Universidad de Buenos Aires
Facultad de Ciencias Exactas y Naturales

Modelización predictiva para la identificación de variables clave y grupos vulnerables en la realización de mamografías en Argentina

Tesis de Licenciatura en Ciencias de Datos

Luz Montserrat

Directora: María Soledad Fernández
Codirectora: Adriana Pérez
Facultad de Ciencias Exactas y Naturales, UBA, 2025

MODELIZACIÓN PREDICTIVA PARA LA IDENTIFICACIÓN DE VARIABLES CLAVE Y GRUPOS VULNERABLES EN LA REALIZACIÓN DE MAMOGRAFÍAS EN ARGENTINA

El cáncer de mama es la principal causa de muerte por cáncer en mujeres en Argentina, y la mamografía es el único método reconocido de detección precoz en población con riesgo promedio. Sin embargo, persisten brechas significativas en el acceso a esta práctica preventiva. Esta tesis analiza las desigualdades en la realización de mamografías en mujeres de 50 a 69 años en Argentina, utilizando técnicas de machine learning aplicadas a los datos de la Encuesta Nacional de Factores de Riesgo 2018 (ENFR). Se construyó una base de datos enriquecida con recodificaciones e índices estructurales y socioeconómicos (utilizando las variables originales de la encuesta). Se entrenaron modelos supervisados (XGBoost, Random Forest y HistGradientBoosting) para identificar variables clave y perfiles con mayor riesgo. Los resultados muestran que factores como el bajo nivel educativo, la cobertura de salud pública exclusiva, las condiciones habitacionales desfavorables y la jurisdicción de residencia, así como determinadas configuraciones familiares y socioeconómicas, están fuertemente asociados a una menor probabilidad de realización de mamografías. Se identificaron subgrupos con alta vulnerabilidad sanitaria que combinan bajos niveles educativos, ausencia de cobertura médica, residencia en localidades pequeñas y precariedad estructural. Estos hallazgos permiten delinear perfiles prioritarios para políticas públicas focalizadas, con el objetivo de avanzar hacia una mayor equidad en salud y una prevención más efectiva del cáncer de mama en la Argentina.

Palabras claves: Mamografía, Cáncer de mama, Machine Learning, Desigualdades, Factores de riesgo, Equidad en salud, Prevención, ENFR.

PREDICTIVE MODELING FOR THE IDENTIFICATION OF KEY VARIABLES AND VULNERABLE GROUPS IN MAMMOGRAPHY SCREENING IN ARGENTINA

Breast cancer is the leading cause of cancer-related death among women in Argentina, and mammography is the only recognized method for early detection in average-risk populations. However, significant gaps persist in access to this preventive practice. This thesis analyzes inequalities in mammography uptake among women aged 50 to 69 in Argentina, using machine learning techniques applied to data from the 2018 National Risk Factor Survey (ENFR). A refined dataset was constructed, including recoded variables and structural and socioeconomic indices (based on the original survey variables). Supervised models (XGBoost, Random Forest, and HistGradientBoosting) were trained to identify key variables and high-risk profiles. The results show that factors such as low educational attainment, exclusive use of public health coverage, poor housing conditions, and province of residence—as well as specific family and socioeconomic configurations—are strongly associated with a lower likelihood of undergoing mammography. Subgroups with high health vulnerability were identified, combining low education levels, lack of medical coverage, residence in small localities, and structural precariousness. These findings make it possible to define priority profiles for targeted public policies, aimed at advancing health equity and more effective breast cancer prevention in Argentina.

Keywords: Mammography, Breast cancer, Machine Learning, Inequalities, Risk factors, Health equity, Prevention, National Risk Factor Survey.

AGRADECIMIENTOS

Quiero agradecer, en primer lugar, a mis papás, Mariano y Silvia, y mis hermanas, Mili y Rochi, por su amor incondicional, su apoyo constante y por estar siempre presentes a lo largo de este camino. Gracias por acompañarme en cada paso, con paciencia, confianza y motivación.

A mis amigas y amigos de la facultad, con quienes compartí miles de clases, entregas, risas y hasta viajes. Sin ustedes, este camino habría sido mucho más difícil, y sin duda, mucho menos feliz. Gracias por estar siempre, por motivarme, escucharme y recordarme que no estaba sola en todo esto. En especial, a Valen y Sol, mis dos superpoderosas, por siempre subirme el ánimo cuando lo necesité, por motivarme a seguir cuando sentía que no podía más, y por estar ahí en cada jornada eterna de estudio. No hay mejor manera de haber transitado esta etapa que junto a ustedes dos.

A la Universidad de Buenos Aires, que tanto me enseñó y me dio un lugar cuando más lo necesitaba. Gracias por haber sido no solo un espacio de formación académica, sino también un lugar de encuentro, reflexión y crecimiento personal. Me enorgullece ser parte de Exactas y haberme formado en una comunidad tan comprometida con el conocimiento, el pensamiento crítico y la educación pública.

A mi directora y a mi codirectora, Sole y Adri, por su guía, su tiempo y su mirada atenta y crítica. Gracias por ayudarme a ordenar mis ideas, por las sugerencias oportunas y por acompañarme con tanta dedicación en este proceso.

Por último, a todas las personas que, de una manera u otra, me acompañaron, me alentaron o estuvieron cerca durante este proceso: gracias. Esta tesis también es un poquito de cada uno de ustedes.

A mi familia, mis amigos y a Exactas.

Índice general

1. INTRODUCCIÓN

El control mundial del cáncer de mama es una cuestión de género, equidad y derechos humanos. Las mujeres desempeñan un papel fundamental en la sociedad; protegerlas del cáncer de mama también protege a sus familias, comunidades y la economía en su conjunto.
– OMS 2022

El cáncer de mama es una de las principales causas de mortalidad en mujeres a nivel mundial. Se trata de una enfermedad en la que células alteradas del tejido mamario se multiplican de forma descontrolada, formando tumores que, si no se detectan y tratan a tiempo, pueden invadir otros órganos a través del proceso de metástasis y provocar la muerte. Aunque existen distintos estadios de progresión, el cáncer de mama en estadio 0 (también conocido como in situ) no representa una amenaza potencial para la vida y puede ser tratado exitosamente si se detecta temprano [1].

