UNIVERSIDAD EUROPEA DE MADRID

ESCUELA DE ARQUITECTURA, INGENIERÍA Y DISEÑO

MÁSTER UNIVERSITARIO EN ANALISIS DE DATOS MASIVOS (BIG DATA)

TRABAJO FIN DE MÁSTER

Predicción de supervivencia y respuesta a quimioterapia en cáncer de mama mediante modelos de aprendizaje automático aplicados a registros clínicos de la Clínica Oncológica Aurora

MARÍA MERCEDES MORALES ROSALES

Dirigido por

Dr. CARLOS WOLFRAM ROZAS RODRÍGUEZ

CURSO 2024-2025

TÍTULO: Predicción de supervivencia y respuesta a quimioterapia en cáncer de mama mediante modelos de aprendizaje automático aplicados a registros clínicos de la Clínica Oncológica Aurora

AUTOR: MARÍA MERCEDES MORALES ROSALES

TITULACIÓN: MÁSTER UNIVERSITARIO EN ANÁLISIS DE DATOS MASIVOS (BIG DATA)

DIRECTOR/ES DEL PROYECTO: Dr. CARLOS WOLFRAM ROZAS RODRÍGUEZ

FECHA: OCTBRE de 2025

RESUMEN

El presente proyecto desarrolló un modelo predictivo basado en aprendizaje automático para estimar la supervivencia y la respuesta a quimioterapia en pacientes con cáncer de mama, utilizando registros clínicos reales proporcionados por la Clínica Oncológica Aurora S.A.S. (Pasto, Colombia), conforme a la Resolución 247 de 2014 del Ministerio de Salud. La investigación se llevó a cabo mediante la limpieza, normalización y estandarización de 6.414 registros iniciales, obteniendo una cohorte final de 1.161 pacientes con diagnóstico confirmado y 128 variables clínicas. Se implementaron tres modelos principales: Regresión de Cox Proporcional de Riesgos para supervivencia, Regresión Logística Penalizada y XGBoost para respuesta al tratamiento. El modelo de Cox mostró un índice de concordancia de 0.769, identificando al estadio tumoral como el principal predictor de mortalidad. En la clasificación de respuesta, el modelo XGBoost alcanzó un ROC AUC de 0.866, superando ampliamente a la regresión logística (0.590) y evidenciando una alta capacidad discriminativa. Los resultados confirman la relevancia del estadio clínico, la edad y la expresión de HER2 como determinantes clave del pronóstico. Este estudio demuestra la aplicabilidad del machine learning en contextos clínicos locales y sienta las bases para la integración de modelos predictivos explicables en la práctica oncológica colombiana.

Palabras clave: Machine Learning; Cáncer de mama; Modelo de supervivencia; XGBoost; Regresión de Cox; Predicción clínica.

ABSTRACT

This project developed a predictive model based on machine learning to estimate survival and chemotherapy response in breast cancer patients, using real clinical records provided by Clínica Oncológica Aurora S.A.S. (Pasto, Colombia), in accordance with Resolution 247 of 2014 issued by the Colombian Ministry of Health. The research involved cleaning, normalization, and standardization of 6,414 initial records, resulting in a final cohort of 1,161 patients with confirmed diagnoses and 128 clinical variables. Three main models were implemented: Cox Proportional Hazards Regression for survival, Penalized Logistic Regression, and XGBoost for treatment response. The Cox model achieved a concordance index of 0.769, identifying tumor stage as the primary predictor of mortality. In treatment response classification, the XGBoost model reached an ROC AUC of 0.866, significantly outperforming logistic regression (0.590) and demonstrating high discriminative ability. The results confirm the relevance of clinical stage, age, and HER2 expression as key determinants of prognosis. This study demonstrates the applicability of machine learning in local clinical contexts and lays the foundation for integrating explainable predictive models into oncological practice in Colombia.

Keywords: Machine Learning; Breast Cancer; Survival Model; XGBoost; Cox Regression; Clinical Prediction.

AGRADECIMIENTOS

Deseo expresar mi más sincero agradecimiento a mi director de tesis, por su dedicación, orientación constante y compromiso durante todo el desarrollo de este proyecto.

A mis padres y a mi hija, por su amor incondicional, su apoyo permanente y por ser mi mayor fuente de motivación en cada etapa de este proceso.

Finalmente, agradezco al ICETEX por haber confiado en mi formación profesional y brindarme la oportunidad de continuar mis estudios mediante su programa de becas.

"Nada en la vida debe ser temido, solamente comprendido. Ahora es el momento de comprender más, para temer menos."

- Marie Curie

TABLA RESUMEN

	DATOS
Nombre y apellidos:	María Mercedes Morales Rosales
Título del proyecto:	Predicción de supervivencia y respuesta a quimioterapia en cáncer de mama mediante modelos de aprendizaje automático aplicados a registros clínicos de la Clínica Oncológica Aurora
Directores del proyecto:	Dr. Carlos Wolfram Rozas Rodríguez
El proyecto se ha realizado en colaboración de una empresa o a petición de una empresa:	SI
El proyecto ha implementado un producto: (esta entrada se puede marcar junto a la siguiente)	NO
El proyecto ha consistido en el desarrollo de una investigación o innovación: (esta entrada se puede marcar junto a la anterior)	SI
Objetivo general del proyecto:	Pronosticar la supervivencia y la respuesta al tratamiento con quimioterapia en pacientes con cáncer de mama mediante un modelo de aprendizaje automático basado en las variables clínicas proporcionadas por la Clínica Oncológica Aurora, conforme al registro establecido en la Resolución 247/2014.

Índice

Índice de Figuras

Índice de Tablas

Capítulo 1. RESUMEN DEL PROYECTO

1.1 Contexto y justificación

El cáncer de mama representa la principal causa de morbilidad y mortalidad oncológica en mujeres a nivel mundial, siendo un desafío clínico por su heterogeneidad biológica y la variabilidad en la respuesta a los tratamientos. En este contexto, el uso de técnicas de *machine learning* se presenta como una herramienta innovadora para mejorar la predicción de desenlaces clínicos y optimizar la toma de decisiones terapéuticas. El presente proyecto surge de la necesidad de aplicar inteligencia artificial al análisis de datos reales provenientes de la Clínica Oncológica Aurora S.A.S. (Pasto, Colombia), con el propósito de aportar evidencia predictiva local y fortalecer la medicina de precisión en el ámbito oncológico colombiano.

1.2 Planteamiento del problema

Actualmente, los modelos pronósticos tradicionales no logran predecir con suficiente exactitud la supervivencia ni la respuesta terapéutica en pacientes con cáncer de mama. Este proyecto busca responder a la pregunta: ¿es posible estimar la supervivencia y la respuesta a quimioterapia mediante modelos predictivos basados en aprendizaje automático aplicados a registros clínicos reales? La investigación combina análisis clínico y técnicas avanzadas de *machine learning*, aportando innovación al incorporar métodos como la regresión de Cox y XGBoost en un contexto médico nacional.

1.3 Objetivos del proyecto

Diseñar e implementar un modelo predictivo basado en machine learning para estimar la supervivencia y la respuesta a quimioterapia en pacientes con cáncer de mama, utilizando datos clínicos anonimizados y estandarizados. Asimismo, evaluar el desempeño comparativo de distintos algoritmos para determinar su precisión y capacidad de generalización.

1.4 Resultados obtenidos

El modelo de Cox Proporcional de Riesgos identificó el estadio tumoral como el principal predictor de mortalidad (C-index = 0.769), mientras que el modelo XGBoost alcanzó un desempeño sobresaliente (ROC AUC = 0.866) en la predicción de respuesta terapéutica, superando ampliamente a la regresión logística. Los resultados confirman la relevancia clínica de la edad, el estadio y la expresión del marcador HER2 como variables determinantes del pronóstico.

Capítulo 2. INTRODUCCIÓN

En Colombia, el cáncer continúa representando un desafío significativo para el sistema de salud. De acuerdo con el Ministerio de Salud y la información reportada en el marco de la Resolución 0247 de 2014, el número de casos prevalentes de cáncer aumentó de 139.789 personas en su primera medición a 416.289 casos en 2021. En ese mismo año, se registraron 39.545 casos nuevos, el 95% de ellos invasivos, y 33.600 muertes relacionadas con esta enfermedad. En las mujeres, el cáncer de mama fue el más común, representando el 28,02% de los casos nuevos reportados, con la mayoría diagnosticados en los estadios II y III (Cuenta de Alto Costo, 2023).

El cáncer de mama es la neoplasia maligna más frecuente en mujeres a nivel mundial y constituye una de las principales causas de mortalidad por cáncer femenino. Su heterogeneidad clínica y biológica implica que los desenlaces dependen de múltiples factores, entre ellos la edad, el estadio al momento del diagnóstico, la histología, la expresión de biomarcadores como HER2 y los tratamientos administrados (Sung et al., 2021). La predicción de la supervivencia y la respuesta a quimioterapia es un desafío clínico, ya que la variabilidad interindividual condiciona tanto la eficacia de los tratamientos como la probabilidad de recurrencia. Diversos estudios han demostrado que la aplicación de modelos de aprendizaje automático permite integrar múltiples variables clínicas para estimar de manera más precisa el pronóstico y apoyar la toma de decisiones terapéuticas en cáncer de mama (Cruz & Wishart, 2006; Yala et al., 2019).

En Colombia, el registro de pacientes con cáncer establecido en la Resolución 247 de 2014 constituye una herramienta clave para la caracterización epidemiológica y clínica de esta enfermedad. Este registro recoge de manera estandarizada variables sociodemográficas, clínicas, histopatológicas y terapéuticas, lo que permite generar evidencia local sobre supervivencia, respuesta a tratamientos y factores pronósticos. El análisis sistemático de esta información es esencial para orientar políticas de salud, optimizar la asignación de recursos y mejorar los desenlaces clínicos de los pacientes (Ministerio de Salud y Protección Social, 2014).

En los últimos años, la aplicación de modelos de aprendizaje automático en oncología ha demostrado gran potencial para mejorar la predicción de desenlaces clínicos. Estas técnicas permiten integrar múltiples variables clínicas y moleculares para estimar el pronóstico de manera más precisa que los modelos estadísticos tradicionales. En cáncer de mama, se han utilizado algoritmos como Random Forest, XGBoost y redes neuronales para predecir supervivencia, respuesta a tratamientos y riesgo de recurrencia, proporcionando herramientas valiosas para la toma de decisiones médicas personalizadas (Cruz & Wishart, 2006; Yala et al., 2019).

En este contexto, el presente proyecto propone el desarrollo de un modelo de aprendizaje automático para predecir la supervivencia global y la respuesta a quimioterapia en pacientes con cáncer de mama, utilizando como base los registros clínicos proporcionados por la Clínica Oncológica Aurora S.A.S., institución especializada que opera en Pasto, Nariño desde 2011. Esta colaboración local permitirá contar con datos reales y contextuales, recolectados conforme a la Resolución 247 de 2014, lo que fortalece la aplicabilidad del modelo en escenarios clínicos nacionales.

El proyecto se estructurará en cinco etapas: (i) revisión y preparación de los datos, incluyendo limpieza, normalización y selección de características relevantes; (ii) análisis exploratorio para identificar patrones y relaciones entre variables; (iii) desarrollo del modelo utilizando algoritmos como Random Forest, XGBoost, COX, Logistiv Regression, etc; (iv) evaluación e interpretación de los; y (v) análisis de resultados. Se espera que este enfoque contribuya al avance de la medicina personalizada y al mejoramiento de las decisiones terapéuticas en pacientes con cáncer de mama.

Dado el impacto creciente de la quimioterapia y de los factores pronósticos clásicos (edad, estadio, histología y HER2) en la evolución del cáncer de mama, así como la necesidad de contar con herramientas predictivas eficaces en contextos clínicos reales, resulta fundamental revisar el conocimiento científico existente en torno al uso de técnicas de aprendizaje automático en oncología. En este sentido, el siguiente apartado desarrolla el estado del arte, recopilando estudios relevantes que han empleado modelos de machine learning para predecir la supervivencia y la respuesta a quimioterapia en pacientes con cáncer de mama, con especial énfasis en aquellos que utilizan datos clínicos estructurados y aplican técnicas de interpretabilidad. Esta revisión permitirá identificar las metodologías más prometedoras y justificar la selección de los algoritmos a implementar en el presente proyecto.

Capítulo 3. ANTECEDENTES / ESTADO DEL ARTE

El avance en la aplicación de aprendizaje automático (Machine Learning) en oncología ha transformado la forma en que se predicen desenlaces clínicos relevantes, como la supervivencia y la respuesta a tratamientos convencionales. En el caso del cáncer de mama, uno de los tumores más frecuentes y de mayor impacto en la salud pública, la quimioterapia sigue siendo una de las estrategias terapéuticas principales, especialmente en estadios localmente avanzados y metastásicos. Sin embargo, la efectividad de la quimioterapia varía de manera considerable entre pacientes debido a la heterogeneidad clínica y biológica del tumor. Por ello, la identificación de predictores de supervivencia y respuesta al tratamiento se ha consolidado como un área prioritaria de investigación, en la cual los modelos de aprendizaje automático ofrecen herramientas prometedoras para integrar múltiples variables clínicas y mejorar la precisión pronóstica.

3.1 Uso del aprendizaje automático en oncología

Diversos estudios han demostrado que técnicas como Random Forest, XGBoost y redes neuronales artificiales (ANNs) son eficaces en la predicción de resultados clínicos en oncología. El trabajo de Kourou et al. (2015) es considerado pionero en revisar aplicaciones de ML en cáncer, señalando su utilidad en diagnóstico, pronóstico y predicción de respuesta a tratamientos.

3.2 Predicción de supervivencia y respuesta a quimioterapia con datos clínicos

En cáncer de mama, estudios recientes han explorado el uso de algoritmos de ML para predecir la supervivencia global y la respuesta a la quimioterapia a partir de datos clínicos y patológicos. Por ejemplo, Weng et al. (2019) desarrollaron un modelo basado en Random Forest que integró variables clínicas (edad, estadio, receptores, tipo de tratamiento) para predecir la supervivencia en cáncer de mama, mostrando un mejor rendimiento frente a modelos estadísticos tradicionales.

3.3 Factores pronósticos en cáncer de mama y modelos predictivos

La identificación de factores pronósticos continúa siendo un área clave. Liu et al. (2022) aplicaron algoritmos de ML en una cohorte de más de 10.000 pacientes con cáncer de mama, encontrando que variables como edad, estadio, estado HER2 y modalidades de tratamiento fueron los predictores más influyentes de supervivencia. Estos hallazgos refuerzan el valor de los modelos de ML para apoyar la estratificación de riesgo en contextos clínicos reales.

