PROYECTO FINAL

APRENDIZAJE AUTOMÁTICO

VALIDACIÓN DE MODELO DE PREDICCIÓN DE INFARTO AGUDO DE MIOCARDIO MEDIANTE APRENDIZAJE AUTOMÁTICO

José Vicente ALZATE GUERRERO (Cód. 22502201)

Soren Fabricius ACEVEDO (Cód.)

Roberto Carlos TIERNO (Cód. 22500842)

UNIVERSIDAD AUTÓNOMA DE OCCIDENTE

FACULTAD DE INGENIERÍA Y CIENCIAS BÁSICAS

MAESTRÍA EN INTELIGENCIA ARTIFICIAL Y CIENCIA DE DATOS

SANTIAGO DE CALI

2025

1. INTRODUCCIÓN

El infarto agudo de miocardio constituye una emergencia médica que resulta de la oclusión súbita de una arteria coronaria, provocando necrosis del músculo cardíaco por isquemia prolongada. Según datos de la Organización Mundial de la Salud, las enfermedades cardiovasculares causan aproximadamente 17.9 millones de muertes anuales, representando el 31% de todas las defunciones globales. En este contexto, el IAM representa una proporción significativa de estos eventos, con una mortalidad que puede reducirse sustancialmente mediante detección temprana e intervención oportuna.

La medicina predictiva ha experimentado un desarrollo exponencial con la implementación de algoritmos de aprendizaje automático en el ámbito cardiovascular. Diversos estudios han reportado modelos predictivos para IAM con precisiones superiores al 90%, promoviendo expectativas sobre su implementación clínica. Sin embargo, una revisión crítica de la literatura revela una problemática fundamental: la mayoría de estos modelos incorporan variables que se obtienen posterior al evento cardíaco, incluyendo biomarcadores de daño miocárdico (troponinas, CK-MB), alteraciones electrocardiográficas específicas del IAM, y parámetros hemodinámicos alterados por el evento coronario.

Esta limitación metodológica compromete significativamente la utilidad clínica de dichos modelos, ya que la verdadera predicción requiere la identificación del riesgo antes de que ocurra el evento. Los algoritmos que utilizan variables post-evento funcionan más como herramientas de confirmación diagnóstica que como instrumentos de predicción temprana, generando una discrepancia entre la precisión reportada y la aplicabilidad clínica real.

La literatura científica actual presenta múltiples aproximaciones al problema de predicción de IAM mediante aprendizaje automático. Los modelos tradicionales de evaluación de riesgo cardiovascular, como el Framingham Risk Score [7][8], las Ecuaciones de Cohorte Agrupada (PCE) [9] y el SCORE2 [10], han establecido las bases para la estratificación de riesgo utilizando variables clásicas: edad, sexo, presión arterial sistólica, colesterol total, colesterol HDL, diabetes, tabaquismo y tratamiento antihipertensivo.

Investigaciones recientes han explorado la integración de técnicas de aprendizaje automático con estos factores de riesgo tradicionales, demostrando mejoras en la capacidad predictiva [11][12]. Estudios utilizando Random Forest, Support Vector Machines y redes neuronales artificiales han reportado áreas bajo la curva ROC entre 0.75 y 0.95 en diferentes poblaciones [13][14]. Sin embargo, la heterogeneidad metodológica y la inclusión frecuente de variables post-evento limitan la comparabilidad y aplicabilidad de estos resultados.

La base de datos MIMIC-IV ha sido extensamente utilizada para el desarrollo de modelos predictivos de mortalidad a 30 días en pacientes con IAM [15][16], mientras que estudios basados en registros electrónicos de salud han explorado la predicción de eventos a 6 meses utilizando hasta 52,000 variables clínicas [17]. No obstante, persiste la necesidad de modelos que utilicen exclusivamente variables pre-evento y que hayan sido validados en poblaciones diversas.

2. ESTADO DEL ARTE

La literatura científica actual presenta múltiples aproximaciones al problema de predicción de IAM mediante aprendizaje automático. Los modelos tradicionales de evaluación de

riesgo cardiovascular, como el Framingham Risk Score [7][8], las Ecuaciones de Cohorte Agrupada (PCE) [9] y el SCORE2 [10], han establecido las bases para la estratificación de riesgo utilizando variables clásicas: edad, sexo, presión arterial sistólica, colesterol total, colesterol HDL, diabetes, tabaquismo y tratamiento antihipertensivo.

Investigaciones recientes han explorado la integración de técnicas de aprendizaje automático con estos factores de riesgo tradicionales, demostrando mejoras en la capacidad predictiva [11][12]. Estudios utilizando Random Forest, Support Vector Machines y redes neuronales artificiales han reportado áreas bajo la curva ROC entre 0.75 y 0.95 en diferentes poblaciones [13][14]. Sin embargo, la heterogeneidad metodológica y la inclusión frecuente de variables post-evento limitan la comparabilidad y aplicabilidad de estos resultados.

La base de datos MIMIC-IV ha sido extensamente utilizada para el desarrollo de modelos predictivos de mortalidad a 30 días en pacientes con IAM [15][16], mientras que estudios basados en registros electrónicos de salud han explorado la predicción de eventos a 6 meses utilizando hasta 52,000 variables clínicas [17]. No obstante, persiste la necesidad de modelos que utilicen exclusivamente variables pre-evento y que hayan sido validados en poblaciones diversas.

3. JUSTIFICACIÓN

La necesidad de desarrollar modelos predictivos verdaderamente tempranos para IAM se fundamenta en múltiples aspectos. Primero, la implementación de estrategias de prevención primaria requiere la identificación de individuos en riesgo antes de que se manifieste el evento coronario. Segundo, la optimización de recursos sanitarios mediante screening dirigido a poblaciones de alto riesgo demanda herramientas predictivas precisas y costo-efectivas. Tercero, el avance hacia la medicina personalizada requiere modelos que integren múltiples variables de riesgo de manera individualizada y dinámica.

La evaluación crítica de algoritmos existentes que presentan limitaciones metodológicas contribuirá a establecer estándares más rigurosos para el desarrollo y validación de modelos predictivos cardiovasculares. Adicionalmente, el desarrollo de modelos basados exclusivamente en variables pre-evento tendrá impacto directo en la práctica clínica, facilitando la implementación de herramientas de screening en atención primaria y especializada.

4. OBJETIVOS

4.1. OBJETIVO GENERAL

Evaluar la efectividad predictiva de un algoritmo reportado para predicción de IAM que utiliza variables post-evento, y desarrollar modelos alternativos de aprendizaje automático basados exclusivamente en variables pre-evento para la predicción temprana de infarto agudo de miocardio.

4.2. OBJETIVOS ESPECÍFICOS

1. Analizar críticamente la metodología y variables utilizadas en el algoritmo objeto de evaluación, identificando las variables post-evento que comprometen su capacidad predictiva real.
2. Desarrollar y entrenar múltiples modelos de aprendizaje automático vistos en clase utilizando exclusivamente variables pre-evento para la predicción de IAM.
3. Comparar el rendimiento predictivo entre el modelo evaluado y los modelos desarrollados mediante métricas estándar de evaluación.
4. Identificar las variables pre-evento con mayor peso predictivo mediante técnicas de selección de características y análisis de importancia.
5. Validar los modelos desarrollados en conjuntos de datos independientes para evaluar su generalización.

5. FUENTES DE DATOS

Habiendo realizado una investigación preliminar sobre el estado del arte de las predicciones de IAM, observamos que, como todo dato clínico, se requerirá un proceso de obtención de datos de pacientes, que para tener un valor legítimo deberá pasar por estudios clínicos de distinto nivel, consentimientos informados por parte de los pacientes, autorización de la autoridad sanitaria y demás cuestiones que implican retrasos que no pueden asumirse en una asignatura tan breve. Por lo tanto, y considerando que más allá de las formalidades, las variables clínicas que se requieren existen más allá de las formalidades, recurrimos al análisis de diversos datasets de uso público, que definen adecuadamente las poblaciones analizadas y que pueden utilizarse para cumplir con los objetivos del proyecto.

