Predicción del riesgo cardiovascular y Prevención de Infartos de Miocardio mediante IA.

Ferran Vidal Rojas

Resumen — El presente trabajo analiza un conjunto de más de 4000 pacientes, con 15 atributos que representan factores de riesgo de cardiopatía coronaria. Mediante inteligencia artificial, se identifican patrones predictivos de infartos de miocardio, abarcando factores demográficos, conductuales y médicos. Para este estudio se utiliza un modelo basado en entidades relacionadas para evaluar la influencia de cada variable en la salud cardíaca. Los resultados permitirán a los profesionales de Atención Primaria poder desarrollar estrategias proactivas para la prevención y el tratamiento de enfermedades cardiovasculares, fomentando una compresión más profunda de los factores de riesgo y estableciendo bases para un futuro más saludable.

Palabras clave.
- Riesgo cardiovascular
- Infarto miocardio
- Enfermedades cardiovasculares
- IA y Big Data en medicina.

Abstract. **Summary of the master's Thesis.** This study analyses a dataset of more than 4,000 patients obtained from Kaggle, containing 15 attributes representing risk factors for coronary heart disease. Using artificial intelligence, predictive patterns for myocardial infarction are identified, covering demographic, behavioural, and medical factors. A model based on related entities is employed to assess the influence of each variable on cardiac health. The results will enable primary care professionals to develop proactive strategies for the prevention and treatment of cardiovascular diseases, fostering a deeper understanding of risk factors and laying the foundation for a healthier future.

Index Terms—Keywords.
- Cardiovascular risk
- Myocardial infarction
- Cardiovascular diseases
- AI and Big Data in Medicine

1. Introducción

Las enfermedades cardiovasculares, en concreto el infarto agudo de miocardio, representan una de las principales causas de morbimortalidad en el mundo actual. La identificación temprana de los factores de riesgo que inducen al desarrollo de estas patologías constituye uno de los trabajos importantes de la medicina preventiva y de la práctica clínica en Atención Primaria.

Por esto, en este trabajo se pretende utilizar y enmarcar las técnicas de **Inteligencia Artificial (IA)** y el **aprendizaje automático (machine learning)** para poder obtener y especificar los pasos a seguir en la predicción del riesgo cardiovascular a 10 años, teniendo el objetivo de contribuir en la detección precoz de pacientes susceptibles de desarrollar cardiopatía coronaria.

Para ello, se ha utilizado un conjunto de datos abiertos provenientes de un estudio epidemiológico de cohortes que incluye información sobre más de 4.000 pacientes y 15 variables clínicas, demográficas y de estilo de vida, que representan habitualmente factores de riesgo para las enfermedades cardiovasculares. Este conjunto corresponde al *Framingham Heart Study dataset,* disponible públicamente en la plataforma Kaggle [1], el cual deriva del histórico **Framingham Heart Study**, iniciado en 1948 en la localidad de Framingham, Massachusetts (EE.UU.), considerado pionero en la identifica-

- *E-mail de contacto: 2005512@uab.cat*
- *Trabajo tutorizado por: Remo Suppi Boldrito*
- *Curso: 2025*

ción de los principales factores de riesgo coronario [2].

A partir de este dataset, se han utilizado diferentes estrategias de preprocesamiento y generación sintética de datos para poder abordar el problema del desbalance de clases, así como técnicas de modelado basadas en algoritmos de aprendizaje automático para identificar patrones predictivos de infarto de miocardio y estimar el riesgo individual de enfermedad coronaria a 10 años.

El infarto agudo de miocardio (IAM) constituye una de las principales causas de morbimortalidad tanto a nivel mundial como en España. A escala global, las enfermedades cardiovasculares son responsables de 17,9 millones de muertes anuales, lo que supone un 32% de todas las defunciones registradas, de las cuales el 85% se deben a infarto de miocardio y accidente cerebrovascular [4,5]. Además, se estima que cada año se producen alrededor de 7,2 millones de nuevos casos de infarto, y más de 200 millones de personas conviven actualmente con enfermedad coronaria, la condición subyacente al IAM [5].

En España, las enfermedades cardiovasculares siguen siendo la primera causa de muerte. En 2022 ocasionaron 119.196 fallecimientos, de los cuales 49.926 correspondieron a cardiopatía isquémica [6]. Cada año se producen entre 70.000 y 75.000 infartos, con una incidencia de 150-160 casos por 100.000 habitantes [7]. Pese a estas cifras, la mortalidad intrahospitalaria es de las más bajas del mundo (3-5% en los IAM con elevación del segmento ST), gracias a la consolidada red asistencial Código IAM y a la generalización del tratamiento con angioplastia primaria, aplicada en más del 90% de los pacientes [8,9].

El perfil del paciente en España muestra una edad media de 65-68 años en varones y 75-78 años en mujeres, con factores de riesgo predominantes como hipertensión arterial, dislipemia, tabaquismo y diabetes [7,9]. Estos datos reflejan que, aunque la mortalidad ha descendido de forma significativa en las últimas décadas, el IAM continúa representando un enorme desafío sanitario y social, que exige tanto estrategias preventivas eficaces como sistemas asistenciales de alta calidad.

El objetivo del estudio es proporcionar un modelo predictivo y clínicamente interpretable, para los profesionales de salud en Atención Primaria, facilitando las estrategias proactivas de prevención y tratamiento de enfermedades cardiovasculares para mejorar así la salud cardiovascular y establecer unas bases para un futuro más saludable.

2. Estado del arte

Utilizar la Inteligencia Artificial (IA) y aprendizaje automático (ML) en la salud ha crecido exponencialmente en los últimos años, impulsado por la creciente disponibilidad de grandes volúmenes de datos sanitarios y el avance en los algoritmos de aprendizaje automático. Particularmente, la predicción del riesgo cardiovascular se ha consolidado como uno de los campos de aplicación más relevantes, por su impacto potencial y la prevención de eventos cínicos graves como el infarto de miocardio.

En diferentes estudios realizados se ha demostrado que los modelos tradicionales de predicción de riesgo, basados en scores clínicos como el Framingham Risk Score o SCORE, presentan limitaciones en los términos de precisión y generalización. Dichos modelos, muy utilizados previamente, se fundamentan en un número reducido de variables y su desarrollo solía basarse en poblaciones concretas, con lo que se limita su aplicabilidad en contextos demográficos y clínicos heterogéneos (13-15).

Por ello, técnicas de aplicación como Machine Learning (ML) y modelos avanzados de IA permite superar estas limitaciones, siendo capaces de procesar grandes cantidades de datos, identificar patrones complejos y no lineales, generando predicciones más precisas y personalizadas. Algoritmos como Random Forest, Gradient Boosting, CatBoost o redes neuronales han demostrado resultados importantes a valorar para la estratificación del riesgo cardiovascular y la prevención de los infartos de miocardio o la enfermedad coronaria, superando a los modelos tradicionales (16,17)

Aunque todo ello es una gran mejoría, el uso de estas técnicas en el ámbito cínico representa una serie de retos importantes. Hay que destacar la necesidad de garantizar la interpretabilidad de los modelos, la gestión adecuada de los datos faltantes, el tratamiento correcto del des-

balance de clases y la validación rigurosa de los resultados para así asegurar la aplicabilidad en la práctica médica.

El presente trabajo aborda estos retos mediante la aplicación de técnicas de generación de datos sintéticos, balanceo de clases y modelos interpretables, con el objetivo de desarrollar un sistema predictivo de riesgo cardiovascular a 10 años que sea preciso, fiable y clínicamente útil en el ámbito de la Atención Primaria.

3. Material y métodos

3.1. Población a estudio y origen de los datos

En el estudio se ha utilizado un conjunto de datos obtenido de la plataforma Kaggle, procedente de un estudio epidemiológico de Framingham (*Framingham Heart Study*), el cual ha servido como base para múltiples investigaciones sobre el riesgo cardiovascular. El dataset incluye información de **4.240 pacientes** y recoge **15 variables** clínicas, demográficas y de estilo de vida, junto con una variable objetivo binaria que indica si el paciente desarrolló **enfermedad coronaria en los siguientes 10 años** (TenYearCHD) [1 - 3].

Los datos son **anonimizados y de acceso público**, por lo que no ha sido necesario el consentimiento informado individual ni la aprobación por parte de un comité ético, de acuerdo con la normativa vigente sobre el uso de datos secundarios de acceso abierto para investigación.

3.2. Variables utilizadas

El conjunto de datos contiene un total de **15 atributos**, distribuidos en variables demográficas, conductuales, antecedentes médicos y parámetros clínicos, además de la variable objetivo [1]. A continuación, se detallan:

1. **Variables demográficas**:
   - **Sex**: Sexo biológico del paciente, codificado como masculino (“M”) o femenino (“F”).
   - **Age**: Edad del paciente (años – Variable continua).
   - **Education:** Nivel educativo del paciente (categórica – valores de 1 a 4).
2. **Variables conductuales**:
   - **is_smoking**: indica si el paciente es fumador actual (“Yes” o “No”).
   - **cigsPerDay**: Número medio de cigarrillos fumados al día (continua).
3. **Antecedentes médicos:**
   - **BPMeds**: Uso de medicación para la presión arterial (Sí/No).
   - **prevalentStroke:** Antecedente de accidente cerebrovascular (Sí/No).
   - **prevalentHyp:** Hipertensión arterial previa (Sí/No).
   - **diabetes**: Diagnóstico previo de diabetes mellitus (Sí/No).
4. **Parámetros clínicos actuales:**
   - **totChol**: Colesterol total (mg/dL).
   - **sysBP**: Presión arterial sistólica (mmHg).
   - **diaBP**: Presión arterial diastólica (mmHg).
   - **BMI**: Índice de masa corporal (kg/m^2).
   - **heartRate**: Frecuencia cardíaca (latidos por minuto).
   - **glucose**: Glucosa en sangre (mg/dL).
5. **Variable objetivo:**
   - **TenYearCHD**: Indica si paciente desarrolló enfermedad coronaria en los siguientes 10 años (binaria: “1” significa “Sí”, “0” significa “No”).

3.3. Procesamiento y Generación de datos

El procesamiento del conjunto de datos original fue una etapa clave en el desarrollo del estudio, ya que implicó la identificación y corrección de problemas estructurales como valores ausentes, variables categóricas no codificadas y desbalance significativo en la variable objetivo. Este preprocesamiento ha sido fundamental para garantizar la calidad y robustez de los modelos predictivos desarrollados posteriormente. [Anexo I].

Unos de los problemas detectados más importantes del dataset es desequilibrio entre clases. Para solucionar este desajuste se recurrió a una técnica de *data augmentation* con la generación de datos sintéticos utilizando el entorno **SDV** (Synthetic Data Vault). Con ello se generaron datos equilibrados a partir de distribuciones aprendidas del conjunto de datos original, preservando la estructura estadística de las variables sin comprometer la privacidad [10]. En el conjunto original, la variable objetivo **TenYearCHD** presen-

taba un marcado desbalance, con alrededor de un **15% de pacientes positivos** frente a un **85% de pacientes negativos**, lo que generaba un riesgo significativo de sesgo en el entrenamiento de los modelos.

Clase	Dataset original (%)	Dataset balanceado con SDV (%)
Negativos (0)	85	51.69
Positivos (1)	15	48.31

3.3.1. Gestión de valores faltantes

Del dataset original (data_cardiovascular_risk.csv), en su primera inspección, se detectaron valores ausentes en variables clínicas y conductuales importantes como **educación**, **cigsPerDay**, **BPMeds**, **totChol**, **BMI** y **glucosa**. La tabla siguiente resume el número de valores nulos identificados [1] [Anexo I].

