Modelo predictivo en el desarrollo de enfermedades cardiovasculares a partir de factores de riesgo

Alejandro Valderrama Cardenas

Master Universitario en Ciencia De Datos

Trabajo Final de Máster
Área 1

Tutor/a de TF
Francesc Julbe López
Profesor/a responsable de la asignatura
Albert Solé Ribalta

Junio 2024

Universitat Oberta de Catalunya

Ficha del Trabajo Final

Título del trabajo:	Modelo predictivo en el desarrollo de enfermedades cardiovasculares a partir de factores de riesgo
Nombre del autor/a:	Alejandro Valderrama Cardenas
Nombre del Tutor/a de TF:	Francesc Julbe López
Nombre del/de la PRA:	Albert Solé Ribalta
Fecha de entrega:	06/2024
Titulación o programa:	Máster Universitario en Ciencia de Datos
Área del Trabajo Final:	M2.878 TFM Análisis predictivo
Idioma del trabajo:	Castellano
Palabras clave	enfermedades cardiovasculares, factores de riesgo, aprendizaje automático

Resumen

Las enfermedades cardiovasculares son la principal causa de muerte a nivel mundial. Dada su importancia en la salud pública, se hace necesario comprender los factores de riesgo que contribuyen al desarrollo de estas enfermedades y cómo prevenirlas y tratarlas de manera efectiva. Este trabajo tuvo como objetivo desarrollar un modelo predictivo de enfermedades cardiovasculares utilizando técnicas de aprendizaje automático y un conjunto de datos de factores de riesgo recopilados por los Centros para el Control y la Prevención de Enfermedades (CDC) de Estados Unidos. Asimismo, se buscaba identificar los factores de riesgo más significativos asociados a estas enfermedades.

Se aplicaron diversas técnicas de muestreo para abordar el desequilibrio de clases y se entrenaron múltiples algoritmos. El modelo Light GBoost con SMOTEENN, entrenado en un conjunto de datos con 29 variables y 353,968 registros, demostró el mejor rendimiento, con un recall de 0.8450 y un AUC-ROC de 0.8139. Se identificaron como factores de riesgo clave la edad avanzada, el sexo masculino, los bajos niveles socioeconómicos (ingresos y educación) y dormir poco. Los resultados destacan la importancia de considerar factores socioeconómicos y de estilo de vida, además de los factores tradicionales, en la evaluación del riesgo cardiovascular. Si bien se reconocen limitaciones como la naturaleza autorreportada de los datos y la ausencia de algunas variables clínicas, este trabajo contribuye al desarrollo de herramientas de predicción más precisas, y enfatiza la necesidad de abordar los determinantes sociales en las políticas de prevención de enfermedades cardiovasculares.

Abstract

Cardiovascular diseases are the leading cause of death worldwide. Given their importance in public health, it is necessary to understand the risk factors that contribute to the development of these diseases and how to prevent and treat them effectively. This work aimed to develop a predictive model for cardiovascular diseases using machine learning techniques and a dataset of risk factors collected by the Centers for Disease Control and Prevention (CDC) in the United States. Furthermore, the goal was to identify the most significant risk factors associated with these diseases.

Various sampling techniques were applied to address class imbalance, and multiple machine learning algorithms were trained. The Light GBoost model with SMOTEENN, trained on a dataset with 29 variables and 353,968 records, demonstrated the best performance, with a recall of 0.8450 and an AUC-ROC of 0.8139. Key risk factors identified included advanced age, male gender, low socioeconomic status (encompassing income and educational attainment), and insufficient sleep duration. The findings underscore the importance of considering socioeconomic and lifestyle factors, in addition to traditional factors, in the assessment of cardiovascular risk. Although limitations such as the self-reported nature of the data and the absence of some clinical variables are acknowledged, this work contributes to the development of more accurate prediction tools and emphasizes the need to address social determinants in cardiovascular disease prevention policies.

Índice

Lista de Figuras

Lista de Tablas

1. Introducción

Este apartado establece el contexto y justificación del tema a tratar, así como los objetivos que se buscan alcanzar en la realización de este trabajo. También se detalla la metodología y la planificación que se tiene previsto seguir, junto con la estructura de los capítulos del documento.

1.1. Contexto y justificación

Las enfermedades cardiovasculares son una de las principales causas de morbilidad y mortalidad a nivel mundial. Según la Organización Mundial de la Salud, son responsables de aproximadamente 17,9 millones de muertes al año, lo que equivale al 31% de todas las muertes a nivel mundial [1]. A medida que la población envejece y aumenta la prevalencia de factores de riesgo como la obesidad, el tabaquismo, la mala alimentación y la falta de actividad física, se espera que esta situación siga creciendo en los próximos años. Dada la importancia en la salud pública, se hace necesario comprender los factores de riesgo que contribuyen al desarrollo de estas enfermedades y como se pueden prevenir y tratar de manera efectiva.

Una de las razones para llevar a cabo este análisis es la diversidad de factores de riesgo involucrados en las enfermedades cardiovasculares. Estos factores incluyen elementos genéticos, demográficos, clínicos y de estilo de vida, así como la exposición a la contaminación del aire y el tabaquismo. Dada esta complejidad, es necesario utilizar enfoques de análisis de datos avanzados para identificar patrones y relaciones subyacentes entre estos factores y el riesgo de desarrollar una enfermedad cardiovascular.

Estas enfermedades afectan a diversos grupos demográficos de manera diferente, y comprender estas diferencias puede ayudar a que las intervenciones preventivas y terapéuticas sean más efectivas para todas las personas. Analizar como varían los factores de riesgo entre diferentes grupos y como estas diferencias afectan la efectividad de las intervenciones puede ayudar a diferenciar los distintos factores y su aplicación de acuerdo al grupo demográfico.

Las comorbilidades, como la obesidad, también pueden ser condiciones que estén relacionadas con las enfermedades cardiovasculares y pueden exacerbar los factores de riesgo y complicar el tratamiento de las personas.

De manera que la justificación de este trabajo se basa en la necesidad de mejorar la comprensión de los factores de riesgo de las enfermedades cardiovasculares y ayudar a desarrollar estrategias efectivas para su prevención y tratamiento.

1.2. Objetivos

Este trabajo tiene como objetivo principal abordar el problema de identificar si una persona puede desarrollar una enfermedad cardiovascular utilizando técnicas de aprendizaje automático. Además, se pretende explorar qué y en qué medida factores de riesgo relacionados con el estilo de vida pueden afectar el riesgo de desarrollar una enfermedad cardiovascular.

Se plantean múltiples objetivos, los cuales se pasan a enumerar a continuación:

1. Desarrollar un modelo predictivo utilizando técnicas de aprendizaje automático para predecir el riesgo de enfermedad cardiovascular.
2. Identificar los principales factores de riesgo asociados con enfermedades cardiovasculares.
3. Investigar como influyen los factores demográficos, como la edad y el género, en el riesgo de desarrollar una de estas enfermedades.
4. Analizar la contribución del estilo de vida, como la actividad física, como uno de los factores determinantes en el desarrollo de una enfermedad cardiovascular.
5. Identificar patrones y relaciones entre las variables de riesgo cardiovascular que permitan una mayor explicabilidad del modelo creado.

1.3. Impacto en sostenibilidad, ético-social y de diversidad

Dimensión sostenibilidad

El coste de la atención y tratamiento de las enfermedades cardiovasculares para los sistemas de salud de los países supone una gran carga económica. Desarrollar estrategias que permitan una mejor prevención, atendiendo a aquellos factores que contribuyen en mayor medida en el desarrollo de una enfermedad cardiovascular, puede propiciar una gestión optima y sostenible de los recursos. En este sentido, el presente trabajo puede tener un impacto positivo en los ODS 8 y ODS 10, los cuales hacen referencia al crecimiento económico y a reducir las desigualdades entre los países, respectivamente.

Por tanto, reducir el coste en este grupo de enfermedades, partiendo de una atención temprana a aquellos factores desencadenantes, permite destinar esos recursos económicos hacia otras áreas de salud pública o en el avance de la consecución de otros objetivos de desarrollo sostenible. Por ejemplo, poner fin a la pobreza (ODS 1), combatir el hambre (ODS 2) o garantizar una educación de calidad (ODS 4).

Dimensión comportamiento ético y de responsabilidad social (RS)

Respecto a esta dimensión, el impacto puede ser positivo y tiene relación con la dimensión de sostenibilidad. Si disminuye el coste para los sistemas de salud, esos recursos pueden trasladarse hacia otros ODS más urgentes o prioritarios como reducir la pobreza (ODS 1), reducir el hambre (ODS 2) o apoyar a otros países para reducir estas desigualdades entre ellos (ODS 10).

Dimensión diversidad, género y derechos humanos

El presente trabajo toma en cuenta estos aspectos y es uno de los objetivos planteados al determinar las diferencias entre distintos grupos demográficos, por lo que atiende a la diversidad, al identificar la raza de la persona y al género (ODS 5). En este sentido, supone un impacto positivo en esta dimensión.

Por último, el objetivo de desarrollo sostenible de salud (ODS 3) se puede considerar transversal a las 3 dimensiones mencionadas. Es en este aspecto donde el presente trabajo puede tener un mayor impacto positivo. Al contribuir a una prevención y atención temprana de las enfermedades cardiovasculares, identificando aquellos factores más determinantes en el desarrollo de estas afecciones. En la medida en que se consiga esto se reducirá el gasto sanitario de los sistemas públicos de salud. Así, este ahorro de recursos puede ayudar a lograr otros objetivos de desarrollo sostenible como reducir el cambio climático (ODS 13) o garantizar energía limpia y asequible (ODS 7). Igualmente, en aquellos países donde no existe esta cobertura sanitaria puede suponer un alto coste o endeudamiento para una familia poder tratar este tipo de enfermedades, lo cual puede desencadenar en una peor calidad de vida o sufrir riesgo de pobreza.

1.4. Enfoque y método

Para abordar este trabajo se utiliza un conjunto de datos que contiene información sobre factores de riesgo para enfermedades cardiovasculares, incluidos datos demográficos, clínicos y de estilo de vida, así como información sobre la prevalencia de enfermedades cardiovasculares en diferentes grupos de población. A partir de este conjunto de datos se van a aplicar técnicas de aprendizaje automático para identificar patrones y relaciones entre las variables.

Para llevar a cabo lo anterior y conseguir los objetivos planteados, se propone la siguiente metodología:

- Recolección de datos: Se recopilaron datos de personas de diversos grupos demográficos, clínicos y de estilo de vida relacionados con enfermedades cardiovasculares.
- Preprocesamiento de datos: Los datos recopilados se someterán a un proceso de limpieza y preprocesamiento. Esto incluye la eliminación de datos faltantes y la corrección de errores, así como la identificación y tratamiento de valores atípicos o extremos.
- Análisis exploratorio de datos: En esta etapa se busca identificar patrones, tendencias y relaciones preliminares entre las variables de riesgo cardiovascular. Esto ayudará a generar hipótesis y a dirigir el enfoque del análisis en etapas posteriores.
- Modelado: Se aplicarán técnicas de aprendizaje automático para descubrir relaciones entre las variables del conjunto de datos.
- Evaluación: Una vez se identifiquen las variables o factores más importantes, se evaluará el modelo para comprobar su rendimiento.

El enfoque y metodología propuesto se puede englobar como una metodología CRISP-DM (Cross Industry Standard Process for Data Mining) para aprovechar su naturaleza cíclica o iterativa entre las fases cuando sea necesario.

1.5. Planificación

El principal recurso necesario para la realización del trabajo es el conjunto de datos que contiene la información de las personas. Además de los recursos informáticos indispensables para realizar el análisis de los datos.

Las tareas que se van a realizar son las siguientes:

- Definición del tema: Concretar la línea de trabajo que se va a llevar a cabo.
- Planificación del trabajo: Listar las tareas a realizar durante el proyecto y estimar el tiempo necesario para llevar a cabo cada una.
- **Redacción PEC1**: Redactar la primera entrega del TFM.
- Búsqueda de bibliografía: Buscar información relacionada y relevante de trabajos anteriores que hayan abordado el tema de factores de riesgo y enfermedades cardiovasculares desde una perspectiva de la ciencia de datos.
- Análisis de las tecnologías y el código disponible: Analizar las herramientas o librerías disponibles para la implementación de los algoritmos. Buscar librerías o código ya existente relacionado con el tema.
- Lectura de bibliografía: Leer y analizar la bibliografía y documentación encontrada.
- Revisar la definición del trabajo: Valorar si es necesario cambiar el enfoque del trabajo final, tras la búsqueda de información, en aquellos aspectos no contemplados en un inicio.
- **Redacción PEC2**: Redactar la segunda entrega del TFM.
- Comprensión del conjunto de datos: Analizar el conjunto de datos para identificar la calidad y tipos de variables de los mismos.
- Preparación de los datos: Tratar los datos para el estudio posterior y garantizar la calidad de los resultados.
- Diseño del algoritmo: Diseñar los modelos y algoritmos que se van a utilizar en el trabajo.
- Realización de pruebas: Evaluar los modelos a partir de varias métricas para medir su rendimiento.
- Implementación del algoritmo: Implementar el modelo que ofrece un mejor rendimiento con base en los resultados de las métricas obtenidas.
- **Documentación PEC3**: Documentar el diseño y la implementación realizada.
- Conclusiones: Identificar las posibles conclusiones obtenidas después de realizar el trabajo.
- **Redacción de la memoria PEC4**: Redactar la memoria final de TFM.
- **Presentación y defensa del proyecto**: Preparar la defensa y presentación del trabajo final.
- **Defensa pública**: Preparar la defensa pública que se llevará a cabo tras la presentación del trabajo final.

En la figura 1.1 se detalla la planificación temporal de las tareas antes señaladas.

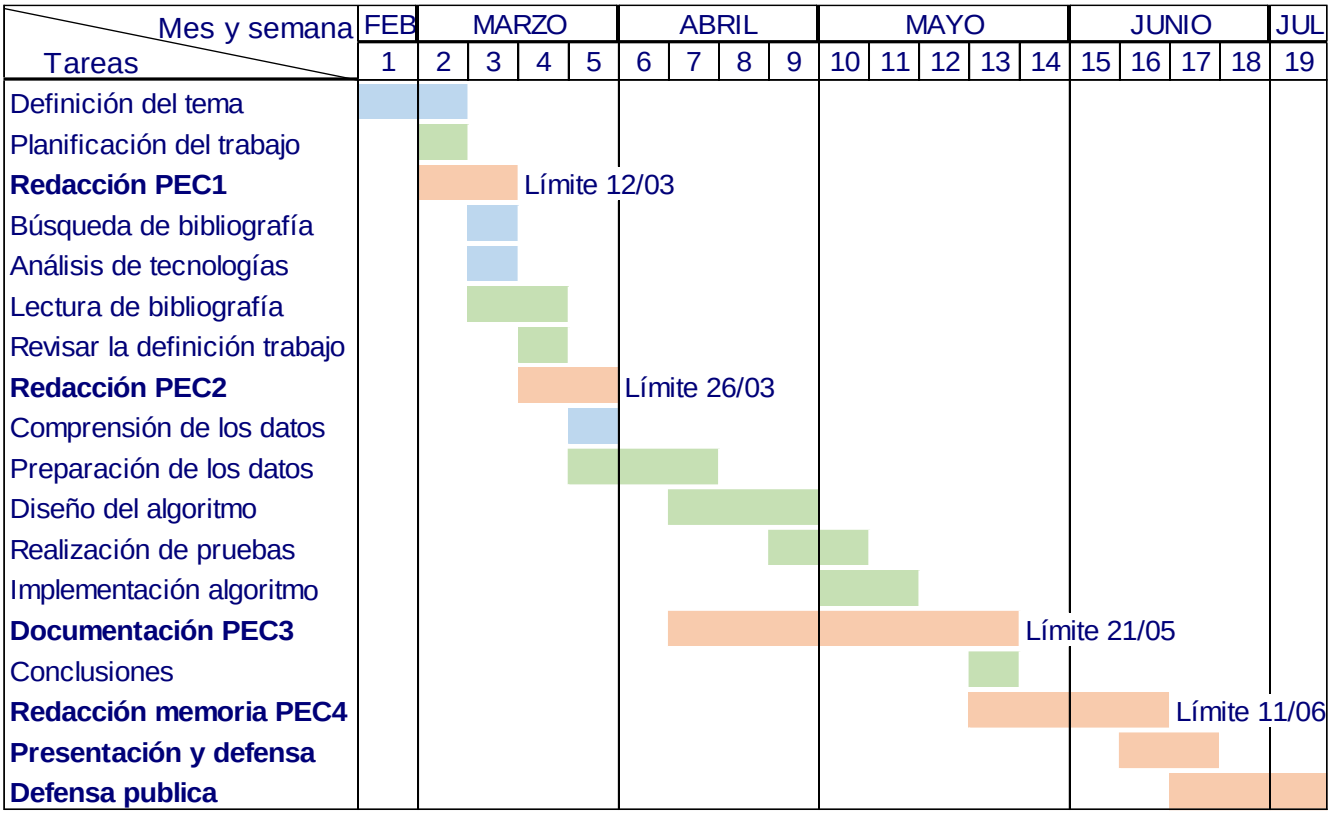


Figura 1.1. Planificación temporal de las tareas.

1.6. Breve sumario de productos obtenidos

Como productos de este trabajo se espera obtener un repositorio que contenga todo el código generado en el proceso de este proyecto. Así como la presente memoria que documenta todo el trabajo realizado.

1.7. Breve descripción de los capítulos de la memoria

Esta memoria está dividida en cinco capítulos, los cuales se pasan a detallar de manera breve a continuación:

- **Capítulo 1**: Se expone una introducción del proyecto junto con los objetivos y la planificación que sirve de guía para el desarrollo y consecución de cada etapa del trabajo.
- **Capítulo 2**: Se resumen los últimos trabajos o investigaciones, así como el estado actual, de los temas que se van a tratar.
- **Capítulo 3**: Se especifica paso por paso el proceso llevado a cabo para la implementación del proyecto. Además de definir y explicar los modelos utilizados.
- **Capítulo 4**: Se presentan los resultados obtenidos del proyecto.
- **Capítulo 5**: Se exponen las conclusiones obtenidas tras la realización del trabajo.
- **Glosario**: Se incluye un listado de términos técnicos y específicos utilizados en el documento.
- **Bibliografía**: Listado de referencias bibliográficas utilizadas durante la realización de este trabajo.
- **Anexo I**: Presenta el anexo de la memoria.

2. Estado del arte

2.1. Introducción

Las enfermedades cardiovasculares representan una carga significativa para la salud pública a nivel mundial, siendo responsables de una proporción sustancial de la morbilidad y mortalidad a nivel global. Desde enfermedades coronarias hasta accidentes cerebrovasculares, estas afecciones afectan no solo la calidad de vida de las personas, sino también los sistemas de salud y la economía en general. La comprensión de los factores de riesgo, la predicción precisa y la prevención efectiva de las enfermedades cardiovasculares resultan necesarios para abordar este importante problema de salud.

La interacción entre factores biológicos, ambientales y sociales desempeña un papel fundamental en el desarrollo y la progresión de las enfermedades cardiovasculares. La influencia de determinantes sociales, como el estatus socioeconómico, en los resultados cardiovasculares ha sido objeto de estudio en numerosas investigaciones [2]. Estos estudios han destacado la importancia de abordar desde distintas perspectivas los factores de riesgo para mejorar la prevención y el manejo de estas enfermedades.

El objetivo de este estado del arte es analizar de manera integral los avances en el análisis predictivo de enfermedades cardiovasculares, centrándose en la identificación de factores de riesgo, los modelos utilizados en la predicción de eventos cardiovasculares y los conjuntos de datos utilizados en los modelos. A través de la revisión de la literatura científica más relevante, se busca proporcionar una visión detallada de las tendencias actuales en la investigación cardiovascular mediante el uso de herramientas de aprendizaje automático y su impacto en la práctica clínica y la salud pública.

La disponibilidad de grandes conjuntos de datos de salud y los avances en tecnologías de análisis de datos han abierto nuevas oportunidades para el desarrollo de modelos predictivos más precisos y sofisticados [3]. Estos avances tienen el potencial de transformar la prevención y el manejo de las enfermedades cardiovasculares, mejorando los resultados clínicos y reduciendo la carga de enfermedad asociada.

El aprendizaje automático permite analizar grandes volúmenes de datos clínicos, identificar patrones complejos y predecir el riesgo individual de eventos cardiovasculares con mayor precisión. Los modelos predictivos basados en inteligencia artificial han demostrado su eficacia en la estratificación del riesgo cardiovascular y en la identificación temprana de posibles complicaciones [4].

Además de los factores biológicos y clínicos tradicionales, los aspectos sociales juegan un papel importante en la aparición y progresión de enfermedades cardiovasculares [5]. El acceso a la atención médica, el entorno socioeconómico, la educación y otros aspectos sociales pueden influir significativamente en la salud cardiovascular de las personas. Por lo tanto, es necesario considerar estos factores al diseñar estrategias de prevención y tratamiento de las enfermedades cardiovasculares.

La implementación de herramientas tecnológicas, como sistemas de información clínica y registros electrónicos de salud, ha facilitado la recopilación y el análisis de datos relevantes para la predicción

de enfermedades cardiovasculares [6]. La capacidad de integrar múltiples fuentes de datos, incluidos datos genéticos, clínicos y de estilo de vida, permite una evaluación más completa del riesgo cardiovascular y una atención más personalizada para los pacientes.

A medida que la investigación en enfermedades cardiovasculares avanza, es fundamental seguir explorando nuevas tecnologías y enfoques para mejorar la predicción y prevención de eventos cardiovasculares [7]. La colaboración interdisciplinaria entre cardiólogos, científicos de datos, epidemiólogos y expertos en salud pública puede permitir desarrollar estrategias innovadoras y basadas en evidencia para abordar la carga global de las enfermedades cardiovasculares.

2.2. Revisión de literatura

2.2.1 Enfermedades cardiovasculares

Las enfermedades cardiovasculares (ECV) constituyen un grupo de trastornos que afectan el corazón y los vasos sanguíneos. Estas enfermedades pueden manifestarse de diversas formas y afectar diferentes componentes del sistema cardiovascular, incluyendo el corazón, las arterias, las venas y los vasos capilares [8].

Tipos de Enfermedades Cardiovasculares [9-10]:

Enfermedad Coronaria: También conocida como enfermedad arterial coronaria, es una afección en la que se acumula placa (depósitos de grasa, colesterol, calcio y otras sustancias) en las arterias coronarias que suministran sangre al corazón. La angina de pecho y el infarto de miocardio son manifestaciones comunes de esta enfermedad.

Hipertensión Arterial: Se caracteriza por una presión arterial elevada en las arterias, lo que puede aumentar el riesgo de enfermedades cardiovasculares como accidentes cerebrovasculares (ictus) y enfermedad coronaria. En 2008, la enfermedad de las arterias coronarias y el ictus representaron el 80 % de las muertes por ECV en los hombres y el 75 % de las muertes por ECV en las mujeres en el mundo [11].

Enfermedad Cerebrovascular: Incluye accidentes cerebrovasculares (ACV) y ataques isquémicos transitorios (AIT), que son causados por la interrupción del flujo sanguíneo al cerebro o por sangrado de un vaso sanguíneo en el cerebro. Comúnmente conocidos como ictus o derrames cerebrales.

Insuficiencia Cardíaca: Es una condición en la que el corazón no puede bombear suficiente sangre para satisfacer las necesidades del cuerpo, lo que puede provocar síntomas como fatiga, dificultad para respirar y acumulación de líquido.

Arritmias Cardíacas: Son trastornos del ritmo cardíaco que pueden causar latidos cardíacos irregulares, demasiado rápidos o demasiado lentos, lo que puede afectar la función cardíaca.

Enfermedades Valvulares: Incluyen afecciones en las válvulas cardíacas que pueden provocar estenosis (estrechamiento) o insuficiencia (fuga) de las válvulas, lo que afecta el flujo sanguíneo dentro del corazón.

2.2.2 Factores de riesgo

Hoy en día, un factor de riesgo se describe como un elemento o característica medible que tiene una relación causal con el incremento de la incidencia de una enfermedad, actuando como factor predictivo independiente y significativo del riesgo de desarrollar una enfermedad [12].

En esta sección, se detallan los factores de riesgo comúnmente asociados con las enfermedades cardiovasculares, aquellos utilizados en los diversos modelos predictivos y aquellos que, podría ser importante tener en cuenta en la evaluación del riesgo cardiovascular.

Entre los factores de riesgos generalmente asociados a ECV e incluidos en modelos predictivos se encuentran los siguientes.

Edad: El envejecimiento es un factor de riesgo no modificable que se incluye en muchos modelos de predicción de riesgo cardiovascular, ya que el riesgo aumenta con la edad debido a cambios fisiológicos y acumulación de factores de riesgo [13-16].

Género: El género o sexo de una persona puede influir en el riesgo de ECV, puesto que existen diferencias biológicas y factores de riesgo específicos para hombres y mujeres [13-17].

Hipercolesterolemia: Niveles altos de colesterol en sangre, especialmente de lipoproteínas de baja densidad (LDL), están asociados con un mayor riesgo de estas afecciones debido a la formación de placas en las arterias [13-16, 18].

Tabaquismo: Fumar cigarrillos es un factor de riesgo modificable que contribuye al desarrollo de ECV al dañar los vasos sanguíneos y aumentar la formación de coágulos [3, 13-16].

Diabetes Mellitus: La diabetes, especialmente la diabetes tipo 2, aumenta el riesgo de enfermedades cardiacas debido a sus efectos en los vasos sanguíneos y el corazón [13-15, 19].

Obesidad: El exceso de peso corporal, en particular la obesidad abdominal, se asocia con un mayor riesgo de ECV debido a la inflamación crónica y otros mecanismos fisiopatológicos. Generalmente medida a partir del Índice de Masa Corporal (IMC) en los modelos de predicción [14-15, 19-20].

Historial Familiar de ECV: La presencia de antecedentes familiares de estas patologías puede aumentar el riesgo de padecerlas, lo que sugiere una predisposición genética [14, 16, 20].

Sedentarismo: La falta de actividad física regular se asocia con un mayor riesgo de enfermedades cardiovasculares, ya que el ejercicio regular es fundamental para mantener la salud cardiaca [18, 21].

Consumo de Alcohol: El consumo excesivo de alcohol puede contribuir al desarrollo de estas afecciones, especialmente en casos de consumo nocivo [22-23].

Niveles de Triglicéridos: Los niveles elevados de triglicéridos en sangre pueden ser un factor de riesgo importante en la predicción del riesgo cardiovascular, especialmente cuando se combinan con otros factores de riesgo [24-25].

Presencia de Enfermedades Autoinmunes: Algunas enfermedades autoinmunes, como la artritis reumatoide, pueden aumentar el riesgo de padecer ECV [14].

Comorbilidades: La presencia de otras condiciones médicas, como la diabetes, la hipertensión o la enfermedad renal, puede aumentar el riesgo de ECV y complicar su manejo [26].

Raza o Etnia: La pertenencia a ciertos grupos étnicos o raciales puede influir en el riesgo de ECV debido a factores genéticos, culturales y socioeconómicos únicos [14].

Alimentación o Dieta: La dieta juega un papel importante en la salud cardiovascular, y el consumo de alimentos ricos en grasas saturadas, sodio y azúcares añadidos puede aumentar el riesgo de enfermedades del corazón [27].

Diversas investigaciones científicas han señalado la existencia de otros factores de riesgo que, si bien han sido validados, no suelen incorporarse en los modelos predictivos. Entre estos, cabe mencionar los siguientes.

Inflamación Crónica: Aunque la inflamación sistémica crónica se ha relacionado con enfermedades cardiovasculares, no siempre se considera en los modelos de predicción de riesgo, a pesar de su importancia en la fisiopatología de estas enfermedades [28].

Estrés: Altos niveles de estrés crónico pueden influir en el desarrollo de ECV debido a sus efectos sobre el sistema cardiovascular [29].

Acceso a Atención Médica: La carencia de servicios médicos adecuados y a tiempo puede influir en el desarrollo y manejo de ECV, lo que destaca la importancia de la equidad en la salud [14, 30].

2.2.3 Modelos predictivos

A continuación, se revisan algunos de los principales modelos existentes, como el modelo Framingham, el Sistema SCORE y el modelo QRISK. Estos modelos se basan en algoritmos de regresión logística para calcular el riesgo de enfermedad cardiovascular. Los cuales consideran como variables predictoras clave la edad, el sexo, la presión arterial, el colesterol y el tabaquismo. El algoritmo de regresión logística combina estos factores de riesgo y les asigna pesos relativos para calcular la probabilidad de que ocurra el evento (enfermedad cardiovascular) dentro de un período de tiempo establecido, generalmente 10 años.

Modelo Framingham (FRS): Es uno de los más conocidos y ampliamente utilizados para predecir el riesgo de enfermedad cardiovascular. Se basa en factores como la edad, el sexo, la presión arterial, el colesterol total, el colesterol HDL, el tabaquismo y la diabetes. Este modelo ha sido fundamental en la evaluación del riesgo cardiovascular en la población general [15, 31].

Sistema SCORE (SCORE): Es un modelo de riesgo cardiovascular desarrollado específicamente para la población europea. Incluye variables como la edad, el sexo, el tabaquismo, la presión arterial sistólica y el colesterol total para estimar el riesgo de eventos cardiovasculares fatales a 10 años. Este modelo ha sido útil en la estratificación del riesgo cardiovascular en Europa [31-32].

Modelo QRISK: Es un enfoque más completo, ya que incorpora una amplia gama de factores de riesgo, incluidos factores socioeconómicos y clínicos, para predecir el riesgo cardiovascular. Además de variables tradicionales como la edad, el sexo y la presión arterial, también considera el IMC, la etnicidad y la presencia de enfermedades crónicas. Este enfoque más holístico ha mejorado la precisión en la predicción del riesgo cardiovascular [31, 33].

Estos modelos tradicionales de evaluación del riesgo cardiovascular, como FRS, SCORE y QRISK, se basan en el algoritmo de regresión logística para estimar el riesgo de eventos cardiovasculares en función de múltiples factores de riesgo. Aunque estos modelos han sido ampliamente utilizados y validados, pueden no capturar relaciones no lineales entre las variables, además de asumir independencia entre ellas. Por tanto, otros algoritmos o enfoques de aprendizaje automático e inteligencia artificial están surgiendo como alternativas para mejorar la precisión de la predicción del riesgo cardiovascular. A continuación, se presentan algunos de estos métodos.

Support Vector Machines (SVM): Son algoritmos de aprendizaje supervisado que se utilizan en clasificación y regresión. En el contexto de ECV, pueden ser útiles para predecir el riesgo cardiovascular al encontrar el hiperplano que mejor separa las diferentes clases de eventos. Son efectivos en espacios de alta dimensionalidad y pueden manejar datos no lineales. Requieren una selección cuidadosa de parámetros y pueden ser computacionalmente costosos [34 p. 710].

Decision Trees (DT): Son algoritmos de aprendizaje supervisado que utilizan una estructura de árbol para representar reglas de decisión. Pueden identificar patrones complejos de interacción entre variables predictoras. Son fáciles de interpretar y pueden manejar datos categóricos y numéricos. Pueden ser propensos al sobreajuste si no se controla la profundidad del árbol [34 p. 675].

Artificial Neural Network (ANN): Son modelos computacionales inspirados en el funcionamiento del cerebro humano. Pueden capturar relaciones no lineales y adaptarse a datos complejos. Requieren grandes cantidades de datos para entrenar y pueden ser difíciles de interpretar [34 p. 801].

Random Forest (RF): Es un algoritmo de aprendizaje supervisado basado en árboles de decisión que combina múltiples árboles para mejorar la precisión predictiva. Se utiliza para identificar patrones complejos en los datos al considerar interacciones no lineales entre múltiples variables predictoras [34 p. 715-716].

Gradient Boosting Machines (GBM): Es un algoritmo de aprendizaje supervisado que construye un modelo predictivo en forma de conjunto de modelos de predicción débiles. Se utiliza para mejorar la precisión predictiva al enfocarse en los errores del modelo anterior y ajustar sucesivamente los modelos subsiguientes [34 p. 719-720].