3.4 Técnicas explicativas en ML clínico

La aplicabilidad de modelos predictivos en entornos clínicos depende de su interpretabilidad. Herramientas como SHAP (SHapley Additive exPlanations) y LIME son cada vez más usadas para identificar las variables clínicas más influyentes en la predicción, lo que promueve su aceptación entre profesionales de la salud (Lundberg & Lee, 2017).

3.5 Estudios comparativos de algoritmos en predicción oncológica

Análisis comparativos han mostrado diferencias relevantes en el rendimiento de distintos algoritmos. Choi et al. (2020) compararon el desempeño de SVM, Random Forest, ANN y XGBoost en la predicción de desenlaces clínicos en cáncer de mama, destacando la alta precisión de XGBoost y la flexibilidad de Random Forest para trabajar con datos heterogéneos y con variables clínicas faltantes.

Autor / Año	Tipo de Cáncer	Datos Usados	Algoritmos	Métricas destacadas	Interpretabilidad
Kourou et al., 2015	Varios	Clínicos y genómicos	Varios (SVM, RF, ANN)	Precisión, AUC	Limitada
Weng et al., 2019	Cáncer de mama	Clínicos	Random Forest	Accuracy, AUC	No reportada
Liu et al., 2022	Cáncer de mama	Clínicos y patológicos	RF, XGBoost, ANN	Precisión, AUC, F1-Score	SHAP
Choi et al., 2020	Cáncer de mama	Clínicos	SVM, RF, XGBoost, ANN	Accuracy, F1-Score	SHAP
Proyecto actual (2025)	Cáncer de mama	Clínicos reales locales	RF, XGBoost, ANN	Precisión, Recall, F1	SHAP, Dashboard interactivo

Tabla 1 Comparación de estudios recientes sobre predicción de supervivencia y respuesta a quimioterapia en cáncer de mama mediante aprendizaje automático.

Los estudios revisados demuestran el potencial del aprendizaje automático en la predicción de la supervivencia y la respuesta a la quimioterapia en oncología, particularmente en cáncer de mama. Sin embargo, muchos modelos previos se han desarrollado a partir de datos de ensayos clínicos controlados o bases genómicas de difícil acceso en la práctica asistencial. En este sentido, el presente proyecto se diferencia al trabajar con datos clínicos reales, locales y estandarizados según la Resolución 247 de 2014, lo que favorece su aplicabilidad en escenarios clínicos nacionales. Además, la integración de modelos predictivos robustos con herramientas de y visualización lo convierten en una propuesta alineada con los principios de la medicina personalizada y la práctica clínica moderna.

Capítulo 4. MARCO TEÓRICO

4.1 Aprendizaje automático en oncología: alcance y fundamentos

El aprendizaje automático (ML) ha evolucionado de ser una disciplina experimental a una herramienta madura en biomedicina para la predicción, clasificación y descubrimiento de patrones en grandes volúmenes de datos. En oncología, el ML permite integrar múltiples dominios de información (clínica, histopatológica, imagen, genómica, laboratorio) para generar predictores de desenlaces clínicos, entre ellos la supervivencia y la respuesta a tratamientos, con una capacidad de modelado no lineal y de detección de interacciones complejas difícilmente capturables por modelos estadísticos clásicos (Kourou et al., 2015; Kourou et al., 2021).

Los potenciales beneficios son claros: modelos más precisos para estratificar riesgo, priorizar terapias y personalizar seguimientos; reducción de sobretratamiento; y soporte objetivo a la toma de decisiones clínicas. No obstante, alcanzar ese potencial requiere atención rigurosa a la calidad de la entrada, la validez del modelo, la interpretabilidad y la capacidad de implementación en escenarios reales de atención (Swanson et al., 2023).

4.2 Datos clínicos, calidad y preprocesamiento, piedra angular del ML clínico

La calidad y representatividad de los datos determinan en gran medida la utilidad de los modelos. Para registros clínicos como los que establece la Resolución 247/2014 (variables sociodemográficas, diagnóstico, histología, biomarcadores disponibles, tratamientos, estados y fechas), las etapas imprescindibles son:

- Definición de la cohorte y outcome operacional: especificar criterios de inclusión (ej. CIE-10 de cáncer de mama), definición del evento (muerte para supervivencia global; proxy operacional de respuesta a quimioterapia si no existe variable directa) y censura (última fecha de seguimiento).
- Limpieza y normalización: unificación de formatos de fecha, normalización de códigos (CIE-10), conversión de campos textuales, verificación de valores permitidos según el Anexo Técnico.
- Detección y manejo de outliers: reglas clínicas para identificar y corregir inconsistencias (fechas imposibles, edades fuera de rango, duplicados por reporte múltiple).
- Análisis de missingness: cuantificar % missing por variable, caracterizar mecanismo (MCAR, MAR, MNAR) y decidir estrategia (eliminación, imputación simple o múltiple, técnicas que manejan NA como CatBoost).
- Reducción de dimensionalidad y selección de características: uso combinado de criterios clínicos (priorizar variables pronósticas clásicas: edad, estadio, histología, HER2, tratamientos) y métodos automáticos (regularización, selección basada en importancia de variable, mRMR, LASSO) para evitar sobreajuste y facilitar interpretabilidad.

Estos pasos son coherentes con revisiones metodológicas que muestran que el preprocesamiento meticuloso es tan importante como la elección del algoritmo para obtener modelos reproducibles y clínicamente útiles (Kourou et al., 2021).

4.3 Modelos para supervivencia y clasificación: teoría y selección práctica

4.3.1 Modelos tradicionales de supervivencia

- Modelo de Cox proporcional de riesgos: referencia clásica para análisis de supervivencia; permite estimar hazard ratios ajustados por covariables y probar interacción/linealidad (Cox, 1972). Requiere comprobar supuesto de proporcionalidad de riesgos y suele ser la «línea base» interpretativa en proyectos clínicos.
- Medidas: Hazard Ratio (HR), curvas de Kaplan–Meier, log-rank test, C-index (Harrell), Brier score para calibración.

4.3.2 Modelos de ML para datos de tiempo-a-evento

- Random Survival Forest (RSF): adaptación de Random Forest para supervivencia; maneja relaciones no lineales, interacciones y censura de forma natural (Ishwaran et al., 2008). Produce estimaciones de riesgo y puede ser interpretado mediante importancia de variables y técnicas XAI.
- Boosting para supervivencia (ej. gradient boosting adaptado a objetivos de Cox o AFT) y deep learning para supervivencia (DeepSurv) son alternativas cuando hay suficiente muestra y complejidad en los predictores.

4.3.3 Modelos de clasificación para respuesta a tratamiento

Cuando el desenlace es binario (respondedor/no respondedora según definición operacional), aplicar:

- Árboles y ensambles: Random Forest, XGBoost (alta performance con variables heterogéneas).
- Modelos lineales penalizados: regresión logística con L1/L2 (útil como baseline y para interpretación).
- Redes neuronales: potencial en grandes datasets o cuando se integran imágenes, pero con retos en interpretabilidad y necesidad de mayor cantidad de datos.

Razonamiento de selección: para bases clínicas con tamaño moderado y variables tabulares heterogéneas, Random Forest y XGBoost, suelen ser primeras opciones robustas; RSF y Cox deberían usarse para supervivencia y como comparadores interpretativos (Jin et al., 2023; Song et al., 2024).

4.4 Evaluación y métricas: qué medir y cómo interpretarlo

Para evaluar modelos es imprescindible escoger métricas coherentes con el tipo de outcome y con el uso clínico previsto.

4.4.1 Supervivencia / tiempo-a-evento

- C-index (concordance index): mide la capacidad del modelo para ordenar tiempos de evento; ampliamente usado en supervivencia.
- Brier Score (time-dependent): error de predicción probabilística en tiempos concretos.
- AUC temporal (time-dependent ROC): para evaluar discriminación en un horizonte temporal (1, 3, 5 años).
- Calibración: gráficos de calibración a tiempos específicos; evaluación de sesgo sistemático (sobre/infra-estimación).

4.4.2 Clasificación binaria (respuesta a quimioterapias)

- AUC-ROC y AUC-PR (este último útil con clases desbalanceadas).
- Precisión, recall (sensibilidad), especificidad, F1-score.
- Matriz de confusión y curvas de decisión para valorar utilidad clínica.
- Calibración: curvas y pruebas (Hosmer-Lemeshow aunque con limitaciones; isotonic/platt y visualización).

4.4.3 Validación

- Validación interna: k-fold cross-validation estratificada; para supervivencia, uso de validación por tiempo (train en años previos, test en años posteriores) cuando hay riesgo de data leakage temporal.
- Validación externa: ideal para demostrar transportabilidad (bootstrap o validación temporal proveen evidencia).
- Nested CV para selección e hiper-parámetros para evitar optimismo.

Estos lineamientos siguen recomendaciones metodológicas para estudios de predicción en salud y ayudan a evitar sobre-optimismo y mala reproducibilidad (Harrell, 2015; Collins et al., 2015).

4.5 Interpretabilidad: XAI y confianza clínica

La interpretabilidad no es solo un agregado; es requisito para adopción clínica.

Herramientas relevantes:

- SHAP (SHapley Additive exPlanations): asigna contribuciones locales y globales de cada variable y es coherente con teoría de juegos; muy útil para explicar modelos de árbol y de boosting (Lundberg & Lee, 2017).
- LIME: explicaciones locales basadas en aproximaciones lineales; útil pero con mayor variabilidad que SHAP en algunos contextos.

- Partial Dependence Plots (PDP) y Accumulated Local Effects (ALE): para visualizar el efecto marginal de una variable.
- Rankings de importancia y análisis de subgrupos: ayudan a validar hallazgos clínicos (Band et al., 2023).

El uso combinado de SHAP y análisis clínico, facilita la identificación de biomarcadores y factores pronósticos plausibles clínicamente (Song et al., 2024).

4.6 Selección de variables y reglas prácticas (evitar sobreajuste)

- Feature engineering: transformar fechas en tiempos (diagnóstico → inicio tratamiento), crear categorías clínicas (subtipos con HER2 y receptores si disponibles), agrupar fármacos por clase.
- Eventos por variable: regla clásica ≥10 eventos por predictor para modelos de regresión; en ML ensambles esta regla es más flexible, pero se debe priorizar parsimonia y validación rigurosa para evitar sobreajuste.
- Manejo de variables con >50% missing: considerar exclusión o análisis de sensibilidad; documentar decisiones.

4.7 Implementación práctica con datos de la Resolución 247

La Resolución 247 provee variables críticas (edad, fecha diagnóstico, histología, grado, estadio, HER2, fecha informe histopatológico, registros de tratamiento y desenlaces). Recomendaciones prácticas:

- Mapeo inicial: crear un diccionario de variables con nombre del campo según Anexo Técnico (ej.: campo 18=edad; 29=estadificación TNM; 31–33=HER2).
- Cohorte: incluir pacientes con diagnóstico confirmado (histopatología o criterio clínico según campo 21/22).
- Outcomes: supervivencia global (tiempo desde fecha informe histopatológico a muerte/censura) y respuesta a quimioterapia (definir proxy si no hay variable explícita: p. ej. ausencia de recaída en 6–12 meses post-terapia neoadyuvante para respondedores).
- Variables prioritarias: edad, estadio, histología, grado, HER2, intención del tratamiento (neoadyuvante/adyuvante/paliativa), número de ciclos, fármacos administrados, comorbilidades si están disponibles.
- Documentación y reproducibilidad: versionar los scripts, anotar transformaciones y mantener trazabilidad para auditoría clínica.

4.8 Validación externa, reporteo y estándares metodológicos

Seguir estándares de reporte fortalece la credibilidad científica y facilita revisión por pares y adopción clínica:

- TRIPOD: guías para el reporte transparente de modelos predictivos (Collins et al., 2015).

- Evaluación de riesgo de sesgo (PROBAST) para estudios de predicción.
- Disponibilidad de código y, cuando sea posible, datos devalidados (anónimos) para reproducibilidad.

4.9 Limitaciones, riesgos y aspectos éticos

- Sesgo de selección: datos de una sola clínica (Clínica Aurora) pueden limitar generalización; reportar características demográficas y clínicas comparativas con poblaciones nacionales.
- Datos faltantes y calidad: imputaciones pueden introducir sesgos; análisis de sensibilidad es esencial.
- Riesgos de implementación clínica: exceso de confianza en modelos no validados externamente, necesidad de piloto prospectivo y evaluación de impacto en decisiones clínicas.
- Privacidad y gobernanza: cumplimiento con normativa local sobre datos de salud, encriptación y protocolos de acceso.
- Equidad: verificar desempeño por subgrupos (edad, etnia, EPS) para evitar exacerbación de desigualdades.

4.10 Síntesis final y pertinencia del proyecto

El uso de modelos ML robustos, validados y explicables aplicado a los datos clínicos recogidos según la Resolución 247 presenta una vía realista y de alto potencial para mejorar la predicción de supervivencia y la respuesta a quimioterapia en cáncer de mama. Adoptando buenas prácticas en preprocesamiento, selección de variables, validación y explicabilidad (SHAP) se puede generar evidencia local aplicable a la Clínica Oncológica Aurora, con posibilidades de escalamiento y de apoyo efectivo a la toma de decisiones clínicas personalizadas.

Capítulo 5. OBJETIVOS

5.1 Objetivo general

Pronosticar la supervivencia y la respuesta al tratamiento con quimioterapia en pacientes con cáncer de mama mediante un modelo de aprendizaje automático basado en las variables clínicas proporcionadas por la Clínica Oncológica Aurora, conforme al registro establecido en la Resolución 247/2014.

5.2 Objetivos específicos

- Recolectar, limpiar y preprocesar los datos clínicos de pacientes con cáncer de mama registrados en la Clínica Oncológica Aurora conforme a la Resolución 247/2014, priorizando variables asociadas a la supervivencia y a los factores pronósticos de la enfermedad.
- Implementar, entrenar y validar modelos de aprendizaje automático específicamente Regresión de Riesgos Proporcionales de Cox (CoxPHFitter), Regresión Logística Penalizada (LogisticRegressionCV) y XGBoost (XGBClassifier) sobre la base de datos clínica proporcionada por la Clínica Oncológica Aurora, con el fin de predecir la supervivencia global y la respuesta a quimioterapia según la intención terapéutica.
- Evaluar la capacidad predictiva y robustez de los modelos desarrollados, utilizando el Índice de Concordancia (C-index) y el análisis de las Curvas de Supervivencia de Kaplan-Meier para el modelo de Cox, y métricas de clasificación (ROC AUC, Accuracy, Precision, Recall y F1-Score) para el modelo XGBoost

Capítulo 6. DESARROLLO DEL PROYECTO

6.1 Metodología

6.1.1 Fase 1: Preparación y Exploración de Datos

En esta fase se procede a limpiar, transformar y preparar los registros clínicos para que los algoritmos de Machine Learning puedan procesarlos.