5.1. DATOS PRIMARIOS:

- Cleveland Heart Disease Dataset: Conjunto clásico con 303 registros y 14 variables cardiovasculares (Disponible en: https://www.kaggle.com/datasets/ritwikb3/heart-disease-cleveland?resource=download)
- Framingham Heart Study Dataset: Datos longitudinales de seguimiento cardiovascular (Disponible en https://www.framinghamheartstudy.org/fhs-for-researchers/data-available-overview/)
- Heart Attack Dataset: Los conjuntos de datos de infarto se recopilaron en el hospital Zheen de Erbil (Irak) entre enero y mayo de 2019. Los atributos de este conjunto de datos son: edad, sexo, frecuencia cardíaca, presión arterial sistólica, presión arterial diastólica, glucemia, CK-MB y troponina, con valores negativos o positivos. Según la información proporcionada, el conjunto de datos médicos clasifica si se trata de un infarto o no. La columna de sexo está normalizada: para hombre se establece en 1 y para mujer en 0 UCI Heart Disease Dataset: Conjunto expandido con múltiples centros médicos. (Disponible en: https://www.kaggle.com/datasets/fatemehmohammadinia/heart-attack-dataset-tarik-a-rashid/data)
- UCI Heart Disease Data: Es una base de datos de referencia de investigadores. Contiene 74 campos de los cuáles sólo se han utilizado 14 de ellos .(Disponible en https://archive.ics.uci.edu/dataset/45/heart+disease)

5.2. VARIABLES DE ESTUDIO PREEVENTO (PREDICTORAS):

- Variables Pre-evento (Predictoras): Demográficas: edad, sexo, etnia; Antropométricas: índice de masa corporal, circunferencia abdominal; Factores de riesgo tradicionales: hipertensión arterial, diabetes mellitus, dislipidemia, tabaquismo actual/previo; Parámetros clínicos basales: presión arterial sistólica/diastólica, frecuencia cardíaca en reposo;
- Variables bioquímicas basales: colesterol total, colesterol LDL, colesterol HDL, triglicéridos, glucemia en ayunas; Historia clínica: antecedentes familiares de cardiopatía isquémica, enfermedad vascular periférica
- Variables electrocardiográficas basales: ritmo sinusal, bloqueos de conducción preexistentes

5.3. VARIABLES POST-EVENTO (A EXCLUIR)

- Biomarcadores de daño miocárdico: troponinas, CK-MB
- Alteraciones electrocardiográficas agudas: elevación del ST, ondas Q patológicas
- Parámetros hemodinámicos durante el evento agudo

5.4. VARIABLE RESULTADO

- Infarto agudo de miocardio confirmado (definido por criterios ESC/AHA)

6. METODOLOGÍA

Se decide realizar un procesamiento de 3 etapas. En la primera etapa se procesan los *datasets* con las correspondientes actividades de Análisis Exploratorio de Datos (EDA). En la segunda parte, se proponen los modelos de predicción con los diferentes *datasets* para poder finalmente evaluar comparativamente los modelos presentados. En la Parte 1, se deberá estandarizar el conjunto de datos para poder establecer una comparación con los demás grupos.

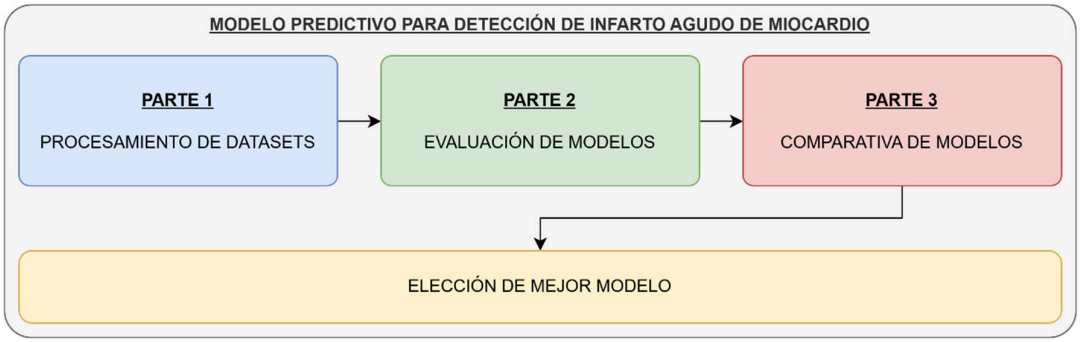


Fig. 1: Metodología utilizada para el tratamiento del problema planteado.

Finalmente, de la evaluación de los modelos se seleccionará el modelo más adecuado para predecir infartos agudos de miocardio.

7. RESULTADOS

7.1. PARTE 1: PREPROCESAMIENTO DE DATOS

7.1.1. Selección de los datasets

Para comenzar con la selección de los datasets, se realizó una búsqueda sistemática de datasets sobre predicción de ataques cardíacos. Existen al respecto más de 17.000 registros que deben filtrarse adecuadamente para contar con información confiable. De todos ellos, se hizo una preselección de 18 datasets que cumplían con los criterios de selección y se evaluaron con el siguiente proceso:

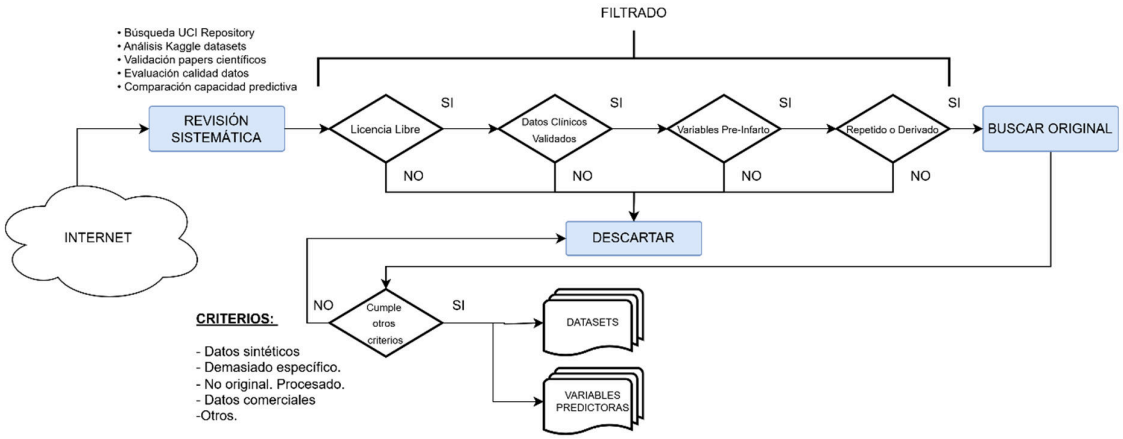


Fig. 2: Proceso de selección y filtrado de los datasets para garantizar su calidad.

Con los criterios de filtrado, se pudieron identificar datasets de alta calidad e integridad, pero, además, se pudieron identificar las variables comunes en varios de ellos que permitieran compararlos y trabajarlos con los mismos criterios. En la revisión sistemática se realizó una primera selección que arrojó los siguientes resultados:

Tabla 1: Resumen de la selección inicial de datos para el procesamiento.

Métrica	Valor
Datasets evaluados inicialmente	18
Fuentes consultadas	6
Papers científicos revisados	12+
Instituciones médicas referenciadas	15+
Países de origen de datos	8
Datasets finalmente seleccionados	3
Tasa de selección	16.7%
Total de registros en datasets seleccionados	71,217
Total de variables únicas analizadas	35
Tiempo de búsqueda sistemática	Proceso exhaustivo

A continuación, se observa la calificación de los tres mejores datasets seleccionados:

Tabla 2: Puntuación obtenida por los tres mejores datasets sobre los que se comparará el dataset original.

Criterio	Peso (%)	Descripción	Dataset 1 (UCI)	Dataset 2 (Combined)	Dataset 3 (Cardiovascular)
Integridad de Datos	25	Completitud y calidad	5/5	4/5	3/5
Cantidad de Información	25	Número de registros	3/5	4/5	5/5
No Repetibilidad	20	Unicidad del dataset	5/5	4/5	4/5
Soporte Científico	20	Papers y validación	5/5	4/5	3/5
Facilidad de Acceso	10	Disponibilidad y licencia	4/5	5/5	5/5
Puntuación Total	100	Promedio ponderado	**92/100**	**88/100**	**82/100**

7.1.2. Descripción inicial de los datasets

UCI Heart Failure Clinical Records

Variable Original	Variable Español	Tipo	Descripción
age	edad	Numérica	Edad del paciente (años)
anaemia	anemia	Binaria	Disminución de glóbulos rojos (0/1)
creatinine_phosphokinase	creatinina_fosfoquinasa	Numérica	Nivel enzima CPK (mcg/L)
diabetes	diabetes	Binaria	Presencia diabetes (0/1)
ejection_fraction	fraccion_eyeccion	Numérica	Porcentaje sangre eyectada (%)
high_blood_pressure	presion_alta	Binaria	Hipertensión (0/1)
platelets	plaquetas	Numérica	Plaquetas en sangre (kiloplaquetas/mL)
serum_creatinine	creatinina_serica	Numérica	Creatinina sérica (mg/dL)
serum_sodium	sodio_serico	Numérica	Sodio sérico (mEq/L)
sex	sexo	Binaria	Género (0=Mujer, 1=Hombre)
smoking	fumador	Binaria	Tabaquismo (0/1)
time	tiempo_seguimiento	Numérica	Días de seguimiento
DEATH_EVENT	evento_muerte	Target	Muerte durante seguimiento (0/1)

