Hipertensión arterial en adultos: análisis de riesgo y clasificación predictiva mediante Random Forest

Arterial hypertension in adults: risk analysis and predictive classification using Random Forest

Daniel Lévano Rodriguez[A], Flor Elizabeth Cerdán León[B], Jesus Inocencio Lopez Rodriguez[C] , Mauricio Antonio Alaluna Godinez[D], Siloh Draguichy Valladares Salas[E] y Diego Pereira Sartori[F]

Resumen— La hipertensión arterial (HTA) ha sido considerada un reto en la salud debido a su impacto en la morbimortalidad cardiovascular y a su frecuente diagnóstico tardío. Se abordó el problema mediante el desarrollo de un modelo de clasificación predictiva basado en Random Forest, con el objetivo de identificar individuos en riesgo de hipertensión a partir de variables clínicas, demográficas y metabólicas. Se utilizó una base de datos proveniente de pacientes en México; mediante la técnica SMOTE-Tomek fue sometida a procesos de limpieza, normalización y balanceo. Se entrenó el modelo con el 70% de los datos y se validó con el 30% restante, aplicando validación cruzada k-fold (k=10). Se evaluó el rendimiento del modelo mediante métricas como precisión, sensibilidad, puntaje F1 y matriz de confusión. Se comparó el modelo con otros métodos como KNN y Decisión Tree. Se alcanzó una exactitud del 98% con el modelo optimizado (127 árboles, profundidad 20) destacando como predictores claves el índice de masa corporal, la presión arterial, la actividad física, el peso y la circunferencia de cintura. Aunque también se evaluaron biomarcadores metabólicos, estos presentaron menor relevancia en la clasificación frente a los predictores antropométricos. Los resultados obtenidos confirman que Random Forest es una herramienta robusta y precisa para la detección temprana del riesgo de hipertensión. Gracias a su integración mediante una API y un formulario interactivo, el modelo resulta accesible incluso para usuarios sin formación técnica, lo que contribuye a estrategias preventivas de salud pública.

Palabras clave: Aprendizaje automático, factores de riesgo, salud pública.

C&T Riqchary

Revista de Investigación en Ciencia y Tecnología
ISSN: 2810-8124 (en línea) / ISSN: 2706-543x
Universidad Nacional Micaela Bastidas de Apurimac – Perú

Vol. 7 Núm. 2 (2025) - Publicado: 19/08/25 - Indexaciones
Número: doi.org/10.57166/riqchary/v7.n2.2025
Páginas: 1- 8 | Recibido 08/07/2025 ; Aceptado 08/12/2025

doi.org/10.57166/riqchary.v7.n2.2025.1

Autores:

A. ORCID iD https://orcid.org/0000-0001-5652-0601
Daniel Lévano Rodriguez, Universidad Nacional Tecnológica de Lima Sur dlevano@untels.edu.pe.

B. ORCID iD https://orcid.org/0000-0001-6747-6335
Flor Elizabeth Cerdán León, Universidad Nacional Tecnológica de Lima Sur fcerdan@untels.edu.pe.

C. ORCID iD https://orcid.org/0009-0008-7225-7223
Jesus Inocencio Lopez Rodriguez, Universidad Nacional Tecnológica de Lima Sur 2223050507@untels.edu.pe.

D. ORCID iD https://orcid.org/0009-0003-9209-5210
Mauricio Antonio Alaluna Godinez, Universidad Nacional Tecnológica de Lima Sur 2223110254@untels.edu.pe.

E. ORCID iD https://orcid.org/0009-0008-0912-1008
Siloh Draguichy Valladares Salas, Universidad Peruana Unión silohvalladares@upeu.edu.pe.

F. ORCID iD https://orcid.org/0009-0004-2899-389X
Diego Pereira Sartori, Universidad Peruana Unión diegopereira@upeu.edu.pe.

Abstract— Arterial hypertension (AH) has been considered a major public health concern due to its impact on cardiovascular morbidity and mortality and its frequent late diagnosis. This study addresses the problem by developing a predictive classification model based on the Random Forest algorithm, aiming to identify individuals at risk of hypertension using clinical, demographic, and metabolic variables. A dataset from patients in Mexico was used and processed through cleaning, normalization, and balancing with the SMOTE-Tomek technique. The model was trained with 70% of the data and validated with the remaining 30%, using 10-fold cross-validation. Its performance was evaluated through metrics such as precision, recall, F1-score, and confusion matrix. The model was compared with other methods such as KNN and Decision Tree. The optimized model (127 trees, depth 20) achieved an accuracy of 98% with body mass index, blood pressure, physical activity, weight, and waist circumference identified as the most relevant predictors. Although metabolic biomarkers were also evaluated, they were less relevant in the classification compared to anthropometric variables. The results confirm that Random Forest is a robust and accurate tool for the early detection of hypertension risk. Thanks to its integration via an API and an interactive form, the model is accessible even to not-technical users, contributing to preventive strategies in public health.

Keywords: Machine learning, risk factors, public health

1 INTRODUCCIÓN

La Hipertensión Arterial (HTA) es un factor de riesgo crucial para las enfermedades cardiovasculares y una de las principales causas de mortalidad a nivel mundial, con un estimado actual del 22% de la población global, quienes padecen esta condición, y menos de una quinta parte de los afectados lleva un adecuado control de su presión arterial [1] [2], lo que incrementa el riesgo de complicaciones graves. Este panorama resalta la necesidad de herramientas efectivas para la detección temprana y el manejo de la HTA, dado que falta conciencia en etapas iniciales contribuye a su sub diagnóstico y a un control insuficiente.

