Predicción de riesgo cardiovascular

Carolina Boix Miralles

Resumen— Las enfermedades cardiovasculares continúan siendo la principal causa de mortalidad mundial. Este trabajo analiza la capacidad predictiva de padecer o no enfermedad cardiovascular con el estudio de distintos algoritmos de inteligencia artificial —Random Forest, regresión logística y redes neuronales profundas— aplicados a un conjunto de datos clínicos de 70K y 11 variables clínicas extraído de una base abierta de la plataforma Kaggle.Tras una fase exhaustiva de preprocesamiento y depuración del conjunto de datos, se procedió al análisis mediante la plataforma BigML, con el objetivo de evaluar el riesgo cardiovascular individual a partir de factores de riesgo tradicionales. El estudio no solo permite aplicar distintos algoritmos de aprendizaje automático, sino también valorar la importancia relativa de cada variable clínica en la predicción del riesgo. Para ello, se emplearon varios tamaños muestrales y diferentes particiones de entrenamiento y prueba, lo que permitió comparar el rendimiento y la estabilidad de los modelos bajo distintas configuraciones, aportando solidez metodológica a los resultados obtenidos.

Palabras clave: Enfermedad Cardiovascular,factores de riesgo,Machine learning,Presión arterial sistólica.

Abstract— Cardiovascular diseases remain the leading cause of mortality worldwide. This study analyzes the predictive capacity of developing or not developing cardiovascular disease by applying various artificial intelligence algorithms —Random Forest, logistic regression, and deep neural networks— to a clinical dataset consisting of 70,000 records and 11 clinical variables, obtained from an open-access database hosted on the Kaggle platform. After an exhaustive phase of data preprocessing and cleaning, the analysis was carried out using the BigML platform, with the aim of evaluating individual cardiovascular risk based on traditional risk factors. The study not only allowed the implementation of different machine learning algorithms, but also made it possible to assess the relative importance of each clinical variable in the prediction of cardiovascular risk. Multiple sample sizes and different training/test partitions were used to compare the performance and stability of the models under various configurations, thereby providing methodological robustness to the results obtained.

Keywords: Cardiovascular disease, risk factors, machine learning, systolic blood pressure.

- *E-mail de contacto: carolinaboix68@gmail.com*
- *Trabajo tutorizado por: Remo Suppi.*
- *Curso: 2025*

1 Introducción

Las enfermedades cardiovasculares (ECV) siguen siendo la principal causa de muerte, aproximadamente 17,9 millones de vidas al año, según los datos de la OMS **[1],** es decir el 32 % de la población mundial. Estos datos reflejan la prevalencia de estas enfermedades,y la necesidad de comprender y controlar los factores de riesgo cardiovascular. Hipertensión, colesterol, tabaquismo, obesidad, inactividad física y diabetes son factores modificables que, si se detectan y gestionan a tiempo, pueden reducir significativamente la probabilidad de eventos cardiovasculares graves. Diversos estudios han demostrado que la mayoría de los factores de riesgo cardiovascular son modificables y prevenibles. Entre los más relevantes se encuentran la hipertensión arterial, los niveles elevados de colesterol, el tabaquismo, la obesidad, la inactividad física y la diabetes tipo 2 **[2], [3]**. Si se detectan y gestionan a tiempo, estos factores pueden reducir la probabilidad de sufrir eventos cardiovasculares graves.

En la práctica clínica, la estimación del riesgo cardiovascular se ha basado históricamente en modelos estadísticos tradicionales, siendo el Framingham Risk Score y el SCORE *(Systematic Coronary Risk Evaluation)* los más ampliamente utilizados. Ambos se desarrollaron a partir de cohortes poblacionales amplias y utilizan ecuaciones de regresión multivariable para predecir el riesgo de eventos cardiovasculares a 10 años. Estas herramientas consideran factores clásicos como la edad, el sexo, la presión arterial, los niveles de colesterol, el tabaquismo y la presencia de diabetes. A pesar de su valor histórico y su aplicabilidad práctica, estos modelos presentan limitaciones relevantes en el contexto actual. En primer lugar, tienden a subestimar el riesgo real en individuos jóvenes con múltiples factores de riesgo, en mujeres y en ciertos grupos étnicos **[4].** En segundo lugar, su enfoque generalista y poblacional impide capturar la complejidad e interacción de variables clínicas individuales. Además, no contemplan factores emergentes como la inflamación, la genética o el entorno psicosocial **[5], lo que** ha generado un debate en torno a su capacidad predictiva.

El avance de la inteligencia artificial (IA) y el aprendizaje automático permite analizar los factores de riesgo cardiovascular,mediante algoritmos predictivos, que pueden procesar grandes volúmenes de datos para identificar patrones que indiquen una alta probabilidad de sufrir enfermedad cardiovascular.

En este estudio se ha realizado un análisis comparativo de distintos algoritmos de inteligencia artificial con el objetivo de predecir la aparición de enfermedad cardiovascular a partir de factores de riesgo clínicos.