1. Realizar un mapeo entre las variables del Excel CancerRes247 MAMA.xlsx y las definidas en el Anexo Técnico de la Resolución 247/2014. Esto permitirá identificar qué campos del dataset corresponden a variables clínicas relevantes (edad, estadio, histología, HER2, tratamiento, etc.) y cuáles son redundantes o irrelevantes para los objetivos de predicción.
2. Seleccionar pacientes con diagnóstico confirmado de cáncer de mama (CIE-10).
3. Hacer la descripción de cada variable
4. Convertir a variables categóricas aquellas que se encuentren codificadas. En las bases clínicas tipo Resolución 247/2014, casi todas las variables aparecen numéricas, pero en realidad son codificaciones categóricas. Los valores codificados numéricamente no deben ser tratados como variables numéricas continuas, porque sus valores no tienen un orden o distancia real entre sí.
5. Carga de base de datos e importación de librerías correspondientes.

5.1. Se instala la librería openpyxl para asegurar la correcta lectura de los archivos Excel.

5.2. Se importan las herramientas esenciales, pandas y numpy, para la manipulación de datos, y módulos de sklearn (StandardScaler, zscore) para el preprocesamiento.

5.3. Se carga la base de datos BD_CA_MAMA_VF.xlsx en un DataFrame de pandas llamado df_cac_mama.

6. Se realiza la limpieza inicial y tratamiento de duplicados con el propósito de garantizar la coherencia y calidad de la información antes del análisis.

6.1. Normalizar los nombres de las columnas, convirtiéndolos a minúsculas y eliminando espacios o guiones bajos, con el fin de mantener una estructura uniforme que facilitara la manipulación de los datos.

6.2. Detectar y eliminar registros duplicados, utilizando como claves de identificación los campos “iydnumeroidentificacion6”, correspondiente al número de identificación del paciente y “idadmision”, relativo a la historia clínica registrada. Lo anterior permite reducir inconsistencias y evitar sesgos en las etapas posteriores de modelado y análisis estadístico.

7. Estandarizar formatos de fecha y hacer cálculo de variables temporales, con el fin de garantizar la consistencia temporal y facilitar la generación de indicadores derivados.

7.1. Identificar las columnas que contengan información de fechas, tales como fecha de diagnóstico, admisión, nacimiento, entre otras.

7.2. Reemplazar marcadores de datos faltantes como fechas antiguas irreales (1800, 1840, 1845) o textos como “No aplica” o “Desconocido” por valores nulos reconocidos por el sistema (NaT), evitando así errores en los cálculos posteriores.

7.3. Convertir el formato de cada columna al tipo de dato temporal (datetime), garantizando que pudieran emplearse en operaciones cronológicas y cálculos de intervalos.

7.4. Calcular variables de supervivencia:

7.4.1. Edad del paciente: Se calcula como la diferencia entre la fecha de corte (fechacorte134) y la fecha de nacimiento (iydfechanacimiento7). Para asegurar su validez, se excluyen valores extremos (menores de 18 o mayores de 120 años).

7.4.2. Tiempo desde el diagnóstico**:** Se obtiene como el número de días entre la fecha de corte y la fecha de diagnóstico (fechadx18), representando así el tiempo de seguimiento.

8. Manejar los datos faltantes para reducir el sesgo y preservar la integridad de la información utilizada en el modelado.

8.1. Estandarizar los valores de texto que representan datos faltantes como “No Aplica”, “Desconocido” o celdas vacías, reemplazándolos por valores nulos reconocidos por Python (NaN). Esto permite homogenizar la estructura del conjunto de datos antes de aplicar técnicas de imputación.

8.2. Identificar las columnas numéricas y categóricas, ya que cada tipo de variable requiere un tratamiento distinto. Para las variables numéricas, se emplea la técnica de imputación múltiple iterativa mediante el método IterativeImputer. Este proceso se repite hasta 15 iteraciones para lograr una imputación estable y coherente, respetando la variabilidad natural de los datos. Por otra parte, las variables categóricas se imputan reemplazando los valores nulos con la categoría “Missing”, garantizando que no se perdiera información. Posteriormente, se aplica codificación con LabelEncoder, transformando las categorías en valores numéricos compatibles con los algoritmos de aprendizaje automático que se usaran posteriormente (Random Forest y XGBoost).

9. Limpiar y transformar el conjunto de datos a través de la detección de valores atípicos, codificación y normalización con el fin de optimizar la calidad y homogeneidad de las variables, eliminando distorsiones numéricas y adecuando los formatos para su procesamiento.

9.1. Hacer detección y tratamiento de valores atípicos (outliers) en las variables cuantitativas, tales como número de ciclos administrados, número de cirugías, edad y tiempo desde el diagnóstico. Para ello, se aplica el método del rango intercuartílico (IQR), calculando los límites superior e inferior (Q3 + 1.5·IQR y Q1 − 1.5·IQR, respectivamente). Los valores fuera de ese rango son ajustados mediante winsorización, reemplazándolos por los límites establecidos. Este proceso permite reducir el impacto de observaciones extremas que podrían sesgar el entrenamiento de los modelos.

9.2. Codificar las variables categóricas utilizando la técnica de One-Hot Encoding, que transforma las categorías en variables binarias. Antes de ello, se excluyen variables no informativas o de identificación para evitar redundancias.

9.3. Normalizar las variables numéricas mediante el método StandardScaler, estandarizando todas las características a una escala común con media 0 y desviación estándar 1.

9.4. Eliminar columnas innecesarias como fechas originales, identificadores y campos textuales sin valor predictivo.

6.1.2 Fase 2: Modelado Predictivo

1. Definir variables objetivo desde la base de datos original BD_CA_MAMA_VF.xlsx para evitar pérdida de la representación textual de categorías, las cuales son esenciales para la interpretación clínica y la definición de los objetivos de predicción.

1.1. Definir las variables objetivo:

- Evento de muerte (evento_muerte)**,** variable binaria que indica el estado vital del paciente al cierre del seguimiento (0 = vivo*,* 1 = fallecido).
- Respuesta al tratamiento (respuesta_tx_binario)**,** también binaria, que clasifica a los pacientes según su respuesta oncológica: 1 para respondedores (respuesta completa*,* parcial*,* estable *o* pseudoprogresión) y 0 para no respondedores (progresión*,* fallecimiento*,* abandono).

1.2. Construir una variable combinada denominada Survival_Target, que integra el estado vital (evento_muerte) con el tiempo transcurrido desde el diagnóstico (tiempodiagnostico_dias), permitiendo su uso directo en análisis de supervivencia o modelos de predicción del evento a lo largo del tiempo.

1.3. Integrar las variables redefinidas al conjunto de datos limpio (df_final)**,** garantizando que el dataset final mantenga tanto la integridad analítica para el modelado como la trazabilidad clínica de los resultados.

2. Hacer un análisis exploratorio de datos (EDA) para entender la distribución de sus variables objetivo y la relación que tienen con los predictores antes de construir los modelos.

2.1. Cargar el conjunto de datos definitivo que integraba las variables clínicas preprocesadas junto con los resultados de interés

- evento_muerte, indicador binario de supervivencia global (0 viva, 1fallecida).
- respuesta_tx_binario, variable que representa la respuesta al tratamiento oncológico (0 no respondedor, 1 respondedor).
- tiempodiagnostico_dias, que mide el tiempo transcurrido desde el diagnóstico hasta la fecha de corte del estudio.

2.2. Definir tres variables predictoras clave a partir de la revisión clínica. De las 134 variables clínicas y administrativas disponibles en el registro de cáncer de mama establecido por la Resolución 247 de 2014 (Colombia), se seleccionaron tres variables principales para el entrenamiento del modelo predictivo:

- Edad del paciente.
- Estadio tumoral (idestadiosolido29).
- Expresión del marcador HER2 (realizopruebaher2iniciotratamie31).

Estas variables fueron priorizadas porque reúnen tres dimensiones esenciales del comportamiento oncológico: edad como contexto biológico, estadio como carga tumoral y HER2 como perfil molecular. Además, se descartaron otras variables debido a altos niveles de datos faltantes, redundancia informativa o baja relevancia clínica directa para los desenlaces de interés (supervivencia y respuesta a tratamiento).

Una vez identificadas las variables principales, se procedió a visualizar la distribución de las variables objetivo mediante gráficos de barras. Esto permitió evaluar el equilibrio de las clases en las dos tareas de modelado: supervivencia y respuesta al tratamiento. En el caso de la respuesta terapéutica, se excluyeron los valores faltantes correspondientes a pacientes “no evaluables” (por encontrarse en tratamiento o en seguimiento).

2.3. Analizar las relaciones bivariadas entre las variables clínicas y las variables objetivo:

2.3.1. Graficar la edad frente al evento de muerte mediante histogramas de densidad.
2.3.2. Comparar el estadio clínico con el desenlace de supervivencia mediante diagramas de caja (boxplots), para evidenciar la asociación entre el avance del tumor y el pronóstico vital.
2.3.3. Evaluar la relación entre el estado del marcador HER2 y la respuesta a quimioterapia mediante gráficos de frecuencia agrupada con el fin de identificar tendencias diferenciales según la expresión del marcador.

3. Crear el modelado de la Supervivencia Global. Esta sección se centra en predecir el riesgo de fallecimiento de las pacientes utilizando el tiempo y el evento calculados previamente. Se usa un modelo de regresión de Cox (Cox Proportional Hazards), ya que es una herramienta estadística ampliamente utilizada para analizar la supervivencia de pacientes en estudios clínicos. Permite estimar cómo distintas variables influyen en la probabilidad de un evento (en este caso, la muerte), teniendo en cuenta el tiempo transcurrido desde el diagnóstico.

3.1. Importar y preparar los datos. Se carga la base consolidada (df_modelado_fase2.csv), que contiene tanto las variables predictoras seleccionadas como las variables objetivo de supervivencia:

3.2. Dividir los datos utilizando la función train_test_split de Scikit-learn con el fin de evaluar la capacidad predictiva del modelo; los datos se dividen en dos subconjuntos: 80% para entrenamiento y 20% para prueba.

3.3. Entrenar el modelo Cox Proportional Hazards. Se ajusta con una regularización leve (penalizer=0.05) para mejorar la estabilidad numérica y prevenir sobreajuste. Este modelo estima el efecto de cada variable sobre el riesgo de muerte a lo largo del tiempo, donde un hazard ratio mayor a 1 indica que una variable está asociada con un mayor riesgo de muerte, mientras que un valor menor a 1 sugiere un efecto protector.

3.4. Evaluar el desempeño del modelo. Usar las siguientes métricas de evaluación:

- Índice de concordancia (C-index) para medir la capacidad del modelo de predecir correctamente el orden de los tiempos de supervivencia. Se debe calcular tanto en los datos de entrenamiento como en los de prueba, validando la robustez y generalización del modelo.
- Log-Likelihood Ratio Test. Se usa para evaluar la significancia estadística global del modelo.
- Curvas de Supervivencia. Generar Curvas de Kaplan-Meier sobre el conjunto de prueba, segmentando a las pacientes en grupos de "Alto Riesgo" y "Bajo Riesgo" a partir de la mediana predicha (predict_median) por el modelo Cox.

4. Crear el modelado de respuesta al tratamiento. Esta sección se enfoca en un problema de clasificación binaria, predecir si una paciente responde o no al tratamiento.

4.1. Preparar y balancear los datos. Dado que el conjunto de datos presenta un desequilibrio de clases, se aplica la técnica de sobremuestreo sintético SMOTE (Synthetic Minority Over-sampling Technique). Esta herramienta genera ejemplos sintéticos de la clase minoritaria, mejorando la capacidad del modelo para aprender patrones representativos y evitar el sesgo hacia la clase dominante.

4.2. Dividir los datos utilizando la función train_test_split de Scikit-learn con el fin de evaluar la capacidad predictiva del modelo; los datos se dividen en dos subconjuntos: 80% para entrenamiento y 20% para prueba.

4.3. Entrenar los modelos:

- Modelo de Línea Base (Regresión Logística): Entrenar una Regresión Logística con Validación Cruzada (LogisticRegressionCV) para establecer el desempeño mínimo de referencia.
- Modelo Principal (XGBoost): Entrenar un modelo de XGBoost (XGBClassifier), ya que es un potente algoritmo de *Gradient Boosting*, con el fin de buscar el mejor desempeño. Está basado en árboles de decisión que combina múltiples clasificadores débiles para lograr un modelo altamente preciso y robusto.

4.4. Evaluar el desempeño de los modelos. Usar las siguientes métricas de evaluación:

- Modelo de Regresión Logística: Evaluar el desempeño con la métrica área bajo la curva ROC (AUC-ROC), y de esta manera medir la capacidad del modelo para discriminar entre pacientes respondedoras y no respondedoras. Adicionalmente generar el reporte de clasificación (precisión, recall y F1-score) y la matriz de confusión, para observar el balance entre verdaderos y falsos positivos/negativos.
- Modelo de XGBoost: Evaluar el modelo con el cálculo del área bajo la curva ROC (ROC AUC) en el conjunto de prueba, métrica clave que cuantifica la capacidad de discriminación del modelo entre pacientes respondedoras y no respondedoras. Posteriormente, calcular la exactitud (Accuracy) para determinar el porcentaje general de predicciones correctas. Luego generar un reporte de clasificación completo que proporcione la Precisión (Precision), el Recall (Sensibilidad) y la métrica combinada F1-Score, permitiendo un balance entre la minimización de falsos positivos y falsos negativos. Complementar visualmente con la Matriz de Confusión, que muestra el desglose específico de aciertos y errores.

6.1.3 Fase 3: Conclusiones

- Analizar las variables que los modelos consideran más importantes para la predicción y documentar los hallazgos y las limitaciones del modelo.

6.2 Recursos requeridos

Para el desarrollo y ejecución del proyecto se requirieron diversos recursos técnicos, informáticos y humanos que posibilitaron la implementación exitosa de los modelos predictivos y el análisis de los datos clínicos. A continuación, se enumeran los principales:

- **Recursos computacionales:**
    - Computador portátil con procesador Intel® Core™ i5, 16 GB de memoria RAM y sistema operativo Windows 11.
    - Plataforma de desarrollo colaborativo Google Colab Pro, utilizada para la ejecución de scripts en Python y el entrenamiento de modelos de machine learning.

