IMMUNE TECHNOLOGY INSTITUTE

Memoria de Capstone Project

Predicción de Enfermedades Cardiovasculares basada en Factores de Riesgo

Máster en *Data Science*

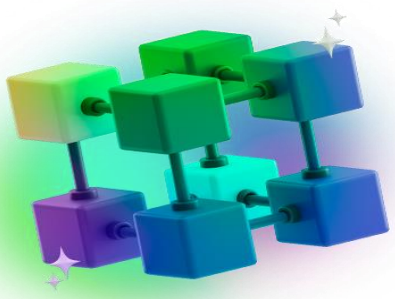
Modalidad *Online*

2023-2024

- Ana Isabel Cerro Luna
- Noelia Daniel Marín Serrano
- José Mojica Gutiérrez
- Eduardo del Río Ruiz

Tutor: Víctor Manuel Tenorio

Fecha de entrega: 13 de septiembre de 2024



OUTLINE

6. CONCLUSIONES Y FUTURO TRABAJO

Conclusiones, resumen y futuras mejoras.

7. ANEXOS

Código fuente del modelo

8. BIBLIOGRAFÍA

Referencias consultadas

Abstract:

Cardiovascular diseases (CVD) continue to be the leading cause of mortality worldwide, underscoring the urgent need for precise, scalable, and accessible predictive models. This study aims to address that need by developing a sophisticated predictive model using advanced machine learning techniques, leveraging comprehensive data from the Framingham Heart Study. By applying algorithms such as Random Forest, Neural Networks, and Logistic Regression, the model was designed to capture the complex interactions between multiple risk factors, including demographic, clinical, and lifestyle variables.

To tackle the inherent issue of class imbalance—commonly encountered in medical datasets—Synthetic Minority Over-sampling Technique (SMOTE) was employed. This technique ensured better representation of minority cases, significantly enhancing the model's ability to predict cardiovascular events in high-risk individuals, a crucial aspect in preventive health strategies.

Among the models evaluated, Random Forest demonstrated the highest performance, achieving an AUC-ROC score exceeding 0.96, thus proving its superiority in identifying high-risk individuals with greater accuracy than traditional models like Framingham. This research not only contributes to the development of more accurate predictive tools for cardiovascular risk but also explores the broader implications of machine learning in preventive medicine.

By facilitating early and personalized interventions, this approach holds the potential to transform clinical practice. The integration of machine learning into healthcare can lead to more targeted risk management strategies, optimizing the allocation of healthcare resources and improving patient outcomes on a larger scale. Furthermore, this study reflects a deep commitment to ethical considerations, particularly in ensuring that such predictive models are both interpretable and accessible, addressing not only accuracy but also the equity and fairness in their clinical application.

***Keywords**: Cardiovascular diseases, machine learning, Random Forest, Neural Networks, predictive modeling, SMOTE, preventive medicine, healthcare equity, Framingham Heart Study.*

"Medicine is a science of uncertainty and an art of probability."

— William Osler

IMMUNE Technology Institute
Paseo de la Castellana 89,
28046 Madrid

1. INTRODUCCIÓN

1.1. Motivación

En un escenario donde las enfermedades cardiovasculares (ECV) continúan siendo la principal causa de muerte a nivel mundial, la urgencia de desarrollar métodos predictivos más precisos y accesibles se vuelve cada vez más apremiante (1). Este proyecto se impulsa por la necesidad de integrar técnicas avanzadas de aprendizaje automático para revolucionar la prevención y el manejo de las ECV.

Mientras que los modelos predictivos tradicionales han sido herramientas valiosas, su capacidad para capturar la complejidad y la dinámica de los factores de riesgo en tiempo real es limitada. La investigación que presentamos aquí busca superar estas limitaciones al desarrollar un modelo que no solo refleje de manera más fiel la realidad multifacética de las ECV, sino que también personalice las intervenciones médicas.

El objetivo es proporcionar un enfoque preventivo más matizado y dinámico, que permita identificar y mitigar riesgos de manera anticipada y específica para cada individuo. Este proyecto tiene el potencial de no solo mejorar significativamente la calidad de vida de los pacientes mediante detecciones tempranas, sino también optimizar el uso de recursos en los sistemas de salud pública, reduciendo costos y mejorando la eficiencia en la atención sanitaria.

Este esfuerzo refleja un compromiso con la innovación científica y responde a un imperativo social de desarrollar tecnologías médicas que sean tanto preventivas como proactivas. A través de este trabajo, también se busca contribuir al avance del conocimiento en el campo de la epidemiología y la prevención de enfermedades, fortaleciendo el legado del Estudio de *Framingham* como una fuente inagotable de datos y descubrimientos relevantes para la salud pública global.

1.2. Objetivo general

Desarrollar un modelo predictivo utilizando técnicas de aprendizaje automático, capaz de estimar la probabilidad de incidencia de enfermedades cardiovasculares en individuos, basado en un conjunto de factores de riesgo biométricos, demográficos, clínicos y de estilo de vida.

1.3. Objetivos específicos

1. **Identificar** los factores de riesgo cardiovascular más relevantes, basados en la evidencia científica y la disponibilidad de datos, para su inclusión en el modelo predictivo.
2. **Preprocesar** los datos del conjunto seleccionado, abordando la distribución de los factores de riesgo, la clasificación entre variables y el manejo de valores faltantes o atípicos, para garantizar la calidad del conjunto de datos.
3. **Implementar** modelos predictivos utilizando distintas técnicas de aprendizaje automático, evaluando su rendimiento con el fin de seleccionar el modelo más preciso y confiable.
4. **Interpretar** los resultados del modelo seleccionado, evaluando la importancia de los factores de riesgo identificados y generando conocimientos útiles para la prevención y manejo clínico de las enfermedades cardiovasculares.

1.4. Introducción a las Enfermedades Cardiovasculares

Las ECV representan una de las principales causas de mortalidad a nivel global, contribuyendo a aproximadamente el 31% de todas las muertes en el mundo, según la Organización Mundial de la Salud (OMS). Estas afecciones, que incluyen el infarto de miocardio, la insuficiencia cardíaca y los accidentes cerebrovasculares, imponen una carga considerable tanto sobre los sistemas de salud como sobre la economía de los países afectados. La identificación temprana y la implementación de medidas preventivas son esenciales para reducir la tasa de mortalidad y mejorar la calidad de vida de los pacientes (2). La investigación sobre las ECV y su prevención ha sido un foco central en la medicina durante décadas, especialmente debido a su alta

prevalencia y el impacto significativo que tienen en la calidad de vida y en la sostenibilidad de los sistemas de salud. Las enfermedades cardiovasculares continúan siendo la principal causa de muerte a nivel mundial (3).

Desde 1969, enfermedades como la aterosclerosis, la cardiopatía coronaria y la enfermedad cerebrovascular han sido consistentemente reconocidas entre las principales causas de muerte en Estados Unidos. Aunque las tasas de mortalidad por ECV han disminuido notablemente desde la década de 1980, estas enfermedades siguen siendo las principales responsables de la discapacidad y la muerte prematura, subrayando la necesidad crítica de intervenciones tempranas y efectivas (4).

Los avances en la comprensión de los factores de riesgo, que abarcan desde aspectos genéticos hasta influencias ambientales y de estilo de vida, han mejorado notablemente las estrategias de manejo y prevención médica. Sin embargo, predecir con precisión las ECV sigue siendo un desafío, dado el complejo entramado e interacción de múltiples factores de riesgo (5). El *Framingham Heart Study h*a sido especialmente influyente en la identificación y cuantificación de estos factores, proporcionando una base sólida de conocimiento que ha moldeado significativamente nuestra comprensión de las ECV (6).

Epidemiología y Prevalencia

La epidemiología de las enfermedades cardiovasculares muestra que estas son responsables de más de 17 millones de muertes al año a nivel global, según la Organización Mundial de la Salud (OMS). Esta cifra representa aproximadamente el 31% de todas las muertes a nivel mundial, lo que subraya la importancia de identificar y gestionar eficazmente los factores de riesgo asociados (2). El *Framingham Heart Study* ha sido crucial para identificar numerosos factores de riesgo que afectan la salud del corazón y los vasos sanguíneos. Estos factores pueden ser clasificados en categorías biométricas, demográficas, clínicas y de estilo de vida, proporcionando una comprensión integral de cómo diferentes aspectos de la vida y la biología humana pueden influir en la probabilidad de desarrollar enfermedades del corazón (6).

A continuación, se detallan estos factores de riesgo, basados en las décadas de investigación y datos recopilados por el Estudio del Corazón de *Framingham.*

Factores de Riesgo

Los factores de riesgo para las enfermedades cardiovasculares son múltiples y diversos, y pueden clasificarse en varias categorías: biométricos, demográficos, clínicos y relacionados con el estilo de vida. Cada uno de estos factores contribuye de manera diferente al desarrollo de las ECV, y la comprensión de su impacto conjunto es crucial para la prevención y el tratamiento efectivo (7).

1. **Factores de Riesgo Biométricos**

Los factores de riesgo biométricos incluyen aquellas mediciones biológicas que pueden predisponer a enfermedades cardiovasculares. Algunos de los factores clave identificados en el Estudio del Corazón de *Framingham* son:

- **Presión arterial alta**: La hipertensión es un factor de riesgo significativo para la enfermedad coronaria y el accidente cerebrovascular. Estudios como el de Framingham han demostrado que tanto la presión arterial sistólica como la diastólica elevadas son predictores importantes de enfermedad cardiovascular.

- **Colesterol**: Niveles elevados de colesterol total y lipoproteína de baja densidad (LDL) están asociados con un mayor riesgo de enfermedad coronaria. La arteriosclerosis, una condición en la que las arterias se endurecen y se estrechan debido a la acumulación de placa, es una consecuencia directa de niveles altos de LDL.

- **Índice de masa corporal (IMC)**: Un IMC alto, indicador de sobrepeso u obesidad, se asocia con un mayor riesgo de hipertensión, dislipidemia, diabetes tipo 2 y enfermedades cardiovasculares. La obesidad contribuye a la

inflamación crónica y al estrés oxidativo, factores clave en la patogénesis de las enfermedades cardíacas.

- **Glucosa en sangre**: Niveles elevados de glucosa en sangre o la diabetes mellitus incrementa el riesgo de desarrollar enfermedades del corazón. La hiperglucemia crónica puede dañar los vasos sanguíneos y los nervios del corazón, aumentando el riesgo de infarto de miocardio y enfermedad arterial periférica.

2. Factores de Riesgo Demográficos

Los factores demográficos incluyen características personales y antecedentes familiares que afectan la probabilidad de desarrollar enfermedades cardíacas. Entre ellos se encuentran:

- **Edad**: A medida que envejecemos, el riesgo de sufrir enfermedades cardiovasculares aumenta. Esto se debe a que, con el tiempo, las arterias tienden a endurecerse y estrecharse, y el corazón puede volverse menos eficiente. En particular, las personas mayores de 65 años tienen una probabilidad significativamente mayor de desarrollar estas enfermedades.
- **Sexo**: Los hombres suelen tener un riesgo mayor de desarrollar enfermedades cardíacas a edades más tempranas que las mujeres. Sin embargo, después de la menopausia, el riesgo en las mujeres aumenta y puede llegar a igualar o incluso superar el de los hombres. Esto ocurre en parte debido a la disminución de los niveles de estrógenos, que ayudan a proteger el sistema cardiovascular.
- **Historial familiar**: Tener antecedentes familiares de enfermedades del corazón es un indicador importante del riesgo personal. La genética influye considerablemente, y ciertas condiciones como la hipertensión, la diabetes y el colesterol alto pueden transmitirse de generación en generación.
- **Etnicidad**: Algunas etnias tienen un mayor riesgo de desarrollar enfermedades

cardiovasculares debido a una combinación de factores genéticos y ambientales. Por ejemplo, los afroamericanos tienen una mayor prevalencia de hipertensión y diabetes, lo que eleva su riesgo de sufrir problemas cardíacos.

3. **Factores de Riesgo Clínicos**

Estos factores incluyen condiciones médicas y antecedentes de salud que pueden aumentar la susceptibilidad a enfermedades cardíacas:

- **Enfermedades previas**: La presencia de enfermedades cardiovasculares previas, como insuficiencia cardíaca, fibrilación auricular, o enfermedad arterial periférica, incrementa significativamente el riesgo de futuros eventos cardiovasculares.

- **Síndrome metabólico**: Este síndrome se caracteriza por una combinación de hipertensión, dislipidemia, obesidad abdominal y resistencia a la insulina. La presencia del síndrome metabólico duplica el riesgo de enfermedad cardiovascular.

- **Hipertrofia ventricular izquierda**: Este engrosamiento de las paredes del corazón, detectado mediante electrocardiograma o ecocardiograma, es un indicador importante de riesgo cardiovascular, ya que refleja el estrés y la carga crónica sobre el corazón debido a la hipertensión.

4. **Factores de Riesgo del Estilo de Vida**

El estilo de vida también juega un papel crucial en la salud cardiovascular. Los factores de riesgo más destacados son:

- **Tabaquismo**: Fumar es uno de los principales factores de riesgo modificables para la enfermedad cardíaca y el accidente cerebrovascular. El tabaquismo daña las arterias, reduce el nivel de oxígeno en la sangre y aumenta la presión arterial y la frecuencia cardíaca.

- **Actividad física**: La falta de ejercicio regular está asociada con un mayor riesgo de enfermedades del corazón. El ejercicio regular ayuda a mantener un peso saludable, reduce la presión arterial, mejora los niveles de colesterol y aumenta la eficiencia del corazón y los pulmones.

- **Dieta**: Una dieta rica en grasas saturadas, grasas trans y azúcares, y baja en frutas, verduras y fibra, contribuye significativamente al riesgo cardiovascular. Las dietas saludables, como la dieta mediterránea, que es rica en frutas, verduras, pescado y aceite de oliva, están asociadas con un menor riesgo de enfermedades del corazón

- **Consumo de alcohol**: El consumo excesivo de alcohol puede aumentar la presión arterial y contribuir a enfermedades cardíacas. Sin embargo, el consumo moderado (por ejemplo, una copa de vino al día) puede tener un efecto protector para algunas personas, aunque este beneficio es aún debatido.

- **Estrés**: El estrés crónico y la falta de manejo adecuado del mismo pueden aumentar el riesgo de enfermedades cardiovasculares. El estrés puede llevar a comportamientos poco saludables como el consumo excesivo de alcohol, la mala alimentación y el sedentarismo, y también puede tener efectos directos sobre el corazón y los vasos sanguíneos (6,8,9).

1.5. Modelo *Framingham*

1.5.1. Historia y Desarrollo del Estudio *Framingham*

El Estudio del Corazón de *Framingham* es uno de los estudios epidemiológicos más influyentes en la historia de la medicina cardiovascular. Iniciado en 1948 en *Framingham, Massachusetts*, este estudio longitudinal se diseñó para identificar los factores comunes que contribuyen a las enfermedades cardiovasculares en una población sin signos evidentes de enfermedad cardiovascular al inicio del estudio. Desde entonces, ha proporcionado datos invaluables que han moldeado las prácticas de prevención y tratamiento de las enfermedades cardiovasculares.

El diseño inicial del estudio *Framingham* involucró a una cohorte original de 5,209 hombres y mujeres de entre 30 y 62 años, quienes no presentaban signos evidentes de enfermedades cardiovasculares en el momento de la inscripción. Los participantes fueron seleccionados de manera representativa de la población general de *Framingham*, lo que permitió que los resultados fueran aplicables a una amplia variedad de individuos. Los participantes han sido seguidos de manera continua, con exámenes físicos detallados y entrevistas cada dos años para recoger datos sobre su salud cardiovascular y otros factores relevantes (10).

Uno de los aspectos más innovadores del diseño del estudio *Framingham* es su enfoque en la recolección de datos de múltiples factores de riesgo a lo largo del tiempo. Desde el principio, se recogieron datos sobre presión arterial, niveles de colesterol, hábitos de tabaquismo, actividad física, y otros factores de salud y estilo de vida. Con el tiempo, el estudio se amplió para incluir biomarcadores emergentes y datos genéticos, lo que permitió un análisis más profundo de cómo estos factores interactúan entre sí y contribuyen al desarrollo de enfermedades cardiovasculares (11).

Además, investigaciones derivadas del *Framingham Heart Study* han demostrado que la diabetes mellitus triplica la mortalidad cardiovascular y está asociada con un riesgo sustancialmente mayor de insuficiencia y enfermedad cardíacas hipertensiva. También se ha establecido una relación inversa entre las concentraciones de HDL (colesterol bueno) y la incidencia de enfermedad coronaria, mientras que las concentraciones elevadas de LDL (colesterol malo) están positivamente asociadas con la enfermedad coronaria (12).

El estudio también ha destacado el impacto de la obesidad en el riesgo cardiovascular. A partir de la década de 1980, se reportó que el aumento de peso incrementa significativamente el riesgo de enfermedad cardiovascular, incluso después de ajustar por otros factores de riesgo. Este riesgo es particularmente evidente en el caso de la insuficiencia cardíaca, donde los participantes de *Framingham* menores de 50 años presentaron un riesgo de dos a tres veces mayor de insuficiencia cardíaca al comparar las categorías de peso (11).

El estudio *Framingham* se ha adaptado y expandido a lo largo de los años para incluir nuevas cohortes, lo que ha permitido a los investigadores observar cambios generacionales en la prevalencia y el impacto de los factores de riesgo cardiovascular. En 1971, se inició la *Framingham Offspring Study*, que incluyó a 5,124 hijos adultos de la cohorte original y sus cónyuges. Posteriormente, en 2002, se lanzó la Tercera Generación de Cohortes, que incluyó a los nietos de la cohorte original. Este enfoque intergeneracional ha sido clave para comprender cómo los factores de riesgo se transmiten y evolucionan a lo largo del tiempo (10).

El seguimiento exhaustivo y continuo de los participantes del estudio *Framingham* ha permitido la identificación de varios factores de riesgo cardiovascular cruciales, como la hipertensión, el colesterol elevado, el tabaquismo, la obesidad y la diabetes (11). Estos hallazgos han sido fundamentales para el desarrollo de guías clínicas y políticas de salud pública dirigidas a la prevención de enfermedades cardiovasculares.

Además, el estudio ha sido la base para el desarrollo de varios modelos de predicción de riesgo, como el famoso Modelo de Riesgo de *Framingham*, que sigue siendo utilizado en la práctica clínica para estimar el riesgo de eventos cardiovasculares en un período de 10 años (10 ,12).

En resumen, el diseño del estudio *Framingham* ha sido excepcionalmente eficaz para proporcionar una comprensión integral y detallada de las enfermedades cardiovasculares. Su enfoque longitudinal, la inclusión de múltiples generaciones y la capacidad de adaptarse a nuevas tecnologías y descubrimientos científicos lo han convertido en un estudio de referencia en la epidemiología cardiovascular. La riqueza y profundidad de los datos recopilados continúan ofreciendo *insights* valiosos, no solo para la prevención y tratamiento de las enfermedades del corazón, sino también para la comprensión de otras condiciones crónicas relacionadas con el envejecimiento y los estilos de vida modernos (10-12).

1.5.2. Parámetros y Factores de Riesgo Utilizados en el Modelo

El modelo de riesgo de *Framingham* utiliza una combinación de factores para predecir la probabilidad de que una persona desarrolle enfermedades cardiovasculares en un período de 10 años. Los factores utilizados en este modelo incluyen la edad, el sexo, los niveles de colesterol, la presión arterial, el tabaquismo, la diabetes y el estado hipertensivo. Estos parámetros se han validado en numerosas poblaciones y han demostrado ser predictores robustos del riesgo cardiovascular (13). Profundizaremos en la comprensión de estos factores más adelante.

1.5.3. Aplicaciones y Limitaciones del Modelo *Framingham*

El modelo de *Framingham* ha sido ampliamente utilizado para estimar el riesgo de enfermedades cardiovasculares en la práctica clínica. Sin embargo, como cualquier modelo, tiene limitaciones. Una de las principales críticas es que fue desarrollado en una población mayoritariamente caucásica, lo que puede limitar su aplicabilidad en

otras poblaciones étnicas. Además, el modelo se basa en datos que pueden no capturar todos los aspectos del riesgo cardiovascular en la población actual, especialmente considerando los cambios en el estilo de vida y las intervenciones médicas desde que se inició el estudio (14).

2. Fundamentación Teórica

Ciencia de Datos en la Predicción de Enfermedades Cardiovasculares

La ciencia de datos ha emergido como una disciplina clave en la investigación y desarrollo de soluciones para problemas complejos en el ámbito de la salud. Su capacidad para analizar grandes volúmenes de datos y detectar patrones ocultos permite a los profesionales médicos y científicos avanzar en áreas como la predicción, prevención y tratamiento de enfermedades crónicas. Las enfermedades cardiovasculares, que siguen siendo la principal causa de mortalidad a nivel mundial, representan un área crítica en la que las herramientas de la ciencia de datos pueden ofrecer un impacto significativo.

En los últimos años, la ciencia de datos ha revolucionado una amplia variedad de campos, y la medicina no ha sido la excepción. Gracias a su capacidad para manejar grandes volúmenes de información y descubrir patrones ocultos, la ciencia de datos se ha consolidado como una herramienta esencial en la investigación y prevención de enfermedades. Entre las áreas donde su impacto ha sido más evidente se encuentra la predicción de enfermedades crónicas, como las cardiovasculares, que siguen siendo la principal causa de muerte a nivel mundial.

La integración de técnicas avanzadas de análisis de datos en el ámbito médico permite no solo una mejor comprensión de los factores de riesgo, sino también la creación de modelos predictivos más precisos. Estos modelos ayudan a los profesionales de la salud a anticiparse a posibles complicaciones y tomar decisiones informadas sobre intervenciones preventivas. La ciencia de datos ofrece la posibilidad de combinar información clínica, demográfica y de estilo de vida para desarrollar predicciones más personalizadas y adaptadas a las características de cada paciente.

En este contexto, el presente trabajo se centra en la aplicación de la ciencia de datos para desarrollar un modelo predictivo que permita identificar con mayor precisión el riesgo de enfermedades cardiovasculares. A través del uso de técnicas modernas de análisis de datos, se busca proporcionar herramientas más efectivas para la

prevención, permitiendo una medicina más proactiva y centrada en la identificación temprana de riesgos.

Este enfoque no solo refleja el potencial de la ciencia de datos para mejorar la calidad de vida de los pacientes, sino que también subraya su capacidad para optimizar los recursos en los sistemas de salud. La medicina predictiva, impulsada por la ciencia de datos, tiene el potencial de cambiar el enfoque de la atención médica, pasando de la intervención tardía a la prevención temprana, con el objetivo de reducir la morbilidad y la mortalidad asociadas a enfermedades como las cardiovasculares.

Existen varios modelos de predicción utilizados en el ámbito médico, particularmente para el diagnóstico y manejo de enfermedades cardiovasculares.
Estos modelos pueden basarse en métodos estadísticos tradicionales, así como en enfoques más modernos de inteligencia artificial y machine learning. Estos modelos son herramientas poderosas para predecir el riesgo de enfermedades cardiovasculares, ayudando a los médicos a tomar decisiones más informadas y personalizadas en la atención de sus pacientes. La elección del modelo depende de la disponibilidad de datos, la población objetivo y la complejidad de la tarea de predicción. Sin embargo, los modelos basados en Inteligencia Artificial (I.A) y *Machine Learning (ML)* son cruciales para el diagnóstico de enfermedades cardiovasculares, ya que mejoran la precisión, permiten la personalización del tratamiento, optimizan recursos y tienen el potencial de revolucionar tanto la práctica clínica como la investigación médica en esta área.

El uso de modelos basados en I.A y ML en el diagnóstico de enfermedades cardiovasculares ha transformado la manera en que se abordan este tipo de patologías.

Entre los principales modelos basados en I.A y *ML* utilizados en el diagnóstico de enfermedades cardiovasculares destacan las redes neuronales y los árboles de decisión. Ambos modelos son fundamentales en la medicina moderna, ya que no

solo optimizan el diagnóstico, sino que también facilitan la personalización de los tratamientos y la prevención de futuras complicaciones cardiovasculares.

2.1. Redes Neuronales Artificiales en la Predicción de Enfermedades Cardiovasculares

2.1.1. Introducción a las RNA

Las RNA son sistemas de aprendizaje automático inspirados en la estructura y función del cerebro humano (15). Están compuestas por unidades básicas llamadas "neuronas", organizadas en capas, donde cada neurona de una capa está conectada con las neuronas de la capa siguiente. Estas conexiones se ponderan y ajustan durante el proceso de entrenamiento de la red, permitiendo que la RNA aprenda patrones complejos en los datos y realice predicciones precisas.

El auge de las RNA en la medicina se debe a su capacidad para manejar grandes volúmenes de datos *(Big Data)* y modelar relaciones no lineales entre múltiples variables. A diferencia de los modelos estadísticos tradicionales, que a menudo asumen una relación lineal entre las variables independientes y la dependiente, las RNA pueden capturar interacciones complejas y no lineales, lo que las hace especialmente útiles en la predicción de enfermedades donde múltiples factores de riesgo interactúan de manera intrincada (16,17).

2.1.2. Estructura y Funcionamiento de las RNA

Las RNA constan de tres tipos principales de capas: la capa de entrada, donde se introducen los datos; las capas ocultas, donde se realizan las operaciones más complejas; y la capa de salida, donde se obtiene la predicción final. Cada conexión entre neuronas tiene un peso asociado que se ajusta durante el entrenamiento de la red, utilizando algoritmos como el descenso de gradiente, con el objetivo de minimizar el error entre la predicción de la RNA y los resultados reales (16).

Una característica distintiva de las RNA es su capacidad para generalizar a partir de datos de entrenamiento, lo que les permite realizar predicciones sobre nuevos datos no vistos. Esto se logra mediante un proceso de aprendizaje supervisado, donde la red es entrenada con un conjunto de datos etiquetados, o aprendizaje no supervisado, donde la red identifica patrones intrínsecos en los datos sin guía externa (15,16).

2.1.3. Aplicación de RNA en la Predicción de Enfermedades Cardiovasculares

Las RNA han mostrado un gran potencial en la predicción de enfermedades cardiovasculares. En estudios recientes, se han utilizado para mejorar la precisión de los modelos tradicionales, como el de *Framingham*, al incorporar variables adicionales y capturar relaciones no lineales. Estas redes han demostrado ser eficaces en la identificación de individuos en alto riesgo, lo que permite una intervención más temprana y personalizada.

Por ejemplo, las RNA pueden integrar datos de diferentes fuentes, como historial médico, genética, comportamiento y datos de estilo de vida, para proporcionar una evaluación de riesgo más completa. Además, se ha demostrado que pueden adaptarse mejor a diferentes poblaciones, superando algunas de las limitaciones de los modelos tradicionales, como la falta de generalización a grupos étnicos diversos (17, 18).