Estudios recientes han comparado directamente Naive Bayes, Random Forest y XG Boost para predecir eventos cardiovasculares, encontrando robustez en métodos de ensamble comparando favorablemente frente a escalas tradicionales como Framingham o SCORE,**[6]** .

Por todo ello los modelos personalizados desarrollados mediante inteligencia artificial permiten una evaluación más específica del riesgo cardiovascular individual, ofreciendo así una predicción más precisa. Como consecuencia, permiten detectar perfiles de alto riesgo que podrían pasar desapercibidos con métodos convencionales, lo que favorece una intervención temprana y dirigida, contribuyendo potencialmente a una reducción significativa de la mortalidad cardiovascular

2 Material y métodos

En el siguiente apartado se describen en detalle los datos utilizados en el estudio, especificando su origen, el volumen total de registros disponibles y la naturaleza de las variables incluidas. Además, se expone el proceso de preprocesamiento de los datos llevado a cabo, con el objetivo de garantizar la calidad del conjunto final y su adecuación para los análisis predictivos posteriores mediante técnicas de inteligencia artificial.

2.1 Datos

La base de datos empleada en este estudio se obtuvo de registros clínicos que documentaban factores de riesgo asociados a enfermedades cardiovasculares. Estos registros contenían tanto información demográfica y clínica de los pacientes, como variables estructuradas recabadas durante las consultas médicas, lo que facilitó la identificación de los casos con y sin diagnóstico de enfermedad cardiovascular. El dataset es un conjunto de datos abierto, disponible públicamente a través de la plataforma Kaggle, proporcionado por el autor Kuzac Kundem y está disponible en https://data.world/kudem.

La base de datos empleada incluía **14** variables con información exhaustiva sobre cada paciente. Gracias a la robustez y el tamaño de la muestra fue posible realizar análisis estadísticos complejos y desarrollar modelos predictivos de machine learning enfocados en el riesgo cardiovascular.

Con un total de 70.000 observaciones, el dataset ofreció una base sólida para el estudio de los factores de riesgo asociados a dichas enfermedades y para el desarrollo de modelos predictivos. No obstante, fue fundamental revisar la presencia de valores atípicos y garantizar una correcta interpretación de cada variable antes de llevar a cabo análisis concluyentes.
Cabe señalar que este conjunto de datos no incluía información recogida a lo largo del tiempo, por lo que no fue posible realizar análisis de tipo longitudinal o de seguimiento temporal.

2.2 Variables

Las variables incluidas en este estudio son:

IdPaciente: Identificador único para cada paciente.
Edad : Expresada en días, lo que permite un cálculo preciso
Sexo: codificada numéricamente (1 = femenino, 2 = masculino).
Altura : Altura medida en centímetros (entero).
Peso : Peso medido en kilogramos (entero).
Sistólica : Lectura de presión arterial sistólica del paciente (entero).
Diastólica : Lectura de presión arterial diastólica del paciente (entero).
Colesterol: nivel de colesterol total leído como mg/dl en una escala de 0 a 5+unidades (entero). Cada unidad denota aumento/disminución de 20 mg/dL.
Glucosa: nivel de glucosa leído como mmol/l en una escala de 0 a 16+ unidades (entero). Cada unidad denota aumento/disminución de 1 mmol/L respectivamente.
Tabaco: Hábito tabáquico (0 = no, 1 = sí).
Alcohol: Consumo de alcohol (0 = no, 1 = sí).
Actividad física: Nivel de actividad (0 = inactivo, 1 = activo).
Cardiopatía: Diagnóstico de enfermedad cardiovascular (0 = no, 1 = sí).
Transformaciones aplicadas a las variables cuantitativas:
Para mejorar la legibilidad y facilitar cálculos posteriores, se aplicaron las siguientes transformaciones:
Edad: Se convirtió de días a años, dividiendo por 365 y redondeando al entero más próximo.
Altura: Se transformó de centímetros a metros, dividiendo entre 100.
IMC: **(**índice de Masa Corporal**)** : Se creó una nueva variable calculada como:

IMC=peso (kg)/(altura (m))2.
El resultado se redondea a dos decimales, utilizando el punto como separador decimal.
Estas transformaciones se realizaron en el dataset antes de cargar los datos al entorno de análisis.

2.3 Preprocesamiento

En el conjunto de datos original, los nombres de los atributos estaban en inglés y, en muchos casos, se encontraban abreviados, lo que podía generar confusión. Por este motivo, se tradujeron todos los nombres de los atributos al castellano y se eliminaron las abreviaciones, con el fin de mejorar la comprensión de la base de datos.
Se identificó una columna denominada index, que no aportaba información relevante para el análisis, por lo que se decidió eliminar.
La columna correspondiente al identificador de los pacientes (idPaciente) comenzaba en el número 0, se modificó para que comenzara en el número 1, asignando así "Paciente 1" al primer registro.

