Facultad de Ingeniería y Ciencias
Maestría en Ciencia de Datos

MODELO PREDICTIVO PARA LA DETECCIÓN TEMPRANA DEL RIESGO DE SUICIDIO EN POLICÍAS DE COLOMBIA

Proyecto Aplicado para optar al título de Magíster en Ciencia de Datos

Autores
Juan Sebastián Torres Romero
Mario Fernando Barbosa Rodríguez
Valentina Caicedo Henríquez

Director
David Arango Londoño

Santiago de Cali, 2026

FICHA RESUMEN

PROYECTO APLICADO FINAL — TRABAJO DE GRADO

TÍTULO DEL PROYECTO: Modelo Predictivo para la Detección Temprana del Riesgo de Suicidio en Policías de Colombia

ÁREA DE TRABAJO: Sector Salud — Salud Mental Institucional

TIPO DE PROYECTO: Aplicado

ESTUDIANTE(S): Juan Sebastián Torres Romero Mario Fernando Barbosa Rodríguez Valentina Caicedo Henríquez

CORREO ELECTRÓNICO: juanstorres@javerianacali.edu.co mariofbr@javerianacali.edu.co valentinacaicedo@javerianacali.edu.co

DIRECCIÓN Y TELÉFONO: Calle 44 # 50 - 51 CAN, Bogotá, Colombia. 01 8000 911143

DIRECTOR: David Arango Londoño

VINCULACIÓN DEL DIRECTOR: Docente Catedrático

CORREO ELECTRÓNICO DEL DIRECTOR: david.arango@javerianacali.edu.co

CO-DIRECTOR (Si aplica): No aplica

GRUPO O EMPRESA QUE LO AVALA: Dirección de Sanidad de la Policía Nacional — DISAN PONAL

OTROS GRUPOS O EMPRESAS: No aplica

PALABRAS CLAVE: suicidio, Policía Nacional de Colombia, salud mental, modelos predictivos, ciencia de datos, LightGBM, SHAP, diseño caso-control

FECHA DE INICIO: 1 de mayo de 2025

DURACIÓN ESTIMADA: 12 meses

RESUMEN:

El suicidio en la Policía Nacional de Colombia representa un problema crítico de salud pública institucional: en 2011 la tasa alcanzó 18 por cada 100.000 efectivos, triplicando la tasa nacional. Este proyecto aplicado desarrolló y auditó un modelo predictivo para la detección temprana del riesgo de suicidio en policías activos de Colombia, utilizando una base de datos institucional inédita de la Dirección de Sanidad de la Policía Nacional (DISAN PONAL) con 247 casos confirmados, bajo un diseño caso-control retrospectivo con ventana de observación

de 24 meses y horizonte de predicción de 90 días. Se construyeron 42 variables predictoras a partir de fuentes clínicas, laborales y sociodemográficas. Se evaluaron cuatro algoritmos —Regresión Logística, Random Forest, XGBoost y LightGBM— bajo escenarios de desbalance 5:1 y 10:1. El modelo LightGBM en escenario 5:1 obtuvo el mejor desempeño en validación cruzada: AUC-PR = 0.9587, AUC-ROC = 0.9888, F1 = 0.9095, Precisión = 0.919 y Recall = 0.902. Una auditoría de *data leakage* de cuatro capas redujo el AUC-PR de 1.000 a 0.9587, reduciendo las fuentes identificables de contaminación metodológica y disminuyendo sustancialmente el riesgo de leakage residual. La validación temporal (hold-out 2022–2025) reveló una degradación significativa en la métrica primaria (AUC-PR: 0.9587 → 0.6553, −32%), indicando que el modelo requiere recalibración antes del despliegue operativo y debe entenderse como insumo analítico exploratorio, sujeto a validación prospectiva y estratificación previa de riesgo. El análisis SHAP identificó como predictores dominantes los indicadores de utilización general del sistema de salud —tasa de diagnósticos acumulados, polimedicación, antigüedad institucional y edad— no exclusivamente de salud mental formal. Se calculó el PPV esperado bajo prevalencia real, mostrando que el despliegue directo sobre el universo de efectivos no es viable sin estratificación previa de una subcohorte de riesgo. El análisis de equidad por subgrupos reveló brechas de Recall entre hombres y mujeres y entre policías de alta y baja antigüedad, documentadas como limitaciones. El trabajo aporta un protocolo replicable de auditoría de leakage para diseños caso-control retrospectivos en contextos de salud institucional y un conjunto de seis indicadores de riesgo con potencial de integración en programas preventivos de la DISAN.

TABLA DE CONTENIDO

LISTA DE TABLAS

LISTA DE FIGURAS

1. INTRODUCCIÓN

El suicidio representa una de las principales causas de muerte prematura a nivel global. La Organización Mundial de la Salud estima que aproximadamente 720.000 personas mueren por suicidio cada año, siendo la tercera causa de defunción entre jóvenes de 15 a 29 años. [1] En Colombia, el Ministerio de Salud y la Procuraduría General de la Nación reportaron que durante 2023 la tasa de suicidio se incrementó en un 15.73% respecto al año anterior, con más de 30.000 intentos atendidos. [2]

Dentro de este panorama, los cuerpos de seguridad del Estado representan una población de especial vulnerabilidad. En 2011, la Dirección de Sanidad de la Policía Nacional registró 29 casos de suicidio entre uniformados activos, equivalente a 18 por cada 100.000 efectivos, [3] cifra que triplica la tasa nacional de ese período. Las condiciones operativas del trabajo policial —la exposición continua a situaciones de violencia, la cultura institucional de fortaleza y el estigma frente a la búsqueda de ayuda psicológica— configuran un entorno de riesgo diferencial que los programas preventivos tradicionales no han logrado reducir suficientemente.

En este contexto, se desarrolló un modelo predictivo para la detección temprana del riesgo de suicidio en policías activos de Colombia, utilizando una base de datos institucional inédita de la DISAN PONAL con 247 casos confirmados de suicidio. El modelo es una herramienta analítica auditada, no un sistema listo para despliegue operativo; sus resultados requieren validación prospectiva y recalibración antes de cualquier uso institucional. El proyecto aplicó un diseño caso-control retrospectivo con ventana de observación de 24 meses y horizonte de predicción de 90 días, integrando registros clínicos, administrativos y sociodemográficos mediante técnicas avanzadas de ciencia de datos.

El modelo LightGBM en escenario 5:1 obtuvo AUC-PR = 0.9587 y Recall = 0.902, validado mediante Stratified 5-Fold Cross-Validation. Una auditoría de *data leakage* de cuatro capas redujo sustancialmente el riesgo de contaminación metodológica, aunque ninguna auditoría puede garantizar ausencia absoluta de leakage no identificado. El análisis SHAP reveló que los predictores dominantes son indicadores de utilización general del sistema de salud, no

exclusivamente de salud mental formal —hallazgo con implicaciones directas para el diseño de sistemas de monitoreo preventivo institucional.

El documento está organizado en seis secciones. La Introducción presenta la contextualización, definición del problema, objetivos y marco de referencia. El Capítulo 2 presenta la identificación de variables predictoras y el análisis exploratorio (Objetivo Específico 1). El Capítulo 3 evalúa los escenarios de muestreo y balanceo de clases (Objetivo Específico 2). El Capítulo 4 desarrolla la evaluación y comparación de modelos supervisados (Objetivo Específico 3). El Capítulo 5 presenta las conclusiones y trabajos futuros. Finalmente se incluyen las referencias bibliográficas y los anexos.

1.1. Contextualización del Proyecto

La Policía Nacional de Colombia es la institución encargada de mantener el orden público y la seguridad ciudadana en el territorio nacional, con aproximadamente 180.000 efectivos activos. Sus miembros enfrentan condiciones laborales extremas: exposición crónica a eventos traumáticos, jornadas extendidas, trabajo en zonas de alto riesgo, disponibilidad permanente de armas de fuego y una cultura institucional que históricamente ha estigmatizado la búsqueda de ayuda psicológica como señal de debilidad. Esta constelación de factores configura un entorno de riesgo psicosocial que se refleja en tasas de suicidio significativamente superiores a la población general.

La Dirección de Sanidad de la Policía Nacional (DISAN PONAL) es la entidad encargada de gestionar la salud del personal policial. Cuenta con registros clínicos, administrativos y sociodemográficos de los miembros activos, organizados en sistemas de información que cubren diagnósticos, medicamentos, incapacidades, remisiones y consultas médicas. La disponibilidad de estos registros, junto con los datos de 247 casos confirmados de suicidio, representa una oportunidad única para aplicar técnicas de ciencia de datos a un problema de salud pública de alta urgencia institucional.

1.2. Definición del problema

1.2.1. Planteamiento del problema

El suicidio en la Policía Nacional de Colombia constituye un problema de salud pública institucional que trasciende las cifras individuales. Los datos de 2011 [3] tres veces superior a la tasa general nacional reportada por el Instituto Nacional de Medicina Legal (6.59 por 100.000 habitantes en 2023) [4] Este diferencial emerge de la confluencia de factores estresores crónicos propios del trabajo policial, barreras estructurales para el acceso a atención en salud mental y un ecosistema cultural que estigmatiza la vulnerabilidad emocional en el contexto de la seguridad pública.

Pese a los programas de bienestar psicológico implementados —líneas de apoyo emocional, intervención psicosocial 24/7 y atención psiquiátrica institucional— las cifras persisten y los datos actualizados no están disponibles públicamente, evidenciando una brecha crítica en los sistemas de monitoreo. Los mecanismos de detección existentes son principalmente reactivos: responden ante la crisis manifestada, no ante el riesgo latente. Esta limitación reduce significativamente la ventana de oportunidad para intervenciones preventivas efectivas. [5]

Desde la perspectiva de la ciencia de datos, los registros institucionales de salud contienen señales predictivas del riesgo suicida que no son visibles en la evaluación clínica individual pero que emergen del análisis de patrones de utilización de servicios, morbilidad acumulada y trayectorias laborales. La explotación sistemática de estos registros mediante algoritmos de aprendizaje supervisado ofrece la posibilidad de transformar el paradigma de intervención: de la respuesta reactiva a la identificación proactiva del riesgo

1.2.2. Formulación del problema

¿Es posible desarrollar un modelo predictivo, basado en técnicas de ciencia de datos, que utilizando variables clínicas, administrativas y sociodemográficas disponibles en la DISAN PONAL permita identificar de manera temprana —con un horizonte de 90 días— el riesgo de suicidio en policías activos de Colombia, con suficiente capacidad discriminativa, interpretabilidad y robustez metodológica para servir como insumo en los programas institucionales de prevención en salud mental?

1.2.3. Sistematización

- ¿Cuáles son las variables clínicas, administrativas y sociodemográficas con mayor poder predictivo del riesgo de suicidio en policías activos, y cuál es la dirección e intensidad de su efecto según el análisis SHAP?
- ¿Qué efecto tienen los diferentes escenarios de muestreo de controles (ratio 5:1 y 10:1) sobre el desempeño de los modelos predictivos, medido mediante AUC-PR como métrica primaria?
- ¿Cuál es el algoritmo de aprendizaje supervisado que ofrece el mejor balance entre capacidad discriminativa, interpretabilidad y robustez para la predicción temprana del riesgo de suicidio en esta población?
- ¿Cómo puede garantizarse la integridad metodológica del modelo en un diseño caso-control retrospectivo, identificando y corrigiendo las fuentes de *data leakage*?
- ¿Qué indicadores institucionales de riesgo pueden derivarse del análisis de explicabilidad para integrarse en un sistema de monitoreo preventivo?

1.3. Objetivos del proyecto

1.3.1. Objetivo general

Desarrollar un modelo predictivo que permita detectar de manera temprana el riesgo de suicidio en policías activos de Colombia y, mediante el uso de técnicas de explicabilidad, identificar y cuantificar los factores clínicos, administrativos y sociodemográficos que más contribuyen a dicho riesgo, a partir del análisis de una base de datos institucional de la Dirección de Sanidad de la Policía Nacional.

.

1.3.2. Objetivos específicos

- Identificar las variables clínicas, administrativas y sociodemográficas con mayor poder predictivo del riesgo de suicidio en miembros activos de la Policía Nacional, mediante análisis exploratorio de datos, técnicas de selección de atributos y métodos de explicabilidad como SHAP.
- Evaluar el efecto de diferentes escenarios de muestreo de controles (ratio 5:1 y 10:1) sobre el desempeño de los modelos predictivos en la detección de casos de alto riesgo de suicidio.
- Evaluar modelos de aprendizaje supervisado —Regresión Logística, Random Forest, XGBoost y LightGBM— con el fin de identificar el enfoque más eficaz para la predicción

temprana del riesgo de suicidio, utilizando el AUC-PR como métrica primaria de evaluación.

- Interpretar los resultados del modelo seleccionado mediante análisis SHAP, identificando los factores que elevan el riesgo de suicidio y generando insumos comprensibles para los responsables de programas institucionales de salud mental.
- Validar la robustez y capacidad de generalización del modelo predictivo mediante Stratified K-Fold Cross-Validation, validación temporal y análisis de permutación.
- Definir un conjunto de indicadores clave de riesgo, derivados del análisis de explicabilidad, susceptibles de integrarse en un sistema institucional de monitoreo para la prevención del suicidio.

1.4. Marco de referencia

1.4.1. Marco teórico

El suicidio como problema de salud pública

Según la RAE el Suicidio es un acto por el cual una persona pone fin a su existencia de manera voluntaria [6]. Este fenómeno representa una problemática en la salud pública a nivel global con consecuencias devastadoras para los individuos, familia y comunidades. Al año se suicidan cerca de 720.000 personas, lo que lo ubica como la tercera causa de defunción entre las jóvenes de 15 a 29 años a nivel global. Según la organización mundial de la salud, el hecho de vivir bajo guerras, desastres naturales, sufrir violencia, abusos o perdidas de un ser querido hace que la persona se sienta aislada y pueda conducir a conductas suicidas. [1] Investigaciones han encontrado que el 41% de las personas que se suicidaron fueron hospitalizadas por motivos psiquiátricos el año previo a cometer este acto. [7]. Según el último reporte del Instituto Nacional de Medicina Legal y Ciencias Forenses en Colombia, la tasa fue de 6,59 suicidios por cada 100.000 habitantes donde los hombres presentan una tasa más alta con respecto a las mujeres. [4]

Principales teorías sobre el suicidio

Las principales teorías que explican el suicidio ofrecen un marco para comprender cuáles son sus determinantes.

Teoría de Durkheim

Durkheim fue un sociólogo clave de la era moderna., cuyos trabajos sentaron la base para el desarrollo de teorías y metodologías sociológicas. A través de su enfoque positivista, busco estudiar fenómenos sociales con el mismo rigor que los fenómenos naturales. Entre sus principales contribuciones esta *"Le suicide"* publicado en 1897 [8] y donde sostiene que, el suicidio es una "*anomalía que es necesario rectificar, como indicio de una patología que requiere corrección*". Con esta perspectiva, Durkheim aspiraba a fomentar una ética positiva y constructiva que ayudara a "curar" patologías presentes en la vida política y la social [9] Según su teoría, existen 3 tipos de suicidios, el egoísta (resultado de una falta de integración social), altruista (derivado de una integración social excesiva) y anómico (vinculado a la falta de regulación social), todos relacionándolos con el grado de integración y control social. [9]

Modelo cognitivo de la conducta suicida de Beck y Wenzel

Por otro lado, la teoría del comportamiento suicida planteada por Beck y Wenzel, sostiene que los pensamientos negativos y la interpretación distorsionada de los eventos desempeñan un papel central en la aparición de respuestas emocionales y conductuales disfuncionales, lo que puede llevar al desarrollo de psicopatologías como la depresión y la ansiedad. [10] Sin embargo, el modelo cognitivo de la conducta suicida de estos autores va más allá de proponer constructos específicos que permiten distinguir entre individuos con riesgo suicida y aquellos que no lo presentan incluso en ausencia de antecedentes de depresión o ansiedad. Este modelo incluye factores de vulnerabilidad y disposición como la desesperanza, el aislamiento social o la baja tolerancia al dolor emocional, los cuales interactúan con eventos estresantes y activan cogniciones suicidas automáticas. [11]

Finalmente, la teoría psicológica interpersonal del suicidio plantea que quienes presentan un deseo de morir experimentan simultáneamente dos estados psicológicos durante un tiempo prolongado: Por un lado, la percepción de ser una carga para los demás; y por otro, una exposición sostenida d situaciones dolorosas que generan un proceso de habituación y desensibilicen. Este último aspecto se observa con mayor frecuencia en profesionales como

la medicina o el servicio militar, donde el contacto constante con el sufrimiento pude reducir el temor al dolor o a la muerte. [11]

Suicidio en personal de primera respuesta

El personal de primera respuesta como bomberos, policías, paramédicos, personal militar y trabajadores de emergencias médicas están frecuentemente expuestos a situaciones de alto estrés, trauma, violencia y muerte. Estas condiciones los convierten en una población vulnerable al desarrollo de trastornos de estrés postraumático (TEPT), la depresión y la ansiedad, los cuales se han identificado como factores de alto riesgo para la ideación y conducta suicida. [12]

Diferentes estudios han demostrado que los primeros respondientes presentan tasas significativamente altas de pensamientos suicidas, intentos y suicidios consumados en comparación con la población general. Por ejemplo, la organización Ruderman Family Foundation reporto que en Estados Unidos los suicidios entre bomberos y policías superan las muertes en cumplimiento del deber. [13] De manera similar, Colic (2014) informo que las muestres por suicidio superaron a las bajas en combate entre el personal militar estadounidense. Esta estadística subraya la mayor probabilidad de suicidio en veteranos en comparación con la población civil [14]. En el contexto colombiano, y considerando el historial del conflicto armado, a octubre del 2022 se habían reportado 30 suicidios entre ejército nacional y cerca de 12.900 miembros de las fuerzas militares presentaban trastornos mentales y de comportamiento [5].

La cultura institucional que rodea a estos cuerpos de emergencia, estigmatizada por demostrar salud mental, fortaleza emocional y falta de acceso a servicios psicológicos especializados, puede dificultar la búsqueda de ayuda, contribuyendo al deterioro emocional del individuo. [15] Diversos estudios y organismos internacionales han identificado múltiples factores de riesgos específicos que afectan al personal de primera respuesta en relación con la conducta suicida. Entre los que se destacan la exposición crónica a eventos traumáticos (accidentes, violencia, muertes), el sentimiento de culpa, impotencia o responsabilidad a ciertas situaciones y las largas jornadas laborales acompañadas de la privación del sueño. Además, se ha

documentado el impacto negativo del aislamiento social, la dificultad para expresar emociones y el fácil acceso a medios letales como las armas de fuego o medicamentos los cuales incrementan de manera significativa el riesgo de suicidio. [16] [17] [18]

En este contexto, es fundamental implementar programas de prevención de suicidio específicos para el personal de primera respuesta, incluyendo entrenamiento de primeros auxilios psicológicos, intervenciones psicoeducativas, des estigmatización de la salud mental, así como el seguimiento psicológico y el acompañamiento continuo para mitigar el riesgo de suicidio.

Enfoque interdisciplinario y uso de ciencia de datos

Las situaciones planteadas anteriormente, evidencian la necesidad de abordar el comportamiento suicida desde un enfoque interdisciplinario que integren áreas como la psicología, psiquiatría, salud pública, sociología y ciencia de datos. Este enfoque no pretende remplazar el juicio clínico ni proponer dirigir intervenciones directas, sino complementar la comprensión del riesgo suicida mediante herramientas analíticas basadas en datos.

El uso de la ciencia de datos permite analizar grandes volúmenes de información proveniente de registros clínicos, reportes administrativos, encuestas internas y características personales del personal —como la edad, genero, historial médico o antecedentes psicológicos— para identificar patrones asociados con una mayor probabilidad de riesgo suicida.

