Ósmosis Revista Médica Estudiantil

ISSN 2992 - 8486

Editado por:
Diego Navarro Espinosa
Instituto de Nutrigenética y Nutrigenómica Traslacional, Departamento de Biología Molecular, Universidad de Guadalajara, México.

***Correspondencia**
Carlos Domínguez-Vargas
Correo:
Carlos.dominguez2546@alumnos.udg.mx

Recibido: 25 de octubre, 2025.
Aceptado: 27 de octubre, 2025.
Publicado: 25 de mayo, 2026.

Cómo citar este artículo:
Domínguez-Vargas C, Chávez-Ramírez A, Domínguez-Vargas K, Ramírez-Flores SJ, Pérez-Castellón MC, Álvarez-Dávalos CM, Márquez-González DJ, Robles-Loera DM, Saldaña-Cruz AM, Martínez-Hernández RM. Aplicación de modelos estadísticos y de machine learning para estimar riesgo de enfermedad cardiovascular. Universidad de Guadalajara, México. Ósmosis Revista Médica Estudiantil. 2026;(5):páginas 44-52.

Aplicación de modelos estadísticos y de machine learning para estimar riesgo de enfermedad cardiovascular

Traditional Risk Factors and Machine Learning Approaches for Cardiovascular Disease Prediction

Carlos Domínguez-Vargas[1], Alejandro Chávez-Ramírez[2], Karen Domínguez-Vargas[3], Samantha Jonnue Ramírez-Flores[1], Miranda Citlali Pérez-Castellón[1], Carina Mariel Álvarez-Dávalos[1], Dante Joel Márquez-González[1], Diana Margarita Robles-Loera[1], Ana Miriam Saldaña-Cruz[4], Ramsés Emiliano Martínez-Hernández[5].

Licenciatura en Médico Cirujano y Partero, Universidad de Guadalajara, Centro Universitario de Ciencias de la Salud (CUCS). [1]
Ingeniería en Sistemas Computacionales, Instituto Tecnológico de Morelia, Tecnológico Nacional de México (TecNM). [2]
Ingeniería informática, Universidad de Guadalajara, Centro Universitario de Ciencias Exactas e Ingeniería (CUCEI). [3]
Departamento de Fisiología, Universidad de Guadalajara, Centro Universitario de Ciencias de la Salud (CUCS). [4]
Licenciatura en Médico Cirujano y Partero, Universidad Michoacana de San Nicolás de Hidalgo, Facultad de Ciencias Médicas y Biológicas "Dr. Ignacio Chávez". [5]

Resumen

Las enfermedades cardiovasculares continúan siendo la principal causa de mortalidad en el mundo, y aunque existen escalas clásicas de predicción de riesgo, como Framingham, su utilidad puede variar según la población estudiada. En este estudio se evaluó el desempeño de dos modelos predictivos basados en aprendizaje automático —regresión logística y Random Forest— utilizando una base de datos abierta de Kaggle con 68,731 registros depurados. Se generaron variables derivadas, como edad e índice de masa corporal, y los modelos fueron entrenados y evaluados mediante métricas como AUC, exactitud, sensibilidad, especificidad y F1 score. La regresión logística presentó mejor rendimiento global (AUC de 0.798), mientras que Random Forest mostró mayor sensibilidad (AUC de 0.780). Los principales factores asociados al riesgo cardiovascular fueron la presión arterial sistólica, la edad, el colesterol y la adiposidad, mientras que la actividad física tuvo un efecto protector leve. Estos resultados sugieren que modelos accesibles y reproducibles pueden ofrecer alternativas útiles para la estratificación del riesgo cardiovascular.

Palabras clave: Enfermedades cardiovasculares; Factores de riesgo; Aprendizaje automático; Regresión logística; Modelos estadísticos; Predicción

Introducción

Las enfermedades cardiovasculares (ECV) continúan siendo la principal causa de morbilidad y mortalidad a nivel global, representando aproximadamente 17.9 millones de muertes cada año, según datos de la Organización Mundial de la Salud [1]. Diversos factores de riesgo tradicionales como la hipertensión arterial, dislipidemia, obesidad y sedentarismo han sido identificados como determinantes clave en el desarrollo de estas patologías [2,3]. Históricamente, modelos clínicos como el score de Framingham han sido utilizados para predecir el riesgo cardiovascular [4]; sin embargo, su aplicabilidad se ha visto limitada por el tipo de población para la que fueron diseñados y por la necesidad de variables clínicas no siempre disponibles en la práctica cotidiana [4,5].

