Universidad Tecnológica de Panamá

Facultad en Ingeniería de Sistemas Computacionales

Integrantes

Ricardo Torres / 8-1026-1230

Abdiel Ortega / 8-1032-656

Elías González / 8-1034-655

Jamal Menchaca / 3-756-1672

Rómel Rodríguez / 8-1036-932

Grupo

1IL124

Curso

Probabilidad Aplicada a TICS

Asignación

Proyecto Final

Facilitador

Juan Marcos Castillo, Phd

2025

Índice

Introducción

Las enfermedades crónicas no transmisibles, en particular las patologías cardiovasculares, se han consolidado como uno de los principales desafíos en el ámbito de la salud pública contemporánea. Su impacto se refleja no solo en las estadísticas de morbilidad y mortalidad, sino también en la calidad de vida de los pacientes y en la creciente demanda de recursos sanitarios. Uno de los mayores retos que presentan estas enfermedades es su evolución silenciosa en etapas tempranas, lo que dificulta su diagnóstico oportuno y limita las posibilidades de intervención preventiva.

En los últimos años, la integración de técnicas estadísticas y computacionales en el análisis de datos clínicos ha abierto nuevas oportunidades para la detección temprana y el estudio de los factores de riesgo asociados a las enfermedades cardiovasculares. A través del uso de bases de datos estructuradas que recopilan información relevante sobre variables fisiológicas (como presión arterial, colesterol, índice de masa corporal) y conductuales (como hábitos de consumo o tipo de personalidad), es posible modelar patrones que revelen relaciones significativas entre los datos del paciente y la probabilidad de desarrollar una condición cardiovascular.

El presente proyecto se enfoca en la exploración de un conjunto de datos clínicos con el objetivo de identificar, mediante herramientas de probabilidad y análisis estadístico, las variables más influyentes en la aparición de enfermedades cardiovasculares. A través de este enfoque, se busca no solo comprender el comportamiento de esta afección en función de los datos disponibles, sino también aportar un marco predictivo que apoye los procesos de diagnóstico precoz y facilite la toma de decisiones clínicas más informadas.

Justificación

Las enfermedades cardiovasculares continúan ocupando un lugar prioritario entre las principales causas de mortalidad a nivel mundial, lo que representa un reto no solo sanitario, sino también económico y social. Su carácter silencioso en etapas iniciales y su alta prevalencia en la población general exigen nuevas estrategias que permitan identificar oportunamente a las personas con mayor riesgo de desarrollarlas. En este contexto, el análisis de datos clínicos se convierte en una herramienta poderosa para anticiparse a los desenlaces negativos y diseñar intervenciones más eficaces.

El presente proyecto se apoya en una base de datos rica en variables fisiológicas y conductuales, como presión arterial, niveles de colesterol, consumo de tabaco y alcohol, antecedentes familiares, obesidad y otros factores clínicos relevantes. Esta información ofrece un panorama integral que permite examinar asociaciones y patrones que podrían no ser evidentes a simple vista, pero que resultan clave en la identificación de perfiles de riesgo relacionados con enfermedades del corazón.

Aplicar herramientas estadísticas y probabilísticas al estudio de estos datos no solo permite sustentar el análisis en evidencia concreta, sino que además contribuye a una comprensión más profunda del impacto individual y combinado de los distintos factores de riesgo. De esta forma, se promueve una aproximación más objetiva y científica que complementa la experiencia médica, facilitando la toma de decisiones fundamentadas en datos reales.

Además, este tipo de análisis resulta especialmente valioso en el contexto académico, ya que permite a los estudiantes desarrollar competencias en el uso de técnicas cuantitativas aplicadas a problemas reales. Asimismo, fortalece la capacidad crítica y analítica, aspectos fundamentales en la formación profesional en áreas relacionadas con la ingeniería, la salud y la ciencia de datos.

En resumen, el desarrollo de este proyecto no solo tiene el potencial de aportar conocimiento útil en materia de prevención de enfermedades cardiovasculares, sino que también representa una oportunidad formativa para aplicar métodos cuantitativos a situaciones del mundo real, promoviendo el uso responsable de la información como una herramienta al servicio de la salud y el bienestar colectivo.

Antecedentes

Las enfermedades cardiovasculares han representado, durante décadas, una de las principales preocupaciones para los sistemas de salud a nivel mundial. A pesar de los avances en medicina, tecnología y prevención, estas afecciones continúan siendo responsables de una elevada tasa de morbilidad y mortalidad en la población adulta. Este panorama ha motivado el desarrollo de múltiples investigaciones orientadas a identificar los factores que inciden en su aparición, con el propósito de promover estrategias preventivas más eficaces.

Estudios clínicos y epidemiológicos han confirmado la existencia de diversos factores de riesgo asociados a la enfermedad coronaria, tales como la hipertensión arterial, los niveles elevados de colesterol LDL, el tabaquismo, el sobrepeso, el consumo excesivo de alcohol y los antecedentes familiares. No obstante, la interacción entre estos factores puede variar de forma significativa entre individuos, lo que hace indispensable un análisis más profundo, que considere tanto su comportamiento aislado como su combinación e interdependencia.

