**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 clasificar mujeres que tienen mayor probabilidad de ser diagnosticadas con 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**

Para el desarrollo del presente trabajo, se empleó la metodología CRISP-DM (Cross Industry Standard Process for Data Mining), la cual fue creada en 1999 por la entonces Comunidad Económica Europea, hoy la Unión Europea. Adicionalmente, es la metodología de referencia más utilizada en el desarrollo de proyectos de minería de datos. Entre sus ventajas se destaca la posibilidad de replicación de proyectos, su independencia en la industria, su neutralidad con respecto a las herramientas y su enfoque en las situaciones de negocio (García, G. 2018).

Las etapas de CRISP-DM aplicadas a este proyecto son: Comprensión del negocio, Comprensión de los datos, Preparación de los datos, Modelado, Evaluación e Implementación.

* 1. **Comprensión del negocio**

En esta fase, se introdujo el problema de la detección temprana del cáncer de mama y el método que se empleará en el trabajo. Se explicó la importancia de reducir la incidencia y mortalidad asociadas al cáncer de mama, alineándose con las políticas de salud pública y las normativas nacionales como la Circular Externa 004 de 2018. Se describió la situación actual de la detección del cáncer de mama en Colombia y los desafíos que enfrentan las Entidades Administradoras de Planes de Beneficio (EAPB) en la gestión del riesgo en salud. Adicional, se establecieron los objetivos a cumplir en el trabajo.

* 1. **Comprensión de los datos**

En primer lugar, se realizó un análisis de la calidad y volumen de los datos. En esta etapa se hizo una revisión de duplicidad, completitud, conformidad, precisión y consistencia de los datos entregados por la EAPB.

Se realizó un análisis exploratorio de los datos el cual se dividió en análisis univariado y bivariado. En la etapa de análisis bivariado se hizo entre la variable dependiente contra las variables independientes y en último lugar un análisis bivariado entre las diferentes variables independientes. La población objetivo consistió en mujeres hasta los 75 años con al menos 2 años de afiliación a la EAPB.

* 1. **Preparación de los datos**

En esta etapa, se realizo la selección de las variables y la limpieza y procesamiento de estas. Se limpiaron los datos manejando los valores faltantes y valores atípicos mediante imputación adecuada. Las variables categóricas fueron codificadas y las numéricas normalizadas. Se crearon nuevas variables a partir de las variables existentes. Se aplicó Análisis de Componentes Principales (PCA) para reducir el tamaño de la base de datos, conservando al menos el 90% de la información original.

* 1. **Modelado**

Se implementaron dos algoritmos de clasificación: K-Nearest Neighbors (KNN) y Regresión Logística. Los modelos fueron entrenados utilizando un 70% de los datos y evaluados con un 30% restante. Se utilizó validación cruzada k-fold para asegurar la robustez de los modelos.

* 1. **Evaluación**

El rendimiento de los modelos fue evaluado utilizando métricas como exactitud, sensibilidad, especificidad, precisión, F1-Score, y área bajo la curva ROC (AUC-ROC). Además, se empleó la curva Lift para analizar la efectividad de los modelos en la clasificación de riesgo. Las matrices de confusión y las curvas ROC y Lift permitieron seleccionar el modelo que mejor balanceó las métricas de desempeño.

* 1. **Implementación**

El modelo seleccionado fue integrado en los sistemas de información de la EAPB, permitiendo la clasificación automática del riesgo de cáncer de mama al ingresar nuevos datos de pacientes. Se desarrolló una interfaz de usuario para que los profesionales de salud puedan visualizar el nivel de riesgo y los factores contribuyentes.

1. **Desarrollo metodológico**
   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.

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**

El conjunto de datos utilizado en este estudio comprende información de 2.190.279 mujeres afiliadas a la Entidad Administradora de Planes de Beneficio (EAPB). Dentro de este conjunto, se identificaron 18.253 registros de mujeres con diagnóstico positivo de cáncer de mama (casos positivos) y 2.172.026 registros de mujeres sin diagnóstico de cáncer de mama (casos negativos).

Se detectaron 66 registros duplicados, los cuales fueron eliminados para garantizar la integridad de los datos.

