UNIVERSIDAD ESTATAL PENÍNSULA
DE SANTA ELENA
FACULTAD DE SISTEMAS Y TELECOMUNICACIONES
INSTITUTO DE POSTGRADO

TITULO DEL TRABAJO DE TITULACIÓN

Modelos de Aprendizaje de máquina para medir el riesgo de contraer enfermedades cardiovasculares

AUTOR

Alejandro Roca, Katherine Viviana

TRABAJO DE TITULACIÓN

Previo a la obtención del grado académico en
MAGISTER EN TECNOLOGÍAS DE LA INFORMACIÓN

TUTOR

Rosero Vásquez, Shendry, Mgtr.

Santa Elena, Ecuador

Año 2024

UNIVERSIDAD ESTATAL PENÍNSULA DE SANTA ELENA
FACULTAD DE SISTEMAS Y TELECOMUNICACIONES
INSTITUTO DE POSTGRADO

TRIBUNAL DE SUSTENTACIÓN

Firmado electrónicamente por:
ALICIA GERMANIA ANDRADE VERA

Ing. Alicia Andrade Vera, Mgtr.
COORDINADORA DEL PROGRAMA

SHENDRY BALMORE ROSERO VASQUEZ

Firmado digitalmente porSHENDRY BALMORE ROSERO VASQUEZ
DN: cn=SHENDRY BALMORE ROSERO VASQUEZ gn=SHENDRY BALMORE c=EC
Motivo:Apruebo este documento
Ubicación:
Fecha:2024-05-02 18:12-05:00

Ing. Shendry Rosero Vásquez, Mgtr.
TUTOR

Firmado electrónicamente por:
DELIA ISABEL CARRION LEON

Ing. Delia Carrión León, Mgtr.
DOCENTE
ESPECIALISTA 1

Firmado electrónicamente por:
JUAN PABLO AMON SALINAS

Ing. Juan Pablo Amón Salinas, Mgtr.
DOCENTE
ESPECIALISTA 2

Firmado electrónicamente por:
MARIA MARGARITA RIVERA GONZALEZ

Abg. María Rivera, Mgtr.
SECRETARIA GENERAL
UPSE

UNIVERSIDAD ESTATAL PENÍNSULA DE SANTA ELENA
FACULTAD DE SISTEMAS Y TELECOMUNICACIONES
INSTITUTO DE POSTGRADO

CERTIFICACIÓN

Certifico que luego de haber dirigido científica y técnicamente el desarrollo y estructura final del trabajo, este cumple y se ajusta a los estándares académicos, razón por el cual apruebo en todas sus partes el presente trabajo de titulación que fue realizado en su totalidad por KATHERINE VIVIANA ALEJANDRO ROCA, como requerimiento para la obtención del título de Magister en Tecnologías de la Información.

TUTOR

SHENDRY BALMORE ROSERO VASQUEZ

Firmado digitalmente porSHENDRY BALMORE ROSERO VASQUEZ
DN: cn=SHENDRY BALMORE ROSERO VASQUEZ gn=SHENDRY BALMORE c=EC
Motivo:Apruebo este documento
Ubicación:
Fecha:2024-05-02 18:13-05:00

Ing. Shendry Rosero Vásquez, Mgtr.

21 días del mes de marzo del año 2024

UNIVERSIDAD ESTATAL PENÍNSULA DE SANTA ELENA
FACULTAD DE SISTEMAS Y TELECOMUNICACIONES
INSTITUTO DE POSTGRADO

DECLARACIÓN DE RESPONSABILIDAD

Yo, **KATHERINE VIVIANA ALEJANDRO ROCA**

DECLARO QUE:

El trabajo de Titulación, **MODELOS DE APRENDIZAJE DE MÁQUINA PARA MEDIR EL RIESGO DE CONTRAER ENFERMEDADES CARDIOVASCULARES**, previo a la obtención del título en Magister en Tecnologías de la Información, ha sido desarrollado respetando derechos intelectuales de terceros conforme las citas que constan en el documento, cuyas fuentes se incorporan en las referencias o bibliografías. Consecuentemente este trabajo es de mi total autoría.

En virtud de esta declaración, me responsabilizo del contenido, veracidad y alcance del Trabajo de Titulación referido.

Santa Elena, a los 21 días del mes de marzo del año 2024

EL AUTOR

Firmado electrónicamente por:
KATHERINE VIVIANA ALEJANDRO ROCA

Katherine Viviana Alejandro Roca

UNIVERSIDAD ESTATAL PENÍNSULA DE SANTA ELENA
FACULTAD DE CIENCIAS DE LA INGENIERÍA
INSTITUTO DE POSTGRADO

CERTIFICACIÓN DE ANTIPLAGIO

Certifico que después de revisar el documento final del trabajo de titulación denominado MODELOS DE APRENDIZAJE DE MÁQUINA PARA MEDIR EL RIESGO DE CONTRAER ENFERMEDADES CARDIOVASCULARES, presentado por el estudiante, KATHERINE VIVIANA ALEJANDRO ROCA fue enviado al Sistema Antiplagio COMPILATIO, presentando un porcentaje de similitud correspondiente al 3%, por lo que se aprueba el trabajo para que continúe con el proceso de titulación.


CERTIFICADO DE ANÁLISIS
magister

tesis - plagio

3%
Textos sospechosos

3% Similitudes
< 1% similitudes entre comillas
0% entre las fuentes mencionadas
< 1% Idiomas no reconocidos

Nombre del documento: tesis - plagio.docx
ID del documento: 98e40769616ecc5acf4fa0a4b25212606d6f6699
Tamaño del documento original: 1,49 MB

Depositante: SHENDRY BALMORE ROSERO VASQUEZ
Fecha de depósito: 21/3/2024
Tipo de carga: interface
fecha de fin de análisis: 21/3/2024

Número de palabras: 14.504
Número de caracteres: 97.243

Ubicación de las similitudes en el documento:

TUTOR

SHENDRY BALMORE ROSERO VASQUEZ

Firmado digitalmente por SHENDRY BALMORE ROSERO VASQUEZ
DN: c=SHENDRY BALMORE ROSERO VASQUEZ gn=SHENDRY BALMORE c=EC
Motivo: Apruebo este documento
Ubicación:
Fecha: 2024-05-02 18:13-05:00

Ing. Shendry Rosero Vásquez, Mgtr.

UNIVERSIDAD ESTATAL PENÍNSULA DE SANTA ELENA
FACULTAD DE SISTEMAS Y TELECOMUNICACIONES
INSTITUTO DE POSTGRADO

AUTORIZACIÓN

Yo, **Katherine Viviana Alejandro Roca**

Autorizo a la Universidad Estatal Península de Santa Elena, para que haga de este trabajo de titulación o parte de él, un documento disponible para su lectura consulta y procesos de investigación, según las normas de la Institución.

Cedo los derechos en línea patrimoniales de artículo profesional de alto nivel con fines de difusión pública, además apruebo la reproducción de este artículo académico dentro de las regulaciones de la Universidad, siempre y cuando esta reproducción no suponga una ganancia económica y se realice respetando mis derechos de autor

Santa Elena, a los 21 días del mes de marzo del año 2024

EL AUTOR

Firmado electrónicamente por:
KATHERINE VIVIANA ALEJANDRO ROCA

Katherine Viviana Alejandro Roca

AGRADECIMIENTO

A mis queridas hijas, Kenia y Karol, por su sacrificio y comprensión durante esta etapa académica. A pesar de los fines de semana dedicados al estudio, ustedes estuvieron a mi lado con amor y apoyo incondicional. Su paciencia y aliento significaron todo para mí.

A mi esposo, Andrés, gracias por ser mi roca durante este viaje. Cada logro que alcanzamos juntos fortalece nuestro vínculo. Tu amor y apoyo constante son mi mayor inspiración.

A mi hermana, Keyla, tu apoyo inquebrantable en momentos decisivos, es invaluable. Tu generosidad y disposición para ayudar nunca serán olvidadas.

A mi tutor de tesis, gracias por su orientación experta y paciencia infinita.

Katherine Viviana, Alejandro Roca

DEDICATORIA

A mi amado padre y mi querida suegra, quienes dejaron huellas imborrables en mi corazón.

Papá, Teófilo Inocente Alejandro Reyes, tu incansable trabajo y dedicación siguen inspirándome cada día.

Suegra, Yolanda Dolores Guaranda Sánchez, tu inquebrantable apoyo en mi etapa académica siempre será recordado con gratitud.

En sus memorias.

Katherine Viviana, *Alejandro Roca*

ÍNDICE GENERAL

ÍNDICE DE TABLAS

ÍNDICE DE FIGURAS

RESUMEN

La detección temprana y precisa de enfermedades cardiovasculares es fundamental para la prevención y tratamiento efectivo de estas condiciones de salud crítica. En esta tesis, se exploró el uso de modelos de aprendizaje automático para mejorar la detección de enfermedades cardiovasculares, centrándose en la optimización del rendimiento del modelo. Se contó con una data de más de 300000 registros de factores de riesgo cardiovascular que mediante el software libre Orange Data Mining facilitó la carga, exploración y comprensión de los datos antes del entrenamiento. Además, de proporcionar los evaluadores de rendimiento entre los modelos seleccionados. Luego de la optimización del modelo más prometedor, mediante técnicas de preprocesamiento, balanceo de clases y validación cruzada, Regresión Logística pasó de un recall de 0.06 a 0.79 a la clase minoritaria. Al combinar estas estrategias, se mejoró la capacidad del modelo para detectar de manera equitativa tanto casos positivos como negativos de enfermedad cardiovascular.

Palabras claves: Enfermedad Cardiovascular, Optimización, Regresión Logística

ABSTRACT

Early and accurate detection of cardiovascular diseases is essential for the prevention and effective treatment of these critical health conditions. In this thesis, the use of machine learning models to improve the detection of cardiovascular diseases was explored, focusing on optimizing model performance. There was data from more than 300,000 records of cardiovascular risk factors that, using the free Orange Data Mining software, facilitated the loading, exploration and understanding of the data before training. In addition, to provide performance evaluators between the selected models. After the optimization of the most promising model, through preprocessing techniques, class balancing and cross-validation, Logistic Regression went from a recall of 0.06 to 0.79 to the minority class. By combining these strategies, the model's ability to equally detect both positive and negative cases of cardiovascular disease was improved.

Keywords: Cardiovascular Disease, Optimization, Logistic Regression

INTRODUCCIÓN

Las enfermedades cardiovasculares (ECV) son la principal causa de muerte en el mundo (Hernández-Martínez et al., 2020). Se considera que una persona sufre de ECV cuando tiene condiciones que afectan la estructura y el funcionamiento del corazón, dentro de este grupo de enfermedades podemos mencionar la cardiopatía coronaria, la hipertensión, la cardiopatía congestiva, las enfermedades cerebrovasculares, la cardiopatía reumática, entre otras (OMS, 2017). Indistintamente de la afectación cardiaca, esta es una epidemia que plantea un gran desafío, ya que tienen un impacto devastador en la salud humana y en los sistemas de salud a nivel mundial.

En cuanto a lo antes expuesto, definir los factores de riesgo cardiovascular (FRC) como una característica biológica, condición y/o comportamiento es primordial para establecer la intensidad de la intervención y las directrices médicas a nivel farmacológico que mengüen la probabilidad de padecer o fallecer a causa de una ECV (Areiza et al., 2018). Cabe mencionar que existen factores que son modificables, tratados o controlados como el tabaquismo, alcoholismo, hábitos alimenticios, obesidad, sedentarismo y otros factores que no son modificables como edad, sexo, predisposiciones genéticas, por lo tanto, mientras más factores presente un paciente mayor es el riesgo de contraer alguna de estas afectaciones (Menéndez, 2023).

La Organización Mundial de la Salud (ONU) y la Organización Panamericana de la Salud (OPS), al indicar que cada año hay más muertes por ECV que por otras causas (PAHO, 2023a). Datos de la OMS indican que estas afectaciones se cobran 17,9 millones de vidas al año (OMS, 2023). Por lo tanto, su estudio y detección temprana representan un desafío de salud global, siendo estas la piedra angular de una prevención y un tratamiento eficaz, que al ser abordadas podría disminuir los índices alarmantes de morbilidad y mortalidad.

En América Latina, se estima que la cantidad de muertes en la región atribuibles a ECV aumentará en más del 60% entre los años 2000 y 2020; en comparación, al aumento en el mundo que es de solo el 5%, convirtiéndola así en la principal causa de muerte en Latinoamérica (López-Jaramillo et al., 2018). Cabe definir que las causas principales de este aumento están relacionadas a los cambios del estilo de vida de los habitantes de la región.

En el caso de los países de LATAM, se calcula que la hipertensión arterial es el principal factor de riesgo de ECV y es el responsable de 1,6 millones de muertes antes de los 70

años (Camafort et al., 2021). Sin embargo, en América Latina los datos disponibles son escasos y difícil de obtener para fines de investigación.

En Ecuador, las enfermedades cardiovasculares son la primera causa de consultas en los establecimientos del Ministerio de Salud Pública (MSP). Entre 2018 y 2022 se registró un promedio anual de 247 000 primeras consultas y casi 1,5 millones de consultas subsecuentes. También son la primera causa de muerte en el país, acumulando el 25% del total de decesos anuales, además se menciona que del 2018 al 2021 el país registró 91271 defunciones por complicaciones cardiovasculares. Los decesos entre los hombres alcanzaron el 53% (CEAP ESPOL, 2023).

Según la encuesta STEPS del 2018(PAHO, 2023b), en Ecuador, se identificó que el 19.8% de la población tenía hipertensión, siendo este un factor de riesgo cardiovascular, de esos el 56.3% no tomaba medicamento para la presión arterial alta, además se determinó que el 37.9% y 25.7% tenía sobrepeso y obesidad, respectivamente que son considerados factores de riesgo modificables (PAHO, 2023b). Estos datos alarmantes, permiten la implementación de medidas para la prevención y el control de las enfermedades no transmisibles como la ECV.

Por otro lado, en Ambato-Ecuador, en el 2019 se realizó un estudio para determinar la prevalencia y asociación de factores de riesgo de ECV en la población de estudio (68.3% mujeres y 31.7% hombres), quedando en evidencias que el consumo de alcohol es un factor de riesgo alto con el 75%, el sedentarismo es del 44%, historia familiar 44% y tabaquismo 37%. Por su parte, se considera esos factores relevantes para padecer algún tipo de enfermedad cardiovascular (Ramos et al., 2021).

Actualmente, el riesgo de contraer enfermedades cardiovasculares debido a factores relacionados con el estilo de vida es un asunto de gran relevancia en el sistema de salud pública y privada; esto debido al aumento de casos con ECV. Del mismo modo, genera una carga económica considerable en comparación con los valores de vida en Ecuador (Castillo et al., 2017), donde una evaluación integral, confiable y válida en atención primaria puede garantizar la prestación de servicios de alta calidad centrados en las necesidades del paciente (Kringos et al., 2019). Por lo tanto, promover cambios positivos en el estilo de vida es una oportunidad de prevenir y reducir la probabilidad de contraer dicha enfermedad que permitirá reducir los índices de mortalidad a nivel mundial (The Texas Heart Institute, 2023).

En conclusión, las bibliografías antes citadas, enfatizan que: “Con un pronóstico oportuno y una consideración exhaustiva del historial médico y el estilo de vida del paciente, es posible predecir las ECV y tomar medidas preventivas para eliminar o controlar esta enfermedad potencialmente mortal” (Saputra et al., 2023). Dado esta problemática, a nivel mundial, se notó la creciente necesidad de diagnósticos precisos y eficientes para diversas enfermedades plantea desafíos significativos debido a la complejidad de los mecanismos subyacentes y la variabilidad de los síntomas entre la población de pacientes (Ahsan et al., 2022). Sin embargo, el avance en el campo del aprendizaje automático (ML), una rama de la inteligencia artificial (IA), ha permitido a investigadores, médicos y pacientes abordar estos desafíos de manera más efectiva.

Para ese fin, los científicos han desarrollado una amplia gama de algoritmos de diagnóstico artificialmente inteligentes para la detección de diversas enfermedades (DinuA & Joseph, 2017), simplificando no solo el proceso de diagnóstico, sino que mejora el análisis de los síntomas clínicos y de laboratorio con datos adecuados (Singh et al., 2021) , reduciendo así la tasa de falsos positivos y permitiendo una efectiva atención médica (Raval et al., 2016).

El presente trabajo, busca ampliar la forma de recopilación y análisis de los datos, al permitir el uso de la tecnología para este fin. Además, de ser base para futuras investigaciones.

Formulación del problema de investigación

General

¿Cómo los modelos de aprendizaje de máquina mejorarían la detección de enfermedades cardiovasculares a partir del análisis de los factores de riesgo?

Específicas

- ¿Cuáles son los modelos predictivos de Aprendizaje de Máquina para el diagnóstico de enfermedades cardiovasculares?
- ¿Cuál es la correlación de los factores de estilo de vida de las personas diagnosticadas con enfermedades cardiovasculares?
- ¿Cuál es la precisión de un modelo de Aprendizaje de Máquina en el diagnóstico de una persona con enfermedades cardiovasculares?

Objetivo General:

Analizar el desempeño de modelos de aprendizaje automático en la estimación de la vulnerabilidad individual al desarrollo de enfermedades cardiovasculares, fundamentando este análisis en el estudio de factores de riesgo específico.

Objetivos Específicos:

- Examinar modelos de aprendizaje automático supervisados para la predicción del riesgo de enfermedades cardiovasculares, basándose en su capacidad de procesamiento y análisis de factores de riesgo relevantes.
- Evaluar la eficacia de los modelos de aprendizaje automático supervisados seleccionados, utilizando métricas de evaluación estandarizadas, tales como precisión, sensibilidad, especificidad y el área bajo la curva (AUC), para la determinación de su rendimiento en la predicción del riesgo de enfermedades cardiovasculares.
- Optimizar el desempeño del modelo de aprendizaje automático supervisado más prometedor mediante la afinación de hiperparámetros y la implementación de técnicas de balanceo de clases para mejorar su generalización y precisión en la predicción de riesgos cardiovasculares en poblaciones diversas.

Planteamiento hipotético

HIPOTESIS

Se postula que los modelos de aprendizaje automático supervisados, al ser entrenados con datos que incluyen factores de riesgo específicos tales como el índice de masa corporal (BMI), historial de tabaquismo, antecedentes de diabetes, edad, entre otros indicadores relevantes, tendrán la capacidad predictiva para identificar la presencia o ausencia de enfermedades cardiovasculares mediante la aplicación de técnicas avanzadas de preprocesamiento y ajuste de hiperparámetros.

CAPÍTULO 1. MARCO TEÓRICO REFERENCIAL

1.1. Revisión de literatura

Las principales tecnologías para predecir enfermedades de manera temprana, explica el autor (ARRUBLA HOYOS, 2022) son: Machine Learning, Big data y la Inteligencia Artificial mediante el desarrollo y evaluación de modelos que apoyen las decisiones médicas. Es por eso que el autor (Mora, 2022) menciona la importancia de encontrar la causalidad en el área de la salud, donde los modelos estadísticos clásicos no logran evidencia de causa y efecto con base sólida, caso contrario ocurre en ML que analiza grandes cantidades de datos con mayor precisión y velocidad para la toma de decisiones.

En medicina los modelos de aprendizaje computacional se han aplicado con éxito, tanto que se ha evidenciado en el aumento de investigaciones y desarrollo de modelos de ML aplicados al diagnóstico médico de diferentes enfermedades (González, 2015). Los beneficios de ML son prometedoras debido a que mejora la detección, diagnóstico y el seguimiento de diferentes enfermedades (Sajda, 2006).

En el trabajo "Socioeconomic factors and machine learning algorithms applied to neglected diseases risk prediction", menciona que el análisis de las relaciones entre las variables socioeconómicas y las enfermedades desantendidas puede ayudar a crear políticas de salud pública para la reducción de casos (Gioia et al., 2022).

Por otro lado, los autores (Ordoñez-Guillen et al., 2023) en su trabajo han evaluado modelos para clasificar los subtipos de diabetes tipo 2 mediante técnicas de Machine Learning. Para ese fin han implementado una metodología de 3 etapas:

- Preprocesamiento de la base de datos.
- Análisis descriptivo para identificar grupos de observaciones.
- Evaluación de modelos de clasificación.

(Saputra et al., 2023) en su estudio "Optimización de hiperparámetros para un sistema de pronóstico basado en datos de enfermedades cardiovasculares" usó un conjunto de datos de 918 pacientes de un hospital de Estados Unidos de entre 28 y 77 años, que mediante modelos de Aprendizaje no supervisado como K-means y Agrupación Jerárquica con el software Orange determinó que el conjunto de datos se puede dividir en 2 grupos, además se determinó que las redes neuronales tuvieron mejor rendimiento con una precisión de 0.90 con respecto a otros algoritmos.

1.2. Desarrollo teórico y conceptual

1.2.1. Estilo de vida

La Organización Mundial de la Salud (OMS) citado por (Rafael Rondanelli & Rafael Rondanelli, 2014) define al estilo de vida como:

> El conjunto de patrones de comportamiento determinados por la interacción entre la persona y la sociedad, además de las condiciones de vida socioeconómicas y ambientales.

La definición de estilo de vida es diversa, sin embargo, según el Glosario de promoción de la salud de la Junta de Andalucía (1986) citado por (Hellín Gómez, 2003) la define como:

> La forma de vida de una persona consiste en reacciones habituales y patrones de comportamiento que se forman durante la socialización. Estas pautas se aprenden en asociación con padres, compañeros, amigos y hermanos o a través de influencias en la escuela, los medios de comunicación, etc.