Detección y eliminación de valores atípicos (*outliers*)
Se aplicaron distintos criterios de exclusión para eliminar valores atípicos o registros con errores de formato, con el objetivo de asegurar la calidad de los datos. Los criterios aplicados fueron:

1. Eliminar pacientes con altura superior a 2 metros o inferior a 1,40 metros.
2. Eliminar pacientes con peso superior a 120 kg o inferior a 40 kg.
3. Eliminar pacientes con presión arterial sistólica superior a 220 mmHg o inferior a 80 mmHg.
4. Eliminar pacientes con presión arterial diastólica superior a 120 mmHg o inferior a 50 mmHg
5. Eliminar pacientes con IMC superior a 50 .

Tras aplicar estos filtros, el conjunto de datos final quedó compuesto por 69.126 pacientes, lo que supuso la eliminación de 874 registros de los 70.000 originales. Posteriormente, se excluyeron los 16.000 registros utilizados previamente para el entrenamiento del modelo predictivo, con el fin de preservar la independencia del análisis y evitar sesgos. Asimismo, se eliminó la variable *ID* al tratarse de un identificador sin valor explicativo. También se descartaron las variables peso y altura, dado que su información ya se encontraba representada en la nueva variable calculada de índice de masa corporal (IMC), clínicamente más significativa. Tras estas modificaciones, el conjunto de datos definitivo quedó compuesto por 53.126 registros con las variables seleccionadas para el análisis predictivo.

3 Análisis de los datos

En este apartado se describen los procesos metodológicos y experimentos computacionales llevados a cabo para el desarrollo del estudio orientado a la predicción del riesgo cardiovascular. El enfoque adoptado combina técnicas de análisis exploratorio, algoritmos de clasificación y métodos de agrupamiento no supervisado, con el objetivo de identificar patrones significativos en los datos clínicos y establecer perfiles de riesgo diferenciados.

3.1 Analisis de clustering

Este estudio de clustering fue desarrollado utilizando código Python , implementado en el entorno colaborativo Google Colab, lo que permitió la ejecución eficiente del código.

Se realizó un primer análisis exploratorio aplicando técnicas de clustering sobre tres subconjuntos de datos secuenciales, de 1K, 5K y 10K registros, respectivamente. En todos los casos se utilizó el algoritmo K-Means, y el método del codo indicó de forma consistente que la segmentación óptima se encontraba en tres clústeres (k = 3), lo que sugiere una estructura interna sólida y bien definida en los datos.

Con el objetivo de verificar la validez de estos resultados y descartar sesgos (bias) derivados del orden secuencial, se repitió el análisis utilizando subconjuntos de 1K, 5K y 10K registros seleccionados aleatoriamente, mediante muestreo aleatorio simple. Los resultados obtenidos coincidieron en todos los casos con los análisis previos, manteniéndose k = 3 como la opción óptima para la consistencia estructural de los datos.

Como prueba adicional se realizó una segmentación alternativa con cuatro clústeres (k = 4), pero esta opción generó un grupo adicional compuesto por pacientes con valores extremos en todas las variables (edad, IMC, presión, glucosa y colesterol), sin aportar información clínicamente útil o interpretable, por lo que esta opción fue descartada.

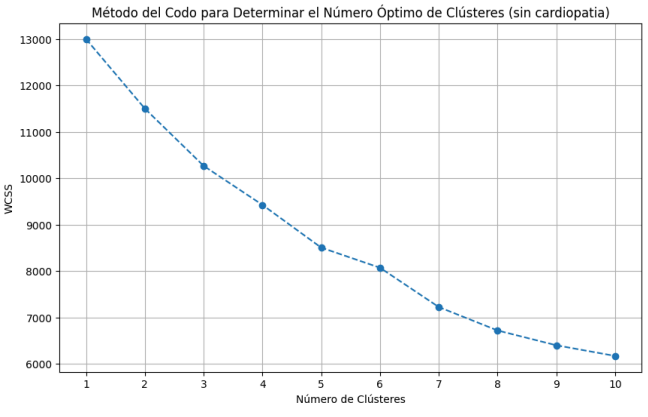

Figura 1. Método del codo para determinar el número óptimo de clústeres en el análisis K-Means, aplicado a pacientes sin diagnóstico de cardiopatía. La curva representa la inercia intra-clúster en función del número de clústeres. El punto de inflexión en *k = 3* sugiere que tres grupos representan la estructura más adecuada del conjunto de datos.

3.2 Análisis de Random Forest

El modelo de Random Forest fue evaluado sobre tres subconjuntos diferenciados de datos de 1K, 5K y 10K de registros. En cada uno de estos grupos se realizaron pruebas con dos particiones de datos distintas: 80 % para entrenamiento y 20 % para test, y 70 /30%, respectivamente. El objetivo principal de este enfoque fue analizar la estabilidad y robustez del modelo en función del tamaño muestral, observando Para llevar a cabo el análisis predictivo mediante el algoritmo Random Forest, se utilizó la plataforma bigML [7], orientada a la aplicación de modelos de aprendizaje automático sin necesidad de programación avanzada. El conjunto de datos clínicos fue previamente preparado en Microsoft Excel y cargado en la plataforma en formato CSV. Una vez importado, bigML aplicó automáticamente un preprocesamiento inicial, incluyendo la codificación de cómo evoluciona el rendimiento del algoritmo a medida que se incrementa el volumen de datos. Esta estrategia permitió evaluar si el aumento del conjunto de entrenamiento tenía un efecto positivo o negativo sobre la capacidad predictiva del modelo, sin anticipar resultados, pero estableciendo una base metodológica sólida y comparable .

