



---

# **Predicción del Riesgo de Cáncer de Mama Basado en Factores Clínicos y Demográficos Mediante Técnicas de Aprendizaje Automático**

**Máster Big Data, Data Science & Inteligencia  
Artificial**

**Tutores:**

Carlos Ortega, Santiago Mota

**Alumno:**

Eva María Villar Álvarez  
77401448P

*18 de septiembre de 2025*

---

# Tabla de Contenido

<b>1. Introducción .....</b>	<b>3</b>
1.1. Metodología CRISP-DM .....	3
1.2. Introducción al contexto del problema.....	3
1.3. Objetivos .....	4
<b>2. Análisis Descriptivo del Dataset .....</b>	<b>4</b>
2.1. Descripción del Dataset.....	4
2.2. Análisis Exploratorio de las Variables (EDA).....	6
<b>3. Preprocesado de Datos (Feature Engineering).....</b>	<b>9</b>
3.1. División Train/Test.....	10
3.2. Feature Transformation .....	10
3.3. Balance de Clases.....	11
3.4. Selección de Variables .....	12
<b>4. Modelos Predictivos de Machine Learning .....</b>	<b>13</b>
4.1. Modelos Utilizados.....	14
4.2. Métricas de Evaluación .....	16
4.3. Búsqueda de Hiperparámetros para Cada Modelo.....	17
<b>5. Comparación y Resultados Finales .....</b>	<b>18</b>
5.1. Evaluación y Comparación de los Modelos .....	18
5.2. Evaluación de la Estabilidad y Robustez de los Modelos .....	19
5.3. Elección del mejor modelo .....	19
<b>6. Interpretación del Modelo Seleccionado .....</b>	<b>19</b>
<b>7. Despliegue del Modelo (Puesta en Producción).....</b>	<b>20</b>
<b>8. Conclusiones y Perspectivas Futuras .....</b>	<b>20</b>
<b>9. Bibliografía.....</b>	<b>22</b>
<b>Anexo A: Análisis Exploratorio de los Datos (EDA).....</b>	<b>23</b>
<b>Anexo B: Selección de Variables .....</b>	<b>37</b>
<b>Anexo C: Búsqueda Hiperparamétrica .....</b>	<b>40</b>
<b>Anexo D: Comparación Entre los Modelos Candidatos.....</b>	<b>56</b>
<b>Anexo E: Interpretabilidad del Modelo Ganador .....</b>	<b>60</b>
<b>Anexo F: Puesta en Producción .....</b>	<b>62</b>
<b>Anexo G: Código del Proyecto Desarrollado en Python.....</b>	<b>64</b>

## 1. Introducción

El cáncer de mama es el cáncer más frecuente en mujeres a nivel mundial. La identificación temprana de factores de riesgo permite orientar estrategias de prevención, optimizar la vigilancia médica e identificar a los grupos con mayor probabilidad de desarrollar la enfermedad. De este modo, se facilita la detección precoz y se contribuye a mejorar las tasas de supervivencia, además de reducir costes sanitarios asociados a tratamientos más agresivos y tardíos.

En este trabajo se analiza la relación entre diversos factores de riesgo y la incidencia del cáncer de mama, integrando dicha información en un modelo predictivo robusto que sirva como herramienta de apoyo en la toma de decisiones clínicas y en la planificación de estrategias de prevención personalizadas. Para ello se emplea el **Breast Cancer Surveillance Consortium (BCSC) Risk Factor Dataset** [1], un conjunto de datos público y diverso que recopila información clínica y demográfica de mujeres de distintos grupos raciales, obtenida mediante cribado mamográfico durante un periodo máximo de diez años.

### 1.1. Metodología CRISP-DM

Para estructurar este trabajo se ha seguido la metodología **CRISP-DM** (*Cross-Industry Standard Process for Data Mining*), ampliamente utilizada en proyectos de ciencia de datos orientados a negocio [2]. Esta metodología guía el proceso desde la comprensión del problema hasta la implementación y monitorización de la solución. Las seis fases de CRISP-DM se aplicaron en este trabajo de la siguiente manera:

- **Comprensión del problema:** Definición del contexto clínico del cáncer de mama y establecimiento de los objetivos del modelo como herramienta de prevención.
- **Comprensión de los datos.** Exploración inicial del “*BCSC Risk Factor Dataset*” para evaluar calidad, estructura y distribución de las variables clínicas y demográficas. Se realizaron análisis descriptivos y visualizaciones (histogramas, gráficos de barras, etc.) para identificar patrones y posibles correlaciones.
- **Preparación de los datos:** Construcción del conjunto de datos definitivo mediante limpieza, imputación de valores faltantes, transformación de variables categóricas (codificación *one-hot* u *ordinal encoding*) y selección de las variables más relevantes. Además, se aplicaron técnicas de *resampling* para balancear las clases y evitar sesgos hacia la clase mayoritaria.
- **Modelado Predictivo:** Desarrollo y ajuste de diferentes modelos de *machine learning* (Regresión Logística, Random Forest, XGBoost, LightGBM y un modelo de *Stacking*). Se aplicó ajuste de hiperparámetros y técnicas para evitar *overfitting*, priorizando métricas de interés clínico, especialmente el recall de la clase positiva.
- **Evaluación:** Comparación de modelos mediante métricas como AUC, recall y curvas ROC. Se valoró también la interpretabilidad y la estabilidad del modelo.
- **Despliegue:** Definición de una herramienta interactiva para la implementación del uso del modelo (p. ej. mediante *Streamlit*), con la posibilidad de ofrecer interpretaciones de las predicciones a nivel individual de la paciente.

CRISP-DM es flexible, lo que permite iterar entre fases según sea necesario, asegurando que los resultados finales estén alineados con los objetivos planteados y ofreciendo una solución robusta y reproducible.

### 1.2. Introducción al contexto del problema

El cáncer de mama es el tipo de cáncer más frecuente en mujeres a nivel mundial, representando aproximadamente el 25% de los nuevos diagnósticos de cáncer [3]. Su detección y prevención temprana son determinantes para mejorar las tasas de supervivencia, ya que aproximadamente la mitad de los casos se asocian a factores de riesgo demográficos y clínicos conocidos.

Determinar el perfil de riesgo individual de cada mujer es fundamental para orientar estrategias preventivas, como, por ejemplo, optimizar la vigilancia médica mediante mamografías de cribado, identificar grupos de riesgo que podrían beneficiarse de seguimiento individualizados priorizando

pruebas genéticas en determinados perfiles, así como aconsejar cambios en hábitos de vida. Una predicción precisa del riesgo constituye un desafío relevante para los oncólogos clínicos, pues no solo facilita la detección temprana en mujeres con mayor probabilidad de desarrollar la enfermedad, sino que también contribuye a reducir costes sanitarios si se ataja a tiempo evitando tratamientos muy agresivos y costosos.

La clasificación mediante aprendizaje supervisado es una de las técnicas de *machine learning* aplicadas en el ámbito médico para la predicción de riesgo. Sin embargo, lograr un buen desempeño en datos reales es un reto, debido al desbalance entre clases, donde la mayoría de los modelos tiende a identificar correctamente la clase mayoritaria mientras puede ignorar la minoritaria (pacientes con cáncer). Diversos algoritmos han demostrado mejorar la precisión en el diagnóstico del cáncer de mama, superando incluso los métodos radiológicos tradicionales en un 11,5 % [4]. Esto evidencia la necesidad de desarrollar procedimientos automatizados de cribado que apoyen la toma de decisiones clínicas.

### 1.3. Objetivos

El objetivo final de este trabajo es construir un modelo predictivo robusto que apoye la toma de decisiones clínicas y contribuya a un seguimiento y tratamiento más personalizados, facilitando la detección precoz del cáncer de mama y la planificación de terapias menos agresivas y costosas.

Los objetivos específicos del trabajo son:

1. Identificar los factores de riesgo clínicos y demográficos con mayor peso predictivo en la aparición del cáncer de mama.
2. Evaluar y comparar distintos modelos de clasificación de *machine learning*.
3. Desarrollar un modelo predictivo que combine alta sensibilidad y aplicabilidad práctica real para estimar la incidencia de cáncer de mama.

## 2. Análisis Descriptivo del Dataset

### 2.1. Descripción del Dataset

El **Breast Cancer Surveillance Consortium (BCSC) Risk Factor Dataset** es un conjunto de datos clínicos anónimos que contiene información agregada sobre factores de riesgo asociados al cáncer de mama en mujeres que se han sometido a mamografías entre 2005 y 2017.

Estos datos son públicos, financiados por agencias federales estadounidenses, incluyendo el *National Cancer Institute* (P01CA154292; U54CA163303), el *Patient-Centered Outcomes Research Institute* (PCS-1504-30370) y la *Agency for Health Research and Quality* (R01 HS018366-01A1), y está orientado a fines académicos e investigación científica. Para más información, puede visitarse <http://www.bcsc-research.org/>. Su uso requiere cumplir las normas éticas y legales establecidas por el consorcio, garantizando la privacidad de las participantes.

En este trabajo, el conjunto de datos se ha utilizado con fines académicos y de demostración de un posible flujo de puesta en producción de modelos. En un escenario real de uso clínico o comercial, sería obligatorio recopilar datos bajo licencias que permitan dicho uso y reentrenar los modelos con esos datos.

El dataset completo contiene **1 522 340 registros**, representando 6 788 436 mamografías. Cada registro representa un examen por mujer, por año-calendario y por edad, y un perfil de combinación de factores de riesgo. Se priorizan las mamografías de cribado frente a las diagnósticas cuando ambas están disponibles para una misma mujer en un mismo año.

Las variables presentes están relacionadas con factores de riesgo clásicos ampliamente estudiados en epidemiología mamaria, tanto clínicos como demográficos. En total contiene **13 atributos**, en el que se incluye una columna "count" que indica cuántas mujeres comparten exactamente esa misma combinación de valores. El conjunto de datos no presenta valores nulos de forma directa. Los valores desconocidos se codifican con "**9: unknown**". La codificación se resume en la Tabla 1 y puede consultarse en la documentación oficial del consorcio [5]:

## Predicción de Cáncer de Mama Basado en Factores Clínicos y Demográficos

Variable original	Nombre en este estudio	Descripción	Codificación	
year	year	Año de observación	2005–2017	
age_group_5_years	age	Grupo de edad (intervalos de 5 años)	1 = 18-29 2 = 30-34 3 = 35-39 4 = 40-44 5 = 45-49 6 = 50-54 7 = 55-59	8 = 60-64 9 = 65-69 10 = 70-74 11 = 75-79 12 = 80-84 13 = >85
race_eth	race	Raza/etnia	1=No hispana blanca (White) 2=No hispana negra (Black) 3=asiática/isleña del Pacífico (AAPI) 4=Nativa americana 5=Hispana (Hisp) 6=Otra/mixta 9=Desconocida	
first_degree_hx	first_degree	Antecedente familiar en 1er grado	0=No 1=Sí 9=Desconocido	
age_menarche	menarche	Edad de la menarquía	0=>14 1=12-13	2=<12 9=Desconocida
age_first_birth	first_birth	Edad del primer parto	0=<20 1=20-24 2=25-29	3=>30 4=Nulípara 9=Desconocida
BIRADS_breast_density	birads	Densidad mamaria (BI-RADS)	1=Casi completamente grasa 2=Densidades fibroglandulares dispersas 3=Heterogéneamente densa 4=Extremadamente densa 9=Desconocida/sistema diferente	
current_hrt	hrt	Uso de terapia hormonal	0=No 1=Sí 9=Desconocido	
menopaus	menopaus	Estado menopáusico	1=Pre/peri-menopáusica 2=Post-menopáusica 3=Menopausia quirúrgica 9=Desconocido	
bmi_group	bmi	Índice de masa corporal, IMC	1=10-24.99 2=25-29.99 3=30-34.99	4=>35 9=Desconocido
biophx	bioph	Biopsia de mama previa o aspiración	0=No 1=Sí 9=Desconocido	
breast_cancer_history	cancer	Cáncer de mama previo diagnosticado	0=No 1=Sí 9=Desconocido	
count	count	Frecuencia	Numérico, número de mujeres con esa combinación de factores de riesgo	

Tabla 1: Variables del BCSC Risk Factor Dataset utilizadas en este estudio. Se indican el nombre original de cada variable, el nombre adaptado en este trabajo, una breve descripción de su significado y la codificación o posibles valores que puede asumir cada variable.

En el dataset se identifican tres tipos de variables (en lo que sigue, se utilizarán solo los nombres simplificados de las variables):

- **Numéricas:** count y year.
- **Categóricas nominales:** race, first\_degree, hrt, menopaus, bioph y cancer.
- **Categóricas ordinales:** age, menarche, first\_birth, birads y bmi

Las variables ordinales mantienen un orden natural en sus categorías, lo que permite analizarlas considerando relaciones de magnitud o rango, a diferencia de las nominales, que representan grupos sin jerarquía implícita.

La variable '**Cancer**' es la variable objetivo (**target**) para los modelos de predicción. En el conjunto original, esta variable toma tres valores con las siguientes frecuencias: 'Yes' (14,6%) 'No' (63,7%) y 'Unknown', (21,7%). Se observa un claro **desbalance de clases**, siendo la clase minoritaria la correspondiente a pacientes con cáncer de mama.

## 2.2. Análisis Exploratorio de las Variables (EDA)

La finalidad de este apartado es explorar la distribución de cada atributo en función de la variable objetivo, con el fin de evaluar visualmente su capacidad de diferenciar entre los distintos diagnósticos de cáncer, así como analizar posibles relaciones entre variables relevantes para la predicción.

Para simplificar el análisis y centrarse en casos con información fiable, se eliminaron los registros cuya variable objetivo "cancer" con valor 9 (*Unknown*). Tras esta depuración, el número de filas se reduce a 1191438. Dada la gran dimensión de la base de datos original, esta eliminación no afecta de manera significativa los objetivos del estudio y permite un análisis más robusto. La **Fig. 1** muestra la proporción final de pacientes en cada clase: 0 (sano) y 1 (cáncer).

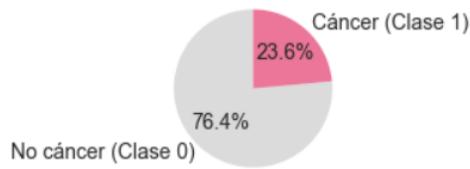


Figura 1: Distribución de la variable objetivo "cancer" tras eliminar los registros con valor "Unknown". Se muestran las proporciones de pacientes sanos (clase 0) y con cáncer (clase 1) en la base de datos depurada

### EDA Variables Numéricas

La variable "year" abarca un periodo de 13 años (2005-2007). Los casos de riesgo están repartidos bastante uniformemente a lo largo del periodo. El máximo de perfiles diferentes (combinaciones de factores de riesgo distintas) se registró en 2009 con 97368. El número total de mujeres por año se mantiene estable en torno a 500000 hasta 2009–2010, para luego mostrar una tendencia decreciente a partir de 2011, alcanzando el mínimo en 2017 con aproximadamente 300000 pacientes. La proporción de mujeres diagnosticadas de cáncer (target = 1) es relativamente constante a lo largo de los años, lo que sugiere que el año de referencia no influye de forma apreciable en la tasa de casos positivos. (Ver **Tabla A1** y **Fig. A1**, en **Anexo A**)

Dado que "year" es únicamente un marcador temporal y no representa un factor de riesgo clínico, **no se utilizará como variable predictiva en la modelización**.

La variable *count* representa la frecuencia de cada combinación de factores de riesgo y no corresponde a un atributo individual. Es decir, indica cuántas pacientes comparten exactamente el mismo perfil de factores de riesgo en el dataset. Su distribución se muestra en la **Tabla A1**.

Al eliminar la variable *year*, los registros correspondientes a un mismo perfil en años distintos se consideraron duplicados, generando 885093 perfiles repetidos. Para solucionar esto, se reconstruyó el dataset agregando los registros por perfil y sumando *count*. Tras esta operación, el dataset final contiene 306345 filas y no presenta duplicados (ver **Tabla A2**)

Cada fila agrupa entre 1 y 15973 mujeres con la misma combinación de factores. Aproximadamente el 25% de las combinaciones son únicas (solo 1 mujer). El percentil 75%, muestra agrupaciones de hasta 8 mujeres con idéntico perfil. En la **Fig. A2** se observa que los perfiles más comunes (valores altos de *count*) son pocos, reflejando que la mayoría de pacientes comparten características frecuentes. En contraste, la gran mayoría de perfiles corresponde a muy pocas pacientes (valores bajos de *count*). Dentro de esta diversidad de combinaciones poco frecuentes, algunas presentan una mayor proporción de casos de cáncer, lo que resalta la importancia de identificar patrones asociados a un mayor riesgo. En resumen, ciertas combinaciones de edad, densidad mamaria, antecedentes familiares y otros factores de riesgo son muy prevalentes en la población, mientras que otras son raras, y en estas combinaciones raras se encuentran tiende a concentrarse un mayor número de casos de cáncer.

Dado que "count" refleja solo la frecuencia de un perfil y no una característica clínica, **tampoco se utiliza como variable predictiva**, aunque resulta útil para analizar la prevalencia de perfiles y la distribución de pacientes en las distintas categorías las variables explicativas (**Fig. A3**)

### EDA Variables Categóricas Nominales

En este apartado se analizan las variables categóricas nominales del conjunto de datos, con el objetivo de comprender cómo se distribuyen los diferentes perfiles de riesgo en la población. Por tanto, aquí lo relevante es la proporción de perfiles en cada categoría y no de pacientes.

Se aplicaron dos tipos de análisis:

- **Univariado**, para describir la distribución de cada variable de manera independiente.
- **Bivariante**, para explorar la relación de las categorías con la variable objetivo (cáncer), utilizando pruebas  $\chi^2$ , tablas de contingencia, mapas de calor (donde los colores más intensos indican mayor asociación) y gráficos comparativos entre la distribución de perfiles y la distribución global de la variable objetivo.

Dado el gran volumen de datos, los valores estadísticos por sí solos pueden no ser concluyentes; y se utilizan como guía para identificar patrones y posibles dependencias relevantes.

Agrupar categorías similares permite asegurar que los resultados reflejen tendencias reales de la población y no estén sesgados por casos aislados. Con el fin de mejorar la representatividad, se agruparon categorías con baja frecuencia (< 5%). En particular, en la variable "race" se fusionó la categoría "Native American" con "Others".

El análisis descriptivo muestra que la mayoría son variables binarias, con excepción de *race* y *menopaus*. Se observa que la raza más representada es la blanca y que predominan las mujeres postmenopáusicas. Asimismo, los perfiles más comunes corresponden a mujeres sin antecedentes familiares en primer grado (*first\_degree*), sin terapia hormonal (*hrt*) y sin biopsias previas (*bioph*) (ver **Tabla A3** y **Figs. A4-A5**).

En el análisis bivariante, la mayoría de variables muestran capacidad para discriminar la variable objetivo (**Figs. A6-A10**), con la excepción de *first\_degree* (**Fig. A8**), que no presenta un patrón diferenciado, comportamiento ya observado en la **Fig. A5**. Los hallazgos más relevantes son:

- **Raza (race)**: la categoría "White" presenta una mayor proporción de casos de cáncer, mientras que "Others" se asocia con ausencia.
- **Menopausia (menopaus)**: el cáncer es más frecuente en mujeres postmenopáusicas, lo que está parcialmente relacionado con la edad.
- **Terapia hormonal (hrt)**: las pacientes con tratamiento hormonal muestran menor incidencia de cáncer de lo esperado.
- **Historial de biopsia (bioph)**: como cabe esperar, las mujeres con biopsias previas presentan una mayor proporción de diagnósticos positivos.

### EDA Variables Categóricas Ordinales

Las variables ordinales son un tipo de variable categórica que, además de agrupar observaciones en categorías, mantienen un orden natural entre ellas. Por ello, para su análisis se adoptó un enfoque doble:

1. **Tratamiento como variables numéricas**: se calcularon estadísticas descriptivas (media, mediana, dispersión) y se evaluó la correlación mediante el coeficiente de Spearman, adecuado para variables ordinales [6].
2. **Tratamiento como variables categóricas**: se examinó la distribución de cada categoría y se aplicaron pruebas de  $\chi^2$  para analizar la relación con la variable objetivo (*cancer*).

Para asegurar representatividad, algunas categorías se recodificaron combinando valores adyacentes, preservando el orden natural. Siguiendo este criterio, la variable *age* se recodificó agrupando las categorías iniciales en "18-39" y las categorías finales en " $\geq 75$ ".

El análisis descriptivo (**Tablas A4–A5 y Figs. A12-15**) muestra que:

- *age* tiene una media en el rango 55–59 años, con distribución relativamente simétrica. El riesgo de cáncer tiende a aumentar con la edad.
- *menarche* y *first\_birth* presentan concentraciones en los valores más bajos, con medianas de 1 (12–13 años) y 2 (25–29 años) respectivamente, indicando que la mayoría de mujeres registradas inicia la menstruación y el primer parto en estos rangos.
- *birads* y *bmi* se distribuyen principalmente en categorías intermedias, con medianas de 2 (densidades fibroglandulares dispersas) y 2 (BMI 25–29,99), reflejando la prevalencia de estos rangos en la población estudiada.
- No se observó correlación significativa entre las variables ordinales analizadas (**Fig. A11**).

En el análisis bivariante, algunas variables ordinales muestran capacidad para discriminar la variable objetivo (**Figs. A16-A20**), destacando:

- *age*: a medida que aumenta, el porcentaje de perfiles con cáncer crece de forma gradual, con un incremento más notable a partir del rango 60–64 años.
- *menarche*: categorías extremas 0 ( $\geq 14$  años) y 2 ( $< 12$  años) muestran un efecto discriminativo leve; los valores faltantes (9) parecen aleatorios y no relacionados con la variable target.
- *birads*: la categoría 4 (“Extremadamente denso”) se asocia con ausencia de cáncer, mientras que el resto de categorías se distribuye de manera equilibrada; los valores faltantes (9) no muestran relación con la variable objetivo.
- *bmi*: la categoría desconocida presenta mayor incidencia de cáncer; el resto de categorías no contribuye significativamente a la discriminación entre presencia o ausencia de la enfermedad.

En general, *age* se perfila como la variable ordinal con mayor capacidad discriminativa (ver **Figs. A12 y A15-A16**), mientras que el resto de ordinales aporta información complementaria pero limitada para predecir *cancer*.

### Medida de Asociación VCramer

Para evaluar la importancia de cada variable con respecto a la variable objetivo, se utilizó el V de Cramer, un estadístico adecuado para variables categóricas nominales y ordinales. Este indicador captura tanto relaciones lineales como no lineales y está acotado entre 0 y 1, donde 0 indica independencia total y 1 dependencia completa.

Los resultados obtenidos son consistentes con el análisis exploratorio previo y permiten ordenar las variables según su poder predictivo (**Fig. A21**):

- **Variables con mayor poder predictivo:** *bioph*, *age*, *menopaus*, *hrt*, *race* y *bmi*.
- **Variables con menor poder predictivo:** *first\_degree*, *first\_birth* y *menarche* ( $V < 0.05$ ; indica que apenas muestran asociación con la variable objetivo y, por tanto, su relevancia predictiva es limitada).

### EDA Valores Faltantes

El estudio de los valores faltantes es un paso esencial en cualquier proyecto de analítica de datos, ya que puede afectar directamente la fiabilidad de los resultados y la capacidad de los modelos predictivos para generalizar. Si no se gestionan adecuadamente, podrían conducir a conclusiones incorrectas.

Se analizaron los valores faltantes del dataset, considerando como tales las categorías etiquetadas como “9” (“unknown”). Este análisis se realizó a dos niveles: variables y observaciones (filas) (**Fig. 2** y **A22**)

- **A nivel de variable:** Ninguna supera el 50% de valores ausentes. Destaca *menarche*, con aproximadamente un 44% de valores faltantes. Eliminarla supondría perder información potencialmente valiosa, su poder predictivo sobre la variable objetivo es limitado.

- **A nivel de observación (fila):** Algunas filas presentan hasta un 75% de variables ausentes; en promedio, los registros contienen un 13% de datos faltantes. Para gestionar esta información, se creó la variable *prop\_missings*, que representa la proporción de valores ausentes por registro (**Tabla A6**). Los registros con más del 30% de valores ausentes presentan cierta relación con la variable objetivo (**Fig. A23**), lo que refuerza la decisión de abordarlos de forma informada, en lugar de imputarlos automáticamente.

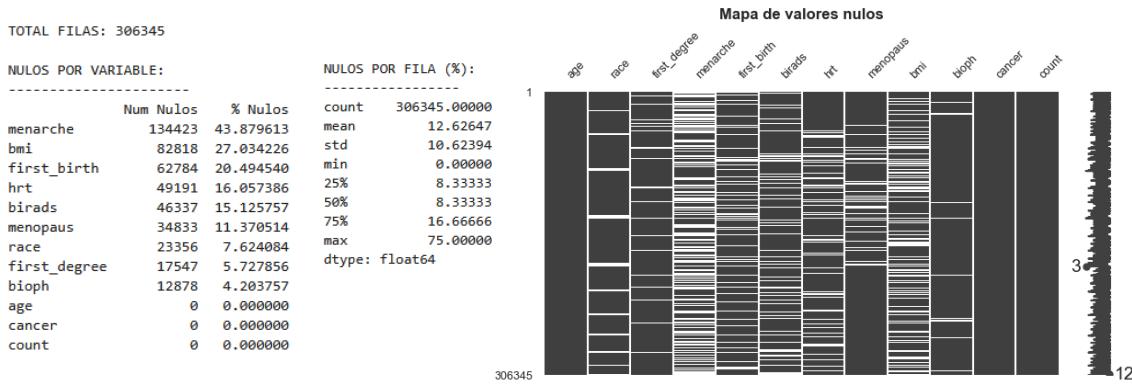


Figura 2: Tabla de distribución de valores faltantes por variable y por registro (Izquierda). Mapa de calor que muestra la posición de los valores ausentes por registro y variable, donde el color blanco indica la presencia de valores faltantes. (Derecha)

Tras este análisis, se realizaron distintas estrategias de imputación según el tipo de variable y patrón de datos faltantes [7-10]:

1. **Eliminación de registros incompletos:** Se eliminaron observaciones con más del 50% de valores perdidos, afectando únicamente a 431 registros (0,14% del total), sin comprometer la representatividad y la calidad de los datos.
2. **Variables categóricas nominales:** se conservó la categoría “Unknown”, ya que en algunos casos la ausencia de datos es informativa (**Figs. A6-A10**).
3. **Variables categóricas ordinales:** Se aplicaron técnicas de imputación multivariante utilizando “*Iterative Imputer*” con árboles de decisión, que permite aprovechar información de múltiples variables y preservar relaciones estructurales de los datos. Este tipo de algoritmos preservan mejor las relaciones entre variables y las estructuras de los datos [7]. Además, se generaron variables binarias que actúan como indicadores de “*Missing*”, permitiendo a los modelos identificar si la ausencia de datos constituye un factor relevante para la predicción.

### 3. Preprocesado de Datos (Feature Engineering)

El preprocesamiento de datos es una etapa crucial en cualquier proyecto de análisis de datos y modelado predictivo, ya que permite preparar la información de manera que los modelos de *machine learning* puedan extraer insights fiables y precisos. El objetivo principal de esta fase es obtener un conjunto de datos limpio y consistente, listo para ser utilizado por los modelos.

En este proyecto, se aplicaron las siguientes técnicas de preprocesamiento:

- **Limpieza de datos:** Se gestionaron los valores faltantes mediante técnicas de imputación, asegurando la integridad del conjunto de datos y relaciones entre variables.
- **Codificación de variables categóricas:** Se aplicó *OneHotEncoding* para variables nominales, y *OrdinalEncoding* para variables con un orden natural, preservando la jerarquía entre categorías. Esto garantiza compatibilidad con los modelos de *machine learning* y mantiene información sobre la estructura de las variables.
- **Balanceo de clases:** Se evaluaron técnicas de *oversampling* y *undersampling* y se analizó su impacto sobre la predicción mediante un modelo simple basado en árbol. La estrategia híbrida RUS+SMOTE resultó ser la más eficaz.

- **Selección de variables:** Se eligieron las variables más relevantes combinando resultados de técnicas automáticas como RFECV, que selecciona características mediante validación cruzada para optimizar el rendimiento del modelo, y la importancia de características estimada por ExtraTreesClassifier, que mide la contribución de cada variable en la reducción de impureza. Estas decisiones se complementaron con los hallazgos del análisis exploratorio, asegurando que se incluyeran las variables con mayor capacidad discriminativa.

### 3.1. Division Train/Test

Las transformaciones de datos deben realizarse después de dividir los datos en conjuntos de entrenamiento (train) y prueba (test), con el fin de:

- **Evitar la contaminación de datos (data leakage):** Si alguna técnica de transformación accede a información del conjunto de prueba durante el ajuste, el modelo podría aprender patrones que no generalizan a datos nuevos, comprometiendo su rendimiento real.
- **Simular un escenario real:** En la práctica, los datos test nunca están disponibles durante el entrenamiento, por lo que esta división simula condiciones reales de uso del modelo.

Previo a la división de los datos, se definió la **variable objetivo (cancer)** y las **variables explicativas**, que incluyen características clínicas y demográficas, así como variables indicadoras de valores faltantes (*\_NA*), para aquellas columnas con información incompleta:

```
['age', 'race', 'bmi', 'first_degree', 'menarche', 'first_birth', 'birads', 'hrt', 'menopaus', 'bioph',  
'prop_missings', 'menarche_NA', 'first_birth_NA', 'birads_NA', 'bmi_NA']
```

Las variables ordinales se trataron como numéricas para preservar su orden natural. Todas comienzan en cero, excepto *bmi* y *birads*, por lo que se les aplicó la corrección correspondiente para homogeneizar la codificación.

El dataset final contiene **305.914 observaciones y 15 variables explicativas**. Para entrenar y evaluar los modelos de manera objetiva, se dividieron las observaciones en conjuntos de entrenamiento y prueba utilizando el **70% para entrenamiento y el 20% para prueba**. La función `train_test_split` se aplicó con el parámetro `stratify` para asegurar que la proporción de clases en los conjuntos de entrenamiento y prueba reflejara la distribución original del dataset. Esta estratificación garantiza que los modelos se entrenen y evalúen bajo condiciones representativas del dataset original, evitando sesgos y facilitando una estimación confiable de su rendimiento.

### 3.2. Feature Transformation

Algunos modelos de clasificación supervisada empleados en este trabajo (Regresión Logística, Random Forest, XGBoost, LightGBM y Extra Trees) no pueden procesar directamente variables categóricas en formato texto. Por ello, estas variables deben transformarse en valores numéricos mediante técnicas como *One-Hot Encoding* para variables nominales y binarias; y *Ordinal Encoding* para variables con jerarquía natural

Se creó un preprocesador que integra todas las transformaciones necesarias y garantiza un flujo secuencial y reproducible (ver **Fig. B1**, en **Anexo B**). Las etapas del preprocesamiento fueron las siguientes:

1. **Imputación de datos faltantes:** Se aplicó *Iterative Imputer* basado en árboles de decisión (*DecisionTreeClassifier*) para completar los valores faltantes en variables categóricas ordinales. Esta técnica preserva las relaciones entre variables y evita que la información perdida comprometa la integridad del dataset. Para más información sobre esta técnica ver referencias [7, 9-10].
2. **Codificación de variables categóricas [11]:**
  - **Variables nominales y binarias:** se aplicó *One-Hot Encoding*., eliminando la primera categoría de cada variable nominal como referencia. Para las binarias (*\*\_NA*) se generó una columna que indica la presencia de la categoría 1, tomando 0 como referencia. Esto previene multicolinealidad y mantiene la consistencia del modelo.