Considerando estos conceptos, se establece que el estilo de vida de una persona se moldea en base a las características individuales, sociales y el medio en que se desenvuelve; siendo esto uno de los principales determinantes de la salud de los habitantes de los países. El estilo de vida no saludable es un aspecto que predomina en las personas diagnosticadas con algún tipo de enfermedad cardiovascular (Suarez Villa et al., 2020).

1.2.2. Tipos de estilo de vida

Para fines investigativos se ha definido varios tipos de estilo de vida, tales como: Estilo de vida saludable, sedentaria, minimalista, activo y eco.

El autor Dellert S. Elliot (1993) tomó en consideración aspectos que afectan a la salud y aquellos que la protegen para definir al Estilo de vida saludable, de igual forma lo describe como “un conjunto de patrones de conducta que caracterizan la manera general de vivir de un individuo o grupo” (Menor Rodríguez, 2017).

Para dar una definición exacta al Estilo de vida sedentaria se considera definir sedentarismo o sedentario como cualidades propias del comportamiento humano donde se evidencia la ausencia de gasto energético en actividades que lo requieren, además acota

finalmente indica que el sedentarismo es la cuarta causa de muerte en el mundo (Sánchez-Guette et al., 2019).

El autor (Sánchez Pérez, 2015) considera que un Estilo de vida activo es sinónimo de saludable, además de traer beneficios en la salud física y mental del individuo. Considera que estos hábitos se adquieren en la infancia y se desarrollan a lo largo de la vida convirtiéndose en comportamientos propios de cada persona.

Por otra parte, el consumo ecológico según (Cervellon y Lindsey, 2011) y citado por (Viñas, 2019) refiere al termino Estilo de vida eco como un fenómeno consumista de productos ecológicos y sostenible que genera bienestar y salud.

Indiferente del estilo de vida de una persona, esto se puede llegar a modificar de acuerdo con las posibilidades del individuo y la necesidad de querer un cambio que genere un desarrollo correcto, para disfrutar una vida plena en años futuros.

1.2.3. Enfermedades Cardiovasculares

Son un conjunto de desórdenes del corazón y de los vasos sanguíneas (Cedeño et al., 2022), además de revelar que:

> "Las enfermedades cardiovasculares resultan las más comunes, graves y de mayor riesgo en términos de mortalidad y morbilidad en gran parte del mundo. Constituyendo un problema de salud prevenible si se tienen en cuenta sus factores de riesgo. (Delgado & Lara, 2021)"

1.2.4. Riesgo Cardiovascular

El riesgo cardiovascular se define como la probabilidad que un individuo pueda desarrollar una enfermedad, evento o accidente de tipo cardiovascular durante un periodo de tiempo específico (Vanuzzo et al., 2008).

1.2.5. Factores de riesgo cardiovasculares

Los factores de riesgo son características o comportamientos que aumentan la probabilidad de desarrollar alguna enfermedad, que al ser abordados de manera eficaz puede disminuir su índice de morbilidad y mortalidad (Croyle & Jemmott, 1991).

Los factores de riesgo cardiovasculares son aquellos signos biológicos o hábitos adquiridos que se presentan con mayor frecuencia en los pacientes con alguna enfermedad cardiovascular (Elizondo, 2020).

Con esa finalidad, se han realizado estudios de carácter investigativo que han identificado variables que aumentan la probabilidad de contraer algún tipo de ECV, es decir que mientras más factores una persona presenta, mayor es el riesgo cardiovascular (The Texas Heart Institute, 2023).

En (Ruiz, 2014) destaca factores modificables (tabaquismo, sedentarismo, obesidad, diabetes, alcoholismo) y no modificables (raza, sexo, edad, antecedentes familiares).

Consumo de Tabaco

Consumir tabaco implica inhalar, exhalar o sostener un producto de tabaco encendido (Ashipala et al., 1d. C.) y que llega a ser adictiva a causa de la sustancia química, nicotina. El estudio de Framingham ha demostrado que los fumadores activos y pasivos tienen un riesgo notable de desarrollar eventos cardiovasculares (Inoue, 2004) caso contrario, dejar de fumar puede demostrar beneficios para la mejora de la salud y el costo de la atención médica.

Consumo de Alcohol

El consumo de alcohol presenta una relación compleja y multifacética con las enfermedades cardiovasculares (CV), donde tanto los patrones de consumo excesivo como los de consumo moderado juegan roles distintos en la salud cardiovascular (Rehm & Roerecke, 2017). Por un lado, episodios de consumo excesivo de alcohol, ya sean ocasionales o crónicos, se asocian con efectos perjudiciales en una amplia gama de enfermedades CV; caso contrario, el consumo ligero a moderado ha demostrado tener efectos protectores, particularmente contra la cardiopatía isquémica y el accidente cerebrovascular isquémico, posiblemente por la mitigación de factores de riesgo hemostáticos y la reducción de la aterosclerosis e inflamación. (Piano, 2017)

Sin embargo, la interpretación de estos beneficios debe hacerse con cautela, equilibrando cuidadosamente los riesgos y beneficios, y considerando factores individuales como variantes genéticas, condiciones socioeconómicas, y posibles interacciones con medicamentos en pacientes cardíacos.

Consumo de frutas y vegetales

La relación entre el consumo de frutas y verduras y las enfermedades cardiovasculares ha sido objeto de varios estudios que destacan su importancia en la prevención de estas enfermedades. Un metanálisis de estudios de cohortes prospectivos encontró una

asociación inversa entre el consumo de frutas y verduras y el riesgo de enfermedad cardiovascular (ECV). Los resultados mostraron que aquellos que consumían 800 g por día de frutas y verduras tenían el riesgo más bajo de ECV. Este hallazgo respalda las recomendaciones de una alta ingesta de frutas y verduras para reducir el riesgo de ECV. (Zhan et al., 2017)

Además, estudios epidemiológicos han indicado que el consumo de frutas está inversamente relacionado con el riesgo de enfermedades cardiovasculares previniendo y facilitando la restauración de la morfología y funciones cardíacas y vasculares después de una lesión. Se han identificado varias frutas, como uva, arándano, granada, manzana, espino y aguacate, que poseen una potente acción protectora cardiovascular. (Zhao et al., 2017). Un análisis de veintitrés estudios con 18047 pacientes con ECV, muestra que un mayor consumo de frutas y verduras totales se asocia significativamente con un menor riesgo de enfermedad coronaria (Gan et al., 2015).

Cáncer

En varios estudios se enfatiza que el cáncer puede influir en el riesgo cardiovascular y cómo los tratamientos pueden afectar la salud del corazón. Uno de los hallazgos revela que mantener una buena salud cardiovascular (CV) puede reducir significativamente el riesgo de desarrollar cáncer en el futuro (Lau et al., 2021).

En cuanto a los sobrevivientes de cáncer, se menciona que tienen un mayor riesgo de desarrollar enfermedades cardiovasculares debido a una serie de efectos secundarios del tratamiento contra el cáncer, como la cardiotoxicidad, y los factores de riesgo cardiovascular modificables, como la hipertensión, la diabetes, la obesidad y el tabaquismo (Agmon Nardi & Iakobishvili, 2018). También se menciona que el cáncer y las enfermedades cardiovasculares comparten factores de riesgo y biología similares, lo que aumenta aún más la probabilidad de problemas cardiovasculares en los sobrevivientes de cáncer (Zheng et al., 2017).

Depresión

La depresión emerge como un factor de riesgo cardiovascular significativo, incrementando el riesgo de eventos cardíacos adversos, incluyendo el infarto agudo de miocardio y la muerte súbita cardíaca (Silverman et al., 2019).

La fisiopatología y el estilo de vida desfavorables en pacientes con depresión contribuyen a su mayor riesgo. Todos estos aspectos, aumentan el riesgo de morbilidad y mortalidad cardiovascular en aproximadamente un 80%. (Penninx, 2017)

Artritis

La artritis inflamatoria (AR) presenta un riesgo cardiovascular (CV) aumentado en un 50%, lo que conlleva una mayor mortalidad y morbilidad. Por tanto, es crucial evaluar y gestionar este riesgo en pacientes con AR (Hannawi & Al Salmi, 2021). Este exceso de riesgo CV está estrechamente ligado a la gravedad de la AR y la inflamación crónica asociada. Estudios epidemiológicos han revelado que los pacientes con AR tienen una mayor probabilidad de desarrollar enfermedad cardíaca isquémica de forma silenciosa, así como de experimentar insuficiencia cardíaca y muerte súbita en comparación con aquellos sin esta condición (DeMizio & Geraldino-Pardilla, 2020).

Diabetes

El incremento en el riesgo cardiovascular se manifiesta tanto en la aparición de arritmias cardíacas como en la muerte súbita cardíaca (MSC), siendo la principal causa de mortalidad entre los pacientes diabéticos (Abuelgasim et al., 2021). A pesar de que se ha reconocido durante décadas que la diabetes es un importante factor de riesgo cardiovascular, las razones exactas de esta asociación aún no se comprenden completamente (Eckel et al., 2021).

Finalmente, estos factores son comunes y sirven para evaluar el riesgo cardiovascular absoluto de la población en general (Lacey et al., 2017).

1.2.6. Políticas de Salud

La intervención eficaz para modificar conductas individuales o colectivas, según (López-Jaramillo et al., 2018) es uno de los puntos a considerar en un modelo de atención de salud integral que ayudaría a disminuir el desarrollo de enfermedades cardiovasculares. Sin embargo, el poco conocimiento de la población sobre los factores de riesgo cardiovascular se evidenció en el estudio "Conocimiento y factores de riesgo cardiovascular en pacientes ambulatorios" realizado por (Areiza et al., 2018) donde se reveló que solo el 43% de los pacientes conoce el tema.

Para esto es indispensable estudios donde se analicen variables sociodemográficas, clínicas-cardiovasculares para determinar los factores asociados a las ECV, como se detalla en (Hernández-Martínez et al., 2020), cuya finalidad es:

> "Crear políticas de salud orientadas a crear entornos propicios para que las opciones saludables se encuentren disponibles y sean asequibles resultan esenciales para motivar a las personas a adoptar y mantener comportamientos saludables" (OMS, 2023)

Adicionalmente, la disponibilidad de servicios para prevenir, diagnosticar y tratar enfermedades es clave para reducir el número de muertes y la discapacidad a causa de enfermedades no transmisibles (OMS, 2020).

1.2.7. Machine Learning

(IBM, 2023) la define como:

> "Una rama de la Inteligencia Artificial que usa los datos y los analiza mediante algoritmos con la finalidad de imitar la forma de pensar del ser humano".

De igual manera, Machine Learning es definido por Arthur Samuel (1952) y citado en (Garcia Dionisio, 2021) como:

> "Campo de estudio que brinda a las computadoras la capacidad de aprender sin estar programadas explícitamente".

El autor (Musa et al., 2022) explica que:

> "El aprendizaje automático (ML) es un subcampo de la IA que puede considerarse como un término general que abarca varios algoritmos que pueden aprender y mejorar automáticamente con la experiencia".

Por otro lado, (González, 2015) explica que el aprendizaje computacional o Machine Learning estudia a los sistemas capaces de aprender a partir de datos, encontrando patrones y regularidades mediante modelos supervisados y no supervisados.

1.2.8. Aprendizaje supervisado

Es un tipo de algoritmo de Machine Learning que utiliza un conjunto de datos conocidos, con el fin de reconocer similitudes y ser capaz de diferenciar los datos con etiquetas diferentes, se definen 2 tipos: de regresión y clasificación (Barra Benavides & Tataje Garcia, 2022).

El ML supervisado se suele usar para problemas de clasificación como identificación de dígitos, diagnostico o detección de fraudes, mientras que los problemas de regresión se usan en las predicciones meteorológicas, crecimiento, expectativa de vida. Otra diferencia es al denominar la variable objetivo la primera es de tipo categórica y la otra es de tipo numérica (Santos, 2021).

Entre los algoritmos encontramos: Arboles de decisión, clasificación de Naive Bayes, Regresión de mínimos cuadrados, regresión logística, support vector machine.

1.2.9. Aprendizaje no supervisado

Este algoritmo no tiene conocimiento previo, y busca organizarlos de alguna manera, entrenando modelo con datos sin procesar y sin etiquetar, entre las utilidades podemos mencionar clusters de datos con similitudes entre características, entender la relación entre diferentes puntos de datos y realizar análisis de datos iniciales (Universidad de Europa, 2022).

1.2.10. Orange Data Mining

Orange es una herramienta de visualización y análisis de datos de código abierto, que está siendo desarrollado en el laboratorio de bioinformática de la Facultad de Ciencias de la Computación e Información de la Universidad de Ljubljana (Eslovenia) (Pronin & Sotnikov, 2022) que ofrece flexibilidad en el preprocesamiento de datos, entrenamiento y prueba de modelos de aprendizaje de maquina (Vaishnav & Rao, 2018).

CAPÍTULO 2. METODOLOGÍA

2.1. Contexto de la investigación

El presente trabajo de investigación se realizará en Santa Elena, Ecuador; que según datos del Instituto Nacional de Estadísticas y Censo (INEC) tiene una población de 308.693 habitantes, donde el 5,7% corresponde a personas de la tercera edad (INEC, 2010).

Santa Elena es una provincia de la Costa Ecuatoriana creada el 7 de noviembre de 2007 (Prefectura Santa Elena, 2009), tiene 3 cantones Santa Elena, La Libertad y Salinas. Al norte limita con la provincia de Manabí, al sur y este con Guayas y al oeste con el Océano Pacífico (Wikipedia, 2023). Latitud: 2.226° y Longitud: 80.858° (IGM, 2017).

Figura 1 Perfil costanero de la Península de Santa Elena

Nota. Por Instituto Geográfico Militar ,2017. Recuperado el 16 de febrero de 2024, de https://www.geoportaligm.gob.ec/portal/

Para el 2017 a nivel nacional se registró un total de 7.404 defunciones por enfermedades cardiovasculares (4.300 enfermedades cerebrovasculares y 3.409 enfermedades hipertensivas) siendo así la primera causa de muerte en nuestro país (INEC SALUD, 2017).

A través de los resultados nacionales del módulo de actividad física y comportamiento sedentario del INEC se detalla que el 88% de los niños y jóvenes(5-17 años) ha realizado actividad física insuficiente, menos de 60 minutos diarios en los últimos 7 días; mientras que el 21.7% de los adultos (18-69 años) tiene una actividad física insuficiente, siendo las mujeres el grupo que más nivel de sedentarismo presenta en esta encuesta (INEC, 2021).

Por otra parte, la Encuesta Nacional de Salud y Nutrición – ENSANUT 2018, indica que 35 de cada 100 niños de entre 5 a 11 años, tiene sobrepeso y obesidad; además otro dato a mencionar es que el 7.6% y 4.3% de niños de 10 a 17 años han consumido alcohol durante los últimos 30 días y han consumido tabaco alguna vez en su vida, respectivamente (INEC, 2018).

2.2. Diseño y alcance de la investigación

Según el autor (Lancheros Florián, 2012), la investigación no experimental se basa en categorías, conceptos, variables, sucesos, comunidades o contextos que se dan sin la intervención directa del investigador, es decir; sin que el investigador altere el objeto de investigación, allí se observan los fenómenos o acontecimientos tal y como se dan en su contexto natural, para después analizarlos. Siendo este el enfoque de la investigación, debido a que se va a predecir a través de un conjunto de datos la probabilidad de que una persona desarrolle una enfermedad cardiovascular.

Explica (Huamani Mantari, 2019) que el alcance de la investigación de tipo correlacional permite determinar la relación estadística entre 2 o más variables cuantitativas sin necesidad de factores externos. Con la finalidad de conocer cuál es la relación de los factores de riesgos para que una persona pueda presentar problemas del corazón y de esta manera tomar medidas preventivas-correctivas.

2.3. Tipo y métodos de investigación

Esta investigación tiene enfoque cuantitativo, porque permite comprender patrones y correlaciones, causa y efecto a partir de la comprobación de hipótesis previamente formuladas (AEL, 2022).

A partir del Método de Investigación Analítico se procederá a la comprensión de los datos mediante procedimiento de análisis estadístico descriptivo de las variables de estudio (Patino & Arbelaz, 2016).

Finalmente, se establecerá una hipótesis que ayudará a negar o afirmar si los algoritmos de aprendizaje de máquina mejora la probabilidad de contraer o no una ECV, a través del Método hipotético-deductivo (Sullca, 2020).

2.4. Población

2.4.1. Población Objetivo

La población está compuesta por 308.855 registros de encuestas telefónicas del Sistema de Vigilancia de Factores de Riesgo del Comportamiento (BRFSS por sus siglas en ingles) que recopila datos relacionadas a la salud, condiciones de salud crónicas y uso de servicios preventivos de los residentes de Estados Unidos (Alphiree (Owner), 2021).

2.4.2. Muestra

Para el siguiente trabajo se dividió el total de los registros de la siguiente manera:

- Entrenamiento: 216.199 registros
- Prueba: 92.656 registros

2.5. Técnicas e instrumentos de recolección de datos

2.5.1. Dataset

Para asegurar la disponibilidad de datos relevantes y de alta calidad para el estudio, se optó por escoger un dataset alojado en el repositorio de Kaggle, que es una plataforma web para científicos de datos y profesionales del aprendizaje de máquina (Kaggle, 2024).

2.5.2. Encuesta a expertos

Conocer la opinión de otros investigadores es la forma más directa de obtener información a través de encuestas profundas sobre el tema (Huamani Mantari, 2019). En este caso, el cuestionario se dirigió a 6 expertos de la salud con la finalidad de conocer cuales son los

factores que inciden o no en la presencia o ausencia de ECV. Además, la encuesta buscó evaluar el conocimiento de los profesionales de la salud sobre el uso de Modelos de Aprendizaje de Máquina para el diagnóstico de enfermedades cardiovasculares.

2.5.3. Trabajos de Investigación

En este trabajo, para obtener información relevante sobre el Aprendizaje de Máquina aplicado a la detección de enfermedades cardiovasculares se consideró la revisión bibliográfica de trabajos, investigaciones o estudios con el fin de descubrir el tipo de modelos utilizados, métricas de rendimiento que ayudó para el análisis y evaluación de los modelos seleccionados para este estudio. Para ese fin se usó Semantic Scholar que es un motor de búsqueda de trabajos académicos(Semantic Scholar, 2024).

2.6. Procesamiento de la evaluación: Validez y confiabilidad de los instrumentos aplicados para el levantamiento de información.

2.6.1. Selección del dataset

Para la selección del dataset se tuvo las siguientes consideraciones:

Tabla 1 Variable del análisis del dataset

Campo	Descripción
Nombre del dataset	Palabra clave: *Cardiovascular Deseases Risk, Heart deseases, cardiovascular risk factors*
Fecha de publicación	último año
Tipo de archivo	.csv
Tamaño del dataset	pequeño
Variables	No contenga presión arterial, niveles de glucosa, niveles de colesterol

Fuente: Elaboración propia, 2024.

Luego del análisis correspondiente, el dataset *Cardiovascular Diseases Risk Prediction Dataset* cumple todas las consideraciones establecidas anteriormente. En la Tabla 2 se detalla información del dataset seleccionado.

Tabla 2 Descripción de las variables del dataset.

Variable	Descripción
General_Health	¿Diría usted que en general su salud es?
Checkup	¿Cuánto tiempo ha pasado aproximadamente desde la última vez que visitó a un médico para un chequeo de rutina?
Excercise	Durante el último mes, además de su trabajo habitual, ¿participó en alguna actividad o ejercicio físico como correr?
Heart_Disease	Encuestados que informaron tener enfermedad coronaria o infarto de miocardio.
Skin_Cancer	Encuestados que informaron tener cáncer de piel.
Other_Cancer	Encuestados que informaron tener algún otro tipo de cáncer.
Depression	Los encuestados que informaron tener un trastorno depresivo.
Diabetes	Encuestados que informaron tener diabetes. En caso afirmativo, qué tipo de diabetes es/era.
Arthritis	Los encuestados que informaron tener artritis.
Sex	Género
Age_Category	Edad
Height(cm)	Tamaño
Weight (kg)	Peso
BMI	Nivel de masa corporal
Smoking_History	Fuma o no
Alcohol_Consumption	¿Cuántos días en el mes tomó al menos un trago de alcohol?
Fruit_Consumption	¿Cuántas veces comes fruta en el mes?
Green_Vegetables_Consumption	¿Cuántas veces comes vegetales en el mes?
FriedPotato_Consumption	¿Cuántas veces comes papas fritas al mes?

Fuente: Elaboración propia, 2024.

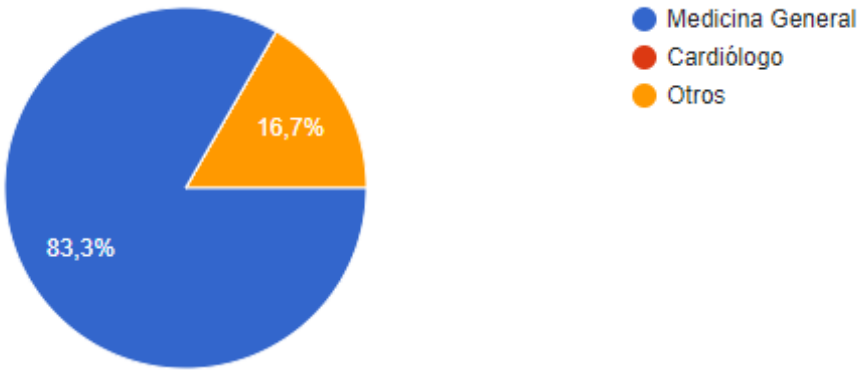
2.6.2. Análisis de la encuesta a expertos

La evaluación de los resultados de la encuesta fue de gran ayuda en el contexto de la investigación. El conocimiento adquirido servirá para discutir e inferir sobre los resultados de los modelos de aprendizaje de máquina y como sería de ayuda para los profesionales de la salud.

A continuación, se presenta el análisis de la entrevista:

1. ¿Cuál es su especialidad médica?

