**Objetivo General**

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

**Objetivos Específicos**

1. **Identificar y Analizar los Factores de Riesgo Principales:**
   * Investigar y seleccionar los factores de riesgo cardiovascular más relevantes que deben incluirse en el modelo predictivo, basándose en evidencia científica y disponibilidad de datos.
2. **Preprocesamiento y Análisis Exploratorio de Datos:**
   * Realizar un análisis exploratorio para entender las características del dataset, incluyendo la distribución de los factores de riesgo, la correlación entre variables y el manejo de valores faltantes o atípicos.
3. **Desarrollo y Validación de Modelos Predictivos:**
   * Aplicar y comparar múltiples algoritmos de aprendizaje automático, como la regresión logística, árboles de decisión y bosques aleatorios, para determinar el más adecuado en términos de precisión, sensibilidad y especificidad.
   * Validar los modelos utilizando técnicas como la validación cruzada y métricas como la curva ROC (Receiver Operating Characteristic, técnica gráfica para evaluar la capacidad de diagnóstico de un sistema clasificador) para asegurar la fiabilidad y generalización del modelo predictivo.
4. **Interpretación de Resultados y Extracción de Conocimientos:**
   * Interpretar los modelos en términos de la importancia de los factores de riesgo y su impacto en la predicción de enfermedades cardiovasculares.
   * Generar conocimientos que puedan ser útiles para la prevención y manejo clínico de las enfermedades cardiovasculares.

**Preguntas de Investigación**

1. ¿Cómo podemos predecir de manera fiable la probabilidad de enfermedades cardiovasculares a largo plazo, utilizando modelos de aprendizaje automático basados en factores de riesgo acumulativos?

**Borrador de introducción**

**Introducción**

Las enfermedades cardiovasculares (ECV) constituyen una de las principales causas de mortalidad a nivel global, posicionándose como una prioridad en la salud pública debido a su gran impacto en la calidad y expectativa de vida. Históricamente, desde 1969[[1]](#footnote-0), patologías como la enfermedad aterosclerótica, la enfermedad coronaria y la enfermedad cerebrovascular han figurado entre las principales causas de muerte en Estados Unidos. Aunque las tasas de mortalidad por ECV han disminuido significativamente desde la década de 1980, estas condiciones siguen siendo líderes en causar discapacidad y muerte prematura, lo que subraya la necesidad de intervenciones efectivas y tempranas.

Los avances en la comprensión de los factores de riesgo, desde genéticos hasta ambientales y de estilo de vida, han permitido mejorar el manejo médico y las estrategias de prevención. Sin embargo, la predicción precisa de ECV sigue siendo un desafío, dada la complejidad y la interacción de múltiples factores de riesgo. El Framingham Heart Study[[2]](#footnote-1) ha sido particularmente influyente en la identificación y cuantificación de estos factores de riesgo, proporcionando información clave que ha moldeado nuestra comprensión de las ECV. Este estudio pionero comenzó en 1948 con el objetivo de identificar los factores de riesgo para la enfermedad coronaria y otras enfermedades cardiovasculares. El Estudio de Framingham comenzó examinando a una muestra de 5,209 adultos de entre 30 y 62 años de edad residentes en Framingham, con el objetivo de identificar factores que pudieran predecir el desarrollo de enfermedades cardiovasculares​​.10 A lo largo de los años, el estudio ha reclutado varias cohortes adicionales, incluyendo los hijos y nietos de los participantes originales, así como muestras más diversas para reflejar mejor la población general​, generando datos longitudinales que han permitido el desarrollo de múltiples modelos predictivos y el descubrimiento de conceptos tan fundamentales como el "perfil de riesgo cardiovascular". Este estudio ha sido fundamental en la identificación de factores como la hipertensión, el colesterol alto y el tabaquismo como principales contribuyentes a la enfermedad cardiovascular. El estudio ha sido pionero en la creación de perfiles de riesgo cardiovascular que se utilizan ampliamente en la práctica clínica para evaluar el riesgo individual de enfermedades cardíacas y cerebrovasculares​