- **Variables ordinales:** se utilizó *Ordinal Encoding* para preservar la jerarquía natural de los niveles. Se definieron explícitamente los niveles para las variables *age* y *prop\_missings* (de menor a mayor grupo de edad o proporción de valores faltantes), mientras que el resto de las variables ordinales, al estar representados por enteros, fueron interpretados automáticamente por el codificador.

El preprocesador se ajustó primero sobre el conjunto de entrenamiento y luego se aplicó al conjunto de prueba, evitando filtración de información (*data leakage*) y asegurando consistencia en la codificación. Se comprobó la ausencia de valores nulos tras la transformación. Las dimensiones finales se indican en la **Tabla 2**.

Tras la transformación, todas las categorías quedaron representadas como valores numéricos. Aunque algunos modelos modernos, como LightGBM, pueden manejar variables categóricas de forma nativa, se mantiene la codificación numérica para garantizar consistencia en todos los modelos utilizados. De esta manera, se unifica el tratamiento de las variables y se asegura que todos los modelos interpreten los datos de manera coherente.

Conjunto de datos	Variables explicativas	Variable objetivo
Entrenamiento	(244731, 24)	(244731,)
Prueba	(61183, 24)	(61183,)

Tabla 2: Dimensiones finales de los conjuntos preprocesados. La tabla muestra el número de observaciones y variables explicativas en los conjuntos de entrenamiento y prueba, así como el tamaño de la variable objetivo,

### 3.3. Balance de Clases

Una vez finalizado el proceso de preprocesamiento y transformación de las variables, los datos están preparados para el entrenamiento de los modelos de *machine learning*. Sin embargo, es necesario abordar el problema del desbalanceo (ver **Fig. 1**), ya que puede afectar la capacidad del modelo para detectar correctamente los casos positivos (*cancer*). El desbalance es un desafío crítico en clasificación, ya que los modelos tienden a favorecer la clase mayoritaria, dificultando la predicción de la clase minoritaria (cáncer positivo). Esto puede llevar a predicciones sesgadas e imprecisas, reduciendo la utilidad del modelo en escenarios reales [12-17].

Para abordar este problema, se evaluaron distintas estrategias de **balanceo de clases**, combinando técnicas de sobremuestreo (*oversampling*) y submuestreo (*undersampling*):

- **SMOTE(Synthetic Minority Over-sampling Technique):** genera instancias sintéticas de la clase minoritaria interpolando entre observaciones existentes, de manera que los datos sintéticos mantengan la estructura del espacio de características.
- **SMOTEENN y SMOTETomek:** combinan *oversampling* de la clase minoritaria con limpieza de la clase mayoritaria mediante técnicas que eliminan ejemplos ruidosos o conflictivos, reduciendo el solapamiento entre clases. SMOTEEN aplica ENN (*Edited Nearest Neighbors*), que elimina ejemplos conflictivos de ambas clases; SMOTETomek elimina pares de ejemplos cercanos de distintas clases, reduciendo el solapamiento.
- **RUS (RandomUnderSampler):** es la técnica de submuestreo más simple y rápida. Elimina aleatoriamente muestras de la clase mayoritaria para evitar que domine el aprendizaje, preservando información relevante de la clase minoritaria, hasta alcanzar la distribución deseada.
- **Combinaciones de RUS y SMOTE:** se implementaron pipelines específicos que aplican primero RUS y posteriormente SMOTE, o viceversa; variando la proporción de reducción de la clase mayoritaria (40% o 50%) y el número de vecinos en SMOTE ( $k=3$  o 5), para analizar su impacto en el desempeño del modelo.

Para comparar estas técnicas, se utilizó un modelo de Random Forest como clasificador base. La literatura indica que, en datos clínicos desbalanceados, Random Forest suele mantener un buen desempeño, detectando correctamente la clase minoritaria sin comprometer la predicción global [18]. Cada técnica se aplicó únicamente al conjunto de entrenamiento (*train*), evitando

filtración de información (*data leakage*), y se evaluó el rendimiento tanto en *train* como en *test* para detectar posibles indicios de sobreajuste. (**Tabla B1**).

Se seleccionaron varias métricas para evaluar cómo afecta cada técnica de balanceo al rendimiento del modelo.

- **Recall (sensibilidad):** mide la proporción de casos positivos correctamente identificados, crítico en la detección de cáncer.
- **F1-Score:** combina recall y precision, equilibrando detección de positivos y exactitud.
- **PR-AUC (Área bajo la curva Precision-Recall):** especialmente útil en datasets desbalanceados, ya que enfatiza la capacidad del modelo de detectar correctamente la clase minoritaria.

Además, se analizó el desempeño tanto en *train* como en *test* para detectar posibles indicios de sobreajuste (si las diferencias de las métricas son significativas) (**Tabla B2**).

La mayoría de las técnicas de balanceo aproximan la proporción de clases al 50/50, excepto SMOTEENN, que alcanza un 45% para la clase positiva. En muchos estudios clínicos y financieros, ratios entre 60/40 o 65/35 ya se consideran un buen equilibrio entre representatividad de la clase minoritaria y reducción de sesgo [12].

El análisis mostró que SMOTE, SMOTETomek y SMOTEENN presentaban métricas de entrenamiento muy altas en *train* pero un desempeño inferior en *test*, indicando sobreajuste debido al exceso de ejemplos sintéticos y posible ruido [12, 15-17]. Por el contrario, la combinación de RUS y SMOTE ofreció un balance aceptable mejor generalización y detección de la clase minoritaria.

Entre las configuraciones evaluadas, RUS+SMOTE5k5 y SMOTE+RUS presenta mejores métricas de desempeño que el resto, compartiendo métricas similares en *test*. No obstante, aunque RUS+SMOTE5k5 presenta un recall de entrenamiento algo más alto (0,919 vs. 0,875), lo que puede sugerir cierto sobreajuste, resulta preferible frente a SMOTE+RUS, ya que esta última elimina información real tras la generación de ejemplos sintéticos, reduciendo la calidad del dataset. En cambio, **RUS+SMOTE5k5** preserva mejor la diversidad de la información y aporta mayor estabilidad, por lo que **se selecciona como la opción final**. Las dimensiones se muestran en la **Tabla B1**.

Es importante destacar que todas las técnicas de *resampling* se aplicaron únicamente al conjunto de entrenamiento, garantizando que la evaluación en *test* refleje el desempeño real del modelo en datos nuevos, evitando así la filtración de información (*data leakage*)

### 3.4. Selección de Variables

En proyectos de *machine learning*, incluir variables irrelevantes puede provocar sobreajuste y disminuir la generalización del modelo. Por ello, en esta fase se seleccionaron las variables más relevantes para la predicción de cáncer de mama utilizando métodos complementarios.

1. **Selección automática con RFECV.** Se aplicó RFECV (*Recursive Feature Elimination with Cross-Validation*), utilizando de estimador un Random Forest de 100 árboles. Se basa en:
  - **Eliminación recursiva de variables (RFE):** elimina iterativamente las variables menos importantes, entrenando un modelo en cada iteración. Las variables con importancia máxima (importance = 1) son las últimas en eliminarse y forman parte del conjunto óptimo de predictores.
  - **Validación cruzada (CV):** evalúa el rendimiento en distintos subconjuntos de los datos (*folds*), garantizando que la selección de variables no dependa de un único muestreo.

El análisis identificó 7 variables óptimas que maximizan el AUC medio (**Fig. B2**): *prop\_missings, first\_birth, birads, bmi, bioph\_1, first\_degree\_1* y *age*. Las variables eliminadas al final del ranking pueden ser descartadas sin comprometer el rendimiento.

2. **Validación con ExtraTreesClassifier.** Para complementar RFECV, se utilizó *ExtraTreesClassifier*, un algoritmo basado en árboles de decisión que introduce mayor

aleatoriedad al elegir los *splits*, lo que lo hace menos propenso al sobreajuste y permite obtener estimaciones más estables de la importancia de las variables. Las 7 variables identificadas por RFECV también aparecieron en el top de importancia de *ExtraTrees*, confirmando la robustez de la selección anterior (**Fig B3**).

3. **Análisis de redundancia y correlación.** Se revisó la matriz de correlación de las variables finales transformadas y balanceadas para identificar posibles variables redundantes (**Fig. B4**). Esto asegura que la selección de variables sea robusta y no incluya información duplicada que pueda afectar la interpretabilidad o el desempeño.
  - No se encontraron pares con correlación > 0,9.
  - Se observó una correlación moderada entre *age* y *menopaus\_2* (0,67), coherente con la edad promedio de aparición de la menopausia.
  - Correlaciones leves entre *prop\_missings* y variables con *flag\_NA* (0,21–0,55).
4. **Consideraciones clínicas.** El análisis exploratorio de datos señaló que algunas variables con menor impacto predictivo, como *hrt\_1* y *race\_White*, según los modelos de selección probados, presentan diferencias significativas entre grupos de interés y son factores de riesgo reconocidos en la literatura. Se decidió mantenerlas en el modelo para preservar la relevancia clínica y epidemiológica, garantizando que el modelo sea interpretable y útil para la toma de decisiones médicas.

Comparando las métricas obtenidas en los diferentes métodos de selección (**Tabla B3**) y considerando la relevancia clínica identificada en el análisis exploratorio, así como las correlaciones entre variables, se seleccionaron 10 variables clave (**Tabla 3**) para el modelo para un total de 230180 registros.

Variables Explicativas Seleccionadas				
bioph_1	age	first_birth	birads	bmi
menarche	first_degree_1	prop_missings,	hrt_1	race_White

Tabla 3: Variables explicativas seleccionadas para el modelo final.

Antes de iniciar la siguiente fase, se transformó la variable objetivo a valores enteros (0 y 1). Aunque las técnicas de balanceo interpretan mejor las clases en formato categórico, los modelos de scikit-learn utilizados (Random Forest, XGBoost, LightGBM) requieren que la variable objetivo sea numérica para calcular correctamente probabilidades y métricas como el AUC.

## 4. Modelos Predictivos de Machine Learning

En esta fase del proyecto se desarrollan y evalúan modelos que permitan cumplir con los objetivos del proyecto. El problema abordado es de **clasificación**, ya que el objetivo consiste en asignar a cada observación una etiqueta o categoría específica, o estimar la probabilidad de pertenencia a dicha categoría, en lugar de predecir un valor continuo. En este caso, el problema consiste en **predecir el riesgo de cáncer de mama**.

Para ello, se entrenaron y compararon distintos algoritmos para identificar aquel que ofrezca el mejor rendimiento y se ajuste mejor al problema planteado. Los modelos evaluados incluyen:

- **Regresión logística (LR)**, captura patrones lineales entre las características y la objetivo.
- **Random Forest (RF)**, permite identificar interacciones y relaciones no lineales mediante la agregación de múltiples árboles (*bagging*).
- **XGBoost (XGB)**, modelo de *boosting* que optimiza el error de predicción de manera secuencial, capturando relaciones complejas entre variables.
- **LightGBM (LGB)**, modelo de *boosting* eficiente y estable, especialmente adecuado para grandes volúmenes de datos y variables heterogéneas.
- **Stacking (STCK)**, modelo ensamblado que combina los anteriores, aprovechando que cada modelo base aporta patrones distintos para mejorar la capacidad predictiva global.

En el caso del modelo stacking, se eligieron modelos base diversos: LR captura patrones lineales, RF identifica interacciones y no linealidades mediante bagging, y LGB gestiona relaciones complejas de manera eficiente. Esta diversidad permite que el modelo ensamblado mejore la predicción combinando fortalezas de cada algoritmo.

Para cada algoritmo, se realizó una búsqueda de hiperparámetros (ver **Anexo C**) con el objetivo de optimizar su rendimiento y garantizar que los modelos seleccionados reflejen la mejor combinación posible entre métricas, capacidad de generalización y robustez.

Dado el tamaño del dataset, se descartó la búsqueda de hiperparámetros para **SVM**, ya que el coste computacional habría sido elevado sin mejoras significativas en comparación con los modelos de *boosting*. De manera similar, se consideraron redes neuronales tipo **MLP**, pero al disponer de solo 10 características, no se esperaban mejoras respecto a modelos clásicos de *machine learning*, ya que la capacidad de las redes para aprender se vería limitada.

#### 4.1. Modelos Utilizados

##### Regresión Logística

La regresión logística (LR) es un modelo de clasificación supervisada ampliamente utilizado en contextos clínicos debido a su capacidad para predecir la probabilidad de un evento binario, como la presencia o ausencia de una enfermedad. En este proyecto, LR se emplea como modelo base, como referencia que permita comparar el desempeño de modelos más complejos.

Ventajas	Desventajas
<b>Interpretabilidad:</b> Los coeficientes permiten cuantificar directamente el efecto de cada variable en la probabilidad del evento, facilitando la interpretación clínica.	<b>Linealidad en la relación:</b> Asume que la relación entre variables predictoras y log-odds del evento es lineal; relaciones no lineales complejas pueden no ser capturadas.
<b>Simplicidad y robustez:</b> fácil de implementar y bajo coste computacional.	<b>Sensibilidad a variables irrelevantes:</b> La inclusión de muchas variables irrelevantes puede reducir la precisión del modelo.
<b>Manejo de desbalance moderado:</b> Con técnicas de <i>resampling</i> o ponderación, puede adaptarse a datasets desbalanceados.	<b>Limitaciones con grandes datasets y alta dimensionalidad</b>

##### Random Forest

Random Forest (RF) es un algoritmo basado en ensambles de árboles de decisión, desarrollado por Leo Breiman en 2001. Cada árbol se entrena de manera independiente y aleatoria, y sus predicciones se combinan mediante votación (para clasificación), reduciendo el riesgo de sobreajuste y mejorando la capacidad de generalización. En contextos clínicos, RF ha demostrado un buen desempeño en la predicción de eventos raros o desbalanceados, [18].

Ventajas	Desventajas
<b>Captura relaciones no lineales</b> y datos desbalanceados.	<b>Menor interpretabilidad</b> que un árbol único o una RL.
<b>Robusto</b> ante valores atípicos y ruido	<b>Computacionalmente costoso</b> para conjuntos de datos muy grandes
<b>Reduce el sobreajuste</b> mediante agregación de múltiples árboles.	
<b>Manejo de datos faltantes y variables mixtas</b> (categóricas y continuas) sin necesidad de escalado.	

## XGBoost

XGBoost (*Extreme Gradient Boosting*) es una técnica de *boosting* basada en árboles de decisión, popular por su alto rendimiento en competiciones de *Machine Learning* y su eficiencia computacional. Desarrollado por Tianqi Chen. Optimiza el método de *gradient boosting* al incorporar técnicas avanzadas que mejoran tanto la velocidad como la precisión del modelo. XGBoost construye árboles de forma secuencial, donde cada nuevo árbol corrige los errores residuales del modelo anterior. Utiliza un enfoque de optimización basado en el descenso del gradiente para ajustar los árboles, lo que permite mejorar el rendimiento del modelo iterativamente.

Ventajas	Desventajas
<p><b>Rendimiento Superior:</b> combina boosting y regularización para maximizar métricas</p> <p><b>Capacidad para modelar relaciones no lineales</b> y valores faltantes.</p> <p><b>Regularización integrada L1 y L2</b> para controlar el sobreajuste.</p> <p><b>Eficiencia:</b> Es rápido en el entrenamiento y la predicción, en grandes datasets y alta dimensionalidad.</p>	<p><b>Interpretabilidad limitada:</b> más difícil de interpretar que modelos lineales o árboles individuales.</p> <p><b>Sensibilidad a hiperparámetros:</b> requiere ajuste cuidadoso de parámetros.</p> <p><b>Computacionalmente costoso:</b> Aunque es eficiente, puede ser más lento que modelos simples en datasets muy grandes si no se optimiza.</p>

## LightGBM

LightGBM (*Light Gradient Boosting Machine*) es un algoritmo de boosting desarrollado por Microsoft, diseñado para manejar grandes volúmenes de datos de manera eficiente y con alta precisión. Construye árboles de manera secuencial para corregir los errores residuales y aplica el descenso del gradiente de forma iterativa para mejorar el modelo. LightGBM introduce innovaciones como el *crecimiento leaf-wise* y el uso de histogramas para dividir los datos, lo que le permite entrenar modelos precisos con menor tiempo de cómputo y menor consumo de memoria. En contextos clínicos, LightGBM ha mostrado niveles de precisión y AUC competitivos o superiores a otros algoritmos de clasificación.

Ventajas	Desventajas
<p><b>Optimizada</b> para manejar grandes volúmenes de datos y alta dimensionalidad.</p> <p><b>Flexibilidad:</b> Soporta tanto variables numéricas como categóricas de manera nativa.</p> <p><b>Regularización avanzada:</b> Dispone de parámetros que reducen el riesgo de sobreajuste incluso en modelos profundos.</p>	<p><b>Sensibilidad a Hiperparámetros:</b> Requiere una cuidadosa configuración de hiperparámetros.</p> <p><b>Menor Interpretabilidad:</b> Similar a otros métodos de boosting, puede ser menos interpretable que modelos basados en árboles individuales, complicando la interpretación de sus predicciones.</p>

## Técnica de Stacking

Stacking (STCK) es una técnica de *ensemble learning* que combina varios modelos base (modelos de primer nivel) con el mismo conjunto de datos de entrenamiento para mejorar la capacidad predictiva final. Luego usa sus predicciones como nuevas características para entrenar un meta-modelo (o modelo de segundo nivel). Este meta-modelo aprende a combinar las predicciones de los modelos base de manera óptima, buscando maximizar el rendimiento global del sistema. Aunque el stacking no garantiza siempre la mejor predicción, generalmente mejora las métricas globales frente a modelos individuales, especialmente si los modelos base son diversos y complementarios.

Ventajas	Desventajas
<b>Mejora del rendimiento:</b> combina modelos para capturar distintos patrones que un único modelo y reducir errores individuales.	<b>Complejidad computacional:</b> entrenar múltiples modelos base + meta-modelo puede ser costoso en tiempo y memoria.
<b>Reducción de sesgo y varianza:</b> corrige errores sistemáticos de modelos simples	<b>Difícil de interpretar:</b> al combinar varios modelos, la interpretabilidad disminuye.
<b>Aprovechar fortalezas de distintos modelos:</b> El meta-modelo aprende de las capacidades de los modelos base, capturando patrones complejos que podrían escapar a modelos individuales.	<b>Riesgo de sobreajuste:</b> si el meta-modelo se entrena con predicciones del mismo conjunto de entrenamiento sin validación cruzada.
<b>Flexibilidad:</b> Se pueden combinar modelos muy distintos, permitiendo aprovechar fortalezas específicas de cada uno.	<b>Dependencia de diversidad:</b> si los modelos base son similares, el stacking no aporta beneficio y puede incluso empeorar la predicción

## 4.2. Métricas de Evaluación

La elección de las métricas, tanto para la búsqueda de hiperparámetros como para la evaluación final, se fundamentó en la naturaleza del problema: la predicción de una enfermedad. En datasets desbalanceados, donde la prevalencia de la enfermedad es mucho menor que la de los casos sanos, la métrica *accuracy* puede resultar engañosa. Un modelo que prediga siempre la clase mayoritaria podría alcanzar una alta *accuracy*, pero sería inútil para identificar pacientes enfermos. Por ello, se seleccionaron métricas más informativas y adecuadas para este contexto clínico:

- **Recall (Sensibilidad):** Mide la capacidad del modelo para identificar correctamente los casos positivos (pacientes con la enfermedad). Minimizar los falsos negativos es crucial, ya que un diagnóstico tardío puede tener consecuencias graves. Maximizar el recall garantiza que la mayor cantidad posible de pacientes con cáncer de mama sea detectada.
- **ROC AUC (Área bajo la curva ROC):** Evalúa la capacidad del modelo para distinguir entre pacientes sanos y enfermos a través de distintos umbrales de decisión. Representa la relación entre la tasa de verdaderos positivos (TPR) y la tasa de falsos positivos (FPR) para diferentes puntos de corte. A diferencia de la *accuracy*, el AUC es robusto frente al desbalance de clases y proporciona una medida de discriminación global; un valor cercano a 1 indica una excelente separación entre clases, mientras que un valor de 0,5 corresponde a un modelo sin capacidad predictiva.

Durante la búsqueda de hiperparámetros, se emplearon ambas métricas de manera complementaria:

- **ROC AUC** como criterio principal de *refit*, asegurando buena capacidad de discriminación general
- **Recall** como métrica adicional, priorizando la detección de pacientes positivos.

Para garantizar una evaluación justa y reproducible, se aplicó un enfoque sistemático:

1. **Validación cruzada estratificada de 5 pliegues (StratifiedKFold):** aplicada sobre los datos de entrenamiento para cada arquitectura durante la búsqueda de hiperparámetros. Esta técnica preserva la proporción de clases en cada partición, generando estimaciones más estables y representativas.
2. **Comparación de métricas entre train y test:** se analizaron recall y AUC de los modelos generados en la búsqueda hiperparamétrica para detectar posibles indicios de sobreajuste.
3. **Evaluación de la estabilidad:** mediante boxplots de los modelos candidatos seleccionados; cajas estrechas reflejan menor variabilidad y mayor consistencia del modelo.
4. **Evaluación final del modelo óptimo:** una vez seleccionado, se calcularon métricas adicionales como recall, AUC, matriz de confusión y curva ROC.

## Umbral de Decisión

Los modelos de clasificación generan probabilidades de pertenencia a cada clase, que deben convertirse en etiquetas binarias aplicando un umbral de decisión (por defecto, 0,5). Sin embargo, este valor puede ajustarse para optimizar el equilibrio entre sensibilidad y especificidad.

Una estrategia consiste en utilizar el **índice de Youden ( $J = TPR - FPR$ )**, que identifica el punto de la curva ROC que maximiza la clasificación correcta de los sujetos. El umbral asociado a este índice ha sido ampliamente utilizado en contextos estadísticos y clínicos, al proporcionar un balance adecuado entre detección de casos positivos y control de falsos positivos [19].

### 4.3. Búsqueda de Hiperparámetros para Cada Modelo

En esta fase se optimizó la configuración interna de los modelos de *machine learning* mediante la búsqueda de hiperparámetros, es decir, aquellos valores que no se aprenden automáticamente a partir de los datos y que determinan el comportamiento del algoritmo. Por ejemplo, en un RF, los hiperparámetros incluyen el número de árboles, la profundidad máxima de cada árbol, etc.

La búsqueda de hiperparámetros consiste en explorar diferentes combinaciones de estos valores con el fin de identificar la configuración que maximiza el rendimiento del modelo. En este proyecto, el objetivo fue:

- Priorizar la detección de pacientes positivos (recall).
- Mantener una alta capacidad de discriminación global (ROC AUC).
- Evitar sobreajuste, asegurando que las mejoras en *train* se reflejaran también en *test*.

Para ello, se evaluaron distintas combinaciones de hiperparámetros de forma sistemática. Los detalles completos de la evaluación se incluyen en el **Anexo C**, mientras que los modelos óptimos y candidatos a modelo ganador se recogen en la **Tabla 4**.

<b>LR</b>	<code>LogisticRegression( C=1, penalty="l2", solver= "lbfgs", random_state=seed, class_weight='balanced')</code>
<b>RF</b>	<code>RandomForestClassifier(bootstrap=True, class_weight="balanced", max_depth=5, n_estimators=100, max_features=0.6, max_samples=0.7, min_samples_leaf=10, random_state=seed)</code>
<b>XGB</b>	<code>XGBClassifier(colsample_bytree=0.8, eta=0.01, gamma=0.2, max_depth=5, min_child_weight=10, subsample= 0.8, n_estimators=100, random_state=seed)</code>
<b>LGBM</b>	<code>LGBMClassifier(boosting_type='gbdt', colsample_bytree=1.0, max_depth=7, learning_rate=0.01, min_child_samples=15, n_estimators=100, reg_alpha = 0.1, subsample=0.6, random_state=seed, verbose=-1, n_jobs=-1, class_weight='balanced')</code>
<b>STCK</b>	<code>modelos_base_stacking = [ ('RF', modelo_RF), ('LR', modelo_LR), ('LGB', modelo_LGB) ]</code> <code>StackingClassifier(estimators=modelos_base_stacking, final_estimator=LogisticRegression(C=0.005, solver='saga', class_weight='balanced', random_state=seed), passthrough=False)</code>

Tabla 4: Configuraciones finales de los modelos seleccionados para predicción de riesgo de cáncer de mama. Se incluyen los hiperparámetros optimizados de cada modelo individual —Logistic Regression (LR), Random Forest (RF), XGBoost (XGB) y LightGBM (LGBM)— así como la composición del modelo de Stacking (STCK) que combina los tres modelos base con una regresión logística como metamodelo final.

## 5. Comparación y Resultados Finales

### 5.1. Evaluación y Comparación de los Modelos

En esta fase se evalúa el grado de acercamiento del modelo a los objetivos, utilizando métricas de rendimiento sobre un conjunto de prueba para cada arquitectura óptima (**Tabla 4**). Los resultados completos, incluyendo curvas ROC, matrices de confusión e informes de clasificación con dos umbrales de decisión distintos (0.5, **Fig. D1** y el umbral definido por el índice de Youden, **Fig. D2**, se presentan en el **Anexo D**. La **Tabla 5** resume las métricas clave, destacando el recall de la clase positiva y el AUC para cada umbral.

Modelos	LR	RF	XGB	LGB	STCK
Recall (0.5)	0.7068	<b>0.7680</b>	0.7537	0.7397	0.7359
AUC (0.5)	0.7473	0.7491	0.7535	0.7528	0.7523
Recall (Youden)	0.7537	<b>0.8070</b>	0.7818	0.7979	0.7792
AUC (Youden)	0.7473	0.7491	0.7535	0.7528	0.7523

Tabla 5: Rendimiento de los modelos seleccionados en términos de Recall y AUC, evaluado con dos umbrales: 0.5 y el umbral óptimo de Youden. Se muestran los valores de cada métrica para los modelos individuales (LR, RF, XGB, LGB) y el modelo de Stacking (STCK).

Para complementar esta evaluación puntual, se realizó **validación cruzada de 5 folds**. Para cada modelo se calcularon las métricas de AUC y recall en cada partición, representándose los resultados mediante boxplots. Esta visualización permite analizar la distribución y variabilidad del rendimiento, así como la capacidad de discriminación (AUC) y sensibilidad (recall) de cada modelo (**Fig. 3**).

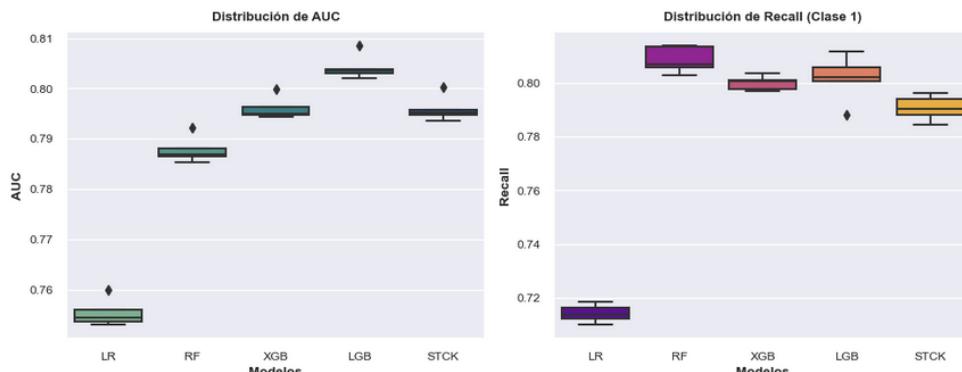


Figura 3: Boxplots de AUC y Recall obtenidos mediante validación cruzada de 5 folds para los modelos comparados.

Los resultados generales muestran que:

- Los modelos finales alcanzaron AUC similares ( $\approx 0.75$ ), con ligera ventaja de XGBoost.
- RF mostró el mejor rendimiento en términos de recall, crítico en un contexto médico.
- En todos los casos, los modelos tienden a **incrementar falsos positivos para capturar un mayor número de positivos**, sacrificando precisión en la clase negativa.

Comparación por modelo:

- **RF**: alcanzó el recall más alto, identificando de manera más eficaz los casos de riesgo.
- **LGB**: obtuvo el mejor AUC, indicando una mayor capacidad de discriminación entre clases.
- **XGB**: presentó un buen equilibrio entre AUC y recall, con un rendimiento consistente.
- **LR**: Rendimiento inferior frente a los modelos basados en árboles, especialmente en recall.
- **STCK**: no aportó mejoras significativas respecto a los modelos individuales, por lo que su complejidad adicional no se justifica en este caso.

## 5.2. Evaluación de la Estabilidad y Robustez de los Modelos

Para evaluar la **estabilidad y robustez**, se variaron las semillas aleatorias (*random\_state*) utilizadas tanto en la partición de los datos como en el proceso de entrenamiento. Este procedimiento permite analizar si los resultados dependen de una configuración específica o, por el contrario, se mantienen consistentes bajo diferentes inicializaciones.

Las métricas observadas mostraron variaciones mínimas entre semillas, lo que confirma que los modelos son robustos frente a la aleatoriedad y las conclusiones son fiables y reproducibles (ver **Figs. D3-D5** para detalles).

## 5.3. Elección del mejor modelo

Tras la evaluación de los distintos modelos candidatos en las anteriores secciones, se concluye que los algoritmos basados en árboles (RF, XGB y LGB) superan de forma consistente a la regresión logística en términos de recall y AUC.

Entre ellos, RF alcanza el mayor recall de la clase positiva, lo que lo convierte en el modelo más eficaz para identificar correctamente los casos relevantes. Por su parte, LGB presenta la mayor AUC, mostrando una capacidad ligeramente superior para discriminar entre clases.

En un contexto médico, el recall es la métrica prioritaria, ya que resulta más crítico identificar todos los casos positivos que minimizar los falsos positivos. Por ello, aunque LGB muestre un AUC superior, la elección se inclina hacia RF. Además, RF presenta ventajas adicionales: mayor interpretabilidad (permite analizar la importancia de las variables y visualizar árboles de decisión), simplicidad de implementación y robustez frente a variaciones aleatorias.

En conclusión, **Random Forest** se selecciona como modelo principal, priorizando la detección de casos positivos en el contexto clínico. LightGBM se mantiene como una alternativa válida en escenarios donde se busque maximizar la discriminación global entre clases.

## 6. Interpretación del Modelo Seleccionado

Para garantizar la transparencia y la confianza en el modelo seleccionado (Random Forest), se aplicó la técnica de SHAP (*SHapley Additive exPlanations*) con el fin de interpretar las predicciones. SHAP se basa en los valores de Shapley de la teoría de juegos y permite cuantificar la contribución de cada variable a la predicción final del modelo. Esta metodología resulta especialmente útil en contextos clínicos, pues ofrece explicaciones tanto a nivel global como local de manera intuitiva:

- **Interpretabilidad global:** muestra la importancia promedio de cada variable en todas las predicciones. De esta forma, es posible identificar qué factores influyen más en el riesgo de cáncer de mama en términos generales. (**Figs. E1-E3**)
- **Interpretabilidad local:** se centra en casos individuales, detallando cómo cada variable concreta ha influido en la predicción de un paciente específico. (**Fig. E4**)

Según las **Fig. E1** y **E2**, la variable más relevante es **bioph**, que explica aproximadamente el 50% de la variabilidad del modelo. Le siguen **edad** (24 %), **raza blanca** (11 %), **terapia hormonal** (7 %) y **proporción de datos ausentes** (6 %). Entre *bioph*, edad y raza blanca se explica el 84 % del comportamiento del modelo, mientras que las nueve primeras variables explican el 100 %.

