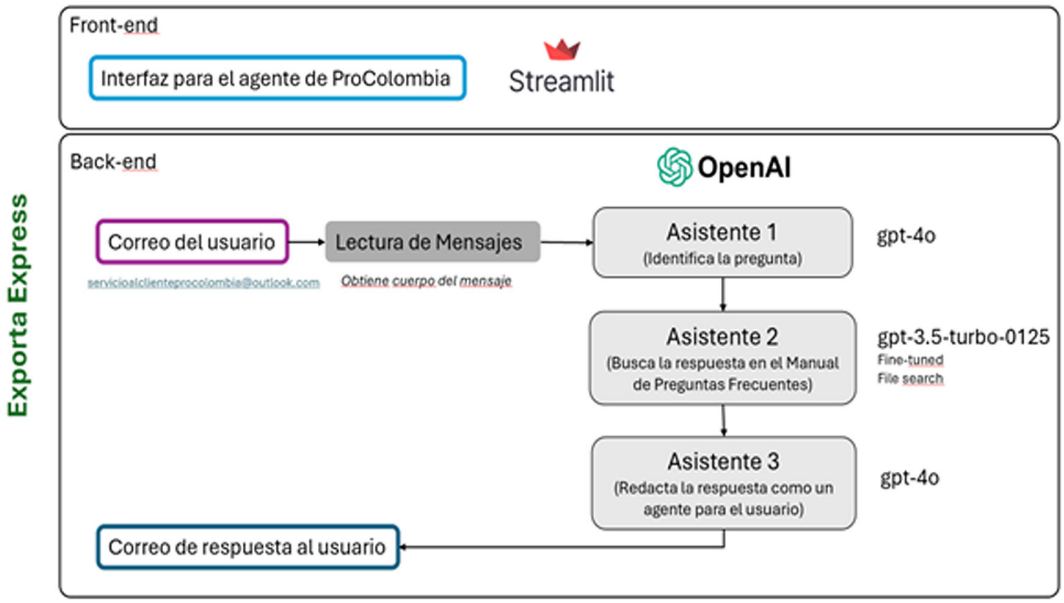
Se inició con la revisión del manual de preguntas frecuentes identificando que requería ciertos ajustes, como la actualización de las referencias de algunas respuestas, la inclusión de preguntas que no estaban, la reclasificación de las categorías como como “Otros” por “Actividades que no son competencia de ProColombia” por indicar un ejemplo, la creación de la sección de “Spam” y la inclusión de la definición cada cadena productiva que atiende ProColombia para darle mayor contexto. Así mismo, se incluyeron los nombres completos de algunas entidades aliadas y se completaron los títulos de las preguntas de acuerdo con el texto de las respuestas para facilitar la búsqueda de las preguntas en el manual y así identificar la respuesta apropiada.

El banco de preguntas y respuestas reales proporcionado por ProColombia reveló que actualmente la Vicepresidencia de Exportaciones remite las respuestas de forma genérica, adjuntando al usuario el texto exacto que trae el manual y no respondiendo de manera personalizada como se busca con la solución propuesta.

Análisis de la información disponible

Para la creación del asistente se diseñó la siguiente arquitectura que se presenta en la ilustración No 3.

Curva ROC



Tasa de falsos positivos (FPR)

Front-end[6]:

Incluye una interfaz gráfica dirigida para el agente de la Vicepresidencia de Exportaciones de ProColombia para que pueda atender más fácilmente las consultas de los usuarios.

Back-end[7]:

Código y uso de asistentes de OpenAI para leer los correos electrónicos de los usuarios, procesar sus datos (nombre, correo electrónico y ubicación), identificar la consulta realizada por el usuario, realizar la búsqueda de la respuesta a la pregunta en el manual de preguntas frecuentes de la Vicepresidencia de Exportaciones, redactar la respuesta para el usuario y remitir dicha respuesta.

Se decidió trabajar con los asistentes de OpenAI para dividir las tareas de procesamiento y especializar a cada uno de estos en la tarea asignada.

Desarrollo de la solución y evaluación

Front-end

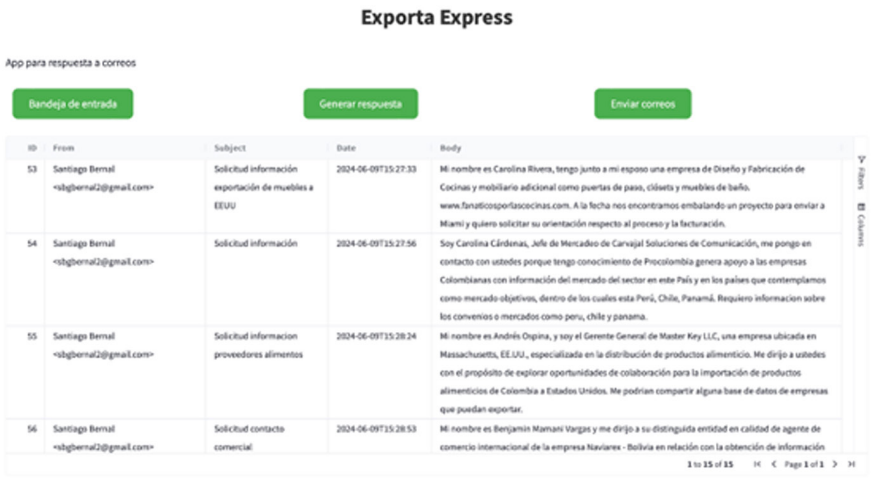
La interfaz amigable para el usuario se creó usando Streamlit[8]. Esta permite al agente supervisar la respuesta brindada por el asistente, revisando el texto generado para realizar ajustes en el caso de que sean necesarios y posteriormente remitir la respuesta al usuario.