Variable	Nulos detectados
Education	10
cigsPerDay	3
BPMeds	5
totChol	4
BMI	1
glucose	53

Para un análisis inicial exploratorio, se optó por eliminar las filas incompletas mediante, *df.dropna(inplace=True)*. No obstante, en fases posteriores, se priorizó la imputación avanzada de valores con el objetivo de conservar la mayor cantidad posible de datos útiles.

3.3.2. Imputación de valores ausentes

Para tratar valores nulos y minimizar la pérdida de información, se utilizó el algoritmo *K-Nearest Neighbors Imputer* (KNN Imputer). Este método permite imputar de forma adecuada a partir de la similitud con observaciones completas, mejorando la representatividad del dataset. Para ello previamente se realizaron las siguientes condiciones: todas las variables se transformaron a formato numérico, las variables categóricas fueron codificadas adecuadamente y se verificó que no existieran inconsistencias o duplicados en los registros. Así, el método permitió estimar los valores faltantes, proporcionando imputaciones coherentes y realistas [11,12].

3.3.3. Codificación variables categóricas

Las variables categóricas se transformaron a variables numéricas.

Variables binarias como **sex**, **is_smoking**, **BPMeds**, **prevalentStroke**, **prevalentHyp** y **diabetes** se codificaron como 0 (No) y 1 (Sí).

La variable **education**, de tipo ordinal, se mantuvo con sus valores (1 a 4).

En los modelos que lo requerían, como Regresión Logística, Random Forest, Gradient Boosting fue necesario esta conversión, pero en el modelo **CatBoost**, diseñado específicamente para manejar datos con variables categóricas, esta transformación no fue utilizada ya que el modelo internamente detecta las categorías y aplica codificaciones propias, con lo que ahorra trabajo y evita errores [13].

3.3.4. Escalado de variables numéricas

Para los modelos sensibles a la escala de los datos, como la regresión logística, se aplicó un proceso de normalización estándar (mediante StandardScaler), que transforma las variables para que presenten media 0 y desviación estándar 1 contribuyendo a mejorar la estabilidad numérica y la interpretación de los coeficientes.

Las variables que fueron sometidas a esta transformación son: **age**, **cigsPerDay**, **totChol**, **sysBP**, **diaBP**, **BMI**, **heartRate** y **glucose**.

3.3.5. Análisis gráfico exploratorio

Las principales variables predictoras mostraron diferencias significativas entre los grupos de pacientes con y sin evento cardiovascular, destacando la edad, la presión arterial sistólica, el colesterol total, la glucosa y la presencia de diabetes.

Para analizar en detalle las distribuciones de las variables se generaron diversas visualizaciones gráficas para identificar los patrones relevantes en los datos: [Anexo I].

- Histogramas de variables numéricas (edad, colesterol, glucosa).
- Boxplots para detectar valores extremos.
- Gráficos de distribución por clase (**TenYearCHD**) para variables como edad, presión arterial y glucosa.
- Matriz de correlación entre variables mediante un heatmap, identificando asociaciones relevantes, como la relación entre edad, presión sistólica y riesgo cardiovascular**.**

Para analizar las variables en relación con el riesgo cardiovascular, se generaron visualizaciones comparativas entre pacientes con y sin eventos coronarios. La **edad** mostró una clara diferencia: los casos positivos se concentraron en mayores de 50 años, reforzando su papel como factor de riesgo. (fig. 1).

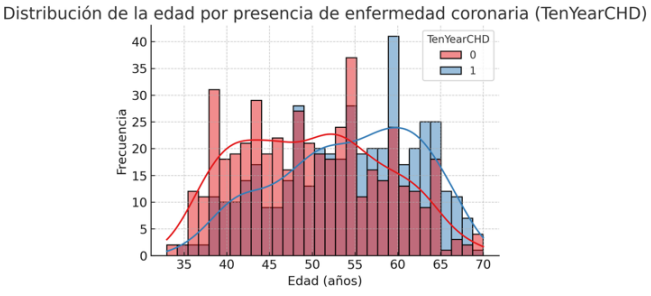


Figura 1. Distribución edad según presencia o ausencia enf. Coronaria

Además, la **matriz de correlación** entre variables clínicas evidenció una asociación moderada entre **edad y presión arterial sistólica**, y correlaciones más débiles entre glucosa, colesterol y otros factores. Estos patrones apoyan su inclusión como predictores relevantes en el modelo (fig.2).

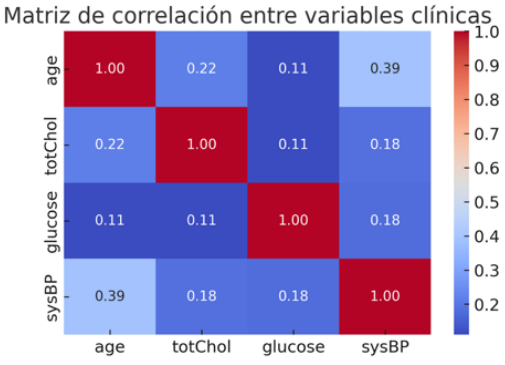


Figura 2. Matriz correlación entre variables clínicas

3.3.6. Generación datos sintéticos y balanceo de clases

La variable objetivo (TenYearCHD) del dataset original mostraba una importante desproporción:

- **Clase 0** (no desarrolla enfermedad): 3.600 pacientes, aproximadamente el 85% de la muestra.
- **Clase 1** (desarrolla enfermedad): 640 pacientes, alrededor del 15% de la muestra.

Esta diferencia evidenció el marcado desbalanceo entre clases, lo que justificó la necesidad de aplicar diversas técnicas para equilibrar la distribución. A lo largo del proyecto se ensayaron diferentes estrategias de balanceo, entre ellas el submuestreo (*undersampling)* de la clase mayoritaria, el sobremuestreo (*oversampling*) de la clase minoritaria, el ajuste del umbral de decisión para priorizar la detección de la clase positiva y la ponderación de clases durante el entrenamiento de los modelos. Si bien estas técnicas aportaron ciertas mejoras, los resultados no fueron completamente satisfactorios, en especial en términos de *recall* (sensibilidad) para la clase minoritaria.

Ante esta limitación, se recurrió a la generación de datos sintéticos mediante la librería **Synthetic Data Vault (SDV).** El procedimiento consistió en separar los datos de la clase positiva (pacientes con enfermedad) y entrenar un modelo generativo específico, a partir del cual se generaron aproximadamente 400 nuevos registros sintéticos.

Paralelamente, la clase negativa se redujo mediante submuestreo hasta unas 450 – 500 observaciones, logrando así un equilibrio entre ambas categorías. La combinación de registros reales y sintéticos permitió conformar un nuevo dataset balanceado y estructuralmente, denominado *data_cardiovascular_balanced.csv,* con un total de 919 observaciones. Esta estrategia mejoró notablemente la representatividad de la clase minoritaria sin recurrir a la duplicación directa de ejemplos ni introducir sesgos artificiales. [Anexo I].

En la siguiente figura se muestra, a modo de ejemplo, la distribución comparativa de clases antes y después del balanceo, evidenciando la mejora en la representatividad de la clase positiva (*TenYearCHD = 1*) (fig.3)

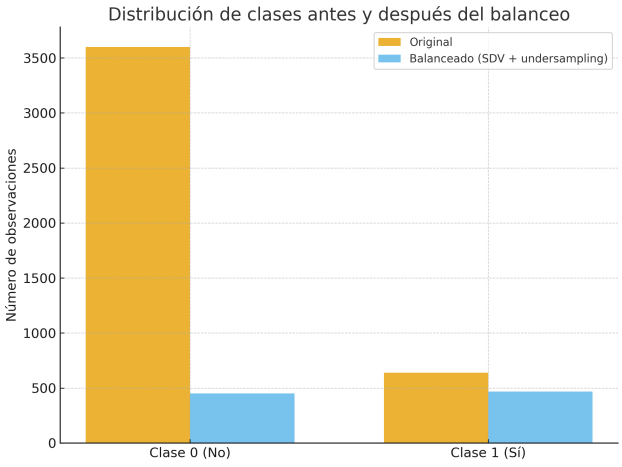


Figura 3. Distribución de clases

En una segunda fase, y con el objetivo de aumentar aún más la cantidad de datos disponibles para el entrenamiento de los modelos, se implementó un proceso adicional de generación sintética utilizando **CTGAN (Conditional Tabular GAN)**. Se entrenaron dos modelos independientes, uno para cada clase, y se generaron 300 registros sintéticos adicionales por categoría, manteniendo el equilibrio entre ambas. Tras combinar los datos reales y los sintéticos, se obtuvo un dataset final denominado *data_cardiovascular_ampliado_imputado.csv*, con un total de 1.519 observaciones balanceadas y sin valores ausentes. [Anexo I].

Fase del dataset	Clase 0 (No)	Clase 1 (Sí)	Total
Original	3.600	640	4.240
Balanceado (SDV + undersampling)	450	469	919
Ampliado (CTGAN)	759	760	1.519

Finalmente, este dataset ampliado fue sometido a un proceso de validación en el que se comprobó la distribución equilibrada entre clases, la conservación de la estructura y relaciones estadísticas observadas en los datos originales, la ausencia de valores nulos o inconsistencias tras la imputación y la representación realista de variables críticas como la edad, la presión arterial y la glucosa. El resultado fue un conjunto de datos estable, balanceado y clínicamente coherente, adecuado para el entrenamiento, validación e interpretación de los modelos predictivos desarrollados en el estudio.

3.4. Modelado y técnicas aplicadas

El estudio se centró en comparar diferentes algoritmos supervisados para estimar el riesgo individual de desarrollar enfermedad coronaria a 10 años. Se priorizó el uso de modelos con capacidad predictiva sólida y, al mismo tiempo, con un buen nivel de interpretabilidad, condición especialmente relevante en entornos sanitarios donde la compresión del modelo resulta clave para su aplicación clínica.

Con el dataset final obtenido se entrenaron y evaluaron distintos modelos de aprendizaje automático. Entre ellos se incluyeron la **Regresión Logística**, ampliamente utilizada en el ámbito médico por su interpretabilidad y facilidad para calcular *odds ratios;* el **Random Forest**, un ensamble de árboles de decisión que permite modelar relaciones no lineales y proporciona estimaciones de importancia de las variables; el **Gradient Boosting**, técnica secuencial de ensamble que corrige errores iterativamente y suele alcanzar alta precisión y robustez; y, finalmente, el **CatBoost**, un algoritmo optimizado para datos tabulares que maneja variables categóricas sin necesidad de preprocesamiento adicional, ofreciendo gran estabilidad y rendimiento, además de interpretabilidad, motivo por el cual fue seleccionado como modelo final del estudio.

Para el entrenamiento y validación de los modelos se aplicaron criterios metodológicos comunes: el dataset se dividió de manera estratificada en conjuntos de entrenamiento (70%) y de prueba (30%); se aplicaron técnicas de escalado y codificación cuando fue necesario; se llevó a cabo el ajuste de hiperparámetros mediante búsqueda en malla (Grid Search) junto con validación cruzada; y se evaluó el rendimiento utilizando métricas estándar como *accuracy*, precision, *recall*, F1-score y área bajo la curva ROC (AUC). Adicionalmente, se analizaron las matrices de confusión para comprender la distribución de los errores y se aplicaron métodos de interpretación basados en la importancia de las variables y valores SHAP, con el objetivo de facilitar la comprensión y validación clínica de los resultados.

4. RESULTADOS

El análisis comparativo de los modelos mostró un rendimiento superior del algoritmo **CatBoost**, que fue seleccionado como modelo final tanto por su elevada capacidad predictiva como por su interpretabilidad clínica.