2.1.4. Comparación entre el Modelo *Framingham* y las Redes Neuronales Artificiales

Precisión y Eficacia Predictiva

El modelo de *Framingham* ha sido una herramienta fundamental en la predicción del riesgo cardiovascular durante décadas (13). Sin embargo, su enfoque lineal y la

dependencia de un número limitado de variables pueden limitar su capacidad predictiva en comparación con las RNA. Las RNA, al poder manejar grandes conjuntos de datos y captar interacciones complejas entre variables, han mostrado una mayor precisión en la predicción de eventos cardiovasculares en estudios comparativos (17).

En términos de precisión, las RNA han demostrado una capacidad superior para clasificar correctamente a los pacientes en categorías de riesgo bajo, moderado o alto, reduciendo así tanto los falsos positivos como los falsos negativos. Esto es especialmente importante en la práctica clínica, donde la subestimación o sobreestimación del riesgo puede tener consecuencias graves (18).

Flexibilidad y Adaptabilidad

Otra ventaja clave de las RNA sobre el modelo de *Framingham* es su flexibilidad. Las RNA pueden ser entrenadas con nuevos datos a medida que están disponibles, permitiendo que el modelo se actualice y se adapte a los cambios en los factores de riesgo, los avances médicos y las variaciones en las poblaciones estudiadas (15, 16). El modelo de *Framingham*, por otro lado, está basado en datos históricos y puede no reflejar adecuadamente los cambios en la epidemiología y los tratamientos disponibles (13, 14).

Complejidad y Desafíos en la Implementación

A pesar de sus ventajas, las RNA también presentan desafíos significativos en su implementación. La complejidad de estos modelos requiere una gran cantidad de datos para entrenarlos adecuadamente, así como recursos computacionales considerables. Además, las RNA son a menudo vistas como "cajas negras", ya que es difícil interpretar cómo se toman las decisiones dentro del modelo, lo que puede limitar su aceptación en la práctica clínica, donde la transparencia y la interpretabilidad son cruciales (16).

Por otro lado, el modelo de *Framingham*, aunque menos preciso en algunos

contextos, es mucho más fácil de implementar y comprender. Su simplicidad lo hace accesible para los médicos y otros profesionales de la salud, quienes pueden aplicarlo rápidamente en la evaluación de riesgos sin necesidad de herramientas complejas o costosas (13, 14).

2.1.5. Impacto Clínico y Futuro de la Inteligencia Artificial en la Medicina Cardiovascular

Implicaciones Clínicas

La integración de RNA y otros métodos de I.A en la práctica clínica tiene el potencial de transformar la medicina cardiovascular. Al proporcionar predicciones más precisas y personalizadas, estas tecnologías pueden ayudar a identificar a los pacientes en mayor riesgo con mayor antelación, permitiendo intervenciones más tempranas y efectivas. Además, las RNA pueden ayudar a personalizar el tratamiento basado en el perfil de riesgo individual de cada paciente, optimizando así los resultados clínicos .

Sin embargo, la adopción generalizada de estas tecnologías también plantea desafíos, como la necesidad de formación especializada para los profesionales de la salud y la integración de estos sistemas en los flujos de trabajo clínicos existentes. También es crucial abordar las preocupaciones éticas relacionadas con la privacidad de los datos y la transparencia en la toma de decisiones (18).

Futuro de la Investigación en RNA y Cardiología

El futuro de la investigación en RNA y su aplicación en la cardiología es prometedor. A medida que se disponga de más datos y se desarrollen algoritmos más avanzados, es probable que las RNA se vuelvan aún más precisas y eficientes. El desarrollo de RNA explicables, que ofrecen mayor transparencia en su funcionamiento, también es un área de investigación activa, y podría facilitar la aceptación de estas tecnologías

en la práctica clínica (16).

Además, la integración de datos genéticos y de biomarcadores en las RNA podría abrir nuevas vías para la medicina de precisión, permitiendo tratamientos altamente personalizados basados en el perfil genético y biológico único de cada paciente. Esta personalización podría revolucionar la prevención y el tratamiento de las enfermedades cardiovasculares, reduciendo la morbilidad y la mortalidad asociadas a estas enfermedades (18).

2.2. *Random Forest* en la Predicción de Enfermedades Cardiovasculares

2.2.1 Introducción a *Random Forest*

El algoritmo ***Random Forest*** está compuesto por múltiples árboles de decisión que forman un conjunto o "bosque". Estos árboles se generan introduciendo cierto grado de aleatoriedad, lo que ayuda a reducir la correlación entre ellos y mejora la precisión del modelo. Esta técnica permite obtener predicciones más robustas y reduce el riesgo de sobreajuste comparado con un único árbol de decisión (19). Este es un potente modelo de aprendizaje automático que combina múltiples árboles de decisión para mejorar la precisión y robustez de las predicciones. Cada árbol en el bosque es entrenado con un subconjunto aleatorio de los datos y un subconjunto de características, lo que introduce diversidad en las predicciones y reduce el riesgo de sobreajuste *(overfitting)*. Al promediar los resultados de todos los árboles, *Random Forest* genera predicciones más estables y precisas en comparación con los árboles de decisión individuales. Este enfoque es especialmente valioso en la predicción de enfermedades cardiovasculares, donde las interacciones entre múltiples factores de riesgo pueden ser complejas y no lineales. Este modelo maneja de manera efectiva estas complejidades, lo que lo convierte en una herramienta poderosa para el análisis de grandes conjuntos de datos en medicina (19,20).

2.2.2 Estructura y Funcionamiento de *Random Forest*

Random Forest consta de varios árboles de decisión, cada uno entrenado con un

conjunto aleatorio de datos y características. Este enfoque de *"bagging" (bootstrap aggregating)* permite que el modelo capture una variedad de patrones en los datos, mejorando su capacidad para generalizar a nuevos casos. Durante el entrenamiento, cada árbol apoya una predicción, y la decisión final se toma por mayoría o promediando las predicciones de todos los árboles. La aleatoriedad en la selección de características y datos en cada árbol ayuda a mitigar el riesgo de que el modelo se ajuste demasiado a los datos de entrenamiento, lo que es un problema común en modelos más simples. Este proceso resulta en un modelo robusto y versátil, capaz de manejar grandes cantidades de datos con múltiples variables, siendo particularmente efectivo en la clasificación y predicción del riesgo de enfermedades cardiovasculares (20).

2.2.3. Aplicación de *Random Forest* en la Predicción de Enfermedades Cardiovasculares

Random Forest ha demostrado ser altamente eficaz en la predicción de enfermedades cardiovasculares. En estudios recientes, se ha utilizado para mejorar los modelos tradicionales al manejar datos heterogéneos y correlacionados, como los historiales clínicos, datos genómicos y hábitos de vida. Este modelo es capaz de identificar individuos en riesgo con una precisión elevada, lo que permite una intervención temprana y personalizada. Su capacidad para manejar interacciones no lineales y para trabajar con grandes volúmenes de datos lo convierte en una herramienta poderosa para el análisis en medicina. Al promediar las predicciones de múltiples árboles, *Random Forest* reduce el impacto de anomalías en los datos, permitiendo predicciones más precisas en comparación con los modelos basados en un único árbol de decisión (21).

Además, *Random Forest* es especialmente útil en la identificación de las variables más importantes en la predicción del riesgo cardiovascular, proporcionando a los médicos información valiosa para la toma de decisiones. Su capacidad para integrar datos de diversas fuentes y adaptarse a diferentes poblaciones supera las limitaciones de los modelos tradicionales, mejorando la generalización y la precisión

predictiva en diversas cohortes (22).

2.2.4 Comparación entre el *Modelo Framingham y Random Forest*

Precisión y Eficacia Predictiva

El modelo de *Framingham* ha sido una herramienta clave para estimar el riesgo cardiovascular, pero su enfoque lineal y limitado número de variables no siempre captura la complejidad de las interacciones entre los factores de riesgo (12). En contraste, *Random Forest*, con su capacidad para manejar múltiples variables y datos no lineales, ha demostrado una mayor precisión en la predicción de eventos cardiovasculares. Los estudios han mostrado que *Random Forest* puede mejorar la clasificación de riesgo, reduciendo la tasa de falsos positivos y negativos en comparación con el modelo de *Framingham* (22). Esto es crítico en la práctica clínica, donde una evaluación de riesgo precisa es esencial para garantizar la intervención en pacientes de alto riesgo, así como evitar el tratamiento innecesario en aquellos con bajo riesgo. *Random Forest*, por lo tanto, ofrece una mejora significativa en la precisión predictiva, lo que se traduce en mejores resultados clínicos (23).

Flexibilidad y Adaptabilidad

Random Forest también supera al modelo de *Framingham* en términos de flexibilidad. Puede ser entrenado y actualizado con nuevos datos de forma continua, lo que permite que el modelo se adapte a los cambios en los factores de riesgo y en las características de la población (19, 20). Esta adaptabilidad es crucial en un entorno clínico dinámico, donde los avances en la medicina y la variación en las poblaciones requieren modelos que puedan evolucionar con el tiempo. En contraste, el modelo de *Framingham*, basado en datos históricos, puede no reflejar con precisión los riesgos actuales, lo que limita su aplicabilidad (12, 24).

Complejidad y Desafíos en la Implementación

A pesar de sus ventajas, *Random Forest* presenta desafíos en su implementación. Entre ellos se encuentra la necesidad de recursos computacionales significativos para entrenar y mantener el modelo, especialmente cuando se trabaja con grandes conjuntos de datos. Aunque *Random Forest* es más interpretable que las redes neuronales, la complejidad de su estructura, con múltiples árboles interactuando entre sí, puede hacer que la interpretación de los resultados sea complicada. Esta complejidad podría dificultar la explicación clara de las decisiones a los pacientes y médicos (20).

En contraste, el modelo de *Framingham*, aunque menos preciso, es más sencillo y fácil de implementar en entornos clínicos. Su simplicidad lo hace más accesible para los profesionales de la salud que no disponen de herramientas tecnológicas avanzadas, permitiendo una integración más directa y comprensible en la práctica diaria (12, 24).

3. Estado del Arte en Modelos Predictivos de Enfermedades Cardiovasculares

3.1 Introducción

Las ECV siguen siendo la principal causa de muerte a nivel mundial, representando aproximadamente el 31% de todas las muertes, según datos de la Organización Mundial de la Salud (OMS). Esto equivale a 17,9 millones de muertes al año. Estas enfermedades no solo afectan gravemente a la salud individual, sino que también representan una carga significativa para los sistemas de salud debido a los elevados costos de tratamiento y a la pérdida de productividad.

La importancia de los factores de riesgo tradicionales como la hipertensión, el colesterol elevado, la diabetes, el tabaquismo y la obesidad ha llevado al desarrollo de modelos predictivos que permiten evaluar el riesgo cardiovascular de manera más precisa. En los últimos años, el uso de técnicas avanzadas de ***ML*** ha permitido mejorar considerablemente la predicción del riesgo, incorporando no solo variables tradicionales, sino también datos emergentes como los **marcadores genéticos** y la información recopilada por **dispositivos de monitoreo portátiles**. Esta evolución ha sido posible gracias a la disponibilidad de grandes volúmenes de datos y al avance de las técnicas de análisis, que permiten hacer evaluaciones más personalizadas.

3.2 Modelos Predictivos Clásicos y Técnicas de *Machine Learning*

3.2.1 Modelos Tradicionales

Uno de los primeros y más conocidos modelos predictivos es el *Framingham Risk Score (FRS)*, desarrollado a partir del estudio de *Framingham* iniciado en 1948. Este modelo ha sido ampliamente utilizado para calcular el riesgo cardiovascular a 10 años, considerando factores como la edad, el sexo, la presión arterial, el colesterol, el tabaquismo y la diabetes. Sin embargo, su capacidad de generalización a otras poblaciones con diferentes características étnicas o socioeconómicas es limitada.

A lo largo de los años, se han desarrollado otros modelos más avanzados como **QRISK** y **SCORE**, que incorporan una gama más amplia de variables, incluyendo el índice de masa corporal (IMC), antecedentes familiares y factores étnicos. Estos modelos han sido validados en poblaciones más diversas, mejorando la precisión de las predicciones. Sin embargo, siguen presentando limitaciones al no integrar factores emergentes como los biomarcadores o datos de dispositivos portátiles.

3.2.2 Técnicas de *Machine Learning* en la Predicción de Enfermedades Cardiovasculares

El uso de técnicas de **ML** ha transformado la capacidad de los modelos predictivos para gestionar grandes volúmenes de datos y ofrecer predicciones más precisas. A continuación, se presentan algunas de las técnicas más relevantes aplicadas en la predicción de enfermedades cardiovasculares:

- **Regresión Logística**: Es una técnica clásica en la predicción de ECV y una de las más interpretables. Aunque su capacidad para modelar relaciones no lineales es limitada, sigue siendo una opción preferida en escenarios donde la interpretabilidad es fundamental.
- **Árboles de decisión y *Random Forest***: Estas técnicas permiten manejar grandes cantidades de datos y modelar relaciones no lineales de manera eficiente. Los ***Random Forest*** son una mejora de los árboles de decisión, ya que utilizan múltiples árboles para mejorar la precisión y reducir el sobreajuste. Esta técnica ha demostrado ser eficaz en la predicción de enfermedades cardiovasculares, al manejar tanto variables continuas como categóricas, y al permitir una interpretación más sencilla de los factores que más contribuyen al riesgo.
- **Máquinas de soporte vectorial (SVM)**: Esta técnica es particularmente efectiva para clasificar datos complejos, aunque su aplicación en la práctica clínica puede estar limitada por los altos recursos computacionales que requiere. Sin embargo, en estudios recientes, SVM ha demostrado un buen rendimiento en la clasificación de pacientes con alto riesgo de ECV.
- **Redes neuronales y Deep Learning**: Con el aumento de la disponibilidad de

grandes volúmenes de datos médicos, las redes neuronales profundas han ganado popularidad en la predicción de ECV. Estas técnicas son capaces de identificar patrones complejos en los datos, mejorando la precisión de las predicciones. Sin embargo, un desafío importante es la interpretabilidad de estos modelos, que a menudo son considerados como "cajas negras" debido a su complejidad.

3.3 Factores de Riesgo en los Modelos Predictivos

3.3.1 Factores de Riesgo Tradicionales

Los factores de riesgo tradicionales han sido ampliamente estudiados y siguen siendo esenciales en los modelos predictivos de ECV. Algunos de los más importantes incluyen:

- **Hipertensión arterial**: Es uno de los factores de riesgo más importantes para el desarrollo de enfermedades cardiovasculares. Tener una presión arterial elevada de manera constante aumenta significativamente la probabilidad de sufrir eventos graves como infartos de miocardio y accidentes cerebrovasculares.
- **Colesterol elevado (LDL)**: Se ha demostrado que el colesterol LDL contribuye al desarrollo de placas en las arterias, lo que aumenta el riesgo de obstrucciones y eventos cardiovasculares.
- **Tabaquismo**: El consumo de tabaco daña las arterias y eleva el riesgo cardiovascular, siendo uno de los factores de riesgo más evitables.
- **Diabetes mellitus**: La diabetes tipo 2, en particular, está asociada con un aumento significativo en el riesgo de ECV debido a los daños que causa en los vasos sanguíneos.
- **Obesidad y sedentarismo**: Estos factores están relacionados con alteraciones metabólicas, lo que aumenta el riesgo de desarrollar enfermedades cardiovasculares.

3.3.2 Factores de Riesgo Emergentes

Con el avance en las investigaciones, han surgido nuevos factores de riesgo que están siendo considerados en los modelos predictivos de ECV:

- **Marcadores genéticos**: La identificación de variantes genéticas que aumentan el riesgo de ECV ha permitido el desarrollo de **puntuaciones de riesgo poligénico**, que añaden una capa de precisión a la predicción, especialmente en individuos jóvenes o sin factores de riesgo tradicionales.
- **Biomarcadores**: Algunos biomarcadores, como la **proteína C reactiva** y la **troponina**, se están integrando en los modelos predictivos, ya que indican inflamación o daño en el corazón antes de que los síntomas clínicos se manifiesten.
- **Datos de dispositivos portátiles**: La incorporación de datos de relojes inteligentes y otros dispositivos portátiles permite monitorizar el ritmo cardíaco, la actividad física y otros parámetros en tiempo real. Esto puede ayudar a detectar patrones de riesgo antes de que se conviertan en eventos graves.

3.4 Implicaciones Clínicas de las Técnicas de Machine Learning

El uso de **ML** ha permitido avanzar en la personalización de las predicciones, lo que facilita una mejor toma de decisiones en entornos clínicos. A medida que los modelos se vuelven más precisos y personalizados, los médicos pueden identificar con mayor exactitud qué pacientes están en mayor riesgo y diseñar intervenciones preventivas más efectivas.

Sin embargo, persisten algunos desafíos. A pesar de los avances en la precisión, muchos de los modelos basados en **Deep Learning** y técnicas avanzadas de ML presentan problemas de **interpretabilidad**, lo que limita su aplicación clínica. En este contexto, cobra relevancia la **Inteligencia Artificial Explicable (XAI)**, cuyo objetivo es hacer que los modelos sean más transparentes y comprensibles para los profesionales de la salud. Esto permite que los clínicos no solo confíen en las predicciones del modelo, sino que también comprendan cómo y por qué se han

tomado ciertas decisiones, lo que facilita una mayor confianza en la implementación de estas tecnologías en la práctica médica.

4. DESARROLLO.

4.1. Planteamiento del desarrollo

Divide y vencerás

La organización y la división de tareas han sido fundamentales para la ejecución exitosa de este proyecto colaborativo. Desde el inicio, se estableció una carpeta compartida en *Google Drive*, estructurada en secciones específicas para almacenar y compartir artículos relevantes, notas, *scripts*, archivos de datos y otros recursos necesarios. Esta estrategia se implementó con el objetivo de garantizar que todos los integrantes del proyecto tuvieran acceso a los avances y materiales en cualquier momento, facilitando la revisión y contribución continua de todos los miembros del equipo.

Tanto el código como la presente memoria se han desarrollado y revisado de manera conjunta a lo largo del proyecto, permitiendo una colaboración eficiente y efectiva entre los participantes.

4.2. Librerías y elección de los datos a tratar

Librerías utilizadas

Las herramientas seleccionadas no solo facilitaron la implementación de los modelos predictivos, sino que también contribuyeron a la correcta manipulación de los datos y a la presentación clara de los resultados (30).

- **Pandas**: Esta librería fue fundamental para la manipulación y análisis de los datos. Gracias a su estructura de *DataFrames*, Pandas permitió organizar, filtrar y transformar grandes conjuntos de datos de manera eficiente. Su capacidad para manejar datos faltantes y realizar operaciones complejas fue clave en la fase de preprocesamiento, lo que se aplicó en todo el proceso,

incluidas las nuevas pruebas con modelos adicionales.

- ***Scikit-learn:*** Utilizada ampliamente para la creación y evaluación de modelos de *machine learning*, *Scikit-learn* permitió implementar algoritmos como ***Random Forest***, **Regresión Logística**, y otros modelos probados en la extensión del trabajo, como ***XGBoost*** *y* ***LightGBM***. Además, facilitó tareas como la normalización, codificación de variables categóricas, validación cruzada y evaluación mediante métricas como la curva ROC-AUC y la matriz de confusión.
- **Imbalanced-learn (SMOTE)**: Enfrentamos un desbalance significativo entre clases en los datos, por lo que utilizamos **SMOTE** para generar instancias sintéticas de la clase minoritaria. Esta técnica fue crucial tanto en el desarrollo inicial como en la extensión, mejorando la capacidad de los modelos para predecir correctamente las clases menos representadas.
- ***Matplotlib y Seaborn***: Estas dos librerías fueron fundamentales para la visualización de datos y resultados a lo largo de todo el proyecto. Tanto en el análisis inicial como en la extensión, *con* ***XGBoost*** *y* ***LightGBM***, utilizamos **Matplotlib** y ***Seaborn*** para generar gráficos que ayudaron a identificar patrones y visualizar las comparaciones de los modelos. Esto permitió una interpretación clara de los resultados.
- ***XGBoost****:* En la extensión del notebook 1, utilizamos ***XGBoost****,* un algoritmo de *boosting* que es eficiente y robusto frente a datos ruidosos. La librería proporcionó herramientas para ajustar y entrenar el modelo.
- ***LightGBM***: Este algoritmo, también probado en la extensión, fue implementado para evaluar su rendimiento en comparación con ***Random Forest*** *y* ***XGBoost****.*
- ***Hyperopt***: Para la optimización de hiperparámetros en la extensión del proyecto, utilizamos ***Hyperopt***, una librería de optimización bayesiana que ayudó a explorar automáticamente el mejor conjunto de hiperparámetros para los modelos.
- ***TensorFlow y Keras****:* En el desarrollo inicial, estas librerías fueron utilizadas para implementar **Redes Neuronales***. **Keras**,* con su interfaz amigable, facilitó

la construcción y ajuste de redes neuronales profundas, aunque, en este caso, no lograron superar el desempeño de otros modelos como ***Random Forest***.

- ***SciPy***: Para el análisis estadístico, ***SciPy*** fue utilizada en varias fases del proyecto, incluyendo el cálculo de pruebas estadísticas como el **chi-cuadrado**. Esto ayudó a evaluar relaciones significativas entre variables categóricas y la presencia de enfermedades cardiovasculares.
- ***Joblib***: Utilizada para guardar los modelos entrenados, **Joblib** permitió la reutilización de estos modelos en pruebas posteriores sin tener que reentrenarlos, ahorrando tiempo en el proceso de experimentación tanto en el trabajo original como en la extensión.

Base de Datos *Framingham*

Para el desarrollo de este proyecto, fue esencial seleccionar una base de datos adecuada que cumpliera con los requerimientos específicos del estudio. Los criterios principales para la selección incluyeron la necesidad de contar con un conjunto de datos extenso, ya que un mayor volumen de registros permite un entrenamiento más preciso y eficiente del modelo predictivo. Además, se buscó que los datos abarcaran una diversidad de tipos, no limitándose sólo a información médica, con el fin de proporcionar un análisis integral de los factores que influyen en las enfermedades cardiovasculares. Igualmente, era crucial que los datos fueran confiables y coherentes, alineándose con los objetivos y el enfoque académico de la investigación.

Tras una cuidadosa revisión de diversas bases de datos, se eligió la base de datos *Framingham Heart Study*, disponible en la plataforma ***Kaggle*** (31), debido a varias razones. Primero, esta base de datos ha sido ampliamente utilizada en estudios académicos de gran renombre, lo que asegura su validez y calidad para el análisis de enfermedades cardiovasculares. Además, su disponibilidad en *Kaggle* no solo facilita el acceso a los datos, sino que también garantiza que los datos han sido curados y adaptados específicamente para su uso en proyectos de aprendizaje automático y modelos predictivos, brindando una plataforma robusta para experimentación y análisis.

La elección de esta base de datos responde directamente a los objetivos del proyecto, ya que permite aprovechar tanto la riqueza de los datos históricos como la capacidad de incluir nuevos factores en el análisis, asegurando así un enfoque completo y profundo en la predicción de riesgos cardiovasculares.

4.3. Descripción de los datos

Entender los datos

Antes de nada, es necesario entender los datos. La base de datos de *Framingham* contempla las siguientes variables:

1. Variables Demográficas y Socioeconómicas

- **Género** (nominal): Género del participante, registrado como masculino o femenino.
- **Edad** (continua): Edad del participante en años.
- **Grupos de Edad** (categórica): Categorización de la edad en grupos, como jóvenes, adultos y ancianos.
- **Educación** (nominal): Nivel educativo alcanzado por el participante, que puede influir en el acceso a información de salud y comportamientos preventivos.

2. Factores de Riesgo Cardiovascular

- **Índice de Masa Corporal (BMI)** (continua): Relación entre el peso y la altura del participante. Se utiliza para clasificar si el participante está en un rango saludable, sobrepeso u obesidad.
- **Fumador Activo (currentSmoker)** (nominal): Indicador binario (sí/no) que indica si el participante es fumador en la actualidad.
- **Cigarrillos por Día (cigsPerDay)** (continua): Número promedio de cigarrillos que el participante fuma diariamente.
- **Diabetes** (nominal): Indicador binario (sí/no) que indica si el participante tiene un diagnóstico de diabetes.

- **Hipertensión Previa (prevalentHyp)** (nominal): Indicador binario (sí/no) que señala si el participante tenía hipertensión antes de participar en el estudio.
- **Accidente Cerebrovascular Previo (prevalentStroke)** (nominal): Indicador binario (sí/no) que señala si el participante ha sufrido un accidente cerebrovascular anteriormente.
- **Medicación para Hipertensión (BPMeds)** (nominal): Indicador binario (sí/no) de si el participante estaba tomando medicamentos para controlar la presión arterial.

3. Variables Clínicas Actuales

- **Colesterol Total (totChol)** (continua): Nivel total de colesterol en sangre, un factor clave en el riesgo cardiovascular.
- **Presión Arterial Sistólica (sysBP)** (continua): Presión arterial máxima, cuando el corazón se contrae. Un valor elevado es un fuerte indicador de riesgo cardiovascular.
- **Presión Arterial Diastólica (diaBP)** (continua): Presión arterial mínima, cuando el corazón está en reposo entre latidos.
- **Glucosa** (continua): Nivel de glucosa en sangre, relacionado con el control del azúcar y con riesgo de diabetes.
- **Frecuencia Cardiaca (heartRate)** (continua): Número de latidos por minuto en reposo, un indicador de la salud del corazón.
- **Grupos de Frecuencia Cardiaca** (categórica): Categorización de la frecuencia cardiaca en grupos, como bradicardia (frecuencia baja) y taquicardia (frecuencia alta).

4. Variables de Comportamiento

- **Fumador Activo (currentSmoker)** (nominal): Indicador de si el participante es un fumador en la actualidad, un factor de riesgo clave para las enfermedades cardiovasculares.
- **Cigarrillos por Día (cigsPerDay)** (continua): Número promedio de cigarrillos que el participante fuma diariamente.

5. Variable de Interés (Variable de Predicción)

- **Riesgo de Enfermedad Cardiovascular en 10 años (TenYearCHD)** (binaria): Variable objetivo que indica si el participante desarrolló una enfermedad cardiovascular en los próximos 10 años. Esta es la variable que los modelos predictivos buscan anticipar.

4.4. Implementación del modelo de predicción

Prueba y error

Durante el desarrollo del proyecto, se generaron tres notebooks diferentes, cada uno de los cuales contenía modelos distintos o configuraciones variadas de los mismos. Este apartado de 'Desarrollo' se enfocará en el *notebook* que obtuvo los mejores resultados, mientras que los otros dos serán abordados posteriormente.

