Ciencia y Educación
(L-ISSN: 2790-8402 E-ISSN: 2707-3378)
Vol. 7 No. 1.1
Edición Especial I 2026

PREDICCIÓN DEL RIESGO CARDIOVASCULAR EN POBLACIÓN JOVEN MEDIANTE APRENDIZAJE AUTOMÁTICO Y DATOS LONGITUDINALES
CARDIOVASCULAR RISK PREDICTION IN YOUNG POPULATIONS THROUGH MACHINE LEARNING AND LONGITUDINAL DATA

Autores: [1]Milton Daniel Chicaiza Criollo, [2]José Renato Cumbal Simba.
[1]ORCID ID: https://orcid.org/0009-0009-8478-0625
[2]ORCID ID: https://orcid.org/0000-0001-8182-5343
[1]E-mail de contacto: dchicaizac@est.ups.edu.ec
[2]E-mail de contacto: rcumbal@ups.edu.ec
Afiliación: [1*2*]Universidad Politécnica Salesiana, (Ecuador).
Artículo recibido: 15 de Enero del 2026
Artículo revisado: 28 de Enero del 2026
Artículo aprobado: 06 de Febrero del 2026
[1]Estudiante de Ingeniería Biomédica egresado de la Universidad Politécnica Salesiana, (Ecuador).
[2]Ingeniero en Electrónica y Telecomunicaciones. Magíster en Gerencia de Sistemas de Información egresado de la Universidad Politécnica Salesiana, (Ecuador). Profesor de la Universidad Politécnica Salesiana, Quito, Ecuador, y miembro del Grupo de Investigación en Telecomunicaciones (GIETEC). Doctorante en Ingeniería, Universidad Pontificia Bolivariana, Medellín, (Colombia).

Resumen

El proyecto se centra en la predicción temprana del riesgo cardiovascular incidente en población joven mediante la evaluación y comparación de tres algoritmos de aprendizaje automático: Regresión Logística, Random Forest y Redes Neuronales. Las Enfermedades Cardiovasculares (ECV) constituyen la principal causa de mortalidad global, y marcadores clínicos tempranos como la hipertensión arterial desempeñan un papel determinante en la evolución futura del riesgo. La evidencia científica indica que los fundamentos de esta patología se presentan en la adultez temprana (18-35 años). Los modelos tradicionales de estratificación de riesgo han mostrado desempeño limitado en población joven. Para abordar esta limitación, se emplea un diseño metodológico basado en datos clínicos longitudinales armonizados de la Encuesta Nacional de Salud y Nutrición del Ecuador (ENSANUT), realizando una comparación sistemática de modelos de aprendizaje automático. El rendimiento predictivo se evalúa mediante métricas como precisión, sensibilidad, F1-score y AUC. Los resultados indican que Random Forest + SMOTE logra el mejor equilibrio entre precisión (PR-AUC: 0.345) y capacidad de detección de casos positivos, mientras que la Regresión Logística + SMOTE alcanza el mayor F1-score (0.404), destacándose por su interpretabilidad clínica. El estudio demuestra que los modelos de aprendizaje automático superan significativamente a los enfoques lineales tradicionales en la identificación temprana del riesgo cardiovascular en adultos jóvenes.

Palabras clave: Riesgo cardiovascular, Población joven, Aprendizaje automático, Datos longitudinales, Hipertensión arterial, Predicción clínica.