K-Nearest Neighbors (K-NN): Es un método de aprendizaje supervisado utilizado para clasificación y regresión. Puede predecir el riesgo cardiovascular basándose en la similitud con los vecinos más cercanos en el espacio de características. Es fácil de entender e implementar, no hace suposiciones sobre la distribución de los datos y puede adaptarse a cambios en el conjunto de datos. Requiere una cantidad significativa de memoria para almacenar todos los datos de entrenamiento y puede ser sensible a valores atípicos [34 p. 705].

2.3. Investigaciones relacionadas

En el presente apartado, se realiza una revisión de las investigaciones relacionadas con el desarrollo de enfermedades cardiovasculares a partir de factores de riesgo. El objetivo es analizar los diferentes modelos predictivos desarrollados en investigaciones previas, con el fin de identificar su efectividad o nuevos enfoques que contribuyan a la creación de un modelo predictivo más preciso y confiable.

Para llevar a cabo esta revisión, se tomaron en cuenta estudios y trabajos académicos que hayan abordado el tema desde diferentes perspectivas, como el análisis de factores de riesgo tradicionales (como la edad, el género, la presión arterial, el colesterol, entre otros) y la incorporación de variables no tradicionales (como indicadores genéticos, datos de estilo de vida y hábitos alimentarios, entre otros).

Asimismo, se analizan las metodologías empleadas en estas investigaciones, evaluando la calidad de los datos utilizados, los algoritmos y técnicas predictivas empleadas, y los resultados obtenidos. Esto puede permitir identificar las fortalezas y limitaciones de los modelos predictivos existentes y proponer nuevas líneas de acción y enfoques que puedan mejorar la precisión y confiabilidad del modelo propuesto en este trabajo.

De acuerdo con Cai Yue et al. (2024) [35], a partir de 20,887 referencias revisadas, analizaron 79 artículos (82.5% entre 2017-2021) en un estudio sistemático que abarco el desarrollo y análisis de 647 modelos predictivos de ECV usando aprendizaje automático, basados en 114 conjuntos de datos. Los conjuntos de datos provinieron principalmente de Europa (27), América (40, en su mayoría de EE. UU.), Asia (27, principalmente de Corea), y Oceanía (5 de Australia), con la notable ausencia de estudios en África. Identificaron 63 artículos centrados en la población general y 16 en subgrupos con enfermedades específicas como la diabetes tipo 2, la hipertensión y las enfermedades renales. La mayoría de los conjuntos de datos provinieron de registros de salud electrónicos (RSE), con una combinación menor de RSE y cuestionarios, o cuestionarios y entrevistas personales. Respecto a la transparencia de los algoritmos y la reproducibilidad de los modelos, identificaron 13 categorías de 66 algoritmos específicos, siendo la regresión logística, random forest y las redes neuronales los más utilizados.

En consecuencia, el presente trabajo acotará el análisis a aquellos estudios que se ajusten al enfoque propuesto para este proyecto, específicamente los 3 trabajos recomendados por su alta replicabilidad y calidad metodológica según lo establecido en la revisión crítica de Cai Yue et al. (2024) [35], así como aquellas contribuciones relevantes publicadas con posterioridad a julio de 2021.

En el trabajo realizado por Lindholm Daniel et al. (2018) [36], el objetivo fue identificar nuevos factores de riesgos para la insuficiencia cardiaca. Se utilizó un modelo de Gradient Boosting Machine (GBM), el cual permitió clasificar a las personas con insuficiencia cardiaca o no, considerando todas las variables disponibles y mejorando la predicción en cada iteración. Se empleó el análisis de regresión de Cox para evaluar la asociación entre las características identificadas y la insuficiencia cardíaca incidente, ajustando variables como la edad, el sexo y otros factores de riesgo establecidos. Se utilizó la plataforma de aprendizaje automático H2O para realizar los análisis y el modelado predictivo. El estudio utilizó datos de más de 500 mil personas de la población general del Reino Unido. Se consideraron 3646 variables en el análisis, de las cuales se identificaron 15 como las más importantes

para la predicción. Además de factores de riesgos establecidos como la enfermedad coronaria y la diabetes tipo 2, se identificaron variables novedosas como la bioimpedancia de las piernas y el volumen medio de reticulocitos. Se evaluó la precisión de los modelos mediante el área bajo la curva ROC (AUROC) con un rendimiento del 0.81. Como conclusión, señalaron que la bioimpedancia de piernas se asocia inversamente con la incidencia de insuficiencia cardiaca en la población general. Donde un modelo simple de medidas exclusivamente no invasivas, que combina la bioimpedancia de la pierna con antecedentes de infarto de miocardio, edad y sexo, proporcionó una capacidad predictiva precisa.

Cho Sang-Yeong et al. (2021) [37] compararon modelos preexistentes de predicción de riesgo cardiovascular como Pooled Cohort Equation (PCE), FRS, SCORE y QRISK3 con otros modelos de aprendizaje automático. Los modelos utilizados incluyeron regresión logística, agregación bootstrap (Bagging), RF, AdaBoost y ANN. El estudio utilizó datos de más de 220 mil personas adultas de Corea. Consideraron 16 variables predictoras; edad, sexo, presión arterial sistólica, colesterol total, colesterol HDL, tabaquismo, diabetes, medicación antihipertensiva y otras 8 variables adicionales utilizadas por QRISK3. Se evaluó la precisión de los modelos mediante AUROC, destacando los algoritmos de redes neuronales y regresión logística con valores 0.751 y 0.749, respectivamente. Como conclusión, señalaron que los algoritmos de aprendizaje automático podrían mejorar el rendimiento en la predicción de riesgo cardiovascular sobre los modelos tradicionales en adultos sanos de Corea. Aunque la magnitud de la mejora resultó modesta, ya que al evaluar los modelos tradicionales en el conjunto de test, los modelos FRS, SCORE y QRISK3 obtuvieron un rendimiento de 0.704, 0.764 y 0.764, respectivamente. Aunque para cada modelo se utilizó un conjunto de datos ligeramente distinto, y en estos casos el mayor rendimiento lo obtuvo el algoritmo de regresión logística 0.785 y el algoritmo de redes neuronales 0.765.

Jiang Yunxing et al. (2021) [38] evaluaron la viabilidad y utilidad de varios modelos de aprendizaje automático en la predicción de riesgos de enfermedad cardiovascular en la población kazaja de China. Además, se buscó identificar los factores de riesgo más relevantes y determinar qué modelo generaba el mejor rendimiento predictivo. Se emplearon 7 algoritmos, incluyendo Logistic Regression (LR), SVM, K-NN, RF, Gaussian Naive Bayes (NB), DT y Extreme Gradient Boosting (XGBoost). Se utilizó el concepto de importancia de variables para identificar los factores de riesgo más significativos. Esto se logró a través de la técnica de “mean decrease impurity” en un modelo de RF. El estudio incluyó 1508 personas, utilizando 22 variables predictoras, entre características sociodemográficas, historial médico, citocinas e índices sintéticos. Los algoritmos con mejor rendimiento fueron LR con un AUROC de 0.872 y SVM con 0.868. Respecto a las conclusiones, el estudio utilizó datos de una población kazaja en China, con una muestra relativamente pequeña pero representativa. Además de los factores de riesgo estándar, las citocinas inflamatorias y otros biomarcadores fueron identificados como factores de riesgo importantes para la predicción de enfermedades cardiovasculares.

Yu Jingzhi et al. (2024) [39] incorporaron el historial longitudinal de los factores de riesgo en la predicción del riesgo a 10 años de enfermedad cardiovascular aterosclerótica (ASCVD), utilizando un modelo de Deep Learning (DL) y evaluaron su rendimiento en comparación con las Ecuaciones Agrupadas de Cohortes (PCE) que se utilizan actualmente en la práctica clínica. Se utilizó el algoritmo Dynamic-DeepHit, que es una combinación de redes neuronales profundas y modelos de riesgo dinámico para el análisis de supervivencia con riesgos competitivos basados en datos longitudinales. El estudio incluyó 15,565 personas, utilizando como variables predictoras sexo, raza, edad, colesterol

total, colesterol HDL, presión arterial sistólica y diastólica, diabetes, tratamiento para hipertensión y tabaquismo. El modelo Dynamic-DeepHit mejoró el rendimiento sobre el PCE, con AUROC de 0.815 y 0.792, respectivamente. La presión arterial sistólica fue identificada como el predictor más importante en el modelo Dynamic-DeepHit, seguida por el colesterol total y la edad. Como conclusión, señalaron que incorporar datos longitudinales de factores de riesgo mediante DL mejoró la predicción del riesgo de ASCVD en comparación con el PCE.

En el estudio realizado por Kim Joung Ouk Ryan et al. (2021) [40], el objetivo fue desarrollar un modelo de predicción de ECV utilizando algoritmos de aprendizaje automático aplicados a los datos del Sistema Nacional de Seguro de Salud de Corea (NHIS-HEALS). Se utilizaron diversos algoritmos como regresión logística, DT, K-NN, RF, GBM, XGBoost, SVM y ANN. Se emplearon los métodos de importancia de características y permutación de importancia para identificar las variables más importantes que contribuyeron al rendimiento del modelo. El estudio incluyó a 9,398 personas, considerando 38 variables relacionadas con factores médicos y de comportamiento. Los algoritmos XGBoost, GBM y RF mostraron la mejor precisión promedio, con valores AUROC de 0.812, 0.812 y 0.811 respectivamente. Otras métricas de rendimiento como matriz de confusión y puntuación F1 también se emplearon para evaluar y comparar los modelos de predicción. El historial previo de enfermedades cardiovasculares se identificó como el factor más importante para el rendimiento del modelo de predicción. Otras variables destacadas incluyeron el colesterol total, el colesterol de lipoproteínas de baja densidad (LDL), la relación cintura-altura y el IMC. Como conclusión, indicaron que es posible una predicción de ECV más fácil y eficiente mediante el uso de algoritmos de aprendizaje automático, evitando así los costos y cargas adicionales asociados a la recopilación de datos de referencia, en comparación con los modelos tradicionales de predicción de riesgo cardiovascular.

Al-Droubi Samer S et al. (2023) [41] tuvieron como objetivo desarrollar modelos predictivos utilizando inteligencia artificial para evaluar el riesgo de enfermedad cardiovascular en pacientes oncológicos. Se utilizaron datos de pacientes anonimizados del Vanderbilt University Medical Center en Nashville, Tennessee, incluyendo pacientes con cáncer de mama, riñón y linfoma de células B, así como aquellos que recibieron inmunoterapia para el tratamiento de melanoma, cáncer de pulmón o cáncer de riñón. Se aplicaron algoritmos de RF y ANN. El estudio analizó a 20 mil personas. Las variables utilizadas incluyeron factores demográficos, historial médico, medicamentos y sobre todo resultados de pruebas de laboratorio. Tanto RF como ANN tuvieron un rendimiento AUROC por encima del 0.90, donde los modelos ANN ofrecieron un rendimiento por encima de RF, llegando al 0.996 en uno de los conjuntos de datos evaluados. Como conclusión, señalaron que los modelos predictivos de aprendizaje automático demostraron ser efectivos para identificar pacientes con riesgo de ECV relacionada con tratamientos contra el cáncer.

Salah Haya y Srinivas Sharan (2022) [42] exploraron en su investigación un marco de aprendizaje automático explicativo para predecir el riesgo de enfermedad cardiovascular a largo plazo entre adolescentes. Se buscó identificar factores de riesgo clave que pudieran ayudar a prevenir la enfermedad cardiovascular en la edad adulta. Los algoritmos utilizados en el estudio fueron DT, RF, XGBoost y Deep Neural Networks (DNN). Estos modelos se utilizaron para predecir el riesgo de enfermedad cardiovascular basándose en una variedad de factores de riesgo recopilados de la Encuesta Nacional de Salud y Desarrollo Adolescente (Add Health) en Estados Unidos. El estudio analizó 14 mil personas, utilizando 36 variables predictoras, incluyendo aspectos sociodemográficos (género, edad, raza), socioeconómicos (educación e ingresos de los padres), estilo de vida (actividad

física, alimentación, consumo de alcohol y tabaco, calidad del sueño, entre otros), salud mental y eventos estresantes. Los modelos que ofrecieron un mejor rendimiento fueron XGBoost y RF con un AUROC de 0.846 y 0.841, respectivamente. Las variables con mayor influencia fueron edad, género, IMC, ingresos de los padres, actividad física, consumo de comida rápida, consumo de tabaco, vida sedentaria y no desayunar; dependiendo del modelo, algunas variables resultaron más relevantes que otras, pero se mantenían como las más importantes. El estudio demostró que los algoritmos de aprendizaje automático podían predecir con precisión el riesgo de ECV en la edad adulta utilizando factores de riesgo de la adolescencia. La interpretabilidad con el método SHAP (SHapley Additive exPlanations) proporcionó información valiosa sobre la importancia y el impacto de los diferentes factores de riesgo. En general, la investigación introdujo un enfoque novedoso para la predicción y prevención temprana del riesgo de ECV utilizando técnicas de aprendizaje automático explicables.

El estudio anterior es el que más se ajusta al enfoque que se va a realizar en el presente trabajo, sobre todo por el desarrollo metodológico, dado que los datos fueron obtenidos a partir de una encuesta nacional. De manera análoga, este trabajo también obtiene los datos a partir de una encuesta nacional, aunque no se limita únicamente a adolescentes.

Quesada Jose A et al. (2019) [43] analizaron y compararon la capacidad predictiva de 15 métodos de aprendizaje automático para estimar el riesgo cardiovascular. Además, compararon estos métodos con las escalas de riesgo SCORE y REGICOR comúnmente utilizadas en la práctica clínica en España. Los modelos utilizados fueron regresión Cox, regresión Cox penalizada, NB, análisis discriminante lineal y cuadrático, regresión logística, regresión logística penalizada, K-NN, SVM lineales y radiales, ANN, DT, Bagging, AdaBoost y RF. Y las escalas de riesgos tradicionales utilizadas fueron SCORE y REGICOR (calibración española del método Framingham). El estudio analizó más de 38 mil personas del Estudio Cardiovascular Valenciano (ESCARVAL). Las variables utilizadas fueron edad, sexo, colesterol total, presión arterial sistólica y consumo de tabaco, para el método REGICOR se añadieron las variables presión arterial diastólica, colesterol HDL y presencia de diabetes. Los tres modelos con mayor capacidad predictiva fueron el análisis discriminante cuadrático, seguido por NB y ANN, con valores AUROC 0.709, 0.708 y 0.704, respectivamente. Además, 10 de los 15 métodos tuvieron mejor capacidad predictiva que SCORE y REGICOR. Y 7 métodos superaron en un 7% la capacidad predictiva de SCORE y REGICOR. Como conclusión, el estudio demostró que varios métodos de aprendizaje automático superaron a las escalas de riesgo tradicionales en la predicción del riesgo cardiovascular. De manera que, los métodos de aprendizaje automático ofrecieron una alternativa prometedora para mejorar la precisión en la predicción del riesgo cardiovascular en comparación con las herramientas tradicionales.

Por último, el trabajo realizado por Alaa Ahmed M et al. (2019) [44] tenía como objetivo desarrollar un modelo de aprendizaje automático para predecir el riesgo de enfermedades cardiovasculares en personas asintomáticas. Asimismo, evaluó si este enfoque mejoraba la precisión predictiva en comparación con los sistemas de puntuación de riesgo convencionales como Framingham, además de considerar si variables no tradicionales podrían aumentar la precisión de la predicción de riesgo de ECV. Los modelos utilizados fueron SVM lineal, RF, ANN, AdaBoost y GBM, además del framework AutoPrognosis, el cual utiliza optimización bayesiana para diseñar modelos de pronóstico clínico. El estudio analizó más de 423 mil personas del Biobank del Reino Unido, utilizando 473 variables, incluyendo historia médica, estilo de vida, medidas físicas, etc. Los modelos con mejor rendimiento fueron; AutoPrognosis, con todas las variables, tuvo un AUROC de 0.774, AutoPrognosis con 7 variables 0.744, y los modelos GBM, AdaBoost y ANN tuvieron un AUROC de 0.769, 0.759 y

0.755, respectivamente. Como conclusiones, el modelo AutoPrognosis demostró poder proporcionar predicciones de riesgo de ECV confiables utilizando únicamente variables, que no requieren pruebas de laboratorio, sobre el estilo de vida y el historial médico de las personas. Donde las variables más predictivas en el modelo fueron edad, género, tabaquismo, ritmo de caminata habitual, autoevaluación general de salud, diagnósticos previos de hipertensión arterial, ingresos, índice de Townsend y edades de los padres al morir. La inclusión de tales variables podría ayudar a proporcionar predicciones de riesgo razonablemente precisas cuando no es viable obtener variables de laboratorio. Un hallazgo notable fue que, además de los factores de riesgo de edad y género bien establecidos, otras dos variables que no son de laboratorio, el "ritmo de caminata habitual" y la "autoevaluación de salud", resultaron ser muy predictivas de los resultados de ECV. Ninguna de estas dos variables se incluye en las herramientas de predicción de riesgo existentes como Framingham.

2.4. Conclusiones

Las enfermedades cardiovasculares (ECV) son la principal causa de muerte en todo el mundo, pero más del 80% es prevenible mediante intervención temprana y cambios en el estilo de vida. La mayoría de los casos de ECV se detectan en la edad adulta, pero los factores de riesgo comienzan a una edad más temprana [42]. Se predice que las ECV causarán más de 23 millones (alrededor del 30,5%) de muertes para 2030 en todo el mundo [45]. Aunque se han reducido las tasas de mortalidad por ECV en regiones de altos ingresos, el 50% de la mortalidad por ECV y el 80% de la carga global de ECV ocurren en países en vías de desarrollo [46].

Los modelos tradicionales, como FRS, SCORE y QRISK, se basan principalmente en datos de poblaciones más grandes y a menudo pasan por alto las variaciones individuales únicas en los factores de riesgo. Esto les impide ofrecer evaluaciones de riesgo personalizadas para pacientes individuales. Además, existe una variabilidad, en los factores utilizados y el nivel de detalle, entre los diferentes puntajes de riesgo tradicionales. Esto resulta en variaciones significativas en el riesgo de ECV calculado para una misma persona. Los criterios de elegibilidad y las definiciones de resultados difieren entre los estudios de validación, lo que puede afectar la calibración de estos modelos y conducir a una sobreestimación o subestimación del riesgo en diferentes poblaciones. La falta de especificación de la etnia en las predicciones de riesgo afecta significativamente la precisión de las evaluaciones de riesgo individual [47].

De la misma manera, en la implementación de estos modelos predictivos multivariables (FRS, SCORE y QRISK), los investigadores han recurrido tradicionalmente a técnicas como las regresiones de Cox o logísticas, basándose en suposiciones de distribución normal, censura aleatoria y correlaciones lineales entre predictores y resultados. Sin embargo, estas suposiciones pueden comprometer la precisión y fiabilidad de los modelos predictivos. Por ejemplo, la distribución de ciertos factores de riesgo puede no ser normal, lo que conduce a predicciones sesgadas. Además, la censura aleatoria o no informativa puede resultar en datos incompletos, afectando la estimación general del riesgo. Otro aspecto es que muchas interacciones entre variables son no lineales, y la suposición de linealidad puede llevar a una representación inadecuada de las relaciones complejas entre predictores y resultados [48]. Este reconocimiento de las limitaciones subraya la necesidad de enfoques más sofisticados y flexibles en la modelización predictiva en el ámbito de la salud.

Una de las principales fortalezas del aprendizaje automático radica en su capacidad para utilizar datos multimodales, lo cual permite analizar datos extensos de diversas fuentes, incluyendo sensores

fisiológicos, secuenciación genómica, pruebas de imagen, tomografía computarizada, ecocardiografía y registros médicos electrónicos. De esta manera, estas técnicas pueden procesar información heterogénea para identificar biomarcadores asociados con subcategorías específicas de enfermedades, mejorando la detección temprana, prediciendo respuestas a medicamentos y ofreciendo información sobre el pronóstico del paciente [49].

El aprendizaje automático ofrece un enfoque prometedor para mejorar la predicción del riesgo cardiovascular, destacando por su capacidad para capturar interacciones complejas entre los diversos factores de riesgo, revelar nuevos predictores y aprovechar la riqueza de los datos multimodales. Estas capacidades pueden permitir evaluaciones de riesgo más precisas y personalizadas en comparación con los modelos tradicionales, lo que en última instancia conduce a una mejor prevención y manejo de las enfermedades cardiovasculares.

Todo lo anterior refleja la necesidad de seguir buscando soluciones y enfoques nuevos que permitan una detección y atención temprana de las ECV. No obstante, estas soluciones deben tomar en cuenta la accesibilidad, evitando depender de pruebas diagnósticas costosas para que puedan ser implementadas en aquellos países cuyos sistemas de salud pública enfrentan limitaciones de recursos. Es en este contexto que el presente trabajo espera contribuir, buscando herramientas que permitan paliar este problema de salud pública global.

3. Materiales y métodos

3.1. Descripción del conjunto de datos

3.1.1 Origen y características

El conjunto de datos utilizado en este proyecto proviene del Sistema de Vigilancia de Factores de Riesgo del Comportamiento (BRFSS, por sus siglas en inglés) de los Centros para el Control y la Prevención de Enfermedades (CDC) de Estados Unidos [50]. El BRFSS es una encuesta telefónica realizada anualmente que recopila información sobre comportamientos de riesgo para la salud, prácticas preventivas y acceso a la atención médica en la población adulta de EE. UU. Los datos corresponden al año 2022.

El archivo se encuentra en formato .xpt, comúnmente utilizado por el software estadístico SAS. Para cargar estos datos en Python, se utilizan las librerías pandas y pyreadstat. El conjunto de datos original contiene 328 variables y se puede descargar desde el siguiente enlace [51].

Sin embargo, muchas de estas variables no son relevantes para el objetivo de este trabajo, que es crear un modelo predictivo para el desarrollo de enfermedades cardiovasculares a partir de factores de riesgo. Algunas variables pueden ser combinaciones o transformaciones de otras variables ya presentes en el conjunto de datos, lo que las hace redundantes para el análisis. Además, el cuestionario del BRFSS incluye preguntas de control para garantizar la calidad de los datos y la coherencia de las respuestas, pero estas variables no aportan información directa sobre los factores de riesgo o las enfermedades cardiovasculares.

Asimismo, las variables que no contienen información variada, es decir, aquellas con todos los valores nulos o con el mismo valor, no contribuyen al modelo predictivo y pueden ser excluidas. Por estas razones, se realizará un proceso de filtrado para seleccionar únicamente aquellas variables que puedan resultar relevantes para este trabajo.

Para obtener más detalles sobre cada una de las variables del conjunto de datos original, se puede consultar el diccionario de datos proporcionado por los CDC en la siguiente fuente [52]. Este diccionario contiene información detallada sobre el nombre, descripción, tipo de datos y valores posibles de cada variable.

3.1.2 Variables y atributos relevantes

De las 328 variables originales, se realiza una selección de aquellas que se consideran relevantes para el objetivo de predecir el desarrollo de enfermedades cardiovasculares a partir de factores de riesgo. Esta selección se basa en diversos criterios, como la validación médica de ciertas variables como factores de riesgo conocidos, la identificación de nuevos factores de riesgo en trabajos anteriores revisados en el estado del arte, y la exploración de posibles nuevos factores dadas las características específicas de este conjunto de datos.

La Tabla 3.1 presenta las variables seleccionadas, junto con su descripción, tipo de datos, rango o categorías, media y desviación estándar (para variables numéricas), y número de valores faltantes. Estas variables se pueden agrupar en diferentes categorías según el tipo de factor de riesgo al que corresponden:

Factores de riesgo no modificables: Edad (_AGEG5YR), Sexo (SEXVAR) y Raza (_IMPRACE).

Factores de riesgo modificables: Tabaquismo (_RFSMOK3), Consumo excesivo de alcohol (_RFDRHV8), Actividad física (EXERANY2), Índice de masa corporal (_BMI5CAT) y Duración del sueño (SLEPTIM1).

Enfermedades crónicas o comorbilidades: Asma (ASTHMA3), Cáncer (CHCOCNC1), Enfisema (CHCCOPD3), Depresión (ADDEPEV3), Fallo renal (CHCKDNY2), Artritis (HAVARTH4) y Diabetes (DIABETE4).

Discapacidades: Sordera (DEAF), Ceguera (BLIND), Deterioro cognitivo (DECIDE), Dificultad para caminar (DIFFWALK), Dificultad para vestirse (DIFFDRES) y Dificultad para hacer recados (DIFFALON).

Factores socioeconómicos y de estilo de vida: Nivel educativo (_EDUCAG), Nivel de ingresos (_INCOMG1), Zona metropolitana de residencia (_METSTAT) y Accesibilidad a la atención médica (MEDCOST1).

Salud general y mental: Percepción de salud general (_RFHLTH), Presencia de problemas de salud mental como estrés y ansiedad (_MENT14D) y Salud física general (_PHYS14D).

COVID-19: Resultado positivo en la prueba de COVID-19 (COVIDPOS)

Además, se incluyen las variables CVDINFR4 (si el participante ha tenido un infarto de miocardio "ataque al corazón" en el pasado), CVDCRHD4 (si al participante se le ha diagnosticado enfermedad coronaria o angina de pecho) y CVDSTRK3 (si el participante ha tenido un accidente cerebrovascular "derrame cerebral o ictus" en el pasado), que se utilizarán para construir la variable objetivo CARDIO, que indica la presencia o ausencia de enfermedades cardiovasculares.

Tabla 3.1. Descripción de variables.

Variable	Descripción	Tipo de datos	Rango/categoría	Media (DE)	Número de Valores faltantes
SEXVAR	Sexo del participante	Categórico	1: Hombre, 2: Mujer	-	0
_AGEG5YR	Grupo de edad en intervalos de 5 años	Categórico	1: 18-24, 2: 25-29, 3: 30-34, 4: 35-39, 5: 40-44, 6: 45-49, 7: 50-54, 8: 55-59, 9: 60-64, 10: 65-69, 11: 70-74, 12: 75-79, 13: 80+, 14[a]	-	0
_BMI5CAT	Categoría de índice de masa corporal (IMC)	Categórico	1: Bajo peso (<18.5), 2: Normal (18.5-24.9), 3: Sobrepeso (25.0-29.9), 4: Obesidad (≥30.0)	-	48,806
_RFSMOK3	Consumo de tabaco	Categórico	1: No, 2: Sí, 9[a]	-	0
_RFDRHV8	Consumo excesivo de alcohol	Categórico	1: No, 2: Sí, 9[a]	-	0
EXERANY2	Realiza actividad física o ejercicio	Binario	2: No, 1: Sí, 7[b], 9[c]	-	2
SLEPTIM1	Horas de sueño	Numérico	1-24, 77[b], 99[c]	7.02(1.5)	3
CVDINFR4	Historial de infarto de miocardio	Binario	2: No, 1: Sí, 7[b], 9[c]	-	4
CVDCRHD4	Historial de enfermedad coronaria	Binario	2: No, 1: Sí, 7[b], 9[c]	-	2
CVDSTRK3	Historial de accidente cerebrovascular	Binario	2: No, 1: Sí, 7[b], 9[c]	-	2
ASTHMA3	Historial de asma	Binario	2: No, 1: Sí, 7[b], 9[c]	-	2
CHCOCNC1	Historial de cáncer	Binario	2: No, 1: Sí, 7[b], 9[c]	-	3
CHCCOPD3	Historial de enfisema o EPOC	Binario	2: No, 1: Sí, 7[b], 9[c]	-	2
ADDEPEV3	Historial de depresión	Binario	2: No, 1: Sí, 7[b], 9[c]	-	7
CHCKDNY2	Historial de enfermedad renal	Binario	2: No, 1: Sí, 7[b], 9[c]	-	2
HAVARTH4	Historial de artritis o enfermedad inflamatoria	Binario	2: No, 1: Sí, 7[b], 9[c]	-	3
DIABETE4	Historial de diabetes	Categórico	1: Sí, 2: Sí (durante el embarazo), 3: No, 4: Prediabetes, 7[b], 9[c]	-	3
DEAF	Sordera o problemas de audición	Binario	2: No, 1: Sí, 7[b], 9[c]	-	18,644
BLIND	Ceguera o problemas de visión	Binario	2: No, 1: Sí, 7[b], 9[c]	-	19,855
DECIDE	Deterioro cognitivo	Binario	2: No, 1: Sí, 7[b], 9[c]	-	20,986
DIFFWALK	Dificultad para caminar o subir escaleras	Binario	2: No, 1: Sí, 7[b], 9[c]	-	22,155

Continúa en la siguiente página

Tabla 3.1

Variable	Descripción	Tipo de datos	Rango/categoría	Media (DE)	Número de Valores faltantes
DIFFDRES	Dificultad para vestirse o bañarse	Binario	2: No, 1: Sí, 7[b], 9[c]	-	22,879
DIFFALON	Dificultad para hacer recados solo	Binario	2: No, 1: Sí, 7[b], 9[c]	-	23,942
_EDUCAG	Nivel educativo	Categórico	1: No termino la secundaria, 2: Graduado de secundaria, 3: No termino la universidad, 4: Graduado de universidad, 9[a]	-	0
_INCOMG1	Nivel de ingresos	Categórico	1: Menos de $15,000, 2: $15,000-$24,999, 3: $25,000-$34,999, 4: $35,000-$49,999, 5: $50,000-$99,999, 6: $100,000-$199,999, 7: $200,000 o más, 9[a]	-	0
_METSTAT	Vive en zona metropolitana	Binario	2: No, 1: Sí	-	9,408
_IMPRACE	Raza o etnia	Categórico	1: Blanco, 2: Negro, 3: Asiático, 4: Indígena, 5: Hispano, 6: Otra raza		0
MEDCOST1	No pudo ver a un médico debido al costo	Binario	2: No, 1: Sí, 7[b], 9[c]	-	4
_RFHLTH	Percepción de salud general	Binario	1: Buena, 2: Mala, 9[a]	-	0
_PHYS14D	Días de mala salud física en los últimos 30 días	Categórico	1: Ningún día, 2: Algunos días, 3: La mayoría de los días, 9[a]	-	0
_MENT14D	Días de mala salud mental en los últimos 30 días	Categórico	1: Ningún día, 2: Algunos días, 3: La mayoría de los días, 9[a]	-	0
COVIDPOS	Resultado positivo en la prueba de COVID-19	Binario	2: No, 1: Sí, 3: Test positivo en casa, 7[b], 9[c]	-	49,235

a. Engloba las categorías “No sabe/No está seguro” y “Se niega a responder”.
b. No sabe/No está seguro.
c. Se niega a responder.

Tras la selección de variables, el conjunto de datos resultante tiene 32 columnas y 445,132 registros.

3.1.3 Preprocesamiento de datos

Se llevan a cabo varias tareas de preprocesamiento y limpieza para abordar problemas comunes como valores faltantes, valores atípicos e inconsistencias en los datos.

La mayoría de las columnas tienen una cantidad relativamente pequeña de valores nulos, que van desde 0 hasta 7. Sin embargo, algunas columnas, como "DEAF", "BLIND", "DECIDE", "DIFFWALK", "DIFFDRES" y "DIFFALON", tienen una cantidad significativa de valores nulos, que oscilan entre 18,644 y 23,942. Las columnas que presentan más problemas son las que se refieren al Índice de

Masa Corporal ("_BMI5CAT") y si han padecido COVID ("COVIDPOS"), ya que tienen alrededor de un 11% de valores nulos del conjunto total de datos.

Para tratar los valores nulos, se utilizó el método de imputación por ecuaciones encadenadas (MICE), puesto que es una técnica empleada en la revisión bibliográfica y ha demostrado ofrecer buenos resultados. Se creó una instancia de IterativeImputer con un máximo de 100 iteraciones y se aplicó al conjunto de datos. Con el fin de evaluar el impacto de esta imputación en los resultados y verificar si introduce factores de riesgo contraintuitivos u omite factores cuya influencia ha sido validada, se realizará una comparación con modelos sin imputación. En el anexo I, se lleva a cabo la predicción eliminando todos los valores nulos para comparar con los resultados obtenidos utilizando la imputación. De esta manera, se podrá determinar si la imputación afecta negativamente el rendimiento de los modelos.