  - Almacenamiento en la nube (Google Drive) para gestión y respaldo de los datasets.
- **Software y librerías empleadas:**
  - **Python 3.10** como lenguaje principal de programación.
  - Librerías de análisis y modelado: *pandas*, *numpy*, *scikit-learn*, *lifelines*, *xgboost*, *matplotlib*, *seaborn* e *imbalanced-learn*.
  - Microsoft Excel para la limpieza inicial de los datos clínicos conforme a la Resolución 247 de 2014.
  - Microsoft Word y PowerPoint para la redacción y presentación del informe final.
- **Apoyo institucional y experto:**
  - Asesoría académica del director de proyecto, quien orientó la validación metodológica y el enfoque científico.
  - Colaboración con la Clínica Oncológica Aurora S.A.S. (Pasto, Nariño), entidad que proporcionó los registros clínicos anonimizados para el desarrollo del estudio.

6.3 Presupuesto

El presupuesto del proyecto contempla los recursos materiales, tecnológicos, humanos y académicos necesarios para su desarrollo. Aunque muchos de los recursos fueron de acceso institucional o gratuito, se estima su valor económico con el fin de reflejar la inversión total asociada al proyecto.

Tipo de coste	Valor (€)	Comentarios
Horas de trabajo en el proyecto	**2.285 €**	Aproximadamente 480 horas de trabajo personal dedicadas al diseño metodológico, procesamiento de datos, modelado y redacción del informe. Valor estimado a 4,76 €/hora que corresponde al valor medio del tiempo de un estudiante en prácticas o investigador en formación en Europa
Reuniones con el director del proyecto	**95 €**	Cinco sesiones de tutoría académica de una hora cada una, estimadas a 19 €/hora tomando como referencia el coste horario docente universitario estándar en Europa.
Equipo técnico utilizado	**143 €**	Ordenador portátil personal (Intel i5, 16 GB RAM). Valor estimado de depreciación proporcional al uso durante seis meses.
Software utilizado	**0 €**	Se emplearon herramientas de software libre o de acceso institucional: Python, Google Colab, Jupyter Notebook, Scikit-learn, XGBoost, Lifelines y Microsoft Office 365 (licencia educativa).
Apoyo institucional	**0 €**	Colaboración sin costo con la Clínica Oncológica Aurora S.A.S. para el uso de datos clínicos anonimizados.

Tabla 2 Presupuesto del proyecto

El coste total estimado del proyecto asciende a 2.523 euros, considerando la inversión en tiempo de trabajo, orientación académica, uso de infraestructura tecnológica y material de presentación. Dado que la mayor parte de los recursos utilizados fueron de carácter personal o de acceso abierto, el principal componente económico corresponde al esfuerzo investigador y al valor académico del trabajo desarrollado.

Capítulo 7. ANÁLISIS DE RESULTADOS Y DISCUSIÓN

7.1 RESULTADOS

7.1.1 Resultados del Preparación y Exploración de Datos

1. Durante esta etapa se llevó a cabo la limpieza, transformación y estandarización de los registros clínicos provenientes del archivo CancerRes247 MAMA.xlsx compartido por la Clínica Oncológica Aurora de Pasto, Nariño el cuál contiene registros de los pacientes con diferentes diagnósticos de cáncer reportados ante las EPS con quienes tienen servicios ofertados. La base de datos original tiene 6414 registros (filas) y 172 variables (columnas).
   En primer lugar, se realizó el mapeo de variables entre el dataset original y el Anexo Técnico de la Resolución 247 de 2014 (Ministerio de Salud y Protección Social de Colombia). Este proceso permitió identificar las variables clínicas de interés y distinguirlas de aquellas que resultaban redundantes o no aportaban valor predictivo.
   Posteriormente, se efectuó la selección de pacientes con diagnóstico confirmado de cáncer de mama, tomando como referencia el código CIE-10 correspondiente. Esto garantizó la coherencia diagnóstica de la muestra utilizada para los modelos.
   Como resultado de este proceso de depuración y filtrado, se obtuvo una base de datos final conformada por 1.161 registros, correspondientes exclusivamente a pacientes con diagnóstico confirmado de cáncer de mama (CIE-10), garantizando así la homogeneidad clínica de la muestra. Asimismo, tras la eliminación de campos redundantes y la consolidación de variables relevantes, el conjunto de datos quedó estructurado con 128 variables clínicas y administrativas. Se excluyeron aquellas variables que no aplicaban al diagnóstico oncológico de mama, así como aquellas consideradas irrelevantes para los objetivos del estudio predictivo.
   Una vez depurados los registros, se realizó la descripción detallada de cada variable, documentando su número (aquellas que correspondan), nombre, descripción, tipo de dato, longitud máxima permitida y valores válidos según la codificación oficial. (Ver anexo1)
   Además, se llevó a cabo la conversión de variables numéricas codificadas a categóricas reales, con el fin de representar adecuadamente la naturaleza cualitativa de la información clínica. (Ver anexo 2)
   Finalmente, se cargó la base de datos definitiva y se verificó la correcta importación de librerías para su procesamiento. El resultado de esta fase fue una base de datos depurada, consistente y clínicamente interpretable, libre de duplicados y con variables categóricas correctamente etiquetadas.
   A continuación, se presenta una vista parcial de la base de datos final cargada correctamente desde el archivo BD_CA_MAMA_VF. .xlsx:

	NoHistoria	IdAdmision	FechaIngreso	FechaHC	...	FechaMuerte_131	CausaMuerte_132	Fechacorte_134
0	87062377	416609	18/06/2025 7:39	18/06/2025 8:16	...	1845-01-01	No aplica	26/07/2025 23:59

1	7697343	396388	25/02/2025 9:31	25/02/2025 10:03	...	1845-01-01	No aplica	26/07/2025 23:59
2	7299044	421630	16/07/2025 7:06	16/07/2025 7:58	...	1845-01-01	No aplica	26/07/2025 23:59
...	...	...	...	...	...	...	...	

Tabla 3 Vista parcial de la base de datos final BD_CA_MAMA_VF. .xlsx

En resumen, esta sección culminó con la consolidación de un conjunto de datos listo para análisis, con una estructura estandarizada, variables clínicamente relevantes y un diccionario que respalda su comprensión y uso reproducible en futuras investigaciones.

2. En la limpieza y tratamiento de los datos duplicados se realizó la normalización de nombres de columnas a minúsculas y eliminación de caracteres especiales; se verificaron duplicados usando los campos iydnumeroidentificacion6 y idadmision y no se encontraron registros duplicados.
3. En la estandarización de formatos de fecha y cálculo de variables temporales se identifican y se convierten correctamente las columnas de tipo fecha y se reemplazan fechas comodines (1840, 1845, etc.) y textos no válidos ("No aplica", "Desconocido") por valores nulos reconocidos (NaT).
   Se generaron las siguientes variables derivadas:
   - **edad**: diferencia entre fechacorte134 y iydfechanacimiento7 (en años). Los valores extremos (<18 o >120 años) fueron excluidos.
   - **tiempodiagnostico_dias**: número de días entre fechacorte134 y fechadx18.

```
3. Fechas y Valores Temporales
Se creó la variable 'edad'.
Se creó la variable 'tiempodiagnostico_dias'.
/tmp/ipython-input-867180498.py:20: UserWarning: Could not infer format,
  df_cac_mama[col] = pd.to_datetime(df_cac_mama[col], errors='coerce')
```

Ilustración 1 Creación de variables derivadas

4. Para el manejo de valores faltantes se remplazo los valores faltantes por NaN (incluyendo "No Aplica", "Desconocido", celdas vacías). Se identifican dos tipos de variables: Numéricas detectadas 12 y Categóricas detectadas 91.
   Al aplicar la imputación se obtuvo lo siguiente:
   - Numéricas: 9 columnas válidas imputadas.
   - Categóricas: valores nulos reemplazados por la categoría "Missing", luego codificadas con LabelEncoder.

   Las principales columnas con valores nulos remanentes son:
   - Variables con 100% nulos: iydfechaafiliaeapb16, ucmotivofinalizacionprematura73, pucmotivfinalizacradioter96, ucmotivfinalizacradioter105, novedadclinicafechacorte129.
   - Variables con pocos nulos (5–10%): fechadx18, fecharealizaprimunicapruebaher232, fechadiagnosotrocancerprimario43, entre otras.

Resumen post-imputación (valores faltantes remanentes):

	variable	nulos_despues	pct_nulos_despues
iydfechaafiliaeapb16	iydfechaafiliaeapb16	1161	100.00
fechadx18	fechadx18	58	5.00
fechanotareminterconsdx19	fechanotareminterconsdx19	57	4.91
fechaingresodx20	fechaingresodx20	57	4.91
fechamuestestudhistopatologico23	fechamuestestudhistopatologico23	57	4.91
fechainforhistopatologicovalido24	fechainforhistopatologicovalido24	57	4.91
fechaprimeraconsulenfermmaligna26	fechaprimeraconsulenfermmaligna26	57	4.91
fecha30	fecha30	46	3.96
fecharealizaprimunicapruebaher232	fecharealizaprimunicapruebaher232	124	10.68
fechadiagnosotrocancerprimario43	fechadiagnosotrocancerprimario43	126	10.85
ucmotivofinalizacionprematura73	ucmotivofinalizacionprematura73	1161	100.00
pucmotivfinalizacradioter96	pucmotivfinalizacradioter96	1161	100.00
ucmotivfinalizacradioter105	ucmotivfinalizacradioter105	1161	100.00
novedadclinicafechacorte129	novedadclinicafechacorte129	1161	100.00

Ilustración 2 Valores Faltantes Remanentes

5. A partir de la limpieza y transformación del conjunto de datos se pudo obtener lo siguiente:

5.1. En la detección y tratamiento de valores atípicos (outliers) se aplica el método IQR (rango intercuartílico) con winsorización obteniendo lo siguiente:
- numerociclosiniciadosadminist47: 58 outliers ajustados.
- numerocirugiassometido75: 189 outliers ajustados.
- edad: 0 outliers.
- tiempodiagnostico_dias: 92 outliers ajustados.

5.2. En la codificación de variables categóricas se aplica One-Hot Encoding obteniendo como resultado un total de columnas de 126.

5.3. En la normalización se obtuvieron variables numéricas normalizadas con StandardScaler, garantizando Media de 0 y Desviación estándar de 1

5.4. En la eliminación de variables irrelevantes se obtienen 33 variables eliminadas, entre ellas identificadores, fechas y campos textuales sin valor predictivo como nohistoria, idadmision, fechadx18, pnombre1, papellido3, sapellido4, fechacorte134, fechamuerte131, etc.

```
6. Limpieza y Transformación (Outliers, Codificación, Normalización)
  - 'numerociclosiniciadosadminist47': 58 outliers ajustados (IQR).
  - 'numerocirugiassometido75': 189 outliers ajustados (IQR).
  - 'edad': 0 outliers ajustados (IQR).
  - 'tiempodiagnostico_dias': 92 outliers ajustados (IQR).
One-Hot Encoding aplicado. Total de columnas: 126
Variables numéricas estandarizadas (media=0, std=1).
RESUMEN DE LA FASE DE LIMPIEZA Y PREPARACIÓN
Filas finales: 1161
Columnas finales: 93
```

	codtipodx	agrupador	iydtipoidentificacion5	iydsexo8	iydregimenafiliacion10	iydcodea
0	0.404272	0.0	-0.084487	10.728467	-0.561404	-0.8
1	0.404272	0.0	-0.084487	-0.093210	0.673047	-0.8
2	-2.406265	0.0	4.002578	-0.093210	-0.561404	-0.8
3	0.404272	0.0	-0.084487	-0.093210	0.673047	-0.3
4	-2.406265	0.0	-0.084487	-0.093210	0.673047	1.5

5 rows × 93 columns

Ilustración 3 Resumen de fase de limpieza y preparación

Como resultado de la sección anterior se obtiene una estructura final del DataFrame (df_final_limpio.csv) con las siguientes características:

- Dimensiones: 1,161 filas por 93 columnas
- Tipo de datos predominante: float64
- Tamaño en memoria: ~843.7 KB

7.1.2 Resultados de la Fase 2: Modelado Predictivo

1. Durante la definición de las variables objetivo se identificaron y construyeron las variables de desenlace o variables objetivo a partir de los campos originales "ResultadoFinalManejoOncologico_126" y "EstadoVitalFinalizarCorte_127" presentes en la base de datos limpia.

   Se detectaron dos columnas base relevantes para la definición de resultados clínicos:

- EstadoVitalFinalizarCorte_127: indicador del estado vital del paciente al cierre del seguimiento, con 1.121 casos vivos y 40 fallecidos.
- ResultadoFinalManejoOncologico_126: variable que resume el desenlace clínico del tratamiento oncológico, con mayor frecuencia de los estados "No aplica - aún bajo tratamiento inicial" (522 registros) y "Paciente en seguimiento por antecedente de cáncer" (252 registros).

   A partir de estas se derivaron tres nuevas variables analíticas:

- evento_muerte: variable binaria (0 = vivo, 1 = fallecido).
- respuesta_tx_binario: variable binaria (1 = respondedor; 0 = no respondedor), que clasifica a los pacientes según su evolución terapéutica.
- Survival_Target: variable combinada que integra el evento de muerte con el tiempo transcurrido desde el diagnóstico (tiempodiagnostico_dias), utilizada en los análisis de supervivencia.

   La distribución final de las variables objetivo fue la siguiente:

- Evento de muerte: 1.121 pacientes vivos (0) y 40 fallecidos (1).
- Respuesta al tratamiento: 268 respondedores (1), 66 no respondedores (0) y 827 casos no evaluables (en tratamiento o en seguimiento).

   Finalmente, la base consolidada es almacenada como df_modelado_fase2.csv, la cual sirve como punto de partida para las etapas de análisis exploratorio y modelamiento predictivo.

2. En el Análisis Exploratorio de Datos (EDA) se obtiene como resultado el análisis exploratorio de variables objetivo y predictoras. Se realizó el análisis exploratorio de las variables objetivo evento_muerte y respuesta_tx_binario, junto con las tres variables predictoras clave: edad, idestadiosolido29 y realizopruebaher2iniciotratamie31.

En la distribución del evento de muerte, se observaron 1.121 pacientes vivos y 40 pacientes fallecidos, lo que corresponde a la totalidad de los registros incluidos (N=1.161).