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

El estudio también ha destacado el impacto de la obesidad en el riesgo cardiovascular. A partir de la década de 1980, se reportó que el aumento de peso incrementa significativamente el riesgo de enfermedad cardiovascular, incluso después de ajustar por otros factores de riesgo. Este riesgo es particularmente evidente en el caso de la insuficiencia cardíaca, donde los participantes de Framingham menores de 50 años presentaron un riesgo de dos a tres veces mayor de insuficiencia cardíaca al comparar las categorías de peso.10

En definitiva, el Framingham Heart Study ha sido crucial para identificar numerosos factores de riesgo que afectan la salud del corazón y los vasos sanguíneos. Estos factores pueden ser clasificados en categorías biométricas, demográficas, clínicas y de estilo de vida, proporcionando una comprensión integral de cómo diferentes aspectos de la vida y la biología humana pueden influir en la probabilidad de desarrollar enfermedades del corazón.

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

**1. Factores de Riesgo Biométricos**

Los factores biométricos son mediciones biológicas que pueden indicar un mayor riesgo de enfermedades cardiovasculares. Entre los más importantes están:

• Presión arterial alta (Hipertensión): La hipertensión es uno de los factores de riesgo más críticos. Una presión arterial consistentemente alta puede dañar las arterias y el corazón, aumentando el riesgo de infarto de miocardio y accidente cerebrovascular. El estudio de Framingham ha demostrado que tanto la presión arterial sistólica como la diastólica elevadas son predictores de enfermedad cardiovascular.

• Colesterol: Los niveles altos de colesterol, particularmente el colesterol LDL (lipoproteína de baja densidad), están estrechamente relacionados con la arteriosclerosis, una condición en la cual las arterias se endurecen y se estrechan debido a la acumulación de placa. El colesterol HDL (lipoproteína de alta densidad), por otro lado, se considera protector ya que ayuda a eliminar el colesterol de las arterias.

• Índice de masa corporal (IMC): Un IMC elevado, indicador de sobrepeso y obesidad, está asociado con un mayor riesgo de hipertensión, dislipidemia, diabetes tipo 2 y enfermedades cardiovasculares. La obesidad contribuye a la inflamación crónica y al estrés oxidativo, factores que también están involucrados en la patogénesis de las enfermedades cardíacas.

• Glucosa en sangre: Los niveles elevados de glucosa en sangre o la diabetes mellitus son fuertes predictores de enfermedades cardiovasculares. La hiperglucemia crónica puede dañar los vasos sanguíneos y los nervios del corazón, incrementando el riesgo de infarto de miocardio y enfermedad arterial periférica.

**2. Factores de Riesgo Demográficos**

Los factores demográficos son características personales y antecedentes familiares que afectan la probabilidad de desarrollar enfermedades cardíacas. Algunos de los factores clave son:

• Edad: El riesgo de enfermedades cardiovasculares aumenta con la edad. Las arterias tienden a endurecerse y estrecharse con el tiempo, y el corazón puede volverse menos eficiente. La incidencia de enfermedades cardiovasculares se incrementa significativamente en personas mayores de 65 años.

• Sexo: Los hombres tienen un mayor riesgo de desarrollar enfermedades del corazón a edades más tempranas en comparación con las mujeres. Sin embargo, después de la menopausia, el riesgo para las mujeres aumenta y eventualmente puede igualar o superar el de los hombres. Esto se debe, en parte, a la disminución de estrógenos, que tienen un efecto protector en el sistema cardiovascular.

• Historial familiar: Tener antecedentes familiares de enfermedad cardiovascular es un fuerte predictor del riesgo personal. La genética juega un papel importante, y ciertas condiciones como la hipertensión, la diabetes y la hipercolesterolemia pueden ser heredadas.

• Etnicidad: Algunas etnias pueden tener un mayor riesgo debido a una combinación de factores genéticos y ambientales. Por ejemplo, los afroamericanos tienen una mayor prevalencia de hipertensión y diabetes, lo que aumenta su riesgo de enfermedad cardiovascular.

