RED IBEROAMERICANA
DE ACADEMIAS DE INVESTIGACIÓN
RED IBAI

Applications of Data Science in Multi, Inter, and Transdisciplinary Environments:

Machine Learning

Editorial Coordinators:
Judith Guadalupe Montero Mora
Ángel Juan Sánchez García
Zoylo Morales Romero

Applications of Data Science in Multi, Inter, and Transdisciplinary Environments: Machine Learning

COORDINADORES

Judith Guadalupe Montero Mora
Ángel Juan Sánchez García
Zoylo Morales Romero

AUTORES

Javier Zair Galán Trujillo, María Sayuri Baltazar Pedraza, Oscar Alberto Dominguez Ortega, Angel Fernando Argüello Ortiz, Cristian Axel Izquierdo López, Laura Yoselin Tonil De La Luz, José Juan Muñoz León, Zoylo Morales Romero, Ángel Martínez Mendez, Eidan Enrique Juarez García, Dalia Abigail Uscanga Gutierrez, Ángel Juan Sánchez García, Mariela Denisse García Jimenez, Victoria Saldaña Hernández, Judith Guadalupe Montero Mora, Alejandro López Becerra, Iara Idalia Zapot Juárez, Jorge Luis Ortiz Libreros, Joshua Iván Olivares Blanco, José Vidal Hernández, Cristian Eduardo Montesino Fernández, Diego Iván González Colorado, Daniela Burgueño Theurel.

TRADUCCIÓN

Daniela Burgueño Theurel
Joana Paloma Lavoignet Trujillo

PROLOGUISTAS:

Angel Fernando Argüello Ortiz
Angel Juan Sánchez García
José Juan Muñoz León

©RED IBEROAMERICANA DE ACADEMIAS DE INVESTIGACIÓN A.C. 2026

EDITA: RED IBEROAMERICANA DE ACADEMIAS DE INVESTIGACIÓN A.C
DUBLÍN 34, FRACCIONAMIENTO MONTE MAGNO
C.P. 91190. XALAPA, VERACRUZ, MÉXICO.
CEL 2282386072
www.redibai.org
redibai@hotmail.com

Sello editorial: Red Iberoamericana de Academias de Investigación, A.C. (978-607-5893)
Primera Edición, Xalapa, Veracruz, México.
No. de ejemplares: 2
Presentación en medio electrónico digital
Formato PDF 6 MB
Fecha de aparición 02/09/2026
ISBN 978-607-5893-96-9

ISBN: 978-607-5893-96-9
9 786075 893969

Xalapa, Veracruz. México a 25 de agosto de 2026

DICTAMEN EDITORIAL

La presente obra fue arbitrada y dictaminada en dos procesos; el primero, fue realizado por el **COMITÉ EDITORIAL RED IBEROAMERICANA DE ACADEMIAS DE INVESTIGACIÓN A.C.** con sede en México; que sometió a los capítulos incluidos en la obra a un proceso de dictaminación a doble ciego para constatar de forma exhaustiva la temática, pertinencia y calidad de los textos en relación a los fines y criterios académicos de la misma, cumpliendo así con la primera etapa del proceso editorial. El segundo proceso de dictaminación estuvo a cargo del **COMITÉ CIENTÍFICO RED IBEROAMERICANA DE ACADEMIAS DE INVESTIGACIÓN A.C.;** donde se seleccionaron expertos en el tema para la evaluación de los capítulos de la obra y se procedió con el sistema de dictaminación a doble ciego. Cabe señalar que previo al envío a los dictaminadores, todo trabajo fue sometido a una prueba de detección de plagio. Una vez concluido el arbitraje de forma ética y responsable y por acuerdo del Comité Editorial y Científico de la Red Iberoamericana de Academias de Investigación A.C. (REDIBAI), se dictamina que la obra **"APPLICATIONS OF DATA SCIENCE IN MULTI, INTER, AND TRANSDISCIPLINARY ENVIRONMENTS: MACHINE LEARNING"** cumple con la relevancia y originalidad temática, la contribución teórica y aportación científica, rigurosidad y calidad metodológica, actualidad de las fuentes que emplea, redacción, ortografía y calidad expositiva.

Dr. Daniel Armando Olivera Gómez
Director Editorial
Sello Editorial: Red Iberoamericana de Academias de Investigación, A.C.
(978-607-5893)
Dublín 34, Residencial Monte Magno
C.P. 91190. Xalapa, Veracruz, México.
Cel 2282386072

Xalapa, Veracruz. México a 02 de septiembre de 2026

CERTIFICACIÓN EDITORIAL

RED IBEROAMERICANA DE ACADEMIAS DE INVESTIGACIÓN A.C. (REDIBAI) con sello editorial **No. 978-607-5893** otorgado por la Agencia Mexicana de ISBN, hace constar que el libro **"APPLICATIONS OF DATA SCIENCE IN MULTI, INTER, AND TRANSDISCIPLINARY ENVIRONMENTS: MACHINE LEARNING"** registrado con el **ISBN 978-607-5893-96-9** fue publicado por nuestro sello editorial con fecha de aparición del 02 de septiembre de 2026 cumpliendo con todos los requisitos de calidad científica y normalización que exige nuestra politica editorial.

Fue evaluado por pares académicos externos y aprobado por nuestro Comité Editorial y Científico.

Todos los soportes concernientes a los procesos editoriales y de evaluación se encuentran bajo el poder Editorial de **RED IBEROAMERICANA DE ACADEMIAS DE INVESTIGACIÓN A.C. (REDIBAI),** los cuales están a disposición de la comunidad académica interna y externa en el momento que se requieran.

La normativa editorial y repositorio se encuentran disponibles en la página **http://www.redibai-myd.org**

Doy fe.
Dr. Daniel Armando Olivera Gómez
Director Editorial
Sello Editorial: Red Iberoamericana de Academias de Investigación, A.C.
(978-607-5893)
Dublín 34, Residencial Monte Magno
C.P. 91190. Xalapa, Veracruz, México.
Cel 2282386072

DICTAMINADORES

Dra. Julia Aurora Montano Rivas
Doctora en Matemáticas. SNI-C.

Dra. Cecilia Cruz López.
Doctora en Investigación Educativa. SNI-C.

Dr. Saúl Domínguez Isidro.
Doctor en Inteligencia Artificial. SNI-1.

Dr. Héctor Xavier Limón Riaño.
Doctor en Inteligencia Artificial. SNI-1.

Dr. Juan Andrés Sánchez García.
Doctor en Arquitectura y Urbanismo. SNI-C

Colaboración entre los cuerpos académicos UV-CA-127 Ingenieria y Tecnología de Software, UV-CA-107 Metodología y Aplicaciones de las Técnicas y Modelos Estadísticos, UV-CA-452 Arquitectura y Urbanismo para el Desarrollo y el CA Aplicaciones de la Ciencia de Datos en Entornos Multi, Inter y Transdisciplinarios, de la Universidad Veracruzana en México.

ÍNDICE

Book Presentation 3

Presentación del libro 5

Technical Presentation 7

Presentación técnica 9

CHAPTER 1: Classifying student academic performance based on the use of social media applying Logistic Regression and Decision Trees 11

Javier Zair Galán Trujillo, María Sayuri Baltazar Pedraza,
Oscar Alberto Dominguez Ortega, Angel Fernando Argüello Ortiz

CHAPTER 2: Prediction of depressive symptoms in older Mexican adults using supervised learning models: Logistic regression and classification trees 24

Cristian Axel Izquierdo López, Laura Yoselin Tonil De La Luz,
José Juan Muñoz León, Zoylo Morales Romero

CHAPTER 3: Prediction of marijuana and THC product consumption using Logistic Regression and Decision Trees 35

Ángel Martínez Mendez, Eidan Enrique Juarez García,
Dalia Abigail Uscanga Gutierrez, Ángel Juan Sánchez García

CHAPTER 4: Prediction of the level of financial well-being using Logistic Regression: analysis of the National Survey on Financial Health (ENSAFI) 2023 49

Mariela Denisse García Jimenez, Victoria Saldaña Hernández,
Judith Guadalupe Montero Mora, Zoylo Morales Romero

CHAPTER 5: Neural Network applied in the National Household Income and Expenditure Survey 59

Alejandro López-Becerra, Iara Idalia Zapot Juárez,
Angel Juan Sánchez García, José Juan Muñoz León

ÍNDICE

CHAPTER 6: Prediction of total income of economic units in Mexico using Supervised Learning models 71

Jorge Luis Ortiz Libreros, Joshua Iván Olivares Blanco, José Vidal Hernández, Angel Fernando Argüello Ortiz

CHAPTER 7: Benefits and characteristics of employment as factors for formal employment in Mexico 79

Cristian Eduardo Montesino Fernández, Diego Iván González Colorado, Daniela Burgueño Theurel, Judith Guadalupe Montero Mora

Book Presentation

It has become increasingly common to speak of data science across virtually all fields of knowledge; however, we rarely pause to define precisely what we mean by it, and, more importantly, to clarify its relationship with those disciplines that have, for decades, been devoted to the analysis of information. Consequently, before presenting the content of this work, we deem it necessary to establish the conception of data science that underpins this volume and the academic endeavours derived from it.

For the purposes of this book, we adopt the definition posited by the National Science Foundation (NSF, 2014), which conceives data science as the discipline dedicated to planning, obtaining, managing, analysing, and inferring knowledge from data. This perspective acknowledges that data, in their raw state, are devoid of meaning; they acquire value only when transformed into actionable information that facilitates the understanding of phenomena, the explanation of processes, and the support of decision-making. Data science is, therefore, a predominantly applied field that integrates foundational principles from statistics, computer science, and mathematics to address real-world problems within the specific contexts in which data are generated. It is neither a mere intersection of disparate bodies of knowledge nor a new label for existing practices, but rather a distinct approach to confronting increasingly complex problems through the integration of diverse tools and analytical perspectives.

A central tenet of this book, and of our broader academic work, is that there exists neither a rupture nor a linear evolutionary progression between statistics and data science. What has genuinely evolved are the problems we seek to resolve. Statistics has consistently aimed to transform data into information through models, inferences, and analytical procedures; this fundamental mandate remains intact. What has changed is the very nature of data: today, they are generated in significantly larger volumes, at unprecedented velocities, across multiple formats, and frequently demand near-real-time analysis. This new reality has rendered indispensable the development of more sophisticated computational tools, machine learning algorithms, robust storage infrastructures, and advanced processing techniques. Although the name of the field has changed, the tools have evolved, and the challenges have expanded, the essence persists: to convert data into actionable knowledge capable of furthering our understanding and improvement of the world.

The seven chapters comprising this work faithfully reflect this conceptual framework. Although they address markedly diverse domains, they all share the common purpose of demonstrating how supervised learning models can effectively support the analysis and comprehension of complex phenomena. The reader will encounter applications examining the impact of social media on academic performance; the prediction of depressive symptoms in older adults; the analysis of marijuana and THC-derivative consumption; the estimation of financial well-being using data from the National Survey on Financial Health (ENSAFI) 2023; the application of neural networks to data from the National Survey of Household Income and Expenditure; the prediction of total income for economic units in Mexico; and the study of employment benefits as factors associated with formal employment. Each chapter addresses a distinct thematic area (education, health, substance use, economics, financial well-being, the labour market, and socioeconomic analysis) and collectively illustrates how data science integrates naturally into multi-, inter-, and transdisciplinary environments to generate evidence that informs decision-making processes.

Beyond the technical results presented herein, this book holds particular significance due to the process through which it was realised. The works were developed by second-year undergraduate students enrolled in the Bachelor's Degree in Data Science at the Universidad Veracruzana. This accomplishment attests to the viability of fostering rigorous research at an early stage of academic training when pedagogical approaches prioritise hands-on work with real data, project-based learning, and sustained academic mentorship. Each chapter thus represents not only an exercise in learning but also a testament to the potential inherent in data science education when it is oriented towards the resolution of concrete, substantive problems.

This endeavour was driven by the Research Group of Applications of Data Science in Multi, Inter and Transdisciplinary Environments, with the institutional support of the Head of the Bachelor's Degree in Data Science (Dr. Judith Guadalupe Montero Mora) and the Dean of the Faculty of Statistics and Informatics at the Universidad Veracruzana (Professor Zoylo Morales Romero). Furthermore, this book is the result of academic collaboration with the Academic Bodies Applied Statistical Methods and Software Engineering and Technology, all of whom are convinced that contemporary challenges necessitate convergent approaches and close interdisciplinary cooperation.

It is our hope that this work serves simultaneously as a showcase of our students' scholarly output and as an invitation to understand data science from a broad, integrative, and profoundly applied perspective. More than a mere collection of algorithms or predictive models, these pages embody a particular philosophy of research, one in which data constitute the point of departure, evidence guides decision-making, and knowledge is constructed through collaborative interdisciplinary dialogue. This is, ultimately, the spirit that inspires this book and that will continue to inform the academic projects we undertake in the future.

Dr. Angel Fernando Argüello Ortiz
Dr. Ángel Juan Sánchez García
Dr. José Juan Muñoz León

Presentación del libro

Hablar de ciencia de datos se ha vuelto frecuente en prácticamente todos los ámbitos del conocimiento, pero pocas veces nos detenemos a definir qué entendemos por ella y, sobre todo, cuál es su vínculo con disciplinas que durante décadas han analizado la información. Por ello, antes de presentar el contenido de esta obra, consideramos necesario establecer la concepción de ciencia de datos que la guía.

Para este libro, retomamos la definición de la National Science Foundation (NSF, 2014): la ciencia de datos es la disciplina dedicada a planificar, obtener, gestionar, analizar e inferir conocimiento a partir de los datos. Esta perspectiva reconoce que los datos, por sí solos, carecen de significado; adquieren valor cuando se transforman en información útil para comprender fenómenos, explicar procesos y apoyar decisiones. La ciencia de datos es, así, un campo eminentemente aplicado que integra fundamentos de la estadística, la computación y las matemáticas para resolver problemas reales en el contexto donde los datos se generan. No se trata de una simple intersección de saberes ni de una nueva etiqueta, sino de una forma distinta de abordar problemas complejos mediante la integración de herramientas y perspectivas.

Una idea central de este libro, y de nuestro trabajo académico, es que no existe una ruptura ni una evolución lineal entre la estadística y la ciencia de datos. Lo que realmente ha evolucionado son los problemas que buscamos resolver. La estadística siempre ha tenido como propósito transformar datos en información mediante modelos, inferencias y procedimientos analíticos; ese cometido permanece intacto. Lo que cambió es la naturaleza de los datos: hoy se generan en volúmenes mucho mayores, a velocidades sin precedentes, en múltiples formatos y, a menudo, requieren análisis casi en tiempo real. Esta nueva realidad hizo indispensable el desarrollo de herramientas computacionales más sofisticadas, algoritmos de aprendizaje automático, infraestructuras de almacenamiento y técnicas avanzadas de procesamiento. Cambió el nombre del campo, evolucionaron las herramientas y se ampliaron los desafíos, pero la esencia sigue siendo la misma: convertir datos en conocimiento útil para comprender y mejorar el mundo.

Los siete capítulos que conforman esta obra reflejan esa concepción. Aunque abordan ámbitos muy diversos, todos comparten el propósito de demostrar cómo los modelos de aprendizaje supervisado apoyan el análisis y la comprensión de fenómenos complejos. El lector encontrará aplicaciones sobre el impacto de las redes sociales en el rendimiento académico; la predicción de síntomas depresivos en adultos mayores; el consumo de marihuana y derivados del THC; la estimación del bienestar financiero con datos de la ENSAFI 2023; el uso de redes neuronales con información de la Encuesta Nacional de Ingresos y Gastos de los Hogares; la predicción del ingreso de unidades económicas en México; y el estudio de los beneficios laborales asociados al empleo formal. Cada capítulo aborda un contexto distinto (educación, salud, consumo de sustancias, economía, bienestar financiero, mercado laboral y análisis socioeconómico) y evidencia cómo la ciencia de datos se integra naturalmente en entornos multi, inter y transdisciplinarios para generar evidencia que oriente la toma de decisiones.

Más allá de los resultados técnicos, este libro tiene un valor especial por el proceso que lo hizo posible. Los trabajos fueron desarrollados por estudiantes de segundo año de la Licenciatura en Ciencia de Datos de la Universidad Veracruzana. Esto demuestra que, cuando la formación privilegia el trabajo con datos reales, el aprendizaje basado en proyectos y el acompañamiento académico, es factible que estudiantes

en etapas tempranas de su carrera realicen investigaciones con un nivel suficiente para compartirse con la comunidad académica. Cada capítulo representa un ejercicio de aprendizaje y, a la vez, una evidencia del potencial de la formación en ciencia de datos cuando se orienta a la resolución de problemas concretos.

Este esfuerzo fue impulsado por el Cuerpo Académico Aplicaciones de la Ciencia de Datos en Entornos Multi, Inter y Transdisciplinarios, con el respaldo de la Jefatura de la Licenciatura en Ciencia de Datos, a cargo de la Dra. Judith Guadalupe Montero Mora, y de la Dirección de la Facultad de Estadística e Informática de la Universidad Veracruzana, a cargo del Mtro. Zoylo Morales Romero. Asimismo, el libro es fruto de la colaboración con los cuerpos académicos Metodología y Aplicaciones de las Técnicas y Modelos Estadísticos, e Ingeniería y Tecnología de Software, convencidos de que los problemas contemporáneos requieren enfoques convergentes y colaboración interdisciplinaria.

Esperamos que esta obra sea, al mismo tiempo, una muestra del trabajo de nuestros estudiantes y una invitación a comprender la ciencia de datos desde una perspectiva amplia, integradora y profundamente aplicada. Más que una colección de algoritmos o modelos predictivos, estas páginas reflejan una forma de entender la investigación: una donde los datos son el punto de partida, la evidencia guía las decisiones y el conocimiento se construye mediante la colaboración entre disciplinas. Ese es, finalmente, el espíritu que inspira este libro y que orientará los proyectos académicos que emprendamos en el futuro.

Dr. Angel Fernando Argüello Ortiz
Dr. Ángel Juan Sánchez García
Dr. José Juan Muñoz León

Technical Presentation

Data science is commonly associated with vast amounts of information, algorithms, and computational tools. However, underlying these resources lies a fundamental idea: to utilise data for the comprehension of phenomena, the identification of patterns, and the generation of information that supports decision-making processes. In this context, machine learning has become firmly established as one of the cornerstones of contemporary data science, particularly when the objective is to analyse complex relationships and make predictions based on empirical data.

Machine learning comprises a set of computational methods that construct models capable of learning patterns from data and subsequently applying them to predict or classify new observations. In supervised models, the very ones employed throughout this work, we have at our disposal a variable of interest (the response) and a set of explanatory variables (the predictors). The model estimates the functional relationship between them using the available cases and subsequently employs this relationship to predict the response on unseen observations. Depending on the nature of the problem, the response may be categorical (classification) or numeric (regression). Techniques such as logistic regression, decision trees, and neural networks address diverse problems under a common logic: learning from data to extract useful and relevant information.

This approach does not imply a rupture with statistics; rather, it denotes an evident continuity. Historically, statistics has had as one of its central tasks the study of data, the identification of relationships, the construction of models, and the performance of inferences. Data science expands upon this work by confronting problems characterised by more voluminous data, generated at higher velocities, originating from heterogeneous sources, and requiring efficient processing. More than a transition, we are witnessing an evolution of the problems, the data, and the available tools. The essence remains: transforming data into information to understand reality. What has changed is the scale and, consequently, the need to incorporate increasingly advanced computational resources.

From this perspective, machine learning constitutes a natural bridge between statistics and data science. Many of its techniques bear a direct relationship with classical statistical procedures (for example, logistic regression or generalised linear models). The difference does not lie in the fundamental purpose, but rather in the manner in which models are constructed, evaluated, and applied to problems involving high dimensionality, large volumes of observations, or complex structures (nonlinearities, interactions, missing data). The incorporation of computational algorithms does not eliminate statistical reasoning; on the contrary, it renders it indispensable for tasks such as variable selection, the appropriate definition of the response, cross-validation, hyperparameter tuning, performance evaluation (using metrics such as accuracy, precision, recall, AUC-ROC, RMSE, etc.), the interpretation of coefficients or predictor importance, and the recognition of limitations (overfitting, bias, variance). Therefore, learning to use algorithms does not mean abandoning statistics, but rather acquiring new tools with which to continue formulating questions and testing hypotheses against the data.

The seven chapters of this work apply these principles to diverse contexts, consistently employing supervised models. The first four chapters utilise logistic regression and decision trees to address: the impact of social media use on academic performance (classification); the prediction of depressive symptoms in older adults; the consumption of marijuana and THC-derived products; and the estimation of financial well-being using

data from the ENSAFI 2023 survey. These studies demonstrate the versatility of a single algorithmic family in responding to questions arising from the social, educational, health, and economic spheres, provided that the problem is correctly formulated and the data are appropriately processed (cleansing, imputation, coding, standardisation, etc.).

The remaining three chapters extend the methodological and thematic spectrum. The fifth chapter incorporates neural networks (multilayer perceptron architectures) applied to the National Survey of Household Income and Expenditure, exploring a technique with a greater capacity to represent nonlinear and highly interactive relationships. The sixth chapter focuses on the prediction of total income for economic units in Mexico using supervised regression models (such as random forests or networks), while the seventh chapter analyses employment benefits as predictors of formal employment, also from a predictive perspective. Taken together, the chapters demonstrate that machine learning does not belong exclusively to a single domain of knowledge; its application is transversal, provided that there exists a question amenable to being addressed through data and a rigorous methodological framework.

A key characteristic of these works is precisely their multi-, inter-, and transdisciplinary nature. The algorithms are similar, or even the same, but the questions and contexts are different. The interpretation of results depends not only on the computational technique, but also on the substantive knowledge of the phenomenon under study, the nature of the variables, and the constraints of the domain. A decision tree may be employed to study academic performance, financial well-being, or formal employment; however, its interpretation, including the identification of nodes, partitioning rules, and variable importance, cannot be detached from the disciplinary context. This interrelationship between method and domain is, in our view, fundamental to training in data science.

Consequently, this book does not present machine learning as a mere catalogue of isolated algorithms, but rather as a set of tools for formulating questions, exploring data, building models, and obtaining evidence on specific problems. Logistic regression, decision trees, and neural networks are, in this sense, means to a broader end: discovering relationships and patterns that help to understand particular phenomena. The quality of the outcome depends not solely on the selected algorithm, but also on the quality of the data, the definition of the problem, the selection of predictors, the validation strategy, the tuning of parameters, and the responsible interpretation of the findings.

In summary, the seven chapters reflect a transversal idea: data science was not born to replace statistics, but rather to continue its purpose within a context of evolved problems and data. Statistics contributes inferential rigour and critical thinking; mathematics provides the foundations of the models; and computing offers the capacity to process information on scales that were previously unapproachable. Machine learning brings together a significant portion of these elements and converts them into applicable tools. We change the tools, we expand the problems, and, on occasion, we also change the name by which we describe our work; however, the essence persists: we begin with data to transform them into information, and we utilise that information to achieve a better understanding of reality.

Dr. Ángel Juan Sánchez García
Dr. Judith Guadalupe Montero Mora
Mtro. Zoylo Morales Romero

Presentación técnica

La ciencia de datos suele asociarse con grandes volúmenes de información, algoritmos y herramientas computacionales. Sin embargo, detrás de estos recursos subyace una idea fundamental: utilizar los datos para comprender fenómenos, identificar patrones y generar información que apoye la toma de decisiones. En este contexto, el aprendizaje máquina se ha consolidado como uno de los pilares de la ciencia de datos contemporánea, especialmente cuando se requiere analizar relaciones complejas y realizar predicciones a partir de datos empíricos.

El aprendizaje máquina comprende un conjunto de métodos computacionales que construyen modelos capaces de aprender patrones a partir de los datos y aplicarlos para predecir o clasificar nuevas observaciones. En los modelos supervisados , los empleados en esta obra, se dispone de una variable de interés (respuesta) y un conjunto de variables explicativas (predictoras). El modelo estima la relación funcional entre ellas utilizando los casos disponibles y, posteriormente, la emplea para predecir la respuesta en observaciones no vistas. Según la naturaleza del problema, la respuesta puede ser categórica (clasificación) o numérica (regresión). Técnicas como la regresión logística, los árboles de decisión y las redes neuronales abordan problemas diversos bajo una lógica común: aprender de los datos para extraer información útil y relevante.