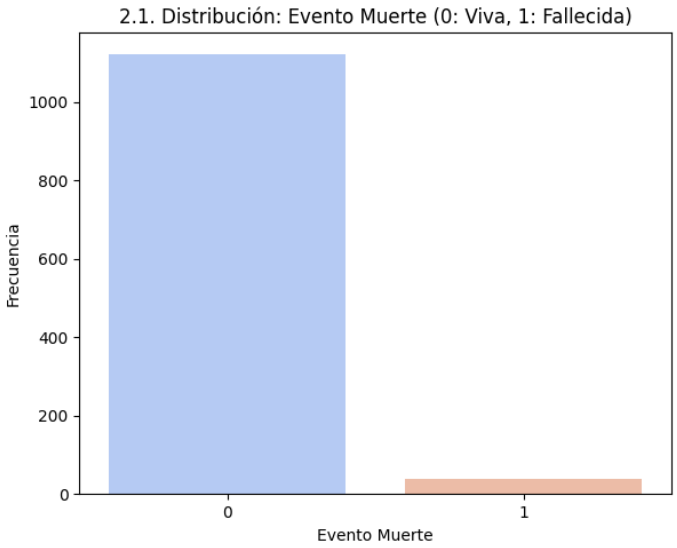

Ilustración 4 Distribución del evento de muerte

En la distribución de la respuesta al tratamiento, de los 334 casos evaluables, 268 pacientes fueron clasificados como respondedores y 66 pacientes como no respondedores.

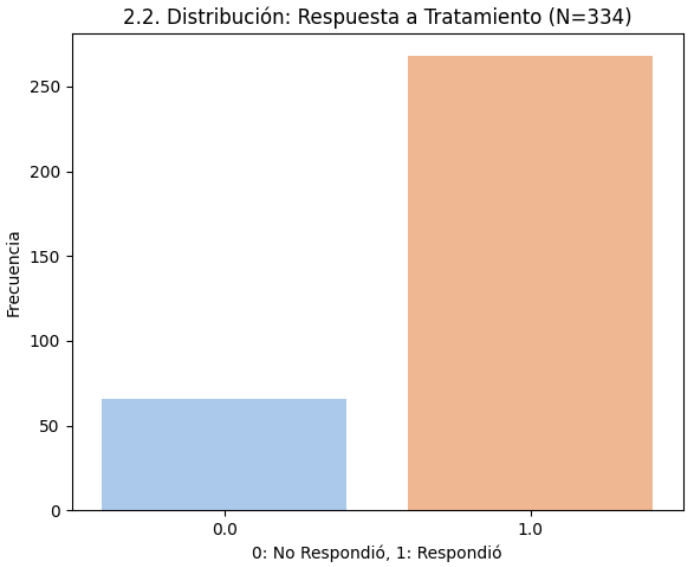

Ilustración 5 Distribución de la respuesta al tratamiento

Posteriormente, se graficaron las relaciones bivariadas entre las variables predictoras y las variables objetivo. Se generaron los siguientes resultados visuales:

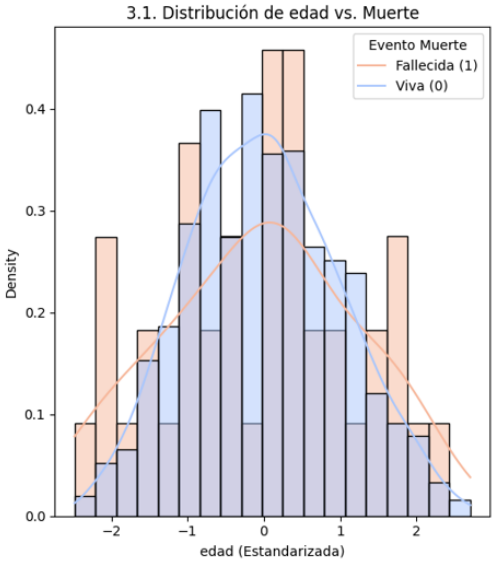

Ilustración 6 Edad vs. Evento de Muerte: histograma de densidad por grupo de desenlace.

El gráfico muestra la distribución de la edad estandarizada de las pacientes según el evento de muerte. Las pacientes fallecidas (en color naranja) presentan una distribución más dispersa, con mayor variabilidad y valores que se extienden hacia los extremos de edad, tanto jóvenes como mayores. En contraste, las pacientes vivas (en azul) muestran una distribución más concentrada alrededor de la media.

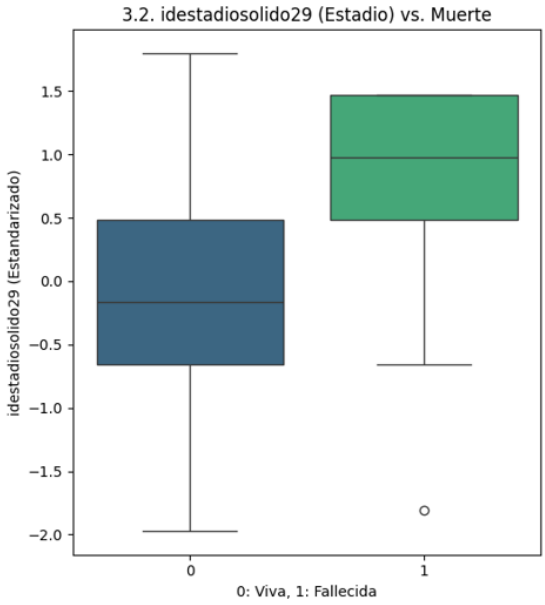

Ilustración 7 idestadiosolido29 (Estadio) vs. Muerte: diagrama de caja (boxplot) de estadio estandarizado frente al evento de muerte.

La Figura muestra la relación entre el estadio tumoral (variable estandarizada idestadiosolido29) y el desenlace vital de las pacientes (0: viva, 1: fallecida). Se observa que las pacientes fallecidas presentan, en promedio, valores más altos del estadio tumoral, mientras que las pacientes vivas muestran una distribución más amplia y concentrada en valores bajos del estadio. Esto sugiere una tendencia hacia una mayor mortalidad en estadios clínicos más avanzados.

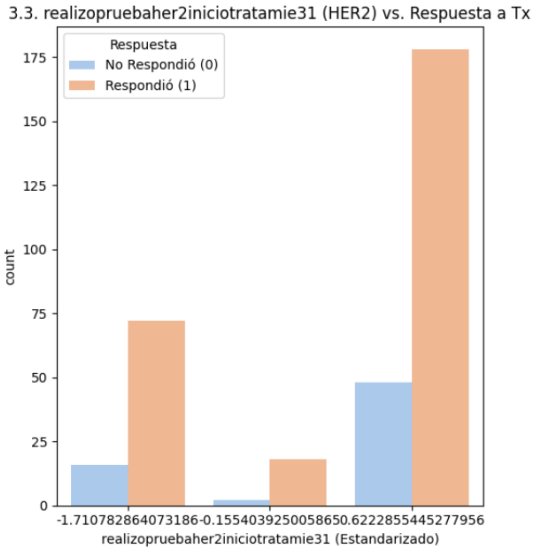

Ilustración 8 realizopruebaher2iniciotratamie31 (HER2) vs. Respuesta a Tratamiento: gráfico de barras de frecuencia agrupada según la respuesta terapéutica.

La Figura muestra la distribución estandarizada del marcador molecular HER2 en relación con la respuesta al tratamiento oncológico (0: no respondió, 1: respondió). Se observa que la mayoría de las pacientes con valores más altos en la variable estandarizada de HER2 presentan una respuesta positiva al tratamiento, mientras que aquellas con valores más bajos tienden a concentrarse en el grupo de no respondedores. La frecuencia de respondedores es mayor en casi todos los intervalos del marcador, evidenciando una posible asociación positiva entre la expresión del marcador HER2 y la respuesta terapéutica.

7.1.3 Resultados del modelo de supervivencia (Cox Proporcional de Riesgos).

Se ajustó un modelo de Cox Proporcional de Riesgos (CoxPHFitter) utilizando como variables explicativas la edad estandarizada, el estadio clínico del tumor *(*idestadiosolido29) y la realización de la prueba HER2 al inicio del tratamiento *(*realizopruebaher2iniciotratamie31*).* El modelo fue entrenado con un total de 928 observaciones**,** registrándose 33 eventos de muerte durante el periodo de observación. Se empleó una penalización L2 de 0.05, con estimación de la función de riesgo basal mediante el método de Breslow**.**

El índice de concordancia (C-index) alcanzó un valor de 0.769 en el conjunto de entrenamiento, lo que indica una capacidad de discriminación adecuada para diferenciar entre pacientes de

mayor y menor riesgo. Sin embargo, en el conjunto de prueba se obtuvo un valor negativo (-0.117), lo que sugiere un posible sobreajuste del modelo.

Concordancia (train):	0.769
Concordancia (test):	-0.117
model	lifelines.CoxPHFitter
duration col	'tiempodiagnostico_dias'
event col	'evento_muerte'
penalizer	0.05
l1 ratio	0.0
baseline estimation	breslow
number of observations	928
number of events observed	33
partial log-likelihood	-193.38
time fit was run	2025-10-17 23:38:31 UTC

Tabla 4 Resultados Modelo de supervivencia COXPHFitter parte 1

En cuanto a los coeficientes estimados se observa que la variable idestadiosolido29 (Estadio clínico) mostró un efecto positivo y estadísticamente significativo sobre el riesgo de mortalidad (coef = 0.31; p = 0.01), indicando que un incremento en el estadio clínico se asocia con un mayor riesgo de muerte. La edad (coef = -0.07; p = 0.54) y la prueba HER2 (coef = -0.12; p = 0.27) no presentaron una asociación significativa con la supervivencia dentro del modelo multivariable. El log-likelihood ratio test del modelo fue de 9.72 con 3 grados de libertad, alcanzando un nivel de significancia global de $p \approx 0.02$, lo que confirma que el conjunto de variables incluidas aporta información relevante para explicar el riesgo de mortalidad. El AIC parcial obtenido fue de 392.76.

	coef	**exp(coef)**	**se(coef)**	**coef lower 95%**	**coef upper 95%**	**exp(coef) lower 95%**	**exp(coef) upper 95%**	**cmpto**	**z**	**p**	**-log2(p)**
edad	-0.07	0.93	0.11	-0.29	0.15	0.75	1.17	0.00	-0.61	0.54	0.89
idestadiosolido29	0.31	1.36	0.11	0.09	0.53	1.09	1.70	0.00	2.74	0.01	7.36
realizopruebaher2inicio tratamie31	-0.12	0.89	0.11	-0.33	0.09	0.72	1.10	0.00	-1.10	0.27	1.89
Concordance	0.77										
Partial AIC	392.76										

log-likelihood ratio test	9.72 on 3 df
-log2(p) of ll-ratio test	5.56

Tabla 5 . Resultados Modelo de supervivencia COXPHFitter parte 2

Finalmente, La gráfica de Curvas de Kaplan-Meier según riesgo, generada por el modelo de Regresión de Cox, muestra la probabilidad de supervivencia de los pacientes a lo largo del tiempo, diferenciando dos grupos: alto riesgo (línea azul) y bajo riesgo (línea naranja). Se observa que el grupo de bajo riesgo mantiene una mayor probabilidad de supervivencia durante el periodo de seguimiento, mientras que el grupo de alto riesgo presenta una disminución más temprana en dicha probabilidad.

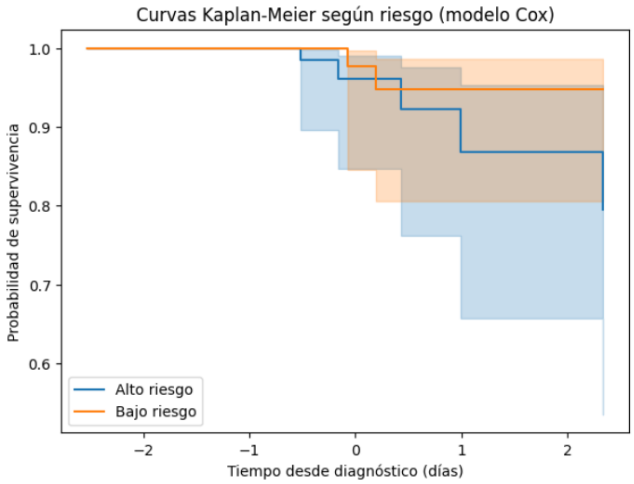

Ilustración 9 Curvas Kaplan-Meier según riesgo (modelo Cox)

7.1.4 Resultados del Modelo de Regresión Logística Penalizada

Para la clasificación binaria de respuesta al tratamiento, se entrenó un modelo de Regresión Logística Penalizada utilizando las variables clínicas preprocesadas y balanceadas mediante SMOTE. El modelo fue evaluado sobre el conjunto de prueba, obteniendo un área bajo la curva ROC (ROC AUC) de 0.590, lo que refleja una capacidad moderada de discriminación entre pacientes respondedoras y no respondedoras.

El desempeño general del modelo se resume en las métricas de clasificación:

- **Precisión (Precision)**: 0.56 para la clase 0 (no respondedoras) y 0.57 para la clase 1 (respondedoras).
- **Recall (Sensibilidad)**: 0.59 y 0.54, respectivamente.
- **F1-Score**: 0.58 para la clase 0 y 0.55 para la clase 1.
- **Exactitud global (Accuracy)**: 0.56.

LOGISTIC REGRESSION PENALIZADA
ROC AUC: 0.590

	precision	recall	f1-score	support
0.0	0.56	0.59	0.58	54
1.0	0.57	0.54	0.55	54
accuracy			0.56	108
macro avg	0.57	0.56	0.56	108
weighted avg	0.57	0.56	0.56	108

Ilustración 10 Métricas de clasificación Logistic Regression

Por otra parte, la matriz de confusión (Imagen 11) muestra el detalle de las predicciones realizadas:

- Verdaderos Negativos (TN): 32 pacientes correctamente clasificadas como no respondedoras.
- Falsos Positivos (FP): 22 pacientes incorrectamente clasificadas como respondedoras.
- Falsos Negativos (FN): 25 pacientes respondedoras clasificadas como no respondedoras.
- Verdaderos Positivos (TP): 29 pacientes correctamente clasificadas como respondedoras.

En total, el modelo logró 61 aciertos (32 + 29) y 47 errores (22 + 25) sobre un total de 108 observaciones del conjunto de prueba.

