

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

2. 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).

3. 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.

3.1. 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.

3.2. 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.

3.3. Preparación de los datos

En esta etapa, se realizó 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.

3.4. Modelado

Se implementaron el algoritmo de clasificación: Regresión Logística. El modelo fue entrenado utilizando un 70% de los datos y evaluado con un 30% restante.

3.5. Evaluación

El rendimiento del modelo 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 del modelo 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.

3.6. 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.

4. Desarrollo metodológico

4.1. 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.

4.2. Comprensión de los datos

4.2.1. 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.

4.2.2. 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.

4.2.3. 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.

4.2.4. 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.

4.2.4.1. 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.

	Valor_IMC	Num_Edad_Menopausia	Num_Edad_Menarca	Edad	Num_Fam_Primer_Grado_Otros	Num_Fam_Segundo_Grado_Otros	Num_Fam_Primer_Grado_CAM	Num_Fam_Segundo_Grado_CAM
count	1948063.000000	31184.000000	888606.000000	2190210.000000	2190210.000000	2190210.000000	2190210.000000	2190210.000000
mean	26.632020	47.651360	11.046907	43.159532	0.052784	0.047167	0.002419	0.001418
std	4.907999	4.585369	2.065677	16.741103	0.254169	0.230054	0.049989	0.038364
min	13.000000	36.000000	9.000000	18.000000	0.000000	0.000000	0.000000	0.000000
25%	23.310000	45.000000	9.000000	29.000000	0.000000	0.000000	0.000000	0.000000
50%	25.964542	48.000000	11.000000	40.000000	0.000000	0.000000	0.000000	0.000000
75%	29.296875	51.000000	13.000000	55.000000	0.000000	0.000000	0.000000	0.000000
max	89.970000	57.000000	17.000000	139.000000	4.000000	3.000000	2.000000	2.000000

Ilustración 1 Descripción 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.

	Ind_CAM	Ind_Frecuencia_Licor	Sexo_Cd	Raza_Desc	Ind_Terapia_Hormonal	Num_Birads	Ind_Ooforectomia_Bilateral	Ind_Ant_Fam_CAM	Ind_Ant_Radio_Torax	Ind_Ant_Fam_Otros_Esperado
count	2190210	1449639	2190210	637948	2190210	483376	2190210	2190210	2190210	2190210
unique	2	2	1	6	2	7	2	2	2	2
top	No	No	F	MESTIZO	No	2	No	No	No	No
freq	2171958	1192433	2190210	478445	2186046	289106	2186211	2182078	1790150	2006987

Ilustración 2 Descripción estadística variables categóricas

4.2.4.2. 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.