Abstract

This research project focuses on early prediction of incident cardiovascular risk in young populations through evaluation and comparison of three machine learning algorithms: Logistic Regression, Random Forest, and Neural Networks. Cardiovascular Disease (CVD) is the leading cause of death worldwide, and early clinical markers such as arterial hypertension play a decisive role in future cardiovascular risk evolution. Scientific evidence indicates that foundations of this pathology emerge in early adulthood (18-35 years). Traditional risk stratification models have shown limited performance in young populations. To address this limitation, a methodological design based on harmonized longitudinal clinical data from Ecuador's National Health and Nutrition Survey (ENSANUT) is employed, performing systematic comparison of machine learning models. Predictive performance is evaluated

using metrics such as accuracy, sensitivity, F1-score, and AUC. Results indicate that Random Forest + SMOTE achieves the best balance between precision (PR-AUC: 0.345) and positive case detection capacity, while Logistic Regression + SMOTE reaches the highest F1-score (0.404), standing out for its clinical interpretability. The study demonstrates that machine learning models significantly outperform traditional linear approaches in early identification of cardiovascular risk in young adults.

Keywords: Cardiovascular risk, Young adults, Machine learning, Longitudinal data, Arterial hypertension, Clinical prediction.

Sumário

O projeto concentra-se na previsão precoce do risco cardiovascular incidente em população jovem mediante avaliação e comparação de três algoritmos de aprendizagem automática: Regressão Logística, Random Forest e Redes Neurais. As Doenças Cardiovasculares (DCV) constituem a principal causa de mortalidade global, e marcadores clínicos precoces como a hipertensão arterial desempenham papel determinante na evolução futura do risco. A evidência científica indica que os fundamentos desta patologia apresentam-se na idade adulta precoce (18-35 anos). Os modelos tradicionais de estratificação de risco mostraram desempenho limitado em população jovem. Para abordar esta limitação, emprega-se um desenho metodológico baseado em dados clínicos longitudinais harmonizados da Pesquisa Nacional de Saúde e Nutrição do Equador (ENSANUT), realizando uma comparação sistemática de modelos de aprendizagem automática. O desempenho preditivo é avaliado mediante métricas como precisão, sensibilidade, F1-score e AUC. Os resultados indicam que Random Forest + SMOTE alcança o melhor equilíbrio entre precisão (PR-AUC: 0.345) e capacidade de detecção de casos positivos, enquanto a Regressão Logística + SMOTE atinge o maior F1-score (0.404), destacando-se pela sua interpretabilidade clínica. O estudo demonstra que os modelos de aprendizagem automática superam significativamente as abordagens lineares tradicionais na identificação precoce do risco cardiovascular em adultos jovens.

Palavras-chave: Risco cardiovascular, População jovem, Aprendizagem automática, Dados longitudinais, Hipertensão arterial, Previsão clínica

Introducción

Las enfermedades cardiovasculares (ECV) representan la principal causa de mortalidad global, generando una demanda creciente de recursos sanitarios y un impacto económico considerable asociado a hospitalizaciones, tratamientos crónicos y pérdida de productividad (WHO, 2023). Comprender las ECV como un proceso acumulativo y progresivo, no como eventos aislados, se vuelve fundamental para la práctica médica moderna y el diseño de estrategias preventivas efectivas (Rajkomar et al., 2019). Investigaciones recientes demuestran que las alteraciones cardiovasculares comienzan en edades tempranas. Factores como el aumento progresivo de la presión arterial, dislipidemia subclínica y aumento del índice de masa corporal aparecen desde la juventud, incluso de manera asintomática (Salah y Srinivas, 2022; Liu et al., 2023). El riesgo cardiovascular se operacionaliza mediante la predicción de hipertensión arterial (HTA) incidente, ya que la presión arterial elevada en edades tempranas es un determinante clave en la progresión futura de enfermedad cardiovascular (Martin et al., 2024). La población joven (18-35 años) constituye un grupo sustancial que determina la evolución del riesgo cardiovascular a largo plazo.