En años recientes, el uso de métodos de aprendizaje automático (machine learning, ML) para la predicción del riesgo cardiovascular ha aumentado de forma considerable, particularmente en modelos entrenados con variables clínicas rutinarias y bases de datos de gran escala. Estudios previos han demostrado que algoritmos como Random Forest, redes neuronales y otros enfoques no lineales pueden alcanzar un desempeño comparable, e incluso superior en algunos escenarios, a los modelos tradicionales [6,7]. No obstante, la ganancia predictiva de estos métodos suele depender del conjunto de

Abstract

La válvula aórtica cuadricúspide es una anomalía congénita extremadamente rara caracterizada por la presencia de cuatro valvas aórticas en lugar de las tres habituales. Su diagnóstico suele realizarse de forma incidental durante estudios de imagen solicitados por otras causas. Se presenta el caso de un hombre de 48 años con antecedentes de hipertensión arterial sistémica, diabetes mellitus tipo 2 e insuficiencia cardiaca con fracción de eyección severamente reducida (15%), hospitalizado por evisceración posoperatoria tras cirugía por diverticulitis complicada. Durante su estancia desarrolló insuficiencia cardiaca aguda descompensada. El ecocardiograma transtorácico identificó incidentalmente una válvula aórtica cuadricúspide con insuficiencia aórtica moderada, dilatación ventricular izquierda severa e insuficiencia mitral y tricuspídea significativas. Además, presentó lesión renal aguda y alteraciones electrolíticas. El manejo incluyó diuréticos intravenosos, inicio de inhibidor de la enzima convertidora de angiotensina y un inhibidor SGLT2, así como corrección hidroelectrolítica, resaltando la importancia del diagnóstico ecocardiográfico integral y del abordaje multidisciplinario.

Keywords: Cardiovascular Diseases; Risk Factors; Machine Learning; Logistic Models; Statistical Models; Predictive Value of Tests

variables disponibles, de la estrategia de validación empleada y, de forma crítica, de la calibración del modelo para estimar riesgos clínicamente interpretables. Asimismo, la heterogeneidad poblacional y la ausencia de validación externa limitan con frecuencia la generalización y la aplicabilidad clínica real de estos enfoques.

El acceso a bases de datos abiertas y de libre uso permite evaluar y validar los factores de riesgo clásicos en poblaciones más amplias, así como aplicar metodologías modernas de análisis predictivo bajo principios de reproducibilidad y transparencia. Más allá de la novedad algorítmica, este enfoque resulta especialmente relevante para estimar el rendimiento alcanzable con variables básicas ampliamente disponibles, discutir el equilibrio entre interpretabilidad y desempeño, y explorar la utilidad clínica de métricas operativas como el ajuste del umbral de decisión. Esta combinación de enfoques epidemiológicos y computacionales puede facilitar el desarrollo de herramientas de estratificación de riesgo adaptables a diversos contextos, especialmente en regiones con recursos limitados.

El objetivo de este estudio transversal fue identificar los factores de riesgo más relevantes para la predicción de ECV en un dataset de acceso abierto, mediante la comparación del desempeño de un modelo estadístico clásico (regresión logística) y un algoritmo de aprendizaje automático (Random Forest). La hipótesis de trabajo plantea que los factores tradicionales continúan siendo altamente relevantes en modelos predictivos actuales y que el uso de métodos de ML puede mejorar el rendimiento discriminativo de forma marginal, sin sacrificar la interpretabilidad clínica.

Métodos

Diseño del estudio y fuente de datos

Se llevó a cabo un estudio transversal de análisis secundario basado en el Cardiovascular Disease Dataset, un conjunto de datos de acceso abierto disponible públicamente en la plataforma Kaggle [8]. El dataset comprende información de más de 70,000 registros de adultos mayores de 18 años, recopilados a partir de mediciones clínicas básicas, características demográficas y variables de estilo de vida. Al tratarse de un análisis de datos anonimizados y de libre acceso, el estudio no requirió aprobación por un comité de ética ni consentimiento informado adicional.