Dado que la imputación por MICE puede introducir valores decimales en algunas columnas, se realizó una conversión de todas las columnas a enteros para mantener la consistencia y compatibilidad con el formato original de los datos. Sobre todo para la columna "_BMI5CAT", que representa las categorías de índice de masa corporal (IMC) y debe tener valores enteros.

Una vez comprobado que ya no existen valores nulos en el conjunto de datos, se creó la variable objetivo "CARDIO" a partir de las variables "CVDINFR4", "CVDCRHD4" y "CVDSTRK3". Esta variable indica si una persona tiene o no una enfermedad cardiovascular basándose en la información proporcionada por las tres variables mencionadas. Se asignó el valor 1 a la variable "CARDIO" si la persona había sufrido un infarto, angina de pecho o un ictus, y el valor 0 si no había sufrido ninguna de estas condiciones.

Después de crear la variable objetivo, se eliminaron las variables "CVDINFR4", "CVDSTRK3" y "CVDCRHD4" del conjunto de datos para evitar la multicolinealidad y la redundancia, ya que la información relevante sobre la presencia o ausencia de enfermedades cardiovasculares estaba capturada en la variable "CARDIO".

A continuación, se analizó la distribución de la única variable numérica en el conjunto de datos, "SLEPTIM1", que representa la cantidad de horas de sueño que una persona obtiene en un período de 24 horas. Se identificaron valores especiales codificados como 77 ("Don't know/Not Sure") y 99 ("Refused"), que representan respuestas en las que los encuestados no están seguros o se niegan a proporcionar la cantidad de horas de sueño. Se decidió eliminar las filas del conjunto de datos que contenían estos valores, ya que no aportan información útil sobre las horas de sueño y podrían introducir ruido y sesgos en el análisis.

Se generaron estadísticas descriptivas, un histograma y un boxplot para visualizar la distribución de la variable "SLEPTIM1". Los resultados mostraron una distribución con valores atípicos que podrían ser tanto reales como errores de entrada. Dado que aún esos valores atípicos podían ser válidos, se decidió categorizar la variable en cinco categorías: sueño muy corto (menos de 5 horas), sueño corto (5 a 6 horas), sueño normal (7 a 8 horas), sueño largo (9 a 10 horas) y sueño muy largo (más de 10 horas). Se creó una nueva variable "SLEPTIM1_cat" categorizada y se eliminó la variable original "SLEPTIM1" del conjunto de datos.

Una vez realizado el análisis descriptivo de las variables categóricas, se procedió a realizar varias recodificaciones y transformaciones en los datos para facilitar su interpretación y hacerlos más adecuados para el análisis y modelado posterior.

En primer lugar, se eliminaron las filas que contenían los valores 7 y/o 9 en varias columnas, como "MEDCOST1", "EXERANY2", "ASTHMA3", entre otras. Estos valores representan respuestas como "No sabe/No está seguro" o "Se niega a responder", y se consideró que no aportaban información útil para el análisis. Dado el gran tamaño del conjunto de datos, eliminar estas filas no tiene un impacto significativo en la proporción de categorías en la variable objetivo "CARDIO".

Sin embargo, para la variable "_INCOMG1", que representa las categorías de ingresos, se decidió mantener todas las categorías, incluyendo los valores 7 y 9, ya que la categoría 9 representa una proporción significativa (15.76%) de las observaciones y su eliminación podría alterar la distribución de los datos y la representatividad de las categorías de ingresos en el análisis.

A continuación, se realizaron varias recodificaciones en las variables para hacerlas más intuitivas y consistentes en su interpretación. Para las variables binarias que indican la presencia o ausencia de una condición, se recodificaron los valores de manera que el valor 2 (correspondiente a "No") se reemplazó por 0, mientras que el valor 1 (correspondiente a "Sí") se mantuvo sin cambios.

En la variable "DIABETE4", se agruparon las categorías de diabetes (1) y prediabetes (4) en una sola categoría (1), mientras que las categorías de ausencia de diabetes (3) y diabetes gestacional (2) se agruparon en una sola categoría (0).

Para la variable "COVIDPOS", se agruparon las categorías relacionadas con la presencia de COVID-19 (valores 1 y 3) en una sola categoría codificada como 1, mientras que la ausencia de COVID-19 (valor 2) se codificó como 0.

En las variables "EXERANY2", "_RFHLTH", "_RFSMOK3" y "_RFDRHV8", se realizaron recodificaciones para que el valor 0 represente la ausencia de un factor de riesgo y el valor 1 represente la presencia de un factor de riesgo.

Estas recodificaciones y transformaciones de las variables categóricas permiten una interpretación más clara y coherente de las variables, facilitan el análisis posterior y hacen que las variables sean más adecuadas para su uso en modelos de aprendizaje automático.

Para ilustrar el impacto de estas transformaciones y recodificaciones en la distribución de las variables y las proporciones de las categorías, se ha generado la Tabla 4.1 que se presenta en detalle en el apartado de resultados (en el apartado 4.1.1). Esta tabla proporcionará una visión más clara de cómo las transformaciones han afectado la estructura y distribución de los datos, y cómo han contribuido a la preparación de los datos para el análisis y modelado posteriores.

3.2. Metodología

3.2.1 Enfoque general y flujo de trabajo

En la Figura 3.2 se muestra el enfoque general de este proyecto, el cual sigue un flujo de trabajo secuencial que abarca las siguientes etapas:

1. **Obtención de los datos**: Se obtuvieron los datos del Sistema de Vigilancia de Factores de Riesgo del Comportamiento (BRFSS) de los Centros para el Control y la Prevención de Enfermedades (CDC) de Estados Unidos.
2. **Preprocesamiento de datos**: Se realizó un preprocesamiento exhaustivo de los datos, que incluyó la imputación de valores nulos, la eliminación de categorías no relevantes, la recodificación de variables y el tratamiento de valores atípicos.
3. **Análisis exploratorio de datos**: Se llevó a cabo un análisis exploratorio de los datos preprocesados para comprender mejor el conjunto de datos, descubrir patrones y relaciones entre las variables, y analizar su relación con la variable objetivo (presencia o ausencia de enfermedades cardiovasculares).
4. **Selección y extracción de características**: Se seleccionaron las variables más relevantes para el modelado predictivo utilizando dos enfoques: selección basada en conocimiento previo (11 variables) y selección del conjunto completo de variables (29 variables).
5. **Entrenamiento y evaluación de modelos**: Se entrenaron y evaluaron diferentes modelos de aprendizaje automático para predecir la presencia o ausencia de enfermedades cardiovasculares. Debido al desbalanceo de clases en la variable objetivo, se aplicaron diversas técnicas de muestreo para mejorar el rendimiento de los modelos.
6. **Interpretación y explicabilidad de los modelos**: Se utilizaron técnicas de importancia de características, como SHAP values, feature importance y permutation importance, para identificar las variables más influyentes en las predicciones de los modelos y proporcionar interpretabilidad y explicabilidad a los resultados.

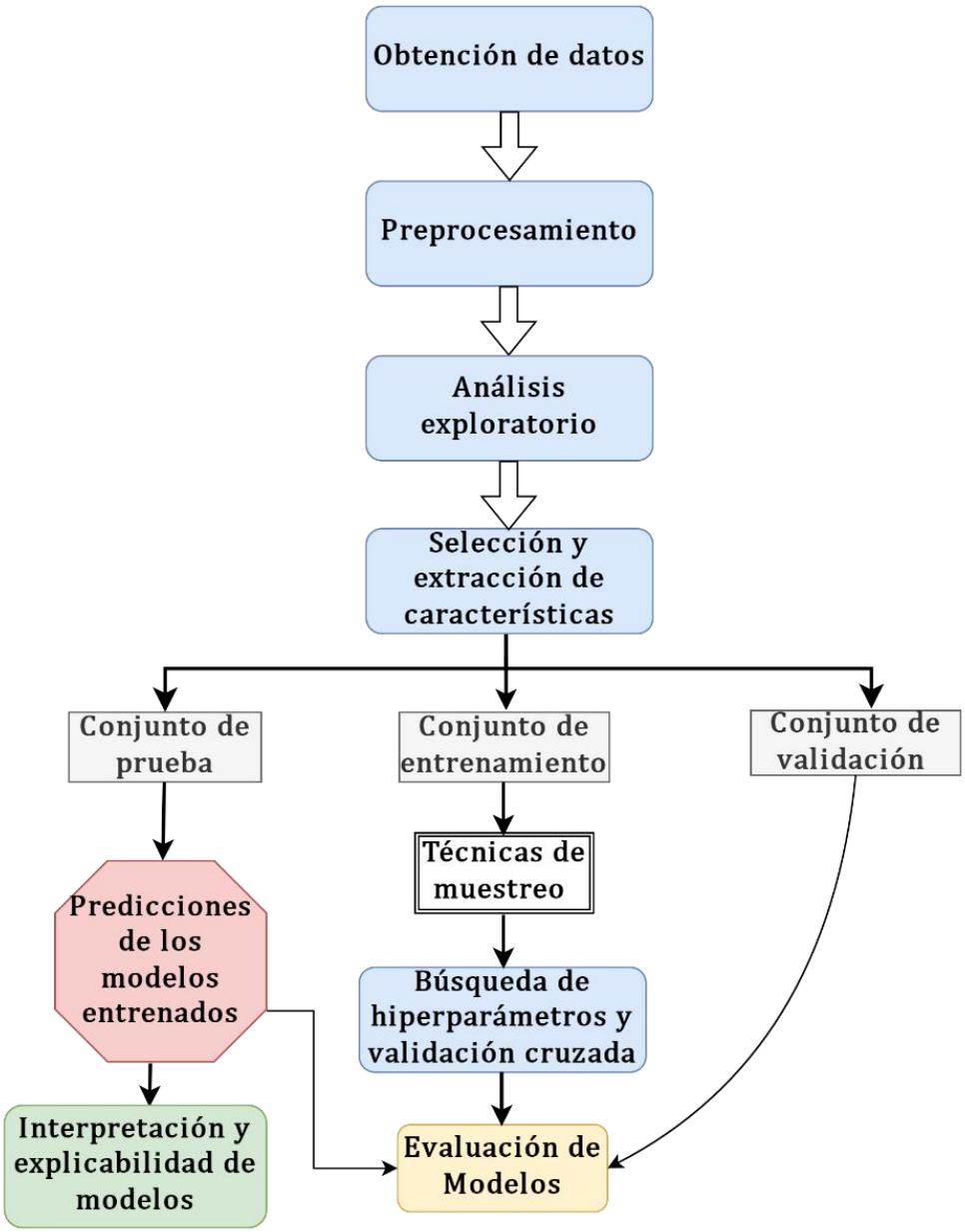


Figura 3.2. Flujo de trabajo de la implementación.

3.2.2 Técnicas de análisis exploratorio de datos

En el análisis exploratorio de datos, se utilizaron diversas técnicas para comprender mejor el conjunto de datos y descubrir patrones y relaciones entre las variables. Estas técnicas incluyeron:

1. Visualizaciones: Se crearon gráficos y visualizaciones para comparar diferentes factores de riesgo con la variable objetivo. Estas visualizaciones permitieron comprender la distribución de las variables y su relación con la presencia o ausencia de enfermedades cardiovasculares.
2. Pruebas estadísticas: Se realizaron pruebas de chi cuadrado para determinar si las relaciones observadas entre las variables y la variable objetivo eran estadísticamente significativas. Estas pruebas se aplicaron tanto para comparaciones entre dos variables como para comparaciones entre más de dos variables.

3. Análisis multivariable: Se exploraron las relaciones entre múltiples variables y la variable objetivo simultáneamente, lo que permitió descubrir patrones más complejos en los datos.

Para llevar a cabo estas técnicas, se crearon cuatro funciones específicas:

- Función para realizar la prueba de chi cuadrado al comparar dos variables categóricas.
- Función para realizar la prueba de chi cuadrado al comparar más de dos variables categóricas.
- Función para crear visualizaciones al comparar dos variables.
- Función para crear visualizaciones al comparar tres variables.

Estas funciones facilitaron el análisis exploratorio y permitieron responder a diversas preguntas sobre la relación entre los factores de riesgo y la presencia de enfermedades cardiovasculares.

3.2.3 Selección y extracción de características

Se utilizaron dos enfoques para la selección y extracción de características:

1. Selección basada en conocimiento previo: Se seleccionaron inicialmente 11 variables que investigaciones previas han señalado como factores relevantes para la predicción de enfermedades cardiovasculares. Este enfoque permitió generar un modelo base de comparación y evitar introducir ruido con un gran número de variables.
2. Selección del conjunto completo de variables: Posteriormente, se entrenaron modelos utilizando todas las variables disponibles (29 en total) para evaluar si la inclusión de variables adicionales mejoraba el rendimiento de los modelos, además de poder descubrir nuevos factores de riesgos al introducir todas las variables.

La Tabla 3.2 proporciona una visión clara de las variables seleccionadas en cada enfoque y el tipo de factor de riesgo que representan, facilitando la comprensión de las decisiones tomadas en la selección de variables y cómo estas podrían afectar los resultados del modelado predictivo.

Tabla 3.2. Variables seleccionadas para cada enfoque.

Enfoque	**Variables seleccionadas**	**Tipo de factor de riesgo**
11 variables	SEXVAR, EXERANY2, CHCCOPD3, CHCKDNY2, HAVARTH4, DIABETE4, _IMPRACE, _AGEG5YR, _BMI5CAT, _RFSMOK3, _RFDRHV8	Diversos (demográficos, de estilo de vida, salud)
Todas las variables	SEXVAR, MEDCOST1, EXERANY2, ASTHMA3, CHCOCNC1, CHCCOPD3, ADDEPEV3, CHCKDNY2, HAVARTH4, DIABETE4, DEAF, BLIND, DECIDE, DIFFWALK, DIFFDRES, DIFFALON, COVIDPOS, _METSTAT, _IMPRACE, _RFHLTH, _PHYS14D, _MENT14D, _AGEG5YR, _BMI5CAT, _EDUCAG, _INCOMG1, _RFSMOK3, _RFDRHV8, SLEPTIM1_cat	Extensivos (incluyen demográficos, de salud, socioeconómicos, de estilo de vida)
Variable Objetivo: CARDIO (Indica la presencia o ausencia de enfermedades cardiovasculares)		

3.2.4 Algoritmos de aprendizaje automático utilizados

En este proyecto, se han empleado diversos algoritmos de aprendizaje automático para abordar la tarea de predecir la presencia o ausencia de enfermedades cardiovasculares. La selección se basa en su capacidad para manejar conjuntos de datos desbalanceados, variables categóricas codificadas numéricamente y capturar relaciones complejas en los datos. A continuación, se listan los algoritmos utilizados; **Regresión Logística** (LR) [53], **Máquinas de Vectores de Soporte** (SVM) [54], **Gaussian Naive Bayes** (GNB) [55], **Gradient Boosting** [56], **XGBoost** [57], **LightGBM** [58], **Árbol de Decisión** [59], **Random Forest** [60], **Red Neuronal** (ANN) [61], **Easy Ensemble** [62], **Voting** [63] y **Stacking** [64].

La elección de estos algoritmos se basa en varios criterios y consideraciones respaldados por la literatura científica. Dado que la variable objetivo presenta un desequilibrio de clases, es importante seleccionar algoritmos que puedan abordar este desafío. Algoritmos como Easy Ensemble y el uso de pesos de clase en modelos como Regresión Logística y SVM han demostrado ser efectivos para mitigar el impacto del desequilibrio de clases.

Los modelos de ensamble, como Gradient Boosting, XGBoost, LightGBM y Random Forest, son conocidos por su capacidad para capturar relaciones complejas en los datos. Estos algoritmos combinan múltiples modelos débiles para crear un modelo final fuerte y pueden manejar espacios de características de alta dimensionalidad.

Dado el tamaño del conjunto de datos y la necesidad de realizar una búsqueda de hiperparámetros, se han considerado algoritmos que ofrezcan un buen equilibrio entre rendimiento y eficiencia computacional. Algoritmos como LightGBM y Random Forest se han seleccionado teniendo en cuenta su capacidad para manejar conjuntos de datos grandes de manera eficiente.

Se ha buscado incluir una variedad de enfoques de aprendizaje automático, desde modelos lineales hasta modelos de ensamble y redes neuronales. Esto permite explorar diferentes perspectivas y capturar diferentes patrones en los datos. La combinación de múltiples enfoques a través de técnicas de ensamble como Voting y Stacking también ha demostrado ser efectiva para mejorar el rendimiento.

Se han incluido modelos simples y consistentes, como Easy Ensemble, que sirven como modelos de control y proporcionan un punto de referencia para el rendimiento mínimo esperado. Estos modelos permiten evaluar si los modelos más complejos ofrecen una mejora significativa en comparación con enfoques básicos. También, se ha considerado la inclusión de modelos ampliamente utilizados en la literatura, como Regresión Logística y SVM, para permitir la comparación con enfoques tradicionales y evaluar si los modelos más avanzados ofrecen una mejora sustancial.

Así mismo, la selección de estos algoritmos se ha realizado teniendo en cuenta la naturaleza del problema, las características del conjunto de datos, la eficiencia computacional y la diversidad de enfoques. Se ha buscado un equilibrio entre modelos robustos, eficientes y capaces de manejar el desequilibrio de clases, al tiempo que se incluyen modelos simples y consistentes como referencia. Esta selección permite explorar diferentes perspectivas y encontrar los modelos más adecuados para predecir la presencia o ausencia de enfermedades cardiovasculares.

3.2.5 Técnicas de muestreo para el desbalanceo de clases

Una vez que los datos se dividieron en conjuntos de entrenamiento, validación y prueba, se abordó el problema del desbalanceo de clases en la variable objetivo mediante la aplicación de diversas técnicas de muestreo al conjunto de datos de entrenamiento. Estas técnicas se utilizaron para entrenar, evaluar e identificar la importancia de las características tanto en el conjunto de datos con 11 variables como en el conjunto de datos con todas las variables disponibles. El objetivo principal fue asegurar que los modelos de aprendizaje automático tuvieran acceso a una representación equilibrada de todas las clases durante el entrenamiento. A continuación, se describen las técnicas de muestreo empleadas.

RandomUnderSampler: Esta técnica de submuestreo aleatorio reduce el número de instancias de la clase mayoritaria para igualar el número de instancias de la clase minoritaria. Esto se logra mediante la eliminación aleatoria de instancias de la clase mayoritaria hasta que se alcance un equilibrio entre las clases.

NearMiss: Es una técnica de submuestreo que selecciona las instancias de la clase mayoritaria que están más cerca de la clase minoritaria en el espacio de características. Existen diferentes versiones de NearMiss, como NearMiss-1, NearMiss-2 y NearMiss-3, que difieren en cómo se calcula la distancia y se seleccionan las instancias.

InstanceHardnessThreshold: Esta técnica de submuestreo se basa en el concepto de "dificultad de instancia" y se aplica al conjunto de entrenamiento. Se calcula la dificultad de cada instancia en función de la frecuencia con la que es clasificada erróneamente por un modelo. Luego, se eliminan las instancias de la clase mayoritaria que tienen una dificultad menor a un umbral específico.

NeighbourhoodCleaningRule: Esta técnica de submuestreo se basa en la identificación y eliminación de instancias de la clase mayoritaria que se consideran "ruidosas" o que invaden el espacio de la clase minoritaria. Para cada instancia, se identifican sus k vecinos más cercanos. Si una instancia de la clase mayoritaria tiene la mayoría de sus vecinos de la clase minoritaria, se considera ruidosa y se elimina. Si una instancia de la clase minoritaria tiene la mayoría de sus vecinos de la clase mayoritaria, esos vecinos se consideran ruidosos y se eliminan.

RandomOverSampler: Esta técnica de sobremuestreo aumenta el número de instancias de la clase minoritaria mediante la duplicación aleatoria de instancias existentes. Replica aleatoriamente las instancias de la clase minoritaria hasta que se alcanza un equilibrio con la clase mayoritaria.

SMOTEN: Es una extensión de la técnica SMOTE diseñada específicamente para datos categóricos. Genera instancias sintéticas para la clase minoritaria mediante la interpolación entre instancias cercanas en el espacio de características categóricas. Además, aplica un enfoque de selección de instancias para identificar y eliminar instancias ruidosas o poco representativas. Utiliza un codificador ordinal para transformar las características categóricas en valores numéricos antes de aplicar el proceso de sobremuestreo.

BorderlineSMOTE: Es una variante de SMOTE que se enfoca en las instancias de la clase minoritaria que se encuentran cerca de la frontera de decisión entre las clases. Genera instancias

sintéticas para la clase minoritaria, utilizando solo las instancias borderline como base para la interpolación.

SMOTEENN: Combina la técnica de sobremuestreo SMOTE con la técnica de submuestreo ENN (EditedNearestNeighbours). Primero, SMOTEENN aplica SMOTE para generar instancias sintéticas para la clase minoritaria. Luego, utiliza ENN para identificar y eliminar las instancias ruidosas o poco representativas tanto de la clase mayoritaria como de la clase minoritaria.

SMOTETomek: Es otra técnica que combina SMOTE con la eliminación de instancias ruidosas utilizando el método Tomek Links. Primero, SMOTETomek aplica SMOTE para generar instancias sintéticas para la clase minoritaria. Luego, identifica los pares de instancias de diferentes clases que son los vecinos más cercanos entre sí (Tomek Links) y elimina estas instancias.

Estas técnicas de muestreo se aplicaron de manera sistemática al conjunto de datos de entrenamiento, tanto para el conjunto de datos con 11 variables como para el conjunto de datos con todas las variables disponibles. Para cada técnica de muestreo, se entrenaron y evaluaron los modelos de aprendizaje automático utilizando los conjuntos de entrenamiento modificados, mientras que los conjuntos de validación y prueba se mantuvieron intactos para una evaluación imparcial del rendimiento del modelo.

Este enfoque exhaustivo permitió evaluar el impacto de cada técnica de muestreo en el rendimiento de los modelos y en la identificación de las características más relevantes para la predicción de la variable objetivo.

La implementación de diversas técnicas de muestreo, al conjunto de entrenamiento, fue necesaria para abordar el desbalanceo de clases y mejorar la capacidad de los modelos para aprender patrones significativos a partir de los datos durante el proceso de entrenamiento. Al equilibrar la representación de las clases y mejorar la calidad de los datos de entrenamiento, se buscó obtener modelos más precisos para la predicción de la presencia o ausencia de enfermedades cardiovasculares.

3.2.6 Métricas de evaluación y validación de modelos

Se han utilizado diversas métricas de evaluación para obtener una visión completa del desempeño de los modelos. A continuación, se listan las métricas utilizadas; **accuracy**, **precision**, **recall** (sensibilidad), **F1-score**, **curva ROC** y **AUC-ROC**, **curva Precision-Recall** y **AUC-PR**, **matriz de confusión**, y **classification report**.

La selección de estas métricas se basa en su idoneidad para el problema de clasificación binaria y su capacidad para manejar conjuntos de datos desbalanceados. La accuracy proporciona una visión general del rendimiento del modelo, pero puede ser engañosa en conjuntos de datos desbalanceados. La precision y el recall se enfocan en la capacidad del modelo para identificar correctamente a los pacientes con enfermedades cardiovasculares, mientras que el F1-score proporciona un equilibrio entre ambas métricas. Las curvas ROC y Precision-Recall, junto con sus áreas bajo la curva (AUC-ROC y AUC-PR), permiten evaluar la capacidad discriminativa de los modelos. La matriz de confusión y el classification report proporcionan una visión detallada del rendimiento del modelo para cada clase.

Además de las métricas de evaluación, se han utilizado técnicas de validación para garantizar la generalización de los modelos y evitar el sobreajuste. En este proyecto, se ha empleado la validación cruzada estratificada con k-folds para evaluar el rendimiento de los modelos durante el proceso de búsqueda de hiperparámetros. La validación cruzada estratificada asegura que cada fold tenga una distribución similar de las clases, lo cual es importante en conjuntos de datos desbalanceados.

Al utilizar múltiples métricas y técnicas de validación, se obtiene una evaluación completa y confiable del rendimiento de los modelos, lo que permite seleccionar los modelos más adecuados para la predicción de enfermedades cardiovasculares.

3.2.7 Búsqueda de hiperparámetros

Se ha realizado una búsqueda de hiperparámetros para 8 modelos seleccionados a partir de los 12 iniciales utilizados al evaluar el conjunto de datos con 11 variables. Los modelos que ofrecieron mejores resultados se utilizaron para la búsqueda de hiperparámetros, empleando la técnica de RandomizedSearchCV.

La elección de RandomizedSearchCV se basó en su eficiencia computacional al muestrear aleatoriamente un subconjunto de combinaciones de hiperparámetros, en lugar de evaluar exhaustivamente todas las combinaciones posibles. Esta técnica permite explorar un espacio de hiperparámetros más amplio de manera eficiente, especialmente cuando se trabaja con conjuntos de datos grandes, como en este caso.

En cuanto a la elección del número de validaciones cruzadas (cv), se utilizan diferentes valores para diferentes modelos. Para Logistic Regression y Gaussian Naive Bayes, se utiliza cv=30, mientras que para el resto de los modelos se utiliza cv=5. Esta decisión se basa en el equilibrio entre la estabilidad de las estimaciones de rendimiento y el costo computacional. Utilizar un mayor número de validaciones cruzadas proporciona estimaciones más estables y confiables del rendimiento del modelo, pero también implica un mayor costo computacional y tiempo de búsqueda.

Además, limitar la búsqueda a 50 iteraciones (n_iter=50) para la mayoría de los modelos es una forma de acotar el tiempo de búsqueda y el costo computacional. Al limitar la búsqueda a 50 iteraciones, se busca encontrar una buena combinación de hiperparámetros en un tiempo razonable, aunque es posible que no se explore exhaustivamente todo el espacio de hiperparámetros.

Por último, otro aspecto importante en la búsqueda de hiperparámetros es la especificación de la métrica de evaluación utilizada para seleccionar los mejores hiperparámetros. En este caso, se establece scoring='f1_weighted' para todos los modelos. La métrica F1-score ponderada (weighted) es relevante debido al considerable desbalanceo de clases en la variable objetivo. La métrica F1-score ponderada tiene en cuenta tanto la precisión como la sensibilidad de cada clase y las pondera según la proporción de muestras en cada clase, asegurando que la clase minoritaria tenga un impacto significativo en la métrica final y reflejando más fielmente el rendimiento del modelo en términos de identificar casos más raros pero importantes.

3.2.8 Herramientas y tecnologías utilizadas

Se han utilizado diversas herramientas y tecnologías para el desarrollo, implementación y evaluación de los modelos. A continuación, se listan las principales herramientas y tecnologías empleadas; **Python** 3.12.2, **Jupyter Notebook** 7.0.8, **Scikit-learn** 1.4.1, **Lightgbm** 4.3.0, **XGBoost** 2.0.3, **Pandas** 2.2.1, **NumPy** 1.26.4, **Matplotlib** 3.8.0, **Seaborn** 0.13.2, **Imbalanced-learn** 0.12.2, **SHAP** 0.45.0, **RandomizedSearchCV**, **feature_importance_**, y **permutation_importance**.

Además de las herramientas y bibliotecas mencionadas anteriormente, se han desarrollado funciones específicas para automatizar y facilitar el proceso de entrenamiento, evaluación y visualización de los modelos. Las funciones entrenar_evaluar_modelos_base y entrenar_evaluar_modelos se encargan de realizar el entrenamiento de los modelos con diferentes técnicas de muestreo, calcular métricas de rendimiento y generar gráficos de curvas ROC y Precision-Recall. Estas funciones permiten una evaluación exhaustiva y comparativa de los modelos bajo diferentes condiciones de muestreo. También se han implementado las funciones graficar_importancia_permutacion y graficar_permutacion_importancia para visualizar la importancia de las variables en los modelos. Estas funciones utilizan tanto la importancia integrada, que se basa en la estructura interna de algunos modelos como los árboles de decisión, como la importancia por permutación, que mide la disminución en el rendimiento del modelo al permutar aleatoriamente los valores de una variable.

Todo el código utilizado en este proyecto, incluyendo los scripts de preprocesamiento, entrenamiento de modelos, evaluación y visualización de resultados, está disponible públicamente en el siguiente repositorio de GitHub: [https://github.com/alvalca/cardio_riskfactors]. Este repositorio proporciona una documentación detallada y permite la reproducibilidad completa de los experimentos y análisis realizados en este trabajo.

4. Resultados

4.1. Análisis exploratorio de datos

4.1.1 Resultados del preprocesamiento

En el capítulo 3, se describieron en detalle los pasos realizados durante el preprocesamiento y limpieza de datos, incluyendo el manejo de valores faltantes mediante imputación por ecuaciones encadenadas, la creación de la variable objetivo "CARDIO", la eliminación de variables redundantes, el tratamiento de valores especiales en la variable "SLEPTIM1" y la recodificación y transformación de variables categóricas.

Como resultado de este proceso, se obtuvo un conjunto de datos limpio y preparado para el análisis y modelado posteriores. La Tabla 4.1 presenta un resumen de las características del conjunto de datos final después del preprocesamiento.

El conjunto de datos resultante consta de 353,968 filas y 30 columnas, donde todas las variables son categóricas. Todas las variables tienen un porcentaje de valores faltantes del 0% debido a la imputación realizada.

Es importante destacar que la variable objetivo "CARDIO" se creó a partir de la combinación de tres variables: "CVDINFR4" (historial de infarto de miocardio), "CVDCRHD4" (historial de enfermedad coronaria) y "CVDSTRK3" (historial de accidente cerebrovascular). La contribución de cada una de estas variables a la presencia de enfermedad cardiovascular (CARDIO = 1) fue la siguiente: CVDINFR4 (5.89%), CVDCRHD4 (6.22%) y CVDSTRK3 (4.48%).

Tabla 4.1. Descripción del conjunto de datos final.