En Perú, la prevalencia estandarizada por edad es del 19,2% [3]. En México, la prevalencia de HTA en adultos fue del 47,8% [4]. En Ecuador presenta una mayor prevalencia de HTA [5]; y se estima que afecta al menos al 19,8% de la población, vinculada a factores como la obesidad, el sedentarismo y los antecedentes familiares [5][6]. Estas cifras reflejan una carga epidemiológica compartida en estos países, y la elevada prevalencia de obesidad y sedentarismo parece estar estrechamente relacionada con el aumento de los casos

de HTA, esto evidencia la urgencia de herramientas predictivas para mejorar la detección temprana y personalizar estrategias preventivas.

Machine Learning (ML) es una rama de la inteligencia artificial. Su enfoque está en desarrollar algoritmos capaces de aprender a partir de datos y tomar decisiones o hacer predicciones sin necesidad de una programación muy compleja [7]. A través del ML se permite mejorar los resultados clínicos mediante modelos predictivos, entre sus principales ventajas se encuentran: Automatización de análisis de datos, detección temprana de patrones anómalos y la posibilidad de integrar múltiples variables en la toma de decisiones clínicas [8].

Random Forest es un algoritmo de aprendizaje automático basado en la construcción de múltiples árboles de decisión durante el entrenamiento y la combinación de sus resultados para mejorar la precisión y reducir el riesgo de sobreajuste [9]. En el contexto de HTA, permite identificar las variables más relevantes que contribuyen al riesgo y detectar patrones complejos y no lineales entre múltiples variables clínicas y sociodemográficas, mejorando la exactitud de la predicción o discriminación clínica [9].

Las métricas de evaluación tienen gran importancia en el desarrollo de modelos de aprendizaje automático, siendo estas la accuracy (exactitud), que mide la proporción de flujos de tráfico adecuadamente clasificados sobre el total; la precisión (precisión), que refleja la fracción de predicciones positivas acertadas entre todas las positivas; recall (sensibilidad), que evalúa la proporción de positivos reales identificados; y el F1-score (puntaje F1), que balancea precisión y sensibilidad. Estas definiciones se derivan de los verdaderos positivos (VP), verdaderos negativos (VN), falsos positivos (FP) y falsos negativos (FN) [10].

$$Precisión = \frac{VP}{VP + FP}$$

$$Sensibilidad = \frac{VP}{VP + FN}$$

$$Exactitud = \frac{VP + VN}{VP + FP + FN + VN}$$

$$PuntajeF1 = 2\frac{(Sensibilidad * Precisión)}{(Sensibilidad + Precisión)}$$

En una investigación previa, utilizaron el algoritmo Random Forest con el objetivo de clasificar a individuos con HTA, enfermedad arterial coronaria (EAC) y aquellos sin enfermedades cardiovasculares (no-ECV), a partir de perfiles metabolómicos plasmáticos [11]. En este estudio se enfoca en el diagnostico cardiovascular en adultos mediante el análisis utilizando aprendizaje automático, dependiendo de variables como aminoácidos, acilcarnitinas, metilargininas, etc [11]. Se observa en este estudio una similitud con el presente trabajo en cuanto al uso de algoritmos de clasificación y resalta la necesidad de integrar datos clínicos para mejorar la predicción del modelo.

Otra investigación fue la realizada por [12], demostró la eficacia del algoritmo Random Forest en la detección temprana de enfermedades cardiovasculares, para la creación de su modelo de clasificación fue necesario un riguroso preprocesamiento de datos, incluyendo la eliminación de valores atípicos y la selección de características relevantes para optimizar el modelo predictivo. La investigación señala que la combinación de técnicas de limpieza de datos y selección de características mejora gradualmente el rendimiento del modelo haciéndolo más confiable.

Bajo este contexto, se consideró que mediante el uso de ML se puede mejorar los resultados clínicos mediante modelos predictivos. La aplicación de estas técnicas de aprendizaje automático permite facilitar la detección de patrones complejos y mejora la capacidad discriminativa en comparación con otros métodos tradicionales [13]. Esto da inicio a cómo desarrollar herramientas de soporte, preventivas y personalizadas, en el manejo de la HTA.

El objetivo de este estudio fue identificar de manera integral los factores de riesgo que incrementan la probabilidad de desarrollar HTA, combinando un enfoque individual con métodos de clasificación Random Forest para optimizar las estrategias de prevención y control de la enfermedad [14]. Se proyecta que el modelo de Random Forest, entrenado con variables clínicas y demográficas, puede predecir el riesgo de HTA con una exactitud superior al 90%. Se espera que el peso y la actividad física sean los predictores más influyentes, dada su conocida relación con el desarrollo de la HTA en esta población.

2 MATERIALES Y METODOLOGÍA

En esta sección se describe la arquitectura de la solución (véase la Figura 1), que abarca desde el preprocesamiento del conjunto de datos (dataset) hasta la interacción con el usuario a través de un formulario, incluyendo el desarrollo del modelo Random Forest y la integración de una API.

En esta investigación, se empleó la metodología CRISP-DM para estructurar el proceso de minería de datos, un estándar ampliamente reconocido que guía la preparación de datos, modelado, evaluación y despliegue [15].

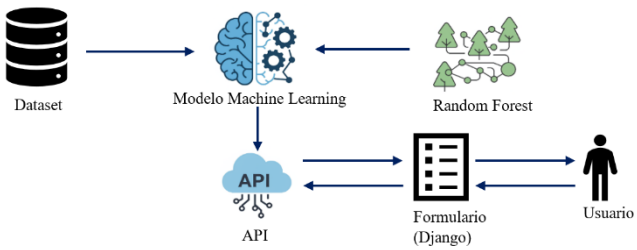

Fig. 1. Arquitectura de la solución.