La **Fig. E3** permiten observar cómo las categorías de las características se relacionan con un mayor riesgo de cáncer de mama. De acuerdo con este gráfico, el modelo identifica como factores de riesgo más relevantes:

- Haber tenido una **biopsia previa**.
- Tener una **edad igual o superior a 45 años** (grupo 5 en adelante).
- Pertenecer al **grupo de raza blanca**.
- **No haber recibido terapia hormonal**.

Cabe destacar que estos hallazgos son **consistentes con las conclusiones del análisis exploratorio de datos (EDA)**, lo que refuerza la coherencia y robustez del modelo.

## 7. Despliegue del Modelo (Puesta en Producción)

El objetivo de esta fase es poner a disposición de los usuarios finales los resultados del modelo predictivo y garantizar su correcto mantenimiento una vez implementado. Tras el desarrollo, entrenamiento y validación del modelo predictivo, se procedió a su implementación en producción mediante una aplicación web interactiva desarrollada con *Streamlit*. Esta implementación transforma el modelo de un prototipo experimental en una herramienta práctica y confiable, lista para su evaluación en entornos clínicos o de investigación.

La aplicación facilita la toma de decisiones individualizadas sobre el riesgo de cáncer de mama y permite generar datos anónimos que contribuyan a mejorar futuras versiones del modelo.

Los pasos metodológicos seguidos fueron los siguientes:

- **Serialización del modelo y preprocesador:** El modelo entrenado y el preprocesador de datos se almacenaron en archivos binarios (*pickle*), asegurando que las transformaciones de las variables y las predicciones sean consistentes con el entrenamiento original.
- **Gestión de datos de pacientes:** Los datos introducidos se guardan de forma anónima en un archivo CSV, creando una base que permite recalibrar y mejorar el modelo en futuras versiones, garantizando privacidad y confidencialidad.
- **Diseño de la interfaz de usuario:** La aplicación permite introducir los datos de cada paciente mediante selectores desplegables, categorizando adecuadamente las variables ordinales y nominales según las reglas definidas en el conjunto de entrenamiento.
- **Predicción en tiempo real:** Cada entrada se transforma automáticamente con el preprocesador y se evalúa mediante el modelo, generando tanto la predicción de riesgo como la probabilidad asociada. Se implementó un umbral óptimo basado en el índice de Youden, asegurando un balance adecuado entre sensibilidad y especificidad.
- **Interpretabilidad del modelo:** Se incorporaron herramientas basadas en SHAP, que permiten visualizar la contribución de las principales variables a cada predicción, aumentando la transparencia y la confianza en los resultados.
- **Presentación y experiencia de usuario:** La aplicación se estructuró con un título descriptivo, subtítulo explicativo y elementos gráficos (como el logo institucional), buscando claridad y profesionalidad en la interacción con el usuario final.

El código que da lugar a este proyecto y el código de la app junto a las instrucciones de ejecución, se pueden consultar en el [Anexo F](#).

## 8. Conclusiones y Perspectivas Futuras

En este trabajo se desarrolló un modelo predictivo para estimar el riesgo de cáncer de mama a partir de datos clínicos y demográficos del dataset "BCSC Risk Factors". Los resultados muestran que es posible identificar patrones de riesgo relevantes mediante técnicas de *machine learning* utilizando únicamente estas variables. Este enfoque puede contribuir a la personalización de la prevención, permitiendo, por ejemplo, ajustar la frecuencia de cribado según el riesgo individual: priorizando a pacientes con mayor riesgo y espaciando controles en aquellos con riesgo reducido.

### Conclusiones Principales:

- Las variables con mayor capacidad predictiva fueron: **edad, raza blanca, terapia hormonal previa, antecedentes de biopsia, índice de masa corporal, historial familiar de primer grado y densidad mamaria**. Este análisis fue corroborado mediante técnicas de explicación de modelos, como SHAP, que confirmaron la relevancia de estos factores.

- La estrategia más eficaz para balancear los datos fue la **combinación de RUS (Random UnderSampler) y SMOTE**, superando a otras alternativas.
- Entre todos los modelos evaluados, **Random Forest** obtuvo el mejor desempeño global en términos de recall y AUC para este dataset.
- La implementación de un **umbral de decisión basado en el índice de Youden** permitió mejorar el **recall hasta 0.81** y el **AUC hasta 0.75**, lo que refuerza la utilidad de ajustar dinámicamente el umbral en contextos clínicos.

#### Limitaciones:

- **Dependencia de la calidad y representatividad del dataset:** los datos provienen mayoritariamente de población estadounidense, lo que podría limitar la generalización a otros contextos demográficos.
- **Gran presencia de valores desconocidos** en variables explicativas (por ejemplo, menarquia), que podría introducir ruido o sesgos residuales y afectar la predicción.
- **Limitación en el número de variables.** Las variables explicativas son muy limitadas para una correcta predicción. Incorporar datos genéticos y biomarcadores podría mejorar la capacidad predictiva.
- **Limitaciones inherentes a Random Forest:** aunque robusto, es menos interpretable que modelos lineales. Hay que tener cuidado porque RF puede estar sesgado hacia la clase mayoritaria en problemas de clasificación desbalanceada, como pueden ser datos clínicos. Para predicciones en tiempo real o en dispositivos con recursos limitados, un RF muy grande puede ser más lento que modelos más simples.

#### Aplicaciones prácticas

- **Soporte a la toma de decisiones médicas:** el modelo puede ayudar a identificar pacientes con riesgo elevado, priorizando cribados y estrategias de prevención.
- **Investigación clínica:** permite analizar patrones de riesgo y explorar relaciones entre variables clínicas y riesgo de cáncer, contribuyendo a estudios epidemiológicos y de poblaciones, así como tratamientos personalizados.
- **Educación y comunicación con pacientes:** visualizaciones SHAP facilitan explicar los factores que más contribuyen al riesgo individual, promoviendo decisiones para mejorar el estilo de vida.
- **Mejora continua del modelo:** los datos de pacientes se registran de forma anónima, generando una base de datos que puede utilizarse para reentrenamiento y refinamiento del modelo en futuras versiones.

#### Perspectiva futura

- **Validación externa** en otras poblaciones independientes y diversas para asegurar robustez y generalización.
- **Integración de variables adicionales**, (ej. genéticas, biomarcadores, hábitos de vida) podría mejorar la predicción y la personalización del riesgo.
- **Mejora continua del modelo** en futuras versiones con los datos adicionales.

## 9. Bibliografía

- [1] Breast Cancer Surveillance Consortium. "Risk Factor Dataset". Accedido en Julio 2025. <https://www.bcrcs-research.org/index.php/datasets/rf/risk-factor-dataset-download>
- [2] P. Haya, "La metodología CRISP-DM en ciencia de datos," *Instituto de Ingeniería del Conocimiento*. 2021. [www.iic.uam.es/innovacion/metodologia-crisp-dm-ciencia-de-datos/](http://www.iic.uam.es/innovacion/metodologia-crisp-dm-ciencia-de-datos/)
- [3] M. F. Kabir and S. Ludwig, "Classification of Breast Cancer Risk Factors Using Several Resampling Approaches," *2018 17th IEEE International Conference on Machine Learning and Applications (ICMLA)*, 2018, pp. 1243-1248, [doi:10.1109/ICMLA.2018.00202](https://doi.org/10.1109/ICMLA.2018.00202)
- [4] S. B. Manir y P. Deshpande, "Critical Risk Assessment, Diagnosis, and Survival Analysis of Breast Cancer," *Diagnostics*, vol. 14, no. 10, p. 984, 2024. [doi: 10.3390/diagnostics14100984](https://doi.org/10.3390/diagnostics14100984)
- [5] Breast Cancer Surveillance Consortium. *Risk Factor Dataset Documentation*. Recuperado de: <https://www.bcrcs-research.org/idx.php/datasets/rf/documentation>
- [6] Y. Cohen y D. Grabois, "Choosing the appropriate correlation coefficient," *Medium*, 27 de enero de 2022. <https://medium.com/@vatvenger/choosing-the-appropriate-correlation-coefficient>
- [7] S. Alam, M. S. Ayub, S. Arora y M. A. Khan, "An investigation of the imputation techniques for missing values in ordinal data enhancing clustering and classification analysis validity," *Decision Analytics Journal*, vol. 9, p. 100341, 2023. [doi:10.1016/j.dajour.2023.100341](https://doi.org/10.1016/j.dajour.2023.100341)
- [8] S. Paghdal, "Handling Missing Values in Categorical Data," *Medium*, 5 de febrero de 2025. <https://medium.com/@paghdalsneh/handling-missing-values-in-categorical-data>
- [9] A. Mazraeh, "Certainly: A Comprehensive Guide to Handling Missing Data in Numerical Datasets," *Medium*, (2025). <https://medium.com/@adnan.mazraeh1993/certainly>
- [10] S. Chowdhury, "The Art of Feature Engineering: A Guide to Handling Missing Values," *Medium*, 25 de octubre de 2024 <https://suparnachowdhury.medium.com/the-art-of-feature-engineering-handling-missing-values>
- [11] S. Subha, "How to handle categorical features," *Medium*, 8 de febrero de 2024. <https://medium.com/@pingsubhak/how-to-handle-categorical-features>
- [12] V. Kumar, G. S. Lalotra, P. Sasikala, D. S. Rajput, R. Kaluri, K. Lakshmann, M. Shoruzzaman, A. Alsufyani y M. Uddin, "Addressing Binary Classification over Class Imbalanced Clinical Datasets Using Computationally Intelligent Techniques," *Healthcare (Basel)*, vol. 10, no. 7, p. 1293, 13 de julio de 2022. [doi.org: 10.3390/healthcare10071293](https://doi.org/10.3390/healthcare10071293)
- [13] M. F. Kabir y S. Ludwig, "Classification of Breast Cancer Risk Factors Using Several Resampling Approaches," *2018 17th IEEE International Conference on Machine Learning and Applications (ICMLA)*, 2018, pp. 1243-1248, [doi: 10.1109/ICMLA.2018.00202](https://doi.org/10.1109/ICMLA.2018.00202)
- [14] R. Gupta, R. Bhargava y M. Jayabalan, "Diagnosis of Breast Cancer on Imbalanced Dataset Using Various Sampling Techniques and Machine Learning Models," *2021 14th International Conference on Developments in eSystems Engineering (DeSE)*, Sharjah, United Arab Emirates, 2021, pp. 162-167, [doi: 10.1109/DeSE54285.2021.9719398](https://doi.org/10.1109/DeSE54285.2021.9719398).
- [15] S. B. Manir y P. Deshpande, "Critical Risk Assessment, Diagnosis, and Survival Analysis of Breast Cancer," *Diagnostics*, vol. 14, no. 10, p. 984, 2024. [doi.org: 10.3390/diagnostics14100984](https://doi.org/10.3390/diagnostics14100984)
- [16] F. Iqey, "Combining Oversampling and Undersampling for Imbalanced Classification: SMOTE + Tomek and SMOTE + ENN," *Medium*, 20 de agosto de 2024. <https://fiquey.medium.com/combining-oversampling-and-undersampling-for-imbalanced-classification-smote-tomek-and-smote-enn>
- [17] N. Appaji, "Balancing Act: Mastering Imbalanced Data with SMOTE and Tomek Link Strategies," *Medium*, 15 de julio de 2024. <https://niranjanappaji.medium.com/balancing-act-mastering-imbalanced-data-with-smote-and-tomek-link-strategies>
- [18] J. Song, Y. Gao, P. Yin, Y. Li, Y. Li, J. Zhang, Q. Su, X. Fu y H. Pi, "The Random Forest Model Has the Best Accuracy Among the Four Pressure Ulcer Prediction Models Using Machine Learning Algorithms," *Risk Management and Healthcare Policy*, vol. 14, pp. 1175-1187, 2021. [doi.org:10.2147/RMHP.S297838](https://doi.org/10.2147/RMHP.S297838)
- [19] M. Hassanzad y K. Hajian-Tilaki, "Methods of determining optimal cut-point of diagnostic biomarkers with application of clinical data in ROC analysis: an update review," *BMC Medical Research Methodology*, vol. 24, p. 84, 2024. <https://doi.org/10.1186/s12874-024-02198-2>

## Anexo A: Análisis Exploratorio de los Datos (EDA)

### ■ EDA: Variables Numéricas

	<b>count</b>	<b>mean</b>	<b>std</b>	<b>min</b>	<b>25%</b>	<b>50%</b>	<b>75%</b>	<b>max</b>	<b>Asimetría</b>	<b>Kurtosis</b>	<b>Rango</b>
<b>count</b>	1191438.0	4.794887	20.373575	1.0	1.0	1.0	3.0	2684.0	37.460536	2891.647015	2683.0
<b>year</b>	1191438.0	2010.780718	3.660390	2005.0	2008.0	2011.0	2014.0	2017.0	0.052841	-1.170511	12.0

Tabla A1: Estadísticos descriptivos de las variables numéricas “count” y “year”, incluyendo media, desviación estándar, mínimo, máximo, cuartiles, así como medidas adicionales de asimetría, curtosis y rango.

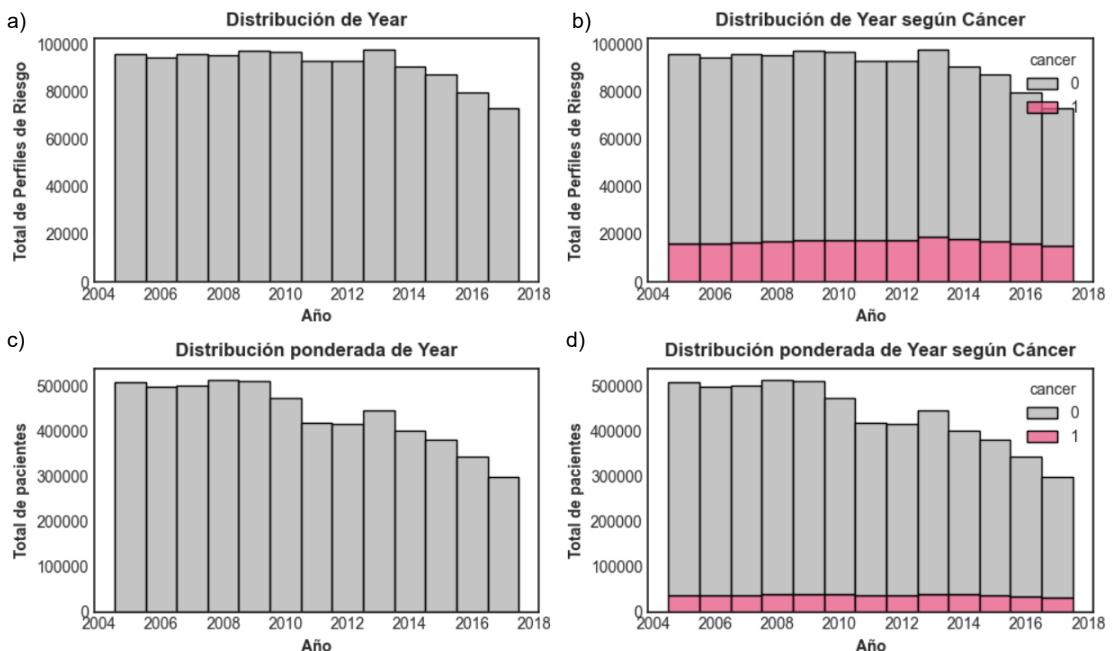


Figura A1: Distribución de la variable “year” en la base de datos: (a) distribución simple de perfiles de riesgo por año, (b) distribución de perfiles de riesgo por año segmentada según la variable objetivo “cáncer”, (c) distribución ponderada por el número de pacientes (“count”) por año y (d) distribución ponderada por pacientes segmentada por “cáncer”.

	<b>count</b>	<b>mean</b>	<b>std</b>	<b>min</b>	<b>25%</b>	<b>50%</b>	<b>75%</b>	<b>max</b>	<b>Asimetría</b>	<b>Kurtosis</b>	<b>Rango</b>
<b>count</b>	306345.0	18.648292	124.00914	1.0	1.0	2.0	8.0	15973.0	42.865351	3509.852782	15972.0

Tabla A2: Estadísticos descriptivos finales de la variable “count”, tras eliminar la variable “year” y reconstruir los perfiles sumando la columna count.

## Predicción de Cáncer de Mama Basado en Factores Clínicos y Demográficos

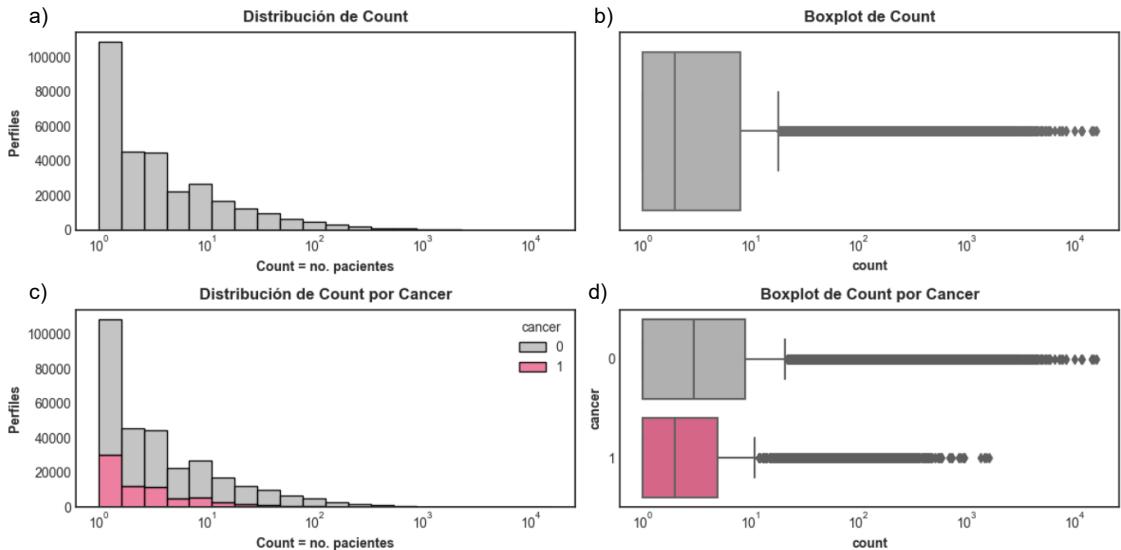


Figura A2: Distribución de la variable “count” en el dataset depurado sin “year”: (a) histograma de la variable “count”, mostrando el número de perfiles por cantidad de pacientes; (b) boxplot de “count”; (c) histograma apilado de “count” segmentado por la variable objetivo “cancer”, y (d) boxplot por categoría “cancer.” Visualizaciones es escala logarítmica.

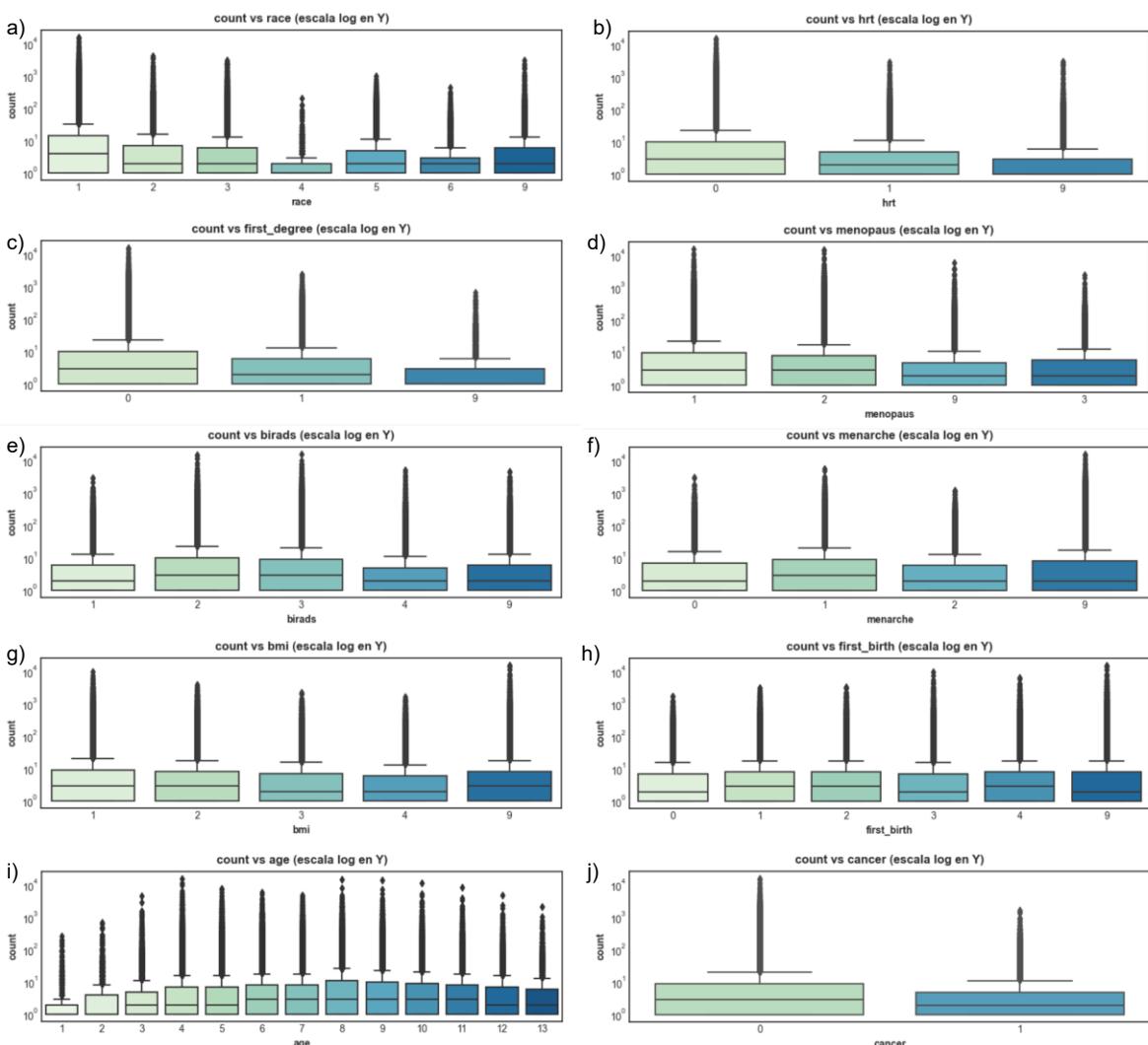


Figura A3: Boxplots de la variable “count” en función de las explicativas: race (a), hrt (b), first\_degree(c), menopaus (d), birads (e), menarche (f), bmi (g), first\_birth (h), age (i) cancer (j). Se utiliza escala logarítmica en el eje Y para representar adecuadamente la dispersión de los valores de pacientes (count) y facilitar la comparación entre categorías.

## ■ EDA: Variables Categóricas Nominales

a)	race	first_degree	hrt	menopaus	bioph	cancer
<b>count</b>	306345	306345	306345	306345	306345	306345
<b>unique</b>	7	3	3	4	3	2
<b>top</b>	1	0	0	2	0	0
<b>freq</b>	118979	190293	223069	175655	162744	234198

b)	race	first_degree	hrt	menopaus	bioph	cancer
<b>count</b>	282989	288798	257154	271512	293467	306345
<b>unique</b>	6	2	2	3	2	2
<b>top</b>	1	0	0	2	0	0
<b>freq</b>	118979	190293	223069	175655	162744	234198

Tabla A3: Resumen estadístico descriptivo de las variables categóricas nominales. Se muestran los resultados: a) Incluyendo todas las categorías, a tal como aparece en los perfiles originales; b) Excluyendo la categoría "9" (desconocido), reemplazada por NaN, para observar la distribución de las categorías conocidas y eliminar el efecto de los valores desconocidos en los perfiles.

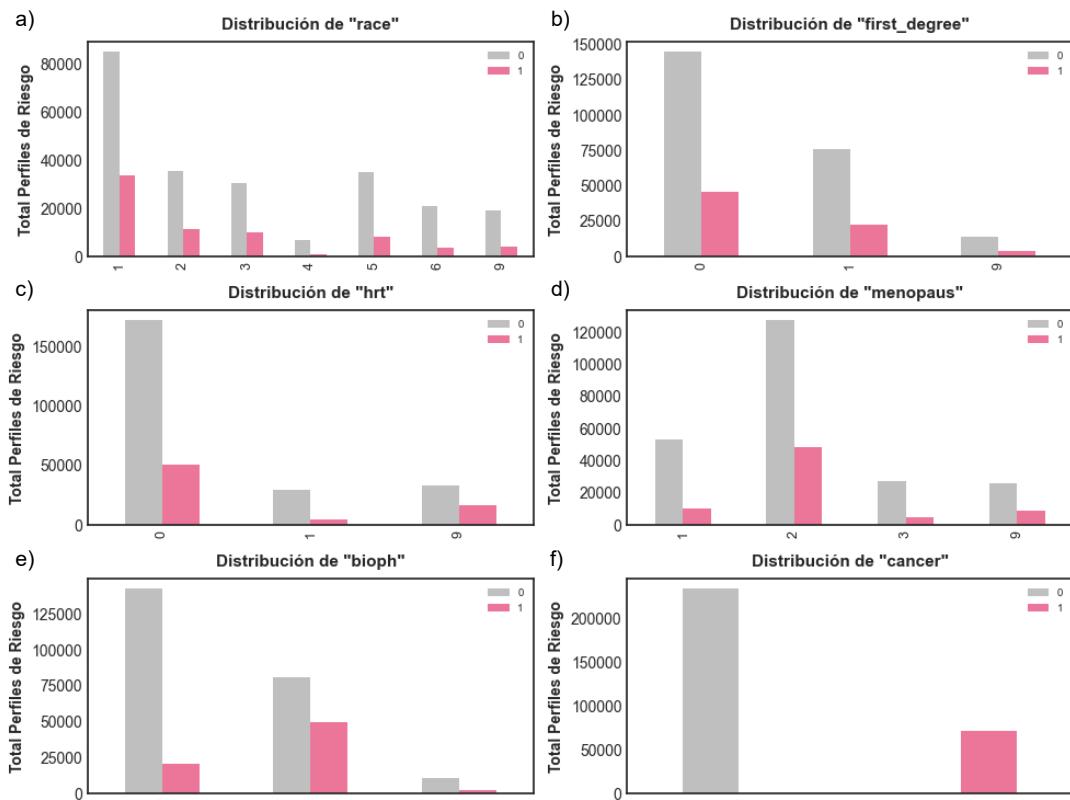


Figura A4: Distribución de los perfiles de riesgo según cada categoría de las variables explicativas categóricas nominales: a), race; b) first\_degree, c) hrt, d) menopaus, e) bioph y f) cancer; segmentadas por la presencia o ausencia de cáncer.

## Predicción de Cáncer de Mama Basado en Factores Clínicos y Demográficos

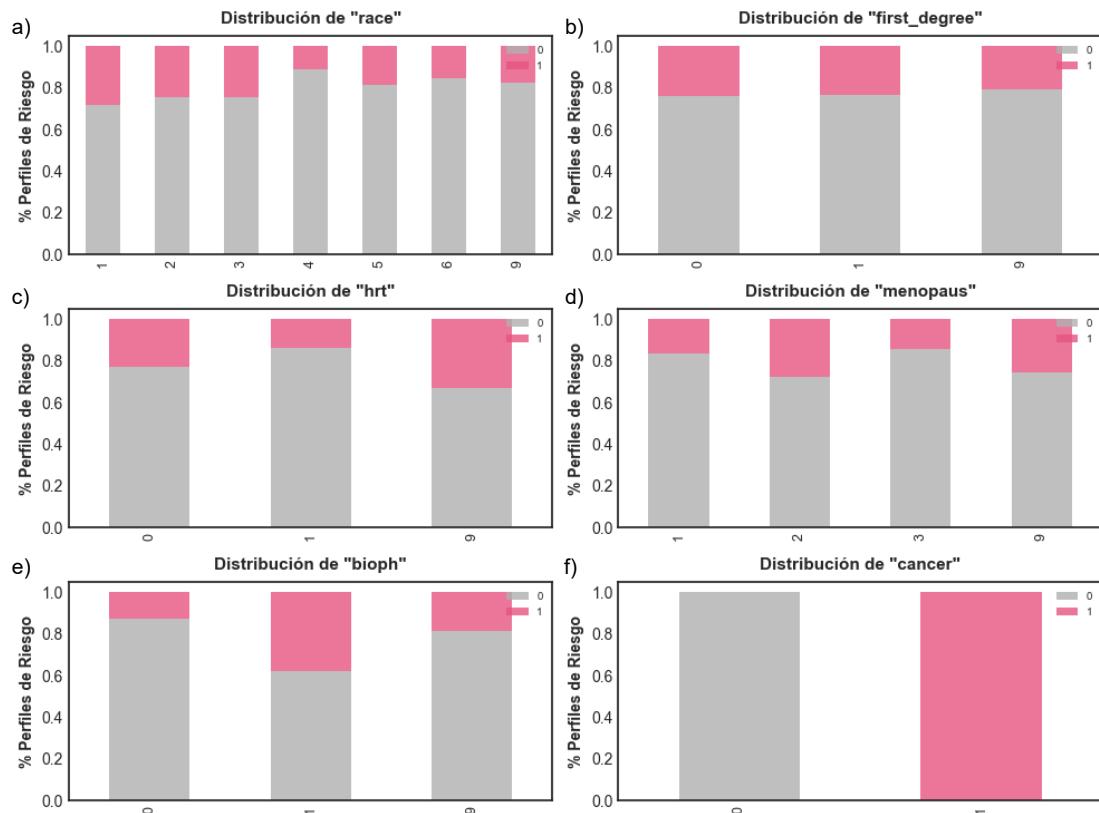


Figura A5: Distribución porcentual de los perfiles de riesgo según cada categoría de las variables nominales: a) race; b) first\_degree, c) hrt, d) menopaus, e) bioph y f) cancer, segmentadas por la presencia o ausencia de cáncer. Cada gráfico de barras apiladas refleja el porcentaje de perfiles dentro de cada categoría, facilitando la comparación relativa entre categorías y la identificación de combinaciones de factores de riesgo más asociadas con la presencia de cáncer.