Variable	Descripción	Tipo de datos	Categorías	Frecuencia (%)
SEXVAR	Sexo del participante	Categórico	1: Hombre, 2: Mujer	1: 47.12, 2: 52.88
_AGEG5YR	Grupo de edad en intervalos de 5 años	Categórico Ordinal	1: 18-24, 2: 25-29, 3: 30-34, 4: 35-39, 5: 40-44, 6: 45-49, 7: 50-54, 8: 55-59, 9: 60-64, 10: 65-69, 11: 70-74, 12: 75-79, 13: 80+	1: 6.17, 2: 5.12, 3: 5.91, 4: 6.59, 5: 6.94, 6: 6.59, 7: 7.77, 8: 8.52, 9: 10.29, 10: 10.99, 11: 10.08, 12: 7.37, 13: 7.65
_BMI5CAT	Categoría de índice de masa corporal (IMC)	Categórico Ordinal	1: Bajo peso (<18.5), 2: Normal (18.5-24.9), 3: Sobrepeso (25.0-29.9), 4: Obesidad (≥30.0)	1: 1.51, 2: 27.64, 3: 39.01, 4: 31.85
_RFSMOK3	Consumo de tabaco	Binario	0: No, 1: Sí	0: 88.33, 1: 11.67
_RFDRHV8	Consumo excesivo de alcohol	Binario	0: No, 1: Sí	0:93.2, 1: 6.8
EXERANY2	Realiza actividad física o ejercicio	Binario	0: Sí, 1: No	0: 77.22, 1: 22.78
ASTHMA3	Historial de asma	Binario	0: No, 1: Sí	0: 85.02, 1: 14:98
CHCOCNC1	Historial de cáncer	Binario	0: No, 1: Sí	0: 88.55, 1: 11.45
CHCCOPD3	Historial de enfisema o EPOC	Binario	0: No, 1: Sí	0: 92.38, 1: 7.62
ADDEPEV3	Historial de depresión	Binario	0: No, 1: Sí	0: 79.27, 1: 20.73
CHCKDNY2	Historial de enfermedad renal	Binario	0: No, 1: Sí	0: 95.5, 1: 4.5
HAVARTH4	Historial de artritis o enfermedad inflamatoria	Binario	0: No, 1: Sí	0: 66.04, 1: 33.96
DIABETE4	Historial de diabetes	Binario	0: No, 1: Sí	0: 84.29, 1: 15.71
DEAF	Sordera o problemas de audición	Binario	0: No, 1: Sí	0: 91.33, 1: 8.67
BLIND	Ceguera o problemas de visión	Binario	0: No, 1: Sí	0:94.94, 1: 5.06
DECIDE	Deterioro cognitivo	Binario	0: No, 1: Sí	0: 88.87, 1: 11.13
DIFFWALK	Dificultad para caminar o subir escaleras	Binario	0: No, 1: Sí	0: 85.07, 1: 14.93
DIFFDRES	Dificultad para vestirse o bañarse	Binario	0: No, 1: Sí	0:96.47, 1: 3.53
DIFFALON	Dificultad para hacer recados solo	Binario	0: No, 1: Sí	0: 92.99, 1: 7.01
_EDUCAG	Nivel educativo	Categórico Ordinal	1: No termino la secundaria, 2: Graduado de secundaria, 3: No termino la universidad, 4: Graduado de universidad	1: 5.05, 2: 23.67, 3: 27.24, 4: 44.04

Continúa en la siguiente página

Tabla 4.1

Variable	Descripción	Tipo de datos	Categoría	Frecuencia (%)
_INCOMG1	Nivel de ingresos	Categórico Ordinal	1: Menos de $15,000, 2: $15,000-$24,999, 3: $25,000-$34,999, 4: $35,000-$49,999, 5: $50,000-$99,999, 6: $100,000-$199,999, 7: $200,000 o más, 9: No sabe/Se niega a responder	1: 4.65, 2: 7.83, 3: 9.75, 4: 11.14, 5: 26.58, 6: 18.35, 7: 5.93, 9: 15.76
_METSTAT	Vive en zona metropolitana	Binario	1: Sí, 2: No	1: 73.3, 2: 26.7
_IMPRACE	Raza o etnia	Categórico Nominal	1: Blanco, 2: Negro, 3: Asiático, 4: Indígena, 5: Hispano, 6: Otra raza	1: 76.04, 2: 7.47, 3: 2.93, 4: 1.51, 5: 9.32, 6: 2.74
MEDCOST1	No pudo ver a un médico debido al costo	Binario	0: No, 1: Sí	0: 92.2, 1: 7.8
_RFHLTH	Percepción de salud general	Binario	0: Buena, 1: Mala	0: 83.38, 1: 16.62
_PHYS14D	Días de mala salud física en los últimos 30 días	Categórico Ordinal	1: Ningún día, 2: Algunos días, 3: La mayoría de los días	1: 61.92, 2: 25.39, 3: 12.69
_MENT14D	Días de mala salud mental en los últimos 30 días	Categórico Ordinal	1: Ningún día, 2: Algunos días, 3: La mayoría de los días	1: 60.58, 2: 26.09, 3: 13.33
COVIDPOS	Resultado positivo en la prueba de COVID-19	Binario	0: No, 1: Sí	0: 66.74, 1: 33.26
SLEPTIM1_cat	Horas de sueño	Categórico Ordinal	1: Muy corto (<5) 2: Corto (5-6), 3: Normal (7-8), 4: Largo (9-10), 5: Muy largo (>10)	1: 3.82, 2: 28.52, 3: 59.47, 4: 7.13, 5: 1.07
CARDIO	Historial de enfermedad cardiovascular	Binario	0: No, 1:Sí	0: 88.68, 1: 11.32

Al examinar los porcentajes de las diferentes categorías en las variables del conjunto de datos final, se pueden resaltar varios aspectos de interés. En primer lugar, la variable "SEXVAR" muestra una distribución bastante equilibrada entre hombres y mujeres, con un 47.12% de hombres y un 52.88% de mujeres. Esta distribución indica que la encuesta logró capturar una representación equitativa de ambos sexos, lo cual es importante para obtener resultados generalizables.

En cuanto a la edad de los participantes, representada por la variable "_AGEG5YR", se observa que los grupos de edad con mayor representación son los de 60-64 años (10.29%), 65-69 años (10.99%) y 70-74 años (10.08%).

Otro aspecto destacable es la prevalencia del sobrepeso y la obesidad en la población estudiada, según lo revelado por la variable "_BMI5CAT". El sobrepeso (39.01%) y la obesidad (31.85%) son las categorías más prevalentes, revelando que una gran proporción de los participantes tienen problemas de peso, además de constatar que hay más personas con obesidad que con peso normal. Donde alrededor del 70% de la población de Estados Unidos tiene sobrepeso u obesidad.

Además, la tabla revela información interesante sobre la prevalencia de otras condiciones de salud. Por ejemplo, el 33.96% de los participantes reportan un historial de artritis o enfermedad inflamatoria, el 15.71% tienen un historial de diabetes, el 14.98% tienen un historial de asma y el 11.45% tienen un historial de cáncer. Estas condiciones crónicas podrían interactuar con los factores de riesgo cardiovascular y contribuir a un mayor riesgo de enfermedades cardiovasculares.

En términos de composición racial, la variable "_IMPRACE" muestra que la mayoría de los participantes son de raza blanca (76.04%), seguidos por hispanos (9.32%) y negros (7.47%). Esta distribución refleja la composición racial de la población general en Estados Unidos. Sin embargo, es importante tener en cuenta que las enfermedades cardiovasculares pueden afectar de manera diferente a diversos grupos raciales y étnicos.

Otro factor relevante es la duración del sueño, representada por la variable "SLEPTIM1_cat". Se observa que la mayoría de los participantes (59.47%) reportan una duración de sueño normal (7-8 horas), mientras que el 28.52% reporta un sueño corto (5-6 horas) y solo el 3.82% reporta muy pocas horas de sueño (<5 horas).

Uno de los aspectos llamativos es la alta prevalencia de problemas de salud mental y física entre los participantes. Uno de ellos es la proporción de participantes que reportan mala salud mental debido a estrés o ansiedad. Según los datos, el 39.42% de los participantes experimentaron días de mala salud mental en el último mes (26.09% algunos días y 13.33% la mayoría de los días). Esta cifra es preocupante, ya que indica que una parte significativa de la población enfrenta dificultades de salud mental, lo cual podría tener implicaciones para su bienestar general y su riesgo de enfermedades cardiovasculares.

De manera similar, la proporción de participantes que reportan días de mala salud física en el último mes también es considerable. El 38.08% de los participantes experimentaron días de mala salud física (25.39% algunos días y 12.69% la mayoría de los días). Esto indica que una parte importante enfrenta problemas de salud física, lo cual podría estar relacionado con la presencia de enfermedades crónicas o factores de riesgo cardiovascular.

Por último, la variable objetivo "CARDIO" revela que el 11.32% de los participantes han tenido alguna enfermedad cardiovascular (infarto de miocardio, enfermedad coronaria o accidente cerebrovascular).

4.1.2 Visualizaciones y gráficos relevantes

En este apartado, se presentan diversas visualizaciones y gráficos como parte del análisis exploratorio de datos. Cada gráfica ha sido diseñada para responder a una pregunta específica sobre el conjunto de datos, con el objetivo de identificar patrones, relaciones y tendencias entre las diferentes variables y la prevalencia de enfermedades cardiovasculares.

A través de estas visualizaciones, se busca obtener una comprensión más profunda de los factores de riesgo y las características asociadas con las enfermedades cardiovasculares. Los gráficos han sido seleccionados para destacar los hallazgos más relevantes y significativos.

¿Cuál es la relación entre la edad y las enfermedades cardiovasculares?

La edad es uno de los factores de riesgo más importantes para las enfermedades cardiovasculares. A medida que las personas envejecen, se producen cambios en el corazón y los vasos sanguíneos que aumentan la probabilidad de desarrollar problemas cardiovasculares. La Figura 4.1 muestra la relación entre la edad y la prevalencia de enfermedad cardiovascular en la población estudiada.

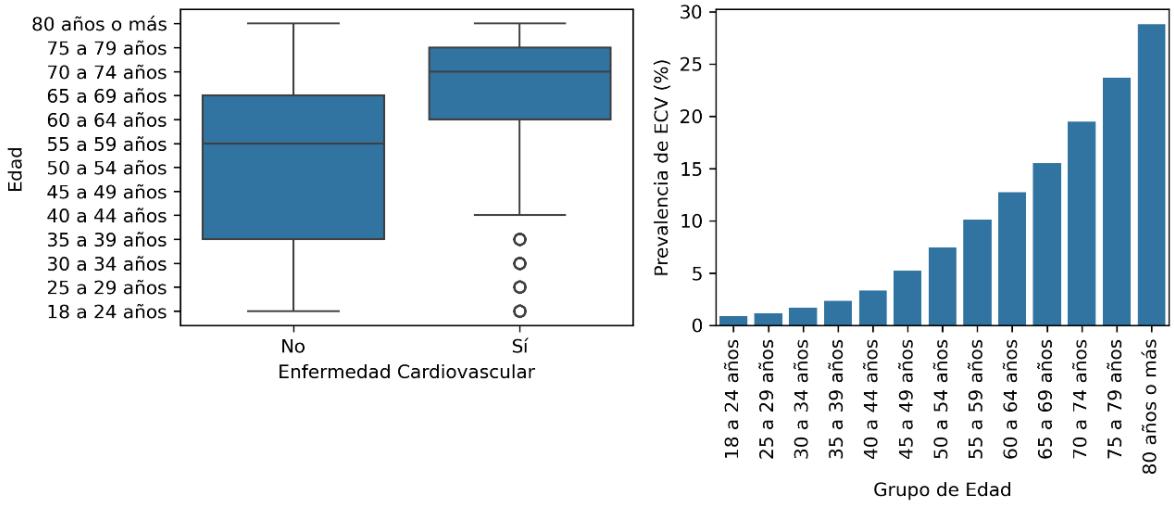


Figura 4.1. Prevalencia de enfermedad cardiovascular por grupo de edad.

La Figura 4.1 muestra un aumento en la prevalencia de enfermedades cardiovasculares con el aumento de la edad. Los grupos de mayor edad presentan una prevalencia significativamente más alta en comparación con los grupos más jóvenes. Este patrón indica que la edad es un factor de riesgo importante para las enfermedades cardiovasculares. Además, esto resalta la importancia de la detección temprana y la gestión de otros factores de riesgo modificables en etapas más tempranas de la vida para mitigar el riesgo a largo plazo.

¿Las personas con obesidad tienen mayor prevalencia de enfermedad cardiovascular que las personas con peso normal?

La obesidad es un factor de riesgo conocido para las enfermedades cardiovasculares, ya que puede contribuir al desarrollo de otros problemas de salud como la hipertensión, la diabetes y los niveles elevados de colesterol. La Figura 4.2 compara la prevalencia de enfermedad cardiovascular entre diferentes categorías de índice de masa corporal (IMC), desde bajo peso hasta obesidad.

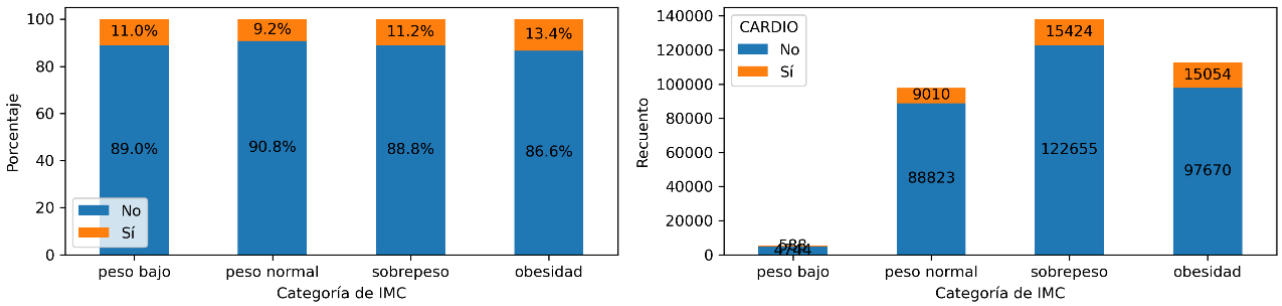


Figura 4.2. Prevalencia de enfermedad cardiovascular por categoría de peso.

La Figura 4.2 muestra la relación entre el índice de masa corporal (IMC) y la prevalencia de enfermedades cardiovasculares. En el gráfico de la izquierda, se observa que la prevalencia de enfermedades cardiovasculares es más alta en las personas con obesidad (13.4%) en comparación con aquellas con peso normal (9.2%) y sobrepeso (11.2%). Este patrón indica que a medida que aumenta el IMC, también lo hace la prevalencia de enfermedades cardiovasculares.

El gráfico de barras apiladas muestra que la prevalencia de enfermedades cardiovasculares aumenta con el IMC, destacando la categoría de obesidad como la más afectada. El gráfico de barras de recuento complementa esta información al mostrar el número absoluto de casos en cada categoría de IMC, lo que permite una evaluación más completa del impacto del IMC, destacando como hay un mayor número de personas con obesidad que con peso normal en el conjunto de datos. Estos datos recalcan que una proporción significativa de la población estudiada tiene un peso corporal excesivo, lo que podría contribuir a una mayor carga de enfermedades cardiovasculares.

¿Hay diferencia en el porcentaje de enfermedad cardiovascular entre hombres y mujeres?

Esta pregunta permite entender cómo el género influye en la prevalencia de enfermedades cardiovasculares. Las diferencias biológicas y de comportamiento entre hombres y mujeres pueden afectar la incidencia y el manejo de estas enfermedades. En la Figura 4.3 se presenta una gráfica que compara la prevalencia de enfermedades cardiovasculares entre hombres y mujeres, proporcionando una visualización clara de cómo el género se relaciona con la salud cardiovascular.

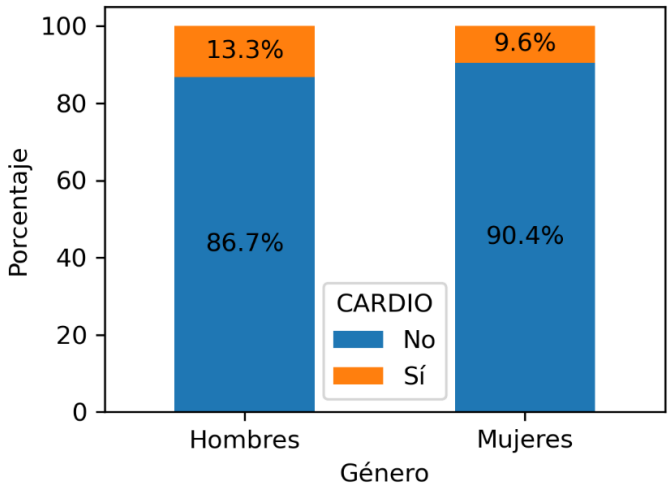


Figura 4.3. Prevalencia de enfermedad cardiovascular por género.

La Figura 4.3 revela una diferencia notable en la prevalencia de enfermedad cardiovascular entre hombres y mujeres. Los hombres presentan una prevalencia del 13.3%, mientras que las mujeres tienen una prevalencia significativamente menor, del 9.6%. Esta diferencia de aproximadamente 3.7 puntos porcentuales sugiere que los hombres tienen un mayor riesgo de desarrollar enfermedades cardiovasculares en comparación con las mujeres. Estos resultados están en línea con la literatura científica existente, que ha identificado diferencias de género en la epidemiología de las enfermedades cardiovasculares [65].

¿Cómo se relaciona el nivel de salud mental (estrés, ansiedad) con la prevalencia de enfermedad cardiovascular?

La salud mental es un aspecto importante a considerar en el contexto de las enfermedades cardiovasculares. Factores como el estrés y la ansiedad pueden tener un impacto significativo en la salud cardiovascular, ya sea de manera directa a través de cambios fisiológicos o indirectamente al influir en los comportamientos de salud. La Figura 4.4 muestra la prevalencia de enfermedad cardiovascular según el nivel de salud mental, clasificado como bueno, regular o malo.

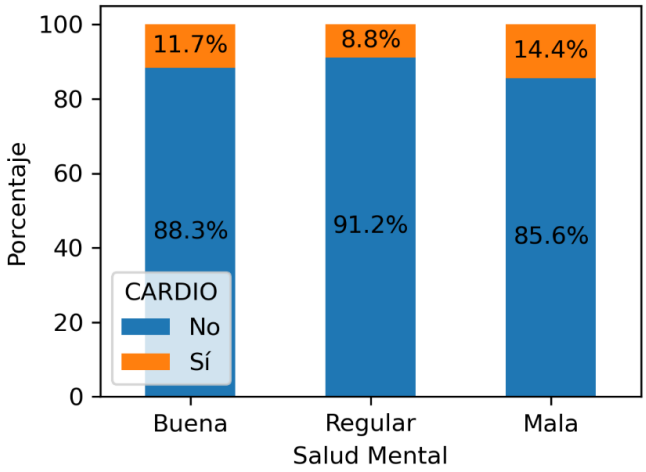


Figura 4.4. Prevalencia de enfermedad cardiovascular según el nivel de salud mental.

La Figura 4.4 revela una relación clara entre el nivel de salud mental y la prevalencia de enfermedad cardiovascular. Las personas con una mala salud mental, que puede incluir niveles altos de estrés y ansiedad, presentan la mayor prevalencia de enfermedad cardiovascular, con un 14.4%.

Estos resultados indican que una mala salud mental, caracterizada por estrés y ansiedad, está asociada con un mayor riesgo de enfermedad cardiovascular. Ya que pueden influir en los comportamientos de salud, como la alimentación poco saludable, el sedentarismo y el tabaquismo, que son factores de riesgo conocidos para las enfermedades cardiovasculares. Las personas con problemas de salud mental también pueden tener dificultades para adherirse a los tratamientos médicos y adoptar estilos de vida saludables.

Es interesante observar que la prevalencia de enfermedad cardiovascular en personas con buena salud mental (11.7%) es ligeramente mayor que en aquellas con salud mental regular (8.8%). Esto puede deberse a la influencia de otros factores de riesgo no capturados en esta gráfica, o a diferencias en la percepción de la salud mental entre los participantes.

¿Fumar se relaciona con mayor prevalencia de enfermedad cardiovascular?

El tabaquismo es un factor de riesgo bien establecido para las enfermedades cardiovasculares. Fumar puede dañar el revestimiento de las arterias, aumentar la inflamación, la coagulación de la sangre y el riesgo de formación de coágulos, lo que contribuye al desarrollo de enfermedades cardíacas y accidentes cerebrovasculares [66]. La Figura 4.5 compara la prevalencia de enfermedad cardiovascular entre fumadores y no fumadores.

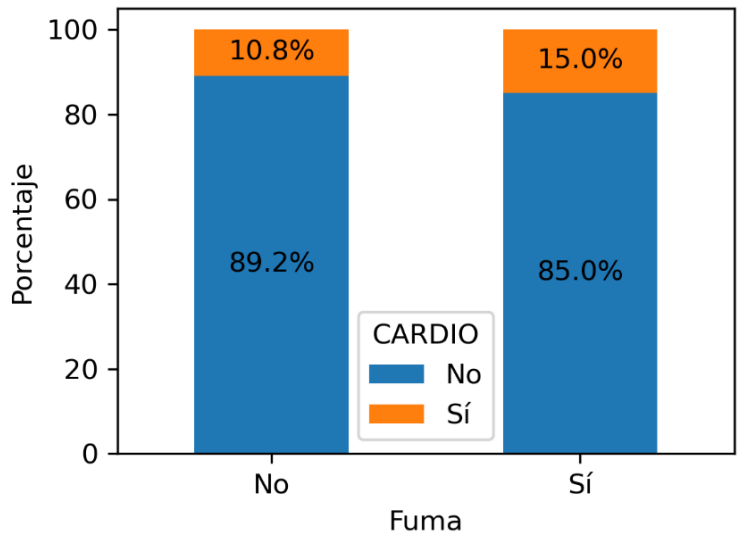


Figura 4.5. Prevalencia de enfermedad cardiovascular según el estado de tabaquismo.

La Figura 4.5 muestra una diferencia notable en la prevalencia de enfermedad cardiovascular entre fumadores y no fumadores. Los fumadores presentan una prevalencia del 15.0%, mientras que los no fumadores tienen una prevalencia significativamente menor, del 10.8%. Esta diferencia de aproximadamente 4.2 puntos porcentuales sugiere que fumar está asociado con un mayor riesgo de enfermedad cardiovascular.

Estos resultados concuerdan con la evidencia científica existente sobre los efectos perjudiciales del tabaquismo en la salud cardiovascular. Los componentes del humo del tabaco, como la nicotina y el monóxido de carbono, pueden aumentar la frecuencia cardíaca, la presión arterial y la demanda de oxígeno del corazón, lo que ejerce una mayor tensión en el sistema cardiovascular. Además, las sustancias químicas del tabaco pueden dañar las paredes de las arterias, promoviendo la acumulación de placa y aumentando el riesgo de aterosclerosis [67].

Es importante destacar que el riesgo cardiovascular asociado con el tabaquismo no se limita solo a los fumadores activos, sino que también afecta a las personas expuestas al humo de segunda mano. La exposición pasiva al humo del tabaco también se ha relacionado con un mayor riesgo de enfermedades cardíacas y accidentes cerebrovasculares.

¿Cómo varía la prevalencia de enfermedad cardiovascular según el nivel educativo y el ingreso?

El nivel educativo y el ingreso son determinantes sociales bien reconocidos de la salud, incluyendo las enfermedades cardiovasculares. Un menor nivel educativo y bajos ingresos a menudo se asocian con un acceso limitado a recursos de salud, mayor exposición a factores de riesgo y adopción de comportamientos menos saludables, lo que puede contribuir a disparidades en la prevalencia de enfermedades cardiovasculares. La Figura 4.6 muestra la prevalencia de enfermedad cardiovascular según el nivel educativo y el rango de ingreso.

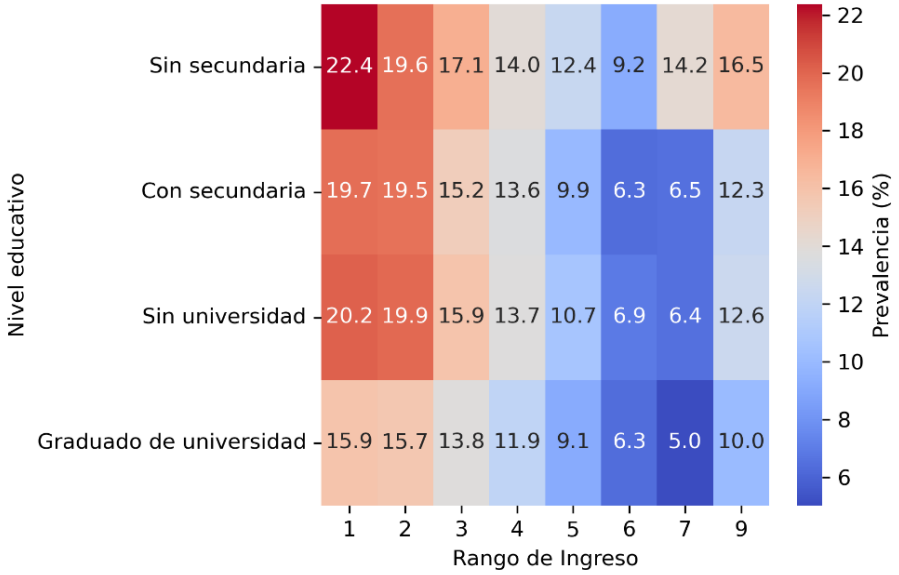


Figura 4.6. Prevalencia de enfermedad cardiovascular según nivel educativo y rango de ingreso.

La Figura 4.6 revela patrones claros en la variación de la prevalencia de enfermedad cardiovascular según el nivel educativo y el ingreso. En general, se observa una mayor prevalencia en los niveles educativos más bajos y los rangos de ingreso inferiores.

Comparando los niveles educativos, aquellos sin educación secundaria tienen la prevalencia más alta en cada rango de ingreso, seguidos por los que tienen educación secundaria. Los individuos sin educación universitaria y los graduados universitarios muestran prevalencias consistentemente más bajas. Por ejemplo, en el rango de ingreso más bajo (1), la prevalencia para aquellos sin secundaria es del 22.4%, en comparación con el 15.9% para los graduados universitarios, una diferencia de 6.5 puntos porcentuales.

Examinando las tendencias según el ingreso, para cada nivel educativo, la prevalencia generalmente disminuye a medida que aumenta el rango de ingreso. Las disparidades son más pronunciadas en los rangos de ingreso inferiores. Tomando el nivel educativo más bajo (sin secundaria), la prevalencia cae del 22.4% en el rango 1 al 9.2% en el rango 6, una reducción de 13.2 puntos. En cambio, para los graduados universitarios, la prevalencia solo disminuye del 15.9% al 6.3% entre esos mismos rangos.

Estos datos indican que tanto el nivel educativo como el ingreso ejercen influencias independientes en la prevalencia de enfermedad cardiovascular, con una interacción donde las disparidades educativas son mayores en los grupos de menores ingresos. La educación puede impactar la salud cardiovascular al moldear el conocimiento sobre la salud, las habilidades y los recursos para adoptar comportamientos saludables.

Mientras tanto, los ingresos afectan el acceso a atención médica de calidad, alimentos nutritivos, vivienda segura y otros recursos que promueven la salud cardiovascular. Desde una perspectiva de equidad en salud, estos resultados subrayan la necesidad de abordar los determinantes sociales subyacentes de las enfermedades cardiovasculares.

Es decir, las personas con menor nivel educativo y menor ingreso pueden tener menos acceso a recursos de salud, menos conocimiento sobre comportamientos saludables y una mayor exposición a factores de riesgo como el tabaquismo y la mala alimentación. Además, el estrés asociado con la inseguridad económica puede contribuir al desarrollo de enfermedades cardiovasculares. La gráfica destaca la importancia de tomar en cuenta las disparidades socioeconómicas en las estrategias de prevención y manejo de enfermedades cardiovasculares.

¿Cómo se relaciona la raza/etnia con la prevalencia de enfermedad cardiovascular según el nivel de ingresos?

La raza/etnia y el nivel socioeconómico son determinantes sociales importantes de la salud, y sus efectos a menudo se entrelazan para dar forma a las disparidades en salud. La raza/etnia puede influir en la salud cardiovascular a través de factores biológicos, socioeconómicos, de comportamiento y ambientales. La Figura 4.7 muestra la prevalencia de enfermedad cardiovascular según la raza/etnia y el rango de ingreso.

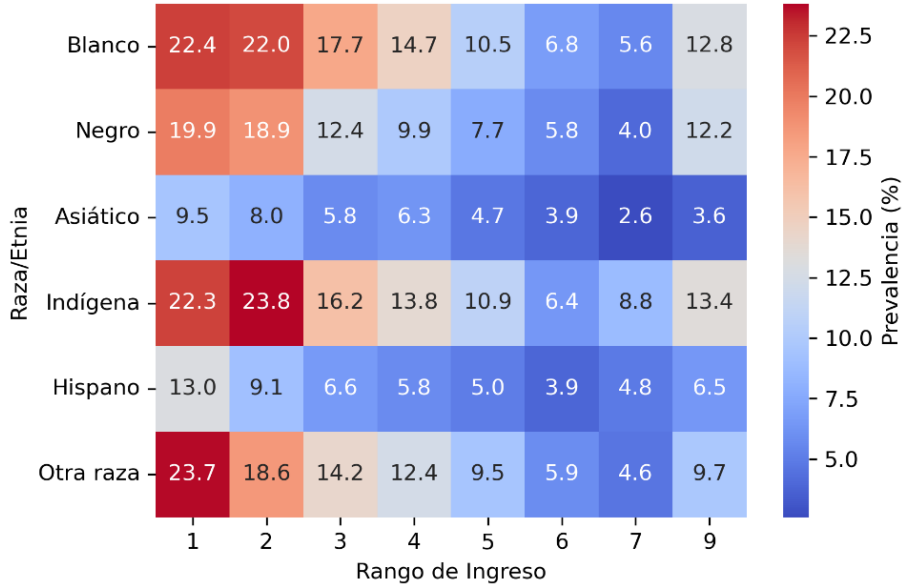


Figura 4.7. Prevalencia de enfermedad cardiovascular según raza/etnia y rango de ingreso.

Al examinar la Figura 4.7, se observan variaciones notables en la prevalencia de enfermedad cardiovascular entre diferentes grupos raciales/étnicos y niveles de ingreso.

En general, las personas negras y las personas indígenas exhiben las prevalencias más altas en casi todos los rangos de ingreso, seguidos por los blancos. Los hispanos y los asiáticos tienden a tener prevalencias más bajas en comparación con otros grupos. Por ejemplo, en el rango de ingreso más bajo (1), los negros tienen una prevalencia del 19.9%, los indígenas del 22.3%, mientras que los hispanos y los asiáticos tienen prevalencias del 13% y 9.5% respectivamente.

Al considerar las tendencias según el ingreso dentro de cada grupo racial/étnico, la prevalencia generalmente disminuye a medida que aumenta el rango de ingreso, pero la magnitud de esta disminución varía. Para los blancos, la prevalencia cae del 22.4% en el rango 1 al 5.6% en el rango

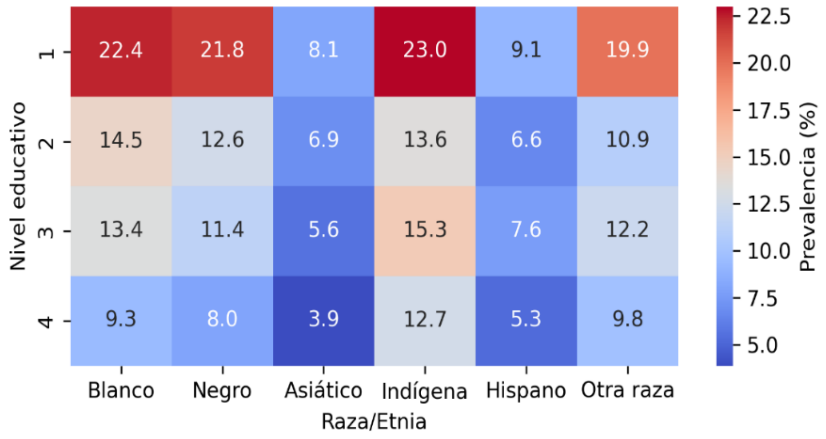
7. En contraste, para los hispanos, la prevalencia solo disminuye del 13% al 4.8% entre esos mismos rangos, una reducción menos pronunciada.

Estos resultados sugieren una compleja interacción entre la raza/etnia y el ingreso en la conformación del riesgo cardiovascular. Si bien un ingreso más alto se asocia con una menor prevalencia dentro de cada grupo racial/étnico, las disparidades entre grupos persisten incluso en los niveles de ingreso más altos. Esto implica que las desventajas socioeconómicas no explican completamente las disparidades raciales/étnicas en la salud cardiovascular.

Varios factores pueden contribuir a estos patrones. Las diferencias en el acceso y la calidad de la atención médica, la discriminación, el estrés crónico, los entornos de vida y trabajo, y los comportamientos de salud moldeados culturalmente pueden desempeñar papeles relevantes. Además, el impacto acumulativo de las desventajas a lo largo de la vida y a través de generaciones puede crear disparidades duraderas. Desde una perspectiva de equidad en salud, estos resultados plantean la necesidad de enfoques que aborden las influencias entrelazadas de la raza/etnia y el estatus socioeconómico. Las políticas de salud pública deberían enfocarse en mejorar el acceso a la atención médica y la educación sobre salud en los grupos socioeconómicamente desfavorecidos y en las minorías raciales/étnicas para reducir la carga de enfermedades cardiovasculares en estas poblaciones.

¿Existe una brecha relevante en la prevalencia de enfermedad cardiovascular entre las personas de diferentes razas/etnias y niveles educativos?