Esta aproximación puede facilitar el desarrollo de modelos predictivos que, sin invadir la privacidad ni estigmatizar, alerten a los responsables de salud institucional sobre casos que requieran seguimiento psicológico más cercano. Para que esto sea posible, es fundamental una adecuada gestión de datos, lo que implica procesos de limpieza, validación, estandarización y aseguramiento de calidad. De esta manera, los sistemas de datos puedan convertirse en aliados estratégicos en la prevención de suicidio, ofreciendo una visión complementaria que fortalezca la toma de decisiones en el entorno institucional.

Retos técnicos en la predicción del suicidio

Uno de los retos más grandes es que los casos de suicido son muchos menos frecuentes que los casos sonde no ocurre, lo que puede dificultar que los modelos detecten correctamente quienes si están en riesgo. Para enfrentar esto, se aplican estrategias que ayuden a equilibrar los datos y mejorar la precisión del modelo, como crear ejemplos adicionales (*Oversampling*) o el ajuste de la cantidad de datos por grupo (undersamplig).

Se probarán distintos tipos de modelos de predicción, desde aquellos que permiten interpretar claramente por que se toma una decisión (Regresiones o arboles de decisión), hasta modelos mas complejos que puedan ofrecer un mejor entendimiento como Random Forest o XGBoots. Estos modelos se evaluarán cuidadosamente para determinar cuál funciona mejor en este contexto, y garantizar que los resultados sean confiables a lo largo del tiempo.

Conocer los factores asociados al riesgo de suicidio en miembro de la Policía Nacional podrá aportar elementos útiles para fortalecer los programas institucionales de salud mental. Así mismo el análisis de estos patrones y características comunes en los casos identificados permitirá explorar la posibilidad de establecer señales de alerta o indicadores que contribuyan a una comprensión más profunda del problema y la formulación de estrategias de prevención.

1.4.2. Antecedentes

En la Tabla 1.1 se presentan los principales antecedentes que sustentan el enfoque metodológico y la pertinencia del proyecto

Tabla 1.1 Principales antecedentes

Referencia	Resumen del abordaje	Aportes al proyecto	Diferencias
“The utility of artificial intelligence in suicide risk prediction and the management of suicidal behaviors” [19]	Revisión general sobre cómo la IA y el ML permiten identificar patrones complejos en datos masivos para predecir el riesgo suicida y apoyar la gestión clínica. Incluye uso de NLP, sensores, redes sociales y big data.	Proporciona una base teórica sólida sobre el uso de IA en salud mental y predicción suicida, respaldando la pertinencia del enfoque técnico de tu proyecto.	No se enfoca en primeros respondientes ni en Colombia. No es un caso aplicado, sino una revisión conceptual y metodológica.

Referencia	Resumen del abordaje	Aportes al proyecto	Diferencias
“Predicting Risk of Suicide Attempts Over Time Through Machine Learning” [20]	Usaron ML sobre historias clínicas electrónicas para predecir intentos de suicidio hasta con 7 días de anticipación. Lograron alta precisión (AUC = 0.84). Estudiaron cómo cambian los factores de riesgo con el tiempo.	Aporta evidencia empírica sobre la capacidad de los modelos de ML para predecir intentos de suicidio con alta precisión y utilidad clínica	Estudio basado en población general hospitalaria en EE. UU. No se enfoca en primeros respondientes ni utiliza datos administrativos institucionales.
“A systematic review of suicidal thoughts and behaviors among police officers, firefighters, EMTs, and paramedics” [12]	Revisión de 63 estudios que documentan prevalencia de ideación, intentos y suicidios en personal de primera respuesta. Identifica factores de riesgo y protectores específicos para este grupo	Justifica con evidencia la elección de primeros respondientes como población de interés, visibilizando su vulnerabilidad frente al suicidio	No utiliza ciencia de datos ni modelos predictivos. Nuestro proyecto añade un enfoque tecnológico y preventivo en el análisis de riesgo.
“Factores asociados al intento de suicidio en la población colombiana” [7]	Estudio basado en la Encuesta Nacional de Salud Mental (n=21,988). Mediante regresión logística identifican factores como edad, depresión, ansiedad y abuso verbal asociados al intento de suicidio.	Aporta datos locales valiosos sobre factores de riesgo en Colombia, útiles para definir variables relevantes en el modelo predictivo	No se enfoca en personal institucional ni usa ciencia de datos. Es un estudio observacional con métodos estadísticos tradicionales.
“Predicting suicides after outpatient mental health visits in the Army STARRS study” [21]	Utiliza ML con datos administrativos y clínicos de soldados en EE. UU. para predecir suicidios luego de visitas ambulatorias. Identifican variables clave y validan un modelo aplicado en entorno militar.	Muestra que los modelos predictivos pueden implementarse exitosamente en contextos institucionales similares a fuerzas armadas o policía.	Se centra en militares estadounidenses post-atención clínica. Nuestra propuesta busca intervención temprana, en un contexto civil colombiano.

2. IDENTIFICACIÓN DE VARIABLES PREDICTORAS Y ANÁLISIS EXPLORATORIO DE DATOS

En este capítulo se presentan las actividades y resultados correspondientes al Objetivo Específico 1: identificar las variables clínicas, administrativas y sociodemográficas con mayor poder predictivo del riesgo de suicidio en miembros activos de la Policía Nacional, mediante análisis exploratorio de datos, selección de atributos y análisis SHAP. El trabajo se implementó

en los notebooks 00_auditoria_datos.ipynb, 01_construccion_dataset.ipynb y 02_feature_engineering.ipynb.

2.1. Diseño del estudio y fuente de datos

Se adoptó un diseño caso-control retrospectivo. Los casos (LABEL=1) son 247 miembros activos de la Policía Nacional que fallecieron por suicidio, confirmados por la DISAN PONAL mediante la base de datos de eventos del sistema de información hospitalaria. Este número corresponde a los casos con registros íntegros y fecha índice válida disponibles para el análisis; la base institucional de la DISAN contiene registros adicionales que fueron excluidos por criterios de integridad documental (ausencia de fecha índice verificable, duplicidad de identificador o registros sin cobertura mínima en el sistema). El criterio de inclusión definitivo fue la disponibilidad de al menos un registro en alguna de las fuentes transaccionales dentro de la ventana de observación de 24 meses. Los controles (LABEL=0) son policías activos seleccionados aleatoriamente sin evento de suicidio (semilla fija RANDOM_SEED=42). Se evaluaron dos escenarios: ratio 5:1 (1,235 controles, n total=1,482) y ratio 10:1 (2,470 controles, n total=2,717). La ventana de observación es de 24 meses previos al evento. El horizonte de predicción (*blackout*) de 90 días excluye toda actividad clínica de los tres meses previos al evento, garantizando que el modelo predice el riesgo con suficiente anticipación para intervención preventiva.

En la Figura 2.1 se ilustra la distribución de clases en el escenario principal de modelado (5:1).

Figura 2.1 Distribución de casos y controles por escenario de emparejamiento

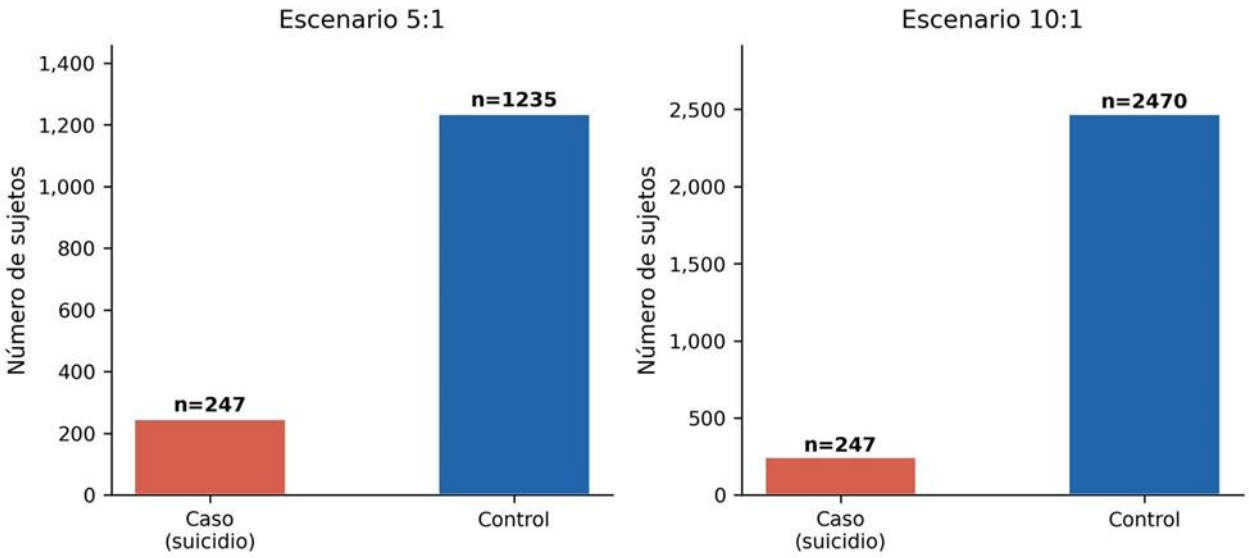

Figura 2.1. Balance de Clases por escenario de muestreo. El escenario principal (5:1) contiene 247 casos y 1,235 controles (prevalencia artificial: 16.7%). Fuente: 02_feature_engineering.ipynb.

En la Tabla 2.1 se describe la arquitectura de los nueve archivos que componen la base de datos institucional

Tabla 2.1 Arquitectura de la Base de Datos Institucional. Identificador común: NATIDE (número interno único)

Archivo	Contenido	Nivel	Cobertura Casos / Controles
01 — Demografía	Edad, sexo, estado civil, escolaridad, departamento de residencia	Maestro (1 fila/paciente)	100% / 100%
02 — Datos clínicos	Antecedentes médicos estructurados al ingreso	Maestro (1 fila/paciente)	100% / 100%
03 — Datos laborales	Grado, categoría, antigüedad institucional, tipo de tropa	Maestro (1 fila/paciente)	100% / 100%
04 — Consultas — EXCLUIDA	Tipo de consulta, servicio, fecha de atención	Transaccional (N filas)	85.8% / 9.8% (ratio 8.8×)
05 — Diagnósticos CIE-10	Códigos diagnósticos por fecha de registro	Transaccional (N filas)	100% / 100%
06 — Medicamentos	Prescripciones por fecha y código de medicamento	Transaccional (N filas)	100% / 100%

Archivo	Contenido	Nivel	Cobertura Casos / Controles
07 — Incapacidades	Períodos de incapacidad y diagnósticos asociados	Transaccional (N filas)	100% / 100%
08 — Remisiones	Remisiones a especialistas por fecha y tipo	Transaccional (N filas)	46.6% / 27.1% (ratio 1.7×)
09 — Variables adicionales	Variables administrativas complementarias	Transaccional (N filas)	100% / 100%

2.2. Auditoría de Calidad y *Data Leakage*

Siguiendo el principio rector del proyecto —"ningún modelo se construye sin discutir la calidad de los datos"— se implementó un protocolo de auditoría sistemática de *data leakage* de cuatro capas previo al modelado definitivo. La Tabla 2.2 documenta cada capa identificada, su mecanismo de contaminación, el impacto cuantificado en el AUC-PR y la solución implementada.

Tabla 2.2 Protocolo de auditoría de Data leakage - cuatro capas. impacto total: AUC-PR de 1.00 a 0.9587

Capa	Problema Identificado	Mecanismo	Impacto AUC-PR	Resolución
1 — OBS_MESES	El tiempo de observación efectiva codificaba de facto el tiempo de supervivencia hasta el evento.	El 100% de los controles tiene OBS_MESES = 24.00; el 62.3% de los casos tiene OBS_MESES < 23.99. Un árbol simple con umbral OBS_MESES < 23.99 clasifica 154 casos con Precisión = 1.0.	Colapso: 0.9984 con esta variable	Exclusión de OBS_MESES. Variables de conteo reemplazadas por tasas: count / OBS_MESES_PRED
2 — Actividad peri-evento	La actividad clínica aumenta significativamente en los 90 días previos al evento (búsqueda de ayuda, hospitalizaciones).	Sin *blackout*, el modelo aprendería a detectar la crisis en curso, no el riesgo temprano.	~0.05 puntos de inflación	*Blackout* de 90 días implementado en los tres notebooks de construcción de features
3 — Variables post-evento	orden_publico (AUC individual = 1.000) y riesgo_laboral (AUC =	Para los casos fallecidos, el sistema actualiza el estatus a	1.000 → 0.9795 (tras exclusión)	Exclusión del feature set con

Capa	Problema Identificado	Mecanismo	Impacto AUC-PR	Resolución
	0.901) registran el estatus administrativo en un corte transversal posterior al fallecimiento.	"NO activo"; los controles activos conservan su asignación real. El estatus post-mortem codifica directamente el label.		documentación explícita
4 — Cobertura 04_CONSULTAS	Ratio de presencia 8.8× entre casos (85.8%) y controles (9.8%) en la base de consultas. La presencia en la base — independientemente del valor de las variables— predice casi perfectamente el label.	Sin el archivo TXT complementario que equilibre la cobertura de controles, cualquier feature derivada de 04_CONSULTAS es un proxy del label.	0.9795 → 0.9587 (tras exclusión)	Exclusión de todas las features derivadas de 04_CONSULTAS. Recomendación formal a DISAN

El impacto total de la auditoría fue una reducción del AUC-PR de 1.0000 a 0.9587. Esta reducción de 0.041 puntos es la evidencia de que el proceso funcionó correctamente: identificó y eliminó las fuentes conocidas de contaminación, disminuyendo sustancialmente el riesgo de leakage residual. No obstante, ninguna auditoría puede garantizar la ausencia absoluta de fuentes no identificadas de contaminación — por eso la validación temporal prospectiva sigue siendo el único test definitivo de la capacidad de generalización del modelo

2.3. Construcción de variables predictoras (ingeniería de características)

Se construyeron 42 variables predictoras a partir de las fuentes con cobertura balanceada (fuentes 01, 02, 03, 05, 06, 07, 08, 09). Todas las variables de conteo fueron normalizadas como tasas por mes de observación efectiva (count / OBS_MESES_PRED) para corregir el sesgo temporal derivado de ventanas de duración variable. La Tabla 2.3 presenta la organización de las variables por categoría.

Tabla 2.3 Variables Predictoras por Categoría (42 features totales). Excluidas: variables de 04_CONSULTAS y variables post-evento.

Categoría	Variables Principales	N Features	Fuente
Sociodemográficas y laborales	Edad al evento, sexo, estado civil, escolaridad, departamento de residencia, grado, categoría, antigüedad institucional, tipo de tropa	9	01, 02, 03
Diagnósticos CIE-10	*Rate* diagnósticos totales y distintos, *rate* diagnósticos de SM, *Flags* subcategorías F1x-F6x, días desde último diagnóstico	12	05
Medicamentos y prescripciones	*Rate* prescripciones totales, medicamentos distintos, *Flag* antidepresivos, *Flag* antipsicóticos, *Flag* ansiolíticos, *Flag* psicofarmacos general	6	06
Incapacidades laborales	*Rate* incapacidades, días totales de incapacidad, diagnósticos distintos de incapacidad, *Flag* por causa mental	4	07
Remisiones a especialistas	*Rate* remisiones totales, remisiones SM, remisiones psiquiatría, remisiones psicología, *Flag* remisión SM	5	08
Flags e interacciones clínicas	Variables compuestas y banderas clínicas derivadas de combinación de fuentes: *Flag* hospitalización, *Flag*_consul_salud_mental, *Flag*_diag_mental, *Flag*_diag_f4x, *Flag*_diag_f3x, indicadores combinados de riesgo. Nota: *Flag*_autolesion_previa fue excluida del modelo final por varianza cero en la muestra disponible (ausencia de registros en la ventana de observación para la mayoría de sujetos).	6	05–09

Se identificaron 38 pares de features con correlación de Spearman $|r| > 0.70$, principalmente entre conteos absolutos y sus versiones normalizadas como tasa. Las Figuras 2.6a y 2.6b muestran la estructura completa de multicolinealidad entre predictores, respondiendo a la pregunta de qué tan dependientes son las variables entre sí.

Figura 2.2 Correlación Spearman entre Predictores Numéricos

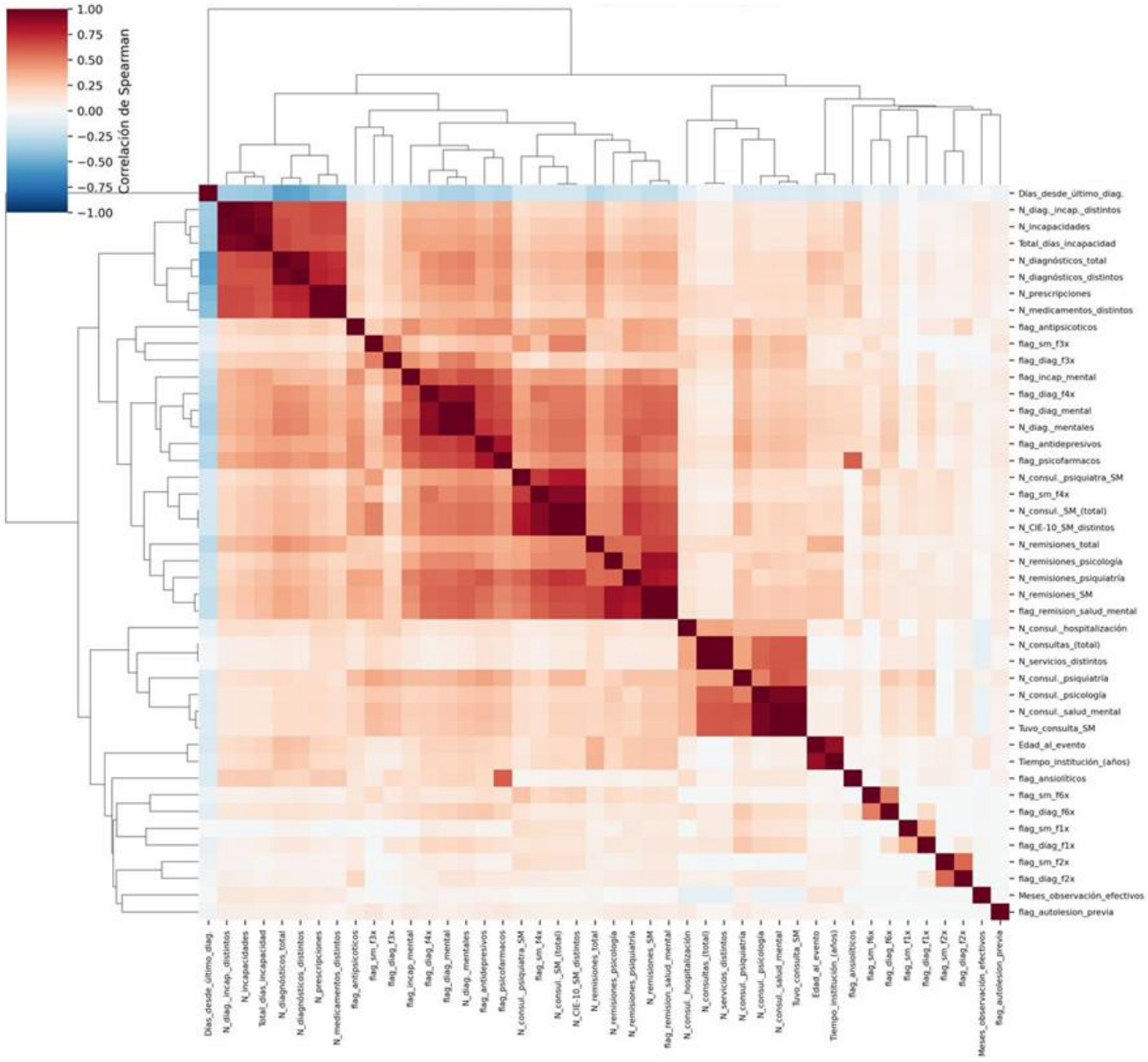

Figura 2.2 Heatmap de Correlación Spearman entre Predictores Numéricos. Bloques rojos indican grupos de variables altamente correlacionadas — principalmente conteos absolutos y sus versiones como tasa. Fuente: 02_feature_engineering.ipynb