Esta aproximación no supone una ruptura con la estadística, sino una continuidad evidente. Históricamente, la estadística ha tenido como tarea central estudiar datos, identificar relaciones, construir modelos y realizar inferencias. La ciencia de datos amplía este trabajo al enfrentar problemas con datos más voluminosos, generados a mayor velocidad, provenientes de fuentes heterogéneas y que requieren procesamiento eficiente. Más que una transición, asistimos a una evolución de los problemas, los datos y las herramientas disponibles. La esencia permanece: transformar datos en información para comprender la realidad. Lo que ha cambiado es la escala y, con ella, la necesidad de incorporar recursos computacionales más avanzados.

Desde esta perspectiva, el aprendizaje máquina constituye un puente natural entre la estadística y la ciencia de datos. Muchas de sus técnicas tienen un vínculo directo con procedimientos estadísticos clásicos (por ejemplo, la regresión logística o los modelos lineales generalizados). La diferencia no radica en el propósito fundamental, sino en el modo en que los modelos se construyen, evalúan y aplican a problemas con alta dimensionalidad, grandes volúmenes de observaciones o estructuras complejas (no linealidades, interacciones, datos faltantes). La incorporación de algoritmos computacionales no elimina el razonamiento estadístico; al contrario, lo vuelve indispensable para tareas como selección de variables, definición de la respuesta, validación cruzada, ajuste de hiperparámetros, evaluación del desempeño (mediante métricas como exactitud, sensibilidad, especificidad, AUC-ROC, RMSE, etc.), interpretación de coeficientes o importancia de predictores, y reconocimiento de limitaciones (sobreajuste, sesgo, varianza). Por tanto, aprender a utilizar algoritmos no significa abandonar la estadística, sino disponer de nuevas herramientas para seguir formulando preguntas y contrastando hipótesis con los datos.

Los siete capítulos de esta obra aplican estos principios a contextos diversos, utilizando siempre modelos supervisados. Los primeros cuatro capítulos emplean regresión logística y árboles de decisión para abordar: el rendimiento académico frente al uso de redes sociales (clasificación); la predicción de síntomas depresivos en adultos mayores; el consumo de marihuana y derivados del THC; y el bienestar financiero

a partir de la ENSAFI 2023. Estos trabajos muestran la versatilidad de una misma familia algorítmica para responder preguntas de áreas social, educativa, sanitaria y económica, siempre que el problema esté bien planteado y los datos sean tratados adecuadamente (limpieza, imputación, codificación, estandarización, etc.).

Los tres capítulos restantes amplían el espectro metodológico y temático. El quinto incorpora redes neuronales (arquitecturas de perceptrón multicapa) aplicadas a la Encuesta Nacional de Ingresos y Gastos de los Hogares, explorando una técnica con mayor capacidad para representar relaciones no lineales y de alta interacción. El sexto se centra en la predicción del ingreso total de las unidades económicas en México mediante modelos de regresión supervisada (como bosques aleatorios o redes), y el séptimo analiza los beneficios laborales como predictores del empleo formal, también desde un enfoque predictivo. En conjunto, los capítulos demuestran que el aprendizaje máquina no pertenece a un único dominio del conocimiento; su aplicación es transversal siempre que exista una pregunta abordable mediante datos y un marco metodológico riguroso.

Una característica clave de estos trabajos es su carácter multi, inter y transdisciplinario. Los algoritmos son similares, pero las preguntas y los contextos son distintos. La interpretación de los resultados no depende únicamente de la técnica computacional, sino también del conocimiento sustantivo del fenómeno estudiado, la naturaleza de las variables y las restricciones del dominio. Un árbol de decisión puede emplearse para estudiar rendimiento académico, bienestar financiero o empleo formal, pero su lectura (identificación de nodos, reglas de partición, importancia de variables) no puede desligarse del contexto disciplinar. Esta interrelación entre método y dominio es, a nuestro juicio, fundamental en la formación en ciencia de datos.

Por ello, este libro no presenta el aprendizaje máquina como un catálogo de algoritmos aislados, sino como un conjunto de herramientas para formular preguntas, explorar datos, construir modelos y obtener evidencia sobre problemas concretos. La regresión logística, los árboles de decisión y las redes neuronales son medios para un fin más amplio: descubrir relaciones y patrones que ayuden a comprender fenómenos determinados. La calidad del resultado no depende solo del algoritmo seleccionado, sino también de la calidad de los datos, la definición del problema, la selección de predictores, la estrategia de validación, el ajuste de parámetros y la interpretación responsable de los hallazgos.

En síntesis, los siete capítulos reflejan una idea transversal: la ciencia de datos no nació para sustituir a la estadística, sino para continuar su propósito en un contexto de problemas y datos evolucionados. La estadística aporta el rigor inferencial y el pensamiento crítico; las matemáticas, los fundamentos de los modelos; y la computación, la capacidad para procesar información a escalas antes inabordables. El aprendizaje máquina reúne estos elementos y los convierte en herramientas aplicables. Cambiamos las herramientas, ampliamos los problemas y, en ocasiones, el nombre con que describimos nuestro trabajo, pero la esencia persiste: partir de los datos para transformarlos en información y utilizar ésta para comprender mejor la realidad.

Dr. Ángel Juan Sánchez García
Dra. Judith Guadalupe Montero Mora
Mtro. Zoylo Morales Romero

CHAPTER 1

Classifying student academic performance based on the use of social media applying Logistic Regression and Decision Trees

Javier Zair Galán Trujillo, María Sayuri Baltazar Pedraza, Oscar Alberto Dominguez Ortega, Angel Fernando Argüello Ortiz

ABSTRACT

The use of social media has become part of students' daily lives by facilitating communication, entertainment, and access to information. However, its intensive use has also been associated with potential effects on academic performance, sleep habits, and student well-being. The objective of this study is to classify the impact on academic performance associated with social media use among students through supervised learning models. For this purpose, the *Students Social Media Addiction* dataset, available on Kaggle, was used. It contains 705 records collected through a survey. The response variable was *AffectsAcademicPerformance*, coded as a binary variable. First, an exploratory data analysis was conducted. Subsequently, a decision tree was fitted as the initial classification model, followed by an ultra-conservative logistic regression model in which variables conceptually too close to the response were excluded to reduce the risk of information leakage. The results show that, even when using general variables related to social media use and student profiles, it is possible to adequately classify the perceived impact on academic performance. These findings highlight the usefulness of such models for identifying risk patterns in students' digital behavior.

INTRODUCTION

Social media use has become a common practice in students' daily lives. These platforms allow users to maintain immediate communication, access information, and engage in activities related to entertainment and learning. However, the increase in time spent using them has raised concerns about their possible effects on academic performance, sleep habits, concentration, and the overall well-being of young people.

In recent years, the study of digital behavior has gained relevance in fields such as education, psychology, and data science. Several studies have indicated that intensive social media use may be associated with digital dependence, changes in sleep patterns, and difficulties maintaining attention during academic activities. In this context, data analysis makes it possible to explore behavioral patterns and develop predictive models that support the identification of factors associated with problematic use of these platforms.

Supervised machine learning provides a useful alternative for addressing this type of problem because it enables models to be trained using previously classified observations. When the response variable is binary and qualitative, as in this study, the problem can be

approached as a classification task. Commonly used methods include decision trees and logistic regression. The former produces interpretable classification rules, whereas the latter estimates the probability of belonging to a specific class.

The problem addressed in this study is the classification of whether a student reports an impact on academic performance associated with social media use. The dataset contains survey-based information on sociodemographic characteristics, academic level, most frequently used platform, average daily hours of use, and variables related to mental health, sleep, social media conflicts, and addiction scores. However, some of these variables may be conceptually too close to the target variable, and their direct inclusion could produce unrealistic results or lead to potential conceptual information leakage.

Based on the above, the objective of this study is to apply supervised learning models to classify the impact on academic performance associated with social media use among students, using the *Students Social Media Addiction* dataset. Specifically, the study seeks to compare a decision tree with an ultra-conservative logistic regression model in which variables that are too closely related to the response are excluded, with the aim of obtaining a more cautious assessment of predictive performance.

The research question guiding this study is as follows: Is it possible to classify whether a student reports an impact on academic performance based on general variables related to social media use and student characteristics? Based on this question, the performance of the models and the usefulness of the variables considered are analyzed to distinguish between students who report an academic impact and those who do not.

METHODOLOGY

Research Design

This study followed a quantitative, applied, and nonexperimental approach because the variables were not manipulated but analyzed using a previously constructed dataset. The study had both descriptive and predictive purposes, as the data structure was first explored and supervised learning models were then applied to classify the impact on academic performance associated with social media use.

The problem was addressed as a binary classification task because the response variable had two categories: students who reported an impact on their academic performance and students who did not. A decision tree was used as the initial model, while an ultra-conservative logistic regression model was used as the main model to reduce the risk of including variables that were conceptually too close to the response variable.

Population

The study population consisted of students included in the *Students Social Media Addiction* dataset, available on the Kaggle platform. This dataset contains survey-based information on social media use habits, sociodemographic characteristics, academic level, and variables related to student well-being. Each record represents the individual response of one student.

The dataset used in this study contains 705 observations and 13 original variables. The students ranged in age from 18 to 24 years and represented different academic levels, including high school, undergraduate, and graduate. Because a secondary data source was used, the analysis was limited to the variables available in the dataset. In addition, the information provided by the source does not describe the sampling procedure in detail; therefore, the results should be interpreted within the context of the data analyzed.

Study Variables

The response variable in this study was *AffectsAcademicPerformance*, which indicates whether students consider that social media use affects their academic performance. For model fitting, this variable was converted into a binary coding scheme, in which a value of 1 represents the Yes category and a value of 0 represents the No category.

The predictor variables available in the dataset include demographic, academic, and behavioral information. These variables include age, gender, academic level, country, average daily hours of use, most frequently used platform, hours of sleep per night, mental health score, relationship status, social media conflicts, and addiction score. The *Student_ID* variable was not included in the models because it serves only as an identifier for each student. The *Country* variable was also excluded from the main model because it contains a large number of categories, which may increase model complexity and make interpretation more difficult.

Variable	Description	Measurement Scale	Coding
Student_ID	Unique identifier assigned to each student in the dataset.	Nominal	It was not included in the models because it only identifies individual records.
Age	Represents the age of the surveyed student.	Ratio	Numerical value between 18 and 24 years.
Gender	Describes the gender reported by the student.	Nominal	Female, Male. For the model, it was transformed using One-Hot Encoding.
AcademicLevel	Indicates the student's academic level.	Ordinal	High School, Undergraduate, Graduate. For the model, it was transformed using One-Hot Encoding.
Country	Indicates the student's country of origin.	Nominal	Variable excluded from the main model due to its large number of categories.
AverageDailyUsage Hours	Represents the average number of hours per day that the student spends using social media.	Ratio	Numerical value expressed in hours.

MostUsedPlatform	Indicates the social media platform most frequently used by the student.	Nominal	Instagram, TikTok, Facebook, WhatsApp, YouTube, Twitter, among others. For the model, it was transformed using One-Hot Encoding.
AffectsAcademicPerformance	Indicates whether the student believes that social media use affects their academic performance.	Nominal	Yes = 1; No = 0. Target variable of the model.
SleepHoursPerNight	Represents the approximate number of hours the student sleeps per night.	Ratio	Numerical value expressed in hours. It was excluded from the ultra-conservative model.
MentalHealthScore	Score associated with the student's perceived mental health.	Ordinal	Discrete numerical values recorded in the dataset. It was excluded from the ultra-conservative model.
RelationshipStatus	Describes the student's relationship status.	Nominal	Single, In Relationship, Complicated. For the model, it was transformed using One-Hot Encoding.
SocialMediaConflicts	Score representing the level of conflict associated with social media use.	Ordinal	Discrete numerical values. It was excluded from the ultra-conservative model.
AddictionScore	Score representing the level of addiction or dependence on social media.	Ordinal	Discrete numerical values. It was excluded from the ultra-conservative model.

Table 1. *Description of variables*

Data Collection

The data were obtained from the *Students Social Media Addiction* dataset, published on the Kaggle platform. The dataset contains information collected through surveys administered to students, which recorded personal characteristics, social media use habits, and variables associated with academic performance, sleep, mental health, and social media addiction.

Because this was a secondary dataset, no information was collected directly. The file was downloaded in CSV format and subsequently processed in Google Colab using the Python programming language. Before applying the models, the dataset structure, variable names, data types, and presence of missing values were examined.

Data Analysis

The analysis was conducted in two stages. In the first stage, an exploratory analysis was performed to describe the dataset and identify differences between students who reported an impact on academic performance and those who did not. Frequencies, percentages, descriptive measures, and differences in means were calculated. In addition, bar charts and boxplots were created to examine the behavior of variables such as age, average daily hours of use, hours of sleep per night, mental health score, social media conflicts, and addiction score.

In the second stage, supervised learning models were applied. First, a decision tree was fitted to identify interpretable classification rules. This model made it possible to determine which variables most clearly separated the classes of the target variable. However, because some variables showed a very close relationship with the response, it was considered necessary to use a more cautious model.

For this reason, the main model was an ultra-conservative logistic regression. The variables *SocialMediaConflicts*, *AddictionScore*, *SleepHoursPerNight*, and *MentalHealthScore* were excluded because they could represent direct consequences of problematic social media use and may be conceptually too close to the target variable. Thus, the model was fitted using only more general variables: *Age*, *Gender*, *AcademicLevel*, *AverageDailyUsageHours*, *MostUsedPlatform*, and *RelationshipStatus*.

Before fitting the model, the target variable was encoded in binary format. Categorical variables were transformed using one-hot encoding to convert them into binary variables compatible with the algorithm. Numerical variables were standardized using *StandardScaler* to prevent differences in scale from influencing model fitting.

The dataset was divided into training and test sets. A total of 75% of the records were used to train the model, while the remaining 25% were used to evaluate its performance. The split was stratified to maintain similar proportions of the target variable classes in both sets. Finally, model performance was evaluated using accuracy, a confusion matrix, precision, recall, F1 score, the receiver operating characteristic (ROC) curve, and the area under the curve (AUC).

RESULTS

Sample Description

The analyzed dataset consisted of 705 records and 13 original variables. Each record represents an individual student's response regarding general characteristics, social media use habits, and perceived effects on academic performance. No missing values were identified in the variables considered; therefore, no imputation procedures or removal of incomplete observations were required.

The response variable in this study was *AffectsAcademicPerformance*, which had two categories: Yes, when the student reported that social media use affected their academic performance, and No, when no such effect was reported. Table 2 presents the distribution of this variable.

Category	Frequency	Percentage
Yes	453	64.26
No	252	35.74
Total	705	100.00

Table 2. *Distribution of the target variable*

As shown, most of the surveyed students reported that social media use affected their academic performance. This distribution confirms that the problem can be addressed as a binary classification task, although the Yes category contains a larger proportion of cases. For this reason, a stratified split was used to preserve similar class proportions in the training and test sets.

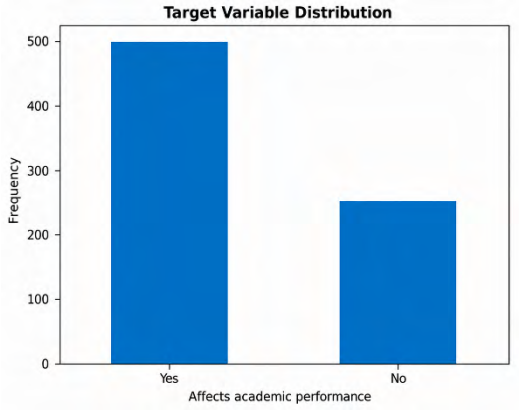

Figure 1. Distribution of the target variable

Descriptive Results

To characterize the dataset, descriptive measures were calculated for the numerical variables. Table 3 presents the mean, median, minimum, and maximum values for age, average daily hours of use, hours of sleep per night, mental health score, social media conflicts, and addiction score.

Variable	Mean	Median	Minimum	Maximum
Age	20.66	21.00	18.00	24.00
AverageDailyUsageHours	4.92	4.80	1.50	8.50
SleepHoursPerNight	6.87	6.90	3.80	9.60
MentalHealthScore	6.23	6.00	4.00	9.00
SocialMediaConflicts	2.85	3.00	0.00	5.00
AddictionScore	6.44	7.00	2.00	9.00

Table 3. *Descriptive statistics of numerical variables*

The students ranged in age from 18 to 24 years, with a mean age of 20.6 years. The average daily social media use was 4.92 hours, although some students reported using these platforms for up to 8.50 hours per day. These results show that the sample generally consisted of young students with substantial daily social media use.

The mean values of the numerical variables were then compared between students who reported an impact on academic performance and those who did not. This comparison made it possible to identify which variables showed the greatest differences between the two groups.

Variable	Mean for No	Mean for Yes	Difference (Yes − No)
AddictionScore	4.60	7.46	2.87
AverageDailyUsage Hours	3.80	5.54	1.73
SocialMediaConflicts	1.78	3.44	1.66
Age	20.92	20.52	-0.40
SleepHoursPerNight	7.81	6.34	-1.47
MentalHealthScore	7.42	5.56	-1.86

Table 4. *Difference in averages between students with and without academic impairment*

Students who reported an impact on academic performance had, on average, higher addiction scores, more hours of daily social media use, and greater levels of social media-related conflict. In contrast, this group also reported fewer hours of sleep and lower mental health scores. These descriptive results should not be interpreted as causal evidence; however, they do show relevant differences between the two groups.

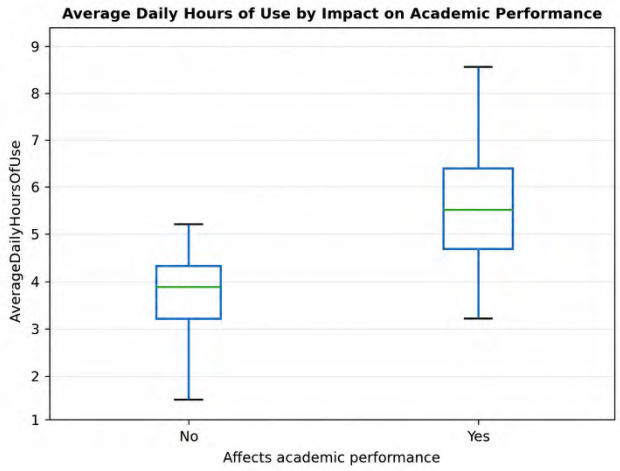

Figure 2. Average hours of daily use according to impact on academic performance

Figure 2 shows that the group of students who reported an impact on academic performance had higher average daily social media use. This result is consistent with the difference in means presented in Table 4, where the Yes group exceeded the No group by approximately 1.73 hours per day.

Decision Tree Results

A decision tree was fitted as the initial classification model. This model made it possible to identify easily interpretable classification rules and determine which variables most clearly separated the classes of the target variable. The tree achieved an accuracy of 0.9981 on the training set and 0.9887 on the test set. Its actual depth was 2, and it had 3 leaves, indicating a simple structure.

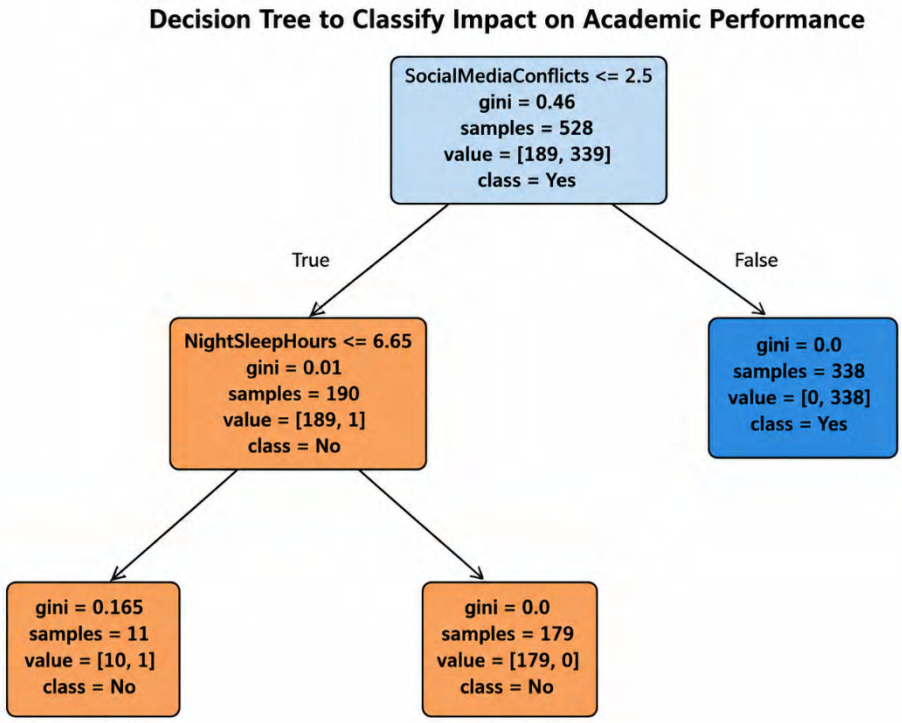

Figure 3. Decision tree for classifying the impact on academic performance

The decision tree primarily used the *SocialMediaConflicts* variable to classify the impact on academic performance. The estimated importance of this variable was 0.9993, whereas *SleepHoursPerNight* had an importance of 0.0007. Although the tree showed high performance, this result should be interpreted with caution because some predictor variables are conceptually very close to the target variable.

Ultra-Conservative Logistic Regression Results

To reduce the risk of conceptual information leakage, an ultra-conservative logistic regression model was fitted. The variables *SocialMediaConflicts*, *AddictionScore*, *SleepHoursPerNight*, and *MentalHealthScore* were excluded because they may be too closely associated with the direct perception of an impact on academic performance. The model was fitted using only *Age*, *Gender*, *AcademicLevel*, *AverageDailyUsageHours*, *MostUsedPlatform*, and *RelationshipStatus*.

The dataset was divided into 75% for training and 25% for testing. Class proportions remained similar in both sets. In the training set, 64.20% of the observations belonged to class 1 and 35.80% to class 0; in the test set, 64.41% belonged to class 1 and 35.59% to class 0.

The model achieved an accuracy of 0.9129 on the training set and 0.8814 on the test set. The difference between these results was not substantial, suggesting that the model did not exhibit severe overfitting. Table 5 presents the main classification metrics.

Class	Precision	Recall	F1-Score	Support
Does Not Affect	0.89	0.76	0.82	63
Affects	0.88	0.95	0.91	114
Accuracy			0.88	177
Macro Average	0.88	0.85	0.87	177
Weighted Average	0.88	0.88	0.88	177

Table 5. Performance metrics of the ultra-conservative logistic regression

The model showed a greater ability to identify students who reported an impact on academic performance, with a recall of 0.95 for the *Yes, affected* class. In contrast, recall for the *No, not affected* class was 0.76, indicating that the model had greater difficulty identifying some cases without a reported academic impact.

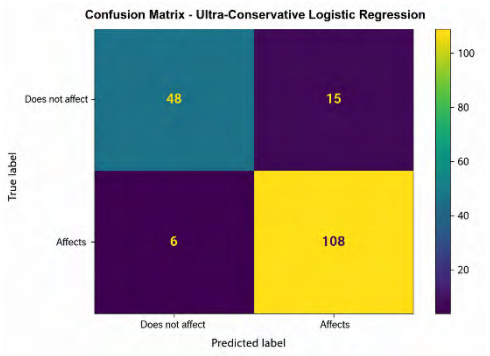

Figure 4. Confusion matrix of the ultra-conservative logistic regression

The confusion matrix shows that the model correctly classified 48 students who did not report an impact on academic performance and 108 students who did. It also produced 15 false positives, in which an impact was predicted even though none was reported, and 6 false negatives, in which the model failed to detect an impact that was reported. Overall, the model made 21 errors on the test set.

