

**Universidad Tecnológica de Panamá**

Facultad de Ingeniería de Sistemas Computacionales

Maestría en Analítica de Datos

**Avance del Proyecto de Investigación**

**“Predicción de Infartos en Indonesia: Un Enfoque Estadístico y Preventivo”**

Asignatura: Modelos Predictivos - 1AN214

Profesor: Juan M. Castillo, PhD

Estudiante: Beliz Raúl Elizondo

Cedula: 8-932-1357

Fecha: 3 de abril de 2025

Centro de Postgrado

Universidad Tecnológica de Panamá

**Introducción**

En el marco de este proyecto, me propuse analizar el riesgo de infarto al miocardio a partir de datos clínicos y sociodemográficos de una población en Indonesia. La motivación principal surge del creciente impacto de las enfermedades cardiovasculares en la salud pública, especialmente en países en vías de desarrollo, donde los recursos para la prevención y el diagnóstico temprano suelen ser limitados. Considerando este contexto, decidí trabajar con una base de datos que reúne información integral sobre salud, estilo de vida y antecedentes médicos, con el objetivo de identificar patrones relevantes y entrenar modelos predictivos que puedan apoyar la toma de decisiones clínicas y de salud pública. A través de este análisis, busco no solo comprender qué factores tienen mayor peso en la ocurrencia de infartos, sino también generar conocimiento aplicable en estrategias preventivas.

**Antecedentes**

Las enfermedades cardiovasculares son la principal causa de muerte en el mundo, representando el 32% del total de fallecimientos anuales según la Organización Mundial de la Salud (WHO, 2021). En países en vías de desarrollo, como Indonesia, esta tendencia ha sido potenciada por la urbanización acelerada, los cambios en el estilo de vida, el incremento de la contaminación ambiental y la transición nutricional hacia dietas más procesadas y con alto contenido calórico (Rahajeng & Tuminah, 2019).

De acuerdo con el Ministerio de Salud de Indonesia, el infarto al miocardio se encuentra entre las cinco principales causas de hospitalización y muerte en adultos mayores de 40 años (Kementerian Kesehatan RI, 2020). Pese a la magnitud del problema, los programas de prevención siguen siendo limitados, especialmente en áreas rurales o con menor acceso a servicios médicos.

En este contexto, el uso de herramientas analíticas y predictivas se presenta como una estrategia innovadora para apoyar el diagnóstico temprano y la gestión del riesgo cardiovascular. Estudios previos han demostrado que los modelos estadísticos y de aprendizaje automático pueden identificar patrones complejos entre múltiples factores de riesgo, mejorando significativamente la precisión en la predicción de eventos cardíacos (Dey et al., 2020; Amin et al., 2019).

**Justificación**

El presente estudio se justifica en la necesidad de desarrollar herramientas que permitan anticipar y reducir el riesgo de infarto al miocardio en contextos con recursos limitados. La predicción temprana mediante modelos analíticos basados en datos clínicos, demográficos y conductuales puede facilitar intervenciones oportunas, disminuir los costos de atención médica y mejorar la calidad de vida de los pacientes (Khan et al., 2020).

La base de datos utilizada proporciona una fuente rica y diversa de información de personas en Indonesia, permitiendo no solo construir modelos predictivos, sino también analizar los factores más determinantes para el desarrollo de enfermedades cardiovasculares. Su tamaño (más de 158,000 registros) y la ausencia de valores nulos la convierten en una base sólida para un enfoque estadístico riguroso.

Desde una perspectiva personal, esta investigación responde también a una motivación íntima: la enfermedad cardiovascular ha afectado a miembros cercanos de mi familia. Este proyecto representa una oportunidad para convertir esa experiencia en aprendizaje, mediante el uso de herramientas científicas que permitan comprender mejor los factores de riesgo y aportar a una prevención más efectiva.

Definición del Problema

El infarto al miocardio constituye una de las principales amenazas para la salud pública en Indonesia, siendo una condición que puede prevenirse en gran medida si se identifican y gestionan oportunamente sus factores de riesgo. Sin embargo, en muchos casos, la detección de riesgo cardiovascular se produce de forma tardía, cuando ya se han manifestado síntomas clínicos avanzados o se ha producido el evento cardíaco. Esta situación se ve agravada por la limitada disponibilidad de herramientas predictivas accesibles para los profesionales de la salud, especialmente en zonas rurales o con menos infraestructura médica (Kementerian Kesehatan RI, 2020).