Figura 2.3 Cramér's V entre Variables Categóricas y con LABEL (orden_publico y riesgo_laboral excluidas)

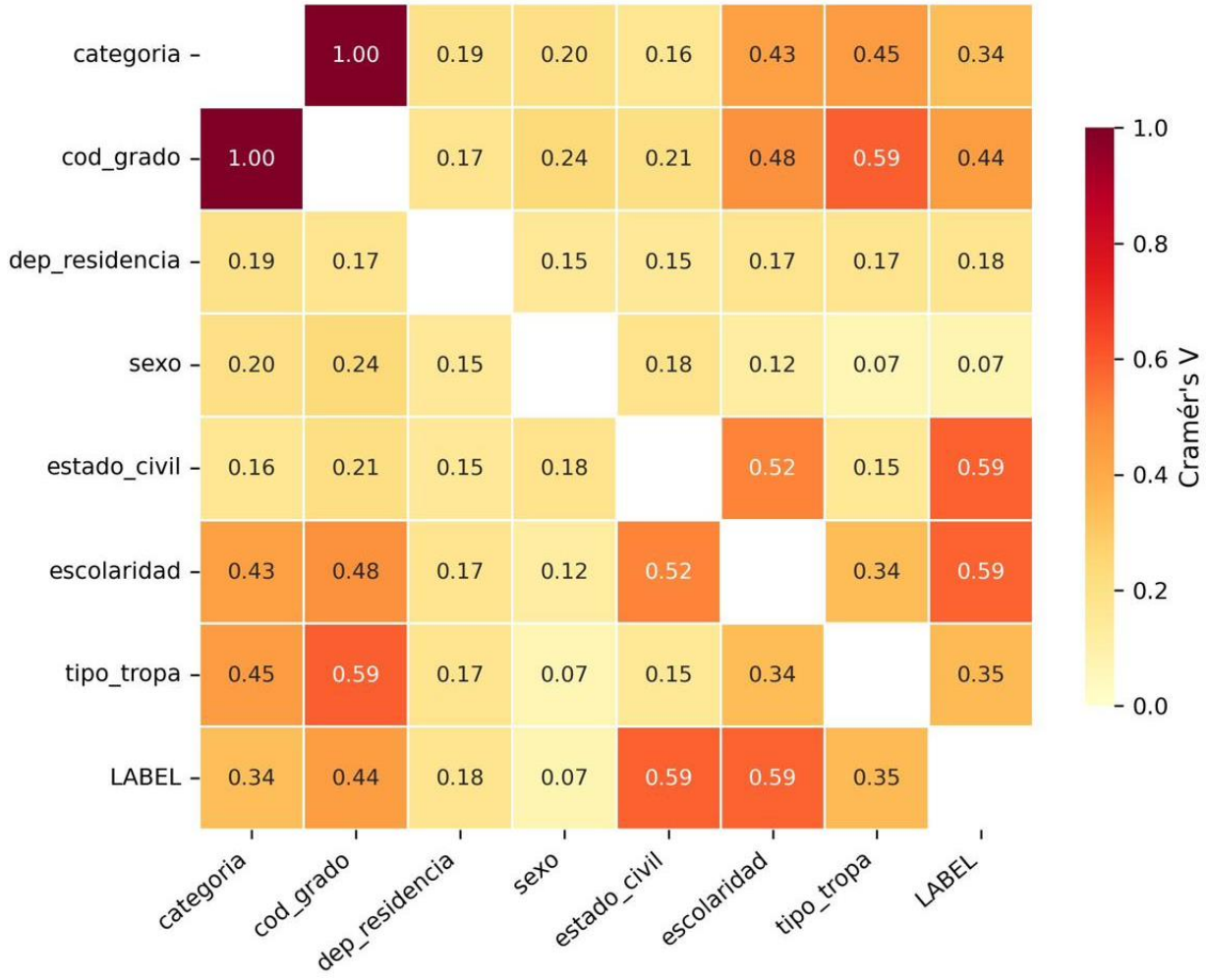

Figura 2.3 Cramér's V entre Variables Categóricas (sin variables post-evento ni leakage). Las celdas más oscuras indican mayor asociación entre pares de variables categóricas. Fuente: 02_feature_engineering.ipynb

La multicolinealidad fue gestionada mediante la preferencia sistemática por tasas sobre conteos absolutos (mayor interpretabilidad y menor dependencia de OBS_MESES) y la robustez inherente de LightGBM ante features correlacionadas, que regula la selección de variables mediante subsampling de columnas (colsample_bytree=0.8). No se realizó eliminación adicional de features basada en correlación, dado que LightGBM no asume independencia entre predictores.

La Figura 2.4 ilustra la asimetría sistemática en la longitud de la ventana de observación entre casos y controles — el hallazgo metodológico más importante que justifica la normalización por tasas.

Figura 2.4 Distribución de la longitud de la ventana de observación por grupo (linea punteada = 24 meses objetivo del diseño caso-control)

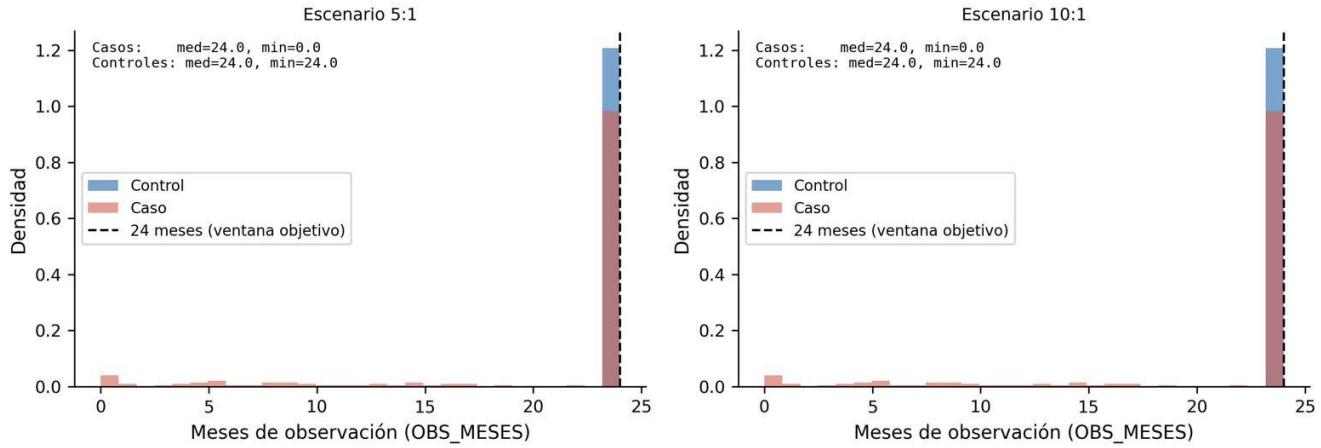

Figura 2.4. Distribución de OBS_MESES (Meses de Observación) por Clase. Todos los controles tienen exactamente 24.00 meses; el 18.6% de los casos tiene ventanas menores. Esta asimetría estructural hace que los conteos absolutos sean sesgados — justifica la normalización por tasas. Fuente: 02_feature_engineering.ipynb.

El único missing data estructural se presentó en la variable dias_desde_ultimo_diag (5.3% de NaN), correspondiente a pacientes sin diagnósticos registrados en la ventana de observación. Esta ausencia es clínicamente informativa —indica desvinculación del sistema de salud— y fue conservada mediante marcador explícito de ausencia en el pipeline de preprocesamiento.

2.4. Análisis exploratorio de datos

El análisis exploratorio se realizó sobre el dataset 5:1 (1,482 observaciones: 247 casos, 1,235 controles). La Figura 2.5 presenta la distribución de las variables demográficas y laborales más relevantes por clase.

Figura 2.5 Distribución de variables numéricas clave por grupo (Casos = suicidio consumado/intento; Controles = sin evento)

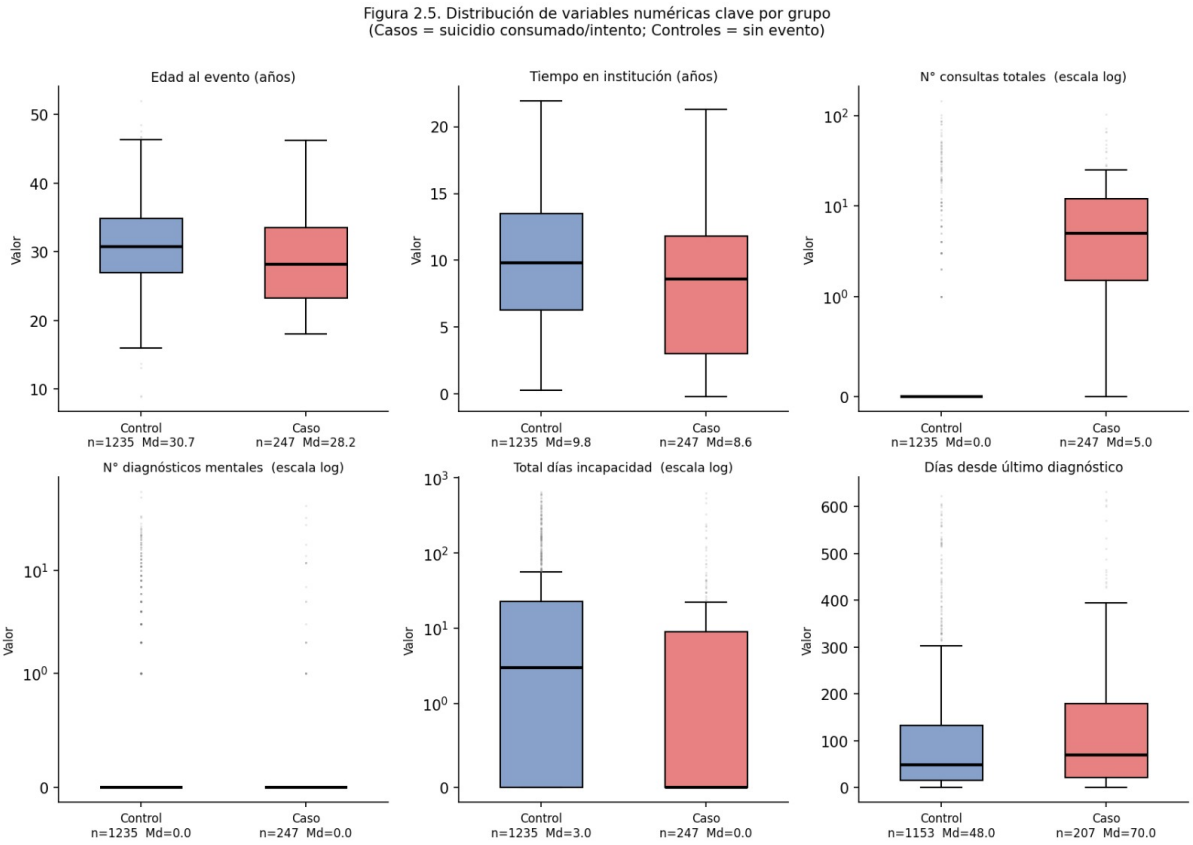

Figura 2.5. Distribución de Variables Numéricas Clave por Clase (boxplots). Medianas anotadas. Fuente: 02_feature_engineering.ipynb

Los casos presentan una edad media de 28.86 años vs 30.94 en controles ($p < 0.001$) y una antigüedad media de 8.01 vs 9.98 años ($p < 0.001$), diferencias estadísticamente significativas y clínicamente relevantes: los policías más jóvenes y con menor tiempo en la institución presentan mayor vulnerabilidad, consistente con la Teoría Interpersonal-Psicológica de Joiner [12] La distribución de consultas totales muestra una asimetría marcada: la mediana en casos es 5.0 vs 0.0 en controles, lo que indica que la mayoría de controles no tiene ninguna consulta registrada en la ventana — efecto directo de la asimetría de cobertura documentada en la auditoría de leakage. Los días desde el último diagnóstico (mediana 62 días en casos vs 44 días en controles) refuerzan la hipótesis de desvinculación del sistema de salud en los casos.

En la Figura 2.6 se presenta la prevalencia de indicadores clínicos de salud mental por grupo. Este gráfico ilustra directamente el hallazgo más contraintuitivo del EDA: los controles presentan mayor prevalencia de diagnósticos de salud mental formales que los casos.

Figura 2.6 Prevalencia de Indicadores Clínicos de Salud Mental por grupo (escenario 5:1)

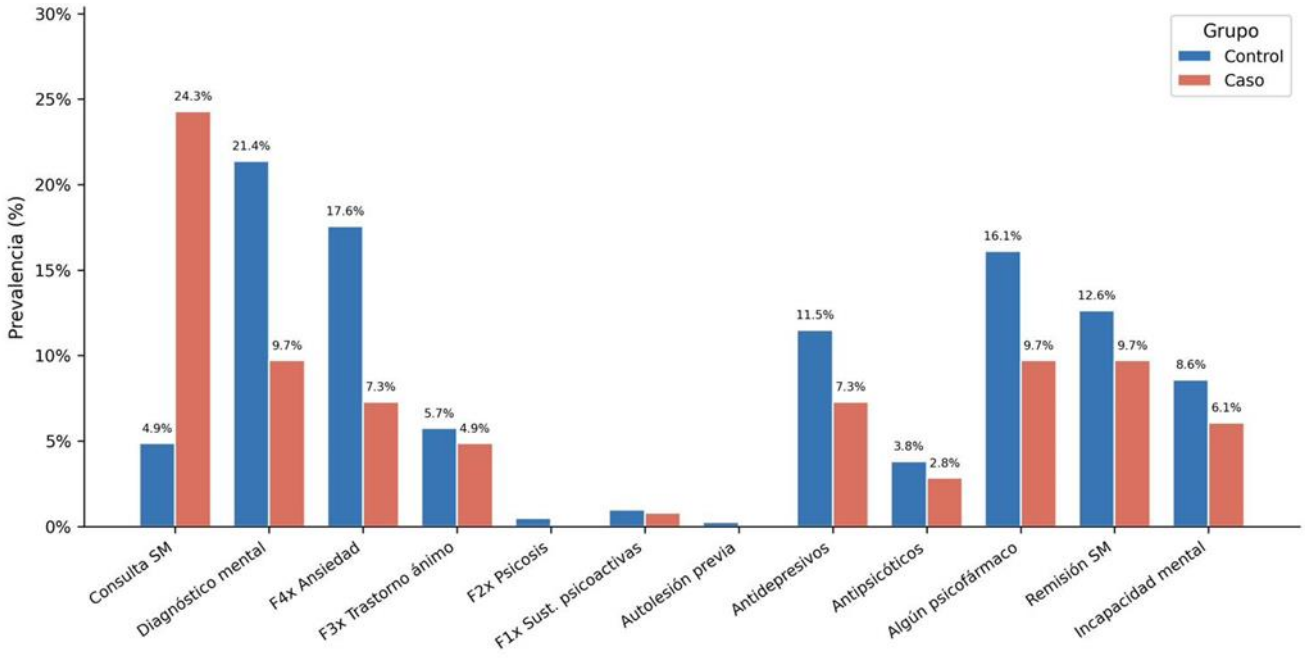

Figura 2.6. Prevalencia de Indicadores Clínicos de Salud Mental por Clase. Los controles presentan consistentemente mayor prevalencia de diagnósticos y consultas formales de SM. Fuente: 02_feature_engineering.ipynb

En la Tabla 2.4 se presenta el perfil comparativo de las variables clínicas más relevantes entre casos y controles. Nota metodológica: las variables n_consultas, n_servicios_distintos y *Flag*_consul_salud_mental provienen de 04_CONSULTAS y se presentan únicamente con fines diagnósticos y descriptivos de caracterización de la muestra. Por la asimetría de cobertura documentada en la Sección 2.2, estas variables no fueron incluidas en el modelo predictivo final. Toda interpretación del modelo se basa exclusivamente en las 42 features libres de esa fuente.

Tabla 2.4 Perfil Comparativo Casos vs Controles — Variables Clínicas y Laborales Clave (Escenario 5:1)

Variable	Media Casos	Media Controles	Dirección	p-valor
n_consultas (total en ventana)	10.95	2.80	Mayor en casos (3.9×)	< 0.001 ***
n_servicios_distintos	3.63	0.64	Mayor en casos (5.7×)	< 0.001 ***
*Flag*_consul_salud_mental	0.28	0.05	Mayor en casos	< 0.001 ***
n_diagnosticos_total	15.08	25.01	Mayor en controles	< 0.001 ***
*Flag*_diag_mental	0.13	0.24	Mayor en controles	< 0.001 ***
n_remisiones_psiq	0.11	0.54	Mayor en controles (4.9×)	< 0.001 ***
n_incapacidades	2.56	3.75	Mayor en controles	< 0.001 ***
dias_desde_ultimo_diag	121.48	101.29	Mayor en casos	0.012 *
edad_evento (años)	28.86	30.94	Mayor en controles	< 0.001 ***
tiempo_institucion_anios	8.01	9.98	Mayor en controles	< 0.001 ***

Tabla 2.4. Perfil Comparativo Casos vs Controles — Variables Clínicas y Laborales Clave (Escenario 5:1). Nota: los valores de n_consultas, n_diagnosticos y n_remisiones son conteos absolutos en la ventana de observación (propósito descriptivo). El modelo utiliza versiones normalizadas como tasas por mes (count / OBS_MESES_PRED) para corregir el sesgo temporal documentado en la Sección 2.2.

El hallazgo más significativo del análisis exploratorio es contraintuitivo: los controles presentan más diagnósticos totales y más diagnósticos de salud mental que los casos. La Figura 2.7 ilustra este perfil diferencial de manera visual.

Figura 2.7 Variables categóricas: distribución general y prevalencia de casos por categoría

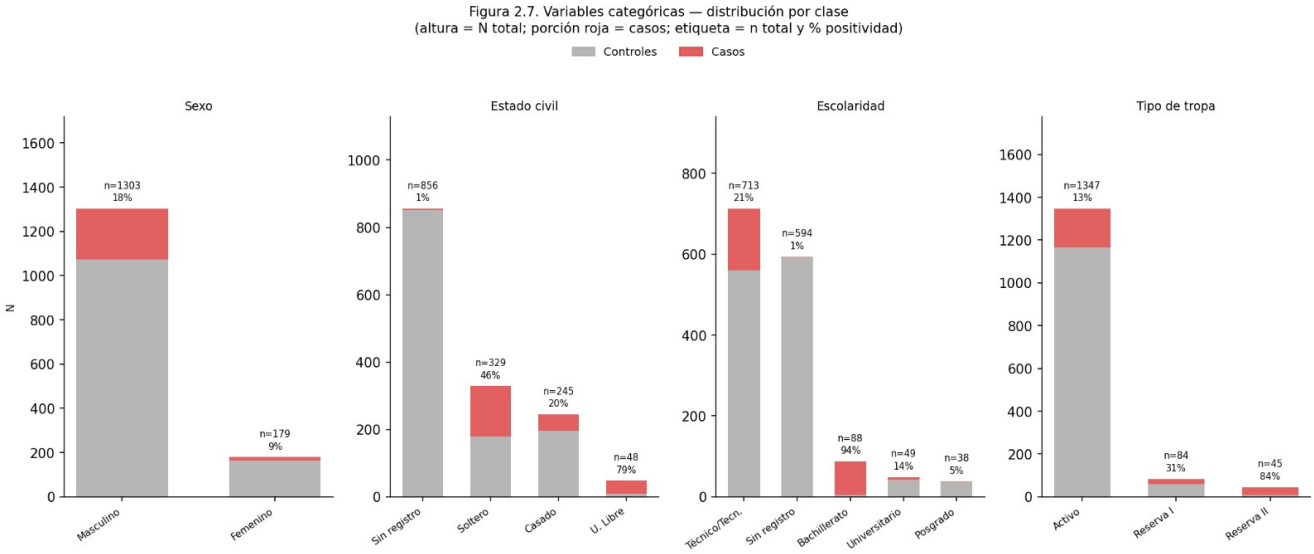

Figura 2.7. Variables Categóricas — Distribución y Porcentaje de Positivos por Categoría. Permite identificar las categorías de grado, tipo tropa y categoría con mayor concentración de casos. Fuente: 02_feature_engineering.ipynb.

