Predicción de enfermedades cardiovasculares mediante aprendizaje automático

Juan Pablo Valencia Martínez y Robert Alexander Blandón Rincón
Universidad de Antioquia
Modelos y Simulación II
Grupo 25

Resumen

El proyecto busca desarrollar un sistema de predicción eficiente y robusto para enfermedades cardiovasculares (ECV), que son la principal causa de muerte global, ya que los métodos de diagnóstico tradicionales son costosos y limitados. Utilizando Aprendizaje Automático (ML), el problema se aborda como una tarea de clasificación supervisada binaria sobre el Cardiovascular Disease Dataset de Kaggle, que contiene datos de 70.000 pacientes y 12 variables clínicas y de hábitos.

El objetivo es construir un modelo predictivo que determine el riesgo de ECV. La revisión del Estado del Arte confirma la utilidad de los enfoques de ML, como Random Forest y Redes Neuronales, los cuales han alcanzado altos niveles de precisión (hasta 84%) en la predicción de ECV, sentando las bases para la implementación y comparación de modelos en este trabajo.

Palabras clave: enfermedades cardiovasculares, aprendizaje automático, predicción, random forest, SVM, redes neuronales.

I. Introducción

Las enfermedades cardiovasculares o ECV se encuentran entre las mayores causas mundiales de mortalidad con un 32% de las muertes aproximadamente según la Organización Mundial de la Salud en el 2023. Esto es algo alto, por lo cual la detección temprana y un buen diagnóstico reducen las complicaciones y mejoran la calidad de vida de las personas.

Sin embargo, estos procedimientos médicos son costosos, invasivos o tardíos, lo cual limita el acceso en ciertos contextos. Haciendo uso del aprendizaje automático dentro del ámbito del Machine Learning, hay herramientas efectivas para analizar grandes volúmenes de datos médicos y generar modelos que puedan predecir patrones asociados a diferentes enfermedades.

Este trabajo aborda un problema de predicción de enfermedades cardiovasculares utilizando una base de datos pública (Cardiovascular Disease Dataset) de Kaggle, implementando y comparando diferentes modelos de aprendizaje supervisado para evaluar la capacidad de clasificación e identificar las variables más relevantes.

El objetivo principal de este proyecto es desarrollar un sistema predictivo eficiente y robusto que contribuya al diagnóstico temprano de las ECV.

II. Descripción del problema

El problema consiste en la construcción de un modelo predictivo capaz de determinar si una persona presenta riesgo de enfermedades cardiovasculares a partir de un conjunto de variables médicas.

A. Composición y limpieza de datos

La base de datos original contiene 70,000 pacientes con 12 variables relevantes. El preprocesamiento incluyó:

Limpieza clínica: Se aplicaron filtros basados en rangos clínicamente posibles:

- Edad: 18-100 años (6,570-36,500 días)
- Altura: 100-250 cm
- Peso: 30-200 kg
- Presión arterial: valores válidos y relación sistólica > diastólica

Balanceo de clases: Mediante submuestreo estratificado se obtuvo un dataset balanceado de 20,000 muestras (10,000 por clase).

Análisis de correlaciones: Todas las variables mostraron correlaciones significativas con la variable objetivo (> 0.01), por lo que se conservaron para el modelado.

B. Variables y codificación

Las variables incluyen:

- **Numéricas**: edad, altura, peso, presión sistólica y diastólica
- **Categóricas**: colesterol (1: normal, 2: alto, 3: muy alto), glucosa (1: normal, 2: alto, 3: muy alto)
- **Binary**: fumar, alcohol, actividad física (0: No, 1: Sí)

C. Enfoque de aprendizaje

Desde el punto de vista de aprendizaje automático, es una tarea de clasificación supervisada binaria, donde el objetivo es predecir la presencia (1) o ausencia (0) de enfermedad cardiovascular.