Heart Failure Prediction Dataset

Variable Original	Variable Español	Tipo	Descripción
Age	edad	Numérica	Edad del paciente (años)
Sex	sexo	Categórica	Género (M/F)
ChestPainType	tipo_dolor_pecho	Categórica	Tipo dolor pecho (TA/ATA/NAP/ASY)
RestingBP	presion_reposo	Numérica	Presión arterial reposo (mm Hg)
Cholesterol	colesterol	Numérica	Colesterol sérico (mg/dl)
FastingBS	glucosa_ayunas	Binaria	Glucosa ayunas >120 mg/dl (0/1)
RestingECG	ecg_reposo	Categórica	ECG reposo (Normal/ST/LVH)
MaxHR	frecuencia_cardiaca_max	Numérica	Frecuencia cardíaca máxima
ExerciseAngina	angina_ejercicio	Binaria	Angina ejercicio (Y/N)
Oldpeak	depresion_st	Numérica	Depresión ST por ejercicio
ST_Slope	pendiente_st	Categórica	Pendiente ST (Up/Flat/Down)
HeartDisease	enfermedad_cardiaca	Target	Enfermedad cardíaca (0/1)

Cardiovascular Disease Dataset

Variable Original	Variable Español	Tipo	Descripción
age	edad_dias	Numérica	Edad en días
gender	genero	Categórica	Género (1=Mujer, 2=Hombre)
height	altura	Numérica	Altura (cm)
weight	peso	Numérica	Peso (kg)

ap_hi	presion_sistolica	Numérica	Presión sistólica (mm Hg)
ap_lo	presion_diastolica	Numérica	Presión diastólica (mm Hg)
cholesterol	colesterol	Categórica	Nivel colesterol (1/2/3)
gluc	glucosa	Categórica	Nivel glucosa (1/2/3)
smoke	fumador	Binaria	Tabaquismo (0/1)
alco	alcohol	Binaria	Consumo alcohol (0/1)
active	actividad_fisica	Binaria	Actividad física (0/1)
cardio	enfermedad_cardiovascular	Target	Enfermedad cardiovascular (0/1)

7.1.3. Comparación entre datasets

A todos los conjuntos de datos que se van a importar, se les realiza un análisis preliminar para compararlos entre sí, y determinar sus fortalezas y debilidades principales.

Pos.	Dataset	AUC Estimado	Fortalezas Principales	Limitaciones
1°	**Dataset Primario**	~0.75-0.85	Signos vitales + glucemia directos	Muestra media, sin comorbilidades
2°	**Heart.csv**	~0.70-0.80	Variables diversas, síntomas específicos	Variables de ECG limitadas
3°	**Cardio Train**	~0.65-0.75	Muestra masiva, variables de estilo de vida	Variables básicas, menos específicas
4°	**Heart Failure**	~0.60-0.70	Comorbilidades específicas	Muestra pequeña, enfoque en falla

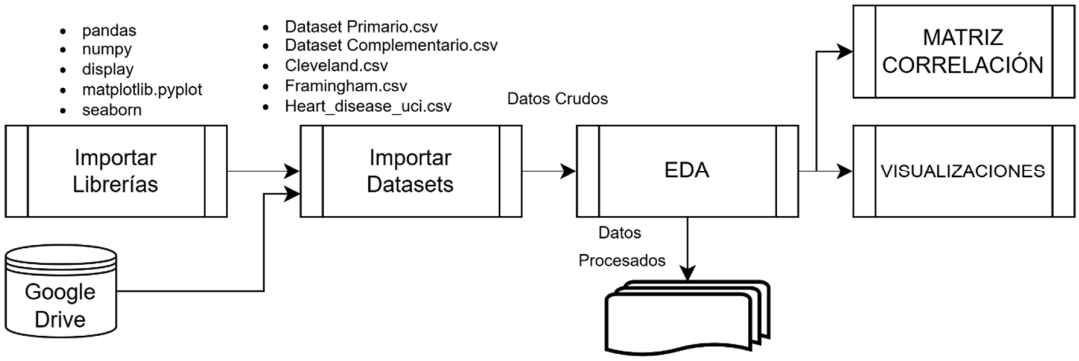
7.1.4. Variables eliminadas y causas

Tabla 3: Eliminación de variables no predictivas y causas de su eliminación.

Dataset Primario	Heart Failure Clinical Records	Heart.csv	Cardio Train
CK-MB: Biomarcador liberado DESPUÉS del daño miocárdico	**creatinine_phosphokinase**: Enzima elevada post-daño (misma que CK-MB)	**Oldpeak**: Depresión ST inducida por ejercicio (posible post-evento)	**id**: No predictivo
Troponina: Biomarcador liberado DESPUÉS del infarto	**ejection_fraction**: Función cardíaca post-falla	**ST_Slope**: Pendiente del segmento ST (posible post-evento)	
	platelets: Alterado post-evento		
	serum_creatinine: Daño renal secundario		
	serum_sodium: Alterado en falla cardíaca		
	serum_sodium: Alterado en falla cardíaca		

7.1.5. Importación de los datasets

Se seleccionan los datos desde el Google Drive y se convierten los datasets a dataframes para poder ser procesados con Google Colab. A los cinco (4) dataframes se les aplica el proceso de Análisis Exploratorio de Datos y con ese procesamiento se obtienen las visualizaciones para facilitar la interpretación de los datos, las respectivas matrices de correlación y los datos limpios y procesados para ser remitidos a la etapa 2 donde servirán de artefacto para la prueba de los modelos de aprendizaje y de clasificación.



7.1.6. Análisis Exploratorio de Datos:

El EDA se basó en el dataset original que fue puesto a evaluación.

7.1.7. Carga y Verificación Inicial del Dataset

Se realiza la carga del "Dataset Primario.csv" con verificación de dimensiones. Se obtiene como resultado un dataframe de 1,319 registros con 9 columnas (Age, Gender, Heart rate, Systolic/Diastolic BP, Blood sugar, CK-MB, Troponin, Result).

7.1.8. Estadísticas Descriptivas

Se analizan las medidas de tendencia central y dispersión para todas las variables numéricas. Se obtienen resultados tales como la Edad promedio: 56.19 años (rango 14-103), la frecuencia cardíaca promedio: 78.34 bpm con outliers extremos (máximo 1,111 bpm) y variables post-infarto altamente variables (CK-MB: 0.321-300, Troponin: 0.001-10.3). Esto último se debe principalmente a que estos biomarcadores varían su valor aumentando en caso de ataque cardíaco, por lo que su valor indica el tiempo desde que inició el ataque, hasta la gravedad del mismo.

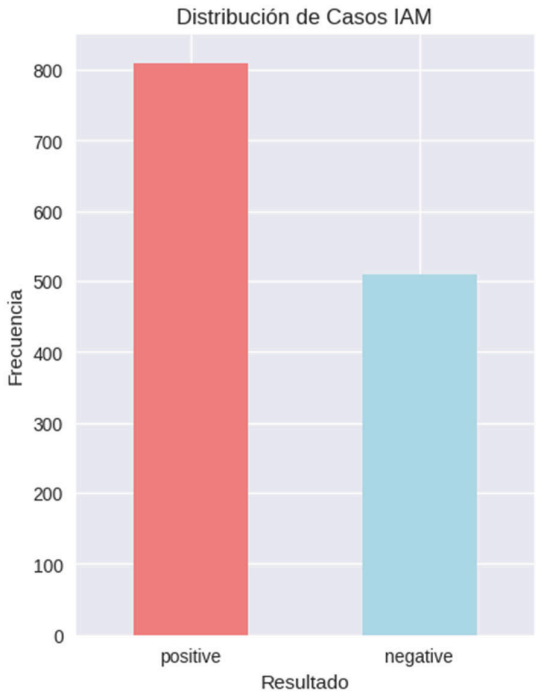
7.1.9. Análisis de Calidad de Datos

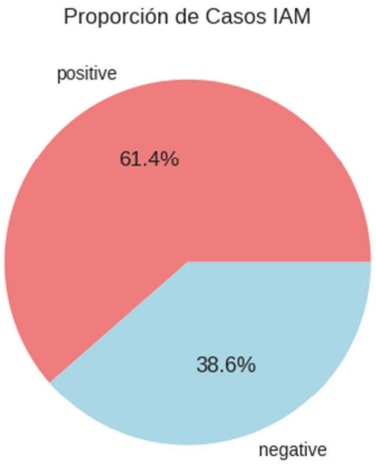
Se ejecuta la verificación de valores faltantes, la detección de duplicados y la identificación de outliers usando método IQR. Se observa un outlier en la frecuencia cardíaca que se elimina ya que un valor de 1111 no es compatible con la vida. Se observa que el dataset no tiene valores faltantes ni datos duplicados y una alta variabilidad en creatin-quinasa y troponina (CK-MB (15.54%), Troponin (19.48%))