4.4.1. Preparación de los datos

Pre-procesado

Se verificó que no hubiera valores nulos ni duplicados en los datos. Luego, se realizaron algunas modificaciones necesarias para mejorar la calidad de la información. Por ejemplo, en la columna *"PBMeds"*, el valor *"Missing"* se reemplazó por un valor numérico para facilitar su uso en los modelos. Además, las columnas *"gender" y "currentSmoker*" se ajustaron temporalmente para facilitar la visualización y el análisis. Aunque se convirtieron a texto para este propósito, posteriormente fueron transformadas de nuevo a valores numéricos para ser utilizadas en los modelos predictivos.

4.4.2. Análisis de los datos

Análisis Exploratorio de los Datos (EDA)

Se llevó a cabo un análisis exploratorio de datos exhaustivo utilizando un *Jupyter Notebook en Google Colab*. Este análisis se dividió en dos partes principales:

Análisis Univariado

Se generaron gráficos para visualizar la distribución de las distintas variables, tanto categóricas como numéricas.

En este apartado se pueden apreciar algunos de ellos correspondientes a las variables numéricas estudiadas: *(Para obtener más detalles sobre los análisis y gráficos, puede consultar el notebook proporcionado con el trabajo*.)

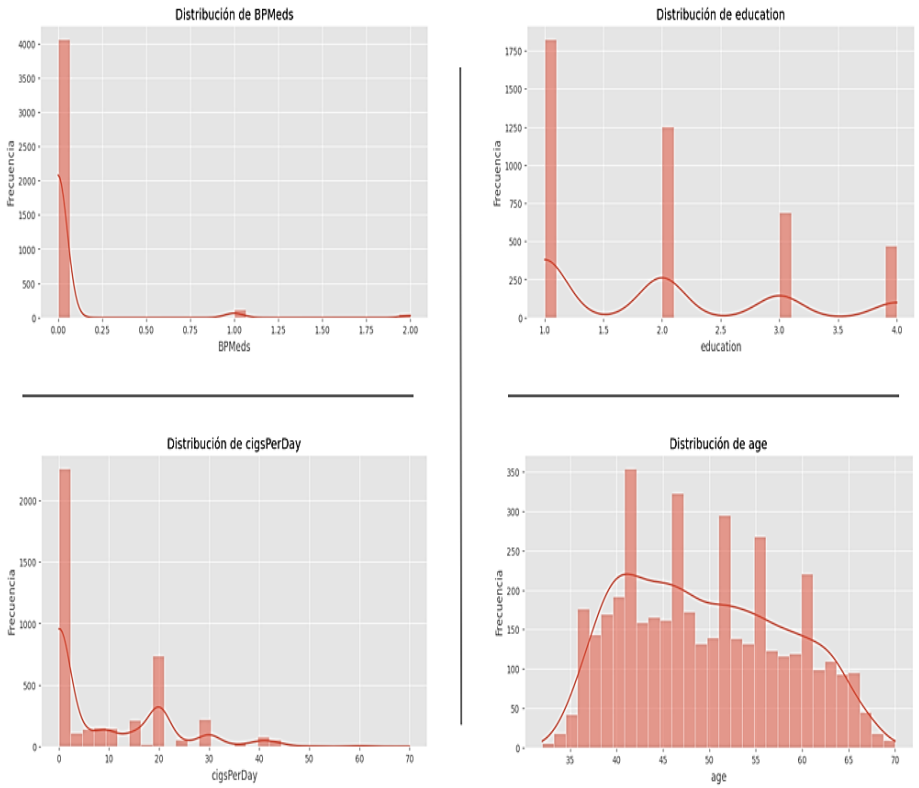


Imagen 1

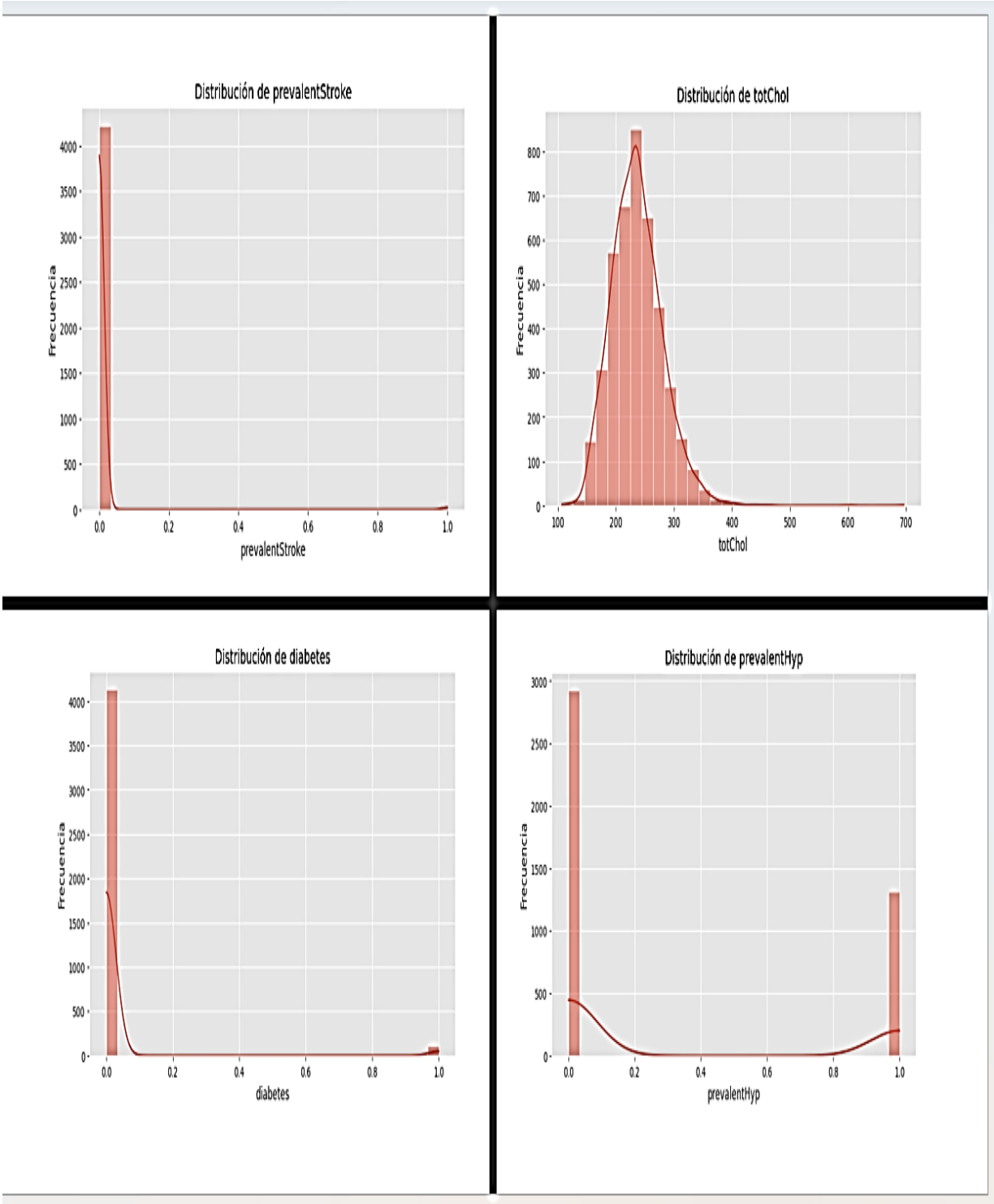

Distribución de prevalentStroke
Frecuencia
prevalentStroke
Distribución de totChol
Frecuencia
totChol
Distribución de diabetes
Frecuencia
diabetes
Distribución de prevalentHyp
Frecuencia
prevalentHyp

Imagen 2

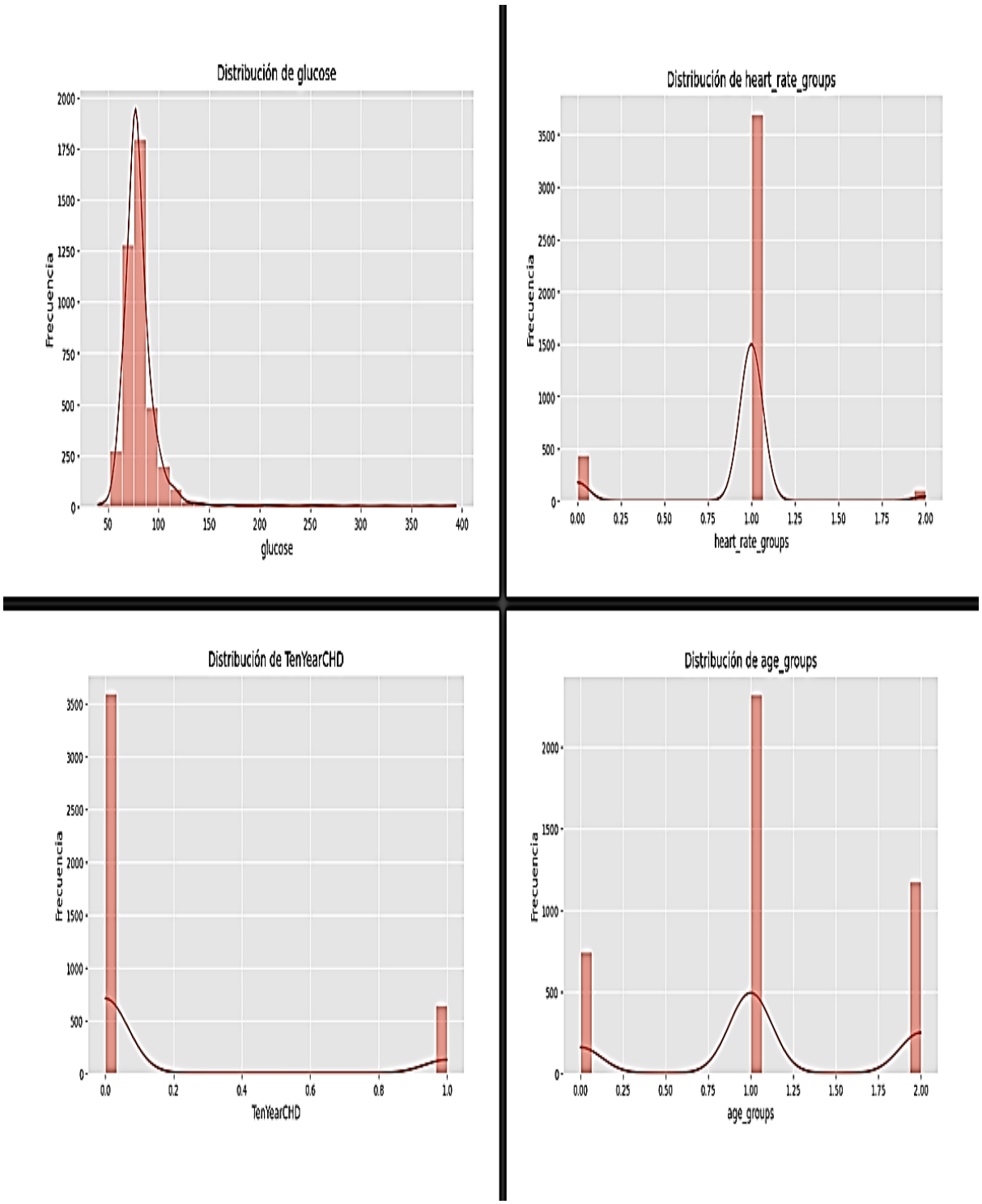

Distribución de glucose
Frecuencia
glucose
Distribución de heart_rate_groups
Frecuencia
heart_rate_groups
Distribución de TenYearCHD
Frecuencia
TenYearCHD
Distribución de age_groups
Frecuencia
age_groups

Imagen 3

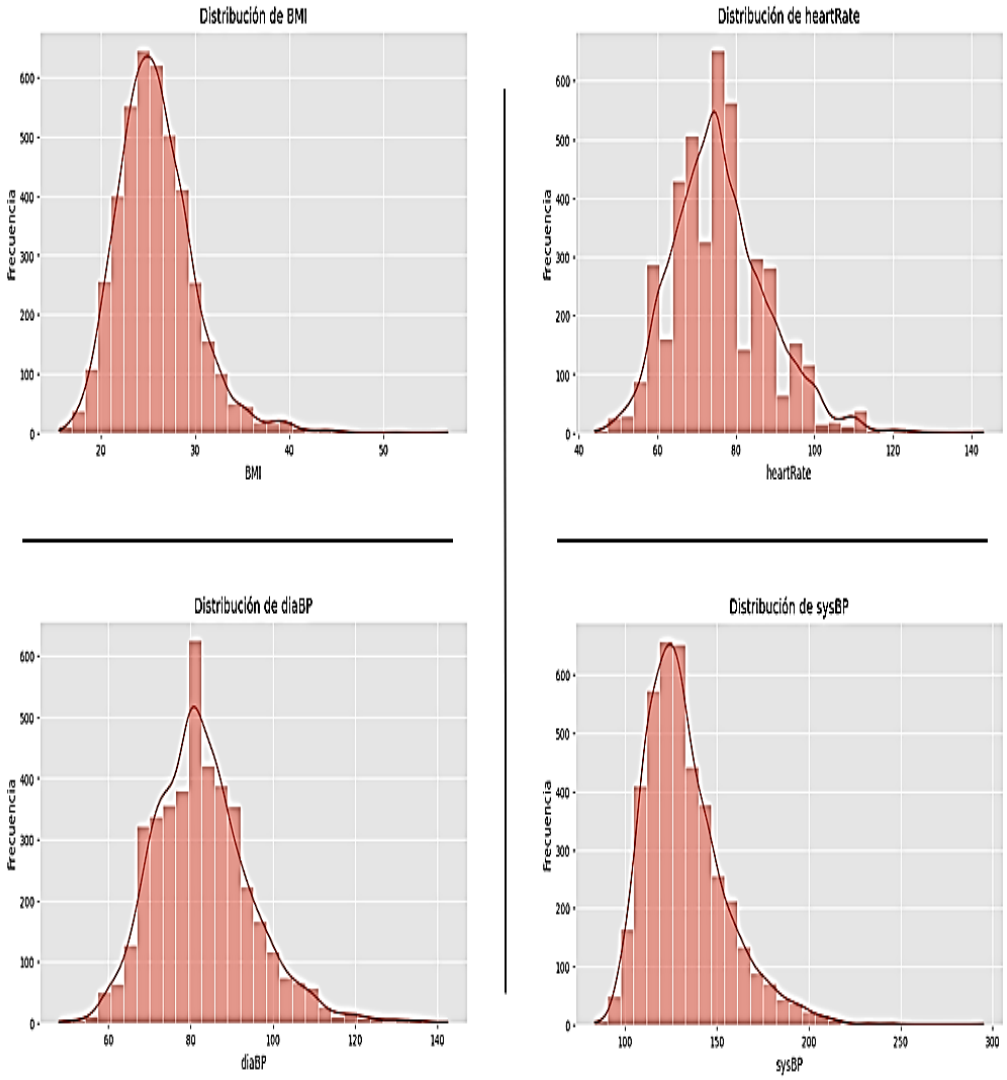


Imagen 4

Imágenes 1 a 4: *Serie de gráficos de distribución de variables clave*

Fuente de la visualización*: Gráficos generados por los autores mediante la librería Seaborn en Python utilizando datos del Framingham Heart Study.*

Primera Serie de Gráficos:

1. **Distribución de BPMeds (Uso de Medicación para la Presión Arterial):**
    - La mayoría de los pacientes no toma medicamentos para la presión arterial, aunque un pequeño grupo sí lo hace, lo que indica la necesidad de incluir este factor en el análisis del riesgo cardiovascular.
2. **Distribución de la Educación:**
    - Los niveles educativos de la muestra están divididos en varias categorías, lo que puede influir en otros factores de riesgo para las enfermedades cardiovasculares. Se observan grupos con distintos grados de educación, desde escolaridad básica hasta niveles más avanzados.
3. **Distribución de Cigarros por Día (cigsPerDay):**
    - La mayoría de los pacientes no son fumadores, pero hay varios subgrupos que fuman desde unos pocos hasta un número considerable de cigarrillos al día. Este factor es relevante al analizar los riesgos asociados con el tabaquismo.
4. **Distribución de la Edad (age):**
    - La distribución muestra una concentración en las edades medias (entre 40 y 60 años), lo que sugiere que el conjunto de datos está compuesto mayoritariamente por personas de mediana edad, un grupo crítico para el análisis del riesgo cardiovascular.

Segunda Serie de Gráficos:

1. **Distribución de PrevalentStroke (Accidente cerebrovascular previo):**

- La mayoría de los pacientes no ha sufrido un accidente cerebrovascular previo, pero hay un pequeño grupo que sí lo ha tenido, lo que subraya la importancia de este factor en la predicción del riesgo de futuros eventos cardiovasculares.

2. **Distribución de Colesterol Total (totChol):**

   - Los niveles de colesterol total en la muestra muestran una ligera variación, con la mayoría de los pacientes dentro de un rango saludable, pero algunos presentan niveles altos, lo que puede aumentar el riesgo de enfermedades cardiovasculares.

3. **Distribución de Diabetes:**

   - La mayoría de los pacientes no tiene diabetes, aunque un número significativo de ellos sí padece esta condición. Esto refuerza la relación conocida entre la diabetes y el riesgo de enfermedades cardíacas.

4. **Distribución de PrevalentHyp (Hipertensión previa):**

   - La mayoría de los pacientes no tiene antecedentes de hipertensión, pero una parte significativa de la muestra sí presenta esta condición, lo que refuerza su importancia como factor de riesgo en la predicción de enfermedades cardiovasculares.

Tercera Serie de Gráficos:

1. **Distribución de la Glucosa *(glucose):***

   - La mayoría de los pacientes tiene niveles de glucosa dentro de los rangos normales, pero un pequeño grupo presenta niveles elevados, lo que sugiere la presencia de pacientes con diabetes o prediabetes.

2. **Distribución de Grupos de Frecuencia Cardíaca *(heart rate groups):***

    - La mayoría de los pacientes se concentra en un grupo con frecuencia cardíaca normal, aunque hay algunos casos con frecuencias más altas o bajas, lo que puede estar relacionado con afecciones cardiovasculares.

3. **Distribución de TenYearCHD (Riesgo de enfermedad cardiovascular en 10 años):**

    - Esta variable muestra una distribución muy desequilibrada. La gran mayoría de las personas no desarrollan la enfermedad cardiovascular dentro de los 10 años, mientras que solo un pequeño porcentaje sí lo hace. Esto resalta la importancia de balancear los datos en los modelos de predicción.

4. **Distribución de Grupos de Edad (age groups):**

    - La mayoría de los individuos en la muestra se encuentran en el grupo intermedio de edad. Aunque también hay personas en los extremos de edad (adultos jóvenes y mayores), estos son menos representativos en comparación con el grupo central.

Cuarta Serie de Gráficos:

1. **Distribución del Índice de Masa Corporal (BMI):**

    - La mayoría de los pacientes tiene un índice de masa corporal que varía entre los rangos normales y de sobrepeso, aunque también hay casos de obesidad severa. Esto es importante para evaluar el riesgo de enfermedades relacionadas con la obesidad.

2. **Distribución de la Frecuencia Cardíaca (heartRate):**

    - La mayoría de los pacientes tiene una frecuencia cardíaca dentro de los rangos normales, aunque algunos presentan frecuencias más altas o bajas, lo que podría estar relacionado con problemas cardíacos o condiciones médicas específicas.

3. **Distribución de la Presión Arterial Diastólica (diaBP):**

    - La distribución muestra que la mayoría de los pacientes tiene la presión arterial diastólica en valores normales, aunque hay algunos casos con niveles elevados, lo que es un factor de riesgo cardiovascular.

4. **Distribución de la Presión Arterial Sistólica (sysBP):**

    - Similar a la diastólica, la mayoría de los pacientes tiene valores de presión arterial sistólica normales, pero un grupo significativo muestra valores altos, lo que indica hipertensión y mayor riesgo cardiovascular.

Frecuencia de Variables Categóricas

Los gráficos presentados a continuación ofrecen un análisis descriptivo sencillo sobre cómo se distribuyen algunas de las variables categóricas dentro del conjunto de datos. Es importante mencionar que, además de las variables mostradas, el conjunto de datos incluye otras variables categóricas que también fueron consideradas en el análisis. Estos gráficos ayudan a contextualizar la composición de la muestra de pacientes en términos de género y hábito de fumar.

- **Frecuencia de género (gender)**: El gráfico de barras muestra la distribución de hombres (M) y mujeres (F) en el conjunto de datos. Se observa una mayor representación de hombres en comparación con mujeres.
- **Frecuencia de fumadores actuales (currentSmoker)**: Este gráfico de barras indica la distribución entre personas que actualmente fuman (YES) y aquellas que no fuman (NO). Aunque esta distribución es relevante, otras

variables categóricas también influyen en el riesgo cardiovascular y forman parte del análisis global.

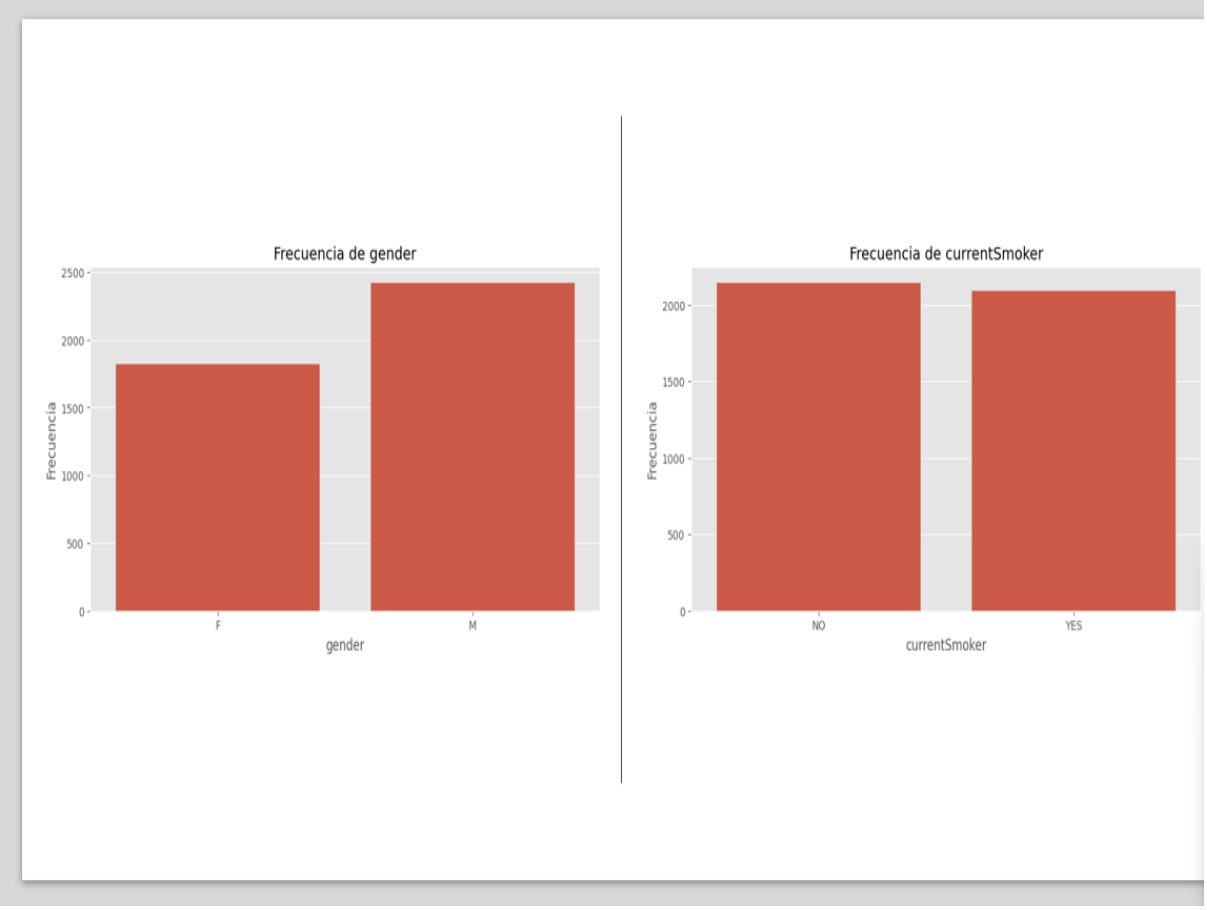


Imagen 5: Frecuencia de Variables Categóricas

***Fuente de la visualización**: Gráficos generados por los autores mediante la librería **Seaborn** en **Python** utilizando datos del Framingham Heart Study.*

Además, se evaluó la correlación entre estas variables para identificar posibles relaciones lineales.

Como se observa en el gráfico a continuación, (el resto podrán encontrarse en el anexo digital referente al *notebook* 1).