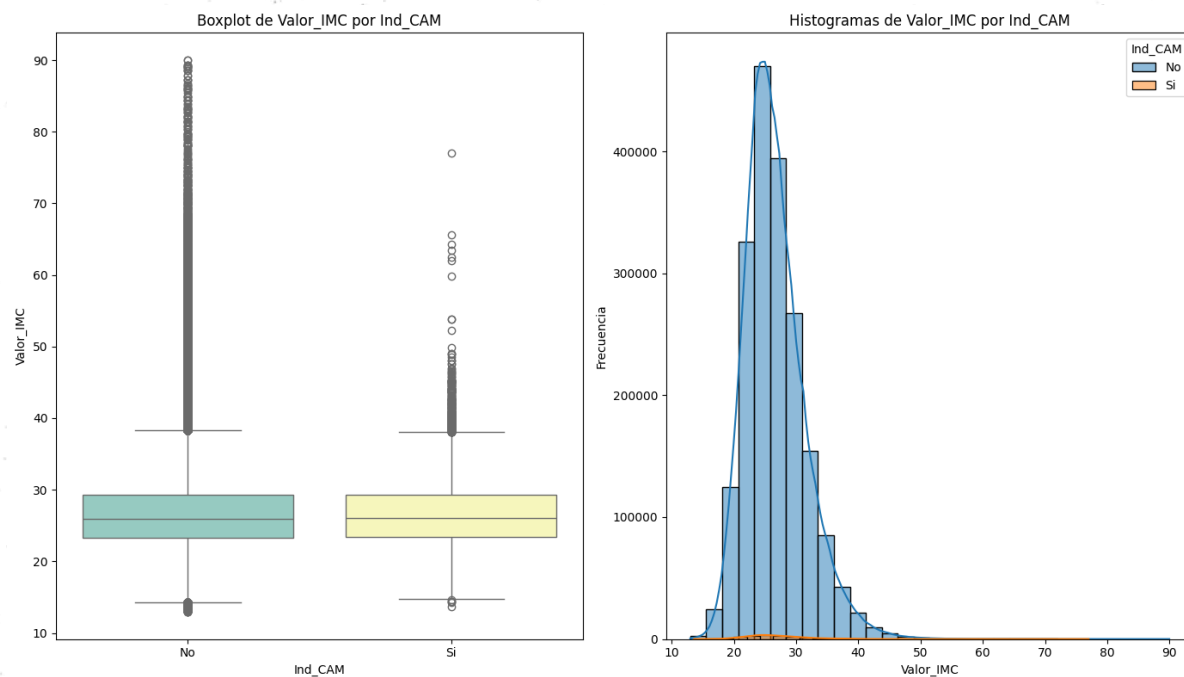


Ilustración 3 Gráfico Valor IMC vs. Indicador cáncer de mama

Universidad EAFIT-Campus principal
Carrera 49 7 Sur 50, avenida Las Vegas
Medellín-Colombia
Teléfonos: (57) (4) 2619500-4489500
Apartado Aéreo: 3300 | Fax: 3120649
Nit: 890.901.389-5

EAFIT Llanogrande
Teléfono: (57) (4) 2619500 exts. 9562-9188
EAFIT Bogotá
Teléfonos: (57) (1) 6114523-6114618
EAFIT Pereira
Teléfono: (57) (6) 3214115

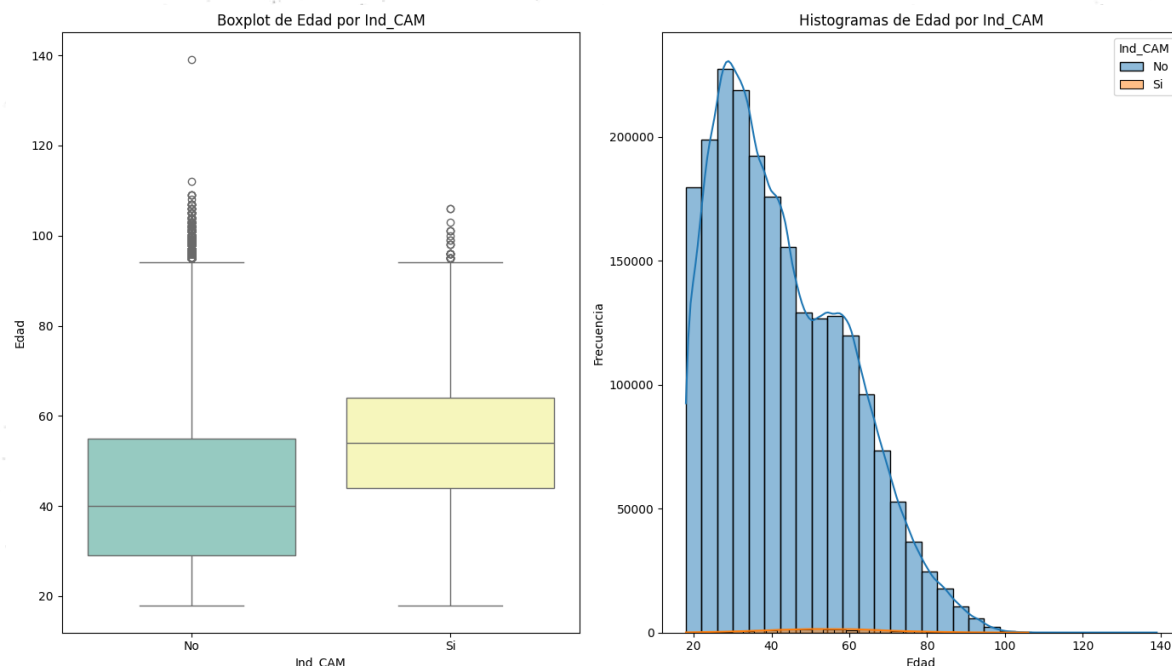


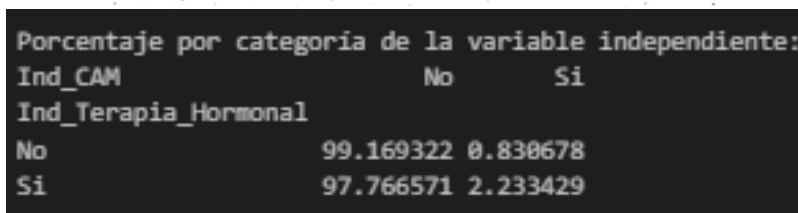
Ilustración 4 Gráfico edad vs. Indicador cáncer de mama

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.