Participantes y criterios de inclusión/exclusión

Se incluyeron todos los registros con datos completos y valores plausibles desde el punto de vista clínico. Se excluyeron aquellos con valores

extremos o erróneos, definidos a priori con base en rangos fisiológicos aceptados, tales como tallas menores a 120 cm, presiones arteriales sistólicas mayores a 250 mmHg o presiones diastólicas menores a 30 mmHg. Tras el proceso de limpieza y control de calidad de los datos, la muestra final consistió en 68,731 individuos. El dataset no presentó valores faltantes, por lo que no fue necesario realizar imputación de datos.

Variables y medidas

La variable de desenlace fue la presencia de enfermedad cardiovascular (cardio: 1 = con ECV, 0 = sin ECV). Las variables predictoras incluidas correspondieron a todas las variables clínicas y demográficas disponibles en el dataset con relevancia epidemiológica documentada para el riesgo cardiovascular, sin aplicar métodos automáticos de selección de variables, con el objetivo de preservar la interpretabilidad clínica de los modelos.

Las variables predictoras se agruparon en:

- ◊ Demográficas: edad (transformada a años), sexo.
- ◊ Antropométricas: talla, peso, índice de masa corporal (IMC).
- ◊ Clínicas: presión arterial sistólica (ap_hi), presión arterial diastólica (ap_lo).
- ◊ Bioquímicas ordinales: colesterol y glucosa (clasificadas como normal, por arriba de lo normal y muy por arriba de lo normal).
- ◊ Estilo de vida: tabaquismo, consumo de alcohol y actividad física.

La edad fue transformada de días a años y el índice de masa corporal se calculó a partir del peso y la talla debido a su mayor interpretabilidad clínica y uso estandarizado en la evaluación del riesgo cardiovascular. Las transformaciones se realizaron mediante las siguientes fórmulas: age_years = age / 365.25 y bmi = weight / (height/100)^2

Las variables continuas fueron estandarizadas cuando fue necesario para el entrenamiento de los modelos, particularmente en la regresión logística, con el fin de asegurar comparabilidad entre coeficientes y estabilidad numérica. Todas las transformaciones y procedimientos de estandarización se ajustaron exclusivamente utilizando los datos del conjunto de entrenamiento y posteriormente se aplicaron al conjunto de prueba, con el objetivo de evitar data leakage.

Manejo del sesgo y validación

Para reducir sesgos derivados de errores de captura, se realizaron controles de plausibilidad clínica previos al análisis. En aquellos registros donde la presión arterial sistólica fue menor que la diastólica (ap_hi < ap_lo), los valores fueron intercambiados, y se eliminaron observaciones con valores fuera de rangos fisiológicos aceptados [9]. El dataset no presentó valores faltantes, por lo que no fue necesario realizar imputación de datos.

Con el fin de prevenir data leakage, todas las transformaciones de variables (incluida la estandarización) se implementaron mediante pipelines ajustados exclusivamente con los datos del conjunto de entrenamiento y posteriormente aplicados al conjunto de prueba [10]. La posible colinealidad entre variables predictoras fue evaluada en el modelo de regresión logística mediante el cálculo del factor de inflación de la varianza (VIF), considerándose valores elevados como indicativos de colinealidad potencial.

Dado que la prevalencia de la variable de desenlace fue cercana al 50%, el problema de clasificación no presentó un desbalance severo de clases, por lo que no se aplicaron técnicas adicionales de reponderación o remuestreo.

Tamaño muestral y partición

No se realizó un cálculo prospectivo del tamaño de muestra debido al carácter secundario del análisis y al uso de un dataset preexistente. El conjunto total de datos fue dividido en un 80% para entrenamiento y un 20% para prueba, utilizando una partición estratificada según la variable de desenlace (cardio), con el fin de conservar la proporción de casos y controles en ambos subconjuntos [11]. La partición se realizó de forma previa a cualquier procedimiento de ajuste o transformación de los datos.

Análisis estadístico

Se entrenaron dos modelos de predicción con enfoques complementarios:

- Regresión Logística (LR): modelo lineal con penalización L2, ampliamente utilizado en epidemiología clínica por su interpretabilidad, entrenado tras la estandarización de las variables continuas [12].

- Random Forest (RF): algoritmo de ensamblado basado en árboles de decisión, capaz de capturar relaciones no lineales y efectos de interacción, entrenado con 400 árboles. El número de árboles fue seleccionado por ofrecer un rendimiento estable sin evidenciar sobreajuste, priorizando la comparabilidad entre modelos sobre la optimización exhaustiva de hiperparámetros. La importancia de las variables se estimó mediante el criterio de Gini [13].