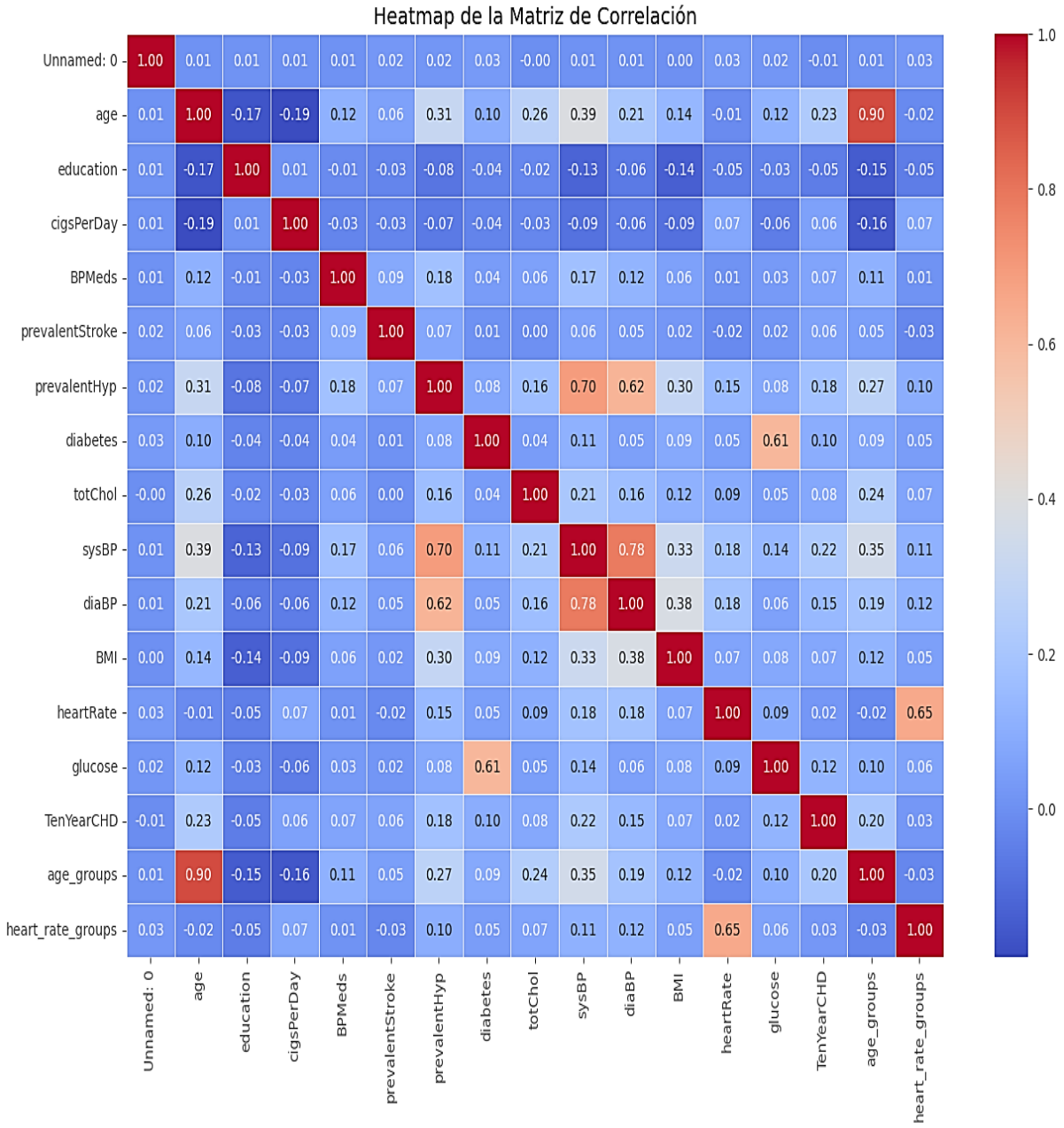


Imagen 6: Matriz de Confusión del Modelo Random Forest

Fuente de la visualización: *Gráfico generado por los autores mediante la librería Seaborn en Python utilizando datos del Framingham Heart Study.*

El *heatmap* o mapa de calor de la matriz de correlación muestra cómo se relacionan las diferentes variables numéricas del conjunto de datos entre sí. Las correlaciones se expresan como un valor entre -1 y 1:

● **Valores cercanos a 1**: Indican una correlación positiva fuerte, lo que significa que, a medida que una variable aumenta, la otra también tiende a aumentar de manera proporcional.

● **Valores cercanos a -1**: Señalan una correlación negativa fuerte, es decir, cuando una variable aumenta, la otra tiende a disminuir.

● **Valores cercanos a 0**: Muestran que hay poca o ninguna relación entre las dos variables, lo que implica que no siguen un patrón claro o predecible entre sí.

Este *heatmap* facilita la identificación de relaciones entre las variables del conjunto de datos, destacando cuáles están más fuertemente relacionadas. Esto es útil para entender qué factores podrían ser más relevantes al momento de predecir eventos cardiovasculares. Por ejemplo, la fuerte correlación entre la presión arterial y la hipertensión previa sugiere que estas variables deben ser consideradas de manera conjunta en los modelos predictivos.

Análisis Bivariado

Se comparó la variable objetivo con las variables numéricas mediante gráficos de cajas, lo que permitió observar la dispersión y la tendencia central de los datos. Asimismo, se realizó una comparación entre la variable objetivo y las variables categóricas utilizando tablas de contingencia y la prueba Chi-Cuadrado, con el fin de evaluar la asociación entre estas variables.

Chi-Cuadrado

La prueba de Chi-cuadrado es una herramienta muy útil para analizar y entender la relación entre dos variables categóricas.

La tabulación cruzada muestra las distribuciones conjuntas de dos variables categóricas, donde las combinaciones de sus categorías se representan en las celdas de la tabla.

El cálculo del valor estadístico de Chi-cuadrado, y su comparación con un valor crítico de la distribución de Chi-cuadrado, permite al investigador determinar si las frecuencias observadas en las celdas son significativamente diferentes de las frecuencias esperadas.

Es importante tener en cuenta que el valor de Chi-cuadrado es muy sensible al tamaño de la muestra: si la muestra es demasiado grande (alrededor de 500 o más), incluso pequeñas diferencias pueden parecer estadísticamente significativas. Además, la distribución de los datos en las celdas también puede influir en los resultados, por lo que es recomendable trabajar con variables categóricas que tengan un número limitado de categorías (32).

Imagen 7

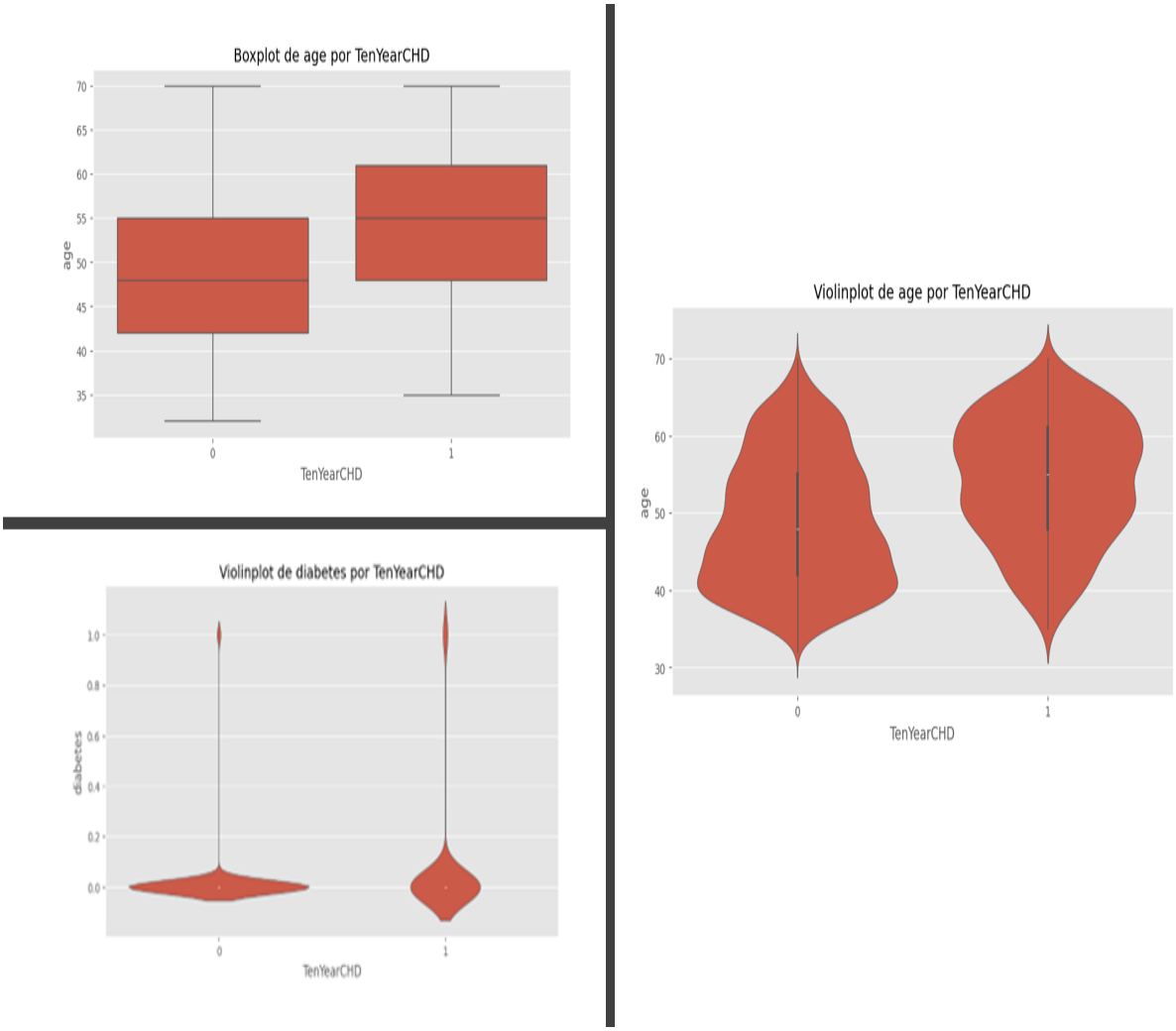


Imagen 8

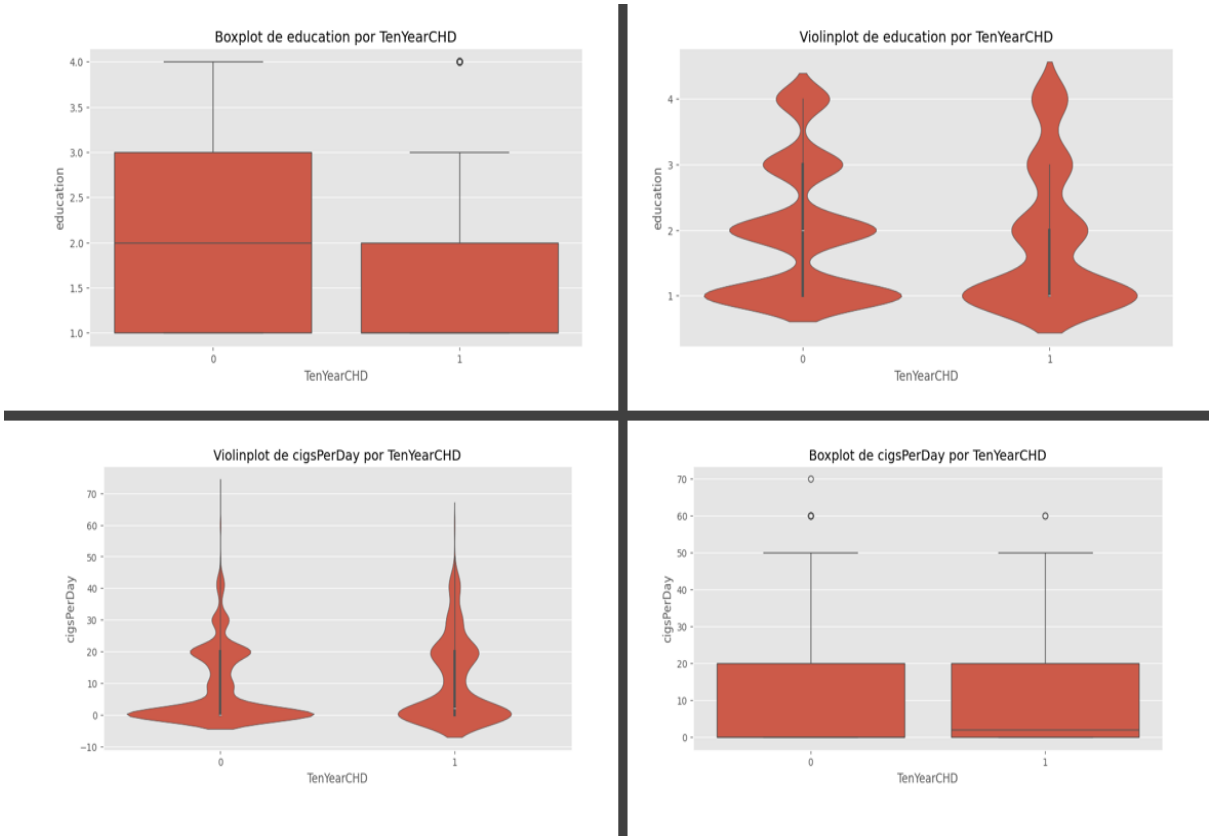


Imagen 9

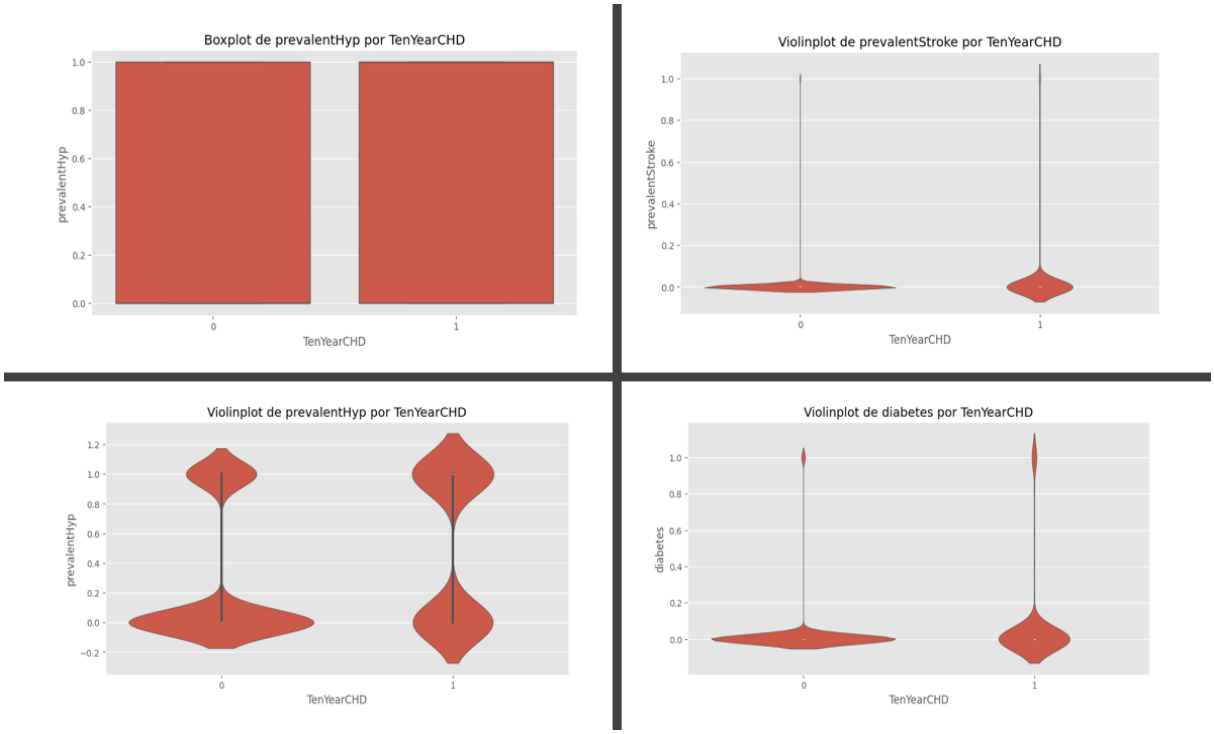
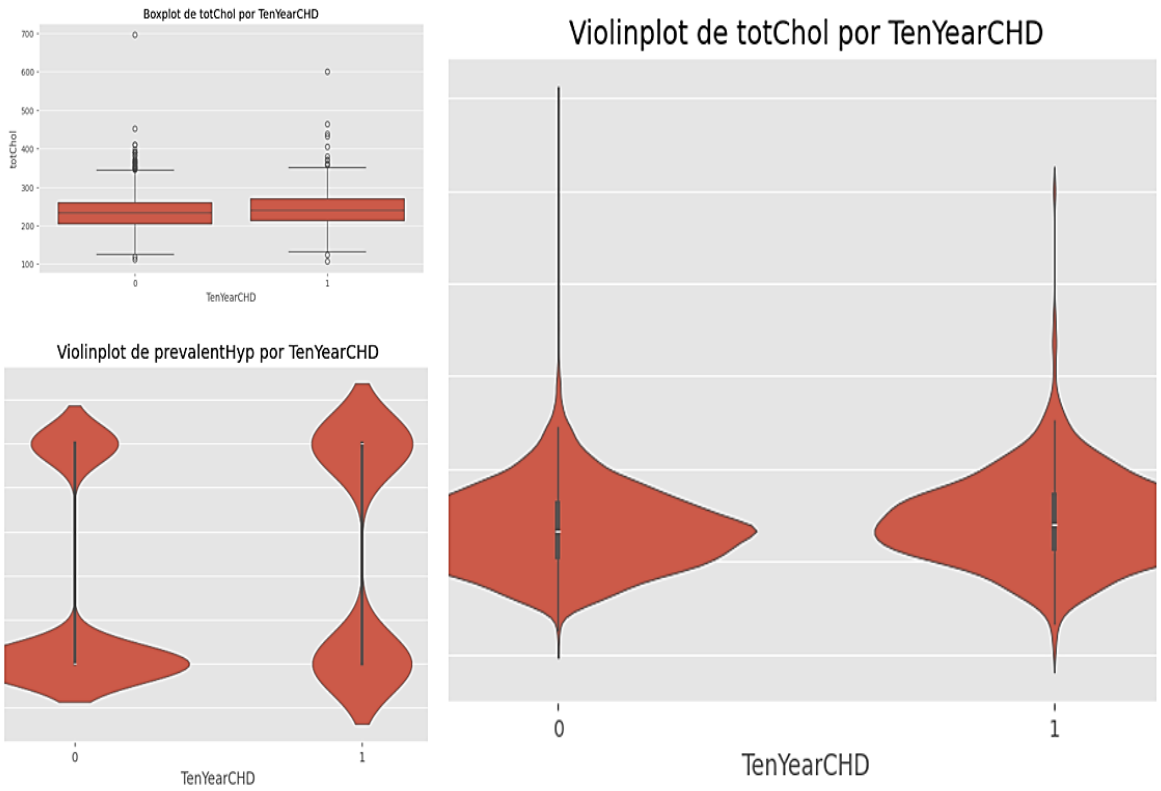


Imagen 10



Imágenes 7-10: *Serie de gráficos boxplot y violinplot que permiten visualizar la distribución de varias variables en relación con la variable de resultado TenYearCHD (la probabilidad de desarrollar una enfermedad cardiovascular en los próximos 10 años).*

Fuente de la visualización: *Gráficos generados por los autores mediante la librería Seaborn en Python utilizando datos del Framingham Heart Study.*

(Para obtener más detalles sobre los análisis y gráficos, puede consultar el notebook proporcionado con el trabajo.)

Descripción de los Gráficos:

Serie 1 (Imagen 7)

- ***Boxplot de Edad por TenYearCHD*:**
- Este gráfico de cajas muestra la distribución de la edad según si el participante desarrolló o no una enfermedad cardiovascular en los próximos 10 años. Se puede observar que las personas que desarrollaron la enfermedad

(TenYearCHD = 1) tienden a ser mayores que las que no la desarrollaron (TenYearCHD = 0).

- ***Violinplot de Edad por TenYearCHD**:*
- Aquí se combina un boxplot y una distribución de densidad. Muestra cómo se distribuyen las edades según si los participantes desarrollaron o no una enfermedad cardiovascular. Las personas con mayor edad parecen tener un riesgo mayor de desarrollar la enfermedad (el gráfico es más ancho en edades avanzadas para TenYearCHD = 1).
- ***Violinplot de Diabetes por TenYearCHD:***
- En este se muestra cómo se distribuyen los casos de diabetes según si los participantes desarrollaron o no una enfermedad cardiovascular en los próximos 10 años. Se observa que la mayoría de los participantes que no desarrollaron la enfermedad no tenían diabetes, mientras que hay una pequeña proporción de personas con diabetes en ambos grupos.

Serie 2 (Imagen 8)

- ***Boxplot de Educación por TenYearCHD:***
- Este gráfico muestra la distribución del nivel educativo entre las personas que desarrollaron enfermedades cardiovasculares (TenYearCHD = 1) y las que no lo hicieron (TenYearCHD = 0). Se puede observar que las personas sin enfermedades cardiovasculares tienden a tener niveles educativos más altos en comparación con aquellas que desarrollaron la enfermedad.
- ***Violinplot de Educación por TenYearCHD:***
- En este se combina un boxplot con la distribución de densidad de los niveles educativos. Muestra que las personas sin enfermedad cardiovascular tienen una distribución más amplia en los niveles educativos más altos, mientras que las personas que desarrollaron la enfermedad tienden a concentrarse en niveles educativos más bajos.
- ***Violinplot de Cigarros por Día (cigsPerDay) por TenYearCHD:***
- Aquí se muestra cómo varía el número de cigarrillos fumados por día entre quienes desarrollaron enfermedades cardiovasculares y quienes no. Aunque la mayoría de las personas fuman menos de 20 cigarrillos al día, se puede ver

una ligera concentración de fumadores en el grupo que desarrolló enfermedades.

- ***Boxplot de Cigarros por Día (cigsPerDay) por TenYearCHD:***
- El gráfico de cajas aquí muestra la dispersión de la cantidad de cigarrillos fumados por día entre los dos grupos (con y sin enfermedad cardiovascular). Los valores son bastante similares entre ambos grupos, con algunos valores más altos que aparecen como posibles valores atípicos.

Serie 3 (Imagen 9)

- ***Boxplot de PrevalentHyp (Hipertensión Previa) por TenYearCHD:***
- En este gráfico se presenta la relación entre la hipertensión previa y el desarrollo de enfermedades cardiovasculares a lo largo de 10 años. La gran mayoría de los participantes, tanto aquellos que desarrollaron la enfermedad como los que no, presentan hipertensión previa, lo que indica que es una condición muy prevalente en la muestra.
- ***Violinplot de PrevalentStroke (Accidente Cerebrovascular Previo) por TenYearCHD:***
- Aquí se visualiza la distribución de accidentes cerebrovasculares previos entre las personas que desarrollaron o no enfermedades cardiovasculares. Los datos sugieren que hay una mayor proporción de individuos con antecedentes de accidente cerebrovascular entre aquellos que desarrollaron la enfermedad.
- ***Violinplot de PrevalentHyp (Hipertensión Previa) por TenYearCHD:***
- En esta visualización se muestra cómo la hipertensión previa está presente tanto en quienes desarrollaron enfermedades cardiovasculares como en quienes no. Sin embargo, parece haber una ligera concentración de casos de hipertensión en el grupo que desarrolló la enfermedad.
- ***Violinplot de Diabetes por TenYearCHD:***
- En cuanto a la diabetes, el gráfico evidencia que existe una mayor incidencia de diabetes en el grupo que desarrolló enfermedades cardiovasculares, lo cual refuerza la idea de que la diabetes es un factor de riesgo significativo en el desarrollo de estas condiciones.

Serie 4 (Imagen 10)

- ***Boxplot de PrevalentHyp (Hipertensión Previa) por TenYearCHD:***
- En éste se presenta la relación entre la hipertensión previa y el desarrollo de enfermedades cardiovasculares a lo largo de 10 años. La gran mayoría de los participantes, tanto aquellos que desarrollaron la enfermedad como los que no, presentan hipertensión previa, lo que indica que es una condición muy prevalente en la muestra.
- ***Violinplot de PrevalentStroke (Accidente Cerebrovascular Previo) por TenYearCHD:***
- Aquí se visualiza la distribución de accidentes cerebrovasculares previos entre las personas que desarrollaron o no enfermedades cardiovasculares. Los datos sugieren que hay una mayor proporción de individuos con antecedentes de accidente cerebrovascular entre aquellos que desarrollaron la enfermedad.
- ***Violinplot de PrevalentHyp (Hipertensión Previa) por TenYearCHD:***
- En esta gráfica se muestra cómo la hipertensión previa está presente tanto en quienes desarrollaron enfermedades cardiovasculares como en quienes no. Sin embargo, parece haber una ligera concentración de casos de hipertensión en el grupo que desarrolló la enfermedad.
- ***Violinplot de Diabetes por TenYearCHD:***
- Respecto a la diabetes, el gráfico evidencia que existe una mayor incidencia de diabetes en el grupo que desarrolló enfermedades cardiovasculares, lo cual refuerza la idea de que la diabetes es un factor de riesgo significativo en el desarrollo de estas condiciones.

Otras visualizaciones

Paralelamente a este análisis se creó una visualización de algunos de los datos en *Power BI* con el mismo propósito de entender los datos y tratar de encontrar patrones en las relaciones entre estos.

4.4.3. Análisis Power BI

El objetivo de este análisis es demostrar las capacidades de PowerBI para explorar y visualizar grandes conjuntos de datos en el ámbito de la salud. Utilizando el ***Dataset de Framingham*** como caso de estudio, se llevó a cabo un análisis exploratorio para identificar los principales factores de riesgo asociados con las enfermedades cardiovasculares. Este análisis no sólo permitió seleccionar el *dataset* como óptimo para el estudio de eventos cardiovasculares, sino que también brindó una base sólida para la prevención y el tratamiento de estas enfermedades.

Limpieza de Datos

En términos de calidad de los datos, el conjunto se encontraba en un estado bastante limpio, con muy pocos valores faltantes en la columna **'*BPMeds*'** (1,25%). Para garantizar la completitud del análisis y evitar sesgos, los valores faltantes se imputaron utilizando la moda, es decir, reemplazando los datos faltantes por el valor más frecuente. Además, se realizaron ajustes en varias columnas para facilitar el análisis, transformando variables como el género, grupos de edad y diversos indicadores binarios (como fumar, diabetes e hipertensión) para permitir un análisis más eficiente.

Column1	gender	age	education	currentSmoker	cigsPerDay	BPMeds	prevalentStroke	prevalentHyp	diabetes	totChol	sysBP	diaBP	BMI	heartRate
0	Femenino	39	4	False	0	0	0	False	False	195	106	70	2697	80
1	Masculino	46	2	False	0	0	0	False	False	250	121	81	2873	95
2	Femenino	48	1	True	20	0	0	False	False	245	1275	80	2534	75
3	Masculino	61	3	True	30	0	0	True	False	225	150	95	2858	65
4	Masculino	46	3	True	23	0	0	False	False	285	130	84	231	85
5	Masculino	43	2	False	0	0	0	True	False	228	180	110	303	77
6	Masculino	63	1	False	0	0	0	False	False	205	138	71	3311	60
7	Masculino	45	2	True	20	0	0	False	False	313	100	71	2168	79
8	Femenino	52	1	False	0	0	0	True	False	260	1415	89	2636	76
9	Femenino	43	1	True	30	0	0	True	False	225	162	107	2361	93

Imagen 11: *Proceso de limpieza de datos en PBI*

Fuente de la visualización: *Gráfico generado por los autores mediante Power BI utilizando datos del Framingham Heart Study*

Metodología y Resultados

El análisis exploratorio se realizó utilizando **Power BI**, aprovechando su interfaz

intuitiva y su capacidad para crear visualizaciones interactivas. Esto permitió una exploración profunda del *dataset*, facilitando la comunicación de los resultados de manera efectiva a públicos no especializados.

Durante el análisis se utilizaron las siguientes técnicas visuales:

- **Histogramas** para evaluar la distribución de las variables numéricas.
- **Diagramas de caja (boxplots)** para comparar grupos y observar la variación de los datos.
- **Gráficos de dispersión** para explorar las relaciones entre variables, como la asociación entre la edad, el Índice de Masa Corporal (IMC) y la presión arterial sistólica.