La visualización de la interfaz se presenta en la ilustración No 4, donde se aprecia que esta app cuenta con tres botones: Bandeja de entrada, Generar respuesta y Enviar correos.

Al seleccionar el botón de Bandeja de entrada, el agente podrá visualizar en una base de datos la información relevante de los correos recientes que hay en el buzón como son el destinatario (from), el tema (subject), la fecha (date) y el cuerpo del mensaje (body).

[6] La parte visible de una aplicación con la que los usuarios interactúan directamente, como botones, formularios y gráficos. Está diseñada para mostrar información y recoger datos del usuario.
[7] Parte no visible del sistema que gestiona la lógica interna, el almacenamiento de datos y la comunicación con servidores. Procesa la información recibida del front-end y envía respuestas.
[8] Es un framework de código abierto que permite crear aplicaciones usando Python. https://streamlit.io/. Para conocer cómo implementar una solución en streamlit se puede revisar la documentación disponible en https://docs.streamlit.io/deploy/streamlit-community-cloud/deploy-your-app

lustración No. 4. Interfaz desarrollada



Luego deberá seleccionar el botón Generar respuesta para visualizar la respuesta generada por Exporta Express. En ese momento podrá realizar la edición de la respuesta si lo considera pertinente.

En el caso de que decida hacerlo, seleccionará el ID que identifica a la consulta y en la casilla “Respuesta2” podrá realizar el ajuste finalizando con el botón Guardar cambios.

lustración No. 4. Interfaz desarrollada

En este momento la respuesta estará lista para su envío al usuario, así que para finalizar el proceso deberá seleccionar el botón Enviar correos.

Back-end

El código en Python inicia con el ingreso al buzón de correo electrónico para abrir los correos no leídos y extraer la información relevante como los datos del usuario (nombre, correo electrónico[9] y ubicación) y el cuerpo del mensaje.

• **Asistente 1:**

Se le asignó la labor de identificar la pregunta realizada por el usuario. Este asistente recibe el cuerpo del mensaje para procesarlo e identificar la pregunta o preguntas que realiza el usuario.

Este asistente trabaja con el modelo GPT -4o y se seleccionó porque actualmente es el modelo más ágil y es 50% más económico con respecto al modelo GPT-4 (OpenAI 2024).

• **Asistente 2:**

Se le asignó la labor de buscar la respuesta a la pregunta formulada por el usuario, pregunta que identificó el asistente 1, en el manual de preguntas frecuentes.

Para el entrenamiento de este asistente se usó fine-tuning[10] porque se requiere que realice la búsqueda de un tema específico y extraiga la respuesta a la pregunta de acuerdo con la información disponible en el manual de preguntas frecuentes. Se trabajó con un banco de preguntas y respuestas esperadas de 160 registros categorizados en 8 temáticas. Esta información se incluyó en un formato JSONL[11] identificando el rol (usuario o sistema) y el contenido para realizar el entrenamiento usando el modelo GPT-3.5-turbo-0125. Adicionalmente se le cargó el manual de preguntas frecuentes para incorporar estos datos a la base de datos de conocimiento del modelo y se limitó a que las respuestas generadas fueran con base en ese documento. Para este asistente se seleccionó el modelo GPT-3.5-

Turbo por la capacidad que tiene de ajustar el modelo para mejorar la generación de resúmenes y la extracción de información precisa de documentos (Peng et al 2023).

• **Asistente 3:** se le asignó la tarea de redactar la respuesta a la pregunta del usuario como lo haría un agente de la Vicepresidencia de Exportaciones. Este asistente recibe la respuesta que el asistente 2 encontró y ajusta su redacción.

Este asistente trabaja con el modelo GPT -4o.

La respuesta generada por el asistente 3 se remite por correo electrónico al usuario.

Con el propósito de revisar la calidad en la generación de respuestas por parte de los asistentes, se midieron las métricas BLEU y ROUGE_L ejecutando los modelos con una base de 15 solicitudes de diferentes temáticas. Se obtuvieron los siguientes resultados:

[9] Se creó el buzón servicioalclienteprocolombia@outlook.com para el desarrollo de la solución.
[10] Para ampliar el paso a paso del proceso de fine-tuning se puede revisar la documentación de OpenAI https://platform.openai.com/docs/guides/fine-tuning
[11] JSON Lines: formato de datos en el que cada línea representa un objeto JSON individual. Es útil para procesar grandes volúmenes de datos de forma secuencial y es comúnmente utilizado en aprendizaje automático. Un JSON individual es un objeto JSON con pares clave-valor que representa una entidad o registro específico. Este formato es ideal para procesar grandes volúmenes de datos de forma secuencial, ya que permite manejar cada objeto de manera independiente, facilitando el procesamiento y almacenamiento.

Tabla No 1

Métrica	ROUGE-L	BLEU
Count	15	15
Mean	0,229	0,110
Std	0,098	0,076
Min	0,078	0,040
25%	0,170	0,063
50%	0,188	0,075
75%	0,295	0,159
Max	0,405	0,265

Fuente: Elaboración propia

Respecto a la métrica ROUGE-L los resultados estuvieron entre 0,078 y 0,405, valores bajos debido a la longitud de las respuestas. Sin embargo, al validar la calidad de la respuesta generada respecto a la respuesta del guía de preguntas frecuentes, la agente de ProColombia encontró consistentes las respuestas generadas y acordes con la respuesta que usualmente ella daría al usuario.

