Decision Support System for Prioritizing Patients with Cardiovascular Risk Using Interpretable Machine Learning

Isaac Zablah[1]; Fiama Garcia[2]; Antonio Garcia Loureiro[3]; Maryori Pereira[4]; Sara Ney[5]; Juan Antonio Hernández[6]; Perla Simons[7]

[1,2,4,6,7]Universidad Nacional Autónoma de Honduras, Honduras, *jose.zablah@unah.edu.hn, fiama.garcia@unah.edu.hn, maryori.pereira@unah.edu.hn, antoniopavon@unah.edu.hn, perla.morales@unah.edu.hn*

[3]Universidad de Santiago de Compostela, España, *antonio.garcia.loureiro@usc.es*

[4]Universidade Federal de Viçosa, Brasil, *sara.ney@ufv.br*

Abstract– ***Cardiovascular disease imposes a high clinical and operational burden on primary care, particularly in resource constrained settings where triage depends on incomplete information and limited consultation capacity. This study proposes and retrospectively validates an interpretable machine learning decision support system to prioritize patients with probable cardiovascular risk. The system integrates routine clinical variables, probabilistic risk estimation, operational thresholds and local explanations for human supervised triage. A secondary retrospective validation was conducted with the open Cleveland Heart Disease dataset from the UCI Machine Learning Repository, compiled in 1988 at a single specialist referral centre among patients already scheduled for coronary angiography, including 303 records and 13 predictors. Four classifiers were compared using five fold cross validation and bootstrap confidence intervals. Random forest achieved the highest discrimination, with ROC AUC of 0.911, while XGBoost offered the best Brier score among machine learning alternatives, 0.123, and was used for operational prioritization. At the selected threshold of 0.599, sensitivity and specificity were both 0.848, prioritizing 161 of 303 patients. The high priority stratum concentrated 68.5% of observed cases. Decision curve analysis, derived from the out of fold confusion matrices, yielded a net benefit of 0.359 at that threshold, above both the prioritize all and the prioritize none strategies across the 0.300 to 0.700 range. Because cohort prevalence reaches 54.5%, the observed positive predictive value of 0.870 would fall to approximately 0.38 at a primary care prevalence of 10%. A head to head comparison against conventional risk scores was not feasible, since the dataset lacks HDL cholesterol, smoking status and an incident outcome. The prototype provides a reproducible framework for cardiovascular triage and an auditable protocol for its own external assessment. Multicentre local validation, recalibration of thresholds and evaluation of acceptability and cost remain preconditions for any clinical deployment.***

Keywords-- ***cardiovascular risk, clinical decision support, decision curve analysis, interpretable machine learning, triage.***

Sistema de apoyo a la decisión para priorizar pacientes con riesgo cardiovascular mediante aprendizaje automático interpretable

Isaac Zablah[1]; Fiama Garcia[2]; Antonio Garcia Loureiro[3]; Maryori Pereira[4]; Sara Ney[5]; Juan Antonio Hernández[6]; Perla Simons[7]
[1,2,4,6,7]Universidad Nacional Autónoma de Honduras, Honduras, *jose.zablah@unah.edu.hn, fiama.garcia@unah.edu.hn, maryori.pereira@unah.edu.hn, antoniopavon@unah.edu.hn, perla.morales@unah.edu.hn*
[3]Universidad de Santiago de Compostela, España, *antonio.garcia.loureiro@usc.es*
[4]Universidade Federal de Viçosa, Brasil, *sara.ney@ufv.br*

Resumen– ***Las enfermedades cardiovasculares representan una carga crítica para los servicios de salud y requieren estrategias oportunas de detección y priorización en atención primaria. Este estudio presenta un sistema interpretable de apoyo a la decisión clínica para ordenar pacientes según probabilidad de enfermedad cardiovascular y generar explicaciones locales útiles para enfermería, medicina general y gestión territorial. Se realizó una validación retrospectiva secundaria con el conjunto Cleveland Heart Disease de UCI, compuesto por 303 registros, 13 predictores clínicos y un desenlace binario. Se compararon regresión logística, bosque aleatorio, gradient boosting y XGBoost mediante validación cruzada de cinco pliegues. El bosque aleatorio obtuvo la mayor AUC ROC (0.911), mientras que XGBoost presentó el mejor Brier entre los modelos de aprendizaje automático (0.123). Con un umbral de 0.599, XGBoost alcanzó sensibilidad y especificidad de 0.848 y priorizó 161 pacientes; el estrato de alto riesgo concentró 68.5% de los casos observados. El análisis de curva de decisión mostró un beneficio neto de 0.359, superior a priorizar a todos o a ninguno entre umbrales de 0.300 y 0.700. Sin embargo, el valor predictivo positivo de 0.870 descendería aproximadamente a 0.38 con una prevalencia de 10%. No fue posible comparar el modelo con escalas convencionales por ausencia de variables esenciales. El sistema constituye un prototipo reproducible que requiere validación multicéntrica local, recalibración de umbrales y evaluación de aceptabilidad y costos antes de su uso clínico.***

Palabras clave-- ***aprendizaje automático interpretable, riesgo cardiovascular, apoyo a la decisión clínica, análisis de curva de decisión, triaje.***

I. Introducción

Las enfermedades cardiovasculares constituyen una de las principales causas de muerte a nivel mundial y concentran una proporción sustancial de muertes prematuras, especialmente en países de ingresos bajos y medios [1]. La respuesta sanitaria recomendada no se limita a la atención hospitalaria de eventos agudos, sino que requiere fortalecer la identificación temprana, el control de factores de riesgo y el manejo longitudinal desde la atención primaria. El paquete técnico HEARTS de la Organización Mundial de la Salud (OMS) propone un abordaje integrado para mejorar la gestión del riesgo cardiovascular en el primer nivel, con protocolos estandarizados, medición adecuada de presión arterial, acceso a medicamentos esenciales, seguimiento basado en equipos y uso operativo de datos [2].

Honduras presenta un contexto particularmente relevante para soluciones de priorización clínica. El *World Heart Observatory* ha reportado una carga significativa de mortalidad cardiovascular en el país, mientras que estudios locales en el occidente hondureño han descrito prevalencias elevadas de hipertensión, diabetes, dislipidemia y subtratamiento de pacientes de alto riesgo [3], [4]. Estos hallazgos sugieren una brecha entre la necesidad clínica y la capacidad instalada para identificar de manera temprana a quienes requieren seguimiento intensificado. La presión sobre centros de salud, médicos generales, enfermería y promotores comunitarios obliga a convertir los datos disponibles en listas de trabajo accionables, trazables y justificables.