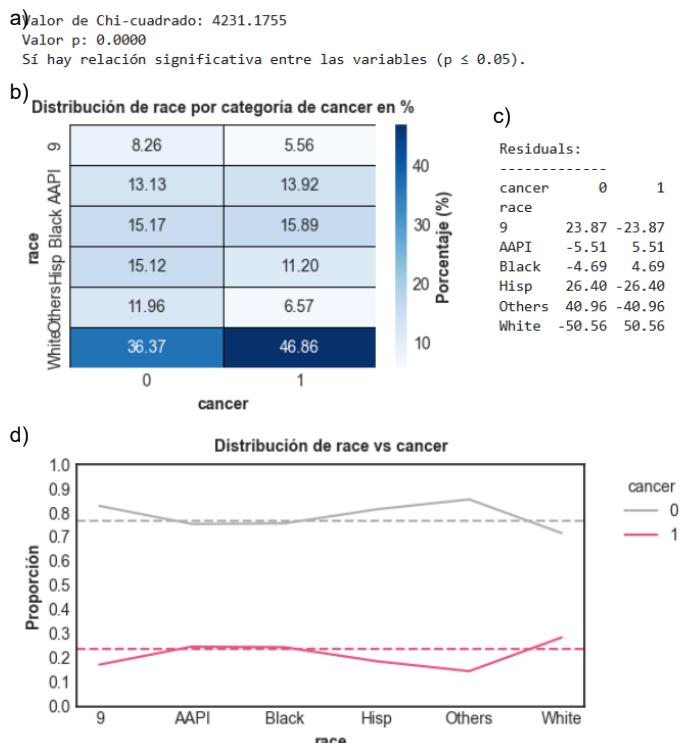


Figura A6: Análisis bivariante entre race y cáncer: a) Resultado del test  $\chi^2$  con su p-valor asociado. b) Mapa de calor de la tabla de contingencia, muestra el porcentaje de perfiles en cada combinación, destacando dónde se concentra un mayor riesgo. c) Residuos estandarizados; permiten identificar combinaciones que se desvían más de lo esperado bajo independencia, ayudando a interpretar la fuerza de la relación. d) Gráficos de líneas comparativos; muestran la distribución de los perfiles respecto a la proporción de cada clase de la variable cáncer; los puntos más alejados de la línea indican una mayor capacidad de la variable race para discriminar la variable objetivo.

## Predicción de Cáncer de Mama Basado en Factores Clínicos y Demográficos

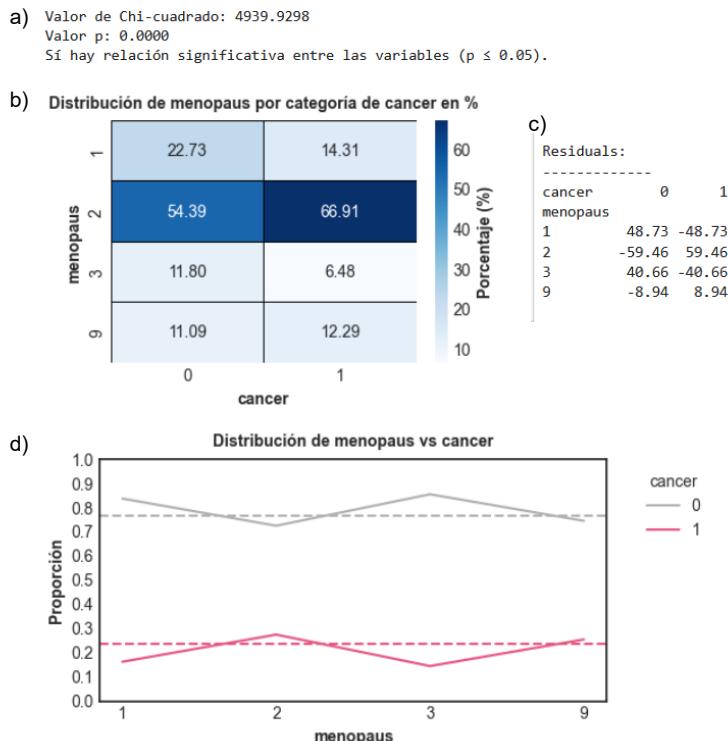


Figura A7: Análisis bivariante entre menopaus y cáncer: a) Resultado del test  $\chi^2$  con su p-valor asociado. b) Mapa de calor de la tabla de contingencia, muestra el porcentaje de perfiles en cada combinación, destacando dónde se concentra un mayor riesgo. c) Residuos estandarizados; permiten identificar combinaciones que se desvian más de lo esperado bajo independencia, ayudando a interpretar la fuerza de la relación. d) Gráficos de líneas comparativos; muestran la distribución de los perfiles respecto a la proporción de cada clase de la variable cáncer; los puntos más alejados de la línea indican una mayor capacidad de la variable menopaus para discriminar la variable objetivo.

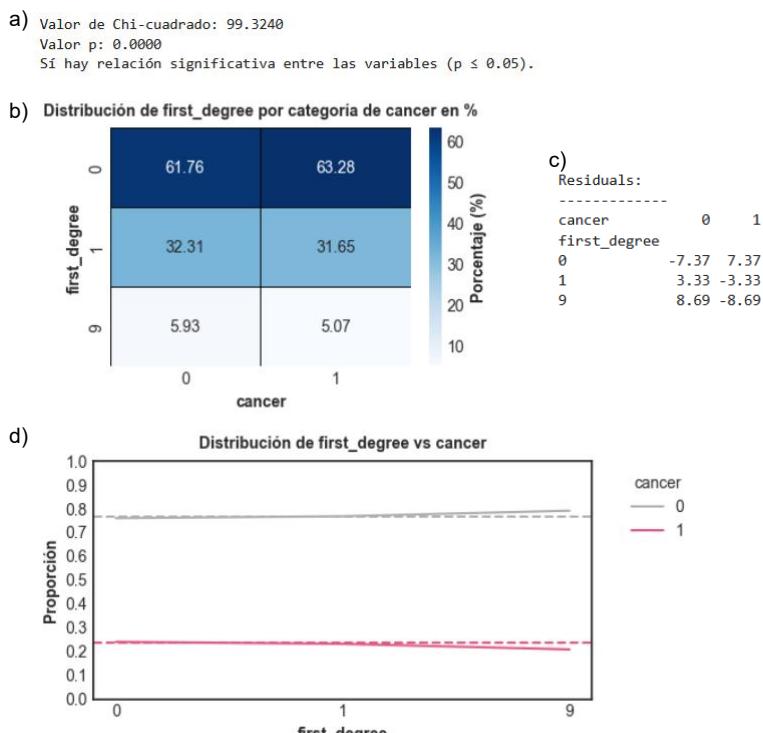


Figura A8: Análisis bivariante entre first\_degree y cáncer: a) Resultado del test  $\chi^2$  con su p-valor asociado. b) Mapa de calor de la tabla de contingencia, muestra el porcentaje de perfiles en cada combinación, destacando dónde se concentra un mayor riesgo. c) Residuos estandarizados; permiten identificar combinaciones que se desvian más de lo esperado bajo independencia, ayudando a interpretar la fuerza de la relación. d) Gráficos de líneas comparativos; muestran la distribución de los perfiles respecto a la proporción de cada clase de la variable cáncer; los puntos más alejados de la línea indican una mayor capacidad de la variable first\_degree para discriminar la variable objetivo.

## Predicción de Cáncer de Mama Basado en Factores Clínicos y Demográficos

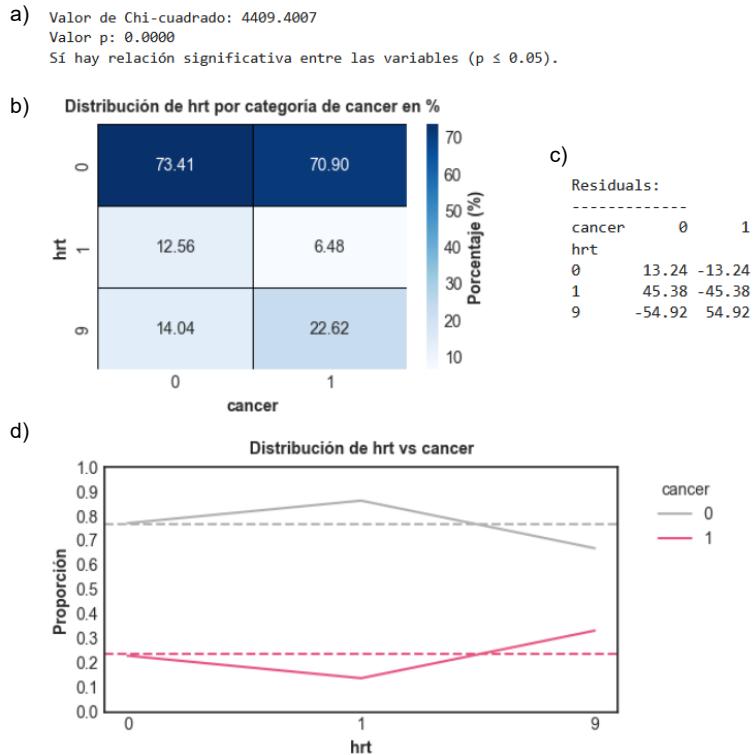


Figura A9: Análisis bivariante entre hrt y cáncer: a) Resultado del test  $\chi^2$  con su p-valor asociado. b) Mapa de calor de la tabla de contingencia, muestra el porcentaje de perfiles en cada combinación, destacando dónde se concentra un mayor riesgo. c) Residuos estandarizados; permiten identificar combinaciones que se desvian más de lo esperado bajo independencia, ayudando a interpretar la fuerza de la relación. d) Gráficos de líneas comparativos; muestran la distribución de los perfiles respecto a la proporción de cada clase de la variable cáncer; los puntos más alejados de la línea indican una mayor capacidad de la variable hrt para discriminar la variable objetivo.

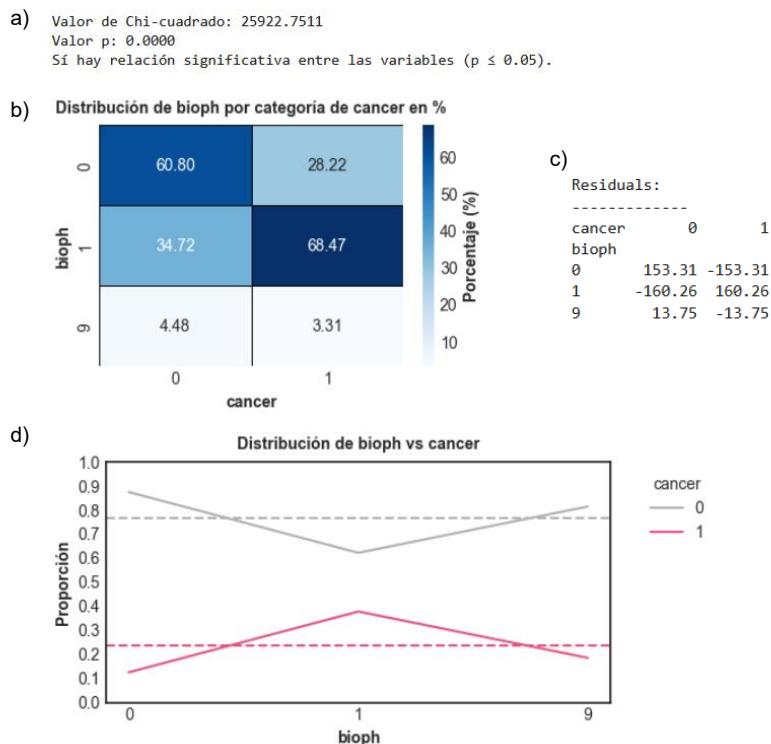


Figura A10: Análisis bivariante entre bioph y cáncer: a) Resultado del test  $\chi^2$  con su p-valor asociado. b) Mapa de calor de la tabla de contingencia, muestra el porcentaje de perfiles en cada combinación, destacando dónde se concentra un mayor riesgo. c) Residuos estandarizados; permiten identificar combinaciones que se desvian más de lo esperado bajo independencia, ayudando a interpretar la fuerza de la relación. d) Gráficos de líneas comparativos; muestran la distribución de los perfiles respecto a la proporción de cada clase de la variable cáncer; los puntos más alejados de la línea indican una mayor capacidad de la variable bioph para discriminar la variable objetivo.

## ■ EDA: Variables Categóricas Ordinales

a)		age	menarche	first_birth	birads	bmi						
	<b>count</b>	306345.000000	306345.000000	306345.000000	306345.000000	306345.000000						
	<b>mean</b>	7.160368	4.491971	3.451834	3.439338	4.085263						
	<b>std</b>	2.684390	4.027745	3.100222	2.503615	3.133181						
	<b>min</b>	1.000000	0.000000	0.000000	1.000000	1.000000						
	<b>25%</b>	5.000000	1.000000	1.000000	2.000000	2.000000						
	<b>50%</b>	7.000000	2.000000	3.000000	3.000000	3.000000						
	<b>75%</b>	9.000000	9.000000	4.000000	4.000000	9.000000						
	<b>max</b>	13.000000	9.000000	9.000000	9.000000	9.000000						
b)	count	mean	std	min	25%	50%	75%	max	Asimetría	Kurtosis	Rango	
	<b>age</b>	306345.0	7.160368	2.684390	1.0	5.0	7.0	9.0	13.0	0.293607	-0.482994	12.0
	<b>menarche</b>	171922.0	0.967218	0.770292	0.0	0.0	1.0	2.0	2.0	0.167581	-1.899166	9.0
	<b>first_birth</b>	243561.0	2.021654	1.452112	0.0	1.0	2.0	3.0	4.0	0.887623	-0.481229	9.0
	<b>birads</b>	260008.0	2.448352	0.944739	1.0	2.0	2.0	3.0	4.0	1.740614	1.948514	8.0
	<b>bmi</b>	223527.0	2.264326	1.090290	1.0	1.0	2.0	3.0	4.0	0.815245	-1.009936	8.0

Tabla A4: Resumen estadístico descriptivo de las variables categóricas ordinales tratadas como numéricas. Se muestran los resultados: a) Incluyendo todas las categorías, a tal como aparece en los perfiles originales; b) Excluyendo la categoría "9" (desconocido), reemplazada por NaN, para observar la distribución de las categorías conocidas y eliminar el efecto de los valores desconocidos en los perfiles.

Correlaciones de Spearman entre variables ordinales

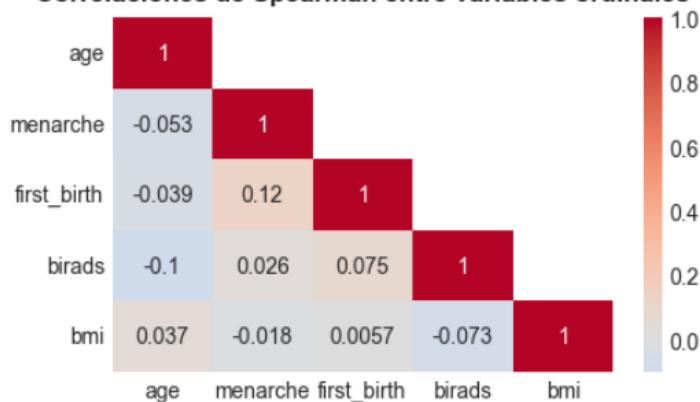


Figura A11: Mapa de calor de las correlaciones de Spearman entre las variables ordinales del conjunto de datos. Un valor cercano a 1 indica correlación positiva, un valor cercano a -1 indica correlación negativa y valores próximos a 0 indican ausencia de correlación

## Predicción de Cáncer de Mama Basado en Factores Clínicos y Demográficos

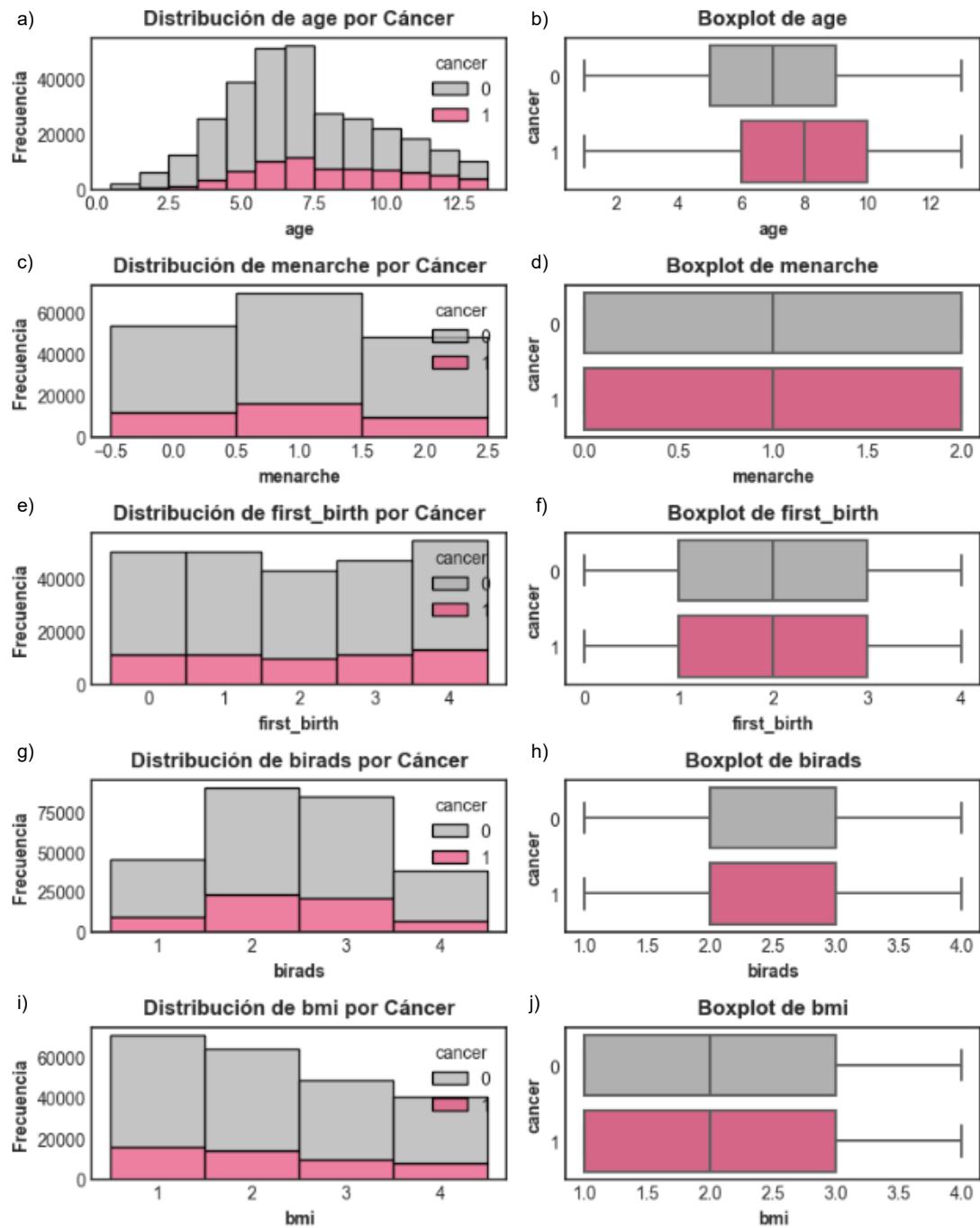


Figura A12: Distribución de las variables ordinales del conjunto de datos en relación con la presencia de cáncer: age (a, b), menarche (c,d), first\_birth (e,f), birads (g, h), bmi (i,j). Columna izquierda: histogramas apilados por categoría de la variable ordinal, mostrando la frecuencia de perfiles con y sin cáncer, lo que permite identificar visualmente cómo varía la proporción de casos según cada nivel de la variable. Columna derecha: boxplots que muestran la relación entre cada variable ordinal y la variable objetivo (cancer), facilitando la identificación de tendencias o diferencias entre categorías.

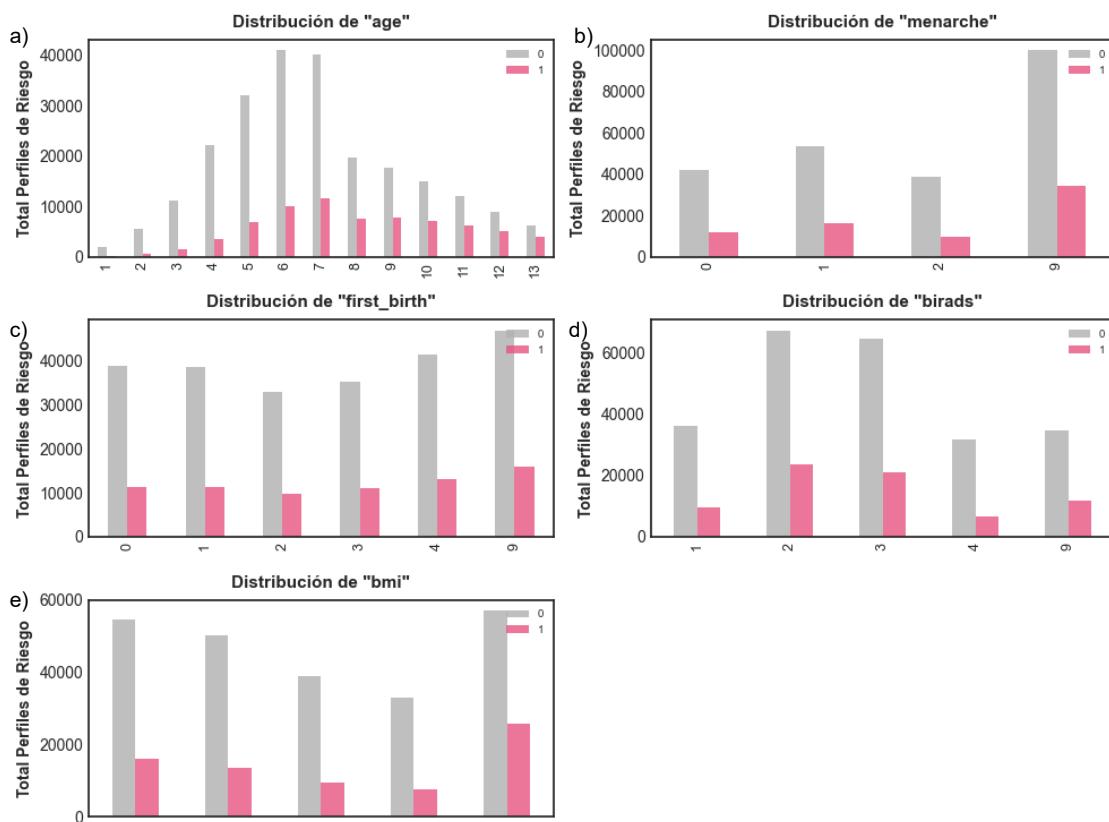
## Predicción de Cáncer de Mama Basado en Factores Clínicos y Demográficos

a)	age	menarche	first_birth	birads	bmi
<b>count</b>	306345	306345	306345	306345	306345
<b>unique</b>	13	4	6	5	5
<b>top</b>	7	9	9	2	9
<b>freq</b>	51929	134423	62784	90691	82818

b)	age	menarche	first_birth	birads	bmi
<b>count</b>	306345	171922.0	243561.0	260008.0	223527.0
<b>unique</b>	13	3.0	5.0	4.0	4.0
<b>top</b>	7	1.0	4.0	2.0	1.0
<b>freq</b>	51929	69728.0	54333.0	90691.0	70526.0

Tabla A5: Resumen estadístico descriptivo de las variables categóricas ordinales tratadas como categóricas. Se muestran los resultados: a) Incluyendo todas las categorías, a tal como aparece en los perfiles originales; b) Excluyendo la categoría "9" (desconocido), reemplazada por NaN, para observar la distribución de las categorías conocidas y eliminar el efecto de los valores desconocidos en los perfiles



## Predicción de Cáncer de Mama Basado en Factores Clínicos y Demográficos

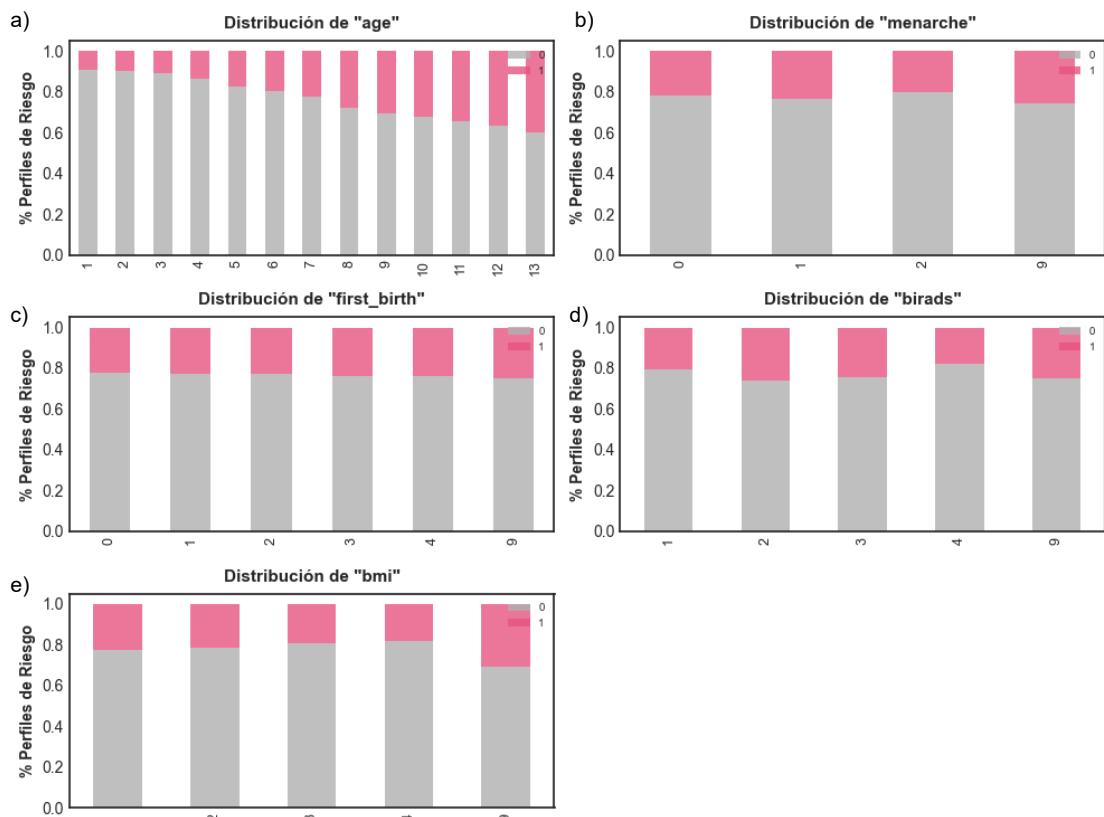


Figura A15: Distribución porcentual de los perfiles de riesgo según cada categoría de las variables nominales: a) age, b) menarche, c) first\_birth, d) birads y e) bmi; segmentadas por la presencia o ausencia de cáncer. Cada gráfico de barras apiladas refleja el porcentaje de perfiles dentro de cada categoría, facilitando la comparación relativa entre categorías y la identificación de combinaciones de factores de riesgo más asociadas con la presencia de cáncer.

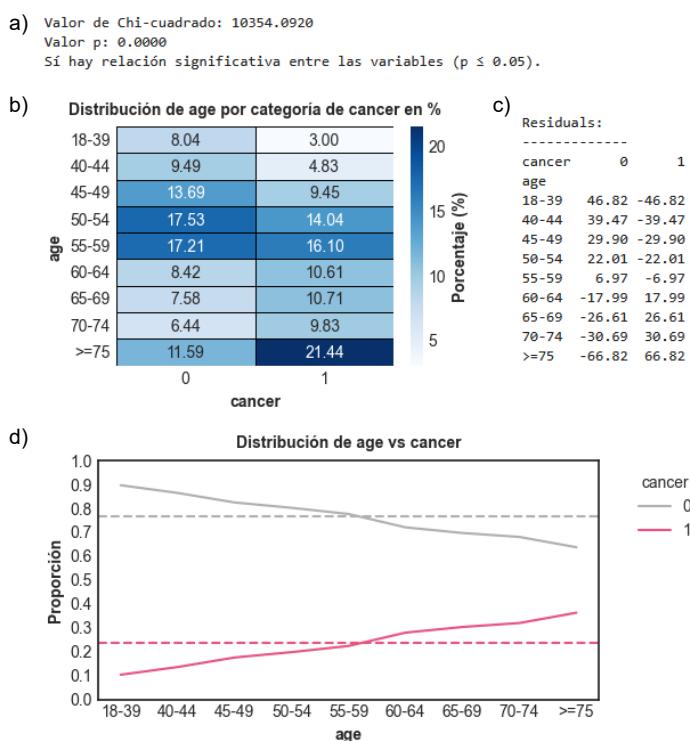


Figura A16 Análisis bivariante entre age y cáncer: a) Resultado del test  $\chi^2$  con su  $p$ -valor asociado. b) Mapa de calor de la tabla de contingencia, muestra el porcentaje de perfiles en cada combinación, destacando dónde se concentra un mayor riesgo. c) Residuos estandarizados; permiten identificar combinaciones que se desvían más de lo esperado bajo independencia, ayudando a interpretar la fuerza de la relación. d) Gráficos de líneas comparativos; muestran la distribución de los perfiles respecto a la proporción de cada clase de la variable cáncer; los puntos más alejados de la línea indican una mayor capacidad de la variable age para discriminar la variable objetivo.

## Predicción de Cáncer de Mama Basado en Factores Clínicos y Demográficos

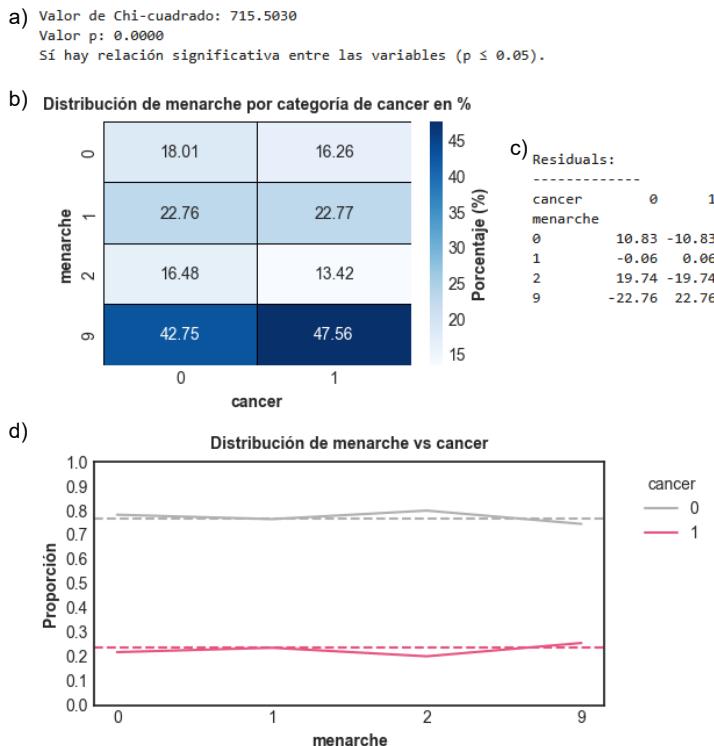


Figura A17 Análisis bivariante entre menarche y cáncer: a) Resultado del test  $\chi^2$  con su p-valor asociado. b) Mapa de calor de la tabla de contingencia, muestra el porcentaje de perfiles en cada combinación, destacando dónde se concentra un mayor riesgo. c) Residuos estandarizados; permiten identificar combinaciones que se desvian más de lo esperado bajo independencia, ayudando a interpretar la fuerza de la relación. d) Gráficos de líneas comparativos; muestran la distribución de los perfiles respecto a la proporción de cada clase de la variable cáncer; los puntos más alejados de la línea indican una mayor capacidad de la variable menarche para discriminar la variable objetivo.