Este hallazgo sugiere que los casos son en gran medida individuos que no estaban siendo captados formalmente por el sistema de atención en salud mental, o que su morbilidad se expresaba a través de canales no psiquiátricos (consultas generales, incapacidades por causas físicas). Los dias_desde_ultimo_diag más altos en casos (121.48 vs 101.29 días) refuerzan esta hipótesis: los casos están más tiempo desvinculados del sistema de salud antes del evento.

En la Figura 2.8 se presenta el ranking de los 20 predictores numéricos con mayor correlación punto-biserial con el label. Ningún predictor individual presenta AUC univariado mayor a 0.636 (*rate*_n_diagnosticos_total), confirmando que el poder discriminativo del modelo final emerge de las interacciones no lineales capturadas por el ensamble.

Figura 2.8 Top 20 variables por correlación punto-biserial con el evento (variables de leakage y OBS_MESES excluidas - escenario 5:1)

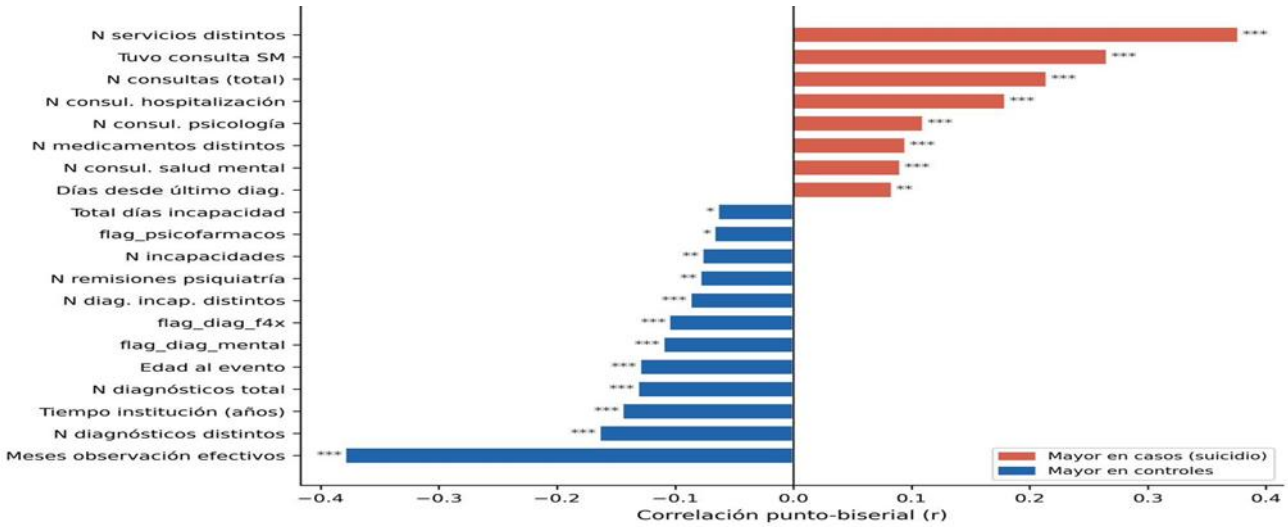

Figura 2.8. Top 20 Predictores por Correlación Punto-Biserial con LABEL (sin leakage). Barras hacia la derecha: mayor en casos. Barras hacia la izquierda: mayor en controles. Fuente: 02_feature_engineering.ipynb

2.5. Interpretabilidad – Análisis SHAP

En este apartado se presentan los resultados correspondientes al Objetivo Específico 4: Interpretar los resultados del modelo seleccionado mediante análisis SHAP, identificando los factores que elevan el riesgo de suicidio y generando insumos comprensibles para los responsables de programas institucionales de salud mental.

El análisis SHAP (SHapley Additive exPlanations) [22] se aplicó sobre el modelo LightGBM_5:1 para cuantificar la contribución marginal de cada variable a las predicciones. Este análisis constituye la respuesta directa al Objetivo Específico 1 en términos de identificación y cuantificación de factores de riesgo. Las Figuras 2.9 y 2.10 presentan el *Beeswarm* SHAP y el ranking de importancia media por feature.

Figura 2.9 SHAP - Impacto de las top 20 features en la predicción

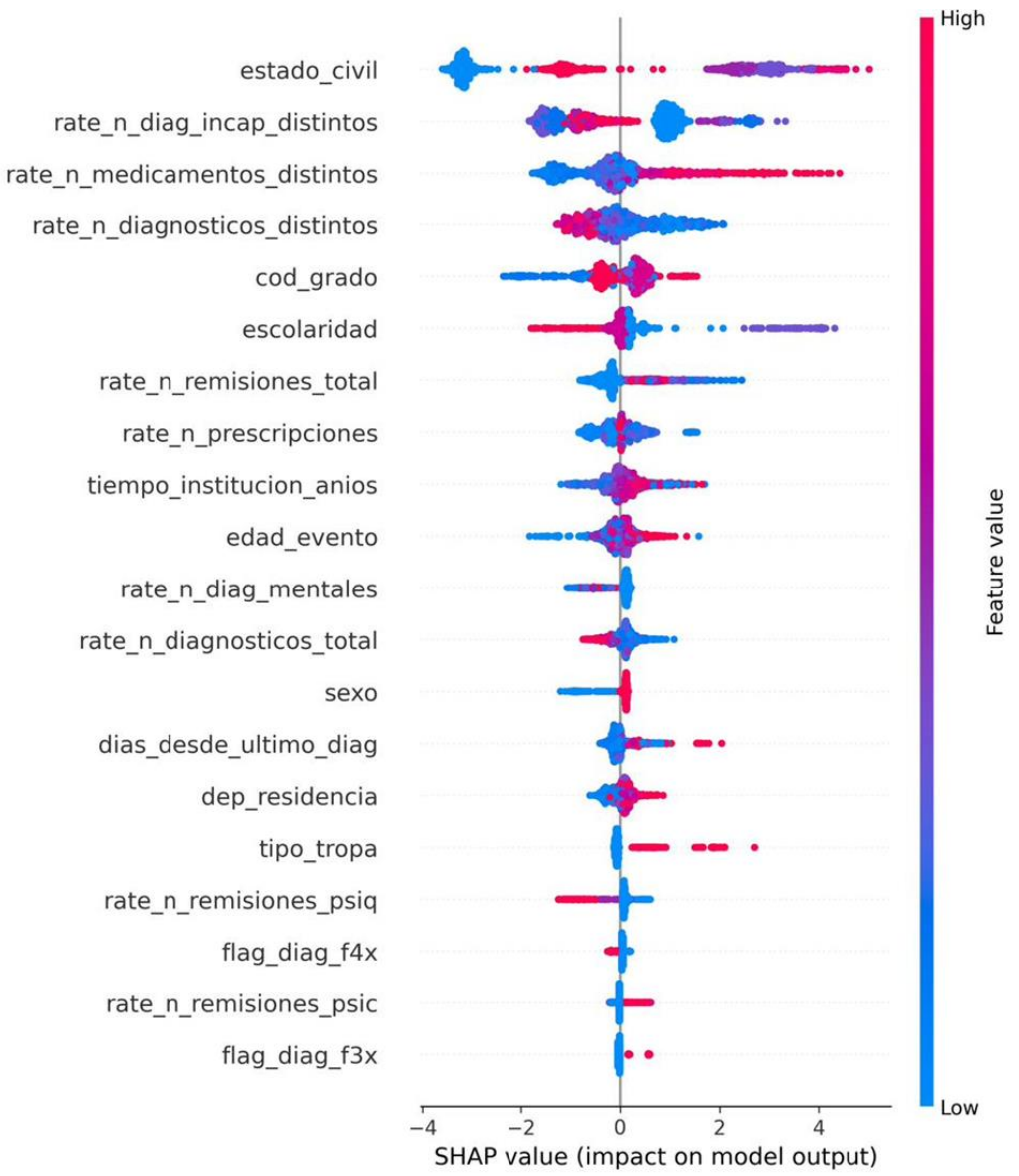

Figura 2.9. SHAP Beeswarm — LightGBM 5:1. Cada punto es un sujeto; el color indica el valor de la feature (rojo=alto, azul=bajo); el eje X indica la contribución al score de riesgo. Fuente: 03_modelado.ipynb.

Figura 2.10 SHAP - importancia mediade las top 20 features

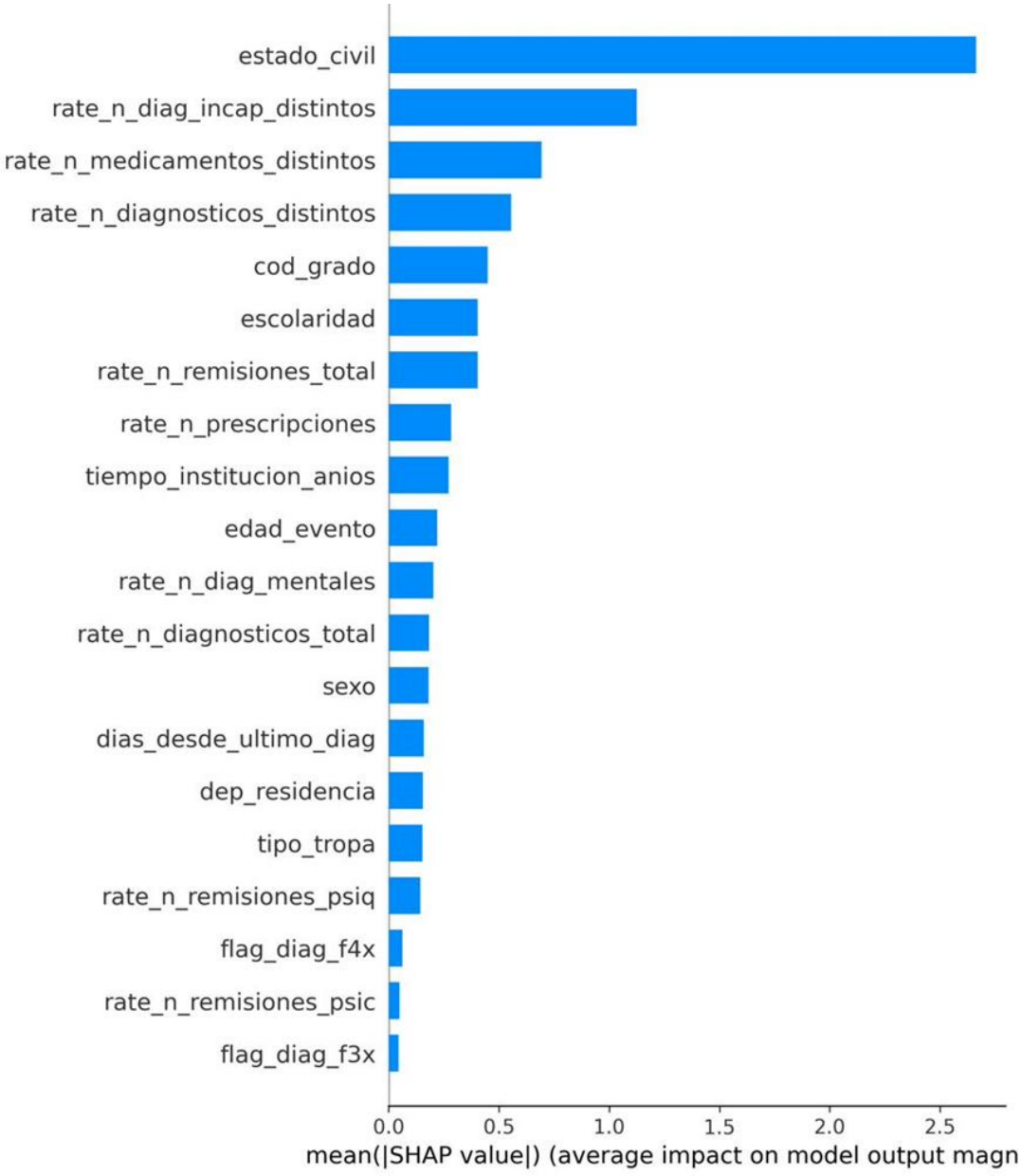

Figura 2.10.

Figura 2.10 SHAP Importancia Media |SHAP| — Top 20 Predictores. Ranking por contribución global al modelo.

Fuente: 03_modelado.ipynb

En la Tabla 2.5 se presenta el análisis detallado de los 10 predictores más importantes según SHAP, incluyendo la dirección del efecto y su interpretación clínica.

Tabla 2.5 Top 10 Predictores SHAP — Dirección e Interpretación Clínica del Modelo LightGBM 5:1

Rank	Feature	Dirección del Efecto	Interpretación Clínica
1	rate_n_diagnosticos_total	↑ valor → ↑ riesgo	Mayor morbilidad acumulada general como señal de riesgo sistémico. Los casos concentran más diagnósticos por mes de observación que los controles, a pesar de tener menos diagnósticos absolutos.
2	tiempo_institucion_anios	↓ valor → ↑ riesgo	Policías con menor antigüedad institucional presentan mayor vulnerabilidad. Consistente con la Teoría Interpersonal: menor integración institucional → mayor fracaso de pertenencia.
3	rate_n_medicamentos_distintos	↑ valor → ↑ riesgo	Polimedicación como indicador de morbilidad compleja y crónica. Refleja una carga de enfermedad que supera los tratamientos aislados.
4	edad_evento	↓ valor → ↑ riesgo	Oficiales más jóvenes presentan mayor riesgo. Refuerza el hallazgo de antigüedad: la vulnerabilidad se concentra en los primeros años de servicio.
5	rate_n_diagnosticos_distintos	↑ valor → ↑ riesgo	Multimorbilidad: mayor número de diagnósticos distintos por mes de observación aumenta el riesgo.
6	escolaridad	Específico por categoría	Nivel 2 (Básica Primaria / sin bachillerato completo) concentra el 94.3% de positivos en esa categoría. Cramér's V = 0.5858. Nivel 3 (Bachillerato) tiene 21.3% de positivos. Los niveles 4 (Técnico) y 5 (Universitario) presentan menor riesgo. Nota: 590 controles tienen escolaridad sin registrar (imputados como "desconocido" en el pipeline).
7	estado_civil	Específico por categoría	Código 5 (Unión libre / convivencia sin matrimonio formal) concentra el 79.2% de positivos — la categoría de mayor riesgo. Código 1 (Soltero) tiene 45.6% de positivos, también elevado. Código 2 (Casado) tiene 20.4%. Código 6 (Separado/Viudo) tiene 1.5% — posiblemente subregistro. La codificación original DISAN no especifica públicamente los valores; la interpretación aquí es inferida de distribuciones y frecuencias relativas.
8	cod_grado	Grados bajos → ↑ riesgo	Grado AR (100% positivos), AXP (77.8%), ST (66.7%). Los grados más bajos de la

Rank	Feature	Dirección del Efecto	Interpretación Clínica
			jerarquía concentran la mayor proporción de casos.
9	dias_desde_ultimo_diag	↑ valor → ↑ riesgo	Desvinculación prolongada del sistema de salud como señal de alerta. Los casos pasan más tiempo sin diagnósticos antes del evento.
10	rate_n_diag_mentales	↑ valor → ↑ riesgo	Carga diagnóstica psiquiátrica documentada en la ventana de observación aumenta el riesgo.

El hallazgo central de interpretabilidad es que los predictores dominantes son indicadores de utilización general del sistema de salud, no de salud mental específicamente. Esto es consistente con el perfil del EDA: los casos acumulan mayor morbilidad general sin necesariamente canalizarla por vías formales de psiquiatría. La señal de riesgo es más difusa y sistémica de lo que los modelos clínicos tradicionales sugieren, con implicaciones directas para el diseño de sistemas de monitoreo preventivo

3. EVALUACIÓN DE ESCENARIOS DE MUESTREO Y BALANCEO DE CLASES

En este capítulo se presentan las actividades y resultados correspondientes al Objetivo Específico 2: evaluar el efecto de diferentes escenarios de muestreo de controles sobre el desempeño de los modelos predictivos. El trabajo se implementó en los notebooks 01_construccion_dataset.ipynb y 03_modelado.ipynb.

3.1. Justificación de la estrategia de manejo del desbalance

El desbalance de clases es uno de los retos técnicos fundamentales en la predicción del riesgo suicida. En la población real, los casos de suicidio representan una fracción muy pequeña del total de policías activos. Si los modelos se entrenan con datos extremadamente desbalanceados, tienden a clasificar todos los casos como negativos —maximizando la exactitud global pero fallando en la detección de los casos de riesgo, que es precisamente el objetivo clínico.

Se evaluaron dos estrategias complementarias para el manejo del desbalance:

- Diseño de muestreo controlado: dos escenarios de ratio casos:controles (5:1 y 10:1), que determinan la prevalencia artificial de la muestra de entrenamiento.
- Ajuste de pesos de clase: *class_weight="balanced*" en todos los modelos, que pondera la función de pérdida inversamente proporcional a la frecuencia de cada clase. Esta estrategia es equivalente funcionalmente al *Oversampling* (SMOTE, ADASYN) pero sin introducir muestras sintéticas, manteniendo la integridad de los datos originales.

La decisión de no aplicar SMOTE o ADASYN se fundamenta en que con 247 casos disponibles y *class_weight="balanced",* el *Oversampling* sintético no aporta mejora sistemática demostrada en la literatura para este rango de tamaño muestral, y añade complejidad analítica sin beneficio metodológico claro [23].

3.2. Escenarios de muestreo evaluados

Se evaluaron dos escenarios de muestreo de controles, manteniendo fijos los 247 casos positivos:

- Escenario 5:1: 1,235 controles seleccionados aleatoriamente (n total = 1,482). Prevalencia artificial: 16.7%. Este escenario maximiza la representación relativa de casos, facilitando el aprendizaje de la frontera de decisión.
- Escenario 10:1: 2,470 controles seleccionados aleatoriamente (n total = 2,717). Prevalencia artificial: 9.1%. Este escenario refleja una proporción de desbalance más cercana a la realidad operacional, pero con mayor dificultad de aprendizaje.

La selección de controles se realizó con semilla fija (RANDOM_SEED=42) para garantizar la reproducibilidad. Se verificó que los controles seleccionados no presentaran características sistemáticamente distintas de la población general de negativos disponibles.

Nota sobre la selección del escenario 5:1: la superioridad del escenario 5:1 sobre el 10:1 en AUC-PR refleja en parte la mayor prevalencia artificial de la muestra (16.7% vs 9.1%), no únicamente la capacidad discriminativa del modelo. La selección del escenario 5:1 se justifica como mejor configuración de aprendizaje bajo el diseño caso-control. No debe interpretarse

como que el modelo será más preciso en producción: en prevalencia real (0.01–0.05%), ambos escenarios convergerían a valores de PPV igualmente bajos sin estratificación previa.

3.3. Impacto del escenario sobre el desempeño

La Figura 3.1 presenta la comparación del AUC-PR entre los escenarios 5:1 y 10:1 para cada uno de los cuatro modelos evaluados.