Además, 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%

Es notable que el 98,58% de los registros de la variable Edad de Menopausia se encuentran vacíos. Sin embargo, esto es coherente ya que el 75% de las mujeres en el dataset tienen una edad igual o inferior a los 55 años, edad en la cual la menopausia es un proceso natural que generalmente ocurre entre los 45 y 55 años, con una edad promedio de 51 años.

Para la conformidad, se verificó que los datos cumplan con los formatos, tipos y valores esperados. Se observó que todas las variables numéricas eran de tipo float64, por lo que se decidió convertirlas a integer, excepto la variable IMC, que se mantiene como float64.

Al revisar los valores permitidos en las variables categóricas, se encontraron valores no válidos en la variable Num\_Birads: 'BIRA', 'BI-R', 'CATE', 'BI R', 'BI -', 'CLAS'. Estos valores fueron reemplazados por NaN para facilitar un análisis más preciso y consistente. Asimismo, en la variable Raza se reemplazaron los valores de 'SIN INFORMACION DESDE LA FUENTE' por vacíos.

Seguido de esto, para evaluar la precisión se analizó la distribución de la edad y se verificó la existencia de edades de menarca y menopausia improbables (menarca menor a 8 años o mayor a 18 años, menopausia menor a 35 años). No se encontraron registros imprecisos.

También se evaluó la consistencia, donde se verificó la coherencia entre el número de familiares afectados y los indicadores de antecedentes familiares. Se detectaron discrepancias en la variable Ind\_Ant\_Fam\_Otros, donde se observó que todos los registros con discrepancias eran "No". Por lo tanto, se eliminó esta variable y se reemplazó por una nueva variable denominada Ind\_Ant\_Fam\_Otros\_Esperado. No se encontraron discrepancias en la variable Ind\_Ant\_Fam\_CAM.

Adicionalmente, se verificó que la Edad de Menarca sea menor que la Edad de Menopausia, y que la Edad sea mayor o igual a ambas. Se encontraron 3 registros con la edad de menopausia mayor a la edad actual, los cuales fueron eliminados para mantener la consistencia de los datos.

Esta revisión de calidad de los datos asegura que el conjunto de datos esté limpio y preparado para las fases subsecuentes de modelado y análisis, garantizando la fiabilidad y precisión de los resultados obtenidos en el proyecto.

* + 1. **Análisis exploratorio de los datos**

El análisis exploratorio de datos (AED) es fundamental para comprender las características y relaciones dentro del conjunto de datos, facilitando la preparación y modelado posterior. Este análisis se divide en dos partes: análisis univariado y análisis bivariado.

* + - 1. **Análisis univariado**

En el análisis univariado, las variables se clasificaron según su tipo (numéricas o categóricas) para examinar sus distribuciones individuales.

Para las variables numéricas, se generaron tablas con estadísticos descriptivos como cantidad de registros, media, mediana, cuartiles y desviación estándar. Además, se crearon histogramas que incluyen la media y la mediana, así como gráficos de caja y bigotes para detectar valores atípicos. Se identificaron más de 46,000 valores atípicos en la variable Valor\_IMC, con valores superiores a 40 y un máximo de 89.97. Se propone imputar estos valores atípicos con la mediana en la etapa de preparación.

Asimismo, la variable Edad presentó más de 2,900 valores atípicos, alcanzando un máximo de 139 años. Dado que el análisis se enfoca en mujeres hasta 75 años, se eliminarán los registros con edades superiores a este umbral.

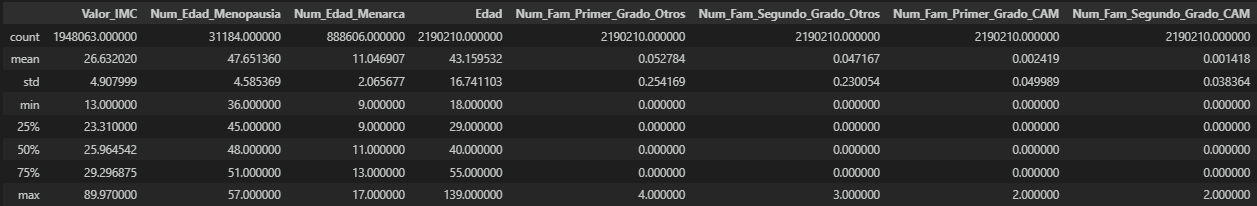


Ilustración 1Descripción estadística variables numéricas

Para las variables categóricas, se elaboraron tablas descriptivas que incluyen cantidad de registros, valores únicos, moda y frecuencia relativa de cada categoría. Este enfoque permitió identificar la distribución y prevalencia de las diferentes categorías dentro de las variables estudiadas.