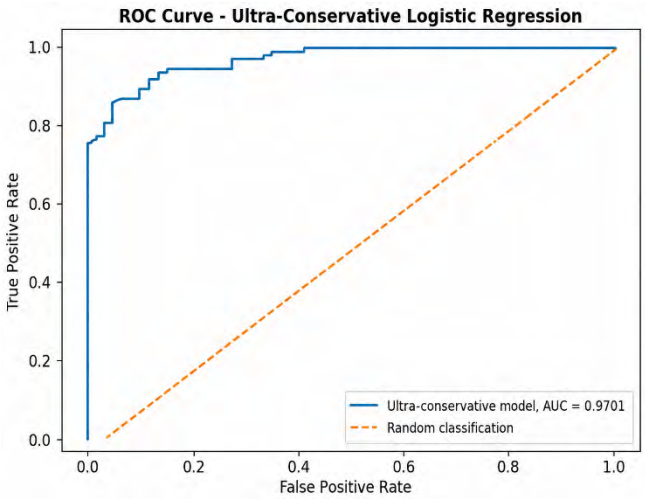

Figure 5. ROC curve of the ultra-conservative logistic regression

The receiver operating characteristic (ROC) curve yielded an area under the curve (AUC) of 0.9701. This value indicates that the model had a strong ability to distinguish between students who reported an impact on academic performance and those who did not. Although the model's performance was high, this result provides a more cautious estimate than those obtained from models that include variables more closely related to the response.

Variable	Coefficient	Odds Ratio
AverageDailyUsageHours	3.0826	21.8142
MostUsedPlatform_TikTok	2.7858	16.2123
MostUsedPlatform_KakaoTalk	2.1401	8.5006
MostUsedPlatform_Instagram	1.6255	5.0812
MostUsedPlatform_Snapchat	1.2679	3.5535
MostUsedPlatform_WhatsApp	0.6226	1.8638
MostUsedPlatform_YouTube	0.0710	1.0736
Gender_Male	0.0364	1.0371

Table 6. Variables that increase the probability of academic impairment

Positive coefficients indicate variables associated with an increase in the estimated probability of belonging to the *Yes, affected* class. Among these variables, *AverageDailyUsageHours* stands out, indicating that daily usage time is one of the most relevant variables in the ultra-conservative model.

Variable	Coefficient	Odds Ratio
AcademicLevel_Undergraduate	-1.2821	0.2775
RelationshipStatus_In Relationship	-0.9863	0.3729
MostUsedPlatform_VKontakte	-0.9779	0.3761
AcademicLevel_High School	-0.7221	0.4857
Age	-0.5289	0.5893
MostUsedPlatform_Twitter	-0.4666	0.6271
RelationshipStatus_Single	-0.2560	0.7742
MostUsedPlatform_LINE	-0.2494	0.7793
MostUsedPlatform_WeChat	-0.2074	0.8127
MostUsedPlatform_LinkedIn	-0.0786	0.9244

Table 7. Variables that decrease the probability of academic impairment

Negative coefficients indicate variables associated with a decrease in the estimated probability of reporting an impact on academic performance. However, categorical variables should be interpreted with caution because their effects are estimated in comparison with a reference category generated during the encoding process.

Model Comparison

Table 8 presents a general comparison of the models considered in the analysis. It includes the decision tree, the full logistic regression model, and the ultra-conservative logistic regression model.

Model	Training Accuracy	Test Accuracy	AUC	Observation
Decision Tree	0.9981	0.9887		High performance, but dependent on variables closely related to the outcome.
Full Logistic Regression	1.0000	1.0000	1.0000	Perfect result; considered to have a risk of conceptual data leakage.
Ultra-Conservative Logistic Regression	0.9129	0.8814	0.9701	Main model because it uses variables less directly related to the outcome.

Table 8. *Comparison of classification models*

The comparison shows that models including variables such as *SocialMediaConflicts*, *AddictionScore*, *SleepHoursPerNight*, and *MentalHealthScore* achieved nearly perfect performance. However, this behavior may be explained by the conceptual proximity of these variables to the target variable. For this reason, the ultra-conservative model was considered more appropriate for the main analysis because it maintained adequate performance without using variables that could introduce information too directly related to the response.

CONCLUSIONS

The objective of this study was to classify the impact on academic performance associated with social media use among students through supervised learning models. The results show that it is possible to distinguish between students who report an impact on academic performance and those who do not by using variables related to social media use and general student characteristics.

The exploratory analysis revealed relevant differences between the two groups. Students who reported an impact on academic performance had higher daily usage time, higher addiction scores, and more social media-related conflicts. They also reported fewer hours of sleep and lower mental health scores, suggesting an association between digital behavior and certain academic and personal factors.

The decision tree made it possible to identify simple classification rules; however, its high performance was associated with variables that were conceptually very close to the target variable. Therefore, an ultra-conservative logistic regression model was fitted after excluding social media conflicts, addiction score, sleep hours, and mental health. This model achieved an accuracy of 0.8814 and an AUC of 0.9701 on the test set, demonstrating adequate classification performance using variables less closely related to the response.

The main contribution of this study is to show that supervised learning models can support the identification of patterns associated with a possible impact of social media use on academic performance. These results may be useful for educational institutions when designing guidance strategies to promote healthy digital habits. However, because the study used a secondary survey-based dataset, causal relationships cannot be established. Future research should incorporate objective academic variables, collect primary data, and compare additional classification algorithms using cross-validation.

LITERATURE

Breiman, L., Friedman, J. H., Olshen, R. A., & Stone, C. J. (1984). Classification and regression trees. Wadsworth International Group.

Cox, D. R., & Snell, E. J. (1989). Analysis of binary data (2nd ed.). Chapman and Hall.

García-Contreras, R., Valle-Cruz, D., & Canales-García, R. A. (2021). Selección organizacional: resiliencia y desempeño de las pymes en la era de la COVID-19. Estudios Gerenciales, 37(158), 73–84. https://doi.org/10.18046/j.estger.2021.158.4291

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., & Duchesnay, É. (2011). Scikit-learn: Machine learning in Python. Journal of Machine Learning Research, 12, 2825–2830.

Robles-Velasco, A., Cortés, P., Muñuzuri, J., & Barbadilla-Martín, E. (2020). Aplicación de la regresión logística para la predicción de roturas de tuberías en redes de abastecimiento de agua. Dirección y Organización, 70, 78–85. https://doi.org/10.37610/dyo.v0i70.570

Zahranusratt. (n.d.). Student social media addiction analysis dataset [Data set]. Kaggle. Recuperado el 15 de junio de 2026, de https://www.kaggle.com/datasets/zahranusratt/student-social-media-addiction-analysis-dataset

CHAPTER 2

Prediction of depressive symptoms in older Mexican adults using supervised learning models: Logistic regression and classification trees

Cristian Axel Izquierdo López, Laura Yoselin Tonil De La Luz, José Juan Muñoz León, Zoylo Morales Romero

ABSTRACT

Depression is one of the most common mental health problems among older adults in Mexico, as it can affect their well-being, autonomy, and quality of life. In response to this issue, the present study aimed to compare two supervised learning models, logistic regression and a classification tree, to identify the presence of depressive symptoms based on sociodemographic variables, health conditions, stress level, social isolation, functioning in activities of daily living, and cognitive performance. The analysis used data from 3,435 participants in the 2021 Mexican Health and Aging Study (MHAS). The presence of depressive symptoms was assessed using the Center for Epidemiologic Studies Depression Scale (CES-D). Before modeling, the data underwent imputation and standardization procedures. Both models were then trained and evaluated using cross-validation to identify the alternative with the best predictive performance. The results showed that logistic regression performed better in distinguishing between participants with and without depressive symptoms. Feelings of loneliness emerged as the variable with the greatest influence on prediction, followed by limitations in instrumental activities of daily living and perceived stress. Overall, the findings suggest that social isolation plays an important role in the early detection of older adults who are more likely to present depressive symptoms. They also show that logistic regression is an effective tool for this type of prediction.

INTRODUCTION

Population aging is one of the most significant demographic changes of the present day, as an increasing number of people are reaching older ages. Although the rise in life expectancy represents an important social and public health achievement, it has also created new challenges for healthcare systems, particularly in relation to mental health care. In this context, depression is recognized as one of the most common disorders among older adults because it can affect their well-being, limit their functioning, and reduce their quality of life.

In Mexico, several studies have analyzed the factors associated with depressive symptoms among older adults. In this regard, Luna-Orozco et al. (2020) used data from the Mexican Health and Aging Study (MHAS) and identified variables associated with a greater likelihood of presenting depressive symptoms, including limitations in activities of daily living, a negative perception of health status, female sex, and low educational attainment. The authors also used logistic regression models to examine the influence of these factors, providing evidence on the elements that may increase the vulnerability of this population group.

Despite these findings, much of the available research has focused on traditional statistical methods aimed primarily at explaining associations between variables. However, there is still limited evidence on the application and comparison of machine learning models to predict the presence of depressive symptoms among older adults in Mexico. This limits the use of tools that could contribute to the earlier identification of individuals at risk.

Based on this perspective, the objective of the study was to develop and compare two supervised learning models, logistic regression and a classification tree, using data from the 2021 Mexican Health and Aging Study (MHAS) to predict the presence of depressive symptoms among older adults in Mexico. Sociodemographic variables, indicators of functioning, perceived health, social isolation, and cognitive performance were considered to determine which model had better predictive ability.

METHODOLOGY

This study used a quantitative, cross-sectional design because the data corresponded to a single measurement conducted in 2021 and the variables were not deliberately manipulated. The study also had a predictive and comparative scope, as its purpose was to develop and compare two supervised learning models: logistic regression and a classification tree. Both models were used to predict the presence of depression among older adults. In addition, the analytical process followed the CRISP-DM methodology (Song et al., 2024), which organized the study into the stages of problem understanding, data preparation, modeling, evaluation, and interpretation of results.

The population consisted of participants in the Cognitive Assessment Module of the 2021 Mexican Health and Aging Study (MHAS). The original dataset included 3,575 records; however, cases with missing information on the dependent variable were excluded, resulting in a final sample of 3,435 older adults. Of this total, 61.1% did not present depression and 38.9% did, reflecting a moderate class imbalance that was considered during the modeling stage.

The dependent variable was obtained from the Center for Epidemiologic Studies Depression Scale (CES-D), which, according to González et al. (2011), is a screening instrument validated for the Mexican population. It measures depressive symptoms through 20 items scored from 0 to 3, with a total score ranging from 0 to 60. Using a cutoff score of 16, the 2021 MHAS generated a variable indicating whether the participant presented depression. This variable was coded dichotomously, with 0 indicating the absence of depressive symptoms and 1 indicating their presence, and it was used as the response variable in this study. The independent variables were grouped into five categories (Table 1).

Variable	Category	Description	Measurement Scale	Coding
MC_CESD_DEPRESSED_21	Dependent	Presence of depressive symptoms according to the CES-D scale	Dichotomous nominal	0 = No; 1 = Yes
GENDER	Sociodemographic	Gender of the study participant	Dichotomous nominal	1 = Male; 2 = Female
AGE_MXCOG_21	Sociodemographic	Age of the study participant in completed years	Ratio	Numerical value expressed in years
MC_N_AIVD_21	Daily Living	Number of limitations in instrumental activities of daily living (IADL)	Ordinal	1 = None; 2 = A little; 3 = Some; 4 = A lot
MC_STRESS_21	Perceived Health	Level of stress related to the COVID-19 pandemic	Ordinal	1 = None; 2 = A little; 3 = Some; 4 = A lot
MC_MEMORY_21	Perceived Health	Self-perception of memory	Ordinal	1 = Excellent; 2 = Very good; 3 = Good; 4 = Fair; 5 = Poor
MC_MEMORY_COMP_21	Perceived Health	Comparison of current memory with memory two years earlier	Ordinal	1 = Better; 2 = About the same; 3 = Worse
MC_LONELY_21	Social Isolation	Feeling of loneliness during the previous week	Dichotomous nominal	0 = No; 1 = Yes
MC_LACK_OF_COMPA_21	Social Isolation	Frequency with which the participant feels a lack of companionship	Ordinal	1 = Never; 2 = Rarely; 3 = Sometimes; 4 = Very often
MC_IGNORED_21	Social Isolation	Frequency with which the participant feels ignored by others	Ordinal	1 = Never; 2 = Rarely; 3 = Sometimes; 4 = Very often
MC_DIFFERENT_ANIMALS_21	Cognitive Tests	Number of different animals mentioned in the verbal fluency test	Ratio	Numerical count
MC_CORRECT_SYMBOLS_21	Cognitive Tests	Number of correct symbols identified in the corresponding test	Ratio	Numerical count
MC_CUBE_PRAXIAS_21	Cognitive Tests	Score obtained in the cube praxis test	Ordinal	0 = Did not meet any criteria; 1 = Met one criterion; 2 = Met two criteria; 3 = Met three criteria; 4 = Met all four criteria

Table 1. *Description of variables*

The National Institute of Statistics and Geography (INEGI), in coordination with the University of Texas Medical Branch, conducted the 2021 Cognitive Assessment Survey for the Mexican Health and Aging Study (MHAS). Its purpose was to obtain information from a subsample of MHAS follow-up participants and to estimate the prevalence and risk factors of dementia and other forms of cognitive impairment among adults aged 58 years and older in Mexico.

In addition, a descriptive analysis was conducted, including the frequency distribution of the dependent variable and a comparison of the means and standard deviations of the quantitative variables according to the presence or absence of depression. Distribution tables were also prepared for the categorical variables to describe the study population and identify preliminary patterns related to depression. Missing values in the predictor variables were imputed using the median of each variable to preserve the available information. The dataset was then divided into 70% for model training and 30% for testing through stratified sampling, which preserved the original class proportions in both subsets. The two models were subsequently trained using class weights inversely proportional to their frequencies to compensate for the imbalance between categories.

Logistic regression was implemented using Elastic Net regularization, which combines L1 and L2 penalties in equal proportions to promote both variable selection and coefficient shrinkage (Kim et al., 2016). Before fitting the model, the predictor variables were standardized to a mean of zero and a standard deviation of one. The hyperparameters were also optimized through a grid search with five-fold stratified cross-validation, using the area under the receiver operating characteristic curve (AUC-ROC) as the selection metric.
Stratified cross-validation consists of dividing the dataset into a specified number of partitions, or folds. In each iteration, one fold is used to evaluate the model, while the remaining folds are used for training. This process is repeated until every partition has served as the test set. In addition, the stratified version preserves the original class proportions within each fold (Corso et al., 2009), which was particularly appropriate in this study because of the imbalance between older adults with and without depressive symptoms.

For logistic regression, different regularization strengths and proportions between the two penalties were explored. For the classification tree, the maximum depth, impurity criterion, minimum number of observations required to split a node, and minimum number required to form a leaf were adjusted. The performance of both models was evaluated in their initial and optimized configurations using accuracy, balanced accuracy, macro F1 score, AUC-ROC, confusion matrices, and classification reports. Balanced accuracy and macro F1 score were included specifically to reduce the effect of class imbalance. In addition, model stability was estimated through five-fold cross-validation on the training set, reporting the mean and standard deviation of the AUC-ROC values obtained during this process.

Finally, the results were interpreted by calculating odds ratios (ORs) for the optimized logistic regression model and Gini impurity-based feature importance for the classification tree. This made it possible to identify risk and protective factors associated with depression among the older adult population studied. In addition, to facilitate interpretation

of the decision rules, a visual representation of the tree was generated and limited to three levels of depth. Data processing and modeling were performed in Python using the pandas, NumPy, and scikit-learn libraries (Baeza Terol, 2025).

RESULTS

The analysis was conducted using a sample of 3,435 older adults from Mexico who participated in the 2021 MHAS Cognitive Assessment Survey, after excluding cases with incomplete information on the depression variable. The distribution of the response variable showed that 61.1% of participants did not present depressive symptoms, corresponding to 2,099 cases, whereas 38.9% did, corresponding to 1,336 cases. This distribution confirms the previously noted class imbalance (Figure 1). The sample included sociodemographic characteristics, indicators of daily functioning, self-perceived health, social isolation, and cognitive performance.

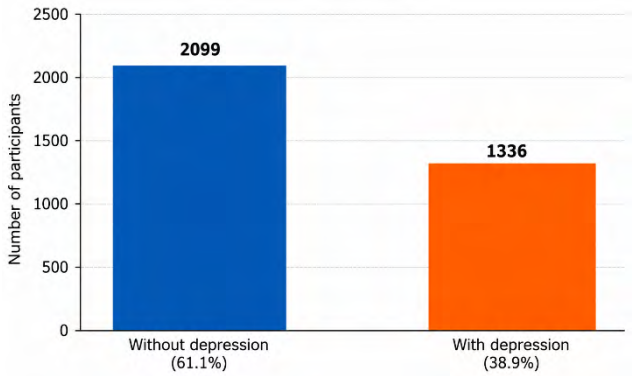

Figure 1. Distribution of the response variable

The descriptive results showed that participants with depression had lower scores on the cognitive tests of verbal fluency and processing speed, as reflected in fewer unique animals named and fewer correct symbols. They also had a slightly higher mean age than participants without depression (Table 2).

Variable	Mean (SD) Total	Mean (SD) Without Depression (0)	Media (Desv) With Depression (1)
AGE_MXCOG_21	72.19 (7.48)	71.71 (7.41)	72.94 (7.52)
MC_DIFFERENT_ANIMALS_21	13.89 (5.02)	14.62 (5.10)	12.75 (4.65)
MC_CORRECT_SYMBOLS_21	18.32 (11.51)	20.30 (12.21)	15.22 (9.53)

Table 2. *Distribution of continuous variables by depression group*

A higher proportion of depression was observed among women than among men. In addition, the percentage of participants with depression increased as the number of limitations in instrumental activities of daily living increased, whereas better performance on the cube praxis test was associated with lower percentages of depression (Table 3).

Variable	Category	N	% of total	% with depression
GENDER	1	1460	42.5	29.9
	2	1975	57.5	45.6
MC_N_AIVD_21	0	2468	71.8	30.6
	1	534	15.5	53.2
	2	282	8.2	66.0
	3	100	2.9	74.0
	4	51	1.5	72.5
MC_CUBE_PRAXIAS_21	0	1130	32.9	48.1
	1	694	20.2	39.9
	2	307	8.9	39.1
	3	228	6.6	33.8
	4	1076	31.3	29.6

Table 3. *Sociodemographic distribution and cognitive performance*

Participants with higher stress levels had higher rates of depression. Similarly, poorer self-perceived memory and perceived decline compared with previous years were associated with a higher frequency of depressive symptoms (Table 4).

Variable	Category	N	% of total	% with depression
MC_STRESS_21	1	1258	36.6	28.0
	2	931	27.1	35.9
	3	673	19.6	40.0
	4	573	16.7	66.5
MC_MEMORY_21	1	30	0.9	13.3
	2	59	1.7	20.3
	3	651	19.0	24.7
	4	2181	63.5	39.4
	5	514	15.0	58.2
MC_MEMORY_COMP_21	1	102	3.0	45.1
	2	1965	57.2	28.4
	3	1368	39.8	53.4

Table 4. *Distribution of subjective health perception and stress levels*

Social isolation variables showed a strong association with depression. Participants who more frequently reported feeling lonely, lacking companionship, or feeling ignored had considerably higher rates of depressive symptoms than those who reported these experiences less frequently (Table 5).

Variable	Category	N	% of total	% with depression
MC_LONELY_21	0	2311	67.3	18.8
	1	1124	32.7	80.2
MC_LACK_OF_COMPA_21	1	2045	59.5	23.1
	2	668	19.4	53.7
	3	448	13.0	60.9
	4	274	8.0	84.7
MC_IGNORED_21	1	2427	70.7	28.4
	2	530	15.4	56.2
	3	334	9.7	68.6
	4	144	4.2	82.6

Table 5. *Distribution of social isolation indicators and their relationship with depression*

In the initial evaluation, logistic regression achieved an accuracy of 81.28% and an AUC-ROC of 0.8837, whereas the classification tree achieved an accuracy of 74.20% and an AUC-ROC of 0.7259. These results demonstrated the superior ability of the logistic regression model to distinguish between the two classes, as shown in the confusion matrices and comparative ROC curve from this initial stage (Figure 2).

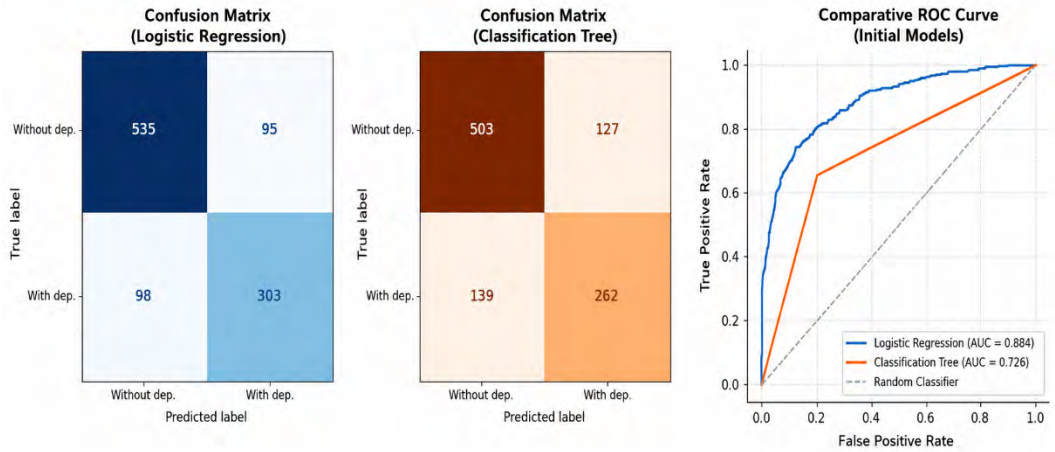

Figure 2. Initial models: confusion matrices and ROC curve

After hyperparameter optimization through grid search and stratified cross-validation, both models achieved similar accuracy, with 81.57% for logistic regression and 82.15% for the classification tree. Despite the narrowing of the accuracy gap, logistic regression maintained a higher AUC-ROC of 0.8834, compared with 0.8585 for the classification tree. Therefore, logistic regression remained the model with the best ability to distinguish between older adults with and without depressive symptoms, as shown in the confusion matrices and comparative ROC curve for this optimized stage (Figure 3).

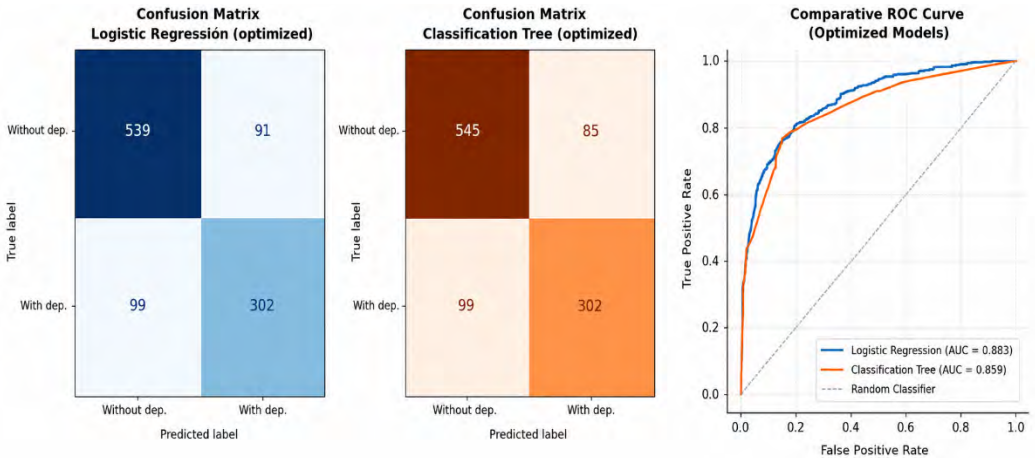

Figure 3. Optimized models: confusion matrices and ROC curve

When the structure of both optimized models was analyzed (Figure 4), feelings of loneliness emerged as the most relevant variable. This predictor had an importance of 0.7395 in the classification tree and an OR of 2.3662 in the logistic regression model, indicating that feeling lonely more than doubled the odds of presenting depressive symptoms. The models also identified perceived stress and difficulties with instrumental activities of daily living as additional risk factors, in contrast to the protective effect associated with better performance in processing speed and constructional praxis.