Figura 2 Gráfico de pastel de la pregunta 1

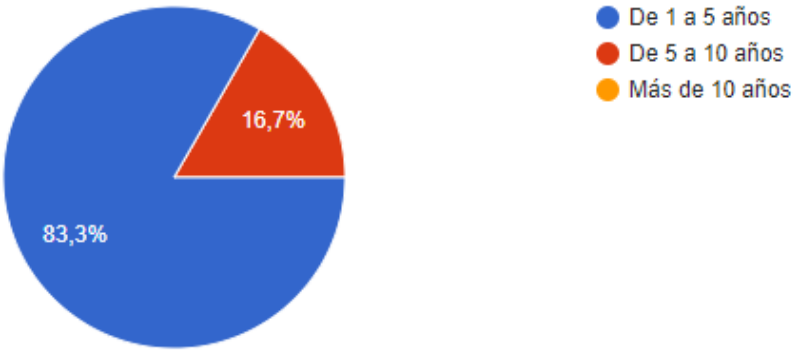


Elaboración propia
Fuente: Datos de la entrevista

Análisis: En la Figura 2 del total de expertos de la salud consultados, el 83,3% son médicos generales mientras que el 16,7% corresponde a otras especialidades.

2. ¿Cuánto tiempo lleva tratando pacientes con enfermedades cardiovasculares?

Figura 3 Gráfico de pastel de la pregunta 2



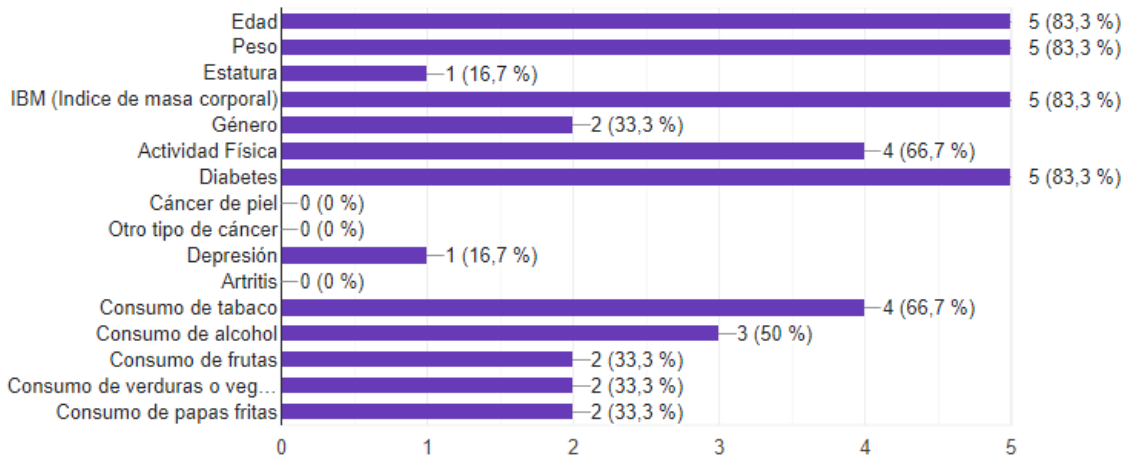
Elaboración propia

Fuente: Datos de la entrevista

Análisis: El 83,3% de los entrevistados que se muestra en la Figura 3 tiene de 1 a 5 años de experiencia tratando a pacientes con enfermedades cardiovasculares y el 16,7% entre de 5 y 10 años.

3. De la lista de factores de riesgo que se muestran a continuación, ¿Qué factores cree que se deben considerar para la detección de enfermedades cardiovasculares? Se puede seleccionar más de una opción.

Figura 4 Gráfico de barras de la pregunta 3

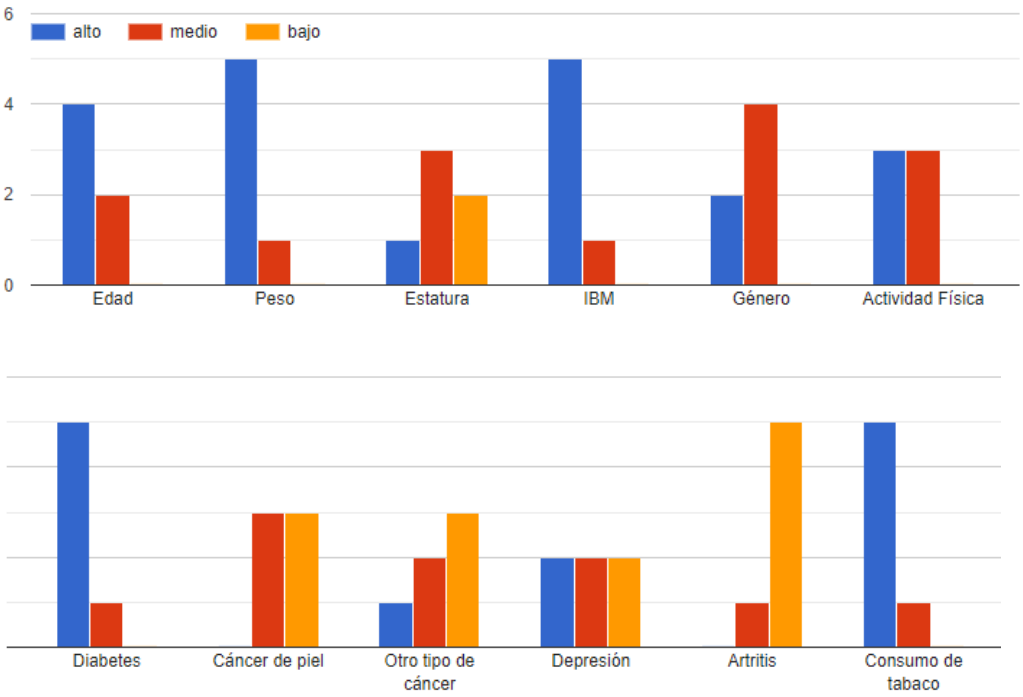


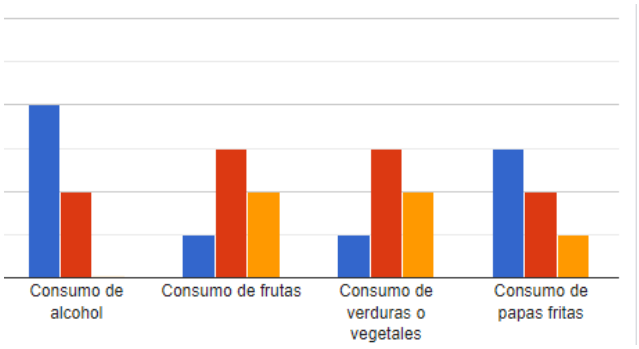
Elaboración propia
Fuente: Datos de la entrevista

Análisis: En la Figura 4 podemos destacar que el 83,3% de especialista indica que los factores de edad, peso, Índice de masa Corporal, Diabetes son factores que se deben considerar al momento de detectar enfermedades cardiovasculares en personas. Por otra parte, la estatura y la depresión es estimada en un 16,7% como factor de riesgo.

4. ¿Cuál es el nivel de influencia de los factores de riesgo para determinar la presencia de enfermedades cardiovasculares?

Figura 5 Gráfico de barra de la pregunta 4



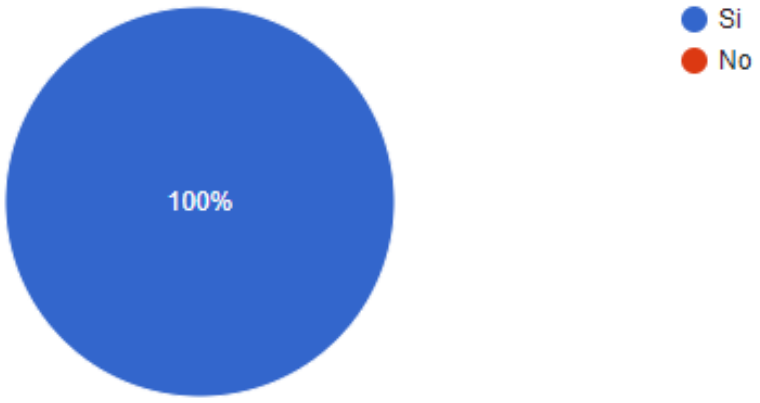


Elaboración propia
Fuente: Datos de la entrevista

Análisis: En la Figura 5 se puede analizar que 5 de cada 6 entrevistados piensa que el peso, el IMC, Diabetes, Consumo de tabaco tiene una influencia alta, mientras que la Depresión tiene influencia baja, media y alta para 2 médicos. Con una influencia media es considerado por 2 médicos como los factores de enfermedad cardiovascular de tipo consumo de alcohol, otro tipo de cáncer, depresión consumo de papas fritas.

5. ¿Conoce sobre el término de Inteligencia Artificial?

Figura 6 Gráfico de pastel de la pregunta 5

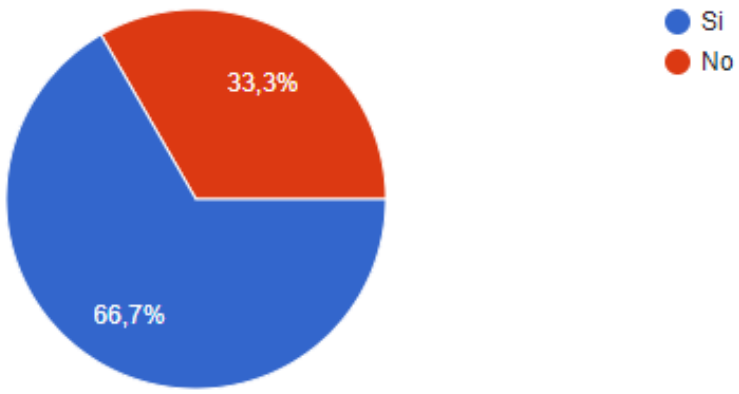


Elaboración propia
Fuente: Datos de la entrevista

Análisis: El 100% de los entrevistados conoce sobre la Inteligencia Artificial.

6. ¿Conoce sobre el término de Aprendizaje de Máquina o Machine Learning?

Figura 7 Gráfico de pastel de la pregunta 6

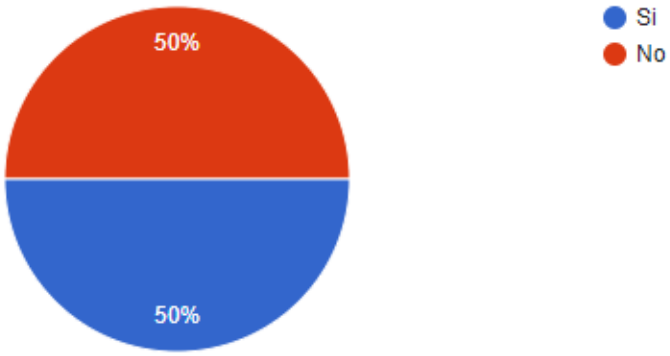


Elaboración propia
Fuente: Datos de la entrevista

Análisis: A pesar de que el 100% de los entrevistados conoce el término Inteligencia Artificial, solo el 66,7% conoce sobre el aprendizaje de máquina mientras que el 33,3% que corresponde a 2 profesionales de la salud no conoce sobre el término lo que podría llevar al desconocimiento de su uso en el área de la medicina.

7. ¿Conoce usted sobre el uso de modelos de Machine Learning para el diagnóstico temprano de enfermedades?

Figura 8 Gráfico de pastel de la pregunta 7

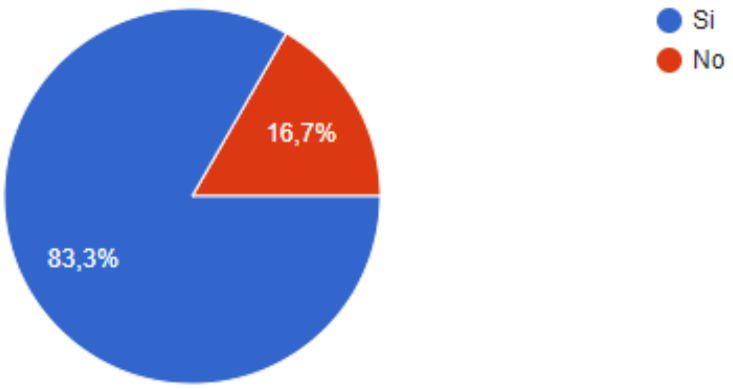


Elaboración propia
Fuente: Datos de la entrevista

Análisis: El 50% de los entrevistados conoce sobre la utilidad de los modelos de aprendizaje de máquina en el diagnostico de enfermedades mientras que la otra mitad no sabe de su utilidad en la medicina diagnóstica y preventiva.

8. ¿Consideraría usted que el uso de modelos de aprendizaje de máquina ayudaría a detectar la presencia de una enfermedad cardiovascular?

Figura 9 Gráfico de pastel de la pregunta 8



Elaboración propia
Fuente: Datos de la entrevista

Análisis: La mayoría de los entrevistados (83,83%) menciona que los modelos de aprendizaje de máquina pueden detectar la probabilidad de que una persona presente una enfermedad cardiovascular.

2.6.3. Análisis de Trabajos de Investigación

Los trabajos seleccionados contendrán las siguientes especificaciones:

Tabla 3 Variables del análisis de Trabajos

Campo	Descripción
Título	Palabras claves: *Machine Learning, Algorithms, Predicting, heart diseases, Diagnosis of Cardiovascular Disease, Performance, Assessment*
Fecha de Publicación	Últimos 5 años
Evaluación del modelo	métricas de evaluación del modelo: Precision, Recall, Puntuación F1, Acurracy
Algoritmos	3 o más algoritmos evaluados.

Fuente: Elaboración propia, 2024.

Finalmente, luego del análisis de la bibliografía disponible, se seleccionaron 4 trabajos que han proporcionado una base teórica y metodológica sólida que contribuyó al área de estudio.

- Machine Learning Algorithms for Predicting and Preventive Diagnosis of Cardiovascular Disease (Sembina et al., 2022).
- Use of Efficient Machine Learning Techniques in the Identification of Patients with Heart Diseases (Ghosh et al., 2021).
- Comparative Performance Assessment Of Machine Learning Algorithms To Predict Cardiovascular Disease (Hemalatha et al., 2023).
- Prediction of Heart Disease using Computational Algorithms (Shobha et al., 2022).

2.7. Metodología de desarrollo

2.7.1. Fase 1: Entendimiento del Problema

En esta primera etapa, se usó Orange Data Mining (Pronin & Sotnikov, 2022) para la carga y análisis exploratorio de los datos (EDA). La carga inicial tuvo el propósito de inspeccionar la estructura y características con el fin de una comprensión profunda de los datos. Por su parte, el análisis exploratorio permitió descubrir nuevos conocimientos a partir de tendencias y relaciones entre las variables con la finalidad de comprender el fenómeno de estudio (Barczewski et al., 2020). A continuación, se detalla los hallazgos obtenidos:

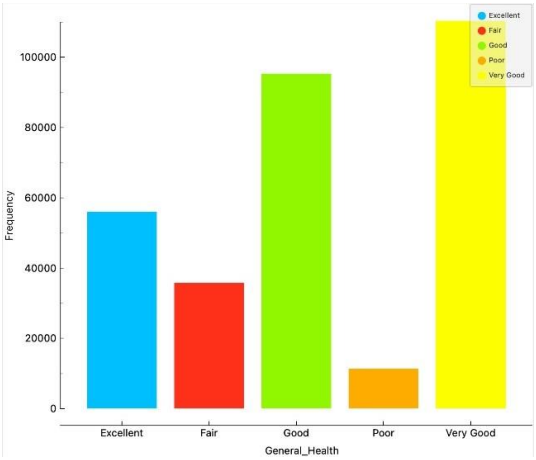
General_Health

Tabla 4 Análisis de la variable General_Health

	Frecuencia	Porcentaje	Porcentaje acumulado
Excellent	55954	18,12	18,12
Fair	35810	11,59	29,71
Good	95364	30,88	60,59
Poor	11331	3,67	64,26
Very Good	110395	35,74	100
Total	308854	100	

Fuente: Elaboración propia, 2024.

Figura 10 Análisis de la variable General_Health



Fuente: Elaboración propia, 2024.

Análisis

Según los datos proporcionados sobre el estado de salud general, se observa que la mayoría de la población presenta una percepción positiva de su salud, ya que la categoría "Very Good" cuenta con la cifra más elevada, alcanzando un total de 110.395 personas. Además, las categorías "Good" y "Excellent" también presentan números considerables, con 95.364 y 55.954 individuos respectivamente. Por otro lado, las categorías que reflejan percepciones menos optimistas, como "Poor" y "Fair", cuentan con menores proporciones, registrando 11.331 y 35.810 personas respectivamente. En conjunto, estos datos sugieren que la mayoría de la población encuestada tiende a evaluar positivamente su estado de salud general, indicando una tendencia favorable en términos de bienestar percibido.

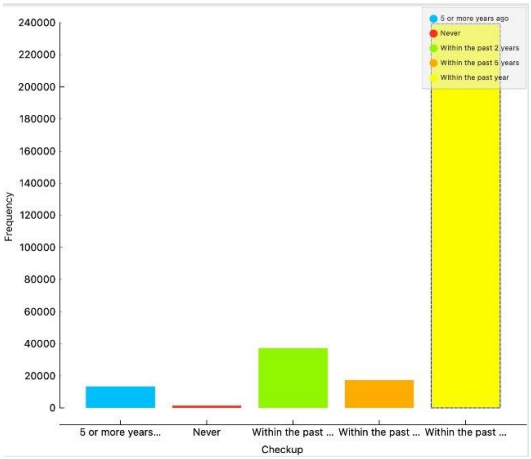
Checkup

Tabla 5 Análisis de la variable Checkup

	Frecuencia	Porcentaje	Porcentaje acumulado
5 or more years ago	13421	4,35	4,35
Never	1407	0,46	4,81
Within the past 2 years	37213	12,05	16,86
Within the past 5 years	17442	5,64	22,50
Within the past year	238371	77,5	100
Total	308854	100	

Fuente: Elaboración propia, 2024.

Figura 11 Análisis de la variable Checkup



Fuente: Elaboración propia, 2024.

Análisis

El 77,5% de los encuestados indicó que su más reciente visita al doctor fue en el último año y que está dentro de las recomendaciones médicas generales. Caso contrario ocurre con el 0,46% que nunca ha asistido a una cita o consulta médica.

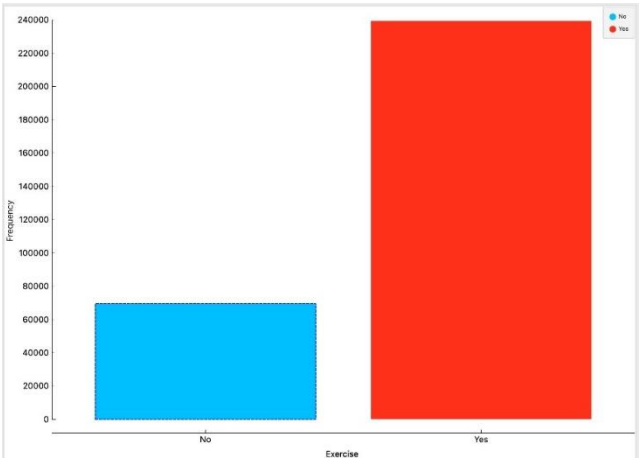
Excercise

Tabla 6 Análisis de la variable Excercise

	Frecuencia	Porcentaje	Porcentaje acumulado
Si	239381	77,51	77,51
No	69473	22,49	100
Total	308854	100	

Fuente: Elaboración propia, 2024.

Figura 12 Análisis de la variable Excercise



Fuente: Elaboración propia, 2024.

Análisis

En el análisis, se evidencia que una gran proporción de la población, representada por 239.381 individuos, participa en algún tipo de actividad física. Por otro lado, 69.473 personas indican no estar involucradas en actividades físicas. Estos resultados sugieren una participación generalizada en la actividad física, aunque un segmento significativo de la población aún no se involucra en este aspecto. La conjunción de estos datos sobre la salud general y la actividad física puede ser valiosa para comprender la relación entre el bienestar percibido y las prácticas de actividad física en esta población específica.

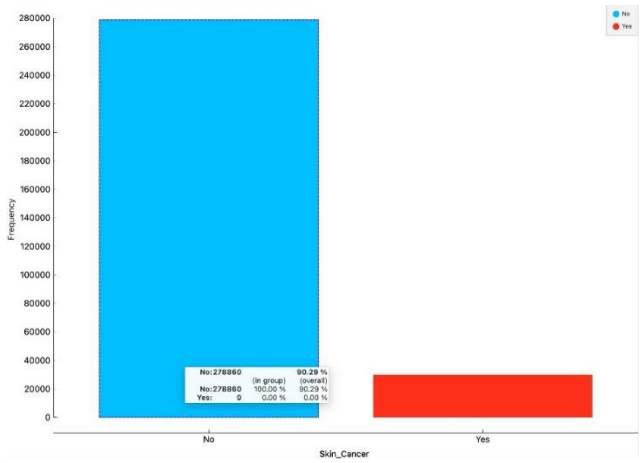
Skin_Cancer

Tabla 7 Análisis de la variable Skin_Cancer

	Frecuencia	Porcentaje	Porcentaje acumulado
Si	29994	9,71	9,71
No	278860	90,29	100
Total	308854	100	

Fuente: Elaboración propia, 2024.

Figura 13 Análisis de la variable Skin_Cancer



Fuente: Elaboración propia, 2024.

Análisis

Los datos sobre el cáncer de piel revelan que un segmento relativamente pequeño de la población, representado por 29,994 individuos, ha experimentado algún tipo de cáncer de piel. Por otro lado, una abrumadora mayoría de 278,860 personas informa no haber sido afectada por esta condición. Estas cifras indican una prevalencia

relativamente baja de cáncer de piel en la muestra, lo que podría sugerir prácticas efectivas de prevención y conciencia en la población. No obstante, es importante considerar que estos datos proporcionan una visión general y que un análisis más detallado, incluyendo factores demográficos y geográficos, podría arrojar luz adicional sobre los patrones y las tendencias asociadas con el cáncer de piel en esta población.