Porcentaje por categoría de la variable independiente:		
Ind_CAM	No	Si
Raza_Desc		
AFROAMERICANO	98.848793	1.151207
BLANCO	97.947286	2.052714
INDÍGENA	98.666667	1.333333
MESTIZO	98.586202	1.493798
MULATO	98.676665	1.323335
ZAMBO	98.291572	1.708428

Ilustración 5 Porcentaje 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%).



Porcentaje por categoría de la variable independiente:		
Ind_CAM	No	Si
Ind_Terapia_Hormonal		
No	99.169322	0.830678
Si	97.766571	2.233429

Ilustración 6 Porcentaje 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.

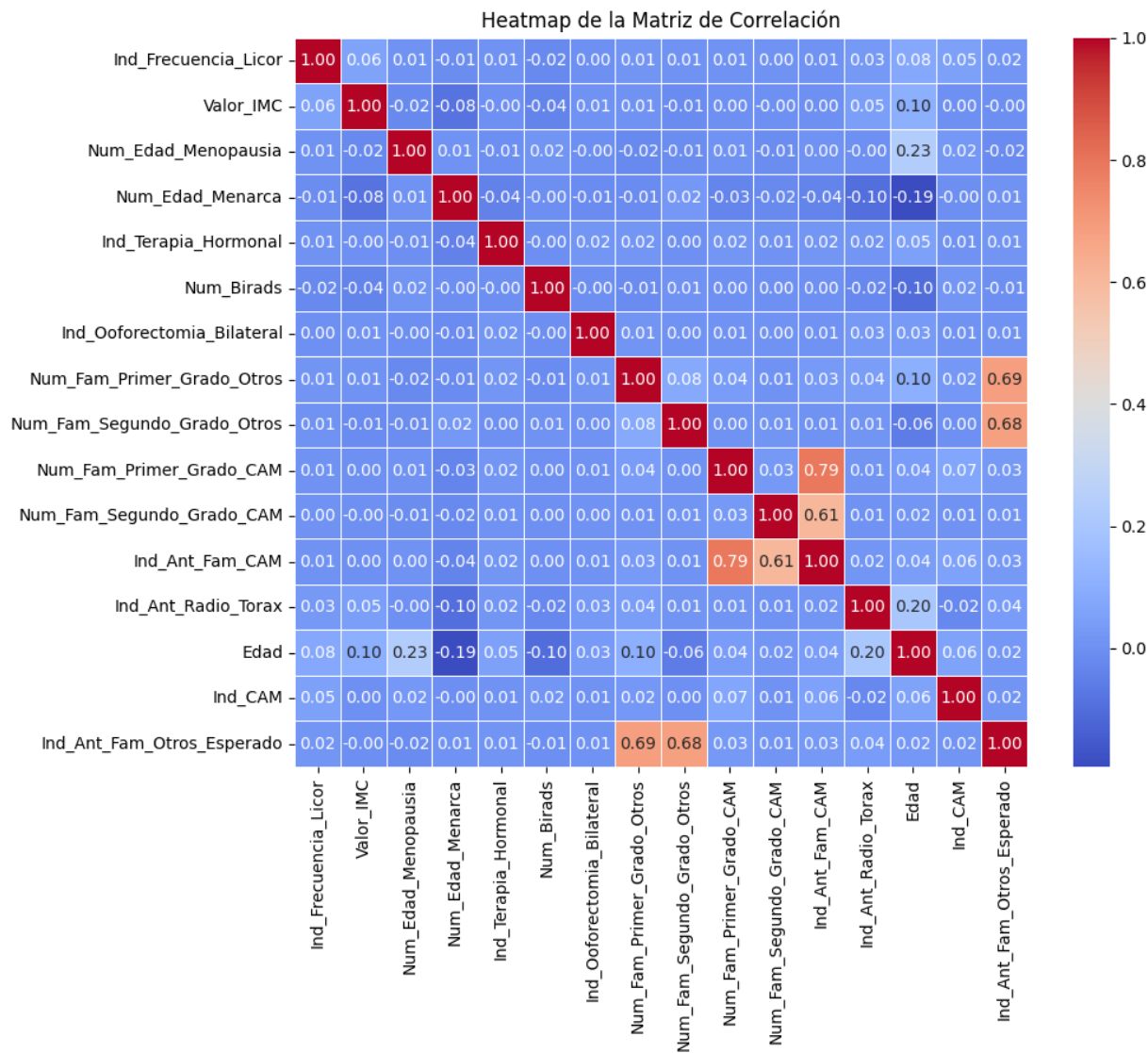


Ilustración 7 Matriz de correlación variables independientes

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.