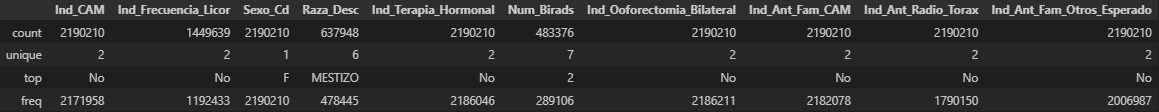


Ilustración 2Descripción estadística variables categóricas

* + - 1. **Análisis bivariado**

El análisis bivariado exploró las relaciones entre cada variable independiente y la variable dependiente

Para las variables numéricas, se realizaron estadísticas descriptivas agrupadas por la categoría de la variable dependiente, complementadas con gráficos de caja y bigotes e histogramas para cada grupo. Además, se efectuaron pruebas de hipótesis para evaluar diferencias significativas entre las categorías. Se observó que no existe una diferencia significativa entre las categorías en relación con el Valor\_IMC, posiblemente influenciada por los valores atípicos previamente identificados. Sin embargo, para otras variables numéricas, se encontraron diferencias significativas.

Gráfico

Descripción generada automáticamente

Ilustración 3Gráfico Valor IMC vs. Indicador cáncer de mama

Gráfico, Gráfico de cajas y bigotes

Descripción generada automáticamente

Ilustración 4Gráfico edad vs. Indicador cáncer de mama

Con las variables categóricas, Se construyeron tablas de frecuencias cruzadas y gráficos de barras apiladas para visualizar la relación entre las variables categóricas y el diagnóstico de cáncer de mama. Por ejemplo, se encontró que la raza blanca presenta un mayor porcentaje de diagnósticos positivos (2.05%), lo cual es consistente con la literatura existente sobre factores de riesgo.

Texto

Descripción generada automáticamente

Ilustración 5Porcentaje por categoría raza vs Indicador cáncer de mama

Asimismo, se observó que las mujeres que han recibido terapia hormonal tienen un mayor porcentaje de diagnósticos positivos de cáncer de mama (2.23%) en comparación con aquellas que no la han recibido (0.83%).

Texto

Descripción generada automáticamente

Ilustración 6Porcentaje por categoría Indicador terapia hormonal vs. Indicador cáncer de mama

Adicionalmente, se realizó un análisis de correlación entre las variables independientes utilizando un heatmap de la matriz de correlación. La mayoría de las variables presentaron correlaciones bajas o nulas, indicando una baja dependencia estadística entre ellas. Excepciones notables incluyen una correlación moderada-alta (0.69) entre el número de antecedentes familiares de primer y segundo grado con cualquier tipo de cáncer y el indicador de antecedentes familiares de cáncer, así como una correlación de 0.23 entre la edad y la menopausia. Estos hallazgos sugieren que, en su mayoría, las variables aportan información única, lo cual es beneficioso para evitar problemas de multicolinealidad en las fases de modelado posteriores.