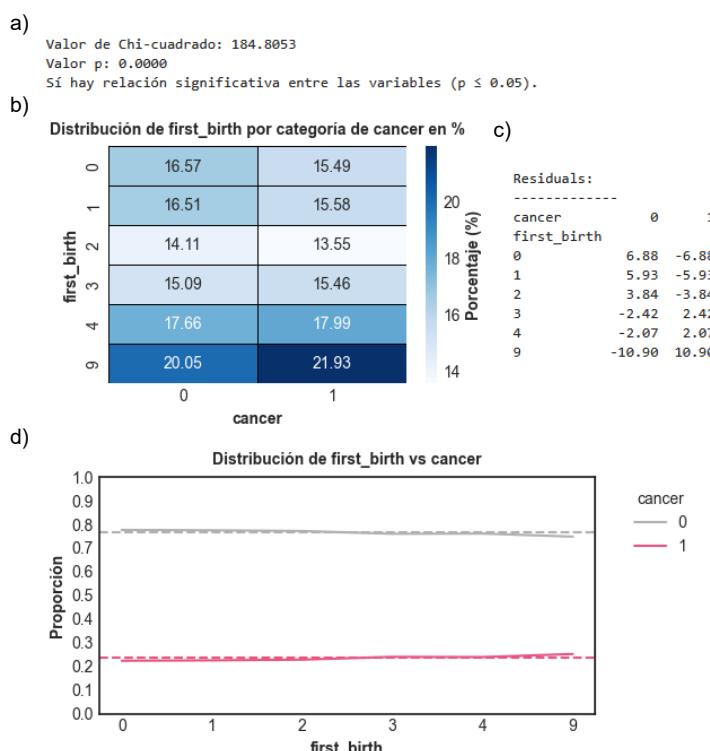


Figura A18 Análisis bivariante entre first\_birth y cáncer: a) Resultado del test  $\chi^2$  con su p-valor asociado. b) Mapa de calor de la tabla de contingencia, muestra el porcentaje de perfiles en cada combinación, destacando dónde se concentra un mayor riesgo. c) Residuos estandarizados; permiten identificar combinaciones que se desvian más de lo esperado bajo independencia, ayudando a interpretar la fuerza de la relación. d) Gráficos de líneas comparativos; muestran la distribución de los perfiles respecto a la proporción de cada clase de la variable cáncer; los puntos más alejados de la línea indican una mayor capacidad de la variable first\_birth para discriminar la variable objetivo.

## Predicción de Cáncer de Mama Basado en Factores Clínicos y Demográficos

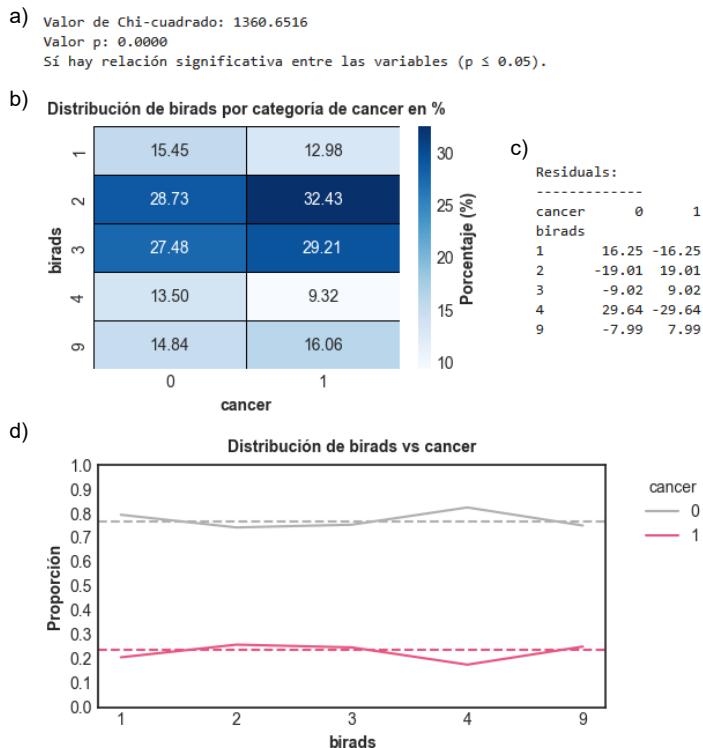


Figura A19 Análisis bivariante entre birads y cáncer: a) Resultado del test  $\chi^2$  con su p-valor asociado. b) Mapa de calor de la tabla de contingencia, muestra el porcentaje de perfiles en cada combinación, destacando dónde se concentra un mayor riesgo. c) Residuos estandarizados; permiten identificar combinaciones que se desvian más de lo esperado bajo independencia, ayudando a interpretar la fuerza de la relación. d) Gráficos de líneas comparativos; muestran la distribución de los perfiles respecto a la proporción de cada clase de la variable cáncer; los puntos más alejados de la línea indican una mayor capacidad de la variable birads para discriminar la variable objetivo.

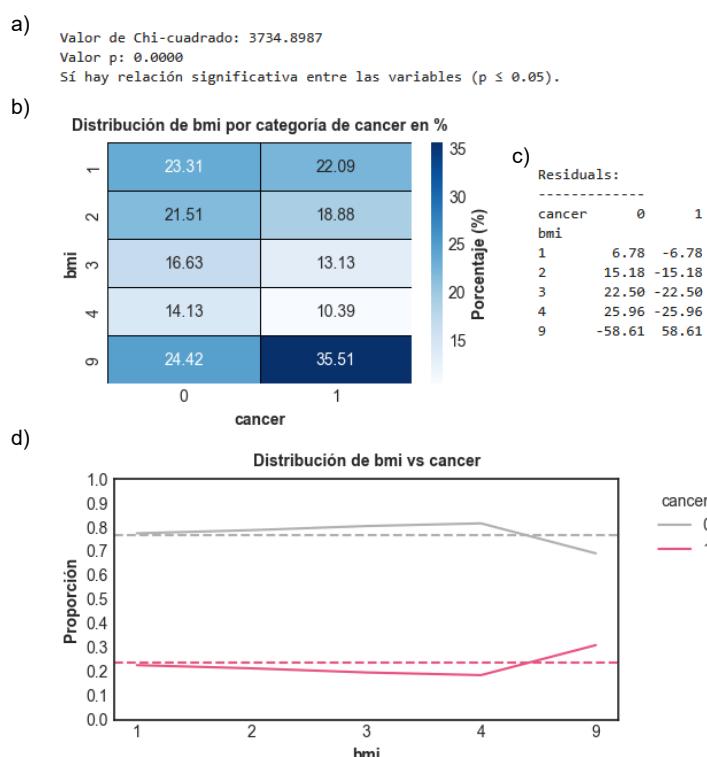


Figura A20 Análisis bivariante entre bmi y cáncer: a) Resultado del test  $\chi^2$  con su p-valor asociado. b) Mapa de calor de la tabla de contingencia, muestra el porcentaje de perfiles en cada combinación, destacando dónde se concentra un mayor riesgo. c) Residuos estandarizados; permiten identificar combinaciones que se desvian más de lo esperado bajo independencia, ayudando a interpretar la fuerza de la relación. d) Gráficos de líneas comparativos; muestran la distribución de los perfiles respecto a la proporción de cada clase de la variable cáncer; los puntos más alejados de la línea indican una mayor capacidad de la variable bmi para discriminar la variable objetivo.

## ■ EDA: VCramer

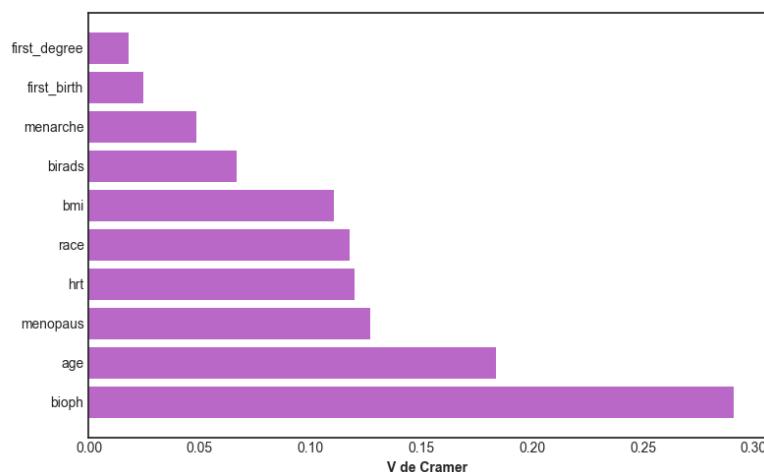


Figura A21: Gráfico de barras que muestran la asociación entre cada variable explicativa y la variable objetivo (cancer) utilizando el coeficiente de VCramer. La longitud de las barras indica la fuerza de la relación, donde valores más altos representan una mayor dependencia entre la variable explicativa y la presencia de cáncer.

## ■ EDA: Valores Faltantes

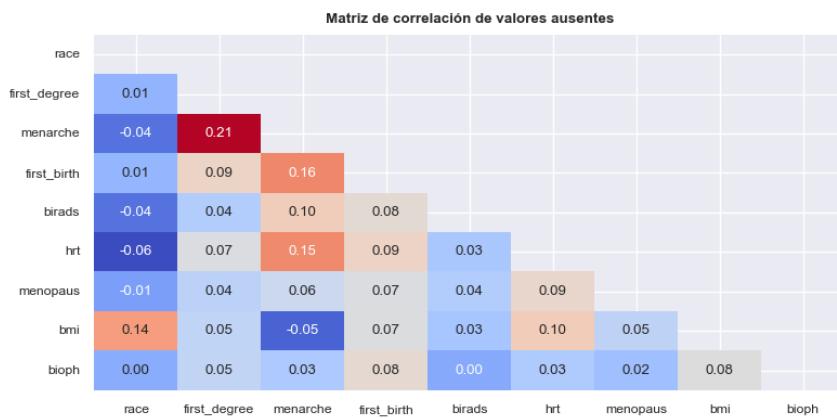


Figura A22: Mapa de calor que muestra la correlación entre los valores faltantes de las diferentes variables del conjunto de datos. Se considera correlación cuando el valor absoluto es superior al 90%.

prop_missings distribution:		
	Count	%
prop_missings		
0.08	98043	32.00
0.17	73910	24.13
0.0	72251	23.58
0.25	39744	12.97
>0.30	22397	7.31

Tabla A6: Distribución de frecuencias de la variable prop\_missing, que representa la proporción de valores faltantes por registro

## Predicción de Cáncer de Mama Basado en Factores Clínicos y Demográficos

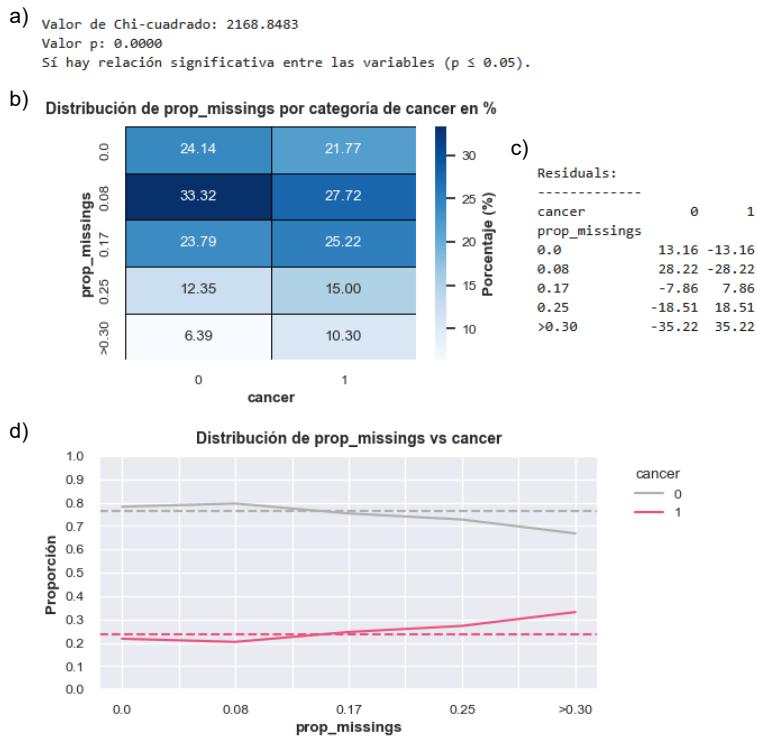


Figura A23 : Análisis bivariante entre prop\_missings y cáncer: a) Resultado del test  $\chi^2$  con su p-valor asociado. b) Mapa de calor de la tabla de contingencia, muestra el porcentaje de perfiles en cada combinación, destacando dónde se concentra un mayor riesgo. c) Residuos estandarizados; permiten identificar combinaciones que se desvían más de lo esperado bajo independencia, ayudando a interpretar la fuerza de la relación. d) Gráficos de líneas comparativos; muestran la distribución de los perfiles respecto a la proporción de cada clase de la variable cáncer; los puntos más alejados de la línea indican una mayor capacidad de la variable prop\_missings para discriminar la variable objetivo

## Anexo B: Selección de Variables

### ■ Feature Transformation

```
# column Transformer
imputer = IterativeImputer(
    estimator=DecisionTreeClassifier(random_state=seed),
    random_state=seed)

preprocessor = ColumnTransformer([
    ("imputer", imputer, cat_w_miss),
    ('OneH_binary', OneHotEncoder(drop='if_binary'), cat_binary),
    ('OneH_nom', OneHotEncoder(drop='first'), cat_nom),
    ("ord_enc", OrdinalEncoder(categories=[age_order, missing_order]), ['age', 'prop_missings']),
], remainder='passthrough')
```

Figura B1: Esquema del preprocesador de datos (pipeline) utilizado en el proyecto.

### ■ Balance de Clases

Técnica	Clase 0 (n)	Clase 1 (n)
None	187186 (76,5%)	57545 (23,5%)
SMOTE	187186 (50%)	187186 (50%)
SMOTEEENN	92314 (55%)	75771 (45%)
SMOTETomek	187067 (50%)	187067 (50%)
RUS+SMOTE5k3	115090 (50%)	115090 (50%)
RUS+SMOTE4k3	143862 (50%)	143862 (50%)
RUS+SMOTE5k5	115090 (50%)	115090 (50%)
RUS+SMOTE4k5	143862 (50%)	143862 (50%)
SMOTE+RUS	112311 (59%)	78618 (41%)

Tabla B1: Distribución de clases tras aplicar técnicas de balanceo en el conjunto de entrenamiento

Técnica	Recall (Train)	Recall (Test)	F1 (Train)	F1 (Test)	PR-AUC (Train)	PR-AUC (Test)
None	0,593	0,170	0,651	0,184	0,815	0,287
SMOTE	0,912	0,210	0,902	0,216	0,981	0,279
SMOTEEENN	1,000	0,677	1,000	0,504	1,000	0,385
SMOTETomek	0,912	0,210	0,902	0,217	0,981	0,279
RUS+SMOTE5k3	0,921	0,414	0,901	0,357	0,978	0,299
RUS+SMOTE4k3	0,919	0,321	0,902	0,300	0,979	0,288
RUS+SMOTE5k5	0,919	0,410	0,901	0,355	0,979	0,301
RUS+SMOTE4k5	0,916	0,315	0,902	0,295	0,980	0,291
SMOTE+RUS	0,875	0,411	0,858	0,355	0,958	0,304

Tabla B2: Comparación de métricas clave (Recall, F1 y PR-AUC) en conjuntos de entrenamiento y prueba para cada técnica de balanceo.

## ■ Selección de Variables

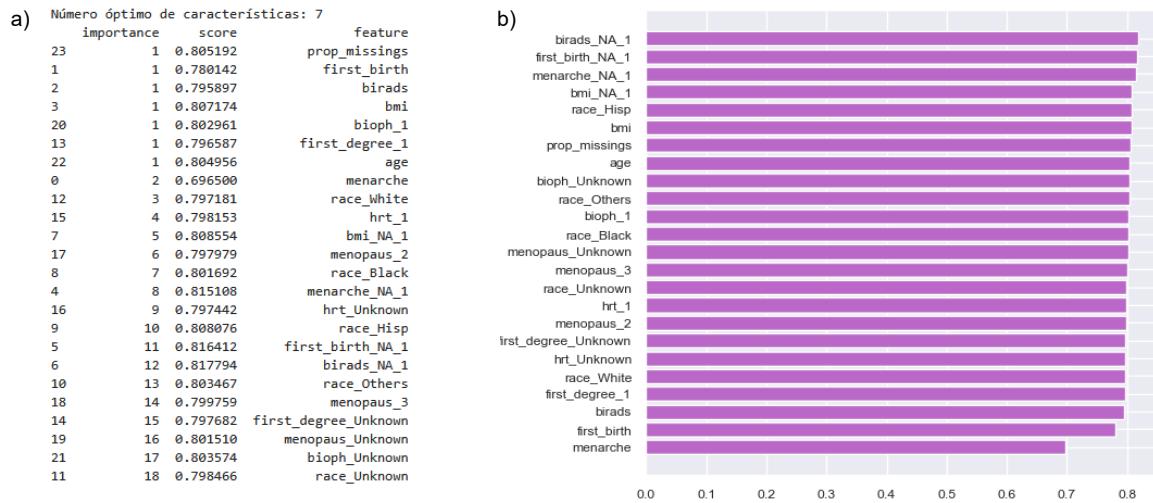


Figura B2: Resultados del proceso de selección de variables con RFECV utilizando RandomForestClassifier como estimador. A la izquierda (a), se presenta el ranking final de RFECV, donde el número de características óptimo fue 7. A la derecha (b), se muestran las puntuaciones medias de validación cruzada (AUC) para cada variable, ordenadas de menor a mayor relevancia.

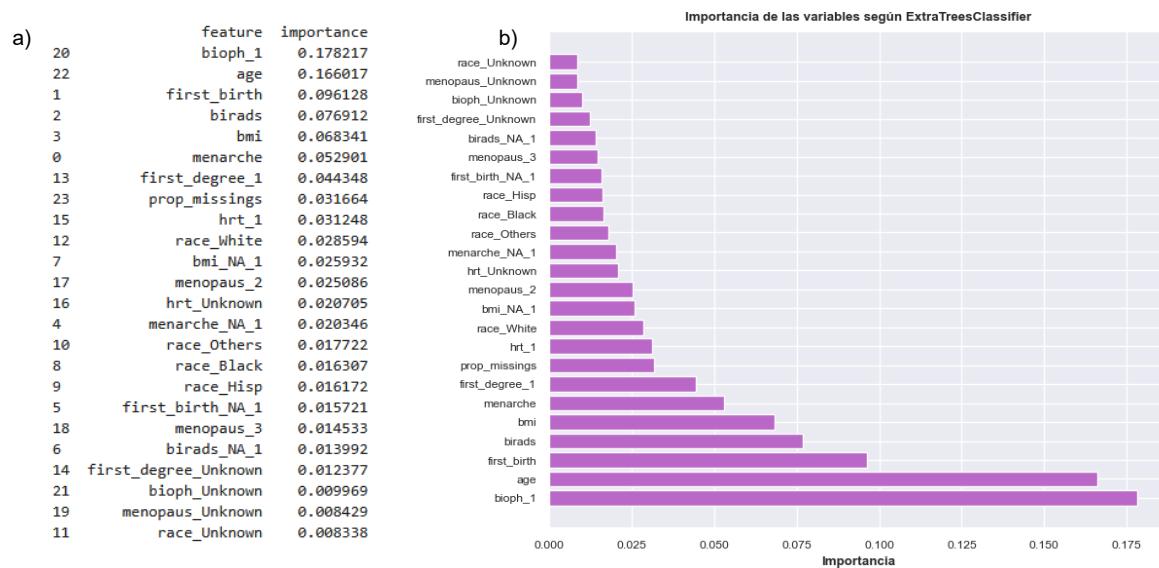


Figura B3: Importancia de las variables según ExtraTreesClassifier. A la izquierda (a), se muestran los valores numéricos de la importancia de cada variable, mientras que a la derecha (b) se visualizan mediante un gráfico de barras horizontales

## Predicción de Cáncer de Mama Basado en Factores Clínicos y Demográficos

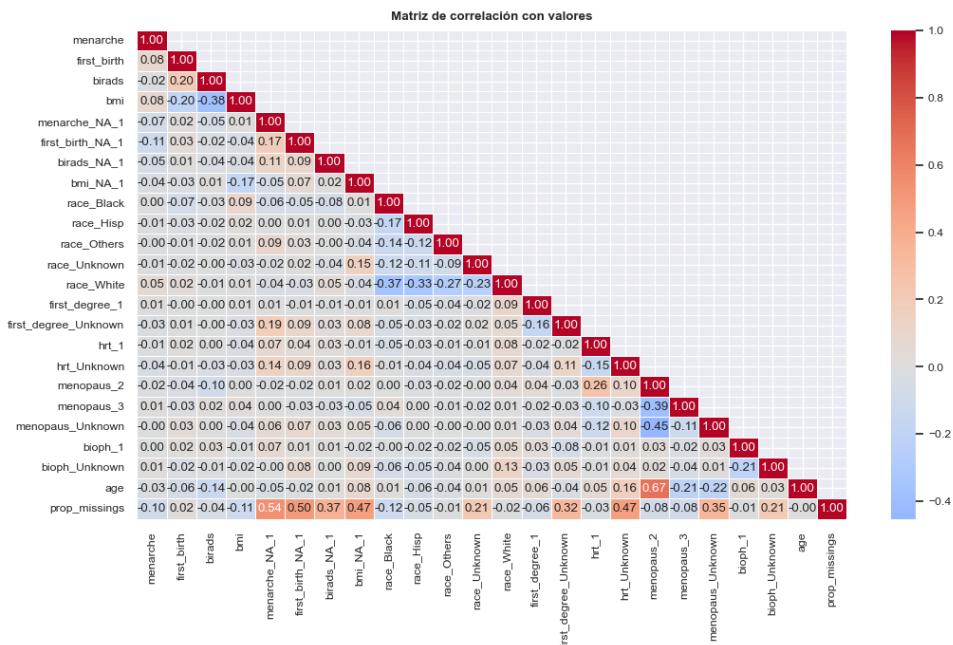


Figura B4: Matriz de correlación entre variables del conjunto de entrenamiento balanceado.

:	feature	RFEcv_Ranking	ExtraTrees_Importance
20	bioph_1	1	0.178217
22	age	1	0.166017
1	first_birth	1	0.096128
2	birads	1	0.076912
3	bmi	1	0.068341
0	menarche	2	0.052901
13	first_degree_1	1	0.044348
23	prop_missings	1	0.031664
15	hrt_1	4	0.031248
12	race_White	3	0.028594
7	bmi_NA_1	5	0.025932
17	menopaus_2	6	0.025086
16	hrt_Unknown	9	0.020705
4	menarche_NA_1	8	0.020346
10	race_Others	13	0.017722
8	race_Black	7	0.016307
9	race_Hisp	10	0.016172
5	first_birth_NA_1	11	0.015721
18	menopaus_3	14	0.014533
6	birads_NA_1	12	0.013992
14	first_degree_Unknown	15	0.012377
21	bioph_Unknown	17	0.009969
19	menopaus_Unknown	16	0.008429
11	race_Unknown	18	0.008338

Tabla B3: Comparación entre la selección de variables realizada mediante RFECV y la importancia estimada por ExtraTreesClassifier. La columna RFEcv\_Ranking indica el orden en que RFECV eliminó las variables (1 = variable seleccionada en el conjunto óptimo), mientras que ExtraTrees\_Importance refleja la contribución de cada variable a la reducción de impureza promedio en los árboles del modelo. Esta comparación permite identificar variables consistentes entre ambos métodos y evaluar su relevancia predictiva y clínica.

## Anexo C: Búsqueda Hiperparamétrica

La búsqueda de hiperparámetros consiste en ajustar de manera óptima los parámetros internos de cada algoritmo con el fin de mejorar su rendimiento. Este proceso no se limita a conseguir la mejor métrica en un conjunto de entrenamiento, sino a garantizar que el modelo sea fiable, estable y capaz de generalizar a nuevos datos.

En el contexto de este trabajo, la **prioridad fue maximizar el Recall**, es decir, la capacidad del modelo de detectar correctamente a los individuos positivos (aquellos en situación de riesgo). En un entorno clínico, este criterio es esencial, ya que un falso negativo podría implicar no detectar a una persona que realmente presenta un riesgo elevado. El AUC se utilizó como métrica complementaria, al reflejar la capacidad global de discriminación del modelo entre clases.

Además, se tuvieron en cuenta dos aspectos clave para garantizar la calidad del modelo:

- **Estabilidad:** consistencia de los resultados a través de las distintas particiones de validación cruzada.
- **Riesgo de sobreajuste:** diferencia entre el rendimiento en entrenamiento y en prueba. Modelos con métricas muy altas, pero con brechas significativas entre *train* y *test* fueron descartados por comprometer su capacidad de generalización.

La metodología aplicada fue la siguiente:

### 1. Búsqueda hiperparamétrica.

- Se aplicó *Grid Search* (búsqueda en rejilla) para explorar de manera sistemática todas las combinaciones de hiperparámetros dentro de un rango definido. Aunque esta técnica resulta computacionalmente costosa, ofrece exhaustividad y precisión.
- Para cada hiperparámetro de cada algoritmo se estableció un conjunto inicial de valores razonables.
- Cada combinación entre los posibles valores de los parámetros fue evaluada mediante validación cruzada de 5 folds, calculando Recall y AUC en cada partición de datos *train*.

### 2. Análisis de resultados mediante visualizaciones

- Evolución de las métricas obtenidas en función de los valores de los hiperparámetros.
- Comparaciones entre los resultados entre *train* y *test* para detectar posibles signos de sobreajuste.
- Boxplots de métricas por fold, lo que permitió evaluar la robustez de cada configuración (menor varianza indica mayor estabilidad).

### 3. Criterios de selección del modelo óptimo

- Maximizar Recall, priorizando la detección de positivos.
- Mantener un AUC elevado, garantizando discriminación fiable entre clases.
- Equilibrio entre Recall y AUC, descartando configuraciones con alto riesgo de sobreajuste o inestabilidad.

### 4. Refinamiento iterativo

En caso de ser necesario, el rango de hiperparámetros se ajustó en torno a las configuraciones más prometedoras, afinando el espacio de búsqueda hasta encontrar un modelo consistente. Además, si procede, se incorporaron nuevos hiperparámetros al análisis, ampliando la exploración para descubrir mejoras adicionales.

A continuación, se detalla para cada modelo el procedimiento reproducible seguido y un grid inicial razonable, manteniendo la misma semilla en todo el proceso para garantizar la comparabilidad de los resultados.

## Búsqueda para la regresión Logística

### 1. Búsqueda inicial: parámetro de regularización C, penalty y solver

```

lr = LogisticRegression(random_state=seed)

param_grid_lr = [
    {
        'C': [0.01, 0.1, 1, 5, 10, 15, 20],
        'penalty': ['l2'],
        'solver': ['lbfgs'],
        'class_weight': ['balanced']
    },
    {
        'C': [0.01, 0.1, 1, 5, 8, 10, 15, 20],
        'penalty': ['l1', 'l2'],
        'solver': ["saga"],
        "class_weight": ["balanced"]
    }
]

grid_lr = GridSearchCV(lr, param_grid_lr, cv=cv5,
                      scoring=scoring_metrics ,
                      refit='roc_auc',
                      n_jobs=-1)

grid_lr.fit(X_train_work, y_train_work)

```

Figura C1: Código del GridSearchCV aplicada al modelo de Regresión Logística, explorando diferentes valores de C, penalizaciones (l1, l2) y solvers (lbfgs, sagal) bajo ponderación balanceada de clases

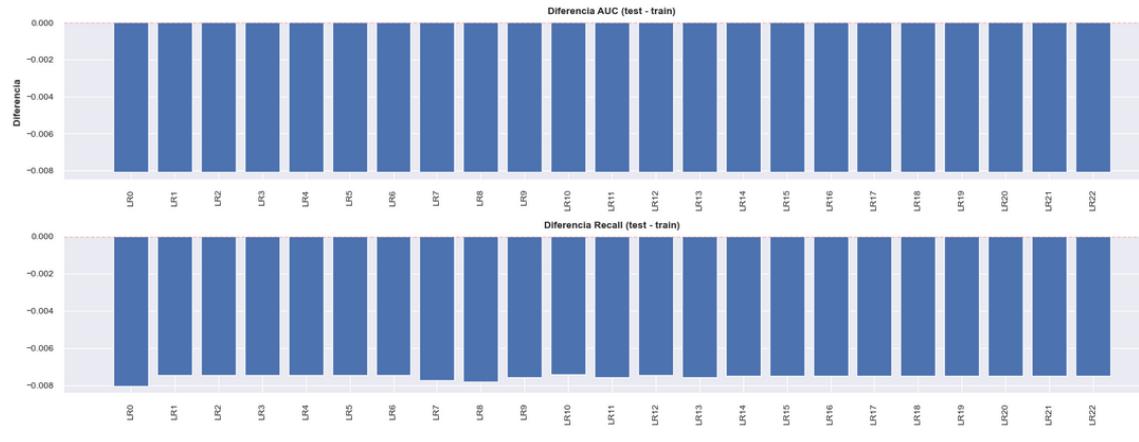


Figura C2: Comparación de la diferencia train-test para diferentes métricas en los modelos obtenidos de la búsqueda inicial para RL.