Respecto a la métrica BLEU los resultados estuvieron entre 0,040 y 0,265, valores bajos dada la definición de la métrica porque busca coincidencias exactas. Aquí nuevamente es pertinente mencionar que el asistente fue diseñado para brindar respuestas personalizadas y no genéricas como está estructurado el manual de preguntas frecuentes.

CONCLUSIONES

El ajuste del modelo mediante fine-tuning ha demostrado ser eficaz para mejorar el rendimiento en la generación de las respuestas en tareas específicas, al entrenar el modelo con consultas reales se logra mayor precisión en las respuestas generadas. Se sugiere ampliar los grupos de temáticas con más ejemplos para así fortalecer el ajuste del modelo.

Uno de los elementos claves en las respuestas de los modelos LLM son las instrucciones de entrada, estas permiten estructurar las respuestas para su posterior uso. En esta estructura de instrucciones se incluyen ejemplos que le permiten al modelo entender las condiciones de uso esperadas, al ampliar estos ejemplos o instrucciones de entrada se ajusta el modelo y se espera obtener mejores respuestas. Por ello, se sugiere incluir una mayor cantidad de ejemplos en las instrucciones para que el modelo mejore la calidad de las respuestas.

Las métricas ROUGE y BLEU no presentan resultados adecuados debido a la longitud de los textos, ya que están diseñadas para evaluar textos cortos y se centran en la medición de la exactitud. Los textos generados eran extensos y presentaron una alta variabilidad en la coincidencia. Se sugiere revisar literatura adicional para buscar métricas más apropiadas que puedan evaluar mejor consistencia y coherencia en textos extensos.

Uno de los retos que surgieron durante el desarrollo de la solución fue la selección del modelo adecuado, ya que cada modelo está optimizado para tareas específicas. Los primeros modelos que probamos no lograron cumplir de manera efectiva con las expectativas, lo que nos obligó a evaluar diferentes opciones antes de encontrar el más adecuado. Este proceso de prueba y ajuste fue crucial para mejorar la precisión en las respuestas generadas.

Para mejorar el desempeño del modelo y la precisión de las citas, el texto de referencia se sugiere dividirlo por temáticas para facilitar que el modelo sea capaz de generar respuestas más precisas y a su vez mejore la calidad de la citación.

La implementación de modelos LLM es útil para agilizar la respuesta de tareas operativas debido a que mejorar el nivel de servicio ya que el usuario tendrá la respuesta a su inquietud en un menor tiempo y libera al agente para que pueda enfocarse en otras tareas.

La solución de atención al usuario mediante LLM tiene la capacidad de ser escalable adaptándose a los requerimientos de la entidad asegurando que todas las consultas sean atendidas de manera oportuna con la capacidad actual del recurso humano asignado a esta labor.

Las herramientas de LLM al estar en constante actualización implica que es necesario evaluar constante los modelos implementados para verificar cambios en funcionalidades de los modelos, estas actualizaciones frecuentemente introducen mejoras en rendimiento, precisión y costo de los modelos.

Referencias

Abdullah, M., Madain, A., & Jararweh, Y. (2022). ChatGPT: Fundamentals, applications and social impacts. In 2022 Ninth International Conference on Social Networks Analysis, Management and Security (SNAMS) (pp. 1-8).

Amatriain, X. (2024). Measuring and Mitigating Hallucinations in Large Language Models: A Multifaceted Approach.

Corchado, J. M., López, S., Garcia, R., & Chamoso, P. (2023). Generative Artificial Intelligence: Fundamentals. ADCAIJ: Advances in Distributed Computing and Artificial Intelligence Journal, 12(1), e31704-e31704.

Kulkarni, A., Shivananda, A., Kulkarni, A., & Gudivada, D. (2023). Applied Generative AI for Beginners. Apress, Berkeley, CA.

Lin, C. Y. (2004). Rouge: A package for automatic evaluation of summaries. In Text summarization branches out.

OpenAI. (2024). Assistants API. OpenAI. https://platform.openai.com/docs/assistants/overview

OpenAI. (2024). Fine-tuning. OpenAI. https://platform.openai.com/docs/guides/fine-tuning

OpenAI. (2024). GPT-4o. OpenAI. https://platform.openai.com/docs/models/gpt-4o

Papineni, K., Roukos, S., Ward, T., & Zhu, W. J. (2002). Bleu: a method for automatic evaluation of machine translation. In Proceedings of the 40th annual meeting of the Association for Computational Linguistics (pp. 311-318).

Peng, A., Wu, M., Allard, J., Kilpatrick, L., & Heidel, S. (2023). GPT-3.5 Turbo fine-tuning and API updates. OpenAI. https://openai.com/index/gpt-3-5-turbo-fine-tuning-and-api-updates/

Snowflake (2024) Streamlit. https://docs.streamlit.io/deploy/streamlit-community-cloud/deploy-your-app

Tamayo, M. (2013). El proceso de la investigación científica (Cuarta Edición ed.). Limusa, México: Editorial Limusa, SA de CV.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., ... & Polosukhin, I. (2017). Attention is all you need. Advances in neural information processing systems, 30.