2.1 Limpieza de datos

Este dataset, que contaba con 4364 registros, se obtuvo a partir de un repositorio de ML, específicamente del sitio web Kaggle, y contiene información demográfica y clínica de pacientes. Las variables registradas incluyeron sexo

(codificado numéricamente), edad y diversos biomarcadores, tales como hemoglobina, ácido úrico, albúmina, colesterol HDL y total, creatinina, glucosa e insulina. La variable dependiente fue el "Riesgo de hipertensión (0 = sin riesgo, 1 = con presencia)". La muestra se obtuvo a partir de registros de pacientes en México, lo que introdujo un sesgo de selección, ya que la mayoría de los datos provenían de centros urbanos [12].

En la limpieza de datos, los criterios de eliminación de registros incluyeron, la eliminación de datos nulos y que se encuentren fuera de su rango. se realizó un análisis exploratorio para evaluar la distribución de los datos y detectar valores atípicos [16]. Se eliminaron datos redundantes que no contribuyen en la predicción del modelo, había valores extremos fuera de los rangos fisiológicos establecidos por cada biomarcador como horas de sueño, actividad total; también la variable actividad total se agrupo en 5 partes para clasificarlos (0 - 4). Los valores nulos de la variable estatura, medida de cintura, tensión arterial, fueron eliminados. El dataset se redujo a 3019 registros, posteriormente se guardó como un archivo CSV con el nombre de "HTA_limpio".

2.2 Balanceo de datos

Se identificó un desbalance en los datos (véase la Figura 2), con una diferencia de 1247 (Con riesgo – Sin riesgo), se observó un predominio de casos positivos frente a casos negativos. En un script de Python, los datos fueron cargados desde el archivo CSV utilizando la librería pandas. Se utilizó la técnica SMOTE-Tomek [16]. Mediante la librería imblearn que combina el sobremuestreo sintético de la clase minoritaria (SMOTE) con la eliminación de muestras cercanas entre clases diferentes (Tomek Links), generando un volumen de entrenamiento más equilibrado. Tras aplicar el método, se pasó de tener 1680 casos positivos y 433 negativos a obtener aproximadamente 1671 muestras en cada clase como se visualiza en la Fig. 2, tras ajustar los datos con dicha técnica. Este nuevo dataframe se guardó con el nombre "HTA _balanceado" para su uso en el entrenamiento del modelo.

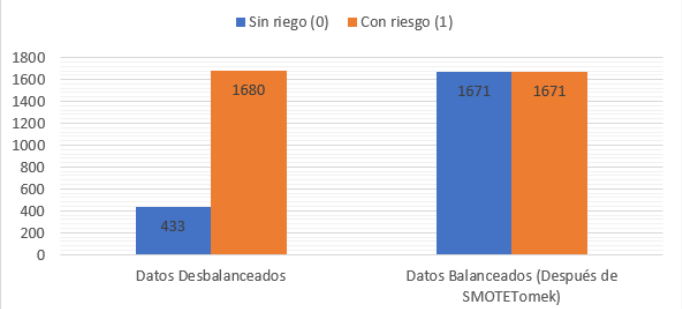

Fig. 2. Comparación y distribución de casos de hipertensión antes y después del balanceo.

2.3 Entrenamiento del modelo

Se empleó un script adicional en Python, utilizando las librerías pandas, numpy y sklearn, se cargaron los datos ya balanceados, se mostró la significancia de cada variable, con el fin de conocer que variables influyen más en el riesgo de HTA. Aquí se seleccionaron las variables más relevantes para el modelo, que se usarán de ahora en adelante.

Otro script de Python, destinado al entrenamiento del algoritmo, utilizando las mismas librerías mencionadas, se cargó el nuevo volumen balanceado, donde se leen las variables seleccionadas para el entrenamiento del modelo. Se separó la variable objetivo (riesgo_hipertension) del resto de las variables predictoras. Posteriormente, los datos fueron divididos en conjuntos de entrenamiento (70%) y prueba (30%) [17] [18] [19].

Para la clasificación del riesgo de hipertensión se implementó el algoritmo RandomForestClassifier utilizando la librería sklearn en Python. Este modelo requería un número de estimaciones (n_estimators) y una profundidad (max_depth), que captura relaciones complejas entre las variables sin sobreajustar el modelo. También fue necesario fijar una semilla (random_state) para que el modelo no varíe cada vez que se ejecute.

La elección del método Random Forest se basa en su capacidad para manejar múltiples variables simultáneamente, su resistencia a los valores atípicos y su eficiencia para identificar relaciones complejas en contextos de clasificación médica [20]. El desempeño del modelo fue validado utilizando matriz de confusión, informe de clasificación y validación cruzada, alcanzando una precisión general del 98% [21] [22].

2.4 Optimización del modelo

La optimización del modelo ha requerido crear un primer modelo con 50 árboles de decisión (n_estimators=50) y una profundidad de 4 (max_depth=4), usando una semilla fija de 42 (random_state=42). El modelo fue entrenado para evaluar su desempeño utilizando el informe de clasificación (classification_report), el cual proporciona métricas como la precisión, la sensibilidad y puntaje F1, tanto para los casos "Con riesgo" como "Sin riesgo".

Se utilizó una matriz de confusión, con el modelo de 50 árboles, para evaluar gráficamente el desempeño del modelo de clasificación. Se emplearon las bibliotecas matplotlib y seaborn, ampliamente utilizadas en Python para la visualización de datos.

