**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*

1. **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.

1. **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).

1. **Metodología**

**Agregar CRISP DM (mirar ejemplo Trabajo de grado Camilo Castro)**

1. **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 y no 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**
     1. **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.
  + 1. **Descripción de las Variables:**

**Variable Dependiente:**

**Ind\_CAM:** 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:**

* **Afiliado\_Id:** Numérica; número de identificación del afiliado en la compañía.
* **Ind\_Frecuencia\_Licor:** Indicador de consumo de alcohol. Binaria; 'Si' para consumo, 'No' para no consumo.
* **Sexo\_Cd:** Sexo. Categórica; se espera que todos los registros correspondan al sexo femenino.
* **Raza\_Desc**: Raza. Categórica; categorías como 'Blanca', 'Afrodescendiente', 'Indígena', 'Mestiza'.
* **Valor\_IMC:** Índice de Masa Corporal (IMC). Numérica continua; kg/m².
* **Num\_Edad\_Menopausia:** Edad de la menopausia. Numérica discreta; aplicable solo a mujeres posmenopáusicas.
* **Num\_Edad\_Menarca:** Edad de la menarca. Numérica discreta; edad en años del primer período menstrual.
* **Ind\_Terapia\_Hormonal**: Indicador de terapia hormonal. Binaria; 'Si' para sí, 'No' para no.
* **Num\_Birads:** Resultado de mamografía (BI-RADS). Categórica ordinal; categorías de '0' a '6'.
* **Ind\_Ooforectomia\_Bilateral:** Indicador de ooforectomía bilateral. Binaria; 'Si' para sí, 'No' para no.
* **Num\_Fam\_Primer\_Grado\_Otros:** Número de familiares de primer grado con cualquier cáncer. Numérica discreta.
* **Num\_Fam\_Segundo\_Grado\_Otros:** Número de familiares de segundo grado con cualquier cáncer. Numérica discreta.
* **Ind\_Ant\_Fam\_Otros**: Indicador de antecedentes familiares con cualquier cáncer. Binaria; 'Si' si existe al menos un familiar afectado, '0' en caso contrario.
* **Num\_Fam\_Primer\_Grado\_CAM:** Número de familiares de primer grado con cáncer de mama**.** Numérica discreta.
* **Num\_Fam\_Segundo\_Grado\_CAM:** Número de familiares de segundo grado con cáncer de mama**.** Numérica discreta.
* **Ind\_Ant\_Fam\_CAM:** Indicador de antecedentes familiares con cáncer de mama**.** Binaria; 'Si' si existe al menos un familiar afectado, 'No' en caso contrario.
* **Ind\_Ant\_Radio\_Torax:** Indicador de radiografía de tórax. Binaria; 'Si' para sí, 'No' para no.
* **Edad:** Numérica continua; edad de la paciente en años al momento del registro correspondiente.

**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.

* + 1. **Calidad y Volumen de los Datos**

**Tamaño del conjunto de datos:** Información de 2.190.279 mujeres afiliadas.

* **Casos positivos:** 18.253 registros de mujeres afiliadas con diagnostico positivo de cáncer de mama.
* **Casos negativos:** 2.172.026 registros de mujeres afiliadas sin diagnostico de cáncer de mama.

**Registros duplicados:** Se detectaron 66 registros duplicados que serán eliminados para garantizar la integridad de los datos.

**Datos faltantes:** Se identificaron variables con datos faltantes que requieren estrategias de imputación o exclusión.

* Ind\_Frecuencia\_Licor: 33,81%
* Valor\_IMC: 11,06%
* Num\_Edad\_Menopausia: 98,58%
* Num\_Edad\_Menarca: 59,43%
* Num\_Birads: 77,87%

Se puede apreciar que 98,58% de los registros de la variable edad de menopausia se encuentra vacío. Sin embargo, dado que el 75% de las mujeres en el dataset tienen una edad igual o inferior a los 55 años, el registro es coherente ya que la menopausia es un proceso natural que suele ocurrir entre los 45 y 55 años, aunque la edad promedio es de 51 años.