Posteriormente, se replicó el análisis sobre esos mismos tres subconjuntos (1K, 5K y 10K), eliminando previamente las variables idPaciente, peso y altura, ya que la información de estas últimas quedaba integrada en la variable derivada IMC. Estas pruebas también se realizaron bajo las configuraciones de 80 /20 % y 70 /30 %, lo que permitió comprobar si la supresión de variables no informativas o redundantes mejoraba el rendimiento del modelo.

Finalmente, tras utilizar los subconjuntos anteriores para entrenamiento (16.000 registros en total), se descartaron esos datos del conjunto original, obteniéndose así un nuevo dataset final compuesto por 53.126 registros. Sobre esta base depurada se repitió el análisis completo con el algoritmo Random Forest, aplicando nuevamente las divisiones 80 /20 % y 70 /30 %. Esta última fase permitió aplicar el modelo a la muestra más amplia disponible, con todas las variables clínicamente relevantes seleccionadas.

3.3 Análisis de Regresión logística

Para complementar el análisis predictivo del riesgo cardiovascular, se implementó un segundo modelo basado en regresión logística. Este modelo, ampliamente utilizado en contextos clínicos por su simplicidad y capacidad interpretativa, fue también desarrollado mediante la plataforma bigML. Al igual que en el caso anterior, el conjunto de datos clínicos fue preparado en Microsoft Excel y exportado en formato CSV, para posteriormente ser cargado y preprocesado automáticamente por la plataforma.

La regresión logística se aplicó sobre los mismos subconjuntos de datos seleccionados 1K, 5K y 10K registros. En cada uno de ellos se probaron dos divisiones distintas de los datos: 80 % para entrenamiento y 20 % para test, y 70 /30 %, con el fin de evaluar la estabilidad del modelo frente a diferentes tamaños de muestra. El objetivo de esta fase fue observar cómo se comportaba el modelo de regresión al aumentar la cantidad de datos disponibles y si ello suponía mejoras o pérdidas en su rendimiento. El enfoque metodológico se mantuvo homogéneo, lo que permite una comparación directa con los otros algoritmos utilizados en el estudio, sin adelantar resultados pero dejando establecida una base sólida para el análisis posterior.

Posteriormente, se repitieron los análisis sobre los mismos tres subconjuntos de 1K, 5K y 10K registros, eliminando previamente las variables idPaciente, peso y altura, por considerarse no predictivas o redundantes, especialmente en el caso de peso y altura, cuya información ya quedaba integrada en la variable derivada IMC. Esta segunda fase de pruebas permitió valorar si la simplificación del modelo mediante la supresión de variables no esenciales impacta en su rendimiento predictivo.

Finalmente, una vez descartados los 16.000 registros utilizados previamente para entrenamiento y pruebas, se trabajó con el conjunto de datos definitivo compuesto por 53.126 registros, sobre el cual se aplicó el modelo de regresión logística utilizando nuevamente las divisiones 80 /20 % y 70 /30 %. Esta etapa permitió realizar un análisis completo sobre el total de datos disponibles, en condiciones más representativas del conjunto poblacional, sentando las bases para la comparación final entre modelos y la extracción de conclusiones.

3.4 Análisis red neuronal profunda

Como tercera técnica de análisis predictivo, se aplicó un modelo de redes neuronales profundas , con el objetivo de explorar el comportamiento de algoritmos más complejos frente a los datos clínicos utilizados en el estudio. Al igual que en los casos anteriores, el desarrollo del modelo se realizó a través de la plataforma bigML, la cual permite configurar modelos de deep learning de forma accesible y sin necesidad de programación directa. Los datos fueron preparados previamente en Microsoft Excel y cargados en la plataforma en formato CSV, siendo sometidos al preprocesamiento automático proporcionado por el entorno.

El modelo fue probado sobre los tres subconjuntos de datos seleccionados aleatoriamente: 1K, 5K y 10K registros. Para cada conjunto, se realizaron pruebas con dos esquemas de partición: 80 % para entrenamiento y 20 % para test, así como 70 /30 %, con el fin de analizar si el rendimiento del modelo mejoraba con un mayor volumen de datos y qué impacto tenía la proporción en el ajuste. Esta fase tuvo un carácter exploratorio, sin anticipar resultados, pero manteniendo una metodología coherente y comparativa con los modelos anteriores, lo que permitirá posteriormente evaluar su eficacia en la

predicción del riesgo cardiovascular.

Finalmente, se realizó el análisis completo sobre el conjunto de datos definitivo, compuesto por 53.126 registros, manteniendo las dos proporciones de partición. Esta última etapa permitió evaluar el comportamiento del modelo con la totalidad de los datos.