El problema abordado en este estudio es la sobrecarga del primer nivel de atención y el triaje clínico insuficiente para pacientes con riesgo cardiovascular. En la práctica, muchas decisiones se toman con información parcial, sin apoyo probabilístico y sin una herramienta que ordene la demanda según riesgo clínico. El resultado puede ser doblemente ineficiente. Por un lado, pacientes de bajo riesgo consumen cupos prioritarios. Por otro, pacientes con probabilidad alta de enfermedad o descompensación permanecen sin evaluación oportuna. Un sistema de apoyo a la decisión puede mejorar la asignación de recursos si entrega una probabilidad calibrada, una clasificación operativa y una explicación clínica comprensible para el usuario final.

El objetivo de este artículo es diseñar y validar retrospectivamente un sistema de apoyo a la decisión para priorizar pacientes con riesgo cardiovascular mediante aprendizaje automático interpretable, , así como delimitar de forma explícita las condiciones bajo las cuales los resultados obtenidos no son transferibles a un entorno asistencial distinto al de la cohorte de desarrollo. El producto científico esperado es un sistema de priorización con validación retrospectiva, diseñado para funcionar como componente complementario al juicio clínico y no como mecanismo autónomo de diagnóstico.

La contribución principal consiste en integrar tres capas técnicas en un flujo reproducible: predicción probabilística, estratificación operativa por umbrales y explicación local orientada al uso clínico.

II. Fundamento tecnológico y clínico

Los modelos clásicos de riesgo cardiovascular, como Framingham, Pooled Cohort Equations, SCORE o MESA, han sido útiles para estratificar poblaciones, pero su desempeño puede variar cuando se aplican fuera de las cohortes que les dieron origen. En América Latina, y particularmente en entornos con limitación de datos longitudinales, la transferencia directa de escalas externas puede producir sesgos de calibración y decisiones subóptimas [4]. A esa limitación se añade una restricción operativa poco discutida: estas escalas exigen predictores que no siempre constan en los registros disponibles. SCORE2 requiere colesterol no HDL y condición de tabaquismo, y las Pooled Cohort Equations incorporan además tratamiento antihipertensivo y diagnóstico de diabetes [5]. La disponibilidad de esas variables determina qué comparaciones son técnicamente ejecutables sobre un conjunto de datos concreto, cuestión que se retoma en la sección III-E. El aprendizaje automático permite explorar relaciones no lineales y combinaciones de variables que no siempre son capturadas por reglas convencionales. Sin embargo, su uso clínico requiere transparencia, auditoría y validación rigurosa.

La literatura reciente sobre modelos predictivos en salud enfatiza que la capacidad discriminativa no basta para justificar implementación. Los modelos deben reportar datos, población, desenlaces, manejo de ausentes, calibración, análisis de umbrales y errores relevantes para el contexto clínico. La guía TRIPOD+AI actualiza las recomendaciones para reportar modelos de predicción basados en regresión o aprendizaje automático [6]. Para sistemas de apoyo a la decisión impulsados por inteligencia artificial en etapas tempranas, DECIDE-AI recomienda describir el flujo de trabajo, la interacción humano sistema, el manejo de salidas, la seguridad y la evaluación exploratoria [7]. En intervenciones evaluadas mediante ensayos, CONSORT-AI añade requisitos para describir la intervención, el entorno de uso y los errores del sistema [8].

Una revisión sistemática y metaanálisis reciente comparó 32 modelos de aprendizaje automático con 26 modelos estadísticos convencionales para predicción cardiovascular a mediano y largo plazo con datos de historias clínicas electrónicas. Los modelos de bosque aleatorio alcanzaron un área bajo la curva (AUC) agrupado de 0.865 (intervalo de confianza (IC) 95%: 0.812–0.917), mientras que los modelos de aprendizaje profundo obtuvieron 0.847 (IC 95%: 0.766–0.927), superando consistentemente a los modelos tradicionales. Otra revisión identificó mejoras modestas pero consistentes en discriminación (ΔAUROC +0.02–0.08) con modelos de árboles de gradiente y redes de supervivencia profunda, junto con barreras persistentes en generalización, interpretabilidad y ausencia de validación externa rigurosa [9], [10].

III. Materiales y métodos

A. *Diseño general del sistema*

Se diseñó un sistema modular de apoyo a la decisión clínica. El primer módulo recibe variables clínicas rutinarias del paciente, como edad, sexo, presión arterial en reposo, colesterol, frecuencia cardiaca máxima, dolor torácico, angina inducida por ejercicio y hallazgos del segmento ST. El segundo módulo normaliza datos, verifica rangos y registra ausencias. El tercer módulo estima una probabilidad individual de enfermedad cardiovascular. El cuarto módulo asigna un estrato operativo, bajo, intermedio o alto, con reglas de priorización ajustables al recurso disponible. Finalmente, el quinto módulo genera explicaciones globales y locales para documentar qué variables influyeron en la priorización. La Figura 1 resume la arquitectura funcional.

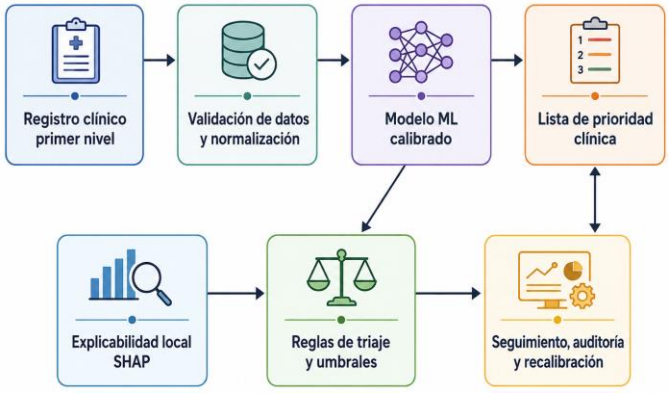


Fig. 1 Arquitectura funcional del sistema de apoyo a la decisión para priorización cardiovascular.

B. *Fuente de datos y variable objetivo*

La validación retrospectiva secundaria usó el conjunto *Heart Disease* de la *UCI Machine Learning Repository*, con licencia *Creative Commons Attribution* 4.0, que integra bases de Cleveland, Hungría, Suiza y VA Long Beach. La versión de Cleveland contiene 303 instancias y 13 predictores ampliamente utilizados en experimentos publicados [11]. Los registros fueron recolectados en 1988 en un centro cardiológico de referencia, entre pacientes derivados a angiografía coronaria por sospecha clínica ya establecida. Ese origen determina dos rasgos que condicionan toda interpretación posterior. El primero es un sesgo de espectro, puesto que la población había sido filtrada previamente por un proceso de derivación especializada. El segundo es un efecto de época, dado que la práctica diagnóstica y terapéutica anterior a la difusión de las troponinas de alta sensibilidad y de la terapia con estatinas difiere de la vigente. La variable

objetivo original representa presencia de enfermedad cardiaca angiográfica, con valores de 0 a 4. En este estudio se usó la codificación binaria disponible en la copia reproducible del archivo de trabajo, donde 1 indica presencia y 0 ausencia de enfermedad. El artículo original asociado a la base evaluó un algoritmo probabilístico para diagnóstico de enfermedad arterial coronaria en diferentes grupos de pacientes [12].

