**Implementación de Modelos de Clasificación para la Detección Temprana del Cáncer de Mama**

**Grupo 3**

*David Alexander Oliva Patiño, Juan José Tobón Vargas, Juan Camilo Herrón Rojas*

**Introducción.**

La gestión integral del riesgo en salud es un enfoque que permite identificar, evaluar y mitigar los riesgos asociados a eventos no deseados y potencialmente evitables que afectan negativamente la salud de los individuos y las poblaciones. Este enfoque se desarrolla a través de varias fases interrelacionadas, que incluyen la identificación, clasificación, intervención, monitoreo y evaluación. Cada una de estas etapas busca no solo entender la probabilidad de ocurrencia de eventos adversos en salud, sino también intervenir de manera oportuna para prevenirlos o minimizar su impacto.

En este contexto, la fase de clasificación adquiere un papel fundamental, ya que permite distinguir entre riesgos poblacionales y riesgos individuales. Mientras que el primero analiza las características generales de una población para determinar patrones de riesgo comunes, el segundo se enfoca en modelos predictivos más precisos, como los índices de salud, para evaluar el riesgo de manera personalizada. Estos índices son herramientas clave que permiten a las organizaciones gestionar de manera efectiva el riesgo individual, considerando desenlaces no deseados que son evitables y que representan una carga significativa tanto para la salud de las personas como para los sistemas de atención.

En Colombia, la Circular Externa 004 de 2018 define el riesgo en salud como la probabilidad de ocurrencia de eventos adversos, incluyendo la enfermedad, su evolución desfavorable o complicaciones derivadas de la misma. Bajo este marco normativo, el desarrollo de herramientas como el índice de salud permite estructurar modelos de clasificación que consideren desenlaces adversos y su prevención, aunque excluyendo elementos como la frecuencia y severidad de los servicios prestados.

En este trabajo se desarrolla un índice de salud específicamente enfocado en el cáncer de mama, una de las principales causas de mortalidad en mujeres y un problema de salud pública prioritario a nivel global y nacional. Este índice tiene como objetivo identificar de manera temprana a las mujeres con mayor riesgo de desarrollar cáncer de mama en la población afiliada a una Entidad Administradora de Planes de Beneficio (EAPB) que opera en Colombia. La identificación y clasificación de estas mujeres mediante modelos predictivos avanzados permitirá orientar intervenciones personalizadas y oportunas que contribuyan a la detección temprana y a la reducción de complicaciones, mejorando así los desenlaces en salud y optimizando el uso de los recursos del sistema.

Este trabajo se centra en el diseño, implementación y evaluación de modelos de clasificación aplicados al índice de salud para cáncer de mama, abordando las particularidades de la población afiliada a la EAPB en estudio. La investigación busca no solo contribuir al entendimiento de los factores asociados al riesgo de cáncer de mama, sino también generar herramientas que potencien la capacidad de las organizaciones para gestionar riesgos de manera efectiva, garantizando mejores resultados en la salud de las mujeres y fortaleciendo las estrategias de prevención en el sistema de salud colombiano.

**Marco Teórico**

La gestión integral del riesgo en salud es un enfoque estratégico que permite a las instituciones de salud identificar, clasificar, intervenir y monitorear los riesgos asociados a eventos adversos en la salud de individuos y poblaciones (Ministerio de Salud y Protección Social, 2018). En este contexto, los índices de salud se han convertido en herramientas clave para identificar riesgos individuales mediante la aplicación de modelos predictivos que permiten diseñar intervenciones oportunas. Estos índices se centran en desenlaces no deseados y evitables, como la aparición de enfermedades o la progresión de estas, y buscan mejorar la eficiencia en el uso de recursos y los resultados en salud (Panamerican Health Organization [PAHO], 2021).

El cáncer de mama, en particular, representa uno de los mayores desafíos de salud pública. Es la principal causa de muerte por cáncer en mujeres a nivel global y una de las más frecuentes en Colombia (Instituto Nacional de Cancerología, 2023). La detección temprana a través de herramientas como tamizajes, diagnóstico temprano y modelos predictivos basados en datos poblacionales ha demostrado ser fundamental para mejorar la supervivencia y reducir complicaciones (World Health Organization [WHO], 2021). Sin embargo, la eficacia de estas herramientas depende de su capacidad para integrar factores de riesgo, como antecedentes familiares, estilo de vida y patrones de acceso a servicios de salud, que permitan una clasificación precisa y personalizada del riesgo (Saslow et al., 2022).

En Colombia, las Entidades Administradoras de Planes de Beneficio (EAPB) tienen la responsabilidad de implementar estrategias de gestión de riesgo basadas en normativas como la Circular Externa 004 de 2018. Esta regulación establece lineamientos para la identificación y gestión de riesgos en salud, promoviendo el uso de herramientas analíticas y modelos de predicción que permitan una gestión proactiva de las poblaciones afiliadas (Ministerio de Salud y Protección Social, 2018). En este sentido, un índice de salud enfocado en cáncer de mama ofrece una oportunidad para fortalecer la detección temprana en poblaciones de riesgo, contribuyendo al cumplimiento de estas normativas y mejorando la calidad de vida de las pacientes.

El diseño de índices de salud efectivos requiere la aplicación de modelos de clasificación basados en aprendizaje automático, los cuales han mostrado un alto potencial para identificar patrones complejos en grandes conjuntos de datos. Estos modelos no solo permiten estratificar a las poblaciones de riesgo, sino también priorizar intervenciones personalizadas que maximicen los beneficios y reduzcan la carga económica y social asociada al cáncer de mama (Gomez et al., 2021).