4.3.Preparación de los datos

Inicialmente, se seleccionaron las variables consideradas relevantes para el análisis, eliminando las variables Sexo y Afiliado_id para enfocar el modelo en factores directamente relacionados con el riesgo de cáncer de mama.

Seguido de esto, se procedió a limpiar las variables seleccionadas, abordando los datos faltantes y los valores atípicos identificados durante el análisis exploratorio. Para los datos faltantes el tratamiento fue el siguiente:

- Ind_Frecuencia_Licor: Se imputó con la moda.
- Raza: Se categorizó como 1 para raza blanca y 0 para otras razas, basándose en estudios que indican un mayor riesgo asociado a esta categoría.
- Valor_IMC: Se imputó con la mediana para manejar los más de 46,000 valores atípicos.
- Num_Edad_Menopausia: Se creó una nueva variable categórica Ind_Menopausia, asignando 1 si la edad era mayor a 50 años y 0 en caso contrario.
- Num_Edad_Menarca: Se reemplazó con la mediana.
- Num_Birads: Se eliminó debido a la presencia de valores no válidos que no podían imputarse de manera consistente.

Por otro lado, para los valores atípicos, calculados con el rango intercuartílico, en el caso del Valor_IMC se reemplazaron con la mediana. Para el caso de la edad se eliminaron los registros con edades superiores a 75 años, alineándose con el objetivo del estudio.

Acto seguido, se crea la variable `Tiempo_Exposicion_Hormonal` con el fin de enriquecer el modelo, dicha variable se refiere a la duración de la exposición hormonal que tuvo o ha tenido la mujer durante su vida. El calculo de dicha variable cumple los siguientes criterios:

- Caso 1: Mujeres con edad < 50 años y sin dato en `Num_Edad_Menopausia`:

$$\text{Tiempo_Exposicion_Hormonal} = \text{Edad} - \text{Num_Edad_Menarca}.$$

- Caso 2: Mujeres con datos en `Num_Edad_Menopausia` y `Num_Edad_Menarca`:

$$\text{Tiempo_Exposicion_Hormonal} = \text{Num_Edad_Menopausia} - \text{Num_Edad_Menarca}$$

- Caso 3: Mujeres con edad > 50 años y sin dato en `Num_Edad_Menopausia`:

$$\text{Tiempo_Exposicion_Hormonal} = 50 - \text{Num_Edad_Menarca}.$$

Tras la creación de esta variable, se eliminó `Num_Edad_Menopausia` para evitar redundancias.

También se creo la variable `Categoria_IMC` que clasifica el `Valor_IMC` en las siguientes categorías:

- Bajo peso: $\text{IMC} < 18.5$
- Normal: $18.5 \leq \text{IMC} < 25$
- Sobrepeso: $25 \leq \text{IMC} < 30$
- Obesidad: $\text{IMC} \geq 30$

Finalizado el proceso anterior, se codificaron las variables categóricas utilizando one-hot encoding, excepto para `Categoria_IMC`, que se codificó con un ordinal encoder para preservar el orden inherente a las categorías.

Las variables seleccionadas para el modelo son:

- `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

Con las variables seleccionadas, se procede a proyectar los datos sobre los eigenvectores asociados a los eigenvalores más altos. Estos eigenvectores representan las direcciones de mayor varianza en los datos, lo que permite reducir el espacio dimensional manteniendo la mayor parte de la información relevante. Este nuevo espacio de menor dimensión está compuesto por combinaciones lineales de las variables originales, y sus componentes principales son mutuamente independientes, es decir, linealmente independientes. Este proceso se conoce como Análisis de Componentes Principales (PCA).

Se realizó una descomposición en valores singulares (SVD) de la matriz de datos, lo que permitió obtener las matrices U , S , y V^T . En este proceso, se verificó que las columnas de la matriz V o V^T traspuesta correspondían a los eigenvectores de la matriz de covarianza de los datos. Esto ocurre porque el SVD y la descomposición espectral de la matriz de covarianza están matemáticamente relacionados: los eigenvectores de la matriz de covarianza son las direcciones principales del espacio que maximizan la varianza, y estas mismas direcciones están contenidas en V durante el SVD.