Este tipo de cáncer afecta a personas del género femenino desde la pubertad hasta edades avanzadas, con presencia en todos los países del mundo. Aunque puede presentarse a cualquier edad adulta, las tasas de incidencia aumentan notablemente a partir de los 50 años, en gran parte debido a factores hormonales relacionados con una menopausia tardía [2]. Si bien se trata de una enfermedad fuertemente vinculada al sexo femenino, entre el 0,5 % y el 1 % de los casos se detecta en varones. Además, cerca de la mitad de los diagnósticos se producen en personas que no presentan factores de riesgo identificables más allá del sexo y la edad [1].

1.1. El Impacto del Cáncer de Mama

En 2022, en todo el mundo se diagnosticaron 2,3 millones de casos de cáncer de mama en mujeres y se registraron cerca de 670.000 defunciones. En ese mismo año, este tipo de cáncer se posicionó como el segundo cáncer más común a nivel global, solo superado por el cáncer de pulmón, y quinto en número de muertes [3]. Según un informe reciente publicado por la Agencia Internacional de Investigación sobre el Cáncer (IARC), se estima que 1 de cada 20 mujeres será diagnosticada con cáncer de mama a lo largo de su vida, y que, si las tendencias actuales continúan, para 2050 los nuevos casos aumentarán un 38 % y las muertes un 68 % [4]. Este crecimiento afectará de forma desproporcionada a los países con bajo Índice de Desarrollo Humano (IDH), un indicador compuesto desarrollado por el Programa de las Naciones Unidas para el Desarrollo (PNUD) [5], lo que pone de relieve las inequidades globales en el acceso al diagnóstico temprano y al tratamiento adecuado. Estas inequidades también se reflejan en el hecho de que, en muchos países con alto IDH, la mortalidad por cáncer de mama ha mostrado una tendencia al descenso durante los últimos 20 años, atribuible a los avances en el tratamiento y a la implementación de programas de tamizaje [1]. Entre 1990 y 2020, 20 países lograron reducir la mortalidad por cáncer de mama en al menos un 2 % anual durante tres años consecutivos. Esto dio lugar

[1] Un tamizaje es una prueba o examen que se aplica a personas asintomáticas para identificar aquellos que tienen un mayor riesgo de padecer una enfermedad o condición que se beneficia de ser detectada y tratada temprano. El objetivo es identificar la enfermedad en una etapa temprana antes de que la persona presente síntomas, lo que permite una intervención más efectiva.

a una reducción global del 40 % en la mortalidad por cáncer de mama en varios países de alto IDH durante ese mismo periodo [6].

En Argentina, el cáncer mamario constituye la primera causa de muerte por cáncer en mujeres. En el año 2022 se registraron 5.750 defunciones en mujeres por esta enfermedad, lo que corresponde a una tasa bruta de 24,4 defunciones cada 100.000 mujeres [7]. Actualmente, se producen 6.100 muertes por esta enfermedad al año y se estima que se producirán más de 22.000 nuevos casos por año, lo cual representa el 32,1 % del total de incidencia de cáncer en Argentina [8].

1.2. Mamografía como Práctica Preventiva

La principal medida para lograr el control del cáncer de mama en el mediano plazo debe centrarse en la detección temprana y en la implementación de tratamientos pertinentes y oportunos. La mamografía es una radiografía de las mamas que permite, a través de imágenes, la detección del cáncer en su fase más temprana, cuando todavía es muy pequeño, no palpable, y la paciente no presenta síntomas ni signos de la enfermedad. Es el único método reconocido para la detección precoz de esta enfermedad en población con riesgo promedio [9]. Este método ha demostrado disminuir la mortalidad por cáncer de mama, ya que permite detectar tumores en etapas tempranas, cuando su tamaño es menor y las probabilidades de tratamiento exitoso son significativamente mayores [10].

En nuestro país, el Programa Nacional de Control de Cáncer de Mama del Instituto Nacional del Cáncer (PNCM - INC) [8] recomienda para la prevención del cáncer de mama a nivel poblacional que todas las mujeres entre los 50 y los 69 años, sin síntomas y sin antecedentes familiares ni personales de cáncer, se realicen una mamografía cada uno o dos años.

Investigaciones recientes han abordado las disparidades en la realización de mamografías desde una perspectiva eco-epidemiológica, analizando cómo factores socioeconómicos se relacionan con la distribución espacio-temporal de la realización de esta práctica en el país (Damiani Quiroz 2024 [11]). Otros estudios han identificado diversas variables asociadas a la realización de mamografías más allá del nivel socioeconómico. Un meta-análisis sistemático llevado a cabo por Mottram et al. (2021) [12] encontró que factores como el estado civil, la educación, la tenencia de vivienda, el estatus migratorio y los antecedentes de resultados previos en mamografías se relacionan significativamente con la asistencia a los controles de detección de cáncer de mama. En particular, en dicho estudio se encontró que las mujeres casadas o que cohabitan tienen mayor probabilidad de realizarse una mamografía en comparación con aquellas solteras o separadas. Además, se ha reportado que las mujeres con un nivel educativo medio o superior presentan una mayor adherencia a los controles que aquellas con un nivel educativo bajo. Estos hallazgos destacan la necesidad de reducir las barreras socioeconómicas para garantizar el acceso equitativo a la realización de mamografías.