**Desarrollo metodológico**

Para desarrollar un índice de salud que permita la detección temprana del cáncer de mama en mujeres afiliadas a una Entidad Administradora de Planes de Beneficio (EAPB) en Colombia, se aplicó la metodología CRISP-DM (Cross-Industry Standard Process for Data Mining). Esta metodología consta de seis fases interrelacionadas: comprensión del negocio, comprensión de los datos, preparación de los datos, modelado, evaluación y despliegue.

1. **Comprensión del Negocio**

El objetivo principal del proyecto es desarrollar un modelo de clasificación que identifique a las mujeres afiliadas a la EAPB con mayor probabilidad de ser diagnosticadas con cáncer de mama, facilitando así intervenciones preventivas y detección temprana. Esto responde a la necesidad de reducir la incidencia y mortalidad asociadas al cáncer de mama, en alineación con las políticas de salud pública y normativas nacionales como la Circular Externa 004 de 2018.

Objetivos Específicos:

* Desarrollar y validar modelos predictivos de clasificación utilizando técnicas de aprendizaje supervisado.
* Implementar un modelo de riesgo frente a la enfermedad de cáncer de mama que permita estratificar el riesgo individual y orientar intervenciones personalizadas.

1. **Comprensión de los Datos**

**Fuentes de Datos:**

La población objetivo son mujeres entre 18 y 75 años que han tenido al menos 2 años de afiliación a los servicios prestados por la EAPB. La EAPB suministra los datos de la población objetivo, basados en un análisis bibliográfico realizado por los médicos de la compañía.

* Registros clínicos de la EAPB: Incluyen historiales médicos, resultados de exámenes y tratamientos previos.
* Datos demográficos: Edad, raza y sexo.
* Antecedentes familiares y personales: Cáncer y condiciones ginecológicas.
* Estilos de vida: Información sobre hábitos como consumo de alcohol.

**Descripción de las Variables:**

**Variable Dependiente:**

**Diagnóstico de cáncer de mama (binaria: Sí/No):**

* **Para casos positivos (Sí):** Mujeres diagnosticadas con cáncer de mama. Las variables independientes corresponden a los datos registrados un año antes del diagnóstico confirmatorio.
* **Para casos negativos (No):** Mujeres sin diagnóstico de cáncer de mama hasta la fecha de corte del 1 de enero de 2024. Las variables independientes corresponden al último registro disponible antes de esta fecha.

**Variables Independientes:**

* Sexo: Categórica; se espera que todos los registros correspondan al sexo femenino.
* Edad: Numérica continua; edad de la paciente en años al momento del registro correspondiente.
* Raza: Categórica; categorías como 'Blanca', 'Afrodescendiente', 'Indígena', 'Mestiza'.
* Índice de Masa Corporal (IMC): Numérica continua; kg/m².
* Indicador de consumo de alcohol: Binaria; '1' para consumo, '0' para no consumo.
* Edad de la menarca: Numérica discreta; edad en años del primer período menstrual.
* Edad de la menopausia: Numérica discreta; aplicable solo a mujeres posmenopáusicas.
* Indicador de terapia hormonal: Binaria; '1' para sí, '0' para no.
* Resultado de mamografía (BI-RADS): Categórica ordinal; categorías de '0' a '6'.
* Indicador de ooforectomía bilateral: Binaria; '1' para sí, '0' para no.
* Número de familiares de primer grado con cualquier cáncer: Numérica discreta.
* Número de familiares de segundo grado con cualquier cáncer: Numérica discreta.
* Número de familiares de primer grado con cáncer de mama: Numérica discreta.
* Número de familiares de segundo grado con cáncer de mama: Numérica discreta.
* Indicador de antecedentes familiares con cualquier cáncer: Binaria; '1' si existe al menos un familiar afectado, '0' en caso contrario.
* Indicador de antecedentes familiares con cáncer de mama: Binaria; '1' si existe al menos un familiar afectado, '0' en caso contrario.
* Indicador de radiografía de tórax: Binaria; '1' para sí, '0' para no.

Consideraciones Temporales:

Para casos positivos: Las variables independientes se extraen de registros correspondientes a un año antes del diagnóstico de cáncer de mama para evitar la inclusión de información que pueda ser consecuencia del diagnóstico (evitando así la fuga de información).

Para casos negativos: Se utiliza el último registro disponible antes de la fecha de corte (1 de enero de 2024), garantizando que los datos reflejen el estado más reciente de las pacientes sin diagnóstico.

**Calidad y Volumen de los Datos:**

**Tamaño del conjunto de datos:** Información de 2,190,279 mujeres afiliadas.

**Referencias bibliográficas**

Gómez, A., Pérez, L., & Rodríguez, M. (2021). Machine learning in breast cancer risk prediction: An overview. Journal of Health Analytics, 12(3), 45-56.

Instituto Nacional de Cancerología. (2023). Estadísticas de cáncer en Colombia 2023. Bogotá: INC.

Ministerio de Salud y Protección Social. (2018). Circular Externa 004: Gestión integral del riesgo en salud. Bogotá: MinSalud.

Panamerican Health Organization [PAHO]. (2021). Comprehensive risk management in public health systems. Washington, D.C.: PAHO.

Saslow, D., et al. (2022). Breast cancer early detection: Evidence-based guidelines. Cancer Journal for Clinicians, 72(5), 112-125.

World Health Organization [WHO]. (2021). Breast cancer early diagnosis and control. Geneva: WHO.