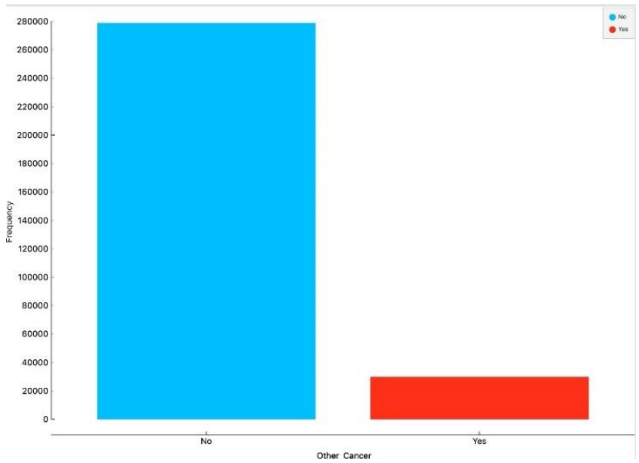
Other_Cancer

Tabla 8 Análisis de la variable Other_Cancer

	Frecuencia	Porcentaje	Porcentaje acumulado
Si	29878	9,67	9,67
No	278976	90,33	100
Total	308854	100	

Fuente: Elaboración propia, 2024.

Figura 14 Análisis de la variable Other_Cancer



Fuente: Elaboración propia, 2024.

Análisis

En cuanto a otros tipos de cáncer, los datos indican que 29.878 personas de la muestra han experimentado algún tipo de cáncer distinto al de piel, mientras que 278.976 personas no han enfrentado esta enfermedad. Estos números reflejan una proporción significativa de la población que ha sido afectada por algún tipo de cáncer, lo que destaca la importancia de la atención y la conciencia en relación con la detección temprana y la

prevención de diversas formas de cáncer. Un análisis más detallado, considerando factores como la edad, el género y los hábitos de vida, podría proporcionar información adicional sobre las características y las posibles causas de esta prevalencia de cáncer en la población estudiada.

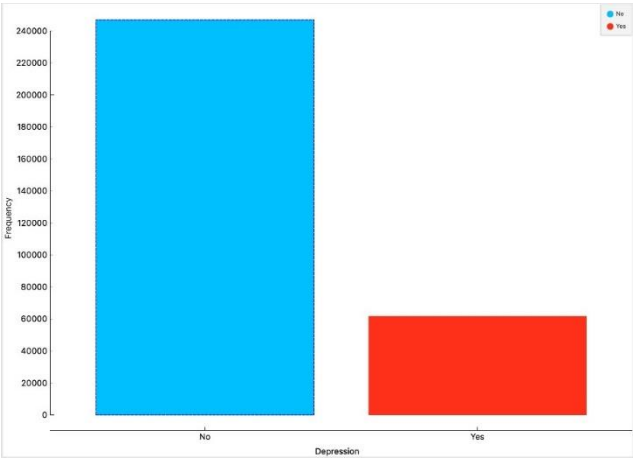
Depression

Tabla 9 Análisis de la variable Depression

	Frecuencia	Porcentaje	Porcentaje acumulado
Si	61910	20,04	20,04
No	246953	79,96	100
Total	308854	100	

Fuente: Elaboración propia, 2024.

Figura 15 Análisis de la variable Depression



Fuente: Elaboración propia, 2024.

Análisis

Los datos relativos a la depresión indican que 61.910 personas de la muestra han experimentado esta condición, mientras que 246.953 individuos no han enfrentado la depresión. Estos números resaltan la importancia de abordar la salud mental en la población, ya que una proporción significativa informa haber experimentado depresión en algún momento. La detección temprana, el acceso a servicios de salud mental y la conciencia pública son factores cruciales para abordar y gestionar la depresión en la comunidad. Un análisis más profundo de estos datos, considerando

factores como el género, la edad y los factores de riesgo, podría proporcionar información valiosa para desarrollar estrategias efectivas de prevención y tratamiento.

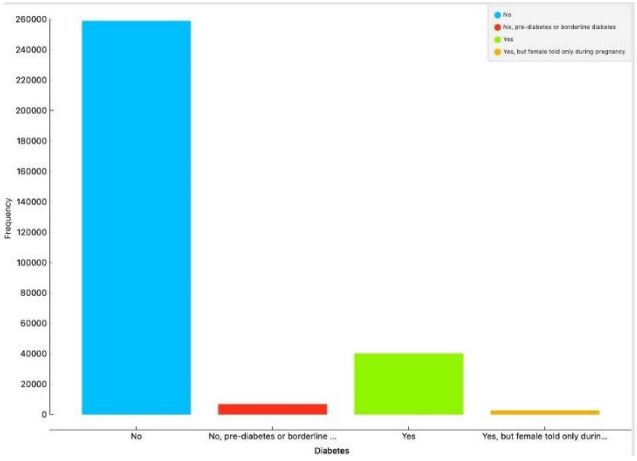
Diabetes

Tabla 10 Análisis de la variable Diabetes

	Frecuencia	Porcentaje	Porcentaje acumulado
Si	40171	13,01	13,01
Si, pero durante el embarazo	2646	0,86	13,87
No	259141	83,90	97,77
No, pre-diabetes	6896	2,23	100
Total	308854	100	

Fuente: Elaboración propia, 2024.

Figura 16 Análisis de la variable Diabetes



Fuente: Elaboración propia, 2024.

Análisis

Los datos relacionados con la diabetes revelan que 40.171 personas de la muestra han sido diagnosticadas con diabetes, mientras que 2.646 indican haber experimentado diabetes únicamente durante el embarazo. Por otro lado, 259.141 personas declaran no tener diabetes, y 6.896 personas informan tener pre-diabetes. Estos resultados resaltan la relevancia de abordar la diabetes y sus distintas manifestaciones en la población, destacando la necesidad de medidas preventivas y estrategias de gestión de la enfermedad. Además, la identificación de casos de diabetes gestacional durante el embarazo destaca

la importancia del cuidado prenatal y la monitorización de la salud materna para prevenir complicaciones. Un análisis más detallado de estos datos, considerando factores como la edad, el estilo de vida y los antecedentes familiares, podría ofrecer una comprensión más completa de la prevalencia y los factores asociados con la diabetes en esta población específica.

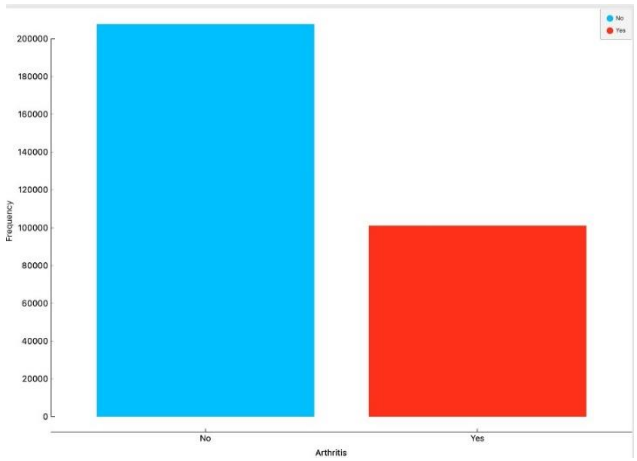
Arthritis

Tabla 11 Análisis de la variable Arthritis

	Frecuencia	Porcentaje	Porcentaje acumulado
Si	101071	32,72	32,72
No	207783	67,28	100
Total	308854	100	

Fuente: Elaboración propia, 2024.

Figura 17 Análisis de la variable Arthritis



Fuente: Elaboración propia, 2024.

Análisis

Los datos sobre la artritis indican que un número significativo de 101.071 personas de la muestra han sido diagnosticadas con esta condición, mientras que 207.783 individuos informan no tener artritis. Estos resultados subrayan la relevancia de la artritis como un problema de salud en la población estudiada y resaltan la necesidad de estrategias de manejo y tratamiento efectivas.

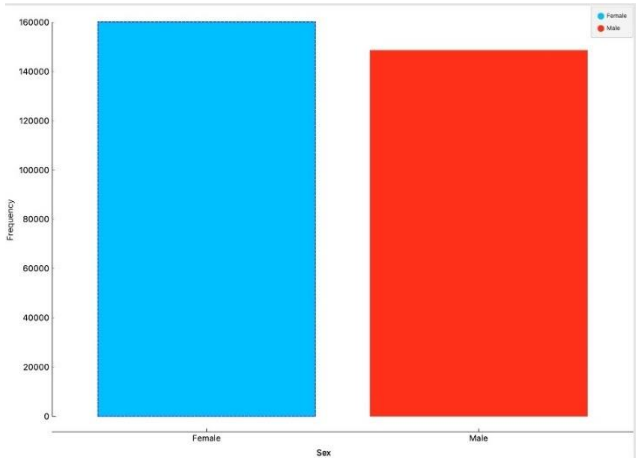
Sex

Tabla 12 Análisis de la variable Sex

	Frecuencia	Porcentaje	Porcentaje acumulado
Femenino	160196	51,87	51,87
Masculino	148658	48,13	100
Total	308854	100	

Fuente: Elaboración propia, 2024.

Figura 18 Análisis de la variable Sex



Fuente: Elaboración propia, 2024.

Análisis

Los datos sobre el género indican que en la muestra hay 160.196 individuos femeninos y 148.658 masculinos, lo que suma un total de 308.854 personas. Esta distribución revela una ligera predominancia de la población femenina en comparación con la masculina.

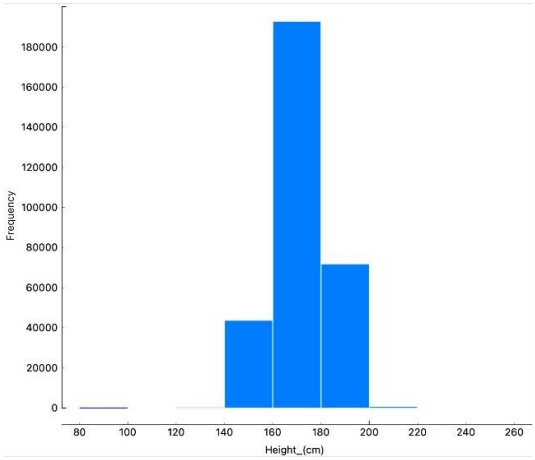
Height(cm)

Tabla 13 Análisis de la variable Height(cm)

	Frecuencia	Porcentaje	Porcentaje acumulado
< 100	17	0,01	0,01
100 – 119	50	0,02	0,03
120 – 139	273	0,09	0,12
140 – 159	43528	14,09	14,21
160 – 179	192693	62,39	76,6
180 – 199	71687	23,21	99,80
200 – 219	587	0,19	99,99
220 – 240	18	0,01	100
> 240	1	0,00	100
Total	308854	100	

Fuente: Elaboración propia, 2024.

Figura 19 Análisis de la variable Height(cm)



Fuente: Elaboración propia, 2024.

Análisis

La distribución de la talla revela una variedad en los rangos, destacando que la mayoría de la población se encuentra en los segmentos de altura entre 160 y 179 (192.693 personas) y 140 a 159 (43.528 personas). Estos datos sugieren una distribución relativamente normal de la altura en la muestra, con la mayoría de las personas ubicadas en rangos considerados como promedio. Los valores extremos, como aquellos por debajo de 100 o por encima de 200, son menos comunes, indicando una menor presencia de casos atípicos en la muestra.

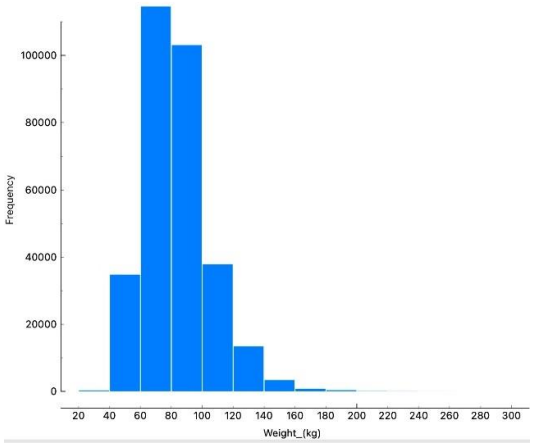
Weight (kg)

Tabla 14 Análisis de la variable Weight (kg)

	Frecuencia	Porcentaje	Porcentaje acumulado
< 40	290	0,094	0,094
40 – 59	34787	11,263	11,357
60 – 79	114608	37,108	48,465
80 – 99	103132	33,392	81,856
100 – 119	37895	12,270	94,126
120 – 139	13381	4,332	98,458
140 – 159	3360	1,088	99,546
160 – 179	808	0,262	99,808
180 – 199	372	0,120	99,928
200 – 219	118	0,038	99,967
220 – 239	73	0,024	99,990
240 – 259	17	0,006	99,996
260 – 279	10	0,003	99,999
>= 280	3	0,001	100
Total	308854	100	

Fuente: Elaboración propia, 2024.

Figura 20 Análisis de la variable Weight (kg)



Fuente: Elaboración propia, 2024.

Análisis

El peso de una persona es considerado un factor de riesgo relevante en las ECV, esto se debe a que esta variable puede ser un indicador del estado de salud general de una persona. En la Tabla 14Tabla 14 observamos que 18,14% de los encuestados tienen un peso mayor de 100Kg, además que la mayor parte de los registros están dentro de un rango de peso de entre 60Kg y menor de 100Kg.

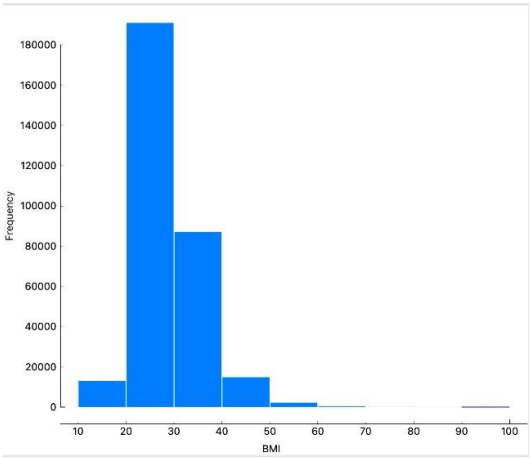
BMI

Tabla 15 Análisis de la variable BMI

	Frecuencia	Porcentaje	Porcentaje acumulado
< 20	13030	4,22	4,22
20 – 29	190975	61,83	66,05
30 – 39	87158	28,22	94,27
40 – 49	14874	4,82	99,09
50 – 59	2254	0,73	99,82
60 – 69	397	0,13	99,95
70 – 79	106	0,03	99,98
80 – 89	45	0,01	100
>= 90	15	0,00	100
Total	308854	100	

Fuente: Elaboración propia, 2024.

Figura 21 Análisis de la variable BMI



Fuente: Elaboración propia, 2024.

Análisis

La distribución del Índice de Masa Corporal (BMI) en la muestra revela patrones significativos en la composición corporal de la población. La mayoría de los individuos presentan un BMI entre 20 y 29, lo que indica una prevalencia considerable de pesos considerados normales. No obstante, la presencia de personas con un BMI inferior a 20 y, particularmente, aquellos con un BMI de 30 o superior, destaca la necesidad de abordar posibles problemas de salud asociados con el bajo peso y la obesidad. La presencia de 15 individuos con un BMI de 90 o superior sugiere una minoría que podría enfrentar desafíos de salud más significativos.

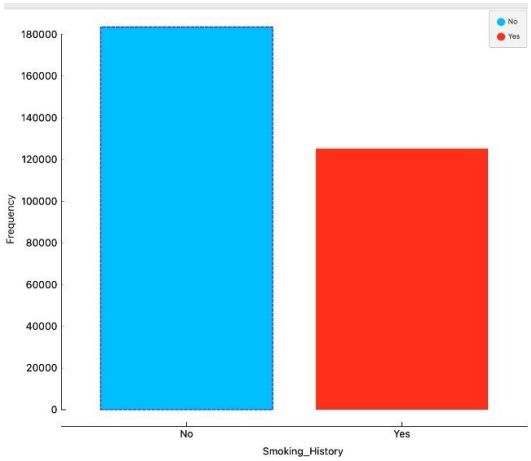
Smoking_History

Tabla 16 Análisis de la variable Smoking_History

	Frecuencia	Porcentaje	Porcentaje acumulado
Si	125264	40,56	40,56
No	183590	59,44	100
Total	308854	100	

Fuente: Elaboración propia, 2024.

Figura 22 Análisis de la variable Smoking_History



Fuente: Elaboración propia, 2024.

Análisis

La información relacionada con el historial de fumador en la muestra revela que 125.264 individuos han afirmado ser fumadores, mientras que 183.590 han declarado no tener antecedentes de tabaquismo. Estos datos proporcionan una instantánea útil para comprender la prevalencia del hábito de fumar en la población estudiada. Un análisis más profundo de estos datos en combinación con otras variables, como el estado de salud general, puede ofrecer una visión integral de los posibles impactos del tabaquismo en la salud de la población.

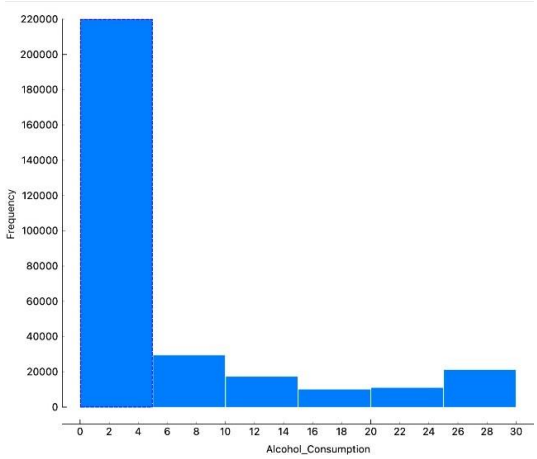
Alcohol_Consumption

Tabla 17 Análisis de la variable Alcohol_Consumption

	Frecuencia	Porcentaje	Porcentaje acumulado
< 5	219831	71,18	71,18
5 – 9	29389	9,52	80,70
10 – 14	17309	5,60	86,30
15 – 19	10655	3,26	89,56
20 – 24	11039	3,57	93,13
25 – 30	21221	6,87	100
Total	308854	100	

Fuente: Elaboración propia, 2024.

Figura 23 Análisis de la variable Alcohol_Consumption



Fuente: Elaboración propia, 2024.

Análisis

La información sobre el historial de consumo de alcohol en la muestra revela patrones diversos en la cantidad de unidades consumidas. La mayoría de los individuos, representados por 219.831 personas, reportan un consumo inferior a 5 unidades de alcohol. Asimismo, se observa un número considerable de personas que consumen entre 5 y 30 unidades, distribuidas en los diferentes rangos. Este análisis destaca la importancia de comprender la variabilidad en los patrones de consumo de alcohol en la población, ya que puede influir en la salud y el bienestar general.

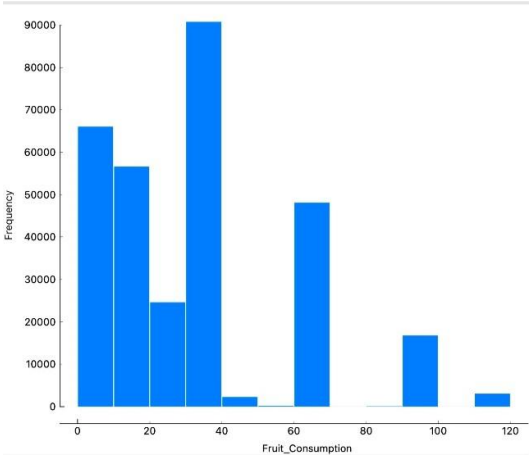
Fruit_Consumption

Tabla 18 Análisis de la variable Fruit_Consumption

	Frecuencia	Porcentaje	Porcentaje acumulado
< 10	66056	21,39	21,39
10 – 19	56679	18,35	39,74
20 – 29	24642	7,98	47,72
30 – 39	90769	29,39	77,11
40 – 49	2290	0,74	77,85
50 – 59	193	0,06	77,91
60 – 69	48146	15,59	93,5
70 – 79	28	0,01	93,51
80 – 89	121	0,04	93,55
90 – 99	16797	5,44	98,99
100 – 109	16	0,01	99,00
110 – 120	3117	1,0	100
Total	308854	100	

Fuente: Elaboración propia, 2024.

Figura 24 Análisis de la variable Fruit_Consumption



Fuente: Elaboración propia, 2024.

Análisis

La información sobre el consumo de frutas en la muestra ofrece una visión detallada de los hábitos alimenticios de la población. La mayoría de los individuos, representados por 90.769 personas, consumen entre 30 y 39 porciones de frutas, indicando una atención relativamente positiva hacia una dieta saludable. Sin embargo, es notorio que hay segmentos de la población con patrones de consumo más bajos, como aquellos que consumen menos de 10 porciones.

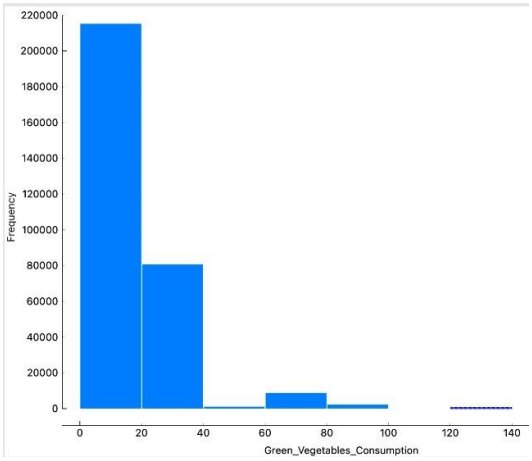
Green_Vegetables_Consumption

Tabla 19 Análisis de la variable Green_Vegetables_Consumption

	Frecuencia	Porcentaje	Porcentaje acumulado
< 20	215247	69,69	69,69
20 – 39	80770	26,15	95,84
40 – 59	1145	0,37	96,21
60 – 79	8816	2,85	99,06
80 – 99	2243	0,73	99,79
100 – 119	15	0,01	99,80
120 – 140	618	0,20	100
Total	308854	100	

Fuente: Elaboración propia, 2024.