Los modelos tradicionales de estimación de riesgo cardiovascular, como Framingham Risk Score o ASCVD, no han sido diseñados para población joven, mostrando desempeño

limitado. En respuesta, la comunidad científica ha desarrollado modelos de aprendizaje automático (Machine Learning, ML) como herramientas complementarias. Algoritmos como Random Forest, regresión logística y redes neuronales han evidenciado mejor desempeño en el reconocimiento de relaciones no lineales y patrones complejos dentro de datos clínicos de alta dimensionalidad (Salah y Srinivas, 2022; Banerjee y Paçal, 2025). Un avance significativo corresponde a la integración sistemática de datos longitudinales, que facilita el estudio de trayectorias individuales mediante mediciones repetidas de variables como presión arterial, perfil lipídico o glucosa. Esta metodología permite detectar variaciones discretas pero clínicamente significativas que anteceden a eventos cardiovasculares, ofreciendo una representación más fiel del riesgo que las aproximaciones transversales (Nguyen et al., 2021; Liu et al., 2023).

Sin embargo, existen pocos estudios que comparen directamente bajo un mismo diseño metodológico modelos clásicos como regresión logística con técnicas más complejas aplicadas específicamente a población joven (Nguyen et al., 2021). Esta carencia dificulta determinar con precisión qué modelo presenta el mejor equilibrio entre precisión predictiva e interpretabilidad clínica. Al no contar con modelos predictivos adaptados a adultos jóvenes, estos pueden ser clasificados sistemáticamente como población de bajo riesgo, aunque desarrollen trayectorias metabólicas negativas que requieran intervenciones tempranas (Martin et al., 2024). El objetivo de esta investigación es comparar el desempeño de tres modelos en la predicción del riesgo cardiovascular operacionalizado mediante HTA incidente en adultos jóvenes, empleando datos clínicos longitudinales de ENSANUT. La pregunta guía es: ¿qué modelo predictivo ofrece mayor precisión y utilidad clínica para identificar riesgo cardiovascular futuro en población joven considerando la evolución temporal de sus factores de riesgo

Materiales y Métodos

Estudio observacional, analítico y comparativo basado en datos longitudinales, orientado al desarrollo, evaluación y comparación de tres modelos predictivos. Este enfoque longitudinal responde a recomendaciones metodológicas actuales que priorizan mediciones repetidas en el tiempo para captar la evolución real del riesgo cardiovascular con mayor precisión que los cortes transversales (Nguyen et al., 2021; Liu et al., 2023). Los datos provienen de la Encuesta Nacional de Salud y Nutrición del Ecuador (ENSANUT), ciclos 2012, 2018 y 2022, que constituyen las fuentes oficiales más completas disponibles a nivel nacional. Estos ciclos comparten diseño muestral probabilístico, metodologías de medición estandarizadas y protocolos clínicos consistentes. Se realizó un proceso de armonización de variables clínicas unificando definiciones operativas, unidades de medida y criterios de inclusión, conservando únicamente aquellas variables presentes de forma consistente en todos los periodos. Este procedimiento permitió construir una base de datos longitudinal coherente para evaluar la evolución temporal de los factores de riesgo cardiovascular en población joven.

La población comprende individuos entre 18 y 35 años. Se excluyeron registros fuera de este rango y observaciones con inconsistencias clínicas (valores fisiológicamente imposibles, registros incompletos). La variable objetivo se definió a partir de mediciones reales de presión arterial, considerando la HTA como marcador clínico temprano del riesgo cardiovascular (Martin et al., 2024; Nguyen et al., 2021). Las

variables seleccionadas incluyen: edad, presión arterial sistólica (PAS), presión arterial diastólica (PAD), glucosa, colesterol total, HDL, LDL y triglicéridos. La selección se fundamenta en su uso recurrente en modelos predictivos de riesgo cardiovascular y su disponibilidad consistente en ENSANUT. La variable objetivo binaria HTA se definió según criterios estándar: PAS ≥ 140 mmHg o PAD ≥ 90 mmHg, alineándose con prácticas habituales en estudios epidemiológicos (Nguyen et al., 2021).