Esta pregunta permite entender cómo los factores socioeconómicos y demográficos interactúan para influir en la salud cardiovascular. Las disparidades en salud entre diferentes grupos raciales y étnicos pueden estar influenciadas por el acceso a recursos, la exposición a factores de riesgo y las diferencias en el comportamiento de salud. Investigaciones anteriores han documentado asociaciones entre estos factores y varias condiciones de salud, incluidas las enfermedades cardiovasculares [68]. De manera que la Figura 4.8 explora la intersección de raza/etnia y educación en la prevalencia de enfermedades cardiovasculares.



.

Figura 4.8. Prevalencia de enfermedad cardiovascular según raza/etnia y nivel educativo.

Al examinar los datos por raza/etnia, se observan disparidades notables. Para cada nivel educativo, los blancos e indígenas consistentemente tienen las prevalencias más altas, mientras que los

asiáticos tienen las más bajas. Por ejemplo, entre aquellos con educación de nivel 1, los blancos tienen una prevalencia del 22.4%, los indígenas del 23%, en contraste con solo 8.1% para los asiáticos. Estas disparidades persisten incluso en los niveles educativos más altos. En el nivel 4, la prevalencia para los blancos es del 9.3%, para los indígenas del 12.7%, pero solo del 3.9% para los asiáticos. Esto sugiere que las disparidades raciales/étnicas en la salud cardiovascular no se explican completamente por las diferencias en educación.

Considerando las diferencias por nivel educativo, se ve un gradiente claro. Para cada raza/etnia, la prevalencia disminuye a medida que aumenta el nivel educativo. Este patrón es más marcado para los blancos, donde la prevalencia cae del 22.4% en el nivel 1 al 9.3% en el nivel 4, una diferencia de más del doble. Sin embargo, la magnitud de este gradiente educativo varía según la raza/etnia.

Estos resultados indican que tanto la raza/etnia como la educación son determinantes importantes de la salud cardiovascular, con efectos que se entrecruzan pero son distintos. Si bien una educación más alta se asocia con una menor prevalencia dentro de cada grupo racial/étnico, no elimina las disparidades entre estos grupos.

Varios factores probablemente contribuyen a estos patrones. Las diferencias en acceso y calidad de la atención médica, discriminación, contextos socioeconómicos y normas culturales en torno a comportamientos de salud pueden explicar estos niveles de prevalencia. Estos hallazgos señalan la importancia de desarrollar estrategias que consideren la interacción entre raza/etnia y educación en la salud cardiovascular. Promover la equidad en la educación es fundamental, pero no suficiente por sí solo. También se necesitan crear entornos que apoyen estilos de vida saludables en diversas comunidades. Las intervenciones deben adaptarse culturalmente, aprovechando las fortalezas y abordando las necesidades específicas de diferentes grupos raciales/étnicos.

¿Hay una combinación preocupante de factores de riesgo de estilo de vida, como fumar, no hacer ejercicio y dormir poco, que se asocie con una prevalencia excepcionalmente alta de enfermedad cardiovascular?

Esta pregunta es importante para entender cómo los comportamientos de salud influyen en la prevalencia de enfermedades cardiovasculares. La American Heart Association ha identificado varios factores de estilo de vida que son esenciales para la salud cardiovascular, incluyendo la actividad física, la exposición a la nicotina y la duración del sueño [69]. A continuación, se presenta la Figura 4.9 que compara la prevalencia de enfermedades cardiovasculares entre diferentes combinaciones de factores de riesgo de estilo de vida, proporcionando una visualización clara de cómo estos comportamientos se relacionan con la salud cardiovascular.

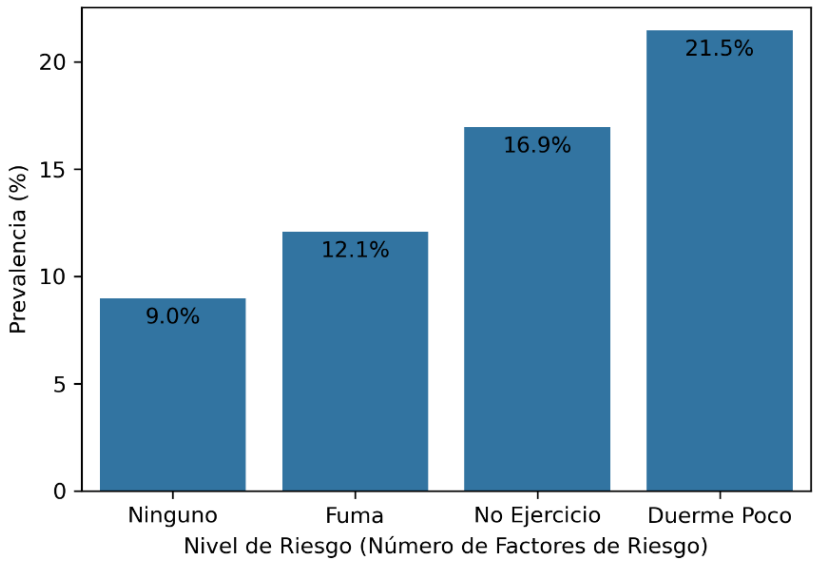


Figura 4.9. Prevalencia de enfermedad cardiovascular según la acumulación de factores de riesgo de estilo de vida.

La Figura 4.9 ilustra cómo la acumulación de factores de riesgo de estilo de vida, como fumar, no hacer ejercicio y dormir poco, se asocia con un aumento importante en la prevalencia de enfermedad cardiovascular.

Cuando una persona no tiene ninguno de estos factores de riesgo (representado por la barra más a la izquierda), la prevalencia de enfermedad cardiovascular es del 9.0%. Sin embargo, al añadir el factor de fumar, la prevalencia aumenta al 12.1%. Esto indica que fumar solo, incluso en ausencia de otros factores de riesgo, puede elevar el riesgo cardiovascular. Al acumular dos factores de riesgo, fumar y no hacer ejercicio, la prevalencia se eleva aún más, alcanzando el 16.9%. Este marcado aumento destaca los efectos aditivos de estos comportamientos poco saludables en la salud cardiovascular. Finalmente, cuando una persona presenta los tres factores de riesgo - fuma, no hace ejercicio y duerme poco - la prevalencia de enfermedad cardiovascular alcanza un significativo 21.5%. Esta observación subraya los graves impactos cardiovasculares de acumular múltiples comportamientos de riesgo.

Estos resultados tienen implicaciones relevantes para la prevención de enfermedades cardiovasculares. Plantean que incluso pequeños cambios de comportamiento, como dejar de fumar o comenzar a hacer ejercicio, pueden reducir apreciablemente el riesgo individual. Además, enfatizan los beneficios cardiovasculares de adoptar un estilo de vida saludable integral que evite la acumulación de múltiples factores de riesgo.

Desde una perspectiva de salud pública, lo anterior señala la importancia de intervenciones y políticas que aborden múltiples comportamientos de riesgo de manera simultánea. Enfoques integrales que promuevan dejar de fumar, aumentar la actividad física y mejorar los patrones de sueño podrían tener un impacto positivo en la reducción de la carga poblacional de enfermedades cardiovasculares.

Al mismo tiempo, es necesario reconocer los determinantes sociales y ambientales más amplios que dan forma a estos comportamientos de riesgo. Abordar las influencias a nivel comunitario y de

políticas, como el acceso a espacios seguros para la actividad física, las normas sociales en torno al tabaquismo y los entornos laborales que promueven un sueño saludable, pueden crear condiciones propicias para elecciones de estilo de vida más saludables.

Por tanto, la gráfica muestra que la acumulación de factores de riesgo de estilo de vida tiene un efecto sinérgico en la prevalencia de enfermedades cardiovasculares. Cada factor de riesgo adicional aumenta significativamente la probabilidad de desarrollar enfermedades cardiovasculares. La falta de sueño, la inactividad física y el tabaquismo son factores de riesgo significativos que, cuando se combinan, elevan considerablemente el riesgo de enfermedades cardiovasculares.

4.2. Evaluación de modelos predictivos

4.2.1 Resultados para el conjunto de datos con 11 variables

Comparación de técnicas de muestreo

En esta sección, se presenta una comparación de las técnicas de muestreo aplicadas al conjunto de datos con 11 variables. Para evaluar el rendimiento de cada técnica de muestreo (9), se entrenaron 8 modelos diferentes con los mejores hiperparámetros encontrados para cada combinación de técnica de muestreo y algoritmo de aprendizaje. Los algoritmos utilizados en este análisis son; **LR**, **GNB**, **Light GBoost**, **RF**, **ANN**, **Easy Ensemble**, **Voting** y **LinearSVC**. Las métricas utilizadas para evaluar el rendimiento de los modelos incluyen precision, recall, F1-score, AUC-PR y AUC-ROC, que son especialmente relevantes para conjuntos de datos desbalanceados.

La Tabla 4.2 muestra los promedios de las métricas de rendimiento para los 8 modelos entrenados con cada técnica de muestreo, así como el tiempo de muestreo correspondiente. Estos promedios proporcionan una visión general del rendimiento de cada técnica de muestreo en todos los modelos, lo que permite una comparación justa y concisa de las técnicas aplicadas.

Tabla 4.2. Comparación de técnicas de muestreo para el conjunto de datos con 11 variables.

Técnica de Muestreo	Recall	Precisión	F1-score	AUC-ROC	AUC-PR	Tiempo de muestreo (min)	Tiempo búsqueda hiperparámetros (min)
RandomUnderSampler	0.7615	0.2426	0.3670	0.8020	0.3387	0.00	49.01
NearMiss	0.7331	0.2256	0.3473	0.7472	0.2475	0.04	48.65
InstanceHardnessThreshold	0.9257	0.1625	0.2763	0.7419	0.2775	0.38	60.15
NeighbourhoodCleaningRule	0.4218	0.3275	0.3484	0.7946	0.3150	1.69	169.90
RandomeOverSampler	0.7557	0.2444	0.3702	0.8001	0.3337	0.00	355.71
SMOTEN	0.7289	0.2373	0.3583	0.7837	0.3200	15.56	559.86
BorderlineSMOTE	0.7486	0.2377	0.3595	0.7920	0.3212	2.34	558.87
SMOTEENN	0.3526	0.3188	0.3256	0.7431	0.2837	1.97	170.90
SMOTETomek	0.7558	0.2401	0.3644	0.7951	0.3300	4.12	414.25

* El resultado de las métricas es el promedio de los 8 modelos.

Para complementar la información presentada en la Tabla 4.2, se ha elaborado la Figura 4.10, que muestra una comparación visual de las métricas de rendimiento seleccionadas (F1-score y AUC-ROC) junto con el tiempo de búsqueda de hiperparámetros para cada técnica de muestreo aplicada al conjunto de datos con 11 variables.

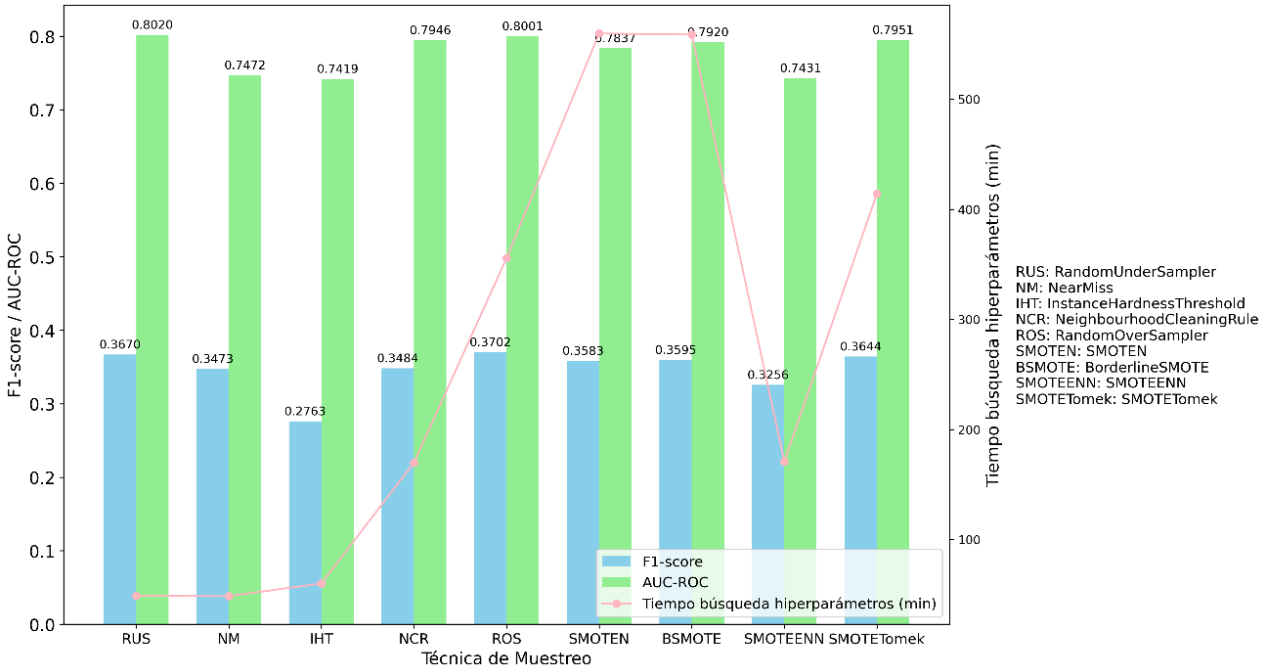


Figura 4.10. Métricas de rendimiento y tiempos por técnica de muestreo (11 variables).

Al examinar la Tabla 4.2 y la Figura 4.10, se pueden extraer varias conclusiones sobre el rendimiento y la eficiencia computacional de las diferentes técnicas de muestreo aplicadas al conjunto de datos con 11 variables. Ambas representaciones de los resultados se complementan para proporcionar una visión completa de las fortalezas y debilidades de cada técnica.

En términos de métricas de rendimiento, tanto la tabla como la figura destacan que las técnicas de sobremuestreo (RandomOverSampler, SMOTEN, BorderlineSMOTE y SMOTETomek) y algunas técnicas de submuestreo (RandomUnderSampler y NearMiss) obtienen valores relativamente altos de Recall y AUC-ROC. Esto indica su capacidad para identificar correctamente la clase minoritaria, en un contexto de desequilibrio de clases. Sin embargo, estas técnicas también presentan valores bajos de Precisión, lo que sugiere que clasifican incorrectamente una proporción considerable de la clase mayoritaria.

Por otro lado, técnicas como InstanceHardnessThreshold y NeighbourhoodCleaningRule muestran un comportamiento opuesto, con valores de Recall y Precisión desequilibrados, lo que indica un sobreajuste hacia una de las clases.

Al considerar las métricas F1-score y AUC-ROC, que proporcionan una evaluación más equilibrada del rendimiento, RandomUnderSampler, RandomOverSampler y SMOTETomek se destacan como las técnicas más efectivas.

Sin embargo, el rendimiento no es el único factor a tener en cuenta al seleccionar una técnica de muestreo adecuada. La eficiencia computacional, representada por los tiempos de muestreo y búsqueda de hiperparámetros, también es importante, especialmente al trabajar con conjuntos de datos grandes como el que se está analizando (alrededor de 350 mil registros).

En este aspecto, RandomUnderSampler se destaca por sus tiempos de muestreo insignificantes y un tiempo de búsqueda de hiperparámetros razonable (49.01 min), mientras que mantiene un buen

equilibrio entre las métricas de rendimiento. RandomOverSampler también ofrece un buen rendimiento, pero su tiempo de búsqueda de hiperparámetros es significativamente mayor (355.71 min), lo que puede ser un factor limitante en términos de recursos computacionales y tiempo.

Otras técnicas como SMOTEN, BorderlineSMOTE y SMOTETomek, aunque ofrecen valores altos de F1-score y AUC-ROC, tienen tiempos de búsqueda de hiperparámetros considerablemente mayores, lo que implica un mayor coste computacional.

En consecuencia, tras analizar tanto la Tabla 4.2 como la Figura 4.10, RandomUnderSampler resulta la técnica de muestreo más adecuada para este conjunto de datos desbalanceado, ya que ofrece un buen equilibrio entre métricas de rendimiento y eficiencia computacional. RandomOverSampler también es una opción viable si se prioriza el rendimiento sobre el coste computacional. Las técnicas de sobremuestreo como SMOTEN, BorderlineSMOTE y SMOTETomek, aunque efectivas en términos de rendimiento, pueden ser menos prácticas debido a sus altos tiempos de búsqueda de hiperparámetros.

Comparación de modelos predictivos

Tras seleccionar RandomUnderSampler como la técnica de muestreo más adecuada para el conjunto de datos con 11 variables, se procede a evaluar el rendimiento de los 8 modelos predictivos entrenados con esta técnica. La Tabla 4.3 presenta las métricas de rendimiento para cada modelo, incluyendo Precision, Recall, F1-score, AUC-ROC y AUC-PR, junto con el tiempo de entrenamiento y el tiempo de búsqueda de hiperparámetros.

Tabla 4.3. Métricas de rendimiento de los modelos predictivos entrenados con RandomUnderSampler (11 variables).

Modelo	Recall	Precisión	F1-score	AUC-ROC	AUC-PR	Tiempo de entrenamiento (min)	Tiempo búsqueda hiperparámetros (min)
Logistic Regression	0.7661	0.2437	0.3697	0.8034	0.34	0.01	1.84
Gaussian Naive Bayes	0.6251	0.2680	0.3752	0.7874	0.32	0.00	0.07
Light GBoost	0.7919	0.2366	0.3644	0.8061	0.35	0.01	1.22
Random Forest	0.7954	0.2347	0.3625	0.8051	0.34	0.06	4.14
Neural Network	0.8135	0.2295	0.3580	0.8060	0.34	2.87	34.86
Easy Ensemble	0.7538	0.2424	0.3669	0.8001	0.34	0.20	3.17
Voting	0.7542	0.2502	0.3757	0.8046	0.34	2.45	-
LinearSVC	0.7916	0.2356	0.3632	0.8025	0.34	0.04	3.71

Para complementar la información presentada en la Tabla 4.3 y obtener una visión más completa del rendimiento de los modelos predictivos entrenados con la técnica de muestreo RandomUnderSampler y el conjunto de datos con 11 variables, se ha elaborado la Figura 4.11. Esta figura muestra las curvas AUC-ROC y AUC-PR para cada modelo, permitiendo una comparación visual de su capacidad para discriminar entre las clases y manejar el desequilibrio presente en los datos. Las curvas AUC-ROC y AUC-PR son herramientas valiosas para evaluar el rendimiento de los modelos de clasificación. La curva AUC-ROC muestra la capacidad del modelo para discriminar entre las clases, mientras que la curva AUC-PR se enfoca en la habilidad del modelo para identificar correctamente la clase minoritaria, lo cual es especialmente relevante en conjuntos de datos desbalanceados. Un modelo con valores más altos de AUC-ROC y AUC-PR indica un mejor rendimiento general.

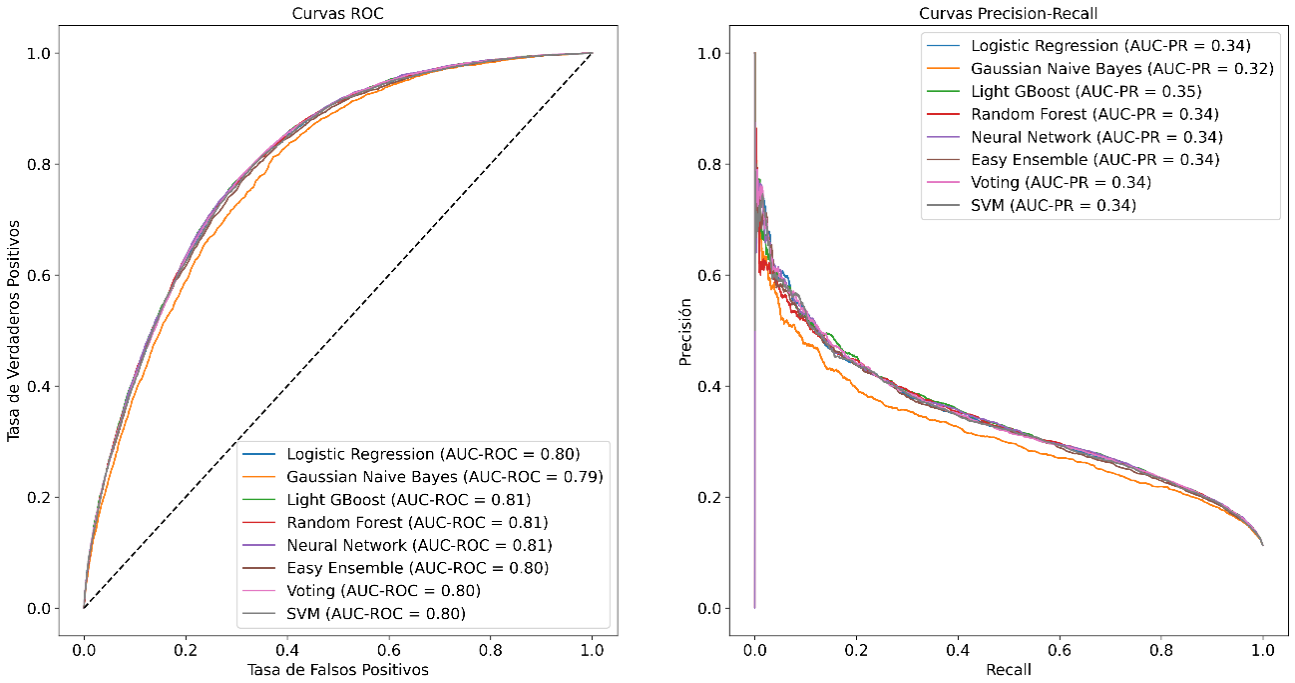


Figura 4.11. Curvas AUC-ROC y AUC-PR para los modelos predictivos entrenados con RandomUnderSampler (11 variables).

Al examinar los resultados presentados en la Tabla 4.3 y la Figura 4.11, se pueden identificar los modelos predictivos que ofrecen el mejor rendimiento para el conjunto de datos con 11 variables, considerando tanto las métricas de rendimiento como el tiempo de entrenamiento y búsqueda de hiperparámetros.

En términos de Recall, se observa que Neural Network obtiene el valor más alto (0.8135), seguido de cerca por Random Forest y Light GBoost (0.7954 y 0.7919, respectivamente). Esto indica su capacidad para identificar correctamente una mayor proporción de instancias de la clase minoritaria.

En cuanto a la Precisión, Gaussian Naive Bayes y Voting presentan los valores más altos (0.2680 y 0.2502, respectivamente), lo que sugiere que una mayor proporción de las instancias clasificadas como positivas por estos modelos son realmente positivas. Sin embargo, sus valores de Recall son relativamente bajos en comparación con otros modelos.

Al considerar el F1-score, que combina Precision y Recall, Voting y Gaussian Naive Bayes obtienen los valores más altos (0.3757 y 0.3752, respectivamente), seguidos de cerca por Logistic Regression (0.3697) e Easy Ensemble (0.3669).

En términos de AUC-ROC y AUC-PR, las diferencias entre los modelos son menos pronunciadas, con valores que oscilan entre 0.7874 y 0.8135 para AUC-ROC, y entre 0.32 y 0.35 para AUC-PR. Esto indica que todos los modelos tienen una capacidad similar para discriminar entre las clases.

Al examinar los tiempos de entrenamiento y búsqueda de hiperparámetros, se observan diferencias significativas entre los modelos. Neural Network tiene el tiempo de entrenamiento más largo (2.87 minutos) y el tiempo de búsqueda de hiperparámetros más extenso (34.86 minutos). Por otro lado,

Gaussian Naive Bayes tiene los tiempos más cortos tanto para entrenamiento como para búsqueda de hiperparámetros (0.00 y 0.07 minutos, respectivamente).

Teniendo en cuenta el equilibrio entre las métricas de rendimiento y los tiempos de entrenamiento y búsqueda de hiperparámetros, Random Forest y Light GBoost son las mejores opciones. Ambos modelos obtienen valores altos de Recall (0.7954 y 0.7919, respectivamente) y valores competitivos de Precisión, F1-score, AUC-ROC y AUC-PR. Además, sus tiempos de entrenamiento y búsqueda de hiperparámetros son relativamente bajos en comparación con otros modelos de alto rendimiento, como Neural Network y Voting.

En conclusión, considerando el análisis realizado y los resultados obtenidos, se considera a Light GBoost como el modelo final para el conjunto de datos con 11 variables. Aunque Random Forest y Light GBoost tienen un rendimiento similar en términos de Recall, Precisión, F1-score y AUC-ROC, Light GBoost presenta una ligera ventaja en cuanto a AUC-PR (0.35 frente a 0.34 de Random Forest). Además, Light GBoost tiene tiempos de entrenamiento y búsqueda de hiperparámetros significativamente menores en comparación con Random Forest (0.01 minutos y 1.22 minutos para Light GBoost, frente a 0.06 minutos y 4.14 minutos para Random Forest, respectivamente). Estos tiempos de entrenamiento y búsqueda de hiperparámetros más cortos hacen que Light GBoost sea más eficiente computacionalmente, lo que puede ser beneficioso cuando se trabaja con conjuntos de datos grandes o cuando se requiere un modelo que pueda ser entrenado y ajustado rápidamente. Por lo tanto, Light GBoost se presenta como la opción más equilibrada y práctica para el conjunto de datos con 11 variables, ofreciendo un buen rendimiento predictivo y una mayor eficiencia computacional.

Para finalizar el análisis del modelo Light GBoost seleccionado para el conjunto de datos con 11 variables, se presentan la matriz de confusión en la Tabla 4.4 y el reporte de clasificación en la Tabla 4.5 obtenidos al evaluar el modelo sobre el conjunto de datos de prueba usando la técnica de muestreo RandomUnderSampler en el conjunto de entrenamiento. Además, se incluyen los mejores hiperparámetros encontrados durante el proceso de ajuste del modelo.

Mejores hiperparámetros para Light GBoost:
{'class_weight': None, 'colsample_bytree': 0.7494566776799121, 'learning_rate': 0.01478556993253947, 'max_depth': 18, 'n_estimators': 205, 'num_leaves': 41, 'subsample': 0.8076987955514452, 'verbose': -1}

Tabla 4.4. Matriz de confusión para el modelo Light GBoost (11 variables).

	Predicción Negativa	**Predicción Positiva**
Negativo	31,726	15,358
Positivo	1,251	4,761

La matriz de confusión muestra el número de instancias clasificadas correcta e incorrectamente por el modelo Light GBoost en el conjunto de datos de prueba. Las filas representan las clases reales, mientras que las columnas representan las clases predichas por el modelo.

Tabla 4.5. Reporte de clasificación para el modelo Light GBoost (11 variables).

	Precision	Recall	F1-score	Support
0	0.96	0.67	0.79	47,084
1	0.24	0.79	0.36	6,012
Accuracy			0.69	53,096
Macro avg	0.60	0.73	0.58	53,096
Weighted avg	0.88	0.69	0.74	53,096

El reporte de clasificación proporciona un desglose detallado de las métricas de rendimiento del modelo Light GBoost para cada clase, así como los promedios macro y ponderado.

La matriz de confusión brinda una visión clara de cómo el modelo Light GBoost está clasificando las instancias en el conjunto de datos de prueba. En este caso, se está trabajando con un problema de clasificación binaria, donde la clase negativa (0) representa a las personas que no desarrollaron enfermedades cardiovasculares, mientras que la clase positiva (1) representa a aquellas que sí las desarrollaron.

Al observar la matriz de confusión, se puede ver que el modelo ha clasificado correctamente una gran cantidad de instancias negativas (31,726) y una proporción considerable de instancias positivas (4,761). Esto indica que el modelo es capaz de identificar correctamente a muchas personas que no desarrollarán enfermedades cardiovasculares, lo cual es un aspecto positivo.

Sin embargo, también se nota que hay un número significativo de falsos positivos (15,358), es decir, personas que fueron clasificadas como propensas a desarrollar enfermedades cardiovasculares, pero que en realidad no las desarrollaron. Esto sugiere que el modelo puede estar sobreestimando el riesgo en algunos casos. Por otro lado, el número de falsos negativos (1,251) es relativamente bajo, lo que significa que el modelo no está pasando por alto a muchas personas que realmente desarrollarán enfermedades cardiovasculares.

Ahora, se analiza el reporte de clasificación para obtener una visión más detallada del rendimiento del modelo. La precisión (precision) para la clase positiva (1) es de 0.24, lo que indica que, de todas las instancias que el modelo clasificó como positivas, solo el 24% realmente pertenecen a esa clase. Esto está directamente relacionado con el alto número de falsos positivos que se observan en la matriz de confusión.

Por otro lado, la sensibilidad (recall) para la clase positiva es de 0.79, lo que significa que el modelo ha identificado correctamente el 79% de las instancias positivas reales. Esto es bastante bueno y sugiere que el modelo es capaz de detectar una gran proporción de las personas que realmente desarrollarán enfermedades cardiovasculares.

El puntaje F1 (f1-score) para la clase positiva es de 0.36, lo cual es una medida que combina la precisión y la sensibilidad. Un puntaje F1 más alto indica un mejor equilibrio entre estas dos métricas.

Es importante tener en cuenta que el conjunto de datos original está muy desbalanceado, con una proporción mucho mayor de instancias negativas que positivas. Esto puede dificultar la tarea del modelo para aprender a clasificar correctamente la clase minoritaria (positiva). A pesar de esto, el modelo Light GBoost ha logrado un recall alto para la clase positiva, lo que es relevante.

En consecuencia, el modelo Light GBoost ha demostrado ser capaz de identificar correctamente a una gran proporción de las personas que no desarrollarán enfermedades cardiovasculares, al tiempo que detecta a la mayoría de las personas que sí las desarrollarán. Sin embargo, el modelo tiende a sobreestimar el riesgo en algunos casos, lo que resulta en un número considerable de falsos positivos. Dada la naturaleza crítica de este problema, donde es importante no pasar por alto a las personas en riesgo, el alto recall para la clase positiva es un aspecto especialmente valioso.

Además de la evaluación del rendimiento, en un apartado posterior de esta memoria se abordará la explicabilidad del modelo final seleccionado, ya sea el modelo entrenado con las 11 variables o el modelo entrenado con todas las variables disponibles. Se identificarán las variables más relevantes que influyen en las predicciones del modelo, lo cual es necesario para comprender cómo el modelo toma sus decisiones y para identificar los factores de riesgo más importantes en el desarrollo de enfermedades cardiovasculares.

4.2.2 Resultados para el conjunto de datos con todas las variables

Comparación de técnicas de muestreo

En esta sección, se presenta una comparación de las técnicas de muestreo aplicadas al conjunto de datos completo, que incluye las 29 variables disponibles. Al igual que en el análisis realizado con el conjunto de datos de 11 variables, se entrenaron 8 modelos diferentes utilizando los mejores hiperparámetros encontrados para cada combinación de técnica de muestreo y algoritmo de aprendizaje. Los algoritmos empleados son los mismos que en la sección anterior: **LR**, **GNB**, **Light GBoost**, **RF**, **ANN**, **Easy Ensemble**, **Voting** y **LinearSVC**. Las métricas de rendimiento evaluadas también son las mismas: precision, recall, F1-score, AUC-PR y AUC-ROC.

La Tabla 4.6 muestra los promedios de las métricas de rendimiento para los 8 modelos entrenados con cada técnica de muestreo aplicada al conjunto de datos completo, junto con el tiempo de muestreo y tiempo de búsqueda de parámetros correspondiente. Estos promedios permiten comparar las técnicas de muestreo aplicadas, proporcionando una visión general del rendimiento de cada técnica en todos los modelos.

Tabla 4.6. Comparación de técnicas de muestreo para el conjunto de datos con todas las variables (29).