Figura 25 Análisis de la variable Green_Vegetables_Consumption



Fuente: Elaboración propia, 2024.

Análisis

La información sobre el consumo de vegetales en la muestra proporciona una visión detallada de los hábitos alimenticios en relación con este grupo alimenticio. La mayoría de los individuos, representados por 215.247 personas, consumen menos de 20 porciones de vegetales, lo que refleja una atención mínima hacia una dieta rica en vegetales. No obstante, se observa una disminución en el número de personas a medida que aumenta la cantidad de porciones, especialmente en los rangos más altos. Este análisis subraya la importancia de fomentar el consumo de vegetales en la población y resalta posibles áreas de intervención para mejorar los hábitos alimenticios.

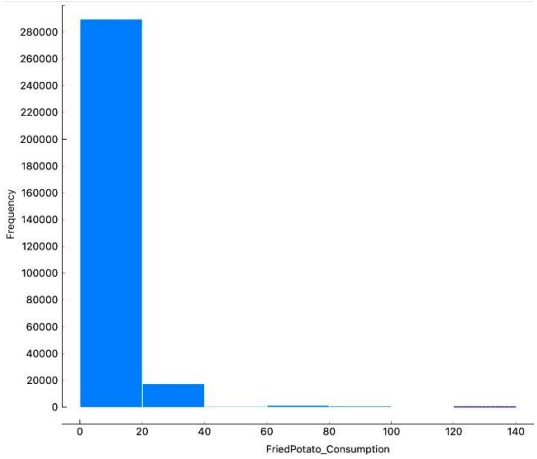
FriedPotato_Consumption

Tabla 20 Análisis de la variable FriedPotato_Consumption

	Frecuencia	Porcentaje	Porcentaje acumulado
< 20	289499	93,73	93,73
20 – 39	17214	5,57	99,3
40 – 59	296	0,10	99,4
60 – 79	1089	0,35	99,75
80 – 99	545	0,18	99,93
100 – 119	3	0	99,93
120 – 140	208	0,07	100
Total	308854	100	

Fuente: Elaboración propia, 2024.

Figura 26 Análisis de la variable FriedPotato_Consumption



Fuente: Elaboración propia, 2024.

Análisis

La información sobre el consumo de papas fritas o sus semejantes en la muestra evidencia patrones distintos de hábitos alimenticios. La mayoría de los individuos, representados por 289.499 personas, consumen menos de 20 porciones, señalando una negativa al consumo de este tipo de alimentación. Sin embargo, se observa una disminución en la cantidad de personas a medida que aumenta la cantidad de porciones, especialmente en los rangos más altos. Estos resultados subrayan la necesidad de abordar los hábitos alimenticios de la población, dada su asociación con riesgos para la salud.

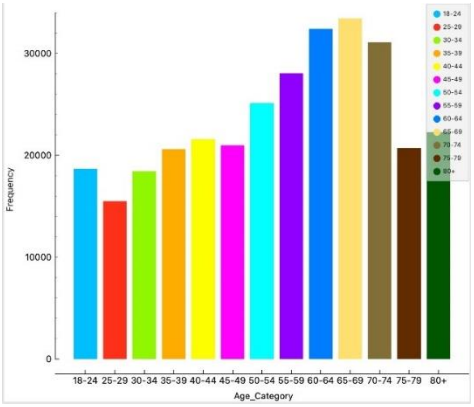
Age_Category

Tabla 21 Análisis de la variable Age_Category

	Frecuencia	Porcentaje	Porcentaje acumulado
18 – 24	18681	6,05	6,05
25 – 29	15494	5,02	11,07
30 – 34	18428	5,97	17,04
35 – 39	20606	6,67	23,71
40 – 44	21595	6,99	30,7
45 – 49	20968	6,79	37,49
50 – 54	25097	8,13	45,62
55 – 59	28054	9,08	54,7
60 – 64	32418	10,50	64,2
65 – 69	33434	10,83	76,03
70 – 74	31103	10,07	86,1
75 – 79	20705	6,70	92,8
>= 80	22271	7,2	100
Total	308854	100	

Fuente: Elaboración propia, 2024.

Figura 27 Análisis de la variable Age_Category



Fuente: Elaboración propia, 2024.

Análisis

La tabla de frecuencias de edades proporciona un panorama detallado de la distribución de la población según grupos etarios. Se observa un patrón ascendente en la cantidad de individuos desde los grupos más jóvenes hasta los de mayor edad. La mayor concentración se encuentra en los rangos de 50 a 74 años, destacando una proporción significativa de la población en etapas intermedias y avanzadas de la vida.

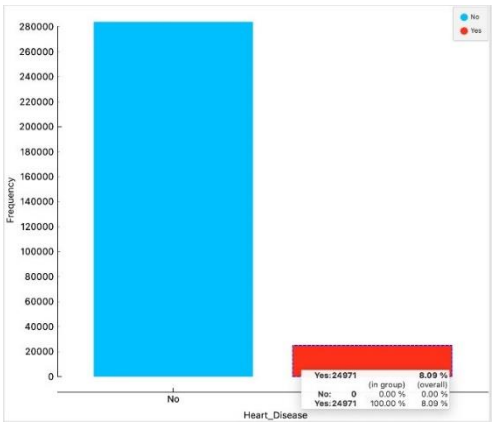
Heart_Disease

Tabla 22 Análisis de la variable Heart_Disease

	Frecuencia	Porcentaje	Porcentaje acumulado
Si	24971	8,09	8,09
No	283883	91,91	100
Total	308854	100	

Fuente: Elaboración propia, 2024.

Figura 28 Análisis de la variable Heart_Disease



Fuente: Elaboración propia, 2024.

Análisis

En la Figura 28 Análisis de la variable Heart_Diseasese evidencia que el 8,09% de los entrevistados presentan alguna enfermedad cardiovascular, mientras que 91,91% no percibe enfermedad cardiovascular alguna. Esto da como resultado el desequilibrio o desbalanceo que existe en la variable objetivo, pues de los 308854 registros, 283883 tiene con valor de clasificación 0 (sin enfermedad cardiovascular) y 24971 tiene valor de clasificación 1 (con enfermedad cardiovascular), es decir que de cada 12 personas al menos 1 presenta algún tipo de enfermedad cardiovascular.

Análisis Bivariado

El análisis bivariado se encarga de examinar la relación interdependiente entre dos variables con el fin de comprender la dinámica de causa y efecto inherente a su interacción, identificar correlaciones y determinar la capacidad predictiva de una variable sobre otra. Esta metodología es esencial en el proceso de esclarecimiento de las relaciones existentes entre variables independientes (predictores) y la variable dependiente (objetivo).

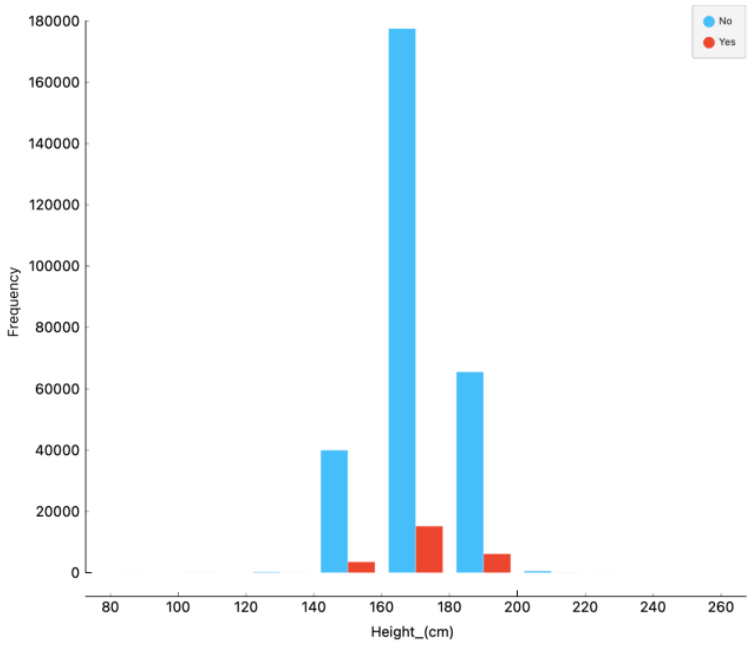
Height_(cm) vs Heart_Disease

Tabla 23 Análisis Height_(cm) vs Heart_Disease

Height	Heart_Disease		
	Si	**No**	**Total**
< 100	15	2	17
100 – 119	2	48	50
120 – 139	24	249	273
140 – 159	3551	39977	43528
160 – 179	15189	177504	192693
180 – 199	6168	65519	71687
200 – 219	35	552	587
220 – 240	0	18	18
> 240	0	1	1
Total	24984	283870	308854

Fuente: Elaboración propia, 2024.

Figura 29 Análisis Height_(cm) vs Heart_Disease



Fuente: Elaboración propia, 2024.

Weight_(Kg) vs Heart_Disease

Tabla 24 Análisis Weight_(Kg) vs Heart_Disease

Weight	Heart_Disease		
	Si	**No**	**Total**
< 40	31	259	290
40 – 59	2086	32701	34787
60 – 79	8107	106501	114608
80 – 99	9161	93971	103132
100 – 119	3797	34098	37895
120 – 139	1311	12070	13381
140 – 159	344	3016	3360
160 – 179	83	725	808
180 – 199	30	342	372
200 – 219	15	103	118
220 – 239	4	69	73
240 – 259	1	16	17
260 – 279	0	10	10
>= 280	1	2	3
Total	24971	283883	308854

Fuente: Elaboración propia, 2024.

BMI vs Heart_Disease

Tabla 25 Análisis BMI vs Heart_Disease

BMI	Heart_Disease		
	Si	No	Total
< 20	810	12220	13030
20 – 29	14090	176885	190975
30 – 39	8365	78793	87158
40 – 49	1438	13436	14874
50 – 59	218	2036	2254
60 – 69	38	359	397
70 – 79	8	98	106
80 – 89	4	41	45
>= 90	0	15	15
Total	24971	283883	308854

Fuente: Elaboración propia, 2024.

Alcohol_Consumption vs Heart_Disease

Tabla 26 Análisis Alcohol_Consumption vs Heart_Disease

Alcohol_Consumption	Heart_Disease		
	Si	No	Total
> 5	19861	199970	219831
5 – 9	1366	28023	29389
10 – 14	864	16445	17309
15 – 19	475	9590	10655
20 – 24	638	10401	11039
25 – 30	1767	19454	21221
Total	24971	283883	308854

Fuente: Elaboración propia, 2024.

Green_Vegetables_Consumption vs Heart_Disease

Tabla 27 Análisis Green_Vegetables_Consumption vs Heart_Disease

Green_Vegetables_Consumption	Heart_Disease		
	Si	No	Total
> 20	197137	18110	215247
20 – 39	6085	74685	80770
40 – 59	82	1063	1145
60 – 79	521	8295	8816
80 – 99	134	2109	2243
100 – 119	0	15	15
120 – 140	39	579	618
Total	203998	104856	308854

Fuente: Elaboración propia, 2024.

Fruit_Consumption vs Heart_Disease

Tabla 28 Análisis Fruit_Consumption vs Heart_Disease

Fruit_Consumption	Heart_Disease		
	Si	No	Total
> 10	6168	59888	66056
10 – 19	4457	52222	56679
20 – 29	1763	22879	24642
30 – 39	7533	83236	90769
40 – 49	173	2117	2290
50 – 59	18	175	193
60 – 69	3336	44810	48146
70 – 79	0	28	28
80 – 89	4	117	121
90 – 99	1301	15496	16797
100 – 109	1	15	16
110 – 120	217	2900	3117
Total	24971	283883	308854

Fuente: Elaboración propia, 2024.

FriedPotato_Consumption vs Heart_Disease

Tabla 29 Análisis FriedPotato_Consumption vs Heart_Disease

FriedPotato_Consumption	Heart_Disease		
	Si	No	Total
> 20	23362	266137	289499
20 – 39	1456	15758	17214
40 – 59	15	281	296
60 – 79	85	1004	1089
80 – 99	40	505	545
100 – 119	0	3	3
120 – 140	13	195	208
Total	24971	283883	308854

Fuente: Elaboración propia, 2024.

General_Health vs Heart_Disease

Tabla 30 Análisis General_Health vs Heart_Disease

General_Health	Heart_Disease		
	Si	No	Total
Excellent	1115	54839	55954
Fair	6789	29021	35810
Good	8643	86721	95364
Poor	3602	7729	11331
Very Good	4822	105573	110395
Total	24971	283883	308854

Fuente: Elaboración propia, 2024.

Checkup vs Heart_Disease

Tabla 31 Análisis Checkup vs Heart_Disease

Checkup	Heart_Disease		
	Si	No	Total
5 or more years ago	342	13079	13421
Never	58	1349	1407
Within the past 2 years	1465	35748	37213
Within the past 5 years	471	16971	17442
Within the past year	22635	216736	238371
Total	24971	283883	308854

Fuente: Elaboración propia, 2024.

Exercise vs Heart_Disease

Tabla 32 Análisis Exercise vs Heart_Disease

Exercise	Heart_Disease		
	Si	No	Total
Si	15967	223414	239381
No	9004	60469	69473
Total	24971	283883	308854

Fuente: Elaboración propia, 2024.

Skin_Cancer vs Heart_Disease

Tabla 33 Análasis Skin_Cancer vs Heart_Disease

Skin_Cancer	Heart_Disease		
	Si	No	Total
Si	4690	25304	29994
No	20281	258579	278860
Total	24971	283883	308854

Fuente: Elaboración propia, 2024.

Other_Cancer vs Heart_Disease

Tabla 34 Análisis Other_Cancer vs Heart_Disease

Other_Cancer	Heart_Disease		
	Si	No	Total
Si	4715	25163	29878
No	20256	258720	278976
Total	24971	283883	308854

Fuente: Elaboración propia, 2024.

Depression vs Heart_Disease

Tabla 35 Análisis Depression vs Heart_Disease

Depression	Heart_Disease		
	Si	No	Total
Si	6101	55800	61910
No	18870	228083	246953
Total	24971	283883	308854

Fuente: Elaboración propia, 2024.

Diabetes vs Heart_Disease

Tabla 36 Análisis Diabetes vs Heart_Disease

Diabetes	Heart_Disease		
	Si	No	Total
Si	8376	31795	40171
Si, pero durante el embarazo	96	2550	2646
No	15705	243436	259141
No, pre-diabetes	794	6102	6896
Total	24971	283883	308854

Fuente: Elaboración propia, 2024.

Arthritis vs Heart_Disease

Tabla 37 Análisis Arthritis vs Heart_Disease

Arthritis	Heart_Disease		
	Si	No	Total
Si	14252	86619	101071
No	10719	197064	207783
Total	24971	283883	308854

Fuente: Elaboración propia, 2024.

Sex vs Heart_Disease

Tabla 38 Análisis Sex vs Heart_Disease

Sex	Heart_Disease		
	Si	No	Total
Femenino	9898	150298	160196
Masculino	15073	133585	148658
Total	24971	283883	308854

Fuente: Elaboración propia, 2024.

Age_Category vs Heart_Disease

Tabla 39 Análisis Age_Category vs Heart_Disease

Age_Category	Heart_Disease		
	Si	No	Total
18 – 24	94	18587	18681
25 – 29	113	15381	15494
30 – 34	201	18227	18428
35 – 39	274	20332	20606
40 – 44	435	21160	21595
45 – 49	678	20290	20968
50 – 54	1181	23916	25097
55 – 59	1991	26063	28054
60 – 64	3012	29406	32418
65 – 69	3823	29611	33434
70 – 74	4561	26542	31103
75 – 79	3752	16953	20705
>= 80	4856	17415	22271
Total	24971	283883	308854

Fuente: Elaboración propia, 2024.

Smoking_History vs Heart_Disease

Tabla 40 Análisis Smoking_History vs Heart_Disease

Smoking_History	Heart_Disease		
	Si	No	Total
Si	14584	110680	125264
No	10387	173203	183590
Total	24971	283883	308854

Fuente: Elaboración propia, 2024.

Análisis

- Se observa un aumento progresivo en la prevalencia de enfermedades cardíacas a medida que avanza la edad. Este patrón sugiere una correlación positiva entre la edad y el riesgo de desarrollar enfermedades cardíacas, siendo especialmente notable el incremento a partir de los 50 años. Esto subraya la importancia de considerar la edad como un factor de riesgo significativo en la predicción de enfermedades cardíacas.
- Los datos muestran que los hombres tienen una mayor prevalencia de enfermedad cardíaca (15,073 casos) en comparación con las mujeres (9,898 casos), indicando

que el sexo podría ser un factor determinante en el riesgo de desarrollar enfermedades cardíacas. Esta diferencia sugiere que las estrategias de prevención y diagnóstico podrían necesitar ser adaptadas según el sexo.

- La prevalencia de enfermedades cardíacas varía significativamente con el estado general de salud reportado, siendo más alta en individuos con un estado de salud "pobre" o "justo" y más baja en aquellos que reportan un estado de salud "excelente" o "muy bueno". Esto refuerza la noción de que la percepción individual del estado de salud puede reflejar factores de riesgo subyacentes para enfermedades cardíacas.
- El análisis revela que la mayor proporción de individuos con enfermedad cardíaca se encuentra en el grupo que consume alcohol moderadamente (> 5 unidades). Sin embargo, es importante notar que este dato por sí solo no establece causalidad, y podría reflejar estilos de vida u otros factores de riesgo asociados.
- Los datos indican que hay una prevalencia significativa de enfermedades cardíacas en individuos con un rango de peso de 80-99 kg, seguido por aquellos en el rango de 60-79 kg. Esto sugiere una posible relación entre el aumento del peso y el riesgo de enfermedad cardíaca. Sin embargo, es crucial considerar que factores como la distribución de la grasa corporal (por ejemplo, adiposidad visceral) y la composición corporal (masa grasa vs. masa muscular) pueden influir en este riesgo.

2.7.2. Fase 2: Preprocesamiento de Datos

Limpieza de datos

La limpieza de datos es un proceso esencial en el análisis de datos; la gestión de valores nulos y vacíos es una parte crucial de este procedimiento. La presencia de datos faltantes puede afectar significativamente la precisión y validez de los resultados obtenidos. Para abordar esta problemática, es fundamental realizar una inspección exhaustiva de los conjuntos de datos, identificando y manejando adecuadamente los valores nulos. Este procedimiento puede involucrar la eliminación de registros con datos faltantes o la imputación de valores utilizando técnicas estadísticas o inferenciales.

Codificación de etiquetas

Otro proceso utilizado es la técnica de codificación de etiquetas "*Label Encoding*", esta técnica implica asignar códigos numéricos o etiquetas a categorías específicas, facilitando así la manipulación y análisis de los datos. Cuando se enfrenta a variables categóricas con valores nulos o desconocidos, la codificación de etiquetas puede ser particularmente útil. Al asignar códigos numéricos a diferentes categorías, se simplifica el tratamiento de valores faltantes y se mejora la eficiencia en las etapas posteriores del análisis de datos.

Análisis de correlación

Las correlaciones proporcionadas sugieren varias asociaciones con la variable de enfermedad cardíaca (Heart_Disease). A continuación, se presenta un análisis breve de cada correlación:

Skin_Cancer (Correlación: 0.090848): Existe una correlación positiva débil entre el cáncer de piel y la enfermedad cardíaca. Este hallazgo puede indicar que las personas con antecedentes de cáncer de piel podrían tener una tendencia ligeramente mayor a desarrollar enfermedades cardíacas, aunque la relación es modesta.

Other_Cancer (Correlación: 0.092387): Al igual que con el cáncer de piel, hay una correlación positiva débil con otros tipos de cáncer y la enfermedad cardíaca. Esto podría sugerir una conexión leve entre el historial de cáncer y la predisposición a enfermedades cardíacas.

Diabetes (Correlación: 0.181072): La correlación más fuerte se observa con la diabetes, indicando una asociación más significativa. Las personas con diabetes tienen una correlación más fuerte con enfermedades cardíacas, lo que respalda la comprensión común de que estas dos condiciones de salud están interrelacionadas.

Arthritis (Correlación: 0.153913): Hay una correlación positiva moderada entre la artritis y la enfermedad cardíaca. Este resultado podría sugerir que las personas con artritis tienen cierta propensión a desarrollar enfermedades cardíacas, aunque la relación no es extremadamente fuerte.

Age_Category (Correlación: 0.229011): La edad se correlaciona positivamente con la enfermedad cardíaca de manera significativa. Esto refleja la comprensión general de que el riesgo de enfermedad cardíaca tiende a aumentar con la edad.

Smoking_History (Correlación: 0.107797): Existe una correlación positiva débil entre el historial de fumar y la enfermedad cardíaca. Esto respalda la relación bien establecida entre el tabaquismo y las enfermedades cardíacas.

Es importante recordar que la correlación no implica causalidad directa, y estos resultados pueden requerir análisis estadísticos adicionales para validar y comprender completamente la naturaleza de las relaciones identificadas. Además, factores adicionales podrían influir en estas asociaciones, y un análisis más profundo podría revelar detalles más específicos sobre las interacciones entre las variables.