III. Estado del arte

Existen diversos estudios recientes que han explorado el uso de algoritmos de aprendizaje automático y profundo para la detección y predicción de enfermedades cardiovasculares. A continuación se resumen algunos de los trabajos más relevantes que emplean enfoques similares al propuesto en este proyecto:

A. Trabajos relacionados

Boix (2025) desarrolló un sistema de predicción basado en modelos de Random Forest y Support Vector Machines (SVM) aplicados al mismo conjunto de datos de Kaggle. El trabajo enfatiza la importancia del preprocesamiento y el balance de clases mediante SMOTE. La validación se realizó con cross-validation de 10 particiones, logrando un desempeño máximo del 73% con Random Forest.

Henao, Gómez y Rojas (2023) propusieron un modelo híbrido de clasificación de riesgo cardiovascular mediante técnicas de selección de características y optimización de hiperparámetros, con Gradient Boosting y K-Nearest Neighbors. La validación cruzada mostró una precisión promedio del 78%, evidenciando la utilidad de los modelos de ensamble en problemas médicos.

Murillo (2022) exploró redes neuronales artificiales, destacando la importancia de la normalización y el ajuste de arquitectura neuronal. El autor alcanzó un F1-score de 0.81, superior a los modelos de regresión logística.

Saridena, Kethar y Saridena (2024) propusieron un modelo de aprendizaje profundo basado en CNN para detección automática de ECV, alcanzando una precisión del 84% y un AUC de 0.90. Este enfoque demuestra la capacidad de las arquitecturas profundas para capturar representaciones complejas de datos clínicos estructurados.

B. Análisis comparativo

Los trabajos revisados emplean diferentes configuraciones y paradigmas de aprendizaje:

- **Configuración**: Aprendizaje supervisado para clasificación binaria
- **Técnicas**: Random Forest, SVM, Gradient Boosting, Redes Neuronales, CNN
- **Validación**: Cross-validation de 5 a 10 particiones
- **Métricas**: Precisión, F1-score, AUC-ROC

Los resultados reportados en la literatura oscilan entre 73% y 84% de precisión, estableciendo una linea de referencia para el presente trabajo.

IV. Entrenamiento y Evaluación de los Modelos

A. Configuración experimental

Para el entrenamiento y evaluación de los modelos, se implementó una metodología de división estratificada de los datos, reservando el 20% para prueba y el 80% para entrenamiento. Sobre el conjunto de entrenamiento se aplicó validación cruzada estratificada de 5 particiones ($k = 5$) para la selección de hiperparámetros mediante Grid Search, utilizando F1-Score macro como métrica de optimización.

El preprocesamiento incluyó estandarización de las variables numéricas (`age`, `height`, `weight`, `ap_hi`, `ap_lo`) mediante `StandardScaler`. Se evaluaron cinco modelos de aprendizaje automático:

- **Modelo paramétrico**: Regresión Logística
- **Modelo no paramétrico**: K-Vecinos Más Cercanos
- **Ensemble de árboles**: Random Forest
- **Red neuronal**: Perceptrón Multicapa
- **Máquina de vectores de soporte**: SVM

Los hiperparámetros evaluados para cada modelo fueron:

Regresión Logística: C [0.1, 1, 10], solver [liblinear, lbfgs], class_weight [balanced, None], max_iter [1000]

K-Nearest Neighbors: n_neighbors [3, 5, 7, 9], weights [uniform, distance], metric [euclidean, manhattan]

Random Forest: n_estimators [100, 200], max_depth [10, 20, None], min_samples_split [2, 5], class_weight [balanced, None]

Support Vector Machine: C [0.1, 1, 10], kernel [linear, rbf], class_weight [balanced, None]

Neural Network: hidden_layer_sizes [(50,), (100,)], activation [relu], alpha [0.0001], learning_rate [constant], max_iter [300]

Las métricas de evaluación incluyeron: **Accuracy**, **F1-Score macro** y **AUC-ROC**.