7.1.10. Limpieza de Datos Atípicos Extremos

Se eliminaron registros con Heart rate > 1000 bpm (fisiológicamente imposibles) debido a que son valores que exceden límites biológicos realistas. Como resultado, se obtiene un dataset limpio manteniendo integridad médica

7.1.11. Análisis de la Variable Objetivo

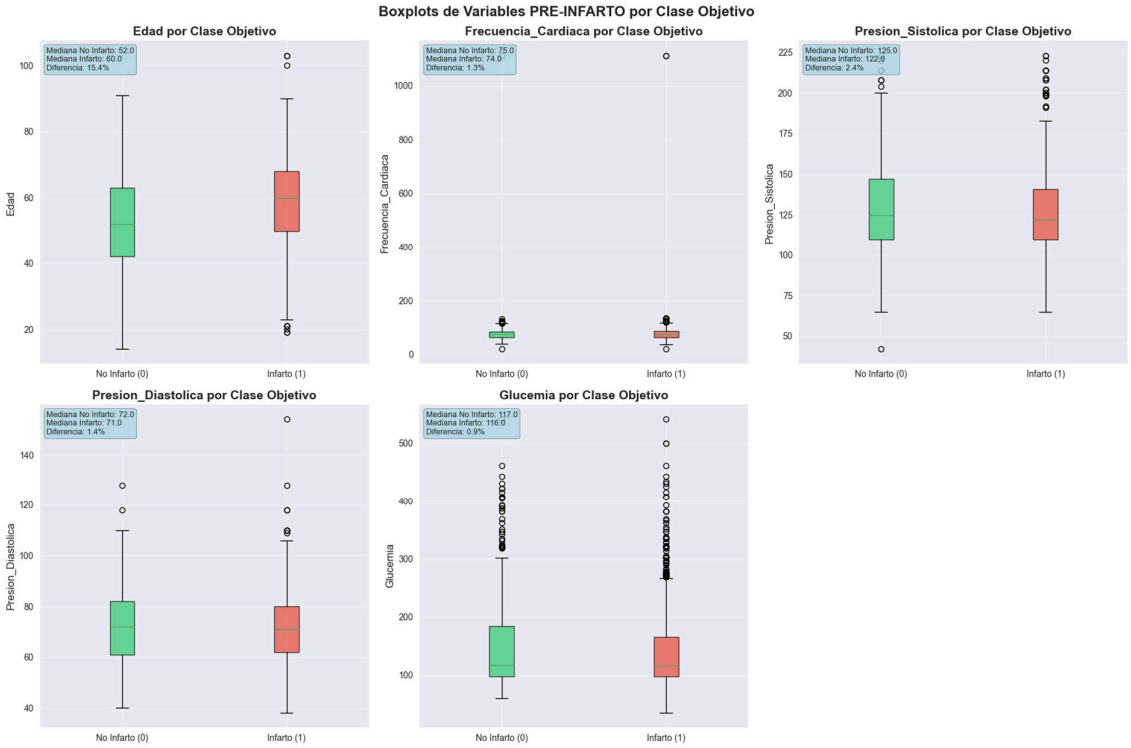




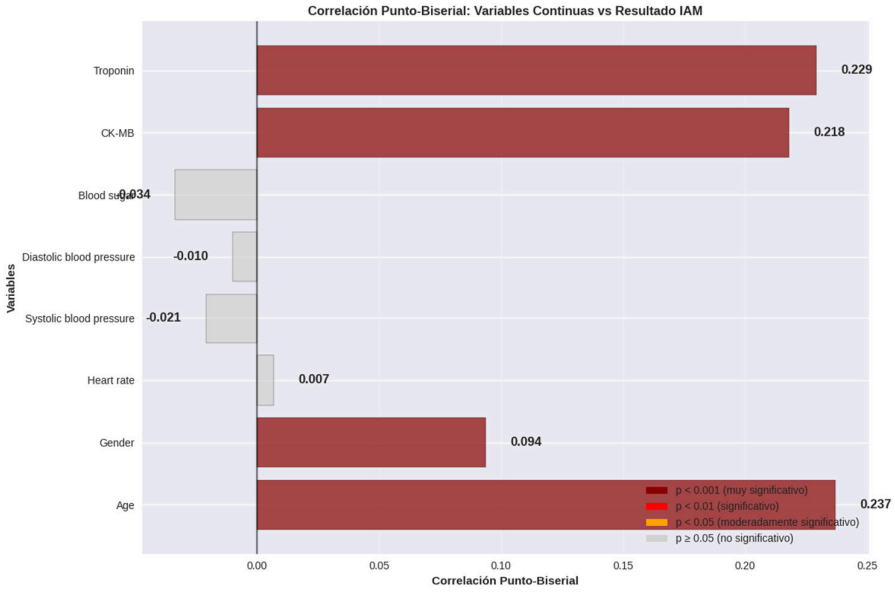
Se analizan las distribuciones de casos positivos/negativos, obteniendo 61.4% casos positivos (810 registros) y 38.6% casos negativos (509 registros). Se observa un dataset desbalanceado pero manejable.

Respecto de las relaciones, se observa que la edad tiene una clara incidencia en el subgrupo con infarto. Asimismo, se observa una clara influencia de la Glucemia elevada y de la presión sistólica elevada (hipertensión) en el subgrupo de los pacientes con infarto. Aunque esto no sea concluyente al momento de determinar la relación inversa (no todo paciente con glucemia alta o hipertensión tiene que terminar en infarto).

Por eso, estos factores que no resultan determinantes pero que tienen una clara relación, son determinados como factores de riesgo.



7.1.12. Matriz de Correlación Punto-Biserial



Considerando que la variable objetivo es de tipo dicotómica, se garantiza un mejor ajuste correlacional cuando se utiliza este tipo de matriz de correlación, (que resulta metodológicamente correcta entre variables continuas y variable dicotómica)
En dicha matriz se obtiene una correlación positiva moderada (≈0.238) para la edad, una correlación muy fuerte para CK-MB y Troponin con significancia estadística y correlaciones débiles pero consistentes con las variables hemodinámicas (presión sistólica y diastólica).

En este punto en particular, cabe destacar que tanto la CK-MB como la Troponina, son variables que aumentan considerablemente ante el daño cardíaco. Es decir, son variables que elevan su valor cuando ya el evento sucedió (variables post-hoc), pero que justamente por eso, carecen de valor predictivo. Además, al ser comparadas o evaluadas en un modelo que desconoce este contexto, generan en el modelo un sesgo temporal ya que se relacionan en forma directa con la variable objetivo. Podrían ellas mismas convertirse en variable objetivo ya que están directamente relacionadas con la ocurrencia de la variable dicotómica objetivo. Si se detecta una elevación de estas variables, es porque ya sucedió el infarto. Y en forma recíproca, si sucedió el infarto, es porque estos biomarcadores (sin importar la magnitud) están elevados.

7.1.13. Análisis Específico de Variables Post-Infarto

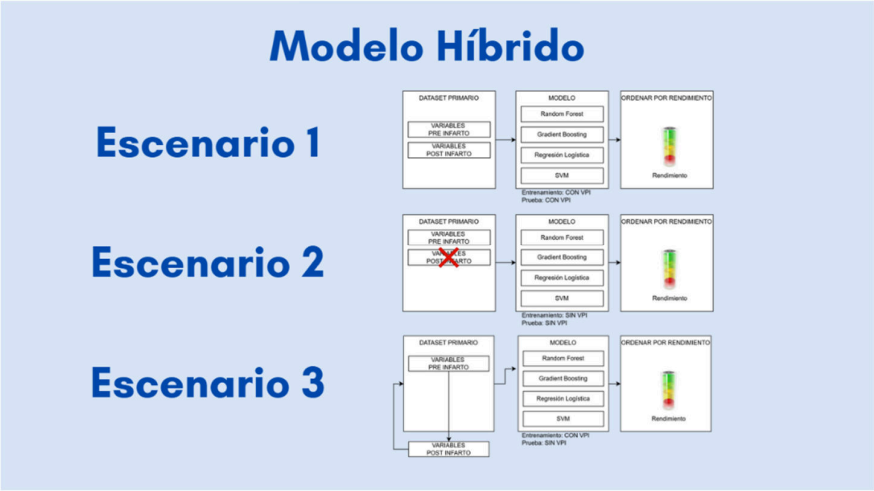
Específicamente para las variables post infarto, se realizan los boxplots comparativos, los histogramas de densidad y el análisis estadístico de diferencias entre grupos. Como resulta lógicamente esperable, la troponin presenta una elevación significativa en casos positivos (ratio >2x) al igual que la CK-MB ya que confirma su elevación post-infarto. Por consiguiente, estas variables presentan una confirmación del sesgo temporal ya que son consecuencia, no predictores.