2.7.3. Fase 3: División del Conjunto de Datos

La división adecuada del conjunto de datos en subconjunto de entrenamiento, validación y prueba es crucial para entrenar y evaluar modelos de aprendizaje automático de manera efectiva. En esta sección, se describe el enfoque utilizado para dividir el conjunto de datos y se justifican las proporciones seleccionadas.

Proporciones Utilizadas

El conjunto de datos se dividió en 2 subconjuntos de la siguiente manera: 70% para entrenamiento, 30% para prueba como sugiere (Vrigazova, 2021). Esta proporción fue elegida con el objetivo de proporcionar un balance entre tener suficientes datos para el entrenamiento del modelo y reservar una cantidad adecuada para evaluar su rendimiento de manera imparcial.

2.7.4. Fase 4: Selección del Modelo

En el trabajo "*Machine Learning Algorithms for Predicting and Preventive Diagnosis of Cardiovascular Disease*" se entrenó 6 algoritmos de aprendizaje de máquina con los datos de 224 pacientes para determinar el mejor en base a métricas de evaluación (Sembina et al., 2022). Dando como resultado que el modelo de Random Forest tiene mejor rendimiento seguido de Logistic Regression como se muestra en Tabla 41.

Tabla 41 Resultados del Trabajo de (Sembina et al., 2022)

Algoritmo	Acurracy	Precision	Recall
Logistic Regression	87.5%	90.62%	87.88%
Support Vector Machine	78.57%	81.25%	87.88%
Naive Bayes	73.21%	53.12%	100.0%
Decision Tree	75.0%	68.75%	84.62%

Random Forest	94.64%	93.75%	96.77%
KNN	75.57%	81.25%	81.25%

Elaboración propia

En el trabajo titulado *"Use of Efficient Machine Learning Techniques in the Identification of Patients with Heart Diseases"* cita la importancia de un diagnóstico preciso y oportuno de las enfermedades cardiovasculares. Para ese fin usaron el conjunto de datos de enfermedades cardiacas de Cleveland para el entrenamiento y evaluación de 3 modelos. Donde Random Forest tiene la mejor Acurracy (99.337%) empleando una validación cruzada de k= 5 fold. (Ghosh et al., 2021)

Tabla 42 Resultados del trabajo (Ghosh et al., 2021)

Algoritmo	Accuracy
Decision Tree	97.17%
Random Forest	99.34%
KNN	90.92%

Elaboración propia

El autor del trabajo *"Comparative Performance Assessment Of Machine Learning Algorithms To Predict Cardiovascular Disease"* enfatiza lo vital de detectar las enfermedades cardiacas antes de que se presenten. Para este estudio se usó las Redes Neuronales, Arboles de Decisión, Naive Bayes, Regresión Logística, Bosques Aleatorios, Máquinas de vector de soporte, K vecinos más cercanos y XG Boost. (Hemalatha et al., 2023)

Los resultados se muestran en la Tabla 43.

Tabla 43 Resultados del trabajo (Hemalatha et al., 2023)

Algoritmos	Precision	Recall	F1-meausure	Acurracy
Decision Tree	0.97	0.96	0.96	95.91
Logistic Regression	0.92	0.98	0.95	94.15
Random Forest	0.97	1.00	0.99	98.25
Naive Bayes	0.92	0.83	0.87	85.96
KNN	0.95	0.88	0.91	90.06
Redes Neuronales	0.96	0.93	0.94	93.57
XG Boost	0.96	0.93	0.94	93.57
SVM	0.92	0.96	0.94	92.98

Elaboración propia

Por otro lado, el trabajo *"Prediction of Heart Disease using Computational Algorithms"* menciona lo difícil que es analizar la gran cantidad de datos del sector de la salud. Es por eso, que aplicar algoritmos de aprendizaje de máquina se vuelve necesario ante este problema mundial. Para este estudio se usó datos sobre enfermedades cardiovasculares de Kaggle con el objetivo de encontrar el algoritmo con mejor precisión. Esto se verifico mediante métricas de rendimiento y matriz de confusión.(Shobha et al., 2022)

Tabla 44 Resultados del trabajo de (Shobha et al., 2022)

Algoritmos	Accuracy	Precision	Recall	F1-Score
Logistic Regression	0.85	0.84	0.91	0.87
KNN	0.76	0.77	0.81	0.79
Decision Tree	0.81	0.87	0.79	0.83
Random Forest	0.90	0.88	0.94	0.91
Naive Bayes	0.86	0.87	0.90	0.94
SVM	0.81	0.81	0.88	0.85

Elaboración propia

Luego del análisis de trabajos similares se consideró el uso de los siguientes algoritmos:

Regresión Logística

Descripción: La Regresión Logística es un algoritmo de clasificación que es especialmente útil para problemas de clasificación binaria. Este algoritmo modela la probabilidad de que una instancia pertenezca a una de las dos clases, basándose en una función logística aplicada a una combinación lineal de las variables predictoras.

Justificación: La elección de la Regresión Logística se fundamenta en su simplicidad, eficiencia computacional y la facilidad de interpretación de los resultados, permitiendo una modelización directa de la probabilidad de presencia o ausencia de enfermedades cardiovasculares. Además, este algoritmo facilita la comprensión de cómo cada factor de riesgo contribuye al resultado, lo cual es crucial para interpretaciones clínicas y recomendaciones de salud preventiva.

Bosques Aleatorios

Descripción: Los Bosques Aleatorios son un método de ensamble que combina múltiples árboles de decisión para mejorar la robustez y precisión de la clasificación o regresión. Cada árbol se entrena con un subconjunto de datos y variables, y el resultado final se obtiene por votación mayoritaria o promedio de las predicciones de todos los árboles.

Justificación: Se seleccionaron los Bosques Aleatorios debido a su capacidad para manejar una gran cantidad de características y su resistencia al sobreajuste, especialmente relevante en contextos con muchos factores predictores, como es el caso de las enfermedades cardiovasculares. La naturaleza del ensamble de múltiples modelos aumenta la precisión de las predicciones y proporciona una medida interna de la importancia de las características, lo cual es invaluable para identificar los principales factores de riesgo. Su eficacia en datasets complejos y su capacidad para manejar tanto variables numéricas como categóricas los hacen particularmente adecuados para esta investigación.

K Vecinos Más Cercanos (KNN)

Descripción: KNN es un algoritmo de aprendizaje supervisado simple pero poderoso que clasifica una nueva instancia basándose en la mayoría de votos de sus 'k' vecinos más cercanos. Es decir, un dato de prueba se asigna a la clase más común entre sus 'k' vecinos más cercanos según una medida de distancia, como la distancia euclidiana. El número de vecinos, 'k', es un parámetro que se debe elegir para optimizar el rendimiento del algoritmo.

Justificación: La elección del algoritmo KNN para este estudio se apoya en su facilidad de implementación y su naturaleza intuitiva, siendo especialmente útil para sistemas de clasificación donde la relación entre los atributos no es lineal. Su capacidad para adaptarse a cambios en los datos en tiempo real lo hace valioso para aplicaciones médicas donde los datos de los pacientes pueden variar significativamente. Además, KNN es útil en casos donde la interpretación de los resultados es importante, ya que la clasificación de una instancia está claramente influenciada por las características de las instancias vecinas más similares. Dado que el riesgo cardiovascular puede ser influido por una combinación de factores que no siempre interactúan de manera lineal, KNN ofrece un método eficaz para capturar estas complejidades sin la necesidad de un modelo subyacente explícito.

Las máquinas de Vectores de Soporte (SVM)

Descripción: Es una técnica de aprendizaje de máquina de clasificación, regresión y detección de outliers, que tiene como objetivo principal encontrar el hiperplano de separación óptimo entre las diferentes clases de datos.

Justificación: La elección de SVM se debe a su gran capacidad para el análisis de enfermedades cardiovasculares, donde la relación entre las variables predictivas y la presencia de ECV se vuelve compleja y no lineal

2.7.5. Fase 5: Entrenamiento del Modelo

El primer escenario de entrenamiento de los modelos a evaluar se realizó mediante Orange Software, que incluye un flujo de trabajo como se muestra en la Figura 30.

Este flujo de trabajo comprende las siguientes etapas:

- Carga del conjunto de datos (Plugin File)
- Visualización de los datos (Plugin Data Table)
- Seleccionar la columna objetivo (Plugin Select Column)
- Prueba y puntaje (Plugin Test & Score)
- Cuatro modelos de aprendizaje:
    - KNN
    - Regresión Logística
    - SVM
    - Random forest
- Tres evaluadores:
    - Matriz de confusión
    - Análisis de región bajo la curva
    - Curva de rendimiento

2.7.6. Fase 6: Configuración de Hiperparámetros

Bosques Aleatorios

Tabla 45 Hiperparámetros Bosques aleatorios

Hiperparámetro	Valor
Número de árboles	10
Profundidad máxima del árbol	5
Número mínimo de muestras requeridas para dividir un nodo	3
Número mínimo de muestras requeridas en cada hoja	5

Fuente: Elaboración propia, 2024.

K Vecinos Más Cercanos (KNN)

Tabla 46 Hiperparámetro KNN

Hiperparámetro	Valor
Número de vecinos	3
Función de ponderación de vecinos	Uniform

Fuente: Elaboración propia, 2024.

SVM

Tabla 47 Hiperparámetro SVM

Hyperparámetro	Valor
C	1
Kernel	RBF

Fuente: Elaboración propia, 2024.

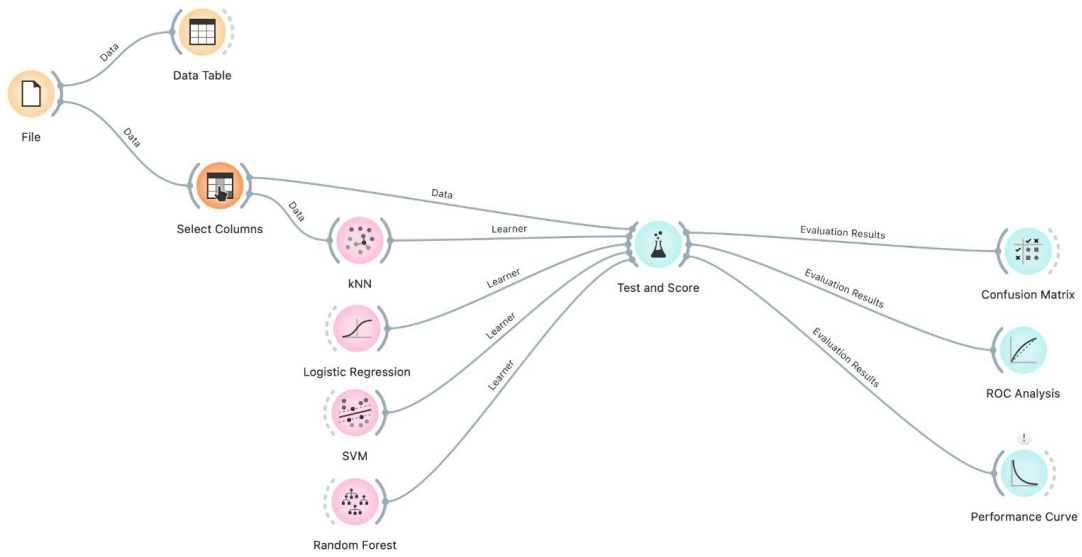
Regresión Logística

Tabla 48 Hiperparámetro Regresión Logística

Hiperparámetro	Valor
Regularización	C=1
Tipo de regularización	Ridge (L2)

Fuente: Elaboración propia, 2024.

Figura 30 Flujo de trabajo en Orange



Fuente: Elaboración propia, 2024.

2.7.7. Fase 7: Evaluación de los Modelos

Las métricas de evaluación seleccionadas son:

Precisión

Expresa la proporción de confianza del modelo en predecir que un individuo es Positivo (Grandini et al., 2020).

$$Precision = \frac{TP}{TP + FP} \quad (1)$$

Recall

Mide la precisión del modelo en encontrar todas las clases positivas en el conjunto de datos (Grandini et al., 2020).

$$Recall = \frac{TP}{TP + FN} \quad (2)$$

Accuracy

Mide la proporción de predicciones correctas realizadas por el modelo sobre el total de predicciones realizadas.

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \quad (3)$$

F1-Score

Métrica para evaluar el rendimiento de un modelo de clasificación, especialmente cuando hay desbalanceo de clases de datos.

$$F1 - Score = 2\left(\frac{precision * recall}{precision + recall}\right) \quad (4)$$

Al final de la Fase 7, se logró elegir el modelo que mejor rendimiento tuvo al momento de detectar casos positivos y negativos, es decir Presencia o Ausencia de Enfermedades Cardiovasculares.

2.7.8. Fase 8: Optimización del Modelo seleccionado

La optimización del modelo involucró 2 escenarios donde se aplicaron tareas de ajuste de hiperparámetro y varias técnicas de preprocesamiento como:

Ingeniería de Características

- PCA para mantener 95% de la varianza y así obtener número de componentes principales.
- La técnica "Selección de características con mejores k características", donde k especifica el número de características que se desean seleccionar en función de una prueba estadística.

Validación cruzada

- Validación cruzada estratificada con k-folds= 5, donde el conjunto de datos se divide 5 veces en partes iguales y realiza el entrenamiento y evaluación del modelo 5 veces.

Balanceo de clases

- Mediante el parámetro class_weight = 'balanced' que ajusta automáticamente los pesos de las clases inversamente proporcionales a sus frecuencias en los datos de entrenamiento.
- SMOTEEN genera de manera sintética nuevas muestras para la clase minoritaria con técnicas de submuestreo para ajustar la proporción entre las clases minoritaria y mayoritaria en el conjunto de datos.

Ajuste de Hiperparámetros

- Búsqueda de hiperparámetros con Grid Search donde se realiza una búsqueda exhaustiva de las combinaciones de hiperparámetros en un espacio definido.

CAPÍTULO 3. RESULTADOS Y DISCUSIÓN

Las pruebas realizadas mediante la herramienta Orange Software muestran un detalle de los criterios de evaluación, obteniendo una comparativa relevante que se muestra en la siguiente tabla.

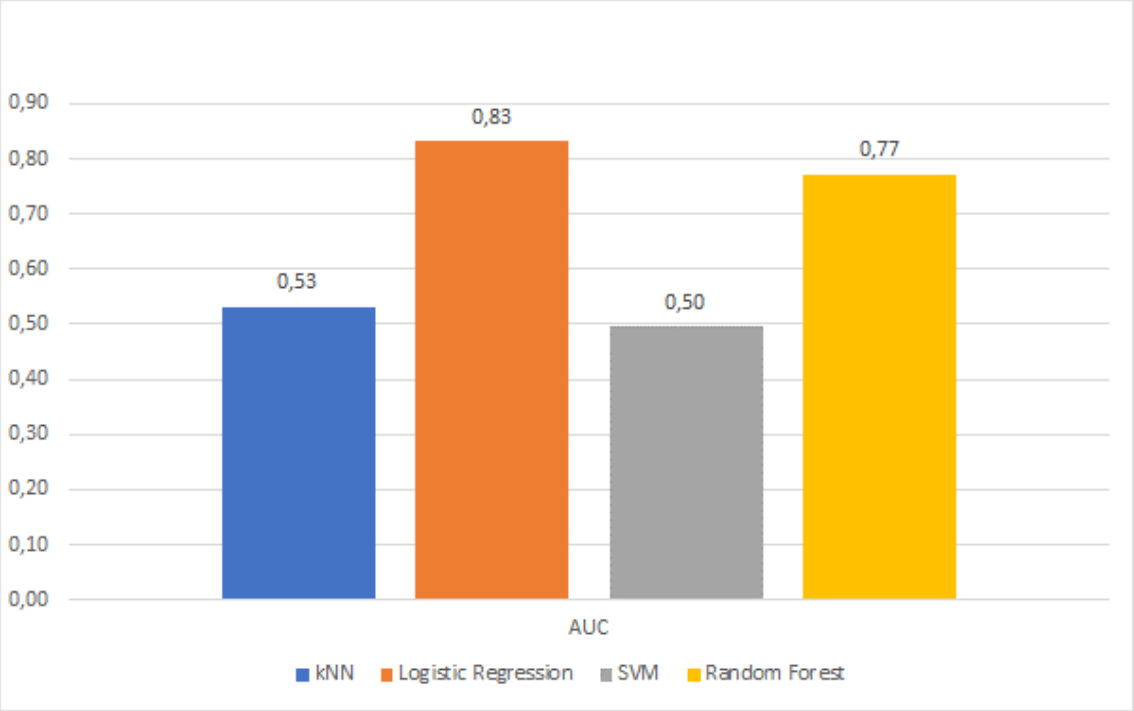
Tabla 49 Métricas de evaluación

Model	AUC	CA	F1	Prec	Recall
kNN	0.532	0.911	0.883	0.863	0.911
Logistic Regression	0.832	0.922	0.893	0.890	0.922
SVM	0.496	0.911	0.882	0.860	0.911
Random Forest	0.771	0.923	0.866	0.852	0.923

Fuente: Elaboración propia, 2024.

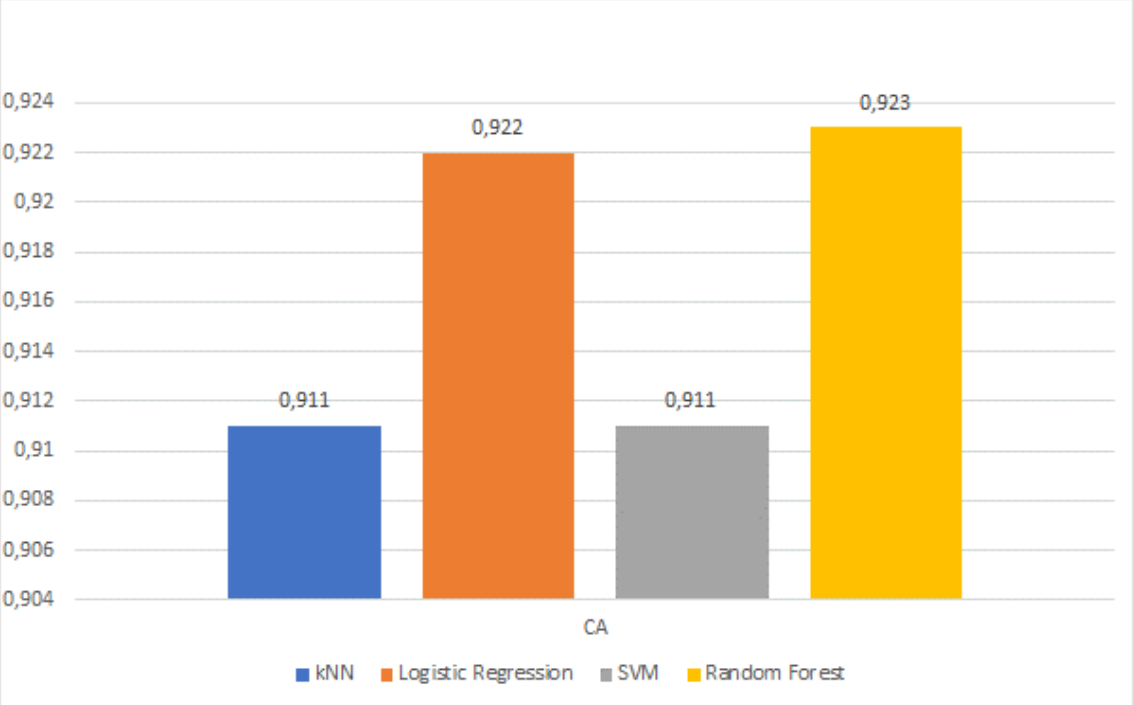
Al analizar los resultados de los diferentes modelos de aprendizaje automático para predecir enfermedades cardiovasculares, se observa que la Regresión Logística y el Bosque Aleatorio destacan como los más prometedores en términos de rendimiento. La Regresión Logística muestra un AUC alto de 0.832 y una precisión de 0.922, lo que sugiere una buena capacidad para distinguir entre casos positivos y negativos. Por otro lado, el Bosque Aleatorio también exhibe un rendimiento sólido, con un AUC de 0.771 y una precisión del 92.3%. Estos resultados indican que ambos modelos son capaces de clasificar correctamente una proporción significativa de muestras. Sin embargo, el SVM muestra un rendimiento inferior con un AUC de 0.496, lo que sugiere dificultades para distinguir entre las clases. A pesar de tener una precisión comparativamente alta de 0.911, su capacidad para identificar correctamente los casos positivos es limitada. Este análisis destaca la importancia de considerar múltiples métricas de rendimiento al evaluar modelos de aprendizaje automático y sugiere que la Regresión Logística y el Bosque Aleatorio son candidatos prometedores para la predicción de enfermedades cardiovasculares.

Figura 31 Análisis comparativo de la métrica AUC



Fuente: Elaboración propia, 2024.