B. Resultados del entrenamiento de Modelos

La Tabla I presenta los resultados en el conjunto de prueba.

TABLE I
RESULTADOS DE LOS MODELOS EN CONJUNTO DE PRUEBA

Modelo	Accuracy	F1-Score	AUC-ROC	Tiempo (s)
Random Forest	0.7385	0.7381	0.7998	300.54
SVM	0.7370	0.7366	0.7907	932.60
Red Neuronal	0.7318	0.7317	0.7971	11.15
Reg. Logística	0.7292	0.7284	0.7932	4.74
K-NN	0.7110	0.7109	0.7679	114.17

El **Random Forest** obtuvo el mejor desempeño global (F1: 0.7381, AUC: 0.7998), seguido por **SVM** (F1: 0.7366). Los mejores hiperparámetros encontrados fueron:

- **Random Forest**: n_estimators=200, max_depth=10, min_samples_split=5, class_weight='balanced'
- **SVM**: C=1, kernel='rbf', class_weight='balanced'
- **Red Neuronal**: hidden_layer_sizes=(100,), activation='relu', alpha=0.0001
- **Reg. Logística**: C=1, solver='lbfgs', class_weight='balanced'
- **K-NN**: n_neighbors=9, weights='uniform', metric='manhattan'

La Tabla II muestra el análisis de sobreajuste.

TABLE II
ANÁLISIS DE SOBREAJUSTE: F1-SCORE CV VS TEST

Modelo	F1 CV	F1 Test	Diferencia
Reg. Logística	0.7284	0.7284	0.0000
K-NN	0.7049	0.7109	-0.0059
Random Forest	0.7315	0.7381	-0.0065
SVM	0.7318	0.7366	-0.0048
Red Neuronal	0.7308	0.7317	-0.0009

Todos los modelos mostraron estabilidad (diferencias < 0.007). En eficiencia, **Regresión Logística** fue el más rápido (4.74s) y **SVM** el más lento (932.60s). La **Red Neuronal** mostró mejor equilibrio rendimiento-eficiencia (11.15s).

La Figura 1 muestra la comparación visual del desempeño.

Los dos mejores modelos (**Random Forest** y **SVM**) se utilizarán en la fase de reducción de dimensionalidad.

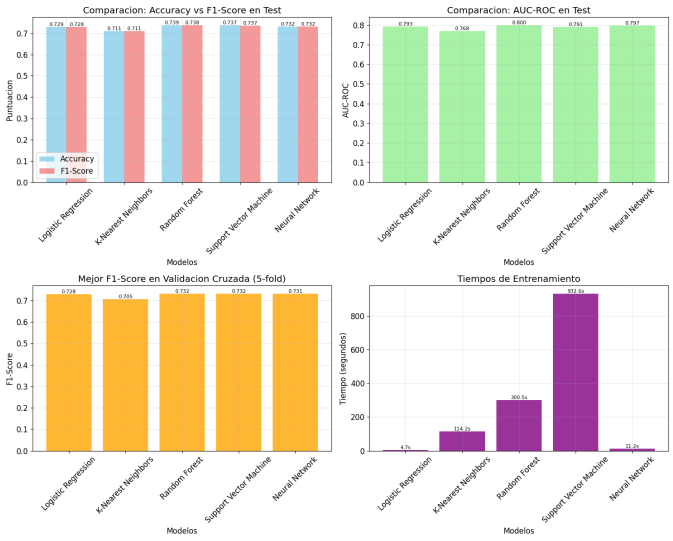

Fig. 1. Comparación del desempeño de modelos

V. Reducción de Dimensión

A. Análisis individual de variables

El análisis de correlación con la variable objetivo reveló que todas las variables presentaban correlaciones significativas superiores a 0.01, por lo que no se identificaron variables candidatas a eliminación. Las variables con mayor correlación fueron presión arterial (ap_hi, ap_lo) y edad, mientras que las variables de hábitos (smoke, alco) mostraron correlaciones menores pero aún significativas.