En este contexto, el uso de bases de datos reales y estructuradas se ha convertido en una herramienta fundamental para el análisis de enfermedades crónicas. A partir del procesamiento estadístico de información clínica proveniente de múltiples pacientes, es posible detectar patrones y asociaciones que no siempre son evidentes en la práctica médica tradicional. Este enfoque ha sido potenciado gracias al desarrollo de técnicas avanzadas provenientes de disciplinas como la estadística, la informática y la ciencia de datos.

El presente proyecto se enmarca en esta línea de trabajo. Utiliza una base de datos con variables clínicas y conductuales para explorar la relación entre ciertos factores y la aparición de enfermedad coronaria. Si bien existen antecedentes en esta área, la iniciativa busca aportar una perspectiva aplicada y contextualizada, poniendo en práctica los conocimientos adquiridos en el curso y generando evidencia que contribuya tanto a la formación académica como al entendimiento de un problema de alta relevancia social.

Definición del Problema

En el estudio de enfermedades cardiovasculares, uno de los principales retos es determinar cuáles son los factores que realmente tienen más influencia en su aparición. Aunque existen muchas variables que pueden estar relacionadas, como la presión arterial, el consumo de tabaco, el colesterol, la obesidad o los antecedentes familiares, no siempre está claro cuáles son los más determinantes o cómo interactúan entre sí. Esta incertidumbre complica el diagnóstico temprano y limita las acciones de prevención.

El conjunto de datos que se analiza en este proyecto incluye diversas variables clínicas y de estilo de vida relacionadas con pacientes, entre ellas: presión arterial sistólica, consumo de tabaco, niveles de colesterol LDL, adiposidad, historial familiar de enfermedades cardíacas, nivel de estrés tipo A, obesidad, consumo de alcohol y edad. También se incluye si la persona ha desarrollado o no enfermedad coronaria. Sin embargo, el simple hecho de tener estos datos no resuelve el problema. La dificultad radica en identificar de manera clara qué variables están más asociadas con la presencia de enfermedad y cuáles no aportan tanto valor predictivo.

Por lo tanto, el problema se centra en analizar esta base de datos para detectar qué factores influyen más fuertemente en la presencia de enfermedad coronaria. Se busca establecer si hay patrones repetitivos entre las personas que presentan esta condición y si es posible determinar combinaciones de variables que permitan reconocer con mayor precisión a individuos en riesgo. Esto implica no solo observar las variables de forma aislada, sino también comprender cómo interactúan entre sí.

Este problema es relevante porque, al no tener una identificación clara de los factores más significativos, se corre el riesgo de tomar decisiones poco efectivas en temas de prevención o diagnóstico. Resolver este problema ayudaría a enfocar la atención médica en los indicadores más importantes y a desarrollar estrategias de intervención más acertadas y personalizadas.

Análisis con diferentes modelos de estocásticos

- *Determinación de la base de datos*

En una primera etapa, utilizamos una base de datos enfocada en enfermedades renales; sin embargo, tras un análisis inicial detectamos que no cumplía con los requisitos necesarios para aplicar modelos de predicción de manera eficiente. Por esta razón, optamos por cambiar de enfoque y seleccionar una base de datos sobre enfermedades cardiovasculares, la cual presentaba variables clínicas más relevantes, mejor calidad en su estructura y mayor compatibilidad con las técnicas de análisis y modelado que deseábamos aplicar.

- *Pre procesamiento y limpieza*

Durante esta etapa, realizamos la limpieza de los datos, eliminando registros incompletos y corrigiendo inconsistencias que pudieran afectar la calidad del análisis. También transformamos variables categóricas, como el tabaquismo y otras condiciones clínicas, en valores numéricos para facilitar su interpretación por los algoritmos de Machine Learning. Adicionalmente, incorporamos en el modelo una función que automatiza este proceso de transformación, permitiendo estandarizar los datos de entrada antes de realizar las predicciones y garantizando mayor consistencia en los resultados.

- *Análisis descriptivo*

Realizamos un análisis exploratorio de la base de datos utilizando estadísticas descriptivas y visualizaciones gráficas, con el objetivo de comprender la distribución de las variables y su comportamiento general. Este análisis nos permitió identificar tendencias, patrones significativos y posibles relaciones entre los factores de riesgo y la presencia de enfermedad cardiovascular, lo que resultó fundamental para orientar la selección de variables y el enfoque del modelo predictivo.

- *Selección de variables*

En lugar de seleccionar únicamente las variables con mayor relevancia, optamos por utilizar la mayoría de las variables en la base de datos para alimentar el modelo Naive Bayes. Esta decisión se basa en que Naive Bayes maneja eficientemente múltiples predictores y considera la independencia condicional entre ellos, lo que permite aprovechar la información completa para mejorar la estimación de probabilidades en la predicción de enfermedades cardiovasculares.