En Argentina, los datos de la Encuesta Nacional de Factores de Riesgo (ENFR) [13] permiten observar estas disparidades. Esta encuesta oficial, realizada periódicamente desde 2005 por el Ministerio de Salud junto al INDEC (Instituto Nacional de Estadística y Censos) [14], releva cada cuatro años información sobre los principales factores de riesgo de enfermedades crónicas no transmisibles (ECNT) en la población adulta. A lo largo de sus distintas ediciones, ha mostrado variaciones en la cobertura de esta práctica entre

diferentes regiones del país y a lo largo del tiempo. La información contenida en esta fuente es particularmente valiosa para caracterizar a las mujeres que no acceden a la mamografía, a pesar de encontrarse dentro del grupo objetivo. En este contexto, surge la necesidad de contar con herramientas analíticas que permitan identificar los factores más relevantes asociados a la realización o no de este estudio, así como segmentar a la población en grupos con distintos niveles de vulnerabilidad frente a esta práctica preventiva.

Como se mencionó, estudios previos han analizado los resultados de realización de mamografía a partir de la última edición de la ENFR ([11], [15], [16]); sin embargo, en esta tesis se propone incorporar la aplicación de herramientas de machine learning en un proceso de análisis de datos, enfoque que, según nuestro conocimiento, no ha sido utilizado previamente para abordar esta problemática.

2. OBJETIVOS

El objetivo general de esta tesis es identificar los factores socioeconómicos a diferentes escalas (individual y hogar) asociados a la probabilidad de que mujeres de entre 50 y 69 años en Argentina se realicen una mamografía al menos cada dos años. Además, se quiere identificar grupos con menor acceso a esta práctica, a fin de proporcionar información clave para el desarrollo de políticas públicas más efectivas para la prevención de cáncer de mama, a partir del análisis de los datos de la ENFR 2018 [17].

2.1. Objetivos Específicos

- **Selección de variables predictoras:** Identificar y seleccionar un conjunto de variables potencialmente asociadas a la realización de mamografías en mujeres de entre 50 y 69 años en Argentina, a partir de los datos disponibles en la Encuesta Nacional de Factores de Riesgo (ENFR) 2018.

- **Construcción de variables socioeconómicas:** Diseñar y construir índices que permitan representar el nivel socioeconómico de las mujeres encuestadas, con el fin de analizar su asociación en la realización de mamografías.

- **Modelado estadístico y predictivo:** Ajustar modelos estadísticos y de machine learning que permitan identificar las principales variables asociadas a la probabilidad de realización de mamografías, evaluando tanto su significación como su aporte predictivo.

- **Comparación de modelos:** Comparar diferentes modelos mediante métricas de desempeño para determinar la estrategia de modelado más adecuada en términos de capacidad predictiva.

- **Identificación de grupos vulnerables:** A partir de los modelos que presenten mejor performance, realizar una segmentación poblacional basada en la realización de mamografías, lo que permitirá identificar grupos de mujeres con baja propensión a realizarse mamografías.

3. METODOLOGÍA

3.1. Fuente de Datos

La Encuesta Nacional de Factores de Riesgo (ENFR) es una encuesta realizada por el Ministerio de Salud y Desarrollo Social de la Nación y el Instituto Nacional de Estadística y Censos (INDEC), de manera periódica cada cuatro años, y constituye un componente central de la Estrategia Nacional de Prevención y Control de Enfermedades No Transmisibles. Además, pertenece al Sistema de Vigilancia de Enfermedades No Transmisibles y del Sistema Integrado de Estadísticas Sociales (SIES). Particularmente, la edición de 2018 fue llevada a cabo durante el último trimestre del año por ambas instituciones: el Ministerio de Salud y Desarrollo Social, a través de la Secretaría de Promoción de la Salud, Prevención y Control de Riesgos y la Dirección Nacional de Promoción de la Salud y Control de Enfermedades No Transmisibles; y el INDEC, mediante la Dirección Nacional de Estadísticas de Condiciones de Vida y la Dirección de Estudios de Ingresos y Gastos de los Hogares, en articulación con las direcciones provinciales de estadística (DPE) de las 24 jurisdicciones del país [13].

El objetivo de la encuesta es proporcionar información válida, confiable y oportuna sobre los factores de riesgo y prevalencias de las principales enfermedades no transmisibles en la población de 18 años o más de la República Argentina. Desde sus inicios en 2005, la encuesta se realiza en hogares de localidades de 5.000 habitantes o más, utilizando un muestreo polietápico por conglomerados que garantiza representatividad nacional y provincial. En este estudio se trabajará con la base de datos correspondiente a la edición 2018 por tratarse de la versión más reciente disponible. La edición prevista para 2022 fue suspendida debido a la emergencia sanitaria derivada de la pandemia de COVID-19, lo que interrumpió la periodicidad cuatrienal con la que se venía realizando. Como consecuencia, la ENFR 2018 continúa siendo, al momento de esta investigación, la fuente más actualizada y completa para el estudio de prácticas preventivas en salud en la población adulta argentina.

Esta edición de la ENFR utilizó un diseño muestral probabilístico y multietápico, con representatividad nacional y provincial para la población de 18 años o más residente en localidades urbanas de 5.000 o más habitantes. El primer paso de la encuesta incluyó una muestra de 49.170 viviendas distribuidas en todas las jurisdicciones del país. Además, ocho grandes aglomerados urbanos fueron considerados dominios de estimación específicos para facilitar la comparabilidad con ediciones anteriores. La distribución geográfica de los puntos de muestreo puede observarse en la Figura 3.1.