El preprocesamiento garantizó la calidad y estabilidad de los modelos. Se filtró la población por rango etario (18-35 años) y se seleccionaron las variables clínicas de interés. Los valores faltantes se abordaron mediante imputación por la mediana, estrategia confiable frente a distribuciones asimétricas, aplicada únicamente a variables predictoras. Las variables PAS y PAD no fueron imputadas en la definición de la etiqueta para evitar sesgos. Opcionalmente se aplicó estandarización mediante StandardScaler para modelos sensibles a la escala. Se evaluaron cuatro modelos de clasificación que contrastan enfoques lineales y no lineales: Regresión Logística, Perceptrón simple, Random Forest y Red Neuronal Multicapa (MLP). El vector de características clínicas es x = [edad, pas, pad, glucosa, col_total, hdl, ldl, trigliceridos]T, y la etiqueta binaria y = hta ∈ {0, 1}. La Regresión Logística se implementó como referente lineal por su transparencia clínica, estimando la probabilidad de HTA mediante función sigmoide:

p ˆ("hta"=1|x)=σ(z)=1/(1+exp⁡(-z))

donde z=β_0+ Σ _(i=1)^8 β_i x_i. La decisión final se obtiene por umbral: y ˆ=I[p ˆ≥τ], con τ=0.5.

El Perceptrón simple se incluyó como modelo neuronal base de baja complejidad para contrastar el beneficio de arquitecturas más profundas. Calcula un puntaje afín y aplica función de decisión:

a=w^T x+b,y ˆ=I[a≥0]

Random Forest representa un enfoque no lineal basado en ensambles. Compuesto por T árboles de decisión, cada árbol produce predicción y ˆ ^((t))∈{0,1} y la salida se determina por votación mayoritaria. En modo probabilístico:

p ˆ("hta"=1|x)=1/T Σ _(t=1)^T y ˆ^((t))

La Red Neuronal Multicapa (MLP) se utilizó como modelo de mayor capacidad para aprender interacciones no lineales. Con una capa oculta, la propagación hacia adelante es:

h=g(W_1 x+b_1),p ˆ("hta"=1|x)=σ(W_2 h+b_2)

donde g(·) es una activación no lineal. Para el ajuste se utilizó entropía cruzada binaria:

L=-[ylog⁡(p ˆ)+(1-y)log⁡(1-p ˆ)]

El conjunto de datos se segmentó estratificadamente (80/20): 80% entrenamiento, 20% prueba. El desbalance de clases se abordó exclusivamente sobre el conjunto de entrenamiento mediante SMOTE, evitando contaminación del conjunto de prueba. La imputación de valores faltantes se aplicó únicamente sobre variables predictoras, no sobre la variable objetivo, preservando la validez del marcador de riesgo. El rendimiento se evaluó mediante validación cruzada estratificada con cinco pliegues (k = 5), utilizando métricas complementarias: matriz de confusión, exactitud, F1-score, ROC-AUC y PR-AUC. Se reportó el desempeño en el

conjunto de prueba y el promedio con desviación estándar durante la validación cruzada, dando mayor importancia a F1 y PR-AUC debido al desequilibrio de clases.

Resultados y Discusión

Desempeño sin balanceo de clases

Sin balanceo, el Accuracy se mantuvo alrededor de 0.76 para varios modelos, pero los valores de F1-score fueron cercanos a cero para Regresión Logística, Random Forest y MLP, indicando desempeño deficiente en detección de la clase positiva (HTA). El Perceptrón presentó F1-score mayor pero con reducción marcada del Accuracy, sugiriendo sesgo hacia la clase minoritaria con mayor cantidad de falsos positivos.

Desempeño con balanceo mediante SMOTE

Al incorporar SMOTE, se observó incremento sustancial en F1-score para todos los modelos, reflejando mayor capacidad para recuperar la clase positiva. La Regresión Logística + SMOTE obtuvo el mejor F1-score (0.404), mientras que Random Forest + SMOTE logró el mayor PR-AUC (0.345), resaltando su capacidad para mantener precisión al incrementar recall bajo desbalance. El MLP + SMOTE presentó desempeño competitivo, confirmando que modelos no lineales capturan patrones complejos en variables clínicas.