ECONOMIA

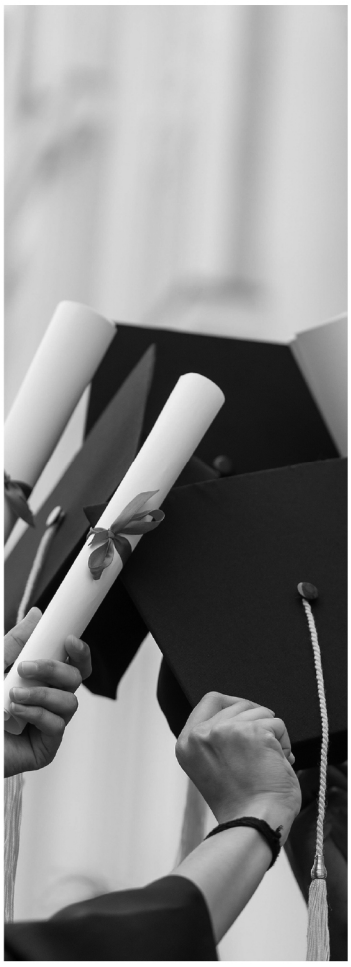
Laboratorio de inteligencia artificial aplicada a Economía

Universidad Externado de Colombia

FACULTAD DE ECONOMÍA

ECONOMIA

Laboratorio de inteligencia artificial aplicada a Economía



Universidad Externado de Colombia
FACULTAD DE ECONOMÍA

ECONOMIA
Laboratorio de inteligencia artificial aplicada a Economía

Rutas inteligentes:

Prediciendo la oferta laboral para transformar políticas públicas aplicadas al mecado de trabajo

Autor

Juan Esteban Londoño Guatibonza juan.londono4@est.uexternado.edu.co

2025

Resumen

Este estudio utiliza técnicas avanzadas de inteligencia artificial, específicamente el modelo Gradient Boosting Regressor, para predecir el número de graduados universitarios según el área educativa. El objetivo principal es proporcionar datos confiables para orientar la planificación estratégica en políticas públicas laborales, facilitando una alineación más eficiente entre educación superior y mercado laboral. La metodología empleada incluye la optimización del modelo mediante búsqueda aleatoria de hiperparámetros y validación cruzada, considerando variables demográficas, económicas y académicas como población, Producto Interno Bruto (PIB), área del conocimiento, matriculados y docentes. Los resultados revelan una elevada precisión predictiva, avalada por métricas robustas (R^2, MSE, MAE). Además, el estudio subraya que la sincronización efectiva de políticas públicas con la oferta laboral puede impulsar el crecimiento económico del país al mejorar las oportunidades de empleo y potenciar la calidad del capital humano.

La relación entre educación superior y mercado laboral es un factor clave en el desarrollo económico y social de cualquier país. En Colombia, existe la necesidad urgente de implementar estrategias que conecten adecuadamente la oferta educativa con las demandas laborales reales, asegurando así oportunidades efectivas para los graduados y contribuyendo al progreso económico nacional. La inteligencia artificial (IA) ofrece una solución innovadora al permitir predicciones precisas sobre la cantidad de graduados en distintas áreas educativas, proporcionando así herramientas esenciales para la planificación estratégica y la formulación de políticas laborales más efectivas.

Este trabajo busca precisamente abordar este desafío mediante la aplicación del modelo Gradient Boosting Regressor, reconocido por su alto rendimiento en tareas predictivas complejas. Al considerar diversas variables relevantes—tales como datos demográficos, indicadores económicos y características propias del sistema educativo—el estudio no solo aporta información sobre el potencial educativo del país, sino que también señala

posibles rutas para mejorar la gestión pública en materia laboral.

A continuación, se describe en detalle la metodología empleada y el análisis llevado a cabo en esta investigación.

Para la realización de este estudio se utilizó un enfoque de inteligencia artificial basado en el modelo Gradient Boosting Regressor, reconocido por su precisión y robustez en problemas predictivos complejos. La base de datos comprende 43,776 observaciones (2018-2023) y 18 variables que recogen información demográfica, económica y académica.

En la fase de preprocesamiento, se llevó a cabo una revisión exhaustiva para identificar valores atípicos y valores faltantes. En particular, las variables con ausencias por debajo del 5% fueron imputadas mediante la mediana, evitando la distorsión que puede generar la media en datos sesgados. Cuando el porcentaje de datos ausentes superó ese umbral, se valoró la pertinencia de excluir la variable o emplear métodos más avanzados (por ejemplo, K-Nearest Neighbors) si su aporte explicativo era significativo. Para atenuar el efecto de valores extremos en variables numéricas, se implementó un escalado robusto.

Además, se trabajó de forma detallada con la variable categórica "AREA_CON" (área del conocimiento) a través de Procesamiento de Lenguaje Natural (PNL), garantizando que cada carrera universitaria quedase asignada a la categoría adecuada. En primer lugar, se extrajo la denominación de cada programa o carrera. Luego, se procedió con técnicas de tokenización para segmentar el texto en palabras clave y posteriormente se aplicó un proceso de clasificación semiautomático, el cual consistía en asociar términos específicos a una de las siguientes áreas:

• **Artes:**

Carreras relacionadas con la expresión creativa (artes plásticas, música, danza, diseño, entre otras).

• **Ciencias Sociales y Humanas:**

Abarca disciplinas centradas en el estudio de la sociedad, cultura y comportamiento humano (sociología, derecho, psicología, historia, etc.).

• **Ciencias Exactas y Naturales:**

Concentra carreras relacionadas con matemáticas, física, química, biología, ingeniería, y áreas afines.

Posteriormente, se vectorizó cada categoría en una representación que pudiera ser procesada por el modelo. Por razones de interpretables y para no perder matices semánticos, se optó por one-hot encoding, el cual transformó estas áreas en columnas binarias (1 o 0) sin imposiciones de un orden jerárquico.

La construcción de cada variable proviene de fuentes oficiales y se ajustó al contexto local:

• **"MATRICULADOS_P"**

Sirvió de proxy para capturar la demanda educativa histórica, considerando la cantidad de estudiantes matriculados en primer curso entre 2014 y 2018.