7.1.14. Ingeniería de Características:

Luego de realizado el EDA, se procedió a la normalización de variables continuas mediante StandardScaler, a la codificación de variables categóricas usando One-Hot Encoding y a la selección de características mediante técnicas univariadas y multivariadas.

7.2. DESARROLLO DE MODELOS

7.2.1. Diseño de Tres Escenarios Experimentales



Se desarrollaron los tres escenarios experimentales con los cuatro modelos de predicción con el propósito de establecer rendimiento máximo teórico y considerando todas las variables (Age, Gender, Heart rate, BP, Blood sugar, CK-MB, Troponin). Luego, en el escenario 2, se descartaron las variables post-infarto, con la intención de evaluar la capacidad de predicción clínica realista y dejando solamente las variables preinfarto para análisis (Age, Gender, Heart rate, BP, Blood sugar). Finalmente se elaboró el tercer escenario, donde se intentó entrenar con conocimiento completo pero evaluar con información limitada mediante un método discutido desde las buenas prácticas pero entendiendo que las variables post-infarto eran también virtualmente variables objetivo (entrenamiento con todas las variables, evaluación solo con pre-infarto)

7.2.2. Preparación de Datos

Se realizó la división train/test (80/20) con estratificación, la normalización con StandardScaler para modelos que lo requieren y la codificación de variable objetivo (0=negative, 1=positive)

7.2.3. Selección de Cuatro Modelos de Machine Learning

Random Forest

Se escogió este modelo porque fue utilizado previamente en otra asignatura (Inferencia Estadística) con excelentes resultados ante escenarios donde otros modelos de selección fallaban. Es un modelo robusto ante outliers y datos desbalanceados, proporciona importancia de características y maneja bien interacciones no lineales.

Como principales ventajas no requiere normalización de datos, resiste bien al overfitting, interpreta importancia de variables y tiene la capacidad de manejarlos datos faltantes internamente.
Y por su parte, como desventajas o debilidades, se nota en el procesamiento que resulta computacionalmente costoso (tardaba 20 segundos en procesar lo que otros modelos procesaban en 4 o 5), es menos interpretable que modelos lineales y tiene sesgos hacia variables con más categorías

Regresión Logística

Este es un modelo que tiene un baseline clásico para clasificación binaria como el problema que estamos tratando. Es rápido, eficiente y altamente interpretable (coeficientes = odds ratios).
Como ventajas de este modelo, podemos citar una interpretabilidad excelente, las probabilidades calibradas, no asume distribución de características y es de los modelos menos propenso al overfitting. También fue utilizado en inferencia estadística oportunamente con buenos resultados y requerimientos de procesamiento bajos.
Como puntos desfavorables, se asume una relación lineal entre variables y log-odds y resulta muy sensible a outliers (por lo que se quitan previamente en la etapa de EDA); requiere normalización de datos y no captura interacciones complejas

Gradient Boosting

Este modelo se está usando abiertamente por su rendimiento y es el estado del arte en muchos problemas de clasificación. Tiene una excelente capacidad predictiva y maneja maneja bien patrones complejos.
Como principales ventajas mencionaremos una muy alta precisión predictiva, su capacidad para manejar automáticamente interacciones y una buena capacidad ante outliers. Tambien proporciona importancia de características.
Como sus principales desventajas es lamentablemente propenso al overfitting si no se regula requiere un procesamiento intensivo, posee muchos hiperparámetros para ajustar y resulta menos interpretable

Support Vector Machine (SVM)

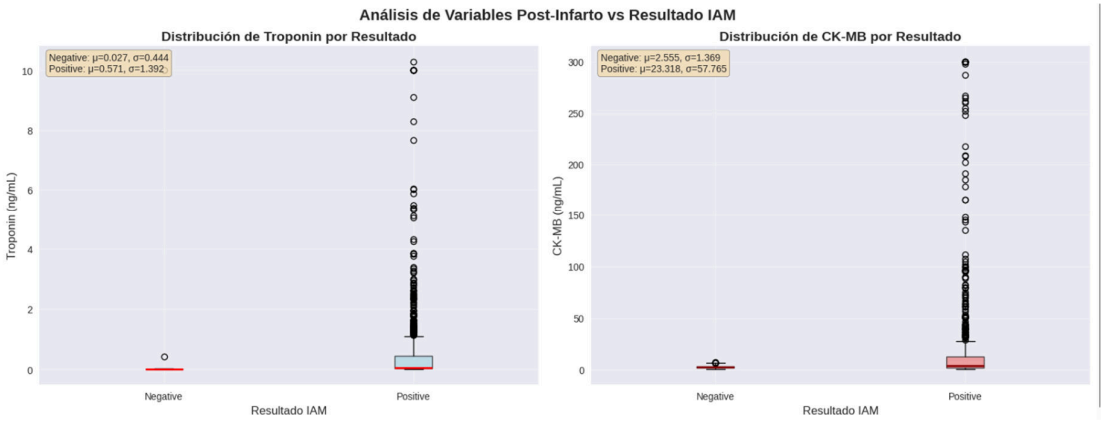
Finalmente, se decidió aplicar el modelo SVM que resulta efectivo en espacios de alta dimensión, suele ser robusto ante overfitting y versátil con diferentes kernels.

Sus principales ventajas son que resulta ser efectivo con pocas muestras, muy versátil (kernel trick), se comporta muy bien ante posibles sobre entrenamientos y cuenta con buenas características ante alta dimensionalidad.

Como puntos en contra, requiere normalización obligatoria que resulta ser un paso adicional en el procesamiento, no proporciona probabilidades directamente, sensible a selección de parámetros y computacionalmente costoso en datasets grandes. Esta última desventaja resultaba sensible en uno de los datasets con 70 mil registros procesados, para lo cual se debió extraer una muestra aleatoria.