La base analizada incluyó 303 registros, 13 predictores y 165 casos positivos, equivalentes a 54.5% del total. Esta prevalencia supera varias veces la esperable en una consulta de primer nivel no seleccionada, hecho cuyas consecuencias sobre el valor predictivo positivo se cuantifican en la sección de discusión. No se identificaron valores ausentes en la copia de archivo separado por comas (CSV) utilizada. La Tabla I resume las variables analíticas. Aunque se trata de una base abierta internacional y no de una cohorte hondureña, resulta pertinente para verificar la consistencia interna del flujo técnico de priorización y para definir el protocolo de una futura validación externa local. No permite, en cambio, estimar el desempeño esperable en establecimientos hondureños de primer nivel.

TABLA I
GRUPOS DE VARIABLES UTILIZADAS

Grupo	Variables	Contenido
Demográficas	age, sex	Edad y sexo codificado
Presión y laboratorio	trestbps, chol, fbs	Presión arterial, colesterol y glucosa
Síntomas y ejercicio	cp, exang, thalach	Dolor torácico, angina y frecuencia cardíaca
Electrocardiografía	restecg, oldpeak, slope	ECG basal y cambios del ST
Marcadores anatómicos	ca, thal	Vasos mayores y talasemia codificada

C. *Preprocesamiento y modelos comparados*

Las variables continuas fueron imputadas por mediana cuando correspondía y estandarizadas para regresión logística. Las variables categóricas codificadas fueron tratadas como entradas discretas, conservando su representación numérica original para modelos basados en árboles. Se compararon cuatro familias de modelos: regresión logística con ponderación de clases, bosque aleatorio, gradient boosting y XGBoost. La regresión logística se incluyó como modelo transparente de referencia, mientras que los modelos de árboles se utilizaron por su capacidad de capturar relaciones no lineales. XGBoost fue incluido por su eficiencia y amplio uso en problemas tabulares [13]. La implementación se realizó en Python con scikit-learn, biblioteca ampliamente utilizada para aprendizaje automático reproducible [14].

La interpretabilidad global se evaluó mediante la importancia de variables derivada del modelo XGBoost (ganancia promedio). La generación de explicaciones locales tipo SHAP ("SHapley Additive exPlanations") se define como componente del diseño del sistema para fases futuras de implementación, pero no fue ejecutada en la validación retrospectiva reportada en este estudio [15], [16]. En un despliegue real, estas explicaciones deberían mostrarse como apoyo al razonamiento clínico, evitando que el usuario interprete la salida como diagnóstico definitivo.

D. *Validación retrospectiva y métrica*

La evaluación se realizó mediante validación cruzada estratificada de cinco pliegues. Se reportaron AUC ROC, AUC de precisión recuperación, exactitud, sensibilidad, especificidad, F1 y Brier score. La incertidumbre se estimó con remuestreo bootstrap sobre predicciones fuera de pliegue, generando intervalos empíricos de 95%. Para el módulo de priorización se analizaron umbrales operativos entre 0.300 y 0.700. El umbral principal fue seleccionado maximizando el índice de Youden bajo una restricción mínima de sensibilidad de 0.800, debido a que el error más costoso en triaje es no priorizar a un paciente con enfermedad probable.

Los estratos operativos se definieron así: riesgo bajo, probabilidad menor de 0.400; riesgo intermedio, probabilidad entre 0.400 y 0.699; riesgo alto, probabilidad igual o mayor que 0.700. Esta definición no pretende sustituir la evaluación clínica, sino ordenar la agenda de seguimiento, orientar llamadas comunitarias, priorizar toma de presión arterial, control de lípidos, electrocardiograma, evaluación médica o referencia según disponibilidad.

Adicionalmente, se realizó un análisis de curva de decisión (Decision Curve Analysis, DCA) para cuantificar el beneficio neto del modelo XGBoost a lo largo del rango de umbrales de probabilidad. El beneficio neto fue calculado como la Fórmula (1):

$$\text{Beneficio neto} = \frac{\text{verdaderos positivos}}{n} - \frac{\text{falsos positivos}}{n} \times \frac{p_t}{1 - p_t} \quad (1)$$

donde p_t es la probabilidad umbral y n el tamaño de la muestra. La curva de decisión fue comparada con las estrategias por defecto de priorizar a todos los pacientes y no priorizar a ninguno. Este análisis permite evaluar si el uso del modelo produce un beneficio clínico real, más allá de las métricas de discriminación y calibración [17], [18]. El beneficio neto se calculó a partir de las matrices de confusión fuera de pliegue correspondientes a cada umbral analizado, de manera que los valores reportados proceden de las mismas predicciones empleadas para estimar sensibilidad y especificidad. Se calcularon además el beneficio neto estandarizado, definido como el cociente entre el beneficio neto y la prevalencia observada, y el número de priorizaciones evitadas por cada 100 pacientes respecto de la estrategia de priorizar a todos [19].

Se realizó una autoevaluación del riesgo de sesgo del estudio utilizando el marco PROBAST+AI, que incluye 16 preguntas de señalización para desarrollo de modelos distribuidas en cuatro dominios: participantes, predictores, desenlace y análisis. El dominio de participantes fue clasificado como riesgo moderado debido al uso de una cohorte internacional retrospectiva no representativa del contexto hondureño. El

dominio de predictores fue clasificado como riesgo bajo, dado que las 13 variables son medibles de forma estandarizada. El dominio de desenlace presenta riesgo moderado porque la variable objetivo se basa en angiografía, un estándar de referencia no disponible en atención primaria. El dominio de análisis fue clasificado como riesgo moderado por el tamaño muestral reducido (303 registros, razón eventos-por-variable de 12.7:1) y la ausencia de validación externa [20].

C. Comparador de referencia y análisis no ejecutados

La regresión logística con ponderación de clases operó como comparador estadístico transparente frente a los modelos de ensamble. No fue posible, en cambio, ejecutar una comparación directa contra escalas clínicas convencionales. Framingham, las Pooled Cohort Equations y SCORE2 requieren colesterol HDL, condición de tabaquismo y tratamiento antihipertensivo, variables ausentes en el conjunto de Cleveland, y estiman el riesgo de eventos incidentes en un horizonte de diez años, mientras que la variable objetivo disponible es la presencia transversal de enfermedad angiográfica. Aplicar esas escalas sobre esta fuente exigiría imputar predictores nunca medidos y equiparar dos desenlaces de naturaleza distinta, procedimiento que generaría una comparación desprovista de validez interpretativa. Por razones análogas, el análisis de equidad se restringió al sexo. La cohorte no registra ruralidad, nivel educativo, aseguramiento ni adscripción territorial, y la estratificación por grupos etarios fragmentaría la muestra en subgrupos demasiado pequeños para una estimación estable.