Además, al proyectar los datos originales sobre las columnas de V , se obtuvieron las mismas proyecciones que al usar los eigenvectores de la matriz de covarianza. Esto confirma que tanto la descomposición SVD como la descomposición espectral son métodos equivalentes para realizar el PCA, ya que ambas identifican las mismas direcciones principales en los datos.

Matriz de covarianzas:

Matriz de covarianzas:

	0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16
0	1.000000	-0.033582	0.013851	-0.012974	0.000690	-0.001241	0.147991	0.000000	0.090586	0.180699	0.055401	-0.000772	0.008312	-0.000274	0.046713	0.000105	0.911988
1	-0.033582	1.000000	-0.009733	0.014673	-0.029602	-0.014900	-0.074185	0.000000	-0.095750	-0.159145	-0.018607	-0.036768	-0.009373	-0.032829	-0.058758	0.005692	-0.029367
2	0.013851	-0.009733	1.000000	0.081711	0.038810	0.008709	0.116520	0.000000	0.101656	0.117722	0.029585	0.016440	0.013959	0.035080	0.044844	0.681274	0.009154
3	-0.012974	0.014673	0.081711	1.000000	0.004134	0.013246	-0.047945	0.000000	-0.057149	-0.035251	0.026055	0.001721	0.003199	0.011334	0.019612	0.685107	-0.014355
4	0.000690	-0.029602	0.038810	0.004134	1.000000	0.031181	0.037499	0.000000	0.034541	0.035873	0.006721	0.018458	0.005107	0.782637	0.012825	0.026251	0.000468
5	-0.001241	-0.014900	0.008709	0.013246	0.031181	1.000000	0.020426	0.000000	0.015390	0.023305	0.005163	0.011290	0.002888	0.621451	0.009004	0.013301	-0.001611
6	0.147991	-0.074185	0.116520	-0.047945	0.037499	0.020426	1.000000	0.000000	0.835406	0.937995	0.076333	0.050609	0.034075	0.041729	0.178780	0.039328	0.129163
7	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	1.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000
8	0.090586	-0.095750	0.101656	-0.057149	0.034541	0.015390	0.835406	0.000000	1.000000	0.714772	0.069069	0.050300	0.034252	0.036074	0.153152	0.022414	0.078459
9	0.180699	-0.159145	0.117722	-0.035251	0.035873	0.023305	0.937995	0.000000	0.714772	1.000000	0.066264	0.044181	0.034759	0.042306	0.166781	0.050042	0.158733
10	0.055401	-0.018607	0.029585	0.026055	0.006721	0.005163	0.076333	0.000000	0.069069	0.066264	1.000000	0.008121	0.002419	0.008592	0.050083	0.039125	0.041385
11	-0.000772	-0.036768	0.016440	0.001721	0.018458	0.011290	0.050609	0.000000	0.050300	0.044181	0.008121	1.000000	0.018373	0.021553	0.017735	0.012384	-0.001756
12	0.008312	-0.009373	0.013959	0.003199	0.005107	0.002888	0.034075	0.000000	0.034252	0.034759	0.002419	0.018373	1.000000	0.005835	0.026028	0.011118	0.006704
13	-0.000274	-0.032829	0.035080	0.011334	0.782637	0.621451	0.041729	0.000000	0.036074	0.042306	0.008592	0.021553	0.005835	1.000000	0.016122	0.028600	-0.000623
14	0.046713	-0.058758	0.044844	0.019612	0.012825	0.009004	0.178780	0.000000	0.153152	0.166781	0.050083	0.017735	0.026028	0.016122	1.000000	0.043227	0.034670
15	0.000105	0.005692	0.681274	0.685107	0.026251	0.013301	0.039328	0.000000	0.022414	0.050042	0.039125	0.012384	0.011118	0.028600	0.043227	1.000000	-0.004251
16	0.911988	-0.029367	0.009154	-0.014355	0.000468	-0.001611	0.129163	0.000000	0.078459	0.158733	0.041385	-0.001756	0.006704	-0.000623	0.034670	-0.004251	1.000000

La matriz de covarianzas tuvo un determinante de cero y un número de condicionamiento extremadamente alto:

Número de condición de la matriz de covarianzas: $4.867325928220324e+16$

El determinante por lo tanto sería cercano a cero o cero, debido a esto, siendo esta matriz singular:

Determinante de la matriz de covarianzas: 0.0

Se considera a PCA como una estrategia efectiva para abordar problemas derivados de matrices de covarianzas mal condicionadas, lo que justifica su exploración en estos escenarios. Una matriz mal condicionada, con un número de condicionamiento elevado o un determinante cercano a cero, indica redundancia entre variables y direcciones de baja varianza que afectan la estabilidad numérica. PCA identifica estas redundancias al calcular los eigenvectores y eigenvalores de la matriz de covarianzas, seleccionando solo las direcciones asociadas a los eigenvalores más grandes, que representan la mayor variabilidad en los datos. Esto elimina direcciones de varianza despreciable, mejora el rango de la matriz y mitiga el mal condicionamiento. Proyectar los datos en este espacio reducido elimina redundancias y correlaciones, simplifica la estructura de los datos y preserva su información más relevante, reduciendo el impacto del ruido y mejorando la estabilidad computacional. Por estas razones, PCA es una herramienta útil para transformar datos de alta dimensionalidad en un formato más manejable y robusto para análisis y modelado.

Por lo tanto, los resultados obtenidos al calcular los componentes principales proyectados sobre los datos necesarios para explicar al menos el 90% de la varianza muestran que el número óptimo de eigenvectores es

diez y asimismo el número de componentes principales que se detallan en la siguiente tabla:

	0	1	2	3	4	5	6	7	8	9
1696732	-0.917673	-0.220904	0.196553	0.454618	-0.869938	-0.537700	-0.274892	-0.204871	-0.843202	1.887623
519581	-0.819513	-0.369647	0.272851	0.667274	0.109196	0.326308	0.196165	0.015597	-0.121700	-0.190209
918747	2.742583	-1.096312	0.364741	0.919393	-0.119404	0.435572	0.808163	0.016429	-0.890739	-0.968079
699063	1.647989	-0.241447	0.190423	-1.953556	-0.328472	-0.420165	-0.300893	-0.078551	-0.697225	1.686384
1531369	-0.386835	-0.444343	0.147918	1.163383	-0.288827	-1.850732	-0.837731	-0.153229	1.785628	-0.911875

Esta puede ser una estrategia para evaluar el modelo y se resalta no será la única, pues también se evaluará con el espacio completo inicial sin la proyección con el fin de comparar. Además se obtuvieron los mismo resultados para descomposición SVD.

4.4. Modelado

Para la fase de modelación, se dividen los datos en un 70% para entrenamiento y un 30% para prueba.

Las variables numéricas del conjunto de datos de entrenamiento son estandarizadas restándoles su media y dividiéndolas por la desviación estándar.

Teniendo la base de datos lista, se procede a modelar.

4.4.1. Regresión logística

4.4.1.1. Modelo con todas las variables

Se realizaron entrenamientos con datos de PCA y para los datos completos:

Los resultados fueron para PCA fueron:

Evaluación de Regresión Logística ajustada:

Accuracy: 0.9919385924420735

Precisión: 0.0

Recall: 0.0

F1-score: 0.0

Matriz de Confusión (Regresión Logística):

	0	1
0	416886	0
1	3388	0

Los resultados sin PCA fueron:

Evaluación de Regresión Logística ajusta

Accuracy: 0.9919195572412284

Precisión: 0.3181818181818182

Recall: 0.002066115702479339

F1-score: 0.004105571847507331

Matriz de Confusión (Regresión Logística)

	0	1
0	416871	15
1	3381	7

Cuando un modelo de regresión logística produce los resultados mostrados en la imagen, con una precisión muy alta, pero con valores de recall, precisión y F1-score muy bajos, es probable que el modelo esté enfrentando problemas de desbalanceo en las clases. Esto significa que la mayoría de las predicciones del modelo están siendo clasificadas como la clase mayoritaria (en este caso, la clase "0"), y el modelo no está

aprendiendo adecuadamente a identificar los casos de la clase minoritaria (la clase "1").