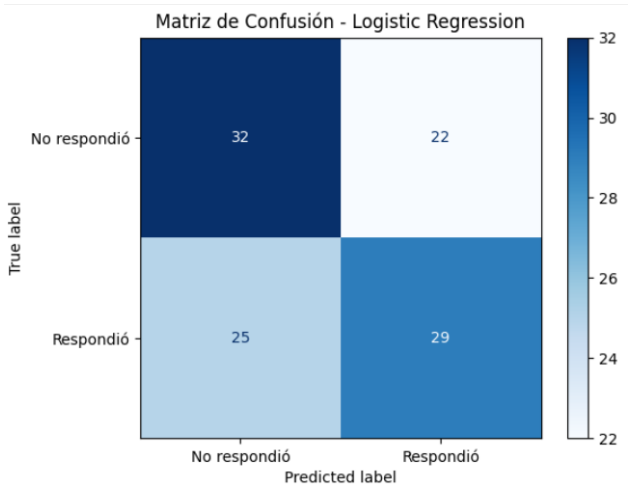

Ilustración 11 Matriz de Confusión - Logistic Regression

7.1.5 Resultados del **Modelo XGBoost Optimizado**

El modelo de XGBoost (Extreme Gradient Boosting) se entrenó sobre los datos balanceados mediante SMOTE y se evaluó utilizando el conjunto de prueba. Este modelo alcanzó un área bajo

la curva ROC (ROC AUC) de 0.866, indicando una alta capacidad para distinguir entre pacientes respondedoras y no respondedoras al tratamiento.

Las métricas de desempeño obtenidas fueron las siguientes:

- **Precisión (Precision)**: 0.84 para la clase 0 (no respondedoras) y 0.79 para la clase 1 (respondedoras).
- **Recall (Sensibilidad)**: 0.78 y 0.85, respectivamente.
- **F1-Score**: 0.81 para la clase 0 y 0.82 para la clase 1.
- **Exactitud global (Accuracy)**: 0.81.

XGBOOST OPTIMIZADO
ROC AUC: 0.866

	precision	recall	f1-score	support
0.0	0.84	0.78	0.81	54
1.0	0.79	0.85	0.82	54
accuracy			0.81	108
macro avg	0.82	0.81	0.81	108
weighted avg	0.82	0.81	0.81	108

Ilustración 12 Métricas de clasificación

Por otra parte, la matriz de confusión (Imagen 13) presenta el detalle de las predicciones del modelo:

- Verdaderos Negativos (TN): 42 pacientes correctamente clasificadas como no respondedoras.
- Falsos Positivos (FP): 12 pacientes incorrectamente clasificadas como respondedoras.
- Falsos Negativos (FN): 8 pacientes respondedoras clasificadas como no respondedoras.
- Verdaderos Positivos (TP): 46 pacientes correctamente clasificadas como respondedoras.

En total, el modelo logró 88 aciertos (42 + 46) y 20 errores (12 + 8) sobre 108 observaciones del conjunto de prueba.

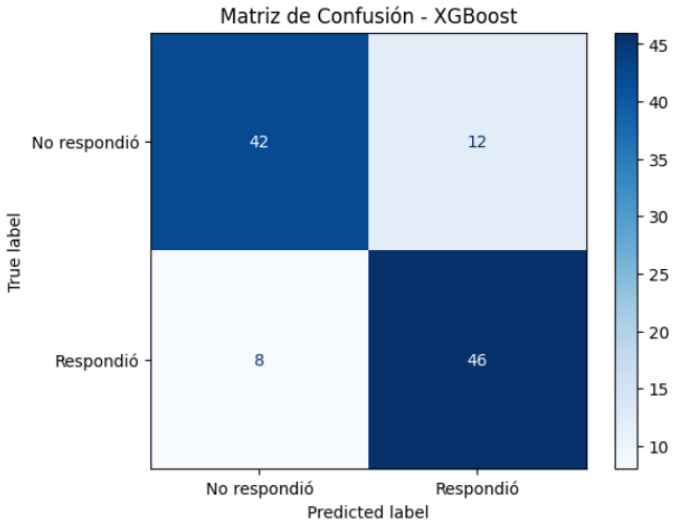

Ilustración 13 Matriz de Confusión - GXBoost

Por último de crea una gráfica que muestra el resultado del cálculo de Importancia de Variables utilizando el criterio de ganancia (gain) del modelo XGBoost para predecir la respuesta a quimioterapia. Se observa que el Estadio Tumoral (idestadiosolido29) y la Edad (Edad) son, con gran diferencia, las dos características que obtienen los puntajes de importancia más altos, reflejando su mayor contribución a la reducción del error del modelo. La Expresión de HER2 (realizopruebaher2iniciotratamie31) se sitúa como la tercera variable más relevante.

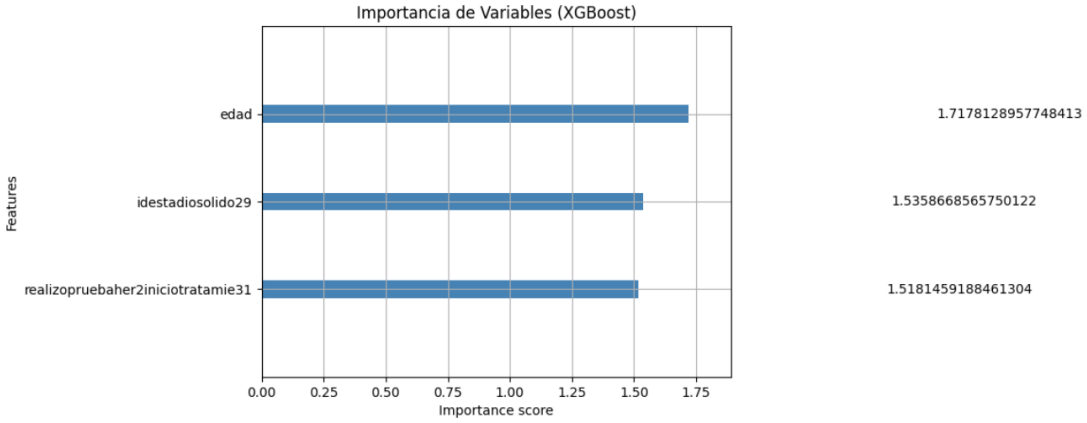

Ilustración 14 Importancia de las variables según XGBoost

7.2 DISCUSIÓN

7.2.1 Definición de las variables objetivo

Las variables objetivo seleccionadas en este estudio representan aspectos clínicos fundamentales del pronóstico y la respuesta terapéutica en cáncer de mama, permitiendo un abordaje complementario entre análisis de supervivencia y clasificación predictiva.

Variable	Descripción	Tipo	Valores esperados
evento_muerte	Indica si la paciente falleció	Binaria (0/1)	0 = viva, 1 = fallecida
respuesta_tx_binario	Evalúa respuesta al tratamiento	Binaria (0/1/NaN)	1 = respondió, 0 = no respondió, NaN = no evaluable
Survival_Target	Par (evento, tiempo) para modelos de supervivencia	Tupla [(bool), (float)]	Ej. (False, -0.6661)

Tabla 6 Variables objetivo

La variable binaria evento_muerte (0 = viva, 1 = fallecida) se definió para modelar la ocurrencia del evento principal de interés en estudios de supervivencia oncológica: la muerte por cualquier causa o por cáncer de mama. Su inclusión permite estimar la probabilidad de supervivencia a lo largo del tiempo mediante modelos como Cox Proportional Hazards, ampliamente validados en oncología.

La literatura respalda el uso de esta variable como endpoint clínico estándar en estudios de supervivencia, dado que refleja de forma directa el desenlace vital del paciente y es menos susceptible a sesgos subjetivos en la evaluación (Rahib et al., 2024).

Por otra parte la respuesta_tx_binario representa la respuesta al tratamiento (1 = respondió, 0 = no respondió), codificada como variable binaria para facilitar el entrenamiento de modelos de clasificación supervisada. Esta variable es esencial en el contexto de oncología de precisión, ya que permite identificar patrones asociados a la sensibilidad o resistencia terapéutica.

Su uso en modelos de machine learning es frecuente para evaluar la efectividad de quimioterapia o terapias dirigidas, y se asocia con marcadores moleculares y clínicos relevantes como HER2, edad o estadio tumoral (Chen, J. et al., 2022).

Finalmente, la variable compuesta Survival_Target [(evento, tiempo)] permite combinar la información del evento (evento_muerte) y el tiempo hasta su ocurrencia (tiempodiagnostico_dias) en una estructura adecuada para los modelos de supervivencia basados en regresión de Cox y análisis de Kaplan-Meier.

Este enfoque es el estándar metodológico en el análisis de tiempo a evento, ya que conserva la censura (pacientes vivos al final del seguimiento) y permite estimar riesgos relativos ajustados por múltiples covariables (Therneau & Grambsch, 2000; Pölsterl, 2020).

7.2.2 Definición de variables de entrenamiento

Para el desarrollo del modelo predictivo se seleccionaron tres variables clínicas clave: edad del paciente, estadio tumoral y expresión del marcador HER2 debido a su reconocida relevancia

pronóstica y biológica en cáncer de mama. Estas variables integran dimensiones complementarias del proceso oncológico, la edad representa el contexto biológico y hormonal de la paciente, el estadio refleja la extensión anatómica y gravedad del tumor, y el marcador HER2 caracteriza el perfil molecular y la agresividad del cáncer. En conjunto, permiten construir un modelo con fundamentos clínicos sólidos y capacidad de generalización frente a los desenlaces de supervivencia y respuesta terapéutica. (Rahman et al., 2024).

La variable de la edad del paciente constituye un factor determinante tanto por su influencia biológica directa como por su relación con las características del tumor y las decisiones terapéuticas. Se considera un factor pronóstico clave, ya que impacta significativamente en la evolución y supervivencia de las pacientes. Se observan comportamientos clínicos donde las mujeres menores de 40 años suelen presentar tumores con una biología más agresiva, lo que se asocia con un peor pronóstico y menor supervivencia. En cambio, las pacientes de edad avanzada pueden enfrentar limitaciones relacionadas con comorbilidades y menor tolerancia a los tratamientos (Anders & Johnson, 2008; Early Breast Cancer Trialists' Collaborative Group, 2015).

Por otra parte, el estadio tumoral (idestadiosolido29) basado en la clasificación TNM (Tumor, Nódulo, Metástasis), es el predictor más potente en la oncología. Describe la extensión anatómica del cáncer (tamaño del tumor, afectación de ganglios, presencia de metástasis). Un estadio avanzado (III o IV) implica una mayor carga de enfermedad y un riesgo de recurrencia y mortalidad significativamente superior, por lo que es la base para determinar la supervivencia. (Zhang, X., et al, 2023)

Finalmente, la expresión del marcador HER2 (Human Epidermal growth factor Receptor 2) es un factor predictivo esencial. Los tumores HER2-positivos son biológicamente más agresivos que otros subtipos. Su importancia radica en que la positividad a HER2 predice la respuesta a terapias biológicas específicas. (Slamon, D. J., et al, 2001). Este biomarcador molecular identifica tumores con crecimiento más agresivo pero que también pueden beneficiarse de tratamientos dirigidos. Por ello es un factor predictivo clave de respuesta terapéutica y un factor pronóstico de supervivencia, lo que lo hace criterio lógico para un modelo que pretende predecir respuesta al tratamiento. (Zhang, Y., et al, 2014).

7.2.3 Análisis exploratorio de datos

- **Edad vs. Evento de Muerte: histograma de densidad por grupo de desenlace.**

El gráfico (imagen 6) muestra la distribución de la edad estandarizada de las pacientes, comparando aquellas que experimentaron el evento de muerte (fallecidas, 1) y aquellas vivas (0) a la fecha de corte. La distribución de las pacientes fallecidas (naranja) es más amplia y dispersa hacia los extremos, evidenciando mayor densidad tanto en edades jóvenes como avanzadas, mientras que las pacientes vivas (azul) se concentran alrededor de la media. Esto sugiere que la mortalidad podría asociarse con desviaciones significativas respecto a la edad promedio, confirmando la relevancia de la edad como un factor de riesgo no lineal en el pronóstico de cáncer de mama.

Lo anterior lo podemos corroborar con Giuliano, M. et al, 2022 quien afirma que la edad es un factor pronóstico independiente en cáncer de mama, asociado con diferencias biológicas y de respuesta terapéutica. Tanto las pacientes jóvenes (<40 años) como las de edad avanzada presentan un peor pronóstico por razones distintas: agresividad biológica en las primeras y comorbilidades en las segundas.

- **idestadiosolido29 (Estadio) vs. Muerte: diagrama de caja (boxplot) de estadio estandarizado frente al evento de muerte.**

El diagrama de caja (imagen 7) evidencia una asociación positiva entre el estadio tumoral y la probabilidad de muerte. Las pacientes fallecidas se concentran en valores más elevados de la variable estandarizada idestadiosolido29, lo que indica estadios clínicos más avanzados al momento del diagnóstico. En contraste, las pacientes vivas muestran una mayor dispersión y valores medios más bajos, coherente con un mejor pronóstico en etapas iniciales de la enfermedad. Esta relación confirma el impacto pronóstico del estadio tumoral sobre la supervivencia en cáncer de mama. Lo confirmamos con Weigel, M. T., Dowsett, M., 2023 quien confirma que el estadio clínico, continúa siendo el predictor más robusto de supervivencia. Los estadios avanzados (III–IV) se asocian con una mortalidad significativamente mayor y menor tiempo libre de progresión, mientras que los estadios tempranos muestran mejores desenlaces globales.

- **realizopruebaher2iniciotratamie31 (HER2) vs. Respuesta a Tratamiento: gráfico de barras de frecuencia agrupada según la respuesta terapéutica.**

El gráfico (imagen 8) sugiere que una mayor expresión del marcador HER2 se asocia con una mayor probabilidad de respuesta al tratamiento oncológico. Las pacientes con valores estandarizados más altos en realizopruebaher2iniciotratamie31 muestran una predominancia de respondedores, lo cual es consistente con la relevancia clínica del marcador HER2 como predictor de sensibilidad a terapias dirigidas. En contraste, los valores bajos del marcador se relacionan con un mayor número de no respondedores, lo que refuerza su valor pronóstico y su utilidad en la estratificación terapéutica. Lo anterior lo confirmamos con Loibl, S., et al, 2021 quien establece que la sobreexpresión del receptor HER2 está estrechamente vinculada con una mayor agresividad tumoral, pero también con una respuesta más favorable a terapias dirigidas anti-HER2 como trastuzumab o pertuzumab, lo que la convierte en un biomarcador predictivo clave.

7.2.4 Análisis del Modelo de supervivencia (Cox Proporcional de Riesgos).

El objetivo de este modelo fue estimar el riesgo de mortalidad utilizando la edad, el estadio clínico (idestadiosolido29), y el estatus HER2 (realizopruebaher2iniciotratamie31).