La elección de estos dos modelos respondió al objetivo del estudio de contrastar un enfoque estadístico clásico, altamente interpretable, frente a un método de aprendizaje automático no lineal ampliamente utilizado, sin introducir complejidad algorítmica adicional que limitara la reproducibilidad y la aplicabilidad clínica.

El desempeño de los modelos se evaluó en el conjunto de prueba mediante las métricas: área bajo la curva ROC (AUC), exactitud, sensibilidad, especificidad, precisión (valor predictivo positivo) y F1 score [14–16]. Asimismo, se construyeron curvas de precisión–recall y se estimó el área promedio bajo dicha curva (Average Precision). La calibración de los modelos se evaluó mediante curvas de calibración y el cálculo del Brier score. Para optimizar la toma de decisiones clínicas, el umbral de clasificación se ajustó utilizando el índice de Youden (sensibilidad + especificidad − 1), y se construyeron matrices de confusión [17].

Todos los análisis se realizaron utilizando Python, con las bibliotecas scikit-learn y pandas.

Evaluación de calibración

La calibración de los modelos predictivos se evaluó mediante curvas de calibración, comparando las probabilidades predichas con las frecuencias observadas en el conjunto de prueba. Adicionalmente, se calculó el Brier score como medida global del error cuadrático medio de las probabilidades predichas, con valores menores indicando mejor calibración del modelo.

No se realizó validación externa ni una optimización exhaustiva de hiperparámetros, dado que el objetivo del estudio fue la comparación metodológica y clínica entre enfoques predictivos basados en variables clínicas básicas, priorizando la reproducibilidad, la interpretabilidad y la aplicabilidad en contextos con recursos limitados.

Declaración sobre el uso de inteligencia artificial

Este manuscrito fue estructurado con el apoyo de una herramienta de inteligencia OpenAI, que fue utilizada para complementar análisis exploratorios de datos y visualizaciones intermedias mediante su función integrada de análisis avanzado de datos Advanced Data Analysis. Todas las decisiones metodológicas, selección de modelos, validación estadística e interpretación de resultados fueron realizadas y verificadas exclusivamente por los autores.

Resultados

Participantes

Tras aplicar los criterios de limpieza y plausibilidad clínica, la muestra final consistió en 68,731 participantes adultos (≥18 años). La prevalencia global de enfermedad cardiovascular (ECV) fue de 49.5%, proporción que se mantuvo en el conjunto de prueba utilizado para la evaluación de los modelos, lo que permitió un adecuado balance de clases para el análisis predictivo [8].

Datos descriptivos y características basales

Las características demográficas, clínicas y antropométricas de la población de estudio se presentan en la Tabla 1, estratificadas según la presencia o ausencia de enfermedad cardiovascular. De manera descriptiva, los participantes con ECV tendieron a ser de mayor edad y presentaron valores más elevados de presión arterial sistólica y diastólica, así como un índice de masa corporal ligeramente superior en comparación con aquellos sin ECV. Asimismo, se observó una mayor proporción de niveles elevados de colesterol en el grupo con ECV.

Otras variables, como talla, peso, glucosa y nivel de actividad física, mostraron diferencias más modestas entre los grupos. Estos hallazgos descriptivos son

Tabla 1. Modelos predictivos de enfermedad cardiovascular

	Umbral Estándar 0.50		Umbral Optimizado (Índice de Youden)	
	Regresión Logística (RL)	Random Forest (RF)	Regresión Logística (RL) Umbral óptimo ≈ 0.464	Random Forest (RF) Umbral óptimo ≈ 0.481
AUC	0.798	0.780		
Exactitud	73.54%	71.96%	73.74%	72.01%
Sensibilidad	67.94%	71.59%	72.64%	73.09%
Especificidad	79.02%	72.34%	74.81%	70.95%
Precisión	0.760	0.717	0.739	0.711
F1 score	0.718	0.716	0.732	0.721

