

**Análisis exploratorio y predictivo de factores asociados al suicidio en Colombia a partir de
la Encuesta Nacional de Salud Mental 2015**

Dicken Santiago Barrera Cuellar

Erika Serrano Bernate

Jesús David Ruiz Peralta

Asesor

Julio Eduardo Mejia Manzano

Universidad Nacional Abierta y a Distancia UNAD

Escuela de Ciencias Básicas, Tecnología e Ingeniería – ECBTI

Especialización en Ciencia de Datos y Analítica

2026

Resumen

El presente estudio analiza los factores asociados al riesgo suicida en la población colombiana utilizando los microdatos de la Encuesta Nacional de Salud Mental (ENSM) 2015. Ante el incremento de la mortalidad por este fenómeno, se aplicó la metodología CRISP-DM para desarrollar modelos de analítica avanzada que superen las limitaciones de los análisis descriptivos tradicionales. Se entrenaron y evaluaron algoritmos de clasificación, incluyendo Regresión Logística, Random Forest y XGBoost, empleando técnicas de imputación y sobremuestreo sintético (SMOTE) para corregir el desbalance de clases sobre el grupo de entrenamiento. Los resultados demuestran que el modelo de Regresión Logística obtuvo el mejor desempeño predictivo (Recall: 0.526, F1-Score: 0.559). Los hallazgos demuestran que la Regresión Logística ofrece una capacidad discriminativa funcional para el tamizaje primario. Aunque el fenómeno suicida es inherentemente complejo y multifactorial, este modelo se posiciona como una herramienta complementaria de apoyo a la decisión clínica, permitiendo priorizar la atención en individuos con perfiles de alta vulnerabilidad.

Palabras clave: Suicidio, Minería de datos, Modelos predictivos, Salud mental, Análisis de datos.

Abstract

This study analyzes the factors associated with suicidal risk in the Colombian population using microdata from the 2015 National Mental Health Survey (ENSM). In response to the increasing mortality related to this phenomenon, the CRISP-DM methodology was applied to develop advanced analytics models capable of overcoming the limitations of traditional descriptive analyses. Classification algorithms, including Logistic Regression, Random Forest, and XGBoost, were trained and evaluated using imputation techniques and synthetic oversampling (SMOTE) to address class imbalance within the training dataset. The results show that the Logistic Regression model achieved the best predictive performance (Recall: 0.526, F1-Score: 0.559). The findings demonstrate that Logistic Regression provides functional discriminative capability for primary screening purposes. Although suicidal behavior is inherently complex and multifactorial, this model stands as a complementary decision-support tool, enabling the prioritization of care for individuals with highly vulnerable profiles.

Keywords: Suicide, Data mining, Predictive models, Mental health, Data analysis.

Tabla de Contenido

Introducción	10
Planteamiento del Problema	11
Pregunta de Investigación.....	13
Justificación	14
Objetivos.....	16
Objetivo General.....	16
Objetivos Específicos	16
Marco de Referencia.....	17
El Suicidio e Intento Suicida como Problema de Salud Pública	17
Situación del Suicidio en Colombia	17
Factores Asociados al Comportamiento Suicida.....	18
Factores Sociodemográficos	18
Factores Sociales y Económicos.....	18
Factores Psicológicos.....	18
Factores Clínicos y de Salud.....	18
Minería de Datos y Machine Learning en el Estudio de Conductas Suicidas.....	19
Marco Normativo.....	20
Metodología del Proyecto Aplicado	22
Fase 1 Comprensión del Negocio (Business Understanding)	23
Fase 2 Comprensión de los Datos (Data Understanding).....	23
Fase 3 Preparación de los Datos (Data Preparation)	23
Fase 4 Modelado (Modeling)	23

Fase 5 Evaluación (Evaluation).....	24
Fase 6 Despliegue (Deployment).....	24
Regresión Logística	24
Random Forest.....	24
XGBoost	25
Resultados.....	26
Flujo de Procesamiento de Datos	26
Filtro de Valores Nulos.....	28
Estandarización del Target	28
Factores de Riesgo.....	29
Sociodemográficas.....	29
Clínicas/Psicológicas	29
Socioeconómicas	29
Correlación	29
Índices de Salud Mental.....	29
Variables de Consumo (Sugeridas).....	29
Resumen Estadístico.....	33
Factor Etario	34
Carga de Salud Mental.....	34
Variables del Entorno	34
Variables Nulas.....	34
Análisis	36
Eliminación de Sesgo por Desbalance.....	38

Preservación de la Estructura Sociodemográfica.....	38
Efecto de la Imputación	38
Calidad del Dato	38
Modelo con Mejor Desempeño	41
Capacidad Predictiva (Recall)	41
Equilibrio General (F1-Score)	42
Interpretabilidad y Estabilidad del Modelo	42
Capacidad Discriminativa (AUC-ROC)	42
Matriz de Confusión	45
Curva ROC	47
Curva Precision – Recall	49
Conclusiones.....	51
Recomendaciones	53
Referencias Bibliográficas	55

Lista de Tablas

Tabla 1 <i>Correlación de Variables del Dataset ENSM con Respecto al Target</i>	30
Tabla 2 <i>Variables Seleccionadas para Entrenamiento</i>	31
Tabla 3 <i>Análisis y Resumen de Cambios</i>	37
Tabla 4 <i>Métricas Obtenidas</i>	40

Lista de Figuras

Figura 1 <i>Fases del Modelo de Proceso CRISP-DM Actual para la Minería de Datos</i>	22
Figura 2 <i>Flujo de Preprocesamiento de Datos</i>	27
Figura 3 <i>Distribución de Intentos Suicida</i>	28
Figura 4 <i>Distribución de Intento Suicida con Balanceo</i>	36
Figura 5 <i>Comparación de Modelos de Clasificación</i>	39
Figura 6 <i>Top 10 Factores Predictores de Riesgo Suicida</i>	44
Figura 7 <i>Matriz de Confusión</i>	45
Figura 8 <i>Curva ROC</i>	47
Figura 9 <i>Curva Precision – Recall</i>	49

Lista de Apéndices

Apéndice A *Video Sustentación*..... 60

Introducción

El intento de suicidio, entendido como una conducta autolesiva con intención de provocar la muerte, pero sin desenlace fatal, se reconoce en Colombia como un evento de notificación obligatoria desde 2016 y constituye uno de los principales predictores del suicidio consumado.

La salud mental en el país ha adquirido una relevancia creciente en las últimas décadas, especialmente a partir de la Encuesta Nacional de Salud Mental de 2015, que permitió dimensionar la magnitud de los trastornos mentales y su estrecha relación con la conducta suicida en la población. Los resultados de este estudio evidenciaron prevalencias significativas de ideación e intentos de suicidio, lo que puso de manifiesto la necesidad de fortalecer las políticas públicas y los sistemas de atención en salud mental.

En consecuencia, el panorama colombiano, sustentado en la evidencia de la encuesta de 2015 y en investigaciones posteriores, refleja la urgencia de articular esfuerzos institucionales, académicos y tecnológicos para enfrentar la problemática del suicidio. La incorporación de nuevas tecnologías, como el aprendizaje automático, modelos predictivos constituyen un marco indispensable para avanzar hacia estrategias de prevención más efectivas y garantizar una atención integral en salud mental.

Planteamiento del Problema

En Colombia, la salud mental y el suicidio representan un reto persistente. Desde el año 2016, en cumplimiento de la ley 1616 de 2013, el Instituto Nacional de Salud (INS), dio inicio a la vigilancia epidemiológica del intento de suicidio. Desde que se inició este proceso las cifras reportadas se han incrementado dramáticamente. Este aumento refleja una problemática de salud pública cada vez más visible y preocupante en el país, el crecimiento de las tasas también pone en evidencia la necesidad de fortalecer las estrategias de prevención y atención psicológica en el país. Pero más allá de las cifras, cada caso representa una persona que atraviesa situaciones complejas y requiere apoyo emocional, familiar e institucional.

En 2015 se reportaron en el SISPRO 1.987 casos de intento de suicidio, en 2016 el SIVIGILA informó que se presentaron 17.587 casos; evidenciando una cifra casi nueve veces más alta de un año al otro, en 2017 la cifra preliminar fue de 25.835 casos, con un incremento del 46,9% con respecto al año anterior, esto podría explicarse por el subregistro que existía del evento (boletín-conducta-suicida, s. f., p. 6).

La Procuraduría General también ha alertado sobre un aumento de suicidios en jóvenes: durante el primer semestre de 2023 se reportaron **3.145 suicidios**, de los cuales **925 ocurrieron en jóvenes** (PROCURADURIA GENERAL DE LA NACION, 2023) Estos datos justifican la importancia de profundizar en los factores que explican la conducta suicida, más allá de los análisis descriptivos tradicionales, para diseñar estrategias preventivas efectivas.

Determinar los factores asociados con este comportamiento complejo es el primer paso en el desarrollo de estrategias de intervención temprana (Buriticá et al., 2023). Es necesario reconocer y abordar diversos factores asociados con el intento de suicidio en pacientes depresivos para desplegar acciones de promoción y prevención, identificación temprana e

intervenciones específicas que impacten en las cifras de suicidio consumado en el país (Betancur Díaz et al., 2023).

Por ello, reducir el número de muertes por suicidio es un imperativo mundial establecido en los Objetivos de Desarrollo Sostenible de las Naciones Unidas y en el Plan de Acción Integral de Salud Mental de la OMS 2013– 2030. Sin embargo, el mundo no está en camino de alcanzar estas metas por lo que se requieren esfuerzos acelerados y sostenidos para prevenir el suicidio y evitar que más personas pierdan la vida innecesariamente (Organización Panamericana de la Salud, 2021). Las políticas y los programas que se plantean la prevención y el control de los comportamientos suicidas producen frecuentemente resultados decepcionantes (Corona-Miranda et al., 2014).

Revisando la literatura y los reportes institucionales, se observa que si bien existen estadísticas nacionales robustas (como las del DANE), faltan estudios que empleen técnicas como minería de datos o machine learning para analizar tendencias complejas. A nivel internacional, se han desarrollado modelos predictivos de suicidio que utilizan diversas variables y muestran buen rendimiento, pero estos esfuerzos no están suficientemente adaptados al contexto colombiano y su realidad social, cultural y educativa.

El problema central es que, aunque existen fuentes de datos nacionales robustas (como los microdatos de la ENSM 2015) y evidencia internacional sobre el uso de machine learning para predicción del riesgo suicida, en Colombia no se han desarrollado modelos analíticos avanzados que identifiquen patrones de riesgo de intento suicida a partir de encuestas de salud mental validadas en el contexto nacional. Esta carencia limita la capacidad de los sistemas de salud, los orientadores escolares y las autoridades sanitarias para desarrollar estrategias de detección temprana basadas en evidencia, más allá del análisis descriptivo convencional. La

ausencia de este tipo de herramientas afecta especialmente a la población colombiana en su conjunto, que enfrenta una combinación de factores de riesgo específicos como la exposición histórica al conflicto armado, la desigualdad socioeconómica y las brechas en el acceso a servicios de salud mental.