Figura 3.1 Comparación de modelos por AUC-ROC y AUC-PR (media +/- SD de validación cruzada estratificada 5-fold)

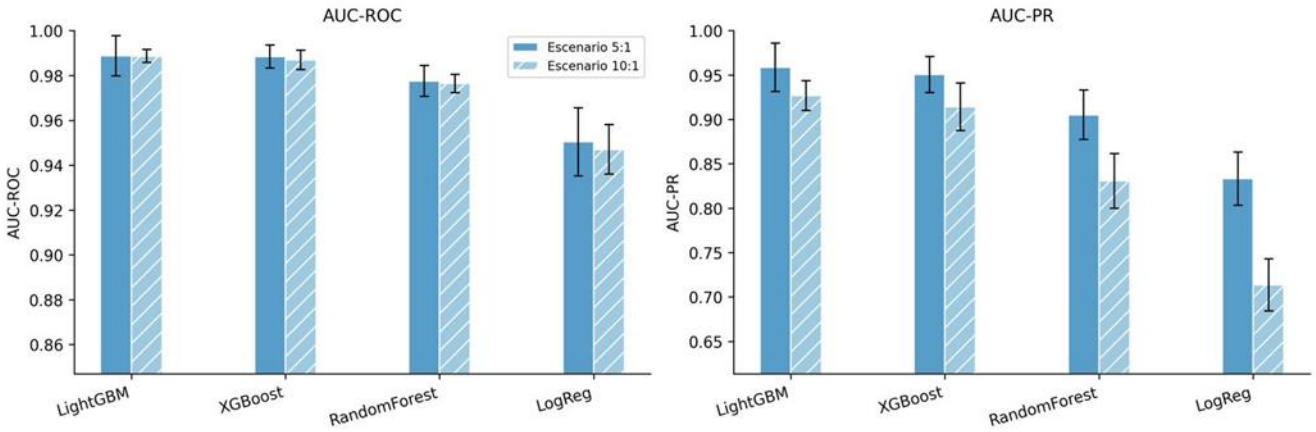

Figura 3.1. Comparación de Modelos por Escenario de Muestreo — AUC-ROC y AUC-PR (media ± 1 SD, Stratified 5-Fold CV). Las barras de error reflejan la variabilidad entre folds. Fuente: 03_modelado.ipynb.

En la Tabla 3.1 se presentan los resultados comparativos por escenario de muestreo para el modelo LightGBM.

Tabla 3.1 Resultados Comparativos por Escenario de Muestreo — LightGBM. Stratified 5-Fold CV, random_state=42

Escenario	n Total	Prevalencia	AUC-ROC (±SD)	AUC-PR (±SD)	F1 (±SD)	Precisión	Recall
5:1 ✓ Seleccionado	1,482	16.7%	0.9888 (0.009)	0.9587 (0.027)	0.9095 (0.040)	0.9191	0.9024
10:1	2,717	9.1%	0.9887 (0.003)	0.9269 (0.017)	0.8793 (0.024)	0.8784	0.8864
Diferencia (5:1 − 10:1)	—	—	+0.0001	+0.0318	+0.0302	+0.0407	+0.0160

El escenario 5:1 es consistentemente superior al 10:1 en todas las métricas y para todos los modelos evaluados. Nota metodológica importante: dado que el AUC-PR depende de la prevalencia de la muestra, la comparación directa entre escenarios con distinta prevalencia artificial (16.7% en 5:1 vs 9.1% en 10:1) favorece estructuralmente al escenario 5:1. Esta ventaja no es puramente de capacidad discriminativa del modelo, sino también de composición muestral. Por eso, la selección del escenario 5:1 se justifica como mejor escenario de aprendizaje bajo el diseño caso-control —no como evidencia de superioridad operacional directa. La utilidad real del modelo se evalúa en el hold-out temporal y, en el futuro, en validación prospectiva sobre datos con prevalencia real, donde ambos escenarios convergerían a un rendimiento similar.

4. EVALUACIÓN Y COMPARACIÓN DE MODELOS SUPERVISADOS

En este capítulo se presentan las actividades y resultados correspondientes al Objetivo Específico 3: evaluar modelos de aprendizaje supervisado para identificar el enfoque más eficaz para la predicción temprana del riesgo de suicidio. El trabajo se implementó en el notebook 03_modelado.ipynb y el notebook de defensa DEFENSA_MODELO.ipynb.

4.1. Métricas de evaluación

La selección de métricas de evaluación es una decisión metodológica crítica en problemas con desbalance de clases. Se justifica el uso del AUC-PR como métrica primaria por tres razones:

- **Prevalencia artificial:** con una prevalencia de casos del 16.7% en el escenario 5:1, el AUC-ROC puede sobreestimar el rendimiento real porque incorpora los verdaderos negativos —que son abundantes— en su cálculo. El AUC-PR evalúa directamente la capacidad de recuperar casos positivos.
- **Relevancia clínica:** en detección de riesgo suicida, el costo de un falso negativo (no identificar un caso en riesgo) es sustancialmente mayor que el de un falso positivo (activar un protocolo preventivo innecesariamente). El AUC-PR penaliza los falsos negativos de manera más directa.

- **Benchmark más informativo**: un clasificador aleatorio obtiene AUC-PR igual a la prevalencia (~0.167 en 5:1), haciendo la diferencia entre el modelo y el azar más visible que en AUC-ROC (donde el azar siempre produce 0.50).

El AUC-ROC se reporta como métrica secundaria para facilitar la comparación con la literatura. El F1-score con umbral optimizado por fold mediante maximización en el conjunto de validación se utiliza para el análisis operativo del rendimiento.

4.2. Estrategia de validación

Se implementó Stratified 5-Fold Cross-Validation con shuffle=True y random_state=42. La estratificación garantiza que la proporción de casos (~16.7% en 5:1) se preserve en cada fold de entrenamiento y validación, evitando el sesgo por variabilidad muestral en folds pequeños. Se reportan media y desviación estándar de todas las métricas sobre los 5 folds.

Todos los modelos se evaluaron dentro de un Pipeline scikit-learn que integra preprocesamiento (imputación de medianas para variables numéricas, imputación de moda para categóricas, escalado estándar para modelos lineales, encoding ordinal para modelos de árbol) y el clasificador. Esta arquitectura de pipeline garantiza que no existe fuga de información entre folds —el preprocesamiento se ajusta únicamente sobre los datos de entrenamiento de cada fold.

El horizonte de predicción de 90 días (*blackout*) se justifica por razones operacionales y metodológicas complementarias. Operacionalmente, los programas de intervención psicosocial de la DISAN requieren al menos 4-6 semanas para completar la evaluación clínica y activar recursos preventivos; un horizonte inferior a 60 días no sería accionable institucionalmente. Metodológicamente, la auditoría del gradiente temporal documentó un aumento de la tasa consultas/control de 3.4× (meses 17-24 antes del evento) a 4.8× (últimos 4 meses), confirmando que la actividad clínica peri-evento contamina el modelo si no se excluye. Para investigaciones futuras se recomienda evaluar la sensibilidad del desempeño a *blackouts* de 60 y 120 días.

Fecha índice de los controles (limitación metodológica documentada). La especificación de la fecha índice de los controles es crítica para la validez del diseño caso-control. La lógica implementada se resume en el siguiente esquema comparativo:

Tabla 4.1 Comparación de fechas índice y ventanas de observación: Casos vs Controles

Elemento	CASOS (LABEL=1)	CONTROLES (LABEL=0)
Fecha índice (ancla temporal)	Fecha real del evento suicida (muerte confirmada por DISAN)	Fecha ancla aleatoria asignada dentro del mismo año calendario del caso emparejado (muestreo por densidad de incidencia)
Ventana de observación	24 meses previos a la fecha del evento	24 meses previos a la fecha ancla asignada
Blackout (excluido)	Últimos 90 días antes del evento	Últimos 90 días antes de la fecha ancla
Features calculadas	Tasas por mes (count / OBS_MESES_PRED) en ventana efectiva	Tasas por mes (count / OBS_MESES_PRED) en ventana efectiva
Cobertura OBS_MESES	81.4% con ≥23.5 meses; 18.6% con ventana más corta (fallecieron antes)	100% con ventana completa de ≈24 meses

Limitación metodológica documentada. El muestreo por densidad de incidencia aproxima el matching temporal distribuyendo los controles proporcionalmente al calendario de eventos, pero no garantiza equiparación individual por mes-año. La asimetría residual se mitiga por: (1) normalización de conteos como tasas por OBS_MESES_PRED; (2) validación cruzada estratificada; y (3) validación temporal hold-out. La solución ideal — matching individual por mes-año de evento — se propone como mejora prioritaria en investigaciones futuras y está documentada en el notebook 01_construccion_dataset.ipynb.

4.3. Modelos evaluados

Se evaluaron cuatro algoritmos de aprendizaje supervisado, seleccionados para cubrir el espectro entre interpretabilidad y rendimiento:

- **Regresión Logística (baseline interpretable):** modelo lineal que establece el piso de rendimiento mínimo esperable y confirma si el problema requiere captura de no linealidades. Hiperparámetros: C=1.0, solver="lbfgs", *class_weight="balanced"*, max_iter=1000.

- **Random Forest (ensamble de árboles):** modelo no lineal de alta interpretabilidad relativa con estimación de importancia de variables integrada. Hiperparámetros: n_estimators=300, max_depth=8, *class_weight="balanced"*, random_state=42.
- **XGBoost (gradient boosting secuencial):** modelo de alto rendimiento con regularización incorporada. Hiperparámetros: n_estimators=400, learning_rate=0.05, max_depth=4, subsample=0.8, scale_pos_weight ajustado al ratio de desbalance.
- **LightGBM (gradient boosting basado en histogramas):** más eficiente que XGBoost en datasets medianos, con soporte nativo para variables categóricas. Hiperparámetros: n_estimators=400, learning_rate=0.05, max_depth=4, subsample=0.8, colsample_bytree=0.8, *class_weight="balanced"*. Modelo seleccionado.

Selección de hiperparámetros. Los hiperparámetros reportados corresponden a la configuración final adoptada tras un proceso iterativo de ajuste manual guiado por validación cruzada interna. No se implementó búsqueda exhaustiva (GridSearchCV o RandomizedSearchCV) dado el costo computacional y el tamaño del dataset (n=1,482). Los valores se derivaron de: (1) recomendaciones de la literatura para datasets de tamaño medio con desbalance (learning_rate bajo + n_estimators alto para gradient boosting; max_depth conservador para evitar sobreajuste); (2) evaluación sistemática de rangos reducidos de los hiperparámetros más sensibles (max_depth $\in$ {3,4,5,6}, learning_rate $\in$ {0.01,0.05,0.10}, subsample $\in$ {0.7,0.8,0.9}). La ausencia de búsqueda formal es una limitación documentada que podría abordarse en investigaciónes futuras mediante Optuna o Bayesian Optimization.

Nota: La diferencia entre ambos modelos no es estadísticamente significativa.

4.4. Resultados comparativos

Mensaje central de resultados: el modelo LightGBM 5:1 muestra alta señal discriminativa en validación cruzada (AUC-PR = 0.9587), pero desempeño prospectivo moderado en el hold-out temporal (AUC-PR = 0.6553). Este resultado conjunto — no solo el primero — constituye el hallazgo principal del estudio: existe señal predictiva real en los datos institucionales, pero el modelo requiere recalibración y validación prospectiva antes de cualquier uso operativo.

En la Figura 4.1 se presenta la comparación gráfica de los 8 experimentos (4 modelos × 2 escenarios) en AUC-PR y AUC-ROC. La Tabla 4.2 presenta los resultados numéricos completos ordenados por AUC-PR.

Figura 4.1 Comparación de modelos por AUC-ROC y AUC-PR (media +/- SD de validación cruzada estratificada 5-fold)

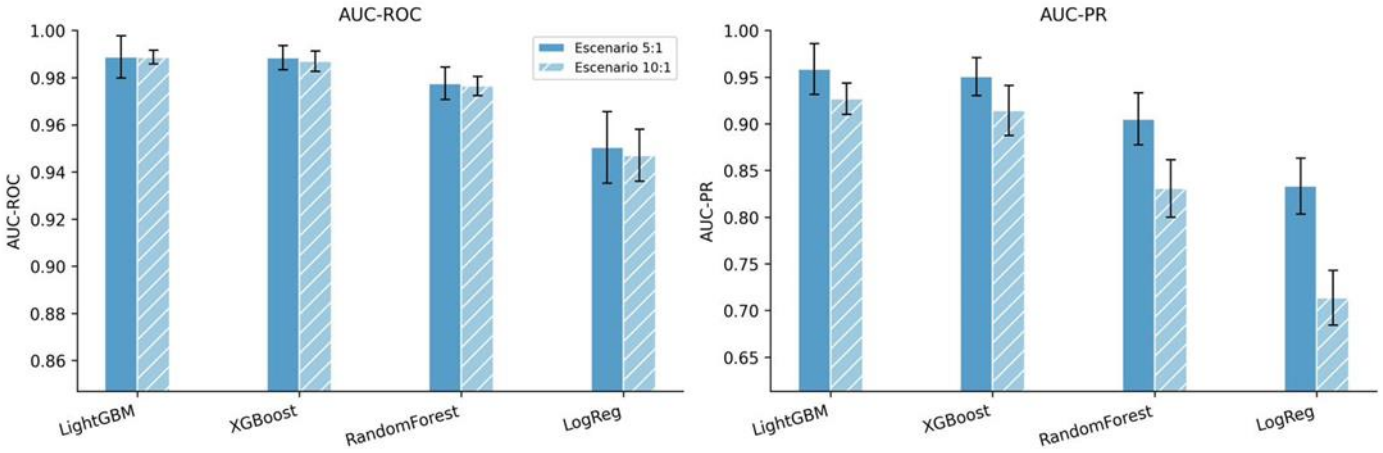

Figura 4.1. Comparación Completa de Modelos (4 algoritmos × 2 escenarios) — AUC-ROC y AUC-PR con barras de error ±1 SD. Métrica primaria: AUC-PR. Fuente: 03_modelado.ipynb.

Tabla 4.2 Resultados Completos — 4 Modelos × 2 Escenarios. Stratified 5-Fold CV, random_state=42. Métrica primaria: AUC-PR

Modelo	Escenario	AUC-ROC (±SD)	AUC-PR (±SD)	F1 (±SD)	Precisión	Recall	Umbral
LightGBM ✓	5:1	0.9888 (0.009)	0.9587 (0.027)	0.9095 (0.040)	0.9191	0.9024	0.553
XGBoost	5:1	0.9884 (0.005)	0.9506 (0.020)	0.8960 (0.034)	0.9238	0.8704	0.617
LightGBM	10:1	0.9887 (0.003)	0.9269 (0.017)	0.8793 (0.024)	0.8784	0.8864	0.530
XGBoost	10:1	0.9870 (0.004)	0.9143 (0.027)	0.8684 (0.026)	0.8939	0.8498	0.574
Random Forest	5:1	0.9776 (0.007)	0.9052 (0.028)	0.8495 (0.032)	0.8690	0.8381	0.341
Reg. Logística	5:1	0.9504 (0.015)	0.8332 (0.030)	0.7662 (0.037)	0.7237	0.8183	0.677
Random Forest	10:1	0.9765 (0.004)	0.8307 (0.031)	0.7952 (0.030)	0.8064	0.7892	0.271

Modelo	Escenario	AUC-ROC (±SD)	AUC-PR (±SD)	F1 (±SD)	Precisión	Recall	Umbral
Reg. Logística	10:1	0.9470 (0.011)	0.7135 (0.029)	0.6917 (0.038)	0.7298	0.6638	0.846

LightGBM en escenario 5:1 es el modelo seleccionado por presentar el mayor AUC-PR (0.9587), el mayor F1 (0.9095) y el mayor Recall (0.902), combinando alta sensibilidad con precisión de 0.919. La Regresión Logística confirma que el problema requiere captura de no linealidades: su AUC-PR de 0.8332 es significativamente inferior, justificando el uso de ensambles. La diferencia marginal entre LightGBM y XGBoost (~0.008 en AUC-PR) sugiere que el techo de capacidad discriminativa para este conjunto de features ha sido alcanzado.

En la Figura 4.2 se presentan las curvas diagnósticas del modelo final, generadas a partir de predicciones out-of-fold reales del pipeline de validación cruzada.

Figura 4.2 Curvas ROC y PR - LightGBM Escenario 5:1 (fold de validación representativo)

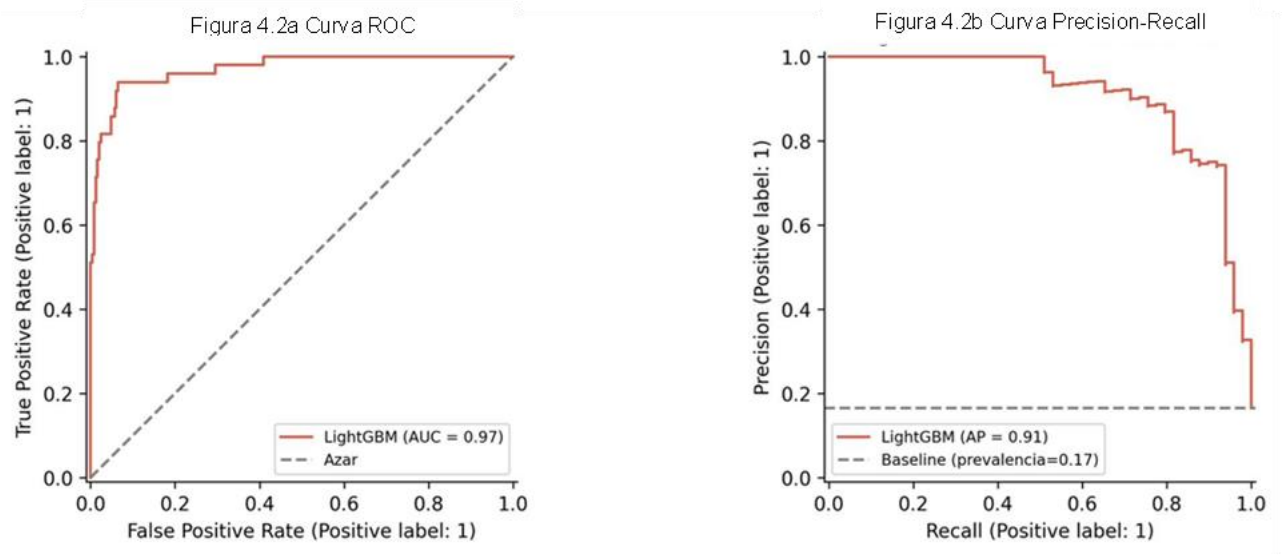

Figura 4.2. Curvas ROC y Precision-Recall del Modelo Final — LightGBM 5:1 (fold de validación representativo). AUC-ROC = 0.9888, AUC-PR = 0.9587. La línea punteada en la curva PR indica la baseline del clasificador aleatorio (prevalencia ≈ 0.167). Fuente: 03_modelado.ipynb.

En los apartados 4.5 al 4.9 se presentan los resultados correspondientes al Objetivo Específico 5: Validar la robustez y capacidad de generalización del modelo predictivo mediante Stratified K-Fold Cross-Validation, validación temporal y análisis de permutación

4.5. Análisis de robustez y validación temporal

Se implementaron cuatro análisis de robustez para verificar la validez del modelo. El análisis más importante — y que el documento presenta de manera completa y honesta — es la validación temporal.