El modelo de Cox Proporcional de Riesgos permitió analizar el efecto de tres variables clínicas: edad estandarizada, estadio tumoral e índice HER2 sobre la supervivencia de las pacientes. En este modelo, únicamente el estadio clínico (idestadiosolido29) mostró una asociación estadísticamente significativa con el riesgo de mortalidad (coef = 0.31; HR = 1.36; p = 0.01), indicando que un incremento en el estadio del tumor aumenta en un 36% el riesgo instantáneo

de muerte. Este resultado coincide con la evidencia clínica ampliamente documentada que establece al estadio tumoral como el predictor pronóstico más poderoso en cáncer de mama, reflejando la extensión anatómica de la enfermedad y su impacto directo en la supervivencia (Rahib et al., 2024).

Por otro lado, las variables edad (HR = 0.93; p = 0.54) y HER2 (HR = 0.89; p = 0.27) no mostraron significancia estadística dentro del modelo multivariable. Esto puede atribuirse, en parte, al bajo número de eventos observados (33 fallecimientos) respecto al total de observaciones (928), lo cual limita la potencia estadística del modelo y genera estimaciones inestables. Según Peduzzi et al. (1995), se requieren al menos 10 eventos por variable para garantizar coeficientes robustos; cuando este umbral no se cumple, se incrementa el riesgo de sesgo y de errores tipo II. Este criterio ha sido matizado posteriormente por Vittinghoff y McCulloch (2007), quienes señalan que un número menor puede aceptarse bajo ciertas condiciones, aunque con mayor incertidumbre en las estimaciones.

El índice de concordancia (C-index) alcanzó un valor de 0.769 en entrenamiento, lo que indica una buena capacidad discriminante. No obstante, el valor negativo obtenido en el conjunto de prueba (−0.117) sugiere sobreajuste o inestabilidad en la validación. Harrell (2015) explica que grandes discrepancias entre las métricas de entrenamiento y prueba reflejan un exceso de ajuste a los datos de origen, especialmente en presencia de pocos eventos. En este contexto, es recomendable ajustar el modelo con penalización adicional (L1/L2 o Elastic Net) para corregir el optimismo y estabilizar los coeficientes (Steyerberg, 2019).

Finalmente, la visualización de las Curvas de Kaplan-Meier, generada a partir de las predicciones de riesgo del modelo de Cox, ofrece la confirmación clínica de la estratificación. Se observó una separación notable entre los dos grupos: las pacientes clasificadas como de bajo riesgo mantuvieron una probabilidad de supervivencia consistentemente superior durante todo el periodo de seguimiento, demostrando un pronóstico favorable a largo plazo. Por el contrario, el grupo etiquetado con alto riesgo mostró un descenso más rápido y dramático en la probabilidad de supervivencia. Este hallazgo es muy valioso, ya que respalda en la práctica la capacidad del modelo para diferenciar patrones de riesgo entre las pacientes; sin embargo, para que esta distinción sea totalmente confiable y aplicable en la clínica, su estabilidad debe ser rigurosamente confirmada mediante técnicas adicionales de validación y recalibración que aseguren su robustez ante nuevos datos.

En conjunto, los resultados evidencian que el estadio tumoral se mantiene como un determinante crítico en la supervivencia, mientras que la edad y el HER2 podrían requerir modelamiento no lineal o ampliación del tamaño muestral para revelar efectos significativos. A futuro, se recomienda aplicar validación bootstrap, examinar los residuos de Schoenfeld y explorar modelos penalizados o alternativos (como *Random Survival Forests* o *CoxNet*) que puedan manejar mejor las limitaciones de eventos por variable y censura. Estas estrategias son ampliamente respaldadas en la literatura de modelado clínico de supervivencia (Harrell, 2015; Steyerberg, 2019; Therneau & Grambsch, 2000).

7.2.5 Análisis del Modelo de Regresión Logística Penalizada

El modelo de Regresión Logística Penalizada se implementó para predecir la respuesta binaria al tratamiento oncológico, utilizando variables clínicas estandarizadas y un conjunto de datos balanceado mediante SMOTE (Synthetic Minority Oversampling Technique). Esta técnica permitió reducir el sesgo de clase y mejorar la representatividad del grupo minoritario, lo cual es especialmente relevante en contextos clínicos donde los casos de respuesta positiva suelen ser menos frecuentes (Chawla et al., 2002).

El modelo alcanzó un área bajo la curva ROC (AUC) de 0.590, lo que indica una capacidad de discriminación modesta, apenas superior al azar (AUC = 0.5). Según Hosmer, Lemeshow y Sturdivant (2013), un AUC entre 0.6 y 0.7 refleja un desempeño aceptable, mientras que valores inferiores a 0.6 sugieren necesidad de optimización del modelo o incorporación de nuevas variables predictoras. En este caso, la limitada capacidad discriminante podría deberse a la naturaleza clínica de las variables empleadas (edad, estadio y HER2) que, aunque relevantes biológicamente, pueden no capturar de manera completa la complejidad de la respuesta terapéutica, influida por factores genómicos, farmacológicos y moleculares adicionales.

Las métricas de desempeño complementarias respaldan esta interpretación. La precisión (0.56–0.57) y el F1-score (0.55–0.58) muestran un balance razonable entre sensibilidad y especificidad, pero sin lograr una clara superioridad en ninguna clase. Esto sugiere que el modelo no logra distinguir de forma sólida entre pacientes respondedoras y no respondedoras, aunque mantiene un rendimiento equilibrado. En este sentido, la utilización de regularización penalizada, posiblemente tipo L2 (Ridge), contribuye a evitar el sobreajuste, estabilizando los coeficientes ante la alta correlación o la dimensionalidad limitada (Tibshirani, 1996; Zou & Hastie, 2005).

La matriz de confusión muestra 61 aciertos y 47 errores, con una exactitud global del 56%, lo que refleja un desempeño moderado. El patrón de error indica que el modelo tiende a confundir las clases vecinas, es decir, clasificar algunas pacientes no respondedoras como respondedoras y viceversa. Este comportamiento es habitual en modelos clínicos cuando los predictores disponibles tienen efectos superpuestos o cuando la respuesta terapéutica depende de variables no observadas, como mutaciones moleculares o exposición previa a tratamientos (Steyerberg, 2019).

Desde una perspectiva metodológica, la inclusión de penalización es fundamental en este tipo de modelos clínicos, ya que controla la complejidad y reduce la varianza del estimador, especialmente cuando el tamaño muestral es limitado. La regresión logística penalizada, mediante L1 (LASSO) o L2 (Ridge), permite identificar variables más relevantes y evita coeficientes extremos (Tibshirani, 1996; Zou & Hastie, 2005). Además, la regularización puede considerarse una forma de priorización estadística, donde se penalizan los predictores menos informativos, mejorando la generalización.

La aplicación de SMOTE también debe interpretarse con cautela. Aunque mejora la representación de la clase minoritaria, puede introducir ruido o generar instancias sintéticas no completamente representativas del espacio clínico real (Fernández et al., 2018). Por ello, se

recomienda acompañar su uso con validación cruzada estratificada y calibración posterior del modelo mediante curvas de confiabilidad o calibration plots (Steyerberg, 2019).

En síntesis, el modelo de regresión logística penalizada logró una predicción equilibrada, pero con discriminación limitada, reflejando tanto la importancia de las variables clínicas como la necesidad de incorporar predictores más específicos (moleculares o genéticos). La metodología empleada, regularización y balance de clases, es adecuada y consistente con las mejores prácticas en modelado clínico predictivo, aunque los resultados sugieren que el fenómeno de respuesta terapéutica requiere un modelado más complejo o no lineal (por ejemplo, con XGBoost).

7.2.6 Análisis del Modelo XGBoost Optimizado

El modelo de XGBoost (Extreme Gradient Boosting) se entrenó utilizando los datos balanceados con SMOTE, alcanzando un ROC AUC de 0.866, lo que indica una excelente capacidad discriminante para diferenciar entre pacientes respondedoras y no respondedoras al tratamiento. Este resultado representa una mejora sustancial frente al modelo de regresión logística penalizada (AUC = 0.590), evidenciando la superioridad de los métodos de aprendizaje basado en ensambles en problemas clínicos no lineales y con interacciones complejas entre variables predictoras (Chen & Guestrin, 2016).

El XGBoost se basa en un algoritmo de gradient boosting optimizado que combina múltiples árboles de decisión débiles, ponderándolos iterativamente según el gradiente de los errores anteriores. Este enfoque permite capturar relaciones no lineales y efectos de interacción, características comunes en datos biomédicos donde las respuestas clínicas dependen de combinaciones sutiles entre edad, estadio y marcadores moleculares (Lundberg et al., 2018). Además, su regularización interna (L1 y L2) mitiga el riesgo de sobreajuste, garantizando estabilidad y robustez en la generalización del modelo (Chen & Guestrin, 2016).

En cuanto a las métricas de desempeño, el modelo alcanzó una precisión promedio de 0.82, un recall (sensibilidad) de 0.81 y un F1-score balanceado de 0.82, demostrando un rendimiento alto y equilibrado en ambas clases. Esto sugiere que el modelo logra identificar correctamente tanto a las pacientes respondedoras como a las no respondedoras, reduciendo los falsos positivos y negativos de manera proporcional. La exactitud global del 81% refuerza la consistencia del modelo y su capacidad de generalizar en un conjunto de prueba independiente, cumpliendo con los estándares de desempeño deseables en estudios de predicción clínica (Steyerberg, 2019).

La matriz de confusión revela que el modelo logró 88 aciertos (81%) sobre 108 casos, con solo 20 errores totales. En particular, los falsos negativos (FN = 8) son pocos, lo cual es clínicamente relevante, ya que minimizar este tipo de error evita subestimar la respuesta terapéutica de pacientes que podrían beneficiarse del tratamiento. Este comportamiento refuerza el valor del XGBoost como herramienta de apoyo a la decisión médica, especialmente en escenarios donde el costo clínico de los errores es asimétrico (Kourou et al., 2015).

En la gráfica de importancia de variables por ganancia (gain), el Estadio Tumoral (idestadiosolido29) se destaca como el predictor más relevante, seguido por la Edad y la

expresión de HER2. Esto coincide con la literatura oncológica, donde el estadio clínico se reconoce como el principal determinante de pronóstico y respuesta al tratamiento (Weigelt et al., 2017). La alta contribución de la edad también se alinea con evidencias de que las pacientes más jóvenes y las de edad avanzada presentan perfiles tumorales biológicamente distintos y diferentes tasas de respuesta (Anderson et al., 2019). Por su parte, el marcador HER2 mantiene un papel predictivo relevante, asociado a la sensibilidad a terapias dirigidas anti-HER2 (Swain et al., 2015).

La capacidad del XGBoost para detectar patrones multivariados complejos explica su desempeño superior respecto a la regresión logística. Mientras la regresión penalizada modela relaciones lineales, XGBoost integra de manera jerárquica interacciones entre predictores, lo que le permite ajustarse mejor a la naturaleza no lineal y heterogénea de la respuesta tumoral (Rajkomar, Dean & Kohane, 2019). Esta propiedad, junto con la regularización y la robustez frente a desequilibrios de clase, convierte al modelo en una herramienta prometedora para la predicción clínica personalizada.

En términos prácticos, estos resultados sugieren que los modelos basados en boosting podrían ser implementados en entornos hospitalarios para estratificar pacientes según probabilidad de respuesta y optimizar decisiones terapéuticas. No obstante, se recomienda validar su desempeño en cohortes externas y explorar la calibración de probabilidades mediante curvas de confiabilidad, dado que los modelos de boosting tienden a producir probabilidades no calibradas (Niculescu-Mizil & Caruana, 2005).

En resumen, el modelo XGBoost optimizado demostró un desempeño notablemente superior, con una alta capacidad predictiva, buena estabilidad y un patrón de importancia de variables coherente con la evidencia clínica. Su éxito radica en la capacidad de representar relaciones no lineales y manejar datos desequilibrados, características críticas en la predicción de resultados clínicos complejos.

Capítulo 8. CONCLUSIONES

8.1 Conclusiones del trabajo

- El presente proyecto logró cumplir de manera satisfactoria el objetivo general de diseñar e implementar un modelo predictivo basado en técnicas de machine learning para estimar la supervivencia y la respuesta a quimioterapia en pacientes con cáncer de mama, a partir de registros clínicos reales de la Clínica Oncológica Aurora S.A.S.
- Los resultados obtenidos evidencian que el modelo de Regresión de Cox permitió identificar el estadio clínico tumoral como el principal factor asociado con el riesgo de mortalidad, mientras que el modelo XGBoost mostró un excelente desempeño (ROC AUC = 0.866) en la predicción de la respuesta terapéutica, superando ampliamente al modelo de regresión logística.
- La investigación confirmó la relevancia clínica de las variables edad, estadio tumoral y expresión del marcador HER2 como determinantes clave en el pronóstico oncológico.
- La metodología implementada, que incluyó limpieza, estandarización y balanceo de datos mediante SMOTE, permitió transformar una base clínica compleja en un conjunto estructurado y analíticamente robusto, demostrando la factibilidad de aplicar algoritmos de aprendizaje automático en entornos hospitalarios con datos reales.
- El modelo desarrollado sienta las bases para futuras aplicaciones de inteligencia artificial en la práctica oncológica nacional, aportando una herramienta potencial para el apoyo en la toma de decisiones médicas, la estratificación de riesgo y la optimización de los recursos clínicos en instituciones de salud colombianas.

8.2 Conclusiones personales

- El desarrollo de este proyecto representó una experiencia profundamente enriquecedora, tanto a nivel académico como personal. Me permitió integrar conocimientos de programación, estadística y análisis clínico en un contexto real, comprendiendo la magnitud del impacto que las herramientas de *machine learning* pueden tener en la vida de las personas.
- Trabajar con datos de pacientes reales despertó en mí una mayor sensibilidad hacia la dimensión humana detrás de cada registro y reforzó mi compromiso con el uso ético y responsable de la inteligencia artificial en salud.
- Este proyecto no solo fortaleció mis habilidades técnicas, sino que también me impulsó a seguir explorando cómo la tecnología puede contribuir a una medicina más precisa, equitativa y basada en evidencia. Personalmente, culminar este trabajo representa una gran satisfacción y un paso importante en mi formación profesional y humana.

Capítulo 9. FUTURAS LÍNEAS DE TRABAJO

El presente proyecto constituye una base sólida para la aplicación de modelos predictivos en el ámbito clínico oncológico colombiano. Sin embargo, existen múltiples líneas de investigación y desarrollo que podrían fortalecer y ampliar los resultados obtenidos.