• **"NACIMIENTOS"**

Se basó en registros de 2005 a 2010 para aproximar la reserva generacional que podría llegar a la educación superior.

• **Variables económicas**

(PIB, INFLACION ANUAL%, INFLACION TOTAL) recogen fluctuaciones del entorno financiero.

• **Variables demográficas**

(POBLACION, DEFUNCIONES) evidencian cambios en la población que inciden en la oferta y la demanda educativa.

• **Variables socioeconómicas**

(HOG_CABECERA, HOG_CPRD, HOG_TOTAL, CASOS_V) permiten comprender factores de calidad de vida y seguridad.

El modelo se optimizó con RandomizedSearchCV para ajustar hiperparámetros como la profundidad y el número de árboles, así como la tasa de aprendizaje. Asimismo, la validación cruzada de 10 particiones (10-fold) permitió evaluar la consistencia de las predicciones en distintos subconjuntos de la muestra. La solidez del modelo en la estimación de graduados universitarios se verificó mediante métricas (R^2, MSE, MAE), lo cual respalda su utilidad para la formulación de políticas públicas laborales dirigidas a alinear la oferta educativa con el mercado de trabajo.

En esta sección se exponen los principales hallazgos del modelo Gradient Boosting Regressor, acompañados de gráficas que ilustran su rendimiento y fiabilidad para predecir el número de graduados universitarios en distintas áreas de conocimiento. A lo largo de esta exposición, se explica la pertinencia de cada métrica y visualización en la formulación de políticas públicas orientadas al crecimiento económico.

Evaluación del Desempeño del Modelo

Métricas Principales (R^2, MSE y MAE): Tras la fase de optimización y validación, se obtuvieron resultados sólidos en el conjunto de test. Un R^2 alto (~0.8959) indica que la mayoría de la variación en el número de graduados se explica mediante las variables incluidas en el modelo. El MSE (~0.1368) y el MAE (~0.2055) reflejan un bajo nivel de error promedio, lo que sugiere que el modelo no solo identifica correctamente las tendencias generales, sino que también aproxima adecuadamente la magnitud de los graduados.

Para la formulación de políticas, una predicción precisa de la oferta de graduados es esencial: un buen ajuste (R^2 elevado) implica que los tomadores de decisiones pueden basarse en estos datos para prever la demanda laboral y planificar recursos en áreas estratégicas del país.

Validación Cruzada (10-Fold): Al dividir la muestra en 10 subconjuntos distintos, se comprobó la estabilidad del modelo. La consistencia observada en las métricas a través de cada fold aumenta la confianza en que las predicciones funcionan no solo para unos pocos municipios o períodos específicos, sino de manera amplia. Esta robustez favorece el diseño de programas de empleo o incentivos que tengan aplicabilidad transversal en diferentes regiones o sectores del país.

Tabla 1: Resultados de Validación Cruzada (escala normalizada)

Fold	R^2	MSE	MAE
Fold 1	0.8122	0.2573	0.2258
Fold 2	0.8289	0.1897	0.2311
Fold 3	0.9024	0.1172	0.1992
Fold 4	0.9168	0.0868	0.1717
Fold 5	0.8804	0.1720	0.2152
Fold 6	0.8550	0.1759	0.2257
Fold 7	0.9139	0.1055	0.1811
Fold 8	0.8961	0.1150	0.2052
Fold 9	0.8701	0.1282	0.2234
Fold 10	0.8773	0.1625	0.1958
Promedio	**0.8753**	**0.1510**	**0.2074**

Fuente: Elaboración propia

Tabla 2: Resultados en el Conjunto de Prueba (escala normalizada)

Conjunto de Datos	R^2	MSE	MAE
Test	0.8959	0.1368	0.2055

Fuente: Elaboración propia

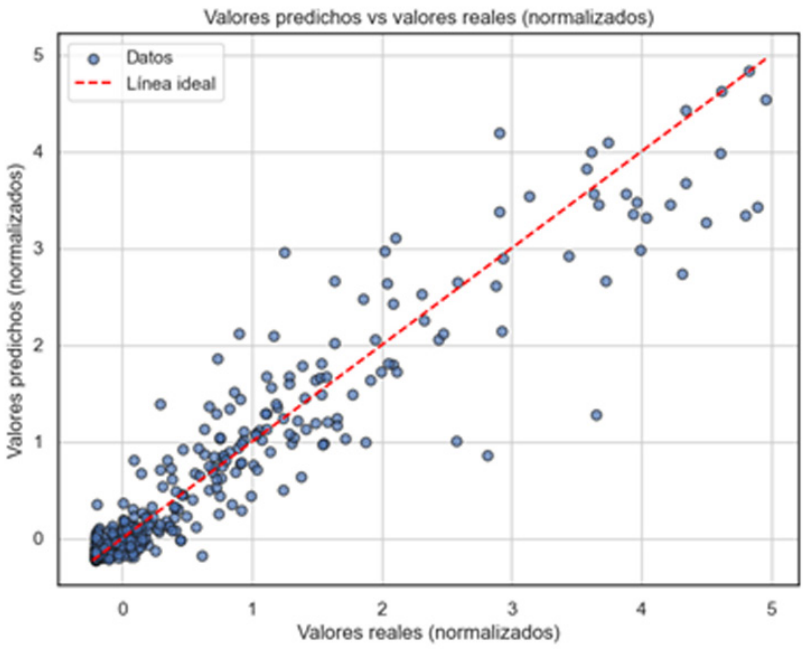
Se muestra en un cuadro o gráfico el promedio de R^2, MSE y MAE para cada iteración de validación cruzada, evidenciando la solidez del modelo y permitiendo comparaciones rápidas entre escenarios. Cuanto más homogéneas sean estas métricas a través de los folds, mayor será la fiabilidad en la toma de decisiones.