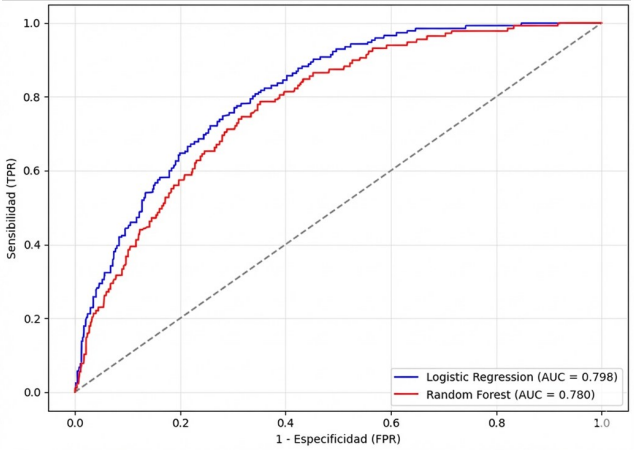
Los valores de exactitud, sensibilidad y especificidad se expresan en porcentaje (%). AUC, precisión y F1 score se presentan como valores entre 0 y 1. AUC: área bajo la curva ROC; RL: regresión logística; RF: Random Forest.

consistentes con patrones reportados previamente en estudios poblacionales sobre riesgo cardiovascular [3,18].

Ambos modelos mostraron un desempeño discriminativo adecuado, con valores de AUC cercanos a 0.80. La regresión logística presentó un AUC ligeramente superior al del modelo Random Forest (0.798 vs. 0.780). Con un umbral estándar de 0.50, la regresión logística mostró mayor exactitud y especificidad, mientras que el modelo Random Forest presentó mayor sensibilidad, lo que indica una mayor capacidad para identificar individuos con enfermedad cardiovascular a costa de un mayor número de falsos positivos.

Tras la optimización del umbral mediante el índice de Youden, ambos modelos mejoraron su equilibrio

Figura 1. Curva ROC para predicción de ECV



Curva ROC para la predicción de enfermedad cardiovascular mediante modelos de regresión logística y random forest. El área bajo la curva (AUC) fue de 0.798 para regresión logística y 0.780 para random forest, indicando buen desempeño discriminativo. El análisis fue realizado en Python (versión 3.10) utilizando las bibliotecas scikit-learn y matplotlib en la plataforma Google Colab.

entre sensibilidad y especificidad, siendo la regresión logística el modelo con mejor desempeño global en términos de exactitud, precisión y F1 score.

Para ambos modelos se generaron matrices de confusión a partir del conjunto de prueba, lo que permitió analizar la distribución de las clasificaciones verdaderas y erróneas. En la regresión logística (LR) con umbral estándar de 0.50, se obtuvieron 4,622 verdaderos positivos (TP), 1,457 falsos positivos (FP), 5,487 verdaderos negativos (TN) y 2,181 falsos negativos (FN). Esto se traduce en una exactitud global del 73.5% y un equilibrio adecuado entre sensibilidad (67.9%) y especificidad (79.0%) [14].

En el caso del modelo Random Forest (RF), la matriz de confusión mostró 4,870 TP, 1,921 FP, 5,023 TN y 1,933 FN, con una exactitud global del 71.9%. Este modelo evidenció una mayor sensibilidad (71.6%) a expensas de una ligera reducción en la especificidad (72.3%), lo cual puede resultar útil en escenarios donde la prioridad clínica sea la detección de casos positivos [13,14]

Tras la optimización del umbral de decisión mediante el índice de Youden, ambos modelos mejoraron su equilibrio entre sensibilidad y especificidad, con una reducción en el número de falsos negativos y un impacto limitado sobre los falsos positivos. Este ajuste fue particularmente relevante para la regresión logística, que alcanzó un mejor balance operativo sin comprometer de forma sustancial su especificidad, incrementando así su utilidad clínica [17].

Curvas de evaluación

Se graficaron las curvas ROC y Precision–Recall (PR) para ambos modelos (Figura 1) [14–16]. En el análisis ROC, tanto la regresión logística como el modelo Random Forest mostraron un buen desempeño discriminativo, con valores de área bajo la curva (AUC) de 0.798 y 0.780, respectivamente, ambos claramente superiores a la línea de referencia. No se estimaron intervalos de confianza para el AUC, lo cual se reconoce como una limitación del análisis.