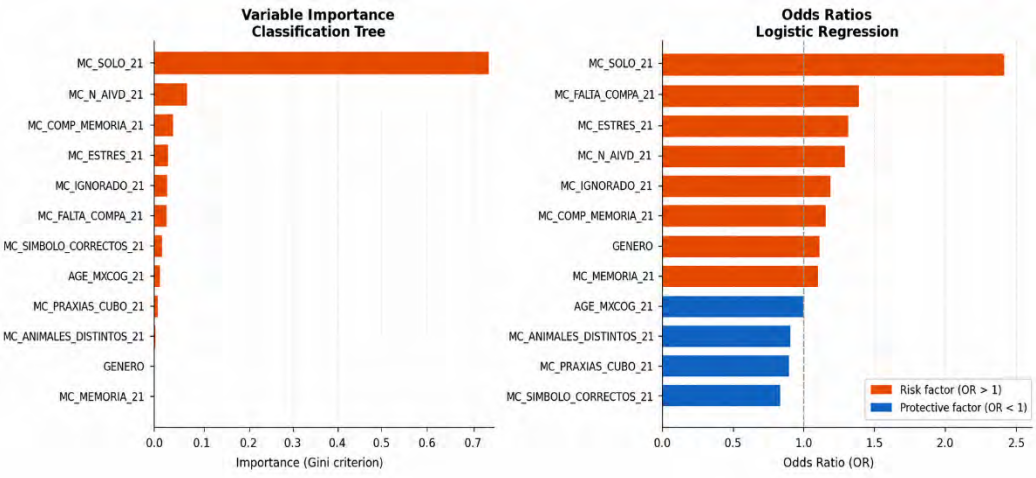

Figure 4. Importance of variables (Tree) and Odds ratio (Regression)

When the tree was limited to three levels of depth, feelings of loneliness appeared as the root node, or first split (Figure 5). This hierarchical position confirms that loneliness was the most discriminative variable for classifying depressive symptoms in this population of older adults.

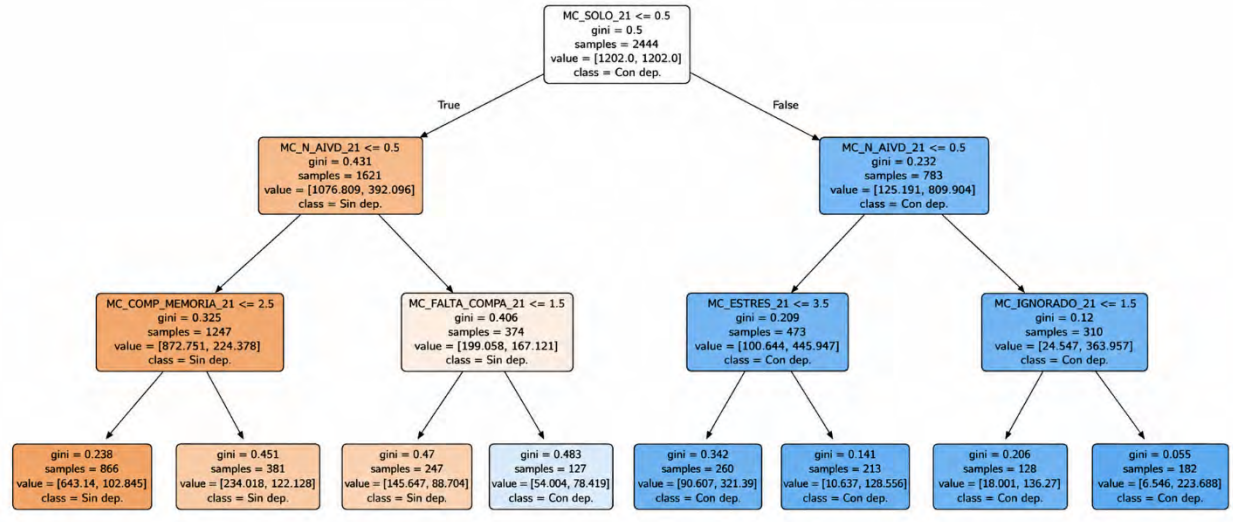

Figure 5. Classification tree (3 levels of depth)

The comparison using stratified cross-validation showed the superiority of logistic regression, which achieved a mean AUC-ROC of 0.8834, compared with 0.8585 for the classification tree. Thus, the logistic regression model was established as the approach with the greatest predictive ability for identifying depressive symptoms in the population studied (Table 6).

Metric	Logistic Regression	Classification Tree
Test Accuracy	0.8157	0.8215
Test Balanced Accuracy	0.8043	0.8091
Test Macro F1-Score	0.8054	0.8110
Test AUC-ROC	0.8834	0.8585
Mean Cross-Validated AUC	0.8751	0.8544
Cross-Validated AUC Standard Deviation	0.0226	0.0266

Table 6. Final comparison of the models

CONCLUSIONS

In conclusion, the comparison of both models for predicting depressive symptoms among older adults in Mexico using data from the 2021 MHAS showed that logistic regression had better overall performance, with an AUC-ROC of 0.8834, compared with 0.8585 for

the classification tree. Thus, logistic regression was established as the model with the greatest discriminative ability. The main finding was that perceived loneliness was the most relevant variable for identifying the presence of depression, based on both its importance in the classification tree (0.7395) and its OR in the logistic regression model (2.3662). This variable was followed by perceived stress and difficulties with instrumental activities of daily living, whereas better cognitive performance acted as a protective factor. Based on these findings, the main contribution of this study is to show that, in addition to the sociodemographic and health variables traditionally examined, indicators of emotional social isolation have considerable predictive value. This expands the available evidence on depression among older adults in Mexico through the use of machine learning models.

Therefore, these results could be used to design brief screening instruments for community health centers, prioritizing questions about loneliness and social isolation as early indicators of risk. This could support the timely referral of older adults to psychological or social support services. Future research should replicate this analysis using longitudinal MHAS data to evaluate the temporal relationship between loneliness and depression. It should also explore other supervised learning models, such as support vector machines or neural networks, to capture interactions between social and cognitive variables.

LITERATURE

Baeza Terol, F. (2025). *Estudio comparativo de las librerías de Machine Learning de Python y Rust*. Universidad Miguel Hernández de Elche. https://dspace.umh.es/bitstream/11000/37163/1/TFG-Baeza%20Terol%2c%20Francesc.pdf

Corso, I., Lorena, C., & Clave, P. (2009). *Aplicación de algoritmos de clasificación supervisada usando Weka*. Universidad Tecnológica Nacional. https://www.investigacion.frc.utn.edu.ar/labsis/Publicaciones/congresos_labsis/cynthia/CNIT_2009_Aplicacion_Algoritmos_Weka.pdf

González-Forteza, C., Torres, C. S., Tapia, A. J., Fernández, I. H., González-González, A., García, F. J., Medina-Mora, M. E., & Mejía, H. F. (2011). Confiabilidad y validez de la escala de depresión CES-D en un censo de estudiantes de nivel medio superior y superior, en la Ciudad de México. *Salud Mental*, *34*(1), 53–59. https://www.scielo.org.mx/scielo.php?pid=S0185-33252011000100007&script=sci_arttext

Instituto Nacional de Estadística y Geografía. (2021). *Encuesta de Evaluación Cognitiva para la ENASEM 2021* [Conjunto de datos]. https://www.inegi.org.mx/programas/ecenasem/2021/#documentacion

Kim, M.-H., Banerjee, S., Park, S. M., & Pathak, J. (2016). Improving risk prediction for depression via Elastic Net regression - Results from Korea National Health Insurance Services Data. *AMIA Annual Symposium Proceedings*, *2016*, 1860–1869. https://pmc.ncbi.nlm.nih.gov/articles/PMC5333336/pdf/2495859.pdf

Luna-Orozco, K., Fernández-Niño, J. A., & Astudillo-García, C. I. (2020). Asociación entre la discapacidad física y la incidencia de síntomas depresivos en adultos mayores mexicanos. *Biomédica: Revista del Instituto Nacional de Salud*, *40*(4), 641–655. https://doi.org/10.7705/biomedica.5398

Song, Y., Chen, L., Liu, H., & Liu, Y. (2024). Machine learning algorithms to predict depression in older adults in China: a cross-sectional study. *Frontiers in Public Health*, *12*, 1462387. https://doi.org/10.3389/fpubh.2024.1462387

CHAPTER 3

Prediction of marijuana and THC product consumption using Logistic Regression and Decision Trees

Ángel Martínez Mendez, Eidan Enrique Juarez García, Dalia Abigail Uscanga Gutierrez, Ángel Juan Sánchez García

ABSTRACT

The use of marijuana and products containing tetrahydrocannabinol (THC) is an important public health issue because of the observed increase in its prevalence and the need to better understand this phenomenon. In Mexico, the 2025 National Survey on Drug, Alcohol, and Tobacco Use (ENCODAT) provides relevant information for its study. The objective of this research was to develop and compare logistic regression and decision tree models to predict a history of marijuana and THC-containing product use based on data from the 2025 ENCODAT. Sociodemographic, employment-related, drug exposure, and social tolerance variables were used for this purpose. Because the response variable was imbalanced, specific strategies were implemented for each model: class weighting in logistic regression and the Synthetic Minority Oversampling Technique (SMOTE) in the decision tree. The models were then evaluated using classification metrics such as accuracy, sensitivity, and specificity. Accuracy was used as the main criterion for comparing model performance, whereas sensitivity was used to assess the models' ability to identify individuals with a history of use and to guide the practical interpretation of the results. The findings showed that both methods were able to classify individuals according to their history of use, although differences in predictive performance were observed. In addition, variables related to drug exposure, age, sex, and the social perception of use provided relevant information for classifying individuals.

Keywords: Supervised machine learning; cannabis; logistic regression; decision tree; ENCODAT 2025; public health.

INTRODUCTION

The use of psychoactive substances continues to be a public health challenge because of its individual and social consequences. International organizations such as the World Health Organization (WHO) and the United Nations Office on Drugs and Crime (UNODC) have reported that substance use continues to increase across different regions of the world, leading to the development of research aimed at better understanding this phenomenon (World Health Organization [WHO], 2024; United Nations Office on Drugs and Crime [UNODC], 2025).

In Mexico, these patterns are monitored primarily through the National Survey on Drug, Alcohol, and Tobacco Use (ENCODAT). Results from the 2025 edition show that the lifetime prevalence of illegal drug use increased from 9.9% in 2016 to 13.1% in 2025. Among these substances, cannabis remained the most prevalent illegal drug, increasing from 8.6% to 12.0% during the same period (CONASAMA et al., 2025).

In addition, the availability of population-based datasets has encouraged the incorporation of machine learning techniques into fields such as public health and epidemiology. These methods make it possible to analyze large volumes of information and identify patterns that may be difficult to detect using traditional statistical approaches (Pedrero et al., 2021). Logistic regression and decision trees are among the most commonly used classification techniques and have proven useful for developing predictive models based on sociodemographic and health-related variables (Dell et al., 2022; Araujo Zarate et al., 2024).

A common issue in the analysis of population surveys is the presence of imbalanced classes, which may affect model performance by favoring the majority class. To address this problem, balancing techniques such as the Synthetic Minority Oversampling Technique (SMOTE) have been developed to improve the representation of the minority class during model training.

In this context, the present study uses data from the 2025 ENCODAT to evaluate the predictive ability of sociodemographic, employment-related, drug exposure, and social tolerance variables regarding the lifetime use of cannabis and THC-containing products through logistic regression and decision tree models.

BACKGROUND

Supervised learning techniques have been used to study factors associated with substance use because of their ability to analyze multiple variables and develop predictive models in the field of health.
Dell et al. (2022) compared logistic regression, decision trees, and random forests to identify factors associated with cannabis use disorder among adults in the United States. The authors found that variables such as age, sex, marital status, and educational level contributed significantly to the classification of individuals. They also observed that tree-based models showed better predictive performance, whereas logistic regression facilitated the interpretation of factors associated with cannabis use.

Also in 2022, Schwebel et al. analyzed factors related to different cannabis use outcomes through classification techniques. Their findings showed that individual and environmental variables may have different effects depending on the behavior being evaluated, highlighting the importance of appropriate variable selection when developing predictive models.

Subsequently, Lansford et al. (2023) identified that demographic and social factors remain strongly associated with problematic alcohol, cannabis, and other substance use during adulthood. Their results emphasize the importance of incorporating social environment variables alongside individual characteristics to better understand these behaviors.

Unlu et al. (2024) developed artificial intelligence-based predictive models using data from population surveys. The authors found that sociodemographic variables and factors related to substance exposure were able to distinguish between cannabis users and nonusers, demonstrating the potential of these methods for population-based data analysis.

In 2025, Leatherdale et al. examined factors associated with cannabis use in a large sample of Canadian youth. The results showed that variables related to the social environment and perceptions of use continued to play an important role in the likelihood of cannabis use, confirming the usefulness of classification models in this field.

Overall, the reviewed literature shows that sociodemographic and social variables, as well as factors related to substance exposure, have the potential to explain or predict behaviors associated with the use of cannabis or THC-containing products. It also demonstrates that techniques such as logistic regression and decision trees are appropriate tools for analyzing data from population-based surveys. However, evidence based on recent data from the Mexican population remains limited, particularly in studies comparing supervised classification models using national survey data.

PROBLEM STATEMENT

The 2025 ENCODAT collects sociodemographic, employment-related, drug exposure, and social tolerance information that may be associated with the use of cannabis and THC-containing products. However, it is not clear which of these variables have sufficient predictive ability to classify individuals according to their history of use.

In addition, the imbalanced distribution of the response variable may affect the performance of classification models and make it more difficult to identify individuals with a history of use. Therefore, it is relevant to evaluate different supervised learning techniques and determine which one provides the best performance for this problem.

STUDY OBJECTIVES

General Objective

To develop and compare logistic regression and decision tree models for predicting lifetime use of cannabis and THC-containing products using data from the 2025 ENCODAT.

Specific Objectives

- To evaluate the predictive ability of sociodemographic, employment-related, drug exposure, and social tolerance variables.
- To apply strategies for addressing class imbalance.
- To compare model performance and identify the variables most relevant to classifying individuals.

Research Questions

Do the variables included in the 2025 ENCODAT make it possible to predict a history of cannabis and THC-containing product use?
How do the evaluated classification models perform when accounting for imbalance in the response variable?
Which of the evaluated models shows the best predictive performance?

MATERIALS AND METHODS

Research Design

A quantitative, observational, cross-sectional study with a predictive scope was conducted. The study was considered quantitative because it used numerical data obtained through a structured survey and applied statistical techniques and machine learning algorithms for analysis. The design was observational because no intervention was performed and the study variables were not manipulated. It was also cross-sectional because the information was collected at a single point in time. Finally, it was classified as predictive because its main purpose was to develop supervised models capable of identifying and predicting THC use.

Study Population

The reference population consisted of individuals aged 12 years and older living in private households in Mexico who participated in the 2025 National Survey on Drug, Alcohol, and Tobacco Use (ENCODAT 2025). The survey included visits to 23,950 households with regional representativeness and aimed to interview 17,591 individuals.

For the present study, the database corresponding to the Individual Questionnaire was used. This instrument was administered directly to each selected household member. The final sample consisted of 14,113 records with valid information for the response variable and the predictor variables included in the analysis.

Data Collection

The data used in this study were obtained from the official website of the 2025 National Survey on Drug, Alcohol, and Tobacco Use (ENCODAT 2025). The survey was conducted by the Ministry of Health of the Government of Mexico in coordination with the National Commission on Mental Health and Addictions (CONASAMA), the Ramón de la Fuente Muñiz National Institute of Psychiatry (INPRFM), and the National Institute of Public Health (INSP), through the Center for Research in Evaluation and Surveys (CIEE).
The official ENCODAT 2025 variable catalog was consulted to identify and select the study variables. This catalog provided the descriptions and coding schemes for the variables used in the analysis.

Study Variables

Response Variable

The dependent variable corresponded to item *di1a* of the 2025 ENCODAT Individual Questionnaire: “Have you ever taken, used, or tried marijuana, also known as weed, cannabis, herb, or *café*, as well as oils or products containing THC, such as cookies, candy, or beverages?” (Secretaría de Salud, 2025). Responses were recoded into a dichotomous variable, where a value of 1 represented a history of cannabis or THC-containing product use and a value of 0 represented no history of use.

Predictor Variables

A total of 16 predictor variables were selected from the official 2025 ENCODAT variable catalog. The selection was based on the researchers’ judgment and included variables

considered potentially related to THC use. Table 1 presents the main characteristics of the variables used to develop the predictive models.

Variable	Description	Measurement Scale	Coding
ds2	Sex assigned at birth	Nominal	Male (1); Female (2)
ds3	Age	Ratio	Age in completed years
ds6	Marital status	Nominal	Married (1); Cohabiting/Consensual union (2); Separated (3); Divorced (4); Widowed (5); Single (6)
ds8	Student status	Ordinal	No, has never attended school (1); No, but previously attended school (2); Yes (3)
ds9	Highest educational level attained	Ordinal	Incomplete primary education (1); Complete primary education (2); Incomplete secondary education (3); Complete secondary education (4); Incomplete high school education (5); Complete high school education (6); Incomplete university education (7); Complete university education (8); Postgraduate education (9)
ds10	Recent employment status	Dichotomous nominal	Yes (1); No (2)
ed1	Previous drug offer	Dichotomous nominal	Yes (1); No (2)
ts1 – ts3	Perceptions of family members, friends, and peers regarding alcohol use	Ordinal	They would approve (1); They would neither approve nor disapprove (2); They would disapprove (3)
ts4 – ts6	Perceptions of family members, friends, and peers regarding illegal drug use	Ordinal	They would approve (1); They would neither approve nor disapprove (2); They would disapprove (3)
ts7 – ts9	Perceptions of family members, friends, and peers regarding high-risk drug use	Ordinal	They would approve (1); They would neither approve nor disapprove (2); They would disapprove (3)

Table 1. Description of predictor variables used in the classification models.
Note: Variables TS1–TS9 were grouped by substance type. The complete wording of the items can be found in the Individual Questionnaire of the ENCODAT 2025.

Statistical Analysis

Data processing and model fitting were performed in Python 3. The pandas and NumPy libraries were used for data manipulation, scikit-learn was used to implement the classification models and calculate performance metrics (Pedregosa et al., 2011), and imbalanced-learn was used to apply the SMOTE technique (Lemaître et al., 2017).

Results were visualized using matplotlib, and the code was developed in a Jupyter Notebook within the Google Colaboratory environment.
Before fitting the classification models, an exploratory data analysis was conducted. Absolute and relative frequencies were calculated for the response variable, and bar charts, pie charts, histograms, and boxplots were created to describe the distribution of the variables and identify possible differences between individuals with and without a history of use.

The dataset was then divided into a training set (80%) and a test set (20%) using stratified random sampling, preserving the original proportions of the response variable categories. The predictor variables were standardized using *z* scores based only on information from the training set to prevent information leakage between subsets (James et al., 2023). Because the response variable categories were imbalanced, the Synthetic Minority Oversampling Technique (SMOTE) was applied to the training set used for the classification tree, generating synthetic observations for the minority class based on nearby real observations in the feature space (Chawla et al., 2002).
To develop the predictive models, a classification tree was implemented using the *DecisionTreeClassifier* algorithm. Hyperparameters were selected through grid search with five-fold stratified cross-validation. Different splitting criteria and levels of tree complexity were evaluated, following the principles of classification and regression trees described by Breiman et al. (2017).
In addition, a binary logistic regression model with Elastic Net regularization was fitted. This approach combines L1 and L2 regularization to reduce overfitting and improve the stability of estimates in the presence of potentially correlated variables (Zou & Hastie, 2005). Model parameters were estimated using the SAGA optimization algorithm (Defazio et al., 2014), and class weighting was incorporated through the class_weight='balanced' parameter, assigning greater weight to the minority class during training.

Model performance was evaluated on the test set using confusion matrices. From these, accuracy, sensitivity, and specificity were calculated. For logistic regression, the area under the receiver operating characteristic curve (AUC-ROC) was also estimated to assess the model's discriminative ability independently of the selected classification threshold (Fawcett, 2006). Accuracy was used as the primary criterion for comparing the models. In addition, sensitivity and specificity were analyzed to evaluate the classification implications of false negatives and false positives, particularly when identifying individuals with a history of use.

RESULTS

Exploratory Analysis

The original dataset consisted of 19,200 records. After selecting the 16 predictor variables and the response variable, and excluding records with missing values in any of these variables, an analytical sample of 14,113 valid observations was obtained. The response variable *di1a* showed a highly imbalanced distribution: 12,264 participants (86.9%) reported that they had never used marijuana or THC-containing products, whereas 1,849 participants (13.1%) reported a history of use. This class imbalance was considered during the modeling stage.

Regarding the descriptive statistics, age (*ds3*) was the only continuous variable in the dataset. The mean age was 33.95 years (median = 33.0; range = 12–65 years), indicating a fairly symmetric distribution that included participants from adolescence through older adulthood (Table 2).

VARIABLE	MEAN	MEDIAN	MINIMUM	MAXIMUN
Age (ds3)	33.95	33.0	12.0	65.0

Table 2. Descriptive statistics for continuous variables
Note: Categorical variables are described using relative frequencies in the text.

Regarding the categorical sociodemographic variables, the sample consisted mostly of men (*ds2*: 53.2% men and 46.8% women). For marital status (*ds6*), the most common category was single (52.9%), followed by married (13.9%) and cohabiting (10.5%). Separated (3.4%), divorced (9.7%), and widowed (2.3%) participants represented smaller proportions. Most participants reported that they were not currently studying (*ds8*: 70.3% not studying and 29.7% studying), and the majority were employed (*ds10*: 77.4% employed and 22.6% not employed). Regarding variable *ed1*, which indicates whether participants had ever been offered drugs, 73.1% responded no, whereas 26.9% responded yes.

For the variables measuring perceived social attitudes toward alcohol use, illegal drug use, and high-risk drug use (*ts1–ts9*), responses showed a clear tendency toward disapproval. Regarding alcohol use, 74.2% of participants perceived that their family would disapprove (*ts1*). This percentage increased to 95.9% for perceived family disapproval of illegal drug use (*ts4*) and to 96.4% for perceived family disapproval of methamphetamine, heroin, or fentanyl use (*ts7*). Perceptions of friends and peers followed the same pattern, although less strongly: 53.6% reported that their best friend would disapprove of alcohol use (*ts2*), and 85.8% reported that their best friend would disapprove of illegal drug use (*ts5*).

The exploratory analysis by class revealed relevant differences between individuals with and without a history of use. The most pronounced difference was observed for *ed1* (“Have you ever been offered drugs?”): 70.5% of participants who reported marijuana use had previously been offered drugs, compared with 20.3% of those without a history of use. This 50-percentage-point difference suggests that direct exposure to drug offers is strongly associated with use. Regarding sex, individuals with a history of use were predominantly male (73.6% men, compared with 50.1% among those without a history of use).With respect to the social environment, participants with a history of use reported lower perceived disapproval from friends regarding alcohol use (32.2% indicated that their friends would disapprove, *ts2*, compared with 56.8% among those without a history of use) and lower perceived disapproval from school or workplace peers regarding illegal drug use (69.2% vs. 83.1%, *ts6*). This indicates that individuals with a history of use perceived a more permissive social environment. Regarding age (Figure 1), although the median was identical in both groups (33 years), participants without a history of use showed greater variability (SD = 16.21 vs. 13.22) and a notably lower first quartile (Q1 = 17 vs. 24). This finding reflects that the group without a history of use included a larger

proportion of adolescents aged 12–17 years, whereas participants with a history of use were more concentrated in the 24–45-year age range.

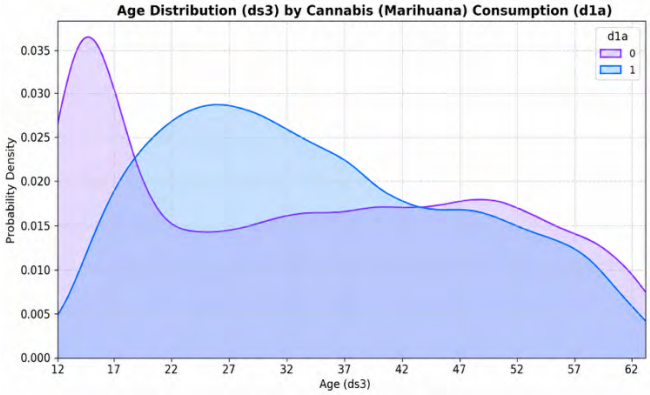

Figure 1. Density graph of the age variable (ds3) according to the class of the response variable di1a**.**

Classification Tree

For training the classification tree, the Synthetic Minority Over-sampling Technique (SMOTE) was applied exclusively to the training set (80% of the sample, *n* = 11,290), balancing both classes to 9,811 observations each. Hyperparameter tuning through grid search (*GridSearchCV*) with five-fold stratified cross-validation evaluated 160 combinations. The optimal configuration used the Gini criterion, a maximum depth of 12 levels, no limit on the number of variables considered at each node, a minimum of 10 samples required to split a node, and a minimum of 5 samples per leaf. This configuration achieved a cross-validation accuracy of 0.8842.