Comparación de valores predichos vs. valores reales

Para evaluar la precisión del modelo Gradient Boosting Regressor, resulta fundamental comparar las predicciones generadas con los datos reales.

A continuación, se presenta el análisis principal: En la Gráfica de Dispersión (Scatter Plot), los valores observados se ubican en el eje X, mientras que las predicciones del modelo se representan en el eje Y.

La línea diagonal indica la predicción perfecta.



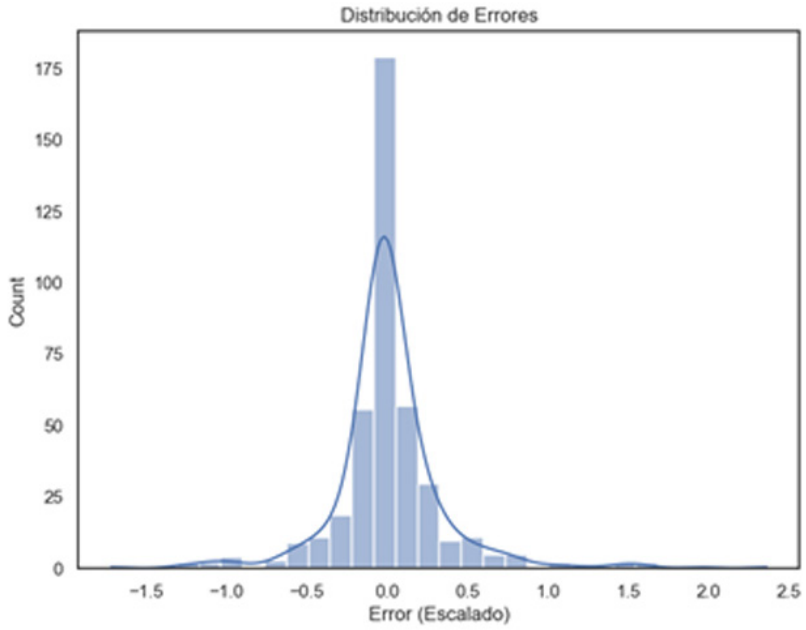
Cuando la mayoría de los puntos se concentra en torno a esta diagonal, se demuestra un elevado nivel de acierto. Una dispersión considerable en ciertas zonas sugiere la necesidad de refinar variables o ajustar la metodología, especialmente si coincide con características sociodemográficas o económicas específicas.
Disponer de predicciones cercanas a los valores reales permite a los entes gubernamentales y educativos focalizar recursos con mayor eficacia. Por ejemplo, si el modelo identifica discrepancias notables en determinadas regiones, pueden asignarse fondos adicionales para infraestructura, becas o programas que fortalezcan la vinculación universidad-empresa. Esta toma de decisiones basada en datos disminuye el riesgo de asignaciones ineficientes y promueve un crecimiento económico más robusto.

Si se detectan municipios con alta densidad poblacional y baja correlación entre valores reales y predichos, esto podría indicar la necesidad de intervenir con urgencia. También se podrían requerir variables adicionales (p. ej., indicadores de desigualdad o inseguridad) para perfeccionar el modelo.

Análisis de errores

El análisis de errores proporciona detalles sobre la diferencia entre valores estimados y observados. A continuación, se describe la principal herramienta para ello:

El Histograma de Errores muestra cómo se distribuye la diferencia (error) entre los valores reales y los predichos por el modelo.

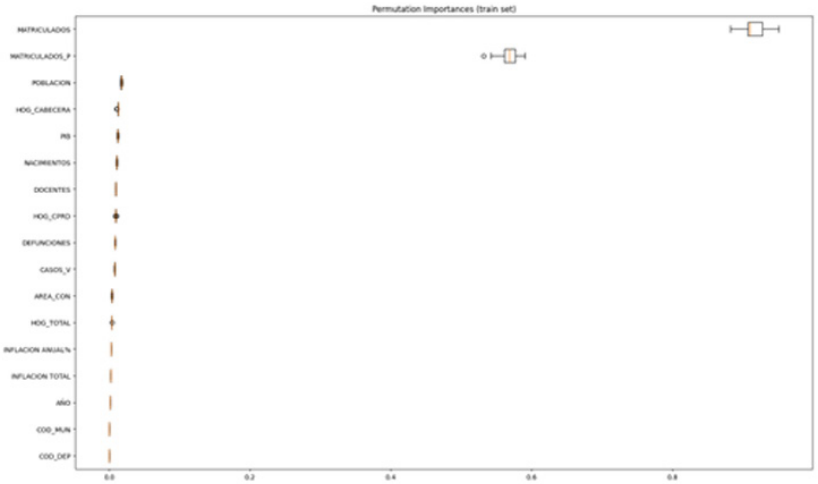


Una distribución concentrada en torno a cero, con baja dispersión, indica que el modelo no presenta sesgos significativos de sobreestimación ni subestimación. Cuando se observan colas alargadas o picos atípicos, podría haber grupos poblacionales o periodos de tiempo que generen mayores desajustes.

Identificar los escenarios con errores más pronunciados (outliers) ayuda a delimitar intervenciones específicas. Un municipio con una economía inestable o un entorno socioeconómico complejo puede requerir soluciones diferenciadas, como incentivos para la retención estudiantil, inversión en seguridad o subvenciones para familias de bajos recursos. El rango intercuartílico y la detección de valores atípicos son métodos útiles para detectar estos casos extremos, en los que el error se desvía de la media de manera significativa.