En la curva Precision–Recall, ambos modelos superaron la línea base correspondiente a la prevalencia de la enfermedad cardiovascular en la muestra (~0.50). La regresión logística mostró un Average Precision ligeramente superior al del modelo Random Forest (0.773 vs. 0.759; Tabla 2), lo cual resulta particularmente relevante en contextos

Tabla 2. Curva Precision–Recall (PR): Average Precision

Regresión Logística	0.773
Random Forest	0.759

clínicos con recursos limitados, donde la precisión diagnóstica es un criterio clave para la toma de decisiones [15].

Ambos modelos superaron claramente la línea base de prevalencia (~0.495), siendo la regresión logística ligeramente superior también en la curva PR, lo cual es relevante en contextos clínicos con recursos limitados, donde la precisión diagnóstica es crucial

[15].

Análisis de variables predictoras

La presión arterial sistólica fue el predictor más fuertemente asociado con la presencia de enfermedad cardiovascular, con un OR de 2.56 por incremento de una desviación estándar (IC 95%: 2.48–2.64; p = 2.1×10^{-298}). La edad y los niveles de colesterol también mostraron asociaciones positivas significativas (p < 10^{-150}).

El peso corporal y la presión arterial diastólica presentaron asociaciones de menor magnitud, aunque estadísticamente robustas. Por el contrario, la actividad física se asoció de forma inversa con la presencia de enfermedad cardiovascular (OR: 0.80; IC 95%: 0.76–0.83; p = 3.1×10^{-43}), lo que sugiere un efecto protector consistente con la literatura previa.

Tabla 3. Regresión Logística

Variable	OR	IC 95%	Valor de p
PA sistólica	2.56	2.48 – 2.64	2.1×10^{-298}
Edad	1.41	1.39 – 1.44	4.6×10^{-176}
Colesterol	1.40	1.60 – 1.70	1.3×10^{-221}
Peso	1.20	1.15 – 1.19	7.9×10^{-65}
PA diastólica	1.11	1.08 – 1.14	2.4×10^{-27}
Actividad física	0.80	0.76 – 0.83	3.1×10^{-43}

En la regresión logística, la presión arterial sistólica fue el predictor más fuertemente asociado con la presencia de enfermedad cardiovascular, con un odds ratio (OR) de 2.56 por incremento de una desviación estándar (IC 95%: 2.48–2.64; p = 2.1×10^{-298}). La edad (OR: 1.41; IC 95%: 1.39–1.44) y los niveles de colesterol (OR: 1.65; IC 95%: 1.60–1.70) también mostraron asociaciones positivas significativas. Por el contrario, la actividad física se asoció de forma inversa con la presencia de enfermedad cardiovascular (OR: 0.80; IC 95%: 0.76–0.83; p = 3.1×10^{-43}), consistente con su papel protector reportado en la literatura previa [3].

Tabla 4. Importancia relativa de las variables en el modelo Random Forest (criterio Gini)

Edad	0.263
PA sistólica	0.164
IMC	0.161
Peso	0.117
PA diastólica	0.085
Colesterol	0.037
Glucosa	0.017

En el modelo Random Forest, la contribución relativa de las variables predictoras se evaluó mediante la importancia basada en el criterio de Gini. La edad fue la variable con mayor peso relativo (Gini: 0.263), seguida de la presión arterial sistólica, el índice de masa corporal y el peso corporal.

A pesar de las diferencias metodológicas entre ambos enfoques, la coincidencia en las variables de mayor relevancia predictiva sugiere una consistencia interna de los modelos en la identificación de factores clave asociados con la enfermedad cardiovascular. Las variables relacionadas con la presión arterial y la adiposidad dominaron la importancia predictiva, en concordancia con la evidencia epidemiológica y con validaciones clínicas previas [2,3,7].

Estas medidas de importancia reflejan la contribución relativa de cada variable a la capacidad predictiva del modelo, pero no implican relaciones causales ni sustituyen la interpretación clínica derivada de modelos inferenciales.

Discusión

Este estudio evaluó y comparó el desempeño de dos enfoques predictivos —regresión logística (LR) y Random Forest (RF)— en la predicción de enfermedad cardiovascular (ECV) utilizando un conjunto de datos de acceso abierto con más de 68,000 registros. Ambos modelos mostraron un buen desempeño discriminativo, con valores de área bajo la curva ROC (AUC) de 0.798 para la LR y 0.780 para el RF, resultados que se sitúan dentro del rango reportado por scores clínicos clásicos como Framingham o QRISK cuando se aplican a poblaciones distintas de aquellas en las que fueron

desarrollados [1,2].