4.4.1.2. KNN

De manera similar se tienen los siguientes resultados con PCA:

```
Evaluación de KNN ajustado:
Accuracy: 0.986815744014619
Precisión: 0.006419073819348923
Recall: 0.004132231404958678
F1-score: 0.005027832644999102
Matriz de Confusión (KNN):
      0      1
0  414719  2167
1   3374    14
```

También se realizó con PCA ajustando los vecinos:

```
Evaluación de KNN ajustado:
Precisión: 0.8883071553228621

      precision    recall  f1-score   support

0         0.98        0.91        0.94        2225
1         0.07        0.24        0.11          67

accuracy          0.89        2292
macro avg         0.52        0.57        0.53        2292
weighted avg      0.95        0.89        0.92        2292

Matriz de Confusión (KNN):
[[2020  205]
 [  51   16]]
```


Sin PCA:

Evaluación de KNN ajustado:
Accuracy: 0.9902991857692839
Precisión: 0.10265282583621683
Recall: 0.02626918536009445
F1-score: 0.0418331374853114
Matriz de Confusión (KNN):

	0	1
0	416108	778
1	3299	89

Este comportamiento puede explicarse por varias razones. En este caso particular, el desbalance de clases es muy marcado, y KNN tiende a favorecer la clase mayoritaria porque la mayoría de los vecinos de cualquier punto estarán en esta clase. Además, los resultados son similares tanto con los datos originales como con PCA, lo que sugiere que la proyección a un espacio de menor dimensionalidad no logró separar de manera significativa las clases en el espacio reducido. Esto puede indicar que la distribución de las clases es inherentemente compleja o que los datos no contienen suficiente información distintiva para el modelo. Por lo tanto, el desempeño limitado de KNN en este caso específico está influenciado por el desbalance de clases, la

naturaleza de los datos, y la incapacidad del modelo para capturar patrones relevantes en un entorno con clases minoritarias tan pequeñas.

4.4.1.3. LDA

Los resultados con PCA fueron:

Evaluación de LDA:

Accuracy: 0.9905371257798484

Precisión: 0.02576489533011272

Recall: 0.004722550177095631

F1-score: 0.007982040409079572

Matriz de Confusión (LDA):

	0	1
0	416281	605
1	3372	16

Sin PCA:

Evaluación de LDA:

Accuracy: 0.990130248361783

Precisión: 0.12450592885375494

Recall: 0.0371900826446281

F1-score: 0.057272727272727274

Matriz de Confusión (LDA):

	0	1
0	416000	886
1	3262	126

Esto puede deberse a que LDA, al basarse en la maximización de la separación entre las medias de las clases mientras minimiza la varianza intraclase, utiliza toda la información disponible en los datos originales para proyectar a un espacio que optimice la discriminación entre clases. Al aplicar PCA previamente, se podría haber perdido información relevante para la separación de clases, ya que PCA selecciona componentes en función de la varianza global de los datos y no necesariamente en función de su capacidad discriminativa. observa un mejor comportamiento en la precisión, sin embargo, aun el desbalanceo de clases es evidente.

4.4.1.4. QDA

Con PCA:

Evaluación de QDA:
Accuracy: 0.9853143425479568
Precisión: 0.05555555555555555
Recall: 0.051357733175914994
F1-score: 0.05337423312883435
Matriz de Confusión (QDA):

	0	1
0	413928	2958
1	3214	174

Sin PCA:

Evaluación de QDA:

Accuracy: 0.875076735653407

Precisión: 0.018376872989723073

Recall: 0.2765643447461629

F1-score: 0.03446373400029425

Matriz de Confusión (QDA):

	0	1
0	366835	50051
1	2451	937

Los resultados obtenidos con QDA fueron significativamente peores, lo cual puede atribuirse a las suposiciones y características particulares del modelo. QDA asume que las clases en los datos siguen distribuciones normales (gaussianas) con matrices de covarianza distintas para cada clase. Si esta suposición no se cumple, lo cual es común en datos reales, el modelo tiende a tener un rendimiento pobre. Además, al ser un modelo que depende de la inversión de las matrices de covarianza, un número elevado de dimensiones o colinealidad entre las variables puede hacer que las matrices sean mal condicionadas o incluso no invertibles.

4.4.1.5. Mejor Modelo y Evaluación

El mejor modelo obtenido hasta ahora se encuentra entre el KNN ajustado, optimizando el número de vecinos y ajustando los umbrales de decisión, junto con el LDA. El KNN con PCA logra superar un 10% de F1-score, mientras que el LDA muestra una precisión notablemente alta, aunque su F1-score se encuentra alrededor del 5%.

Este desempeño puede explicarse por las diferencias en las características y suposiciones de ambos modelos. En el caso del KNN, su capacidad de adaptarse a la estructura local de los datos le permite capturar algunas relaciones en la clase minoritaria, especialmente tras el ajuste de umbrales, que mejora el balance entre precisión y recall. Además, el uso de PCA ayuda a reducir el ruido y concentrar la varianza relevante en un espacio de menor dimensión, lo que facilita el proceso de clasificación en un modelo sensible como KNN. Sin embargo, su dependencia de la distancia puede hacerlo vulnerable a desequilibrios en las clases, lo que limita su rendimiento general.