Tabla 1. *Comparación global del desempeño de los modelos evaluados*

Modelo	F1-score	PR-AUC	Observación
Regresión Logística + SMOTE	0.404	0.341	Mejor F1; interpretable
Random Forest + SMOTE	0.393	0.345	Mejor PR-AUC
MLP + SMOTE	0.397	0.340	Desempeño competitivo
Sin balanceo	< 0.05	< 0.34	No recomendado

Fuente: Elaboración propia

Comparación global

La Regresión Logística + SMOTE destaca por su mejor F1-score y ventaja de interpretabilidad, haciéndola atractiva para implementación clínica. Random Forest + SMOTE ofrece el mejor PR-AUC, siendo opción sólida cuando se prioriza discriminación de clase positiva a distintos umbrales. El MLP + SMOTE mantiene desempeño consistente, aunque su interpretabilidad es menor y requiere mayor cuidado en entrenamiento.

Curvas de aprendizaje y matrices de confusión

Las curvas de pérdida para MLP y SGD mostraron rápida disminución inicial y posterior estabilización, con brechas pequeñas entre entrenamiento y validación, indicando convergencia sin sobreajuste severo. Las matrices de confusión revelaron que Random Forest alcanza mejor equilibrio en clasificación de ambas clases. El MLP presentó mayor número de falsos negativos. Desde perspectiva clínica, los falsos negativos (casos de HTA no detectados) representan limitación relevante al retrasar intervenciones preventivas tempranas. Se evaluó Random Forest con umbral optimizado (t = 0.19), priorizando reducción de falsos negativos. Esto evidenció incremento significativo en detección de casos positivos de HTA, a costa de aumento controlado de falsos positivos. Este compromiso resulta clínicamente deseable en prevención primaria, donde el costo de no identificar un caso real supera al de realizar evaluaciones adicionales.

Importancia de variables

La importancia estimada por Random Forest mostró que PAS y PAD concentran la mayor relevancia, coherente con la definición clínica de HTA y su rol como predictoras principales. Las variables metabólicas (glucosa y perfil lipídico) contribuyen complementariamente,

aportando información adicional sobre el perfil cardiometabólico.

Comparación entre modelos clásicos y de aprendizaje automático

Al comparar directamente modelos clásicos con enfoques de ML, se observa brecha clara en el equilibrio entre desempeño predictivo y flexibilidad. La regresión logística ofrece estructura interpretable y sencilla, pero su rendimiento se limita por relaciones no lineales. Random Forest demostró robustez, manteniendo alta capacidad predictiva a pesar de correlación entre variables y datos imputados. Además, permitió analizar importancia relativa de variables clínicas, facilitando interpretación de factores con mayor impacto en riesgo cardiovascular. Este aspecto resulta especialmente valioso en el ámbito clínico, donde la interpretabilidad es requisito clave para adopción de modelos predictivos. La red neuronal multicapa alcanzó rendimiento comparable o superior en términos de AUC, aunque su naturaleza de "caja negra" limita la interpretación directa. Mayor desempeño predictivo no siempre implica adopción inmediata en práctica médica si no se acompaña de mecanismos adecuados de explicación.

Impacto del preprocesamiento y datos longitudinales

Las técnicas de preprocesamiento tuvieron impacto importante. La imputación mediante mediana permitió mantener mayor número de registros sin introducir sesgos extremos. SMOTE aplicado únicamente sobre entrenamiento ayudó a mitigar desbalance de clases sin comprometer validez del conjunto de prueba. El enfoque longitudinal capturó la evolución temporal de factores de riesgo, ofreciendo representación más fiel del proceso de desarrollo de ECV. Este análisis facilita identificación de trayectorias de riesgo que pueden pasar desapercibidas en enfoques transversales. Estos resultados coinciden con estudios previos que destacan el valor de datos longitudinales en predicción de eventos cardiovasculares, particularmente en población joven donde los cambios metabólicos se manifiestan gradualmente.