Se encontró la selección más óptima del modelo usando la tasa de error, mediante un gráfico generado en Python usando la librería sklearn y matplotlib, y se descubrió (véase la Figura 3) que, si se usaban una cantidad mayor de hasta aproximadamente 129 árboles, la tasa de error es menor.

https://doi.org/10.57166/riqchary.v7.n2.2025.1

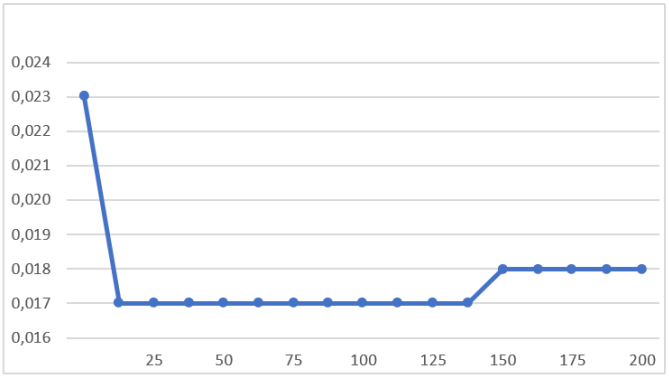

Fig. 3. Tasa de error vs Número de árboles en Random Forest.

Considerando la tasa de error, se optó un modelo con 127 árboles de decisión (n_estimators=127) y una profundidad máxima de 20 (max_depth=20), y se usó una semilla fija de 42 (random_state=42). Se compararon las métricas entre los dos modelos para comprobar el modelo más óptimo, siendo el de 127 árboles.

2.5 Comparación del modelo

Tomando en cuenta la profundidad máxima y la semilla fija empleados en el modelo de random forest, se comparó en otros códigos de Python modelos que usan tanto KNN como Árbol de Decisión, utilizando las librerías pandas, numpy, sklearn y matplotlib. De esta manera se comprobó el modelo más adecuado, siendo el de Random Forest.

2.6 Predicción y despliegue

Para la generación de la API se empleó el framework Flask, utilizando las librerías flask, threading y time, donde se carga el modelo, se definen los datos esperados de la API y su respuesta. Esta define una ruta en la cual se le podrá enviar solicitudes con datos en formato JSON procesados en tiempo real.

Se simularon entradas de prueba a través de esta API para evaluar el funcionamiento del modelo y se verificó un adecuado desempeño para predecir el riesgo de hipertensión. Se creó un formulario para acceder al modelo y la API creada, empleando el framework Django, en el cual se le pueden colocar los datos pedidos, que transformará los datos en formato JSON y los enviará al API, donde serán analizados por el modelo y dará una respuesta, que la API devuelve al formulario para su visualización.

El formulario desarrollado con Django (mostrado en la figura 4) permite procesar entradas y devolver resultados con los datos colocados, calcular ciertas variables (como el índice de masa corporal) necesarias para el modelo con otras variables del formulario, optimizando el tiempo, sea dinámico y más sencillo de usar. Este formulario mostrará directamente el resultado (Con riesgo o Sin riesgo), tiene una función enviar un reporte, que se envía por correo en un PDF, con los datos colocados con el resultado del paciente.

Fig. 4. Visualización del formulario.

3 Resultados

3.1 Significancia de los factores de riesgo

El resultado de las variables más influyentes para el modelo, como se observa en la Figura 4, son el índice de masa corporal con un 26.21%, la tensión arterial con un 21.84%, la medida de cintura con un 14.61%, el peso con un 11.09%, la actividad física con un 5.94%, la estatura con un 2.92%, las horas de sueño con un 2.42%, el sexo con un 2.33%, la edad con un 2.24%, la HbA1c (Hemoglobina glucosilada) con un 1.13%, la glucosa promedio con un 0.99%, la insulina con un 0.76%, el colesterol LDL con un 0.70%, los triglicéridos con un 0.67%, la glucosa con un 0.62%.

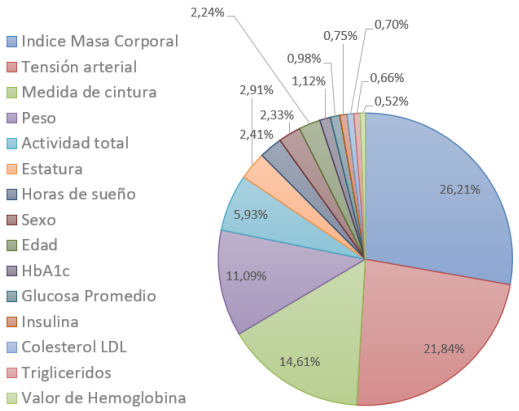

Fig. 5. Importancia de los factores de riesgo.

3.2 Comparación y selección del modelo

El modelo inicial, con 50 árboles de decisión, como indica la Tabla 1, se obtiene una exactitud de 96%, para casos "Con riesgo" la precisión de 97% y una sensibilidad del 95%; para casos "Sin riesgo" la precisión es de 95% y una sensibilidad del 97%.

TABLA 1
Valores de las métricas de clasificación del modelo con 50 árboles.

Métrica	Precisión	Sensibilidad	Puntaje F1
Sin riesgo	0.95	0.97	0.96
Con riesgo	0.97	0.95	0.96
Exactitud	-	-	0.96

Los resultados de la matriz de confusión, tal como se aprecia en la Tabla 2 muestra que los 479 casos clasificados como "Sin riesgo" (VN+FP), 465 fueron correctamente identificados, y de los 524 casos "Con riesgo" (FN+VP), 500 fueron correctamente predichos. Se obtuvieron 14 falsos positivos, y 24 falsos negativos.

TABLA 2
Valores obtenidos de la matriz de confusión con 50 árboles.

n_estimators=50		Predicción	
		Sin riesgo	Con riesgo
Valores reales	Sin riesgo	VN = 465	FP = 14
	Con riesgo	FN = 24	VP = 500

El modelo mejorado, con 127 árboles de decisión (véase Tabla 3), obtuvo una exactitud de 98%, para casos "Con riesgo" la precisión de 99% y una sensibilidad del 97%; para casos "Sin riesgo" la precisión es de 97% y una sensibilidad del 99%.

TABLA 3.
Valores de las métricas de clasificación del modelo optimizado.