**Conformidad:** La conformidad se refiere a si los datos cumplen con formatos, tipos y valores esperados.

Al hacer un análisis de los tipos de datos, se encontró que en general todos los datos se encontraban bien. Dado que todas las variables numéricas eran de tipo float64, se decide cambiarlas por integer, exceptuando el Valor del IMC que se conserva como float64.

Por otro lado, al realizar una validación de valores permitidos en las variables categóricas, nos encontramos con valores en la variable Num\_birads que no cumplen con las categorías (0-6).

Dichas categorías encontradas fueron: 'BIRA', 'BI-R', 'CATE', 'BI R', 'BI -', 'CLAS' las cuales no corresponden a las categorías estándar de BI-RADS. Es posible que sean:

* **Errores de tipeo:** Simples errores al ingresar los datos.
* **Abreviaturas no convencionales:** Abreviaturas utilizadas en un contexto específico que no son ampliamente reconocidas.
* **Categorías personalizadas:** Categorías creadas para un estudio particular, que no siguen el sistema BI-RADS estándar.

Por lo tanto, se decide reemplazar estos valores por vacíos (NaN) para facilitar un análisis más preciso y consistente.

Igualmente, la variable en la variable raza se decide reemplazar los valores de 'SIN INFORMACION DESDE LA FUENTE' por vacíos.

Seguido de esto, se analiza que las edades se encuentren bien. Es decir, que la edad de la menarca sea para todo el caso menor que la edad de la menopausia; que la edad sea mayor o igual a la edad de menarca y que la edad sea mayor o igual a la edad de la menopausia.

En este último, nos encontramos que 3 registros tienen la edad de la menopausia mayor a la edad, por lo que se decide eliminar dichos registros.

**Precisión:** La precisión evalúa si los datos son correctos y representan fielmente la realidad.

Para el análisis de la precisión se analizo la distribución de la edad, se buscó la existencia de la edad de menarca improbable (menor a 8 años o mayor a 18 años) y edad de menopausia improbable (menor a 35 años). En este caso, no encontraron registros imprecisos.

**Consistencia:** La consistencia garantiza que los datos son uniformes y coherentes en todo el conjunto.

Se realizo una verificación de la consistencia entre el Número de Familiares y los Indicadores de Antecedentes Familiares. Lo que se busca es verificar que en los casos donde el número de familiares es diferente de 0 o None (es decir, hay al menos un familiar afectado), el indicador correspondiente sea **Si**, y en caso contrario, sea **No**.

En este caso, nos encontramos que la variable Ind\_Ant\_Fam\_Otros tenía discrepancia. Dado que todos los valores donde hay discrepancia con la variable original son "No", se decide eliminar la variable Ind\_Ant\_Fam\_Otros y reemplazarla por la variable nueva llamada Ind\_Ant\_Fam\_Otros\_Esperado.

La misma revisión se hizo para la variable Ind\_Ant\_Fam\_CAM, en este caso no se encontraron discrepancias.

**Análisis Exploratorio de Datos**

**Análisis univariado**

Para el análisis univariado se dividen las variables según su tipo (numéricas o categóricas).

Para las variables numéricas, se analiza realiza una tabla con estadísticos generales como cantidad de registros, media, mediana, cuartiles y desviación estándar. Adicional, se realiza un histograma agregando la media y la mediana de los datos, y un gráfico de caja y bigotes para detectar outliers.

En este análisis, nos encontramos con que la variable Valor\_IMC cuenta con más de 46.000 valores atípicos. Valores que corresponden a un IMC superior a un valor de 40, alcanzo un máximo de 89,97. Se propone en la etapa de preparación imputar los valores atípicos con el valor de la mediana.

Asimismo, la variable edad tiene más de 2.900 valores atípicos, registrando un valor máximo de 139 años. Sin embargo, dado que el objetivo es modelar solo para mujeres hasta los 75 años, se eliminará en la etapa de preparación los registros con una edad mayor a 75 años.