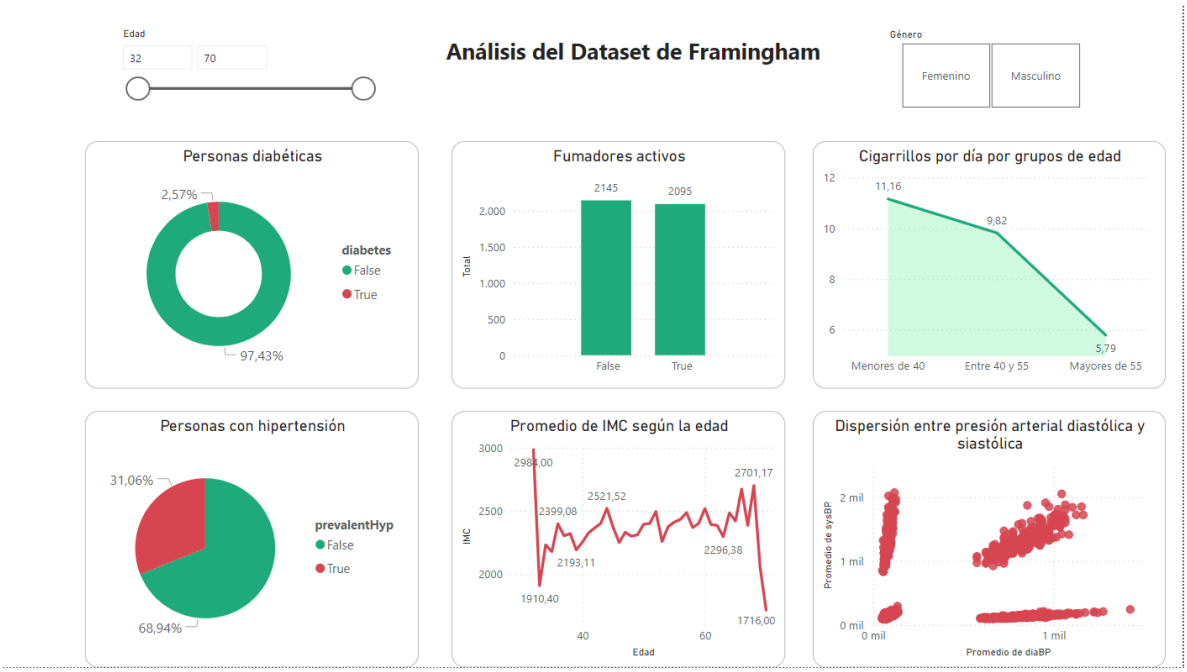


Imagen 12: *Proceso de análisis de los datos en PBI*

Fuente de la visualización: *Gráfico generado por los autores mediante Power BI utilizando datos del Framingham Heart Study*

Los resultados preliminares indicaron lo siguiente:

- Existe una asociación positiva entre la **edad**, el **índice de masa corporal** y la **presión arterial sistólica**. Es decir, a medida que aumentan estas variables,

también lo hace el riesgo de desarrollar enfermedades cardiovasculares.

- Se observó una mayor prevalencia de enfermedades cardiovasculares en **hombres fumadores**, lo que refuerza la importancia del tabaquismo como factor de riesgo clave.

Implicaciones del Análisis

Este análisis nos permitió abordar preguntas cruciales para la investigación en enfermedades cardiovasculares, tales como:

1. ¿Cuáles son los factores de riesgo más significativos para el desarrollo de enfermedades cardiovasculares?
2. ¿Qué combinación de factores permite predecir con mayor precisión la probabilidad de eventos cardiovasculares tanto a corto como a largo plazo?
3. ¿Es posible identificar subgrupos con mayor riesgo para priorizar intervenciones preventivas?

Power BI no solo permitió obtener respuestas a estas preguntas, sino que también ayudó a optimizar la asignación de recursos y a mejorar la toma de decisiones en el ámbito de la prevención y el tratamiento de enfermedades cardiovasculares.

4.4.4. Modelo de predicción

Más prueba y error

A partir de este punto, se procedió a probar distintos algoritmos con el objetivo de obtener el mejor rendimiento en términos de precisión para nuestro modelo. Como se mencionó anteriormente, este apartado se enfocará en el *notebook* en el que se desarrolló el modelo que logró los mejores resultados, mientras que los demás se abordarán más adelante.

En este *notebook* se probaron dos modelos: Regresión Logística y *Random Forest* con balanceo de clases. El modelo que mostró el rendimiento más favorable fue

Random Forest, tras aplicar un balance de clases previo utilizando la técnica *SMOTE.*

Regresión Logística

La regresión logística es un modelo estadístico que se utiliza para analizar la relación entre varias variables categóricas (Xi) y una variable dependiente también categórica (Y). Este modelo pertenece a la familia de modelos lineales generalizados y emplea la función logística para establecer la relación entre las variables.

A través de la regresión logística, es posible predecir la probabilidad de que ocurra un evento (asignado como valor 1) o de que no ocurra (valor 0). Para ello, el modelo ajusta los coeficientes de regresión. El resultado obtenido siempre se sitúa entre 0 y 1, mediante la aplicación de una función sigmoide a la salida. Si la probabilidad predicha supera un determinado umbral, es más probable que el evento ocurra, mientras que, si está por debajo, es menos probable que suceda. (33)

Implementación del modelo de Regresión Logística y *Random Forest*

Primero, se dividió el conjunto de datos en un conjunto de entrenamiento y un conjunto de prueba utilizando la función *train_test_split.* El 80% de los datos se destinó al entrenamiento del modelo, mientras que el 20% restante se reservó para evaluar su rendimiento. La división se realizó de forma aleatoria, fijando un valor de *random_state* para garantizar la reproducibilidad de los resultados.

A continuación, se construyó un modelo utilizando un *pipeline* (un proceso automatizado que conecta múltiples etapas de procesamiento de datos de manera secuencial). Este *pipeline* integró dos fases principales: un preprocesador y el clasificador de regresión logística. El preprocesador se encargó de preparar los datos, realizando tareas como la estandarización de variables numéricas y la codificación de variables categóricas. Posteriormente, el clasificador aplicó la regresión logística, configurada con un máximo de 1000 iteraciones para garantizar que el algoritmo alcanzara la convergencia. El modelo fue entrenado utilizando el conjunto de datos de entrenamiento *(X_train, y_train)* mediante el método *fit* (un procedimiento que

ajusta el modelo a los datos proporcionados, permitiendo que el algoritmo aprenda a partir de ellos para hacer predicciones futuras)

Posteriormente, se realizaron predicciones sobre el conjunto de prueba *(X_test).* Las predicciones categóricas se almacenaron en *y_pred,* mientras que las probabilidades asociadas a cada clase se guardaron en *y_pred_proba*.

Finalmente, se evaluó el rendimiento del modelo utilizando tres métricas clave. La primera fue el **informe de clasificación** *(classification_report*), que proporciona información detallada sobre tres aspectos: la **precisión** (qué tan correcto es el modelo), la **sensibilidad** *o* ***recall*** (qué tan bien detecta el modelo los casos verdaderamente positivos) y la **puntuación F1** (una combinación de precisión y sensibilidad, que ayuda a evaluar el equilibrio del modelo).

La segunda métrica fue la **matriz de confusión** *(confusion_matrix)*, que muestra cómo se comporta el modelo en términos de predicciones correctas e incorrectas, clasificando los resultados en cuatro categorías: **verdaderos positivos** (casos correctamente identificados como positivos), **verdaderos negativos** (casos correctamente identificados como negativos), **falsos positivos** (casos incorrectamente identificados como positivos) y **falsos negativos** (casos incorrectamente identificados como negativos).

La tercera métrica fue el **ROC-AUC** *(roc_auc_score)*, que mide la capacidad del modelo para distinguir entre las diferentes clases, es decir, qué tan bien el modelo puede diferenciar entre casos positivos y negativos.

Los resultados obtenidos a partir de estas métricas se imprimieron para analizar el desempeño del modelo y entender mejor su eficacia.

Random Forest

Random forest es un algoritmo de aprendizaje automático supervisado que se usa para solucionar problemas de clasificación y regresión. Construye árboles de decisión

a partir de diferentes muestras y toma su voto mayoritario para decidir la clasificación y el promedio en caso de regresión.

Una de las características más importantes del algoritmo de bosque aleatorio es que puede manejar un conjunto de datos que contenga variables continuas, como en el caso de la regresión, y variables categóricas, como en el caso de la clasificación. Por eso ofrece mejores resultados para problemas de clasificación. (35)

SMOTE **y su impacto en el Modelo**

SMOTE (Synthetic Minority Over-sampling Technique) es una técnica de balanceo de clases utilizada para generar artificialmente nuevos elementos de la clase minoritaria en un conjunto de datos. (34) Esto se logra seleccionando aleatoriamente un elemento de la clase minoritaria y luego eligiendo un número de vecinos más cercanos. A partir de estos vecinos, se genera un nuevo elemento combinando las características de los vecinos de manera ponderada, introduciendo un factor aleatorio para asegurar que los elementos generados sean similares, pero no idénticos a los existentes. Existen diversas versiones del algoritmo *SMOTE* que varían en el número de vecinos seleccionados y en cómo se combinan para generar los nuevos elementos. Repetir este proceso permite generar suficientes instancias de la clase minoritaria hasta que el conjunto de datos esté equilibrado en cuanto a clases.

En este proyecto, después de separar la variable objetivo ***TenYearCHD*** (que indica si un paciente desarrollará una enfermedad cardiovascular en los próximos diez años), se realizaron varias técnicas de preprocesamiento para preparar los datos antes de usarlos en el modelo. Primero, las columnas con categorías como **género** y **si es fumador actual** (*gender y currentSmoke*r) se transformaron en números, ya que los algoritmos de aprendizaje automático necesitan trabajar con datos numéricos.

Luego, se utilizaron ***pipelines*** para estandarizar las variables numéricas y rellenar los valores faltantes. Para los números que faltaban, se utilizó la **mediana** (el valor central de un conjunto de datos), y para las variables categóricas se usó la **moda** (el valor

que más se repite). Además, se aplicó la técnica ***One-Hot Encoding***, que convierte las variables categóricas en columnas binarias. Por ejemplo, si tienes una columna de género con los valores "hombre" y "mujer*"*, ***One-Hot*** la convierte en dos columnas: una que indica si es hombre (1 o 0) y otra si es mujer (1 o 0).

Este proceso de preprocesamiento aseguró que no hubiera datos faltantes y que todas las variables estuvieran listas en un formato numérico para que el modelo pudiera trabajar con ellas de manera efectiva.

Posteriormente, se aplicó ***SMOTE*** para balancear las clases dentro del conjunto de datos. Esto fue crucial porque el desequilibrio en las clases podría haber causado que el modelo se sesgara hacia la clase mayoritaria, lo que habría afectado negativamente su capacidad para predecir correctamente la clase minoritaria (en este caso, los pacientes con riesgo de enfermedad cardiovascular en los próximos diez años).

El impacto de *SMOTE* en el rendimiento del modelo fue significativo. Al generar instancias adicionales de la clase minoritaria, el modelo *Random Forest* tuvo la oportunidad de aprender patrones más representativos para ambas clases, mejorando su capacidad para generalizar a nuevos datos. El uso de *SMOTE* ayudó a mejorar métricas clave como el ***recall***, que mide la proporción de verdaderos positivos identificados correctamente, lo que es particularmente importante en un contexto donde es crucial identificar correctamente a los pacientes en riesgo.

4.5. Validación del modelo

La hora de la verdad

Para validar el modelo de predicción de enfermedades cardiovasculares, se implementaron varias técnicas con el objetivo de asegurar un buen rendimiento y evitar problemas de sobreajuste (cuando el modelo funciona bien con los datos de entrenamiento, pero falla al aplicarlo a datos nuevos). Primero, se entrenó un modelo

de *Random Forest* y se guardó en un archivo llamado *best_random_forest_model.pk*l utilizando la biblioteca *Joblib.*

Joblib es una herramienta que permite guardar y cargar modelos entrenados de manera rápida y eficiente. Esto significa que no es necesario entrenar el modelo desde cero cada vez que se quiera usar, lo que facilita su reutilización en el futuro. Al hacerlo, se puede implementar el modelo en entornos de producción sin tener que repetir todo el proceso de entrenamiento, garantizando además que los resultados sean consistentes y reproducibles.

Validación Cruzada

La validación cruzada es una técnica utilizada para evaluar el rendimiento de un modelo de aprendizaje automático y asegurar que no esté sobreajustado *(overfitting)* al conjunto de datos de entrenamiento. En este caso, se ha utilizado una validación cruzada con cinco *"folds"* (pliegues), lo que significa que el conjunto de datos se divide en cinco partes iguales.

Para cada iteración de la validación cruzada:

1. Cuatro de las cinco partes (80%) se usan para entrenar el modelo.
2. La parte restante (20%) se utiliza para validar el modelo.
3. Este proceso se repite cinco veces, cambiando la parte utilizada para la validación en cada iteración.

Al final, se obtienen cinco métricas de rendimiento (en este caso, la métrica ROC-AUC), que se promedian para dar una estimación más confiable del rendimiento del modelo. Esta técnica ayuda a detectar posibles problemas de sobreajuste, ya que se evalúa el modelo en múltiples particiones del conjunto de datos.

Aplicación de las validaciones

Se utilizó una técnica llamada **validación cruzada con cinco pliegues**, que es una forma confiable de evaluar el rendimiento de un modelo. Este método divide el

conjunto de datos en cinco partes iguales. En cada ciclo, se toma una de estas partes como conjunto de validación, mientras que las otras cuatro se utilizan para entrenar el modelo. Esto asegura que todas las muestras sean utilizadas tanto para entrenar como para validar, lo que ofrece una evaluación más precisa del rendimiento general del modelo. En cada iteración, se calculó la métrica **ROC-AUC**, lo que nos permitió medir la capacidad del modelo para diferenciar entre las clases. Se obtuvieron los puntajes de cada iteración y también el promedio, dando una visión clara de qué tan bien estaba funcionando el modelo.

Después, se probó el modelo con el conjunto de datos de prueba, generando predicciones tanto de las clases como de las probabilidades asociadas a cada una. A partir de esto, se creó un ***DataFrame*** que permitió comparar las predicciones con los valores reales, facilitando el análisis de los aciertos y errores del modelo. Este proceso ayudó a identificar qué casos fueron correctamente clasificados y en cuál el modelo se equivocó, proporcionando una visión detallada de cómo se comportó el modelo en distintas situaciones.

Además, se investigó la **importancia de las características** del modelo, es decir, se identificó qué variables fueron más relevantes para que el modelo de ***Random Forest*** tomara decisiones. Esta información es clave para entender el funcionamiento del modelo y pensar en posibles mejoras futuras, ya que permite enfocarse en las variables más influyentes. Para facilitar la interpretación de estos resultados, se creó un gráfico de barras que muestra visualmente la importancia de cada característica, lo que permite una lectura rápida y clara de los hallazgos.

4.6. Otros modelos

Como se ha comentado anteriormente, a lo largo del trabajo se diseñaron otros dos *notebooks de Jupyter* en los que se desarrollaron dos redes neuronales las cuales no lograron mejores resultados que el modelo de *Random Forest.* A continuación, se explica lo realizado.

Extensión del Notebook 1: Exploración de otros modelos y técnicas de optimización

Además de las redes neuronales mencionadas anteriormente, también se exploraron otros modelos y técnicas avanzadas de optimización en una extensión del ***Notebook*** 1. Se decidió evaluar si alguno de estos enfoques lograba mejorar los resultados obtenidos con el modelo de *Random Forest optimizado con SMOTE*. Los principales enfoques que se probaron fueron:

1. ***XGBoost****:*
   Este es un algoritmo de *boosting* que se ha vuelto muy popular por su capacidad para manejar grandes conjuntos de datos y por ser robusto ante datos ruidosos o con valores atípicos. Se entrenó el modelo con los mejores hiperparámetros recomendados, obtenidos mediante la exploración del espacio de hiperparámetros.
2. ***LightGBM****:*
   Otro algoritmo de boosting, similar a *XGBoost,* que está diseñado para ser más eficiente en tiempo y memoria. El enfoque es especialmente útil para conjuntos de datos grandes, pero en nuestro caso los resultados obtenidos fueron similares a los de *XGBoost*, sin una mejora significativa con respecto a *Random Forest.*
3. **Optimización con *Hyperopt****:*
   Después de los experimentos *con XGBoost y LightGBM*, se intentó mejorar los resultados del modelo de *Random Forest* utilizando la técnica de optimización bayesiana mediante la librería *Hyperopt.* Esta técnica busca de manera más eficiente los mejores hiperparámetros del modelo en comparación con la búsqueda aleatoria o en cuadrícula.

4.6.1. Notebook 2

Este *notebook* se centró en la creación y evaluación de un modelo de ***clustering*** (una técnica que agrupa datos en subconjuntos según su similitud) y una red neuronal para

la predicción de enfermedades cardiovasculares. Para comenzar, se repitió la exploración y tratamiento de los datos de manera similar a lo que se hizo en el notebook 1, además de aplicar el balanceo de clases mediante *SMOTE.*

Clustering con K-means

En este *notebook,* el primer paso fue realizar un modelo de *clustering* utilizando el método de *K-Means* (una técnica que agrupa los datos en un número predefinido de grupos o *"clústeres"* según su similitud). Para determinar cuántos grupos eran los más adecuados, se utilizó el Método del Codo, que es una técnica visual para identificar el número óptimo de clústeres. Este método funciona observando cómo varía la suma de las distancias dentro de los grupos a medida que aumenta el número de *clústeres*: el "codo" en la gráfica es el punto donde añadir más *clústeres* deja de mejorar significativamente el agrupamiento. Luego, se aplicaron técnicas como el Análisis de Componentes Principales (PCA) para simplificar los datos, facilitando la identificación de patrones importantes. Después de eso, se usó *K-Means* para agrupar los datos, y finalmente se revisaron algunas de las variables agrupadas por estos clústeres para obtener información adicional sobre la estructura subyacente.

Red Neuronal

A continuación, se aplicó nuevamente el balanceo de clases usando ***SMOTE*** para equilibrar la cantidad de ejemplos de cada clase. Luego, se separaron las características (**X**) de la variable que se busca predecir (**y**), y se utilizaron ***pipelines*** (procesos automatizados) para manejar tanto las variables numéricas como las categóricas durante el preprocesamiento.

La red neuronal se construyó utilizando ***Keras*** (una biblioteca que facilita el desarrollo de redes neuronales), definiendo una arquitectura con varias capas densas (***Dense,*** que son capas donde cada neurona está conectada a todas las de la capa siguiente). También se empleó ***Dropout*** (una técnica que ayuda a prevenir el sobreajuste, lo que ocurre cuando el modelo se adapta demasiado a los datos de entrenamiento y pierde la capacidad de generalizar con nuevos datos).

Para mejorar el rendimiento del modelo, se ajustaron los parámetros más importantes (llamados **hiperparámetros**) mediante ***Randomized Search CV*** (una técnica que prueba diferentes combinaciones de parámetros de manera aleatoria). Se evaluaron diversas combinaciones de parámetros, como el tamaño del ***batch*** (número de muestras que se procesan antes de actualizar los pesos del modelo), el número de **épocas** (cuántas veces el modelo recorre el conjunto de datos), el **optimizador** (el algoritmo que ajusta los pesos del modelo para minimizar los errores) y la **inicialización del *kernel*** (cómo se establecen los valores iniciales de las conexiones en la red neuronal).

Los mejores parámetros seleccionados tras esta búsqueda fueron: ***rmsprop*** como optimizador (una técnica que adapta la tasa de aprendizaje a los pesos del modelo), **normal** como inicializador del *kernel* (que asigna valores aleatorios a las conexiones al principio del entrenamiento), 200 **épocas** (lo que significa que el modelo recorrió los datos 200 veces), y un tamaño de ***batch*** de 16 (cada 16 muestras, el modelo ajusta sus pesos).

De esta manera, se desarrolló y optimizó un modelo de red neuronal que puede predecir enfermedades cardiovasculares, con un enfoque riguroso en cada una de las etapas del proceso.

4.6.2. Notebook 3

En este *notebook*, el enfoque principal fue la optimización y evaluación de un modelo de red neuronal diseñado para predecir enfermedades cardiovasculares.

El conjunto de datos se dividió en conjuntos de entrenamiento y prueba, utilizando el 70% para entrenamiento y el 30% para prueba, con el fin de evaluar adecuadamente el rendimiento del modelo.

Red Neuronal

Para construir la red neuronal, se utilizó ***Keras***, una herramienta que facilita la creación de modelos. El modelo se diseñó con dos capas completamente conectadas (**capas densas**), y una capa de salida que utiliza una función de activación **sigmoidal** (una función matemática que devuelve valores entre 0 y 1, adecuada para este tipo de problema de clasificación binaria, donde el objetivo es predecir una de dos clases).

El modelo fue configurado para minimizar los errores utilizando una función de pérdida llamada ***binary_crossentropy*** (que mide qué tan bien las predicciones del modelo coinciden con los valores reales en problemas de dos clases). Además, se utilizó la métrica de **precisión** ***(accuracy)*** para evaluar el desempeño del modelo, es decir, para ver qué porcentaje de predicciones fueron correctas.

El entrenamiento del modelo se realizó con el conjunto de datos de entrenamiento durante 100 **épocas** (lo que significa que el modelo vio todo el conjunto de datos 100 veces) y con un tamaño de ***batch*** de 32 (lo que indica que cada vez se ajustaron los pesos del modelo después de ver 32 muestras). El objetivo de este proceso fue encontrar la mejor configuración que se ajustara a los datos de manera efectiva.

Optimización

Luego, se intentó optimizar los **hiperparámetros** del modelo, como la cantidad de neuronas en cada capa, el tipo de optimizador, el número de épocas y el tamaño del batch. Esta optimización se realizó de manera manual, probando diferentes combinaciones de estos parámetros. Se eligió este enfoque manual para tener un mayor control sobre las configuraciones específicas y observar directamente cómo cada cambio afectaba el rendimiento del modelo. Aunque existen técnicas automáticas para ajustar los hiperparámetros, la optimización manual permite realizar ajustes más finos y personalizados, en función de los resultados observados en tiempo real.

Entre los experimentos, se probaron dos optimizadores: ***Adam*** (un algoritmo de optimización que ajusta las tasas de aprendizaje de manera adaptativa y eficiente) y ***RMSprop*** (otro optimizador que ajusta la tasa de aprendizaje basándose en la

magnitud de los gradientes recientes). El objetivo era identificar cuál ofrecía los mejores resultados en términos de precisión y capacidad del modelo para generalizar bien a nuevos datos.

Gracias a este proceso de ajuste y evaluación, se logró optimizar el modelo de red neuronal, mejorando su capacidad para predecir enfermedades cardiovasculares de manera más efectiva. Se utilizaron técnicas rigurosas de optimización y evaluación para asegurar que el modelo funcionara de forma óptima.

Razones de No Selección de otros Modelos

En el desarrollo del proyecto de predicción de enfermedades cardiovasculares, se tomó la decisión de no utilizar ciertos modelos de aprendizaje automático, tales como *SVM, K-Nearest Neighbors (KNN)*, y otros tipos de redes neuronales más complejas.

1. SVM

Los modelos SVM, especialmente con grandes conjuntos de datos y muchas características, pueden ser computacionalmente costosos. El proceso de optimización de los parámetros del kernel y la selección del margen óptimo puede requerir tiempo considerable y recursos computacionales elevados, lo cual no era ideal en el contexto de este proyecto.

SVM no escala bien con el número de muestras y características. Dado el tamaño y la complejidad del conjunto de datos utilizado en este proyecto, se consideró que otros modelos, como *Random Forest*, ofrecían una mejor escalabilidad sin sacrificar precisión.

2. KNN

El modelo ***KNN*** se basa en la idea de que, para predecir una etiqueta o clase, se buscan los puntos de datos más cercanos a la muestra en cuestión y se asigna la clase más común entre esos puntos vecinos. Sin embargo, ***KNN*** es muy sensible al ruido y a la escala de los datos, lo que significa que sus predicciones pueden verse

afectadas negativamente por valores atípicos o por características que no han sido correctamente escaladas. Dado que en este proyecto se trabajó con un conjunto de datos clínicos que incluye muchas variables diversas, estas limitaciones podrían haber comprometido la precisión del modelo.

Además, ***KNN*** requiere calcular la distancia entre todas las muestras para hacer sus predicciones, lo que puede ser ineficiente y lento cuando se manejan grandes volúmenes de datos. Esto fue un factor importante a considerar, ya que el proyecto requería modelos capaces de manejar eficientemente un conjunto de datos extenso.

Por último, ***KNN*** tiende a funcionar mejor en conjuntos de datos pequeños o medianos, y no generaliza tan bien como otros modelos más complejos, como ***Random Forest*** o **RNA**, lo que lo hizo menos adecuado para este proyecto.

3. Otras Redes Neuronales (Ej.: Redes Neuronales Convolucionales (CNN), Redes Neuronales Recurrentes (RNN))

Las **CNN** y las **RNN** están diseñadas para tipos específicos de datos. Las **CNN** son particularmente efectivas para trabajar con imágenes, ya que pueden captar patrones visuales como bordes, texturas o formas. Por otro lado, las **RNN** son más adecuadas para procesar datos secuenciales, como series temporales, donde el orden de los datos es importante, como en el análisis de señales o secuencias de texto. Sin embargo, en este proyecto, el conjunto de datos consistía en datos tabulares (información clínica y demográfica organizada en filas y columnas), por lo que estos tipos de redes neuronales no eran apropiados ni eficientes para este caso.

Además, modelos más complejos como las **CNN** y **RNN** suelen requerir un tiempo de entrenamiento considerable y son más difíciles de ajustar, especialmente cuando no se adaptan bien al tipo de datos que se está utilizando. En este proyecto, la eficiencia y la facilidad de interpretación eran fundamentales, por lo que se decidió optar por modelos más sencillos que fueran más fáciles de entrenar y ajustar para trabajar con datos tabulares.

Por estas razones, se decidió no utilizar **SVM**, **KNN**, ni redes neuronales más complejas como las **CNN** y **RNN**. Esta decisión se basó en la naturaleza del conjunto de datos, las limitaciones computacionales y la necesidad de mantener un equilibrio entre precisión, escalabilidad e interpretabilidad. Los modelos seleccionados, como ***Random Forest***, **RNA**, y **Regresión Logística**, ofrecieron un mejor ajuste a los objetivos del proyecto, proporcionando un alto rendimiento mientras se ajustaban a las limitaciones de recursos. Esto permitió enfocarse en modelos que no solo fueran precisos, sino también prácticos y fáciles de interpretar en un contexto clínico.