Por su parte, el cuestionario por autorreporte está compuesto por un **Bloque Hogar** y un **Bloque Individual**. El primero releva información sobre el entorno del hogar y sus integrantes. Las dimensiones abordadas incluyen:

- Características de la vivienda (CV)
- Características del hogar (CH)
- Ingreso total del hogar (IH)

Figura 3.1: Distribución geográfica de los puntos de muestreo de la ENFR 2018 en localidades de 5.000 habitantes o más. (Fuente: ENFR 2018 Nota técnica [18])

- Características de la jefatura del hogar
- Situación laboral del jefe del hogar (SL)
- Características del respondente del Bloque Individual

En cuanto al Bloque Individual, se realizó el cuestionario a una persona adulta (de 18 años o más) seleccionada aleatoriamente dentro del hogar. Este bloque relevó información detallada sobre su estado de salud y los principales factores de riesgo asociados a las enfermedades no transmisibles. Las temáticas abordadas fueron:

- Situación laboral (SL)
- Salud general (SG)
- Actividad física (AF)
- Consumo de tabaco y exposición al humo de tabaco ajeno (TA)
- Hipertensión arterial (HA)
- Peso corporal (PC)
- Alimentación (AL)
- Colesterol (CO)
- Consumo de alcohol (CA)
- Diabetes (DI)
- Lesiones (LE)
- Prácticas preventivas (PP): solo para mujeres
- Prevención de cáncer colorrectal (CC): solo para personas de 50 años o más
- Mediciones físicas y antropométricas (MA)
- Mediciones bioquímicas (MQ)

En conjunto, ambos bloques de información permiten construir un panorama integral sobre las condiciones sociales y de salud de la población adulta residente en localidades de 5.000 o más habitantes en Argentina. La ENFR 2018, por su riqueza temática, representatividad y cobertura nacional, constituye una fuente adecuada para abordar los objetivos de esta tesis, centrados en el análisis predictivo del acceso a mamografías en el país.

Cabe destacar que este trabajo se basa exclusivamente en los datos autorreportados de la encuesta, por lo que no se consideran las mediciones físicas, antropométricas ni bioquímicas incluidas en otros módulos. Esta elección metodológica permite enfocarse en variables accesibles y comparables en el tiempo, facilitando la exploración de asociaciones relevantes y la construcción de modelos predictivos sobre la realización de mamografías.

3.2. Definición de la Muestra

El análisis de este trabajo se enfocó exclusivamente en mujeres de entre 50 y 69 años en Argentina, dado que el Programa Nacional de Control de Cáncer de Mama (PNCM) del Instituto Nacional del Cáncer (INC), dependiente del Ministerio de Salud de la Nación, recomienda la realización de una mamografía cada dos años para mujeres de este grupo etario, sin síntomas ni antecedentes personales o familiares de la enfermedad, como medida de prevención poblacional [8]. Esta recomendación se fundamenta en que la mamografía es el único método reconocido para la detección precoz del cáncer de mama en mujeres con riesgo promedio.

La base original de la ENFR 2018 cuenta con un total de 29.224 registros, es decir, 29.224 personas que contestaron la encuesta de los 49.170 hogares seleccionados. A partir de esta base, se realizó un proceso de depuración y selección de casos, mediante el cual se filtraron primero los registros correspondientes al sexo femenino, y luego aquellos dentro del rango etario de 50 a 69 años, definido como población objetivo para la estrategia de tamizaje mamográfico. Adicionalmente, se excluyeron los registros con respuesta "No sabe/No contesta" (código 99) en la variable dependiente `control_mamografia`, por no aportar información útil para los modelos predictivos. En total, se eliminaron 25 casos por este motivo, lo que representa aproximadamente el 0,53 % de la submuestra. Dado que se trata de un porcentaje muy bajo, esta exclusión no compromete la representatividad de la muestra analítica. Este procedimiento de selección se resume en el diagrama de flujo de la Figura 3.2.

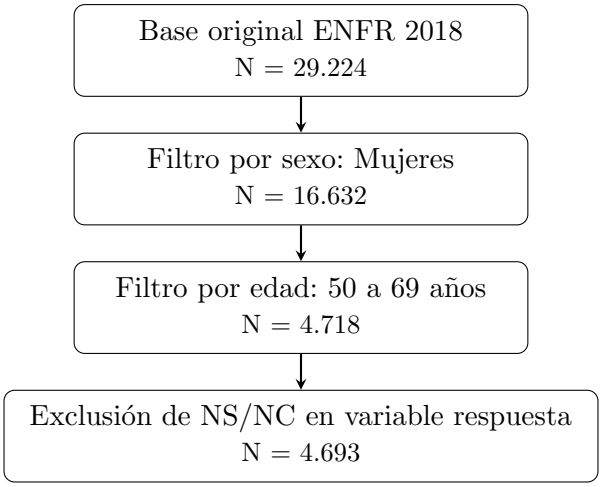


Figura 3.2: Diagrama de flujo del proceso de selección muestral según criterios de inclusión y disponibilidad de respuesta.

El resultado de este recorte es una submuestra de 4.693 mujeres de entre 50 y 69 años, que constituye la población analizada en este estudio. Esta subpoblación representa el grupo de mayor interés para la implementación de políticas públicas de prevención en cáncer de mama.

3.3. Análisis Exploratorio de Datos y Variables Involucradas

La base de la ENFR 2018 incluye un total de 284 variables. Estas no corresponden únicamente a respuestas directas del cuestionario, sino que también incluyen identificadores técnicos, factores de expansión, variables recodificadas por el equipo encuestador, codificaciones internas, y datos geográficos u operativos asociados al relevamiento. Por ello, parte del trabajo consistirá en identificar, seleccionar, transformar o recodificar (en caso que corresponda) aquellas variables relevantes para los objetivos analíticos de este estudio.

3.3.1. Variable Respuesta

En la sección de *Prácticas Preventivas* de la ENFR 2018 se encuentra la variable respuesta, denominada `control_mamografia`, que tuvo valor 1 si la mujer se realizó una mamografía en los últimos dos años, 2 si no y 99 si no sabe o no contesta. Esta variable fue construida a partir de la recodificación de las siguientes variables originales del cuestionario:

`bipp01` "*¿Alguna vez se hizo una mamografía?*"
Opciones: 1. Sí;
2. No;
99. Ns/Nc.