Por otro lado, para las variables categóricas, se realiza una tabla descriptiva con estadísticos como cantidad de registros, valores únicos, moda y frecuencia relativa de cada una de las categorías.

**Análisis bivariado**

Para el análisis bivariado, se realiza inicialmente un análisis de cada una de las variables según su tipo (numéricas o categóricas) frente a la variable dependiente.

Para las variables numéricas, se analiza las estadísticas descriptivas del análisis univariado pero esta vez agrupando por la categoría de la variable dependiente. Igualmente, se realiza un gráfico de caja y bigotes e histograma para cada categoría. Por ultimo, se realiza una prueba de hipótesis para evaluar diferencias entre las categorías.

En este caso, encontramos que no hay diferencia significativs entre las categorías versus el valor del IMC. En el resto de casos si se encuentran diferencias significativas entre las categorías. Es posible que el comportamiento que presenta el valor del IMC esté siendo afectado por los valores atípicos, se revisará este supuesto más adelante cuando se haga la corrección de los registros.

Para las variables categóricas, se realiza una tabla de frecuencias cruzadas y se realiza un gráfico de barras apiladas.

En este caso, nos encontramos con que variables como la raza blanca tiene un porcentaje por categoría mayor (2,05%) para ‘Si’ en el indicador de cáncer de mama, siendo esto coherente con la literatura ya que este es un factor de riesgo estudiado.

Esto también se presenta con el indicador de terapia hormonal, donde el porcentaje se evidencia un mayor porcentaje de mujeres con el indicador de cáncer de mama que cuentan con la marca de terapia hormonal (2.23%) frente a un (0,83%) que tienen el indicador de cáncer de mama y no tienen la marca de terapia hormonal.

Adicional, para complementar el análisis bivariado, se realizó un análisis de cada una de las variables independientes frente a las otras.

En este análisis, se utilizó un heatmap de la matriz de correlación para explorar las relaciones lineales entre variables clínicas, demográficas y familiares en un conjunto de datos sobre cáncer de mama. Se observó que la mayoría de las variables presentan correlaciones bajas o nulas, lo que indica una baja dependencia estadística entre ellas, salvo algunas excepciones, como la correlación moderada-alta entre los antecedentes familiares de primer y segundo grado (0.69) y entre la edad y la menopausia. Estos hallazgos sugieren que las variables, en su mayoría, aportan información única, lo cual resulta útil para evitar problemas de multicolinealidad en análisis posteriores.

Se evidencia que hay relación lineal fuerte entre las variables de número de antecedentes familiares con las variables de indicadore de antecedentes familiares. Por lo tanto, la mejor opción es quedarse solo con la variable indicador de antecedentes familiares.

Se puede hacer una descripción del análisis univariado y bivariado.

Con respecto a los outliers. Para la variable IMC se decide imputar los valores faltantes y los valores atípicos con la mediana.

Para la variable edad, dado que la idea es modelar mujeres hasta los 75 años, se eliminan los registros donde la edad sea mayor a este valor.

Para la variable Edad menopausia: Imputar el valor promedio a aquellas mujeres mayores que no tengan registro. Mirar si se hace un promedio para las que ind cam es si y no

Para la edad de menarca, colocar el valor promedio.

Variables categóricas:

Imputar los valores nulos con la moda

* Crear una nueva variable: Esta variable se llama Expo\_hormonal: la idea es calcular los años en que la mujer tuvo exposición hormonal. El calculo sería edad – edad menarca (cuando la edad sea menor a 50 años y no tenga dato en edad menopausia), edad menopausia – edad menarca (si tiene las dos). 50 – edad menarca (Si tiene mas de 50 años y no tiene dato de edad menopausia).

Lo que sigue es hacer el análisis bivariado. Mirar como las variables dependientes están distribuidas versus la variable dependiente. Seria repetir el análisis univariado pero separando los datos entre Ind\_Cam Si y No.