4 Resultados

Los resultados se organizan en cuatro partes: primero, un análisis exploratorio mediante *clustering* que validó la estructura y calidad de datos antes de entrenar los modelos predictivos. Segundo, la comparación de tres algoritmos —Random Forest, regresión logística y redes neuronales profundas— mediantes sus curvas de ROC, analizando su capacidad discriminativa. Tercero un examen de las matrices de confusión de cada modelo, que permite valorar su rendimiento en términos de sensibilidad y especificidad . Finalmente se realiza una comparativa de los resultados del peso de los factores de riesgo en la predicción de enfermedad.

4.1 Clustering

La persistencia de k = 3 en todas las pruebas aporta valor clínico añadido, ya que permite clasificar a los pacientes en tres perfiles diferenciados de riesgo cardiovascular: bajo, moderado y alto. Esta segmentación facilita la estratificación del riesgo, optimizando la toma de decisiones clínicas e intervenciones preventivas adaptadas a cada grupo. Así, el análisis de clustering se presenta no sólo como una herramienta exploratoria, sino como un recurso aplicable en contextos clínicos reales, capaz de traducir grandes volúmenes de datos en conocimiento accionable y personalizado.

Variable	Clúster 0	Clúster 1	Clúster 2
Sistólica	127.6	122.28	137.48
Colesterol	1.41	1.07	1.52
Edad	51.22	51.74	54.78
IMC	26.83	25.25	31.82
Cardiopatía	0.48	0.39	0.71

Figura 2. La tabla muestra la distribución de variables clave según los tres clústeres identificados, asociados a riesgo cardiovascular bajo (clúster 1), medio (clúster 0) y alto (clúster 2). Se observan diferencias marcadas entre grupos.

4.2 Rendimiento algoritmos

El estudio de las curvas ROC marca el inicio del análisis comparativo del rendimiento de los modelos predictivos aplicados al conjunto de datos definitivo. Tanto el algoritmo Random Forest como el modelo de Deepnets mostraron un desempeño prácticamente equivalente, alcanzando valores de AUC de 0.8066 y 0.8064, respectivamente. Estos resultados reflejan una notable capacidad discriminativa para predecir el riesgo cardiovascular. Por su parte, la regresión logística obtuvo un valor de AUC de 0.7983, ligeramente inferior, pero aún dentro de un rango alto de precisión. La proximidad en los valores de AUC entre Random Forest y Deepnets evidencia que ambos algoritmos logran identificar patrones clínicamente relevantes con eficacia similar. A pesar de su simplicidad, la regresión logística también presenta un rendimiento competitivo, lo que respalda su utilidad en contextos clínicos. En conjunto, estos resultados refuerzan la validez del uso de técnicas de aprendizaje automático como herramientas complementarias a los métodos tradicionales en la estratificación del riesgo cardiovascular.

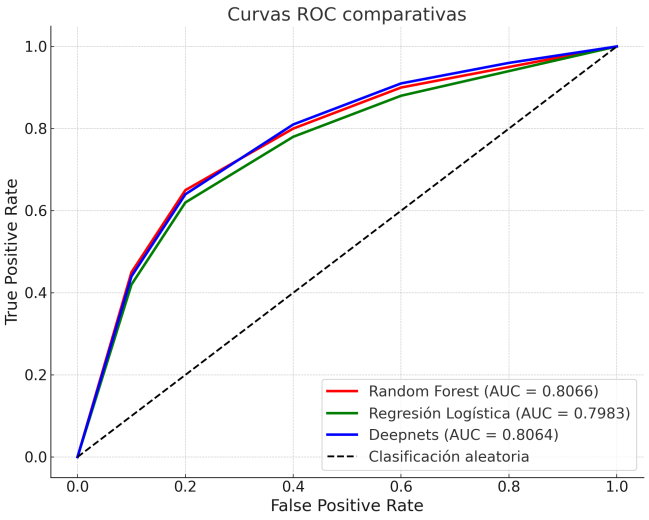

Figura 3 Curvas ROC comparativas obtenidas mediante los algoritmos Random Forest, Regresión Logística y Deepnets . La línea discontinua representa la clasificación aleatoria. Random Forest (AUC = 0.8066) y Deepnets (AUC = 0.8064), y la regresión logística (AUC = 0.7983).

Figura 4. Árbol de decisión derivado de un modelo Random Forest, utilizado para predecir el riesgo de enfermedad cardiovascular. Cada nodo representa una división basada en una variable clínica (como IMC, edad o colesterol), con colores que identifican la característica empleada. El grosor del camino indica la ruta más representativa o frecuente en las decisiones del modelo

4.3 Comparativa matriz confusión

A continuación, se realizó un análisis comparativo con las métricas clave de los tres modelos evaluados: Random Forest, Regresión Logística y Deepnets. Este análisis permite observar de forma comparativa el rendimiento de cada algoritmo en la predicción del riesgo cardiovascular, considerando indicadores como la precisión, el recall, el F1-Score y la exactitud (accuracy).