B. Extracción de características lineal

Se aplicó Análisis de Componentes Principales (PCA) para evaluar la reducción lineal de dimensionalidad. El análisis de varianza explicada mostró que se requieren 7 componentes principales para capturar el 90% de la varianza original (Tabla III).

TABLE III
VARIANZA EXPLICADA POR COMPONENTES PRINCIPALES

Componente	Varianza Individual	Varianza Acumulada
1	0.318	0.318
2	0.205	0.523
3	0.144	0.667
4	0.109	0.776
5	0.081	0.857
6	0.041	0.898
7	0.032	0.931
8	0.026	0.957
9	0.025	0.982
10	0.012	0.994
11	0.006	1.000

El criterio seleccionado fue 7 componentes, logrando una reducción del 36.4% en la dimensionalidad (de 11 a 7 variables). Los resultados de evaluar los mejores modelos con PCA se presentan en la Tabla IV.

Se observa una ligera disminución en el desempeño con PCA: Random Forest perdió 0.0050 puntos en F1-Score, mientras que SVM mantuvo un desempeño muy similar con apenas 0.0009 puntos de diferencia.

TABLE IV
RESULTADOS DE MODELOS CON REDUCCIÓN PCA

Modelo	Técnica	Accuracy	F1-Score	AUC-ROC
Random Forest	Original	0.7385	0.7381	0.7998
Random Forest	PCA	0.7332	0.7331	0.7924
SVM	Original	0.7370	0.7366	0.7907
SVM	PCA	0.7362	0.7357	-

C. Extracción de características no lineal

Se evaluó UMAP (Uniform Manifold Approximation and Projection) como técnica de reducción no lineal, utilizando el mismo número de componentes (7) para comparación equitativa. Los resultados (Tabla V) muestran un impacto más significativo en el desempeño.

TABLE V
RESULTADOS DE MODELOS CON REDUCCIÓN UMAP

Modelo	Técnica	Accuracy	F1-Score	AUC-ROC
Random Forest	Original	0.7385	0.7381	0.7998
Random Forest	UMAP	0.7255	0.7253	0.7864
SVM	Original	0.7370	0.7366	0.7907
SVM	UMAP	0.7145	0.7138	-

UMAP resultó en mayores pérdidas de rendimiento: Random Forest disminuyó 0.0128 puntos en F1-Score y SVM 0.0227 puntos. La Figura 2 muestra la comparación visual entre técnicas.

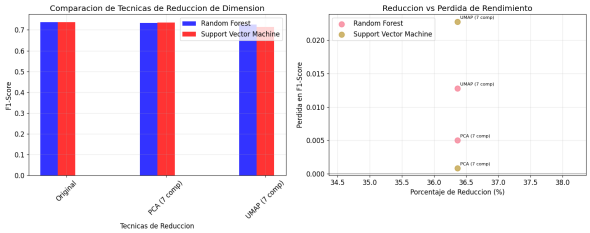

Fig. 2. Comparación de técnicas de reducción de dimensionalidad

D. Discusión y conclusiones

La evaluación comparativa (Tabla VI) revela que ninguna técnica de reducción mejoró el desempeño de los modelos originales.

TABLE VI
COMPARATIVA FINAL DE TÉCNICAS DE REDUCCIÓN

Técnica	Reducción	Mejor F1	Pérdida vs Original
Original	0%	0.7381	-
PCA (7 comp)	36.4%	0.7357	-0.0024
UMAP (7 comp)	36.4%	0.7253	-0.0128

Las posibles razones para estos resultados incluyen:

- Las 11 variables originales contienen información discriminativa relevante que se pierde en la reducción
- El dataset no presenta alta redundancia entre variables

La recomendación final es **mantener todas las variables originales**, ya que la reducción dimensional no proporciona beneficios en términos de rendimiento predictivo, a pesar de lograr una reducción del 36.4% en la dimensionalidad.