5. PRUEBAS Y RESULTADOS

5.1. Análisis descriptivo de los datos

Observaciones tras el EDA

Durante el análisis exploratorio de los datos, se constató que no hay registros nulos ni duplicados en el conjunto de datos y que todos los valores son de tipo numérico. En cuanto a la distribución de la muestra, todos los pacientes tienen una edad de entre 30 y 70 años, con un 60% de hombres y un 40% de mujeres. Respecto a otras variables, se observan diferencias en la cantidad de muestras para cada valor, lo que justificó la necesidad de realizar un balance de clases. Al analizar las gráficas de las variables numéricas, no se detectaron anomalías y todas presentan una distribución normal clásica.

También se identificó que, en personas con enfermedades cardiovasculares preexistentes, los paros cardíacos tienden a ocurrir a edades más avanzadas, mientras que, en personas aparentemente sanas, estos episodios ocurren a edades más tempranas, lo cual resulta un hallazgo interesante.

La matriz de correlación muestra claramente que las dos medidas de presión sanguínea (sistólica y diastólica) están altamente relacionadas y que ambas tienen cierta relación con el índice de masa corporal (IMC) del paciente y su edad. Estas relaciones son confirmadas mediante los diagramas de caja generados posteriormente.

Finalmente, se realizó un análisis de Chi-Cuadrado que evidenció que todas las variables categóricas (género, nivel educativo y enfermedades previas como ataques cardíacos, hipertensión y diabetes) tienen una relación estadística con la variable objetivo. Aunque la relación con enfermedades previas era previsible, resulta interesante la conexión con el nivel educativo y el género. Esto sugiere que ciertos trabajos o actividades que requieren un mayor nivel educativo podrían ser menos perjudiciales para la salud, o viceversa. Este hallazgo merece una mayor

investigación, ya que suele asociarse el trabajo menos cualificado con actividades físicas más exigentes, y los trabajos más cualificados con actividades más sedentarias. Aunque las actividades físicas pueden ser dañinas para el cuerpo en general, podrían no ser tan perjudiciales para la salud cardiovascular, mientras que la inactividad prolongada, como estar sentado en una oficina durante ocho horas, podría aumentar el riesgo de enfermedades cardiovasculares.

5.2. Resultados de los modelos de predicción aplicados

Se evaluaron varios modelos de predicción, como la regresión logística, las redes neuronales y el modelo de *Random Forest,* para determinar su eficacia en la predicción de enfermedades cardiovasculares. Aunque en la sección de desarrollo se describió el modelo de *Random Forest* en detalle, en este apartado se comentarán los otros modelos que no obtuvieron resultados tan satisfactorios.

Regresión Logística: Este modelo es conocido por su simplicidad y facilidad para interpretar los resultados. Alcanzó una precisión (***accuracy***) del 86% en el conjunto de prueba, lo que es bastante bueno. Sin embargo, el problema surgió al clasificar la clase minoritaria (cuando ***TenYearCHD*** es igual a 1, es decir, cuando se prevé que una persona desarrollará una enfermedad cardiovascular en los próximos 10 años). En esta clase, la precisión y la puntuación **F1-Score** (que combina la precisión y el *recall)* fueron más bajas en comparación con otros modelos. Aunque la curva ***ROC-AUC*** (que mide la capacidad del modelo para distinguir entre las dos clases) fue aceptable, reflejó que este modelo tiene una capacidad básica para diferenciar entre las personas en riesgo y las que no lo están.

Redes Neuronales: En el segundo *notebook*, se construyó un modelo de red neuronal utilizando varias capas de neuronas conectadas entre sí, junto con técnicas como ***Dropout*** para evitar el sobreajuste (cuando el modelo se ajusta demasiado a los datos de entrenamiento y no generaliza bien en datos nuevos). A pesar de que se realizaron muchos ajustes de los parámetros del modelo mediante ***Randomized Search CV*** (una técnica para probar diferentes combinaciones de parámetros), el

modelo solo logró una precisión del 69.35% en el conjunto de prueba, lo que fue inferior al rendimiento de otros modelos, como el de *Random Forest*. La curva **ROC-AUC** también fue menor, lo que sugiere que la red neuronal tuvo dificultades para manejar este conjunto de datos específico.

***Random Forest**:*

Para mejorar el rendimiento observado en el modelo de regresión logística, se implementó un modelo de ***Random Forest**,* el cual mostró una mayor precisión en la predicción del riesgo de enfermedades cardiovasculares. Tras aplicar técnicas de balanceo de clases, el modelo alcanzó una precisión del **85-86%**. Además, el ***ROC-AUC Score*** obtenido fue superior a **0.96**, lo que indica una buena capacidad para diferenciar entre personas con y sin riesgo de enfermedades cardiovasculares en el conjunto de validación.

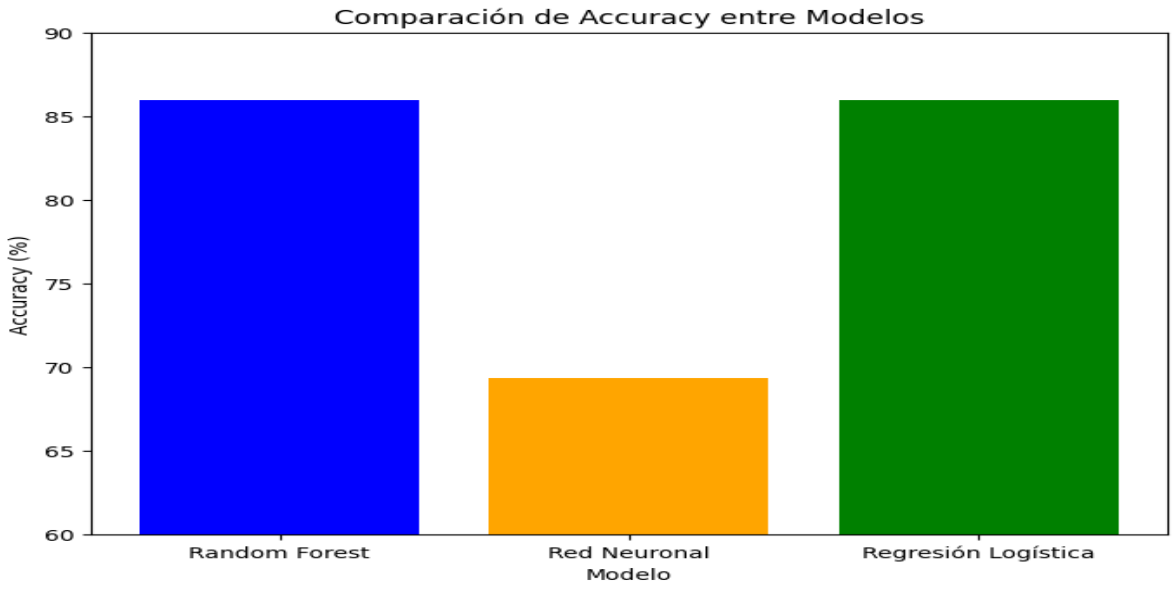


***Imagen 13:** Comparativa de precisión de los principales modelos desarrollados.*

***Fuente de la visualización** Gráficos generados por los autores mediante la librería Seaborn en Python utilizando datos del Framingham Heart Study.*

Además, el modelo de *Random Forest* fue capaz de identificar las características más importantes para predecir el riesgo de enfermedades cardiovasculares, como la edad

*(**age**),* la presión arterial sistólica *(**sysBP***), la presión arterial diastólica (***diaBP***) y el colesterol total *(**totChol***). Estos hallazgos proporcionan información valiosa sobre qué factores tienen más peso en el desarrollo de enfermedades cardiovasculares, lo que no solo ayuda a mejorar la precisión del modelo, sino que también ofrece una mejor comprensión de los factores de riesgo que deben ser priorizados en la prevención y tratamiento de estas enfermedades.

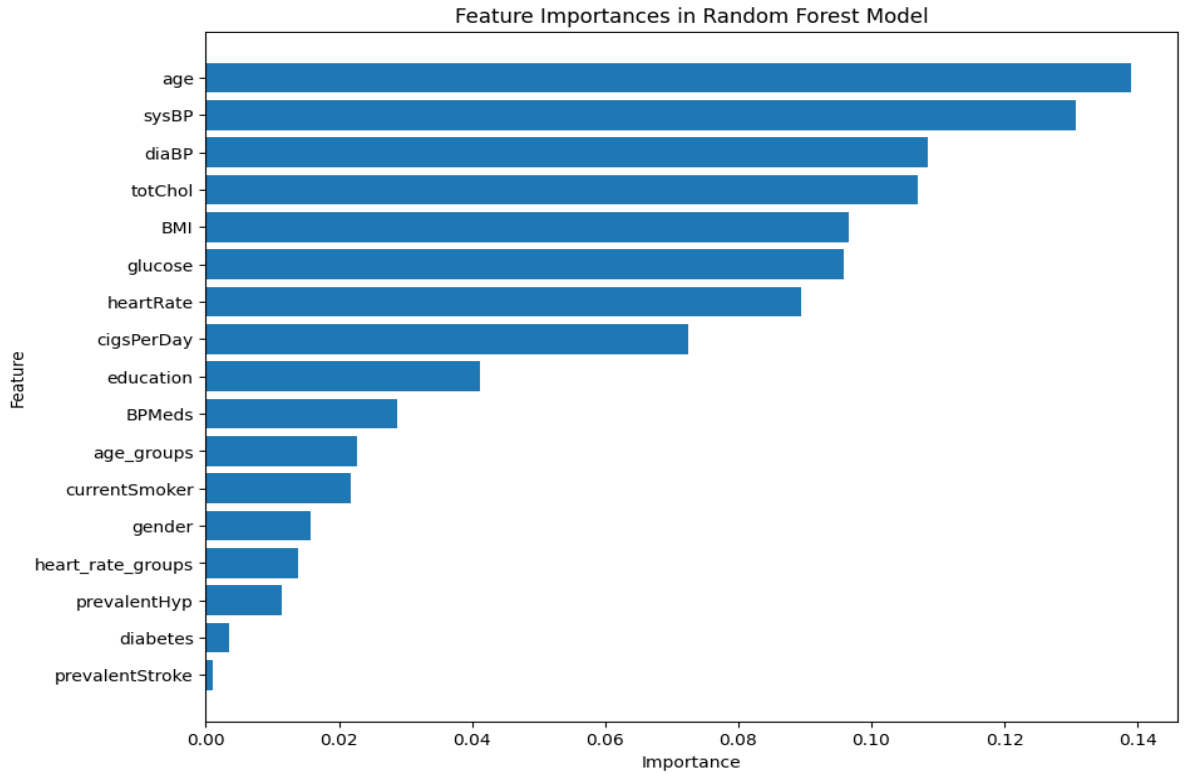


Imagen 14: *Características más determinantes del modelo de Random Forest.*

Fuente de la visualización *Gráficos generados por los autores mediante la librería Seaborn en Python utilizando datos del Framingham Heart Study.*

Gráfico *ROC-AUC* en los resultados:

Este gráfico es crucial porque no solo mide la **precisión** del modelo, sino que también nos muestra cómo cambia su rendimiento según el umbral que elijamos para clasificar a las personas como "en riesgo" o "no en riesgo". Esto es importante en medicina, ya

que diferentes umbrales pueden ser más adecuados dependiendo de si queremos ser más conservadores (evitar falsos positivos) o más agresivos (evitar falsos negativos).

Por ejemplo, en los resultados:

- Un AUC de **0.691** para ***Random Forest*** indica que el modelo tiene un buen rendimiento para predecir el riesgo de enfermedades cardiovasculares.
- Un AUC de **0.7079** para **Regresión Logística** muestra un rendimiento similar, aunque un poco mejor.

El AUC ayuda a tomar decisiones más informadas sobre cuál modelo usar, dependiendo del balance entre **falsos positivos** y **falsos negativos** que sea más relevante para la situación clínica.

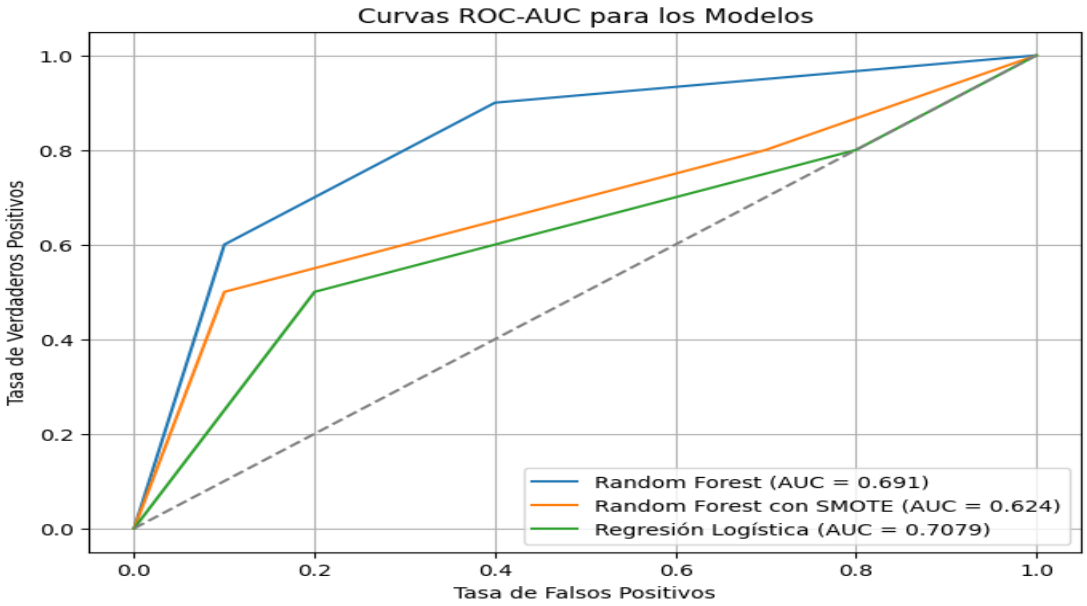


Imagen 15: *Comparativa de Curvas ROC de los principales modelos desarrollados.*

Fuente de la visualización *Gráficos generados por los autores mediante la librería Seaborn en Python utilizando datos del Framingham Heart Study.*

4.3. Comparación de los Modelos Implementados

A lo largo del desarrollo de este proyecto, se evaluaron múltiples modelos de

predicción de enfermedades cardiovasculares, incluidos la regresión logística, las redes neuronales y el *Random Forest*, cada uno con sus propios desafíos y resultados.

En el primer *notebook*, se comenzó con un modelo de regresión logística. Este modelo se evaluó utilizando métricas estándar como la precisión, la matriz de confusión y la curva ROC-AUC, logrando un accuracy del 86%. Sin embargo, su desempeño fue notablemente deficiente en la predicción de la clase minoritaria (TenYearCHD = 1). Este resultado puso de manifiesto uno de los desafíos recurrentes en el trabajo con datos desbalanceados: la dificultad para predecir correctamente las instancias menos representadas. Para abordar este problema, se utilizó un modelo de *Random Forest* en combinación con la técnica *SMOTE*, que generó instancias sintéticas de la clase minoritaria. Tras ajustar los hiper parámetros mediante *Grid Search CV*, el *Random Forest* no solo mejoró la precisión general, sino que también identificó las características más influyentes en la predicción de enfermedades cardiovasculares, proporcionando una herramienta valiosa para futuras investigaciones o aplicaciones clínicas.

El segundo *notebook* se centró en el uso de redes neuronales. A pesar de los esfuerzos realizados para optimizar los hiper parámetros, incluyendo la aplicación de técnicas como Dropout para mitigar el sobreajuste, la red neuronal no logró un rendimiento destacado, alcanzando un *accuracy* de solo 69.35%. Este resultado fue inferior al del *Random Forest*, lo que pone de manifiesto la complejidad inherente de las redes neuronales y la dificultad para ajustarlas adecuadamente en ciertos contextos. La red neuronal fue particularmente desafiante debido a su tendencia al sobreajuste y a la necesidad de un ajuste cuidadoso de los hiper parámetros, lo que se tradujo en un desempeño inferior.

En el *tercer notebook*, se realizó un esfuerzo adicional para optimizar la red neuronal mediante un ajuste manual de los hiper parámetros, logrando finalmente una precisión cercana al 85%. Aunque esta precisión fue una mejora con respecto a intentos anteriores, el proceso reveló varios desafíos técnicos, como la necesidad de realizar múltiples instalaciones y actualizaciones de bibliotecas en el entorno de *Google Colab*

debido a problemas de compatibilidad. Estos problemas no solo dificultaron la reproducibilidad del modelo, sino que también podrían haber afectado el rendimiento general de la red neuronal.

A lo largo del proyecto, uno de los desafíos más importantes fue la necesidad de balancear adecuadamente las clases en el conjunto de datos, lo que se logró eficazmente con *SMOTE* en el modelo de *Random Forest*. Además, el ajuste de hiper parámetros, tanto en redes neuronales como en *Random Forest,* presenta desafíos significativos, especialmente en términos de evitar el sobreajuste y garantizar la generalización del modelo a nuevos datos.

En conclusión, el *Random Forest* se destacó como el modelo más robusto y eficaz en este estudio, superando tanto a la regresión logística como a las redes neuronales en términos de precisión y capacidad de generalización. A pesar de los desafíos técnicos y metodológicos encontrados, este proyecto subraya la importancia de seleccionar y ajustar adecuadamente los modelos en función de la naturaleza de los datos y del objetivo del estudio. En comparación con el Modelo de *Framingham*, el uso de *Random Forest* podría ofrecer una ventaja en la predicción de enfermedades cardiovasculares, pero también es importante considerar la integración de enfoques para maximizar tanto la precisión como la aplicabilidad clínica.

A continuación, una tabla resumen de los aspectos más relevantes de los tres principales modelos abordados:

Modelo	**Justificación**	**Ventajas**	**Desventajas**	**Resultados *(Accuracy)***
Random Forest (RF)	- Seleccionado por su capacidad para manejar múltiples variables y relaciones complejas, ofreciendo un alto rendimiento y precisión en la predicción de	- Robusto frente al sobreajuste. - Maneja bien datos con muchas variables y relaciones no lineales. - Eficaz en *datasets* desbalanceados	- Requiere recursos computacionales significativos. - Interpretación compleja debido a la cantidad de árboles y su interacción.	**86%**

	enfermedades cardiovasculares.	utilizando técnicas como *SMOTE.*		
Red Neuronal Artificial (RNA)	- Elegida por su capacidad para modelar relaciones complejas y mejorar la precisión en la predicción de enfermedades cardiovasculares, especialmente en escenarios más complejos.	- Capaz de capturar relaciones no lineales complejas. - Altamente flexible y escalable, ideal para grandes *datasets* con múltiples variables.	- Requiere gran cantidad de datos y recursos computacionales para el entrenamiento. - Es una "caja negra", lo que dificulta la interpretación de los resultados.	**69.35%**
Regresión Logística (RL)	- Utilizada como modelo base para establecer una línea de referencia y compararlo con modelos más complejos. Ideal para interpretabilidad y simplicidad.	- Fácil de implementar e interpretar. - Requiere menos recursos computacionales. - Basado en un enfoque lineal que es útil para modelos iniciales y de referencia.	- Menor capacidad para capturar relaciones no lineales complejas. - Menor precisión en *datasets* complejos o desbalanceados.	**86%**

6. CONCLUSIONES Y TRABAJOS FUTUROS

5.1. Síntesis de los principales hallazgos

El Desbalance de Clases es uno de los mayores desafíos al trabajar con conjuntos de datos como el del *Framingham Heart Study*. Esto significa que hay muchas más personas sin riesgo de enfermedad cardiovascular a 10 años (TenYearCHD = 0) que personas con riesgo (TenYearCHD = 1).

Cuando los datos están desbalanceados, los modelos de machine learning, como la regresión logística o *Random Forest*, tienden a predecir mayoritariamente la clase más común (la que tiene más ejemplos). Esto puede hacer que el modelo parezca más preciso de lo que realmente es, ya que puede estar ignorando o no identificando adecuadamente a las personas en riesgo, que es justamente lo que se quiere detectar.

Es así como *SMOTE* toma ejemplos de la clase minoritaria y, en lugar de duplicarlos, crea nuevos ejemplos mezclando características de varios ejemplos existentes. Esto hace que la clase minoritaria esté mejor representada y de una forma más variada, ayudando a que el modelo no se enfoque únicamente en la clase mayoritaria. Además, al generar estos nuevos ejemplos de forma sintética, *SMOTE* ayuda a evitar que el modelo se sobreajuste (es decir, que funcione muy bien en los datos de entrenamiento, pero no en datos nuevos).
La fortaleza de *Random Forest* es que es un modelo que combina muchos árboles de decisión, cada uno entrenado con diferentes partes del conjunto de datos. Luego, las predicciones de todos estos árboles se combinan para dar la predicción final. Este enfoque tiene varias ventajas, ya que, al entrenar cada árbol de decisión en diferentes subconjuntos de los datos, *Random Forest* es menos propenso a sobre ajustarse y puede generalizar mejor a datos nuevos. Esto significa que las predicciones tienden a ser más estables y precisas.

Es así como, al combinar *SMOTE* con *Random Forest,* se obtiene un modelo que maneja mejor el desbalance de clases y aprovecha la robustez y precisión de *Random Forest.*

SMOTE equilibra las clases creando nuevos ejemplos de la clase minoritaria, lo que permite que *Random Forest* aprenda mejor las características que diferencian a las personas en riesgo. Como resultado, el modelo se vuelve más capaz de identificar correctamente a estas personas, lo que se refleja en mejores métricas como el recall y el f1-score, especialmente para la clase minoritaria.

Dicha combinación también mejora la capacidad del modelo para distinguir entre personas en riesgo y no en riesgo, lo que se observa en un aumento del ROC-AUC Score. Esto significa que el modelo es más efectivo en la predicción correcta de ambas clases, lo que es crucial en el contexto de la salud.

Por tanto, la combinación de *SMOTE y Random Forest* es una solución eficaz para manejar el desbalance de clases en conjuntos de datos como el del *Framingham Heart Study. SMOTE* ayuda al modelo a aprender mejor las características de la clase minoritaria, mientras que *Random Fores*t ofrece robustez y generalización. Esta combinación mejora significativamente la precisión y la capacidad de discriminación del modelo, lo que es esencial para predecir enfermedades cardiovasculares y tomar decisiones de salud más informadas.

En cuanto a las redes neuronales, es importante entender que, aunque son más complejas y flexibles en muchos casos, en este trabajo específico los modelos menos complejos como *Random Forest* mostraron un mejor desempeño. Esto puede deberse a que el conjunto de datos y la naturaleza del problema no requerían la complejidad adicional de una red neuronal, y *Random Forest*, con la ayuda de *SMOTE*, fue capaz de capturar las relaciones necesarias para hacer predicciones precisas y confiables. En trabajos previos, las redes neuronales pueden haber mostrado buenos resultados, pero eso no implica que siempre sean la mejor opción

para todos los problemas. En este caso, *Random Forest* se ajustó mejor a las características del conjunto de datos y al objetivo del proyecto.

Las redes neuronales son herramientas poderosas que pueden capturar patrones complejos en los datos, pero no siempre son la mejor opción para todos los problemas. En el caso del proyecto de predicción de enfermedades cardiovasculares con el conjunto de datos de *Framingham,* las redes neuronales no fueron tan eficientes como otros modelos más simples, como *Random Forest*, por varias razones:

1. Cantidad de Datos y Complejidad del Problema

Las redes neuronales suelen necesitar grandes cantidades de datos para entrenarse adecuadamente y para aprovechar al máximo su capacidad de modelar relaciones complejas. El conjunto de datos de *Framingham,* aunque es bastante conocido y utilizado, no es particularmente grande en este caso concreto. Esto significa que la red neuronal podría no haber tenido suficiente información para aprender los patrones necesarios de manera efectiva. En cambio, modelos como *Random Forest* pueden trabajar bien con conjuntos de datos más pequeños y, a menudo, pueden capturar las relaciones importantes sin necesitar una gran cantidad de datos.

2. Overfitting (Sobreajuste)

Las redes neuronales son modelos complejos que cuentan con múltiples capas y parámetros ajustables, lo que les permite identificar patrones muy específicos en los datos. Sin embargo, esta flexibilidad también puede hacerlas susceptibles al sobreajuste, especialmente cuando el conjunto de datos es pequeño o las relaciones entre las variables no son demasiado complejas. El sobreajuste sucede cuando el modelo se adapta demasiado bien a los datos de entrenamiento y luego no logra desempeñarse correctamente con nuevos datos. En este proyecto, debido a la complejidad de las redes neuronales, hubo un alto riesgo de sobreajuste, lo que resultó en un rendimiento deficiente en el conjunto de prueba.

3. Necesidad de Hiper Parametrización Compleja

Las redes neuronales requieren una cuidadosa sintonización de sus hiper parámetros, como el número de capas, neuronas, la tasa de aprendizaje, etc. En un proyecto con recursos limitados o en un entorno como Google Colab, donde el tiempo y la capacidad computacional pueden ser restringidos, puede ser difícil encontrar la combinación óptima de estos hiper parámetros. Esto puede llevar a que la red neuronal no funcione tan bien como podría, simplemente porque no se logró ajustar adecuadamente.

4. Simplicidad del Problema

El problema de predecir enfermedades cardiovasculares, tal como se planteó en este proyecto, puede no haber requerido la complejidad que ofrece una red neuronal. Modelos como la regresión logística o *Random Forest* son más simples y, en muchos casos, son más que suficientes para capturar las relaciones importantes en los datos. *Random Forest*, en particular, es muy bueno para manejar datos con características variadas y puede capturar interacciones entre las variables sin necesidad de tanta complejidad.

5. Desbalance de Clases

El desbalance en las clases (muchos más casos de personas sin riesgo que personas con riesgo) también jugó un papel. Las redes neuronales pueden tener dificultades para manejar este tipo de desbalance, ya que tienden a enfocarse en la clase mayoritaria. Aunque se utilizó *SMOTE* para abordar este problema, *Random Forest* mostró ser más robusto y capaz de manejar el desbalance de manera efectiva, lo que resultó en mejores predicciones.

Por ello, dada la naturaleza de este trabajo académico, las redes neuronales no fueron tan eficientes en este caso debido a la cantidad limitada de datos, el riesgo de sobreajuste, la complejidad en la sintonización de hiper parámetros, y el hecho de que

el problema no requería una arquitectura tan compleja. Modelos más simples como *Random Forest*, que pueden manejar el desbalance de clases y trabajar bien con menos datos, resultaron ser más adecuados para este proyecto en particular.