Pregunta de Investigación

¿De qué manera la aplicación de técnicas de minería de datos y modelos de machine learning sobre los microdatos de la ENSM 2015 puede contribuir a la identificación y clasificación de factores asociados al intento suicida en la población colombiana, como insumo para el desarrollo de sistemas de alerta temprana en salud mental?

Justificación

El análisis del riesgo suicida en Colombia ha adquirido una relevancia crítica debido al incremento sostenido en las tasas de mortalidad e intentos reportados. Según datos del DANE (2024), las muertes por suicidio han mostrado repuntes significativos en diversos grupos etarios, lo que plantea un desafío de salud pública sin precedentes. Informes técnicos del Ministerio de Salud corroboran que esta tendencia es especialmente alarmante en la población juvenil (MinSalud, 2018), lo que exige una evolución en los métodos de detección y prevención.

Esta investigación aporta un valor estratégico tanto al ámbito científico como al educativo al proponer un modelo predictivo basado en datos reales. En la actualidad, el Machine Learning (ML) se ha consolidado como una herramienta esencial para el sistema de salud, permitiendo la transición de análisis descriptivos hacia la detección temprana y proactiva de trastornos mentales (Castrillon et al., 2022). La integración de algoritmos de aprendizaje automático busca complementar la evaluación clínica tradicional, optimizando la precisión en la identificación de individuos en riesgo mediante el análisis de grandes volúmenes de datos que los métodos convencionales no logran procesar con la misma agilidad (Tang et al., 2024; Atmakuru et al., 2025).

Para el desarrollo de este modelo, se utilizan los microdatos de la Encuesta Nacional de Salud Mental (ENSM) 2015. La elección de esta fuente se fundamenta en que representa el último conjunto de datos con representatividad nacional, profundidad temática y validación oficial disponible en Colombia hasta la fecha. Si bien el Ministerio de Salud y Protección Social anunció el inicio de la quinta edición de la ENSM en julio de 2025 (Ministerio de Salud y Protección Social, 2025), esta sigue desarrollándose en fase de recolección de datos en 2026.

Finalmente, este estudio tiene el potencial de transformar la política pública y la práctica docente. Al demostrar la eficacia de la analítica predictiva, los resultados podrán servir como evidencia para que los Ministerios de Educación y Salud diseñen estrategias conjuntas de prevención más eficaces, orientadas a sistemas de alerta temprana que mitiguen el impacto del suicidio en la sociedad colombiana.

Objetivos

Objetivo General

Analizar los factores asociados al intento suicida en la población colombiana mediante modelos predictivos, integrando variables sociodemográficas, clínicas y ambientales de fuentes oficiales y literatura especializada.

Objetivos Específicos

Consolidar un conjunto de datos confiable sobre suicidio en Colombia, mediante la identificación, integración, depuración y preprocesamiento de bases de datos oficiales de salud.

Analizar las variables relevantes vinculadas al suicidio mediante técnicas estadísticas y de análisis de datos, con el propósito de identificar patrones, tendencias y relaciones significativas.

Validar el desempeño de los modelos predictivos utilizando métricas apropiadas y procedimientos de validación, para establecer su capacidad de generalización y aplicabilidad en escenarios reales.

Marco de Referencia

El Suicidio e Intento Suicida como Problema de Salud Pública

El suicidio y el intento suicida constituyen uno de los principales desafíos de salud pública a nivel mundial. La Organización Mundial de la Salud (WHO, 2025).

El intento suicida se define como una conducta autolesiva llevada a cabo con intención de provocar la propia muerte, pero sin desenlace fatal (WHO, 2014). A diferencia del suicidio consumado, el intento suicida es un evento observable y registrable en vida del individuo, lo que lo convierte en la variable de análisis central de los modelos predictivos orientados a la detección temprana. La OPS (2021) estima que, por cada suicidio consumado, ocurren 8 intentos, lo que dimensiona la magnitud epidemiológica del fenómeno y justifica su estudio independiente.

Situación del Suicidio en Colombia

En Colombia las cifras muestran una tendencia creciente. Entre 2015 y 2023 la tasa de suicidios pasó de 5,1 a 6,2 muertes por cada 100.000 habitantes, según cifras preliminares del DANE (2024). Por otro lado, de acuerdo con el Instituto Nacional de Salud (2025), la incidencia de intento de suicidio tuvo un aumento progresivo desde el 2020 hasta el 2023, años después de los cuales se observa un decremento para luego adoptar un comportamiento estable a 2025. Esta tendencia preocupa particularmente porque los intentos de suicidio y su consumación representan una alta carga emocional y económica para la sociedad en general.

Los estudios realizados en el país señalan diversos factores específicos para el contexto nacional. Aparicio Castillo et al. (2020) identificaron que los pacientes con ideación suicida persistente tienen una probabilidad 4 veces mayor de hacer dos o más intentos. Castro Moreno et al. (2023) concluyeron que los pacientes masculinos, con historial de depresión, antecedentes de enfermedades crónicas y exposición a alta carga laboral tenían mayor riesgo de suicidio.

Otros estudios nacionales (Betancur Díaz et al., 2023; Buriticá et al., 2023; Cuello Villamil et al., 2023)-señalaron factores como baja escolaridad, conflictos de pareja, uso de sustancias psicoactivas, desempleo, pobreza y antecedentes psiquiátricos.

Factores Asociados al Comportamiento Suicida

A través de la literatura en general-se distinguen-múltiples factores asociados al suicidio que pueden agruparse en al menos cuatro grupos:

Factores Sociodemográficos

Hombres en edad adulta tienen alta probabilidad de suicidio consumado, en comparación con otros grupos poblacionales (Castro Moreno et al., 2023).

En adolescentes, el riesgo de comportamiento suicida es mayor en mujeres, de niveles socioeconómicos bajos (Buriticá et al., 2023).

Factores Sociales y Económicos

No recibir ingresos, no tener pareja como desencadenante de intento de suicidio (Betancur Díaz et al., 2023).

Personas solteras tienen mayor riesgo de intento y los conflictos de pareja o rupturas amorosas son factores desencadenantes (Cuello Villamil et al., 2023).

Factores Psicológicos

Depresión mayor, ansiedad y trastornos afectivos como factores de riesgo (Ntakolia et al., 2025).

Ideación suicida persistente (Aparicio Castillo et al., 2020).

Factores Clínicos y de Salud

Enfermedades crónicas (Castro Moreno et al., 2023).

Acceso limitado a servicios de salud mental y no contar con redes de apoyo (Varma & Batterham, 2025).

En adición, literatura reciente coincide en que los factores asociados al comportamiento suicida son dinámicos y varían según la población, el grupo etario, la composición social, las características clínicas, en general el contexto modifica las variables predictoras. Por ejemplo, los factores predictores difieren entre adolescentes de distintas etapas del desarrollo (Kusuma et al., 2024), entre regiones con particularidades sociales específicas (Lyu et al., 2025). Además, revisiones más amplias señalan que los factores de riesgo no son universales y deben interpretarse de manera culturalmente contextualizada (Atmakuru et al., 2025).

Minería de Datos y Machine Learning en el Estudio de Conductas Suicidas

La complejidad en el fenómeno estudiado ha motivado el uso de técnicas de minería de datos y modelos de machine learning (ML) para identificar patrones y mejorar la predicción del riesgo de suicidio. Estas técnicas han permitido analizar grandes volúmenes de información, encontrando relaciones con variables que aumentan la probabilidad de eventos suicidas, así como lo señala Atmakuru et al. (2025) en una revisión de 156 estudios.

En consiguiente, diversos estudios han empleado modelos tanto clásicos como avanzados para la predicción del riesgo suicida. Entre los algoritmos más utilizados se encuentran la regresión logística, los bosques aleatorios (Random Forest), métodos basados en gradientes como XGBoost, los cuales han mostrado desempeños significativos en tareas de clasificación en este problema (Ntakolia et al., 2025; Tang et al., 2024). También se han aplicado modelos de redes neuronales y técnicas de aprendizaje profundo para capturar relaciones entre conjuntos de datos clínicos y sociodemográficos (Atmakuru et al., 2025). Esta diversidad refleja la amplitud de enfoques de abordaje y las múltiples perspectivas analíticas de este problema.

Marco Normativo

El estudio de los factores asociados al suicidio en Colombia se sustenta en un conjunto de normas que reconocen la salud mental como un derecho fundamental y un asunto prioritario de salud pública. Estas disposiciones permiten no solo la atención clínica, sino también el análisis estadístico y epidemiológico, como el que se realiza a partir de la Encuesta Nacional de Salud Mental 2015.

En primer lugar, la Constitución Política de Colombia (1991) establece en su artículo 49 que la salud es un servicio público a cargo del Estado, lo cual implica la obligación de garantizar el acceso a servicios de promoción, prevención y atención en salud (Art. 49 *Constitución Política 1 de 1991 Asamblea Nacional Constituyente - Gestor Normativo*, s. f.). Este principio constituye la base para el desarrollo de políticas relacionadas con la salud mental y la conducta suicida.

Un avance fundamental en el marco normativo es la Ley 1616 de 2013, conocida como la Ley de Salud Mental la cual reconoce la salud mental como un derecho fundamental y de interés público. Esta ley establece que el Estado debe garantizar la promoción, prevención, diagnóstico, tratamiento y rehabilitación de los trastornos mentales, bajo un enfoque integral y basado en derechos humanos (*Ley 1616 de 2013 Congreso de la República de Colombia*, s. f.). Además, define la salud mental como un proceso dinámico que permite a las personas desarrollar sus capacidades y contribuir a la sociedad, lo cual es clave para entender los factores protectores y de riesgo asociados al suicidio.

Asimismo, esta ley incorpora la necesidad de sistemas de información y vigilancia epidemiológica en salud mental, incluyendo la conducta suicida como evento de interés en salud pública. Esto respalda el uso de fuentes de datos como la Encuesta Nacional de Salud Mental, ya

que promueve la generación de conocimiento para la toma de decisiones y el diseño de políticas públicas.

Por otra parte, la Ley 1751 de 2015 (Ley Estatutaria de Salud) fortalece este marco al reconocer la salud como un derecho fundamental autónomo, garantizando el acceso oportuno, eficaz y de calidad a los servicios de salud, incluyendo la salud mental (Pérez, 2014). Esta normativa enfatiza la dignidad humana y la integralidad de la atención, lo que implica abordar el suicidio no solo desde una perspectiva clínica, sino también social y preventiva.

En conclusión, el marco normativo colombiano ofrece una base sólida para el análisis del suicidio desde un enfoque integral, al articular el derecho a la salud, la política pública en salud mental y los sistemas de información epidemiológica. Esto permite que estudios basados en datos, como el análisis exploratorio y predictivo, contribuyan a la comprensión del fenómeno y al diseño de estrategias efectivas de prevención.