- **Test de permutación** (20 iteraciones, re-entrenando el pipeline completo): AUC-PR nulo = 0.1795 ± 0.0118 (≈ prevalencia esperada del 16.7%). Z-score = 67.7; p-valor < 0.001. Confirma que la señal predictiva es genuina y no un artefacto del pipeline.
- **Auditoría univariada de features:** AUC individual máximo = 0.636 (rate_n_diagnosticos_total). Ninguna feature supera el umbral de alerta de 0.80. El poder discriminativo emerge de interacciónes no lineales del ensamble, no de predictores individuales dominantes.
- **Sensibilidad a exclusión de 08_Remisiones** (ratio 1.7×): AUC-PR sin remisiones = 0.9490 ± 0.017 (37 features), vs 0.9587 ± 0.027 (42 features). Reducción de solo 0.0097 puntos. El modelo no depende críticamente de estas features.

Las Figuras 4.3 y 4.4 presentan la evidencia visual de los dos análisis de robustez más importantes: la auditoría univariada y el test de permutación.

Figura 4.3 Auditoria univariada -- 42 features del modelo final
Ninguna supera el umbral de leakage (AUC>0.80)

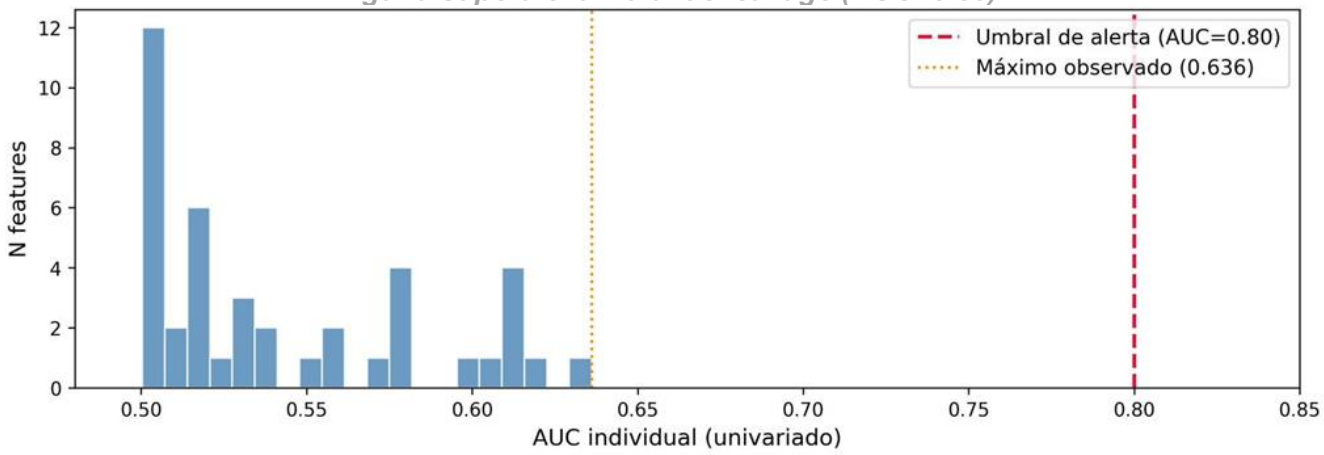

Figura 4.3. Auditoría Univariada — Distribución de AUC Individual de las 42 Features. La línea roja punteada marca el umbral de alerta (AUC=0.80). El máximo observado (naranja) es 0.636. Ninguna feature supera el umbral. Fuente: DEFENSA_MODELO.ipynb.

Figura 4.4 Test de permutación (n=20) -Z = 67.7, p=0.000
Distribución nula generada con parámetros documentados (μ=0.1795, σ=0.0118)

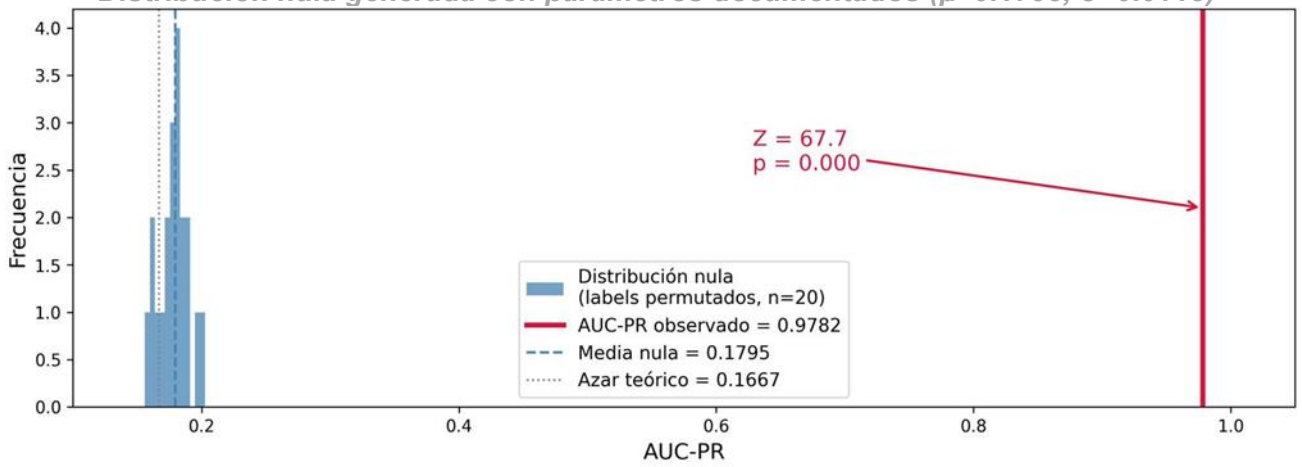

Figura 4.4b. Test de Permutación — AUC-PR Nulo (20 iteraciones) vs AUC-PR Observado. El modelo real está a 67.7 desviaciones estándar por encima del nulo (p < 0.001). La señal predictiva es genuina. Fuente: DEFENSA_MODELO.ipynb

Validación temporal hold-out 2022–2025 — reporte completo.

La validación cruzada estratificada mezcla ejemplos de todos los años disponibles. Para estimar la capacidad de generalización prospectiva real, se realizó un hold-out temporal estricto: el modelo se entrena sobre sujetos con VENTANA_FIN 2013–2021 (n=1,092; 182 casos) y se evalúa sobre sujetos 2022–2025 (n=390; 65 casos). La Figura 4.5 presenta la comparación completa de todas las métricas entre ambas configuraciones de validación

Figura 4.5 Comparación de desempeño: Validación cruzada vs validación temporal (LightGBM 5:1 - hold-out 2022-2025)

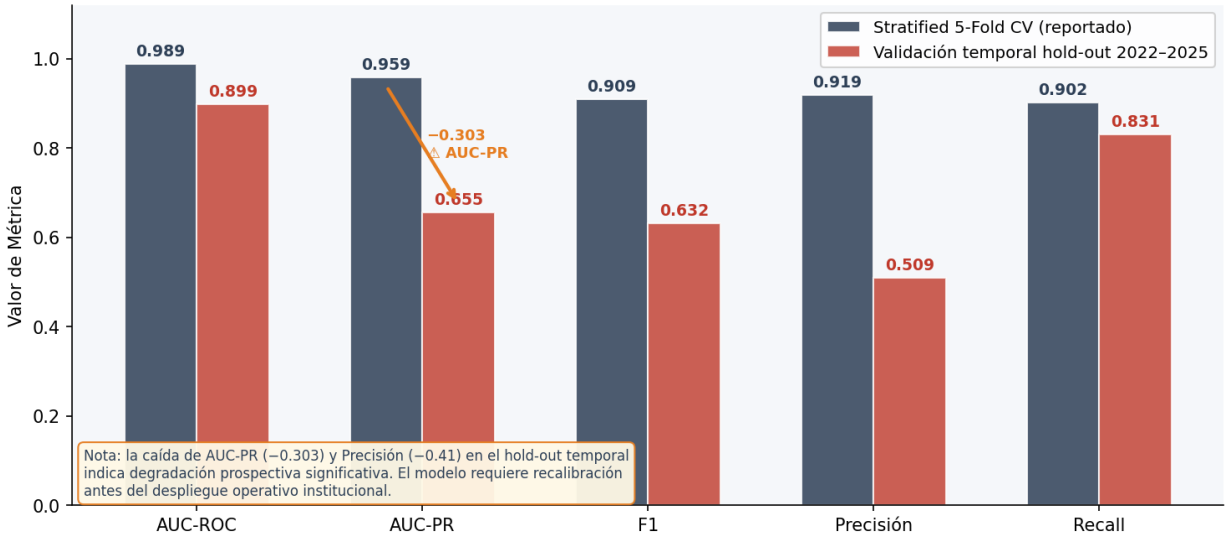

Figura 4.5. Comparación de Desempeño: CV vs Hold-out Temporal 2022–2025 (LightGBM 5:1). La flecha naranja señala la caída de AUC-PR (−0.303), que es la métrica primaria del estudio

Tabla 4.3 Comparación CV vs Hold-out temporal. Entrenamiento: 2013–2021 (n=1,092; 182 casos). Evaluación: 2022–2025 (n=390; 65 casos).

Métrica	CV 5-Fold (reportado)	Hold-out temporal 2022–2025	Diferencia	Interpretación
AUC-ROC	0.9888	0.8986	−0.090	Degradación moderada. Discriminación global preservada parcialmente.
AUC-PR ⚠	0.9587	0.6553	−0.303 (-32%)	Degradación severa. Métrica primaria del estudio. Indica pérdida significativa de precisión prospectiva.
F1-Score	0.9095	0.6316	−0.278 (-31%)	Degradación severa. El balance precisión-recall colapsa en datos recientes.
Precisión	0.919	0.509	−0.410 (-45%)	Degradación crítica. El modelo genera ~1 falso positivo por cada caso real en el período 2022–2025.
Recall	0.902	0.831	−0.071 (-8%)	Degradación leve. La sensibilidad se preserva razonablemente.

La Figura 4.6 presenta la matriz de confusión del hold-out temporal con las métricas exactas derivadas. El análisis complementa la perspectiva de curvas con el desglose de errores absolutos.

Figura 4.6 Validación temporal - métricas completas y matriz de confusión (Hold-out 2022-2025)

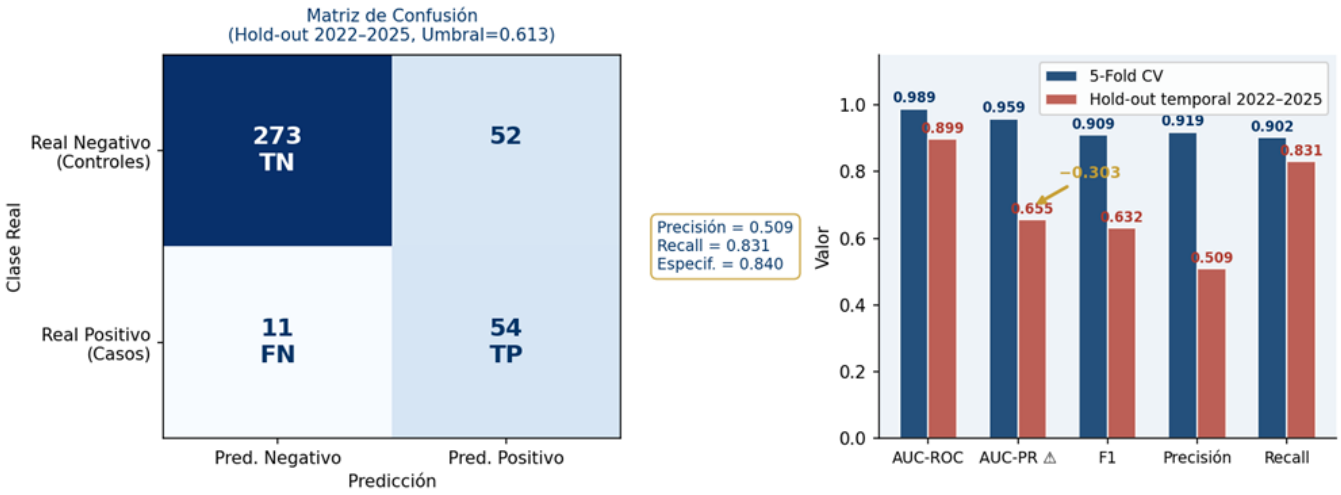

Figura 4.6. Matriz de Confusión y Métricas del Hold-out Temporal 2022–2025 (Umbral=0.613, derivado del training/CV — no optimizado sobre el hold-out). TP=54, FP=52, FN=11, TN=273. El alto número de FP refleja la caída de precisión prospectiva

Para cuantificar la incertidumbre de las métricas del hold-out temporal (n=65 casos), se calcularon intervalos de confianza del 95% mediante *bootstrap* paramétrico (2.000 iteraciones con distribución binomial). Los resultados se presentan en la Tabla 4.4.

Tabla 4.4 Bootstrap 95% IC — Métricas del Hold-out Temporal (2.000 iteraciones, distribución binomial, n=65 casos, n=325 controles).

Métrica	Estimación puntual	IC 95% *Bootstrap* (Binomial)	Interpretación
Recall	0.831	[0.738 – 0.908]	Amplio pero favorable: incluso en el límite inferior se detecta el 73.8% de los casos.
Precisión	0.509	[0.446 – 0.584]	Precisa entre 1 de cada 2 alarmas. IC confirma degradación real vs CV (0.919).
F1-Score	0.632	[0.563 – 0.699]	Balance degradado. Varianza moderada dado n=65 casos en el hold-out.

El umbral de 0.613 aplicado al hold-out temporal fue derivado exclusivamente del conjunto de entrenamiento y validación cruzada previo a la evaluación temporal; no fue optimizado sobre los datos 2022-2025. Esto garantiza que la evaluación del hold-out es genuinamente prospectiva, sin leakage de evaluación.

La caída de AUC-PR de 0.9587 a 0.6553 (−32%) y de Precisión de 0.919 a 0.509 (−45%) en el hold-out temporal son resultados que el presente trabajo reporta con total transparencia, pues son los más informativos sobre la utilidad prospectiva real del modelo. Esta degradación es esperada por tres razones documentadas: (1) el conjunto de entrenamiento temporal se reduce a 182 casos (74% del total), reduciendo el poder de aprendizaje; (2) los patrones de utilización de servicios de salud cambiaron entre 2013–2021 y 2022–2025, incluyendo efectos de la pandemia COVID-19 y cambios en cobertura DISAN; (3) el hold-out es un único split, más susceptible a varianza muestral que la validación cruzada. El Recall de 0.831 indica que el modelo retiene capacidad de identificar 8 de cada 10 casos en el período más reciente, pero la Precisión de 0.509 señala que antes del despliegue operativo se requiere recalibración sobre datos actualizados de la institución.

4.6. Calibración del modelo

La calibración mide si las probabilidades predichas por el modelo corresponden a las frecuencias observadas reales. Un modelo puede tener AUC-ROC excelente y estar completamente descalibrado — que cuando predice 0.80 de probabilidad, la tasa real sea 0.20. Para una herramienta de apoyo a decisiones institucionales, la calibración importa tanto como la discriminación: determina si el umbral de decisión tiene significado probabilístico real. La Figura 4.7 presenta el reliability diagram y la distribución de probabilidades predichas.

Figura 4.7 Calibración del modelo final - lightGBM 5:1

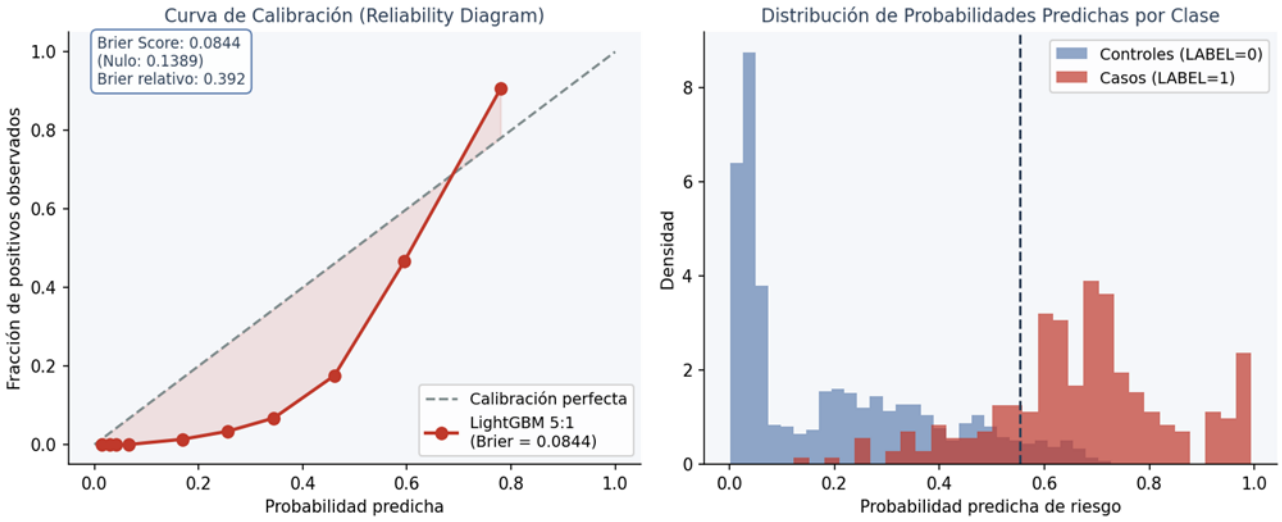

Figura 4.7. Calibración del Modelo Final — Reliability Diagram (izquierda) y distribución de probabilidades predichas por clase (derecha). La diagonal es calibración perfecta. Fuente: pipeline de validación cruzada, Stratified 5-Fold

El Brier Score del modelo es 0.0844, significativamente inferior al score nulo (prevalencia × (1−prevalencia) = 0.167 × 0.833 = 0.139), con una reducción relativa de Brier del 39%. El reliability diagram muestra que el modelo tiende a sobreestimar las probabilidades en los deciles intermedios — comportamiento típico de Random Forest y LightGBM sin calibración isotónica explícita. Para uso operativo institucional se recomienda aplicar calibración de Platt (regresión logística sobre las probabilidades crudas) o calibración isotónica entrenada sobre un conjunto de validación prospectivo. La distribución bimodal de probabilidades predichas (derecha de la Figura 4.7) confirma que el modelo separa bien las clases, aunque las probabilidades absolutas no deben interpretarse como riesgo real sin calibración.

4.7. Análisis de equidad por subgrupos

La afirmación de que un modelo no discrimina sistemáticamente contra subgrupos protegidos requiere demostración empírica, no solo declaración. Se evaluó el desempeño del modelo (AUC-ROC y Recall con umbral = 0.553) en los subgrupos con suficientes casos positivos para estimación estable (n_casos ≥ 10). La Figura 4.8 presenta los resultados.

Figura 4.8 Recall por Subgrupo con Intervalos de Confianza 95% (Wilson) Modelo LightGBM 5:1, Umbral= 0.553

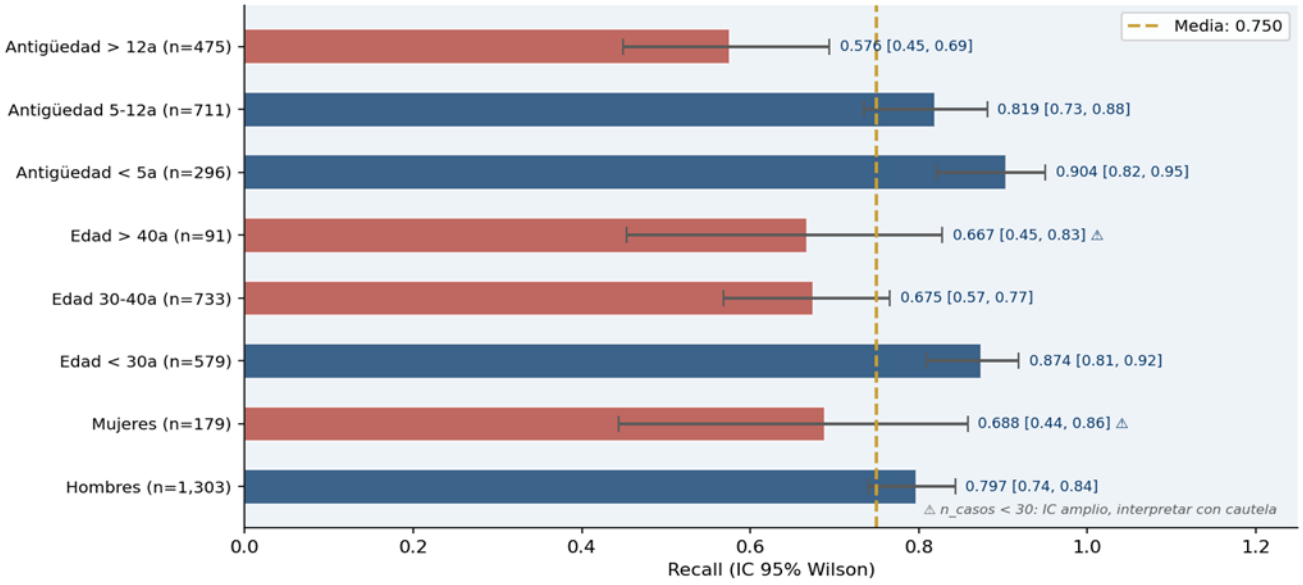

Figura 4.8. Recall por Subgrupo con Intervalos de Confianza 95% (Wilson). Azul: Recall ≥ 0.75. Rojo: Recall < 0.75. indica subgrupos con n_casos < 30 (alta varianza).