`bipp02` "*¿Cuándo fue la última vez que se hizo una mamografía?*"
Opciones: 1. Menos de 1 año;
2. De 1 a 2 años;
3. Más de 2 a 3 años;
4. Más de 3 años;
99. Ns/Nc.

Se definió como valor positivo (`control_mamografia` = 1) a aquellas mujeres que respondieron afirmativamente en la variable `bipp01` (es decir, indicaron que sí se realizaron alguna vez una mamografía) y que, en la variable `bipp02`, señalaron que la última mamografía fue realizada hace menos de dos años (opciones 1 o 2). En los demás casos, se asignó el valor 2, salvo cuando el valor era 99 en alguna de las dos preguntas, lo que se resolvió asignando un 99 en la nueva variable, indicando que la persona no sabe o no contesta alguna de las dos preguntas. Esto implica que para la construcción de la variable `control_mamografia` se consideró como negativo toda aquella mujer que no se realizó una mamografía en los dos últimos años. Posteriormente, la variable fue recodificada para el análisis en formato binario, asignando el valor 1 a los casos donde la mujer se realizó una mamografía y 0 a los que no, excluyendo las observaciones con valor 99 (Ns/Nc).

3.3.2. Selección de Variables Independientes

Para reducir la dimensionalidad del conjunto de variables predictoras y enfocar el análisis en aquellas variables potencialmente informativas, se llevó a cabo una selección combinando criterios teóricos y herramientas estadísticas exploratorias.

Depuración Manual Inicial

El primer paso en la selección de variables independientes consistió en una depuración manual preliminar. De las 284 variables disponibles en la base de datos original (ENFR 2018), se seleccionaron 108 que se consideraron relevantes para el objetivo del estudio.

La elección se realizó con el propósito de evitar redundancias y reducir la complejidad dimensional del conjunto de datos. Se excluyeron variables que contenían información superpuesta con otras ya incluidas, que presentaban un elevado porcentaje de valores nulos, o que estaban dirigidas a subpoblaciones no pertinentes para el estudio. A su vez, se descartaron algunas variables que, si bien contenían información relevante, eran redundantes con otras ya sintetizadas o presentaban categorías demasiado específicas que dificultaban su análisis con técnicas automatizadas. Este trabajo también requirió una revisión manual del contenido de las variables, con el fin de asegurar su adecuación a los objetivos analíticos del estudio y su compatibilidad con los métodos de modelado utilizados.

Evaluación Automática Complementaria

Como segunda parte en la selección de variables, se aplicaron métodos automáticos complementarios para refinar el conjunto de variables:

- Para las variables categóricas, se empleó el *coeficiente V de Cramér* [19] con el objetivo de cuantificar la fuerza de asociación entre cada predictor y la variable dependiente. Este coeficiente toma valores entre 0 y 1, donde 0 indica ausencia de asociación y 1 refleja una relación perfecta, lo que permite identificar aquellas variables categóricas con mayor capacidad explicativa.
- Para las variables continuas, se aplicó un análisis de varianza unidireccional (ANOVA de un factor) con el objetivo de evaluar diferencias significativas entre los grupos definidos por la variable dependiente. Se utilizaron los p-valores como criterio de significancia estadística, y se complementó el análisis con el coeficiente eta cuadrado (η^2) como medida del tamaño del efecto.

Este procedimiento fue complementado con una evaluación interpretativa orientada a asegurar la consistencia conceptual y la relevancia práctica del conjunto final de variables. En esta etapa se consideraron aspectos como la aplicabilidad general de cada variable dentro de la población objetivo, su pertinencia teórica y su adecuación a los objetivos analíticos del estudio. Se evitó incorporar variables que, aun siendo estadísticamente relevantes, se encontraban restringidas a subgrupos particulares o presentaban limitaciones de interpretación — como ciertas preguntas dirigidas exclusivamente a personas con diagnóstico de diabetes. De este modo, la selección de variables a incorporar buscó garantizar cobertura uniforme, consistencia conceptual y robustez metodológica del modelo predictivo.

Este conjunto de variables depurado sirvió como base para los procedimientos de preprocesamiento y modelado desarrollados en las siguientes etapas del estudio.

3.3.3. Análisis Exploratorio de Datos

Se llevó a cabo un análisis exploratorio de datos con el objetivo de comprender las distribuciones de las variables, detectar valores atípicos y evaluar la necesidad de transformaciones o recodificaciones antes del modelado. Esta etapa, al realizarse posteriormente

a la selección manual y automática de variables, permitió refinar aún más el conjunto de variables predictoras.

Para las **variables continuas**, se calcularon medidas de tendencia central y dispersión (media y desvío estándar), con el fin de observar patrones generales y posibles asimetrías o valores extremos. En el caso de las **variables categóricas**, se analizaron las distribuciones mediante frecuencias absolutas y relativas, identificando modalidades poco frecuentes o categorías que podían ser agrupadas para facilitar el análisis.

En esta etapa, además, se analizaron de manera preliminar las asociaciones entre ciertas variables seleccionadas y la variable objetivo (realización de mamografía en los últimos dos años). Esto permitió observar patrones de interés y posibles tendencias en los datos, lo cual permite priorizar algunas variables por sobre otras en los análisis posteriores.

3.3.4. Recodificación y Construcción de Índices

En la última etapa de selección de variables predictoras, se aplicaron dos estrategias complementarias: por un lado, la recodificación de variables originales de la base de datos, y por otro, la construcción de índices compuestos a partir de múltiples indicadores. Ambas estrategias tuvieron como objetivo mejorar la calidad del dataset para el análisis predictivo, reduciendo dimensionalidad, mejorando la interpretación y asegurando la cobertura de aspectos estructurales del nivel socioeconómico y las condiciones materiales de los hogares.