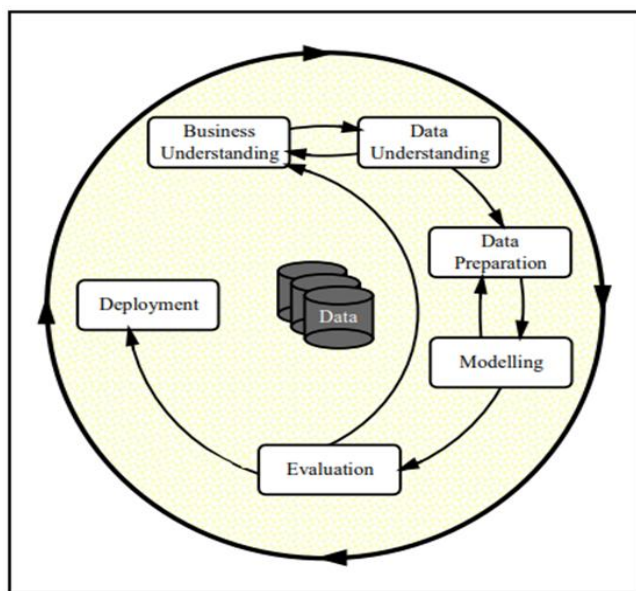
Metodología del Proyecto Aplicado

La metodología del proyecto se basa en el estándar CRISP-DM (Cross-Industry Standard Process for Data Mining), un marco para el desarrollo de proyectos de minería de datos y análisis predictivo. Este enfoque garantiza un proceso iterativo, sólido y preciso que abarca todas las etapas, desde la comprensión del contexto hasta la validación del modelo predictivo.

El círculo exterior de la Figura 1 simboliza la naturaleza cíclica de la minería de datos. La minería de datos no termina una vez implementada la solución. Las lecciones aprendidas durante el proceso y a partir de la solución implementada pueden generar nuevas preguntas de negocio, a menudo más específicas (Wirth & Hipp, s. f.).

Figura 1

Fases del Modelo de Proceso CRISP-DM Actual para la Minería de Datos



Nota. CRISP-DM: Towards a standard process model for data mining

Fase 1 Comprensión del Negocio (Business Understanding)

Se definirá el problema analítico como la identificación de factores asociados al intento suicida en la población colombiana a partir de los microdatos de la Encuesta Nacional de Salud Mental (ENSM) 2015. En esta fase se establecerá el alcance del proyecto, su pertinencia en salud pública y el valor del modelo como apoyo para la comprensión de patrones de riesgo y la detección temprana.

Fase 2 Comprensión de los Datos (Data Understanding)

Se realizará una revisión de los microdatos de la ENSM 2015 para comprender su estructura, cobertura, organización por módulos y calidad de la información. Además, se identificarán los archivos de adolescentes y adultos, la variable `identificador_persona` como llave de integración, y se explorarán la distribución de `intento_suici`, los valores faltantes y el desbalance de clases para evaluar la viabilidad del análisis y del modelado predictivo.

Fase 3 Preparación de los Datos (Data Preparation)

Se realizará un análisis exploratorio de datos para determinar la distribución, la calidad, la presencia de valores atípicos y las correlaciones iniciales de las variables. Se aplicarán métodos de limpieza de datos para eliminar los valores faltantes y resolver las inconsistencias. Posteriormente se procederá a la normalización o estandarización de las variables de las variables numéricas y codificación de las variables categóricas. Finalmente se hará creación de variables para tener en cuenta los factores de riesgo complejos identificados en el marco teórico.

Fase 4 Modelado (Modeling)

Se utilizarán algoritmos de Machine Learning para la tarea de clasificación del riesgo suicida. Los modelos se seleccionarán en función de su eficiencia en la literatura científica para el análisis de la conducta suicida. Se entrenarán y evaluarán inicialmente la Regresión Logística,

Random Forest y métodos basados en gradiente como XGBoost (Atmakuru et al., 2025; Ntakolia et al., 2025).

Fase 5 Evaluación (Evaluation)

Se comprobará el rendimiento de los modelos desarrollados, tal como se plantea en los objetivos específicos. Se emplearán métricas clave de clasificación como: Precisión, Recall, F1-Score y Área bajo la curva ROC (AUC). El modelo que demuestre el mejor equilibrio entre sensibilidad y especificidad será el seleccionado.

Fase 6 Despliegue (Deployment)

Finalmente se creará un informe final con las conclusiones y recomendaciones claras. Se detallará como el modelo seleccionado puede servir como herramienta de apoyo en sistemas de alerta temprana para profesionales de la salud y orientadores escolares.

El proyecto se desarrollará utilizando el lenguaje de programación Python, el uso se debe a que ofrece un ecosistema robusto de librerías para todas las fases de CRISP-DM. Para la preparación de los datos tomaremos las librerías Pandas y Numpy, Para el modelado usaremos la librería Scikit-learn con los modelos de Regresión Logística, Random Forest y XGBoost:

Regresión Logística

Es altamente interpretable y ofrece una base sólida para entender la contribución lineal de cada factor de riesgo en la probabilidad de la conducta suicida.

Random Forest

Minimiza el sobreajuste y maneja bien las variables no lineales y la interacción entre ellas, lo cual es fundamental dado que el fenómeno del suicidio es multicausal.

XGBoost

Conocido por su alta precisión en tareas de clasificación al combinar múltiples árboles de decisión de forma secuencial.

Resultados

Siguiendo la metodología CRISP-DM, la primera fase de resultados se centró en la consolidación de una fuente de datos con la robustez necesaria para el entrenamiento de modelos de Machine Learning.

Tras evaluar las bases de datos de Estadísticas Vitales del DANE, se determinó que, aunque se precisa en la caracterización del suicidio consumado, carecen de la dimensionalidad necesaria (variables predictoras y casos de control "no evento") para un modelo de clasificación binaria. En consecuencia, se seleccionó la Encuesta Nacional de Salud Mental (ENSM) 2015, la cual ofrece una profundidad temática validada en el contexto colombiano.

Dada la estructura de la ENSM, donde la información de un mismo individuo se distribuye en módulos complementarios, se realizó un proceso de integración mediante la columna *identificador_persona*.

Inicialmente se planteó el estudio en población adolescente; sin embargo, el análisis exploratorio reveló un desbalance crítico (solo 88 casos de intento suicida registrados).

Para garantizar la capacidad de generalización del modelo, se optó por unificar los microdatos de adolescentes y adultos, consolidando un dataset de 799 registros válidos con un balance de clases más manejable (317 casos positivos de intento y 482 negativos).

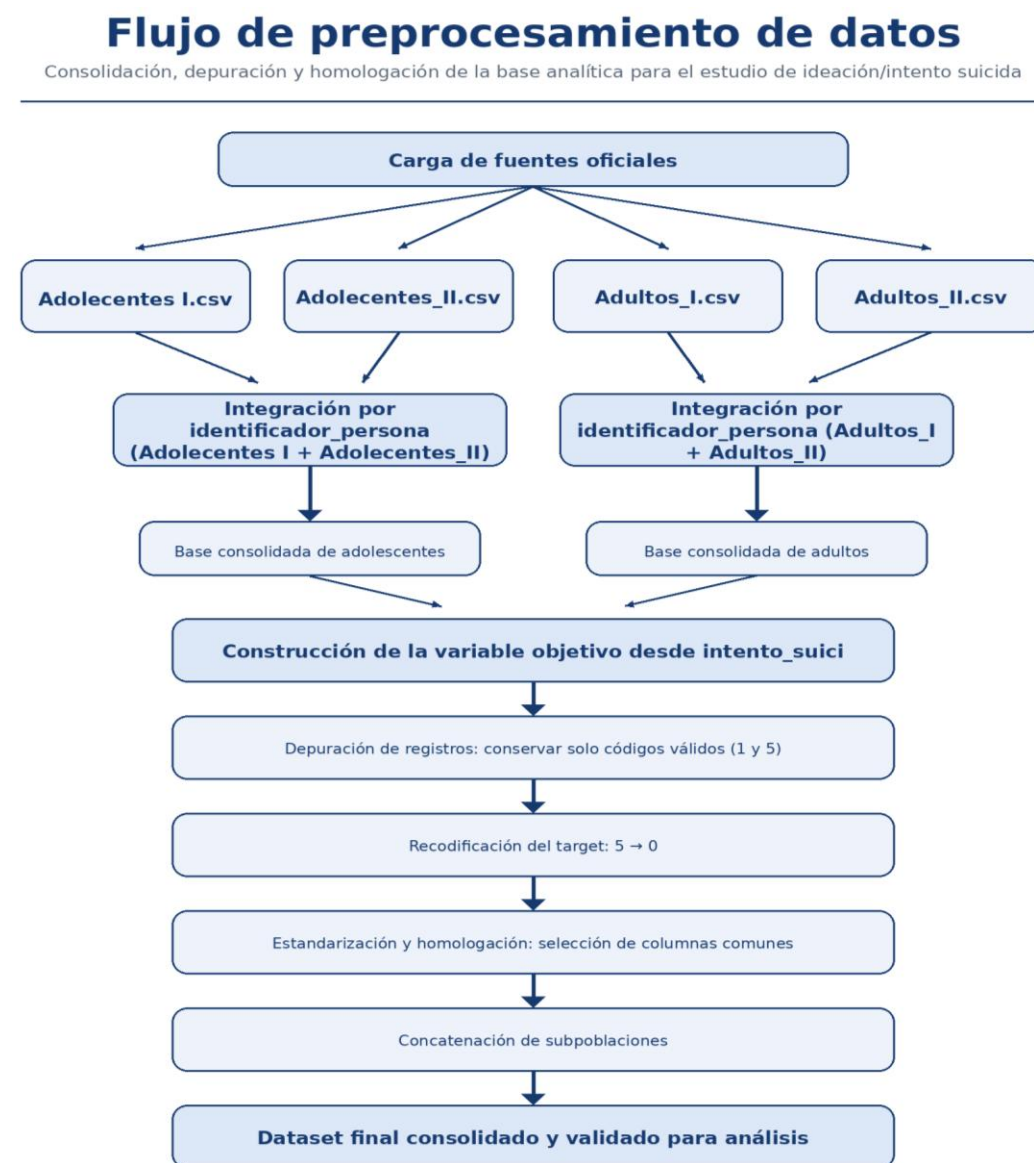
Flujo de Procesamiento de Datos

En la siguiente ilustración se muestra el flujo seguido para este proceso de preprocesamiento de datos, donde se tuvo en cuenta los distintos grupos poblacionales,

integración de archivos mediante identificadores únicos y posterior estandarización de variable objetivo.