Importancia de características

El modelo Gradient Boosting Regressor permite evaluar la relevancia de cada variable en el resultado final. Uno de los métodos más ilustrativos es el Permutation Importance, que mide cuánto se resiente la calidad del modelo al alterar aleatoriamente los valores de una variable.

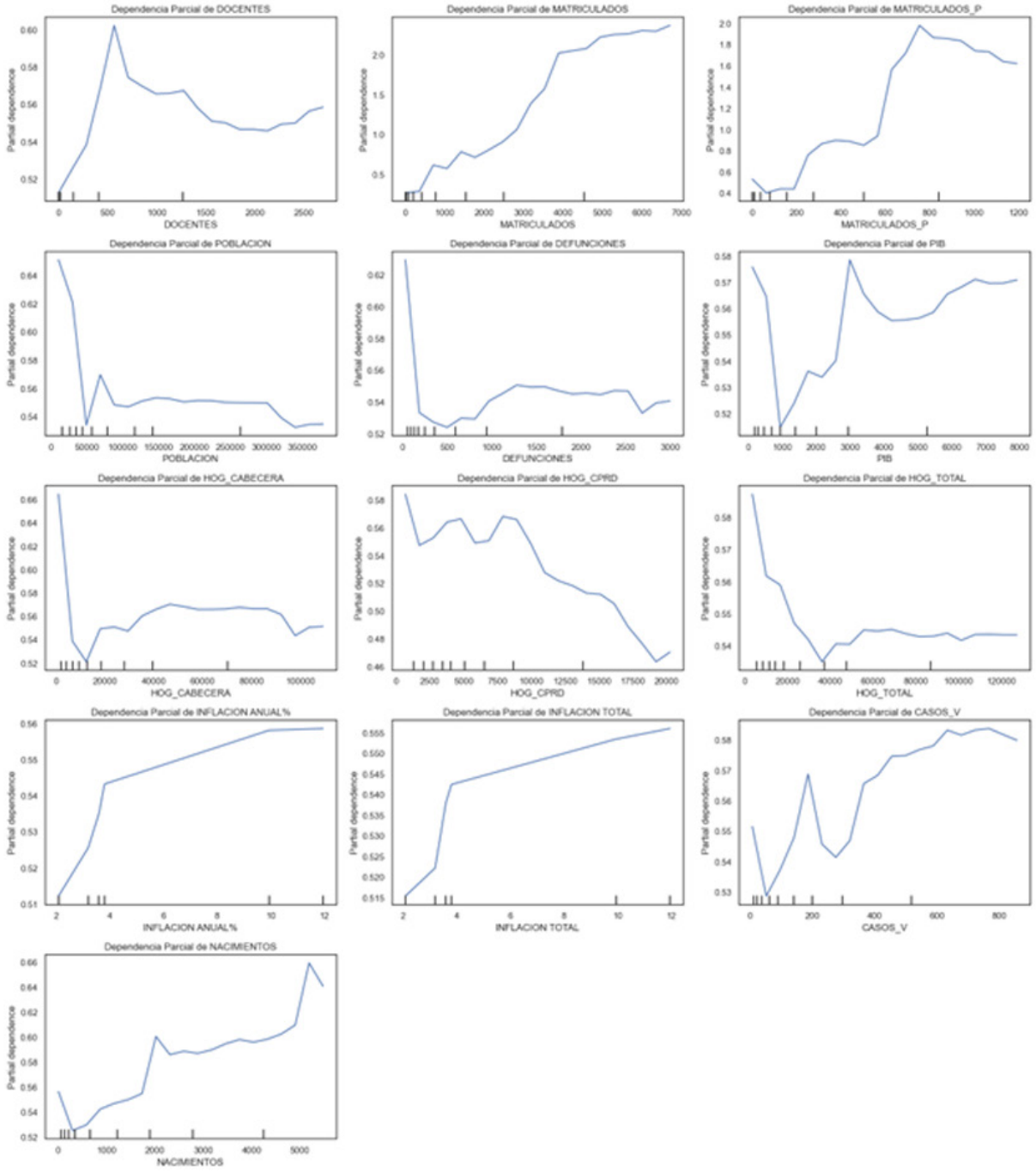


Variables demográficas, como la población total (POBLACION), suelen incidir de forma destacada en el número de graduados. Indicadores económicos (PIB, INFLACION) también muestran un peso significativo, evidenciando la relación entre el dinamismo financiero y las oportunidades educativas. También los Factores académicos (DOCENTES, MATRICULADOS) influyen directamente, dado que reflejan la capacidad de las instituciones y la demanda existente.

Si se detecta un impacto notorio de CASOS_V (casos de violencia) o DEFUNCIONES, se enfatiza el rol que tienen la seguridad y la calidad de vida en la formación superior.

Análisis de dependencia parcial y explicación local

Además de conocer la importancia global de las variables, es conveniente profundizar en la forma en que cada una afecta la predicción, tanto a nivel global como en casos puntuales, con las Gráficas de Dependencia Parcial (Partial Dependence Plots).



• DOCENTES:

Un aumento moderado suele correlacionarse con más graduados; no obstante, tras cierto límite, el efecto tiende a estabilizarse.

• MATRICULADOS y MATRICULADOS_P:

Una mayor cantidad de estudiantes inscritos anticipa un incremento proporcional en graduados, aunque pueden presentarse excepciones según la región.

• POBLACION y NACIMIENTOS:

En zonas con alta densidad demográfica o un mayor número de nacimientos, el modelo tiende a predecir más graduados.