**3. Factores de Riesgo Clínicos**

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

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

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

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

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

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

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

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

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

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

• Estrés: El estrés crónico y la falta de manejo adecuado del mismo pueden aumentar el riesgo de enfermedades cardiovasculares. El estrés puede llevar a comportamientos poco saludables como el consumo excesivo de alcohol, la mala alimentación y el sedentarismo, y también puede tener efectos directos sobre el corazón y los vasos sanguíneos. 11

El Framingham Heart Study y otros estudios han sido cruciales para establecer modelos de riesgo, pero la evolución en el aprendizaje automático ofrece una nueva oportunidad para mejorar la precisión de estas predicciones.

Este proyecto se propone desarrollar y validar un modelo predictivo avanzado que utiliza algoritmos de aprendizaje automático para estimar la probabilidad de incidencia de enfermedades cardiovasculares, basándose en una integración detallada de factores de riesgo metabólicos, de estilo de vida, y ambientales. A través de un análisis exhaustivo de datos y la aplicación de modelos estadísticos complejos, el estudio busca ofrecer un enfoque más dinámico y personalizado en la predicción de riesgos cardiovasculares, superando las limitaciones de los modelos tradicionales basados en escalas estáticas y puntos de corte fijos.

Este enfoque innovador no solo tiene el potencial de mejorar la detección temprana y la personalización del manejo clínico de pacientes en riesgo, sino que también puede contribuir significativamente al desarrollo de políticas de salud pública más efectivas, centradas en la prevención de las ECV a nivel global.

[[3]](#footnote-2)

**Sección de Motivación**

En un contexto donde las enfermedades cardiovasculares (ECV) siguen siendo la principal causa de mortalidad mundial, la necesidad de avanzar en métodos predictivos precisos y accesibles es más urgente que nunca. Este proyecto está motivado por la posibilidad de integrar técnicas avanzadas de aprendizaje automático[[4]](#footnote-3) para transformar la prevención y el manejo de las ECV. La investigación actual sugiere que los modelos predictivos tradicionales, aunque útiles, a menudo no capturan la complejidad y la dinámica de los factores de riesgo en tiempo real. Al desarrollar un modelo que sí lo haga, este estudio aspira no solo a fortalecer las capacidades diagnósticas, sino también a personalizar la intervención médica, lo que podría resultar en una mejora significativa en la calidad de vida de los pacientes y una optimización de recursos en sistemas de salud pública. Este enfoque no solo refleja un compromiso académico con la innovación científica, sino que también responde a un imperativo social de desarrollar tecnología médica que sea tanto preventiva como proactiva.

La creciente prevalencia de las enfermedades cardiovasculares (ECV) y su impacto en la mortalidad global requieren urgentemente herramientas de diagnóstico más efectivas. Este proyecto está motivado por el objetivo de aplicar métodos de aprendizaje automático para mejorar la precisión de los modelos predictivos actuales. Al integrar datos multidimensionales en tiempo real, aspiramos a proporcionar una predicción más matizada y dinámica de los riesgos de ECV. Este enfoque práctico no solo tiene el potencial de mejorar la detección temprana y personalizada de enfermedades, sino también de optimizar la asignación de recursos en la atención sanitaria, haciendo un uso más eficiente de las intervenciones preventivas y terapéuticas.

El desarrollo de un modelo predictivo eficaz podría permitir intervenciones más tempranas y personalizadas, reduciendo así la incidencia y la mortalidad asociadas a las enfermedades cardiovasculares. Además, este trabajo contribuirá al avance del conocimiento científico en el campo de la epidemiología y la prevención de enfermedades, consolidando el legado del Estudio de Framingham como una fuente inagotable de datos y hallazgos relevantes.