A pesar del avance de la tecnología y el crecimiento de los datos clínicos disponibles, aún existen brechas importantes en la aplicación práctica de modelos predictivos en contextos reales de salud. Muchos sistemas de predicción no están adaptados a la realidad sociodemográfica y conductual de poblaciones específicas, como la indonesia, lo que limita su precisión y aplicabilidad. Además, gran parte de los análisis previos se enfocan en modelos complejos con poca interpretabilidad clínica, dificultando su adopción por profesionales no especializados en analítica.

Frente a esta realidad, surge la necesidad de desarrollar un modelo predictivo accesible, preciso y contextualizado, que permita identificar con anticipación a los individuos con mayor riesgo de sufrir un infarto al miocardio. Este modelo debe integrar variables clínicas, demográficas y de estilo de vida, y ser construido sobre una base de datos representativa de la población indonesia. De este modo, se pretende no solo mejorar la capacidad de detección temprana, sino también aportar evidencia útil para el diseño de estrategias preventivas más eficaces en salud pública.

**Análisis Predictivo**

1. **Determinación de la Base de Datos**

La base de datos utilizada, titulada "Heart Attack Prediction Indonesia", se compone de 158,355 registros y 28 variables. Esta fuente se caracteriza por su alta amplitud y diversidad, lo cual permite abordar el riesgo de infarto al miocardio desde una perspectiva multidimensional. Las variables incluyen:

* Datos Demográficos: Edad, género, región, nivel de ingresos, entre otras.
* Variables Clínicas: Presión arterial, niveles de colesterol total, LDL, HDL, glucosa en ayunas, etc.
* Factores de Estilo de Vida y Antecedentes Médicos: Hábitos de actividad física, historial familiar, hipertensión, entre otros.

La elección de esta base se fundamenta en la representatividad de la muestra y en la calidad de los registros, ya que se constató la ausencia de valores nulos y se identificaron registros anómalos (por ejemplo, valores negativos en LDL) que fueron tratados oportunamente.

1. **Pre-procesamiento y Limpieza**

El proceso de pre-procesamiento es esencial para garantizar la calidad del análisis predictivo. En nuestro código se realizaron los siguientes pasos:

* Conversión de Variables:

Se transformaron las variables de tipo character a factor, lo que permite manejarlas adecuadamente en modelos estadísticos. Además, las variables numéricas con pocos niveles (por ejemplo, aquellas con menos de 10 valores únicos) se convirtieron en factores, ya que representan categorías discretas.

* Conversión de la Variable Objetivo:

La variable que indica si se produjo un infarto se convirtió en factor con niveles "No" y "Yes", facilitando la interpretación y el modelado.

* Verificación de Datos Faltantes y Outliers:

Se utilizó la función colSums(is.na(data)) para confirmar la integridad de los datos. Por otra parte, se generaron boxplots para identificar posibles valores atípicos y, en el caso de encontrar registros con errores (por ejemplo, valores negativos en colesterol LDL), se procedió a su tratamiento o eliminación para evitar sesgos.

* Generación de Variables Dummy y Escalado:

Las variables categóricas se transformaron a variables dummy para que todos los predictores sean numéricos, y se aplicó un escalado (centrado y escalado) a los datos, lo que es especialmente importante para modelos lineales como la regresión logística.

* División en Conjuntos de Entrenamiento y Prueba:

Se realizó una partición estratificada (80% para entrenamiento y 20% para prueba) para garantizar que la proporción de clases se mantuviera constante, lo que ayuda a evaluar de manera más confiable el desempeño de los modelos.

1. **Análisis Descriptivo**

El análisis descriptivo se ejecutó con el fin de entender la distribución y el comportamiento de las variables, así como su relación con la ocurrencia de infartos:

* Distribución de Variables Numéricas:

Se generaron histogramas para cada variable numérica, lo que permitió visualizar la distribución de cada uno de los indicadores (por ejemplo, edad, niveles de colesterol, glucosa en ayunas) y detectar posibles sesgos o asimetrías en la data.

* Boxplots para Identificar Outliers:

Los boxplots se utilizaron para identificar valores extremos en las variables numéricas. Este análisis visual es fundamental para detectar errores de registro o valores atípicos que puedan afectar el modelado.