- *Selección de modelos*

Optamos por el modelo Naive Bayes debido a su simplicidad, facilidad de interpretación y su eficaz desempeño con datos tanto categóricos como continuos. Además, su capacidad para proporcionar estimaciones probabilísticas lo convierte en una herramienta ideal para nuestro objetivo de cuantificar el riesgo de enfermedad cardiovascular, facilitando una toma de decisiones más informada y fundamentada.

Conclusiones

El estudio permitió identificar relaciones claras entre algunas variables y el riesgo de enfermedad. Factores como el tabaco, la edad y el historial familiar destacaron por su impacto. Los resultados mostraron patrones consistentes en los datos analizados.

El análisis descriptivo facilitó la comprensión inicial de los datos. Se identificaron tendencias y valores que luego sirvieron para el análisis predictivo. Ambos enfoques se complementaron y aportaron valor al trabajo.

El proceso de selección de variables fue clave para obtener buenos resultados. Se descartaron datos irrelevantes y se priorizaron los más influyentes. Esto mejoró la precisión del modelo aplicado.

El enfoque del proyecto demostró que es posible predecir eventos de salud con datos. Los resultados permiten pensar en aplicaciones similares en otros contextos. La estructura usada fue efectiva para el objetivo planteado.

El trabajo en equipo permitió una mejor organización y distribución de tareas. Cada etapa fue desarrollada con una metodología clara y ordenada. Los objetivos del proyecto se cumplieron según lo propuesto.

Recomendaciones y futuros estudios

- **Explorar modelos predictivos avanzados:** Se recomienda probar algoritmos más potentes que Naive Bayes, como Random Forest, XGBoost, SVM o redes neuronales profundas (DNN, CNN-LSTM). Estos modelos han demostrado una mayor capacidad para capturar relaciones no lineales y mejorar la precisión en la predicción del riesgo cardiovascular.
- **Aplicar técnicas de selección e interpretación de variables:** Es fundamental emplear métodos como PCA, LASSO o SHAP para identificar las variables más relevantes y reducir la complejidad del modelo. Esto mejora no solo el rendimiento predictivo, sino también la interpretabilidad y la eficiencia computacional.
- **Manejar el desbalance de clases en los datos:** Dado que la prevalencia de enfermedad cardiovascular suele ser inferior a la de personas sanas, se recomienda utilizar técnicas como SMOTE o ADASYN para balancear las clases. Esto ayuda a evitar sesgos y mejora la capacidad del modelo para detectar casos positivos.
- **Ampliar el conjunto de variables utilizadas:** Se sugiere enriquecer la base de datos con nuevas variables clínicas, conductuales y biológicas, como marcadores genéticos, niveles de actividad física (medidos por dispositivos), factores psicológicos (estrés, depresión) y antecedentes sociales. Estos factores pueden aumentar significativamente el poder predictivo del modelo.
- **Validar el modelo en entornos clínicos reales:** Además de realizar validaciones cruzadas internas, se recomienda implementar el modelo en contextos reales de atención médica para evaluar su efectividad en la práctica. Esto puede incluir su integración en sistemas de apoyo al diagnóstico, dashboards de riesgo o herramientas digitales de monitoreo.
- **Fomentar el uso académico e interdisciplinario del modelo:** Se propone utilizar este tipo de proyectos como herramientas formativas en programas de ingeniería, ciencia de datos y salud pública. Incorporar casos clínicos reales en entornos educativos fortalece el aprendizaje práctico y promueve el desarrollo de soluciones tecnológicas con impacto social

Bibliografía

Colesterol malo (LDL). (2017). *Blood, Heart and Circulation*. https://medlineplus.gov/spanish/ldlthebadcholesterol.html

¿Cuáles son los valores normales de la presión arterial? (2021, noviembre 19). Genfar. https://www.genfar.com/te-cuidamos/cuales-son-los-valores-normales-de-la-presion-arterial/

FEC. (s/f). *¿Por qué el consumo de alcohol es perjudicial para mi corazón?* Fundación Española del Corazón. Recuperado el 29 de julio de 2025, de https://fundaciondelcorazon.com/dudas/600-ipor-que-el-alcohol-es-perjudicial-para-mi-corazon.html

La carne roja y el riesgo de enfermedad del corazón. (s/f). NIH MedlinePlus Magazine. Recuperado el 29 de julio de 2025, de https://magazine.medlineplus.gov/es/art%C3%ADculo/la-carne-roja-y-el-riesgo-de-enfermedad-del-corazon

Nyantudre, A. (2024). *Cardiovascular Diseases Dataset* [Data set].

TABACO Y CORAZÓN: ¿Cómo afecta el tabaco a nuestro corazón? (s/f). Blogs Quirónsalud. Recuperado el 29 de julio de 2025, de https://www.quironsalud.com/blogs/es/corazon/tabaco-corazon-afecta-tabaco-corazon

Anexos

Base de datos

La base de datos utilizada para realizar este estudio ha podido ser ejecutada en formato Excel. Para acceder a ella presione ctrl + Click(Con el mouse) sobre la palabra “Base de datos.”

Base de datos.

Descripción de cada columna de datos

Descripción

Análisis descriptivo

Análisis estocástico

Script