Métrica	Precisión	Sensibilidad	Puntaje F1
Sin riesgo	0.97	0.99	0.98
Con riesgo	0.99	0.97	0.98
Exactitud	-	-	0.98

Los valores obtenidos de la matriz de confusión (véase la Tabla 4) de los 480 casos reales clasificados como "Sin riesgo" (VN+FP), 476 fueron correctamente identificados, los 524 casos "Con riesgo" (FN+VP), 510 fueron identificados como verdaderos positivos; estos resultados evidencian que los falsos positivos se redujeron de 14 a 4, y los falsos negativos de 24 a 14.

TABLA 4
Valores obtenidos de la matriz de confusión con 127 árboles.

n_estimators=127		Predicción	
		Sin riesgo	Con riesgo
Valores reales	Sin riesgo	VN = 476	FP = 4
	Con riesgo	FN = 14	VP = 510

En la Figura 6 se compara que el modelo con 127 árboles difiere con el modelo con 50 árboles, mostrando una diferencia del 2% en precisión, sensibilidad y puntaje F1 para ambas clases. Con respecto a la exactitud general, pasa de 96% a 98%.

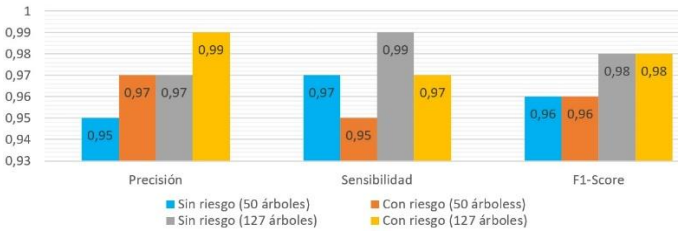

Fig. 6. Comparación entre el modelo con 50 árboles vs modelo con 127 árboles.

El modelo entrenado con KNN, como se indica en la Tabla 5, obtiene una exactitud de 92%; para casos "Con riesgo" la precisión de 98% y una sensibilidad del 85%; para casos "Sin riesgo" la precisión es de 87% y una sensibilidad del 98%.

TABLA 5
Valores de las métricas de clasificación del modelo KNN.

Métrica	Precisión	Sensibilidad	Puntaje F1
Sin riesgo	0.87	0.98	0.92
Con riesgo	0.98	0.85	0.91
Exactitud	-	-	0.92

Los resultados obtenidos con el modelo Decision Tree son semejantes a los obtenidos por el modelo con Random Forest (mostrado en la Tabla 3), por lo tanto, se comparó sus curvas AUC y ROC, como se muestra en la Figura 7. El modelo Decision Tree tiene un valor AUC de 0.998 y el modelo Random Forest un valor AUC de 0.976.

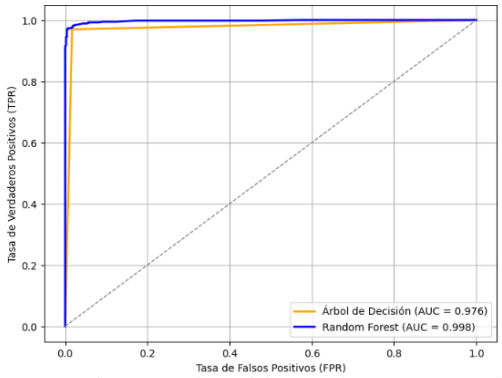

Fig. 7. Comparación entre el modelo Random Forest vs el modelo Decision Tree.

Se realizó la validación cruzada del modelo Random Forest y del modelo Decision Tree, como se muestra en la Tabla 6. Random Forest obtuvo una precisión de 98.05% con una desviación estándar del 0.31%, mientras que, el modelo Decision Tree obtuvo una precisión promedio de 97.48% con una desviación estándar del 0.97%.

TABLA 6.
Valores obtenidos de la validación cruzada.

Modelos	Precisión	Desviación estándar
Random Forest	98.05%	0.31%
Decision Tree	97.48%	0.97%

4 DISCUSIONES

4.1 Análisis de resultados y comparación con estudios previos

El análisis de la importancia de las variables reveló que, en el conjunto total de variables, los factores con mayor peso predictivo fueron la masa corporal, actividad total, peso, tensión arterial y medida de cintura. El análisis enfocado en los biomarcadores (véase Figura 4), observó que ciertos parámetros metabólicos (por ejemplo, HbA1c, colesterol, triglicéridos, insulina, etc.) tuvieron una importancia relativamente menor de lo esperado. Esto sugiere que los indicadores relacionados con el estado físico (masa corporal, peso y medidas antropométricas) son determinantes clave para la clasificación del riesgo hipertensivo. No hay que descartar la utilidad de biomarcadores, pues futuras investigaciones podrían explorar el impacto de estos biomarcadores en combinación con otras variables clínicas para mejorar la capacidad predictiva del modelo) [23] [24] [25].

Los resultados obtenidos demuestran que el modelo Random Forest, como parte de los modelos de aprendizaje automático, supera significativamente a métodos tradicionales, como la regresión logística y KNN, en la detección de patrones complejos presentes en los datos clínicos. Un estudio reciente realizado por Kandil et al. Demostró que los métodos tradicionales tienen una menor capacidad predictiva o discriminatoria en comparación con algunos enfoques más avanzados, lo que puede respaldar el uso de algoritmos de ML en el análisis del riesgo de hipertensión [26] [27] [28].