Recodificación de Variables

Se identificaron variables con un número elevado de categorías o con categorías poco frecuentes y se procedió a su recodificación o agrupación. Esta transformación permitió simplificar el análisis, reducir la dimensionalidad y mejorar la estabilidad estadística de las estimaciones. Las principales recodificaciones fueron:

1. **Índice de hacinamiento**

   El hacinamiento fue calculado como el cociente entre la cantidad de personas del hogar (`cant_componentes`) y la cantidad de ambientes de uso exclusivo declarados en la vivienda (`bhcv02`), dos variables que se encontraban a disposición en la ENFR. Este cálculo sigue la definición oficial del INDEC, que considera **hacinamiento crítico** cuando un hogar presenta más de tres personas por ambiente exclusivo [20, 21].

   Para este estudio, además de la categoría oficial, se propuso una clasificación más detallada del índice de hacinamiento, con el objetivo de captar matices dentro de las condiciones habitacionales. Esta recodificación permitió categorizar los hogares en cuatro niveles, los cuales se pueden observar en la Tabla 3.1.

   Es importante aclarar que únicamente la categoría de **hacinamiento crítico** forma parte de las definiciones oficiales utilizadas por el INDEC en sus informes de condiciones de vida [21]. Las restantes categorías (holgado, sin hacinamiento, moderado) fueron incorporadas a modo de adaptación metodológica, con fines analíticos y de interpretación. Estas adaptaciones buscan ofrecer una caracterización más matizada de las condiciones habitacionales dentro del universo bajo estudio.

Tabla 3.1: Clasificación del índice de hacinamiento adaptada a partir de criterios del INDEC y organismos internacionales.

Categoría	**Índice de hacinamiento**	**Definición**
Hogar holgado	< 1	Hogares donde cada persona dispone de un ambiente propio o más.
Sin hacinamiento	≥ 1 y ≤ 2	Hogares con un nivel adecuado de espacios disponibles.
Hacinamiento moderado	> 2 y ≤ 3	Hogares con cierto grado de congestión espacial.
Hacinamiento crítico	> 3	Hogares con grave insuficiencia de ambientes disponibles.

2. Nivel educativo

Las variables originales `nivel_instruccion` y `nivel_instruccion_j`, que indican el máximo nivel educativo alcanzado por el respondente y por el jefe o jefa de hogar, fueron recategorizadas en tres grandes categorías con el objetivo de simplificar el análisis, mejorar la estabilidad del modelo predictivo y capturar mejor las desigualdades educativas con relevancia en términos de acceso a recursos y servicios.

Las categorías originales de las variables eran las siguientes:

1. Sin instrucción
2. Primario incompleto
3. Primario completo
4. Secundario incompleto
5. Secundario completo
6. Terciario o universitario incompleto
7. Terciario o universitario completo
8. Educación especial

Se agruparon los niveles de educación básica baja e incompleta —que incluyen desde la ausencia total de escolaridad hasta el secundario incompleto— dentro de una categoría de *bajo nivel educativo*, ya que reflejan trayectorias formativas limitadas, generalmente asociadas a mayores barreras en el acceso al mercado laboral formal y a servicios de salud [22]. Los niveles de secundario completo y terciario o universitario incompleto se consolidaron en una categoría intermedia, representando un *medio nivel educativo*. Finalmente, se mantuvo como categoría aparte el grupo con nivel terciario o universitario completo, dada su marcada asociación con mejores condiciones socioeconómicas, habitacionales y sanitarias, llamando a esta categoría *alto nivel educativo*. Esta agrupación se puede ver en la Tabla 3.2.

Esta estrategia de agrupamiento sigue prácticas reconocidas en estudios sociales y censales tanto a nivel nacional como internacional [23, 24], y permite capturar de manera más robusta las diferencias de contexto educativo en el análisis multivariado.

Tabla 3.2: Recategorización de las variables `nivel_instruccion` y `nivel_instruccion_j`.

Categoría	Descripción	Códigos originales
1	Bajo nivel educativo	1, 2, 3, 4
2	Medio nivel educativo	5, 6
3	Alto nivel educativo	7

La categoría original 8, educación especial (referida a la modalidad educativa para personas con discapacidad) fue excluida del análisis en la recategorización, ya que no podía ser asignada con fundamentos sólidos a ninguna de las tres categorías anteriores sin introducir ambigüedades conceptuales.

3. Tipo de combustible para cocinar

La variable `bhcv06` recoge información sobre el tipo principal de combustible utilizado en el hogar para cocinar. Las categorías originales eran:

1. Gas de red
2. Gas de tubo o garrafa
3. Kerosene, leña o carbón
4. Otro

Dado que las dos últimas categorías registran una frecuencia muy baja en la muestra, se decidió agruparlas en una sola categoría residual para evitar problemas en el análisis estadístico. Esta decisión responde a criterios prácticos de consolidación de clases poco representadas, sin implicar una asimilación conceptual entre las opciones agrupadas.

La recodificación aplicada fue la siguiente:

Tabla 3.3: Recategorización de la variable `bhcv06` (tipo de combustible para cocinar).

Categoría	Descripción
1	Gas de red
2	Gas de tubo o garrafa
3	Kerosene, leña, carbón u otros combustibles alternativos

4. Barreras para el consumo de frutas y verduras

La variable `barreras_fyv` identifica el principal motivo declarado por el cual la persona encuestada no consume la cantidad recomendada de frutas y verduras. Las opciones originales eran:

1. Factores condicionantes individuales
2. Factores condicionantes del entorno
3. Factores económicos
4. Consume la cantidad que considera adecuada
5. Otros

Al igual que la recategorización de la variable anterior, dado que había dos categorías, "Factores del entorno" y "Otros", que presentaban una frecuencia muy baja en la muestra (1,2 % y 0,69 % respectivamente), se decidió agruparlas en una única categoría residual. Esta recodificación permite mejorar la estabilidad de los modelos sin perder información relevante, ya que ambas opciones representan barreras de tipo externo o circumstantial, poco reportadas. Se mantuvieron separadas las respuestas sobre factores individuales y económicos, por su mayor peso y valor interpretativo. Por otro lado, la categoría "consume lo adecuado" se conservó como referencia explícita de ausencia de barreras.