Modelo	Accuracy	Precision	Recall	F1-score	AUC
Regresión Logística	0,77	0,78	0,76	0,77	0,82
Random Forest	0,77	**0,81**	0,76	0,78	0,83
Gradient Boosting	0,77	0,77	**0,79**	0,78	0,83
CatBoost	**0,78**	**0,78**	**0,78**	**0,78**	**0,97**

En la **Tabla** anterior se presentan las métricas globales de cada modelo, incluyendo exactitud, precisión, sensibilidad, F1-score y AUC. Entre todos ellos, CatBoost alcanzó los valores más altos, con un **AUC de 0,979** reflejando una excelente capacidad discriminativa para estimar el riesgo de enfermedad coronaria a 10 años.

La comparación de curvas ROC evidencia el rendimiento diferencial, siendo CatBoost el que logra la mayor área bajo la curva [Anexo II], (fig.4).

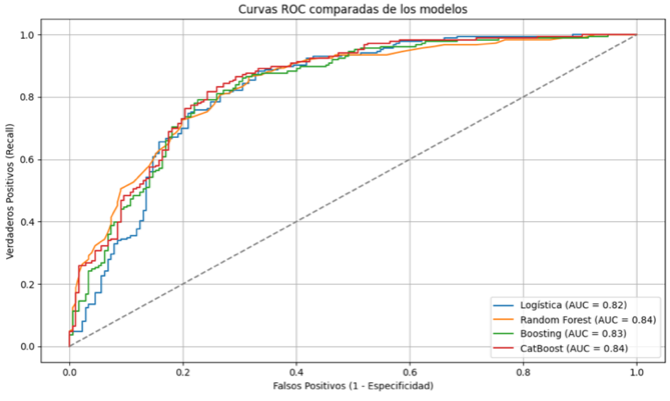


Figura 4. Curva ROC comparativa de los modelos

La matriz confirma un adecuado equilibrio entre falsos positivos y negativos, mostrando que el modelo clasifica correctamente la mayoría de los pacientes con y sin riesgo (fig.5).

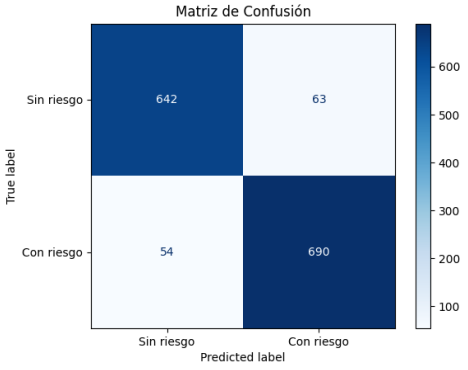


Figura 5. Matriz de confusión – CatBoost

Las variables más influyentes en la predicción fueron: **edad, presión arterial sistólica, hipertensión previa, colesterol total, cigarrillos por día y glucosa**, lo que coincide con factores de riesgo cardiovasculares bien establecidos en la literatura médica (fig.6).

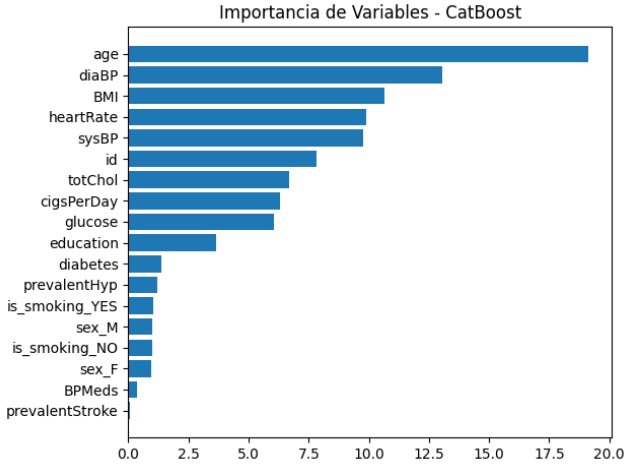


Figura 6. Importancia variables según CatBoost

Los gráficos SHAP permiten comprender cómo los valores altos o bajos de cada predictor afectan al riesgo estimado. Por ejemplo:

- **Edad avanzada**: se asocia directamente con mayor riesgo, reflejando el deterioro vascular progresivo.

- **Presión sistólica elevada**: indicador temprano de daño cardiovascular.
- **Colesterol total y glucosa altos**: evidencian disfunción metabólica.
- **Consumo de tabaco e hipertensión previa**: aumentan de forma importante la probabilidad de infarto.

(fig. 7).

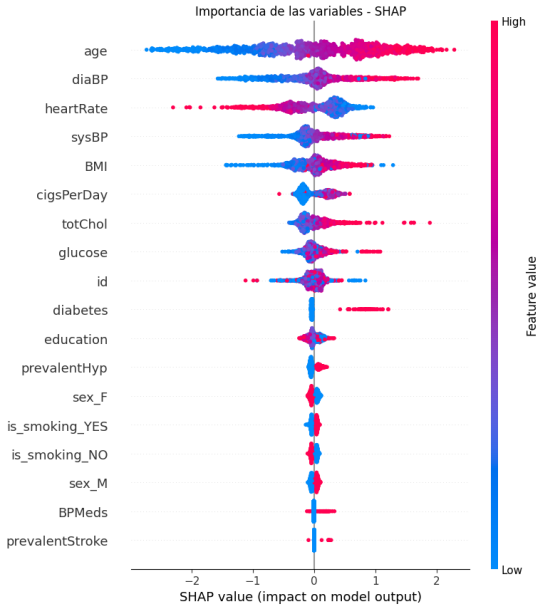


Figura 7. Interpretación SHAP - CatBoost

Por otro lado, se identificaron **factores protectores** asociados a menor riesgo relativo, como la **ausencia de antecedentes de ictus o diabetes**, mantener la presión arterial y niveles lipídicos en rangos normales, y un índice de masa corporal no excesivo.

En conjunto, estos resultados respaldan que CatBoost es un modelo robusto, preciso e interpretable, capaz de identificar patrones clínicos relevantes para la **estratificación del riesgo cardiovascular en Atención Primaria.** (fig.8)

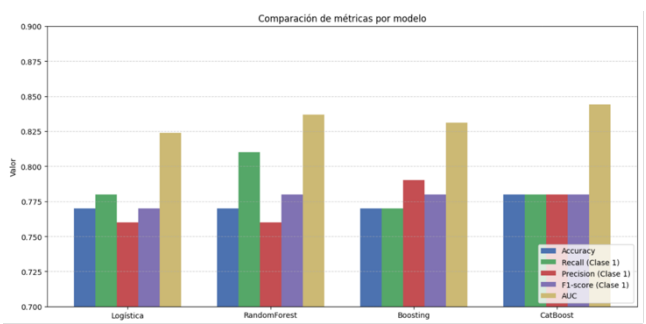


Figura 8. Comparación métricas por modelo

5. Discusión y limitaciones

En el estudio se confirma que el algoritmo **CatBoost** constituye una de las alternativas más sólidas para la predicción del riesgo cardiovascular a 10 años. Su rendimiento, cuantificado por un área bajo la curva ROC (AUC) de 0,979, sugiere una capacidad discriminativa superior para distinguir entre pacientes que desarrollarán enfermedad coronaria y aquellos que no. Este valor, notablemente alto, está por encima del de muchos de los modelos reportados en la literatura médica para tareas similares e indica un potencial significativo para su aplicación clínica. Este resultado adquiere mayor relevancia si se considera que los factores de riesgo analizados (hipertensión, diabetes, obesidad, entre otros) son ampliamente prevalentes en la población y habitualmente gestionados en Atención Primaria.

Más allá del rendimiento numérico, la elección de CatBoost se justifica por ventajas técnicas y prácticas. Su capacidad para manejar **variables categóricas de forma nativa** elimina la necesidad de procedimientos de codificación complejos, con lo que simplifica el flujo de trabajo y reduce el riesgo de errores. Asimismo, sus **herramientas de interpretabilidad**, como la importancia de variables o las gráficas de dependencia parcial, permiten no solo predecir, sino también **comprender los mecanismos subyacentes a cada estimación**. Esta transparencia es fundamental en entornos clínicos, donde la validación de las predicciones frente al juicio experto del médico resulta imprescindible.

La aplicación de este modelo en Atención Primaria podría transformar el abordaje del riesgo cardiovascular, permitiendo pasar de un sistema reactivo a uno proactivo y personalizado. Integrado en la historia clínica electrónica como un sistema de apoyo a la decisión clínica (CDSS), el modelo podría trabajar en segundo plano, señalando de forma automática a los pacientes con mayor riesgo tras cada consulta.

Con ello se generarían múltiples beneficios:

1. **Priorización de intervenciones**: concentrar recursos en los pacientes con mayor necesidad, optimizando la prevención secundaria y los programas de rehabilitación.

2. **Motivación de los pacientes**: usar predicciones visuales y objetivas como herramienta educativa para promover cambios en los estilos de vida.

3. **Eficiencia del sistema**: mejorar la asignación de recursos sanitarios hacia las personas con mayor necesidad, mejorando así la eficiencia del sistema.

El modelo no solo demuestra un alto grado de precisión, sino que también ofrece un apoyo clínico práctico y operativo para los médicos de familia en la práctica diaria de prevención cardiovascular.

Pese a los resultados positivos, hay que reconocer las limitaciones de este trabajo. El modelo se entrenó y validó sobre un único conjunto de datos, lo que reduce su validez externa y su capacidad de generalización. Por lo que debe ser validado en cohortes multicéntricas e independientes con diferentes características demográficas y prevalencia de enfermedad. Además, el alto AUC (~98) podría estar parcialmente condicionado por las características específicas del dataset, lo que obliga a interpretarlo con cautela, requiriendo verificación en entornos del mundo real con datos más ruidosos.

Por limitaciones de tiempo, no se implementaron técnicas de validación cruzada (ej. **K-fold cross-validation**), que hubieran permitido evaluar con mayor robustez el riesgo de sobreajuste. Al igual que la ausencia de validación en cohortes multicéntricas e independientes limita la extrapolación a escenarios clínicos reales, donde los datos suelen ser más heterogéneos y ruidosos.

6. Conclusión

Los datos confirman que el infarto de miocardio es un problema de salud pública de primer orden. A nivel mundial, su carga es abrumadora. En España, aunque la mortalidad ajustada ha disminuido de forma notable gracias a un sistema de atención excelente y rápido (Código IAM), el número absoluto de casos y muertes sigue siendo elevado debido al envejecimiento de la población y la persistencia de factores de riesgo como la hipertensión, la diabetes y la obesidad.

En la **prevención primaria** (promover estilos de vida saludables) sigue siendo asignatura pendiente para reducir la incidencia. Trabajos como el presente pueden contribuir a mejorar la capacidad de los profesionales de Atención Primaria para anticiparse a la aparición de eventos coronarios. El modelo CatBoost se ha revelado como una herramienta predictiva robusta, con un AUC de 0,979, lo que indica una capacidad discriminativa superior a la de muchos modelos previamente descritos en la literatura médica.

Además de su rendimiento, CatBoost ofrece ventajas prácticas: maneja variables categóricas de forma nativa, reduce la necesidad de preprocesamiento complejo y dispone de herramientas de interpretabilidad que permiten comprender los factores de riesgo clave, facilitando su integración en la práctica clínica. Esta combinación de precisión y transparencia refuerza su valor potencial como apoyo en la toma de decisiones clínicas en el primer nivel asistencial.

Por limitación en el tiempo, no se ha podido aplicar una validación más exhaustiva mediante técnicas como **k-fold cross – validation**, que permitiría valorar mejor el posible sobreajuste (overfitting) de los algoritmos y reforzar la robustez de los resultados. Este aspecto debería contemplarse en futuros trabajos para afianzar las conclusiones.