Figura 2

Flujo de Preprocesamiento de Datos



De esta manera, se obtuvo el dataset unificado que contaba originalmente con 1.560 columnas, lo que representaba un riesgo de muchas variables o características. Luego se aplicaron los siguientes criterios de exclusión:

Filtro de Valores Nulos

Se eliminaron las variables con más del 50% de datos faltantes, reduciendo el conjunto a 696 variables.

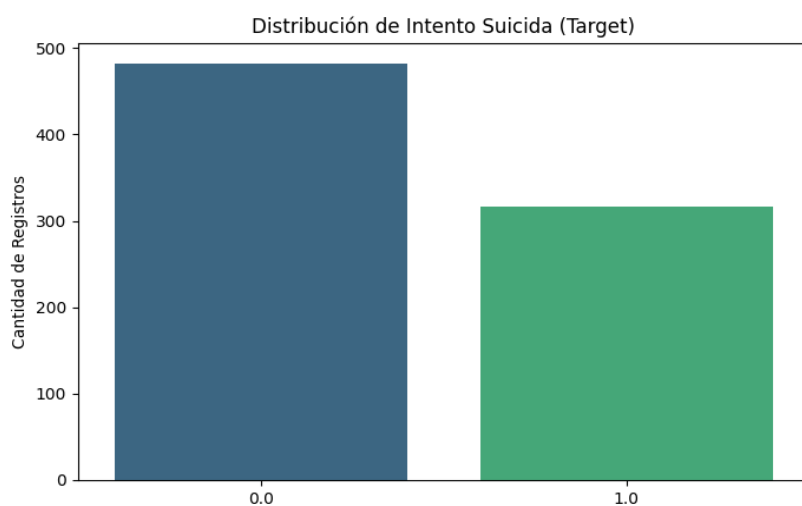
Estandarización del Target

Se recodificó la variable *intento_suici*, transformándola en una estructura binaria donde (1) representa la presencia de intento suicida y (0) la ausencia (originalmente codificada como (5) en la ENSM).

En la figura 3, se demuestra que existe una proporción importante de casos positivos, 317 personas con intento suicida no es minoría pequeña, por tanto, el modelo a implementar tiene suficientes ejemplos para aprender patrones reales.

Figura 3

Distribución de Intentos Suicida



Factores de Riesgo

Para la definición de los factores de riesgo definitivos, se implementó un enfoque híbrido que combina la evidencia científica (literatura) con la evidencia estadística (correlación).

Se cruzaron variables identificadas en una tesis doctoral (Artieda Urrutia, 2017) y artículos especializados como: (Castro Moreno et al., 2023); (Aparicio Castillo et al., 2020); (Buriticá et al., 2023) que han demostrado relevancia en el contexto nacional. Las dimensiones seleccionadas incluyen:

Sociodemográficas

Edad, sexo, nivel educativo, estado civil y área residencial (*rural/urbana*).

Clínicas/Psicológicas

Presencia de depresión mayor (*d_mddh30*), ansiedad generalizada (*d_gad30*) y bipolaridad.

Socioeconómicas

Problemas económicos graves (*m11_bp31*) y pérdida de empleo (*m11_bp30*).

Correlación

Se calculó la correlación de Pearson respecto al target para identificar patrones no evidentes en la literatura, pero significativos en el dataset actual. Destacan variables de diagnóstico global y recurrencia:

Índices de Salud Mental

Dimensión (escala de sano a trastorno) y *numero_trastornos_vida*.

Variables de Consumo (Sugeridas)

sqr_sug_global_2 y *sqr_sug_psicosis2*, que vinculan el consumo de sustancias en los últimos 12 meses con el riesgo detectado.

A continuación, se presenta el cruce de variables seleccionadas. El criterio de inclusión final se basa en la presencia de la variable en la literatura especializada o una correlación de Pearson superior a 0.15 con el intento suicida.

Tabla 1

Correlación de Variables del Dataset ENSM con Respecto al Target

Variable (ENSM)	Descripción Técnica	Correlación
planear_suici	Ideación suicida previa	0,4277
sd4	Experiencias traumáticas (Evento B)	0,4252
sqr_sug_global_2	Consumo sustancias últimos 12 meses	0,1960
dimension	Escala: Sano / Problema / Trastorno	0,1906
sqr_sug_psicosis2	Indicador de riesgo por psicosis	0,1840
numero_trastornos_12m	Sumatoria de trastornos en la vida (ult. 12 meses)	0,1748
sqr_sug_dep_ans_20	Consumo alcohol (Últimos 12 meses)	0,1698
numero_trastornos_vida	Sumatoria de trastornos en la vida	0,1541

Tabla 2*Variables Seleccionadas para Entrenamiento*

Dimensión	Variable (ENSM)	Descripción Técnica	Fuente
Demográfica	Age	Edad del individuo	
	Sex	Sexo (Género)	Castro Moreno et
	rural_urb	Área de residencia (Rural/Urbano)	al. (2023); Buriticá et al.
	NivelEducativo	Grado de instrucción alcanzado	(2023)
	EstadoCivil	Situación sentimental/legal	
Social	m2_p56	Participación en grupos (Religiosos, Deportivos, etc.)	Betancur Díaz et
	m2_p27	Ocupación principal del individuo	al. (2023)
Socioeconómica	m11_bp31	Problemas económicos graves	Betancur Díaz et
	m11_bp30	Pérdida de empleo o jubilación	al. (2023); Cuello Villamil et al. (2023)
Salud Mental	d_mddh30	Depresión mayor (30 días)	Aparicio Castillo
	d_gad30	Ansiedad generalizada (30 días)	et al. (2020); Buriticá et al. (2023)

	d_bipolarI30	bipolaridad 30 días	(Artieda Urrutia, 2017)
	Dimension	Escala: Sano / Problema / Trastorno	Correlación
	numero_trastornos_12m	Sumatoria de trastornos en la vida (ult. 12 meses)	Correlación
	numero_trastornos_vida	Sumatoria de trastornos en la vida	Correlación
	sqr_sug_global_2	Consumo sustancias últimos 12 meses	Correlación
Consumo	sqr_sug_dep_ans_20	Consumo alcohol (Últimos 12 meses)	Correlación
	sqr_sug_psicosis2	Indicador de riesgo por psicosis	Correlación
	m14_p1	Frecuencia de consumo de alguna bebida alcohólica	Buriticá et al. (2023)
Experiencias	sd4	Experiencias traumáticas (Evento B)	Correlación
	planear_suici	Ideación suicida previa	Correlación

Es importante aclarar que las variables seleccionadas no deben interpretarse como causas directas del intento suicida. En el marco de un modelo de clasificación supervisada, estas variables son factores asociados que aportan poder discriminativo al algoritmo, consistentes con

una concepción multicausal del fenómeno (WHO, 2025; Atmakuru et al., 2025). Su validez es específica para la población colombiana captada por la ENSM 2015 y no puede generalizarse sin validación adicional a otras poblaciones o períodos temporales.

Un caso que merece análisis particular es la variable `planear_suici` (ideación suicida previa). Su inclusión como predictor está respaldada por la literatura: los pacientes con ideación suicida persistente tienen una probabilidad significativamente mayor de realizar dos o más intentos Aparicio Castillo et al. (2020). Sin embargo, esta variable presenta una proximidad conceptual con el desenlace que puede introducir cierto grado de circularidad: la ideación suicida es parte del espectro de la conducta suicida y no constituye un factor de riesgo externo e independiente.

En el contexto de este modelo, su inclusión se justifica porque representa un estado clínico observable y evaluable que precede al intento en la mayoría de los casos documentados, y su omisión reduciría la capacidad predictiva del algoritmo de forma sustancial. No obstante, su peso en el modelo debe interpretarse con cautela y en ningún caso debe utilizarse de forma aislada como único criterio de alerta en contextos clínicos.

Después de seleccionar las variables o factores que serán determinantes para el modelo de clasificación, se realizó un análisis descriptivo exhaustivo del dataset consolidado, el cual cuenta con 799 registros y 21 predictores seleccionados bajo criterios de literatura y relevancia estadística.

Resumen Estadístico

El resumen estadístico revela hallazgos clave sobre la población estudiada.

La variable dependiente (intento suicida) presenta un desbalance moderado que es crucial documentar para justificar el uso posterior de técnicas de sobremuestreo (SMOTE).

Clase Negativa (0.0), 482 registros, equivalentes al 60.33% de la muestra.

Clase Positiva (1.0), 317 registros, que representan el 39.67%. Este volumen de casos positivos es significativamente alto para un modelo de salud mental, lo que proporciona una base de entrenamiento robusta para identificar patrones de riesgo.

Aunque el desbalance no puede considerarse severo, la predominancia de la clase “No intento” podía inducir sesgos en el entrenamiento del modelo, favoreciendo predicciones hacia la clase mayoritaria.

Factor Etario

La muestra abarca un rango desde los 12 hasta los 92 años, con una media de 38.8 años ($SD=16.8$). Esto permite que el modelo tenga capacidad de generalización en diferentes etapas del ciclo vital, desde la adolescencia hasta la vejez.

Carga de Salud Mental

Se observa que el número promedio de trastornos en la vida es de 0.57, con casos máximos de hasta 3 trastornos concurrentes. Variables como d_mddh30 (Depresión) y d_gad30 (Ansiedad) muestran una tendencia hacia valores altos en la escala (media cercana a 5), lo cual es consistente con la literatura que identifica estos diagnósticos como predictores críticos.

Variables del Entorno

La variable $sd4$ (experiencias traumáticas) muestra una media de 3.61, lo que indica una presencia considerable de eventos estresantes en la muestra, factor que la literatura nacional destaca como desencadenante de conductas autolesivas.

Variables Nulas

El análisis de nulos identificó áreas que requieren tratamiento mediante imputación.

Ocupación (m2_p27). Registra 387 nulos, lo cual es coherente con los flujos de respuesta de la ENSM 2015, donde esta pregunta depende de condiciones laborales previas.

Dimensiones de Salud Mental. Variables como *numero_trastornos_12m* (272 nulos) y *sqr_sug_dep_ans_20* (249 nulos) presentan faltantes que serán tratados mediante imputación por la moda o mediana para evitar la pérdida de información valiosa durante el entrenamiento del modelo.