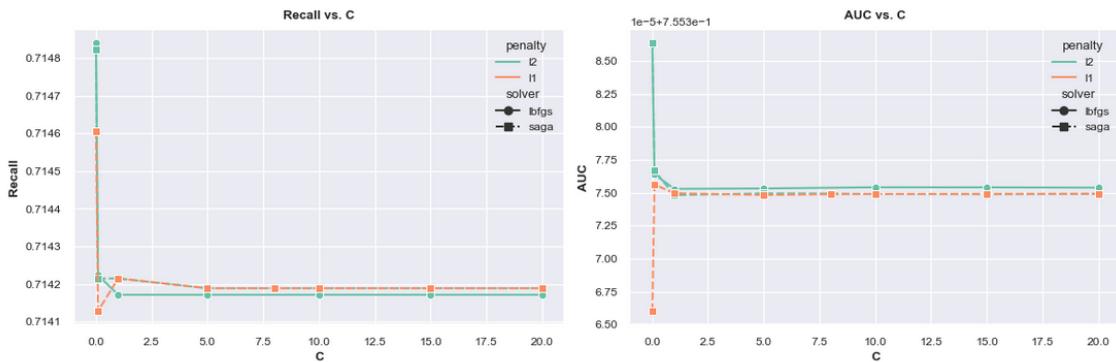


Figura C3: Recall promedio (izquierda) y Auc promedio (derecha) resultantes de la validación cruzada para la RL en función del parámetro de regularización C y comparaciones con las penalizaciones (l1, l2) y solvers (lbfgs, sagal)

## Predicción de Cáncer de Mama Basado en Factores Clínicos y Demográficos

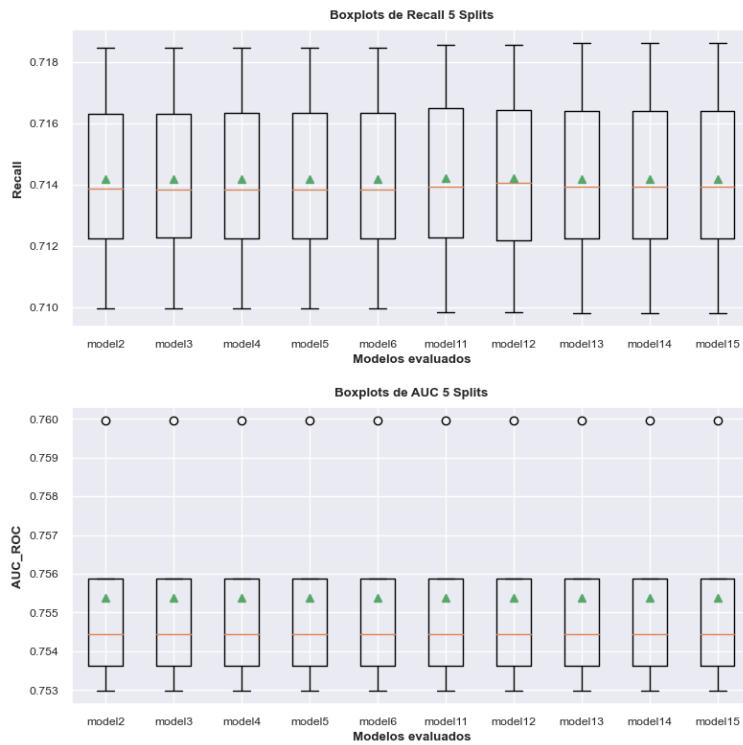


Figura C4: Boxplots de desempeño de los modelos seleccionados tras la búsqueda en LR: distribución de Recall (arriba) y AUC (abajo) obtenidos en los distintos splits de validación cruzada. Se representan únicamente los 10 mejores modelos menos sobreentrenados y que maximizan las métricas.

```
AUC para cada modelo:
LR2 | AUC_Train: 0.7554 | AUC_Test: 0.7473 | Diferencia: -0.0081
LR3 | AUC_Train: 0.7554 | AUC_Test: 0.7473 | Diferencia: -0.0081
LR4 | AUC_Train: 0.7554 | AUC_Test: 0.7473 | Diferencia: -0.0081
LR5 | AUC_Train: 0.7554 | AUC_Test: 0.7473 | Diferencia: -0.0081
LR6 | AUC_Train: 0.7554 | AUC_Test: 0.7473 | Diferencia: -0.0081
LR11 | AUC_Train: 0.7554 | AUC_Test: 0.7473 | Diferencia: -0.0081
LR12 | AUC_Train: 0.7554 | AUC_Test: 0.7473 | Diferencia: -0.0081
LR13 | AUC_Train: 0.7554 | AUC_Test: 0.7473 | Diferencia: -0.0081
LR14 | AUC_Train: 0.7554 | AUC_Test: 0.7473 | Diferencia: -0.0081
LR15 | AUC_Train: 0.7554 | AUC_Test: 0.7473 | Diferencia: -0.0081

Recall para cada modelo:
LR2 | Recall_Train: 0.7143 | Recall_Test: 0.7068 | Diferencia: -0.0075
LR3 | Recall_Train: 0.7143 | Recall_Test: 0.7068 | Diferencia: -0.0075
LR4 | Recall_Train: 0.7143 | Recall_Test: 0.7068 | Diferencia: -0.0075
LR5 | Recall_Train: 0.7143 | Recall_Test: 0.7068 | Diferencia: -0.0075
LR6 | Recall_Train: 0.7143 | Recall_Test: 0.7068 | Diferencia: -0.0075
LR11 | Recall_Train: 0.7143 | Recall_Test: 0.7067 | Diferencia: -0.0076
LR12 | Recall_Train: 0.7143 | Recall_Test: 0.7068 | Diferencia: -0.0075
LR13 | Recall_Train: 0.7143 | Recall_Test: 0.7067 | Diferencia: -0.0076
LR14 | Recall_Train: 0.7143 | Recall_Test: 0.7068 | Diferencia: -0.0075
LR15 | Recall_Train: 0.7143 | Recall_Test: 0.7068 | Diferencia: -0.0075
```

Figura C5: Desempeño de los modelos de Regresión Logística seleccionados (LR2–LR15) en términos de AUC y Recall. Para cada modelo se muestran las métricas en el conjunto de entrenamiento y prueba, así como la diferencia entre ambos.

Todos los modelos presentan prácticamente el mismo rendimiento, con valores de *Mean AUC* y *Mean Accuracy* casi idénticos tanto en los conjuntos de entrenamiento como de prueba. Esto indica que su rendimiento es equivalente entre ellos. En este contexto, conviene aplicar el principio de parsimonia, seleccionando el modelo más simple y con menor riesgo de *overfitting*: Modelo ganador: LR2

```
LogisticRegression(C=1, penalty="l2", solver= "lbfgs",
random_state=seed, class_weight="balanced")
```

## Búsqueda para Random Forest

1. **Búsqueda inicial:** número de estimadores, profundidad de árbol y muestras por hoja.

```

rf = RandomForestClassifier(random_state= seed)

param_grid_rf = {
    "bootstrap": [True],
    "n_estimators": [100, 200, 300, 400],
    "max_depth": [5, 10, 20, 30],
    "min_samples_leaf": [2, 5, 10, 20],
    # "max_features": [0.6, 0.7, 0.8, 0.9, 1.0],
    # "max_samples": [0.6, 0.7, 0.9],
    "class_weight": ["balanced"]
}

grid_rf = GridSearchCV(rf, param_grid_rf, cv=cv5,
                       scoring=scoring_metrics,
                       refit="roc_auc",
                       n_jobs=-1)
    
```

Figura C6: Definición de la rejilla de hiperparámetros para la búsqueda en GridSearchCV del modelo RF. Se exploran distintos valores de número de estimadores (`n_estimators`), profundidad máxima del árbol (`max_depth`) y número mínimo de muestras por hoja (`min_samples_leaf`), utilizando bootstrap activado y ponderación de clases balanceada.

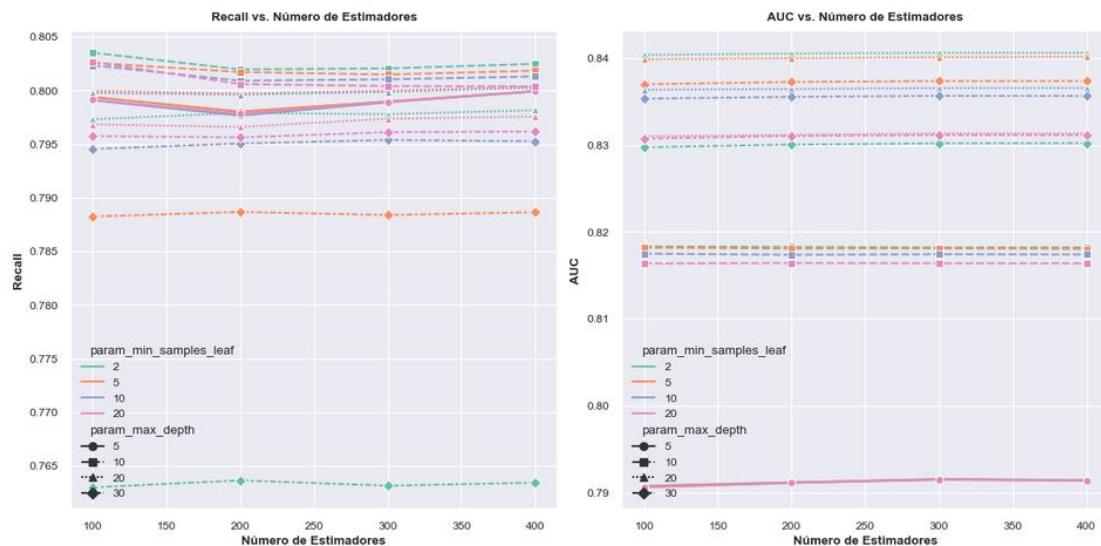


Figura C7: Recall promedio (izquierda) y AUC promedio (derecha) resultantes de la validación cruzada para RF en función del número de estimadores (`n_estimators`). Se comparan distintos valores de `min_samples_leaf` (colores) y `max_depth` (estilos de línea).

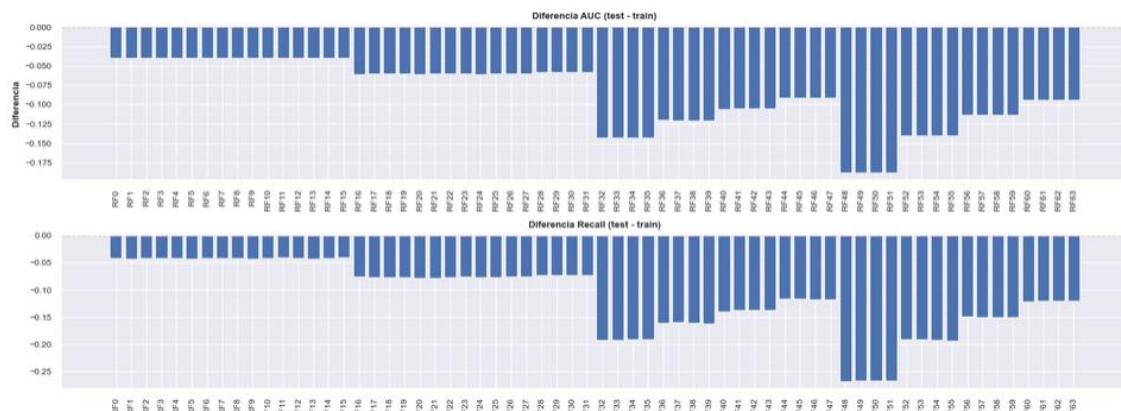


Figura C8 : Comparación de la diferencia train-test para diferentes métricas en los modelos obtenidos de la búsqueda inicial para RF.

## 2. Segunda búsqueda: fracción de muestra y fracción de las variables explicativas

```

param_grid_rf = {
    "bootstrap": [True],
    "n_estimators": [100],
    "max_depth": [5],
    "min_samples_leaf": [10],
    "max_features": [0.6, 0.7, 0.8, 0.9, 1.0],
    "max_samples": [0.6, 0.7, 0.8, 0.9, 1.0],
    "class_weight": ["balanced"]
}
    
```

Figura C9: Nueva definición de la malla de hiperparámetros para Random Forest, fijando  $n\_estimators=100$ ,  $max\_depth=5$  y  $min\_samples\_leaf=10$ , y explorando distintas proporciones de  $max\_features$  y  $max\_samples$  con bootstrap activado y ponderación de clases balanceada.

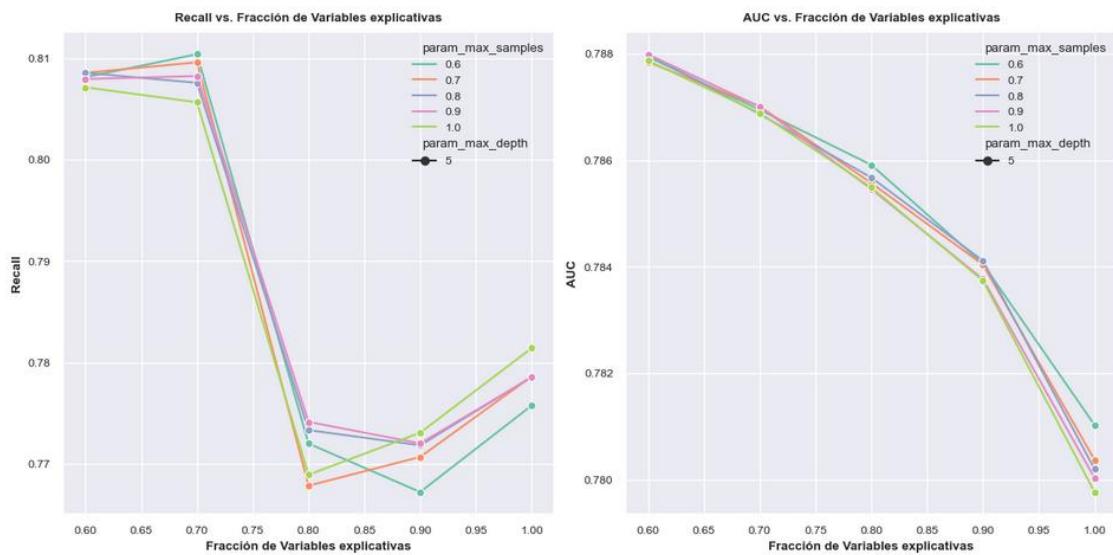


Figura C10: Recall promedio (izquierda) y AUC promedio (derecha) resultantes de la validación cruzada para RF en función de la fracción de variables explicativas utilizadas ( $max\_features$ ). Se comparan distintas fracciones de muestras ( $max\_samples$ , colores) y se mantiene fija la profundidad máxima ( $max\_depth$ )

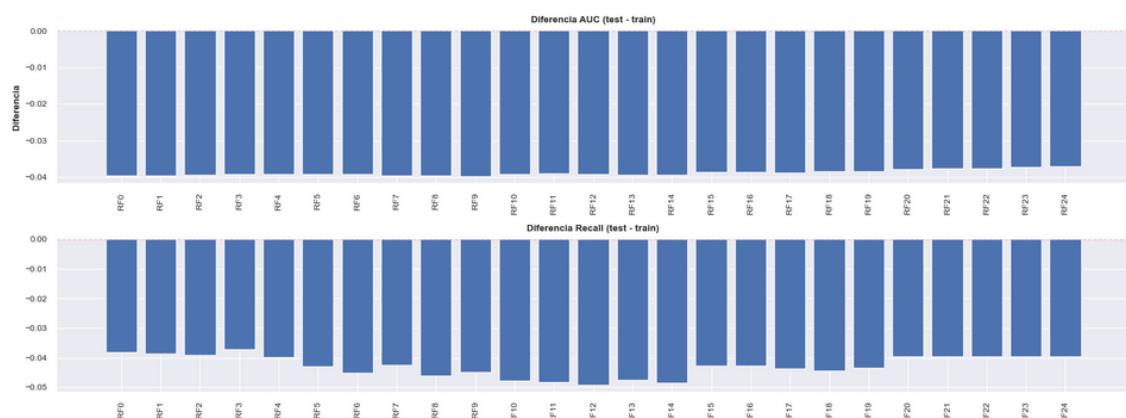


Figura C11: Comparación de la diferencia train-test para diferentes métricas en los modelos obtenidos de la búsqueda segunda búsqueda en RF.

## Predicción de Cáncer de Mama Basado en Factores Clínicos y Demográficos

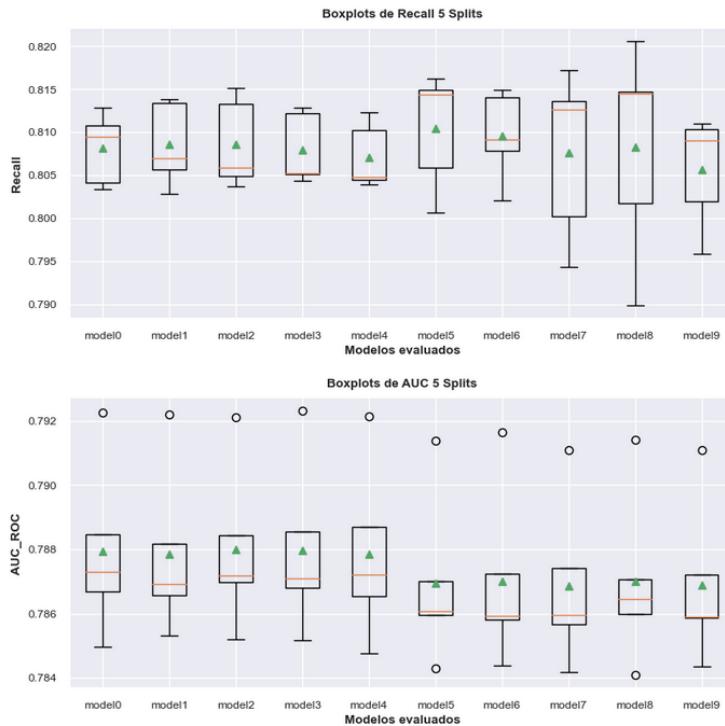


Figura C12: Boxplots del rendimiento de los modelos seleccionados tras la búsqueda para RF: distribución de Recall (arriba) y AUC (abajo) obtenidos en los distintos splits de validación cruzada. Se representan únicamente los 10 mejores modelos menos sobreentrenados y que maximizan las métricas

```
AUC para cada modelo:
RF0 | AUC_Train: 0.7885 | AUC_Test: 0.7490 | Diferencia: -0.0395
RF1 | AUC_Train: 0.7887 | AUC_Test: 0.7491 | Diferencia: -0.0396
RF2 | AUC_Train: 0.7883 | AUC_Test: 0.7490 | Diferencia: -0.0393
RF3 | AUC_Train: 0.7881 | AUC_Test: 0.7489 | Diferencia: -0.0392
RF4 | AUC_Train: 0.7882 | AUC_Test: 0.7490 | Diferencia: -0.0392
RF5 | AUC_Train: 0.7874 | AUC_Test: 0.7482 | Diferencia: -0.0392
RF6 | AUC_Train: 0.7876 | AUC_Test: 0.7483 | Diferencia: -0.0393
RF7 | AUC_Train: 0.7878 | AUC_Test: 0.7483 | Diferencia: -0.0395
RF8 | AUC_Train: 0.7879 | AUC_Test: 0.7483 | Diferencia: -0.0395
RF9 | AUC_Train: 0.7876 | AUC_Test: 0.7479 | Diferencia: -0.0397
```

```
Recall para cada modelo:
RF0 | Recall_Train: 0.8076 | Recall_Test: 0.7695 | Diferencia: -0.0381
RF1 | Recall_Train: 0.8067 | Recall_Test: 0.7680 | Diferencia: -0.0387
RF2 | Recall_Train: 0.8077 | Recall_Test: 0.7686 | Diferencia: -0.0391
RF3 | Recall_Train: 0.8047 | Recall_Test: 0.7675 | Diferencia: -0.0372
RF4 | Recall_Train: 0.8068 | Recall_Test: 0.7678 | Diferencia: -0.0398
RF5 | Recall_Train: 0.8070 | Recall_Test: 0.7639 | Diferencia: -0.0430
RF6 | Recall_Train: 0.8008 | Recall_Test: 0.7557 | Diferencia: -0.0450
RF7 | Recall_Train: 0.8096 | Recall_Test: 0.7671 | Diferencia: -0.0425
RF8 | Recall_Train: 0.8018 | Recall_Test: 0.7557 | Diferencia: -0.0461
RF9 | Recall_Train: 0.8048 | Recall_Test: 0.7598 | Diferencia: -0.0450
```

Figura C13: Desempeño de los modelos de RF seleccionados en términos de AUC y Recall. Para cada modelo se muestran las métricas en el conjunto de entrenamiento y prueba, así como la diferencia entre ambos.

Los modelos RF5 a RF9 presentan una mayor diferencia entre *train* y *test* en *Recall*, por lo que se descartan. Por otro lado, los modelos RF0 a RF4 muestran un rendimiento prácticamente idéntico. Entre ellos, RF1 se destaca por ser el modelo más sencillo y robusto, manteniendo un *Recall* elevado. Modelo ganador: RF1

```
RandomForestClassifier( bootstrap=True, class_weight="balanced",
                        n_estimators=100, max_depth=5,
                        max_features=0.6, max_samples=0.7,
                        min_samples_leaf=10, random_state=seed)
```

## Búsqueda para XGBoost

1. **Búsqueda inicial:** número de estimadores, profundidad de árbol y peso mínimo de los hijos

```
xgb = XGBClassifier(random_state=seed)

param_grid_xgb = {
    'n_estimators': [100, 200, 300],
    'max_depth': [3, 5, 10, 15, 20],
    'min_child_weight': [1, 3, 5, 10, 15],
    #'gamma': [0.1, 0.5, 1] # reducción de pérdida
    #'Learning_rate': [0.01, 0.1, 0.3]
    #'subsample': [0.6, 0.8, 1.0], #% rows
    #'colsample_bytree': [0.6, 0.8, 1.0], #% features
}

grid_xgb = GridSearchCV(xgb, param_grid_xgb, cv=cv5,
                        scoring=scoring_metrics ,
                        refit='roc_auc',
                        n_jobs=-1)
```

Figura C14: Configuración de la búsqueda GridSearchCV para el modelo XGBoost (XGBClassifier), explorando distintos valores de número de estimadores (`n_estimators`), profundidad máxima de los árboles (`max_depth`) y peso mínimo de los hijos (`min_child_weight`).

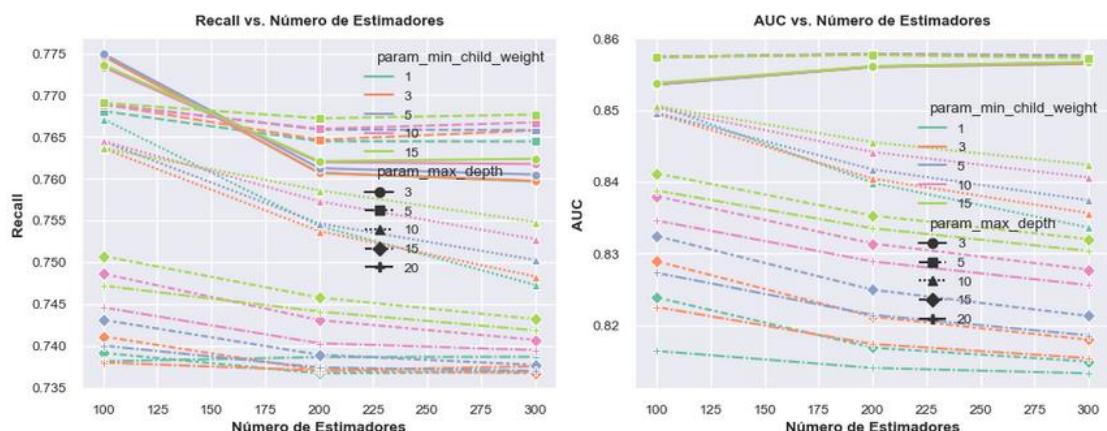


Figura C15: Recall promedio (izquierda) y AUC promedio (derecha) resultantes de la validación cruzada para XGBoost en función del número de estimadores (`n_estimators`). Se comparan distintos valores de `min_child_weight` (colores) y `max_depth` (estilos de línea)



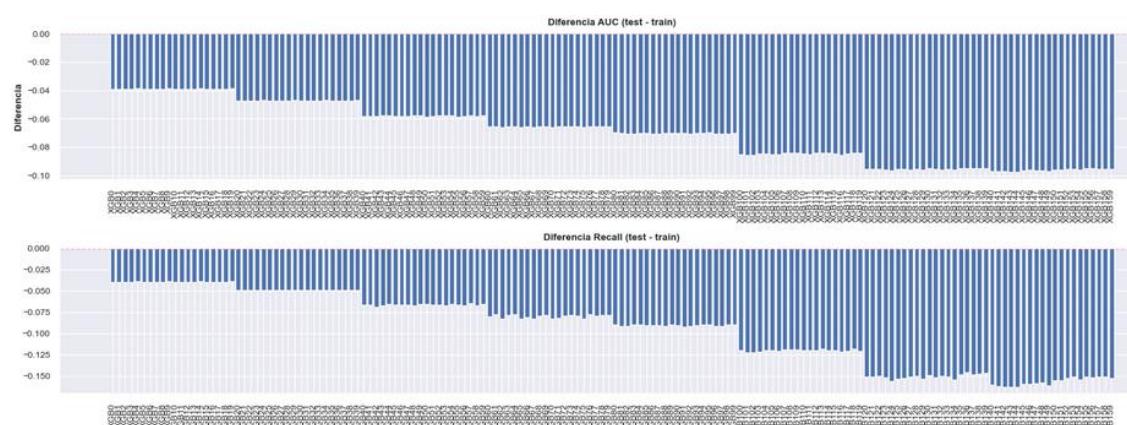
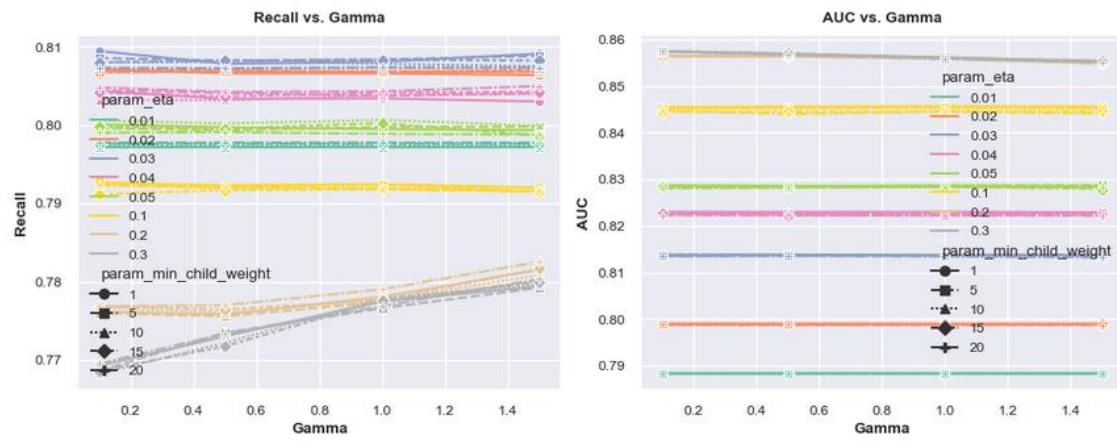
Figura C16: Comparación de la diferencia train-test para diferentes métricas en los modelos obtenidos de la búsqueda inicial en XGBoost.

## 2. Segunda búsqueda: peso de las variables, gamma y peso mínimo de los hijos

```

param_grid_xgb = {
    'n_estimators': [100],
    'max_depth': [5],
    'min_child_weight': [1, 5, 10, 15, 20],
    'eta': [0.01, 0.02, 0.03, 0.04, 0.05, 0.1, 0.2, 0.3] , # peso de las variables
    'gamma': [0.1, 0.5, 1, 1.5] # reducción de pérdida
    #'subsample': [0.6, 0.8, 1.0], % rows
    #'colsample_bytree': [0.6, 0.8, 1.0], % features
}
    
```

Figura C17 Nueva definición de la malla de hiperparámetros para XGboost, fijando `n_estimators`=100 y `max_depth`=5, y explorando distintos valores de `min_child_weight`, `eta` (tasa de aprendizaje o peso de las variables) y `gamma` (reducción mínima de pérdida para realizar un split).



## 3. Tercera búsqueda: fracción de variables, fracción de registros y peso mínimo de los hijos

```

param_grid_xgb = {
    'n_estimators': [100],
    'max_depth': [5],
    'min_child_weight': [1, 3, 5, 10, 15],
    'eta': [0.01], # peso de las variables
    'gamma': [0.2], # reducción de pérdida
    'subsample': [0.6, 0.7, 0.8, 0.9, 1.0], # % rows
    'colsample_bytree': [0.6, 0.7, 0.8, 0.9, 1.0], # % features
}
    
```

Figura C20: Nueva definición de la malla de hiperparámetros para XGBoost fijando `n_estimators`=100, `max_depth`=5, `min_child_weight` variable, `eta`=0.01 y `gamma`=0.2. Se exploran distintas fracciones de filas (`subsample`) y de características (`colsample_bytree`)

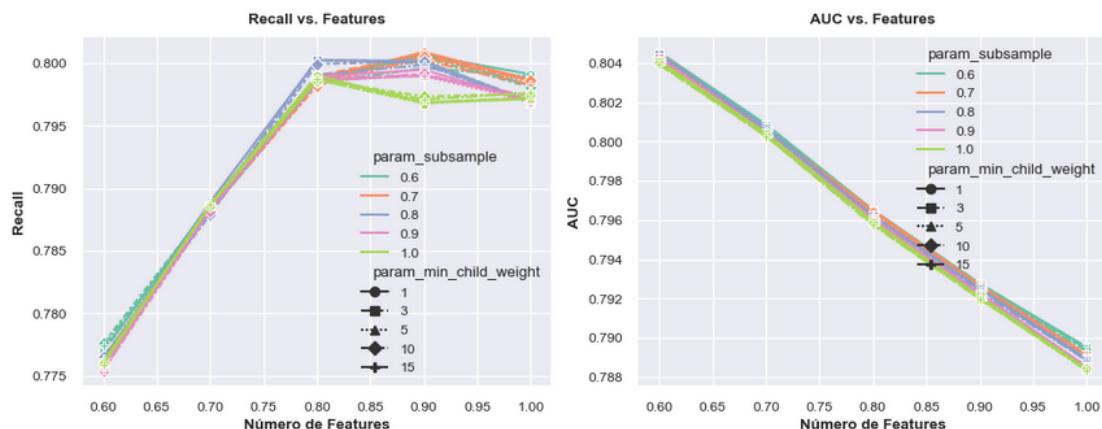
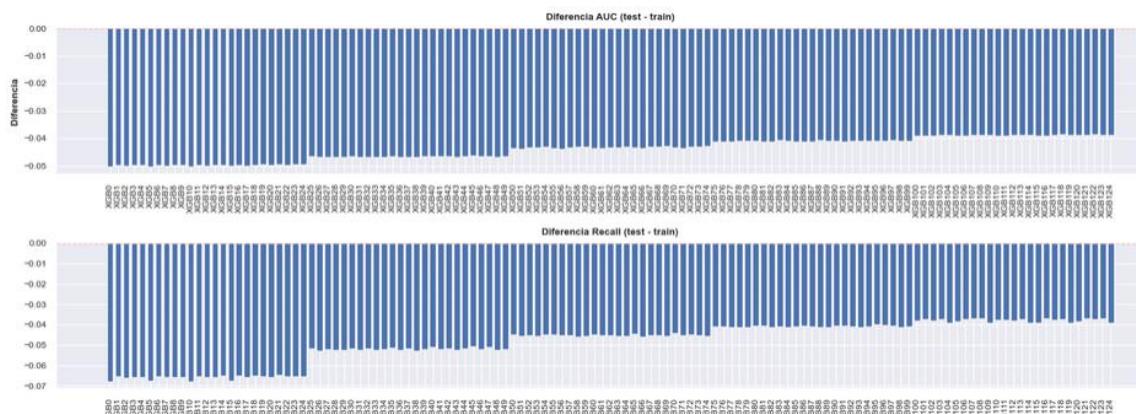


Figura C21: Recall promedio (izquierda) y AUC promedio (derecha) resultantes de la validación cruzada para XGBoost en función de la fracción de características utilizadas (`colsample_bytree`). Se comparan distintas fracciones de filas (`subsample`, colores) y distintos valores de `min_child_weight` (estilos de línea)



## Predicción de Cáncer de Mama Basado en Factores Clínicos y Demográficos

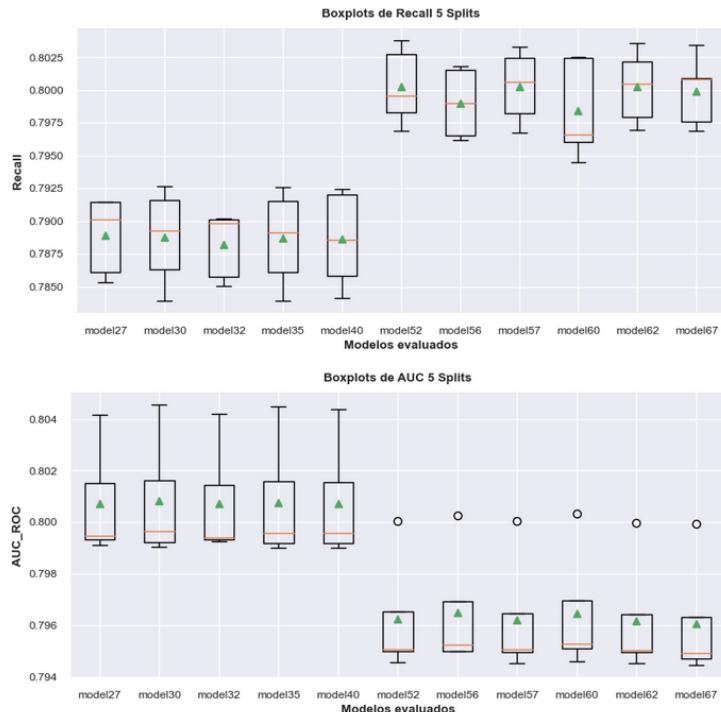


Figura C23: Boxplots del rendimiento de los modelos seleccionados tras la búsqueda para XGBoost: distribución de Recall (arriba) y AUC (abajo) obtenidos en los distintos splits de validación cruzada. Se representan únicamente los 11 mejores modelos menos sobreentrenados y que maximizan las métricas

```
AUC para cada modelo:
XGB27 | AUC_Train: 0.8015 | AUC_Test: 0.7546 | Diferencia: -0.0469
XGB30 | AUC_Train: 0.8014 | AUC_Test: 0.7547 | Diferencia: -0.0467
XGB32 | AUC_Train: 0.8014 | AUC_Test: 0.7546 | Diferencia: -0.0469
XGB35 | AUC_Train: 0.8013 | AUC_Test: 0.7547 | Diferencia: -0.0467
XGB40 | AUC_Train: 0.8012 | AUC_Test: 0.7547 | Diferencia: -0.0465
XGB52 | AUC_Train: 0.7970 | AUC_Test: 0.7535 | Diferencia: -0.0435
XGB56 | AUC_Train: 0.7973 | AUC_Test: 0.7535 | Diferencia: -0.0438
XGB57 | AUC_Train: 0.7970 | AUC_Test: 0.7535 | Diferencia: -0.0434
XGB60 | AUC_Train: 0.7973 | AUC_Test: 0.7537 | Diferencia: -0.0436
XGB62 | AUC_Train: 0.7969 | AUC_Test: 0.7536 | Diferencia: -0.0434
XGB67 | AUC_Train: 0.7968 | AUC_Test: 0.7535 | Diferencia: -0.0432

Recall para cada modelo:
XGB27 | Recall_Train: 0.7899 | Recall_Test: 0.7377 | Diferencia: -0.0521
XGB30 | Recall_Train: 0.7895 | Recall_Test: 0.7377 | Diferencia: -0.0518
XGB32 | Recall_Train: 0.7900 | Recall_Test: 0.7381 | Diferencia: -0.0520
XGB35 | Recall_Train: 0.7892 | Recall_Test: 0.7376 | Diferencia: -0.0516
XGB40 | Recall_Train: 0.7888 | Recall_Test: 0.7378 | Diferencia: -0.0510
XGB52 | Recall_Train: 0.7993 | Recall_Test: 0.7539 | Diferencia: -0.0455
XGB56 | Recall_Train: 0.7992 | Recall_Test: 0.7536 | Diferencia: -0.0455
XGB57 | Recall_Train: 0.7992 | Recall_Test: 0.7539 | Diferencia: -0.0454
XGB60 | Recall_Train: 0.7989 | Recall_Test: 0.7548 | Diferencia: -0.0449
XGB62 | Recall_Train: 0.7992 | Recall_Test: 0.7539 | Diferencia: -0.0453
XGB67 | Recall_Train: 0.7992 | Recall_Test: 0.7537 | Diferencia: -0.0454
```

Figura C24: Desempeño de los modelos de Xgboost seleccionados en términos de AUC y Recall. Para cada modelo se muestran las métricas en el conjunto de entrenamiento y prueba, así como la diferencia entre ambos.