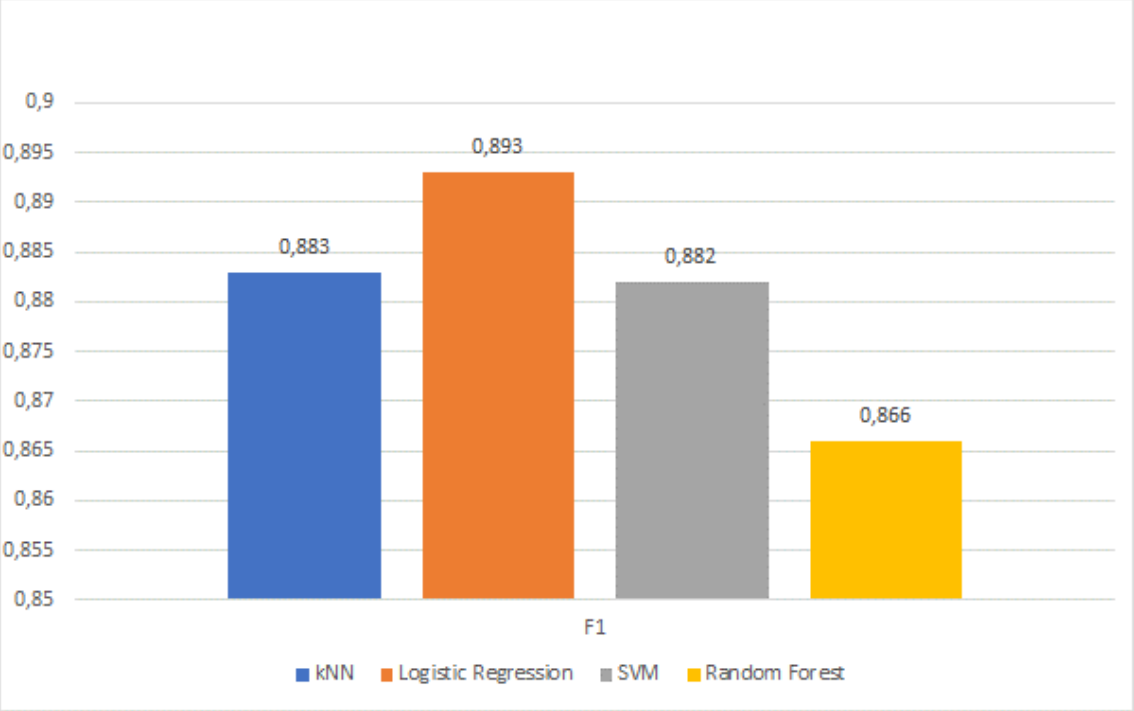
Figura 32 Análisis comparativo de la métrica CA



Fuente: Elaboración propia, 2024.

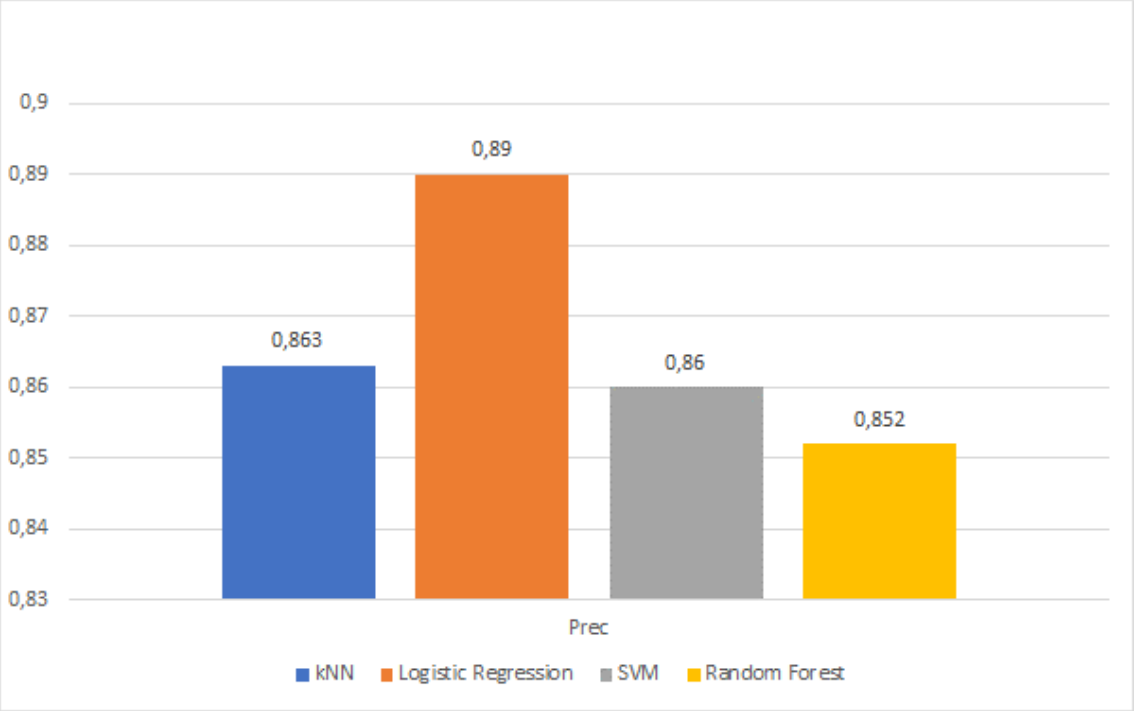
Por otro lado, la Regresión Logística (RL) tiene un F1-score de 0.893 (**Figura 33**), lo que indica una buena armonización entre precisión y recall. Esto significa que el modelo es capaz de identificar correctamente un alto porcentaje de los casos positivos (recall), al tiempo que minimiza los falsos positivos (precisión). Esta capacidad equilibrada es crucial en la predicción de enfermedades cardiovasculares, donde tanto la detección temprana de casos positivos como la minimización de diagnósticos erróneos son igualmente importantes. Muy cerca le siguen, KNN y SVM con 0.883 y 0.882 respectivamente. El modelo Random Forest mostró un rendimiento ligeramente inferior con un F1-score de 0.866. Estos hallazgos sugieren que, aunque todos los modelos demostraron habilidades competitivas en la tarea de clasificación, la Regresión Logística exhibió la mejor capacidad de generalización según la métrica F1-score en el conjunto de datos evaluado.

Figura 33 Análisis comparativo de la métrica F1 Score



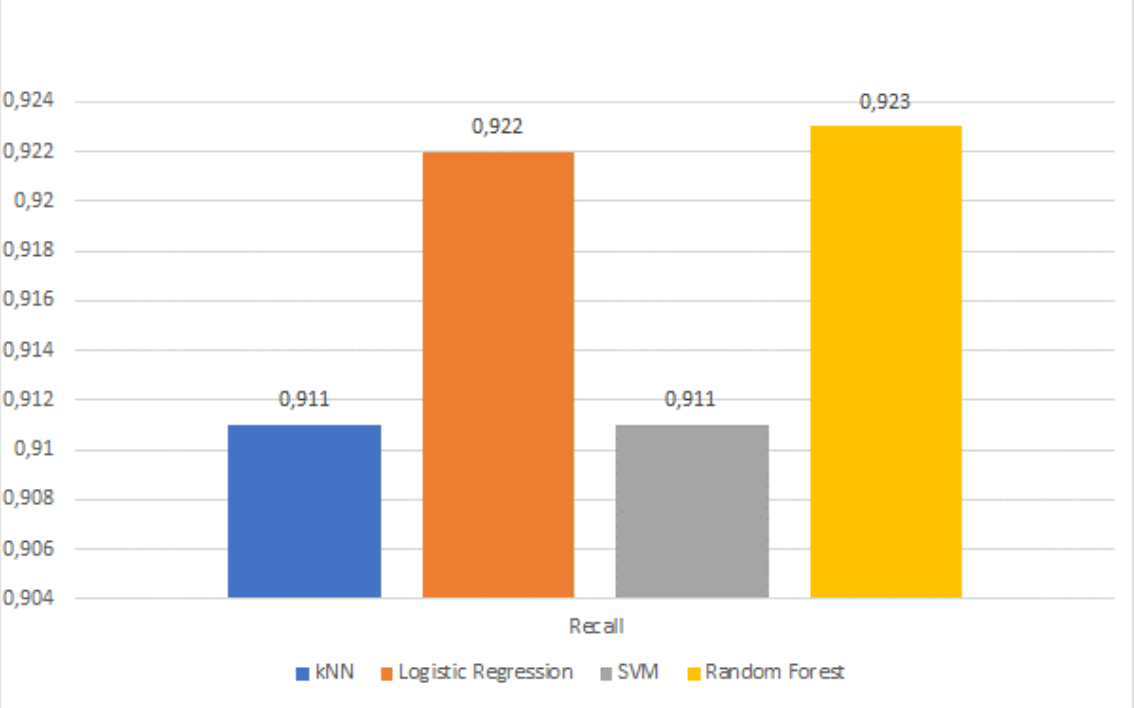
Fuente: Elaboración propia, 2024.

Figura 34 Análisis comparativo de la métrica Precisión



Fuente: Elaboración propia, 2024.

Figura 35 Análisis comparativo de la métrica Recall



Fuente: Elaboración propia, 2024.

Es fundamental reconocer que estas diferencias pueden deberse a variaciones en los conjuntos de datos, características específicas y enfoques metodológicos. A pesar de las discrepancias, la consistencia en la superioridad de mis modelos en términos de precisión sugiere la posibilidad de que las particularidades del conjunto de datos utilizado se alineen de manera más favorable con los algoritmos de aprendizaje de máquina utilizados.

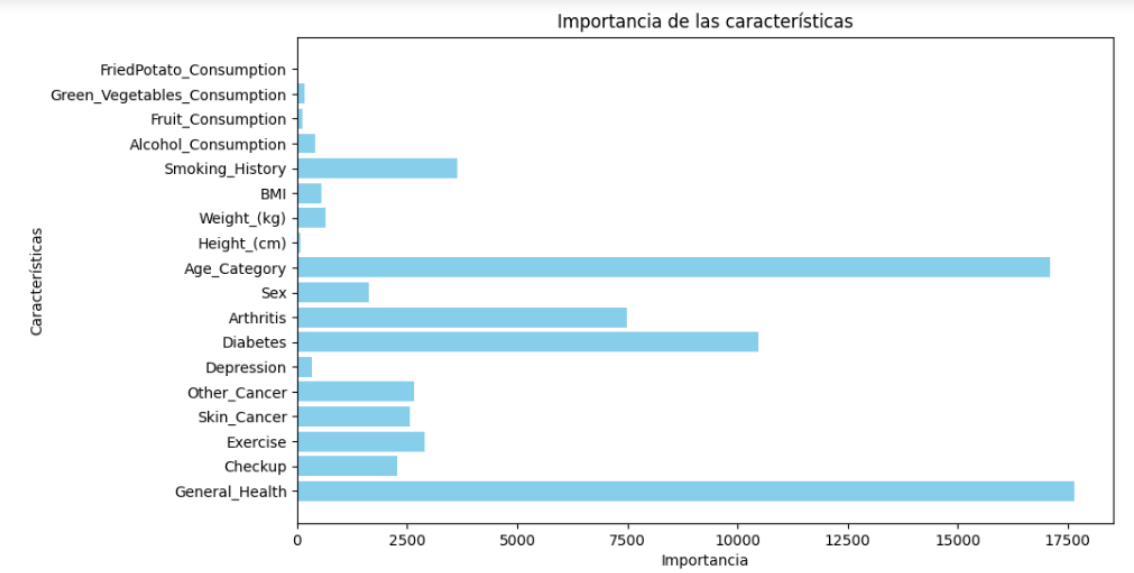
La evaluación de los modelos utilizados demuestra un comportamiento regular por parte de Regresión Logística, por lo tanto, se realizó un análisis más exhaustivo antes del entrenamiento, que van desde tareas de limpieza de datos, normalización, ingeniería de características, además de tomar en cuenta aspectos importantes del conjunto de datos como el desequilibrio de clases, y el ajuste de hiperparámetros. Luego de la optimización del modelo:

Tabla 50 Configuración final Regresión Logística Escenario 1 y 2

Hiperparámetro	Valor
Regularización	C=0.001
Tipo de regularización	Ridge (L2)

Fuente: Elaboración propia, 2024.

Figura 36 Análisis de la Importancia de las características Escenario 1



Fuente: Elaboración propia, 2024.

Tabla 51 Informe de Clasificación antes de la optimización con datos de evaluación

	Precision	Recall	F1-Score
0	**0.92**	**0.99**	**0.96**
1	**0.52**	**0.06**	**0.11**
Acurracy			**0.92**

Fuente: Elaboración propia, 2024

Tabla 52 Informe de Clasificación después de la optimización con datos de evaluación- Escenario 1

	Precision	Recall	F1-Score
0	**0.97**	**0.72**	**0.83**
1	**0.20**	**0.79**	**0.32**
Acurracy			**0.72**

Fuente: Elaboración propia, 2024.

Tabla 53 Informe de Clasificación después de la optimización con datos de evaluación- Escenario 2

	Precision	Recall	F1-Score
0	**0.96**	**0.76**	**0.85**
1	**0.20**	**0.65**	**0.30**
Acurracy			**0.75**

Fuente: Elaboración propia, 2024

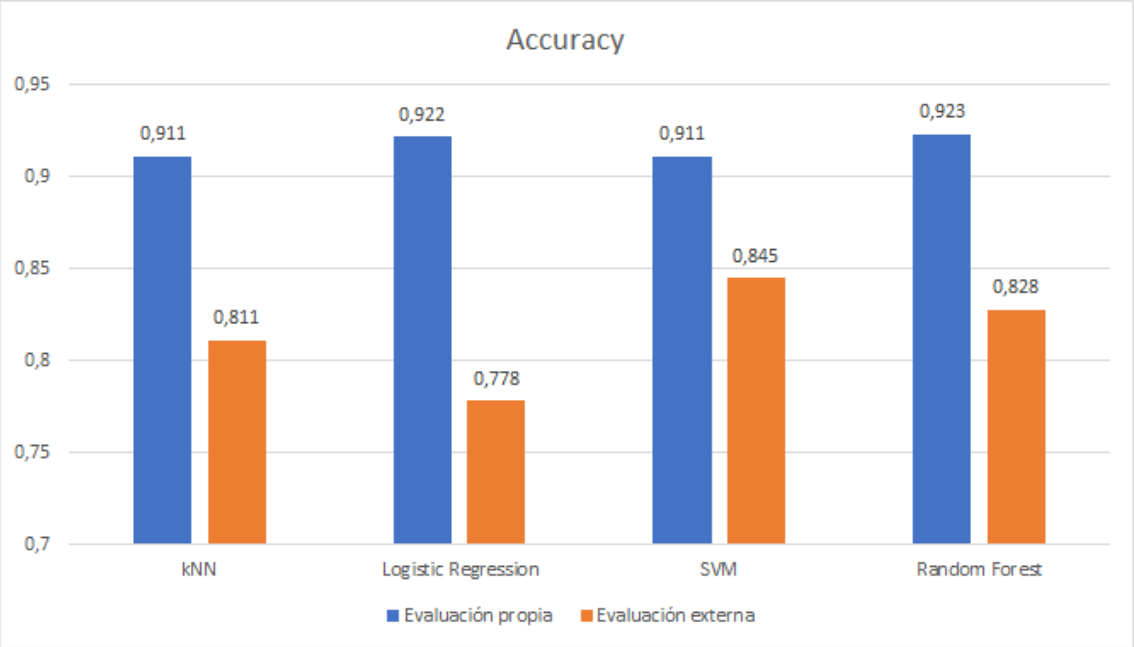
Los resultados muestran que, antes de la optimización del modelo, se observaba un alto desequilibrio en la capacidad de clasificación entre las clases positivas y negativas de enfermedades cardiovasculares, lo que resultaba en una baja precisión y recall para la clase 1. Sin embargo, tras la implementación de técnicas de optimización, como la validación cruzada y el balanceo de clases, se evidenció una mejora significativa en el desempeño del modelo, particularmente en términos de F1-Score y recall para ambas clases. A pesar de una ligera disminución en la precisión global del modelo, se logró un equilibrio más adecuado entre la precisión y el recall, lo que indica una mayor capacidad para identificar de manera precisa los casos positivos de enfermedades cardiovasculares. Estos hallazgos respaldan la eficacia de la optimización del modelo mediante técnicas de preprocesamiento para mejorar su capacidad predictiva y su utilidad en la detección temprana de enfermedades cardiovasculares.

Discusión

La comparación entre los resultados de efectividad de los modelos de aprendizaje de máquina para el diagnóstico de enfermedades cardiovasculares, presentados en los gráficos anteriores y en los proporcionados por (Sun, 2022) resaltan notables discrepancias y similitudes. En términos de Regresión Logística, mi estudio muestra un rendimiento significativamente superior, con una precisión de 0.922, en comparación con

el valor de 0.7784 reportado por el autor. Esto sugiere una capacidad más sólida de la Regresión Logística en mi conjunto de datos específico para clasificar correctamente los casos. Además, la comparación revela que, en general, los modelos evaluados en este proyecto: k-Nearest Neighbors (kNN), Support Vector Machine (SVM) y Random Forest, exhiben una efectividad superior en términos de precisión, destacándose principalmente en el caso de Random Forest como se muestra en la siguiente figura.

Figura 37 Análisis con estudio externo



Fuente: Elaboración propia, 2024.

En el análisis comparativo del rendimiento de la regresión logística en la predicción de enfermedades cardiovasculares, se observa una notable variabilidad en los resultados entre diferentes estudios. En nuestro propio estudio, la regresión logística muestra una precisión del 97% para la clase negativa, lo que indica una eficacia considerable en la identificación de casos de ausencia de enfermedades cardiovasculares. Sin embargo, el recall para la clase positiva es sustancialmente inferior, alcanzando solo el 72%. Esta discrepancia en la capacidad de identificar correctamente los casos positivos de enfermedades cardiovasculares sugiere limitaciones en la sensibilidad del modelo. Al comparar estos resultados con otros estudios, se observa una tendencia similar en algunos casos, donde la regresión logística muestra un buen rendimiento en términos de precisión, pero puede presentar desafíos en la identificación de casos positivos. Por ejemplo, el estudio de (Sembina et al., 2022) también encontró que la regresión logística tenía una

alta precisión pero un recall ligeramente inferior al 90%, lo que indica una capacidad limitada para detectar casos positivos de enfermedades cardiovasculares en comparación con otros algoritmos como Random Forest.

Por otro lado, otros estudios resaltan el potencial de la regresión logística en la predicción de enfermedades cardiovasculares. (Hemalatha et al., 2023) encontraron que la regresión logística mostraba una precisión del 92% y un recall del 98%, lo que sugiere un rendimiento robusto en la identificación de casos positivos. Estos resultados contrastan con los obtenidos en nuestro estudio y resaltan la importancia de considerar las diferencias en los conjuntos de datos y enfoques metodológicos entre estudios al interpretar el rendimiento de la regresión logística en la predicción de enfermedades cardiovasculares. En general, si bien la regresión logística puede demostrar una precisión sólida en algunos casos, su capacidad para identificar de manera efectiva los casos positivos de enfermedades cardiovasculares puede variar y podría beneficiarse de estrategias adicionales de optimización y refinamiento del modelo.

CONCLUSIONES

El presente trabajo ha proporcionado la evaluación de modelos de aprendizaje de máquina para medir la susceptibilidad de una persona a contraer enfermedades cardiovasculares por medio de sus factores de riesgo conocidos. Para la elección de los modelos se realizó a través de un proceso meticuloso donde se determinó que la Regresión Logística es un modelo eficiente y de clara interpretación a partir de sus resultados. Por su parte los Bosques Aleatorios han demostrado ser capaces de manejar la complejidad y el ruido de los datos, asegurando una evaluación precisa. La aplicación de KNN ha mostrado la importancia de considerar la proximidad de los casos en el espacio de características para entender mejor la complejidad de las relaciones entre factores de riesgo, mientras que SVM demostró ser crucial para clasificar eficazmente los casos en presencia de relaciones complejas y no lineales entre dichos factores.

Al medir el rendimiento de los modelos seleccionados, se ha podido observar diferencias significativas en su capacidad para predecir la susceptibilidad de una persona a contraer enfermedades cardiovasculares. De los modelos analizados, Regresión Logística demostró un rendimiento superior con un AUC DE 0.832, una exactitud de 0.922, una puntuación de F1 de 0.893, una precisión de 0.890 y una sensibilidad de 0.922. Estos resultados sugieren la capacidad del modelo en distinguir entre pacientes con y sin enfermedades cardiovasculares. En el caso de KNN y SVM mostraron un valor de AUC menor, 0.5532 y 0.496 respectivamente, esto indica una reducida capacidad para clasificar casos positivos y negativos. Por su parte, los bosques aleatorios presentaron un desempeño equilibrado con un AUC de 0.771 y una exactitud de 0.923 a pesar de una puntuación de 0.866 que es ligeramente inferior a los otros modelos, reflejando su facultad para tratar datos complejos y relaciones no lineales. En este sentido, estos hallazgos son prometedores y respaldan la importancia de utilizar modelos predictivos avanzados, en la identificación temprana y el manejo de las enfermedades cardiovasculares como lo indican los especialistas. Sin embargo, también se destaca la necesidad de investigaciones adicionales para validar estos resultados en diferentes poblaciones y entornos clínicos, así como para explorar el potencial de otros modelos de aprendizaje automático en este contexto.

Con la aplicación de técnicas de optimización se mejoró algunas métricas, pero también introdujo cambios significativos en otras. Es esencial realizar un análisis más detallado

para comprender las razones detrás de estas variaciones y determinar si los ajustes adicionales son necesarios para mejorar el rendimiento general del modelo. En este sentido, la precisión para la clase 0 mejoró después de la optimización, la precisión global del modelo disminuyó del 0.92 al 0.72. Esto se debe al balance de las clases y la capacidad que tiene el modelo en detectar correctamente cada una de ellas, mostrando una mejoría notable en la capacidad de detectar correctamente presencia de ECV, pues su índice aumenta de 0.06 a 0.79.

Por su parte, los especialistas consideran la importancia de contextualizar los resultados de los modelos en diferentes escenarios, pues, si bien hay una notable discrepancia entre los valores de precisión y la sensibilidad obtenida por los modelos, luego de evaluar con datos reales el resultado muestra un índice de confianza muy alto comparando las predicciones con los diagnósticos en base a experiencia, lo que revela la capacidad de este en proporcionar una evaluación precisa y útil en la práctica médica diaria.

RECOMENDACIONES

Luego de la evaluación de los modelos de aprendizaje de máquina se subraya la importancia de la selección adecuada del modelo de aprendizaje de máquina, considerando tanto la naturaleza de los datos como el objetivo específico del estudio, para mejorar la predicción y prevención de enfermedades cardiovasculares en poblaciones. La integración de estos modelos, cada uno con sus fortalezas únicas, proporciona una plataforma sólida para el avance en la identificación de individuos en riesgo, enfatizando la relevancia de una aproximación holística y multidimensional en la investigación cardiovascular.