En comparación con el estado del arte, nuestros resultados con todas las variables (F1: 0.7381) son competitivos con los reportados en la literatura, que oscilan entre 0.73 y 0.84, validando la efectividad del enfoque propuesto.

VI. Conclusiones

El desarrollo de este proyecto permitió implementar y evaluar un sistema de predicción de enfermedades cardiovasculares mediante técnicas de aprendizaje automático, alcanzando los objetivos planteados y obteniendo resultados competitivos con el estado del arte.

Los principales hallazgos y contribuciones de este trabajo son:

- **Desempeño predictivo robusto**: El modelo de Random Forest obtuvo el mejor desempeño global con un F1-Score de 0.7381 y AUC-ROC de 0.7998, demostrando una capacidad discriminativa significativa para la predicción de ECV.
- **Comparación exhaustiva de modelos**: Se evaluaron cinco familias de algoritmos, identificando que los modelos de ensemble (Random Forest) y máquinas de vectores de soporte (SVM) superan a metodos tradicionales como regresión logística y K-NN en este problema específico.
- **Análisis de reducción dimensional**: La evaluación de técnicas de reducción (PCA y UMAP) reveló que mantener todas las variables originales produce el mejor rendimiento, despite de lograr una reducción del 36.4% en dimensionalidad. Esto sugiere que la información contenida en las 11 variables clínicas es importante y no redundante.
- **Metodología robusta**: La implementación de validación cruzada estratificada, grid search para optimización de hiperparámetros y evaluación con múltiples métricas aseguró la validez y reproducibilidad de los resultados.
- **Contribución al diagnóstico temprano**: El sistema desarrollado alcanza un nivel de precisión competitivo con métodos reportados en el estado del arte, ofreciendo una herramienta potencial para apoyo al diagnóstico médico no invasivo.

Las limitaciones del estudio incluyen la dependencia de un único dataset, aunque de tamaño considerable (70,000 pacientes), y la no inclusión de variables clínicas adicionales como historial familiar o marcadores bioquímicos específicos.

En conclusión, este proyecto demuestra la viabilidad de utilizar aprendizaje automático para la predicción de enfermedades cardiovasculares, con resultados que se alinean con el estado del arte y que podrían integrarse en sistemas de apoyo a la decisión clínica para mejorar la detección temprana de esta condición de salud globalmente relevante.

VII. REFERENCIAS

[1] Kaggle, *Cardiovascular Disease Dataset*, 2023. Disponible en: https://www.kaggle.com/datasets/sulianova/cardiovascular-disease-dataset

[2] C. Boix, *Predicción de enfermedades cardiovasculares mediante aprendizaje automático*, Trabajo de Fin de Grado, Universitat Autònoma de Barcelona, España, 2025. Disponible en: https://ddd.uab.cat/pub/trerecpro/2025/318701/Boix_Carolina_TFG.pdf

[3] J. Henao-Ruiz, L. Gómez-Zapata, y C. Rojas-Londoño, "Modelo híbrido para la estimación del riesgo cardiovascular mediante técnicas de Machine Learning," *Ingeniería y Competitividad*, vol. 25, no. 2, Universidad del Valle, 2023. Disponible en: https://revistaingenieria.univalle.edu.co/index.php/ingenieria_y_competitiv

[4] C. A. Murillo Vivanco, *Diseño de un modelo para la predicción de ataques cardíacos mediante técnicas de aprendizaje automático*, Trabajo de Titulación, Universidad Técnica de Machala, Ecuador, 2022. Disponible en: https://repositorio.utmachala.edu.ec/bitstream/48000/19919/1/TTFIC-2022-IS-DE00046.pdf

[5] A. Saridena, A. Saridena, y J. Kethar, *Cardiovascular Disease Detection using Deep Learning*, ResearchGate, 2024. Disponible en: https://www.researchgate.net/publication/378272534_A_Supervised_Deep