IV. RESULTADOS

A. Desempeño predictivo

La Tabla II presenta el desempeño comparativo. El bosque aleatorio alcanzó la mayor AUC ROC, 0.911, con intervalo bootstrap de 0.881 a 0.940. XGBoost obtuvo AUC ROC de 0.904 y el menor Brier score entre los modelos de aprendizaje automático, 0.123, con intervalo de 0.105 a 0.145. La regresión logística mostró desempeño competitivo, con AUC ROC de 0.895 y sensibilidad de 0.909 al umbral de 0.500, lo que confirma que una línea base estadística interpretable puede ser sólida en conjuntos tabulares pequeños. Las diferencias entre familias de modelos fueron menores que la amplitud de sus intervalos bootstrap, que se solapan en su totalidad, por lo que los datos no sustentan la superioridad estadística de ninguna de ellas. No obstante, el componente XGBoost fue retenido para la demostración operativa por su equilibrio entre calibración, sensibilidad y especificidad al umbral seleccionado.

La Figura 2 muestra que todos los modelos superaron claramente la línea de no discriminación. La diferencia entre métodos fue moderada, un resultado esperable en una base pequeña y con predictores clínicos altamente informativos. La Figura 3 muestra la calibración por cuantiles. La calibración no fue perfecta, pero XGBoost presentó el menor error cuadrático probabilístico, lo que favorece su uso como sistema de priorización cuando la salida principal es una probabilidad y no una etiqueta dicotómica rígida.

TABLA II
DESEMPEÑO DE MODELOS DE VALIDACIÓN CRUZADA

Modelo	AUC ROC	AUC PR	Exactitud	Sensibilidad	Especificidad	F1	Brier
Regresión logística	0.895 (0.849 - 0.927)	0.901 (0.860 - 0.936)	0.848 (0.795 - 0.885)	0.909 (0.859 - 0.954)	0.775 (0.698 - 0.841)	0.867 (0.826 - 0.903)	0.127 (0.103 - 0.156)
Bosque aleatorio	0.911 (0.881 - 0.940)	0.919 (0.884 - 0.951)	0.838 (0.795 - 0.884)	0.885 (0.825 - 0.932)	0.783 (0.707 - 0.839)	0.856 (0.808 - 0.898)	0.127 (0.111 - 0.143)
Gradient boosting	0.898 (0.861 - 0.928)	0.901 (0.857 - 0.939)	0.799 (0.756 - 0.842)	0.873 (0.820 - 0.918)	0.710 (0.636 - 0.792)	0.825 (0.781 - 0.864)	0.127 (0.107 - 0.149)
XGBoost	0.904 (0.864 - 0.932)	0.903 (0.846 - 0.942)	0.818 (0.776 - 0.852)	0.891 (0.847 - 0.929)	0.732 (0.651 - 0.790)	0.842 (0.803 - 0.878)	0.123 (0.105 - 0.145)

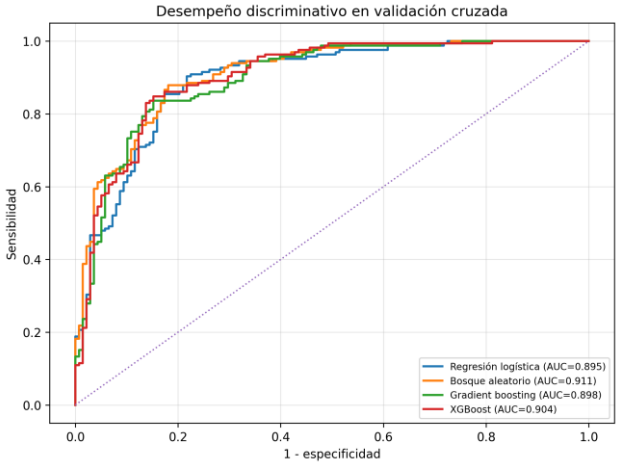


Fig. 2 Curvas ROC de los modelos evaluados mediante predicciones fuera de pliegue.

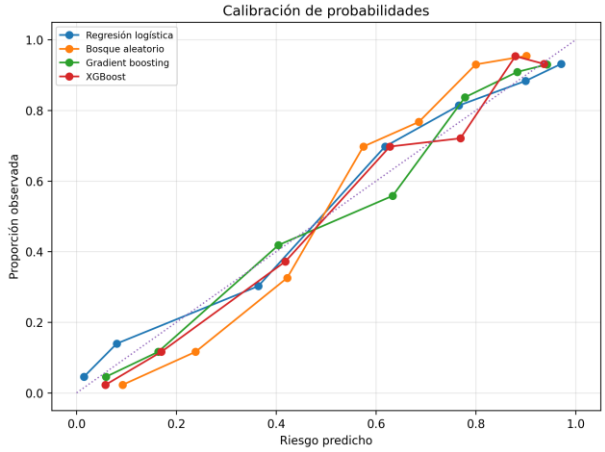


Fig. 3 Curvas de calibración por cuantiles de los modelos evaluados.

B. *Umbral operativo y priorización*

El umbral principal seleccionado para XGBoost fue 0.599. A ese punto, el sistema alcanzó sensibilidad de 0.848, especificidad de 0.848, valor predictivo positivo de 0.870 y valor predictivo negativo de 0.824. Operativamente, priorizó 161 pacientes, equivalentes a 53.1% de la cohorte, con 140 verdaderos positivos y 21 falsos positivos. Para un centro de salud saturado, esta lógica permite convertir una demanda inicial no ordenada en una lista de atención escalonada.

La Tabla III evidencia el intercambio entre cobertura y precisión. Con umbral de 0.300 se prioriza 69.0% de pacientes y se obtiene sensibilidad de 0.958, apropiada para campañas de búsqueda activa cuando la capacidad de seguimiento es amplia. Con umbral de 0.700 se prioriza 42.6% y la especificidad aumenta a 0.877, útil cuando la consulta médica o el electrocardiograma son recursos críticos. El umbral de 0.599 representa un punto balanceado para triaje rutinario.

TABLA III
UMBRALES OPERATIVOS DE PRIORIZACIÓN

Umbral	% Priorizado	Sensibilidad	Especificidad
0.300	69.000	0.958	0.630
0.400	65.700	0.927	0.667
0.500	60.700	0.891	0.732
0.599	53.100	0.848	0.848
0.600	52.800	0.842	0.848
0.700	42.600	0.679	0.877

La estratificación por probabilidades produjo tres grupos clínicamente diferenciados, como muestra la Figura 4. El estrato bajo incluyó 104 pacientes, con 12 casos observados y riesgo promedio de 0.147. El estrato intermedio incluyó 69 pacientes, con 40 casos y riesgo promedio de 0.573. El estrato alto incluyó 130 pacientes, con 113 casos y riesgo promedio de 0.862. Este último grupo concentró 68.5% de todos los casos, lo que respalda su utilidad como cola prioritaria de intervención.