Asimismo, se han generado gráficas adicionales correspondientes al rendimiento de los distintos algoritmos evaluados. Aunque no se incluyen en el cuerpo principal del trabajo para no sobrecargar la sección de resultados, estas se han incorporado en un **anexo**, al que se remite para una visión más completa del proceso comparativo. [Anexo II].

7. LÍNEAS FUTURAS DE INVESTIGACIÓN

Para afianzar y expandir estos hallazgos se proponen varias líneas de trabajo:

- **Validación externa** en cohortes multicéntricas y diversas desde el punto de vista demográfico y epidemiológico.

- **Estudios piloto de implementación** en centros de salud para evaluar la usabilidad, la aceptación clínica y el impacto real en los resultados de salud.

- **Integración técnica mediante APIs seguras** en los sistemas de historia clínica electrónica, garantizando interoperabilidad y protección de datos.

- **Desarrollo de interfaces intuitivas** que muestren las predicciones y su explicación de manera clara y accionable dentro del flujo de trabajo habitual del médico.

VALORACIÓN CRÍTICA DEL AUTOR

El desarrollo de este TFM ha permitido comprobar cómo la **Inteligencia Artificial y el Big Data en Salud** pueden aportar valor real a la práctica clínica, siempre que se apliquen con rigor metodológico y manteniendo la perspectiva de utilidad práctica para los profesionales sanitarios.

El estudio presenta **limitaciones metodológicas importantes**, como son la dependencia de un único dataset y la falta de validación cruzada (k-fold), lo que obliga a interpretar los resultados con cautela.

El modelo debe ser considerado como una **propuesta inicial y prometedora**, que requiere consolidarse mediante validación en cohortes más amplias y diversas, así como estudios piloto en entornos reales de Atención Primaria. Solo entonces se podrá determinar su verdadero impacto en la prevención cardiovascular.

Este TFM refuerza la idea de que la IA no sustituye al juicio clínico, sino que lo **complementa y potencia**, abriendo posibilidades para una medicina más predictiva, preventiva y personalizada.

VALORACIÓN FINAL

El **infarto de miocardio** continúa siendo un desafío prioritario de salud pública, tanto a nivel mundial como en España, donde la carga de enfermedad sigue siendo elevada a pesar de los avances en la asistencia urgente y en la reducción de la mortalidad. La persistencia de factores de riesgo como la hipertensión, la diabetes, la obesidad y el tabaquismo evidencia la necesidad de reforzar estrategias de **prevención primaria** más efectivas.

En este contexto, el algoritmo **Catboost** ha demostrado ser una **herramienta predictiva robusta y clínicamente interpretable**. Su capacidad para manejar variables categóricas de forma nativa, junto con su interpretabilidad mediante métricas de importancia y análisis SHAP, lo posicionan como un modelo con gran potencial de **integración práctica en Atención Primaria.**

Podemos indicar que el modelo se proyecta como un aliado estratégico en la valoración del riesgo cardiovascular y en la mejora de la toma de decisiones clínicas, facilitando un abordaje más proactivo, preventivo y personalizado de la enfermedad coronaria.

Bibliografía

1. Aasheesh200. Framingham Heart Study dataset Kaggle; 2019. Disponible en: https://www.kaggle.com/datasets/aasheesh200/framingham-heart-study-dataset

2. Dawber TR, Meadors GF, Moore FE. Epidemiological approaches to heart disease: The Framingham Study. Am J Public Health Nations Health. 1951;41(3):279-86. doi:10.2105/AJPH.41.3.279

3. Framingham Heart Study. Cardiovascular disease risk prediction models. [Internet]. Disponible en: https://framinghamheartstudy.org/

4. World Health Organization (WHO). Cardiovascular diseases (CVDs) [Internet]. Geneva: WHO; 2021 Disponible en: https://www.who.int/news-room/fact-sheets/detail/cardiovascular-diseases-(cvds)

5. GBD 2019 Diseases and Injuries Collaborators. Global burden of 369 diseases and injuries in 204 countries and territories, 1990–2019: a systematic analysis for the Global Burden of Disease Study 2019. Lancet. 2020;396(10258):1204-22. doi:10.1016/S0140-6736(20)30925-9

6. Instituto Nacional de Estadística (INE). Defunciones según la Causa de Muerte. Año 2022 [Internet]. Madrid: INE; 2023. Disponible en: https://www.ine.es

7. Cequier Á, et al. Registro Codi IAM. Resultados de la atención del infarto agudo de miocardio en España. Rev Esp Cardiol. 2019;72(5):349-57. doi: 10.1016/j.recesp.2018.06.015

8. Sociedad Española de Cardiología (SEC). Informe RECALCAR 2023. Recursos y Calidad en Cardiología [Internet]. Madrid: SEC; 2023 Disponible en: https://secardiologia.es/recalcar

9. Sociedad Española de Cardiología (SEC). Las cardiopatías isquémicas agudas provocan 15.000 ingresos al año [Internet]. Madrid: SEC; 2022 [citado 12 sep 2025]. Disponible en: https://secardiologia.es

10. Patki N, Wedge R, Veeramachaneni K. The Synthetic Data Vault. In: 2016 IEEE Internacional Conference on Data Science and Advanced Analytics (DSAA). Montreal: IEEE; 2016. p. 399-410. doi: 10.1109/DSAA.2016.49

11. Van Buuren S, Groothuis-Oudshoorn K. mice: Multivariate imputation by chained equations in R. J Stat Softw. 2011;45(3):1-67. doi:10.18637/jss.v045.i03

12. Dorogush AV, Ershov V, Gulin A. CatBoost: gradient boosting with categorical features support. *arXiv preprint*. 2018; arXiv:1810.11363.

13. D'Agostino RB, Vasan RS, Pencina MJ, Wolf PA, Cobain M, Massaro JM, et al. General cardiovascular risk profile for use in primary care: the Framingham Heart Study. *Circulation*. 2008;117(6):743-53.

14. Conroy RM, Pyörälä K, Fitzgerald AP, Sans S, Menotti A, De Backer G, et al. Estimation of ten-year risk of fatal cardiovascular disease in Europe: the SCORE project. *Eur Heart J*. 2003;24(11):987-1003.

15. Hippisley-Cox J, Coupland C. Development and validation of QRISK3 risk prediction algorithms to estimate future risk of cardiovascular disease: prospective cohort study. *BMJ*. 2017;357:j2099.

16. Weng SF, Reps J, Kai J, Garibaldi JM, Qureshi N. Can machine-learning improve cardiovascular risk prediction using routine clinical data? *PLoS One*. 2017;12(4):e0174944.

17. Ambale-Venkatesh B, Yang X, Wu CO, Liu K, Hundley WG, McClelland R, et al. Cardiovascular event prediction by machine learning: the Multi-Ethnic Study of Atherosclerosis. *Circ Res*. 2017;121(9):1092-101.

18. Pedregosa F, Varoquaux G, Gramfort A, Michel V, Thirion B, Grisel O, et al. Scikit-learn: Machine learning in Python. J Mach Learn Res. 2011;12:2825-30.

19. Kunstmann S, Gaínza D. Estrategias de prevención y detección de factores de riesgo cardiovascular. Rev. Med Clin Condes. 2010;21(5):697-704.

20. Orozco-Beltrán D, Brotons Cuixart C, Banegas JR, et al. Recomendaciones preventivas cardiovasculares. Actualización PAPPS 2022. Aten Primaria. 2022;54(102444). doi: 10.1016/j.aprim.2022.102444

21. Organización Mundial de la Salud (OMS). Prevención de las enfermedades cardiovasculares: directrices sobre la evaluación y reducción del riesgo cardiovascular. Ginebra: OMS; 2007.

22. Ministerio de Salud de Argentina. Guía para la prevención de las enfermedades cardiovasculares. 1.ª ed. Buenos Aires: Ministerio de Salud; 2012.

23. Sociedad Española de Cardiología. Guía de práctica clínica para la prevención de las enfermedades cardiovasculares. Rev. Esp. Cardiol Supl. 2021;21(Esp):1-50.

24. Cabré JJ, Martín F, Costa B, et al. Metabolic syndrome as a cardiovascular disease risk factor: patients evaluated in primary care. BMC Public Health. 2008; 8:251. doi:10.1186/1471-2458-8-251

25. Ambroselli D, Masciulli F, Romano E, et al. New advances in metabolic syndrome, from prevention to treatment: the role of diet and food. Nutrients. 2023;15(640). doi:10.3390/nu15030640

26. Guía de prevención de enfermedades cardiovasculares para el primer nivel de atención. Ministerio de Salud de la Nación Argentina. 2ª ed. Buenos Aires; 2018.

27. American Heart Association. Heart Disease and Stroke Statistics—2023 Update. Circulation. 2023;147(8):e93–e210.

AGRADECIMIENTOS

Agradezco la formación recibida por parte del profesorado del Máster en Inteligencia Artificial y Big Data en Salud, así como la orientación y seguimiento continuo del tutor del trabajo, Remo Suppi Boldrito. También he de expresar toda mi gratitud a mi familia por su gran apoyo incondicional a lo largo de la formación y en el desarrollo de este TFM.

ANEXO I.

PREPROCESAMIENTO INICIAL

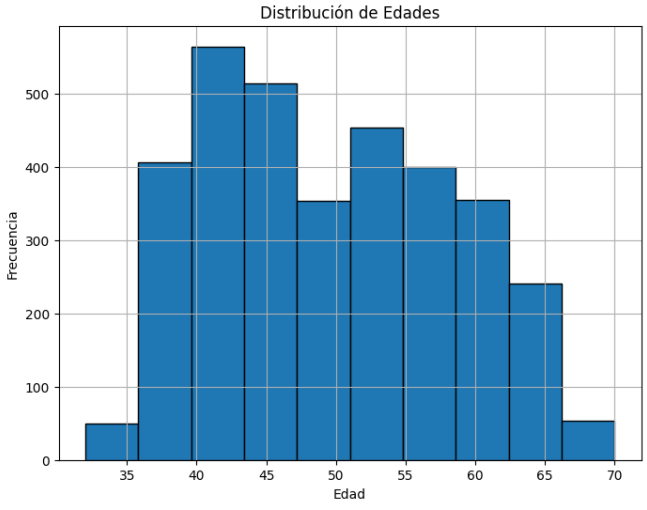
1. Visualización Dataset: data_cadiovascular_risk.csv (Python3)

	id	age	education	sex	is_smoking	cigsPerDay	...	sysBP	diaBP	BMI	heartRate	glucose	TenYearCHD
0	0	64	2.0	F	YES	3.0	...	148.0	85.0	NaN	90.0	80.0	1
1	1	36	4.0	M	NO	0.0	...	168.0	98.0	29.77	72.0	75.0	0
2	2	46	1.0	F	YES	10.0	...	116.0	71.0	20.35	88.0	94.0	0
3	3	50	1.0	M	YES	20.0	...	158.0	88.0	28.26	68.0	94.0	1
4	4	64	1.0	F	YES	30.0	...	136.5	85.0	26.42	70.0	77.0	0