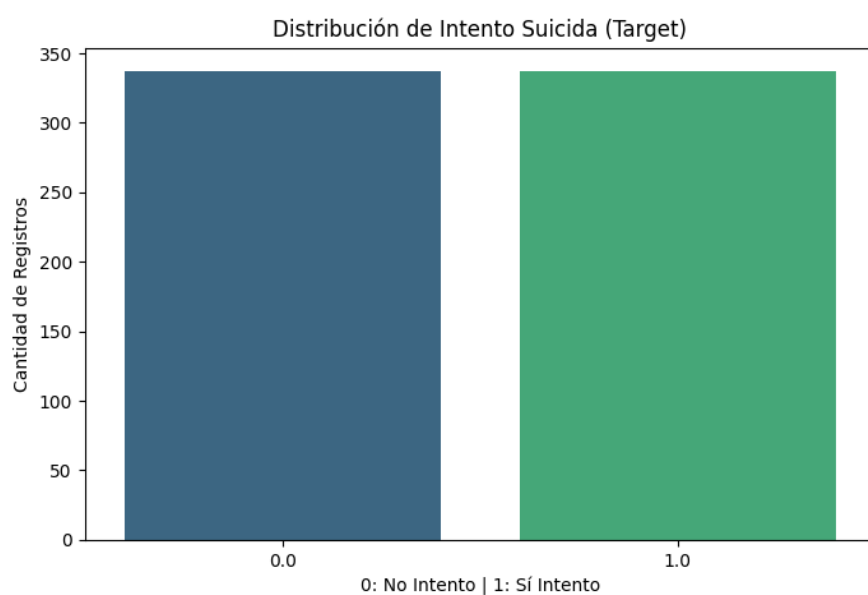
Una vez identificadas estas particularidades en la distribución etaria, las cargas de salud mental y los vacíos de información en variables críticas como ocupación y trastornos, se procedió a la fase de Preparación de Datos. Este paso fue fundamental para mitigar el riesgo de sesgo en los algoritmos, aplicando técnicas de imputación estadística para tratar los nulos detectados y el algoritmo SMOTE para corregir el desbalance del target. El objetivo fue transformar el dataset crudo en una base de conocimiento robusta que conserve la representatividad de la población colombiana.

Para la aplicación de SMOTE se siguió un protocolo metodológico estricto para evitar la fuga de información (data leakage), uno de los errores más comunes y de mayor impacto en los estudios de machine learning aplicados a salud. Concretamente, el sobremuestreo sintético se aplicó después de la partición train/test y exclusivamente sobre el conjunto de entrenamiento (70% del dataset original). El conjunto de prueba (30% restante) no fue sometido a ninguna transformación de balanceo y conservó la distribución de clases original de la ENSM 2015. Al respetar la secuencia correcta (particionar, imputar, balancear en entrenamiento), las métricas de resultados reflejan el desempeño genuino del modelo ante datos con distribución real, que es la condición que encontraría el sistema en una implementación real de alerta temprana.

En la figura 4, se presenta la distribución de clases y posteriormente la aplicación del balanceo mediante SMOTE. Aunque la diferencia entre clases no era extrema, se consideró pertinente realizar el balanceo debido a que los algoritmos de clasificación tienden a favorecer la clase dominante, reduciendo la sensibilidad en la detección de casos positivos de riesgo suicida.

Figura 4

Distribución de Intento Suicida con Balanceo



Análisis

Como resultado de este tratamiento, la Tabla 3 presenta un análisis comparativo entre el dataset original y el conjunto de entrenamiento balanceado después de los procesos de imputación y sobremuestreo sintético. Los resultados muestran que el balanceo logró equilibrar las clases sin generar distorsiones significativas en las distribuciones originales de variables relevantes como edad, nivel educativo, trastornos de salud mental y experiencias traumáticas.

Tabla 3*Análisis y Resumen de Cambios*

Métrica / Variable	Dataset Original (N=799)	Dataset Balanceado (N=674)	Observación
			Balance perfecto
Balance Target (0 / 1)	60,3% / 39,7%	50,0% / 50,0%	alcanzado mediante SMOTE.
Media de Edad (age)	38,80 años	38,24 años	Variación mínima tras imputación y balanceo.
Desviación Edad (std)	16,87	16,71	La dispersión se mantuvo estable.
Nivel Educativo (mean)	1,75	1,76	Sin cambios significativos.
Trastornos Vida (mean)	0,57	0,44	Ligera disminución por imputación de nulos a la moda (0).
Exp. Traumáticas (sd4)	3,61	3,55	Consistencia total en el factor de riesgo.
Valores Nulos (Total)	1.884 nulos detectados	0	Limpieza completa del dataset.

Eliminación de Sesgo por Desbalance

El target pasó de una proporción 60/40 a una distribución equitativa. Esto es fundamental para que el modelo no se sesgue hacia la clase mayoritaria (no intento) y mejore su capacidad de detectar el riesgo real (clase 1).

Preservación de la Estructura Sociodemográfica

VARIABLES CRÍTICAS COMO LA EDAD (*age*), EL SEXO (*sex*) Y EL ESTADO CIVIL (*EstadoCivil*) MUESTRAN MEDIAS Y DESVIACIONES ESTÁNDAR CASI IDÉNTICAS. ESTO CONFIRMA QUE LOS REGISTROS SINTÉTICOS CREADOS POR SMOTE SON COHERENTES CON LA POBLACIÓN COLOMBIANA ORIGINAL.

Efecto de la Imputación

LA COLUMNA *m2_p27 (Ocupación)*, QUE ANTES TENÍA 387 NULOS, AHORA ESTÁ COMPLETA. SE OBSERVA UN AJUSTE EN SU MEDIA DEBIDO A QUE LOS VALORES FALTANTES FUERON REEMPLAZADOS POR LA MODA (LA OCUPACIÓN MÁS FRECUENTE), LO QUE ESTABILIZA EL PREDICTOR PARA EL ALGORITMO DE MACHINE LEARNING.

Calidad del Dato

EL CHEQUEO DE NULOS ARROJA `Series([], dtype: int64)`, LO QUE SIGNIFICA QUE YA NO EXISTEN VACÍOS EN LAS 21 COLUMNAS PREDICTORAS, GARANTIZANDO QUE ALGORITMOS COMO XGBOOST Y RANDOM FOREST PROCESEN EL 100% DE LA MUESTRA BALANCEADA.

FINALMENTE, SE IMPLEMENTARON TRES ALGORITMOS DE APRENDIZAJE SUPERVISADO PARA IDENTIFICAR LOS FACTORES ASOCIADOS AL RIESGO SUICIDA. LOS RESULTADOS SE COMPARARON BAJO UN ESQUEMA DE PARTICIÓN DE DATOS 70/30.

Regresión Logística. Se utilizó como modelo base por su alta interpretabilidad de los coeficientes.

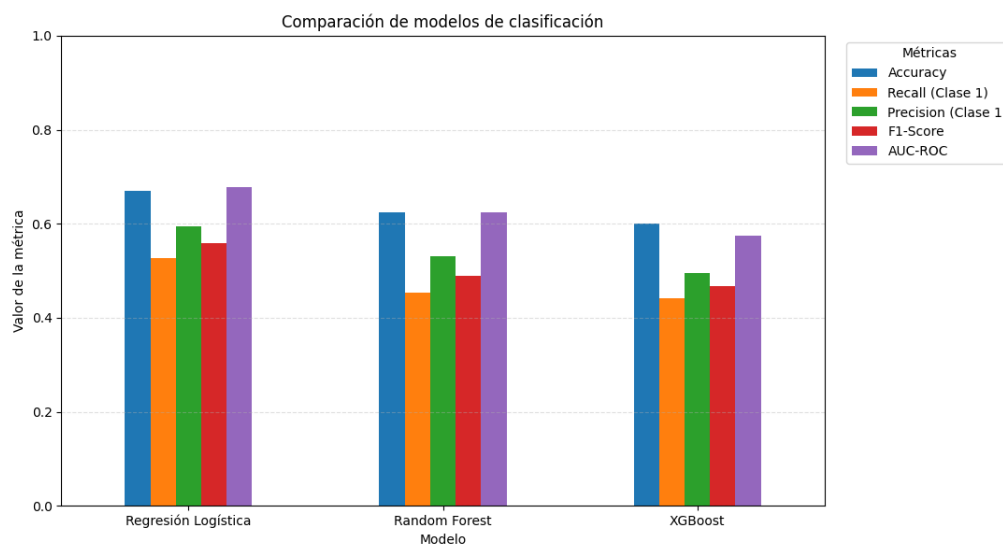
Random Forest. Implementado para capturar interacciones no lineales entre variables sociodemográficas y clínicas, minimizando el sobreajuste.

XGBoost. Algoritmo de aumento de gradiente seleccionado por su eficiencia computacional y precisión superior en tareas de clasificación complejas.

Se entrenaron los tres algoritmos de aprendizaje supervisado sobre el train balanceado. El objetivo principal fue maximizar la detección de casos positivos (intento suicida), priorizando la métrica de Recall (Sensibilidad) sobre la precisión global, dado que en contextos de salud mental el costo de un falso negativo (no detectar a alguien en riesgo) es significativamente más alto que el de un falso positivo. Ver figura 5.

Figura 5

Comparación de Modelos de Clasificación



A continuación, se presentan las métricas obtenidas tras la evaluación con el conjunto de prueba (30% de la muestra):

Tabla 4

Métricas Obtenidas

Modelo	Accuracy	Recall (Clase 1)	F1-Score	AUC-ROC
Regresión Logística	0,671	0,526	0,559	0,679
Random Forest	0,625	0,452	0,489	0,624
XGBoost	0,600	0,442	0,467	0,575

Estos resultados obtenidos deben interpretarse como un desempeño metodológicamente coherente con la complejidad del fenómeno estudiado. El intento suicida es un evento multifactorial, dinámico y de difícil predicción, por lo que incluso en la literatura internacional los modelos de machine learning presentan desempeños variables. En una revisión y metaanálisis sobre modelos predictivos de ideación, intento y muerte por suicidio, Kusuma et al. (2022) reportaron un AUC agrupado de 0,86, sensibilidad de 0,66 y especificidad de 0,87. En este sentido, la diferencia obtenida frente a estos resultados puede explicarse por las características de la ENSM 2015, que corresponde a una encuesta transversal basada en autorreportes y no a un sistema de seguimiento clínico longitudinal.

En consecuencia, el modelo debe entenderse como un prototipo analítico de **apoyo** exploratorio, capaz de evidenciar asociaciones relevantes entre variables de salud mental e intento suicida, pero con limitaciones para su uso operativo inmediato. Para mejorar su desempeño futuro sería necesario incorporar datos más recientes, variables longitudinales, información clínica detallada, antecedentes de atención en salud mental, ajuste de umbrales de

clasificación, validación externa y técnicas de inteligencia artificial explicable que fortalezcan la interpretación de los resultados.

Modelo con Mejor Desempeño

la Regresión Logística presentó el mejor desempeño global, con un Accuracy de 0,671, Recall de 0,526, F1-Score de 0,559 y AUC-ROC de 0,679. Aunque estos valores no corresponden a un modelo de alta capacidad predictiva, sí muestran que las variables seleccionadas contienen señales asociadas al intento suicida y permiten una clasificación superior al azar. Este resultado es coherente con la dificultad inherente de predecir eventos relacionados con la conducta suicida, los cuales dependen de múltiples factores clínicos, sociales, económicos, familiares y contextuales.