Después, hacer un análisis multivariado.

Básicamente matriz de correlación, PCA,

1. **Preparación de los datos**

* Hacer un Análisis de las componentes principales (PCA) para modelar con un base de datos de menor dimensión y optimizar los tiempos de procesamiento.

La fase de **Preparación de los Datos** es una etapa crítica dentro de la metodología CRISP-DM, ya que implica transformar los datos brutos en un conjunto de datos limpio y estructurado, adecuado para el análisis y modelado. Esta etapa garantiza que los datos utilizados sean de alta calidad, relevantes y estén en el formato correcto para construir modelos predictivos confiables y precisos.

**Selección de Datos**

En primer lugar, se seleccionan las variables que se consideran pueden ser relevantes para el análisis. En este caso, se elimina la variable Sexo y afiliado\_id inicialmente.

**Limpieza de Datos**

Acto seguido, se procede a realizar una limpieza de las variables inicialmente seleccionadas para llevar a modelar.

Para esto, lo primero que se hace es dar tratamiento a los datos faltantes.

Con respecto a los datos faltantes, se realiza el siguiente tratamiento:

- Ind\_Frecuencia\_Licor: Se imputa con la moda.

- Raza: Dado que según los estudios la raza con mayor riesgo es raza blanca, se decide categorizar dicha variable donde se deje 1 cuando sea raza blanca, 0 de lo contrario.

- Valor\_IMC: Se decide imputar el valor con la mediana.

- Num\_Edad\_Menopausia: Se decide crear una variable que la reemplace por una variable categórica llamada Ind\_Menopausia donde sea 1 cuando la edad sea mayor a 50 años, 0 en caso contrario.

- Num\_Edad\_Menarca: Se decide reemplazar por el valor de la mediana.

- Num\_Birads: Se decide eliminar la variable dado que no hay una forma de imputarla con un criterio valido.

Con respecto a los outliers encontrados en las variables Edad y Valor\_IMC, se toman las siguientes decisiones:

- Valor\_IMC: Reemplazar los outliers con la mediana.

- Para edad: Dado que el objetivo de estudio es para mujeres hasta los 75 años, se eliminan los registros con una edad superior.

**Construcción de las variables**

La creación de la variable Tiempo\_Exposicion\_Hormonal es esencial para tu modelo de clasificación, ya que representa la duración de la exposición hormonal de cada mujer, un factor potencialmente significativo en el riesgo de desarrollar cáncer de mama.

Criterios para Calcular Tiempo\_Exposicion\_Hormonal

1. Caso 1:
   * Condición: Edad < 50 años y no tiene dato en Num\_Edad\_Menopausia.
   * Cálculo: Tiempo\_Exposicion\_Hormonal = Edad - Num\_Edad\_Menarca.
2. Caso 2:
   * Condición: Tiene datos tanto en Num\_Edad\_Menopausia como en Num\_Edad\_Menarca.
   * Cálculo: Tiempo\_Exposicion\_Hormonal = Num\_Edad\_Menopausia - Num\_Edad\_Menarca.
3. Caso 3:
   * Condición: Edad > 50 años y no tiene dato en Num\_Edad\_Menopausia.
   * Cálculo: Tiempo\_Exposicion\_Hormonal = 50 - Num\_Edad\_Menarca.

Teniendo esta nueva variable, se procede a eliminar la variable Num\_Edad\_Menopausia

También, se decide crear la variable Categoria\_IMC la cual cumple las siguientes condiciones:

- Bajo peso: IMC < 18.5

- Normal: 18.5 ≤ IMC < 25

- Sobrepeso: 25 ≤ IMC < 30

- Obesidad: IMC ≥ 30

**Estandarizar variables númericas.**

Ahora, lo que se busca es estandarizar las variables númericas. Estas son 'Valor\_IMC', 'Edad', 'Tiempo\_Exposicion\_Hormonal'. Para esto, se utiliza la técnica minMax.

**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.