* Análisis de Variables Categóricas:

Se creó un gráfico de barras a partir de las variables categóricas, mostrando la frecuencia de cada categoría. Esto permitió identificar, por ejemplo, la distribución de género, región y niveles de ingreso dentro de la muestra.

* Comparación de Variables Según el Estado de Infarto:

Se realizaron gráficos comparativos (boxplots) en los que se analizaron las diferencias en las distribuciones de las variables numéricas entre los grupos de pacientes con y sin infarto. Este análisis inicial ayudó a identificar variables que podrían tener un impacto significativo, como la presión arterial sistólica, la edad y los niveles de colesterol.

* Matriz de Correlación:

Se calculó la correlación de Spearman entre las variables numéricas y la variable objetivo (convertida a 0 y 1). Esto permitió cuantificar la asociación de cada predictor con el riesgo de infarto, orientando la selección de variables relevantes para el modelado.

1. **Selección de Variables**

La selección de variables se realizó combinando criterios estadísticos y visuales:

* Criterio Visual y Estadístico:

A través de los gráficos comparativos y el análisis de correlación, se identificaron las variables que presentaban mayores diferencias entre los grupos (con y sin infarto). Variables como la edad, la presión arterial sistólica, el colesterol LDL y la glucosa en ayunas mostraron relaciones significativas con la variable objetivo.

* Importancia en Modelos Basados en Árboles:

La utilización del modelo Random Forest permitió obtener medidas de importancia de variables (como el MeanDecreaseGini). Esta herramienta cuantificó la contribución de cada variable al desempeño predictivo del modelo, reforzando la selección de aquellas que consistentemente se relacionaron con el riesgo de infarto.

* Reducción de Dimensionalidad:

Se descartaron o fusionaron aquellas variables con baja varianza o con alta colinealidad, lo que contribuyó a simplificar el modelo y mejorar su interpretabilidad sin sacrificar la precisión.

1. **Selección de Modelos**

Se exploraron y compararon diversos enfoques de modelado supervisado para identificar el método que mejor se ajuste al problema:

* Regresión Logística:

Se utilizó como modelo base por su alta interpretabilidad. La regresión logística permite obtener coeficientes que indican el efecto de cada variable sobre la probabilidad de sufrir un infarto, facilitando la interpretación clínica.

* Random Forest:

Este método de ensamble aprovecha múltiples árboles de decisión para capturar relaciones no lineales y manejar interacciones complejas entre variables. Además, ofrece una medida de importancia de cada predictor, lo que resulta útil para identificar los principales factores de riesgo.

* XGBoost:

Como modelo de boosting, XGBoost combina varios modelos débiles para obtener un modelo fuerte. Este enfoque mostró un desempeño superior en términos de AUC y equilibrio entre sensibilidad y especificidad, y es especialmente útil para manejar desequilibrios en la variable objetivo.

Cada uno de estos modelos fue evaluado mediante un conjunto de prueba independiente (20% de los datos) y se compararon utilizando métricas como la matriz de confusión, la curva ROC y el AUC. Aunque las curvas ROC de los modelos resultaron muy similares, la elección final se apoyó en criterios de interpretabilidad y robustez. La regresión logística facilita la interpretación de coeficientes y la comprensión de la relación entre variables, mientras que Random Forest y XGBoost capturan de manera más efectiva las interacciones y relaciones no lineales.

En conjunto, el análisis predictivo se orienta a identificar estadísticamente cuáles son las variables más asociadas al riesgo de infarto. Los hallazgos permiten no solo comprender el impacto de cada factor (como la edad, presión arterial y niveles de colesterol) sino también sentar las bases para el diseño de posibles estrategias preventivas o cláusulas médicas que orienten la atención y monitoreo de los pacientes con mayor riesgo. Este enfoque integral es esencial para desarrollar herramientas predictivas aplicables en entornos clínicos y de salud pública, contribuyendo a la mejora en la detección temprana y en la implementación de medidas preventivas.