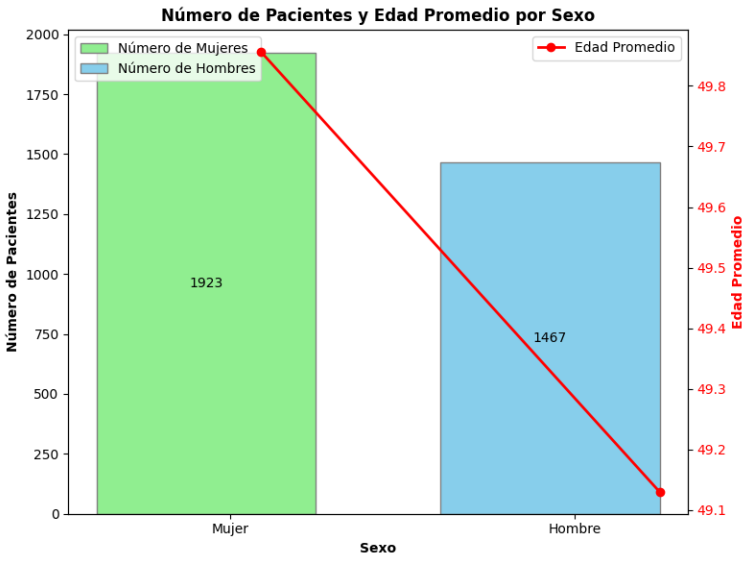
2. Revisión valores faltantes

```
>>> print(df.isnull().sum())
id                  0
age                 0
education          87
sex                 0
is_smoking          0
cigsPerDay         22
BPMeds             44
prevalentStroke     0
prevalentHyp        0
diabetes            0
totChol            38
sysBP               0
diaBP               0
BMI                14
heartRate           1
glucose           304
TenYearCHD          0
dtype: int64
```

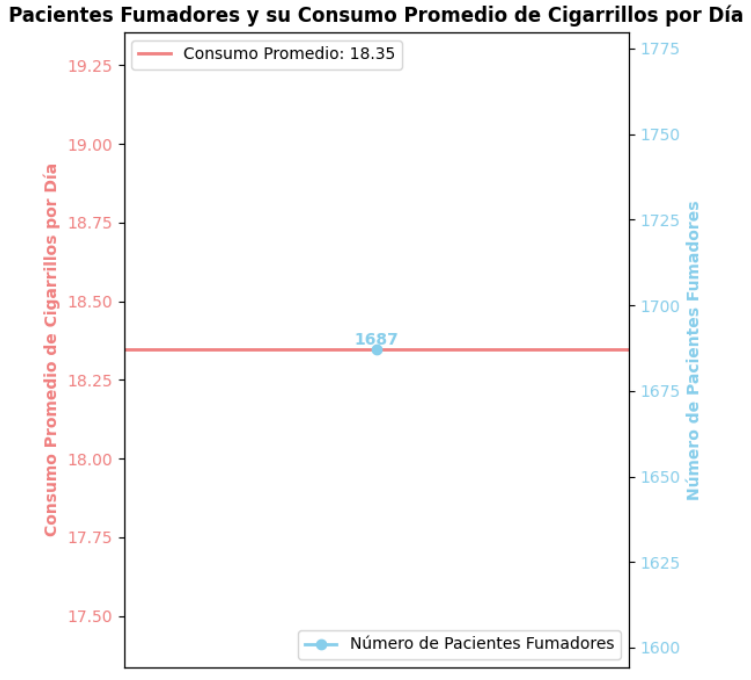
3. Distribución Edad (inicial)



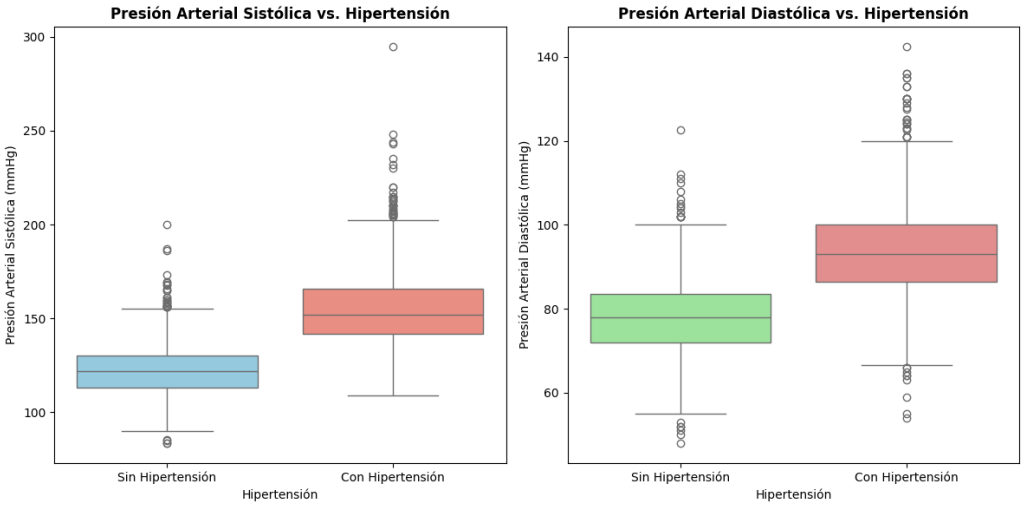
4. Número de pacientes y edad promedio por sexo



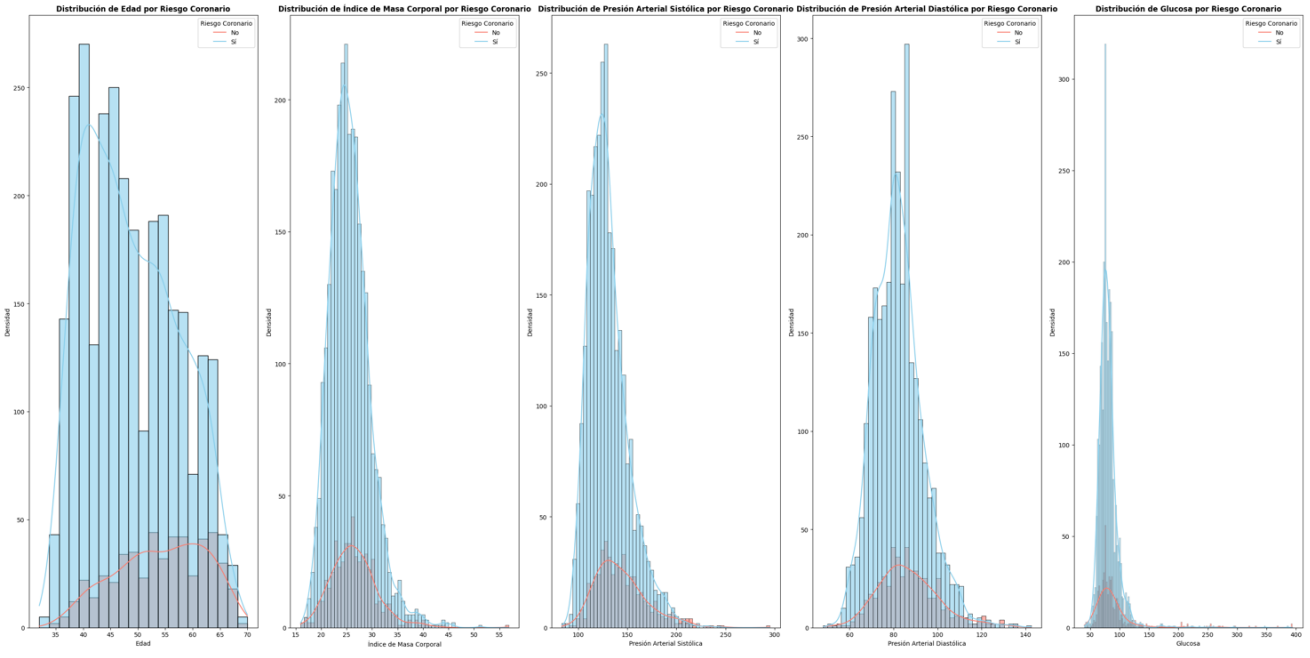
5. Pacientes Fumadores y su consumo promedio de cigarrillos por día



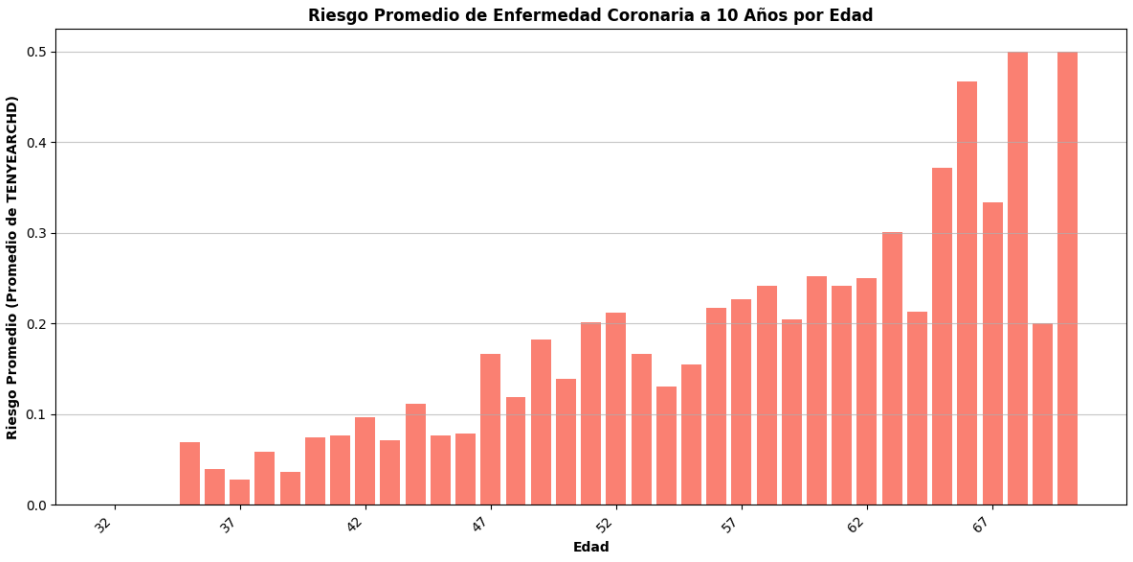
6. Box plot presión arterial sistólica / diastólica