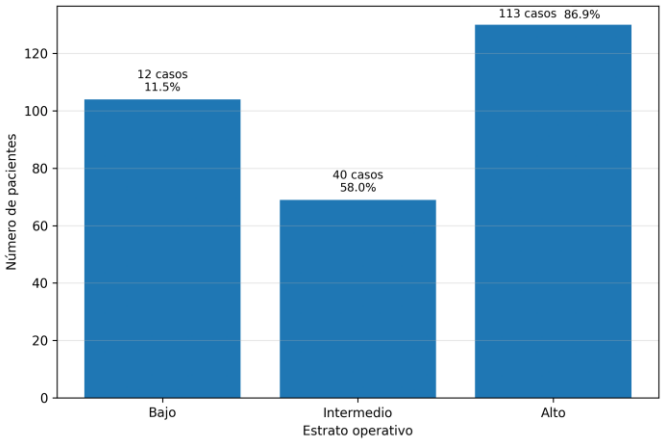


Fig. 4 Distribución de pacientes por estrato operativo y proporción observada de casos.

C. *Análisis de curva de decisión*

El beneficio neto del modelo XGBoost superó a las dos estrategias por defecto en todo el rango de umbrales examinado, según se resume en la Tabla IV. En el umbral operativo de 0.599 el beneficio neto fue 0.359, frente a −0.136 de la estrategia de priorizar a todos los pacientes y 0 de la estrategia de no priorizar a ninguno. Esa magnitud equivale a identificar cerca de 36 casos verdaderos por cada 100 pacientes evaluados sin incurrir en falsos positivos adicionales, con un beneficio neto estandarizado de 0.659. Respecto de priorizar a todos, el modelo evita alrededor de 33 priorizaciones innecesarias por cada 100 pacientes sin perder casos verdaderos. El cruce entre la curva del modelo y la de priorizar a todos ocurre por debajo de 0.300, lo que indica que la ventaja del modelo se concentra en escenarios donde la capacidad de seguimiento es limitada y el umbral de intervención es alto. Conviene subrayar que el análisis se estimó sobre las mismas predicciones fuera de pliegue utilizadas en el desarrollo, por lo que constituye una medida de consistencia interna y previsiblemente optimista del beneficio clínico.

D. *Interpretabilidad*

La Figura 5 resume la importancia global del modelo XGBoost. Las variables con mayor contribución fueron tipo de dolor torácico, talasemia, número de vasos mayores coloreados por fluoroscopia, angina inducida por ejercicio, depresión ST relativa al reposo, pendiente del segmento ST y frecuencia cardiaca máxima. Este patrón es clínicamente plausible porque agrupa síntomas, respuesta al ejercicio y hallazgos electrocardiográficos. Para implementación en primer nivel, el sistema no debería exponer tecnicismos estadísticos al usuario, sino una explicación breve, por ejemplo: prioridad alta por dolor torácico compatible, angina inducida por ejercicio y alteraciones del ST.

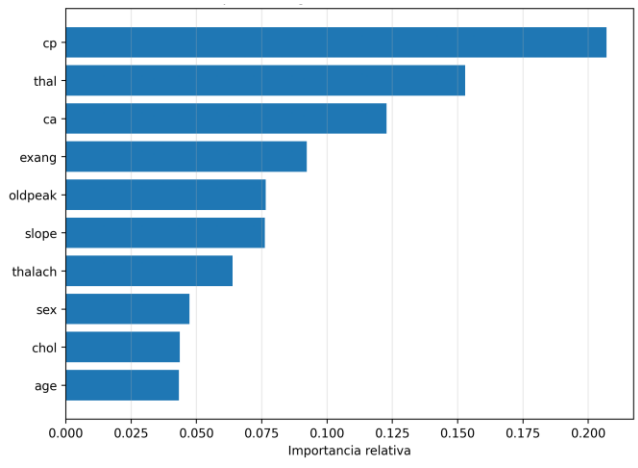


Fig. 5 Importancia global de variables del modelo XGBoost usado para priorización.

E. *Análisis de equidad por sexo*

Se evaluó el desempeño del modelo XGBoost de forma desagregada por sexo. En el subgrupo masculino (n=206), la AUC ROC fue 0.886 y el Brier score fue 0.137. En el subgrupo femenino (n=97), la AUC ROC fue 0.905 y el Brier score fue 0.094. La diferencia en AUC ROC entre subgrupos fue de -0.019 a favor del subgrupo femenino (intervalo bootstrap 95%: -0.097 a 0.067). Dado que ese intervalo contiene el cero, los datos no sustentan ni descartan un desempeño diferencial por sexo: el estudio carece de potencia para dirimirlo. Se calcularon la tasa de paridad de oportunidades (equalized odds) y la paridad predictiva entre grupos. La razón de equalized odds fue 0.232, con una diferencia máxima absoluta entre sensibilidad y tasa de falsos positivos de 0.092, valor determinado por el cociente de tasas de falsos positivos entre subgrupos, magnitud reducida en ambos grupos y por ello inestable, mientras que la diferencia máxima absoluta entre las tasas correspondientes fue de 0.092. La discrepancia entre ambas métricas ilustra que las razones resultan poco informativas cuando el denominador es pequeño, como ocurre en el subgrupo femenino (n=97). La razón de paridad predictiva, estimada mediante el valor predictivo positivo entre grupos, fue 0.987. Esta evaluación no constituye una auditoría de equidad completa, pero ofrece transparencia mínima sobre el comportamiento diferencial del modelo, siguiendo recomendaciones recientes sobre equidad en aprendizaje automático clínico [21], [22]. Una auditoría completa exigiría estratificación por edad, ruralidad, aseguramiento y acceso efectivo a servicios, con tamaños de subgrupo suficientes para una estimación estable, condición que esta cohorte no satisface.

V. Discusión

Los resultados muestran que un sistema de apoyo a la decisión basado en aprendizaje automático interpretable puede ordenar pacientes con probabilidad cardiovascular usando variables clínicas relativamente accesibles. La magnitud de la AUC ROC, cercana a 0.90 en los principales modelos, sugiere capacidad discriminativa suficiente para un prototipo de priorización. Sin embargo, la interpretación debe ser prudente. La base Cleveland fue construida en un entorno diagnóstico especializado y no reproduce la epidemiología, el acceso, la mezcla de pacientes ni la calidad de datos del primer nivel hondureño. Por tanto, el hallazgo central no es que el modelo esté listo para uso clínico local, sino que el flujo técnico de priorización es reproducible, auditable y transferible a una validación externa con datos nacionales.