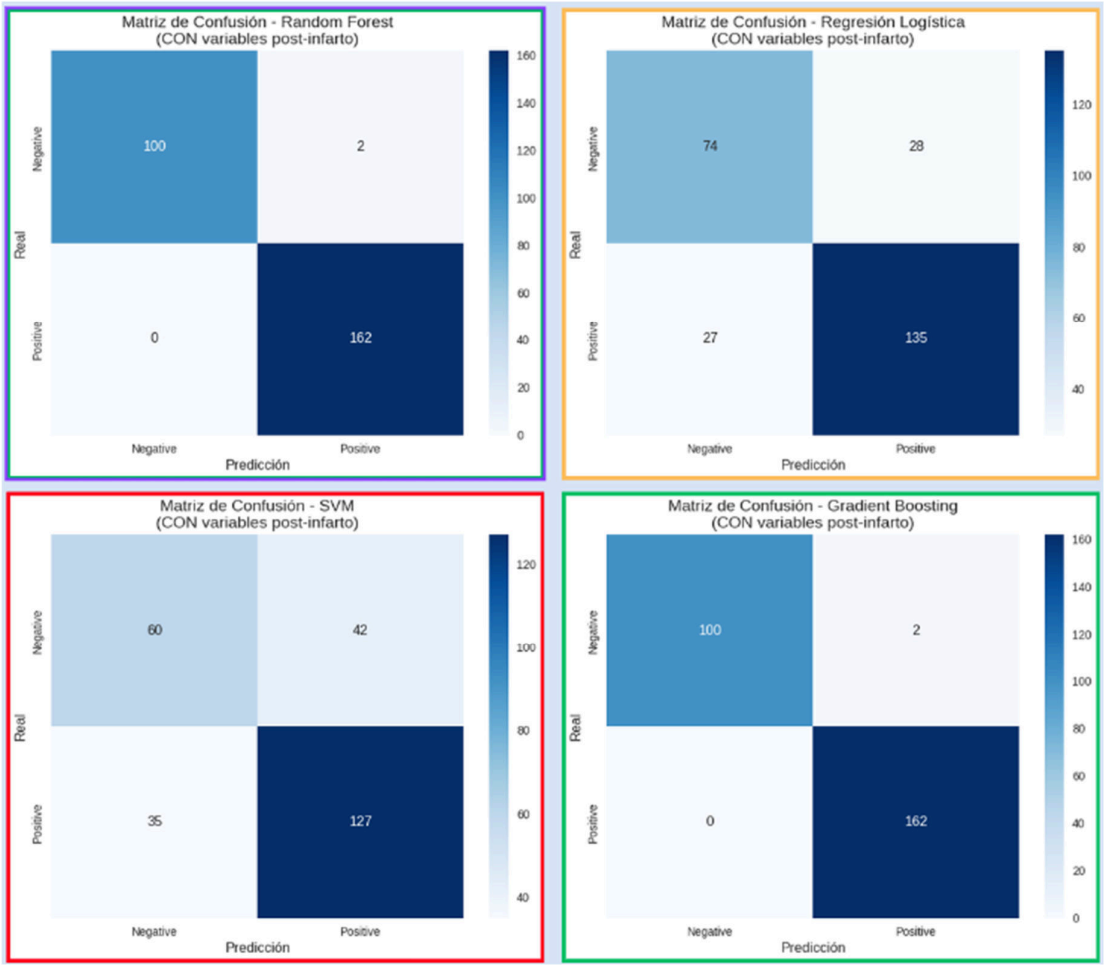
7.3. MODELO HÍBRIDO

Teniendo en cuenta las características del problema a resolver, y considerando que los mejores resultados se habían obtenido en el dataset original considerando las variables post infarto (troponina y creatina kinasa MB), se decidió hacer la evaluación interna del dataset, evaluando tres escenarios diferentes: en primer lugar el dataset original como se presenta con todas sus variables (inclusive las VPI); otro escenario directamente sin las VPI y un tercer escenario donde los modelos se entrenaran CON las variables post infarto, pero se probara sin éstas. De esa forma, se podría lograr un entrenamiento rico ya que el modelo aprende de casos completos con todos los biomarcadores; una evaluación realista, porque se evalúa solo con información disponible antes del infarto y una aplicación práctica ya que se simula un sistema entrenado con datos históricos completos pero aplicado en tiempo real. Si bien es una práctica no recomendada (ya que lo que se evalúa es el funcionamiento del modelo con datos incompletos) elimina el "sesgo temporal", resultando interesante medir cuánta pérdida de precisión ocurre cuando no están disponibles las variables que suceden post-hoc, o post evento.
En el siguiente gráfico, se observa que, si bien resulta imposible medir la Troponina y la Creatina Kinasa en forma previa, sus valores se ven alterados (elevados) indefectiblemente cuando sucede un ataque cardíaco, e inclusive, de acuerdo a su concentración, puede estimarse el tiempo en que comenzó el infarto o inclusive su gravedad. En caso de poder considerar estas variables como "objetivo", sabiendo que se elevan sistemáticamente ante eventos cardíacos positivos, virtualmente podríamos recurrir a predecir su alteración y el infarto con variables medibles en forma previa a la ocurrencia del evento.



7.3.1. Análisis con las Variables Post Infarto

	MÉTRICAS			
	Random Forest	Regresión Logística	Gradient Boosting	SVM
Exactitud (Accuracy):	0.9924	0.7917	0.9924	0.7083
Precisión (Precision)	0.9878	0.8282	0.9878	0.7515
Sensibilidad (Recall)	1.0000	0.8333	1.0000	0.7840
F1-Score	0.9939	0.8308	0.9939	0.7674
AUC-ROC	0.9982	0.8893	0.9872	0.8103

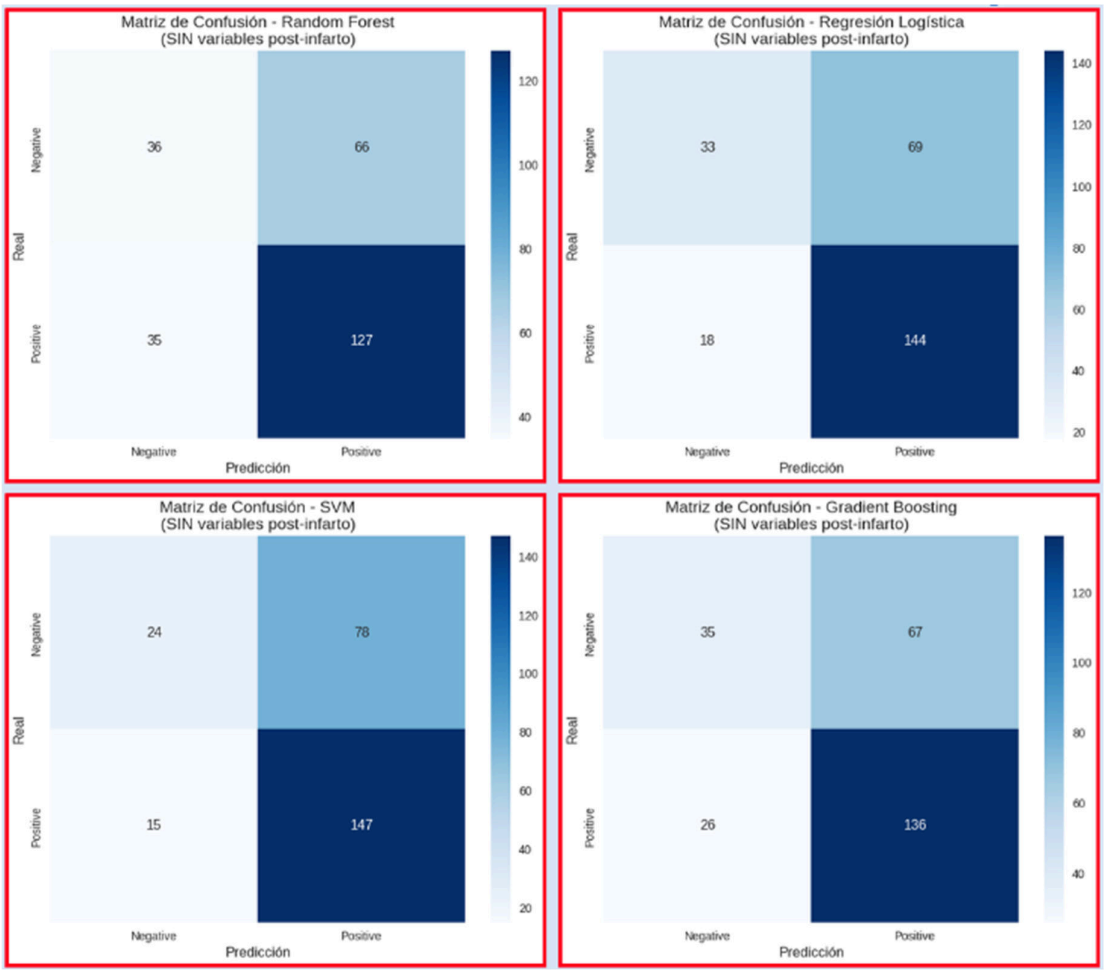


Con las variables post infarto, resulta evidente que la predicción debería dar alta en todos los modelos, sin embargo, aparecen dos modelos (Regresión Logística y SVM) que brindan un resultado medio y pobre de análisis, aún con variables post infarto.

7.3.2. Análisis sin las Variables Post Infarto

	MÉTRICAS			
	Random Forest	Regresión Logística	Gradient Boosting	SVM
Exactitud (Accuracy):	0.6174	0.6705	0.6477	0.6477
Precisión (Precision)	0.6580	0.6761	0.6700	0.6533
Sensibilidad (Recall)	0.7840	0.8889	0.8395	0.9074
F1-Score	0.7155	0.7680	0.7452	0.7597
AUC-ROC	0.5891	0.6604	0.6211	0.6135