La regresión logística presentó un mejor desempeño global en términos de AUC, exactitud y especificidad, mientras que el modelo Random Forest mostró una mayor sensibilidad, lo que refleja una mayor capacidad para identificar individuos con ECV, a costa de un incremento en el número de falsos positivos. El ajuste del umbral de decisión mediante el índice de Youden permitió mejorar el equilibrio operativo entre sensibilidad y especificidad en ambos modelos, siendo particularmente relevante para la LR, que alcanzó un mejor balance clínico sin comprometer de forma sustancial su especificidad [3].

Comparación con la literatura existente

Los factores de riesgo identificados como más relevantes —presión arterial sistólica, edad, niveles de colesterol y medidas de adiposidad— son consistentes con la evidencia epidemiológica acumulada y con los lineamientos de prevención cardiovascular emitidos por entidades como la American Heart Association y la European Society of Cardiology [4–6]. Asimismo, la actividad física mostró una asociación inversa significativa con la presencia de ECV, en concordancia con estudios poblacionales previos que documentan su papel protector [7,8].

Desde el punto de vista metodológico, estos resultados respaldan observaciones previas que señalan que los modelos de aprendizaje automático, como Random Forest, pueden capturar relaciones no lineales y complejas entre variables, aunque no necesariamente superan a los modelos estadísticos clásicos cuando se dispone de un conjunto de variables clínicas básicas, bien definidas y con fuerte respaldo fisiopatológico [9–11].

Limitaciones del estudio

Este estudio presenta varias limitaciones. En primer lugar, el dataset utilizado proviene de una fuente única y carece de información detallada sobre su contexto geográfico y temporal, lo que limita la generalización externa de los hallazgos [12]. En segundo lugar, las variables de estilo de vida, como tabaquismo, consumo de alcohol y actividad física, fueron autorreportadas, lo que introduce un riesgo de sesgo de medición [13,18]. Además, no se incluyeron biomarcadores avanzados, como fracciones lipídicas específicas o marcadores inflamatorios, que podrían mejorar la discriminación y calibración de los modelos [14].

Si bien se incorporó una evaluación de calibración mediante curvas de calibración y Brier score, la ausencia de validación externa impide extrapolar directamente estos resultados a otras poblaciones, por lo que su aplicación clínica debe interpretarse con cautela [15].

Validez externa y aplicabilidad clínica

A pesar de estas limitaciones, los hallazgos sugieren que es posible desarrollar modelos predictivos confiables utilizando datos abiertos y variables clínicas ampliamente disponibles, lo cual resulta particularmente relevante en contextos con recursos limitados. La regresión logística, al combinar un desempeño adecuado con alta interpretabilidad, continúa siendo una herramienta competitiva y clínicamente útil para la estratificación del riesgo cardiovascular [3,10].

Por su parte, el uso de modelos como Random Forest puede resultar ventajoso en escenarios donde se priorice la sensibilidad diagnóstica y se disponga de recursos computacionales suficientes, siempre que sus resultados se interpreten como herramientas complementarias y no sustitutas del juicio clínico [11,16].

Conclusión

Este estudio muestra que es posible desarrollar modelos predictivos confiables para la estimación del riesgo de enfermedad cardiovascular (ECV) utilizando datos de acceso abierto y métodos estadísticos y computacionales accesibles. Tanto la regresión logística como el algoritmo Random Forest alcanzaron un desempeño discriminativo adecuado, con valores de área bajo la curva ROC (AUC) de 0.798 y 0.780, respectivamente.

La regresión logística se consolidó como el modelo con mejor desempeño global y mayor interpretabilidad clínica, mientras que el modelo Random Forest presentó una sensibilidad superior, característica que puede resultar útil en escenarios donde se prioriza la detección de casos, aun a costa de un mayor número de falsos positivos. Este contraste resalta la importancia de considerar no solo el rendimiento global, sino también el contexto clínico de aplicación al seleccionar un modelo predictivo.

Los principales factores asociados con la presencia

de ECV —presión arterial sistólica, edad, niveles de colesterol y medidas de adiposidad— reafirmaron su relevancia en la predicción del riesgo cardiovascular. Asimismo, la actividad física mostró una asociación inversa significativa con la ECV, en concordancia con la evidencia epidemiológica disponible.