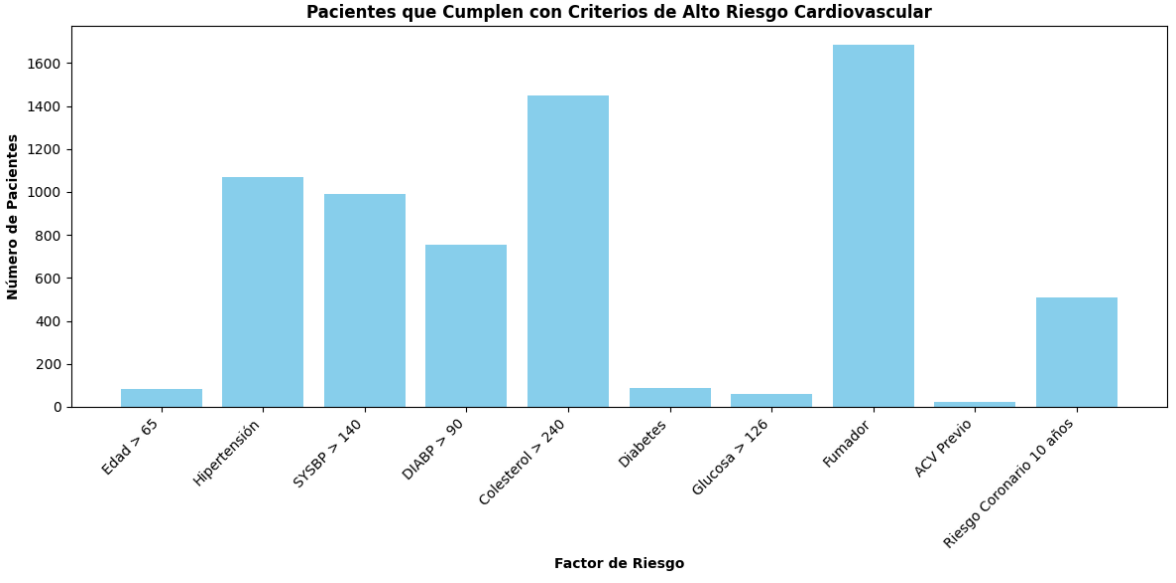
7. Distribución por variables



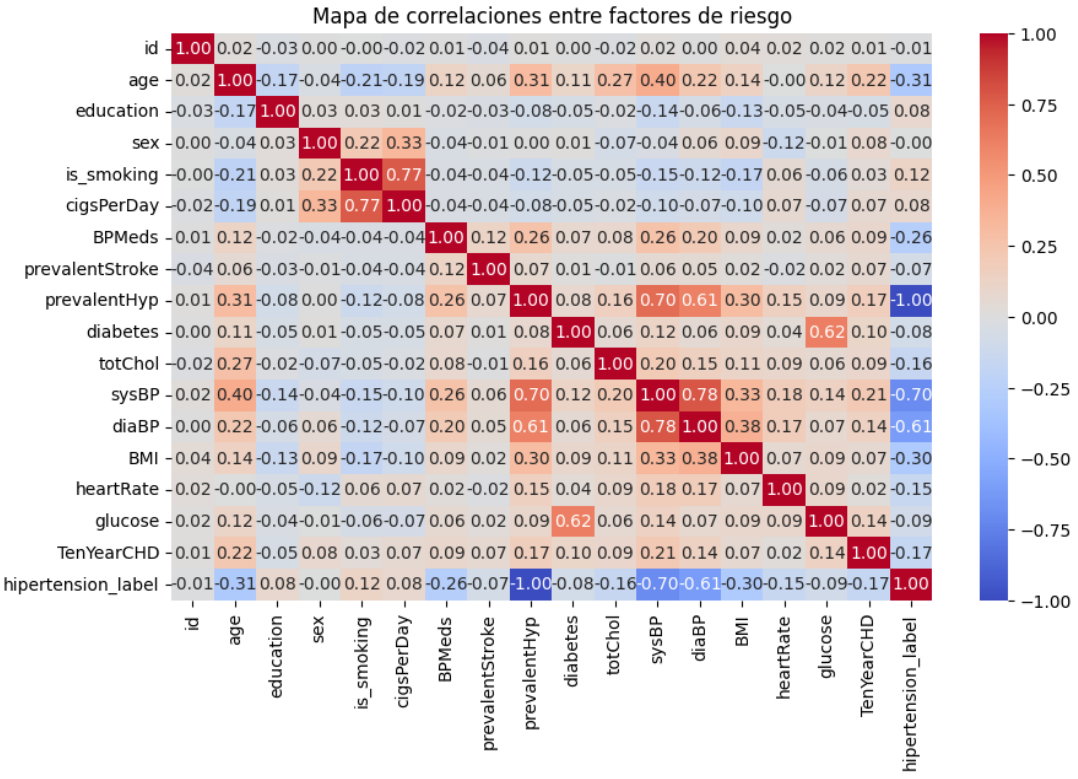
8. Riesgo promedio enfermedad coronaria a 10 años por Edad



9. Gráfico de criterios individuales



10. Matriz correlación (inicial)



11. Precisión y reporte de clasificación (inicial)

Precision: 0.8584070796460177

	precision	Recall	f1-score	Support
0	0.87	0.99	0.92	581
1	0.53	0.09	0.16	97
accuracy			0.86	678
macro avg	0.70	0.54	0.54	678
weighted avg	0.82	0.86	0.81	678

AUC-ROC: 0.7064

Recall: 0.0928

F1-score: 0.1579

XGBoost - AUC-ROC: 0.6512411945277428

XGBoost - Precisión: 0.8510324483775811

Gradient Boosting - AUC-ROC: 0.687882605532587

Gradient Boosting - Precisión: 0.8480825958702065

BALANCEAR CLASES. NUEVOS VALORES. SDV.

1. Nueva base de datos balanceada: data_cardiovascular_balanced.csv

	id	age	education	sex	is_smoking	cigsPerDay	BPMeds	prevalentStroke	prevalentHyp	diabetes	totChol	sysBP	diaBP	BMI	heartRate	glucose	TenYearCHD
0	2509	64	1.0	F	NO	0.0	0.0	0	1	0	312.0	160.0	82.0	27.59	140.0	94.0	0
1	3135	36	2.0	F	NO	0.0	0.0	0	0	0	209.0	107.0	73.5	21.59	75.0	73.0	0
2	1738	52	1.0	F	NO	0.0	0.0	0	0	0	245.0	131.0	80.0	32.04	80.0	81.0	0
3	330	41	2.0	M	YES	30.0	0.0	0	0	0	293.0	115.0	77.5	26.26	85.0	57.0	0
4	1878	60	1.0	M	YES	10.0	0.0	0	1	0	217.0	167.0	109.0	24.86	95.0	72.0	1

2. Resultados dos clases en TenYearCHD

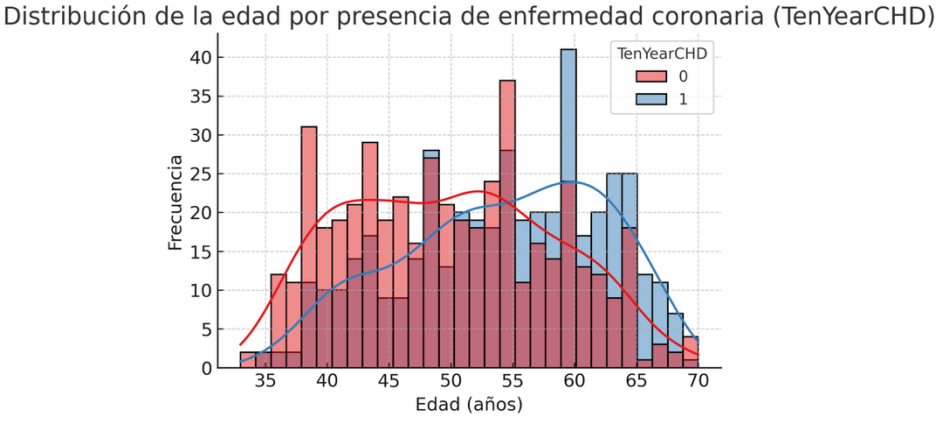
TenYearCHD	count
0	475
1	444

dtype: int64

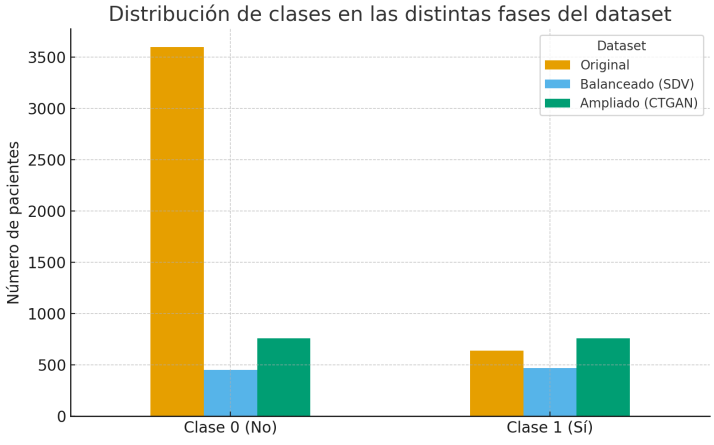
3. Posterior codificar variables categóricas y eliminar filas con valores faltantes

```
Clase 0 original: 405
Clase 1 original: 444
```

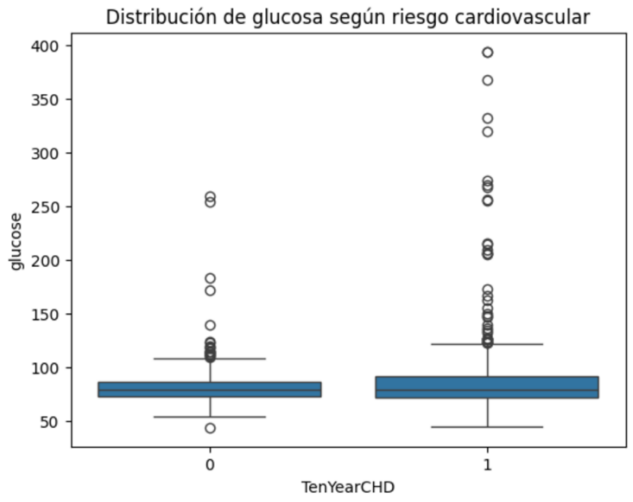
4. Comparación distribución por clase (TenYearCHD)



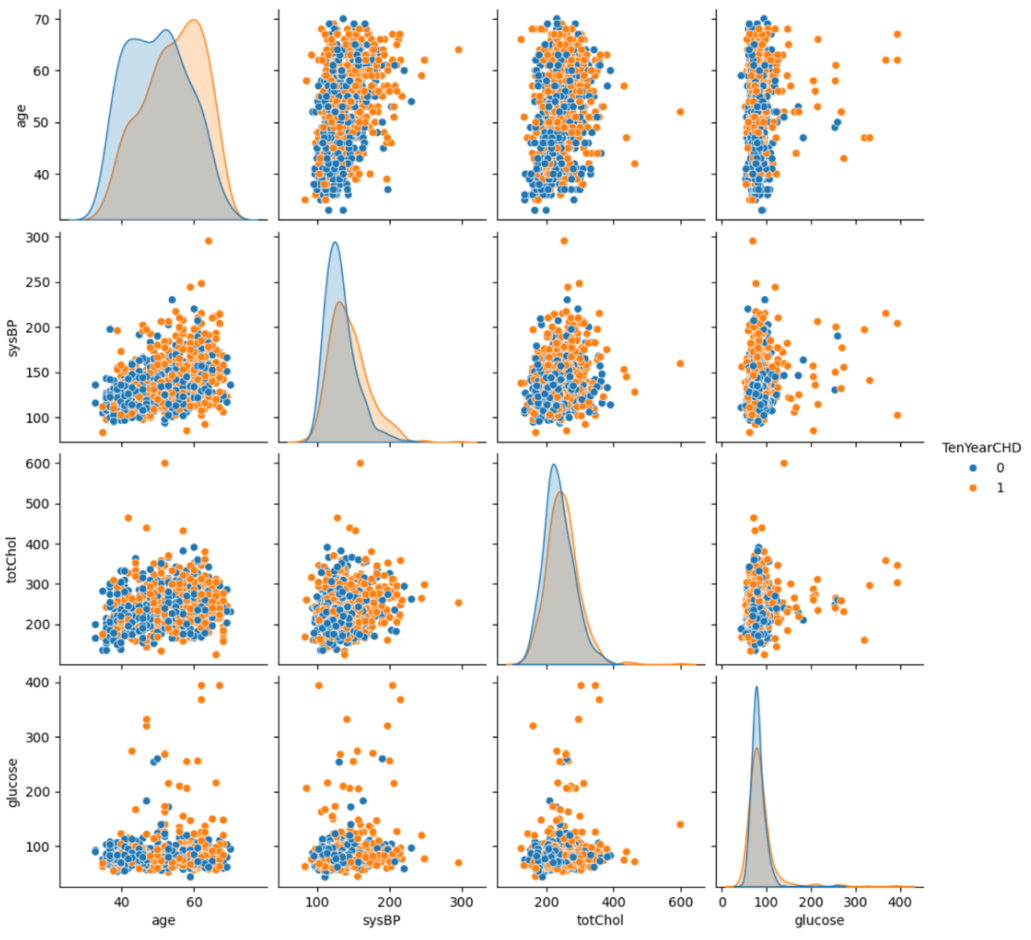
5. Distribución clases fases dataset



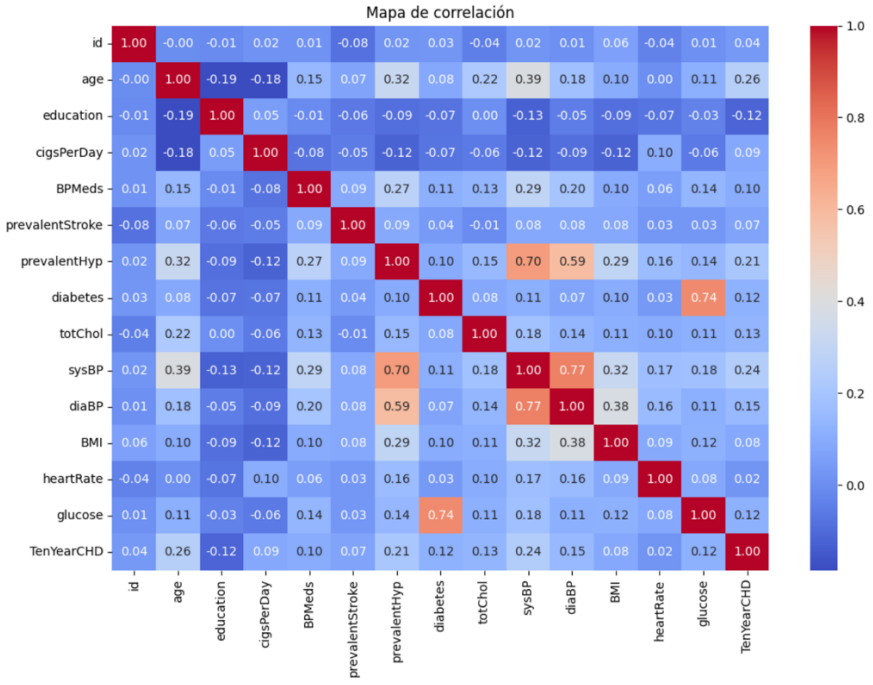
6. Distribución glucosa según riesgo cardiovascular