Capacidad Predictiva (Recall)

La Regresión Logística presentó el valor más alto de Recall para la clase positiva, con un resultado de 0,526, lo que indica que el modelo logró identificar correctamente el 52,6 % de los individuos con antecedente de intento suicida. Aunque este desempeño es modesto, resulta relevante porque fue obtenido sobre un conjunto de prueba sin balanceo artificial, lo que permite una estimación más realista de la capacidad de generalización del modelo. En salud mental, el error de mayor costo es el falso negativo: no detectar a una persona en riesgo real puede derivar en la omisión de una intervención oportuna, con consecuencias irreversibles. Un falso positivo, en cambio, genera una evaluación clínica adicional pero no produce daño directo. Por esta razón, el Recall se define como la métrica de referencia primaria, criterio consistente con la literatura especializada (Tang et al., 2024; Ntakolia et al., 2025).

Equilibrio General (F1-Score)

La Regresión Logística obtuvo el mejor F1-Score entre los modelos evaluados, con un valor de 0,559. Este resultado muestra que el modelo logró el mejor equilibrio relativo entre precisión y sensibilidad, aunque todavía presenta limitaciones importantes para una aplicación operativa. El F1-Score confirma que el modelo tiene cierta capacidad para clasificar casos positivos, pero también evidencia que existe margen de mejora, especialmente en la reducción de falsos negativos y en el fortalecimiento de la precisión de las alertas generadas.

Interpretabilidad y Estabilidad del Modelo

A diferencia de lo esperado inicialmente, los modelos más complejos, como Random Forest y XGBoost, no superaron a la Regresión Logística en el conjunto de prueba. Esto sugiere que, con el tamaño muestral disponible y las variables seleccionadas, un modelo lineal e interpretable ofreció mayor estabilidad y mejor capacidad de generalización. En un contexto sensible como la salud mental, esta característica representa una ventaja metodológica, ya que permite comprender de forma más clara la relación entre los factores incluidos y la probabilidad estimada de intento suicida.

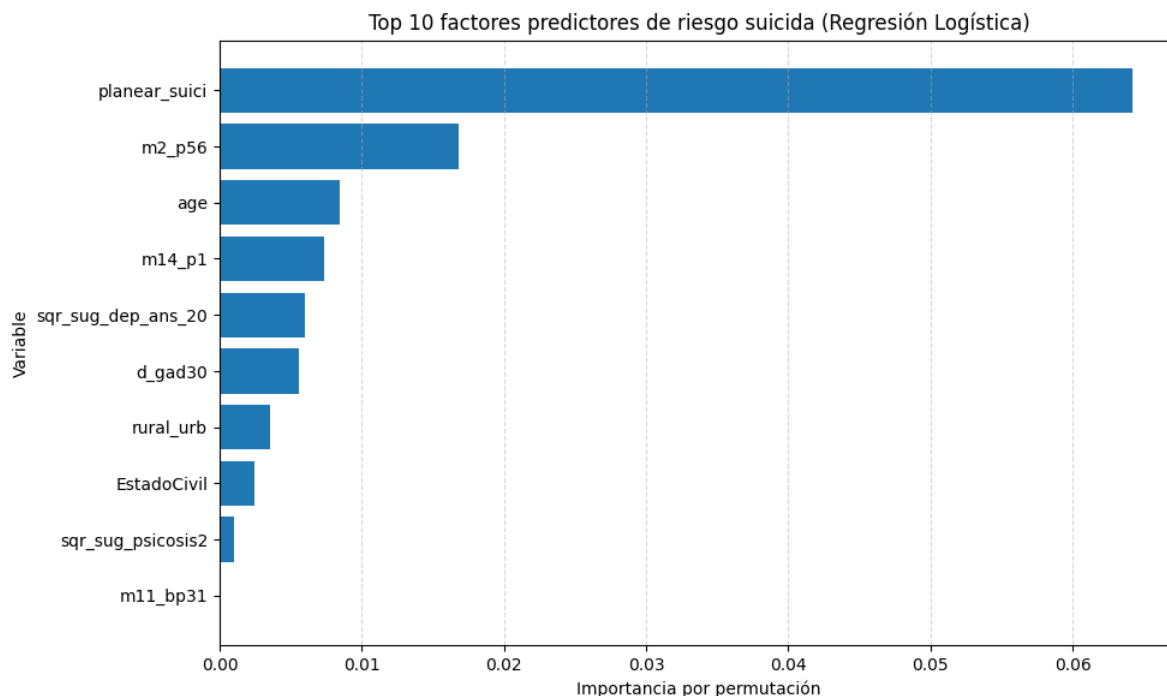
Capacidad Discriminativa (AUC-ROC)

La Regresión Logística también presentó el mayor AUC-ROC, con un valor de 0,679, lo que indica una capacidad discriminativa moderada para diferenciar entre personas con y sin antecedente de intento suicida. Este valor se encuentra por encima del azar, pero no alcanza un nivel suficiente para considerar el modelo como una herramienta diagnóstica o de alerta temprana autónoma. En consecuencia, los resultados deben interpretarse como evidencia exploratoria de asociaciones relevantes en la ENSM 2015 y como una base metodológica para

futuros modelos que incorporen datos más recientes, variables longitudinales, información clínica detallada y validación externa.

Figura 6

Top 10 Factores Predictores de Riesgo Suicida



La importancia relativa de las variables predictoras reportada en la Figura 6 fue estimada mediante importancia por permutación sobre el modelo principal, Regresión Logística. Esta técnica evalúa cuánto disminuye el desempeño del modelo cuando se altera aleatoriamente cada variable, por lo que permite identificar cuáles predictores aportan mayor capacidad discriminativa al clasificador.

La variable con mayor importancia fue `planear_suici`, lo cual es consistente con la literatura que reconoce la ideación y planificación suicida como antecedentes próximos del intento. Sin embargo, esta variable debe interpretarse con especial cautela, debido a su cercanía conceptual con el desenlace. No representa un factor externo independiente, sino una señal clínica dentro del espectro de la conducta suicida.

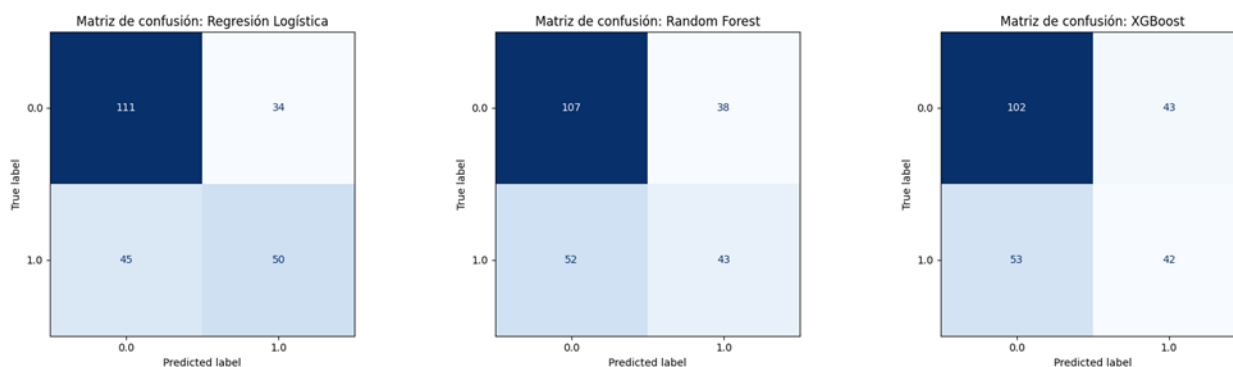
También aparecen variables sociales, sociodemográficas y clínicas como m2_p56, age, m14_p1, sqr_sug_dep_ans_20, d_gad30, rural_urb, EstadoCivil y sqr_sug_psicosis2. Este patrón refuerza la interpretación multidimensional del fenómeno: el poder predictivo no depende de una sola variable, sino de la combinación de señales relacionadas con salud mental, trayectoria vital, contexto social y condiciones personales.

Con el propósito de evaluar el desempeño de los modelos predictivos desarrollados, se presentan a continuación diversas representaciones gráficas que permiten analizar de manera integral su comportamiento. El uso de estas herramientas gráficas facilita la interpretación de los resultados, permitiendo identificar fortalezas, limitaciones y la capacidad de los modelos para discriminar entre individuos con y sin riesgo de intento suicida.

Matriz de Confusión

Figura 7

Matriz de Confusión



La matriz de confusión del modelo principal, Regresión Logística, permite observar con mayor detalle los aciertos y errores de clasificación. Sobre el conjunto de prueba sin balanceo

artificial, el modelo clasificó correctamente 111 casos de la clase negativa y 50 casos de la clase positiva. A su vez, registró 34 falsos positivos y 45 falsos negativos.

Estos resultados muestran que el modelo logra reconocer una proporción relevante de los individuos sin antecedente de intento suicida, pero tiene una capacidad más limitada para detectar los casos positivos. En términos prácticos, identificó correctamente 50 de los 95 casos positivos presentes en el conjunto de prueba, lo que equivale a un Recall de 0,526.

Desde la perspectiva de salud pública, los 45 falsos negativos constituyen la principal limitación del modelo, dado que representan personas con antecedente de intento suicida que no fueron clasificadas como positivas. Por esta razón, el modelo no puede considerarse una herramienta suficiente para la toma de decisiones clínicas, sino un insumo exploratorio que requeriría validación, ajuste de umbrales y acompañamiento profesional.

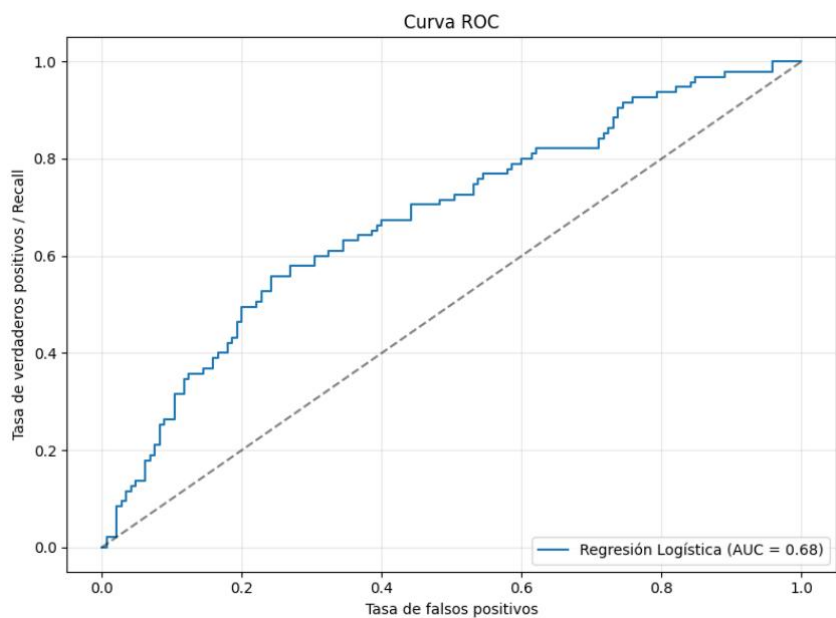
La presencia de 34 falsos positivos también debe ser considerada. Aunque este tipo de error puede generar evaluaciones adicionales innecesarias, en contextos de prevención del suicidio suele ser menos crítico que omitir casos reales. Sin embargo, un sistema de alerta institucional debería equilibrar sensibilidad y precisión para evitar sobrecargar los servicios de salud mental.