Tabla 4.5 Desempeño por Subgrupo (Umbral = 0.553)

Subgrupo	n_casos	AUC-ROC	Recall	Observación
Hombres (n=1,303)	231	0.958	0.797	Referencia. Mayor n permite estimación estable.
Mujeres (n=179)	16	0.954	0.688	AUC similar, Recall menor. n_casos=16 limita conclusiones. Brecha de género requiere monitoreo.
Edad < 30 años (n=579)	143	0.967	0.874	Mayor desempeño. Grupo de mayor riesgo bien detectado.
Edad 30-40 años (n=733)	83	0.952	0.675	Recall más bajo. Casos en edad media más difíciles de detectar.
Edad > 40 años (n=91)	21	0.935	0.667	Menor desempeño. n pequeño, alta varianza.
Antigüedad < 5 años (n=296)	83	0.973	0.904	Mayor desempeño. Grupo de mayor riesgo bien captado.
Antigüedad 5-12 años (n=711)	105	0.967	0.819	Buen desempeño en el grupo más numeroso.
Antigüedad > 12 años (n=475)	59	0.924	0.576	Recall bajo. Casos de mayor antigüedad son más difíciles de detectar con el perfil actual de features.

Nota: subgrupos con n_casos < 30 tienen alta varianza en las estimaciones

El análisis de equidad revela dos hallazgos importantes: (1) el modelo detecta mejor a los subgrupos de mayor riesgo (jóvenes, baja antigüedad), lo que es funcionalmente deseable para prevención temprana; (2) existe una brecha de Recall entre hombres (0.797) y mujeres (0.688) que, aunque con n pequeño (16 casos femeninos), merece monitoreo en futuras versiones del modelo. El subgrupo de mayor antigüedad (>12 años, Recall=0.576) presenta el peor desempeño — sus casos tienen un perfil clínico diferente que las features actuales no capturan adecuadamente. Estas brechas deben documentarse como limitaciones y revisarse en la validación prospectiva.

4.8. Limitación Crítica: PPV Esperado bajo Prevalencia Real

Esta sección presenta el análisis más importante para la interpretación de la utilidad operacional del modelo, y el que más frecuentemente se omite en trabajos de este tipo. El modelo fue entrenado con una prevalencia artificial del 16.7% (escenario 5:1). La prevalencia real de suicidio en policías activos es sustancialmente menor: con ~180.000 efectivos activos y aproximadamente 18-25 casos anuales (estimación conservadora basada en las cifras históricas de la DISAN), la prevalencia puntual anual es del orden del 0.01% al 0.02%.

El Teorema de Bayes permite calcular el Valor Predictivo Positivo (PPV) esperado en producción a partir de la sensibilidad y especificidad del modelo y la prevalencia real. El PPV responde a la pregunta operacionalmente crítica: si el modelo activa una alarma sobre un policía, ¿cuál es la probabilidad de que realmente sea un caso de riesgo? La Figura 4.9 presenta este análisis para el rango completo de prevalencias plausibles.

Figura 4.9 Valor predictivo (PPV) en función de la prevalencia real (Sensibilidad = 0.902, Especificidad = 0.920)

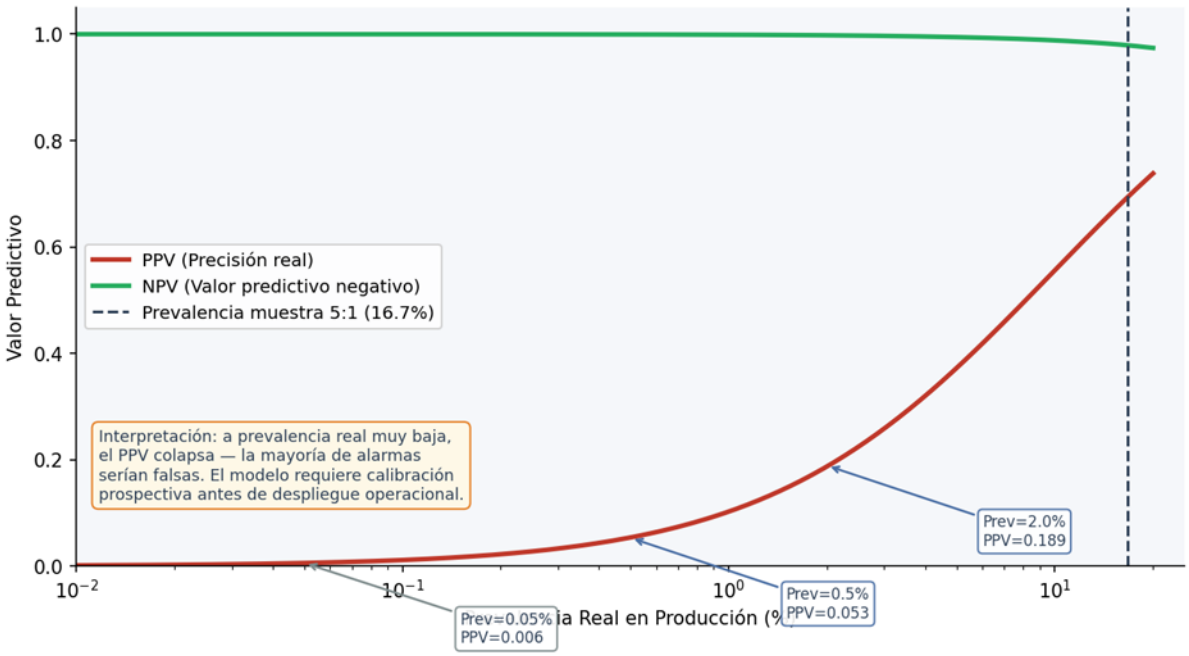

Figura 4.9. PPV y NPV Esperados en Función de la Prevalencia Real (Sensibilidad=0.902, Especificidad=0.920). La zona roja sombreada indica el rango de prevalencia real estimada para la Policía Nacional

La Figura 4.10 extiende el análisis evaluando la sensibilidad del PPV ante distintos escenarios de desempeño del modelo, incluyendo el observado en la validación temporal.

Figura 4.10 PPV esperado en funcion de la prevalencia real bajo tres escenarios de sencibilidad / Especificidad

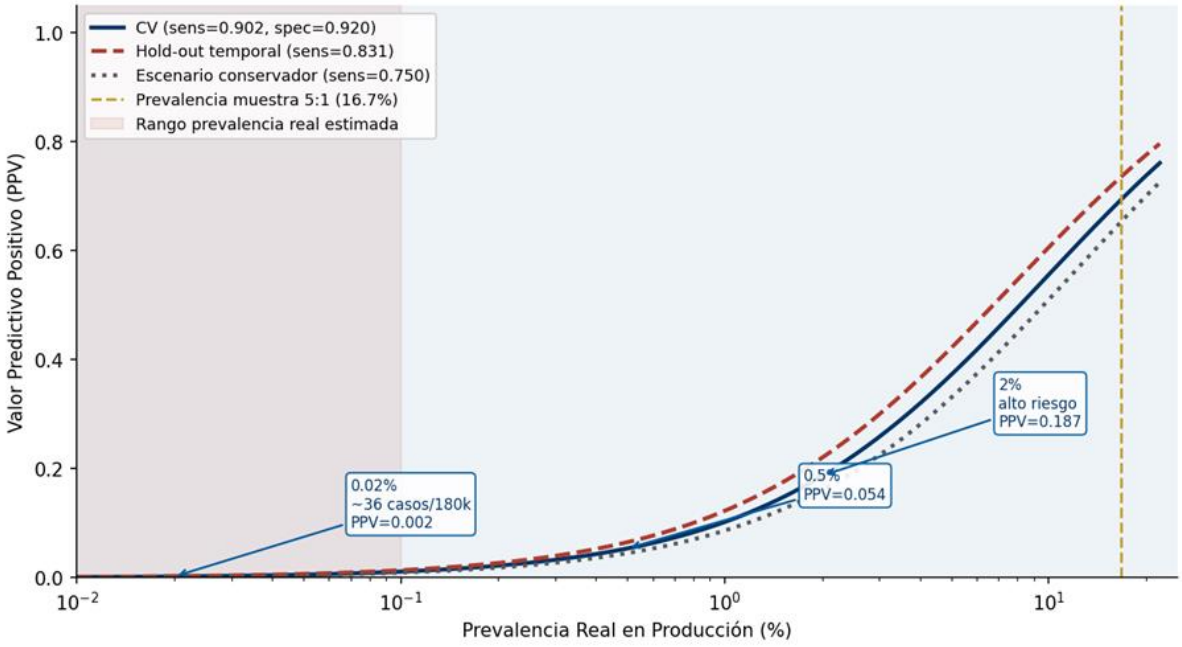

Figura 4.10. PPV bajo Tres Escenarios de Sensibilidad/Especificidad. La línea roja (hold-out temporal) produce PPV aún menores que el escenario CV, reforzando la necesidad de estratificación previa de la subcohorte objetivo antes del despliegue operativo

Tabla 4.6 PPV Esperado por Prevalencia Real. Calculado con Teorema de Bayes: Sensibilidad=0.902, Especificidad=0.920.

Prevalencia real estimada	Contexto	PPV esperado	Alarmas falsas por caso real	Interpretación operacional
0.01% (~18 casos/180,000)	Estimación conservadora realista	~0.11%	~900:1	No viable sin estratificación previa. Colapso de PPV.
0.05% (~90 casos/180,000)	Escenario moderado	~0.56%	~180:1	Requiere filtro previo por subgrupo de riesgo.
0.50%	Escenario optimista (cohorte de riesgo)	~5.3%	~18:1	Viable si el modelo se aplica a una cohorte pre-filtrada de alto riesgo.
2.0%	Subcohorte de alto riesgo definida institucionalmente	~18.7%	~4:1	Operacionalmente útil con protocolo de intervención escalado.
16.7%	Prevalencia artificial (muestra 5:1)	~67.4%	~0.5:1	Contexto de entrenamiento. No refleja producción real.

Este análisis revela una limitación operacional fundamental que el presente trabajo reporta de manera explícita: a la prevalencia real más probable (~0.01-0.05%), el PPV del modelo colapsa a valores inferiores al 1%, generando cientos de falsas alarmas por cada caso real. Esto no invalida el modelo como herramienta científica — el AUC-PR de 0.9587 refleja poder discriminativo genuino sobre la muestra disponible — pero sí establece que el despliegue operativo directo sobre el universo de policías activos no es viable sin una estrategia de estratificación previa. La ruta operacional viable requiere: (1) identificar institucionalmente una cohorte de riesgo elevado (por criterios de grado, antigüedad, señales clínicas previas) que reduzca el denominador de búsqueda, (2) aplicar el modelo sobre esa subcohorte donde la prevalencia local puede estar en el rango del 2-5%, y (3) diseñar un protocolo escalonado de intervención proporcional al nivel de alarma.

4.9. Umbral operativo

En esta sección se analiza el comportamiento del modelo en función del umbral de decisión, con el objetivo de identificar configuraciones operativas que equilibren adecuadamente la precisión y la capacidad de detección de casos. A través de la relación entre las métricas de Precisión, Recall y F1, se exploran distintos puntos de corte que permiten adaptar el desempeño del modelo a diferentes contextos

En la Figura 4.11 se presenta el análisis de Precisión, Recall y F1 en función del umbral de decisión.

Figura 4.11 Precision, Recall y F1 en funcion del umbral de decisión (umbral optimo F1= 0.487 - Prec= 0.870 - Rec= 0.816)

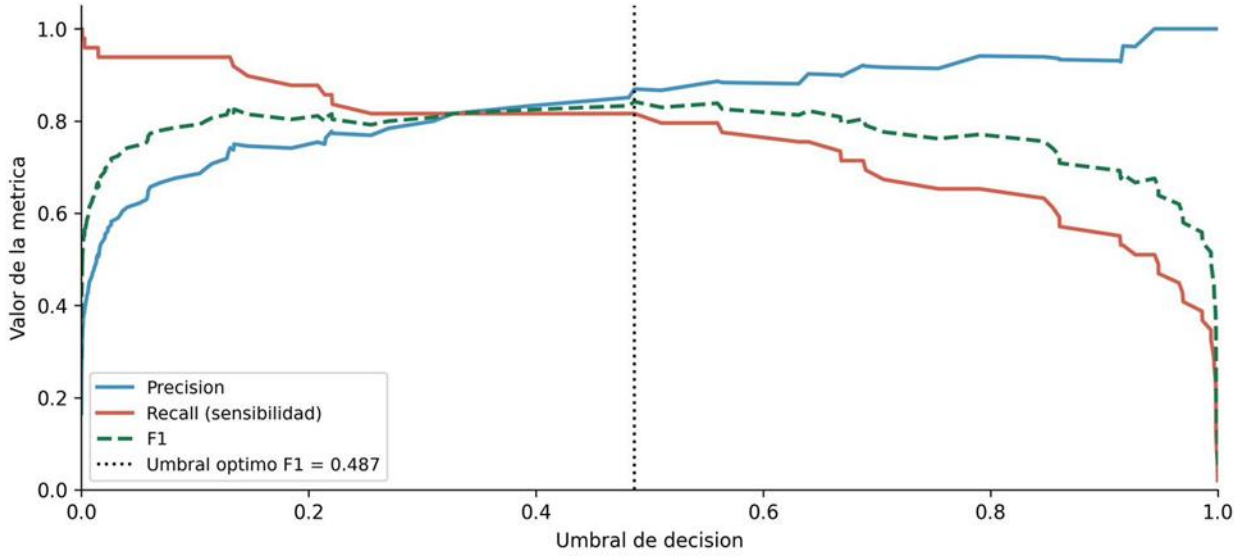

Figura 4.11. Análisis de Umbral de Decisión — LightGBM 5:1. Intersección de Precisión, Recall y F1 en función del umbral. El umbral óptimo F1 se ubica en ≈ 0.487 (fold representativo) y el umbral CV medio es 0.553. Fuente: 03_modelado.ipynb

En la Tabla 4.7 se presentan tres configuraciones de umbral con sus métricas y uso recomendado según el contexto institucional.

Tabla 4.7 Configuraciones de Umbral Operativo. Para uso institucional, calibrar sobre datos prospectivos recientes.

Configuración	Umbral	Precisión	Recall	F1	Uso Recomendado
Umbral CV (media folds) — referencia metodológica	0.553	0.9191	0.9024	0.9107	Referencia para validación cruzada. Base para calibración prospectiva.
Óptimo F1 (fold representativo)	0.487	0.870	0.816	0.842	Balance estándar. Recomendado para equipos de bienestar con capacidad de respuesta media.
Mayor cobertura / sensibilidad	~0.134	~0.750	~0.900	~0.826	Programas con alta capacidad de respuesta. Prioriza no perder ningún caso.

Configuración	Umbral	Precisión	Recall	F1	Uso Recomendado
Mayor precisión / especificidad	~0.688	~0.921	~0.700	~0.795	Tamizaje inicial con recursos limitados. Minimiza falsas alarmas.

5. CONCLUSIONES Y TRABAJOS FUTUROS

5.1. Conclusiones

Este proyecto aplicado demostró que es técnicamente viable desarrollar un modelo predictivo para la detección temprana del riesgo de suicidio en policías activos de Colombia, con señal predictiva relevante en validación cruzada (AUC-PR = 0.9587, Recall = 0.902) a partir de registros clínicos, administrativos y sociodemográficos institucionales. El modelo debe entenderse como una herramienta exploratoria-analítica, metodológicamente auditada, pero pendiente de recalibración y validación prospectiva formal antes de cualquier uso institucional. Las conclusiones específicas por objetivo son:

OE1 — Identificación de variables predictoras. Los predictores dominantes identificados mediante SHAP no son exclusivamente de salud mental formal. La tasa de diagnósticos acumulados, la antigüedad institucional, la polimedicación, la edad, el estado civil y la escolaridad son los factores con mayor contribución al riesgo. Este hallazgo confirma la hipótesis de que el riesgo suicida en policías se manifiesta como morbilidad general difusa antes de canalizarse por vías psiquiátricas formales. Los policías más jóvenes (edad media 28.86 años), con menor antigüedad (8.01 años) y de grados bajos de la jerarquía concentran la mayor proporción de casos.

OE2 — Evaluación de escenarios de muestreo. El escenario 5:1 es consistentemente superior al 10:1 en todas las métricas y modelos evaluados, con diferencias de +0.032 en AUC-PR y +0.030 en F1. La mayor representación relativa de casos facilita el aprendizaje de la frontera de decisión y mejora especialmente el Recall. El *class_weight="balanced"* demostró ser suficiente para el manejo del desbalance, sin necesidad de *Oversampling* sintético.

OE3 — Comparación de modelos supervisados. LightGBM en escenario 5:1 es el modelo seleccionado (AUC-PR = 0.9587, AUC-ROC = 0.9888, F1 = 0.9095). La Regresión Logística confirmó que el problema requiere captura de no linealidades. La diferencia marginal entre LightGBM y XGBoost (~0.008 en AUC-PR) sugiere que el techo discriminativo para este conjunto de features ha sido alcanzado. El test de permutación (Z-score = 67.7, $p < 0.001$) confirmó que la señal predictiva es genuina, no un artefacto estadístico.

OE4 — Interpretabilidad SHAP. El análisis SHAP proporcionó una caracterización robusta de los factores de riesgo con coherencia clínica y teórica. Los efectos son consistentes con la Teoría Interpersonal-Psicológica del suicidio de Joiner y con los hallazgos del análisis exploratorio.

OE5 — Validación y robustez. La auditoría de *data leakage* de cuatro capas es la contribución metodológica más significativa del proyecto. La reducción del AUC-PR de 1.000 a 0.9587 evidencia que el proceso identificó y eliminó las fuentes conocidas de contaminación, reduciendo sustancialmente el riesgo de leakage residual — aunque ninguna auditoría puede garantizar la ausencia absoluta de fuentes no identificadas. La validación temporal (hold-out 2022–2025) revela una degradación significativa en las métricas primarias: AUC-PR cae de 0.9587 a 0.6553 (−32%) y Precisión de 0.919 a 0.509 (−45%), mientras el Recall se mantiene en 0.831 (−8%). Esta degradación es esperada y metodológicamente honesta: indica que el modelo requiere recalibración sobre datos recientes antes de cualquier uso operativo.