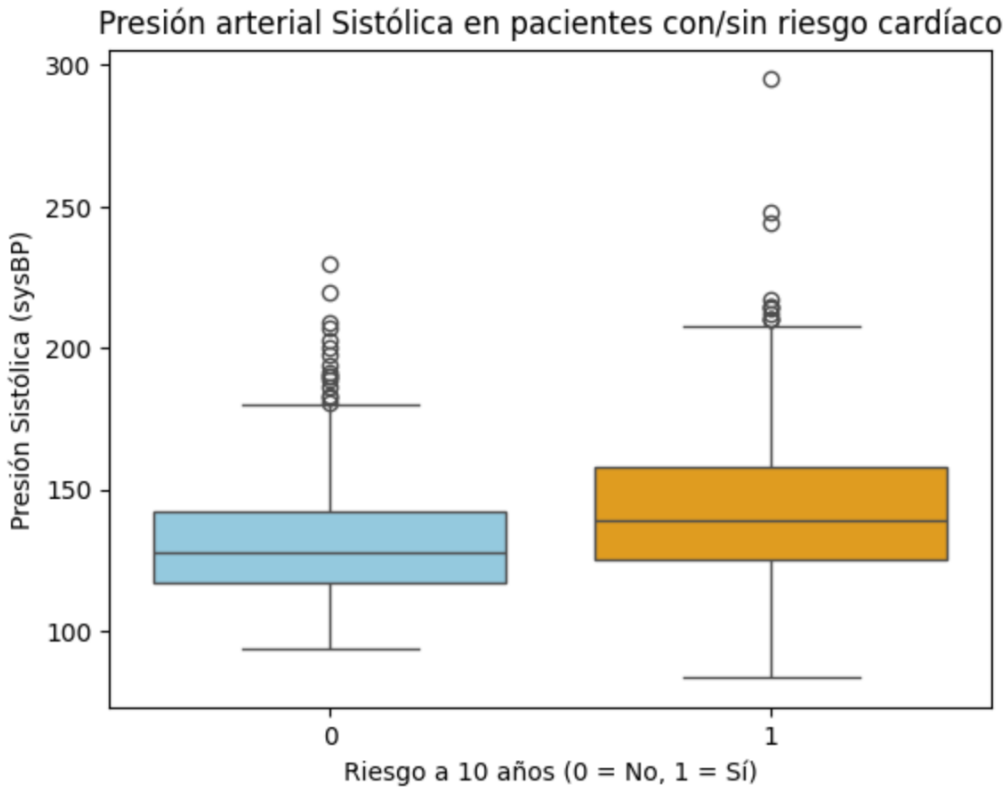
7. Pairplot (distribuciones cruzadas)



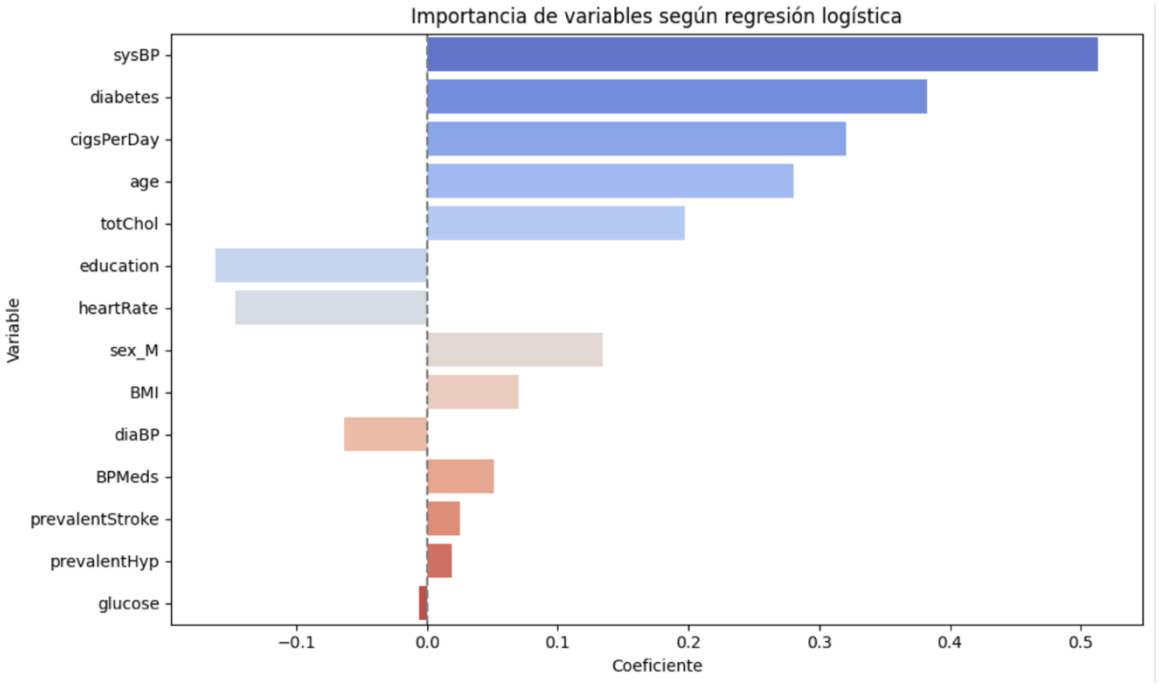
8. Correlación (heatmap)



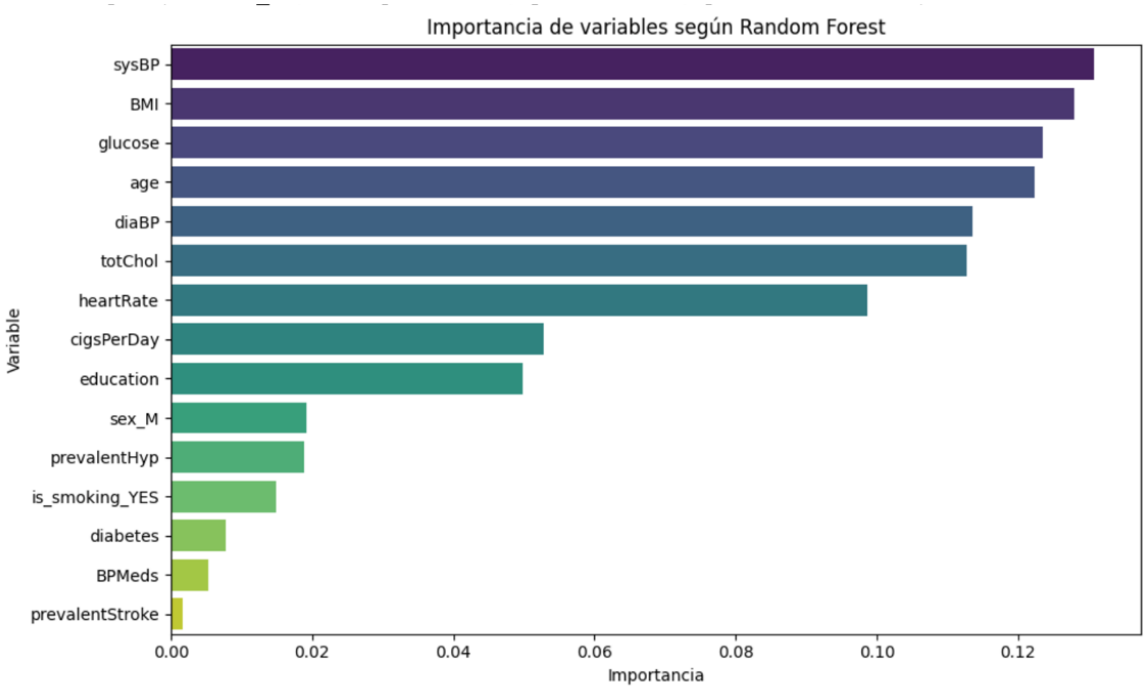
9. Presión arterial sistólica en pacientes con/sin riesgo cardíaco