Structurally, the final model reached the maximum depth of 12 levels and contained 841 nodes, of which 421 were leaves or terminal nodes. The variable *ed1* provided the best initial split between the classes. Figure 2 presents the graphical representation of the classification tree.

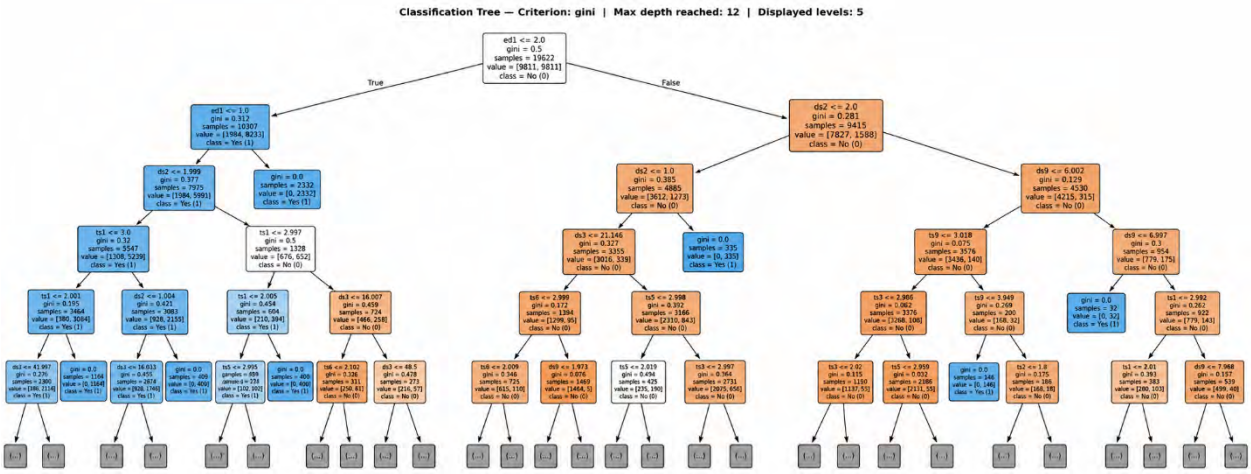

Figure 2. Graphical representation of the Classification Tree on the test set
Note: For analytical clarity and to avoid visual clutter in the document, the graphical representation of the model was strictly limited to its first 5 decision levels.

When evaluated on the test set (*n* = 2,823), the classification tree achieved an accuracy of 0.8402, a sensitivity of 0.4054, and a specificity of 0.9058. The low sensitivity reflects the model's difficulty in correctly identifying individuals with a history of use, as it classified 220 of the 370 actual positive cases in the test set as negative (Figure 3). Regarding feature importance, the three variables that contributed most to reducing impurity were age (*ds3*, 7.18%), having been offered drugs (*ed1*, 58.08%), and sex (*ds2*, 12.45%). Together, these three variables accounted for 77.71% of the model's total feature importance, highlighting the influence of sociodemographic factors and direct drug exposure on the prediction of use.

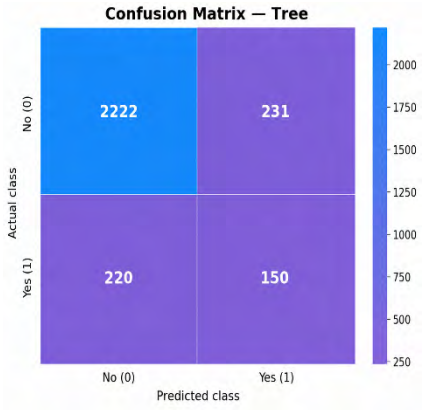

Figure 3. Confusion matrix of the Classification Tree on the test set.

Logistic Regression With Class Weighting

The logistic regression model was fitted using a class-weighting strategy (class_weight='balanced'), which assigns each observation a weight inversely proportional to the frequency of its class in the training set without generating synthetic observations. The hyperparameters used were Elastic Net regularization, which combines L1 and L2 penalties, with *C* = 0.01, *l1_ratio* = 0.75, and the SAGA optimizer, using a maximum of 2,000 iterations. On the test set, the model achieved an accuracy of 0.7807, a sensitivity of 0.7162, a specificity of 0.7905, and an AUC-ROC of 0.8198. It correctly identified 265 of the 370 individuals with a history of use in the test set (Figure 4).

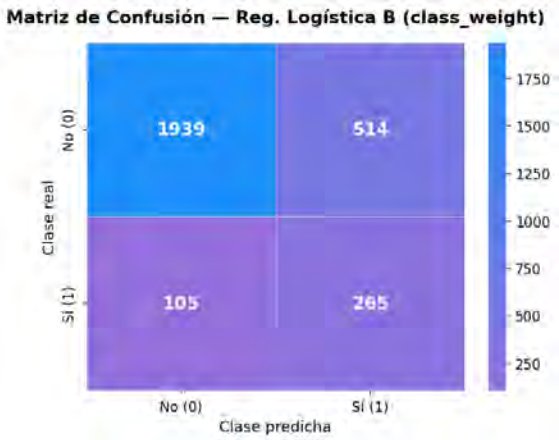

Figure 4. ROC curve of the Logistic Regression with class weighting (AUC = 0.820).

The ROC curve (Figure 5) shows a satisfactory discriminatory capacity, far surpassing the random classifier.

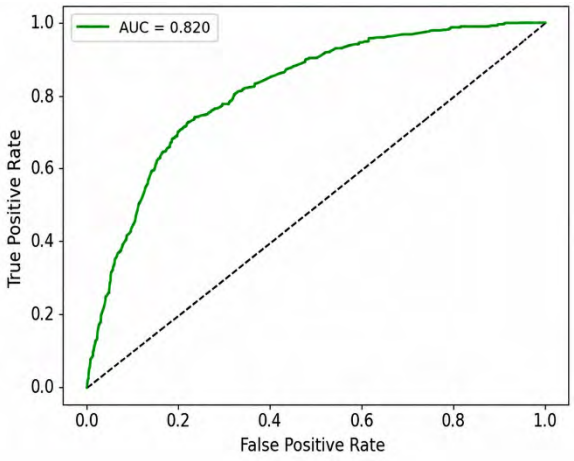

Figure 5. ROC curve of the Logistic Regression with class weighting (AUC = 0.820).

The analysis of the model coefficients (Figure 6) identified the factors most strongly associated with marijuana or THC-containing product use. Elastic Net regularization reduced four predictors (*ds3*, *ds6*, *ts4*, and *ts9*) to zero, indicating that they did not provide relevant discriminatory information under the applied regularization. The variable with the greatest effect in the model was *ed1* (having been offered drugs; $\beta = -0.83$, OR = 0.43). Individuals who had never been offered drugs had approximately 57% lower odds of having used marijuana, a result consistent with the exploratory analysis, which showed a 50-percentage-point difference between groups. Active employment status (*ds10*, OR = 0.74) and male sex (*ds2*, OR = 0.75) also showed negative associations with use. Among the positive predictors, educational attainment (*ds9*, OR = 1.09) had the largest magnitude, suggesting a slight association between higher educational attainment and the likelihood of use, possibly mediated by greater social exposure and increasing age.
Perceived family disapproval of alcohol use (*ts1*, OR = 0.86) and perceived disapproval from friends (*ts2*, OR = 0.90) showed a moderate protective effect. Similarly, family disapproval of illegal drug use (*ts5*, OR = 0.90) also acted as a protective factor, suggesting that social norms within the immediate environment regarding both alcohol and drugs consistently reduce the likelihood of marijuana use, regardless of whether they originate from family members or friends.

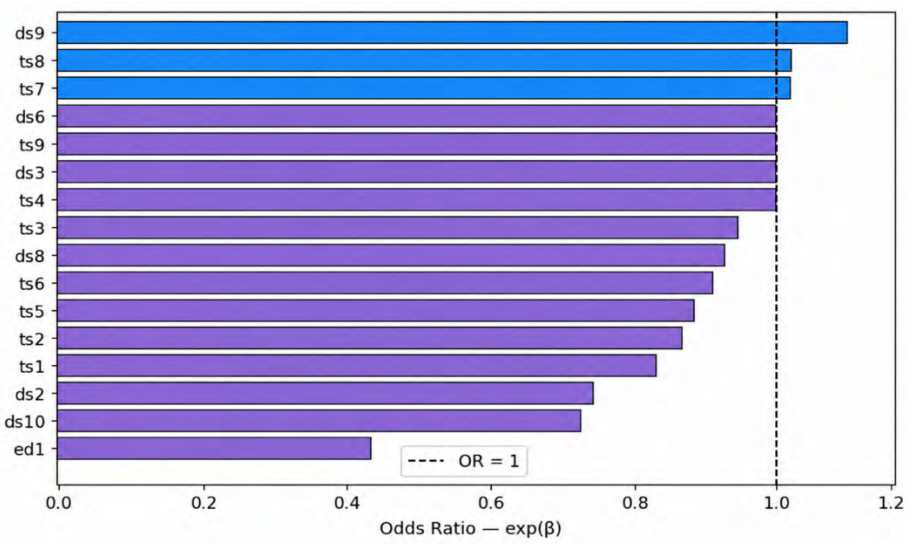

Figure 6. Odds Ratios of the Logistic Regression with class weighting.
Note: Blue bars: OR > 1 (higher probability of consumption); purple bars: OR < 1 (lower probability of consumption). The dashed line indicates OR = 1 (no effect).

Model Comparison

Table 3 summarizes the performance metrics of both models evaluated on the same test set, allowing a direct comparison despite the use of different strategies to address class imbalance. The classification tree, trained using SMOTE-balanced data, achieved higher overall accuracy (0.8402 vs. 0.7807) and higher specificity (0.9058 vs. 0.7905), making it more conservative when classifying individuals as nonusers. In contrast, logistic regression with class weighting achieved higher sensitivity (0.7162 vs. 0.4054) and an AUC-ROC of 0.8198, indicating better overall discriminatory ability and substantially more effective identification of individuals with a history of use.

Model	Accuracy	Sensitivity	Specificity	AUC-ROC
Classification Tree (SMOTE)	0.8402	0.4054	0.9058	—
Logistic Regression (class_weight)	0.7807	0.7162	0.7905	0.8198

Table 3. Performance metrics evaluated for the models in the test set

Both models address class imbalance through different approaches and produce complementary results. The main difference lies in the implicit classification threshold associated with each strategy: SMOTE modifies the training distribution so that the tree learns from balanced classes, whereas class weighting penalizes errors involving the minority class directly within the logistic regression loss function. In a public health context, where correctly identifying individuals with a history of marijuana use is a priority for guiding preventive interventions, logistic regression with class weighting represents the more practically useful alternative because of its higher sensitivity.

CONCLUSIONS

The results indicate that the variables included in the 2025 ENCODAT have predictive ability for identifying a history of marijuana and THC-containing product use. The evaluated models also showed different behaviors when addressing the classification problem, particularly because of the imbalance in the response variable. Overall, the comparison showed that both models are viable alternatives for this type of problem, although their performance may vary depending on the metric considered and the specific objective of the classification task.

Among the main findings, sociodemographic, drug exposure, and social tolerance variables provided useful information for classifying individuals according to their history of use. In addition, the model development process showed that imbalance in the response variable is an important factor to consider when comparing and interpreting the performance of different classification methods.

One of the main contributions of this study is the application of supervised learning techniques to recent data from the Mexican population obtained through the 2025 ENCODAT. The findings may serve as a reference for future research aimed at developing predictive models using data from national surveys and other public health-related datasets.

The limitations of the study include the variables available in the survey and those selected for the analysis, as other factors potentially associated with marijuana use were not considered. Variables related to mental health, family and social environments, peer influence, risk perception, and the history of other substance use have been identified in previous research as relevant factors for understanding this behavior. In addition, the models were developed exclusively using information from the 2025 ENCODAT; therefore, the results may not necessarily be generalized to other populations or contexts without further validation.

Finally, future studies should incorporate variables related to psychological, family, and social factors, as well as evaluate other machine learning algorithms and different strategies for addressing class imbalance. This would help strengthen the predictive ability of the models and expand the available knowledge on the use of classification techniques in the study of substance use.

LITERATURE

Araujo Zarate, P., Martínez Lobo, D., y Contreras Chávez, J. (2024). Modelos predictivos en la clasificación de nacidos vivos y mortinatos: un estudio comparativo entre técnicas de machine learning y regresión logística en función de variables sociodemográficas y clínicas. *Revista Ciencias Biomédicas, 13*(4), 175–189. https://doi.org/10.32997/rcb-2024-4940

Breiman, L., Friedman, J. H., Olshen, R. A., & Stone, C. J. (2017). *Classification and regression trees*. CRC Press. https://doi.org/10.1201/9781315139470

Chawla, N. V., Bowyer, K. W., Hall, L. O., & Kegelmeyer, W. P. (2002). SMOTE: Synthetic Minority Over-sampling Technique. *Journal of Artificial Intelligence Research, 16*, 321–357. https://doi.org/10.1613/jair.953

Comisión Nacional de Salud Mental y Adicciones, Instituto Nacional de Psiquiatría Ramón de la Fuente Muñiz, & Instituto Nacional de Salud Pública. (2025). *Encuesta Nacional de Consumo de Drogas, Alcohol y Tabaco 2025: Resumen ejecutivo*. Secretaría de Salud. https://encuestas.insp.mx/repositorio/encuestas/ENCODAT2025/informes.php

Defazio, A., Bach, F., & Lacoste-Julien, S. (2014). SAGA: A fast incremental gradient method with support for non-strongly convex composite objectives. *Advances in Neural Information Processing Systems, 27*, 1646–1654. https://doi.org/10.48550/arXiv.1407.0202

Dell, N. A., Vaughn, M. G., Prasad Srivastava, S., Alsolami, A., & Salas-Wright, C. P. (2022). Correlates of cannabis use disorder in the United States: A comparison of logistic regression, classification trees, and random forests. *Journal of Psychiatric Research, 151*, 590–597. https://doi.org/10.1016/j.jpsychires.2022.05.021

Fawcett, T. (2006). An introduction to ROC analysis. *Pattern Recognition Letters, 27*(8), 861–874. https://doi.org/10.1016/j.patrec.2005.10.010

James, G., Witten, D., Hastie, T., & Tibshirani, R. (2023). *An introduction to statistical learning: With applications in Python* (2nd ed.). Springer. https://doi.org/10.1007/978-3-031-38747-0

Lansford, J. E., Goulter, N., Godwin, J., McMahon, R. J., Dodge, K. A., Crowley, M., Pettit, G. S., Bates, J. E., & Lochman, J. E. (2023). Predictors of problematic adult alcohol, cannabis, and other substance use: A longitudinal study of two samples. *Development and Psychopathology, 35*(4), 2028–2043. https://doi.org/10.1017/S0954579422000670

Leatherdale, S. T., Battista, K., Patte, K. A., MacKillop, J., & Bélanger, R. (2025). Using decision trees to examine risk profiles for cannabis use among large samples of underage youth before and after cannabis legalization in Canada. *Addictive Behaviors Reports, 22*, 100632. https://doi.org/10.1016/j.abrep.2025.100632

Lemaître, G., Nogueira, F., & Aridas, C. K. (2017). Imbalanced-learn: A Python toolbox to tackle the curse of imbalanced datasets in machine learning. *Journal of Machine Learning Research, 18*(17), 1–5. http://jmlr.org/papers/v18/16-365.html

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., & Duchesnay, É. (2011). Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research, 12*, 2825–2830. https://jmlr.org/papers/v12/pedregosa11a.html

Pedrero, V., Reynaldos-Grandón, K., Ureta-Achurra, J., y Cortez-Pinto, E. (2021). Generalidades del Machine Learning y su aplicación en la gestión sanitaria en Servicios de Urgencia. *Revista Médica de Chile, 149*(2), 248–254. https://doi.org/10.4067/S0034-98872021000200248

Schwebel, F. J., Richards, D. K., Pfund, R. A., Joseph, V. W., & Pearson, M. R. (2022). Using decision trees to identify salient predictors of cannabis-related outcomes. *Journal of Psychoactive Drugs, 54*(5), 419–428. https://doi.org/10.1080/02791072.2021.2014081

United Nations Office on Drugs and Crime. (2025). *World Drug Report 2025*. United Nations Office on Drugs and Crime. https://www.unodc.org/unodc/data-and-analysis/world-drug-report-2025.html

Unlu, A., Hakkarainen, P., Karjalainen, K., & Subasi, A. (2024). Risk prediction model for cannabis use with artificial intelligence approach. *Journal of Substance Use, 29*(6), 1077–1084. https://doi.org/10.1080/14659891.2023.2242961

World Health Organization. (2024). *Global status report on alcohol and health and treatment of substance use disorders*. World Health Organization. https://www.who.int/publications/i/item/9789240096745

Zou, H., & Hastie, T. (2005). Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society: Series B, 67*(2), 301–320. https://doi.org/10.1111/j.1467-9868.2005.00503.x

CHAPTER 4

Prediction of the level of financial well-being using Logistic Regression: analysis of the National Survey on Financial Health (ENSAFI) 2023

Mariela Denisse García Jimenez, Victoria Saldaña Hernández, Judith Guadalupe Montero Mora, Zoylo Morales Romero

ABSTRACT

Financial health is a component of individual well-being that includes the ability to meet financial obligations, maintain savings, and achieve future stability. In Mexico, the National Institute of Statistics and Geography (INEGI) conducted the 2023 National Survey on Financial Health (ENSAFI), which collected information from 20,448 individuals aged 18 years and older. This study developed a logistic regression model to predict financial well-being using a variable selection pipeline. The target variable was regrouped into a binary outcome: low well-being (levels 1 and 2) and high well-being (levels 3 and 4). The model was estimated using maximum likelihood with 15 predictor variables and achieved a cross-validated AUC of 0.878 using five-fold cross-validation, an accuracy of 0.793, and a McFadden's R^2 of 0.370, indicating a good model fit. The most important predictor variables belonged to five dimensions: ability to respond to unexpected expenses, saving habits, perceived financial self-confidence, emotional stress, and socioeconomic indicators. The implications for public policy are discussed based on odds ratios and 95% confidence intervals.

Keywords: financial well-being, logistic regression, ENSAFI, INEGI.

INTRODUCTION

Financial well-being is a multifaceted construct that reflects an individual's ability to meet current financial obligations, maintain stability when facing unexpected events, and plan for the future with confidence. In Mexico, where a portion of the population works under informal conditions and has limited access to financial products, measuring and predicting financial well-being is important for developing financial inclusion and economic education programs.

According to Vallejo-Trujillo and Martínez (2016), financial well-being is an indicator of a person's overall well-being and reflects their level of financial education. It can be defined as a mental or physical response that also provides greater confidence in an individual's ability to repay debts and generate income.

According to a study conducted by Ortiz et al. (2019), financial well-being has a positive and significant correlation with life satisfaction. In addition, the authors identified income and the number of children living in the household as predictors of financial well-being, along with variables related to physical and mental health.

The 2023 National Survey on Financial Health (ENSAFI), published by the National Institute of Statistics and Geography (INEGI, 2023), collects information on income,

expenses, debt, savings, financial habits, perceptions, and financial stress among the adult population in Mexico. The survey includes a variable called *Well-Being Level* (*NIV_BIENES*), which consists of four ordinal levels. Predicting this ordinal variable from direct measures makes it possible to identify the observable factors associated with it and estimate the magnitude of their effects.

To analyze the determinants of financial well-being using ENSAFI data, a method is required that not only classifies individuals according to their level of financial well-being but also makes it possible to identify the influence of each variable on the response. Logistic regression meets both requirements because its coefficients indicate the probability of belonging to one group or another based on an individual's characteristics. In addition, maximum likelihood estimation makes it possible to obtain confidence intervals for statistical inference (Agresti, 2015).

The objective of this study is to develop a parsimonious logistic regression model, validated through stratified cross-validation, to predict the binary level of financial well-being among the adult population in Mexico. The research question is as follows: Which variables from the ENSAFI questionnaire are significantly associated with the level of financial well-being, and to what extent do they influence it?

The proposed hypothesis is that financial behavior factors and positive perceptions have an effect on financial well-being that is comparable to or greater than that of objective factors such as income or age.

METHODOLOGY

This was a quantitative, observational study with a predictive approach. A secondary analysis was conducted using the public 2023 ENSAFI dataset. The methodological design was organized into five phases:

1. Preprocessing and recategorization of the target variable.
2. Filtering variables based on missing data.
3. Statistical univariate screening.
4. Variable selection using Lasso (L1) regularization.
5. Estimation of the final model through maximum likelihood, including odds ratios and 95% confidence intervals.

The target population consisted of individuals aged 18 years or older residing in private households throughout Mexico. The survey included information from 20,448 individuals across the 32 federal entities.

The *NIV_BIENES* variable, which originally contained four ordinal categories, was recategorized into a binary variable for the logistic regression model. Class 0, representing low well-being, grouped the original levels 1 and 2, whereas class 1, representing high well-being, grouped levels 3 and 4. This recategorization produced nearly balanced classes: 10,056 observations with low financial well-being (49.2%) and 10,392 with high financial well-being (50.8%).

Starting with the 253 variables included in the questionnaire, a stepwise variable selection pipeline was applied. The 15 final variables selected by the Lasso model are described in Table 1.

Variable	Description	Measurement Scale	Coding

Variable	Descripción	Escala de medición	Codificación	
P7_6_1	Capacity to pay for an emergency using savings		Liquidity	Ordinal 1-3
P7_9_2	Confidence to plan your financial future		Self-confidence	Ordinal 1-5
P6_11_3	In the last month, he made pending purchases.		Consumption	Nominal 1-2
P8_4	Level of financial concern (0-10)		Stress	Ratio 0-10
P7_10_09	He believes he will get the things he wants		Orientation	Ordinal 1-5
P8_2_2	She has felt sad about her financial situation.	Emotional stress	Nominal 1-2	
P6_9	Sufficient income to cover expenses without going into debt	Liquidity	Nominal 1-2	
INCOME_M	Monthly income from work	Socioeconomic	Ratio (MXN)	
P7_9_3	Confidence in decision-making regarding financial products	Self-confidence	Ordinal 1-5	
P6_3	Do you currently have savings account(s)	Saving	Nominal 1-2	
P7_10_01	You can control your financial situation	Control	Ordinal 1-5	
P6_11_2	Saved money or increased their savings during the past month	Saving	Nominal 1-2	
P7_9_1	Confidence in managing money on a daily basis	Self-confidence	Ordinal 1-5	
AGE_V	Verified age of the person surveyed	Socioeconomic	Ratio (años)	
P7_10_02	When he makes plans, he carries them out	Control	Ordinal 1-5	

Table 1. Description of variables

Given the length of the questionnaire, a stepwise selection process was conducted. In the first phase, folio identifier variables, screening-question variables, and items *P7_7* and *P7_8* were excluded. These items are components used to construct the *NIV_BIENES* variable, and their inclusion could cause overfitting in the model. In addition, variables with more than 40% missing values were removed, reducing the dataset to 182 variables.

In the second phase, bivariate statistical tests were applied: the chi-square test for variables with 15 or fewer unique categories and point-biserial correlation for variables with a wider range of values. Only variables with a *p* value below .05 were selected, resulting in 173 candidate variables. In addition, one variable from each pair with a Pearson correlation greater than .80 was removed. Finally, in the third phase, an L1-penalized logistic regression model (Lasso, $C = 0.05$) was fitted using the 50 highest-ranked variables from the univariate screening. This method forces the coefficients of less relevant variables to exactly zero, thereby performing automatic variable selection and resulting in 15 final variables.

Data were obtained from the public 2023 ENSAFI dataset downloaded from the INEGI website. The survey was conducted through face-to-face household interviews across Mexico using structured questionnaires administered by trained personnel. Missing values were imputed using the median of each variable.