Comparado a un modelo desarrollado en una investigación previa [11], dicho grupo alcanzó una precisión del 80% en la clasificación multiclase y del 91% en la clasificación binaria entre casos de ECV y no-ECV. También, en otros trabajos [29], se utilizó un modelo de Árbol de Decisión para la predicción de enfermedades cardiacas, el cual alcanzó una exactitud cercana al 79,78%, lo que evidenció unas limitaciones en la detección de patrones complejos y una mayor susceptibilidad al sobreajuste. En contraste, el modelo presentado en este estudio logró una exactitud del 98%, como se evidencia en la Tabla 3, la cual se ha confirmado nuestra hipótesis y ha dado un excelente desempeño del modelo si queremos saber el riesgo de HTA de un paciente. Este desempeño indica que el modelo puede servir como una herramienta útil al momento de tomar decisiones con respecto a los pacientes y su condición actual.

Otro estudio [12] desarrollo un modelo de predicción de enfermedad cardiovascular utilizando Random Forest y un dataset clínicos del repositorio de Kaggle, alcanzando una exactitud del 99% en la clasificación binaria de presencia o ausencia de enfermedad cardíaca. Si bien su modelo abarcó un espectro más amplio de patologías cardiovasculares, el modelo del presente artículo se enfocó exclusivamente en la predicción del riesgo de HTA, alcanzando una exactitud del 98% y métricas por clase altamente equilibradas (precisión de 99% y sensibilidad de 97% en la clase "Con riesgo"; precisión de 97% y sensibilidad de 99% en la clase "sin riesgo") (véase tabla 3).

4.2 Limitaciones y mejoras a futuro

A pesar de la diferencia en los objetivos clínicos, ambos estudios comparten un nivel metodológico riguroso y resultados de alto rendimiento, lo que posiciona nuestra propuesta como una alternativa igualmente sólida y competitiva como modelo de soporte clínico en el ámbito de la salud.

El desempeño del modelo es significativamente prometedor, aun así, se recomienda explorar mejoras que incluyan:

a. Incorporación de métricas adicionales y validaciones externas que demuestren su utilidad en la evaluación del riesgo de sobreajuste.
b. Ampliación de la muestra hacia diferentes contextos poblacionales para aumentar la validez externa de los hallazgos, no solo de estratos sociales sino también de grupos étnicos variados. Estudios como [30], han resaltado la necesidad de validar estos modelos en muestras heterogéneas antes de su aplicación clínica.

Los presentes hallazgos del estudio llegan a confirmar la utilidad de modelos de ML, especialmente el modelo Random Forest, en la identificación tanto tardía como temprana del riesgo de presión arterial alta. Esta herramienta en su utilidad ayuda con la mejora en precisión de la clasificación y también llega a ofrecer un potencial significativo para la implementación de estrategias preventivas en salud pública.

Aunque el modelo desarrollado presenta un rendimiento sobresaliente con resultados satisfactorios, es importante reconocer que los datos empleados provienen de una población específica, lo que puede limitar la aplicación del modelo a otros contextos geográficos, además, se omitió diferentes atributos del modelo debido a su baja significancia para la predicción. Sin embargo, esto puede cambiar con futuras investigaciones, que amplíen la base de datos y permitan tanto el uso de mayores atributos en el modelo como el abarcar una población mayor para obtener mejores resultados en el modelo.

Estudios como [31], han demostrado que el autocuidado y la educación en salud pueden mejorar significativamente el control de la presión previniendo su elevación y reduciendo la probabilidad de desarrollar HTA en adultos. Se combina un enfoque individual con métodos de clasificación Random Forest para optimizar las estrategias de control de la enfermedad mediante el análisis predictivo con enfoques preventivos [31] [32].

La precisión del modelo utilizado está respaldada por métricas adicionales (AUC-ROC, sensibilidad, precisión), se puede explicar gracias a la calidad del preprocesamiento de los datos y la adecuada selección de hiperparámetros. Sin embargo, se debe señalar que el modelo fue entrenado con una muestra proveniente exclusivamente de centros urbanos de México, lo cual podría inducir un sesgo de selección y limitar la generalización de los resultados a poblaciones de distinto estrado social. En este sentido, estudios como los de Delgado-Galeano han resaltado la necesidad de validar estos

modelos en muestras heterogéneas antes de su aplicación clínica [30].

5 Conclusiones y Recomendaciones

5.1 Conclusiones

Los resultados obtenidos en la investigación evidencian que los objetivos propuestos fueron alcanzados. Se identificaron de manera integral los principales factores de riesgo asociados al desarrollo de la HTA, desarrollando un modelo predictivo basado en Random Forest con una exactitud del 98%, validado mediante métricas especializadas como precisión, sensibilidad y puntaje F1. En conjunto al modelo se implementó un sistema funcional mediante una API integrada a un formulario interactivo, lo que evidencia su aplicabilidad práctica en escenarios reales en contextos clínicos y comunitarios.

El análisis de la importancia de las variables reveló que factores antropométricos, como el índice de masa corporal, la tensión arterial, el nivel de actividad física, el peso corporal y la medida de cintura, fueron más determinantes en la predicción del riesgo que los biomarcadores. Esto sugiere que las estrategias de prevención deberían enfocarse más en la promoción activa de estilos de vida saludables. Sin embargo, el modelo también contempla variables metabólicas, lo que refuerza su carácter integral en la evaluación del riesgo.

Las métricas obtenidas confirman la robustez del modelo, y el bajo margen de error evidenciado en la matriz de confusión refuerza su utilidad como herramienta de tamizaje y soporte clínico. El despliegue y la integración de este modelo en sistemas de salud pública corresponde con el objetivo de desarrollar una estrategia prometedora para la detección precoz y la prevención de enfermedades crónicas como la HTA. Se contribuye a la eficiencia de los sistemas de salud pública y particular gracias a la disponibilidad de herramientas predictivas accesibles, basadas en evidencia.

5.2 Recomendaciones

Para futuras investigaciones, se recomienda ampliar la base de datos con estudios a poblaciones más amplias y de diferentes contextos tanto geográficos como socioeconómicos para obtener mejores resultados, además de explorar la integración de modelos y técnicas avanzadas para optimizar la capacidad predictiva.