	Accuracy	F1-Score	Precision	Recall
Deepnets	73.85	0.7522	70.98	79.99
Regresión Logística	73.12	0.7458	70.25	79.48
Random Forest	74.2	0.7501	72.19	78.07

Figura 5. Comparativa de desempeño entre modelos predictivos. La tabla recoge los valores obtenidos para cada métrica de evaluación, permitiendo identificar las fortalezas de cada enfoque. Random Forest destaca en precisión general (74.20 %) y exactitud, mientras que Deepnets ofrece un equilibrio sólido entre F1-score y recall. La regresión logística, aunque más sencilla, mantiene resultados competitivos.

4.4 Peso de los factores de riesgo

Con el objetivo de comprender qué factores influyen con mayor peso en la predicción del riesgo cardiovascular, se analizaron las variables más relevantes según la importancia otorgada por los modelos DeepNet y Random Forest. En ambos casos, la variable presión arterial sistólica fue identificada como el factor más determinante, aunque con diferente peso relativo (39.46 % en DeepNet frente a 24.92 % en Random Forest).

En el modelo DeepNet, la colesterolemia (18.60 %) y la presión diastólica (17.65 %) completan el trío principal, mientras que Random Forest sitúa como variables destacadas la edad (22.4 %) y el IMC (20.65 %). Estas diferencias reflejan cómo cada algoritmo interpreta los patrones de riesgo en función de su arquitectura interna.

El análisis comparativo permite valorar la solidez de las variables clínicas clásicas en el entrenamiento de modelos predictivos, y ofrece una visión complementaria sobre la influencia relativa de cada factor en función del algoritmo utilizado.

Variable	Deepnets (%)	Random Forest (%)
Sistólica	39.46	24.92
Colesterol	18.6	9.94
Diastólica	17.65	9.07
Edad	9.71	22.4
IMC	3.8	20.65
Alcohol	3.23	1.84

Figura 6 Importancia relativa de las variables clínicas más significativas en la predicción del riesgo cardiovascular según los modelos de Deepnets y Random Forest. Se muestran los seis factores más influyentes identificados .

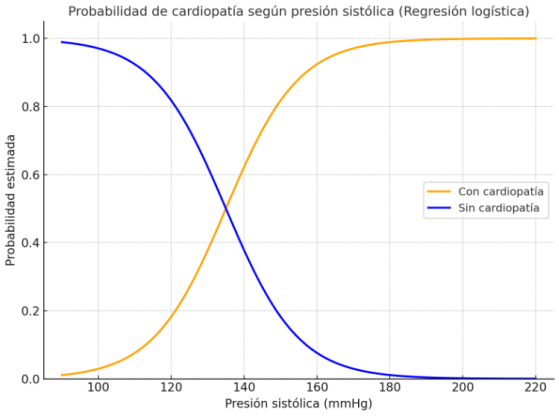

Figura 7. Distribución estimada del riesgo de cardiopatía en función de la presión arterial sistólica según el modelo de regresión logística. Se observa un punto de inflexión en torno a los 135 mmHg, a partir del cual la probabilidad de padecer una cardiopatía aumenta significativamente.

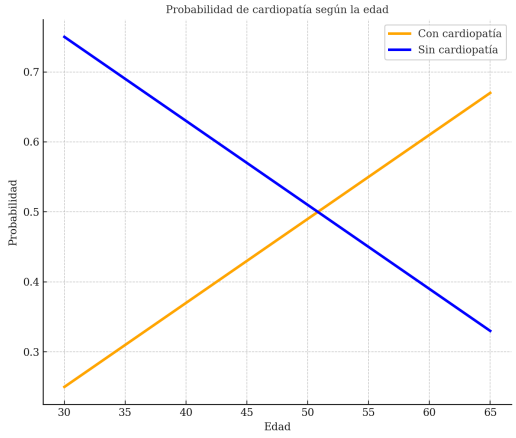

Figura 8. Relación entre edad y probabilidad de sufrir una cardiopatía según el modelo de regresión logística. La probabilidad de padecer una cardiopatía aumenta progresivamente con la edad. La curva azul indica la probabilidad de no presentar cardiopatía, con un comportamiento inverso. Esta visualización permite entender el peso específico de la edad como factor predictivo del riesgo cardiovascular.

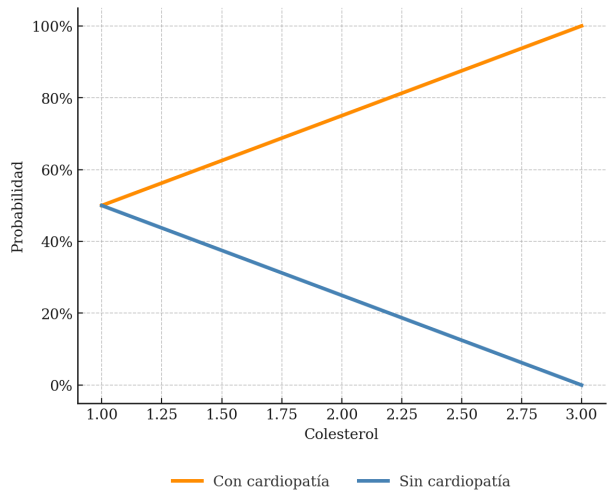

Figura 9. Probabilidad de sufrir una cardiopatía en función del nivel de colesterol, según el modelo de regresión logística. La probabilidad de padecer cardiopatía presenta un aumento progresivo del riesgo conforme se incrementa el nivel de colesterol.