The final logistic regression model was estimated through maximum likelihood using the Python package statsmodels. Standardized coefficients were obtained, and the results were reported as odds ratios with their corresponding 95% confidence intervals, indicating the effect of each variable on financial well-being.
Model fit was assessed using McFadden's R^2, with values above .30 considered indicative of a good fit. In addition, the model's predictive performance was evaluated using accuracy, macro F1 score, and the area under the receiver operating characteristic curve through five-fold stratified cross-validation. This procedure helped ensure that the results did not depend on a particular division of the data.

RESULTS

The sample consisted of 20,448 observations. The binary distribution of the target variable was nearly balanced: 10,056 observations corresponded to low financial well-being (49.2%), and 10,392 corresponded to high financial well-being (50.8%). This balance reduced the risk of bias toward the majority class and made resampling techniques unnecessary. Figure 1 illustrates this distribution.

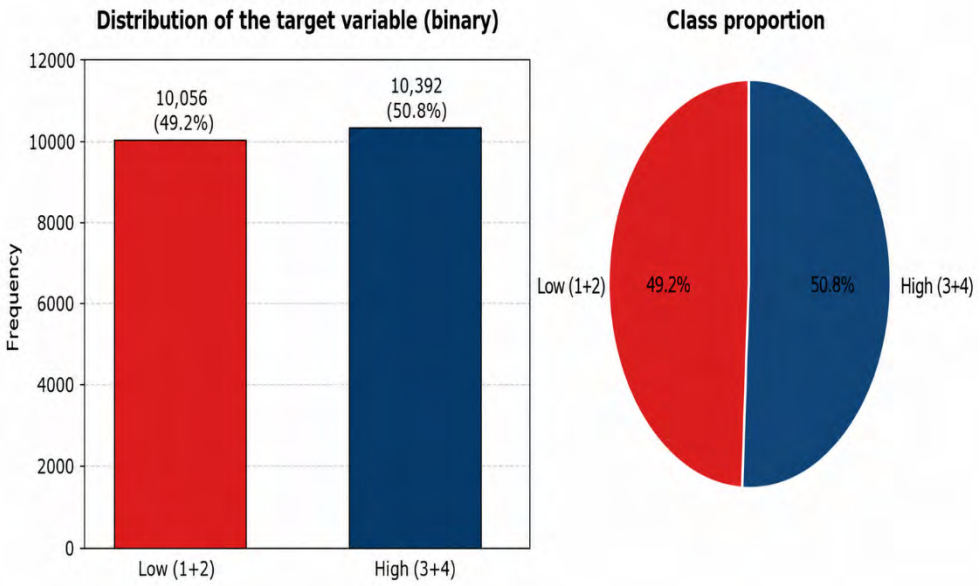

Figure 1. Distribution of the target variable (level of financial well-being)

Table 2 presents the performance metrics of the logistic regression model using the 15 variables selected through Lasso. The model was estimated on the full sample and evaluated using five-fold cross-validation to reduce the risk of overfitting.

Metric	Value	Interpretation
Accuracy (5-Fold CV)	0.793	Approximately 8 out of 10 individuals were correctly classified.
Macro F1-Score (5-Fold CV)	0.792	Balance between precision and recall across both classes.
AUC-ROC (5-Fold CV)	0.878	Good discriminative ability.
Sensitivity	78.7%	High well-being cases correctly identified.
Specificity	80.6%	Low well-being cases correctly identified.
McFadden's R^2	0.370	Excellent model fit (values above 0.30 are considered excellent).

Table 2. Performance metrics of the logistic regression model

Table 3 presents the standardized coefficients, odds ratios, and 95% confidence intervals for the 15 variables included in the model. All coefficients were statistically significant ($p <$ .001). The odds ratio indicates how the odds of belonging to the high financial well-being class change for each one-standard-deviation increase in a predictor variable, while holding all other variables constant.

Variable	β	OR	IC 95%	p	Brief Description
P7_6_1	+0.551	1.735	[1.668–1.803]	<0.001	Able to cover a financial emergency using savings
P7_9_2	+0.447	1.564	[1.489–1.643]	<0.001	Confidence in planning for the financial future
P6_11_3	+0.366	1.442	[1.379–1.507]	<0.001	Made previously postponed purchases during the past month
P7_10_09	+0.328	1.389	[1.328–1.452]	<0.001	Believes they will obtain the things they want
P8_4	+0.321	1.378	[1.320–1.439]	<0.001	Level of financial concern
P6_9	+0.254	1.289	[1.238–1.341]	<0.001	Sufficient income without going into debt
P6_11_2	+0.227	1.254	[1.197–1.315]	<0.001	Saved money or increased savings during the past month
P7_10_01	+0.219	1.245	[1.190–1.302]	<0.001	Control over their financial situation
P7_9_3	+0.208	1.231	[1.181–1.284]	<0.001	Confidence in making financial decisions
P6_3	+0.197	1.218	[1.170–1.268]	<0.001	Has active savings accounts
P7_9_1	+0.181	1.199	[1.144–1.257]	<0.001	Confidence in managing money
P7_10_02	+0.111	1.117	[1.067–1.170]	<0.001	Follows through on the plans they make
AGE_V	+0.141	1.151	[1.107–1.196]	<0.001	Verified age (years)
INCOME_M	−0.331	0.718	[0.674–0.766]	<0.001	Monthly employment income (MXN)
P8_2_2	−0.295	0.744	[0.715–0.775]	<0.001	Has felt sad because of their financial situation

Table 3. Standardized Coefficients, Odds Ratios and 95% Confidence Intervals

The variables with the largest positive odds ratios were *P7_6_1* (OR = 1.735, 95% CI [1.668, 1.803]) and *P7_9_2* (OR = 1.564, 95% CI [1.489, 1.643]). The first result indicates that, for each one-standard-deviation increase in the ability to cover unexpected expenses using savings, the odds of high financial well-being increased by 73.5%. The second result

shows that confidence in planning for one's financial future was the predictor with the second-largest effect, highlighting the role of the subjective dimension of well-being.
The variables with negative coefficients were *INGRESO_M* (OR = 0.718) and *P8_2_2* (OR = 0.744). The negative coefficient for income may be explained by the fact that additional income no longer independently differentiates perceived financial well-being after the other variables are held constant. The *P8_2_2* variable indicates that having felt sad because of one's financial situation reduced the odds of high financial well-being by 25.6%, highlighting the importance of the emotional dimension among the predictor variables.

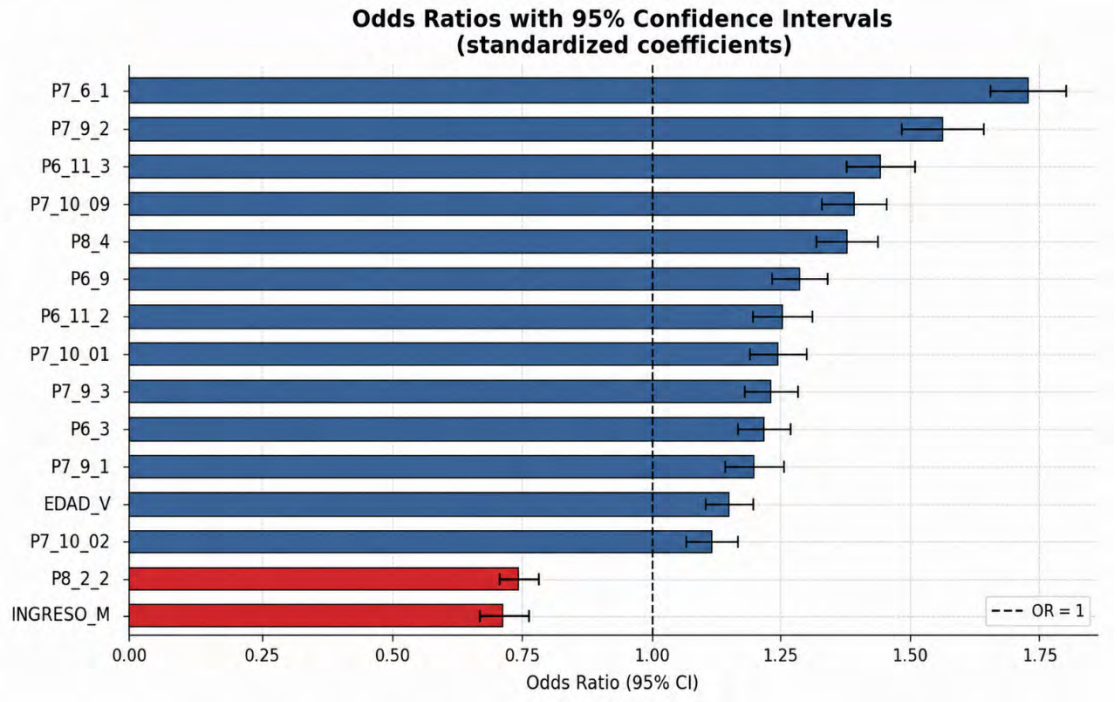

Figure 2. *Odds Ratios with 95% Confidence Intervals*

Figure 3 presents the confusion matrices in both case counts and percentages. The model correctly classified 80.6% of the high financial well-being cases (sensitivity) and 78.7% of the low financial well-being cases (specificity), showing a balanced classification pattern.

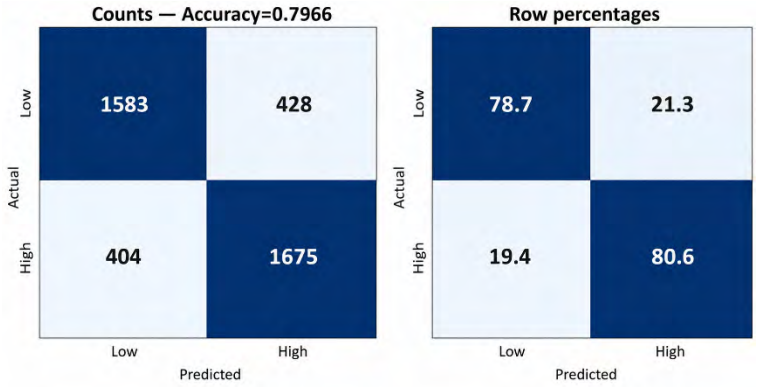

Figure 3. *Confusion matrix of the model*

Figure 4 presents the receiver operating characteristic (ROC) curve (AUC = 0.878) and the precision-recall curve (AP = 0.877). The AUC falls within the range considered good, between 0.80 and 0.90. In addition, the precision-recall curve remained well above the baseline value of 0.508 throughout, confirming the model's ability to distinguish between the two classes across different decision thresholds.

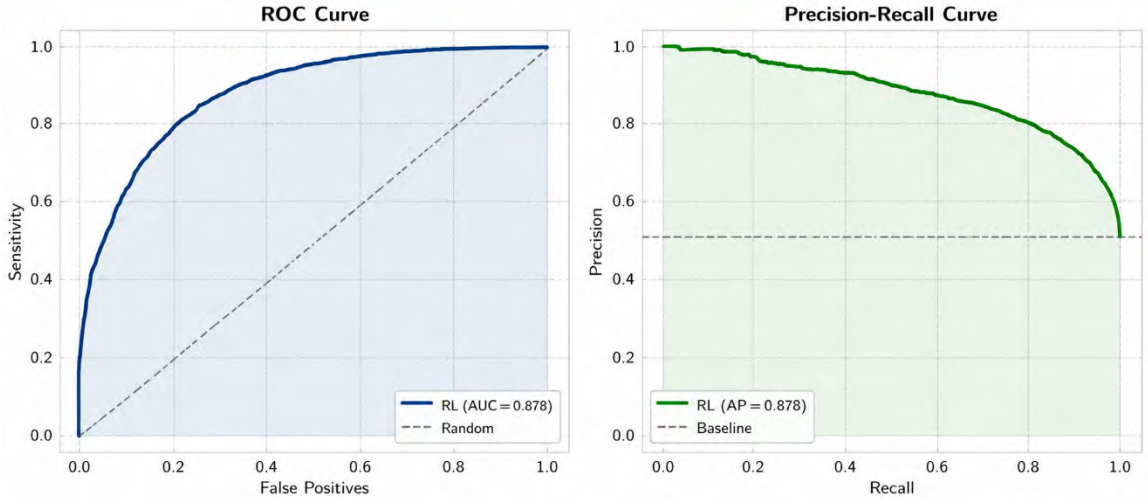

Figure 4. *ROC Curve and Recall Accuracy of the model*

The 15 significant predictor variables were grouped into five conceptual dimensions. The liquidity and payment capacity dimension (*P7_6_1*, *P6_9*) reflects whether an individual has the resources to cover unexpected expenses and meet financial needs without taking on debt. The saving habits dimension (*P6_11_2*, *P6_3*) captures active saving behaviors. The financial self-confidence dimension (*P7_9_1*, *P7_9_2*, *P7_9_3*, *P7_10_01*, *P7_10_02*, *P7_10_09*) groups perceptions related to financial control and planning. The emotional stress dimension (*P8_2_2*, *P8_4*) reflects subjective distress associated with one's financial situation. Finally, the socioeconomic dimension (*INGRESO_M*, *EDAD_V*) captures objective characteristics of the respondent.

CONCLUSIONS

This study demonstrates that it is possible to predict the level of financial well-being among the adult population in Mexico with good performance using a logistic regression model based on 15 variables selected through a Lasso pipeline from the 2023 ENSAFI. The model achieved a cross-validated AUC of 0.878, an accuracy of 0.793, and a McFadden's R^2 of 0.370, indicating excellent fit in the context of this study. In addition, five-fold cross-validation confirmed the stability of the results.

The main finding is that the most influential predictors were not strictly economic variables but rather factors related to financial behavior and perception. The ability to cover unexpected expenses using savings was the strongest predictor, with an odds ratio of 1.735, followed by confidence in planning for the financial future and the ability to make pending purchases during the previous month. These findings support the hypothesis that

financial education interventions and the development of emergency savings may have a greater effect on perceived financial well-being than marginal increases in income.

The emotional dimension of financial well-being also proved important. Feeling sad for financial reasons (*P8_2_2*, OR = 0.744) significantly reduced the likelihood of high financial well-being, showing a stronger effect than objective variables such as income or age and some behavioral factors. Consequently, this finding suggests the importance of incorporating indicators of financial mental health into financial support and inclusion programs.

The methodological pipeline used in this study, which combines univariate screening and Lasso regularization before maximum likelihood estimation, is appropriate for large-scale surveys containing many candidate variables. It may also be preferable to the traditional stepwise method because of its greater statistical stability. Future research should explore Elastic Net regularization, which combines L1 and L2 penalties, incorporate interactions among the main predictors, and conduct a fairness analysis to determine whether the model presents systematic bias according to sex, age, or region.

LITERATURE

Agresti, A. (2015). Foundations of linear and generalized linear models. Wiley.

Instituto Nacional de Estadística y Geografía (2023). Encuesta Nacional sobre Salud Financiera (ENSAFI) 2023: Diseño muestral y metodológico. INEGI.

Ortiz, E. S., Lobos, G., & Guevara, D. (2019). Factores Determinantes del Bienestar Financiero y su Relación con la Calidad de Vida en una Muestra de Profesionales de Guayaquil, Ecuador. *Información tecnológica*, *30*(1). http://www.scielo.cl/scielo.php?script=sci_abstract&pid=S0718076420190001001 21&lng=es&nrm=iso&tlng=es

Tibshirani, R. (1996). Regression shrinkage and selection via the Lasso. Journal of the Royal Statistical Society B, 58(1), 267–288.

Vallejo-Trujillo, L. S., & Martínez-Rangel, M. G. (2016). Perfil de Bienestar Financiero: Aporte Hacia la Mejora de una Comunidad Inteligente. *Investigación administrativa*, *45*(117). http://www.scielo.org.mx/scielo.php?script=sci_abstract&pid=S2448-76782016000100006&lng=es&nrm=iso&tlng=es

CHAPTER 5

Neural Network applied in the National Household Income and Expenditure Survey

Alejandro López-Becerra, Iara Idalia Zapot Juárez, Angel Juan Sánchez García, José Juan Muñoz León

ABSTRACT

The 2024 National Survey of Household Income and Expenditure (ENIGH) provides a statistical overview of the patterns, amounts, and distribution of income and expenditure in Mexico, along with sociodemographic information and data on housing infrastructure (INEGI, 2025). The general objective of this study was to evaluate the technical feasibility and performance of a multiclass artificial neural network for the automated classification of household socioeconomic strata. The project followed the Knowledge Discovery in Databases (KDD) process, characterized by the large-scale exploration of raw microdata without prior feature selection. For this purpose, a pipeline was developed in Python using a column transformer to apply standard scaling (*StandardScaler*) to numerical variables and one-hot encoding (*OneHotEncoder*) to categorical variables. Subsequently, a multilayer perceptron (*MLPClassifier*) with two hidden layers of 64 and 32 neurons and early stopping was trained. Among the main results, the model achieved an overall accuracy of 62.58% on the test set. This performance shows that, although the algorithm automates pattern extraction, the high similarity in consumption dynamics among middle socioeconomic classes imposes a natural predictive limit. These findings support both the potential and the limitations of applying neural networks to sociodemographic databases.

INTRODUCTION

According to the National Center for Education Statistics (2008), socioeconomic status is a composite economic and sociological indicator used to classify individuals or groups according to their relative position within a social structure. In Mexico, the most robust primary source for this purpose is the 2024 National Survey of Household Income and Expenditure (ENIGH), developed by the National Institute of Statistics and Geography (INEGI, 2025). This instrument provides a multidimensional statistical overview of the patterns, sources, and distribution of household financial resources while simultaneously collecting sociodemographic, occupational, and housing infrastructure microdata. In the field of data science, the ENIGH represents a large, high-dimensional information ecosystem suitable for the implementation of advanced machine learning algorithms.

The study of social stratification and its influence on human dynamics has been widely addressed through traditional analytical and econometric approaches. One example is the work conducted by Arán Filippetti (2012), who carried out a quantitative, nonexperimental, and cross-sectional study to evaluate the predictive relationship between socioeconomic status and cognitive processes during the school years. In that study, socioeconomic status served as an independent variable to explain significant differences in performance through multivariate statistical models.In addition, recent studies by Cepeda Villasana et al. (2024) have used the 2024 ENIGH to analyze the interaction of sociodemographic

variables, such as the age, sex, and educational level of the head of household, in determining household per capita income. Their findings demonstrate the informational and analytical potential of these data for understanding socioeconomic dynamics in Mexico.

Previous studies such as these show that classical methodologies generally assume an analytical direction in which socioeconomic status is already predefined and is commonly used to predict or explain other phenomena. At this point, the present study identifies a methodological gap and proposes a reversal of the traditional approach. Rather than treating socioeconomic status as the cause or predictor variable, this project establishes it as the response variable. In this way, machine learning is used to determine whether the raw ENIGH data contain sufficient mathematical signal to infer household socioeconomic classification inversely and automatically. This shift toward a predictive approach supports the feasibility of addressing complex classification problems through advanced supervised learning techniques.

The problem addressed in this study arises from a gap in technical knowledge and seeks to determine whether household data contain enough information for a neural network to distinguish among socioeconomic strata. By establishing socioeconomic stratum as the response variable and using all survey columns as predictors, except those excluded beforehand, the algorithm operates in a high-dimensional and highly noisy setting. The goal is not to design a perfect model but to evaluate the capabilities, scope, and limitations of a “black box” neural network when applied to real sociodemographic data, where the strong similarity in consumption patterns across certain population groups imposes a natural limit on the model’s predictive learning.

The objective of this study was to evaluate the performance and feasibility of a multiclass neural network, specifically a multilayer perceptron, for classifying the socioeconomic stratum of households in Mexico as the response variable. The process was based on the Knowledge Discovery in Databases (KDD) methodology and used indicators from the 2024 ENIGH without prior variable selection.

The following research question was proposed:

- Is it possible to classify the socioeconomic stratum of households in Mexico using sociodemographic and economic-financial variables?

Based on this question, the study was conducted under the hypothesis that all indicators functioned as external manifestations of a complex internal structure, allowing the neural network to identify predictive patterns and approximate the response variable.

METHODOLOGY

The study population consisted of households surveyed in the National Survey of Household Income and Expenditure between August 21 and November 28, 2024, for a total of 105,728 private dwellings. This survey compiles the most relevant information for each participating household. Each record represents one household and includes key variables detailing quarterly income from employment, transfers, remittances, or rent; expenditures on goods and services such as food, health care, education, and

transportation; the demographic characteristics of household members; and housing conditions.

Variable	Description	Measurement Scale	Coding
size_loc	Locality size	Nominal (Ordinal)	Classified into four main categories: 1 = Localities with 100,000 or more inhabitants; 2 = Localities with 15,000 to 99,999 inhabitants; 3 = Localities with 2,500 to 14,999 inhabitants; 4 = Localities with fewer than 2,500 inhabitants
str_socio	Socioeconomic stratum	Nominal (Ordinal)	1 = Low; 2 = Lower-middle; 3 = Upper-middle; 4 = High
type_hou	Household type	Nominal	1 = One-person household; 2 = Nuclear; 3 = Extended; 4 = Composite; 5 = Co-resident
sexo_hh	Sex of the household head	Nominal	1 = Male; 2 = Female
Age_hh	Age of the household head	Ratio	Numerical (Discrete)
educa_hh	Formal education of the household head	Nominal (Ordinal)	01 = No schooling; 02 = Preschool; 03 = Incomplete primary education; 04 = Complete primary education; 05 = Incomplete secondary education; 06 = Complete secondary education; 07 = Incomplete upper-secondary education; 08 = Complete upper-secondary education; 09 = Incomplete university education; 10 = Complete university education; 11 = Postgraduate education
tot_mem	Number of household members	Ratio	Numerical (Discrete)
males	Male household members	Ratio	Numerical (Discrete)
females	Female household members	Ratio	Numerical (Discrete)
Olders	Household members aged 18 years or older	Ratio	Numerical (Discrete)
youngers	Household members aged 17 years or younger	Ratio	Numerical (Discrete)
p12_64	Household members aged 12 to 64 years	Ratio	Numerical (Discrete)
p65mas	Household members aged 65 years or older	Ratio	Numerical (Discrete)
employed	Employed household members	Ratio	Numerical (Discrete)
recip_inc	Income recipients	Ratio	Numerical (Discrete)
recip_employed	Employed income recipients	Ratio	Numerical (Discrete)

inc_curr	Total current income	Ratio	Numerical (Continuous)
incemploy	Employment income	Ratio	Numerical (Continuous)
employ	Income from dependent employment	Ratio	Numerical (Continuous)
salary	Wage and salary income	Ratio	Numerical (Continuous)
over_inc	Overtime income	Ratio	Numerical (Continuous)
commissions	Income from commissions and tips	Ratio	Numerical (Continuous)
y-end_bonus	Year-end bonus income	Ratio	Numerical (Continuous)
employcomp	Income from employment compensation	Ratio	Numerical (Continuous)
otra_rem	Other remuneration	Ratio	Numerical (Continuous)
remu_in-kind	In-kind remuneration	Ratio	Numerical (Continuous)
businesses	Income from independent businesses	Ratio	Numerical (Continuous)
nonagrop	Income from non-agropecuarios businesses	Ratio	Numerical (Continuous)
industrial	Industrial business income	Ratio	Numerical (Continuous)
comercial	Commercial business income	Ratio	Numerical (Continuous)
services	Income from services	Ratio	Numerical (Continuous)
agrop	Income from agropecuarios businesses	Ratio	Numerical (Continuous)
agricul	Agricultural income	Ratio	Numerical (Continuous)
pecuarios	Livestock income	Ratio	Numerical (Continuous)
breedingact	Income from breeding activities	Ratio	Numerical (Continuous)
fishing	Fishing income	Ratio	Numerical (Continuous)
others_jobs	Income from other jobs	Ratio	Numerical (Continuous)
rentalsprop	Income from property rentals	Ratio	Numerical (Continuous)
profits	Income from profits	Ratio	Numerical (Continuous)
rental	Rental income	Ratio	Numerical (Continuous)
transfer	Income from total transfers	Ratio	Numerical (Continuous)
pension	Retirement or pension income	Ratio	Numerical (Continuous)