Los modelos XGBoost XGB52 a XGB67 presentan diferencias mínimas entre los valores de *train* y *test*, indicando buena generalización. Entre ellos, XGB67 se destaca como el más robusto, mostrando el mejor equilibrio y rendimiento para ambas métricas evaluadas (AUC y Recall). Modelo ganador: XGB67

```
XGBClassifier(n_estimators=100, eta=0.01, gamma=0.2, max_depth=5,
               min_child_weight=10, subsample= 0.8, colsample_bytree=0.8,
               random_state=seed)
```

## Búsqueda para LightGBM

### 1. Búsqueda inicial: estimadores, profundidad máxima y número mínimo de muestras por hoja

```

lgb = LGBMClassifier(random_state=seed, verbose=-1, n_jobs=-1, class_weight='balanced')

param_grid_lgb = {
    "n_estimators": [100, 150, 200, 250, 300],
    "max_depth": [3, 5, 7, 10],
    "min_child_samples": [1, 5, 10, 15],
    # "learning_rate": [0.1, 0.05, 0.01],
    # "reg_alpha": [0.1],
    # "reg_lambda": [0.1, 0.5, 1.0],
    # "subsample": [0.6, 0.7, 0.8, 0.9, 1.0],
    # "colsample_bytree": [0.6, 0.7, 0.8, 0.9, 1.0],
    # "boosting_type": ['gbdt', 'dart'],
}

grid_lgb = GridSearchCV(lgb, param_grid_lgb, cv=cv5,
                        scoring=scoring_metrics,
                        refit='roc_auc',
                        n_jobs=-1)
    
```

Figura C25: Configuración de la búsqueda GridSearchCV para el modelo LightGBM (LGBMClassifier), explorando distintos valores de número de estimadores (n\_estimators), profundidad máxima (max\_depth) y número mínimo de muestras por hoja (min\_child\_samples).

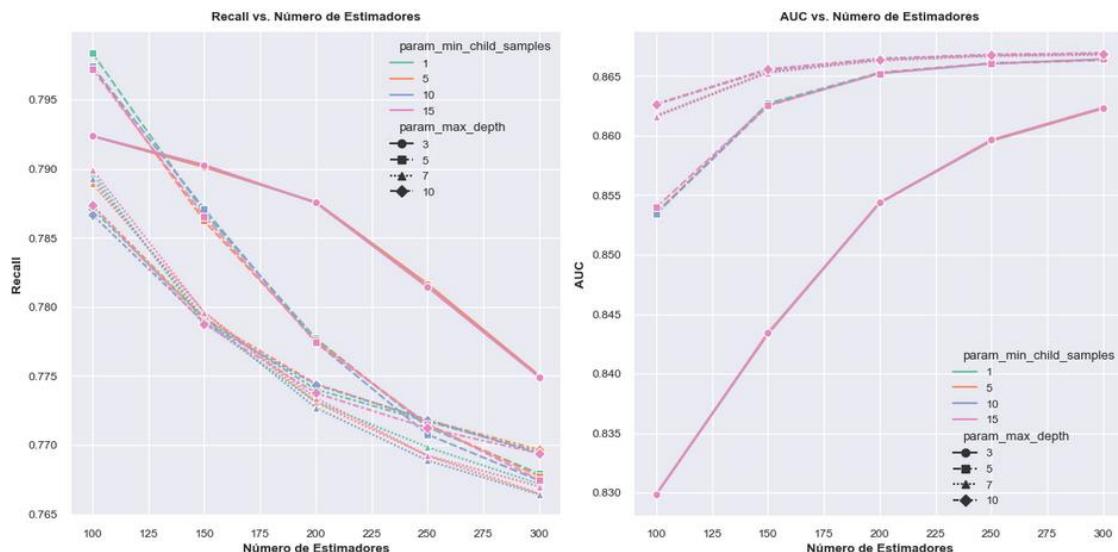


Figura C26: Recall promedio (izquierda) y AUC promedio (derecha) resultantes de la validación cruzada para LightGBM en función del número de estimadores (n\_estimators). Se comparan distintos valores de min\_child\_samples (colores) y max\_depth (estilos de línea)

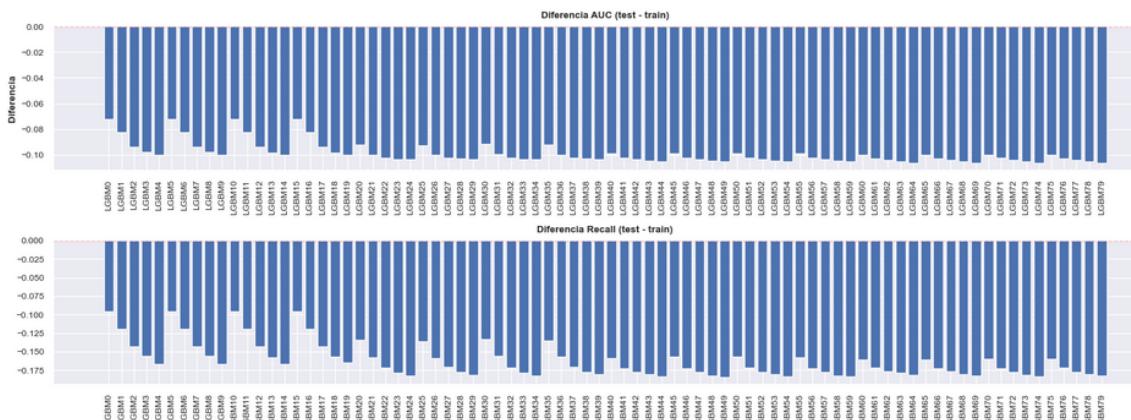


Figura C27: Comparación de la diferencia train-test para diferentes métricas en los modelos obtenidos de la primera búsqueda en LightGBM

## 2. Segunda búsqueda: learning rate y regularizaciones alpha (L1) y lambda (L2)

```

param_grid_lgb = {
    "n_estimators": [100],
    "max_depth": [7],
    "min_child_samples": [15],
    "learning_rate": [0.01, 0.02, 0.04, 0.06, 0.08, 0.1, 0.2],
    "reg_alpha": [0, 0.1, 0.5, 1],
    "reg_lambda": [0, 0.1, 0.5, 1],
    # "subsample": [0.6, 0.7, 0.8, 0.9, 1.0],
    # "colsample_bytree": [0.6, 0.7, 0.8, 0.9, 1.0],
    # "boosting_type": ['gbdt', 'dart'],
}
    
```

Figura C28: Nueva definición de la malla de hiperparámetros para LightGBM, fijando `n_estimators`=100, `max_depth`=7 y `min_child_samples`=15, y explorando distintos valores de `learning_rate`, `reg_alpha` y `reg_lambda`.

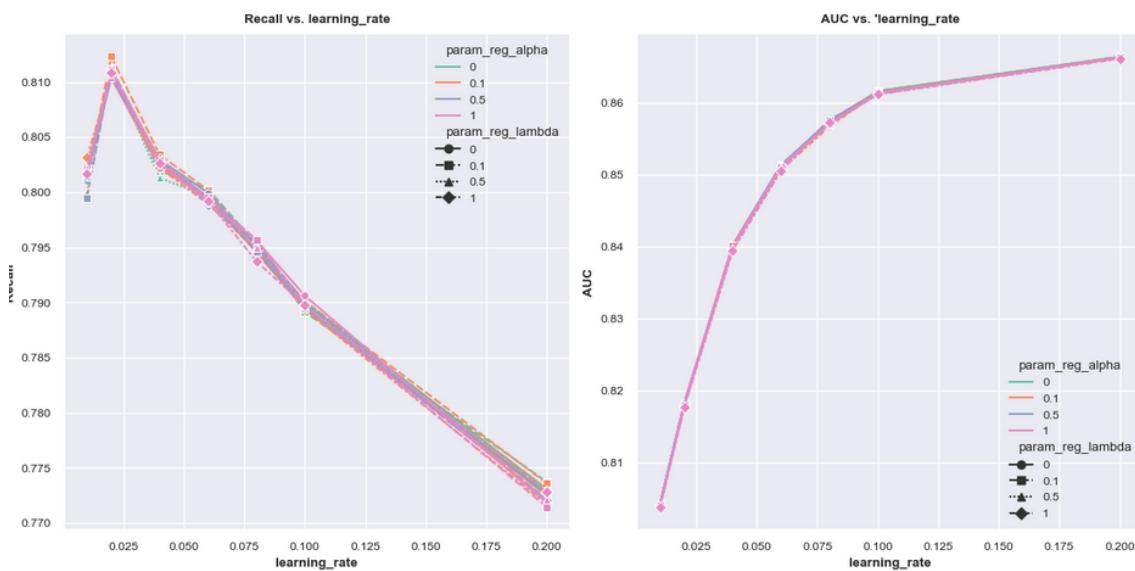


Figura C29: Recall promedio (izquierda) y AUC promedio (derecha) resultantes de la validación cruzada para LightGBM en función de la tasa de aprendizaje (`learning_rate`). Se comparan distintos valores de regularización L1 (`reg_alpha`, colores) y L2 (`reg_lambda`, estilos de línea)



Figura C30: Comparación de la diferencia `train-test` para diferentes métricas en los modelos obtenidos de la segunda búsqueda en LightGBM

### 3. Tercera búsqueda: fracción de variables, fracción de muestras y tipos de boosting

```
param_grid_lgb = {
    "n_estimators": [100],
    "max_depth": [7],
    "min_child_samples": [15],
    "learning_rate": [0.01],
    "reg_alpha": [0.1],
    # "reg_Lambda": [0],
    "subsample": [0.6, 0.7, 0.8, 0.9, 1.0],
    "colsample_bytree": [0.6, 0.7, 0.8, 0.9, 1.0],
    "boosting_type": ['gbdt', 'dart'],
}
```

Figura C31: Nueva definición de la malla de hiperparámetros para LightGBM fijando `n_estimators`=100, `max_depth`=7, `min_child_samples`=15, `learning_rate`=0.01 y `reg_alpha`=0.1. Se exploran distintas fracciones de filas (`subsample`) y de características (`colsample_bytree`), así como los tipos de boosting (`gbdt` y `dart`),

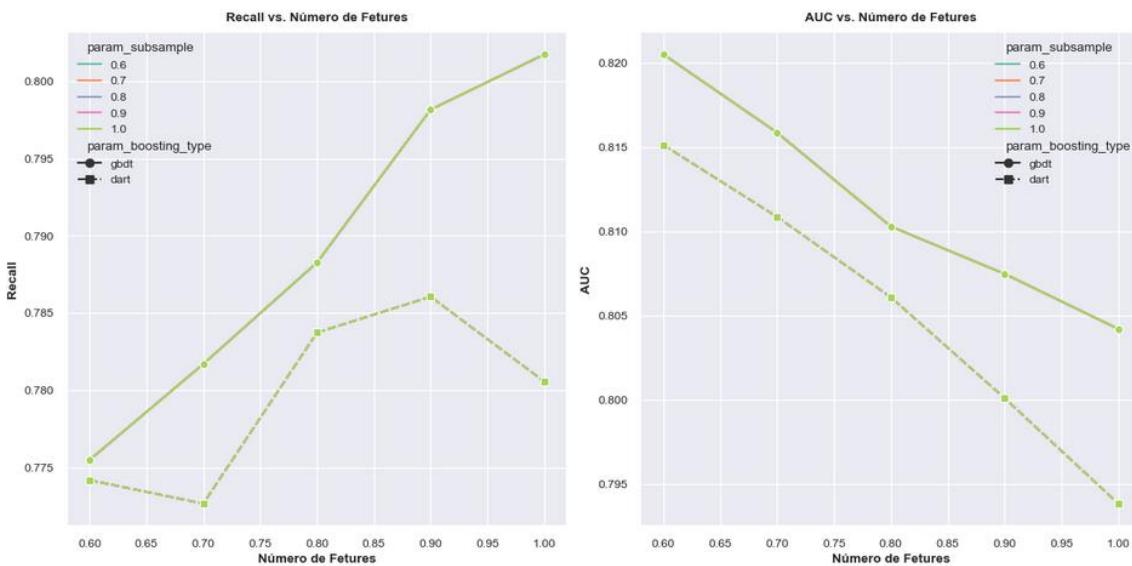


Figura C32: Recall promedio (izquierda) y AUC promedio (derecha) resultantes de la validación cruzada para LightGBM en función del número de características utilizadas (`colsample_bytree`). Se comparan distintas fracciones de filas (`subsample`, colores) y los tipos de boosting (`boosting_type`, estilos de línea)

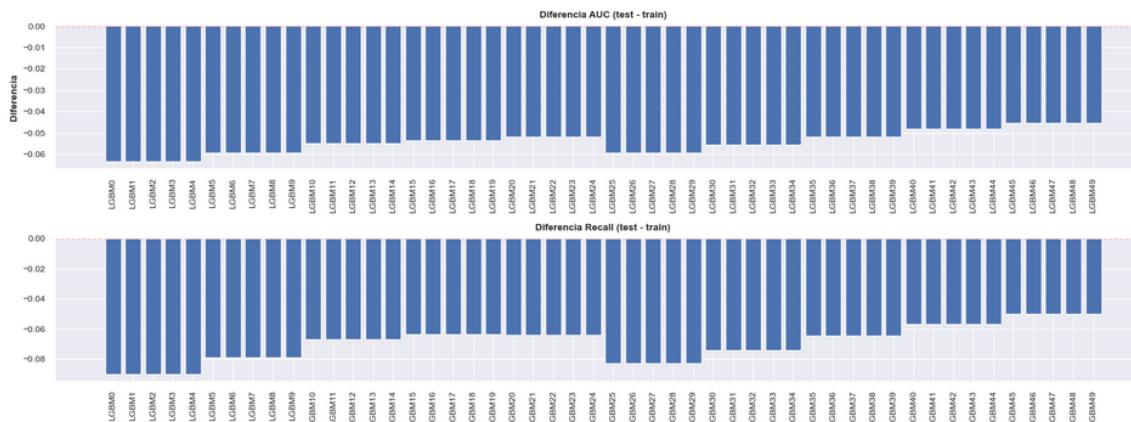


Figura C33: Comparación de la diferencia `train-test` para diferentes métricas en los modelos obtenidos de la tercera búsqueda en LightGBM

## Predicción de Cáncer de Mama Basado en Factores Clínicos y Demográficos

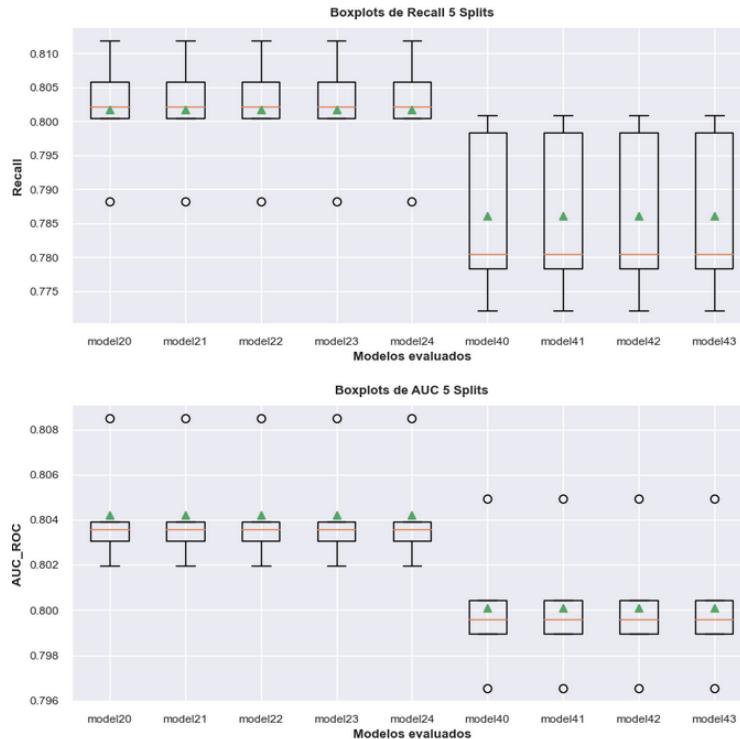


Figura C34: Boxplots del rendimiento de los modelos seleccionados tras la búsqueda para LightBGM: distribución de Recall (arriba) y AUC (abajo) obtenidos en los distintos splits de validación cruzada. Se representan únicamente los 9 mejores modelos menos sobreentrenados y que maximizan las métricas

AUC para cada modelo:  
 LGB20 | AUC\_Train: 0.8046 | AUC\_Test: 0.7528 | Diferencia: -0.0518  
 LGB21 | AUC\_Train: 0.8046 | AUC\_Test: 0.7528 | Diferencia: -0.0518  
 LGB22 | AUC\_Train: 0.8046 | AUC\_Test: 0.7528 | Diferencia: -0.0518  
 LGB23 | AUC\_Train: 0.8046 | AUC\_Test: 0.7528 | Diferencia: -0.0518  
 LGB24 | AUC\_Train: 0.8046 | AUC\_Test: 0.7528 | Diferencia: -0.0518  
 LGB40 | AUC\_Train: 0.8004 | AUC\_Test: 0.7525 | Diferencia: -0.0480  
 LGB41 | AUC\_Train: 0.8004 | AUC\_Test: 0.7525 | Diferencia: -0.0480  
 LGB42 | AUC\_Train: 0.8004 | AUC\_Test: 0.7525 | Diferencia: -0.0480  
 LGB43 | AUC\_Train: 0.8004 | AUC\_Test: 0.7525 | Diferencia: -0.0480

Recall para cada modelo:  
 LGB20 | Recall\_Train: 0.8036 | Recall\_Test: 0.7397 | Diferencia: -0.0639  
 LGB21 | Recall\_Train: 0.8036 | Recall\_Test: 0.7397 | Diferencia: -0.0639  
 LGB22 | Recall\_Train: 0.8036 | Recall\_Test: 0.7397 | Diferencia: -0.0639  
 LGB23 | Recall\_Train: 0.8036 | Recall\_Test: 0.7397 | Diferencia: -0.0639  
 LGB24 | Recall\_Train: 0.8036 | Recall\_Test: 0.7397 | Diferencia: -0.0639  
 LGB40 | Recall\_Train: 0.7794 | Recall\_Test: 0.7226 | Diferencia: -0.0568  
 LGB41 | Recall\_Train: 0.7794 | Recall\_Test: 0.7226 | Diferencia: -0.0568  
 LGB42 | Recall\_Train: 0.7794 | Recall\_Test: 0.7226 | Diferencia: -0.0568  
 LGB43 | Recall\_Train: 0.7794 | Recall\_Test: 0.7226 | Diferencia: -0.0568

Figura C35: Desempeño de los modelos de LightGBM seleccionados en términos de AUC y Recall. Para cada modelo se muestran las métricas en el conjunto de entrenamiento y prueba, así como la diferencia entre ambos

Los modelos LightGBM LGB40 a LGB43 presentan diferencias menores entre *train* y *test*, indicando buena generalización, pero su *Recall* resulta menos robusto. Aunque muestran mayor estabilidad, los modelos LGB20–LGB24 ofrecen un *Recall* superior, que es la métrica prioritaria en este caso. Entre ellos, LGB20 destaca por su menor variabilidad, combinando un buen desempeño con consistencia entre los distintos *splits* de validación. Modelo ganador: LGB20

```
LGBMClassifier(n_estimators=100, boosting_type='gbdt', colsample_bytree=1.0,
               learning_rate=0.01, max_depth=7, min_child_samples=15,
               reg_alpha = 0.1, subsample=0.6, random_state=seed,
               verbose=-1, n_jobs=-1, class_weight='balanced')
```

## Búsqueda para Stacking

### 1. Búsqueda inicial: parámetro de regularización C y solver

```

# Elegimos una regresión (Lógistica pues estamos en clasificación) como metamodelo.
# Es un modelo que de una forma muy simple nos permite combinar modelos base.
# Ayuda a la interpretabilidad final
meta_modelo = LogisticRegression(random_state=seed)

# Construimos el stacking: para ello, se entrena cada modelo base sobre datos train,
# y las predicciones que se consiguen en CV se usan para entrar el metamodelo
stacking = StackingClassifier(
    estimators=modelos_base_stacking,
    final_estimator=meta_modelo,
    cv=cv5, n_jobs=1,
    # Si se activa (True), se pasan las características originales al meta-modelo
    passthrough=False
)

# Tuneo de los hiperparámetros propios del stacking.

param_grid = {
    'final_estimator__C': [0.001, 0.005, 0.01, 0.03, 0.06, 0.1, 0.5, 1.0, 2, 5],
    'final_estimator__solver': ['lbfgs', 'saga'],
    'final_estimator__class_weight': ['balanced']
}

grid_search = GridSearchCV(estimator=stacking,
                           param_grid=param_grid,
                           cv=cv5,
                           n_jobs=1,
                           scoring=['recall', 'roc_auc'],
                           refit='roc_auc',
                           verbose=5)

```

Figura C36: Código del stacking classifier utilizado: los modelos base se entrenan sobre los datos de entrenamiento y sus predicciones en validación cruzada sirven como entradas para el metamodelo de regresión logística. Se realiza una búsqueda en rejilla de hiperparámetros sobre el metamodelo, evaluando distintos valores de regularización (C), solvers (lbfgs y saga) y ponderación de clases (balanced).

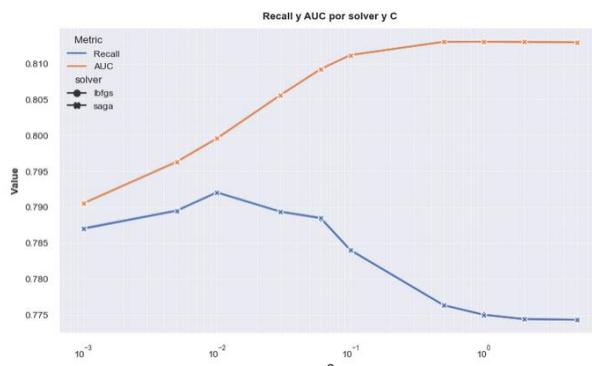


Figura C37: Recall promedio (izquierda) y AUC promedio (derecha) resultantes de la validación cruzada para el metamodelo de regresión logística en el stacking classifier en función del parámetro de regularización C (escala logarítmica) y del solver utilizado (lbfgs o saga).

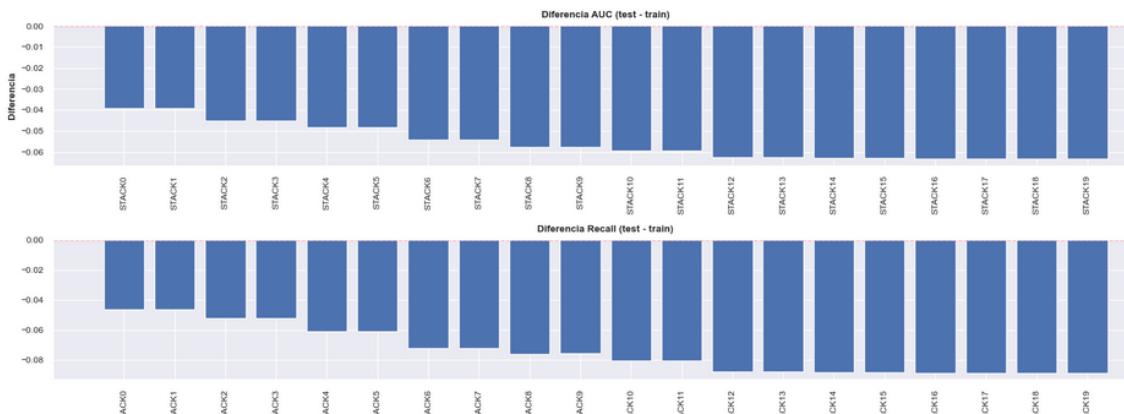


Figura C38: Comparación de la diferencia train-test para diferentes métricas en los modelos obtenidos en la búsqueda para el metamodelo de regresión logística en el stacking classifier

## Predicción de Cáncer de Mama Basado en Factores Clínicos y Demográficos

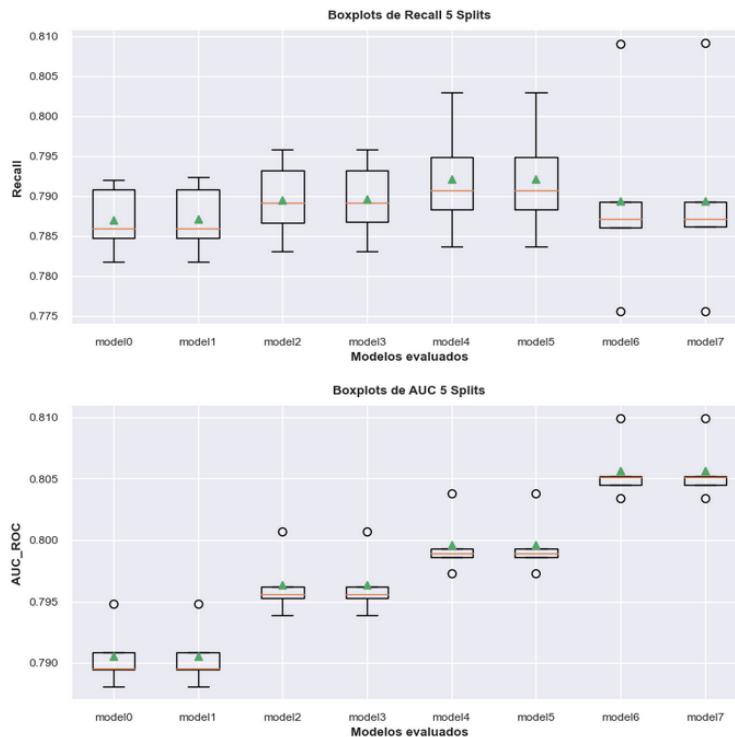


Figura C39: Boxplots del rendimiento de los modelos seleccionados tras la búsqueda para el metamodelo del stacking: distribución de Recall (arriba) y AUC (abajo) obtenidos en los distintos splits de validación cruzada. Se representan únicamente los 8 mejores modelos menos sobreentrenados y que maximizan las métricas

```
AUC para cada modelo:
STACK0 | AUC_Train: 0.7914 | AUC_Test: 0.7519 | Diferencia: -0.0394
STACK1 | AUC_Train: 0.7914 | AUC_Test: 0.7519 | Diferencia: -0.0394
STACK2 | AUC_Train: 0.7974 | AUC_Test: 0.7523 | Diferencia: -0.0450
STACK3 | AUC_Train: 0.7974 | AUC_Test: 0.7523 | Diferencia: -0.0450
STACK4 | AUC_Train: 0.8007 | AUC_Test: 0.7525 | Diferencia: -0.0483
STACK5 | AUC_Train: 0.8007 | AUC_Test: 0.7525 | Diferencia: -0.0483
STACK6 | AUC_Train: 0.8068 | AUC_Test: 0.7524 | Diferencia: -0.0544
STACK7 | AUC_Train: 0.8068 | AUC_Test: 0.7524 | Diferencia: -0.0544

Recall para cada modelo:
STACK0 | Recall_Train: 0.7869 | Recall_Test: 0.7404 | Diferencia: -0.0465
STACK1 | Recall_Train: 0.7869 | Recall_Test: 0.7404 | Diferencia: -0.0465
STACK2 | Recall_Train: 0.7879 | Recall_Test: 0.7359 | Diferencia: -0.0520
STACK3 | Recall_Train: 0.7879 | Recall_Test: 0.7359 | Diferencia: -0.0520
STACK4 | Recall_Train: 0.7907 | Recall_Test: 0.7297 | Diferencia: -0.0609
STACK5 | Recall_Train: 0.7906 | Recall_Test: 0.7297 | Diferencia: -0.0609
STACK6 | Recall_Train: 0.7900 | Recall_Test: 0.7177 | Diferencia: -0.0723
STACK7 | Recall_Train: 0.7901 | Recall_Test: 0.7177 | Diferencia: -0.0723
```