La consecuencia de esa diferencia admite cuantificación. La discriminación medida por AUC ROC se mantiene relativamente estable ante cambios de prevalencia, pero el valor predictivo positivo no. Con sensibilidad y especificidad de 0.848, el valor predictivo positivo observado de 0.870 corresponde a una prevalencia de 54.5%, propia de una población derivada a angiografía. Si ese mismo punto de operación se aplicara en consulta de primer nivel con una prevalencia de 10%, el valor predictivo positivo descendería a aproximadamente 0.38, y con una prevalencia de 5% a cerca de 0.23. En otras palabras, entre tres y cuatro de cada cinco pacientes priorizados no presentarían la condición buscada. El cálculo no invalida el flujo técnico, pero delimita con precisión qué debe recalibrarse antes de cualquier uso asistencial: el umbral operativo y los cortes de estrato, y no solamente los coeficientes del modelo.

A la distancia contextual se suma una distancia temporal. Los registros datan de 1988 y anteceden a la incorporación rutinaria de troponinas de alta sensibilidad, a la difusión de la terapia con estatinas y a los cambios en el perfil poblacional de factores de riesgo ocurridos en las tres décadas siguientes. Predictores como el número de vasos coloreados por fluoroscopia reflejan además una práctica diagnóstica que la angiografía por tomografía computarizada ha desplazado de forma parcial. El conjunto conserva valor como banco de pruebas metodológico ampliamente documentado y comparable entre estudios, pero no como representación de la población que hoy acude a un centro de salud.

La contribución práctica del sistema está en su diseño operativo. A diferencia de un clasificador que solo emite positivo o negativo, el prototipo produce una probabilidad, un estrato, un umbral ajustable y una explicación. Esta combinación permite adaptar la herramienta a diferentes escenarios de capacidad. En una brigada comunitaria, el umbral puede bajar para aumentar sensibilidad. En una clínica con cupos cardiológicos limitados, puede subir para priorizar especificidad. En un programa municipal de enfermedades no transmisibles, el estrato alto puede alimentar una lista semanal de llamadas, visitas domiciliarias, confirmación de presión arterial y control de laboratorio.

El análisis de curva de decisión aporta un argumento de naturaleza distinta al de la discriminación. Un modelo puede separar bien las clases y, aun así, no ofrecer ventaja frente a decidir por defecto. En este caso el beneficio neto superó ambas estrategias de referencia en todo el rango examinado, lo que sostiene la coherencia interna de la propuesta de priorización. La magnitud del beneficio, sin embargo, depende de la prevalencia de la cohorte y fue estimada sobre las mismas predicciones empleadas en el desarrollo, de modo que debe leerse como evidencia de consistencia y no como demostración de utilidad clínica.

La ausencia de comparación frente a escalas convencionales merece una precisión. No se trata de un análisis omitido, sino de un análisis inejecutable sobre esta fuente de datos, por las razones expuestas en la sección III-E. La regresión logística cumplió la función de comparador transparente y su

desempeño, con AUC ROC de 0.895, resultó indistinguible del de los modelos de ensamble una vez considerada la amplitud de los intervalos bootstrap. Ese resultado tiene una lectura práctica relevante para entornos con recursos computacionales y humanos limitados: en conjuntos tabulares pequeños con predictores clínicos informativos, la ganancia atribuible a los métodos de ensamble puede no compensar su costo en opacidad, mantenimiento y dependencia de bibliotecas externas. La comparación pertinente frente a Framingham, las Pooled Cohort Equations o SCORE2 corresponde a la fase de validación local, cuando se disponga de los predictores que esas escalas requieren y de desenlaces incidentes verificados.

Desde una perspectiva de salud pública, la utilidad del sistema no se limita a clasificar individualmente a los pacientes según probabilidad de enfermedad cardiovascular. Su valor potencial también reside en transformar datos clínicos rutinarios en acciones organizadas de prevención, vigilancia epidemiológica, seguimiento longitudinal y planificación territorial. La identificación temprana de personas con mayor probabilidad de enfermedad puede apoyar intervenciones dirigidas a individuos, familias y comunidades con mayor vulnerabilidad, así como facilitar el monitoreo de la evolución del riesgo y la evaluación del impacto de las acciones desarrolladas por los equipos de atención primaria.

Es importante señalar que las dos variables con mayor contribución al modelo, número de vasos mayores coloreados por fluoroscopia y tipo de talasemia, son hallazgos diagnósticos especializados que no se obtienen de forma rutinaria en el primer nivel de atención. Este desacoplamiento entre las variables más influyentes del modelo y las variables accesibles en la práctica cotidiana implica que un modelo local debería reconfigurarse con predictores como índice de masa corporal, perímetro abdominal, hemoglobina glicosilada, tabaquismo, antecedentes familiares y presión arterial sistólica y diastólica. Futuros trabajos deberían evaluar si un modelo restringido a variables de atención primaria mantiene capacidad discriminativa suficiente para justificar su implementación como herramienta de triaje.

Desde la perspectiva de gestión sanitaria, el sistema puede integrarse como un tablero de priorización, no como una aplicación aislada. Su valor de negocio público radica en reducir colas clínicas mal ordenadas, disminuir citas de baja prioridad en agendas críticas, orientar compras de insumos diagnósticos y justificar intervenciones territoriales ante tomadores de decisión. Para escalamiento, el componente más importante no es el algoritmo, sino la gobernanza de datos, la interoperabilidad con registros clínicos y el protocolo de respuesta ante cada estrato. Sin esta capa organizacional, el modelo puede convertirse en una puntuación sin efecto real en resultados sanitarios.

La transparencia es una condición de uso, no un atributo accesorio. El personal de salud debe poder conocer qué variables influyeron en la priorización y cuándo ignorar la recomendación. Un paciente con síntomas agudos debe seguir rutas de urgencia, aunque el modelo estime bajo riesgo. De igual forma, un paciente con datos incompletos, embarazo, comorbilidad grave o signos de alarma requiere evaluación clínica directa. El sistema debe registrar anulaciones clínicas, falsos negativos, falsos positivos y cambios de umbral, pues esa auditoría permite recalibrar el modelo y mejorar seguridad.