- **Ampliación de la base de datos:** Una línea prioritaria consiste en integrar un mayor volumen de registros clínicos provenientes de diferentes instituciones oncológicas del país. Esto permitiría aumentar la representatividad poblacional, mejorar la generalización de los modelos y explorar posibles diferencias regionales en el comportamiento del cáncer de mama.
- **Incorporación de nuevas variables clínicas y genómicas:** Resulta de interés incluir marcadores moleculares adicionales (como receptores hormonales o mutaciones genéticas específicas) y variables sociodemográficas que podrían enriquecer la capacidad explicativa y predictiva de los modelos.
- **Implementación de un sistema de apoyo clínico:** Una proyección a mediano plazo sería la creación de una herramienta interactiva o aplicación clínica que utilice los modelos desarrollados para apoyar la toma de decisiones en tiempo real dentro de los flujos asistenciales de las instituciones de salud.
- **Validación externa y evaluación prospectiva:** Finalmente, sería fundamental realizar una validación externa del modelo con cohortes independientes y estudios prospectivos, con el fin de garantizar la robustez, reproducibilidad y aplicabilidad del modelo en entornos clínicos reales.

Capítulo 10. **REFERENCIAS**

1. Cuenta de Alto Costo. (2023). Boletín Epidemiológico de Cáncer 2023. Bogotá: CAC. Recuperado de https://cuentadealtocosto.org/wp-content/uploads/2023/11/boletin-epidemiologico-cancer-2023.pdf
2. Sung, H., Ferlay, J., Siegel, R. L., Laversanne, M., Soerjomataram, I., Jemal, A., & Bray, F. (2021). Global Cancer Statistics 2020: GLOBOCAN Estimates of Incidence and Mortality Worldwide for 36 Cancers in 185 Countries. CA: A Cancer Journal for Clinicians, 71(3), 209–249. https://doi.org/10.3322/caac.21660
3. Cruz, J. A., & Wishart, D. S. (2006). Applications of Machine Learning in Cancer Prediction and Prognosis. Cancer Informatics, 2, 59–77. https://doi.org/10.1177/117693510600200030
4. Ministerio de Salud y Protección Social. (2014). Resolución 247 de 2014. Por la cual se establece el reporte para el registro de pacientes con cáncer en Colombia. Bogotá, Colombia: MinSalud. Recuperado de https://www.minsalud.gov.co/Normatividad_Nuevo/Resoluci%C3%B3n%200247%20de%202014.pdf
5. Yala, A., Lehman, C., Schuster, T., Portnoi, T., & Barzilay, R. (2019). A Deep Learning Mammography-based Model for Improved Breast Cancer Risk Prediction. Radiology, 292(1), 60–66. https://doi.org/10.1148/radiol.2019182716
6. Kourou, K., Exarchos, T. P., Exarchos, K. P., Karamouzis, M. V., & Fotiadis, D. I. (2015). Machine learning applications in cancer prognosis and prediction. Computational and Structural Biotechnology Journal, 13, 8–17. https://doi.org/10.1016/j.csbj.2014.11.005
7. Weng, W. H., Wagholikar, K. B., McCray, A. T., Szolovits, P., & Choi, Y. I. (2019). Medical subdomain classification of clinical notes using a machine learning-based natural language processing system. BMC Medical Informatics and Decision Making, 19(1), 1–13. https://doi.org/10.1186/s12911-019-0745-5
8. Liu, J., Huang, Z., Yu, J., Li, S., & Li, X. (2022). Machine learning-based prediction of breast cancer survival using clinical and pathological features. Frontiers in Oncology, 12, 869024. https://doi.org/10.3389/fonc.2022.869024
9. Lundberg, S. M., & Lee, S. I. (2017). A unified approach to interpreting model predictions. Advances in Neural Information Processing Systems, 30 (NIPS 2017), 4765–4774. https://arxiv.org/abs/1705.07874
10. Choi, E., Lee, H., Park, T., et al. (2020). Machine learning–based prediction models for breast cancer prognosis. Scientific Reports, 10, 13938. https://doi.org/10.1038/s41598-020-70732-0
11. Band, S. S., Yarahmadi, A., Hsu, C.-C., Biyari, M., Sookhak, M., Ameri, R., Dehzangi, I., Chronopoulos, A. T., & Liang, H.-W. (2023). Application of explainable artificial intelligence in medical health: A systematic review of interpretability methods. Informatics in Medicine Unlocked, 40, 101286. https://doi.org/10.1016/j.imu.2023.101286
12. Collins, G. S., Reitsma, J. B., Altman, D. G., & Moons, K. G. (2015). Transparent reporting of a multivariable prediction model for individual prognosis or diagnosis (TRIPOD): the TRIPOD statement. Annals of Internal Medicine.
13. Cox, D. R. (1972). Regression models and life-tables. Journal of the Royal Statistical Society. Series B (Methodological), 34(2), 187–220.

14. Harrell, F. E., Jr. (2015). Regression modeling strategies: With applications to linear models, logistic and ordinal regression, and survival analysis (2nd ed.). Springer. https://doi.org/10.1007/978-3-319-19425-7

15. Ishwaran, H., Kogalur, U. B., Blackstone, E. H., & Lauer, M. S. (2008). Random survival forests. The Annals of Applied Statistics, 2(3), 841–860. https://doi.org/10.1214/08-AOAS169

16. Jin, Y., Lan, A., Dai, Y., Jiang, L., & Liu, S. (2023). Development and testing of a random forest-based machine learning model for predicting events among breast cancer patients with a poor response to neoadjuvant chemotherapy. European Journal of Medical Research, 28(1), 394. https://doi.org/10.1186/s40001-023-01575-0

17. Kourou, K., Exarchos, K. P., Papaloukas, C., Sakaloglou, P., Exarchos, T., & Fotiadis, D. I. (2021). Applied machine learning in cancer research: A systematic review for patient diagnosis, classification and prognosis. Computational and Structural Biotechnology Journal, 19, 5546–5555. https://doi.org/10.1016/j.csbj.2021.10.001

18. Song, X., Chu, J., Guo, Z., Wei, Q., Wang, Q., Hu, W., Wang, L., Zhao, W., Zheng, H., Lu, X., & Zhou, J. (2024). Prognostic prediction of breast cancer patients using machine learning models: a retrospective analysis. Gland Surgery, 13(9), 1575–1587. https://doi.org/10.21037/gs-24-331

19. Swanson, K., Wu, E., Zhang, A., Alizadeh, A. A., & Zou, J. (2023). From patterns to patients: Advances in clinical machine learning for cancer diagnosis, prognosis, and treatment. Cell, 186(8), 1772–1791. https://doi.org/10.1016/j.cell.2023.03.002

20. Anders, C. K., & Johnson, R. H. (2008). Breast cancer in the young woman. The Breast, 17(Suppl 1), S19–S21.

21. Early Breast Cancer Trialists' Collaborative Group. (2015). Effects of chemotherapy and hormonal therapy in older women with early breast cancer: an overview of randomized trials. The Lancet, 385(9984), 2307-2319.

22. Slamon, D. J., Leyland-Jones, B., Shak, S., Fuchs, H., Paton, V., Carter, A., ... & Norton, L. (2001). Use of chemotherapy plus a monoclonal antibody against HER2 for metastatic breast cancer that overexpresses HER2. The New England Journal of Medicine, 344(11), 783–792. https://www.nejm.org/doi/pdf/10.1056/NEJM200103153441101

23. Rahib, L., Miller, K. D., & Siegel, R. L. (2024). Cancer statistics, 2024: Progress and challenges in survival outcomes. CA: A Cancer Journal for Clinicians, 74(1), 7–33. https://acsjournals.onlinelibrary.wiley.com/doi/10.3322/caac.21818

24. Chen, J., et al. (2022). Machine learning models based on immunological genes for predicting response to neoadjuvant therapy in breast cancer. Frontiers in Immunology, 13, 948601.

25. Therneau, T. M., & Grambsch, P. M. (2000). Modeling Survival Data: Extending the Cox Model. Springer. https://doi.org/10.1007/978-1-4757-3294-8

26. Pölsterl, S. (2020). Survival analysis for deep learning: State-of-the-art and challenges. arXiv preprint arXiv:2003.13003. https://arxiv.org/abs/2003.13003

27. Rahman, H. A., Nazhirah Zaim, S. N., Suhaimei, U. S., & Jamain, A. A. (2024). Prognostic factors associated with breast cancer-specific survival from 1995 to 2022: A systematic review and meta-analysis of 1,386,663 cases from 30 countries. Diseases, 12(6), 111

28. Zhang, X., Chen, R., Li, J., et al. (2023). Prognostic and predictive factors for breast cancer-specific survival: a meta-analysis. Medicine, 102(18), e33456.

29. Zhang, Y., Liu, S., & Lv, C. (2014). Breast cancer survival defined by the ER/PR/HER2 subtypes and a surrogate classification in the California Cancer Registry. Breast Cancer Research and Treatment, 147(2), 497-506.

30. Giuliano, M., Trivedi, M. V., Schiff, R., & Osborne, C. K. (2022). Age and breast cancer: How much does it matter? Journal of Clinical Oncology, 40(5), 486–495.
31. Weigel, M. T., & Dowsett, M. (2023). The prognostic value of TNM staging and its integration into survival prediction in breast cancer. The Lancet Oncology, 24(3), 261–273.
32. Loibl, S., Poortmans, P., Morrow, M., Denkert, C., & Curigliano, G. (2021). Breast cancer. The Lancet, 397(10286), 1750–1769. https://doi.org/10.1016/S0140-6736(20)32381-3
33. Peduzzi, P., Concato, J., Feinstein, A. R., & Holford, T. R. (1995). Importance of events per independent variable in proportional hazards regression analysis. Journal of Clinical Epidemiology, 48(12), 1503–1510. https://doi.org/10.1016/0895-4356(95)00048-8
34. Vittinghoff, E., & McCulloch, C. E. (2007). Relaxing the rule of ten events per variable in logistic and Cox regression. American Journal of Epidemiology, 165(6), 710–718. https://doi.org/10.1093/aje/kwk052
35. Steyerberg, E. W. (2019). Clinical Prediction Models: A Practical Approach to Development, Validation, and Updating (2nd ed.). Springer. https://link.springer.com/book/10.1007/978-3-030-16399-0
36. Chawla, N. V., Bowyer, K. W., Hall, L. O., & Kegelmeyer, W. P. (2002). SMOTE: Synthetic Minority Over-sampling Technique. Journal of Artificial Intelligence Research, 16, 321–357. https://doi.org/10.1613/jair.953
37. Hosmer, D. W., Lemeshow, S., & Sturdivant, R. X. (2013). Applied Logistic Regression (3rd ed.). Wiley. https://onlinelibrary.wiley.com/doi/book/10.1002/9781118548387
38. Tibshirani, R. (1996). Regression Shrinkage and Selection via the Lasso. Journal of the Royal Statistical Society: Series B (Methodological), 58(1), 267–288. https://doi.org/10.1111/j.2517-6161.1996.tb02080.x
39. Zou, H., & Hastie, T. (2005). Regularization and variable selection via the Elastic Net. Journal of the Royal Statistical Society: Series B (Statistical Methodology), 67(2), 301–320. https://doi.org/10.1111/j.1467-9868.2005.00503.x
40. Fernández, A., García, S., Galar, M., Prati, R. C., Krawczyk, B., & Herrera, F. (2018). Learning from Imbalanced Data Sets. Springer. https://link.springer.com/book/10.1007/978-3-319-98074-4
41. Anderson, W. F., Rosenberg, P. S., & Matsuno, R. K. (2019). Age-specific trends in breast cancer incidence in the United States. JAMA Network Open, 2(3), e190313. https://doi.org/10.1001/jamanetworkopen.2019.0313
42. Chen, T., & Guestrin, C. (2016). XGBoost: A scalable tree boosting system. Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 785–794. https://doi.org/10.1145/2939672.2939785
43. Lundberg, S. M., Erion, G., Chen, H., DeGrave, A., Prutkin, J. M., Nair, B., Katz, R., Himmelfarb, J., Bansal, N., & Lee, S.-I. (2018). Explainable AI for trees: From local explanations to global understanding. Nature Machine Intelligence, 2(1), 56–67. https://doi.org/10.1038/s42256-019-0138-9
44. Niculescu-Mizil, A., & Caruana, R. (2005). Predicting good probabilities with supervised learning. Proceedings of the 22nd International Conference on Machine Learning, 625–632. https://doi.org/10.1145/1102351.1102430
45. Rajkomar, A., Dean, J., & Kohane, I. S. (2019). Machine learning in medicine. New England Journal of Medicine, 380(14), 1347–1358. https://doi.org/10.1056/NEJMra1814259
46. Swain, S. M., Baselga, J., Kim, S. B., Ro, J., Semiglazov, V., Campone, M., ... & Cortés, J. (2015). Pertuzumab, trastuzumab, and docetaxel in HER2-positive metastatic breast cancer

(CLEOPATRA study). New England Journal of Medicine, 372(8), 724–734. https://doi.org/10.1056/NEJMoa1413513

47. Weigelt, B., Reis-Filho, J. S., & Swanton, C. (2017). Breast cancer molecular stratification: From intrinsic subtypes to integrative clusters. Genome Medicine, 9(1), 57. https://doi.org/10.1186/s13073-017-0446-8

Capítulo 11. **ANEXOS**

11.1 Anexo 1 DICCIONARIO BASE DE DATOS.

Este archivo en formato Excel contiene el diccionario de variables correspondiente a la Resolución 247 de 2014 del Ministerio de Salud y Protección Social de Colombia. En él se especifican los nombres, descripciones, tipos de dato, longitudes y valores permitidos para cada campo de la base de datos clínica de cáncer de mama. Este documento sirve como referencia técnica para la interpretación correcta de las variables utilizadas en el modelado.

https://github.com/MMMR1193/DICCIONARIO-BASE-DE-DATOS

11.2 Anexo 2 BD_CA_MAMA_VF

Archivo en formato Excel (.xlsx) que corresponde a la base de datos depurada y estandarizada utilizada en el desarrollo del proyecto. Contiene 1.161 registros de pacientes con diagnóstico confirmado de cáncer de mama y 128 variables clínicas y administrativas.

https://github.com/MMMR1193/BD_CA_MAMA_VFBD_CA_MAMA_VF

11.3 Anexo 3 CÓDIGO BD_CAC_MAMA_VERSION_FINAL

Archivo en formato Notebook de Python (.ipynb) que integra todo el código ejecutado durante la metodología del proyecto. Incluye las etapas de carga, limpieza, preprocesamiento, generación de variables, entrenamiento de modelos de supervivencia y clasificación, así como la evaluación de resultados. Este anexo constituye la evidencia técnica del desarrollo computacional del estudio

https://github.com/MMMR1193/BD_CAC_MAMA_VERSION_FINAL