La recodificación final fue la siguiente:

Tabla 3.4: Recategorización de la variable `barreras_fyv` (barreras para el consumo de frutas y verduras).

Nueva categoría	**Descripción**	**Códigos originales incluidos**
1	Factores individuales	1
2	Factores económicos	3
3	Factores de entorno u otros	2, 5
4	No presenta barreras (consume adecuado)	4

Otras variables recodificadas fueron por ejemplo `bhcv08` (fuente de agua del hogar), que se simplificó en una categoría binaria según el acceso a red pública o no, y `tipo_hogar` (estructura familiar), que se reorganizó en cinco categorías más sintéticas a partir de los códigos originales del cuestionario. Estas transformaciones buscaron facilitar la inclusión de estas variables en los modelos sin perder su valor informativo.

Construcción de Índices Compuestos

Por otro lado, se realizó un desarrollo de nuevas variables a partir de múltiples ítems de la encuesta, con el objetivo de representar de manera más integral dimensiones como la vulnerabilidad estructural o el nivel socioeconómico. Estas variables no se encontraban directamente disponibles en la ENFR 2018, por lo que fue necesario construirlas a partir de criterios metodológicos consolidados.

1. **Índice de Necesidades Básicas Insatisfechas (NBI)**

   El índice de Necesidades Básicas Insatisfechas (NBI) se construyó siguiendo la metodología oficial del INDEC [25], que identifica como hogares con NBI a aquellos que presentan al menos una de las siguientes privaciones estructurales:

   - **NBI 1 — Vivienda:** Hogares[1] que habitan piezas en inquilinato, hoteles o pensiones, viviendas no destinadas a fines habitacionales, viviendas precarias u "otros tipos de vivienda". Se excluyen casas, departamentos y ranchos.
   - **NBI 2 — Condiciones sanitarias:** Hogares que no poseen baño ni letrina.

[1] Aquí el término Hogar hace referencia a los habitantes del hogar.

- **NBI 3 — Hacinamiento:** Hogares con más de 3 personas por habitación exclusiva.
- **NBI 4 — Asistencia escolar:** Hogares con al menos un niño de 6 a 12 años que no asiste a la escuela.
- **NBI 5 — Capacidad de subsistencia:** Hogares con 4 o más personas por miembro ocupado y jefe/a sin educación primaria completa.

Con la ENFR 2018 fue posible reconstruir de forma fiel las tres primeras dimensiones, utilizando las siguientes variables:

- **Tipo de vivienda (NBI 1):** Para identificar hogares con privación habitacional, se utilizó la variable `bhcv01`, que clasifica el tipo de vivienda en el que reside el hogar y se excluyeron las categorías de casa, casilla (asimilada a rancho) y departamento.
- **Condiciones sanitarias (NBI 2):** Esta dimensión se reconstruyó a partir de la variable `bhcv09`, que pregunta si el hogar cuenta con baño o letrina. Aquellos hogares que respondieron negativamente (categoría 2) fueron clasificados como en situación de privación.
- **Hacinamiento (NBI 3):** El indicador de hacinamiento crítico se construyó utilizando la variable construida previamente. Un hogar presenta hacinamiento crítico si el índice es superior a 3, es decir, si hay más de tres personas por cada ambiente disponible para uso exclusivo del hogar.

No fue posible reconstruir directamente las dimensiones de asistencia escolar (**NBI 4**) y capacidad de subsistencia (**NBI 5**), ya que la ENFR no incluye información sobre la edad ni asistencia escolar de los integrantes del hogar, ni sobre la ocupación de todos sus miembros. Sin embargo, se propuso una **variable proxy** que busca capturar situaciones de vulnerabilidad estructural similares a las cubiertas por esos indicadores.

El **proxy** se definió como 1 (privación) cuando se cumplen simultáneamente las siguientes condiciones:

- El jefe o jefa del hogar **no completó la educación primaria** o directamente **no tiene instrucción**.
- El jefe o jefa del hogar se encuentra **desocupado/a**.

Aunque no se dispone de información sobre la ocupación o educación del resto de los integrantes del hogar, el estatus laboral y la educación del jefe han sido ampliamente utilizados como proxy de capacidad de generación de ingresos del grupo familiar. Este enfoque no reemplaza la definición oficial del NBI, pero permite capturar dimensiones sociales comparables a las cubiertas por los indicadores **NBI 4** y **NBI 5**.

Finalmente, se definió como **hogar con NBI** a aquel hogar que presenta al menos una privación en cualquiera de las dimensiones reconstruidas (NBI 1, 2, 3 o proxy para 4/5).

2. Nivel Socioeconómico (NSE)

La ENFR 2018 incluye una variable denominada `quintil_uc`, que clasifica a los hogares en cinco grupos según el ingreso ajustado por unidad consumidora. Aunque útil como indicador distributivo, esta variable presenta limitaciones para el análisis estructural: dentro de un mismo quintil pueden convivir situaciones sociales muy distintas, y no existe un anclaje conceptual con los niveles de clase comúnmente utilizados en estudios de mercado o comunicación política. Por ello, se buscó construir una variable de Nivel Socioeconómico (NSE) que permitiera una segmentación más precisa de la población.

El NSE es una medida ampliamente utilizada en estudios sociales, de consumo y opinión pública para clasificar a los hogares según su posición relativa en la estructura social. A diferencia del ingreso directo, el NSE busca captar dimensiones más amplias del capital social y económico, integrando variables como educación, ocupación, tenencia de bienes y condiciones habitacionales.