**Metodología**

1. **Recopilación de Datos:**
   * **Fuente de Datos:** Utilizaremos un dataset consolidado que incluye información demográfica, biométrica, clínica y sobre hábitos de vida de los pacientes. Este conjunto de datos, proveniente de bases de datos públicas como el Framingham Heart Study, ha sido seleccionado por su riqueza de variables y su relevancia clínica comprobada.
   * **Selección de Variables:** Se elegirán variables basadas en su relevancia clínica y disponibilidad, con un enfoque específico en aquellas que han demostrado un fuerte vínculo con el riesgo cardiovascular en estudios anteriores.
2. **Análisis Exploratorio de Datos (EDA):**
   * **Estadísticas Descriptivas:** Calcularemos medidas de tendencia central y dispersión para cada variable, identificando distribuciones anómalas y valores atípicos que puedan influir en la precisión del análisis.
   * **Correlaciones:** Analizaremos las correlaciones entre diferentes factores de riesgo para identificar posibles interacciones y dependencias que podrían afectar los resultados del modelo.
3. **Preprocesamiento de Datos:**
   * **Limpieza de Datos:** Aplicaremos técnicas de imputación o eliminación para manejar valores faltantes, asegurando la integridad de nuestro dataset para el análisis.
   * **Transformación de Variables:** Normalizaremos o estandarizaremos las variables numéricas para mitigar cualquier sesgo debido a la escala de los datos y convertiremos variables categóricas en formatos adecuados para su análisis mediante técnicas como el one-hot encoding.
4. **Desarrollo de Modelos Predictivos:**
   * **Selección de Modelos:** Experimentaremos con varios modelos de aprendizaje automático, seleccionando aquellos que mejor se adapten a nuestro tipo de datos y objetivo de estudio, como la regresión logística, árboles de decisión y bosques aleatorios.
   * **Entrenamiento de Modelos:** Dividiremos el dataset en conjuntos de entrenamiento y prueba, utilizando técnicas como la validación cruzada para afinar los parámetros y evitar el sobreajuste.
   * **Evaluación de Modelos:** Los modelos serán evaluados utilizando métricas clave como la precisión, la sensibilidad, la especificidad y el área bajo la curva ROC, seleccionando el modelo que ofrezca el mejor rendimiento.

5. **Interpretación de Resultados**

- Importancia de las variables: análisis de la importancia de las variables para identificar los factores de riesgo más significativos y su impacto en la predicción de enfermedades cardiovasculares.

- Validación del Modelo: Validación del modelo final con datos externos (si están disponibles) para asegurar la aplicabilidad en diferentes poblaciones.

**Estado del Arte en Modelos Predictivos de Riesgo Cardiovascular**

1. **Modelos de Riesgo Tradicionales:** Los estudios tradicionales, como los referenciados en los enlaces sobre las guías de la ESC y otros análisis de riesgo cardiovascular, han establecido modelos basados en factores de riesgo bien conocidos (hipertensión, colesterol, diabetes, tabaquismo, etc.). Estos modelos son efectivos para poblaciones generales pero a menudo no capturan la variabilidad individual ni factores emergentes que pueden influir en el riesgo cardiovascular de una persona.[[5]](#footnote-4)
2. **Avances Tecnológicos en Predicción:[[6]](#footnote-5)** Con el avance de las tecnologías de big data y aprendizaje automático, estudios recientes han comenzado a explorar modelos más complejos que integran una mayor cantidad de variables, incluyendo biomarcadores genéticos y datos de estilo de vida obtenidos a través de dispositivos de seguimiento personal. Esto permite una predicción más personalizada y dinámica.[[7]](#footnote-6)
3. **Modelos Integrativos:** [[8]](#footnote-7)Algunos de los estudios proporcionados sugieren un enfoque integrativo, combinando modelos estadísticos tradicionales con técnicas de machine learning para mejorar la precisión de las predicciones. Esta integración permite explorar nuevas formas de entender la interacción entre múltiples factores de riesgo.
4. **Eficacia y Aplicabilidad Clínica:** La revisión de la literatura también destaca la importancia de la validación clínica de estos modelos. No basta con desarrollar un modelo predictivo; es crucial que su eficacia y precisión sean probadas en entornos clínicos para garantizar que los resultados sean aplicables y útiles en la práctica médica real.

Dado que el proyecto es académico y se basará en un dataset público de Kaggle, es importante mantener expectativas realistas sobre lo que se puede lograr. Vamos a enfocarnos en modelos de aprendizaje automático que sean apropiados para el alcance del proyecto y el tipo de datos disponibles.