scholarship	Scholarship income	Ratio	Numerical (Continuous)
donations	Income from donations	Ratio	Numerical (Continuous)
remittance	Remittance income	Ratio	Numerical (Continuous)
bene_gov	Benefits from government programs	Ratio	Numerical (Continuous)
transf_hou	Transfers from other households	Ratio	Numerical (Continuous)
trans_inst	Transfers from institutions	Ratio	Numerical (Continuous)
estim_dwe	Estimated rental value of the dwelling	Ratio	Numerical (Continuous)
others_inc	Other current income	Ratio	Numerical (Continuous)
exp_mon	Total monetary expenditure	Ratio	Numerical (Continuous)
food	Expenditure on food, beverages, and tobacco	Ratio	Numerical (Continuous)
food_in	Food consumed within the household	Ratio	Numerical (Continuous)
cereals	Expenditure on cereals	Ratio	Numerical (Continuous)
meat	Expenditure on meat	Ratio	Numerical (Continuous)
fish	Expenditure on fish and seafood	Ratio	Numerical (Continuous)
milk	Expenditure on milk and dairy products	Ratio	Numerical (Continuous)
eggs	Expenditure on eggs	Ratio	Numerical (Continuous)
oils	Expenditure on oils and fats	Ratio	Numerical (Continuous)
tubers	Expenditure on tubers	Ratio	Numerical (Continuous)
vegetables	Expenditure on vegetables and legumes	Ratio	Numerical (Continuous)
fruit	Expenditure on fruit	Ratio	Numerical (Continuous)
sugar	Expenditure on sugar and honey	Ratio	Numerical (Continuous)
coffee	Expenditure on coffee, tea, and chocolate	Ratio	Numerical (Continuous)
spices	Expenditure on spices and seasonings	Ratio	Numerical (Continuous)
others_food	Expenditure on other food products	Ratio	Numerical (Continuous)
beverages	Expenditure on alcoholic and non-alcoholic beverages	Ratio	Numerical (Continuous)
food_out	Food consumed outside the household	Ratio	Numerical (Continuous)
tabacco	Expenditure on tobacco	Ratio	Numerical (Continuous)
Cloth_foot	Expenditure on	Ratio	Numerical (Continuous)

	clothing and footwear		
clothing	Expenditure on clothing	Ratio	Numerical (Continuous)
footwear	Expenditure on footwear	Ratio	Numerical (Continuous)
housing	Expenditure on housing and services	Ratio	Numerical (Continuous)
housingrent	Expenditure on housing rent	Ratio	Numerical (Continuous)
proptax_main	Expenditure on property tax and maintenance	Ratio	Numerical (Continuous)
water	Expenditure on water services	Ratio	Numerical (Continuous)
electricity	Expenditure on electricity and fuels	Ratio	Numerical (Continuous)
cleaning	Expenditure on cleaning products	Ratio	Numerical (Continuous)
careserv	Expenditure on household care services	Ratio	Numerical (Continuous)
utensils	Expenditure on household utensils	Ratio	Numerical (Continuous)
equipment	Expenditure on furniture and household equipment	Ratio	Numerical (Continuous)
healthcare	Expenditure on healthcare	Ratio	Numerical (Continuous)
outpa_care	Expenditure on outpatient care	Ratio	Numerical (Continuous)
hosp_care	Expenditure on hospital care	Ratio	Numerical (Continuous)
medic_prod	Expenditure on medicines and healthcare products	Ratio	Numerical (Continuous)
transport	Expenditure on transportation and communications	Ratio	Numerical (Continuous)
public	Expenditure on public transportation	Ratio	Numerical (Continuous)
intercity	Expenditure on intercity transportation	Ratio	Numerical (Continuous)
acqui_vehi	Expenditure on vehicle acquisition	Ratio	Numerical (Continuous)
maintenance	Expenditure on vehicle maintenance	Ratio	Numerical (Continuous)
spareparts	Expenditure on automotive spare parts	Ratio	Numerical (Continuous)
fuel	Expenditure on vehicle fuel	Ratio	Numerical (Continuous)
commun	Expenditure on communications	Ratio	Numerical (Continuous)
educa_expen	Expenditure on education and	Ratio	Numerical (Continuous)

	recreation		
educational	Expenditure on educational services	Ratio	Numerical (Continuous)
expen	Expenditure on recreation	Ratio	Numerical (Continuous)
pack_travel	Expenditure on travel packages	Ratio	Numerical (Continuous)
personal	Expenditure on personal care	Ratio	Numerical (Continuous)
personal_care	Expenditure on personal care products	Ratio	Numerical (Continuous)
acces_pers	Expenditure on personal accessories	Ratio	Numerical (Continuous)
others_exp	Other miscellaneous expenditures	Ratio	Numerical (Continuous)
transf_prov	Expenditure on transfers provided	Ratio	Numerical (Continuous)
finan_tot	Total financial inflows	Ratio	Numerical (Continuous)
withdra_inv	Withdrawals from investments	Ratio	Numerical (Continuous)
loans	Loans received	Ratio	Numerical (Continuous)
others_finan	Other financial inflows	Ratio	Numerical (Continuous)
finanout_tot	Total financial outflows	Ratio	Numerical (Continuous)
pay_hous	Housing fee payments	Ratio	Numerical (Continuous)
mater_serv	Expenditure on materials and services	Ratio	Numerical (Continuous)
deposits	Deposits into savings and investments	Ratio	Numerical (Continuous)
loans_third	Loans granted to third parties	Ratio	Numerical (Continuous)
pay_card	Credit card payments	Ratio	Numerical (Continuous)
debt	Debt payments	Ratio	Numerical (Continuous)
balance	Capital balance	Ratio	Numerical (Continuous)
others_outflows	Other financial outflows	Ratio	Numerical (Continuous)

The descriptive analysis examined three aspects: class distributions, logarithmic transformations, and contingency and correlation analyses. Class distribution is particularly important because the model needs to learn from all classes in a balanced manner. Otherwise, it may reach inaccurate conclusions by creating nonexistent relationships, thereby affecting its generalizability. For this reason, the frequencies of the four socioeconomic strata were analyzed to identify possible imbalances in the dependent variable.

In addition, macro-level financial variables, such as current income and total monetary expenditure, tend to show positive skewness. Therefore, temporary logarithmic transformations were applied to improve visualization and more clearly identify the boundaries separating the socioeconomic strata.

The contingency and correlation analyses involved creating cross-tabulations to assess the percentage influence of categorical variables on socioeconomic stratum. In addition, a linear correlation matrix was generated for the continuous variables to rule out the inclusion of features with high structural multicollinearity.

The study adopted a quantitative, nonexperimental approach with a predictive scope. Its objective was to develop a model capable of distinguishing among four socioeconomic strata: low, lower-middle, upper-middle, and high, using a combination of quantitative and qualitative variables. Several predictive models were tested, including decision trees, logistic regression, and discriminant analysis. However, the best results were obtained with an artificial neural network architecture.

A neural network has a large number of adjustable hyperparameters, which can produce many different results. Identifying the best combination to achieve optimal performance is one of the most difficult and computationally demanding challenges. In this study, the neural network was evaluated using five-fold cross-validation. This means that each configuration was trained and assessed on five different subsets of the data in rotation. In addition, the model was trained after standardizing the numerical variables and applying one-hot encoding to the categorical variables.

For neural network hyperparameter optimization, a search space covering four main dimensions was defined using a random search method (*RandomizedSearchCV*). For the hidden-layer architecture (*hidden_layer_sizes*), configurations ranged from simple single-layer topologies with 32 or 64 neurons to more complex two- and three-layer architectures, including (64, 32), (64, 32, 16), (128, 64), and (128, 64, 32). These configurations were evaluated to examine the effect of network depth and width on representational capacity. For the initial learning rate (*learning_rate_init*), values of 0.01, 0.005, 0.001, and 0.0001 were tested to calibrate the convergence speed of the gradient descent algorithm and avoid both excessive oscillation and overly slow convergence.

For batch size (*batch_size*), groups of 32, 64, and 128 samples per iteration were considered. This made it possible to examine the trade-off between training stability and the computational cost associated with each weight update. Finally, to reduce overfitting through the L2 regularization term (*alpha*), penalties of 0.01, 0.001, 0.0001, and 0.00001

(1e-05) were applied to the network weights. The purpose was to identify the level of regularization that provided the best balance between fit to the training data and the model's generalization ability.

RESULTS

The analysis was conducted using a cleaned sample of 91,414 households. The distribution of the socioeconomic stratum variable (*EstSo*) showed a substantial class imbalance. Stratum 2 contained the largest share of the sample, accounting for 50.7% of all households, whereas stratum 4, representing the highest socioeconomic status, accounted for only 8.9%.

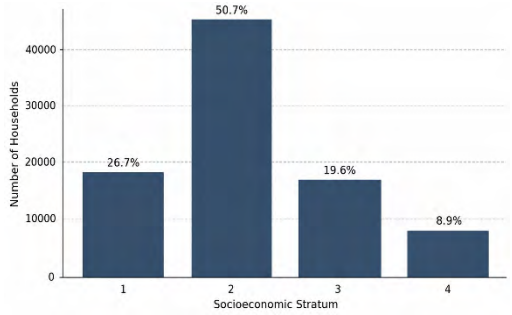

Figure 1. Distribution of the variable Socioeconomic stratum

Among the predictor variables, income and expenditure increased consistently across the *EstSo* categories. However, food expenditure remained relatively stable across the four strata. Sociodemographic variables, such as the educational level of the head of household and the degree of urbanization, showed a direct relationship with *EstSo*. Postgraduate education and residence in large cities were associated with the highest probability of belonging to the upper socioeconomic stratum.

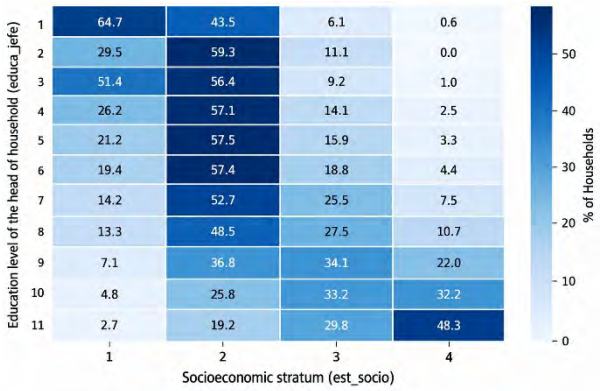

Figure 2. Percentage of socioeconomic stratum given educational level

Finally, the correlation analysis revealed structural multicollinearity between total monetary expenditure and its subcomponents. This finding suggests that household socioeconomic hierarchy is fundamentally associated with educational and geographic characteristics, while financial variables operate more as direct consequences of these factors than as independent causes.

The optimization process showed that socioeconomic classification did not require a highly complex deep learning architecture. The intermediate network with two hidden layers of 64 and 32 neurons consistently outperformed larger topologies, such as the architecture with 128, 64, and 32 neurons, during cross-validation. The model achieved an overall accuracy of 62.6% on the test set. Stratum 2, the majority class, obtained the highest recall, correctly identifying more than 80% of the actual cases. Stratum 1 showed high precision, above 70%, suggesting that the model was effective at identifying lower socioeconomic classes, although it had greater difficulty distinguishing between them. Stratum 3 represented the greatest challenge, with precision of 48%, recall of 39%, and an F1 score of 0.43. This decline in performance indicates that the upper-middle socioeconomic class has highly diffuse economic boundaries that make classification more difficult. This difficulty may also have been intensified by the substantial class imbalance, as stratum 3 was positioned between classes with markedly different frequencies.

Despite being the minority class, representing less than 9% of the total sample, the model achieved a notably balanced performance when classifying stratum 4, with 58% precision and 60% recall (F1 score = 0.59). This result suggests that higher-income households have sufficiently distinctive and atypical predictive characteristics for the neural network to identify them correctly, despite the challenge posed by class imbalance.

CONCLUSIONS

The results of this project show that the boundaries separating socioeconomic strata in Mexico are highly diffuse. In practical terms, it is difficult to distinguish one class from another unless the gap extends beyond a single stratum. For example, stratum 1 could be distinguished relatively easily from stratum 3, whereas the differences between strata 2 and 3 were more difficult to identify. The neural network and its performance metrics reflect this social dynamic: the model was able to distinguish broad socioeconomic groups but showed greater confusion between neighboring strata. These findings indicate that social classes cannot be defined precisely using sociodemographic or economic-financial variables in isolation. Instead, a close and multidimensional analysis is necessary to identify the boundaries between them, even when those boundaries remain unclear.

This study makes it clear that valuable information can continue to be extracted from this dataset, either through a detailed statistical comparison of each variable across the four social groups or by testing new training methods with predictive models. From a social science perspective, the model's limited performance in predicting stratum 3 is particularly relevant because it opens the possibility of investigating the structural reasons why this social class is so unstable and difficult to define. The model also showed that social class is strongly related to locality size, with rural areas showing very limited representation of strata 3 and 4. This pattern reflects a serious inequality problem and supports the need to examine rural exclusion from qualitative and spatial perspectives: To what extent does the lack of infrastructure, such as hospitals, higher education institutions, and internet access, prevent wealth accumulation even when individuals work hard?

Finally, the project demonstrates that supervised learning models are useful for understanding certain phenomena, identifying patterns, and clarifying social problems that are not immediately visible. Making these barriers visible through data is an essential step toward developing and implementing effective solutions and, ultimately, improving social mobility.

LITERATURE

Arán Filippetti, V. (2012). Estrato Socioeconómico y Habilidades Cognitivas en Niños Escolarizados: Variables Predictoras y Mediadoras. *Interdisciplinaria*, *29*(1), 129-149. http://www.scielo.org.ar/scielo.php?script=sci_arttext&pid=S1668-70272012000100007

Berdichevsky, G., Kelty, K., Straubel, J. B., & Toomre, E. (2006). The tesla roadster battery system. Tesla Motors, 1(5), 1-5.

National Center for Education Statistics. (2008). *S - Glossary*. Condition of Education. https://web.archive.org/web/20080412230538/http://nces.ed.gov/programs/coe/glossary/s.asp

Tesla, N. (2007). My inventions: the autobiography of Nikola Tesla. A Distant Mirror.

elcic, I., & Stryker, C. (s.f.). *¿Qué es el ajuste de hiperparámetros?* IBM. Recuperado el 14 de junio de 2026, de https://www.ibm.com/mx-es/think/topics/hyperparameter-tuning

CHAPTER 6

Prediction of total income of economic units in Mexico using Supervised Learning models

Jorge Luis Ortiz Libreros, Joshua Iván Olivares Blanco, José Vidal Hernández, Angel Fernando Argüello Ortiz

ABSTRACT

The analysis of income among economic units is a fundamental component for understanding the behavior of Mexico's productive sector, particularly given the growing availability of structured information from the 2019 National Economic Census conducted by the National Institute of Statistics and Geography (INEGI). The objective of this study was to identify the supervised learning model with the greatest predictive ability for estimating the total income of economic units registered in the National Economic Census. Using a quantitative predictive design, the analysis was conducted on a cleaned dataset of 4,808 economic units obtained from an original database of 7,370 records. Three feature selection techniques were applied: the ANOVA *F* test, mutual information, and Pearson correlation. These methods identified 11 predictor variables related to employed personnel, productive inputs, and economic performance indicators. Four supervised learning algorithms were evaluated: Ridge regression, a regression tree, support vector regression (SVR), and a neural network. Hyperparameters were optimized through grid search with cross-validation. The results showed that the neural network achieved the best predictive performance, with a coefficient of determination of R^2 = 0.7410, a mean absolute error (MAE) of 903.20, and a root mean square error (RMSE) of 2,957.15 thousand pesos. These findings demonstrate its greater ability to capture the complex nonlinear relationships present in nationwide economic census data.

INTRODUCTION

General Context of the Topic

The National Institute of Statistics and Geography (INEGI) produces databases that make it possible to examine the structure and behavior of Mexico's productive sector in detail. In this context, the National Economic Census is one of the most comprehensive statistical instruments for the systematic collection of information on formal economic units in Mexico. It includes variables related to employment, production, inputs, value added, total income, and other areas relevant to the analysis of the business sector.

The use of data analysis techniques, particularly machine learning, has provided researchers and decision-makers with tools capable of processing large volumes of information and extracting predictive patterns that conventional statistical methods may not always capture. In the field of applied economics, combining census data with supervised learning models makes it possible to analyze and estimate variables of economic interest.

BACKGROUND

The application of machine learning methods to economic analysis has been driven by both increased computational capacity and the growing availability of large economic and administrative datasets. In a field where variables may show nonlinear relationships, interactions, and high-dimensional structures, supervised learning algorithms have demonstrated strong predictive performance and, in some cases, have outperformed more traditional econometric models, particularly in regression tasks involving skewed distributions and extreme outliers.

Among the techniques that have received the most attention in economic regression problems are decision trees and their ensemble variants, support vector regression (SVR), and artificial neural networks. Multilayer perceptron models have shown advantages in capturing complex patterns in economic tabular data because they can approximate nonlinear functions through hidden layers and nonlinear activation functions, such as the rectified linear unit. This property is especially relevant when working with economic variables whose relationship with the target variable cannot be adequately described by a linear or low-degree polynomial function.

Regression models such as Ridge regression have been widely used in econometrics to address multicollinearity and overfitting when the number of predictors is large. Unlike ordinary linear regression, Ridge regression applies a penalty to the magnitude of the coefficients, reducing estimator variance at the cost of introducing controlled bias. However, this model retains the assumption of a linear relationship between the predictors and the target variable, which limits its ability to model nonlinear phenomena such as business income in heterogeneous populations.

Selecting predictor variables before model training is recognized as a critical step in any predictive modeling process. Combining selection techniques such as the ANOVA *F* test, mutual information, and Pearson correlation makes it possible to identify subsets of variables that are relevant from different statistical perspectives. The first evaluates linearly explained variability, the second detects both linear and nonlinear dependencies, and the third quantifies the strength and direction of linear associations. The combined results of these three methods support the final selection of predictors and reduce the risk of including irrelevant variables that could compromise model generalization.

PROBLEM STATEMENT

Despite the information provided by INEGI's National Economic Census, the estimation and prediction of total income among economic units through machine learning models remains an insufficiently explored area. Traditional statistical methods, such as ordinary linear regression, assume linear relationships among variables and face limitations when dealing with skewed distributions, extreme outliers, and complex covariance structures, all of which are inherent characteristics of business income data at the census level.
The total income variable (*1000A*) from INEGI's National Economic Census shows a right-skewed distribution, with a mean of 1,666.34 thousand pesos and a median of 117.50 thousand pesos. This reflects the coexistence of microenterprises with minimal income and large-scale economic units with exceptionally high income within the same dataset.

This pattern not only poses a challenge for model fitting but also makes it necessary to use techniques capable of capturing the variability of the phenomenon without assuming rigid structures. Classical linear models are limited for this purpose, which justifies exploring machine learning algorithms with greater expressive capacity.

Predictor selection becomes a challenge when working with high-dimensional datasets, as in the case of the National Economic Census, whose original version contains 30 variables and 7,370 records.

An unsupported selection of predictors may result in unstable, overfitted models with limited generalization ability. Therefore, it is necessary to apply feature selection methods that identify the subset of variables with the greatest predictive relevance and ensure consistency in the modeling process across different criteria.

Research Objective

The objective of this study is to identify the supervised learning model with the best predictive performance for estimating the total income (*I000A*) of registered economic units. Four algorithms were compared: Ridge regression, a regression tree, support vector regression (SVR), and a neural network. The models were trained using a previously selected set of predictor variables to determine which algorithm achieved the best results based on the evaluation metrics considered.

Hypothesis

Based on the nonlinear distribution of income among economic units and the recognized ability of neural network models to approximate more complex functions, the following hypothesis was proposed: nonlinear machine learning models, particularly the neural network, will achieve superior predictive performance, measured by the coefficient of determination (R^2), compared with linear and semilinear models applied to the same National Economic Census dataset. This expectation is based on the assumption that total income has relationships with the predictor variables that exceed the modeling capacity of methods based on strict linearity assumptions.

METHODOLOGY

Research Design

This study followed a quantitative predictive design based on data from INEGI's 2019 National Economic Census. Supervised machine learning techniques were applied to predict the total income (*I000A*) of economic units using a set of variables selected through exploratory analysis.

Population

The original dataset contained 7,370 records and 30 variables corresponding to economic units in Mexico. After removing duplicate records and observations with missing values in the response variable, the final sample consisted of 4,808 valid economic units for analysis.

Study Variables

The response variable was *I000A* (Total Income), a continuous quantitative variable representing the total income of each economic unit. The predictor variables were selected through a preliminary exploratory analysis using three feature selection techniques: the ANOVA *F* test, mutual information, and Pearson correlation. The 11 variables selected based on their consistency across the three methods are presented in Table 1.

Variable	Description	Measurement Scale	Coding
J000A	Value added (wealth generated)	Ratio	Wealth generated by the economic unit after deducting intermediate consumption
H001A	Total employed personnel	Ratio	Total number of people who worked in the economic unit during the period
H001D	Category D personnel	Ratio	Individuals classified under category D within the workforce
H010C	Subtype of dependent personnel	Ratio	Dependent personnel classified under subtype C within the economic unit
A121A	Consumption of type 2 inputs	Ratio	Amount spent on type 2 inputs during the period
A131A	Consumption of type 3 inputs	Ratio	Amount spent on type 3 inputs during the period
Q010A	Productivity indicator	Ratio	Relationship between the value generated and the human resources employed
Q020A	Efficiency indicator	Ratio	Efficiency with which the economic unit transforms inputs into value added
Q030A	Operational indicator	Ratio	Operational performance of the economic unit in its main activity
Q400A	Aggregate indicator	Ratio	Consolidated performance of the economic unit expressed as a reference value
Q900A	Overall indicator	Ratio	Overall performance of the economic unit considering all its activities

Table 1. Description of selected predictor variables

Exploratory Data Analysis

An exploratory data analysis was conducted to identify the variables with the greatest explanatory ability for total income (*I000A*). The structure of the dataset was first reviewed by verifying the presence of missing values, duplicate records, and the nature of the variables.

The first technique applied was the analysis of variance (ANOVA *F* test), which measures a variable's ability to explain variability in the target variable by comparing explained and

residual variation. The second technique was mutual information, which makes it possible to detect both linear and nonlinear relationships between the predictor variables and the target variable. Finally, Pearson correlation was calculated to evaluate the strength and direction of the linear relationship between each explanatory variable and total income.

Data Collection and Preparation

The data were obtained from INEGI's 2019 National Economic Census. After removing duplicate records and observations without a value for the response variable, missing values in the predictor variables were imputed using the median, a robust strategy in the presence of outliers. The dataset was divided into a training set (80%, n = 3,846) and a test set (20%, n = 961).

Data Analysis

Four supervised learning models were evaluated for the regression problem: Ridge regression, a regression tree (*DecisionTreeRegressor*), support vector regression (SVR), and an artificial neural network (*MLPRegressor*). The hyperparameters of each model were optimized through grid search (*GridSearchCV*) with five-fold cross-validation (*KFold*, shuffle=True, random_state=42), using the coefficient of determination (R^2) as the optimization metric. Models requiring feature scaling—Ridge regression, SVR, and the neural network—were implemented within a pipeline that applied standardization through *StandardScaler* before model fitting.

The evaluation metrics used to compare model performance on the test set were mean absolute error (MAE), root mean square error (RMSE), and the coefficient of determination (R^2).

RESULTS

Data Description

The analyzed sample consisted of 4,808 economic units from the National Economic Census. The response variable, total income (*I000A*), showed a strongly right-skewed distribution, with a mean of 1,666.34 thousand pesos and a median of 117.50 thousand pesos, reflecting substantial heterogeneity in the size of the economic units. The descriptive statistics are presented in Table 2.

Variable	Mean	Median	Minimum	Mximum	N
I000A (Total Income)	1,666.34	117.50	0.00	814,927.00	4,808

Table 2. Descriptive statistics of the response variable

Variable Selection

The three methods used to select the most relevant predictor variables produced consistent results. The selected variables included measures related to employed personnel (*H001A*, *H001D*, *H010C*), economic indicators (*Q010A*, *Q020A*, *Q030A*, *Q400A*, *Q900A*), and variables associated with wealth generation and productive inputs (*J000A*, *A121A*, *A131A*). Based on the agreement across the three criteria, a subset of 11 variables was selected to preserve different economic dimensions while reducing redundancy.