En conjunto, la matriz de confusión confirma que la Regresión Logística ofrece una capacidad funcional pero limitada para clasificar casos asociados al intento suicida. Su utilidad se ubica principalmente en el plano exploratorio y metodológico, como base para futuras versiones con más datos, mejores variables y validación externa.

Curva ROC

Figura 8

Curva ROC



La curva ROC (Receiver Operating Characteristic) presentada en la Figura 8 permite evaluar la capacidad discriminativa del modelo Regresión Logística para diferenciar entre individuos con y sin riesgo de intento suicida bajo distintos umbrales de clasificación.

En la gráfica, el eje vertical representa la tasa de verdaderos positivos (TPR o Recall), mientras que el eje horizontal muestra la tasa de falsos positivos (FPR). La línea diagonal punteada corresponde al comportamiento de un clasificador aleatorio, cuyo desempeño equivale a una probabilidad del 50% de acertar. Por tanto, mientras más alejada se encuentre la curva ROC de esta diagonal y más cercana esté al extremo superior izquierdo, mejor será la capacidad del modelo para discriminar correctamente entre ambas clases.

El modelo de Regresión Logística alcanzó un valor AUC-ROC de 0,68. En términos prácticos, esto significa que existe una probabilidad del 68% de que el modelo asigne una puntuación de riesgo mayor a un individuo con antecedente de intento suicida que a uno seleccionado al azar sin dicho antecedente. Si bien este valor se sitúa por debajo de modelos no lineales más complejos, representa un desempeño considerablemente superior al azar, confirmando que el algoritmo logra capturar patrones significativos en las variables sociodemográficas, clínicas y económicas analizadas.

La trayectoria de la curva muestra que el modelo permite incrementar la sensibilidad sin un crecimiento excesivamente abrupto de los falsos positivos en los tramos iniciales. En el contexto de la salud mental, esta característica es relevante, ya que se suele priorizar la detección temprana de casos vulnerables (minimizar falsos negativos), asumiendo un margen controlado de falsas alertas para asegurar la cobertura preventiva.

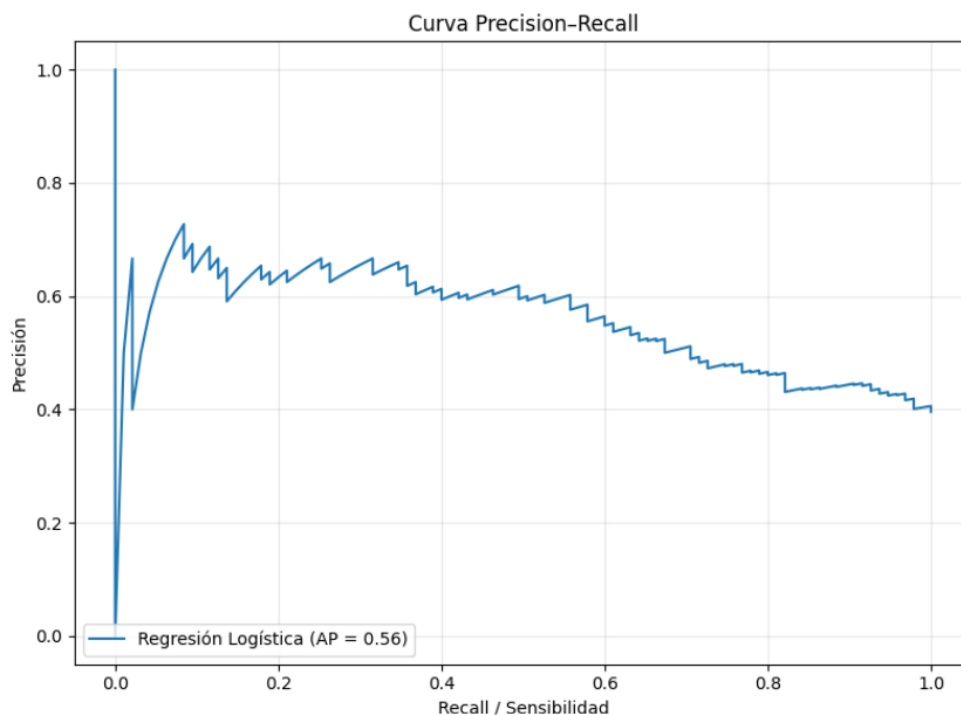
Es importante destacar que el intento suicida es un fenómeno multifactorial y dinámico, cuya predicción mediante datos transversales de encuestas poblacionales representa un desafío técnico elevado. Por ello, un AUC de 0,68 se considera un resultado funcional y consistente con la literatura científica en psicología y psiquiatría forense, donde la predictibilidad suele verse limitada por la naturaleza subjetiva y situacional de la conducta humana.

En conclusión, la curva ROC demuestra que la Regresión Logística posee una capacidad discriminativa útil para apoyar procesos de tamizaje y priorización de casos. No obstante, dado que el modelo no es perfecto, debe interpretarse como una herramienta complementaria de apoyo a la decisión clínica y no como un mecanismo de diagnóstico definitivo, reconociendo la complejidad inherente al fenómeno del riesgo suicida.

Curva Precision – Recall

Figura 9

Curva Precision – Recall



La curva Precision–Recall presentada en la Figura 9 permite evaluar el desempeño del modelo de Regresión Logística enfocándose específicamente en la detección de la clase positiva (individuos con riesgo de intento suicida). Esta métrica es fundamental en contextos de salud pública donde existe un desbalance de clases y el interés primordial radica en la identificación precisa de los casos de riesgo.

En esta representación, el eje horizontal muestra el Recall (sensibilidad), que mide la capacidad del modelo para capturar la mayor cantidad posible de casos positivos reales. El eje vertical representa la Precisión, que indica qué proporción de las alertas generadas por el modelo corresponden efectivamente a individuos con antecedentes de intento suicida.

El modelo de Regresión Logística obtuvo un Average Precision (AP) de 0,56. Este valor resume el comportamiento global de la curva y refleja un desempeño moderado pero informativo. La gráfica muestra que el modelo logra mantener niveles de precisión aceptables en rangos de Recall bajos; sin embargo, se observa una fluctuación inicial y un descenso progresivo de la precisión a medida que se intenta capturar una mayor proporción de la población en riesgo.

Este comportamiento evidencia el desafío intrínseco de la predicción en salud mental: para lograr una alta sensibilidad (detectar a la mayoría de las personas en riesgo), el modelo tiende a generar un número considerable de falsos positivos. En la práctica clínica, este intercambio suele considerarse aceptable, bajo la premisa de que el costo humano de omitir un caso de riesgo real (falso negativo) supera el costo operativo de realizar evaluaciones de seguimiento a individuos identificados preventivamente.

La curva Precision–Recall complementa los hallazgos de la curva ROC. Mientras que la ROC mide la capacidad de separación general entre clases, el AP de 0,56 ofrece una visión más conservadora y realista sobre la eficacia del modelo en la identificación directa de la clase vulnerable. Los resultados sugieren que, aunque la Regresión Logística captura tendencias significativas, la naturaleza multifactorial y la baja prevalencia relativa del intento suicida imponen un límite a la precisión que un modelo lineal puede alcanzar.

En conjunto, estos resultados confirman que el modelo posee una capacidad funcional para el tamizaje primario, permitiendo priorizar grupos poblacionales para intervenciones preventivas. No obstante, el nivel de precisión alcanzado refuerza la necesidad de utilizar esta herramienta como un sistema de apoyo a la decisión y no como un criterio diagnóstico único, debiendo ser integrada dentro de una evaluación clínica integral y multidimensional.

Conclusiones

La consolidación de la Encuesta Nacional de Salud Mental 2015 como fuente de datos demostró ser viable y metodológicamente justificada. La integración de los módulos de adolescentes y adultos permitió construir un dataset de 799 registros con 21 predictores seleccionados, decisión que fue necesaria ante el bajo número de casos positivos en el segmento adolescente ($n = 88$) y que amplió el alcance del análisis a personas de 12 a 92 años. Este punto de partida es fundamental para interpretar los hallazgos siguientes: los resultados son representativos de la población colombiana general captada por la ENSM 2015, no de un grupo etario específico.

Sobre esa base, el análisis exploratorio y estadístico confirmó la naturaleza multidimensional del riesgo de intento suicida. La importancia por permutación del modelo principal mostró que `planear_suici`, asociada a la planificación suicida previa, fue la variable con mayor aporte discriminativo. Este resultado debe interpretarse con cautela, debido a su cercanía conceptual con el intento suicida. También se identificaron variables sociales y sociodemográficas como `m2_p56` (participación en grupos sociales, religiosos o deportivos), `age` (edad), `rural_urb` (área de residencia) y `EstadoCivil` (situación sentimental o legal); así como variables clínicas y de salud mental como `sqr_sug_dep_ans_20` (tamizaje de síntomas depresivos y ansiosos), `d_gad30` (ansiedad generalizada en los últimos 30 días), `sqr_sug_psicosis2` (indicador sugestivo de síntomas psicóticos) y `m14_p1` (frecuencia de consumo de bebidas alcohólicas).

Estos hallazgos no implican causalidad; corresponden a factores asociados con utilidad discriminativa para la clasificación predictiva y deben interpretarse dentro del contexto específico de la ENSM 2015.

La comparación de modelos evidenció que la Regresión Logística fue el algoritmo con mejor desempeño global bajo el flujo metodológico corregido. Este modelo obtuvo Accuracy de 0,671, Recall de 0,526, F1-Score de 0,559 y AUC-ROC de 0,679, superando a Random Forest y XGBoost en todas las métricas evaluadas. Este resultado sugiere que, con el tamaño muestral disponible y las variables seleccionadas, un modelo lineal e interpretable generalizó mejor que modelos de mayor complejidad.

Los resultados demuestran la viabilidad técnica de desarrollar un prototipo predictivo del intento suicida a partir de encuestas nacionales de salud mental, pero también evidencian límites importantes. Con una sensibilidad de 52,6%, la Regresión Logística logró identificar una proporción relevante de casos positivos, aunque dejó sin detectar casi la mitad de ellos. Por tanto, el modelo debe entenderse como una herramienta exploratoria de apoyo al análisis y no como un sistema clínico autónomo de alerta temprana.