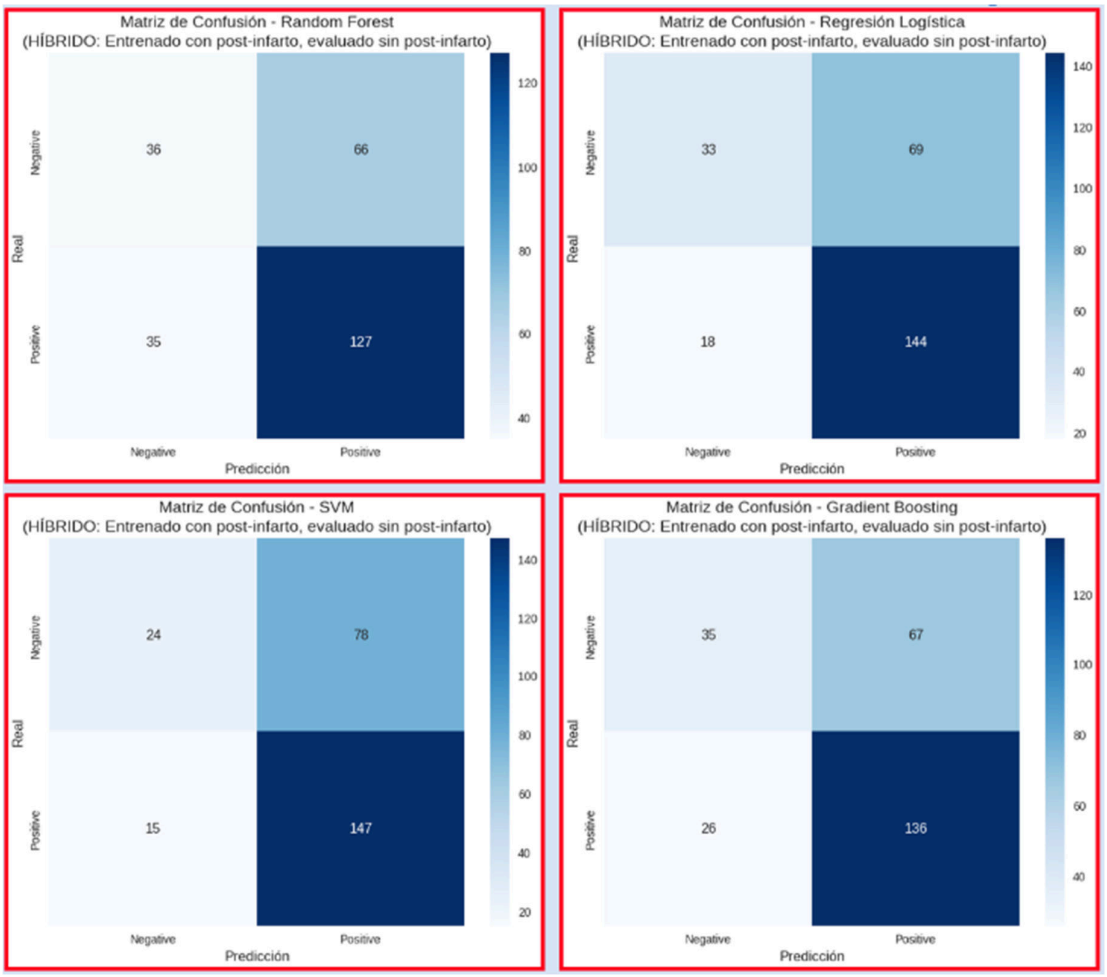
Cuando excluimos las variables post infarto del análisis (que, como hemos dicho tienen una relación directamente proporcional con la variable objetivo y se elevan si y solo si hay un infarto), ninguno de los modelos responde adecuadamente, brindando resultados pobres.



7.3.3. Análisis Híbrido (Entrenando con VPI – Probando sin VPI)

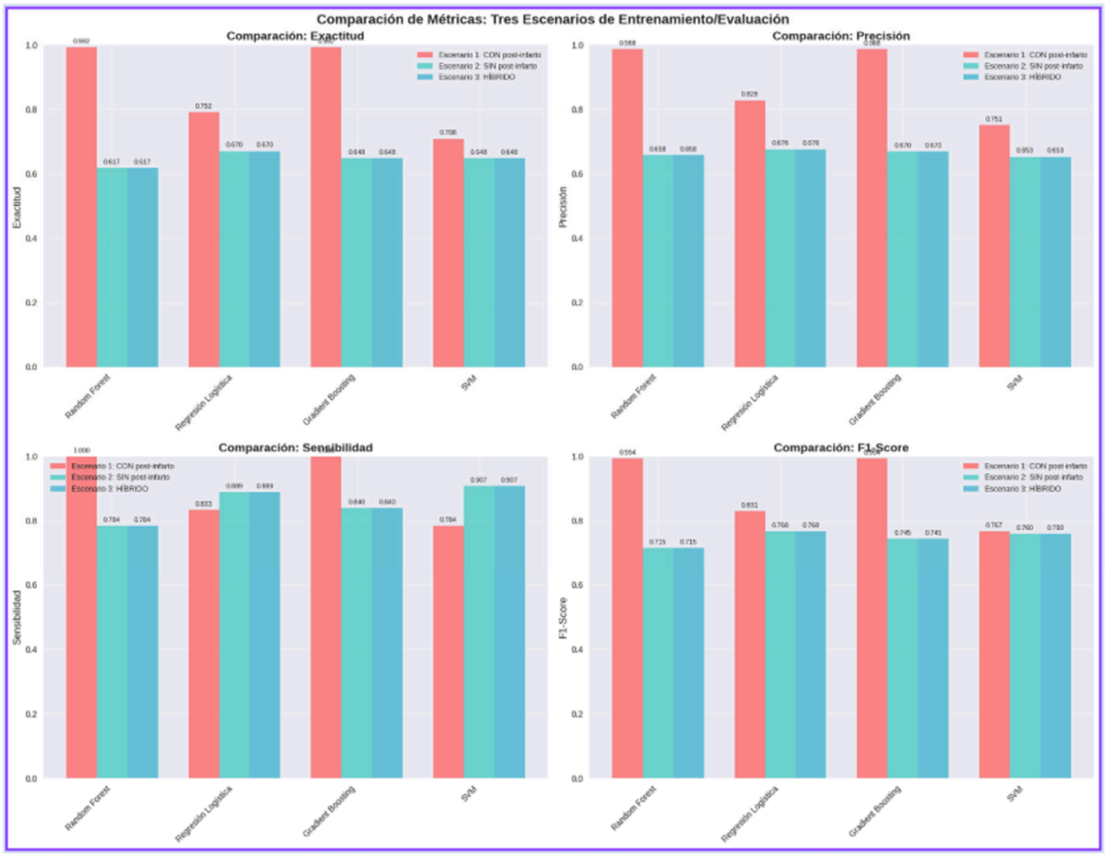
	MÉTRICAS			
	Random Forest	Regresión Logística	Gradient Boosting	SVM
Exactitud (Accuracy):	0.6174	0.6705	0.6477	0.6477
Precisión (Precision)	0.6580	0.6761	0.6700	0.6533
Sensibilidad (Recall)	0.7840	0.8889	0.8395	0.9074
F1-Score	0.7155	0.7680	0.7452	0.7597
AUC-ROC	0.5891	0.6604	0.6211	0.6135

En este caso, se presenta el caso híbrido donde se entrena el modelo con todas las variables, incluidas las de preinfarto y se prueba sólo con las variables pre infarto. El resultado obtenido es exactamente igual al que se da cuando se quitan las variables post infarto del entrenamiento. Es decir, la predictibilidad del modelo se define por las variables post infarto que, no son predictoras, introducen un sesgo temporal y no pueden medirse indirectamente.



7.3.4. Comparativa entre modelos y escenarios

	ESCENARIO 1 (CON POST-INFARTO)	ESCENARIO 2 (SIN POST-INFARTO)	ESCENARIO 3 (HÍBRIDO)
ACCURACY	Random Forest (0.9924)	Regresión Logística (0.6705)	Regresión Logística (0.6705)
PRECISION	Random Forest (0.9878)	Regresión Logística (0.6761)	Regresión Logística (0.6761)
RECALL	Random Forest (1.0000)	SVM (0.9074)	SVM (0.9074)
F1	Random Forest (0.9939)	Regresión Logística (0.7680)	Regresión Logística (0.7680)



8. RECOMENDACIONES DE USO

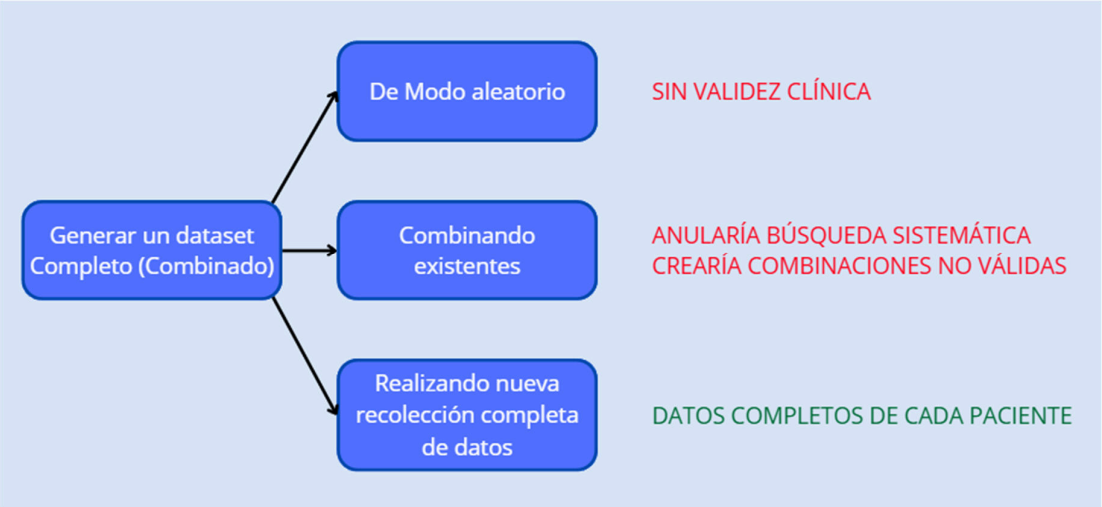
Finalmente, al no tener una utilidad propia ni un modelo predictor válido, se realizaron las combinaciones con otros datasets que incorporan algunas variables poblacionales, geográficas y de hábitos (y algunas post infarto también) que potencian la predictibilidad del dataset original. Sin embargo, surgieron planteos sumamente válidos que deben tenerse en cuenta y que se mencionan en la siguiente sección “Discusión”