**Regresión Logística[[9]](#footnote-8):**

* + Un modelo estadístico simple pero poderoso para problemas de clasificación binaria. Es ideal para comenzar el análisis de datos de enfermedades cardiovasculares, especialmente cuando se busca entender la relación entre factores de riesgo y la probabilidad de incidencia de la enfermedad.
  + La regresión logística es fácil de implementar, interpretar y requiere menos recursos computacionales en comparación con modelos más complejos. Además, proporciona coeficientes que pueden interpretarse directamente en términos de odds ratio para cada factor de riesgo.

1. **Árboles de Decisión y Bosques Aleatorios (Random Forests):**
   * Estos modelos son útiles para manejar datasets con un número significativo de variables y pueden manejar tanto datos numéricos como categóricos sin necesidad de transformación previa.
   * Los árboles de decisión proporcionan una representación visual fácil de entender de cómo se toman las decisiones de clasificación, lo que es útil para presentaciones académicas. Los bosques aleatorios, por otro lado, mejoran la precisión y la estabilidad del modelo a través de un enfoque de ensamblaje.
2. **Máquinas de Soporte Vectorial (SVM):**
   * SVM es efectivo para encontrar límites de decisión claros entre dos clases.
   * Aunque pueden ser computacionalmente intensivos con datasets muy grandes, en un contexto académico y con un dataset de tamaño moderado de Kaggle, SVM puede ser manejable y proporcionar resultados robustos.
3. **Validación del Modelo:**
   * Independientemente del modelo elegido, la validación cruzada debe ser utilizada para evaluar la eficacia del modelo de manera robusta.
   * La validación cruzada es un método estándar en proyectos académicos que ayuda a asegurar que los resultados del modelo son generalizables y no simplemente ajustados a un subconjunto específico de datos.

1. National Heart, Lung, and Blood Institute. (2004). The Framingham Heart Study: 50th anniversary report. Bethesda, MD: National Institutes of Health. [↑](#footnote-ref-0)
2. Dawber, T. R., Mead, F. S., & Kannel, W. B. (1951). The Framingham Study: An epidemiological investigation of cardiovascular disease. American Journal of Public Health, 41(3),

   10 Mahmood, S. S., Levy, D., Vasan, R. S., & Wang, T. J. (2014). The Framingham Heart Study and the epidemiology of cardiovascular disease: a historical perspective. *The lancet*, *383*(9921), 999-1008. [↑](#footnote-ref-1)
3. 11 Andersson, C., Nayor, M., Tsao, C. W., Levy, D., & Vasan, R. S. (2021). Framingham heart study: JACC focus seminar, 1/8. *Journal of the American College of Cardiology*, *77*(21), 2680-2692. [↑](#footnote-ref-2)
4. Li, Y., Yang, Y., Han, Y., & Song, L. (2019). Application of machine learning algorithms in predicting cardiovascular diseases: A systematic review and meta-analysis. Journal of the American Heart Association, 8(4), e011312. doi: 10.1161/JAHA.118.011312 [↑](#footnote-ref-3)
5. Aznar-Benitah, J., López-Jiménez, F., Moreno-Rius, C., & de la Torre-Hernández, J. (2021). Guía ESC 2021 sobre prevención cardiovascular primaria: ¿qué hay de nuevo?. Revista Española de Cardiología, 74(9), 737-747. [↑](#footnote-ref-4)
6. Varo, J., & Sanmartín, C. (2009). Capacidad predictiva de las funciones de riesgo cardiovascular. Revista Española de Cardiología, 62(11), 1333-1344. [↑](#footnote-ref-5)
7. Chan, J. S., & Parikh, N. I. (2019). Machine learning in cardiovascular disease prediction. Circulation: Cardiovascular Imaging, 12(11), e008080. [↑](#footnote-ref-6)
8. Un nuevo modelo de predicción del riesgo cardiovascular a 10 años. Revista Cubana de Cardiología y Cirugía Cardiovascular, 39(3). http://scielo.sld.cu/scielo.php?script=sci\_arttext&pid=S0864-21252023000300015 [↑](#footnote-ref-7)
9. Hosmer, D. W., Lemeshow, S., & Sturdivant, R. X. (2013). Applied logistic regression (3rd ed.). Wiley. [↑](#footnote-ref-8)