Técnica de Muestreo	Recall	Precisión	F1-score	AUC-ROC	AUC-PR	Tiempo de muestreo (min)	Tiempo búsqueda hiperparámetros (min)
RandomUnderSampler	0.7671	0.2660	0.3934	0.8252	0.3800	0.00	54.06
NearMiss	0.8272	0.1482	0.2513	0.6985	0.2425	0.63	54.06
InstanceHardnessThreshold	0.9271	0.1782	0.2988	0.8012	0.3250	1.15	93.83
NeighbourhoodCleaningRule	0.5417	0.3547	0.4032	0.8244	0.3787	21.56	195.53
RandomeOverSampler	0.7407	0.2700	0.3934	0.8157	0.3625	0.00	492.23
SMOTEN	0.5423	0.2674	0.3563	0.7660	0.3037	45.03	701.98
BorderlineSMOTE	0.6897	0.2357	0.3509	0.7738	0.2950	31.96	562.22
SMOTEENN	0.8238	0.2290	0.3582	0.8076	0.3412	44.82	419.84
SMOTETomek	0.7100	0.2338	0.3499	0.7768	0.2962	87.23	588.84

* El resultado de las métricas es el promedio de los 8 modelos.

Para complementar la información presentada en la Tabla 4.6, se ha elaborado la Figura 4.12, que muestra una comparación visual de las métricas de rendimiento seleccionadas (F1-score y AUC-ROC) junto con el tiempo de búsqueda de hiperparámetros para cada técnica de muestreo aplicada al conjunto de datos completo, que incluye las 29 variables.

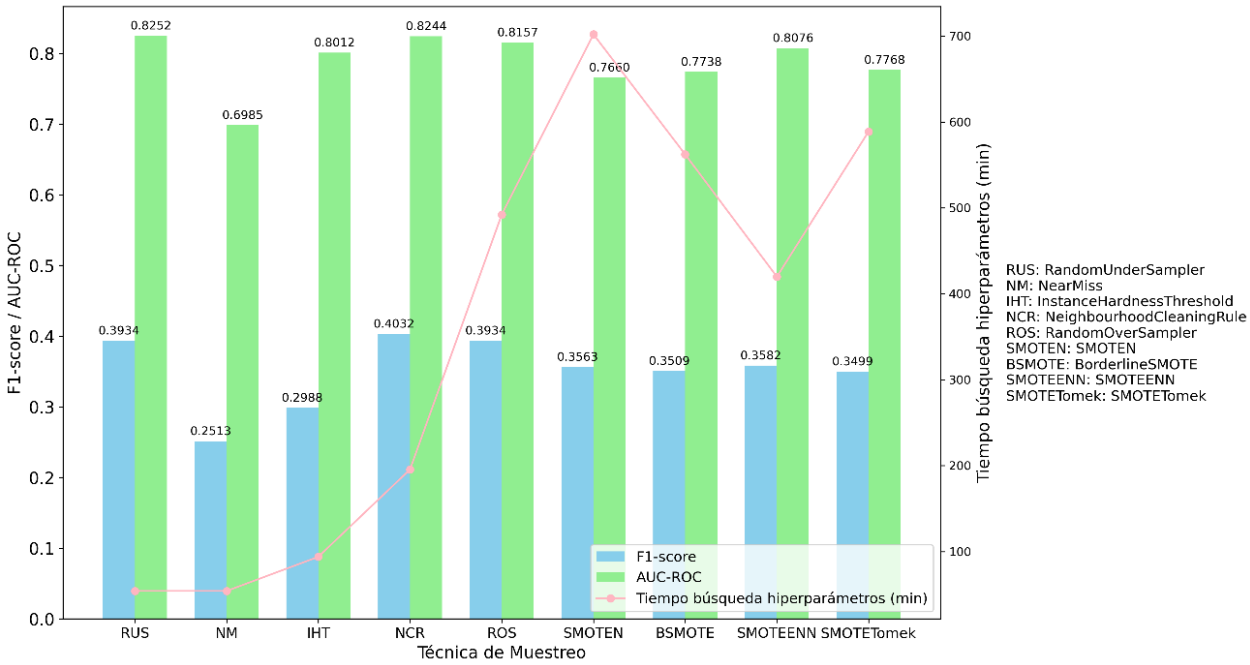


Figura 4.12. Métricas de rendimiento y tiempos por técnica de muestreo (29 variables).

La Tabla 4.6 y la Figura 4.12 ofrecen una perspectiva completa sobre el rendimiento y la eficiencia computacional de las técnicas de muestreo cuando se aplican al conjunto de datos completo, que incluye las 29 variables disponibles. Estos resultados permiten evaluar el impacto de considerar un conjunto más amplio de características en la capacidad predictiva y el coste computacional de los modelos. La combinación de la información presentada en ambos formatos facilita una comprensión integral de los puntos fuertes y las limitaciones de cada técnica en este contexto específico.

En términos de métricas de rendimiento, tanto la tabla como la figura destacan que RandomUnderSampler, NeighbourhoodCleaningRule y SMOTEENN obtienen valores relativamente altos de AUC-ROC (superiores a 0.80) y un buen equilibrio entre Precision y Recall, reflejado en valores de F1-score más altos en comparación con las demás técnicas. Esto indica su capacidad para identificar correctamente la clase minoritaria en un contexto de desequilibrio de clases, manteniendo un buen desempeño general.

Por otro lado, técnicas como NearMiss e InstanceHardnessThreshold, a pesar de tener valores altos de Recall, presentan una Precisión baja, lo que resulta en valores inferiores de F1-score y AUC-PR. Esto sugiere un sobreajuste hacia la clase minoritaria y una clasificación incorrecta de una proporción considerable de la clase mayoritaria.

Al considerar la eficiencia computacional, representada por los tiempos de muestreo y búsqueda de hiperparámetros, RandomUnderSampler se destaca por su eficiencia, ya que no requiere tiempo adicional para el muestreo y presenta uno de los tiempos más bajos de búsqueda de hiperparámetros (54.06 minutos). En contraste, técnicas como SMOTEN, BorderlineSMOTE y SMOTEENN tienen

tiempos de muestreo y búsqueda de hiperparámetros considerablemente más altos, lo que implica un mayor coste computacional.

No obstante, tras examinar detenidamente la Tabla 4.6 y la Figura 4.12, se puede concluir que InstanceHardnessThreshold y SMOTEENN son las dos técnicas de muestreo más adecuadas para el conjunto de datos con todas las variables, teniendo en cuenta que el objetivo es predecir la presencia de una enfermedad cardiovascular en un conjunto de datos altamente desbalanceado.

InstanceHardnessThreshold se destaca por su alta capacidad predictiva, reflejada en un valor de Recall de 0.9271, lo que indica su habilidad para identificar correctamente a los pacientes con la enfermedad. Aunque su Precisión es relativamente baja (0.1782), en el contexto de una enfermedad cardiovascular, especialmente en pacientes mayores de 55 años, podría ser aceptable tener un número elevado de falsos positivos a cambio de no pasar por alto casos reales de la enfermedad. El tiempo de búsqueda de hiperparámetros de InstanceHardnessThreshold (93.83 minutos) es razonable considerando su rendimiento.

Por otro lado, SMOTEENN ofrece un equilibrio más balanceado entre Precision (0.2290) y Recall (0.8238), lo que resulta en un modelo más robusto y menos propenso a sesgos hacia una clase en particular. Además, SMOTEENN obtiene valores altos de AUC-ROC (0.8076) y AUC-PR (0.3412), lo que indica un buen rendimiento general del modelo. Aunque su tiempo de búsqueda de hiperparámetros es considerablemente mayor (419.84 minutos) en comparación con técnicas como RandomUnderSampler, este coste computacional adicional puede justificarse por su capacidad para generar un modelo más equilibrado y confiable.

Finalmente, considerando la importancia de obtener un modelo confiable y equilibrado para la predicción de enfermedades cardiovasculares, se selecciona SMOTEENN como la técnica de muestreo final para el conjunto de datos con todas las variables. A pesar de su mayor tiempo de procesamiento, SMOTEENN proporciona un modelo más fuerte y menos sesgado.

Comparación de modelos predictivos

Tras seleccionar SMOTEENN como la técnica de muestreo más adecuada para el conjunto de datos con todas las variables (29), se procede a evaluar el rendimiento de los 8 modelos predictivos entrenados con esta técnica. La Tabla 4.7 presenta las métricas de rendimiento para cada modelo, incluyendo Precision, Recall, F1-score, AUC-ROC y AUC-PR, junto con el tiempo de entrenamiento y el tiempo de búsqueda de hiperparámetros.

Tabla 4.7. Métricas de rendimiento de los modelos predictivos entrenados con SMOTEENN (29 variables).

Modelo	Recall	Precisión	F1-score	AUC-ROC	AUC-PR	Tiempo de entrenamiento (min)	Tiempo búsqueda hiperparámetros (min)
Logistic Regression	0.8515	0.2207	0.3506	0.8108	0.35	0.01	17.88
Gaussian Naive Bayes	0.8084	0.2299	0.3580	0.8046	0.34	0.00	0.86
Light GBoost	0.8450	0.2287	0.3599	0.8139	0.35	0.03	8.59
Random Forest	0.8084	0.2344	0.3634	0.8034	0.32	1.71	74.26
Neural Network	0.7658	0.2372	0.3622	0.7897	0.30	14.73	257.40
Easy Ensemble	0.8413	0.2252	0.3553	0.8112	0.35	3.12	41.43
Voting	0.8204	0.2353	0.3657	0.8147	0.36	12.72	-
LinearSVC	0.8495	0.2206	0.3502	0.8115	0.35	0.02	19.42

Para complementar la información presentada en la Tabla 4.7 y obtener una visión más completa del rendimiento de los modelos predictivos entrenados con la técnica de muestreo SMOTEENN y el conjunto de datos con todas las variables (29), se ha elaborado la Figura 4.13. Esta figura muestra las curvas AUC-ROC y AUC-PR para cada modelo, permitiendo una comparación visual de su capacidad para discriminar entre las clases y manejar el desequilibrio presente en los datos.

La Figura 4.13 presenta dos gráficos que permiten evaluar y comparar el desempeño de los modelos predictivos desde diferentes perspectivas. El primer gráfico muestra las curvas AUC-ROC, que indican la capacidad de cada modelo para discriminar entre las clases a medida que se ajusta el umbral de clasificación. Un modelo con una curva AUC-ROC más cercana a la esquina superior izquierda del gráfico se considera mejor, ya que maximiza la tasa de verdaderos positivos mientras minimiza la tasa de falsos positivos. El segundo gráfico presenta las curvas AUC-PR, que son especialmente útiles cuando se trabaja con conjuntos de datos desbalanceados. Estas curvas muestran la relación entre la precisión y la sensibilidad (recall) para diferentes umbrales de clasificación, y un modelo con una curva AUC-PR más alta indica un mejor rendimiento en la identificación de la clase minoritaria.

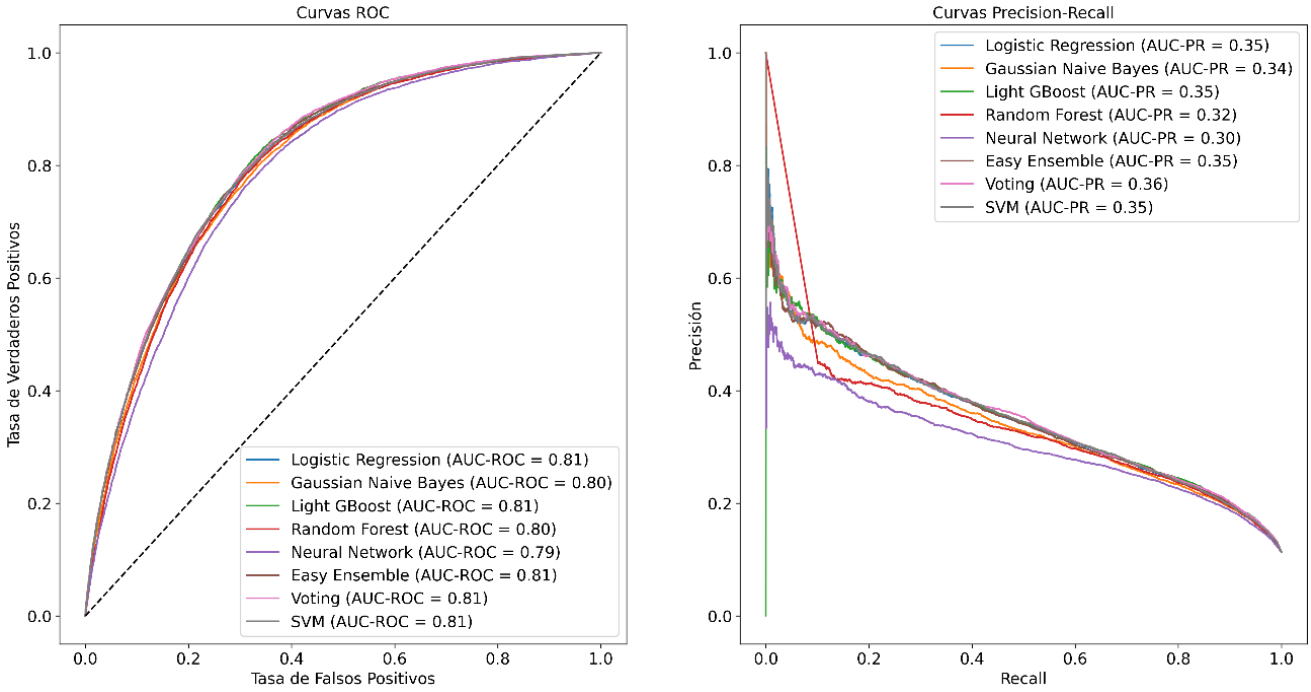


Figura 4.13. Curvas AUC-ROC y AUC-PR para los modelos predictivos entrenados con SMOTEENN (29 variables).

Analizando los resultados presentados en la Tabla 4.7 y las curvas ROC y Precision-Recall de la Figura 4.13, se pueden extraer varias conclusiones sobre el rendimiento de los modelos predictivos entrenados con la técnica de muestreo SMOTEENN y el conjunto de datos con todas las variables (29).

En términos de métricas de rendimiento, se observa que la mayoría de los modelos alcanzan valores similares de AUC-ROC, oscilando entre 0.79 y 0.81. Esto indica que, en general, los modelos tienen una buena capacidad para discriminar entre las clases. Sin embargo, al examinar las métricas de Precision, Recall y F1-score, se evidencian algunas diferencias entre los modelos.

Los modelos Logistic Regression, Light GBoost, Easy Ensemble y LinearSVC obtienen los valores más altos de Recall (superiores a 0.84), lo que sugiere una mayor capacidad para identificar correctamente la clase minoritaria. No obstante, estos modelos también presentan valores de Precision relativamente bajos (alrededor de 0.22), lo que implica una tasa elevada de falsos positivos.

Por otro lado, Random Forest, Neural Network y Voting muestran un mejor equilibrio entre Precision y Recall, con valores de F1-score ligeramente superiores a los demás modelos. Esto indica que estos modelos logran un compromiso más adecuado entre la identificación de la clase minoritaria y la minimización de los falsos positivos.

Al examinar las curvas ROC y Precision-Recall, se confirma que los modelos tienen un rendimiento similar en términos de AUC-ROC, con curvas muy próximas entre sí. Sin embargo, en las curvas Precision-Recall, se observa que Logistic Regression, Gaussian Naive Bayes, Light GBoost, Easy Ensemble y LinearSVC presentan un mejor desempeño, con valores de AUC-PR de 0.35.

Considerando los tiempos de entrenamiento y búsqueda de hiperparámetros, se destacan Logistic Regression, Gaussian Naive Bayes, Light GBoost y LinearSVC por su eficiencia computacional, con tiempos significativamente menores en comparación con los demás modelos.

Teniendo en cuenta todos los aspectos analizados, se considera a Light GBoost como el modelo final. Este modelo obtiene valores altos de Recall y AUC-PR, lo que indica una buena capacidad para identificar la clase minoritaria, aspecto fundamental en el contexto de la predicción de enfermedades cardiovasculares. Además, Light GBoost presenta tiempos de entrenamiento y búsqueda de hiperparámetros relativamente bajos, convirtiéndolo en una opción eficiente desde el punto de vista computacional.

Como último paso en el análisis del modelo Light GBoost elegido para el conjunto de datos con las 29 variables, se muestran la matriz de confusión en la Tabla 4.8 y el reporte de clasificación en la Tabla 4.9. Estos resultados se obtuvieron al evaluar el modelo sobre el conjunto de datos de prueba, habiendo aplicado la técnica de muestreo SMOTEENN en el conjunto de entrenamiento. Adicionalmente, se proporcionan los mejores hiperparámetros identificados durante la fase de ajuste del modelo, los cuales fueron utilizados para su entrenamiento final.

Mejores hiperparámetros para Light GBoost:
{'class_weight': None, 'colsample_bytree': 0.8465638144810189, 'learning_rate': 0.07200487242466924, 'max_depth': 6, 'n_estimators': 89, 'num_leaves': 90, 'subsample': 0.8983173108622504, 'verbose': -1}

A continuación, se muestra la matriz de confusión del modelo Light GBoost entrenado con los mejores hiperparámetros y la técnica de muestreo SMOTEENN. Esta matriz permite visualizar el desempeño del modelo en términos de las predicciones correctas e incorrectas para cada clase.

Tabla 4.8. Matriz de confusión del modelo Light GBoost con SMOTEENN (29 variables).

	Predicción Negativa	**Predicción Positiva**
Negativo	29,949	17,135
Positivo	932	5,080

Para complementar la información proporcionada por la matriz de confusión, se presenta el reporte de clasificación del modelo Light GBoost. Este reporte incluye métricas detalladas de rendimiento, como Precision, Recall y F1-score, para cada clase, así como los promedios ponderados y no ponderados.

Tabla 4.9. Reporte de clasificación del modelo Light GBoost con SMOTEENN (29 variables).

	Precision	Recall	F1-score	Support
0	0.97	0.64	0.77	47,084
1	0.23	0.84	0.36	6,012
Accuracy			0.66	53,096
Macro avg	0.60	0.74	0.56	53,096
Weighted avg	0.89	0.66	0.72	53,096

Al examinar la matriz de confusión del modelo Light GBoost con SMOTEENN en la Tabla 4.8, se observa que el modelo ha clasificado correctamente una gran proporción de casos negativos (29,949) y una cantidad considerable de casos positivos (5,080). Esto indica que el modelo es efectivo en la identificación de personas que no están en riesgo de desarrollar enfermedades cardiovasculares, lo cual es un aspecto favorable.

No obstante, también se aprecia un número significativo de falsos positivos (17,135), lo que implica que el modelo ha clasificado erróneamente a un grupo de personas como propensas a desarrollar enfermedades cardiovasculares, cuando en realidad no lo son. Este resultado señala que el modelo está sobreestimando el riesgo en ciertos casos. Por otra parte, la cantidad de falsos negativos (932) es relativamente baja, indicando que el modelo no está pasando por alto a muchas personas que realmente están en riesgo de desarrollar enfermedades cardiovasculares.

El reporte de clasificación en la Tabla 4.9 proporciona una visión más completa del desempeño del modelo. La precisión para la clase positiva (1) es de 0.23, lo que significa que, de todos los casos que el modelo ha clasificado como positivos, solo el 23% pertenecen realmente a esa clase. Este valor está directamente vinculado con la cantidad elevada de falsos positivos observados en la matriz de confusión.

En contraste, el recall para la clase positiva es de 0.84, indicando que el modelo ha identificado correctamente el 84% de los casos positivos reales. Este valor es bastante alto y resalta que el modelo tiene una buena capacidad para detectar a la mayoría de las personas que realmente están en riesgo de desarrollar enfermedades cardiovasculares.

El F1-score para la clase positiva es de 0.36, una métrica que combina la precisión y el recall. Un F1-score más elevado indica un mejor equilibrio entre estas dos métricas.

También hay que tener en cuenta que el conjunto de datos original presenta un desbalance significativo, con una proporción mucho mayor de casos negativos en comparación con los positivos. Esta característica puede dificultar la tarea del modelo para aprender a clasificar correctamente la clase minoritaria (positiva). A pesar de ello, el modelo Light GBoost ha logrado obtener un recall alto para la clase positiva, lo cual es un aspecto positivo en este contexto.

Por tanto, el modelo Light GBoost ha demostrado ser capaz de identificar correctamente a una gran proporción de personas que no están en riesgo de desarrollar enfermedades cardiovasculares, al

tiempo que detecta a la mayoría de las personas que sí lo están. A pesar de la tendencia a sobreestimar el riesgo en algunos casos, lo que resulta en un número considerable de falsos positivos, el alto recall para la clase positiva es un aspecto destacable, dado que minimiza la posibilidad de pasar por alto a las personas que realmente necesitan atención médica preventiva.

4.2.3 Selección del modelo final

En este apartado, se presenta la selección del modelo final para la predicción de enfermedades cardiovasculares, basándose en los resultados obtenidos en los apartados 4.2.1 y 4.2.2. En estos apartados, se evaluaron y compararon diferentes técnicas de muestreo y modelos predictivos para dos conjuntos de datos: uno con 11 variables y otro con todas las variables disponibles (29). El objetivo principal era identificar el modelo más adecuado para predecir el desarrollo de enfermedades cardiovasculares a partir de factores de riesgo, teniendo en cuenta el desequilibrio de clases en la variable objetivo.

Tras un análisis de los resultados, se seleccionó el modelo Light GBoost con la técnica de muestreo SMOTEENN, entrenado en el conjunto de datos con todas las variables, como el modelo final. Esta decisión se fundamenta en varios criterios, incluyendo el rendimiento del modelo, su interpretabilidad, complejidad y la naturaleza informativa de las variables adicionales.

En términos de rendimiento, el modelo Light GBoost con SMOTEENN y todas las variables muestra una mejora notable en varias métricas esenciales en comparación con el modelo Light GBoost con RandomUnderSampler (RUS) y 11 variables. Específicamente, el modelo con todas las variables alcanza un recall de 0.8450, lo que indica una alta capacidad para identificar correctamente a los pacientes con enfermedad cardiovascular. Aunque la precisión es ligeramente menor (0.2287 frente a 0.2366). Además, el modelo con todas las variables obtiene un AUC-ROC de 0.8139, superando al modelo con 11 variables (0.8062). Esta métrica es especialmente relevante, ya que indica la capacidad del modelo para discriminar entre las clases a diferentes umbrales de clasificación. Un AUC-ROC más alto indica que el modelo con todas las variables tiene una mejor capacidad discriminatoria general.

Dado que el modelo con todas las variables mejora el rendimiento en comparación con el modelo de 11 variables, resulta adecuado seleccionarlo y profundizar en la identificación de las variables que tienen un mayor impacto en el desempeño del modelo. Este análisis permitirá obtener información valiosa sobre los factores de riesgo más significativos y potencialmente descubrir nuevas variables de interés.

La tabla 4.10 presenta una comparación de las métricas clave (F1-score, precision, recall, AUC-ROC y AUC-PR) de los modelos Light GBoost con RUS (11 variables) y con SMOTEENN (todas las variables). Esta tabla permite una comparación visual directa del rendimiento de ambos modelos, destacando la superioridad del modelo con todas las variables en términos de recall y AUC-ROC.

Tabla 4.10. Comparación de métricas de rendimiento entre los modelos Light GBoost con RUS (11 variables) y SMOTEENN (todas las variables).

Modelo	Recall	Precisión	F1-score	AUC-ROC	AUC-PR	Tiempo de entrenamiento (min)	Tiempo búsqueda hiperparámetros (min)
Light GBoost RUS (11vars)	0.7919	0.2366	0.3644	0.8061	0.35	0.01	1.22
Light GBoost SMOTEENN (todas las variables)	0.8450	0.2287	0.3599	0.8139	0.35	0.03	8.59

Además del rendimiento, otro factor importante a considerar es la interpretabilidad y complejidad del modelo. Aunque el modelo con todas las variables es más complejo debido al mayor número de características, su capacidad para proporcionar información más detallada sobre los factores de riesgo es fundamental en el contexto de este trabajo. Al incluir un conjunto más amplio de variables, este modelo permite explorar y descubrir relaciones y patrones que podrían pasar desapercibidos con un modelo más limitado. Esto es especialmente importante en el contexto de la predicción de enfermedades cardiovasculares, donde la identificación de nuevos factores de riesgo puede tener implicaciones significativas para la prevención y el tratamiento.

Sin embargo, es necesario reconocer las posibles desventajas del modelo seleccionado. El uso de todas las variables puede aumentar la complejidad del modelo y el tiempo de entrenamiento, lo que puede ser un factor limitante en algunos casos. Además, el modelo con todas las variables presenta un número considerable de falsos positivos. No obstante, dado el contexto de la predicción de enfermedades cardiovasculares, donde es más importante identificar correctamente a los pacientes en riesgo, se considera que los beneficios del modelo con todas las variables superan estas desventajas.

Por tanto, el modelo Light GBoost con SMOTEENN entrenado en el conjunto de datos con todas las variables se selecciona como el modelo final debido a su mejor rendimiento general, su capacidad para proporcionar información más detallada sobre los factores de riesgo y su potencial para identificar nuevos factores de riesgo. Aunque este modelo presenta algunas desventajas, como una mayor complejidad y un número elevado de falsos positivos, se considera que estos inconvenientes son aceptables dado el objetivo principal de identificar correctamente a los pacientes con riesgo de enfermedad cardiovascular.

4.3. Interpretación de resultados

En el apartado anterior, se seleccionó el modelo Light GBoost con la técnica de muestreo SMOTEENN, entrenado en el conjunto de datos con todas las variables, como el modelo final para la predicción de enfermedades cardiovasculares. Este modelo demostró un rendimiento superior en términos de métricas clave, como el recall y el AUC-ROC, y se destacó por su capacidad para manejar el desequilibrio de clases presente en los datos.

Una vez seleccionado el modelo final, es necesario ir más allá de las métricas de rendimiento y profundizar en la interpretación de los resultados. Este proceso de interpretación es esencial para comprender cómo el modelo toma sus decisiones y cuáles son los factores de riesgo más influyentes en la predicción de enfermedades cardiovasculares.

En este apartado, se llevará a cabo un análisis exhaustivo de los factores de riesgo más influyentes utilizando tres técnicas diferentes: SHAP values, feature importance y permutation importance. Cada una de estas técnicas proporciona una perspectiva única sobre la importancia de las variables predictoras y su contribución al rendimiento del modelo.

Además, se discutirán los hallazgos obtenidos y se evaluará su relevancia en el contexto de la predicción de enfermedades cardiovasculares. Así mismo, se explorarán las implicaciones de los factores de riesgo identificados.

Por último, se abordarán las limitaciones y consideraciones del estudio, reconociendo los posibles sesgos, la generalización de los resultados y las áreas que requieren investigación adicional. Esta discusión puede permitir identificar oportunidades de mejora en futuros trabajos.

4.3.1 Análisis de los factores de riesgo más influyentes

Resultados de SHAP values

Para obtener una comprensión más profunda de cómo el modelo Light GBoost con SMOTEENN toma sus decisiones y cuáles son los factores de riesgo más influyentes, se utilizó la técnica de SHAP (Shapley Additive Explanations) values. Los SHAP values son un método de explicación de modelos de aprendizaje automático que asigna a cada característica un valor de importancia para una predicción particular. Estos valores se basan en la teoría de juegos y proporcionan una medida de la contribución de cada característica al resultado final del modelo.

Para el modelo Light GBoost seleccionado, se calcularon los valores SHAP para todas las instancias del conjunto de datos de prueba. Estos valores permiten cuantificar la importancia de cada característica predictora y determinar si su contribución fue favorable o desfavorable para la predicción de enfermedades cardiovasculares.

La Figura 4.14 muestra el gráfico de importancia global de las características según los SHAP values. En este gráfico, las características están ordenadas de arriba hacia abajo por su importancia promedio absoluta. Cada punto representa una instancia del conjunto de datos, y el color indica si el valor de la característica es alto (rojo) o bajo (azul) para esa instancia. Este gráfico permite identificar las características que tienen el mayor impacto global en las predicciones del modelo.

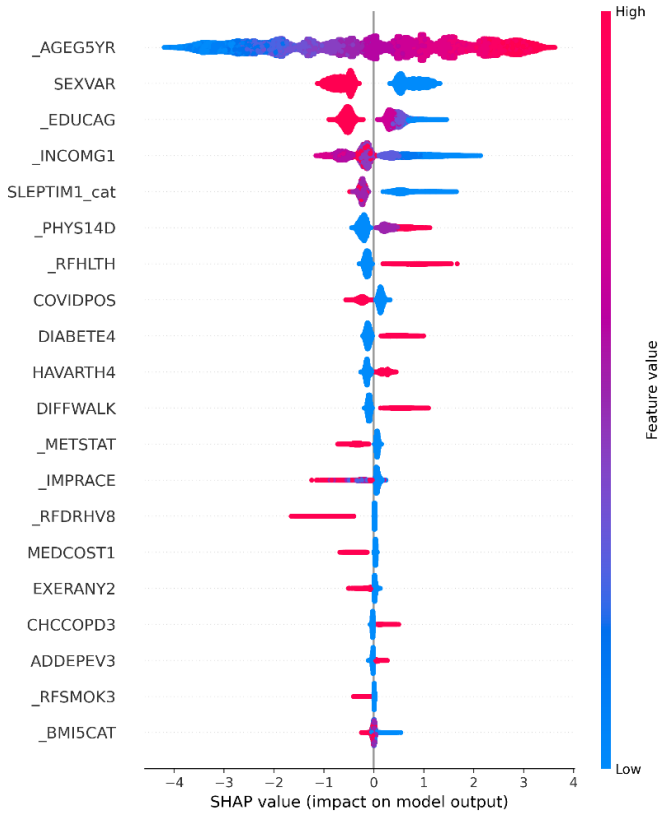


Figura 4.14. Gráfico de importancia global de las características según los SHAP values.

Analizando la Figura 4.14, que muestra la importancia global de las características según los SHAP values para el modelo Light GBoost, se pueden extraer varias conclusiones relevantes sobre los factores de riesgo más influyentes en la predicción de enfermedades cardiovasculares.

En primer lugar, se observa que la edad (AGEG5YR) es la característica con el mayor impacto global en las predicciones del modelo. La distribución de los puntos indica que tanto los valores altos (rojo) como los bajos (azul) de edad tienen una influencia significativa en el resultado final. Esto resalta que la edad es el factor de riesgo más determinante, y que tanto las personas mayores como las más jóvenes pueden tener un riesgo diferente de desarrollar enfermedades cardiovasculares.

La segunda característica más importante es el sexo (SEXVAR). La distribución de los puntos muestra que ser mujer (valores altos, en rojo) tiene un impacto negativo en el modelo, mientras que ser hombre (valores bajos, en azul) tiene un impacto positivo. Esto indica que el sexo es un factor de riesgo significativo, y que los hombres pueden tener un mayor riesgo de enfermedades cardiovasculares en comparación con las mujeres.

Además de los factores demográficos, los aspectos socioeconómicos también desempeñan un papel importante, como el nivel educativo (EDUCAG) y el nivel de ingresos (INCOMG1). La distribución de los puntos sugiere que niveles de educación e ingresos más bajos están asociados con un mayor riesgo de enfermedades cardiovasculares.

Los hábitos de vida saludables también emergen como factores relevantes. La duración del sueño (SLEPTIM1_cat) y los días de mala salud física en el último mes (PHYS14D) exhiben una influencia

notable en las predicciones del modelo. Esto destaca la importancia de mantener un sueño adecuado y una buena salud física para prevenir el riesgo de enfermedades cardiovasculares.

Además, factores como la percepción general de salud (RFHLTH) y la dificultad para caminar (DIFFWALK) también muestran una influencia notable en las predicciones del modelo. Esto indica que la salud autopercibida y la capacidad funcional son aspectos importantes a considerar en la evaluación del riesgo cardiovascular.

Es interesante notar que algunas características relacionadas con enfermedades crónicas, como la diabetes (DIABETE4) y la artritis (HAVARTH4), exhiben una influencia relativamente menor en comparación con otros factores. Esto no significa que estas condiciones no sean relevantes, sino que su impacto global es menos pronunciado en este modelo específico.

Resultados de feature importance y permutation importance

Para obtener una comprensión más completa de la importancia de las características en el modelo Light GBoost, se presentan los resultados de dos técnicas adicionales: feature importance y permutation importance. Mientras que la feature importance mide la contribución de cada característica al rendimiento predictivo del modelo durante el proceso de entrenamiento, la permutation importance evalúa la importancia de las características al medir la disminución en el rendimiento del modelo cuando se permutan aleatoriamente los valores de una característica específica.