VI. Modelo de implementación y sostenibilidad

Se propone una implementación por fases. La fase 1 consiste en validación retrospectiva local y multicéntrica con registros anonimizados de establecimientos de primer nivel, incorporando presión arterial, antecedentes, diabetes, tabaquismo, índice de masa corporal, lípidos cuando estén disponibles y eventos clínicos. La fase 2 corresponde a una evaluación silenciosa prospectiva, donde el sistema calcula prioridades sin modificar decisiones clínicas y se documenta el grado de concordancia entre la prioridad calculada y la asignada por el personal. La fase 3 habilita un uso supervisado por personal de salud, con tablero de pacientes, bitácora de explicación y registro de decisiones. La fase 4 evalúa impacto operativo, tiempos de atención, proporción de pacientes de alto riesgo contactados, adherencia al seguimiento y derivaciones oportunas. . Esta última fase incorpora, además, las tres dimensiones exigibles a cualquier afirmación de utilidad clínica. La primera es la aceptabilidad de los usuarios, medida con instrumentos validados de usabilidad percibida y entrevistas al personal de enfermería y medicina general. La segunda es la frecuencia y el motivo de las anulaciones de la recomendación del sistema, indicador directo de confianza y de seguridad. La tercera es un análisis de costo-consecuencia que estime el costo incremental por caso de alto riesgo identificado y efectivamente contactado, comparado con la práctica de asignación de agenda vigente. Sin estas mediciones, el beneficio neto estimado en la sección IV-C permanece como una propiedad del modelo y no como un efecto sobre el sistema de atención

La implementación supervisada debe acompañarse de una estrategia formal de capacitación del talento humano en salud. Esta capacitación debería incluir el manejo técnico de la herramienta, la interpretación clínica de probabilidades y estratos de riesgo, la identificación de situaciones en las que debe prevalecer el juicio clínico, y los principios éticos relacionados con inteligencia artificial, protección de datos personales y uso responsable de la información clínica. El sistema debe ser presentado como apoyo al razonamiento profesional y no como sustituto de la valoración realizada por médicos, enfermería u otros integrantes del equipo de salud.

La sostenibilidad tecnológica puede basarse en una arquitectura ligera. El modelo puede ejecutarse en un servidor institucional, una red local o una instancia en nube de bajo

costo. La interfaz puede ser web, con autenticación por rol, exportación de listas y control de auditoría. El mantenimiento requiere monitorear deriva de datos, recalibración semestral o anual y revisión clínica de umbrales. En términos de negocio social, la solución puede licenciarse como software institucional, ofrecerse como servicio a municipalidades o integrarse a programas universitarios de innovación pública, siempre bajo una política de datos que preserve confidencialidad y evite decisiones automatizadas no supervisadas.

VII. Limitaciones

Este estudio tiene limitaciones importantes. Primero, la validación fue retrospectiva y secundaria, basada en una base internacional abierta. No debe interpretarse como validación clínica para Honduras. Segundo, el tamaño muestral fue reducido, con 303 registros, lo que limita estabilidad de métricas y estimación de calibración. Tercero, los datos fueron recolectados en 1988 en un único centro de referencia, entre pacientes ya derivados a angiografía coronaria, lo que introduce sesgo de espectro y efecto de época; la prevalencia resultante de 54.5% excede varias veces la de una consulta de primer nivel no seleccionada, con las consecuencias sobre el valor predictivo positivo descritas en la sección de discusión. Cuarto, algunas variables de la base Cleveland, como fluoroscopia o talasemia, no están disponibles rutinariamente en el primer nivel de atención, por lo que un modelo local debería usar variables más accesibles o trabajar con módulos escalonados. Quinto, la variable objetivo representa enfermedad angiográfica y no eventos cardiovasculares prospectivos. Sexto, la equidad por sexo, edad, ruralidad y pertenencia territorial no pudo evaluarse con profundidad por ausencia de variables contextuales. Séptimo, no se ejecutó comparación frente a escalas de riesgo convencionales, por las razones de disponibilidad de predictores y de naturaleza del desenlace expuestas en la sección III-E; en consecuencia, el estudio no autoriza ninguna afirmación de superioridad respecto de los instrumentos actualmente en uso. Octavo, el análisis de curva de decisión se estimó sobre predicciones fuera de pliegue de la misma cohorte de desarrollo y sin validación externa, por lo que sobreestima de manera previsible el beneficio alcanzable en condiciones reales.

Además, la transferencia del modelo al contexto hondureño requiere considerar diferencias epidemiológicas, socioculturales, territoriales y organizacionales que no están representadas en la base Cleveland. Por ello, una futura validación externa local, preferentemente multicéntrica y con muestreo de establecimientos de distinta capacidad resolutiva y ubicación geográfica, debería incorporar variables clínicas accesibles en el primer nivel de atención, así como determinantes sociales y contextuales relevantes para la población donde se pretende implementar la herramienta. Este proceso permitiría recalibrar el modelo, evaluar su aplicabilidad real y reducir el riesgo de decisiones subóptimas derivadas de una cohorte de desarrollo no representativa.

La razón de eventos por variable fue de 12.7 (165 eventos para 13 predictores), un valor que cumple el criterio mínimo convencional de 10 para regresión logística pero que puede ser insuficiente para modelos de ensamble con mayor flexibilidad paramétrica. Estudios recientes sugieren que los modelos basados en árboles de gradiente requieren razones EPV sustancialmente mayores para lograr estimaciones estables de desempeño, especialmente cuando se reportan métricas de calibración [23]. Esta limitación refuerza la necesidad de interpretar las métricas reportadas como estimaciones preliminares sujetas a alta variabilidad muestral.

En futuras investigaciones se recomienda construir una cohorte local y multicéntrica, retrospectiva y prospectiva con desenlaces clínicos definidos, comparar el modelo con escalas de riesgo convencionales, una vez disponibles los predictores que estas requieren, evaluar calibración por subgrupos, incorporar análisis de decisión clínica, estimar costos evitados y medir aceptabilidad por usuarios. La validación externa debe realizarse antes de cualquier despliegue con impacto asistencial.

VIII. Conclusiones

Se desarrolló y validó retrospectivamente un prototipo de sistema de apoyo a la decisión para priorizar pacientes con riesgo cardiovascular mediante aprendizaje automático interpretable. En la base Cleveland Heart Disease, los modelos evaluados alcanzaron desempeño discriminativo alto, con AUC ROC entre 0.895 y 0.911. XGBoost permitió una priorización operativa balanceada con umbral de 0.599, sensibilidad de 0.848, especificidad de 0.848 y concentración de 68.5% de los casos en el estrato alto. El análisis de curva de decisión, derivado de las matrices de confusión fuera de pliegue, mostró un beneficio neto de 0.359 en ese umbral, superior a las estrategias de priorizar a todos y de no priorizar a ninguno en todo el rango examinado. El sistema ofrece una ruta técnica viable para transformar registros clínicos en listas de priorización explicables, ajustables y auditables. Estos resultados acreditan la consistencia interna del flujo propuesto, pero no constituyen evidencia de utilidad clínica en Honduras: proceden de una cohorte recolectada en 1988, en un único centro especializado y con una prevalencia muy superior a la del primer nivel de atención. Su aplicación real en Honduras exige validación externa multicéntrica, comparación directa con escalas convencionales, recalibración local, integración con procesos de atención primaria, capacitación del talento humano, protocolos de actuación por nivel de riesgo y gobernanza estricta de datos.