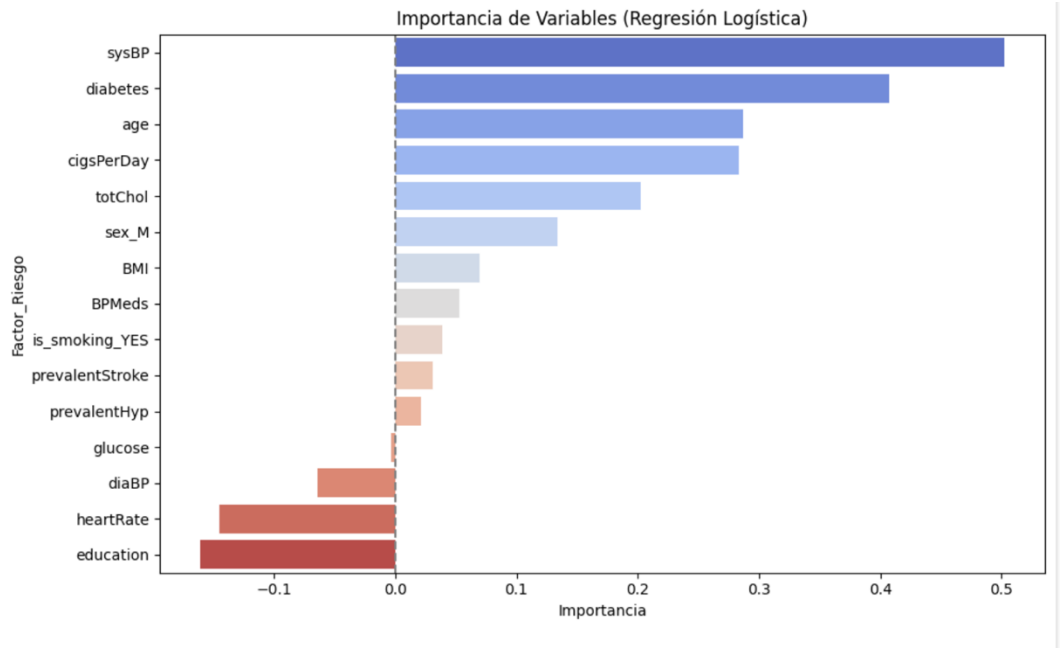
10. Importancia de las variables



11. Importancia variables según Random Forest



12. Gráfico importancia variables valoradas



13. Revaloración métrica por clases

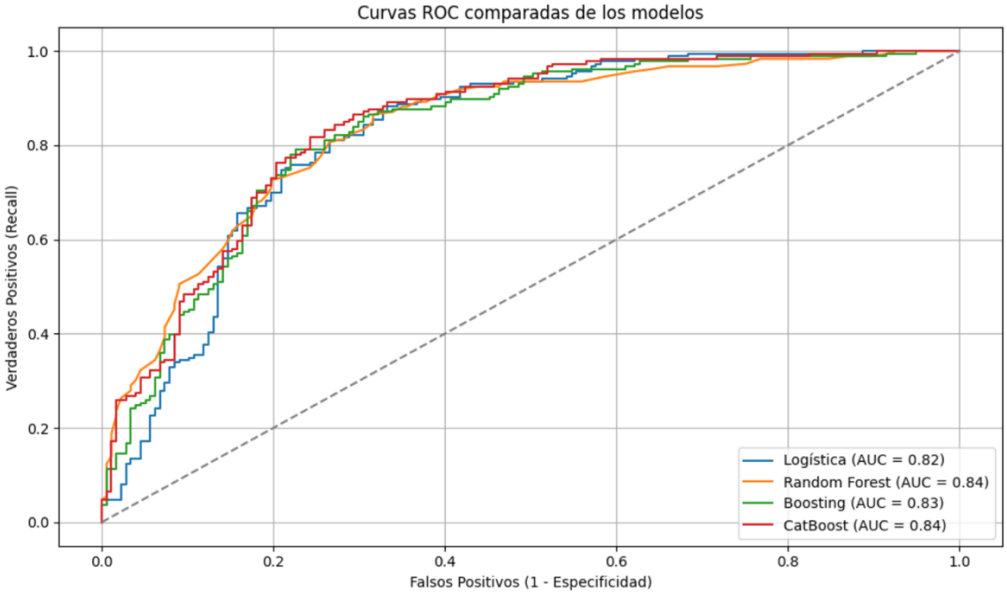
```
Matriz de Confusión:
 [[75 47]
 [38 95]]
```

	precision	recall	f1-score	support
0	0.66	0.61	0.64	122
1	0.67	0.71	0.69	133
accuracy			0.67	255
macro avg	0.67	0.66	0.66	255
weighted avg	0.67	0.67	0.67	255

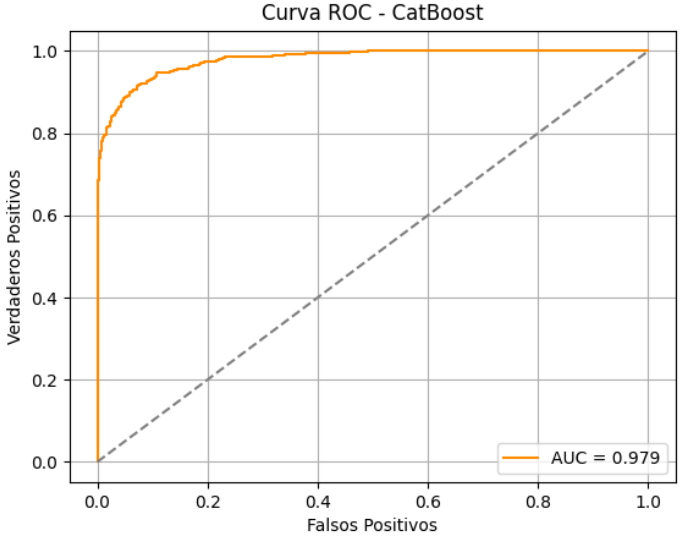
ANEXO II. GRÁFICAS DE LOS MODELOS EVALUADOS

1. Curvas ROC comparativas

Permiten evaluar la capacidad discriminativa de los algoritmos para predecir el riesgo cardiovascular a 10 años.



Métrica	Accuracy	Recall	Preci-sion	AUC
CatBoost	0.78	0.78	0.78	0.979



2. Matrices de confusión

Muestran el desempeño en términos de clasificaciones correctas e incorrectas (verdaderos positivos, falsos positivos, verdaderos negativos y falsos negativos).

Figura A1.

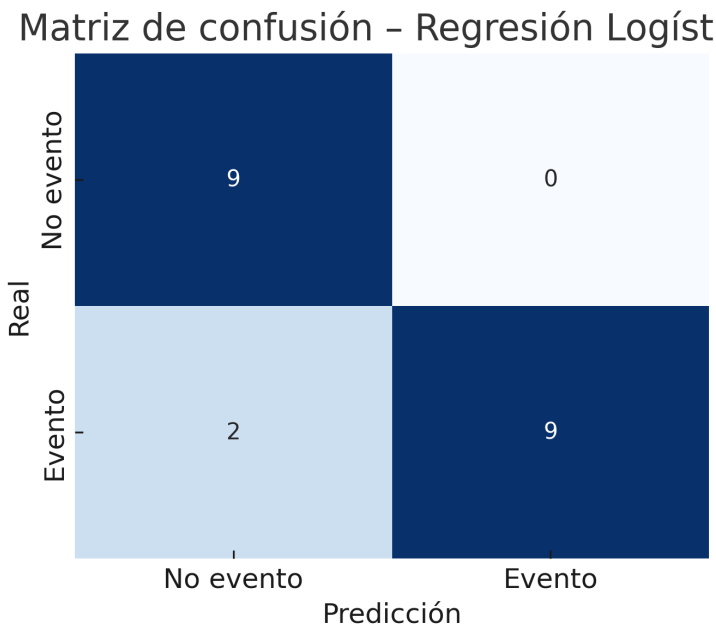


Figura A2.

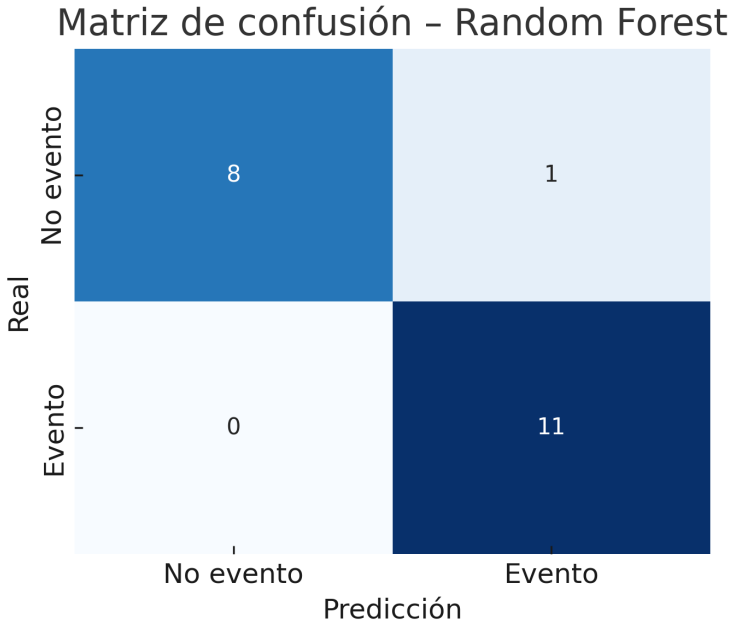


Figura A3.

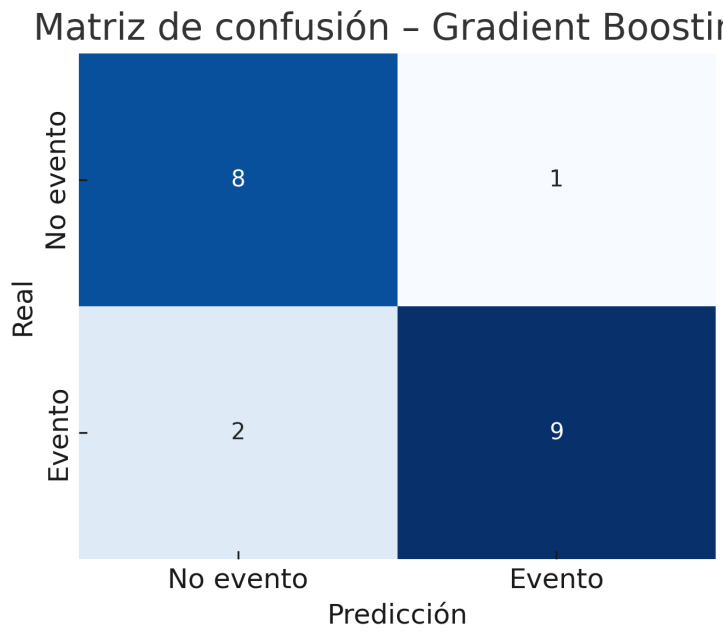
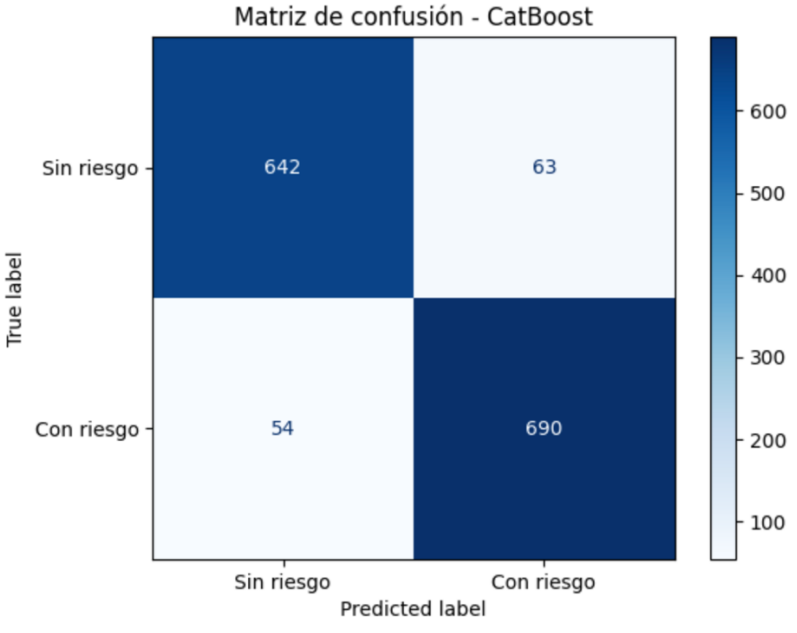


Figura A4.



3. Importancia de variables

Gráficos que ilustran qué variables aportaron mayor peso a la predicción en cada modelo, contribuyendo a su interpretación clínica.

Figura A9. Importancia de variables – Random Forest

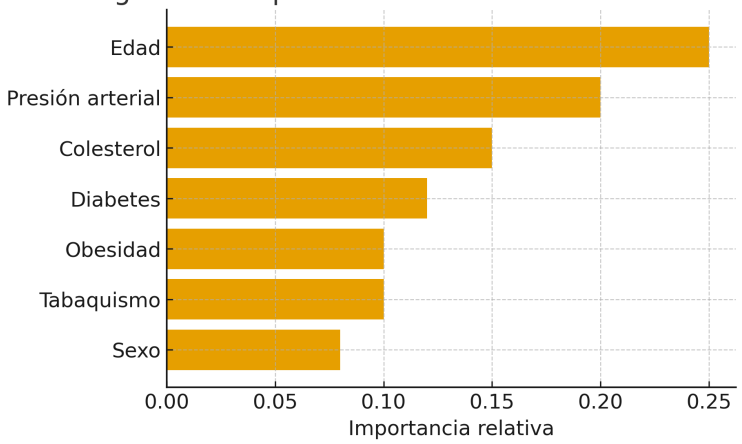


Figura A10. Importancia de variables – Gradient Boosting