Finalmente, es necesario reconocer que la utilidad de estos hallazgos está acotada por limitaciones que el estudio asume con transparencia: los datos son transversales y de 2015, la variable objetivo depende de autodeclaración con riesgo de subregistro, la variable `planear_suici` presenta proximidad conceptual con el desenlace y el sobremuestreo se realizó con SMOTE sobre variables categóricas codificadas numéricamente. No obstante, estas restricciones no invalidan el aporte del estudio: demostrar que los microdatos de encuestas nacionales de salud mental pueden procesarse y modelarse de manera rigurosa constituye una contribución metodológica concreta para futuras investigaciones y para la toma de decisiones basada en evidencia.

Recomendaciones

Los hallazgos del estudio generan recomendaciones en distintos niveles que conforman una ruta progresiva desde el fortalecimiento metodológico hasta la implementación institucional responsable.

En el plano metodológico, la prioridad inmediata es fortalecer la interpretabilidad y la evaluación del modelo. Dado que la Regresión Logística fue el algoritmo con mejor desempeño, se recomienda complementar su análisis con coeficientes estandarizados, importancia por permutación y herramientas de inteligencia artificial explicable como SHAP o LIME. Esto permitiría comprender con mayor claridad qué variables contribuyen a cada alerta y evitar interpretaciones simplistas de los factores asociados al intento suicida.

En el plano de la implementación, los resultados no justifican todavía la adopción de un sistema autónomo de alerta temprana. La Regresión Logística puede utilizarse como base exploratoria por su transparencia e interpretabilidad, pero cualquier piloto institucional debería estar acompañado de protocolos de confirmación clínica, supervisión profesional y evaluación continua de falsos negativos y falsos positivos.

A nivel de política pública, los hallazgos refuerzan la necesidad de fortalecer la recolección de información sobre planificación suicida, experiencias traumáticas, síntomas emocionales, consumo, condiciones sociales y acceso a servicios de salud mental. La futura disponibilidad de los microdatos de la nueva ENSM permitiría recalibrar el modelo, evaluar la estabilidad temporal de los predictores y mejorar la capacidad de generalización.

Como línea de mejora futura, se recomienda incorporar datos longitudinales, historia clínica, antecedentes de atención en salud mental, eventos recientes, redes de apoyo, ajuste sistemático de hiperparámetros, comparación con modelos usando otras técnicas de balanceo.

Estas acciones podrían mejorar la capacidad predictiva y reducir la tasa de falsos negativos en la detección de intentos suicidas.

Finalmente, cualquier implementación del modelo debe estar enmarcada en un protocolo ético que garantice la confidencialidad de los datos, el consentimiento informado y la supervisión institucional del sistema en producción. Las alertas generadas no pueden utilizarse con fines discriminatorios ni como sustituto de la evaluación clínica, y dado que un modelo entrenado con datos de 2015 degradará su rendimiento con el tiempo, establecer mecanismos de monitoreo continuo es una condición de responsabilidad para cualquier institución que decida adoptarlo.

Referencias Bibliográficas

- Aparicio Castillo, Y. A., Blandón Rodríguez, A. M., & Chaves Torres, N.-M. (2020). High prevalence of two or more suicide attempts associated with suicidal ideation and mental disease in Colombia 2016. *Revista Colombiana de Psiquiatría (English Ed.)*, 49(2), 95–100. <https://doi.org/10.1016/j.rcpeng.2018.07.002>
- Artieda Urrutia, P. (2017). *Características sociodemográficas, clínicas y psicométricas de una población de individuos que intentan el suicidio*. <https://repositorio.uam.es/handle/10486/680469>
- Atmakuru, A., Shahini, A., Chakraborty, S., Seoni, S., Salvi, M., Hafeez-Baig, A., Rashid, S., Tan, R. S., Barua, P. D., Molinari, F., & Acharya, U. R. (2025). Artificial intelligence-based suicide prevention and prediction: A systematic review (2019–2023). *Information Fusion*, 114, 102673. <https://doi.org/10.1016/J.INFFUS.2024.102673>
- Betancur Díaz, M. A., Garzón Ladino, A. L., Vallejo Silva, A., & Murillo Pinto, M. (2023). Factors Associated with Suicidal Intent in Economically Active Patients Diagnosed with Major Depression | Factores asociados con intento suicida en pacientes económicamente activos con diagnóstico de depresión mayor. *Revista Colombiana De Psiquiatria*, 52(2), 87–92. <https://doi.org/10.1016/j.rcp.2021.01.003>
- Boletín-conducta-suicida. (s. f.). Recuperado 26 de abril de 2026, de <https://www.minsalud.gov.co/sites/rid/Lists/BibliotecaDigital/RIDE/VS/PP/ENT/boletin-conducta-suicida.pdf>
- Buriticá, I. T., Arango, M. del M., Vélez, I., Estrada, S., Sierra-Hincapié, G., & Restrepo, D. (2023). Factores asociados con el comportamiento suicida en adolescentes. Estudio

poblacional. *Revista Colombiana de Psiquiatría*, 52, S36–S43.

<https://doi.org/10.1016/J.RCP.2021.08.005>

Castrillon, S. O., Marín, L. M. G., Villegas, H. H. J., & Escobar, C. C. P. (2022). Machine learning aplicado en la clasificación y predicción de la depresión: Una revisión sistemática. *RISTI - Revista Iberica de Sistemas e Tecnologias de Informacao*, (Special Issue 47), 363–375. <https://investigaciones-pure.udemedellin.edu.co/es/publications/machine-learning-aplicado-en-la-clasificaci%C3%B3n-y-predicci%C3%B3n-de-la-/>

Castro Moreno, L. S., Fuertes Valencia, L. F., Pacheco García, O. E., & Muñoz Lozada, C. M. (2023). Risk factors associated with suicide attempt as predictors of suicide, Colombia, 2016–2017. *Revista Colombiana de Psiquiatría (English Ed.)*, 52(3), 176–184.

<https://doi.org/10.1016/j.rcpeng.2021.03.005>

Constitución Política 1 de 1991 Asamblea Nacional Constituyente—Gestor Normativo. (s. f.).

Recuperado 26 de abril de 2026, de

<https://www.funcionpublica.gov.co/eva/gestornormativo/norma.php?i=4125>

Corona-Miranda, B., Alfonso-Sagué, K., Hernández-Sánchez, M., & Lomba-Acevedo, P. (2014). *Artículo Original*.

Cuello Villamil, A., Camelo Ciro, A., & Castro Molinares, S. (2023). Caracterización del intento de suicidio en un municipio de Colombia durante tres años. *Diversitas*, 19(2), 85–106.

<https://doi.org/10.15332/22563067.9371>

DANE. (2024). *Boletín técnico - Estadísticas Vitales (EEVV)*.

Estadísticas Vitales (EEVV). (2023, March).

https://www.dane.gov.co/files/investigaciones/poblacion/pre_estadisticasvitales_IVtrim_2022pr.pdf

- Instituto Nacional de Salud. (2025). *Boletín Epidemiológico Semanal. Semana epidemiológica 36: Intento de suicidio.*
- Kusuma, K., Larsen, M., Quiroz, J. C., Gillies, M., Burnett, A., Qian, J., & Torok, M. (2022). The performance of machine learning models in predicting suicidal ideation, attempts, and deaths: A meta-analysis and systematic review. *Journal of Psychiatric Research, 155*, 579–588. <https://doi.org/10.1016/J.JPSYCHIRES.2022.09.050>
- Kusuma, K., Larsen, M., Quiroz, J. C., & Torok, M. (2024). Age-stratified predictions of suicide attempts using machine learning in middle and late adolescence. *Journal of Affective Disorders, 365*, 126–133. <https://doi.org/10.1016/J.JAD.2024.08.043>
- Ley 1616 de 2013 Congreso de la República de Colombia. (s. f.). Recuperado 26 de abril de 2026, de <https://www.alcaldiabogota.gov.co/sisjur/normas/Norma1.jsp?i=51292>
- Lyu, J., Wang, C., Gao, Z., & Zhang, J. (2025). The concise machine learning prediction models for suicide attempt in China: Based on demographic and social factors. *Journal of Affective Disorders, 388*, 119634. <https://doi.org/10.1016/J.JAD.2025.119634>
- Ministerio de Salud y Protección Social. (2025, July 8). *El Ministerio de Salud y Protección Social da inicio a la quinta Encuesta Nacional de Salud Mental en los 32 departamentos del país.* <https://www.minsalud.gov.co/CC/Noticias/2025/Paginas/inicio-quinta-encuesta-nacional-de-salud-mental.aspx>.
<https://www.minsalud.gov.co/CC/Noticias/2025/Paginas/inicio-quinta-encuesta-nacional-de-salud-mental.aspx>
- MinSalud. (2018). *Boletín de salud mental Conducta suicida Subdirección de Enfermedades No Transmisibles.*
- Nacional, E. (2021). *Prevención de la Conducta Suicida en Colombia.*

- Ntakolia, C., Yotsidi, V., Rannou, I., & Gournellis, R. (2025). Interpretable Machine Learning approach for predicting clinically significant suicide risk: A case study of patients with major depressive disorder in Greece. *Psychiatry Research*, 351, 116607.
<https://doi.org/10.1016/J.PSYCHRES.2025.116607>
- Pérez, G. G. (2014). *ARTÍCULO 2o. NATURALEZA Y CONTENIDO DEL DERECHO FUNDAMENTAL A.*
- Prevención del suicidio - OPS/OMS / Organización Panamericana de la Salud. (2021).
https://www.paho.org/en/topics/suicide-prevention?page=3&utm_source=PANTHEON_STRIPPED
- PROCURADURIA GENERAL DE LA NACION. (2023). *SEGUIMIENTO A ACCIONES EN SALUD MENTAL POR PARTE DE LAS ENTIDADES TERRITORIALES CERTIFICADAS EN EDUCACIÓN – ETC.*
<https://www.procuraduria.gov.co/Documents/2025/Julio/Informe%20ejecutivo%20salud%20mentalcolegios%20-%202023.pdf>
- Tang, H., Miri Rekavandi, A., Rooprai, D., Dwivedi, G., Sanfilippo, F. M., Boussaid, F., & Bennamoun, M. (2024). Analysis and evaluation of explainable artificial intelligence on suicide risk assessment. *Scientific Reports*, 14(1), 1–13. <https://doi.org/10.1038/S41598-024-53426-0;SUBJMETA>
- Varma, S., & Batterham, P. J. (2025). Self-recognition and formal help-seeking for mental health conditions: A cross-sectional study. *Journal of Affective Disorders*, 388, 119612.
<https://doi.org/10.1016/J.JAD.2025.119612>
- WHO. (2014). *Preventing suicide A global imperative A global imperative.*
www.who.int

WHO. (2025, March 25). *Suicide*. <https://www.who.int/news-room/fact-sheets/detail/suicide>

Wirth, R., & Hipp, J. (s. f.). *CRISP-DM: Towards a Standard Process Model for Data Mining*.

Apéndices

Apéndice A

Video Sustentación

<https://youtu.be/CQOI-xWb92k>