La Figura 4.15 muestra tanto la feature importance como la permutation importance para el modelo Light GBoost con SMOTEENN, entrenado en el conjunto de datos con todas las variables. En el gráfico de la izquierda, las características están ordenadas de arriba hacia abajo por su importancia relativa según la feature importance, representada por la longitud de las barras. En el gráfico de la derecha, se presenta la permutation importance, donde las características están ordenadas por su impacto en el rendimiento del modelo al ser permutadas.

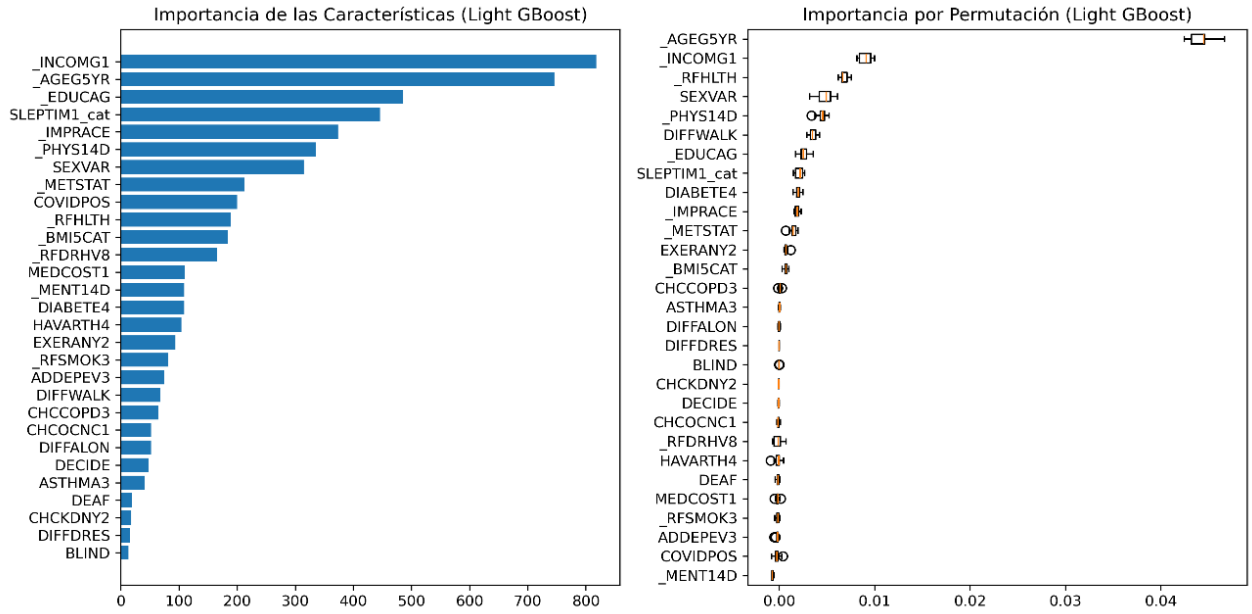


Figura 4.15. Importancia y permutación de las características para el modelo Light GBoost.

En el gráfico de importancia de características, se puede observar que las variables más importantes para el modelo Light GBoost son INCOMG1 (nivel de ingresos), AGEG5YR (edad agrupada en intervalos de 5 años), EDUCAG (nivel educativo), SLEPTIM1_cat (categoría de tiempo de sueño) e IMPRACE (raza/etnia). Estas características tienen los valores más altos de importancia, lo que sugiere que son los factores de riesgo más relevantes en la predicción del modelo. Estas variables reflejan la influencia de los factores socioeconómicos y de estilo de vida en el riesgo cardiovascular.

Cabe destacar que la edad (AGEG5YR) y el nivel de ingresos (INCOMG1) son las dos características más importantes según esta métrica, lo que coincide con los hallazgos del análisis de SHAP values. Esto refuerza la idea de que la edad y los factores socioeconómicos desempeñan un papel muy importante en el riesgo de enfermedades cardiovasculares.

Por otro lado, el gráfico de importancia por permutación muestra una perspectiva ligeramente diferente. En este caso, las características más importantes son AGEG5YR (edad), INCOMG1 (nivel de ingresos), RFHLTH (percepción de salud general), SEXVAR (sexo) y PHYS14D (días de mala salud física). Estas variables son las que más contribuyen al rendimiento del modelo cuando se las permuta, lo que indica su relevancia en la predicción.

Es interesante observar que la percepción de salud (RFHLTH) y el sexo (SEXVAR) tienen una mayor importancia según la métrica de permutación en comparación con la importancia de características. Esto sugiere que, aunque estas variables no sean las más importantes en términos de ganancias de división, su permutación tiene un impacto significativo en el rendimiento del modelo.

Así mismo, algunas características, como la diabetes (DIABETE4) y la artritis (HAVARTH4), tienen una importancia relativamente menor según ambos métodos. Esto indica que, aunque estas condiciones son factores de riesgo conocidos para las enfermedades cardiovasculares, su impacto en el modelo específico es menos pronunciado en comparación con otras variables.

Tanto la importancia de características como la importancia por permutación resaltan la relevancia de factores como la edad, los ingresos, el nivel educativo, el tiempo de sueño, la salud física, la percepción de salud y el sexo en la predicción de enfermedades cardiovasculares. Estos resultados coinciden en gran medida con los hallazgos del análisis de SHAP values, brindando una comprensión más sólida de los factores de riesgo clave.

4.3.2 Discusión de los hallazgos

Los resultados obtenidos en este trabajo, utilizando el modelo Light GBoost con la técnica de muestreo SMOTEENN y todas las variables disponibles, proporcionan información valiosa sobre los factores de riesgo más influyentes en la predicción de enfermedades cardiovasculares. A través de técnicas de interpretación como los SHAP values, la feature importance y la permutation importance, se han identificado características clave que contribuyen al riesgo cardiovascular.

Entre los hallazgos más destacados, se encuentra la importancia de factores demográficos como la edad y el sexo. La edad se ha identificado consistentemente como uno de los predictores más fuertes, lo que concuerda con el conocimiento médico existente sobre el aumento del riesgo cardiovascular a medida que las personas envejecen. Además, el sexo también ha demostrado ser un factor importante, con los hombres presentando un mayor riesgo en comparación con las mujeres.

Otro aspecto relevante son los factores socioeconómicos, como el nivel de ingresos y el nivel educativo. Los resultados sugieren que un menor nivel socioeconómico está asociado con un mayor riesgo de enfermedades cardiovasculares. Esto resalta la importancia de considerar las desigualdades socioeconómicas en la prevención y el manejo de estas enfermedades.

Además, se han identificado factores de estilo de vida, como el tiempo de sueño y la actividad física, como predictores significativos. Dormir pocas horas y tener más días de mala salud física parecen estar asociados con un mayor riesgo cardiovascular. Estos hallazgos respaldan la importancia de promover hábitos de sueño saludables y fomentar la actividad física regular como medidas preventivas.

La percepción general de salud también ha demostrado ser un factor influyente. Una peor percepción de la propia salud se asocia con un mayor riesgo de enfermedades cardiovasculares, lo cual señala que la autopercepción de la salud puede ser un indicador valioso a tener en cuenta en la evaluación del riesgo.

Otros factores, como vivir en una zona metropolitana y tener dificultades para caminar, también han mostrado una influencia notable en las predicciones del modelo. Estos hallazgos plantean la necesidad de considerar una amplia gama de factores, más allá de los tradicionales, en la evaluación del riesgo cardiovascular.

Para contextualizar los resultados obtenidos en este estudio, es útil compararlos con otros trabajos similares. Un estudio riguroso realizado en el Reino Unido por Weng et al. (2017) [4] utilizó un conjunto de datos similar por extensión y aplicó diferentes algoritmos de aprendizaje automático para predecir el riesgo cardiovascular. La Tabla 4.11 muestra los resultados de ese estudio.

Tabla 4.11. Resultados del estudio de Weng et al. (2017) sobre predicción de riesgo cardiovascular en el Reino Unido.

Algorithms	**Cases Incorrect (False Negative)**	**Total CVD Cases**	**Total Non-Cases**	**Sensitivity (True Positive)**	**Specificity (True Negative)**	**Positive Predictive Value (PPV)**	**Negative Predictive Value (NPV)**
ACC/AHA Model	2,761	7,404	75,585	62.7%	70.3%	17.1%	95.1%
ML: Random Forest	2,570	7,404	75,585	65.3%	70.5%	17.8%	95.4%
ML: Logistic Regression	2,437	7,404	75,585	67.1%	70.7%	18.3%	95.6%
ML: Gradient Boosting Machines	2,407	7,404	75,585	67.5%	70.7%	18.4%	95.7%
ML: Neural Networks	2,406	7,404	75,585	67.5%	70.7%	18.4%	95.7%

A pesar de las diferencias en las poblaciones y los conjuntos de datos, se observan algunas similitudes notables en las métricas de rendimiento. En este trabajo, la precisión (Positive Predictive Value, PPV) fue de 22.87%, ligeramente superior a los valores entre 17.1% y 18.4% obtenidos por los distintos modelos en el estudio de Weng et al. Y una sensibilidad del 84.5% (ver Tabla 4.10), mientras que en el estudio de Weng et al. oscilo entre 62.7% y 67.5%. De hecho, si se hubiera utilizado una precisión similar para hacer los resultados más comparables, se habría empleado el modelo entrenado con la técnica de muestreo InstanceHardnessThreshold, que obtuvo una precisión

media de 17.8% entre los 8 modelos evaluados. Sin embargo, al usar esta técnica de muestreo, se logró una sensibilidad mucho mayor en este trabajo, alcanzando un 92.7%.

En ambos estudios se obtuvieron valores de precisión relativamente bajos, lo cual es común en problemas de clasificación con un alto desequilibrio de clases. Estas similitudes resaltan la consistencia de los resultados obtenidos en diferentes contextos y refuerzan la validez de las métricas de rendimiento alcanzadas en este trabajo. Además, la comparación con otros estudios permite contextualizar los hallazgos y evaluar su generalización a diferentes poblaciones y entornos clínicos.

Sin embargo, es importante destacar que este trabajo se centra en identificar los factores de riesgo más influyentes y explorar su importancia relativa, mientras que el estudio de Weng et al. se enfocó principalmente en comparar el rendimiento de diferentes algoritmos de aprendizaje automático con un modelo tradicional (ACC/AHA). Además, mientras el estudio de Weng et al. se realizó en la población del Reino Unido, este proyecto se enfoca en la población de Estados Unidos.

La comparación con otros estudios también resalta la importancia de considerar las diferencias en las poblaciones y los sistemas de atención médica. Aunque los resultados pueden variar entre diferentes países y contextos, la identificación de factores de riesgo clave y la aplicación de técnicas de aprendizaje automático para mejorar la predicción del riesgo cardiovascular son aspectos comunes y relevantes. Estos resultados destacan la capacidad de los modelos de aprendizaje automático desarrollados en este trabajo para lograr un alto nivel de sensibilidad en la predicción de enfermedades cardiovasculares, incluso con una precisión comparable a la de otros estudios relevantes en el campo.

Por otro lado, es importante tener en cuenta las limitaciones del estudio, como la naturaleza observacional de los datos y la posibilidad de factores de confusión no medidos. Además, la generalización de los resultados a otras poblaciones y contextos debe hacerse con cautela.

4.3.3 Limitaciones y consideraciones del estudio

El presente estudio ha abordado la predicción de enfermedades cardiovasculares utilizando técnicas de aprendizaje automático y ha explorado diferentes estrategias para manejar el desbalanceo de clases en el conjunto de datos. A pesar de los resultados prometedores obtenidos, es fundamental reconocer las limitaciones y consideraciones del trabajo.

En primer lugar, los datos utilizados provienen de una encuesta telefónica que recopila información sobre comportamientos de riesgo para la salud, prácticas preventivas y acceso a la atención médica en la población adulta de EE. UU. Aunque el BRFSS es una fuente valiosa de datos, es importante tener en cuenta que la información se basa en autoinformes de los participantes, lo que puede estar sujeto a sesgos de memoria y deseabilidad social. Esto puede afectar la precisión de ciertas variables, especialmente aquellas relacionadas con hábitos de vida y antecedentes médicos. Además, la naturaleza transversal de la encuesta limita la capacidad para establecer relaciones causales entre los factores de riesgo y las enfermedades cardiovasculares.

Otra consideración importante es que los datos utilizados corresponden al año 2022, lo que significa que reflejan el estado de salud y los comportamientos de riesgo de la población en un momento

específico. Es posible que ciertos factores de riesgo y patrones de enfermedad hayan cambiado con el tiempo, y los resultados pueden no ser completamente generalizables a otros períodos.

Así mismo, aunque el conjunto de datos utilizado es amplio y representativo de la población adulta de EE. UU., es posible que no capture todas las variables relevantes para predecir el riesgo cardiovascular. En particular, el modelo desarrollado en este proyecto no incluye información directa sobre los niveles de colesterol y la presión arterial alta, que son factores de riesgo bien establecidos para las enfermedades cardiovasculares. La inclusión de preguntas específicas sobre estos aspectos en la encuesta podría haber mejorado aún más la capacidad predictiva del modelo.

Otra limitación del estudio es que se centra en la población adulta de Estados Unidos, y los resultados pueden no ser directamente extrapolables a otras poblaciones o contextos culturales. Las diferencias en los factores de riesgo, los estilos de vida y los sistemas de atención médica pueden influir en la aplicabilidad de los hallazgos a nivel internacional.

Además, es importante señalar que el modelo desarrollado en este trabajo se basa en un conjunto de datos desequilibrado, con una proporción mucho menor de casos positivos (personas con enfermedades cardiovasculares) en comparación con los casos negativos. Aunque se aplicaron técnicas de muestreo y se utilizaron métricas de evaluación apropiadas para abordar este desequilibrio, es posible que el modelo tenga limitaciones para generalizar a poblaciones con una prevalencia diferente de enfermedades cardiovasculares.

A pesar de estas limitaciones, los hallazgos de este trabajo muestran la efectividad en la aplicación de técnicas de aprendizaje automático para predecir el riesgo cardiovascular e identificar los factores de riesgo más influyentes.

Por último, es importante reconocer que este trabajo se basa en datos observacionales y no establece relaciones causales entre los factores de riesgo y las enfermedades cardiovasculares. Los resultados deben interpretarse como asociaciones y no como prueba de causalidad. Se necesitan estudios prospectivos y ensayos clínicos para confirmar las relaciones causales y evaluar la eficacia de las intervenciones basadas en los factores de riesgo identificados.

Este trabajo destaca el potencial de los modelos de aprendizaje automático para mejorar la predicción del riesgo cardiovascular y resalta la importancia de considerar una amplia gama de factores de riesgo, incluyendo características demográficas, socioeconómicas y de estilo de vida. De igual manera, es esencial tomar en cuenta las limitaciones y las consideraciones mencionadas al interpretar los resultados. La naturaleza de los datos de la encuesta telefónica y la ausencia de algunos factores de riesgo clave pueden haber influido en la capacidad predictiva de los modelos.

5. Conclusiones

Se desarrolló con éxito un modelo predictivo utilizando técnicas de aprendizaje automático para predecir el riesgo de enfermedad cardiovascular. El modelo Light GBoost con la técnica de muestreo SMOTEENN, entrenado en el conjunto de datos con todas las variables disponibles (29), demostró el mejor rendimiento general, con valores altos de recall (0.8450) y AUC-ROC (0.8139). Esto indica la capacidad del modelo para identificar correctamente a las personas con riesgo de padecer estas enfermedades.

Se identificaron los principales factores de riesgo asociados con enfermedades cardiovasculares utilizando técnicas de interpretación de modelos como SHAP values, feature importance y permutation importance. La edad, el sexo, los factores socioeconómicos (nivel de ingresos y educación), el tiempo de sueño, la salud física y la percepción general de salud se destacaron consistentemente como las características más influyentes en la predicción del riesgo cardiovascular.

Se investigó la influencia de los factores demográficos, como la edad y el género, en el riesgo de desarrollar enfermedades cardiovasculares. Los resultados mostraron que la edad es el factor de riesgo más determinante, con un aumento del riesgo a medida que las personas envejecen. Además, se observó que los hombres tienen un mayor riesgo en comparación con las mujeres. Estos hallazgos concuerdan con la literatura médica existente y enfatizan la relevancia de considerar los factores demográficos en la estratificación del riesgo y la personalización de las estrategias preventivas.

Los resultados obtenidos subrayan la necesidad de considerar no solo los factores de riesgo tradicionales, como la edad y el sexo, sino también aspectos socioeconómicos y de estilo de vida en la evaluación del riesgo cardiovascular. Se analizó la contribución del estilo de vida, como la actividad física. Los hallazgos sugieren que dormir pocas horas y tener más días de mala salud física están asociados con un mayor riesgo cardiovascular, lo que resalta la importancia de promover hábitos de sueño saludables y fomentar la actividad física regular como medidas preventivas.

Este trabajo también ha puesto de manifiesto la necesidad de abordar las disparidades en salud asociadas con factores como la raza/etnia y el nivel socioeconómico. Donde se observaron diferencias significativas en la prevalencia de enfermedades cardiovasculares entre diferentes grupos raciales/étnicos, niveles de ingresos y niveles educativos. Específicamente, se encontró que los bajos niveles de ingresos y de educación contribuyen a un mayor riesgo de estas patologías.

Se identificaron patrones y relaciones entre las variables de riesgo cardiovascular que permiten una mayor explicabilidad del modelo creado. El análisis de SHAP values y la importancia de las características revelaron interacciones complejas entre factores demográficos, socioeconómicos y de estilo de vida en la determinación del riesgo cardiovascular.

Además de los objetivos específicos, es importante destacar otros aspectos relevantes del proyecto. Se utilizó un conjunto de datos amplio y representativo de la población adulta de EE. UU., lo que aumenta la generalización de los resultados. Sin embargo, se reconocen limitaciones como la naturaleza de los datos utilizados, que fueron proporcionados directamente por los participantes del estudio, y la ausencia de algunas variables clínicas clave, como los niveles de colesterol y la presión arterial.

La comparación con otros estudios similares, como el de Weng et al. (2017) [4], resalta la consistencia de los resultados obtenidos en diferentes contextos y refuerza la validez de las métricas de rendimiento alcanzadas en este trabajo. Esto destaca el potencial de los modelos de aprendizaje automático para mejorar la predicción del riesgo cardiovascular en diferentes poblaciones.

El proyecto abordó de manera exhaustiva el desafío del desequilibrio de clases en el conjunto de datos, evaluando múltiples técnicas de muestreo y seleccionando SMOTEENN como la más adecuada. Esto demuestra la importancia de considerar y abordar este aspecto en problemas de clasificación con clases desbalanceadas.

Aunque el modelo desarrollado presenta un número considerable de falsos positivos, en el contexto de la predicción de enfermedades cardiovasculares, se considera aceptable priorizar la identificación correcta de los pacientes en riesgo (alta sensibilidad) sobre la precisión. Esto permite una detección temprana y la implementación de estrategias preventivas.

Este proyecto ha demostrado la efectividad de las técnicas de aprendizaje automático en la predicción del riesgo de enfermedades cardiovasculares, identificando los factores de riesgo más influyentes y proporcionando información útil para la prevención y el manejo de estas afecciones. A pesar de las limitaciones reconocidas, los resultados obtenidos sientan las bases para el desarrollo de herramientas de predicción más precisas y personalizadas. A nivel de salud pública, los hallazgos enfatizan la trascendencia de abordar los factores de riesgo modificables, como el estilo de vida, y de considerar los determinantes sociales de la salud en las políticas y programas de prevención de enfermedades cardiovasculares. La integración cuidadosa de estos modelos en los sistemas de atención médica, junto con el juicio clínico experto, puede contribuir a mejorar la prevención y el tratamiento de las enfermedades cardiovasculares y, en última instancia, a reducir la carga de estas afecciones en la sociedad.

Glosario

Aterosclerosis: Es el depósito y la infiltración de lípidos (grasas) en las paredes de las arterias. Este hecho engrosa progresivamente las arterias, haciendo que pierdan su elasticidad y se produzca una insuficiencia del riego sanguíneo.

Bioimpedancia: Es un método para estimar la composición corporal, en particular la grasa corporal y la masa muscular, donde una corriente eléctrica débil fluye a través del cuerpo.

Censura aleatoria: Ocurre cuando los individuos pueden abandonar el estudio o experimentar un evento de interés (como la muerte) por causas que no están relacionadas con el evento de interés.

Citocinas: Son un grupo de proteínas y glucoproteínas producidas por diversos tipos celulares que actúan fundamentalmente como reguladores de las respuestas inmunitaria e inflamatoria.

Comorbilidad: Presencia de dos o más enfermedades al mismo tiempo en una persona.

Fisiopatológico: Mecanismos por los cuales se originan las distintas enfermedades.

Morbilidad: Es un estado enfermo, de discapacidad, o mala salud debido a cualquier causa.

Prevalencia: En medicina, es la proporción de personas de una población que presentan cierta enfermedad, afección o factor de riesgo en un momento específico o durante un periodo determinado.

Reticulocitos: Son glóbulos rojos que aún se están desarrollando. También se les conoce como glóbulos rojos inmaduros.

Townsend Index: Es una medida de privación material dentro de una población, basada en cuatro variables: desempleo, propiedad de automóvil, propiedad de vivienda y hacinamiento en el hogar.

Triglicéridos: Son un tipo de grasas que circulan en la sangre. Son el tipo más frecuente de grasas en el cuerpo.

Bibliografía

[1] World Health Organization. Cardiovascular diseases (CVDs) (who.int) [Internet]. Geneva: World Health Organization; 2021 [Consulta: 01 de marzo de 2024]. Disponible en: https://www.who.int/news-room/fact-sheets/detail/cardiovascular-diseases-(cvds)

[2] Schultz WM, Kelli HM, Lisko JC, et al. Socioeconomic Status and Cardiovascular Outcomes: Challenges and Interventions. Circulation [Internet]. 2018 Jun 18 [Consulta: 13 de marzo de 2024];137(20):2166-2178. Disponible en: https://doi.org/10.1161/CIRCULATIONAHA.117.029652

[3] Karatzia L, Aung N, Aksentijevic D. Artificial intelligence in cardiology: Hope for the future and power for the present. Frontiers in Cardiovascular Medicine [Internet]. 2022 Oct 13 [Consulta: 13 de marzo de 2024];9:945726. Disponible en: https://doi.org/10.3389/fcvm.2022.945726

[4] Weng SF, Reps J, Kai J, Garibaldi JM, Qureshi N. Can machine-learning improve cardiovascular risk prediction using routine clinical data?. PLoS One [Internet]. 2017 Abr 4 [Consulta: 13 de marzo de 2024];12(4):e0174944. Disponible en: https://doi.org/10.1371/journal.pone.0174944

[5] Martínez-García M, Salinas-Ortega M, Estrada-Arriaga I, Hernández-Lemus E, García-Herrera R, Vallejo M. A systematic approach to analyze the social determinants of cardiovascular disease. PLoS One [Internet]. 2018 Ene 25 [Consulta: 13 de marzo de 2024];13(1):e0190960. Disponible en: https://doi.org/10.1371/journal.pone.0190960

[6] Mathur P, Srivastava S, Xu X, Mehta JL. Artificial Intelligence, Machine Learning, and Cardiovascular Disease. Clinical Medicine Insights: Cardiology [Internet]. 2020 Sep 9 [Consulta: 13 de marzo de 2024];14:1179546820927404. Disponible en: https://doi.org/10.1177/1179546820927404

[7] Sun X, Yin Y, Yang Q, Huo T. Artificial intelligence in cardiovascular diseases: diagnostic and therapeutic perspectives. European Journal of Medical Research [Internet]. 2023 Jul 21 [Consulta: 13 de marzo de 2024];28(1):242. Disponible en: https://doi.org/10.1186/s40001-023-01065-y

[8] Thiriet M. Cardiovascular Disease: An Introduction. Vasculopathies [Internet]. 2019 Feb 19 [Consulta: 13 de marzo de 2024];8:1-90. Disponible en: https://doi.org/10.1007/978-3-319-89315-0_1

[9] Olvera Lopez E, Ballard BD, Jan A. Cardiovascular Disease. [Actualizado 2023 Ago 22]. In: StatPearls [Internet]. Treasure Island (FL): StatPearls Publishing. 2024 Jan [Consulta: 13 de marzo de 2024]. Disponible en: https://www.ncbi.nlm.nih.gov/books/NBK535419/

[10] American Heart Association. What is cardiovascular disease? (heart.org) [Internet]. Dallas: American Heart Association; 2024 Ene 9 [Consulta: 13 de marzo de 2024]. Disponible en: https://www.heart.org/en/health-topics/consumer-healthcare/what-is-cardiovascular-disease

[11] Mendis, Shanthi, Puska, Pekka, Norrving, B, World Health Organization, World Heart Federation. et al. Global atlas on cardiovascular disease prevention and control. Geneva: World Health Organization. 2011 [Consulta: 13 de marzo de 2024]. Disponible en: https://iris.who.int/handle/10665/44701

[12] O'Donnell CJ, Elosua R. Factores de riesgo cardiovascular. Perspectivas derivadas del Framingham Heart Study [Cardiovascular risk factors. Insights from Framingham Heart Study].

Revista Española de Cardiología [Internet]. 2008 Mar [Consulta: 13 de marzo de 2024];61(3):299-310. Disponible en: https://doi.org/10.1157/13116658

[13] Goff DC Jr, Lloyd-Jones DM, Bennett G, et al. 2013 ACC/AHA guideline on the assessment of cardiovascular risk: a report of the American College of Cardiology/American Heart Association Task Force on Practice Guidelines [published correction appears in Circulation. Publicación original 2013 Nov 12;129(25 Suppl 2):S74-5]. Circulation [Internet]. 2014 Jun 24 [Consulta: 14 de marzo de 2024];129(25 Suppl 2):S49-S73. Disponible en: https://doi.org/10.1161/01.cir.0000437741.48606.98

[14] Hippisley-Cox J, Coupland C, Vinogradova Y, et al. Predicting cardiovascular risk in England and Wales: prospective derivation and validation of QRISK2. BMJ (Clinical research ed.) [Internet]. 2008 Jun 26 [Consulta: 14 de marzo de 2024];336(7659):1475-1482. Disponible en: https://doi.org/10.1136/bmj.39609.449676.25

[15] D'Agostino RB Sr, Vasan RS, Pencina MJ, et al. General cardiovascular risk profile for use in primary care: the Framingham Heart Study. Circulation [Internet]. 2008 Ene 22 [Consulta: 14 de marzo de 2024];117(6):743-753. Disponible en: https://doi.org/10.1161/CIRCULATIONAHA.107.699579

[16] Ridker PM, Buring JE, Rifai N, Cook NR. Development and validation of improved algorithms for the assessment of global cardiovascular risk in women: the Reynolds Risk Score [published correction appears in JAMA. 2007 Abr 4;297(13):1433]. JAMA [Internet]. 2007 Feb 14 [Consulta: 14 de marzo de 2024];297(6):611-619. Disponible en: https://doi.org/10.1001/jama.297.6.611

[17] Yi J, Wang L, Guo X, Ren X. Association of Life's Essential 8 with all-cause and cardiovascular mortality among US adults: A prospective cohort study from the NHANES 2005-2014. Nutrition, Metabolism, and Cardiovascular Diseases : NMCD [Internet]. 2023 Ene 27 [Consulta: 14 de marzo de 2024];33(6):1134-1143. Disponible en: https://doi.org/10.1016/j.numecd.2023.01.021

[18] Nedkoff L, Briffa T, Zemedikun D, Herrington S, Wright FL. Global Trends in Atherosclerotic Cardiovascular Disease. Clinical Therapeutics [Internet]. 2023 Oct 31 [Consulta: 14 de marzo de 2024];45(11):1087-1091. Disponible en: https://doi.org/10.1016/j.clinthera.2023.09.020

[19] Ortega, Francisco B., Lavie, Carl J., and Blair, Steven N. Obesity and cardiovascular disease. Circulation Research [Internet]. 2016 May 27 [Consulta: 14 de marzo de 2024]; 118 (11) 1752-1770. Disponible en: https://doi.org/10.1161/CIRCRESAHA.115.306883

[20] Hershman DL, Till C, Shen S, et al. Association of Cardiovascular Risk Factors With Cardiac Events and Survival Outcomes Among Patients With Breast Cancer Enrolled in SWOG Clinical Trials. Journal of Clinical Oncology : official journal of the American Society of Clinical Oncology [Internet]. 2018 Mar 27 [Consulta: 14 de marzo de 2024];36(26):2710-2717. Disponible en: https://doi.org/10.1200/JCO.2017.77.4414

[21] Larsson SC, Åkesson A, Wolk A. Primary prevention of stroke by a healthy lifestyle in a high-risk group. Neurology [Internet]. 2015 Jun 2 [Consulta: 14 de marzo de 2024];84(22):2224-2228. Disponible en: https://doi.org/10.1212/WNL.0000000000001637

[22] Patra J, Taylor B, Irving H, et al. Alcohol consumption and the risk of morbidity and mortality for different stroke types--a systematic review and meta-analysis. BMC Public Health [Internet]. 2010 May 18 [Consulta: 14 de marzo de 2024];10:258. Disponible en: https://doi.org/10.1186/1471-2458-10-258

[23] Goldstein LB, Bushnell CD, Adams RJ, et al. Guidelines for the primary prevention of stroke: a guideline for healthcare professionals from the American Heart Association/American Stroke

Association [published correction appears in Stroke. 2011 Feb;42(2):e26]. Stroke [Internet]. 2010 Dic 2 [Consulta: 14 de marzo de 2024];42(2):517-584. Disponible en: https://doi.org/10.1161/STR.0b013e3181fcb238

[24] Chareonrungrueangchai K, Wongkawinwoot K, Anothaisintawee T, Reutrakul S. Dietary Factors and Risks of Cardiovascular Diseases: An Umbrella Review. Nutrients [Internet]. 2020 Abr 15 [Consulta: 14 de marzo de 2024];12(4):1088. Disponible en: https://doi.org/10.3390/nu12041088

[25] Bhatt DL, Steg PG, Miller M, et al. Cardiovascular Risk Reduction with Icosapent Ethyl for Hypertriglyceridemia. The New England Journal of Medicine [Internet]. 2019 Ene 3 [Consulta: 15 de marzo de 2024];380(1):11-22. Disponible en: https://doi.org/10.1056/NEJMoa1812792

[26] Lip GYH, Genaidy A, Estes C. Cardiovascular disease (CVD) outcomes and associated risk factors in a medicare population without prior CVD history: an analysis using statistical and machine learning algorithms. Internal and Emergency Medicine [Internet]. 2023 Jun 9 [Consulta: 15 de marzo de 2024];18(5):1373-1383. Disponible en: https://doi.org/10.1007/s11739-023-03297-6

[27] Bechthold A, Boeing H, Schwedhelm C, et al. Food groups and risk of coronary heart disease, stroke and heart failure: A systematic review and dose-response meta-analysis of prospective studies. Critical Reviews in Food Science and Nutrition [Internet]. 2019 [Consulta: 15 de marzo de 2024];59(7):1071-1090. Disponible en: https://doi.org/10.1080/10408398.2017.1392288

[28] Shivappa N, Godos J, Hébert JR, et al. Dietary Inflammatory Index and Cardiovascular Risk and Mortality-A Meta-Analysis. Nutrients [Internet]. 2018 Feb 12 [Consulta: 15 de marzo de 2024];10(2):200. Disponible en: https://doi.org/10.3390/nu10020200

[29] Lee TK, Wickrama KAS, O'Neal CW. How Early Stressful Life Experiences Combine With Adolescents' Conjoint Health Risk Trajectories to Influence Cardiometabolic Disease Risk in Young Adulthood. Journal of Youth and Adolescence [Internet]. 2021 May 4 [Consulta: 15 de marzo de 2024];50(6):1234-1253. Disponible en: https://doi.org/10.1007/s10964-021-01440-0

[30] Mosquera PA, San Sebastian M, Waenerlund AK, Ivarsson A, Weinehall L, Gustafsson PE. Income-related inequalities in cardiovascular disease from mid-life to old age in a Northern Swedish cohort: A decomposition analysis. Social Science & Medicine [Internet]. 2016 Ene [Consulta: 15 de marzo de 2024];149:135-144. Disponible en: https://doi.org/10.1016/j.socscimed.2015.12.017

[31] Damen JA, Hooft L, Schuit E, et al. Prediction models for cardiovascular disease risk in the general population: systematic review. BMJ (Clinical research ed.) [Internet]. 2016 May 16 [Consulta: 15 de marzo de 2024];353:i2416. Disponible en: https://doi.org/10.1136/bmj.i2416

[32] SCORE2 working group and ESC Cardiovascular risk collaboration. SCORE2 risk prediction algorithms: new models to estimate 10-year risk of cardiovascular disease in Europe. European Heart Journal [Internet]. 2021 Jun 13 [Consulta: 15 de marzo de 2024];42(25):2439-2454. Disponible en: https://doi.org/10.1093/eurheartj/ehab309

[33] Hippisley-Cox J, Coupland C, Vinogradova Y, et al. Predicting cardiovascular risk in England and Wales: prospective derivation and validation of QRISK2. BMJ (Clinical research ed.) [Internet]. 2008 Jun 26 [Consulta: 15 de marzo de 2024];336(7659):1475-1482. Disponible en: https://doi.org/10.1136/bmj.39609.449676.25

[34] Russell S, Norvig P. Artificial Intelligence: A Modern Approach, Global Edition. 4th ed. Harlow: Pearson Education Limited; 2021. 1168 p.