Futuros estudios podrían explorar la combinación de este enfoque con modelos más avanzados, como redes neuronales profundas o aprendizaje por refuerzo, para seguir optimizando su capacidad predictiva y adaptabilidad clínica.

Referencias

[1] N. K. Toala-Lino, Y. Peñaherrera Moran, y I. G. Parrales-Pincay, «Hipertensión arterial como factor predisponente de insuficiencia renal en adultos.», *MQRInvestigar* , vol. 7, n.º 1, pp. 367-389, ene. 2023, doi: 10.56048/MQR20225.7.1.2023.367-389.

[2] World Health Organization, «OMS | Hipertensión». Accedido: 9 de agosto de 2025. [En línea]. Disponible en: https://www.who.int/news-room/fact-sheets/detail/hypertension

[3] K. Romero Jares y R. Centeno Quispe, «Perú: Encuesta Demográfica y de Salud Familiar - ENDES 2022», may 2025. Accedido: 9 de agosto de 2025. [En línea]. Disponible en: https://www.inei.gob.pe/media/MenuRecursivo/publicaciones_digitales/Est/Lib1898/libro.pdf

[4] I. Campos-Nonato *et al.*, «Prevalencia, tratamiento y control de la hipertensión arterial en adultos mexicanos: resultados de la Ensanut 2022», *Salud Publica Mex*, vol. 65, pp. s169-s180, jun. 2023, doi: 10.21149/14779.

[5] A. N. Zavala-Hoppe, T. E. Zambrano-Flores, L. H. Vivar-Medina, y J. E. Fuentes-Parrales, «Epidemiología y factores de riesgo de la hipertensión arterial en los países de Latinoamérica y Europa», *MQRInvestigar* , vol. 8, n.º 1, pp. 1371-1389, feb. 2024, doi: 10.56048/MQR20225.8.1.2024.1371-1389.

[6] A. L. Pico Pico, E. Y. Reyes Reyes, D. A. Anchundia Alvia, y M. D. L. Á. Cobos Moreno, «Comportamiento epidemiológico de la hipertensión arterial en el Ecuador», *RECIMUNDO: Revista Científica de la Investigación y el Conocimiento*, vol. 7, n.º 4, pp. 299-307, 2023, doi: 10.26820/recimundo/7.(4).oct.2023.299-307.

[7] A. Shafizadeh *et al.*, «Machine learning-enabled analysis of product distribution and composition in biomass-coal co-pyrolysis», *Fuel*, vol. 355, p. 129464, ene. 2024, doi: 10.1016/J.FUEL.2023.129464.

[8] F. Plazzotta, D. Luna, y F. González Bernaldo de Quirós, «Sistemas de Información en Salud: Integrando datos clínicos en diferentes escenarios y usuarios», *Rev Peru Med Exp Salud Publica*, vol. 32, n.º 2, pp. 343-351, abr. 2015, doi: 10.17843/rpmesp.2015.322.1630.

[9] J. H. Chen, X. L. Wang, y F. Lei, «Data-driven multinomial random forest: a new random forest variant with strong consistency», *J Big Data*, vol. 11, n.º 1, pp. 1-32, dic. 2024, doi: 10.1186/S40537-023-00874-6/FIGURES/7.

[10] R. M. Alzoman, M. J. F. Alenazi, L. Belli, G. Ferrari, M. Martalò, y L. Davoli, «A Comparative Study of Traffic Classification Techniques for Smart City Networks», *Sensors*, vol. 21, n.º 14, p. 4677, jul. 2021, doi: 10.3390/S21144677.

[11] N. E. Moskaleva *et al.*, «Target Metabolome Profiling-Based Machine Learning as a Diagnostic Approach for Cardiovascular Diseases in Adults», *Metabolites*, vol. 12, n.º 12, p. 1185, dic. 2022, doi: 10.3390/METABO12121185/S1.

[12] K. Sumwiza, C. Twizere, G. Rushingabigwi, P. Bakunzibake, y P. Bamurigire, «Enhanced cardiovascular disease prediction model using random forest algorithm», *Inform Med Unlocked*, vol. 41, p. 101316, ene. 2023, doi: 10.1016/J.IMU.2023.101316.

[13] J. L. Speiser, «A random forest method with feature selection for developing medical prediction models with clustered and longitudinal data», *J Biomed Inform*, vol. 117, p. 103763, may 2021, doi: 10.1016/J.JBI.2021.103763.

[14] Y. Surichaqui Gómez y J. A. Mori Castro, «Estilo de

Vida y su Relación con el Estado Nutricional en pacientes Adultos Mayores con Hipertensión Arterial en el Hospital de Huaycán de Lima», *Ciencia Latina Revista Científica Multidisciplinar*, vol. 7, n.º 4, pp. 9069-9089, sep. 2023, doi: 10.37811/CL_RCM.V7I4.7609.

[15] C. Schröer, F. Kruse, y J. M. Gómez, «A Systematic Literature Review on Applying CRISP-DM Process Model», *Procedia Comput Sci*, vol. 181, pp. 526-534, ene. 2021, doi: 10.1016/J.PROCS.2021.01.199.

[16] Y. Yao, Y. He, y H. Ou, «Missing Data Imputation Method Combining Random Forest and Generative Adversarial Imputation Network», *Sensors 2024, Vol. 24, Page 1112*, vol. 24, n.º 4, p. 1112, feb. 2024, doi: 10.3390/S24041112.

[17] M. Alyami *et al.*, «Predictive modeling for compressive strength of 3D printed fiber-reinforced concrete using machine learning algorithms», *Case Studies in Construction Materials*, vol. 20, p. e02728, jul. 2024, doi: 10.1016/J.CSCM.2023.E02728.