5 Discusión

En la discusión se abordará la comparación de los resultados obtenidos con estudios previos similares. Asimismo, se analizará las limitaciones del estudio.

5.1 Rendimiento estudio mismos datos

El presente estudio coincide con la investigación publicada por Kırboğa y Küçüksille (2023) **[8]** en el uso de la misma base de datos abierta para la predicción del riesgo cardiovascular mediante algoritmos de aprendizaje automático. Sin embargo, existen diferencias metodológicas sustanciales que justifican una comparación crítica de los resultados.

Mientras que el artículo antes mencionado empleó siete modelos sin realizar preprocesamiento exhaustivo —incluyendo Random Forest, XGBoost y regresión logística—, este trabajo ha apostado por una estrategia más rigurosa de limpieza, transformación y selección de datos. Como resultado, los modelos aquí implementados (Random Forest, Deepnets y regresión logística) mostraron métricas superiores, especialmente en el modelo Random Forest, con un AUC de 0.8066 frente al 0.773 reportado por Kırboğa y Küçüksille.

La ausencia de preprocesamiento en el estudio mencionado puede haber limitado el rendimiento de sus modelos, dado que el sesgo por *outliers* o valores mal categorizados puede afectar negativamente la capacidad predictiva. Además, aunque en su caso el XGBoost ofreció el mejor rendimiento (AUC de 0.803), la mejora obtenida aquí mediante Random Forest sugiere que un tratamiento adecuado de los datos puede igualar o incluso superar modelos más complejos.

Finalmente, ambos estudios coinciden en señalar como factores clave la presión sistólica, el colesterol y la edad.

5.2 limitaciones estudio

En primer lugar, si bien se aplicó un riguroso proceso de preprocesamiento y limpieza de datos, la base de datos utilizada es de naturaleza secundaria y abierta, por lo que no se tiene control sobre el proceso de recogida de los datos originales ni sobre posibles sesgos de muestreo.

En segundo lugar, aunque se exploraron diferentes tamaños muestrales y modelos (Random Forest, regresión logística y Deepnets), no se incorporaron otros algoritmos potencialmente competitivos como XGBoost o LightGBM, que podrían haber aportado un rendimiento superior. Del mismo modo, por razones de tiempo y alcance del trabajo, no se realizó un ajuste hiperparamétrico avanzado (tuning) de los modelos, lo cual podría mejorar significativamente los resultados.

Otra limitación importante es que las variables disponibles se centran únicamente en factores clínicos y de estilo de vida, sin incorporar datos genéticos, socioeconómicos o hábitos dietéticos que también influyen en el riesgo cardiovascular. Finalmente, si bien los modelos mostraron buenas métricas de rendimiento, su validación se hizo únicamente con datos internos y no con una cohorte externa, lo que limita la generalización de los resultados a otras poblaciones o contextos clínicos reales.

6 Conclusión

El presente estudio cumple con los objetivos propuestos y permite determinar el riesgo cardiovascular utilizando diferentes modelos de IA.

Se considera Random Forest como el modelo más robusto de todos, aunque el análisis con la red neuronal alcanza valores muy similares de AUC (0.8066 y 0.8064, respectivamente),Random Forest ofrece mayor interpretabilidad y transparencia en sus decisiones, permitiendo conocer cómo cada variable contribuye al resultado final. La coherencia entre los modelos y la validez clínica de las variables más influyentes respaldan la fiabilidad del análisis.

El uso de técnicas de limpieza, normalización y segmentación de los datos contribuyó significativamente a mejorar los resultados respecto a otros estudios similares[8]. Los resultados obtenidos posicionan este trabajo como una contribución sólida y replicable.

.En resumen, y basándome en mi larga experiencia laboral en Atención primaria considero que la la integración de IA en la estratificación del riesgo cardiovascular no sólo es viable, sino que mejora la precisión diagnóstica en comparación con métodos tradicionales, abriendo la puerta a aplicaciones clínicas más personalizadas a partir de variables clínicas habituales, optimizando la toma de decisiones en la consulta.

7 Lineas abiertas

A partir de los resultados de este estudio, se abren diversas posibilidades para investigaciones futuras que permitan ampliar y perfeccionar el uso de inteligencia artificial en la predicción del riesgo cardiovascular.

Aunque este estudio ha evaluado diversos conjuntos de datos y considerado el riesgo de sobreajuste en los modelos,futuras investigaciones deberìan profundizar en este aspecto mediantes técnicas de validación más robustas, como *K-fold cross-validation* u otros métodos avanzados, para garantizar una mayor generalización de los resultados.

Asimismo, la incorporación de datos longitudinales permitiría analizar la evolución temporal del riesgo y construir modelos de predicción más dinámicos. También resultaría valioso trabajar con bases de datos más amplias, que incluyan hábitos alimenticios, antecedentes familiares, tratamiento medicamentoso, u

otras condiciones clínicas, con el fin de aumentar la validez de los resultados.