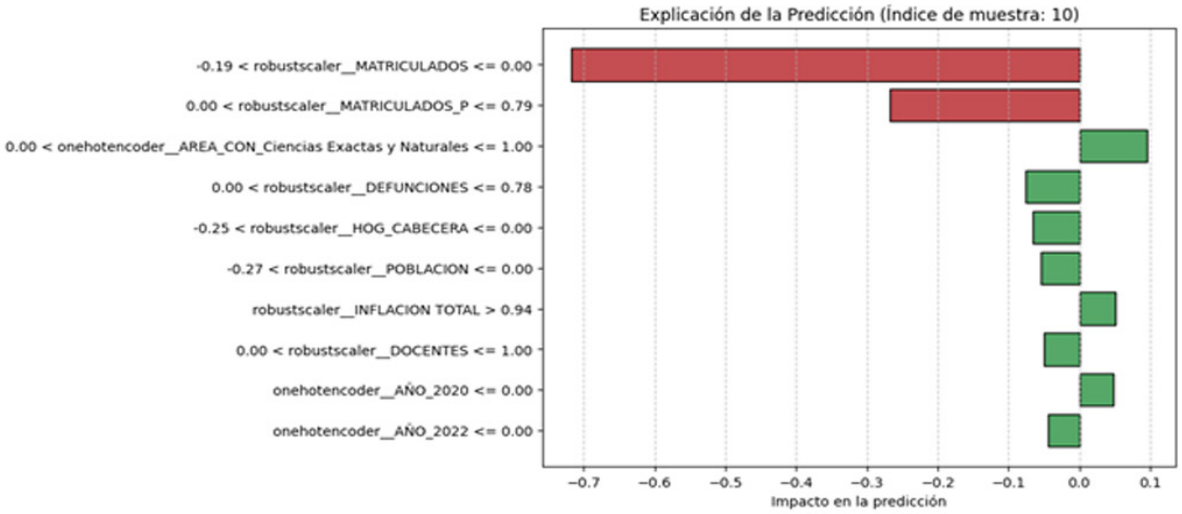
• PIB e INFLACION:

Un crecimiento económico moderado impulsa la formación superior, pero la inflación desbordada puede frenar este desarrollo.

• HOGARES (HOG_CABECERA, HOG_CPRD, HOG_TOTAL) y DEFUNCIONES:

Cambios en la configuración de hogares o variaciones en la tasa de mortalidad afectan en menor medida, pero no deben ignorarse.

Mediante métodos como SHAP o LIME, se comprende cómo cada variable influye en la predicción de un caso específico.



• Un alto número de MATRICULADOS en un municipio podría reducir la predicción de graduados si se combina con otros factores adversos, como infraestructuras limitadas o escasez de docentes especializados.

• La pertenencia a un área académica en particular (p. ej., Ciencias Exactas y Naturales) puede elevar las proyecciones de graduados, sobre todo cuando existen laboratorios y recursos adecuados.

• La población y los hogares en cabecera, combinados con índices de seguridad o bajo nivel de defunciones, suelen favorecer la persistencia de los estudiantes hasta la graduación.

• Una INFLACION TOTAL moderada implica costos educativos relativamente estables, mientras que, en casos de inflación elevada, es común que disminuya la continuidad académica.

En síntesis, el análisis detallado de las características globales y locales de la modelo habilita una toma de decisiones más inteligente y focalizada. Tanto las instituciones de educación superior como las autoridades pueden identificar

áreas críticas donde se necesitan recursos adicionales, minimizar el impacto de la inflación o priorizar zonas con alto crecimiento poblacional. Con ello, se propicia un uso estratégico de los recursos y se potencia el impacto positivo en la formación de graduados, contribuyendo al desarrollo económico y social.

El uso de técnicas avanzadas de inteligencia artificial, en particular el modelo Gradient Boosting Regressor, permite estimar con alta precisión el número de graduados universitarios y, a su vez, orientar con mayor eficacia las políticas públicas. Gracias a la combinación de análisis global (métricas de desempeño, validación cruzada, importancia de características) y local (gráficas de dependencia parcial, SHAP o LIME), se logra una visión más completa de los factores que inciden en el volumen de graduados. Esto habilita la planificación estratégica en áreas críticas —como infraestructura educativa, seguridad y apoyo socioeconómico— promoviendo un uso óptimo de los recursos y evitando planes genéricos que ignoren la diversidad regional.

En este sentido, el foco en variables clave (PIB, población, número de matriculados, entre otras) sienta las bases para intervenciones específicas, potenciando áreas con alta demanda y mejorando las condiciones donde existan brechas o riesgos, como la violencia o la inflación desbordada. De igual modo, el hallazgo de que ciertas disciplinas (por ejemplo, Ciencias Exactas y Naturales) tienen una correlación positiva con la formación de capital humano de calidad, impulsa la inversión en laboratorios y programas de apoyo. Adicionalmente, la incorporación de explicaciones locales refuerza la pertinencia de adaptar las acciones a cada municipio o región, lo que fomenta la equidad en el acceso y permanencia en la educación superior.

En última instancia, la alineación de la oferta formativa con las demandas laborales respaldada por datos rigurosos aporta beneficios tanto para el individuo, al favorecer su empleabilidad, como para el país, al robustecer el crecimiento económico.

De esta manera, se configura un círculo virtuoso en el cual la formación universitaria de calidad y la aplicación efectiva de políticas focalizadas generan un impacto social y productivo sostenible.


IA
ECONOMIA
Laboratorio de inteligencia artificial aplicada a Economia