El presente trabajo aporta evidencia de que, aun en ausencia de biomarcadores avanzados, es posible desarrollar herramientas de estratificación de riesgo cardiovascular con rendimiento comparable al de scores clínicos clásicos. Además, ofrece una base sólida para el diseño futuro de scores clínicos simplificados y para la exploración de modelos más avanzados de aprendizaje automático, como XGBoost o redes neuronales profundas. En conjunto, estos resultados apoyan la integración de enfoques tradicionales y modernos en la construcción de soluciones predictivas orientadas a mejorar la prevención y detección temprana de enfermedades cardiovasculares, especialmente en contextos de recursos limitados.

No obstante, dado que los modelos no fueron validados en cohortes externas, sus resultados deben interpretarse con cautela al extrapolarlos a otros contextos poblacionales, y se requiere validación adicional antes de su implementación clínica.

Agradecimientos

Este trabajo fue posible gracias al acceso a la base de datos pública Cardiovascular Disease Dataset disponible en Kaggle, por parte de Svetlana Ulianova.

Conflicto de intereses

Los autores declaran no tener conflictos de interés.

Financiamiento

No se recibió financiamiento específico para este trabajo.

Bibliografía

1. World Health Organization. Cardiovascular diseases (CVDs). [Internet]. 2024 [cited 2025 Oct 24]. Available from: https://www.who.int/news-room/fact-sheets/detail/cardiovascular-diseases-(cvds)
2. Yusuf S, Hawken S, Ôunpuu S, Dans T, Avezum A, Lanas F, et al. Effect of potentially modifiable risk factors associated with myocardial infarction in 52 countries (the INTERHEART study): case-control study. Lancet. 2004;364(9438):937–952.
3. Arnett DK, Blumenthal RS, Albert MA, et al. 2019 ACC/AHA guideline on the primary prevention of cardiovascular disease. J Am Coll Cardiol. 2019;74(10):e177–e232.
4. D'Agostino RB Sr, Vasan RS, Pencina MJ, et al. General cardiovascular risk profile for use in primary care: the Framingham Heart Study. Circulation. 2008;117(6):743–753.
5. Karmali KN, Lloyd-Jones DM. Adding a new dimension to cardiovascular risk prediction: moving from scores to curves. JAMA. 2014;311(21):2176–2177.
6. Weng SF, Reps J, Kai J, Garibaldi JM, Qureshi N. Can machine-learning improve cardiovascular risk prediction using routine clinical data? PLoS One. 2017;12(4):e0174944.
7. Goldstein BA, Navar AM, Carter RE. Moving beyond regression techniques in cardiovascular risk prediction: applying machine learning to address analytic challenges. Eur Heart J. 2017;38(23):1805–1814.
8. Cardiovascular Disease Dataset. Kaggle [Internet]. 2018 [cited 2025 Oct 24]. Available from: https://www.kaggle.com/datasets/sulianova/cardiovascular-disease-dataset
9. Steyerberg EW. Clinical prediction models: a practical approach to development, validation, and updating. Springer; 2019.
10. Varma S, Simon R. Bias in error estimation when using cross-validation for model selection. BMC Bioinformatics. 2006;7:91.
11. Kuhn M, Johnson K. Applied predictive modeling. Springer; 2013.
12. Hosmer DW, Lemeshow S, Sturdivant RX. Applied logistic regression. 3rd ed. Wiley; 2013.
13. Breiman L. Random Forests. Mach Learn. 2001;45:5–32.
14. Altman DG, Bland JM. Diagnostic tests 3: receiver operating characteristic plots. BMJ.

1994;309(6948):188.

15. Powers DMW. Evaluation: from precision, recall and F-measure to ROC, informedness, markedness & correlation. J Mach Learn Tech. 2011;2(1):37–63.
16. Chicco D, Jurman G. The advantages of the Matthews correlation coefficient (MCC) over F1 score and accuracy in binary classification evaluation. BMC Genomics. 2020;21(1):6.
17. Youden WJ. Index for rating diagnostic tests. Cancer. 1950;3(1):32–35.
18. Rosenman R, Tennekoon V, Hill LG. Measuring bias in self-reported data. Int J Behav Healthc Res. 2011 Oct;2(4):320-332. doi: 10.1504/IJBHR.2011.043414. PMID: 25383095; PMCID: PMC4224297.