Por último, se podría estudiar la integración de estas herramientas en sistemas clínicos reales, evaluando su impacto sobre la toma de decisiones, la eficiencia del diagnóstico precoz y la reducción de la morbimortalidad en la población.

8 Bibliografia

[1] *World Health Organization (WHO). (2023). Cardiovascular diseases (CVDs): Fact sheet.*

[2] *Fernández-San-Martín, M. I., et al. (2014). The effectiveness of lifestyle interventions to reduce cardiovascular risk in patients with severe mental disorders: Meta-analysis of intervention studies. Community Mental Health Journal, 50(1), 81–95*

[3] *Gómez-García, F., et al. (2022). Primary and secondary cardiovascular disease prevention through lifestyle interventions: A systematic review and meta-analysis. Frontiers in Cardiovascular Medicine, 9, 1010528*

[4] *D'Agostino, R. B., Vasan, R. S., Pencina, M. J., Wolf, P. A., Cobain, M., Massaro, J. M., & Kannel, W. B. (2008). General cardiovascular risk profile for use in primary care: The Framingham Heart Study. Circulation, 117(6), 743–753. https://doi.org/10.1161/CIRCULATIONAHA.107.699579*

[5] *Piepoli, M. F., Hoes, A. W., Agewall, S., Albus, C., Brotons, C., Catapano, A. L., … & ESC Scientific Document Group. (2016). 2016 European Guidelines on cardiovascular disease prevention in clinical practice. European Heart Journal, 37(29), 2315–2381. https://doi.org/10.1093/eurheartj/ehw106*

[6] *A. Prajapati et al., "Heart Disease Prediction Using Various Machine Learning Algorithms," SSRN, 2022. DOI: 10.2139/ssrn.4117242..*

[7] https://bigml.com/

[8] *Yildiz, B. S., Cinar, A., Cetin, M., Karaca, O., & Altun, I. et al. (2023). Identifying cardiovascular disease risk factors in adults with explainable artificial intelligence. Computers in Biology and Medicine, 168, 107632. https://doi.org/10.1016/j.compbiomed.2023.107632*

[9] *V. K. Yarasuri, R. P. Kumar, S. V. Babu y T. V. Vijaylaxmi, "Developing Machine Learning Models for Cardiovascular Disease Prediction," en Proc. 2022 2nd Asian Conference on Innovation in Technology (ASIANCON), Bengaluru, India, 2022, pp. xx–yy, doi: 10.1109/ASIANCON55314.2022.9908772*An Empirical Study of Machine Learning (ML) Algorithms in the Perspective of Cardiovascular Disease (CVD) Prediction..

[10] *V. R. Camm, L. Delgado y P. C. Lee, "Machine Learning Model Predicting the Likelihood of a Patient Developing Cardiovascular Disease Based on Their Medical History and Risk Factors," Am J Biomed Sci & Res, vol. 18, no. 1, pp. 1–7, 2023, doi: 10.34297/ajbsr.2023.18.002429.*

[11] T. Y. Wong, X. Y. Cheung, H. H. Tham, et al., "Artificial intelligence in predictive cardiovascular risk modelling across multi-ethnic populations: A systematic review," *Lancet Digit Health*, vol. 5, no. 7, pp. e404–e416, 2023, doi: 10.1016/S2589-7500(23)00094-1.

[12] *S. Castel-Feced, S. Malo, I. Aguilar-Palacio, C. Feja-Solana, J. A.*

Casasnovas, L. Maldonado y M. J. Rabañaque-Hernández, "Influence of cardiovascular risk factors and treatment exposure on cardiovascular event incidence: Assessment using machine learning algorithms," PLoS One, vol. 18, no. 11, Art. e0293759, Nov. 2023, doi: 10.1371/journal.pone.0293759.

[13] L. Marsh, D. Khan, M. Zhang, et al., "Artificial Intelligence in Cardiovascular Clinical Trials: Applications and Ethical Considerations," J Am Coll Cardiol, vol. 84, no. 20, pp. 2051–2062, Nov. 2024, doi: 10.1016/j.jacc.2024.08.069.

[14] T. M. Oudejans, M. A. Smits, J. J. P. Kastelein, et al., "Risk Factor Clusters and Cardiovascular Disease in High-Risk Patients: The UCC-SMART Study," BMJ Open, vol. 11, no. 8, e058400, 2021.

[15] S. Kask-Flight, K. Durak, K. Suija, A. Rätsep y R. Kalda, "Reduction of cardiovascular risk factors among young men with hypertension using an interactive decision aid: cluster-randomized controlled trial," BMC Cardiovasc Disord., vol. 21, no. 1, Art. 543, 2021, doi: 10.1186/s12872-021-02339-1.

[16] C. M. Petersein, K. Jabbour, A. Kamath Mulki, et al., "Metabolic Syndrome: An Updated Review on Diagnosis and Treatment for Primary Care Clinicians," J Prim Care Community Health, vol. 15, 2024 Jan–Dec, Article 21501319241309168,