AGRADECIMIENTOS

Agradecemos el apoyo recibido del Instituto de Investigación en Ciencias Médicas y Derecho a la Salud (ICIMEDES) de la Facultad de Ciencias Médicas (FCM) y la Dirección de Investigaciones Científicas, Humanísticas y Tecnológicas (DICIHT) todas ellas de la Universidad Nacional Autónoma de Honduras (UNAH).

REFERENCIAS

[1] [1] World Health Organization, “Cardiovascular diseases (CVDs),” WHO Fact Sheet, Jul. 31, 2025. [Online]. Available: https://www.who.int/news-room/fact-sheets/detail/cardiovascular-diseases-(cvds). Accessed: May 14, 2026.

[2] World Health Organization, HEARTS: Technical Package for Cardiovascular Disease Management in Primary Health Care: Risk-Based CVD Management. Geneva, Switzerland: WHO, 2020. [Online]. Available: https://www.who.int/publications/i/item/9789240001367. Accessed: May 14, 2026.

[3] World Heart Federation, “Heart health in Honduras,” World Heart Observatory, 2025. [Online]. Available: https://world-heart-federation.org/world-heart-observatory/countries/honduras/. Accessed: May 14, 2026.

[4] E. E. Montalvan Sanchez et al., “Cardiovascular risk assessment in the resource limited setting of Western Honduras: An epidemiological perspective,” Int. J. Cardiol. Heart Vasc., vol. 27, Art. no. 100476, Apr. 2020, doi: 10.1016/j.ijcha.2020.100476.

[5] SCORE2 Working Group and ESC Cardiovascular Risk Collaboration, “SCORE2 risk prediction algorithms: New models to estimate 10-year risk of cardiovascular disease in Europe,” Eur. Heart J., vol. 42, no. 25, pp. 2439–2454, Jul. 2021, doi: 10.1093/eurheartj/ehab309.

[6] G. S. Collins et al., “TRIPOD+AI statement: Updated guidance for reporting clinical prediction models that use regression or machine learning methods,” BMJ, vol. 385, Art. no. e078378, Apr. 2024, doi: 10.1136/bmj-2023-078378.

[7] B. Vasey et al., “Reporting guideline for the early-stage clinical evaluation of decision support systems driven by artificial intelligence: DECIDE-AI,” Nat. Med., vol. 28, no. 5, pp. 924–933, May 2022, doi: 10.1038/s41591-022-01772-9.

[8] X. Liu et al., “Reporting guidelines for clinical trial reports for interventions involving artificial intelligence: The CONSORT-AI extension,” Nat. Med., vol. 26, no. 9, pp. 1364–1374, Sep. 2020, doi: 10.1038/s41591-020-1034-x.

[9] T. Liu, A. Krentz, L. Lu, and V. Curcin, "Machine learning based prediction models for cardiovascular disease risk using electronic health records data: Systematic review and meta-analysis", *Eur. Heart J. Digit. Health*, vol. 6, no. 1, pp. 7–22, Oct. 2024, doi: 10.1093/ehjdh/ztae080

[10] T. Liu, A. J. Krentz, Z. Huo, and V. Ćurčin, "Opportunities and challenges of cardiovascular disease risk prediction for primary prevention using machine learning and electronic health records: A systematic review", *Rev. Cardiovasc. Med.*, vol. 26, no. 4, Art. no. 37443, Apr. 2025, doi: 10.31083/RCM37443.

[11] A. Janosi, W. Steinbrunn, M. Pfisterer, and R. Detrano, Heart Disease. Irvine, CA, USA: UCI Mach. Learn. Repository, 1988, doi: 10.24432/C52P4X.

[12] R. Detrano, A. Jánosi, W. Steinbrunn, M. Pfisterer, J. Schmid, S. Sandhu, K. Guppy, S. Lee, and V. Froelicher, “International application of a new probability algorithm for the diagnosis of coronary artery disease,” Am. J. Cardiol., vol. 64, no. 5, pp. 304–310, Aug. 1989, doi: 10.1016/0002-9149(89)90524-9.

[13] T. Chen and C. Guestrin, “XGBoost: A scalable tree boosting system,” in Proc. 22nd ACM SIGKDD Int. Conf. Knowl. Discov. Data Min. (KDD), San Francisco, CA, USA, 2016, pp. 785–794, doi: 10.1145/2939672.2939785.

[14] F. Pedregosa et al., “Scikit-learn: Machine learning in Python,” J. Mach. Learn. Res., vol. 12, pp. 2825–2830, 2011.

[15] S. M. Lundberg and S.-I. Lee, “A unified approach to interpreting model predictions,” in Adv. Neural Inf. Process. Syst., vol. 30, Long Beach, CA, USA, 2017, pp. 4765–4774.

[16] S. Hur *et al.*, "Comparison of SHAP and clinician friendly explanations reveals effects on clinical decision behaviour," *npj Digit. Med.*, vol. 8, Art. no. 578, 2025, doi: 10.1038/s41746-025-01958-8.

[17] A. J. Vickers and E. B. Elkin, "Decision curve analysis: A novel method for evaluating prediction models", *Med. Decis. Making*, vol. 26, no. 6, pp. 565–574, Nov. 2006, doi: 10.1177/0272989X06295361.

[18] L. Zhao *et al.*, "Understanding decision curve analysis in clinical prediction model research", *Postgrad. Med. J.*, vol. 100, no. 1185, pp. 512–515, Jun. 2024, doi: 10.1093/postmj/qgae027.

[19] A. J. Vickers, B. van Calster, and E. W. Steyerberg, “Net benefit approaches to the evaluation of prediction models, molecular markers, and diagnostic tests,” BMJ, vol. 352, Art. no. i6, Jan. 2016, doi: 10.1136/bmj.i6.

[20] K. G. M. Moons *et al.*, "PROBAST+AI: An updated quality, risk of bias, and applicability assessment tool for prediction models using regression or artificial intelligence methods", *BMJ*, vol. 388, Art. no. e082505, Mar. 2025, doi: 10.1136/bmj-2024-082505.

[21] S. V. Chinta *et al.*, "AI-driven healthcare: A review on ensuring fairness and mitigating bias", *PLOS Digit. Health*, vol. 4, no. 5, Art. no. e0000864, May 2025, doi: 10.1371/journal.pdig.0000864

[22] M. Hoche, O. Mineeva, G. Rätsch, E. Vayena, and A. Blasimme, "What makes clinical machine learning fair? A practical ethics framework", *PLOS Digit. Health*, vol. 4, no. 3, Art. no. e0000728, Mar. 2025, doi: 10.1371/journal.pdig.0000728.

[23] R. D. Riley, J. Ensor, K. I. E. Snell, F. E. Harrell, G. P. Martin, J. B. Reitsma, K. G. M. Moons, G. Collins, and M. van Smeden, “Calculating the sample size required for developing a clinical prediction model,” BMJ, vol. 368, Art. no. m441, Mar. 2020, doi: 10.1136/bmj.m441.