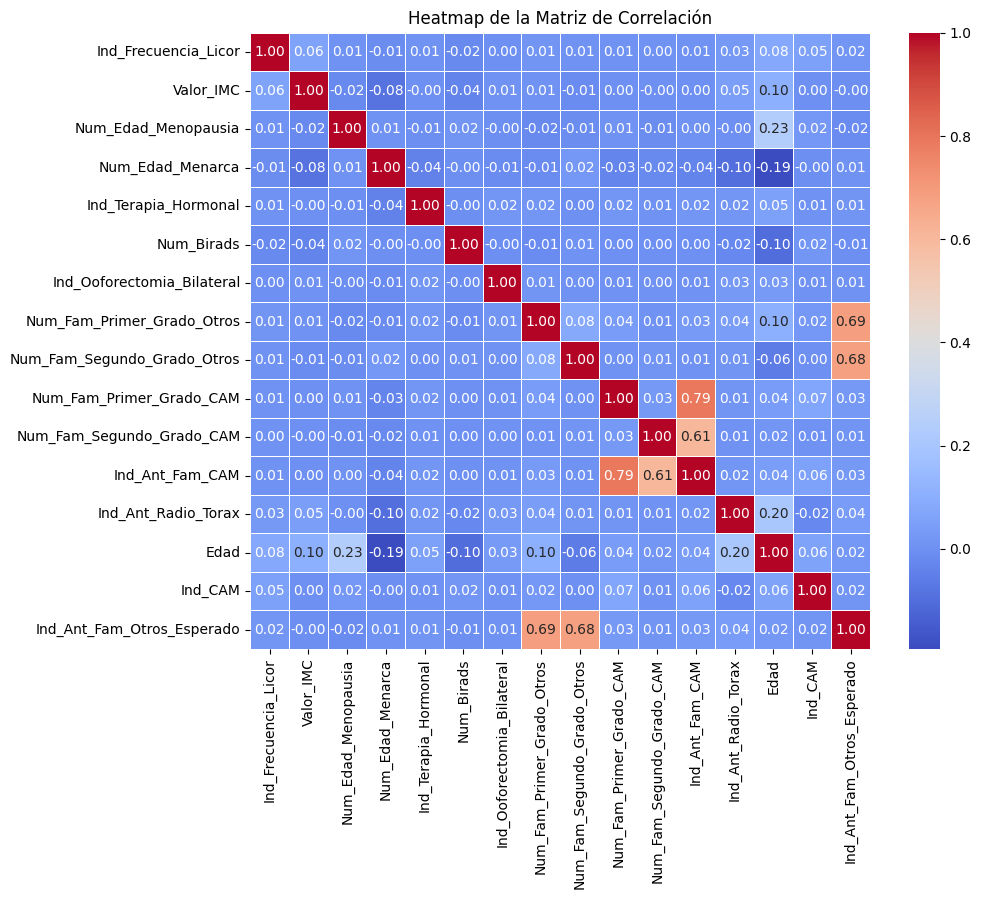


Ilustración 7Matriz de correlación variables independientes

Este análisis exploratorio de datos proporciona una comprensión profunda de las características individuales y las relaciones entre variables dentro del conjunto de datos, estableciendo una base sólida para las etapas de preparación y modelado en el desarrollo del índice de salud para la detección temprana del cáncer de mama.

* 1. **Preparación de los datos**
  2. **Modelado**
  3. **Evaluación**

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).

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.

**Formateo de datos.**

Lo que sigue, es codificar las variables categóricas. Para esto, se utiliza one hot encoding exceptuando a categoría imc para la cual se implementa ordinal encoder.

**Variables finales para modelar:**

Después del proceso de limpieza de datos, las variables seleccionadas son: 'Valor\_IMC', 'Num\_Edad\_Menarca', 'Num\_Fam\_Primer\_Grado\_Otros',

'Num\_Fam\_Segundo\_Grado\_Otros', 'Num\_Fam\_Primer\_Grado\_CAM',

'Num\_Fam\_Segundo\_Grado\_CAM', 'Edad', 'Ind\_CAM', 'Ind\_Raza\_Blanca',

'Ind\_Menopausia', 'Tiempo\_Exposicion\_Hormonal',

'Ind\_Frecuencia\_Licor\_1', 'Ind\_Terapia\_Hormonal\_1',

'Ind\_Ooforectomia\_Bilateral\_1', 'Ind\_Ant\_Fam\_CAM\_1',

'Ind\_Ant\_Radio\_Torax\_1', 'Ind\_Ant\_Fam\_Otros\_Esperado\_1',

'Categoria\_IMC\_Encoded'

**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.

* García, G. (2018). *Metodologías de minería de datos: CRISP-DM y su aplicación en proyectos empresariales*. Revista de Ingeniería y Tecnología, 15(2), 34-45.
* 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.