[35] Cai Y, Cai YQ, Tang LY, et al. Artificial intelligence in the risk prediction models of cardiovascular disease and development of an independent validation screening tool: a systematic review. BMC Medicine [Internet]. 2024 Feb 5 [Consulta: 15 de marzo de 2024];22(1):56. Disponible en: https://doi.org/10.1186/s12916-024-03273-7

[36] Lindholm D, Fukaya E, Leeper NJ, Ingelsson E. Bioimpedance and New-Onset Heart Failure: A Longitudinal Study of >500 000 Individuals From the General Population. Journal of the American Heart Association [Internet]. 2018 Jun 29 [Consulta: 15 de marzo de 2024];7(13):e008970. Disponible en: https://doi.org/10.1161/JAHA.118.008970

[37] Cho SY, Kim SH, Kang SH, et al. Pre-existing and machine learning-based models for cardiovascular risk prediction. Scientific Reports [Internet]. 2021 Abr 26 [Consulta: 15 de marzo de 2024];11(1):8886. Disponible en: https://doi.org/10.1038/s41598-021-88257-w

[38] Jiang Y, Zhang X, Ma R, et al. Cardiovascular Disease Prediction by Machine Learning Algorithms Based on Cytokines in Kazakhs of China. Clinical Epidemiology [Internet]. 2021 Jun 9 [Consulta: 15 de marzo de 2024];13:417-428. Disponible en: https://doi.org/10.2147/CLEP.S313343

[39] Yu J, Yang X, Deng Y, et al. Incorporating longitudinal history of risk factors into atherosclerotic cardiovascular disease risk prediction using deep learning. Scientific Reports [Internet]. 2024 Ene 31 [Consulta: 15 de marzo de 2024];14(1):2554. Disponible en: https://doi.org/10.1038/s41598-024-51685-5

[40] Kim JOR, Jeong YS, Kim JH, Lee JW, Park D, Kim HS. Machine Learning-Based Cardiovascular Disease Prediction Model: A Cohort Study on the Korean National Health Insurance Service Health Screening Database. Diagnostics (Basel, Switzerland) [Internet]. 2021 May 25 [Consulta: 15 de marzo de 2024];11(6):943. Disponible en: https://doi.org/10.3390/diagnostics11060943

[41] Al-Droubi SS, Jahangir E, Kochendorfer KM, et al. Artificial intelligence modelling to assess the risk of cardiovascular disease in oncology patients. European Heart Journal - Digital Health [Internet]. 2023 May 8 [Consulta: 15 de marzo de 2024];4(4):302-315. Disponible en: https://doi.org/10.1093/ehjdh/ztad031

[42] Salah H, Srinivas S. Explainable machine learning framework for predicting long-term cardiovascular disease risk among adolescents. Scientific Reports [Internet]. 2022 Dic 19 [Consulta: 15 de marzo de 2024];12(1):21905. Disponible en: https://doi.org/10.1038/s41598-022-25933-5

[43] Quesada JA, Lopez-Pineda A, Gil-Guillén VF, et al. Machine learning to predict cardiovascular risk. International Journal of Clinical Practice [Internet]. 2019 Jul 1 [Consulta: 15 de marzo de 2024];73(10):e13389. Disponible en: https://doi.org/10.1111/ijcp.13389

[44] Alaa AM, Bolton T, Di Angelantonio E, Rudd JHF, van der Schaar M. Cardiovascular disease risk prediction using automated machine learning: A prospective study of 423,604 UK Biobank participants. PLoS One [Internet]. 2019 May 15 [Consulta: 15 de marzo de 2024];14(5):e0213653. Disponible en: https://doi.org/10.1371/journal.pone.0213653

[45] Hassannejad R, Mansourian M, Marateb H, et al. Developing Non-Laboratory Cardiovascular Risk Assessment Charts and Validating Laboratory and Non-Laboratory-Based Models. Global Heart [Internet]. 2021 Sep 2 [Consulta: 15 de marzo de 2024];16(1):58. Disponible en: https://doi.org/10.5334/gh.890

[46] Roth GA, Johnson C, Abajobir A, et al. Global, Regional, and National Burden of Cardiovascular Diseases for 10 Causes, 1990 to 2015. Journal of the American College of Cardiology [Internet].

2017 Jul 4 [Consulta: 15 de marzo de 2024];70(1):1-25. Disponible en: https://doi.org/10.1016/j.jacc.2017.04.052

[47] Shishehbori, F. and Awan, Z., “Enhancing Cardiovascular Disease Risk Prediction with Machine Learning Models”, arXiv e-prints [Internet], 2024 Feb 9 [Consulta: 15 de marzo de 2024]. Disponible en: https://doi.org/10.48550/arXiv.2401.17328

[48] You J, Guo Y, Kang JJ, et al. Development of machine learning-based models to predict 10-year risk of cardiovascular disease: a prospective cohort study. Stroke and Vascular Neurology [Internet]. 2023 Dic 29 [Consulta: 15 de marzo de 2024];8(6):475-485. Disponible en: https://doi.org/10.1136/svn-2023-002332

[49] MacEachern SJ, Forkert ND. Machine learning for precision medicine. Genome [Internet]. 2021 Abr [Consulta: 15 de marzo de 2024];64(4):416-425. Disponible en: https://doi.org/10.1139/gen-2020-0131

[50] Centers for Disease Control and Prevention. CDC - About BRFSS (cdc.gov) [Internet]. Atlanta: Centers for Disease Control and Prevention; 2024 [Consulta: 08 de mayo de 2024]. Disponible en: https://www.cdc.gov/brfss/about/index.htm

[51] Centers for Disease Control and Prevention. CDC - 2022 BRFSS Survey Data and Documentation (cdc.gov) [Internet]. Atlanta: Centers for Disease Control and Prevention; 2024 [Consulta: 08 de mayo de 2024]. Disponible en: https://www.cdc.gov/brfss/annual_data/annual_2022.html

[52] Centers for Disease Control and Prevention. 2022 BRFSS Codebook CDC (cdc.gov) [Internet]. Atlanta: Centers for Disease Control and Prevention; 2024 [Consulta: 08 de mayo de 2024]. Disponible en: https://www.cdc.gov/brfss/annual_data/2022/zip/codebook22_llcp-v2-508.zip

[53] Cox, D. R. (1958). The Regression Analysis of Binary Sequences. Journal of the Royal Statistical Society. Series B (Methodological). 1958 Jul;20(2):215-242. Disponible en: https://www.jstor.org/stable/2983890

[54] Cortes C, Vapnik V. Support-Vector Networks. Machine Learning [Internet]. 1995;20(3):273-97. Disponible en: https://link.springer.com/article/10.1007/BF00994018

[55] John GH, Langley P. Estimating Continuous Distributions in Bayesian Classifiers. In: Besnard P, Hanks S, editors. UAI [Internet]. Morgan Kaufmann; 1995. p. 338-45. Disponible en: http://www.isle.org/~langley/papers/flex.uai95.pdf

[56] Friedman JH. Greedy function approximation: a gradient boosting machine. Annals of Statistics. 2001;29(5):1189-1232. Disponible en: https://jerryfriedman.su.domains/ftp/trebst.pdf

[57] Chen T, Guestrin C. XGBoost: A Scalable Tree Boosting System. In: Krishnapuram B, Shah M, Smola AJ, Aggarwal CC, Shen D, Rastogi R, editors. KDD [Internet]. ACM; 2016. p. 785-94. Disponible en: https://dl.acm.org/doi/pdf/10.1145/2939672.2939785

[58] Ke G, Meng Q, Finley T, Wang T, Chen W, Ma W, et al. LightGBM: A Highly Efficient Gradient Boosting Decision Tree. In: Guyon I, Luxburg UV, Bengio S, Wallach H, Fergus R, Vishwanathan S, et al., editors. Advances in Neural Information Processing Systems 30 [Internet]. Curran Associates, Inc.; 2017. p. 3146-3154. Disponible en: http://papers.nips.cc/paper/6907-lightgbm-a-highly-efficient-gradient-boosting-decision-tree.pdf

[59] Breiman L, Friedman JH, Olshen RA, Stone CJ. Classification and Regression Trees. Wadsworth; 1984.

[60] Breiman L. Random forests. Machine learning. 2001;45:5-32. Disponible en: https://www.stat.berkeley.edu/~breiman/randomforest2001.pdf

[61] Rumelhart DE, Hinton GE, Williams RJ. Learning representations by back-propagating errors. Nature [Internet]. 1986;323(6088):533-536. Disponible en: https://www.nature.com/articles/323533a0

[62] Liu X-Y, Wu J, Zhou Z-H. Exploratory Undersampling for Class-Imbalance Learning. IEEE Transactions on Systems, Man, and Cybernetics Part B [Internet]. 2009;39(2):539-50. Disponible en: https://cs.nju.edu.cn/zhouzh/zhouzh.files/publication/tsmcb09.pdf

[63] Kuncheva LI. Combining Pattern Classifiers: Methods and Algorithms. Wiley; 2004.

[64] Wolpert DH. Stacked generalization. Neural networks. 1992;5(2):241-259. Disponible en: https://www.sciencedirect.com/science/article/abs/pii/S0893608005800231

[65] Bots SH, Peters SAE, Woodward M. Sex differences in coronary heart disease and stroke mortality: a global assessment of the effect of ageing between 1980 and 2010. BMJ Global Health [Internet]. 2017 Mar 27 [Consulta: 15 de mayo de 2024];2(2):e000298. Disponible en: https://doi.org/10.1136/bmjgh-2017-000298

[66] Gallucci, G., Tartarone, A., Lerose, R., Lalinga, A. V., & Capobianco, A. M. Cardiovascular risk of smoking and benefits of smoking cessation. Journal of thoracic disease [Internet]. 2020 Jul [Consulta: 15 de mayo de 2024];12(7), 3866–3876. Disponible en: https://doi.org/10.21037/jtd.2020.02.47

[67] Fernández de Bobadilla, J., Sanz de Burgoa, V., Garrido Morales, P., López de Sá, E., & en representación de los investigadores del estudio RETRATOS. Riesgo cardiovascular: evaluación del tabaquismo y revisión en atención primaria del tratamiento y orientación sanitaria. Atencion primaria [Internet]. 2011 Nov [Consulta: 15 de mayo de 2024]; 43(11), 595–603. Disponible en: https://doi.org/10.1016/j.aprim.2010.10.005

[68] Havranek EP, Mujahid MS, Barr DA, et al. Social Determinants of Risk and Outcomes for Cardiovascular Disease: A Scientific Statement From the American Heart Association. Circulation [Internet]. 2015 Sep 1 [Consulta: 15 de mayo de 2024];132(9):873-898. Disponible en: https://doi.org/10.1161/CIR.0000000000000228

[69] Lloyd-Jones DM, Allen NB, Anderson CAM, et al. Life's Essential 8: Updating and Enhancing the American Heart Association's Construct of Cardiovascular Health: A Presidential Advisory From the American Heart Association. Circulation [Internet]. 2022 Aug 2 [Consulta: 15 de mayo de 2024];146(5):e18-e43. Disponible en: https://doi.org/10.1161/CIR.0000000000001078

ANEXO I

En este anexo, se presenta una comparación entre los resultados obtenidos utilizando la imputación por ecuaciones encadenadas (MICE) y los resultados obtenidos eliminando todos los valores nulos del conjunto de datos original. El objetivo es determinar si la imputación MICE afecta negativamente el rendimiento de los modelos.

Para llevar a cabo esta comparación, se aplicó la técnica de muestreo RandomUnderSampler tanto al conjunto de datos con las 11 variables seleccionadas como al conjunto de datos con todas las variables disponibles. En ambos casos, se utilizaron dos versiones del conjunto de datos: una con imputación MICE, que contaba con 353,968 registros, y otra sin valores nulos, donde se eliminaron todas las instancias que contenían valores faltantes, resultando en un conjunto de datos de 320,437 registros.

Los resultados de los modelos se presentan en las siguientes tablas:

- Tabla A1.1: Rendimiento de los modelos con imputación MICE (conjunto de datos con 11 variables)
- Tabla A1.2: Rendimiento de los modelos sin valores nulos (conjunto de datos con 11 variables)
- Tabla A1.3: Rendimiento de los modelos con imputación MICE (conjunto de datos con todas las variables)
- Tabla A1.4: Rendimiento de los modelos sin valores nulos (conjunto de datos con todas las variables)

La Tabla A1.1 es la misma que aparece en el documento principal, la cual corresponde a la Tabla 4.3, pero se vuelve a incluir en el anexo para facilitar la comparación respecto a las métricas de rendimiento en el conjunto sin imputación.

Los modelos se entrenaron y evaluaron utilizando las mismas técnicas y métricas descritas en la sección principal del estudio. Se compararon los resultados obtenidos con la imputación MICE y los resultados obtenidos eliminando los valores nulos para determinar el impacto de la imputación en el rendimiento de los modelos.

Tabla A1.1. Rendimiento de los modelos con imputación MICE (conjunto de datos con 11 variables)

Modelo	Recall	Precisión	F1-score	AUC-ROC	AUC-PR	Tiempo de entrenamiento (min)	Tiempo búsqueda hiperparámetros (min)
Logistic Regression	0.7661	0.2437	0.3697	0.8034	0.34	0.01	1.84
Gaussian Naive Bayes	0.6251	0.2680	0.3752	0.7874	0.32	0.00	0.07
Light GBoost	0.7919	0.2366	0.3644	0.8061	0.35	0.01	1.22
Random Forest	0.7954	0.2347	0.3625	0.8051	0.34	0.06	4.14
Neural Network	0.8135	0.2295	0.3580	0.8060	0.34	2.87	34.86
Easy Ensemble	0.7538	0.2424	0.3669	0.8001	0.34	0.20	3.17
Voting	0.7542	0.2502	0.3757	0.8046	0.34	2.45	-
LinearSVC	0.7916	0.2356	0.3632	0.8025	0.34	0.04	3.71

Tabla A1.2. Rendimiento de los modelos sin valores nulos (conjunto de datos con 11 variables)

Modelo	Recall	Precisión	F1-score	AUC-ROC	AUC-PR	Tiempo de entrenamiento (min)	Tiempo búsqueda hiperparámetros (min)
Logistic Regression	0.7626	0.2457	0.3716	0.7983	0.33	0.00	1.45
Gaussian Naive Bayes	0.6151	0.2685	0.3738	0.7827	0.31	0.00	0.05
Light GBoost	0.7730	0.2456	0.3728	0.8001	0.33	0.01	1.04
Random Forest	0.7743	0.2443	0.3714	0.7993	0.33	0.10	3.38
Neural Network	0.7410	0.2534	0.3777	0.8007	0.33	0.96	29.24
Easy Ensemble	0.7550	0.2472	0.3725	0.7963	0.32	0.06	2.75
Voting	0.7354	0.2562	0.3800	0.7999	0.33	0.81	-
LinearSVC	0.7692	0.2444	0.3710	0.7982	0.33	0.00	2.94

Tabla A1.3. Rendimiento de los modelos con imputación MICE (conjunto de datos con todas las variables)

Modelo	Recall	Precisión	F1-score	AUC-ROC	AUC-PR	Tiempo de entrenamiento (min)	Tiempo búsqueda hiperparámetros (min)
Logistic Regression	0.7758	0.2693	0.3999	0.8286	0.39	0.01	4.01
Gaussian Naive Bayes	0.6059	0.2859	0.3885	0.7991	0.33	0.00	0.19
Light GBoost	0.7972	0.2627	0.3952	0.8310	0.39	0.01	1.90
Random Forest	0.8032	0.2600	0.3929	0.8294	0.38	0.08	7.69
Neural Network	0.8531	0.2377	0.3719	0.8315	0.39	0.54	27.03
Easy Ensemble	0.7620	0.2674	0.3959	0.8240	0.38	0.65	8.79
Voting	0.7425	0.2828	0.4096	0.8287	0.39	0.71	-
LinearSVC	0.7961	0.2608	0.3929	0.8281	0.39	0.04	4.45

Tabla A1.4. Rendimiento de los modelos sin valores nulos (conjunto de datos con todas las variables)

Modelo	Recall	Precisión	F1-score	AUC-ROC	AUC-PR	Tiempo de entrenamiento (min)	Tiempo búsqueda hiperparámetros (min)
Logistic Regression	0.7599	0.2703	0.3988	0.8209	0.37	0.01	3.46
Gaussian Naive Bayes	0.5878	0.2958	0.3935	0.8006	0.34	0.00	0.13
Light GBoost	0.7827	0.2625	0.3932	0.8228	0.37	0.01	1.89
Random Forest	0.7901	0.2599	0.3910	0.8214	0.37	0.15	7.39
Neural Network	0.8167	0.2486	0.3812	0.8225	0.37	0.35	23.70
Easy Ensemble	0.7539	0.2657	0.3929	0.8168	0.36	0.27	5.19
Voting	0.7655	0.2691	0.3982	0.8217	0.37	0.43	-
LinearSVC	0.7687	0.2657	0.3949	0.8203	0.37	0.09	4.07

Después de examinar los resultados presentados en las Tablas A1.1, A1.2, A1.3 y A1.4, se pueden extraer varias conclusiones sobre el impacto de la imputación MICE en el rendimiento de los modelos.

En primer lugar, al comparar las Tablas A1.1 y A1.2, que corresponden a los modelos entrenados con el conjunto de datos de 11 variables, se observa que la imputación MICE no solo no afecta negativamente el rendimiento, sino que incluso lo mejora ligeramente en la mayoría de los modelos. Por ejemplo, el modelo Light GBoost con imputación MICE (Tabla A1.1) obtiene un AUC-ROC de 0.8061 y un F1-score de 0.3644, mientras que sin imputación (Tabla A1.2), estos valores son de 0.8001 y 0.3728, respectivamente. Este patrón se repite en otros modelos, como Random Forest y Neural Network, donde la imputación MICE conduce a mejoras marginales en las métricas de rendimiento.

Por otro lado, al analizar las Tablas A1.3 y A1.4, correspondientes a los modelos entrenados con todas las variables, se observa un impacto aún más significativo de la imputación MICE. En este caso, todos los modelos con imputación MICE (Tabla A1.3) obtienen valores superiores en las

métricas de rendimiento en comparación con los modelos sin imputación (Tabla A1.4). Por ejemplo, el modelo Light GBoost con imputación MICE alcanza un AUC-ROC de 0.8310 y un F1-score de 0.3952, mientras que sin imputación, estos valores son de 0.8228 y 0.3932, respectivamente. Esta mejora se observa de manera consistente en todos los modelos cuando se utiliza la imputación MICE en el conjunto de datos con todas las variables.

Estos resultados indican que la imputación MICE no solo no perjudica el rendimiento de los modelos, sino que en realidad resulta beneficiosa al mejorar las métricas de evaluación. Una posible explicación para este fenómeno es que la imputación MICE permite aprovechar toda la información disponible en el conjunto de datos, incluyendo aquellas instancias con valores faltantes que de otro modo serían excluidas. Al estimar los valores faltantes de manera inteligente, la imputación MICE proporciona a los modelos una visión más completa de los datos, lo que les permite capturar mejor las relaciones y patrones subyacentes.

Además, es interesante notar que la mejora en el rendimiento es más pronunciada cuando se utilizan todas las variables en comparación con el conjunto de datos de 11 variables. Esto destaca que la imputación MICE es especialmente útil cuando se trabaja con un gran número de variables, ya que permite aprovechar al máximo la información disponible y mejorar la capacidad predictiva de los modelos.

Por tanto, los resultados presentados en las tablas del Anexo I respaldan la utilización de la imputación MICE como una técnica efectiva para manejar los valores faltantes en el conjunto de datos. La imputación MICE no solo no afecta negativamente el rendimiento de los modelos, sino que en realidad conduce a mejoras en las métricas de evaluación, especialmente cuando se trabaja con un gran número de variables.

Para complementar el análisis del impacto de la imputación MICE en el rendimiento de los modelos, se presentan a continuación las gráficas de SHAP values para los modelos con y sin imputación MICE, tanto para el conjunto de datos con 11 variables como para el conjunto de datos con todas las variables.

Las gráficas de SHAP values permiten visualizar la importancia y dirección de la influencia de cada variable en las predicciones de los modelos. Al comparar las gráficas de SHAP values de los modelos con y sin imputación MICE, se puede identificar si hay cambios significativos en la importancia o dirección de los factores de riesgo.

La Figura A1.1 muestra los SHAP values para el modelo entrenado con el conjunto de datos de 11 variables y con imputación MICE, mientras que la Figura A1.2 presenta los SHAP values para el modelo entrenado con el mismo conjunto de variables pero sin valores nulos.

Por otro lado, la Figura A1.3 muestra los SHAP values para el modelo entrenado con todas las variables y con imputación MICE, mientras que la Figura A1.4 presenta los SHAP values para el modelo entrenado con todas las variables pero sin valores nulos.

Estas figuras permiten realizar una comparación visual de la importancia y dirección de las variables en los modelos con y sin imputación MICE, y determinar si la imputación introduce factores de riesgo contraintuitivos o si omite factores cuya influencia ha sido validada. Para lo cual se va a utilizar Light

GBoost como modelo para evaluar las variables más importantes en las predicciones en las 4 figuras siguientes, dado que fue el modelo utilizado en los resultados finales.

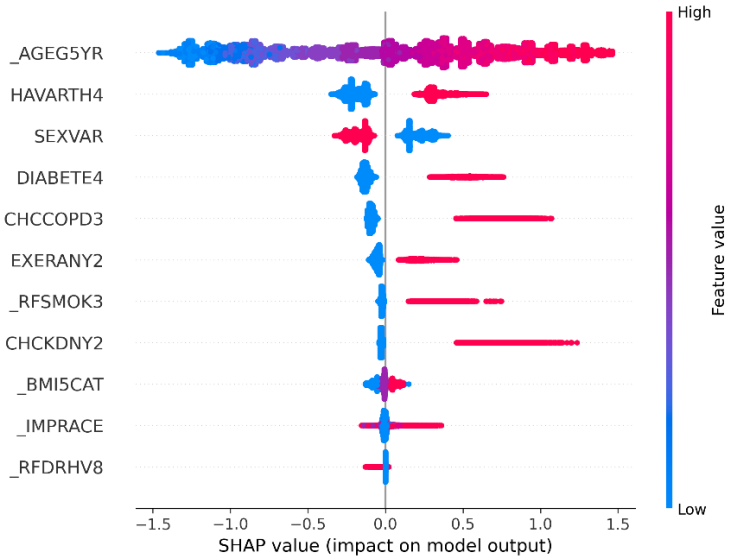


Figura A1.1. SHAP values para el modelo con imputación MICE (conjunto de datos con 11 variables)

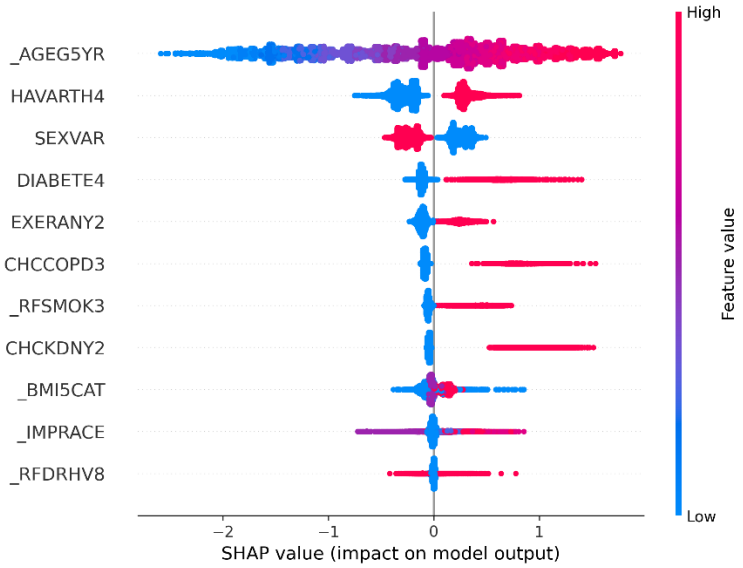


Figura A1.2. SHAP values para el modelo sin valores nulos (conjunto de datos con 11 variables)

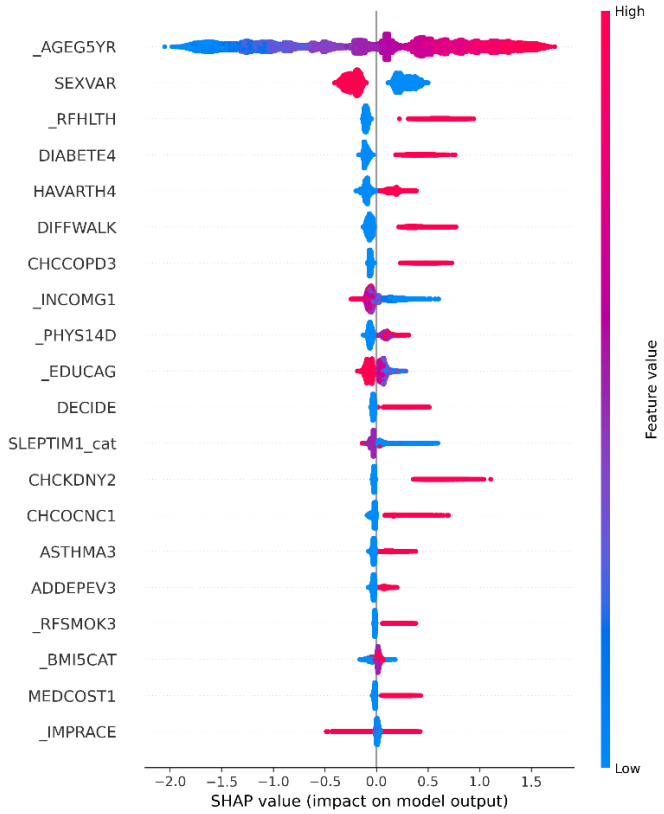


Figura A1.3. SHAP values para el modelo con imputación MICE (conjunto de datos con todas las variables)

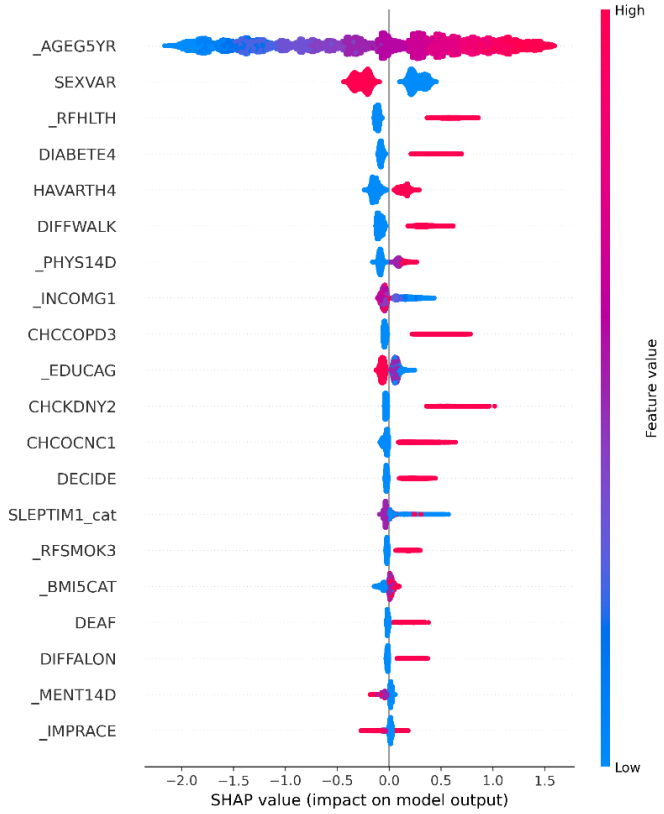


Figura A1.4. SHAP values para el modelo sin valores nulos (conjunto de datos con todas las variables)

Después de examinar las gráficas de SHAP values para los modelos con y sin imputación MICE, tanto para el conjunto de datos con 11 variables como para el conjunto de datos con todas las variables, se pueden hacer las siguientes puntualizaciones.

En primer lugar, al comparar las Figuras A1.1 y A1.2, que corresponden a los modelos entrenados con el conjunto de datos de 11 variables, se observa que la imputación MICE no introduce cambios significativos en la importancia y dirección de las variables. En ambas figuras, las variables más influyentes son AGEG5YR, HAVARTH4, SEXVAR y DIABETE4, y su dirección de influencia (positiva o negativa) se mantiene consistente. Esto indica que la imputación MICE no altera sustancialmente la interpretación de los factores de riesgo cuando se trabaja con un conjunto reducido de variables.

Por otro lado, al comparar las Figuras A1.3 y A1.4, correspondientes a los modelos entrenados con todas las variables, se observan algunas diferencias más notables. Aunque las variables más importantes (AGEG5YR, SEXVAR, RFHLTH, DIABETE4) se mantienen consistentes en ambas figuras, la imputación MICE parece introducir algunas variables adicionales con una influencia considerable, como CHCCOPD3, EDUCAG y DECIDE. Esto apunta que la imputación MICE puede ayudar a capturar la influencia de variables que podrían haber sido omitidas debido a los valores faltantes cuando se trabaja con un conjunto de datos más completo.

Es importante destacar que, en todas las figuras, no se observan factores de riesgo contraintuitivos o la omisión de factores cuya influencia ha sido validada. Las variables más relevantes, como la edad (AGEG5YR), el sexo (SEXVAR), la presencia de enfermedades crónicas (DIABETE4, HAVARTH4) y la percepción de la salud (RFHLTH), son consistentes con el conocimiento médico establecido sobre los factores de riesgo de enfermedades cardiovasculares.

Además, la dirección de la influencia de las variables se mantiene consistente en todas las figuras. Por ejemplo, valores altos de AGEG5YR (correspondientes a edades avanzadas) tienen un impacto positivo en la predicción de enfermedades cardiovasculares, mientras que valores bajos de RFHLTH (correspondientes a una percepción de salud buena) tienen un impacto negativo.

De manera que el análisis de las gráficas de SHAP values respalda la utilización de la imputación MICE como una técnica efectiva para manejar los valores faltantes en el conjunto de datos. La imputación MICE no introduce factores de riesgo contraintuitivos ni omite factores validados, y puede ayudar a capturar la influencia de variables adicionales cuando se trabaja con un conjunto de datos más completo. Además, la consistencia en la dirección de la influencia de las variables en todas las figuras refuerza la robustez de los resultados obtenidos con la imputación MICE.