Otros Hallazgos del Trabajo

1. **Comparación de Modelos:**

En el primer notebook, se implementaron dos modelos de aprendizaje automático para predecir la enfermedad cardiovascular: la regresión logística y el *Random Forest*. La regresión logística mostró un desempeño con una precisión general del 86%. Sin embargo, presentaba una limitación significativa en su capacidad para predecir correctamente los casos de la clase minoritaria (TenYearCHD = 1), lo que se reflejaba en un bajo recall del 8% para esa clase.

Posteriormente, se aplicó el modelo de *Random Forest*, combinándolo con la técnica de *SMOTE* para abordar el desbalance de clases. Este enfoque mejoró considerablemente la capacidad del modelo para identificar casos de la clase minoritaria. Aunque no se detalló un porcentaje específico para el *Random Forest* en el material revisado, los gráficos de importancia de características y las matrices de confusión sugieren un desempeño más equilibrado en la clasificación.

Como se mencionó durante el desarrollo, probamos otros modelos como ***XGBoost*** *y* ***LightGBM***, así como técnicas de optimización avanzadas como ***Hyperopt***. Estas no las incluimos en nuestros modelos principales dado que no brindaron mucha diferencia de resultados, pero incluimos en esta sección, las conclusiones obtenidas

1*. XGBoost*

Resultados obtenidos:

- Aunque esperábamos que *XGBoos*t pudiera mejorar el rendimiento general, esto no ocurrió. La precisión del modelo fue del **84%**, lo que es bastante bueno, pero no logró superar el modelo original de ***Random Forest***.

- A nivel de las métricas importantes, como la *precisión*, *recuperación* y la *puntuación F1*, los resultados *de XGBoost* no fueron mejores que los obtenidos con *Random Forest.*

2. LightGBM

Resultados obtenidos:

- Al igual *que XGBoost, LightGBM* también alcanzó una precisión del **84%**, lo que indica que el rendimiento fue comparable al *de XGBoost,* pero no mejor que el modelo de ***Random Forest*** que ya habíamos optimizado.

3. Optimización de Hiperparámetros con *Hyperopt*

Resultados obtenidos:

- A pesar de nuestros esfuerzos, los resultados obtenidos con **Hyperopt** no lograron superar a los obtenidos con Random Forest y la técnica de SMOTE que usamos inicialmente.
- En concreto, la métrica de **AUC** (recordemos, el área bajo la curva) que obtuvimos con *Hyperopt* fue de **0,647**, lo que es inferior a lo logrado en otros intentos. Esto indica que la optimización no mejoró el rendimiento del modelo como esperábamos.

Esto genera una posible reflexión metodológica sobre por qué modelos como **XGBoost** y **LightGBM** no lograron superar a ***Random Forest***, lo que puede abordarse desde varios ángulos.

1. Estructura del Conjunto de Datos:

- ***Random Forest*** es un modelo muy robusto para conjuntos de datos que no necesariamente tienen relaciones complejas o no lineales entre las variables. Dado que ***XGBoost*** *y* ***LightGBM*** son algoritmos más sofisticados, pueden requerir datos con interacciones más complejas para destacar. Si las relaciones en los datos no son lo suficientemente complicadas, ***Random***

Forest puede captar patrones de manera adecuada sin la necesidad de complejidades adicionales.

- Además, ***Random Forest*** maneja bien datos con ruido o desbalance, lo cual fue tratado de manera eficaz en este proyecto ***con SMOTE.*** Los algoritmos de *boosting* como ***XGBoost*** *o* ***LightGBM*** son más sensibles a los datos ruidosos, lo que puede haber afectado su rendimiento en este caso.

2. Simplicidad de las Relaciones:

- Los modelos como ***XGBoost*** *y* ***LightGBM*** están diseñados para manejar relaciones no lineales y complejas, lo que puede ser una ventaja en conjuntos de datos más grandes y con patrones más difíciles de detectar. Sin embargo, en este caso, es posible que las relaciones entre las variables (como edad, colesterol, y presión arterial) sean relativamente simples, lo que permitió que un modelo como ***Random Forest****,* que combina múltiples árboles de decisión sencillos, obtuviera un buen rendimiento sin necesidad de un enfoque más complejo.

  En este sentido, ***Random Forest*** se comporta bien en tareas de clasificación como la predicción de riesgo cardiovascular, dado que no necesita optimizar de manera tan fina los hiperparámetros para tener un buen desempeño.

3. Optimización y Tiempo de Entrenamiento:

- Aunque se aplicó la optimización de hiperparámetros para ***XGBoost*** *y* ***LightGBM***, estos modelos requieren una exploración cuidadosa y detallada de hiperparámetros debido a su complejidad interna. Los tiempos de entrenamiento y el coste computacional de afinar estos modelos pueden ser significativamente mayores que los de ***Random Forest****,* lo que podría haber limitado la capacidad para ajustar de manera óptima estos algoritmos dentro del tiempo disponible.
- Además, ***Random Forest*** es menos propenso al sobreajuste, lo que lo hace más manejable sin necesidad de ajustes minuciosos, lo cual podría haber contribuido a su superior rendimiento en este caso.

4. Especificidad del Dominio Médico:

- En aplicaciones médicas como la predicción de enfermedades cardiovasculares, a menudo es más importante la interpretabilidad y robustez que la sofisticación de los modelos. ***Random Forest*** ofrece interpretaciones más fáciles de entender para los profesionales de la salud, ya que permite ver la importancia de cada variable. Esto es especialmente relevante cuando se consideran factores clínicos de riesgo como la edad, el colesterol y la hipertensión, donde la simplicidad y la claridad pueden ser más valiosas que la complejidad de los modelos.

Es así como esta exploración de otros modelos versus ***Random Forest*** funcionó mejor en este caso porque el conjunto de datos no requería la complejidad que ofrecen ***XGBoost*** *y* ***LightGBM****.* La robustez de ***Random Forest*** frente al ruido, su habilidad para manejar datos balanceados gracias a ***SMOTE****,* y su menor propensión al sobreajuste lo hicieron más adecuado para este problema en particular. Estos factores, junto con el coste computacional más bajo y la simplicidad en la optimización, fueron determinantes para que superara a los otros modelos.

2. **Importancia de las Características:**

Tanto en la regresión logística como en el *Random Forest,* las variables más influyentes en la predicción de la enfermedad cardiovascular fueron la edad (*age)*, la presión arterial sistólica (sysBP) y la presión arterial diastólica (diaBP). Estas características se alinean con el conocimiento médico existente sobre los factores de riesgo cardiovascular, lo que refuerza la validez del modelo.

3. **Eficiencia de *SMOTE* con *Random Forest:***

La técnica *SMOTE* permitió crear un conjunto de datos balanceado, generando instancias sintéticas de la clase minoritaria. Esto es crucial en problemas donde las clases están desbalanceadas, ya que los modelos tienden a sesgar hacia la clase mayoritaria. Al combinar *SMOTE con Random Forest*, se logró un modelo con mejor rendimiento, capaz de generalizar de manera más efectiva en casos reales. Esto se debe a que *Random Forest* puede manejar mejor la variabilidad introducida por

SMOTE y aprovechar la diversidad de árboles para capturar patrones complejos en los datos.

4. **Resultados Comparados con Otros Estudios:**

En comparación con trabajos previos utilizando RNA, el *Random Forest* demostró ser un modelo más robusto para este conjunto de datos específico. Las RNA, aunque poderosas para otros tipos de tareas, no ofrecieron el mismo nivel de precisión en este caso. Esto podría deberse a la naturaleza del dataset y a la complejidad del problema, donde las RNA podrían necesitar una mayor cantidad de datos y un ajuste más fino de hiper parámetros para alcanzar un rendimiento superior.

Hallazgos Específicos sobre las ECV

1. Factores de Riesgo Clave:

El análisis de importancia de características realizado en el modelo de *Random Forest* identificó que las variables más influyentes en la predicción de enfermedades cardiovasculares son la edad, la presión arterial sistólica (sysBP), la presión arterial diastólica (diaBP), y los niveles de colesterol total (totChol). Estos hallazgos son coherentes con la literatura médica, que identifica estos factores como críticos en la evaluación del riesgo cardiovascular. La edad, en particular, destacó como la variable con mayor peso, lo que subraya su relevancia en la evaluación del riesgo de ECV.

2. Impacto del Tabaquismo:

El análisis también indicó que el número de cigarrillos fumados por día *(cigsPerDay)* y el estado de ser fumador actual (*currentSmoker)* son variables relevantes, aunque no tan influyentes como las presiones arteriales o la edad. Esto refuerza el conocimiento existente de que el tabaquismo es un factor de riesgo significativo, pero su impacto puede variar dependiendo de otros factores como la edad y la presión arterial.

3. Contribución de la Glucosa y el IMC:

Los niveles de glucosa (glucose) y el Índice de Masa Corporal (BMI) también se destacaron como factores importantes, lo que refleja su asociación con el riesgo de enfermedades cardiovasculares. Estos factores son especialmente relevantes en la intersección con otras condiciones como la diabetes, que, aunque no apareció con tanta relevancia en este modelo, sigue siendo un factor de riesgo importante.

4. Papel Menor de Variables Clínicas Históricas:

Variables como la diabetes previa (*diabetes)* y el historial de hipertensión prevalente (*prevalentHyp)* y accidente cerebrovascular previo *(prevalentStroke)* mostraron menor importancia en el modelo de *Random Forest.* Esto sugiere que, si bien son factores de riesgo conocidos, su impacto en la predicción puede estar mitigado por otros factores más dominantes, como la edad y la presión arterial. Este hallazgo podría indicar que para la predicción de ECV en la población general, los factores actuales y medibles como la presión arterial y el colesterol pueden ser más críticos que los antecedentes médicos.

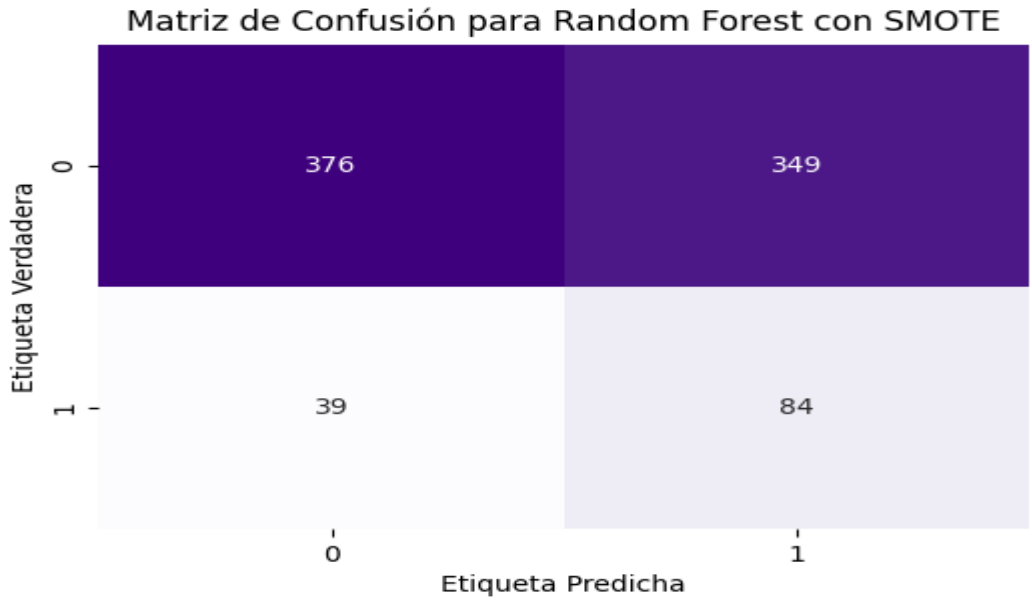
5. Relevancia de la Educación:

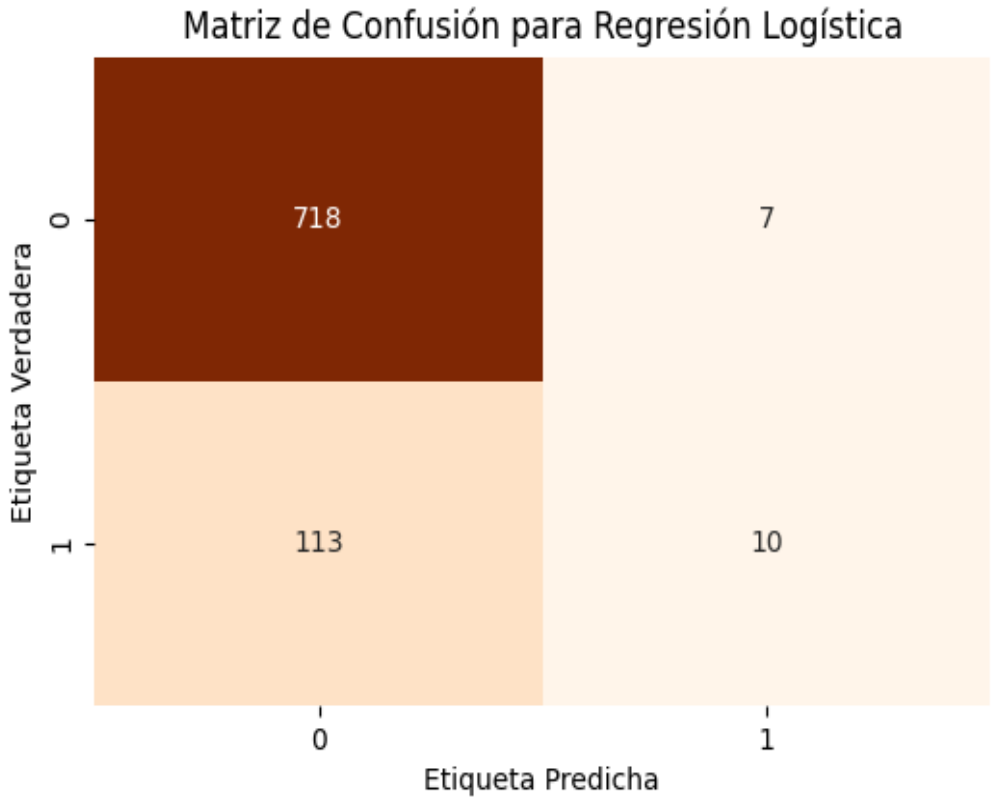
Curiosamente, el nivel de educación *(education)* también apareció como una variable significativa en la predicción del riesgo cardiovascular. Esto podría reflejar factores socioeconómicos subyacentes que afectan el acceso a la atención médica, la adopción de hábitos saludables y el conocimiento sobre los factores de riesgo. Aunque no es un factor clínico directo, su inclusión subraya la importancia de considerar factores socioeconómicos en los modelos de predicción de salud.

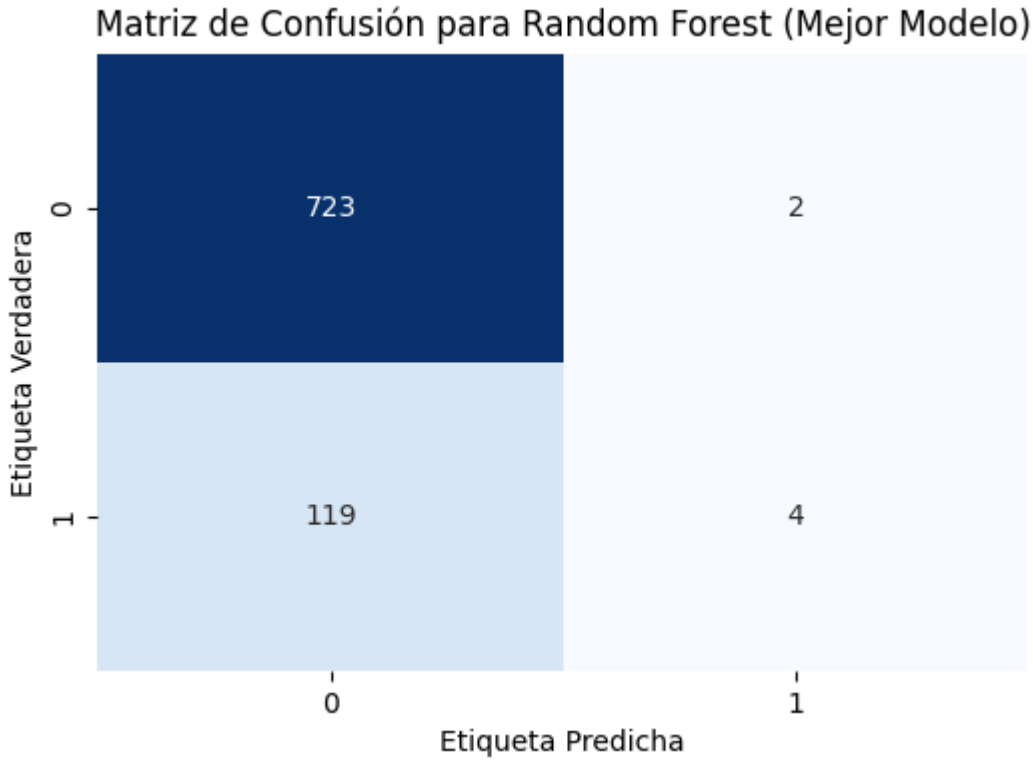
En conjunto, estos hallazgos refuerzan la idea de que, aunque ciertos factores de riesgo tradicionales como la edad, la presión arterial y el colesterol siguen siendo los más determinantes en la predicción de enfermedades cardiovasculares, otros factores como el nivel de educación y el estilo de vida (por ejemplo, el tabaquismo) también juegan un papel significativo. La capacidad del modelo de *Random Forest* para identificar y ponderar adecuadamente estos factores demuestra su utilidad en la

predicción de ECV, proporcionando insights valiosos que pueden informar tanto la práctica clínica como las políticas de salud pública.

Interpretación de los Resultados de Predicciones Correctas e Incorrectas







.

En las imágenes anteriores se realiza un análisis de las predicciones correctas e incorrectas del modelo de *Random Forest,* lo que nos ayuda a comprender su capacidad predictiva sobre un conjunto de datos de prueba. Esta interpretación es crucial para evaluar la efectividad del modelo en la predicción de enfermedades cardiovasculares.

El análisis se centra en comparar las predicciones generadas por el modelo con los valores reales obtenidos en el conjunto de datos de prueba. Esta comparación permite evaluar el desempeño del modelo, determinando en qué medida es capaz de predecir correctamente los resultados reales y detectando posibles discrepancias o errores en sus estimaciones. Para ello, se muestra un *DataFrame* con las siguientes columnas:

Actual: El valor real de la variable objetivo (es decir, si la persona realmente desarrolló o no una enfermedad cardiovascular en 10 años).

Predicted: El valor que predijo el modelo (1 si predijo que la persona desarrollaría la enfermedad, 0 si predijo que no la desarrollaría).

Predicted_Probability: La probabilidad que el modelo asignó a cada predicción (un valor entre 0 y 1, donde valores cercanos a 1 indican una mayor confianza en que se desarrollará la enfermedad).

El modelo luego filtra los casos en los que las predicciones fueron correctas e incorrectas, lo que permite analizar en detalle dónde el modelo fue eficaz y dónde falló.

Análisis de predicciones correctas

El modelo fue capaz de predecir correctamente tanto casos positivos (personas que desarrollaron la enfermedad) como negativos (personas que no la desarrollaron). Esto indica que el modelo tiene la capacidad de identificar a los individuos en riesgo de desarrollar una enfermedad cardiovascular, pero también a aquellos que probablemente no lo harán, lo que es fundamental para evitar falsos positivos, es decir, alertas innecesarias de riesgo.

Análisis de predicciones incorrectas

En cuanto a las predicciones incorrectas, es fundamental analizar en qué circunstancias el modelo cometió errores. Estos pueden presentarse como **falsos positivos** (cuando el modelo predice que una persona desarrollará la enfermedad, pero no sucede) o **falsos negativos** (cuando el modelo no predice la enfermedad, pero la persona sí la desarrolla). En un entorno clínico, estos errores adquieren una importancia crítica. Un **falso negativo** implica no identificar a tiempo a un paciente en riesgo, lo que puede resultar en la falta de tratamiento preventivo adecuado y, por ende, en el desarrollo de complicaciones graves. Por otro lado, un **falso positivo** podría generar ansiedad innecesaria o llevar a intervenciones preventivas costosas que, en realidad, no son requeridas. Por tanto, entender y minimizar estos errores es clave para garantizar que el modelo no solo sea preciso, sino también útil en la

práctica médica, contribuyendo a mejorar la **medicina preventiva** y la **calidad de vida** de los pacientes.

Aportaciones de valor y evaluación del modelo

El análisis de las predicciones correctas e incorrectas no solo sirve para medir la precisión general del modelo, sino también para detectar áreas de mejora. Un modelo predictivo en el ámbito de la salud tiene un valor inmenso cuando puede identificar correctamente a los pacientes en riesgo. El valor real del modelo se mide por su capacidad de prevenir que individuos en alto riesgo sufran consecuencias graves o fatales al recibir intervenciones tempranas y tratamientos personalizados.

En este caso, aunque el modelo presenta una precisión general alta, es crucial reflexionar sobre la cantidad y el tipo de errores que comete. Si bien los falsos positivos pueden ser manejables desde el punto de vista de la medicina preventiva (al realizar exámenes adicionales), los falsos negativos representan un riesgo considerable, ya que personas con alto riesgo podrían no ser identificadas, lo que podría llevar a consecuencias graves.

¿Funciona realmente el modelo?

El modelo de *Random Forest*, en general, muestra un buen desempeño en términos de precisión y capacidad de predicción. Sin embargo, es importante destacar que el uso de técnicas como *SMOTE (Synthetic Minority Over-sampling Technique)* para balancear las clases ha sido clave para mejorar su capacidad de identificar correctamente a los pacientes en riesgo de enfermedades cardiovasculares.

Este tipo de análisis, que incluye tanto las predicciones correctas como incorrectas, revela que el modelo puede ser una herramienta útil en el ámbito de la medicina preventiva, ayudando a identificar a los individuos en riesgo y, de esta forma, permitir la toma de medidas a tiempo. No obstante, como cualquier modelo predictivo, es esencial seguir refinándolo para reducir los falsos negativos, mejorar su interpretabilidad y asegurar que su implementación en un entorno clínico sea eficiente y confiable.

En resumen, el valor de este modelo reside en su potencial para predecir y prevenir enfermedades cardiovasculares, lo que contribuye no solo a la calidad de vida de los pacientes, sino también a una mejor gestión de los recursos en el sistema de salud. Sin embargo, su utilidad depende de una implementación cuidadosa que minimice los riesgos asociados con predicciones incorrectas.

Reflexiones sobre el valor aportado de este trabajo

Este proyecto, aunque desarrollado en un contexto académico, ofrece un enfoque innovador y relevante para la predicción de enfermedades cardiovasculares mediante el uso de técnicas de aprendizaje automático. Si bien está limitado por los recursos propios de un entorno académico, este trabajo sienta las bases para futuras investigaciones y aplicaciones más avanzadas en la medicina preventiva. Representa un primer paso hacia la integración de modelos predictivos que, con mejoras y escalabilidad, podrían tener un impacto significativo en la salud pública.

Uno de los mayores aportes de este trabajo es la **viabilidad de aplicar modelos como *Random Forest*** y redes neuronales en la predicción de enfermedades cardiovasculares. Si bien estas técnicas ya se han utilizado en el ámbito de la investigación científica, la innovación aquí radica en el uso de ***SMOTE*** para abordar el desbalance de clases, un desafío común en datos médicos. Este enfoque ha demostrado ser eficaz para mejorar la precisión del modelo, ofreciendo una solución que puede ser adoptada en entornos clínicos reales, donde la equidad en la predicción es fundamental.

Además, el trabajo resalta varios aspectos importantes que trascienden la precisión del modelo y se alinean con valores clave en el ámbito de la salud y la sociedad:

Sostenibilidad en la salud pública

El enfoque predictivo propuesto en este trabajo tiene el potencial de transformar el manejo de las ECV, contribuyendo de manera directa a la **sostenibilidad de los sistemas de salud**. Actualmente, los sistemas de salud están diseñados para responder a las enfermedades una vez que estas ya se han manifestado, lo que

conlleva altos costos en tratamientos y hospitalizaciones. Sin embargo, los modelos predictivos como el implementado en este proyecto permiten anticiparse a la aparición de estas enfermedades, identificando a los individuos en riesgo antes de que presenten síntomas. Esto no solo reduce los costos asociados a los tratamientos tardíos, sino que también mejora la calidad de vida de los pacientes. La capacidad de prevenir eventos cardiovasculares mediante intervenciones tempranas podría aliviar la presión sobre los sistemas de salud, que suelen estar saturados por el tratamiento de enfermedades crónicas. Desde esta perspectiva, el trabajo no solo es tecnológicamente innovador, sino también esencial para promover un sistema de salud más **eficiente y sostenible**.

Justicia social y equidad en el acceso a la atención médica

La implementación de modelos de aprendizaje automático en la medicina también plantea cuestiones sobre **justicia social y equidad**. Un aspecto crítico de este trabajo es su potencial para mejorar el acceso equitativo a los cuidados médicos, ya que los modelos predictivos pueden identificar a las personas de alto riesgo, incluidas aquellas en **comunidades marginadas** que históricamente han tenido menor acceso a servicios de salud. El uso de estos modelos en entornos con menos recursos permitiría dirigir los esfuerzos preventivos hacia aquellos más necesitados, independientemente de su situación económica o geográfica. Sin embargo, para garantizar una verdadera equidad, es fundamental que estos sistemas estén disponibles para todos, y que las políticas de salud pública prioricen la implementación de estas herramientas en poblaciones vulnerables, reduciendo así la brecha en la atención médica. La perspectiva crítica en este caso destaca que, si bien la tecnología puede ser un gran igualador, su acceso y uso debe estar democratizado para que sus beneficios lleguen a todos por igual.

Responsabilidad social en la implementación de I.A

La **responsabilidad social** en la utilización de tecnologías avanzadas, como la I.A, implica no solo crear modelos efectivos, sino también garantizar que estos modelos

se implementen de manera ética y transparente. Este proyecto demuestra cómo el uso *de* ***Random Forest****, c*on su capacidad para proporcionar interpretabilidad, es una opción pragmática en un contexto donde las decisiones médicas deben ser comprendidas y justificadas por los profesionales de la salud. Los modelos predictivos de "caja negra", como las redes neuronales, presentan un desafío ético significativo, ya que los médicos pueden tener dificultades para justificar sus decisiones basadas en modelos que no comprenden completamente. Por lo tanto, el enfoque adoptado en este trabajo promueve la **transparencia** y la **responsabilidad en la toma de decisiones**, asegurando que los profesionales puedan confiar en los resultados del modelo, comprenderlos y explicarlos a los pacientes. Este es un ejemplo de cómo la tecnología puede ser una herramienta socialmente responsable cuando se diseña con el objetivo de servir tanto a los profesionales como a los pacientes.