Figura C40: Desempeño de los modelos de stacking seleccionados en términos de AUC y Recall. Para cada modelo se muestran las métricas en el conjunto de entrenamiento y prueba, así como la diferencia entre ambos

Los modelos 2 y 3 muestran un mejor equilibrio entre Recall y AUC, con diferencias mínimas entre *train* y *test*, lo que indica un desempeño estable y buena capacidad de generalización, comparable al resto de los modelos, pero con menor riesgo de sobreajuste. Entre ellos, el modelo 3 destaca por presentar menor desviación estándar en Recall, ofreciendo así mayor consistencia en la métrica más relevante. Modelo Ganador: STCK3

```
StackingClassifier(estimators=modelos_base_stacking,
                  final_estimator=LogisticRegression(C=0.005, solver='saga',
                                                     class_weight='balanced', random_state=seed),
                  passthrough=False)
```

## Anexo D: Comparación Entre los Modelos Candidatos

### ■ Comparación en el punto de corte 0.5 (seed = 12345)

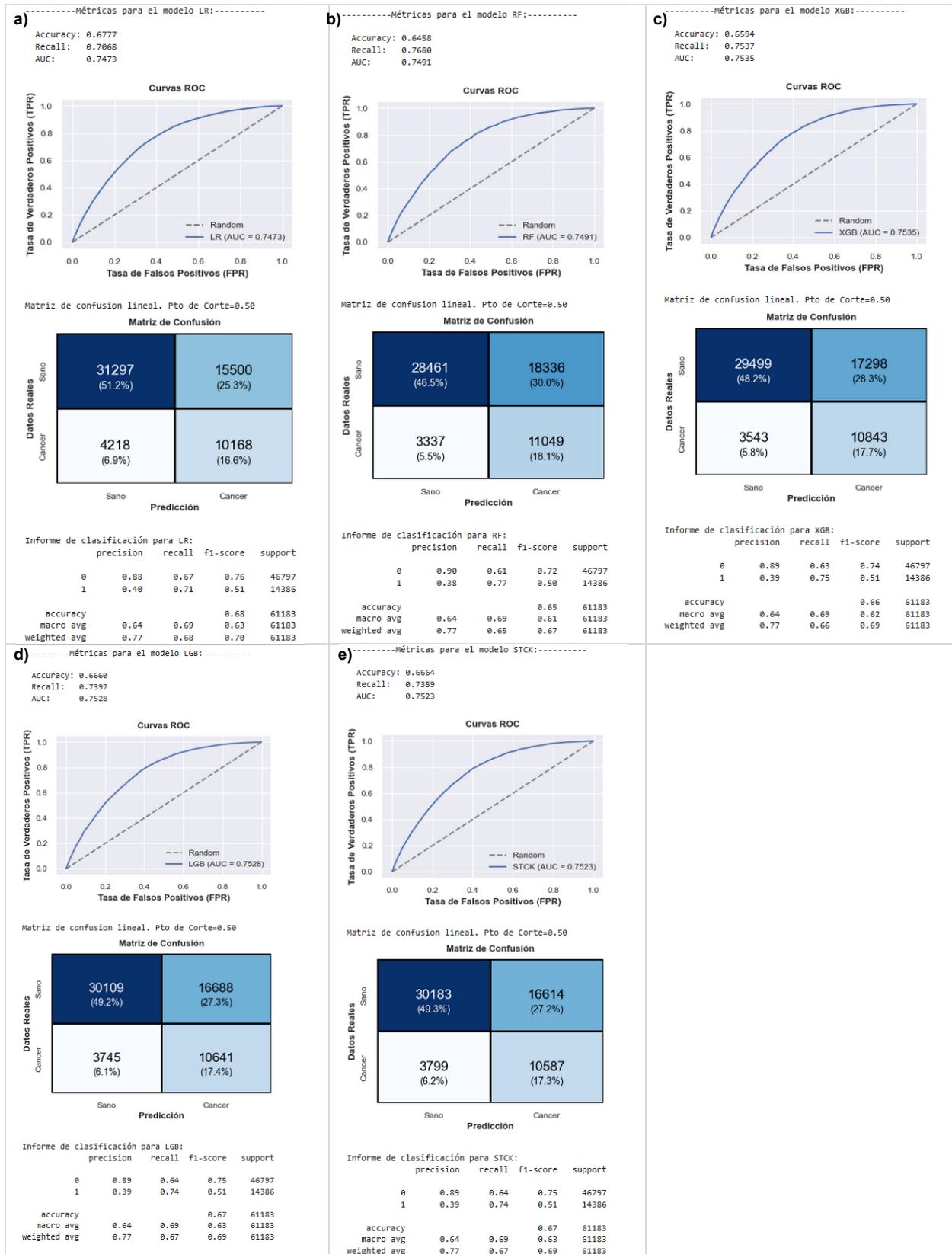


Figura D1: Resultados de la evaluación en test para los modelos seleccionados. Se muestran las métricas principales (Accuracy, Recall, AUC), la curva ROC con el área bajo la curva correspondiente, la matriz de confusión para el umbral de decisión (por defecto 0.5) y el informe de clasificación detallado. Modelos: a) RL, b) RF, c) XGB, d) LGB y e) STCK

# Predicción de Cáncer de Mama Basado en Factores Clínicos y Demográficos

## ■ Comparación en el índice de Youden (seed = 12345)

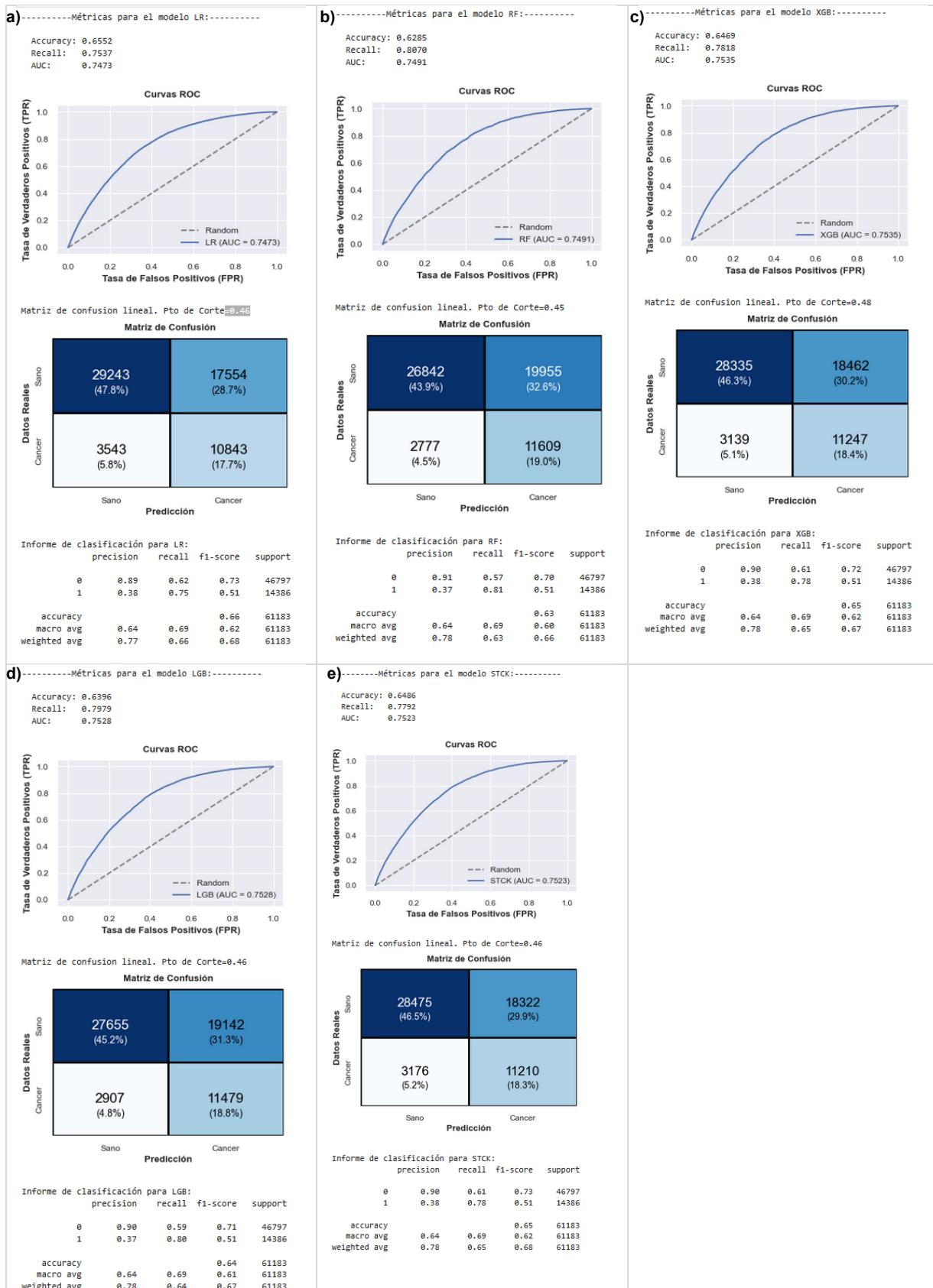


Figura D2: Resultados de la evaluación en test para los modelos seleccionados. Se muestran las métricas principales (Accuracy, Recall, AUC), la curva ROC con el área bajo la curva correspondiente, la matriz de confusión para el umbral de Youden y el informe de clasificación detallado. Modelos: a) RL, b) RF, c) XGB, d) LGB y e) STCK

## Predicción de Cáncer de Mama Basado en Factores Clínicos y Demográficos

### ■ Comparación con otra semilla (seed = 12545)

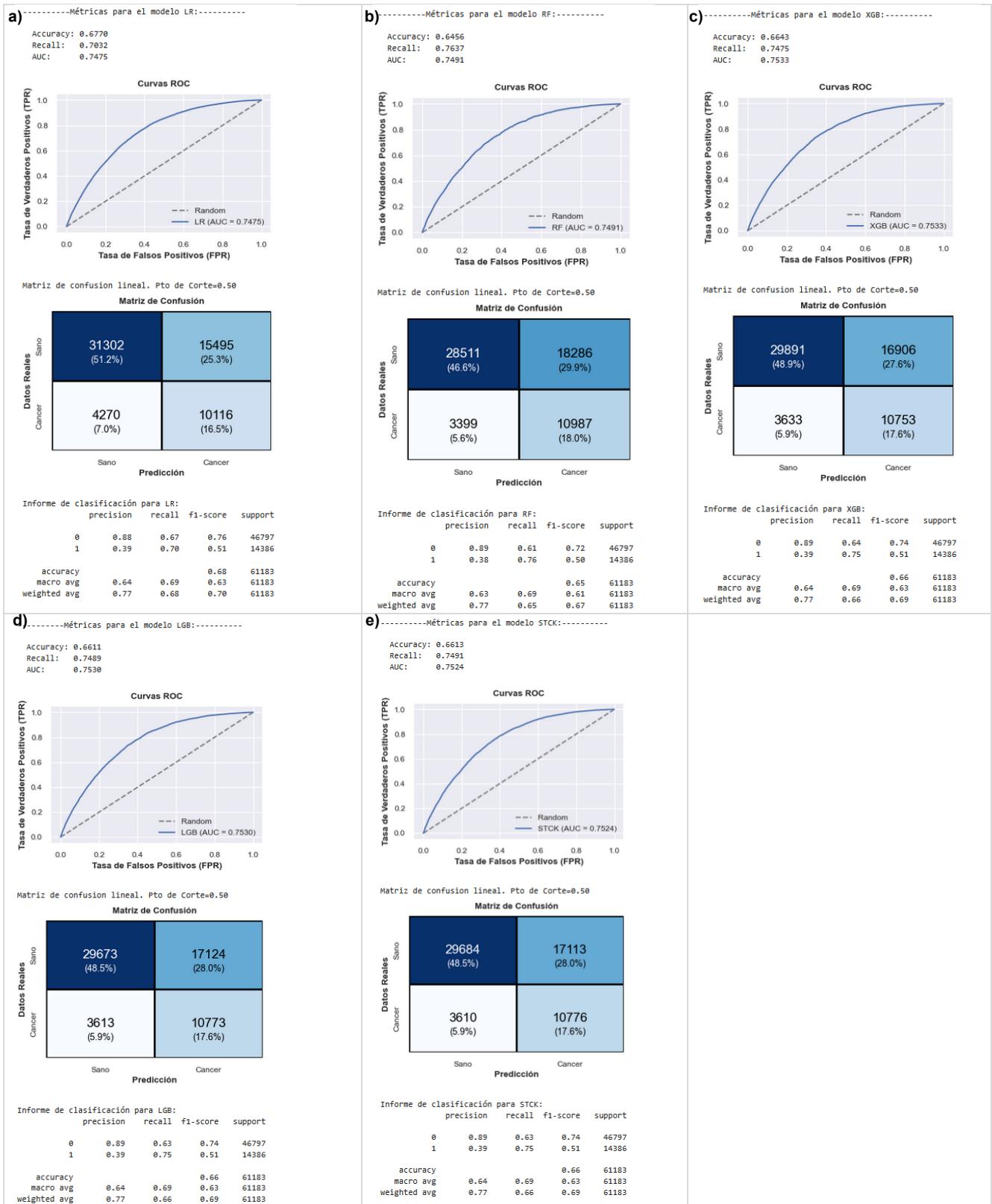


Figura D3: Resultados de la evaluación en el conjunto de test para los modelos seleccionados, utilizando una semilla aleatoria distinta en la partición de los datos y entrenamiento de los modelos. Se incluyen las métricas principales (Accuracy, Recall, AUC), la curva ROC y la matriz de confusión para un umbral estándar.

## Predicción de Cáncer de Mama Basado en Factores Clínicos y Demográficos

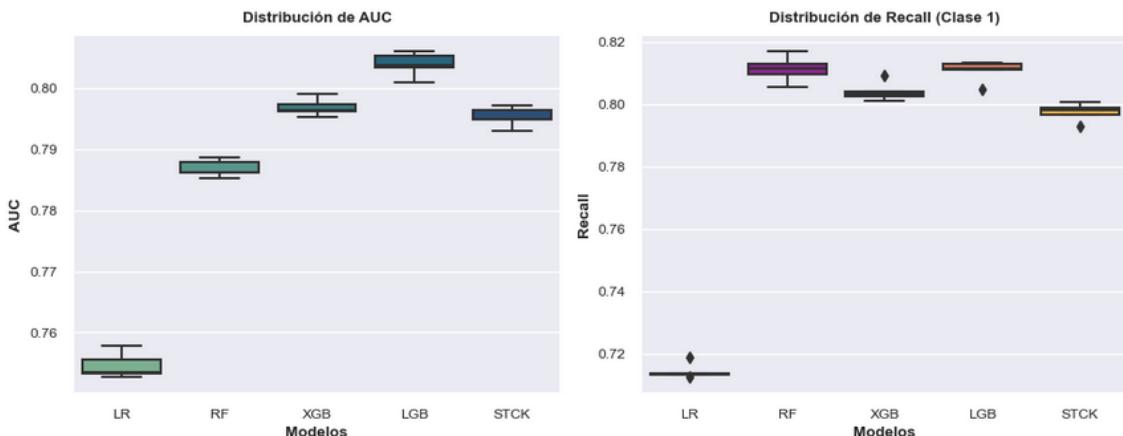


Figura D4: Distribución de las métricas AUC y Recall obtenidas mediante validación cruzada de 5 folds para los modelos comparados. El entrenamiento y la partición de los datos se realizaron con una semilla distinta (seed = 12545),

LR	AUC_Train: 0.7546	AUC_Test: 0.7475	Diferencia: -0.0070
LR	Recall_Train: 0.7143	Recall_Test: 0.7032	Diferencia: -0.0111
<hr/>			
RF	AUC_Train: 0.7869	AUC_Test: 0.7491	Diferencia: -0.0378
RF	Recall_Train: 0.8064	Recall_Test: 0.7637	Diferencia: -0.0427
<hr/>			
XGB	AUC_Train: 0.7977	AUC_Test: 0.7533	Diferencia: -0.0444
XGB	Recall_Train: 0.7996	Recall_Test: 0.7475	Diferencia: -0.0522
<hr/>			
LGB	AUC_Train: 0.8044	AUC_Test: 0.7530	Diferencia: -0.0515
LGB	Recall_Train: 0.8103	Recall_Test: 0.7489	Diferencia: -0.0615
<hr/>			
STCK	AUC_Train: 0.7967	AUC_Test: 0.7524	Diferencia: -0.0443
STCK	Recall_Train: 0.7992	Recall_Test: 0.7491	Diferencia: -0.0502

Figura D5: Rendimiento de los modelos finales en los conjuntos de entrenamiento (train) y prueba (test). Para cada modelo se reportan las métricas de Recall y AUC, junto con la diferencia entre ambos conjuntos, con el fin de evaluar posibles indicios de sobreajuste.

## Anexo E: Interpretabilidad del Modelo Ganador

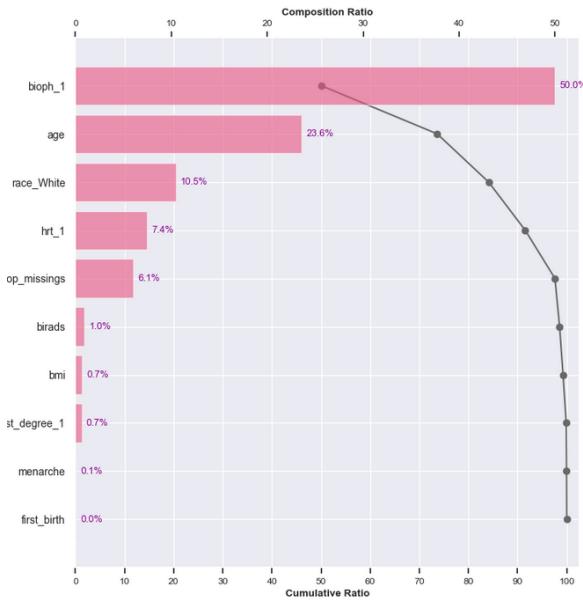


Figura E1: Contribución de las características del modelo evaluado mediante valores SHAP. El gráfico combina barras horizontales que muestran el porcentaje de contribución individual de cada feature (Composition Ratio, eje superior) y la contribución acumulada (Cumulative Ratio, eje inferior). Las variables se ordenan de mayor a menor impacto, y se destacan los valores porcentuales al final de cada barra para facilitar la interpretación visual de las variables más influyentes.

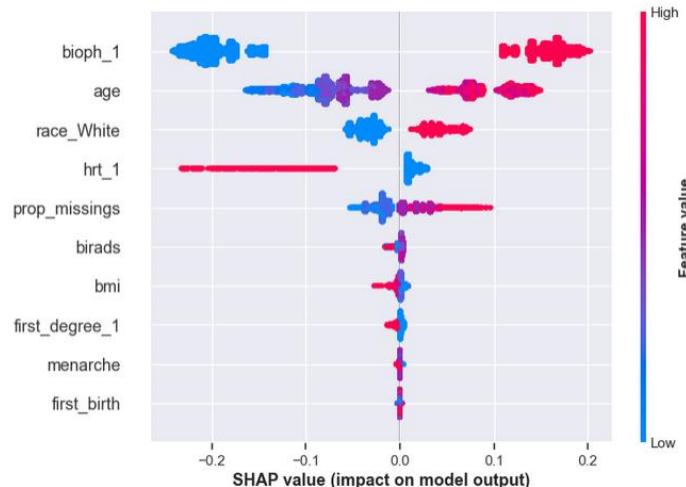


Figura E2: Distribución del impacto de cada característica en las predicciones del modelo para la clase 1, evaluada mediante valores SHAP. Cada punto corresponde a un paciente y su color indica el valor de la característica (rojo: alto, azul: bajo). El eje X mide el impacto en la predicción: valores positivos se asocian a mayor riesgo de cáncer de mama, mientras que valores negativos se relacionan con menor probabilidad.

## Predicción de Cáncer de Mama Basado en Factores Clínicos y Demográficos

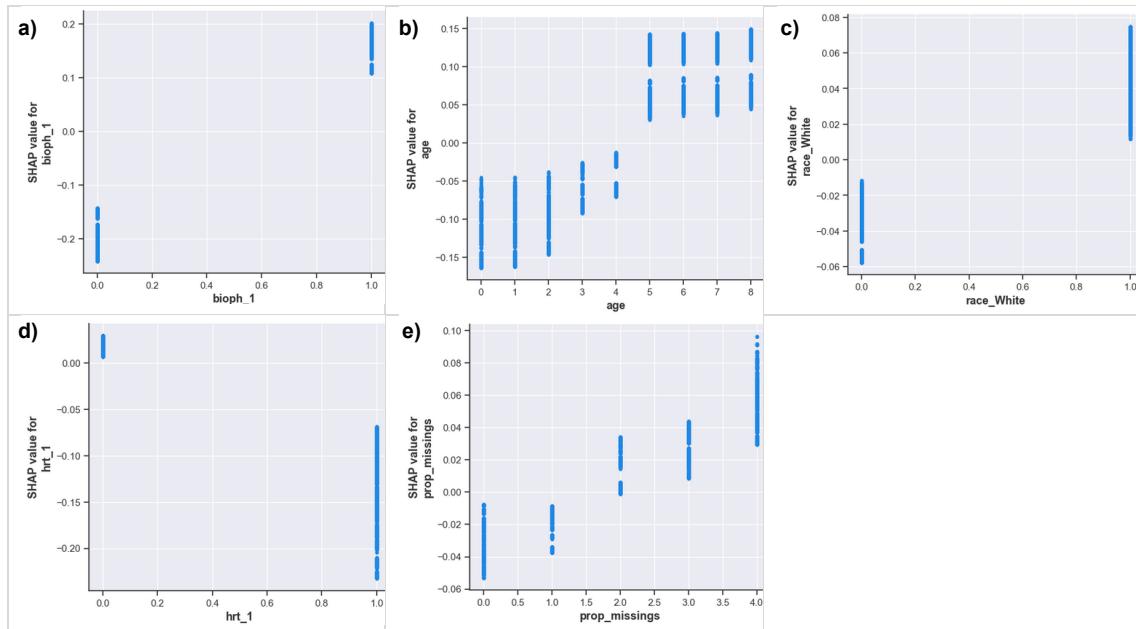


Figura E3: Gráficos de dependencia de las principales características seleccionadas: a) bioph\_1, b) age, c) race\_White, d) hrt\_1 y d) prop\_missings). Cada gráfico muestra cómo el valor de la variable impacta en la predicción del modelo, evaluada mediante valores SHAP, permitiendo visualizar la relación entre la variable y la salida predicha, así como posibles interacciones con otras variables

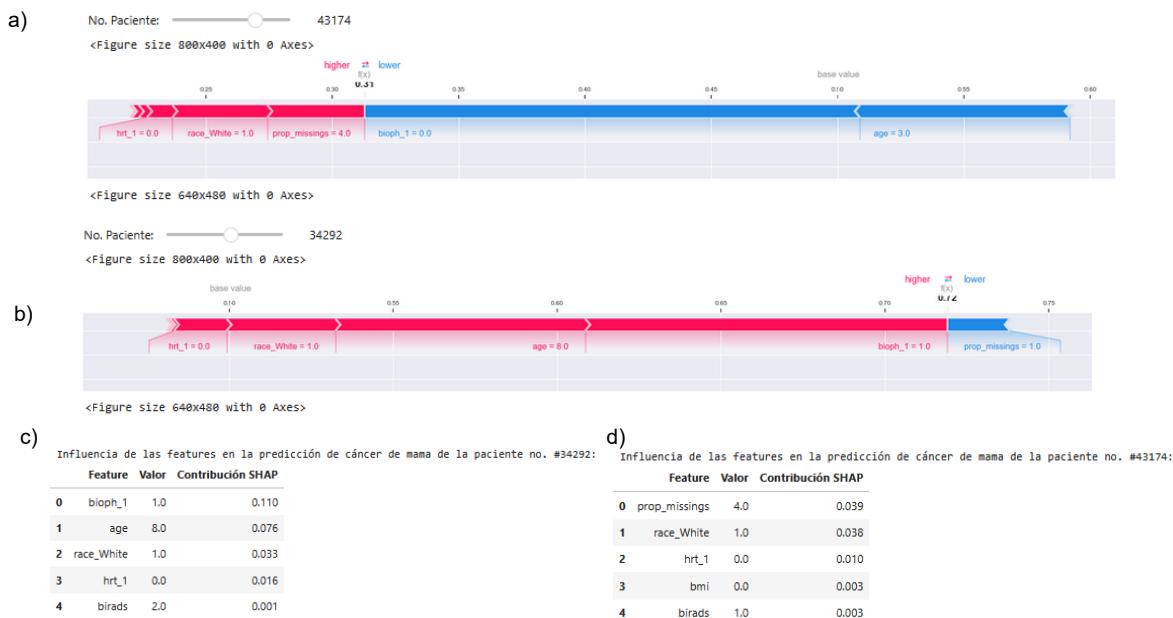


Figura E4: Gráficos de fuerza de SHAP interactivo para pacientes individuales, mostrando como ejemplos los pacientes: a) 43174 y b) 34292. Indica cómo las características contribuyen a la predicción de cáncer de mama. Cada barra indica la magnitud y dirección del impacto de cada variable en la predicción del modelo para ese paciente, y se destacan las variables con mayor contribución, junto con sus valores y aportes específicos. La magnitud se muestra en una tabla para cada paciente : c) 43174 y d) 34292

## Anexo F: Puesta en Producción

Para poner en funcionamiento la **aplicación web de predicción de riesgo de cáncer de mama**, se siguieron los siguientes pasos:

### 1. Preparación del entorno

- Asegurarse de contar con todas las librerías y dependencias descritas en el archivo de requerimientos (requirements.txt).
- Activar el entorno virtual correspondiente (por ejemplo, con conda o venv).

### 2. Ubicación de archivos necesarios:

Todos los archivos deben colocarse en la misma carpeta para que la aplicación funcione correctamente. Los archivos necesarios son:

- app.py: script principal de la aplicación.
- Win\_model.pkl : modelo entrenado serializado.
- preprocessor.pkl: objeto procesador serializado.
- logo.jpg: imagen para la cabecera de la aplicación.

### 3. Ejecución de la aplicación:

- Abrir una terminal en la carpeta que contiene los archivos.
- Activar el entorno virtual previamente definido.
- Ejecutar el siguiente comando: streamlit run app.py
- Esto abrirá automáticamente la aplicación en el navegador predeterminado.

A continuación, se muestra la interfaz de la aplicación web. Esta interfaz permite introducir los datos de los pacientes mediante cajas de selección. Estos datos se guardan para añadirlos a la base de datos, y obtener la predicción de riesgo de cáncer de mama:

The screenshot shows the homepage of the 'BreastHealth Predictor' web application. At the top, it says 'BreastHealth Predictor' and 'Predicción personalizada de salud mamaria'. To the right is a small image of a doctor wearing a white coat and a pink ribbon. Below the header, there's a section titled 'Introduce los datos de la paciente:' (Enter patient data). This section contains several dropdown menus for inputting patient information:

Grupo de edad de la paciente	Raza / etnia	Familiares de primer grado
Edad 55-59	Blanca	No
Edad menarquia	Edad primer parto	Densidad mamaria (BIRADS)
Edad >14	Edad >30	Casi totalmente grasa
Terapia hormonal	Estado menopáusico	Grupo IMC
No	Post-menopáusica	30-34.99
Biopsia previa	Historial previo de cáncer de mama	
No	No	

At the bottom left is a button labeled 'Guardar paciente y predecir' (Save patient and predict).

Figura F1: Interfaz web inicial de la herramienta de producción, desarrollada con Streamlit, donde se introducen los datos de la paciente para la predicción de riesgo de cáncer de mama.

## Predicción de Cáncer de Mama Basado en Factores Clínicos y Demográficos

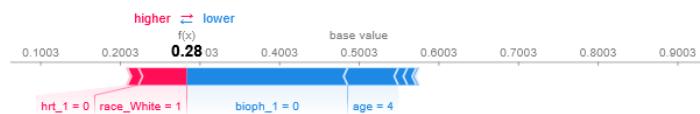
Al pulsar el botón “**Guardar y predecir**”, la aplicación registra correctamente los datos introducidos y muestra un mensaje confirmando que la información ha sido guardada exitosamente. A continuación, el programa proporciona **el tipo de riesgo y la probabilidad asociada** para la paciente. Además, se incluye un **gráfico SHAP** que explica cómo cada característica ha influido en la predicción individual, permitiendo visualizar los factores que más contribuyen al riesgo detectado. Esta funcionalidad facilita la interpretación de la predicción y puede ser utilizada por profesionales clínicos para **tomar decisiones informadas** sobre el seguimiento y posibles medidas preventivas para la paciente.

### Resultado:

Paciente guardado correctamente!

Riesgo bajo. Probabilidad: 28.44%

### Explicación SHAP para la paciente:



### Top características más influyentes en la predicción:

	Feature	Valor	Contribución SHAP
0	<code>race_White</code>	1	0.061
1	<code>hrt_1</code>	0	0.014
2	<code>first_degree_1</code>	0	0.001
3	<code>menarche</code>	0	0
4	<code>first_birth</code>	3	0

Figura F2: Pantalla de la aplicación tras pulsar el botón “Guardar y predecir”, mostrando el mensaje de confirmación de guardado, el tipo de riesgo y probabilidad de la paciente, y un gráfico SHAP que indica cómo cada característica contribuye a la predicción individual, facilitando la interpretación clínica y el seguimiento preventivo.

## Anexo G: Código del Proyecto Desarrollado en Python

Todo el código desarrollado para el proyecto se encuentra disponible en un **repositorio público de GitHub**, perteneciente a la alumna, cuyo enlace se incluye a continuación. Este repositorio permite acceder de manera organizada a todos los archivos utilizados durante el desarrollo y puesta en producción del modelo de predicción de riesgo de cáncer de mama.

El repositorio contiene:

1. **Jupyter Notebook**: donde se desarrolló todo el proyecto en Python, incluyendo el análisis exploratorio de datos, feature engineering, la selección y evaluación de modelos, y la interpretación de resultados mediante SHAP.
2. **Archivo de requerimientos (requirements.txt)**: que recoge todas las librerías y dependencias necesarias para ejecutar el proyecto en cualquier entorno compatible con Python.
3. **Video explicativo**: que describe paso a paso el desarrollo del proyecto, mostrando la metodología seguida, la construcción de modelos y la interpretación de resultados.
4. **Carpeta de la aplicación web**: que incluye todos los elementos necesarios para la puesta en producción de la interfaz virtual, entre los que se encuentran:
  - o app.py: script principal de la aplicación.
  - o logo.jpg: imagen utilizada en la cabecera de la aplicación.
  - o Win\_model.pkl: modelo entrenado y serializado.
  - o preprocessor.pkl: objeto preprocesador serializado.

**Enlace al repositorio:**

[https://github.com/EvaTartaruga/TFM\\_Eva\\_Villar\\_Alvarez\\_Prediccion\\_Cancer\\_Mama](https://github.com/EvaTartaruga/TFM_Eva_Villar_Alvarez_Prediccion_Cancer_Mama)