[18] S. Khasim, H. Ghosh, I. S. Rahat, K. Shaik, y M. Yesubabu, «Deciphering Microorganisms through Intelligent Image Recognition: Machine Learning and Deep Learning Approaches, Challenges, and Advancements», *EAI Endorsed Transactions on Internet of Things*, vol. 10, nov. 2023, doi: 10.4108/EETIOT.4484.

[19] G. R. Ren *et al.*, «Machine Learning Predicts Recurrent Lumbar Disc Herniation Following Percutaneous Endoscopic Lumbar Discectomy», *Global Spine J*, vol. 14, n.º 1, pp. 146-152, ene. 2024, doi: 10.1177/21925682221097650.

[20] Y. Cooli y C. Mahesh, «Recursive Parallel Partition Random Forest for Medical Disease Classification», *International Journal of Intelligent Engineering and Systems*, vol. 14, n.º 5, p. 2021, doi: 10.22266/ijies2021.1031.11.

[21] J. H. Chung *et al.*, «Random forest identifies predictors of discharge destination following total shoulder arthroplasty», *JSES Int*, vol. 8, n.º 2, pp. 317-321, mar. 2024, doi: 10.1016/J.JSEINT.2023.04.003.

[22] O. Nikolaychuk, J. Pestova, y A. Yurin, «Wildfire Susceptibility Mapping in Baikal Natural Territory Using Random Forest», *Forests 2024, Vol. 15, Page 170*, vol. 15, n.º 1, p. 170, ene. 2024, doi: 10.3390/F15010170.

[23] J. J. Hidalgo Flores, M. A. Guerrero Dueña, y R. García Rodríguez, «La obesidad como factor de riesgo de la hipertensión arterial», *Revista Científica Higía de la Salud*, vol. 5, n.º 2, pp. 2021-2033, dic. 2021, doi: 10.37117/HIGIA.V1I5.576.

[24] D. A. Altamirano Lojano, R. Alvarez Ochoa, J. P. Garcés-Ortega, y G. Cordero Cordero, «Índice de masa corporal e Hipertensión Arterial en Adultos», *Revista Multidisciplinaria Investigación Contemporánea*, vol. 2, n.º 1, pp. 102-131, ene. 2024, doi: 10.58995/REDLIC.IC.V2.N1.A57.

[25] M. De Jesús Sosa-Martínez, I. León-Lozano Jair, Y. García-Jiménez, B. Garduño-Orbe, A. J. Lagarza-Moreno, y G. Juanico-Morales, «Frecuencia de dislipidemias y determinación del riesgo cardiovascular en pacientes con hipertensión arterial sistémica», *Atención Familiar*, vol. 30, n.º 4, pp. 245-250, dic. 2023, doi: 10.22201/fm.14058871p.2023.486536.

[26] P. D. F. Isles, «A random forest approach to improve estimates of tributary nutrient loading», *Water Res*, vol. 248, p. 120876, ene. 2024, doi: 10.1016/J.WATRES.2023.120876.

[27] H. Kandil, A. Soliman, N. S. Alghamdi, J. R. Jennings, y A. El-Baz, «Using Mean Arterial Pressure in Hypertension Diagnosis versus Using Either Systolic or Diastolic Blood Pressure Measurements», *Biomedicines*, vol. 11, n.º 3, p. 849, mar. 2023, doi: 10.3390/BIOMEDICINES11030849.

[28] B. K. Meher, M. Singh, R. Birau, y A. Anand, «Forecasting stock prices of fintech companies of India using random forest with high-frequency data», *Journal of Open Innovation: Technology, Market, and Complexity*, vol. 10, n.º 1, p. 100180, mar. 2024, doi: 10.1016/J.JOITMC.2023.100180.

[29] M. Ahad, D. Padha, y H. Sharma, «Machine Learning in Cardiology: a Survey of Early Detection Models for Heart Diseases», *IJFMR - International Journal For Multidisciplinary Research*, vol. 5, n.º 3, may 2023, doi: 10.36948/IJFMR.2023.V05I03.3113.

[30] M. Delgado-Galeano, «Historia de la hipertensión arterial: revisión narrativa», *Salud UIS*, vol. 55, n.º 1, may 2023, doi: 10.18273/SALUDUIS.55.E:23043.

[31] Y. Anggriani Utama, «Pengaruh Self Management pada Pasien Hipertensi: Sebuah Tinjauan Sistematis», *Jurnal Ilmiah Universitas Batanghari Jambi*, vol. 23, n.º 1, pp. 422-429, feb. 2023, doi: 10.33087/JIUBJ.V23I1.3528.

[32] O. G. Montero Cadena, G. J. Guzmán Kure, R. C. Acosta Bravo, y M. B. Peñafiel Peñafiel, «Principales factores de riesgo de la hipertensión arterial», *RECIMUNDO*, vol. 7, n.º 2, pp. 89-97, jul. 2023, doi: 10.26820/recimundo/7.(2).jun.2023.89-97.

[33] B. E. Rıchar Wılliam, G. V. Ana Lucıla †, A. Escobar Magaly, N. A. Lucıa, J. A. Zevallos Vıllodas, y C. R. Castro Galarza, «LA NO ADHERENCIA AL TRATAMIENTO ANTIHIPERTENSIVO Y FACTORES ASOCIADOS: UNA REVISIÓN», *Advances in Science and Innovation*, vol. 1, n.º 1, pp. 45-52, dic. 2022, doi: 10.61210/ASI.V1I1.5.

APÉNDICE

Data set original fue obtenido a partir de: https://www.kaggle.com/datasets/frederickfelix/hipertension-arterial-mxico

Repositorio en Github de los modelos y el formulario desarrollado: https://github.com/JesusG753/Hipertension_Arterial.git