Conclusiones

De los resultados mostrados, su análisis y discusión, se pueden obtener las siguientes conclusiones sobre la predicción del riesgo cardiovascular en población joven mediante aprendizaje automático:

- Los modelos de aprendizaje automático con balanceo de clases (SMOTE) superan significativamente a los enfoques sin balanceo, incrementando el F1-score de <0.05 a valores superiores a 0.39, demostrando su efectividad para detectar casos positivos de hipertensión arterial en población joven.
- La Regresión Logística + SMOTE alcanza el mejor F1-score (0.404) y ofrece alta interpretabilidad clínica, constituyendo la opción más equilibrada para implementación en contextos donde la explicabilidad del modelo es prioritaria.
- Random Forest + SMOTE obtiene el mejor PR-AUC (0.345), destacándose como la alternativa más robusta cuando se requiere optimizar la discriminación de casos positivos a diferentes umbrales de decisión, particularmente útil en prevención primaria.
- El uso de datos longitudinales armonizados de ENSANUT (2012, 2018, 2022) permite capturar la evolución temporal de factores de riesgo cardiovascular, superando las limitaciones de los enfoques transversales tradicionales en la identificación de trayectorias clínicas relevantes.

- La presión arterial sistólica y diastólica se confirman como los predictores más importantes según el análisis de importancia de variables en Random Forest, validando su rol determinante en la predicción temprana de riesgo cardiovascular en adultos jóvenes.
- El ajuste del umbral de decisión (de 0.5 a 0.19 en Random Forest) permite priorizar la reducción de falsos negativos, estrategia clínicamente deseable en prevención primaria donde el costo de no detectar un caso real supera al de realizar evaluaciones adicionales.

Agradecimiento

Se agradece a la Universidad Politécnica Salesiana y al Grupo de Investigación en Telecomunicaciones (GIETEC) por el apoyo brindado en el desarrollo de esta investigación.

Referencias Bibliográficas

Banerjee, T. y Paçal, İ. (2025). A systematic review of machine learning in heart disease prediction. *Turkish Journal of Biology*, *49*, 600. https://pmc.ncbi.nlm.nih.gov/articles/PMC12614364/

Liu, W., Laranjo, L., Klimis, H., Chiang, J., Yue, J., Marschner, S., Quiroz, J., Jorm, L. y Chow, C. (2023). Machine-learning versus traditional approaches for atherosclerotic cardiovascular risk prognostication in primary prevention cohorts: A systematic review and meta-analysis. *European Heart Journal*, 310-322.

Martin, S., Aday, A., Almarzooq, Z., Anderson, C., Arora, P., Avery, C., Baker-Smith, C., Gibbs, B., Beaton, A., Boehme, A., Commodore-Mensah, Y., Currie, M., Elkind, M., Evenson, K., Generoso, G., Heard, D., Hiremath, S., Johansen, M., Kalani, R. y Palaniappan, L. (2024). 2024 heart disease and stroke statistics: A report of US and global data from the American Heart Association. *Circulation*, *149*, E347–E913. https://pubmed.ncbi.nlm.nih.gov/38264914/

Nguyen, H., Vasconcellos, H. y Keck, K. (2021). Longitudinal clinical data analysis for cardiovascular risk prediction using machine learning. *Journal of Medical Systems*, *45*(8), 78.

Rajkomar, A., Dean, J. y Kohane, I. (2019). Machine learning in medicine. *New England Journal of Medicine, 380*, 1347–1358.

Salah, H. y Srinivas, S. (2022). Explainable machine learning framework for predicting long-term cardiovascular disease risk among adolescents. *Scientific Reports*, *12*, 12.

World Health Organization. (2023). World heart report 2023. World Health Organization.