7. **Conclusiones**

consistente El análisis predictivo realizado ha permitido identificar de forma robusta los principales factores de riesgo asociados al infarto al miocardio en la población de Indonesia, destacándose variables como la edad, la presión arterial sistólica, el colesterol LDL, la glucosa en ayunas y el historial familiar, las cuales se manifestaron de manera a través de distintos enfoques modelísticos. La combinación de técnicas de preprocesamiento y limpieza de datos, junto con un exhaustivo análisis descriptivo y la utilización de modelos predictivos (regresión logística, Random Forest y XGBoost), ha evidenciado que, a pesar de que las curvas ROC resultaron similares en términos de desempeño global, cada modelo aportó elementos complementarios: la regresión logística ofreció interpretabilidad a través de sus coeficientes, mientras que los modelos basados en árboles capturaron interacciones complejas y proporcionaron medidas de importancia de variables que refuerzan la solidez de los hallazgos. En conjunto, estos resultados sustentan la validez de la metodología y destacan la relevancia de los factores de riesgo identificados, ofreciendo una base estadística sólida para futuras intervenciones preventivas en el ámbito clínico y de salud pública.

8. **Recomendaciones y Futuros Estudios**

Se recomienda la implementación de modelos predictivos interpretables, como la regresión logística, en entornos clínicos para la identificación temprana de pacientes de alto riesgo, complementados con protocolos de seguimiento e intervenciones focalizadas que atiendan factores críticos como la hipertensión y niveles elevados de colesterol y glucosa. Además, se sugiere la integración de estos hallazgos en políticas de salud que prioricen la prevención y el tratamiento oportuno, mientras que futuros estudios deberían ampliar la base de datos incluyendo variables adicionales (por ejemplo, indicadores genéticos y ambientales) y explorar técnicas avanzadas de interpretación de modelos (como SHAP o LIME) para profundizar en la comprensión de la influencia de cada predictor, además de validar los modelos en estudios prospectivos o de cohortes para evaluar su impacto a largo plazo.

**Anexos**

1. Base de datos

<https://github.com/belizondo98/Predicci-n-de-Infartos-en-Indonesia-Un-Enfoque-Estad-stico-y-Preventivo/blob/main/heart_attack_prediction_indonesia.csv>

1. Descripción de cada columna de datos:

<https://github.com/belizondo98/Predicci-n-de-Infartos-en-Indonesia-Un-Enfoque-Estad-stico-y-Preventivo/blob/main/Tabla%20de%20Columnas.pdf>

1. Análisis descriptivo

<https://github.com/belizondo98/Predicci-n-de-Infartos-en-Indonesia-Un-Enfoque-Estad-stico-y-Preventivo/blob/main/Analisis%20Descriptivo>

1. Análisis predictivo:

<https://github.com/belizondo98/Predicci-n-de-Infartos-en-Indonesia-Un-Enfoque-Estad-stico-y-Preventivo/blob/main/Analisis%20Predictivo>

1. Script de código o archivos de dónde se hizo el análisis

<https://github.com/belizondo98/Predicci-n-de-Infartos-en-Indonesia-Un-Enfoque-Estad-stico-y-Preventivo/blob/main/Codigo%20Final%202.R>

Bibliografía

Amin, M. S., Chiam, Y. K., & Bakar, M. S. A. (2019). A review of heart disease prediction systems using data mining and machine learning techniques. *Journal of Physics: Conference Series*, 1339(1), 012097. https://doi.org/10.1088/1742-6596/1339/1/012097

Dey, D., Slomka, P. J., Leeson, P., Comaniciu, D., & Shrestha, S. (2020). Artificial intelligence in cardiovascular imaging: JACC state-of-the-art review. *Journal of the American College of Cardiology*, 76(11), 1318–1335. https://doi.org/10.1016/j.jacc.2020.07.065

Khan, Y., Ali, S., & Shahbaz, M. (2020). A review on the applications of machine learning in heart disease diagnosis. *Health Informatics Journal*, 26(3), 1458–1480. https://doi.org/10.1177/1460458219892860

Kementerian Kesehatan Republik Indonesia. (2020). *Profil Kesehatan Indonesia Tahun 2019*. Jakarta: Pusat Data dan Informasi. https://pusdatin.kemkes.go.id/

Rahajeng, E., & Tuminah, S. (2019). Prevalence of risk factors for non-communicable diseases in Indonesia. *Health Science Journal of Indonesia*, 10(1), 19–26. https://doi.org/10.22435/hsji.v10i1.1316

WHO. (2021). *Cardiovascular diseases (CVDs)*. <https://www.who.int/news-room/fact-sheets/detail/cardiovascular-diseases-(cvds)>