Uso	Usar	Complementar con
Predicción Temprana de Infarto	Dataset Primario (variables fisiológicas directas)	Heart.csv (síntomas adicionales)
Screening Poblacional	Cardio Train (variables antropométricas y estilo de vida)	Dataset Primario (signos vitales)
Pacientes con Comorbilidades	Heart Failure (factores de riesgo específicos)	Heart.csv (síntomas cardiovasculares)

9. DISCUSIÓN:

El dataset por sí solo puede presentar una gran capacidad predictiva para las herramientas digitales con las que se evalúa. Presenta variables post infarto que son confirmadoras (mas no predictoras de un evento cardíaco). Pero se observa que los resultados de los modelos mejoran cuando se agregan otras variables como los hábitos, costumbres y perfiles

genéticos, raciales o geográficos entre otros. Pero lamentablemente no se cuenta con un dataset que incluya toda esta información.



Por consiguiente, existen tres alternativas: la más válida consiste en realizar una prueba completa con todas las variables pre infarto que pueden potenciar la predictibilidad del dataset original. Pero lamentablemente no se cuenta con el tiempo o recursos para adelantar este estudio que sería interesante por su propio alcance.
Si ese dataset se generara aleatoriamente, se perdería la validez clínica toda vez que muchas de las variables se asignarían en forma aleatoria que no necesariamente representa la realidad poblacional de la muestra. Y, si por otro lado, se combinaran los registros existentes, no solamente no se ajustaría a un escenario real (aunque lo acercaría), pero se generarían combinaciones discutibles ya que puede haber dos pacientes de la misma edad y con las mismas costumbres pero con resultados sumamente diferentes.

10. CONCLUSIONES

- Cuando se trate de conjuntos de datos que sean procesados en proyectos académicos, debe tenerse en cuenta que dichos datos provengan al mismo tiempo de fuentes confiables y que hayan sido generados en el ámbito de un proyecto académico. De esta forma se podrá garantizar la trazabilidad de la información y la validez de las conclusiones.
- El dataset primario que se sometió a comparativas resultó ser la mejor alternativa para predicción específica de infarto utilizando las variables pre-infarto, aunque para los modelos utilizados, los resultados de ese dataset solo fue realmente pobre con cualquiera de ellos. Los demás datasets podrían brindar al primario de mayor robustez, al consolidar los resultados previstos con el agregado de variables post-infarto e inclusive con el agregado de resultados fatales, que mejoran la predictibilidad del dataset primario.
- Al eliminar los biomarcadores post-infarto se podría mejorar la validez clínica y se mejora la predicción a corto plazo, aunque requiere un dataset consistente que permita perfilar mejor los grados de riesgo clínico. Así como está la información, el dataset tiene una función clínica orientadora, pero requiere indefectiblemente de los biomarcadores post infarto para tener una validez clínica.

- Todos los modelos pre-infarto son aplicables e implementables en la consulta profesional, al menos como orientadores de riesgo cardíaco. Tienen la ventaja de que no requieren pruebas costosas de laboratorio y resultan ideales para detección temprana y prevención.

11. REFERENCIAS

[1] World Health Organization, "Cardiovascular diseases (CVDs) - Fact sheet," 2021. [Online]. Available: https://www.who.int/news-room/fact-sheets/detail/cardiovascular-diseases-(cvds)

[2] K. Thygesen et al., "Fourth universal definition of myocardial infarction (2018)," J. Am. Coll. Cardiol., vol. 72, no. 18, pp. 2231-2264, Oct. 2018.

[3] J. Liu et al., "Machine learning models to predict 30-day mortality for critical patients with myocardial infarction: a retrospective analysis from MIMIC-IV database," Front. Cardiovasc. Med., vol. 11, Art. no. 1368022, Sep. 2024.

[4] S. A. Mortazavi et al., "Prediction of myocardial infarction from patient features with machine learning," Front. Cardiovasc. Med., vol. 9, Art. no. 754609, Mar. 2022.

[5] K. K. Weng et al., "Machine learning to predict the likelihood of acute myocardial infarction," Circulation, vol. 140, no. 11, pp. 899-909, Sep. 2019.

[6] M. Ahmadi et al., "Prediction of the fatal acute complications of myocardial infarction via machine learning algorithms," Front. Cardiovasc. Med., vol. 11, Art. no. 1402503, May 2024.

[7] P. W. Wilson et al., "Prediction of coronary heart disease using risk factor categories," Circulation, vol. 97, no. 18, pp. 1837-1847, May 1998.

[8] R. B. D'Agostino Sr. et al., "Validation of the Framingham coronary heart disease prediction scores: results of a multiple ethnic groups investigation," JAMA, vol. 286, no. 2, pp. 180-187, Jul. 2001.

[9] D. C. Goff Jr. et al., "2013 ACC/AHA guideline on the assessment of cardiovascular risk: a report of the American College of Cardiology/American Heart Association Task Force on Practice Guidelines," Circulation, vol. 129, no. 25 Suppl 2, pp. S49-73, Jun. 2014.

[10] SCORE2 working group and ESC Cardiovascular risk collaboration, "SCORE2 risk prediction algorithms: new models to estimate 10-year risk of cardiovascular disease in Europe," Eur. Heart J., vol. 42, no. 25, pp. 2439-2454, Jul. 2021.

[11] J. S. Alonso-González et al., "Machine learning-based myocardial infarction bibliometric analysis," Front. Med., vol. 12, Art. no. 1477351, Feb. 2025.

[12] A. Rajkomar et al., "Machine learning in medicine," N. Engl. J. Med., vol. 380, no. 14, pp. 1347-1358, Apr. 2019.

[13] B. Ambale-Venkatesh et al., "Cardiovascular event prediction by machine learning: the multi-ethnic study of atherosclerosis," Circ. Res., vol. 121, no. 9, pp. 1092-1101, Oct. 2017.

[14] A. Khera et al., "Machine learning to predict the likelihood of acute myocardial infarction," Circulation, vol. 140, no. 11, pp. 899-909, Sep. 2019.

[15] A. E. Johnson et al., "MIMIC-IV, a freely accessible electronic health record dataset," Sci. Data, vol. 7, Art. no. 206, Dec. 2020.

[16] T. Pollard et al., "Machine learning-based prediction of mortality in acute myocardial infarction with cardiogenic shock," Front. Cardiovasc. Med., vol. 11, Art. no. 1402503, 2024.

[17] S. M. Grundy et al., "Prediction of incident myocardial infarction using machine learning applied to harmonized electronic health record data," BMC Med. Inform. Decis. Mak., vol. 20, Art. no. 254, Oct. 2020.
[18] F. Conroy et al., "Performance of the Framingham risk models and pooled cohort equations for predicting 10-year risk of cardiovascular disease: a systematic review and meta-analysis," BMC Med., vol. 17, Art. no. 109, Jun. 2019.
[19] J. Marrugat et al., "Validation of the general Framingham Risk Score (FRS), SCORE2, revised PCE and WHO CVD risk scores in an Asian population," Int. J. Cardiol. Cardiovasc. Risk Prev., vol. 18, Art. no. 200187, Sep. 2023.
[20] R. B. D'Agostino et al., "General cardiovascular risk profile for use in primary care: the Framingham Heart Study," Circulation, vol. 117, no. 6, pp. 743-753, Feb. 2008.
[21] Chicco, D., Jurman, G. Machine learning can predict survival of patients with heart failure from serum creatinine and ejection fraction alone. BMC Medical Informatics and Decision Making 20, 16 (2020). DOI: 10.1186/s12911-020-1023-5
[22] Detrano, R. et al. International application of a new probability algorithm for the diagnosis of coronary artery disease. American Journal of Cardiology, 1989.
[23] UCI Machine Learning Repository. Heart Disease Data Set. https://archive.ics.uci.edu/ml/datasets/heart+disease
[24] Kaggle Public Datasets. Heart Failure Prediction Dataset. https://www.kaggle.com/datasets/fedesoriano/heart-failure-prediction
[25] Kaggle Public Datasets. Cardiovascular Disease Dataset. https://www.kaggle.com/datasets/sulianova/cardiovascular-disease-dataset