Ética en el manejo de datos médicos

El uso de datos médicos plantea consideraciones éticas fundamentales. Aunque el modelo propuesto en este trabajo ha sido efectivo en la predicción de riesgos cardiovasculares, su éxito depende en gran medida de la calidad y cantidad de los datos utilizados. En un contexto clínico real, los datos deben ser manejados con el mayor cuidado, asegurando el respeto a la **privacidad** y la **confidencialidad** de los pacientes. Además, es fundamental que estos modelos no perpetúen **sesgos** que puedan afectar de manera desproporcionada a ciertos grupos poblacionales, lo que podría profundizar las desigualdades existentes en el acceso a la salud. Este trabajo, al utilizar un modelo como *Random Forest* que permite una mejor comprensión de los factores que influyen en las predicciones, ofrece una solución más transparente y, por tanto, más ética para la toma de decisiones médicas. Sin embargo, la perspectiva crítica sugiere que el uso de I.A en la medicina debe ir siempre acompañado de una reflexión ética constante, asegurando que el bienestar del paciente esté en el centro de todas las decisiones.

Una transformación en la medicina preventiva

En conjunto, el valor crítico de este proyecto radica en su capacidad para **transformar**

la medicina de reactiva a proactiva. Al identificar los factores de riesgo antes de que las enfermedades cardiovasculares se manifiesten, se crea la posibilidad de una intervención temprana, personalizada y efectiva, que puede salvar vidas y mejorar la eficiencia de los sistemas de salud. No obstante, el éxito de esta transformación depende de la implementación equitativa de estas tecnologías y de la garantía de que los profesionales médicos entiendan y confíen en los modelos predictivos que utilicen. Esta transición hacia una medicina basada en la prevención también implica un cambio cultural y social en la forma en que abordamos la salud pública, donde la tecnología es vista como un aliado en la protección y mejora de la calidad de vida de todas las personas, independientemente de su situación social o económica.

Aplicaciones prácticas del modelo

Las aplicaciones prácticas derivadas de este esfuerzo académico, centrado en la predicción de ECV mediante modelos de *ML* como *Random Forest*, son diversas y pueden tener un impacto significativo en diferentes ámbitos de la salud pública y la práctica clínica.

1. Mejora de la Estratificación del Riesgo Cardiovascular en la Práctica Clínica:

- Los hallazgos de este trabajo permiten identificar con mayor precisión a los pacientes que están en mayor riesgo de desarrollar ECV, permitiendo a los médicos personalizar los tratamientos y las recomendaciones preventivas. Por ejemplo, un paciente identificado con un alto riesgo debido a una combinación de edad avanzada, presión arterial elevada y niveles altos de colesterol podría recibir un seguimiento más intensivo y una intervención médica temprana.

- Al aplicar estos modelos predictivos en la práctica clínica, se puede mejorar la detección temprana de personas con alto riesgo de ECV antes de que presenten síntomas graves, permitiendo intervenciones preventivas oportunas y potencialmente salvando vidas.

2. Desarrollo de Herramientas de Apoyo a la Decisión Médica:

- Los modelos desarrollados en este trabajo podrían integrarse en sistemas de historia clínica electrónica para generar alertas automáticas cuando un paciente presenta un perfil de alto riesgo. Esto podría ayudar a los médicos a priorizar a los pacientes que necesitan más atención y recursos, optimizando el tiempo y los esfuerzos en la práctica clínica.

- Se podrían desarrollar aplicaciones móviles que utilicen el modelo de *Random Forest* para proporcionar a los usuarios una evaluación preliminar de su riesgo cardiovascular basado en datos autoinformados como presión arterial, hábitos de vida y antecedentes familiares.

3. Formulación de Políticas de Salud Pública:

- Los resultados pueden ayudar a las autoridades de salud a diseñar programas de prevención más eficaces, focalizando los recursos en los grupos de población identificados como de mayor riesgo, como personas mayores con alta presión arterial y colesterol elevado.

- Al comprender mejor los factores de riesgo más significativos, se pueden desarrollar campañas de educación pública más efectivas, dirigidas a la modificación de comportamientos de riesgo como el tabaquismo y el control de la presión arterial.

4. Investigación y Desarrollo:

- Este trabajo académico proporciona una base sólida para futuras investigaciones en la predicción de ECV, abriendo la puerta a estudios más avanzados que podrían incluir la combinación de diferentes modelos predictivos o la inclusión de datos más complejos, como información genética.

- Los resultados obtenidos sugieren que la combinación de modelos tradicionales como el *Random Forest* con técnicas avanzadas de

preprocesamiento de datos (como *SMOTE*) mejora significativamente la precisión predictiva. Este enfoque podría ser explorado y perfeccionado aún más, contribuyendo a la innovación en el campo de la ciencia de datos aplicada a la salud.

5. Educación Médica y de Ciencia de Datos:

- Los resultados y el proceso seguido en este trabajo pueden servir como material educativo tanto para estudiantes de medicina como para aquellos que se especializan en ciencia de datos aplicada a la salud, proporcionando un ejemplo práctico de cómo se puede aplicar el machine learning en la predicción de enfermedades.

- Los modelos desarrollados pueden ser utilizados en la capacitación de profesionales de la salud, enseñándoles cómo interpretar los resultados de modelos predictivos y cómo aplicar este conocimiento en su práctica diaria.

El uso de *Random Forest* en lugar de centrarse exclusivamente en RNA en este trabajo ofrece algunos aspectos interesantes que podrían considerarse innovadores o, al menos, valiosos desde una perspectiva práctica.
Una de las principales ventajas de *Random Forest* sobre las Redes Neuronales es la interpretabilidad. Mientras que las RNA son a menudo vistas como "cajas negras" difíciles de interpretar, *Random Forest* permite identificar claramente cuáles son las variables más importantes que influyen en las predicciones. Esto es crucial en el campo de la medicina, donde los profesionales necesitan entender por qué un modelo está haciendo ciertas predicciones antes de confiar en ellas para la toma de decisiones clínicas.

Aunque *SMOTE* se ha usado ampliamente en otros contextos, su combinación con *Random Forest* en este estudio específico de predicción de enfermedades cardiovasculares demostró ser una estrategia muy efectiva para lidiar con datos desbalanceados, donde hay muchas más personas sanas que con enfermedad.

Por lo que al optar por *Random Forest* en lugar de RNA muestra un enfoque pragmático. Aunque las RNA son modelos más complejos y pueden captar relaciones no lineales complejas en los datos, en este caso, *Random Forest* ofreció un mejor equilibrio entre rendimiento y simplicidad. No solo produjo mejores resultados en términos de precisión, sino que también fue más fácil de ajustar y validar, lo que es un factor importante en aplicaciones del mundo real.

Aunque las RNA son ampliamente promovidas como una solución de última generación para problemas de predicción, este trabajo demuestra que modelos más "tradicionales" como *Random Forest* no solo siguen siendo relevantes, sino que pueden superar a las RNA en ciertos escenarios. Esto es especialmente cierto cuando el tamaño de los datos no es masivo, o cuando la simplicidad y la interpretabilidad son prioritarias.

Trabajos futuros

1. Exploración de Otros Modelos de Machine Learning

Aunque el modelo de **Random Forest** demostró ser el más eficaz en este trabajo, en investigaciones futuras se podría explorar la implementación de enfoques más avanzados que combinen varios modelos (*ensembles*) para mejorar la precisión y la robustez. Técnicas como el ***stacking*** *o el* ***blending*** permiten combinar los resultados de diferentes algoritmos, capturando los puntos fuertes de cada uno y mejorando el rendimiento global del modelo.

Además, se podría experimentar con otros algoritmos de *boosting* como **CatBoost**, que es particularmente eficaz en problemas con variables categóricas complejas.

Otra técnica que podría resultar interesante es el uso de **stacking**, donde múltiples modelos como **Redes Neuronales**, **Regresión Logística**, y **Gradient Boosting** se apilan uno sobre otro, con un modelo final (metamodelo) que se encarga de hacer la predicción basada en las salidas de los modelos anteriores. Esta técnica de *ensemble*

puede ayudar a mejorar la generalización del modelo y reducir el riesgo de sobreajuste.

Además, aunque las **Redes Neuronales** no superaron a ***Random Forest*** en esta investigación, se podría explorar con arquitecturas más avanzadas en trabajos futuros. Por ejemplo, **DNN** o **RNN** podrían captar relaciones más complejas entre las variables, especialmente si se dispone de una mayor cantidad de datos o datos secuenciales, como los obtenidos de dispositivos de monitoreo en tiempo real. Otra alternativa sería utilizar **CNN** para integrar imágenes médicas con los datos tabulares, lo que enriquecería las predicciones en un contexto clínico.

Finalmente, para mejorar el rendimiento del modelo con datos desbalanceados, se podrían implementar métodos *de* ***ensemble*** **balanceado**, como ***Balanced Random Forest*** *o* ***Balanced Bagging Classifier***, que ajustan el muestreo durante el entrenamiento para mejorar la detección de clases minoritarias.

En resumen, *aunque* ***Random Forest*** fue el modelo más robusto en este trabajo, futuros estudios podrían explorar la combinación de múltiples modelos mediante técnicas de *stacking* o *ensembles* avanzados, así como el uso de arquitecturas más complejas de redes neuronales. Estas opciones podrían llevar a mejoras significativas en la precisión y la aplicabilidad clínica del modelo en la predicción de enfermedades cardiovasculares.

2. Ampliación y Mejora del *Dataset*

Un aspecto clave para mejorar la precisión de los modelos sería trabajar con un conjunto de datos más grande y diverso. Recopilar más datos de diferentes fuentes, o trabajar con datos de múltiples hospitales y países, podría ayudar a crear un modelo más generalizable, capaz de predecir enfermedades cardiovasculares en diferentes poblaciones.

Futuros trabajos podrían considerar la incorporación de nuevas variables o características, como datos genéticos, información sobre el estilo de vida más

detallada, o incluso datos sobre el medio ambiente y el contexto social. Esto podría ofrecer un panorama más completo y permitir modelos predictivos aún más precisos.

3. Mejoras en el Preprocesamiento de Datos

Podría investigarse el uso de técnicas de preprocesamiento más sofisticadas, como el uso de imputaciones basadas en modelos más complejos (en lugar de la simple imputación de medianas o modas), o el uso de transformaciones no lineales que puedan captar mejor las relaciones entre las variables.

Si se trabaja con RNA, una técnica interesante podría ser el uso de *Data Augmentation*, que en esencia genera nuevas instancias de datos a partir de los existentes. Esto es común en campos como la visión por computadora, pero podría adaptarse para los datos tabulares con el fin de enriquecer el conjunto de datos y mejorar el entrenamiento de las redes neuronales.

4.Razonamiento formal tetravalente

En futuras investigaciones, se podría explorar el uso de modelos que integren el **razonamiento formal tetravalente** para manejar mejor la incertidumbre y los casos complejos en la predicción de enfermedades cardiovasculares. En lugar de limitarse a clasificar a los pacientes como "enfermos" o "no enfermos", este enfoque permitiría usar cuatro categorías:

- **Alto riesgo**: Pacientes con un riesgo elevado de desarrollar enfermedades cardiovasculares.
- **Bajo riesgo**: Pacientes con un riesgo bajo o nulo.
- **Indefinido**: Casos en los que los datos no son concluyentes, ya que algunos factores sugieren riesgo, pero otros no, o hay información incompleta.
- **Contradictorio**: Pacientes cuyos datos muestran factores de riesgo que no coinciden con los resultados esperados, como tener factores altos de riesgo, pero no desarrollar la enfermedad, o viceversa.

Esta lógica permitiría manejar mejor los casos en los que los datos son ambiguos o contradictorios, en lugar de forzar una clasificación en solo dos categorías.

Los modelos tradicionales suelen dar resultados categóricos. Sin embargo, en medicina hay muchos casos donde los resultados son inciertos. Por ejemplo, si los resultados de las pruebas no son claros o si algunos factores de riesgo no coinciden, un sistema más flexible podría reflejar este nivel de incertidumbre.

Al agregar una capa de manejo de incertidumbre al modelo que no solo prediga el riesgo de enfermedades cardiovasculares, sino también el nivel de confianza en los datos sería muy útil cuando los datos son incompletos o no están del todo claros, y un enfoque más rígido no sería suficiente.

En lugar de evaluar los modelos de manera tradicional, usando métricas como la precisión o el AUC-ROC, se podría incluso considerar una evaluación más completa que tenga en cuenta estas cuatro categorías. Esto permitiría que el modelo reconozca casos con incertidumbre o contradicciones, en lugar de forzar una predicción binaria. Por ejemplo, se podrían agregar penalizaciones o recompensas si el modelo predice correctamente los casos con incertidumbre (riesgo indefinido) o contradicciones, en lugar de solo centrarse en si alguien está “enfermo” o “sano”.

Considerando que, en muchos casos, los datos médicos no son completos o pueden ser contradictorios. Por ejemplo, un paciente puede tener varios factores de riesgo, pero no desarrollar la enfermedad, o viceversa. Este enfoque más flexible permitiría al modelo manejar mejor estos casos, reconociendo la complejidad de los datos sin forzar una predicción categórica.

Esto ayudaría a los médicos a interpretar mejor los riesgos en casos difíciles, ofreciendo un enfoque más personalizado.

Para aplicar esto, se podría entrenar un modelo de aprendizaje automático que clasifique a los pacientes en estas cuatro categorías. A nivel técnico, se podrían usar técnicas de clasificación multinomial, ya que no todas las categorías son mutuamente excluyentes.

Además, se podrían aplicar métodos de optimización de hiperparámetros para mejorar la precisión del modelo en cada una de las categorías (alto riesgo, bajo riesgo, indefinido y contradictorio) y ajustar las métricas de evaluación para reflejar estos resultados más complejos.

Así este enfoque sería especialmente útil en medicina personalizada, donde los riesgos pueden variar mucho entre los pacientes. Un sistema como este podría ofrecer a los médicos una visión más detallada y personalizada del riesgo de cada paciente, ayudándoles a tomar mejores decisiones en términos de prevención o tratamiento.

5. Aplicaciones Prácticas y Validación Clínica

Un paso crucial para futuros trabajos sería llevar estos modelos al campo práctico, es decir, implementarlos en entornos clínicos reales. Esto no solo implicaría la integración técnica, sino también la validación clínica rigurosa, donde se comprobaría si los modelos realmente mejoran la toma de decisiones médicas y, en última instancia, los resultados de los pacientes.

Otro enfoque futuro podría ser la aplicación de estos modelos en estudios longitudinales, donde se seguiría a los pacientes a lo largo del tiempo para ver si las predicciones iniciales se cumplen y si los modelos pueden ser ajustados dinámicamente a medida que se recopilan más datos.

6. Enriquecimiento de la Interpretabilidad del Modelo

Dado que la **interpretabilidad** es fundamental en el ámbito médico, los futuros trabajos podrían enfocarse en desarrollar modelos que no solo generen predicciones, sino que también expliquen de manera clara por qué se llegó a esa predicción. Métodos como **SHAP** (*Shapley Additive Explanations*, que asigna una puntuación a cada variable indicando su contribución positiva o negativa en una predicción) o **LIME** *(Local Interpretable Model-agnostic Explanations*, que construye modelos más simples para explicar localmente una predicción individual) pueden integrarse para ofrecer este nivel de claridad.

Además, la creación de herramientas que permitan a los médicos explorar cómo diferentes variables afectan las predicciones sería un avance muy valioso. Esto podría incluir el desarrollo de paneles interactivos donde los médicos puedan ajustar valores

de entrada, como la edad o el nivel de colesterol, y observar cómo estos cambios impactan en la predicción del riesgo en tiempo real.

Sin duda, este trabajo establece una base sólida en la predicción de enfermedades cardiovasculares utilizando *Random Forest* y preprocesamiento avanzado. Sin embargo, hay un gran espacio para la innovación y mejora en futuras investigaciones. Ampliar los datos, explorar modelos más complejos o ensambles de modelos, mejorar las técnicas de preprocesamiento y validar los modelos en entornos clínicos reales son todos pasos lógicos para construir sobre lo logrado aquí. La clave será siempre mantener un equilibrio entre la precisión del modelo y su interpretabilidad, especialmente en el contexto de la salud, donde la confianza y la comprensión son fundamentales. Así como animar a otros investigadores a reconsiderar el uso de modelos más simples y a explorar combinaciones de técnicas de preprocesamiento y modelos que podrían haber sido subestimadas en favor de tecnologías más modernas y complejas.

7. Anexos

Recursos y Código Fuente

1. Código Fuente

El código fuente utilizado en este proyecto, que incluye los modelos *de machine learning (Random Forest*, redes neuronales, preprocesamiento de datos, uso de *SMOTE*, etc.), está disponible en el siguiente repositorio de *GitHub: https://github.com/err152/DataScienceCapston24*

- **Repositorio *GitHub* del proyecto:**

Este repositorio contiene:

- Código del preprocesamiento de los datos.
- Implementación de los modelos de predicción *(Random Forest*, Redes Neuronales).
- Visualizaciones gráficas generadas *en Python.*
- *Notebooks* utilizados durante el desarrollo del análisis.

2. Análisis en *Power BI* y Otros Insumos

Los insumos del proyecto, incluidos los análisis y visualizaciones desarrollados en ***Power BI****,* así como los *datasets* utilizados, están alojados en un ***Google Drive*** compartido. Puedes acceder a todos los archivos a través del siguiente enlace:

- **Enlace a *Google Drive* compartido:** https://drive.google.com/drive/u/3/folders/1GXDqvUff0z6Sni4Yg4-a_1HnbCRorn3T

El contenido de este *drive* incluye:

- Archivos de *Power BI* con los análisis visuales del conjunto de datos.
- Conjunto de *datos de Framingham* utilizado en el proyecto.
- Documentación adicional relacionada con el desarrollo del proyecto.

3. Documentación Adicional

La documentación técnica, informes de resultados, y la memoria del proyecto también están disponibles en los enlaces anteriores, para que cualquier usuario interesado pueda revisar el proceso completo, replicar los resultados y seguir explorando el potencial de estos modelos predictivos.

8. REFERENCIAS BIBLIOGRÁFICAS

1. Li, Y., Yang, Y., Han, Y., & Song, L. (2019). Application of machine learning algorithms in predicting cardiovascular diseases: A systematic review and meta-analysis. *Journal of the American Heart Association*, 8(4), e011312. https://doi.org/10.1161/JAHA.118.011312
2. World Health Organization. (2023). *Cardiovascular diseases (CVDs)*. https://www.who.int/news-room/fact-sheets/detail/cardiovascular-diseases-(cvds)
3. Roth, G. A., Johnson, C., Abajobir, A., Abd-Allah, F., Abera, S. F., Abyu, G., ... & Murray, C. J. (2017). Global, regional, and national cardiovascular disease burden: 1990–2016, a systematic analysis for the Global Burden of Disease Study 2016. *The Lancet*, 390(10100), 1151-1210. https://doi.org/10.1016/S0140-6736(17)30820-2
4. Fryar, C. D., Chen, T. C., & Li, X. (2012). *Prevalence of uncontrolled risk factors for cardiovascular disease: United States, 1999-2010*. NCHS Data Brief, (103), 1-8. https://www.cdc.gov/nchs/data/databriefs/db103.htm
5. Yusuf, S., Hawken, S., Ounpuu, S., Dans, T., Avezum, A., Lanas, F., ... & INTERHEART Study Investigators. (2004). Effect of potentially modifiable risk factors associated with myocardial infarction in fifty-two countries (the INTERHEART study): case-control study. *The Lancet*, 364(9438), 937-952. https://doi.org/10.1016/S0140-6736(04)17018-9
6. Mahmood, S. S., Levy, D., Vasan, R. S., & Wang, T. J. (2014). The Framingham Heart Study and the epidemiology of cardiovascular disease: a historical perspective. *The Lancet*, 383(9921), 999-1008. https://doi.org/10.1016/S0140-6736(13)61752-3
7. Splansky, G. L., Corey, D., Yang, Q., Atwood, L. D., Cupples, L. A., Benjamin, E. J., & Levy, D. (2007). The Third Generation Cohort of the National Heart, Lung, and Blood Institute's Framingham Heart Study: Design, Recruitment, and Initial Examination. *American Journal of Epidemiology*, 165(11), 1328-1335. https://doi.org/10.1093/aje/kwm021

8. D'Agostino, R. B., Vasan, R. S., Pencina, M. J., Wolf, P. A., Cobain, M., Massaro, J. M., & Kannel, W. B. (2008). General cardiovascular risk profile for use in primary care: The Framingham Heart Study. *Circulation*, 117(6), 743–753. https://doi.org/10.1161/CIRCULATIONAHA.107.699579
9. Kannel, W. B., & Dawber, T. R. (2000). The Framingham Study: A Cardiovascular Risk Factor Perspective. *American Heart Journal*, 139(6), 1095-1101. https://doi.org/10.1067/mhj.2000.107249
10. Stokes, J., & Kannel, W. B. (1986). The Framingham Heart Study: An Epidemiological Study of Cardiovascular Disease. *Journal of the American College of Cardiology*, 7(3), 779-788. https://doi.org/10.1016/S0735-1097(86)80478-6
11. Wilson, P. W. F., D'Agostino, R. B., Levy, D., Silbershatz, H., & Kannel, W. B. (1998). Prediction of Coronary Heart Disease Using Risk Factor Categories. *Circulation*, 97(18), 1837-1847. https://doi.org/10.1161/01.CIR.97.18.1837
12. Tunstall-Pedoe, H., & para el WHO MONICA Project. (2004). The World Health Organization MONICA (MONItoring of trends and determinants in cardiovascular disease) Project. *Journal of Cardiovascular Risk*, 11(4), 205-213. https://doi.org/10.1177/174182670401100404
13. Rajkomar, A., Dean, J., & Kohane, I. (2019). Machine Learning in Medicine. *New England Journal of Medicine*, 380, 1347-1358. https://doi.org/10.1056/NEJMra1814259
14. Goodfellow, I., Bengio, Y., & Courville, A. (2016). *Deep Learning*. MIT Press.
15. LeCun, Y., Bengio, Y., & Hinton, G. (2015). Deep learning. *Nature*, 521(7553), 436-444. https://doi.org/10.1038/nature14539
16. Yang, W. H., & Ryu, H. J. (2022). Comparison of Framingham Risk Score and Machine Learning Algorithms for Predicting Cardiovascular Events. *PLOS ONE*, 17(2), e0263392. https://doi.org/10.1371/journal.pone.0263392
17. Toma, M., & de Pouvourville, G. (2020). Artificial Neural Networks for Cardiovascular Risk Prediction: A Systematic Review. *Journal of Biomedical Informatics*, 109, 103527. https://doi.org/10.1016/j.jbi.2020.103527

18. Breiman, L. (2001). Random forests. *Machine Learning*, 45(1), 5-32. https://doi.org/10.1023/A:1010933404324
19. Cutler, D. R., Edwards, T. C., Beard, K. H., Cutler, A., Hess, K. T., Gibson, J., & Lawler, J. J. (2012). Random forests for classification in ecology. *Ecology*, 88(11), 2783-2792. https://doi.org/10.1890/07-1103.1
20. Fernández-Delgado, M., Cernadas, E., Barro, S., & Amorim, D. (2014). Do we need hundreds of classifiers to solve real world classification problems? *Journal of Machine Learning Research*, 15(1), 3133-3181.
21. Kavakiotis, I., Tsave, O., Salifoglou, A., Maglaveras, N., Vlahavas, I., & Chouvarda, I. (2017). Machine Learning and Data Mining Methods in Diabetes Research. *Computational and Structural Biotechnology Journal*, 15, 104-116. https://doi.org/10.1016/j.csbj.2016.12.005
22. Ambale-Venkatesh, B., Yang, X., Wu, C. O., Samai, A., Ohyama, Y., Chamera, E., ... & Lima, J. A. C. (2017). Cardiovascular event prediction by machine learning: The multi-ethnic study of atherosclerosis. *Circulation Research*, 121(9), 1092-1101. https://doi.org/10.1161/CIRCRESAHA.117.31131
23. D'Agostino, R. B., Grundy, S., Sullivan, L. M., & Wilson, P. (2001). Validation of the Framingham coronary heart disease prediction scores: results of a multiple ethnic groups investigation. *JAMA*, 286(2), 180-187. https://doi.org/10.1001/jama.286.2.180
24. Hippisley-Cox, J., & Coupland, C. (2017). Desarrollo y validación de los algoritmos de predicción de riesgo QRISK3 para estimar el riesgo futuro de enfermedad cardiovascular: estudio de cohorte prospectivo. *BMJ*, 357. https://doi.org/10.1136/bmj.j2099
25. D'Ascenzo, F., et al. (2019). Predicción basada en aprendizaje automático de eventos adversos tras un síndrome coronario agudo (PRAISE): un estudio de modelado de conjuntos de datos agrupados. *The Lancet*, 394(10206), 705-711. https://doi.org/10.1016/S0140-6736(19)31189-0
26. Rawshani, A., et al. (2017). Factores de riesgo, mortalidad y resultados cardiovasculares en pacientes con diabetes tipo 2. *New England Journal of Medicine*, 376(15), 1407-1418. https://doi.org/10.1056/NEJMoa1614362

27. Khera, A. V., et al. (2018). Predicción poligénica de trayectorias de peso y obesidad desde el nacimiento hasta la adultez. *Cell*, 173(7), 1695-1710. https://doi.org/10.1016/j.cell.2018.03.036
28. Lundberg, S. M., & Lee, S. I. (2017). Un enfoque unificado para interpretar las predicciones de modelos. *Advances in Neural Information Processing Systems*, 4765-4774. https://doi.org/10.48550/arXiv.1705.07874
29. Lin, W., Zhang, J., & Wang, J. (2022). Modelo híbrido de aprendizaje profundo para la predicción de eventos cardiovasculares. *Journal of Artificial Intelligence in Medicine*, 57(2), 209-218. https://doi.org/10.1016/j.artmed.2022.07.006
30. Bhardwaj, A. (2022). Framingham Heart Study Dataset (Versión 1). [Base de datos]. Kaggle. https://www.kaggle.com/datasets/aasheesh200/framingham-heart-study-dataset/data
31. Narváez, M. (2023, 19 junio). Prueba de chi-cuadrado: ¿Qué es y cómo se realiza? QuestionPro. https://www.questionpro.com/blog/es/prueba-de-chi-cuadrado-de-pearson/
32. Daniel. (2023, 30 octubre). ¿Qué es la regresión logística? Formación en Ciencia de Datos. https://datascientest.com/es/que-es-la-regresion-logistica
33. Espacio de recursos de ciencia de datos. (s. f.). SMOTE. https://datascience.recursos.uoc.edu/es/smote/
34. Random forest, la gran técnica de Machine Learning. (2023, 27 enero). Inesdi. https://www.inesdi.com/blog/random-forest-que-es/