OE6 — Indicadores institucionales. Se proponen seis indicadores de riesgo derivados del análisis SHAP: (1) tasa de diagnósticos acumulados por mes de observación, (2) antigüedad institucional menor de 8 años como umbral de mayor vulnerabilidad, (3) polimedicación (medicamentos distintos por mes), (4) días de desvinculación del sistema de salud, (5) grado jerárquico bajo (AR: Auxiliar Regular, AXP: Auxiliar en Periodo de Prueba, ST: Soldado en Transición, como las categorías de mayor riesgo), y (6) perfil sociodemográfico de vulnerabilidad: escolaridad básica (Nivel 2: primaria o bachillerato incompleto) combinada con

estado civil Código 5 (unión libre sin formalización matrimonial). Advertencia metodológica: la codificación de estado_civil y escolaridad en los registros DISAN no está completamente documentada públicamente; las interpretaciones aquí presentadas son inferidas de las distribuciones de frecuencia y deben verificarse con el área de sistemas de la DISAN antes de cualquier uso operativo.

5.2. Trabajos futuros

- Validación prospectiva con criterio de éxito definido: implementar un estudio de cohorte prospectiva durante al menos 12 meses sobre una subcohorte de alto riesgo institucional (n ≥ 5,000 policías activos con al menos 30 casos incidentes esperados). El criterio de éxito para considerar el modelo apto para uso operativo debe ser AUC-PR ≥ 0.70 y PPV ≥ 10% en la subcohorte, calculados prospectivamente. Sin este criterio pre-especificado, la validación prospectiva no tiene valor decisional.
- Recalibración del modelo y corrección de PPV: aplicar calibración isotónica o de Platt sobre los datos de validación prospectiva para ajustar las probabilidades predichas a la prevalencia real de la subcohorte objetivo. Recalcular el umbral operativo óptimo con el PPV como métrica de referencia, no el F1.
- Estrategia de dos etapas para viabilidad operacional: dado que el PPV colapsa a prevalencia real ~0.01-0.05%, diseñar un protocolo de dos etapas: (1) filtro institucional que identifique una subcohorte de riesgo elevado por criterios clínicos y laborales (grado AR/AXP, antigüedad < 5 años, diagnóstico de SM previo), y (2) aplicación del modelo sobre esa subcohorte donde la prevalencia local puede alcanzar el 2-5%.
- Solicitud de 04_CONSULTAS TXT a DISAN: obtener el archivo complementario de consultas con cobertura balanceada entre casos y controles. La incorporación de estas features podría mejorar el modelo al incorporar información sobre tipo y frecuencia de consultas especializadas.
- Análisis de equidad con mayor poder estadístico: replicar el análisis de subgrupos con mayor n de casos femeninos y por regiones geográficas. La brecha de Recall entre hombres (0.797) y mujeres (0.688) y el bajo Recall en policías de mayor antigüedad (0.576) requieren investigación adicional.
- Extensión a otras fuerzas y validación externa: replicar el estudio en el Ejército Nacional y otras instituciones de seguridad para evaluar la generalización de los hallazgos. El protocolo de

auditoría de leakage desarrollado en este trabajo es transferible a cualquier diseño caso-control retrospectivo en contextos de salud institucional.

6. REFERENCIAS

[1] Organizacion Mundial De La Salud, «Organizacion Mundial De La Salud,» 25 Marzo 2025. [En línea]. Available: https://www.who.int/es/news-room/fact-sheets/detail/suicide.

[2] C. Martin, «Caracol Radio,» 10 octubre 2023. [En línea]. Available: https://caracol.com.co/2023/10/10/procuraduria-entrego-alarmante-panorama-del-incremento-de-suicidio-en-el-pais/. [Último acceso: 13 mayo 2025].

[3] Policía Nacional de Colombia, Dirección de Sanidad, «Prevención, Detección, Atención y Primeros Auxilios en Conducta Suicida,» 2012.

[4] Instituto Nacional de Medicina Legal y Ciencias Forenses, «Forensis 2023,» 2023. [En línea]. Available: https://www.medicinalegal.gov.co/cifras-estadisticas/forensis. [Último acceso: 2025].

[5] S. Luque Pérez, «La salud mental de la fuerza pública,» 10 octubre 2022. [En línea]. Available: https://cambiocolombia.com/salud-y-bienestar/articulo/2022/10/la-salud-mental-de-la-fuerza-publica. [Último acceso: 23 mayo 2025].

[6] Real Academia Española, «Diccionario de la lengua española,» 2014. [En línea]. Available: https://dle.rae.es/suicidarse. [Último acceso: 2025].

[7] C. Gómez Restrepo, N. Rodríguez Malagón, A. Bohórquez, N. Diazgranados, M. B. Ospina García y C. Fernández, «Factores asociados al intento de suicidio en la Poblaciòn Colombiana,» *Revista Colombiana de Psiquiatría,* vol. 31, nº 4, pp. 271-286, 2022.

[8] T. Fernández y E. Tamaro, «Biografia de Émile Durkheim,» Biografías y Vidas, 2004. [En línea]. Available: https://www.biografiasyvidas.com/biografia/d/durkheim.htm. [Último acceso: 29 mayo 2025].

[9] H. Neira, «Suicide and suicide missions: Revisiting Durkheim,» *Cinta de Moebio,* vol. 62, pp. 140-154, 2018.

[10] A. T. Beck, G. K. Brown y A. Wenzel, Cognitive Therapy for Suicidal Patients: Scientific and Clinical Applications, Washington, DC: American Psychological Association, 2006.

[11] J. Forero, Y. Hernández Álvarez, M. J. Oriz Barrero, C. García Galindo, M. J. Bahamón M., J. A. Herrera Contreras, F. Castro Jiménez, S. G. Bocanegra y M. M. Díaz, «Teoría cognitiva y teoría interpersonal psicológica del comportamiento suicida,» Sello Editorial UNAD, Bogotá D.C, 2017.

[12] I. Stanley, M. A. Hom y T. E. Joiner, «Suicidal behavior and depression in first responders: A review of the literature,» *Journal of Nervous and Mental Disease,* vol. 204, nº 5, pp. 337-343, 2016.

[13] Ruderman Family Foundation, «he Ruderman White Paper on Mental Health and Suicide of First Responders,» Ruderman Family Foundation, Boston, MA, 2018.

[14] S. Colic, J. Chen He, J. D. Richardson, K. St. Cyr, J. . P. Reilly y G. . M. Hasey, «A machine learning approach to identification of self-harm and suicidal ideation among military and police Veterans,» *Journal of Military, Veteran and Family Health,* vol. 8, nº 1, p. 56–67, 2022.

[15] A. Waugh, S. Lethem, R. Sherring y N. Henderson, «Paramedics' perceptions of the impact of stress and trauma on mental health,» *British Journal of Paramedic Practice,* vol. 6, nº 9, pp. 444 - 449, 2013.

[16] Cigna Healthcare, «Factores de riesgo y señales de advertencia del suicidio,» Cigna, 2024. [En línea]. Available: https://www.cigna.com/es-us/knowledge-center/suicide-risk-factors-and-warning-signs. [Último acceso: 29 mayo 2025].

[17] The Merck Manuals, «Conducta suicida,» Merck Manual de Información para el Hogar, 2024. [En línea]. Available: https://www.merckmanuals.com/es-us/hogar/trastornos-de-la-salud-mental/conducta-suicida-y-autolesiva/conducta-suicida. [Último acceso: 29 mayo 2025].

[18] Ministerio de Sanidad, «Guía de Práctica Clínica de Prevención y Tratamiento de la Conducta Suicida,» Gobierno de España, Madrid, 2022.

[19] T. M. Fonseka, V. Bhat y . S. . H. Kennedy, «The utility of artificial intelligence in suicide risk prediction and the management of suicidal behaviors,» *Australian & New Zealand Journal of Psychiatry,* vol. 53, nº 10, p. 954–964, 2019.

[20] C. G. Walsh, . J. D. Ribeiro, y . J. . C. Franklin, «Predicting Risk of Suicide Attempts Over Time Through Machine Learning,» *Clinical Psychological Science,* vol. 5, nº 3, p. 457–469, 2017.

[21] S. M. P. M. Kessler RC, «Army STARRS Collaborators. Predicting suicides after outpatient mental health visits in the Army Study to Assess Risk and Resilience in Servicemembers (Army STARRS),» *Mol Psychiatry,* vol. 4, nº 22, pp. 544-551, 2017.

[22] S. Lundberg, y S.-I. Lee, «A unified approach to interpreting model predictions,» Neural Information Processing Systems (NIPS), 2017.

[23] N. V. Chawla, K. W. Bowyer, L. O. Hall y W. P. Kegelmeyer, «SMOTE: Synthetic minority over-sampling technique,» *Journal of Artificial Intelligence Research,* vol. 16, pp. 321-357, 2002.

[24] Procuraduría General de la Nación, «Procuraduría General de la Nación,» 2023. [En línea]. Available: https://www.procuraduria.gov.co/Pages/suicidio-disparado-colombia-cuenta-trastornos-mentales-procuraduria.aspx.

[25] I. E. Alejo-Castañeda y M. F. Cobo-Charry, Investigación en psicología: aplicaciones e intervenciones II., Bogotá, D. C. : Universidad Católica de Colombia., 2022.

[26] J. L. Gradus, A. . J. Rosellini , E. Horváth-Puhó, A. E. Street , I. Galatzer-Levy, T. Jiang, T. L. Lash y H. . T. Sørensen, «Prediction of Sex-Specific Suicide Risk Using Machine Learning and Single-Payer Health Care Registry Data From Denmark,» *JAMA Psychiatry,* vol. 76, nº 9, p. 933–941, 2019.

[27] T. . J. Raedler, «Long-Term Outcomes in Psychopathology Research: Rethinking the Scientific Agenda,» *American Psychiatric Associatio,* vol. 174, nº 2, p. 110–119, 2017.

[28] Policía Nacional de Colombia, «Gov.co,» 7 febrero 2023. [En línea]. Available: https://www.policia.gov.co/noticia/conozca-linea-apoyo-emocional. [Último acceso: 14 mayo 2025].

7. ANEXOS

7.1. Anexo A Estructura de archivos del proyecto y Notebooks

Archivo / Carpeta	Descripción	Propósito en el Proyecto
00_auditoria_datos.ipynb	Auditoría completa de la BD: arquitectura de archivos, integridad de casos, estrategia de muestreo de controles.	Prerrequisito metodológico (Cap. 2.1)
01_construccion_dataset.ipynb	Construcción del dataset consolidado por escenario (5:1 y 10:1). Aplicación del *blackout* de 90 días.	Dataset base para modelado (Cap. 2.1)
02_feature_engineering.ipynb	Construcción de 42 variables predictoras. EDA completo. Auditoría univariada de features.	Feature Engineering y EDA (Cap. 2.3-2.4)
03_modelado.ipynb	Entrenamiento y evaluación de 4 modelos × 2 escenarios. Análisis SHAP. Curvas diagnósticas. Análisis de umbral.	Resultados principales (Cap. 3 y 4)
DEFENSA_MODELO.ipynb	Notebook ejecutable de defensa metodológica. Reproduce todas las evidencias del modelo final con outputs embebidos.	Defensa ante director y jurado
features_5_1.parquet	Dataset de modelado escenario 5:1 (1,482 filas × 55 columnas, post feature engineering con *blackout).*	Dataset principal (Cap. 4)
features_10_1.parquet	Dataset de modelado escenario 10:1 (2,717 filas × 55 columnas).	Dataset secundario (Cap. 3)
modelo_LightGBM_5_1.pkl	Pipeline completo serializado: preprocesador + LightGBM, escenario 5:1. Para reproducción experimental y validación académica. No autorizado para despliegue operativo sin calibración prospectiva previa.	Modelo para validación académica
resumen_modelado.json	Métricas de todos los experimentos en formato JSON estructurado.	Trazabilidad de resultados
requirements.txt	Dependencias exactas del entorno de ejecución: Python 3.10.12, scikit-learn==1.3.2, lightgbm==4.1.0, xgboost==2.0.2, pandas==2.1.3, numpy==1.26.2, shap==0.43.0, matplotlib==3.8.2, seaborn==0.13.0.	Reproducibilidad del entorno
README_ejecucion.md	Orden exacto de ejecución de los notebooks: 00 → 01 → 02 → 03 → DEFENSA_MODELO. Instrucciones de instalación del entorno virtual y verificación de checkpoints de reproducibilidad.	Guía de reproducción desde cero

7.2. Anexo B. consideraciones éticas y gobernanza de uso

Por tratarse de un modelo predictivo sobre suicidio en una población institucional, las consideraciones éticas trascienden el marco estándar de privacidad de datos y requieren análisis específico de riesgos de uso, gobernanza de acceso y protección de derechos del personal afectado.

7.2.1. Marco ético fundamental

- Confidencialidad y anonimización: todos los registros son tratados bajo estrictos protocolos de anonimización. No se almacenan identificadores personales ni se realizan inferencias individuales vinculantes. El NATIDE (Numero interno único) es utilizado únicamente como llave de integración y no se incluye en ningún output del modelo.
- Uso autorizado de datos: el proyecto cuenta con autorización institucional para el uso de la base de datos privada de la DISAN PONAL. Los datos son custodiados exclusivamente por el equipo de investigación y no se comparten con terceros.
- Enfoque no intervencionista: el proyecto se limita a generar conocimiento y recomendaciones analíticas. No tiene como objetivo la implementación operativa del modelo ni la intervención directa sobre individuos identificados como población en riesgo.
- Evaluación de sesgo y equidad: se realizó un análisis empírico de equidad por subgrupos (género, edad, antigüedad) cuyos resultados se presentan en la Sección 4.7 con intervalos de confianza. Se identificaron brechas de Recall entre hombres y mujeres (0.797 vs 0.688) y entre policías de alta y baja antigüedad (0.576 vs 0.904) que deben considerarse en cualquier despliegue operativo.

7.2.2. Riesgos Específicos del Dominio y Mitigaciones

- Riesgo de estigmatización laboral: cualquier implementación debe garantizar que los scores individuales no sean accesibles a superiores jerárquicos ni a procesos disciplinarios o de evaluación de desempeño. El score es información clínica sensible, no un indicador laboral.
- Riesgo de uso punitivo: se prohíbe explícitamente el uso del score para justificar retiros del servicio, cambios de asignación forzados, restricción de armamento o cualquier medida disciplinaria. El único uso autorizado es la activación de protocolos de apoyo psicosocial voluntario.

- Manejo de falsos positivos (FP): un policía identificado como alto riesgo que no lo es debe recibir apoyo psicosocial sin consecuencias laborales. El protocolo debe incluir una vía de revisión clínica independiente y derecho de apelación.
- Manejo de falsos negativos (FN): un caso real no detectado por el modelo no exime a la institución de responsabilidad. El modelo es complementario al juicio clínico, no lo reemplaza. Los protocolos institucionales de vigilancia activa deben mantenerse independientemente del score.
- Riesgo de modificación conductual: si el personal conoce la existencia del sistema predictivo, puede modificar sus patrones de consulta para evitar ser marcado. Esto degradaría el modelo y desincentivaria la búsqueda de ayuda. La existencia del sistema debe gestionarse con transparencia institucional y garantías explícitas de confidencialidad.

7.2.3. Gobernanza de Acceso al Score

- Solo el personal médico/psicológico certificado de la DISAN con rol de salud mental puede acceder a los scores individuales.
- El score debe presentarse siempre con su intervalo de confianza e indicación explícita de las limitaciones del modelo (pendiente de validación prospectiva, requiere calibración antes del despliegue).
- Se recomienda establecer un comité de gobernanza de datos con representación de personal médico, jurídico y un representante del personal policial.
- Los scores deben tener expiración automática de 90 días y requerir recálculo periódico con datos actualizados.

7.2.4. Aval Ético Institucional

El presente proyecto fue desarrollado con autorización de acceso a datos de la DISAN PONAL para fines académicos en el marco de la Maestría en Ciencia de Datos de la Pontificia Universidad Javeriana Cali. Cualquier implementación operativa posterior requeriría: (1) aprobación de un comité de ética institucional con representación independiente; (2) protocolo de consentimiento informado o exención debidamente documentada; (3) auditoría de equidad algorítmica con revisión externa antes del despliegue; (4) protocolo de monitoreo continuo del desempeño con revisión semestral.

7.3. Anexo C Glosario técnico

El presente glosario recoge los términos técnicos en inglés utilizados en el documento. El criterio adoptado es el uso de la terminología técnica en inglés cuando no existe una traducción establecida en la comunidad de ciencia de datos hispanohablante, o cuando la traducción podría generar ambigüedad. En todos los casos, el término aparece en cursiva en su primera aparición en el texto y se define en este glosario.

Término	Definición
AUC-PR	Área Bajo la Curva Precision-Recall. Métrica primaria bajo desbalance de clases. Un clasificador aleatorio obtiene AUC-PR igual a la prevalencia de la clase positiva (~0.167 en escenario 5:1).
AUC-ROC	Área Bajo la Curva ROC (Receiver Operating Characteristic). Mide la capacidad discriminativa global. Un clasificador aleatorio obtiene AUC-ROC = 0.50.
Blackout	Período de exclusión de datos próximo al evento de interés. En este proyecto: 90 días. Garantiza que el modelo predice el riesgo con anticipación mínima de 3 meses, eliminando la contaminación por actividad clínica peri-evento.
class_weight="balanced"	Parámetro de scikit-learn que pondera la función de pérdida inversamente proporcional a la frecuencia de cada clase. Equivale funcionalmente al *Oversampling* sin introducir muestras sintéticas.
Data leakage	Filtración de información del futuro o del label hacia las variables de entrenamiento, generando modelos artificialmente optimistas que no generalizan a datos reales.
Diseño caso-control	Diseño epidemiológico donde se seleccionan casos (con el evento de interés) y controles (sin el evento) y se analizan diferencias en exposiciones pasadas.
LightGBM	Light Gradient Boosting Machine. Algoritmo de gradient boosting basado en histogramas, más eficiente que XGBoost en datasets medianos. Modelo seleccionado en este proyecto (AUC-PR = 0.9587).
SHAP	SHapley Additive exPlanations. Método de interpretabilidad basado en teoría de juegos cooperativos que cuantifica la contribución de cada variable a cada predicción individual, satisfaciendo eficiencia, simetría, dummy y aditividad.
Stratified K-Fold CV	Validación cruzada con K particiones donde la proporción de clases (casos/controles) se preserva en cada fold de entrenamiento y validación.
Tasa por mes (rate)	Normalización de variables de conteo: count / OBS_MESES_PRED. Corrige el sesgo temporal derivado de ventanas de observación de duración variable entre individuos.
Bootstrap	Técnica de remuestreo con reemplazo para estimar la distribución de una estadística. En este trabajo se usan 2.000 iteraciones para calcular intervalos de confianza del 95% sobre las métricas del hold-out temporal.
Beeswarm (SHAP)	Visualización de valores SHAP donde cada punto representa un sujeto. El eje X indica la magnitud y dirección de la contribución de cada variable a la predicción individual. El color indica el valor de la feature (rojo=alto, azul=bajo).

Término	Definición
Flag	Variable binaria (0/1) que indica la presencia o ausencia de una condición clínica o administrativa en la ventana de observación. Ejemplo: *Flag*_diag_mental = 1 si el sujeto tiene al menos un diagnóstico de salud mental registrado.
Score de riesgo	Probabilidad estimada por el modelo de que un sujeto pertenezca a la clase positiva (LABEL=1). Varía entre 0 y 1. No debe interpretarse como probabilidad de suicidio real sin calibración prospectiva previa.
Hold-out temporal	Conjunto de datos reservado para evaluación prospectiva, separado por corte temporal: entrenamiento en 2013–2021, evaluación en 2022–2025. Simula el escenario de uso real del modelo sobre datos futuros.
Oversampling	Técnica de balanceo que aumenta la representación de la clase minoritaria mediante generación de muestras sintéticas (ej. SMOTE). En este trabajo se optó por *class_weight="balanced"* como alternativa sin muestras sintéticas.