A pesar de los resultados obtenidos, se recomienda realizar una exploración exhaustiva de las características presente en el dataset, que puede llevar a la incorporación de nuevas variables con una fuerte relación con el problema de estudio o la aplicación de ingeniería de características con el fin de capturar mejor la complejidad de los datos. También es importante comprender y analizar a profundidad los resultados de los diferentes modelos porque estos pueden proporcionar información valiosa y significativa asociados con las enfermedades cardiovasculares.

Dada la importancia de una alta sensibilidad en la detección de casos positivos, se recomienda enfocar los esfuerzos en mejorar esta métrica para garantizar una identificación efectiva de los casos de interés. Para lograrlo, se sugiere explorar técnicas adicionales de ajuste de hiperparámetros y estrategias de balanceo de clases que puedan ayudar a reducir la cantidad de falsos negativos y, por lo tanto, mejorar la capacidad del modelo para detectar correctamente los casos positivos. Además, es crucial continuar monitoreando y evaluando regularmente el rendimiento del modelo, realizando ajustes según sea necesario. La mejora continua en la sensibilidad del modelo no solo fortalecerá su capacidad para identificar adecuadamente los casos positivos, sino que también aumentará su utilidad y relevancia en aplicaciones prácticas.

Es importante tener en cuenta el costo computacional al aplicar técnicas como la validación cruzada y grid_search en el entrenamiento de modelos de aprendizaje automático. Estas técnicas pueden ser computacionalmente intensivas, especialmente cuando se utilizan en conjuntos de datos grandes o con modelos complejos que requieren una búsqueda exhaustiva de hiperparámetros. Por lo tanto, se recomienda considerar cuidadosamente la capacidad computacional disponible y el tiempo necesario para

completar estos procesos. Además, es útil explorar alternativas más eficientes, como el uso de técnicas de validación cruzada estratificada o el ajuste de hiperparámetros de forma más selectiva para reducir el tiempo de computación sin comprometer la calidad del modelo. Al hacerlo, se puede maximizar la eficiencia del proceso de modelado y optimización de hiperparámetros, lo que permite una investigación más ágil y resultados más rápidos.

REFERENCIAS

Abuelgasim, E., Shah, S., Abuelgasim, B., Soni, N., Thomas, A., Elgasim, M., & Harky, A. (2021). Clinical overview of diabetes mellitus as a risk factor for cardiovascular death. *Reviews in Cardiovascular Medicine*, *22*(2), 301. https://doi.org/10.31083/j.rcm2202038

AEL. (2022). *Universidad de Colima*. https://recursos.ucol.mx/tesis/investigacion.php

Agmon Nardi, I., & Iakobishvili, Z. (2018). Cardiovascular Risk in Cancer Survivors. *Current Treatment Options in Cardiovascular Medicine*, *20*(6), 47. https://doi.org/10.1007/s11936-018-0645-8

Ahsan, M. M., Luna, S. A., & Siddique, Z. (2022). Machine-Learning-Based Disease Diagnosis: A Comprehensive Review. *Healthcare*, *10*(3), 541. https://doi.org/10.3390/healthcare10030541

Alphiree (Owner). (2021). *Cardiovascular Diseases Risk Prediction Dataset*. https://www.kaggle.com/datasets/alphiree/cardiovascular-diseases-risk-prediction-dataset

Areiza, M., Osorio, E., Ceballos, M., & Amariles, P. (2018). Conocimiento y factores de riesgo cardiovascular en pacientes ambulatorios. *Revista Colombiana de Cardiología*, *25*(2), 162-168. https://doi.org/10.1016/j.rccar.2017.07.011

Ashipala, D., Tomas, N., Medusalem, J. M. H., Ashipala, D., Tomas, N., & Medusalem, J. M. H. (1d. C., enero 1). *Smoking: A Biopsychosocial Perspective* (smoking) [Chapter]. Https://Services.Igi-Global.Com/Resolvedoi/Resolve.Aspx?Doi=10.4018/978-1-7998-2139-7.Ch006; IGI Global. https://www.igi-global.com/gateway/chapter/www.igi-global.com/gateway/chapter/252421

Barczewski, A., Bezerianos, A., & Boukhelifa, N. (2020). How Domain Experts Structure Their Exploratory Data Analysis: Towards a Machine-Learned Storyline. *Extended Abstracts of the 2020 CHI Conference on Human Factors in Computing Systems*, 1-8. https://doi.org/10.1145/3334480.3382845

Camafort, M., Alcocer, L., Coca, A., Lopez-Lopez, J. P., López-Jaramillo, P., Ponte-Negretti, C. I., Sebba-Barroso, W., Valdéz, O., & Wyss, F. (2021). Registro Latinoamericano de monitorización ambulatoria de la presión arterial (MAPA-LATAM): Una necesidad urgente. *Revista Clínica Española*, *221*(9), 547-552. https://doi.org/10.1016/j.rce.2021.02.002

Castillo, N., Malo, M., Villacres, N., Chauca, J., Cornetero, V., de Flores, K. R., Tapia, R., & Ríos, R. (2017). Metodología para la estimación de costos directos de la atención integral para enfermedades no trasmisibles. *Revista Peruana de Medicina Experimental y Salud Pública*, *34*, 119-125. https://doi.org/10.17843/rpmesp.2017.341.2774

CEAP ESPOL. (2023). *Ecuador acumula pacientes con enfermedades cardiovasculares | CEAP :: Centro de Estudios Asia-Pacífico*. https://ceap.espol.edu.ec/es/content/ecuador-acumula-pacientes-con-enfermedades-cardiovasculares

Croyle, R. T., & Jemmott, J. B. (1991). *Psychological Reactions to Risk Factor Testing* (J. A. Skelton & R. T. Croyle, Eds.; pp. 85-107). Springer US. https://doi.org/10.1007/978-1-4613-9074-9_5

DeMizio, D. J., & Geraldino-Pardilla, L. B. (2020). Autoimmunity and Inflammation Link to Cardiovascular Disease Risk in Rheumatoid Arthritis. *Rheumatology and Therapy*, *7*(1), 19-33. https://doi.org/10.1007/s40744-019-00189-0

DinuA, J., & Joseph, F. (2017). *A study on Deep Machine Learning Algorithms for diagnosis of diseases*. https://www.semanticscholar.org/paper/A-study-on-Deep-Machine-Learning-Algorithms-for-of-DinuA-Joseph/513ea1b5f2bbee25522f8c07636d7447fbd9ea5c

Eckel, R. H., Bornfeldt, K. E., & Goldberg, I. J. (2021). Cardiovascular disease in diabetes, beyond glucose. *Cell Metabolism*, *33*(8), 1519-1545. https://doi.org/10.1016/j.cmet.2021.07.001

Gan, Y., Tong, X., Li, L., Cao, S., Yin, X., Gao, C., Herath, C., Li, W., Jin, Z., Chen, Y., & Lu, Z. (2015). Consumption of fruit and vegetable and risk of coronary heart disease: A meta-analysis of prospective cohort studies. *International Journal of Cardiology*, *183*, 129-137. https://doi.org/10.1016/j.ijcard.2015.01.077

Ghosh, P., Azam, S., Karim, A., Jonkman, M., & Hasan, Md. Z. (2021). Use of Efficient Machine Learning Techniques in the Identification of Patients with Heart Diseases. *2021 the 5th International Conference on Information System and Data Mining*, 14-20. https://doi.org/10.1145/3471287.3471297

Grandini, M., Bagli, E., & Visani, G. (2020, agosto 13). *Metrics for Multi-Class Classification: An Overview*. arXiv.Org. https://arxiv.org/abs/2008.05756v1

Hannawi, S., & Al Salmi, I. (2021). *[PDF] Cardiovascular Risk in Rheumatoid Arthritis: Literature Review | Semantic Scholar*. https://www.semanticscholar.org/paper/Cardiovascular-Risk-in-Rheumatoid-Arthritis%3A-Review-Hannawi-Hannawi/62495e04928cf6609ebe9fe20734c6a632cdf77e

Hemalatha, S., Kavitha, T., Niruba, D., Nandhakumar, S., & Venkatesh, R. (2023). Comparative Performance Assessment Of Machine Learning Algorithms To Predict Cardiovascular Disease. *2023 International Conference on Computer Communication and Informatics (ICCCI)*, 1-9. https://doi.org/10.1109/ICCCI56745.2023.10128547

Hernández-Martínez, J. C., Varona-Uribe, M., & Hernández, G. (2020). Prevalencia de factores asociados a la enfermedad cardiovascular y su relación con el ausentismo laboral de los trabajadores de una entidad oficial. *Revista Colombiana de Cardiologia*, *27*(2), 109-116. https://doi.org/10.1016/j.rccar.2018.11.004

Huamani Mantari, S. (2019). Habilidades de investigación pedagógica en los docentes de primaria. *Universidad Nacional de Tumbes*. https://repositorio.untumbes.edu.pe/handle/20.500.12874/1641

IGM. (2017). *Geoportal Ecuador – Infraestructura de Datos Espaciales*. https://www.geoportaligm.gob.ec/portal/

INEC. (2010). *Población y Demografía*. Instituto Nacional de Estadística y Censos. https://www.ecuadorencifras.gob.ec/censo-de-poblacion-y-vivienda/

INEC. (2018). *Salud, Salud Reproductiva y Nutrición* |. https://www.ecuadorencifras.gob.ec/salud-salud-reproductiva-y-nutricion/

INEC. (2021). *Actividad Física y Sedentarismo* |. https://www.ecuadorencifras.gob.ec/actividad-fisica-y-sedentarismo/

INEC SALUD. (2017). *INEC SALUD*. https://www.ecuadorencifras.gob.ec/documentos/web-inec/Sitios/inec_salud/index.html

Inoue, T. (2004). Cigarette Smoking as a Risk Factor of Coronary Artery Disease and its Effects on Platelet Function. *Tobacco Induced Diseases, 2*(1), 2. https://doi.org/10.1186/1617-9625-2-2

Kringos, D., Nuti, S., Anastasy, C., Barry, M., Murauskiene, L., Siciliani, L., & De Maeseneer, J. (2019). Re-thinking performance assessment for primary care: Opinion of the expert panel on effective ways of investing in health. *European Journal of General Practice, 25*(1), 55-61. https://doi.org/10.1080/13814788.2018.1546284

Lancheros Florián, L. C. (2012). *Investigación no Experimental*. https://repositorio.konradlorenz.edu.co/handle/001/2317

Lau, E. S., Paniagua, S. M., Liu, E., Jovani, M., Li, S. X., Takvorian, K., Suthahar, N., Cheng, S., Splansky, G. L., Januzzi, J. L., Wang, T. J., Vasan, R. S., Kreger, B., Larson, M. G., Levy, D., De Boer, R. A., & Ho, J. E. (2021). Cardiovascular Risk Factors Are Associated With Future Cancer. *JACC: CardioOncology*, *3*(1), 48-58. https://doi.org/10.1016/j.jaccao.2020.12.003

López-Jaramillo, P., López-López, J., Rey, J. J., & Camacho, P. A. (2018). *EPIDEMIOLOGÍA Y DISTRIBUCIÓN REGIONAL*.

Menéndez, S. S. (2023). *Enfermedades Cardiovasculares*.

OMS. (2017). *Enfermedad pulmonar obstructiva crónica (EPOC)*. https://www.who.int/es/news-room/fact-sheets/detail/cardiovascular-diseases-(cvds)

OMS. (2023). *Enfermedades cardiovasculares*. https://www.who.int/es/health-topics/cardiovascular-diseases

PAHO. (2023a). *Enfermedades cardiovasculares—OPS/OMS | Organización Panamericana de la Salud*. https://www.paho.org/es/temas/enfermedades-cardiovasculares

PAHO. (2023b). *Informe de Ecuador: Mejorando la salud cardiovascular desde comunidades locales hasta el nivel nacional con un enfoque participativo - OPS/OMS | Organización Panamericana de la Salud*. https://www.paho.org/es/noticias/16-5-2023-informe-ecuador-mejorando-salud-cardiovascular-desde-comunidades-locales-hasta

Patino, J. D. P., & Arbelaz, I. C. L. (2016). GESTIÓN HUMANA DE ORIENTACIÓN ANALÍTICA: UN CAMINO PARA LA RESPONSABILIZACIÓN. *Revista de Administração de Empresas*, *56*(1), 101-113. https://doi.org/10.1590/S0034-759020160109

Penninx, B. W. J. H. (2017). Depression and cardiovascular disease: Epidemiological evidence on their linking mechanisms. *Neuroscience & Biobehavioral Reviews*, *74*, 277-286. https://doi.org/10.1016/j.neubiorev.2016.07.003

Piano, M. (2017). *Alcohol Research: Current Reviews;*

Prefectura Santa Elena. (2009). *Provincialización*. https://www.santaelena.gob.ec/index.php/provincializacion/23santa-elena/santa-elena

Pronin, S., & Sotnikov, A. (2022). Using the Orange platform for data analysis. *Bulletin of Kharkov National Automobile and Highway University*, *99*, 131. https://doi.org/10.30977/BUL.2219-5548.2022.99.0.131

Ramos, M., Tinajero, M., Monge Moreno, A. M., López, P., & Galarraga, E. (2021). Factores de riesgo cardiovascular en estudiantes de la Universidad Técnica de

Ambato, Ecuador. *GICOS: Revista del Grupo de Investigaciones en Comunidad y Salud*, *6*(4), 23-36.

Raval, D. Y., Bhatt, D. N., Kumhar, M., Parikh, V., & Vyas, D. (2016). *Medical Diagnosis System Using Machine Learning*. https://www.semanticscholar.org/paper/Medical-Diagnosis-System-Using-Machine-Learning-Raval-Bhatt/29bb49b42f15d92ad4a9d2ee60d52cea9f2bd6a9

Rehm, J., & Roerecke, M. (2017). Cardiovascular effects of alcohol consumption. *Trends in Cardiovascular Medicine*, *27*(8), 534-538. https://doi.org/10.1016/j.tcm.2017.06.002

Semantic Scholar. (2024). *Semantic Scholar | AI-Powered Research Tool*. https://www.semanticscholar.org/

Sembina, G., Aitim, A., & Shaizat, M. (2022). Machine Learning Algorithms for Predicting and Preventive Diagnosis of Cardiovascular Disease. *2022 International Conference on Smart Information Systems and Technologies (SIST)*, 1-5. https://doi.org/10.1109/SIST54437.2022.9945708

Shobha, V., Smitha, C., & Kodipalli, A. (2022). Prediction of Heart Disease using Computational Algorithms. *2022 International Conference on Distributed Computing, VLSI, Electrical Circuits and Robotics (DISCOVER)*, 287-292. https://doi.org/10.1109/DISCOVER55800.2022.9974702

Silverman, A. L., Herzog, A. A., & Silverman, D. I. (2019). Hearts and Minds: Stress, Anxiety, and Depression: Unsung Risk Factors for Cardiovascular Disease. *Cardiology in Review*, *27*(4), 202-207. https://doi.org/10.1097/CRD.0000000000000228

Singh, P., Singh, N., Singh, K. K., & Singh, A. (2021). *Diagnosing of disease using machine learning*. 89-111. https://doi.org/10.1016/B978-0-12-821229-5.00003-3

Suarez Villa, M. E., Navarro Agamez, M. D. J., Caraballo Robles, D. R., López Mozo, L. V., Recalde Baena, A. C., Suarez Villa, M. E., Navarro Agamez, M. D. J., Caraballo Robles, D. R., López Mozo, L. V., & Recalde Baena, A. C. (2020). Estilos de vida relacionados con factores de riesgo cardiovascular en estudiantes Ciencias de la Salud. *Ene*, *14*(3). https://scielo.isciii.es/scielo.php?script=sci_abstract&pid=S1988-348X2020000300007&lng=es&nrm=iso&tlng=es

Sullca, P. R. D. la C. (2020). El hipotético-deductivismo en la explicación de las ciencias sociales. *Horizonte de la Ciencia*, *10*(18), Article 18. https://doi.org/10.26490/uncp.horizonteciencia.2020.18.430

Sun, W. (2022, marzo). *Using Machine Learning Approach to Identify and Analyze High Risks Patients with Heart Disease*. 2022 International Conference on Biotechnology, Life Science and Medical Engineering. https://doi.org/10.23977/blsme.2022028

The Texas Heart Institute. (2023). *Factores de riesgo cardiovascular*. The Texas Heart Institute. https://www.texasheart.org/heart-health/heart-information-center/topics/factores-de-riesgo-cardiovascular/

Vaishnav, D., & Rao, B. R. (2018). Comparison of Machine Learning Algorithms and Fruit Classification using Orange Data Mining Tool. *2018 3rd International Conference on Inventive Computation Technologies (ICICT)*, 603-607. https://doi.org/10.1109/ICICT43934.2018.9034442

Vanuzzo, D., Pilotto, L., Mirolo, R., & Pirelli, S. (2008). [Cardiovascular risk and cardiometabolic risk: An epidemiological evaluation]. *Giornale Italiano Di Cardiologia (2006)*, *9*(4 Suppl 1), 6S-17S.

Vrigazova, B. (2021). The Proportion for Splitting Data into Training and Test Set for the Bootstrap in Classification Problems. *Business Systems Research Journal*, *12*(1), 228-242.

Zhan, J., Liu, Y.-J., Cai, L.-B., Xu, F.-R., Xie, T., & He, Q.-Q. (2017). Fruit and vegetable consumption and risk of cardiovascular disease: A meta-analysis of prospective cohort studies. *Critical Reviews in Food Science and Nutrition*, *57*(8), 1650-1663. https://doi.org/10.1080/10408398.2015.1008980

Zhao, C.-N., Meng, X., Li, Y., Li, S., Liu, Q., Tang, G.-Y., & Li, H.-B. (2017). Fruits for Prevention and Treatment of Cardiovascular Diseases. *Nutrients*, *9*(6), 598. https://doi.org/10.3390/nu9060598

Zheng, H. C., Onderko, L., & Francis, S. A. (2017). Cardiovascular Risk in Survivors of Cancer. *Current Cardiology Reports*, *19*(7), 64. https://doi.org/10.1007/s11886-017-0873-7

ANEXOS

Anexo 1: Formato de encuesta a especialistas de la salud

Objetivo: Establecer cuáles son los factores que inciden en la presencia o ausencia de ECV y medir el nivel de conocimiento de los profesionales de la salud sobre el uso de Modelos de Aprendizaje de Máquina para el diagnóstico de enfermedades cardiovasculares.

1. ¿Cuál es su especialidad médica?
    a. Medicina General
    b. Cardiología
    c. Otros
2. ¿Cuánto tiempo lleva tratando pacientes con enfermedades cardiovasculares?
    a. De 1 a 5 años
    b. De 5 a 10 años
    c. Más de 10 años
3. De la lista de factores de riesgo que se muestran a continuación, ¿Qué factores cree que se deben considerar para la detección de enfermedades cardiovasculares? Se puede seleccionar más de una opción.
    a. Edad
    b. Peso
    c. Estatura
    d. ICM
    e. Género
    f. Actividad Física
    g. Diabetes
    h. Cáncer de la piel
    i. Otro tipo de cáncer
    j. Depresión
    k. Artritis
    l. Consumo de tabaco
    m. Consumo de alcohol
    n. Consumo de frutas

o. Consumo de verduras o vegetales
p. Consumo de papas fritas

4. ¿Cuál es el nivel de influencia de los factores de riesgo para determinar la presencia de enfermedades cardiovasculares?

	Alto	Medio	Bajo
Edad			
Peso			
Estatura			
ICM			
Género			
Actividad Física			
Diabetes			
Cáncer de piel			
Otro tipo de cáncer			
Depresión			
Artritis			
Consumo de tabaco			
Consumo de alcohol			
Consumo de frutas			
Consumo de verduras o legumbres			
Consumo de papas fritas			

5. ¿Conoce sobre el término de Inteligencia Artificial?
    a. Si
    b. No
6. ¿Conoce sobre el término de Aprendizaje de Máquina o Machine Learning?
    a. Si
    b. No
7. ¿Conoce usted sobre el uso de modelos de Machine Learning para el diagnóstico temprano de enfermedades?
    a. Si
    b. No
8. ¿Consideraría usted que el uso de modelos de aprendizaje de máquina ayudaría a detectar la presencia de una enfermedad cardiovascular?
    a. Si
    b. No

9. De acuerdo a su experiencia, examine el siguiente caso y deduzca si esta persona puede o no presentar enfermedades cardiovasculares.

Edad(años)	68
Sexo	Femenino
Peso(Kg)	96
Talla(cm)	160
IMC	38
Estado de salud	regular
Ultima visita al médico	hace 1 año
Ejercicio	no
Cáncer de piel	no
Otro tipo de cáncer	no
Depresión	no
Diabetes	no
Artritis	si
Fuma	si
Consumo de alcohol (días al mes)	24
Consumo de frutas (veces al mes)	0
Consumo de verduras (veces al mes)	0
Consumo de papas fritas o similares (veces al mes)	3

   a. Si
   b. No

10. De acuerdo a su experiencia, examine el siguiente caso y deduzca si esta persona puede o no presentar enfermedades cardiovasculares.

Edad(años)	52
Sexo	Masculino
Peso(Kg)	95
Talla(cm)	168
IMC	34
Estado de salud	Muy Buena
Ultima visita al médico	hace 1 año
Ejercicio	si
Cáncer de piel	no
Otro tipo de cáncer	no
Depresión	no
Diabetes	si
Artritis	si
Fuma	no
Consumo de alcohol (días al mes)	3
Consumo de frutas (veces al mes)	60
Consumo de verduras (veces al mes)	0
Consumo de papas fritas o similares (veces al mes)	4

   a. Si
   b. No