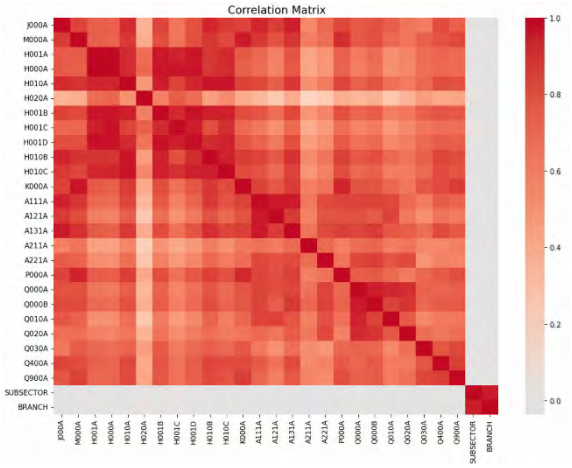

Figure 1. *Heat map of the correlation matrix of the analyzed economic variables.*

Model Comparison

All four evaluated models showed competitive performance, with R^2 values above 0.65 in every case. Table 3 summarizes the metrics obtained on the test set, along with the best hyperparameters identified through grid search.

Model	MAE	RMSE	R^2	Best configuration
Ridge Regression	1,004.76	3,172.90	0.7018	alpha = 0.1
Regression Tree	956.43	3,414.05	0.6548	max_depth=10, min_samples_leaf=2
SVR	880.09	3,165.69	0.7032	C=100, gamma=scale, kernel=linear
Neural Network (MLP)*	**903.20**	**2,957.15**	**0.7410**	**(100,), relu, alpha=0.001**

Table 3. Comparison of performance metrics by model

Selected Model

The artificial neural network (MLP) achieved the best performance, with an R^2 of 0.7410, an MAE of 903.20, and an RMSE of 2,957.15. It also produced the lowest root mean square error among all the evaluated models. The architecture selected through grid search consisted of one hidden layer with 100 neurons, the rectified linear unit (ReLU) activation function, and an L2 regularization term (*alpha*) of 0.001. The model was trained for a maximum of 5,000 iterations using a fixed random seed (random_state=42) to ensure the reproducibility of the results.

The grid search evaluated the following hyperparameter combinations: hidden-layer sizes of (20,), (50,), (100,), and (50, 25); ReLU and hyperbolic tangent (*tanh*) activation functions; and *alpha* regularization values of 0.0001, 0.001, and 0.01. The optimal configuration was hidden_layer_sizes=(100,), activation="relu", and alpha=0.001. The

single-hidden-layer architecture with the ReLU activation function was effective in capturing the nonlinear relationships present in the economic data, outperforming the other algorithms in terms of both R^2 and RMSE.

CONCLUSIONS

The objective of this study was to identify the supervised learning model with the greatest predictive ability for estimating the total income of economic units included in INEGI's National Economic Census. Through a structured process of variable selection, data preparation, and algorithm comparison using cross-validation, the artificial neural network (MLP) with one hidden layer of 100 neurons, a ReLU activation function, and L2 regularization (*alpha* = 0.001) achieved the best performance. The model explained 74.10% of the variability in total income (R^2 = 0.7410) and produced the lowest root mean square error (RMSE = 2,957.15).

Among the most relevant findings, variables related to employed personnel, productive inputs, and economic performance indicators provided the greatest amount of information for predicting total income. This suggests that company size and workforce structure, together with productive activity, are fundamental determinants of income levels, a finding consistent with the economic theory of production.

The main contribution of this study lies in the systematic application of multiple feature selection techniques and the rigorous comparison of four supervised learning algorithms using real data from the Mexican economic sector. The findings demonstrate that neural network models can capture complex nonlinear relationships that linear models, such as Ridge regression, cannot model with the same level of accuracy.

From a practical perspective, the results may be useful for economic planning and policy institutions seeking to estimate the performance of economic units based on observable characteristics. This could support the targeting of assistance programs, the identification of businesses at risk, and the forecasting of tax revenue. In addition, the developed model could be integrated into economic monitoring systems to update estimates as new census data become available.Future research should explore ensemble algorithms such as random forests, gradient boosting, and XGBoost, which often provide additional improvements in regression problems involving economic variables. It is also recommended to extend the analysis to the classification of economic sectors (*SECTOR*), apply dimensionality reduction techniques, and evaluate model stability across different data partitions. Finally, examining the relative importance of the variables through interpretability methods such as SHAP or permutation importance would help provide a deeper understanding of the underlying economic mechanisms.

LITERATURE

Google Colaboratory. (2026). *Google Colaboratory*. Google. https://colab.research.google.com/

Guyon, I., & Elisseeff, A. (2003). An introduction to variable and feature selection. *Journal of Machine Learning Research, 3*, 1157–1182.

Harris, C. R., Millman, K. J., van der Walt, S. J., Gommers, R., Virtanen, P., Cournapeau, D., Wieser, E., Taylor, J., Berg, S., Smith, N. J., Kern, R., Picus, M., Hoyer, S., van Kerkwijk, M. H., Brett, M., Haldane, A., del Río, J. F., Wiebe, M., Peterson, P., ... Oliphant, T. E. (2020). Array programming with NumPy. *Nature, 585*(7825), 357–362. https://doi.org/10.1038/s41586-020-2649-2

Hunter, J. D. (2007). Matplotlib: A 2D graphics environment. *Computing in Science & Engineering, 9*(3), 90–95. https://doi.org/10.1109/MCSE.2007.55

Instituto Nacional de Estadística y Geografía. (2020). *Censos Económicos 2019: Resultados definitivos*. INEGI. https://www.inegi.org.mx/programas/ce/2019/

McKinney, W. (2010). Data structures for statistical computing in Python. In S. van der Walt & J. Millman (Eds.), *Proceedings of the 9th Python in Science Conference* (pp. 56–61). https://doi.org/10.25080/Majora-92bf1922-00a

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., & Duchesnay, É. (2011). Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research, 12*, 2825–2830.

Rumelhart, D. E., Hinton, G. E., & Williams, R. J. (1986). Learning representations by back-propagating errors. *Nature, 323*(6088), 533–536. https://doi.org/10.1038/323533a0

Van Rossum, G., & Drake, F. L. (2009). *Python 3 reference manual*. CreateSpace. https://www.python.org/

Waskom, M. L. (2021). Seaborn: Statistical data visualization. *Journal of Open Source Software, 6*(60), 3021. https://doi.org/10.21105/joss.03021

CHAPTER 7

Benefits and characteristics of employment as factors for formal employment in Mexico

Cristian Eduardo Montesino Fernández, Diego Iván González Colorado, Daniela Burgueño Theurel, Judith Guadalupe Montero Mora

ABSTRACT

Labor formality is a fundamental component of workers' well-being because it is associated with access to benefits, job stability, and social protection mechanisms. In Mexico, the persistence of high levels of labor informality has generated interest in identifying the factors that distinguish workers with and without an employment contract. The objective of this study was to classify workers according to the presence or absence of an employment contract through supervised learning techniques and to identify the labor and business characteristics with the greatest influence on this classification. Data from the 2024 National Survey of Household Income and Expenditure (ENIGH) were used, resulting in a final sample of 114,025 observations. A decision tree model was implemented and achieved an accuracy of 87.55%. The feature importance analysis showed that receiving an annual bonus was the main factor associated with having an employment contract, followed by company classification, life insurance, hours worked, and company size.

Keywords: labor formality, decision tree, supervised learning, classification, employment contract.

INTRODUCTION

Labor formality is one of the main topics of discussion in Mexico's labor market because it is directly related to the country's economic stability and social well-being. Workers with formal employment generally have access to multiple benefits established by law that contribute to improving their living conditions. In contrast, labor informality continues to represent a challenge for the country's economic and social development because, according to data from the National Institute of Statistics and Geography (INEGI, 2024), a substantial proportion of the employed or economically active population works under informal conditions.

In recent years, several studies have examined the factors associated with labor formality and informality using information from national surveys and administrative records. International organizations such as the International Labour Organization (ILO) have indicated that variables related to company size, economic sector, employment conditions, and access to labor benefits play a fundamental role in determining employment formality (ILO, 2023).

At the same time, the development of machine learning techniques has made it possible to address classification and prediction problems involving socioeconomic phenomena. Among these techniques, decision trees have been widely used in fields such as economics, health, education, and the labor market because of their ability to identify complex patterns and generate easily interpretable decision rules (James et al., 2021). In this regard, decision trees also make it possible to determine which variables have the

greatest ability to explain or predict a phenomenon of interest, facilitating the interpretation of results by researchers and decision-makers.

However, despite the availability of statistical information on labor conditions in Mexico, it remains important to identify which employment characteristics are most strongly associated with having an employment contract. In particular, it is relevant to examine the role of labor benefits, hours worked, and company characteristics in explaining employment formality.

Therefore, the objective of this study is to classify workers according to the presence or absence of an employment contract by developing a decision tree model using data from the National Survey of Household Income and Expenditure (ENIGH). The study also seeks to identify the labor and company characteristics with the greatest influence on worker classification in order to better understand the factors associated with labor formality in Mexico.

METHODOLOGY

This study followed a quantitative approach with descriptive and predictive purposes. Supervised learning techniques were applied to develop a classification model capable of distinguishing between workers with and without an employment contract based on characteristics related to their employment conditions and the companies where they work.

The data were obtained from the National Survey of Household Income and Expenditure (ENIGH), conducted by the National Institute of Statistics and Geography (INEGI). This survey is one of the main sources of statistical information in Mexico because it collects sociodemographic, economic, and labor-related information on households and their members.

The original dataset consisted of 164,325 records corresponding to surveyed workers. However, as part of the data-cleaning process described below, only records with missing values in the target variable related to the existence of an employment contract were removed. This exclusion was justified because such records could correspond to an employer, a self-employed individual, or a business owner rather than an employee. After this stage, the final analytical sample consisted of 114,025 observations.The response variable in this study was the existence of an employment contract, classified into two categories: workers with an employment contract and workers without one. All other variables were treated as predictors. Labor benefit variables included annual bonus and life insurance. Employment and company characteristics included whether the respondent supervised personnel, remuneration status, hours worked, whether the respondent had business partners, and the type, size, and economic activity of the company.

Database Variable	Variable	Description	Measurement Scale	Coding
contract	Employment contract	Indicates whether the worker has an employment contract	Dichotomous nominal	1 = Yes; 2 = No

personnel	Personnel	Indicates whether the worker supervises or employs personnel	Dichotomous nominal	1 = Yes; 2 = No
pay	Payment arrangement	Describes the worker's remuneration status	Nominal	1 = Receives payment; 2 = Unpaid work in a household business; 3 = Unpaid work in another person's business
pres_2	Year-end bonus	Indicates whether the worker receives a year-end bonus as an employment benefit	Dichotomous nominal	02 = Receives a year-end bonus; Blank = Does not receive a year-end bonus
pres_9	Life insurance	Indicates whether the worker receives life insurance as an employment benefit	Dichotomous nominal	09 = Receives life insurance; Blank = Does not receive life insurance
hwork	Hours worked	Number of hours worked during the reference period	Ratio	Values ranging from 1 to 999 hours
class_comp	Company classification	Type of organization in which the worker is employed	Nominal	1 = Independent, personal, or family business; 2 = Private company; 3 = Government institution; 4 = Non-governmental institution
size_comp	Company size	Number of people employed by the company	Ordinal	1 = 1 person; 2 = 2–5 people; 3 = 6–10 people; 4 = 11–15 people; 5 = 16–20 people; 6 = 21–30 people; 7 = 31–50 people; 8 = 51–100 people; 9 = 101–250 people; 10 = 251–500 people; 11 = 501 or more people; 12 = Does not know
econoact	Economic activity	Main economic activity of the company or business	Nominal	1 = Industrial; 2 = Commercial; 3 = Services; 4 = Agricultural activities; 5 = Animal breeding and husbandry activities; 6 = Gathering activities; 7 = Reforestation and tree logging; 8 = Hunting and animal capture activities; 9 = Fishing activities
partners	Business partners	Indicates whether there are business partners involved in the business or economic activity	Dichotomous nominal	1 = Yes; 2 = No

Table 1. Description of variables

Once the variables had been defined, a descriptive analysis was conducted during preprocessing to examine their general characteristics and verify the quality of the information used. The data were then cleaned by removing records with missing values in the target variable. In addition, categorical variables were transformed using one-hot encoding so that they could be incorporated into the classification model.

For the predictive analysis, a decision tree was implemented to classify workers according to the presence or absence of an employment contract. The model used the Gini impurity

criterion to identify the optimal data splits. A pruning procedure was also applied through the complexity parameter (*ccp_alpha*) to obtain a simpler and more interpretable model, which was then evaluated to support the corresponding conclusions.

RESULTS

Description of the Sample and Data Analyzed

The original database of workers surveyed in the National Survey of Household Income and Expenditure (ENIGH) consisted of 164,325 records. After the data-cleaning process, in which records with missing values in the response variable indicating the presence or absence of an employment contract were removed, the final sample consisted of 114,025 records. The sample was then divided into two subsets: the training set included 79,817 observations, whereas the test set contained 34,208.

Descriptive Results

Statistic	Hours Worked
Observations	164,325
Mean	40.44
Standard Deviation	18.70
Minimum	1
Median	43
Maximum	168

***Table 2**. Descriptive statistics of hours worked*

The descriptive analysis identified relevant characteristics related to the employment and company conditions of the surveyed workers. The mean number of hours worked per week was 40.44, with a median of 43 hours, indicating that most workers had schedules close to a standard workweek. The minimum recorded value was 1 hour and the maximum was 168 hours, reflecting substantial variability in working hours. The distribution suggests that a large proportion of employed individuals worked around 40 hours per week. However, the presence of extreme values may indicate unusually long work schedules, data-entry errors, or misunderstandings when responding to the survey.

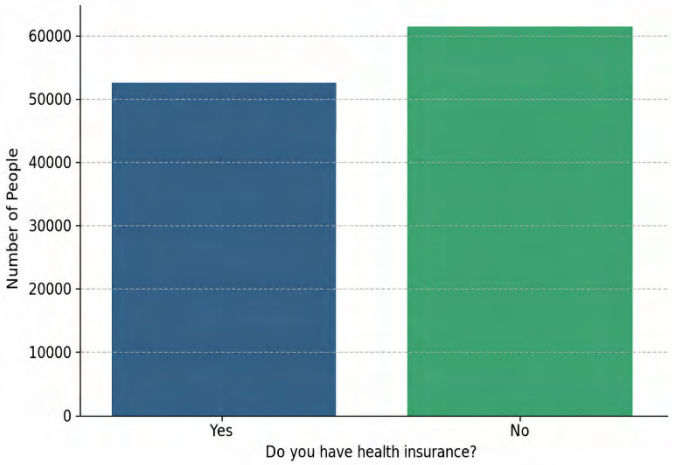

***Figure 1.** Distribution of the Contract variable*

The frequency chart showed a relatively balanced distribution between the two categories, which is favorable for training classification models because it reduces problems associated with class imbalance.
Similarly, substantial variability was observed in company size, ranging from establishments with only a few workers to large organizations with hundreds of employees. This characteristic is particularly relevant because company structure and capacity are often associated with the availability of labor benefits, hiring mechanisms, and levels of formality. Therefore, analyzing these variables helps provide a better understanding of the factors that influence the presence or absence of employment contracts among workers in Mexico.

Therefore, the descriptive results demonstrate the diversity of Mexico's productive sector, where small family businesses and independent enterprises coexist with large private companies and public institutions. This reality reflects the complexity of the Mexican labor market and the presence of different forms of work organization, each with particular characteristics regarding hiring practices, access to benefits, and employment conditions.

Model Results

Therefore, the descriptive results demonstrate the diversity of Mexico's productive sector, where small family businesses and independent enterprises coexist with large private companies and public institutions. This reality reflects the complexity of the Mexican labor market and the presence of different forms of work organization, each with particular characteristics regarding hiring practices, access to benefits, and employment conditions.

Class	Accuracy	Recall	F1-score
With contract (1)	0.86	0.88	0.87
No contract (2)	0.89	0.88	0.89
Global average	0.88	0.88	0.88

Table 3. *Decision tree classification metrics*

The results show that the model had adequate predictive ability, with balanced performance across both categories and all metric values close to 0.88. A feature importance analysis was also conducted to identify the factors with the greatest influence on worker classification. The annual bonus benefit (*pres_2*) was the most relevant variable, accounting for 73.61% of the model's total importance. It was followed by company classification (*clas_emp*) and the life insurance benefit (*pres_9*).

Variable	Importancia
Year-end bonus (pres_2)	0.7361
Government institution classification (class_comp_3)	0.0756
Private company classification (class_comp_2)	0.0655
Life insurance (pres_9)	0.0612
Hours worked (hwork)	0.0224
Company size (size_comp_2)	0.0151

Table 4. *Most important variables of the model*

The findings suggest that labor benefits are the most important factors for distinguishing between formal and informal employment. In particular, receiving an annual bonus was the main indicator associated with the existence of an employment contract.

To reduce model complexity and prevent overfitting, a pruning procedure was applied through cross-validation using the complexity parameter (*ccp_alpha*). The optimal value obtained was 0.000526. As a result, the pruned tree had a new maximum depth of 7 levels and 16 terminal nodes.

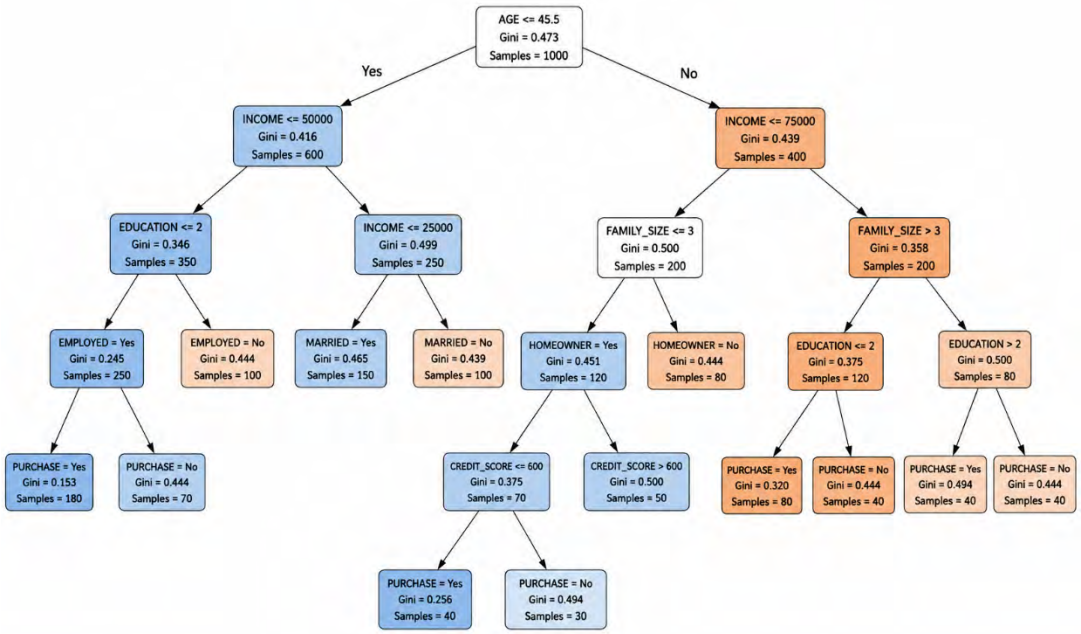

Figure 2. Visualization of the final tree (Pruned)

According to Figure 2, the model's first split was based on whether the worker received an annual bonus, indicating that this variable had the greatest ability to distinguish between workers with and without an employment contract. Variables related to company classification, access to life insurance, and organization size then further refined the classification.

Overall, the model showed that workers who received labor benefits, particularly an annual bonus and life insurance, as well as those employed by larger companies or organizations with more formal structures, were more likely to have an employment contract. In contrast, the absence of these characteristics was associated with a greater likelihood of belonging to the group of workers without a contract. Finally, the pruned model achieved an accuracy of 87.55%, which was very close to that of the original tree. This indicates that the tree structure could be simplified without losing predictive ability.

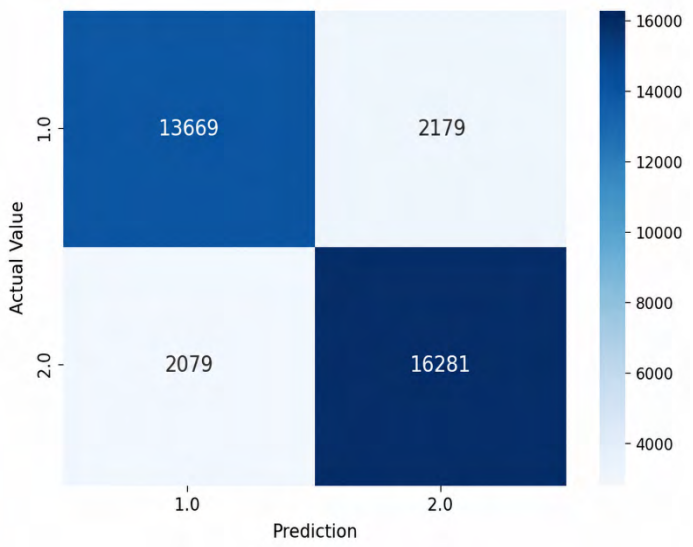

***Figure 3**. Confusion matrix of the tree.*

The confusion matrix showed balanced model performance across both categories. A total of 13,669 workers in class 1 (with a contract) and 16,281 workers in class 2 (without a contract) were classified correctly, whereas the numbers of misclassified observations were relatively low at 2,179 and 2,079, respectively. These results indicate that the model had an adequate ability to distinguish between workers with and without an employment contract, which is consistent with the overall accuracy of approximately 88%.

Algorithm Comparison

In addition, a multilayer perceptron (MLP) neural network was implemented to compare its performance with that of the decision tree.

Model	Accuracy
Decision Tree	0.8776
Pruned Decision Tree	0.8755
Neural Network (MLP)	0.8804

Table 5. Comparison of models

The MLP achieved the highest accuracy, although the difference compared with the decision tree was minimal. Therefore, the decision tree retains an advantage as a more suitable alternative for the objectives of the study because, in addition to providing similar predictive performance, it allowed the variables associated with employment formality to be identified and interpreted directly.

CONCLUSIONS

The results show that employment benefits are the main factors associated with employment formality. In particular, the year-end bonus was the most important variable in the model, followed by company classification and the availability of life insurance. These findings suggest that access to employment benefits is one of the most relevant elements for distinguishing workers who have a formal employment relationship from those who face more vulnerable working conditions.

Likewise, the characteristics of the company where the worker is employed were also relevant within the classification model. The variable related to company classification made an important contribution to predicting the existence of an employment contract, suggesting that the type of organization in which the worker is employed is associated with employment formality.

In the predictive analysis, both the decision tree and the neural network showed very similar levels of performance. However, priority was given to interpreting the results obtained through the decision tree because it clearly identified the most relevant variables and made it possible to understand the rules used to classify workers. This was essential, as it was the main objective of the study.

The main contribution of this study is to demonstrate that machine learning techniques can serve as a useful tool for gaining a better understanding of the factors associated with employment formality in Mexico. The findings make it possible to identify which characteristics are most strongly related to the existence of an employment contract. This information may support the development of strategies and policies aimed at improving working conditions and expanding workers' access to employment benefits.

Finally, future research should incorporate variables related to sociodemographic characteristics, such as age, sex, and educational level, among others. It would also be advisable to evaluate more advanced classification algorithms. This could contribute to a deeper understanding of the factors associated with employment formality and support the development of more suitable models.

LITERATURE

Instituto Nacional de Estadística y Geografía. (2025). Encuesta Nacional de Ingresos y Gastos de los Hogares (ENIGH) 2024. INEGI. https://www.inegi.org.mx/programas/enigh/nc/2024/

Instituto Nacional de Estadística y Geografía. (2024). *Encuesta Nacional de Ocupación y Empleo (ENOE): Indicadores de ocupación y empleo, diciembre de 2024* [Boletín de prensa]. INEGI. https://www.inegi.org.mx/contenidos/saladeprensa/boletines/2024/ENOE/ENOE2024_12.pdf

Organización Internacional del Trabajo. (2022). Panorama laboral 2022: América Latina y el Caribe. Oficina Regional de la OIT para América Latina y el Caribe. https://mexico.un.org/sites/default/files/2023-02/panorama_laboral_Am%C3%A9ricaLatinayelCaribe_2022_OIT.pdf

ISBN: 978-607-5893-96-9