Por otro lado, el LDA, a pesar de su alta precisión, obtiene un F1-score bajo debido a su limitada capacidad para capturar la clase minoritaria. Este modelo asume distribuciones gaussianas con la misma matriz de covarianza para todas las clases, lo que puede no ajustarse completamente a la realidad de los datos. Aunque funciona bien en clasificaciones lineales y logra identificar correctamente la mayoría de los ejemplos de la clase mayoritaria, el desbalance de clases y la baja representación de la clase minoritaria reducen su capacidad para lograr un recall y un F1-score altos.

En resumen, tanto el KNN ajustado como el LDA tienen ventajas y limitaciones particulares, y su desempeño refleja cómo las características del modelo interactúan con las propiedades específicas del conjunto de datos, como el desbalance de clases y la distribución de las variables.

4.5.Despliegue

4.5.1. Ciclo de vida de los datos y procesamiento

Para soportar las necesidades de almacenamiento y procesamiento de datos, se emplea la infraestructura en la nube proporcionada por Microsoft Azure.

Específicamente:

Universidad EAFIT-Campus principal
Carrera 49 7 Sur 50, avenida Las Vegas
Medellín-Colombia
Teléfonos: (57) (4) 2619500-4489500
Apartado Aéreo: 3300 | Fax: 3120649
Nit: 890.901.389-5

EAFIT Llanogrande
Teléfono: (57) (4) 2619500 exts. 9562-9188
EAFIT Bogotá
Teléfonos: (57) (1) 6114523-6114618
EAFIT Pereira
Teléfono: (57) (6) 3214115

- Azure Databricks: Utilizado para el procesamiento de datos, entrenamiento de modelos y gestión de pipelines de datos. Databricks ofrece un entorno escalable y optimizado para análisis de big data y machine learning.
- Azure Data Lake Storage (ADLS): Empleado para el almacenamiento de datos estructurados, permitiendo manejar grandes volúmenes de información con alta eficiencia y seguridad.
- Teradata: Utilizado como data warehouse de la compañía, donde se realiza el proceso de ETL (Extract, Transform, Load) para la ingestión mensual de datos.

Los datos utilizados en este proyecto son proporcionados por la Entidad Administradora de Planes de Beneficio (EAPB) y se consultan directamente en Azur. La ingesta de datos se realiza mediante un proceso de ETL (Extract, Transform, Load) en el data warehouse de la compañía, Teradata. Este proceso se ejecuta de manera batch una vez al mes, extrayendo y transformando los datos necesarios para el modelado predictivo. La periodicidad mensual asegura que los modelos se actualicen con información reciente sin sobrecargar los recursos de procesamiento de Databricks.

4.5.2. Almacenamiento de los datos

Los datos procesados se almacenan en Azure Data Lake Storage (ADLS), aprovechando su capacidad para manejar grandes volúmenes de datos estructurados de manera eficiente.

4.5.3. Frameworks de procesamiento

Para el procesamiento y análisis de los datos, se emplea Azure Databricks. Este framework permite gestionar los pipelines de datos y los modelos entrenados de manera eficiente. Los pipelines automatizados en Databricks aseguran que los datos se procesen de forma consistente y reproducible, facilitando la actualización periódica de los modelos con nuevos datos ingresados.

4.5.4. Despliegue del modelo

Los resultados de los modelos se exportan en formato .parquet, un formato columnar eficiente para el almacenamiento y la transferencia de datos. Estos archivos se transfieren a una aplicación de gestión de pólizas en SAP, permitiendo que los profesionales de salud accedan a las predicciones de riesgo de cáncer de mama directamente desde su sistema de gestión. Este enfoque simplifica la integración de los resultados del modelo en los procesos operativos existentes de la EAPB.

5. Conclusiones

6. 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.

Universidad EAFIT-Campus principal
Carrera 49 7 Sur 50, avenida Las Vegas
Medellín-Colombia
Teléfonos: (57) (4) 2619500-4489500
Apartado Aéreo: 3300 | Fax: 3120649
Nit: 890.901.389-5

EAFIT Llanogrande
Teléfono: (57) (4) 2619500 exts. 9562-9188
EAFIT Bogotá
Teléfonos: (57) (1) 6114523-6114618
EAFIT Pereira
Teléfono: (57) (6) 3214115