En Argentina, una de las clasificaciones más utilizadas es la propuesta por la Sociedad Argentina de Investigadores de Marketing y Opinión (SAIMO) [26], que agrupa a la población en cinco estratos:

- **ABC1:** clase alta.
- **C2:** clase media alta.
- **C3:** clase media típica o baja.
- **D1:** clase baja superior.
- **D2E:** clase baja.

Esta clasificación permite una segmentación más cualitativa y estructural que la simple distribución por quintiles. Sin embargo, cuando no se dispone de todos los indicadores requeridos para calcular el algoritmo completo del NSE —como en el caso de la ENFR 2018—, es posible construir aproximaciones empíricas utilizando el ingreso del hogar como variable proxy.

Se optó por una estrategia basada en la distribución empírica del ingreso dentro de la muestra, replicando la estructura de proporciones del NSE publicada por el SAIMO para el año 2018. Dichas proporciones fueron obtenidas a partir de un reporte técnico basado en datos de la Encuesta Permanente de Hogares (EPH) y en el algoritmo oficial desarrollado por el SAIMO y CEIM [27]. A partir de la variable de ingreso mensual del hogar (`bhih01`), se ordenaron todos los casos con valores válidos y se asignaron los niveles socioeconómicos de acuerdo con las proporciones poblacionales históricas:

- **ABC1:** 5,8 %
- **C2:** 18,2 %
- **C3:** 30,2 %
- **D1:** 32,5 %
- **D2E:** 13,3 %

Estos cortes se aplicaron secuencialmente sobre la muestra ordenada por ingreso, garantizando que la distribución resultante respetara las proporciones originales reportadas por el SAIMO.

3. Índice material

Con el objetivo de captar de forma sintética las condiciones estructurales de las viviendas, se construyó un índice de materiales de la vivienda a partir de un conjunto de variables habitacionales relevadas en la ENFR 2018 en la sección *Caracteristicas de la Vivienda*. Se incluyeron las siguientes dimensiones:

- **Material del piso** (`bhcv03`): distingue entre pisos de alta calidad (cerámica, madera), materiales intermedios (cemento o ladrillo fijo) y materiales precarios (tierra o ladrillo suelto).
- **Material del techo** (`bhcv04`): clasifica la cubierta en materiales sólidos (losas, tejas), intermedios (chapas con cubierta) o precarios (cartón, caña, barro).
- **Presencia de cielorraso** (`bhcv05`): indicador de terminación interior de la vivienda.
- **Tipo de combustible para cocinar** (`bhcv06`): se interpreta como indicador indirecto del acceso a infraestructura energética.
- **Ubicación del acceso al agua** (`bhcv07`): se distingue si el agua se encuentra dentro o fuera de la vivienda.
- **Fuente de provisión de agua** (`bhcv08`): diferencia entre agua corriente de red pública y fuentes no seguras (pozo, perforación, aljibe).
- **Tipo de inodoro o letrina** (`bhcv10`): indicador de calidad del saneamiento intradomiciliario.
- **Destino del desagüe del inodoro** (`bhcv11`): permite distinguir entre conexión a red cloacal, cámara séptica, pozo o desagüe primitivo.

Antes de realizar la construcción de este índice, en las variables utilizadas se reemplazaron por valores faltantes aquellas respuestas no válidas o residuales como "no sabe", "no contesta" u "otros". Todas las variables fueron luego convertidas a tipo categórico.

Se aplicó un Análisis de Correspondencias Múltiples (MCA) sobre las variables categóricas seleccionadas, con el objetivo de sintetizar en un único eje latente la variabilidad compartida entre los distintos indicadores habitacionales. Esta técnica, ampliamente utilizada para reducir dimensionalidad en conjuntos de variables cualitativas, permite identificar patrones comunes que explican la estructura subyacente de las condiciones materiales de los hogares.

Del análisis se retuvo únicamente la primera componente, que fue interpretada como un índice sintético. Este eje resume la posición relativa de cada hogar en función de sus condiciones habitacionales: los valores más bajos del componente indican viviendas con mejores características estructurales, mientras que los más altos reflejan mayor precariedad. Para facilitar su interpretación, el índice fue reescalado a un rango de valores entre 0 y 1, donde 1 indica mayor precariedad.

De manera complementaria, el índice continuo fue recategorizado en terciles. Se asignaron tres niveles —*alto*, *medio* y *bajo*— que permiten clasificar a los hogares según su situación relativa de los materiales de la vivienda dentro de la muestra. No obstante, se conservaron tanto la versión continua del índice como su versión categórica en terciles, con el objetivo de evaluar posteriormente su utilidad diferencial como predictoras de la variable respuesta en los modelos supervisados.

3.4. Elección del Dataset Final

Con el objetivo de predecir la variable respuesta `control_mamografia` y analizar los factores que se asocian con la realización de esta práctica preventiva, se construyeron tres versiones preliminares del dataset. Cada una presentó una estrategia distinta de selección y transformación de variables, y fueron comparadas en términos de completitud, interpretabilidad y potencial explicativo. A continuación, se describen las tres versiones consideradas:

- **Versión 1 — Dataset reducido con recodificaciones e índices:** Esta versión incluyó únicamente las variables recodificadas —tanto las generadas en este trabajo como las provistas por los responsables del diseño de la ENFR— y los índices construidos a partir de variables originales. Se excluyeron, por tanto, las variables utilizadas para generar estas nuevas transformaciones. Esto permitió reducir la dimensionalidad del dataset, priorizando indicadores más sintéticos y conceptualmente robustos. También se conservaron las variables que no requerían recodificación ni fueron utilizadas en la construcción de otros indicadores.

- **Versión 2 — Dataset con variables originales sin transformaciones:** En esta segunda versión se adoptó un enfoque opuesto al anterior: se incluyeron todas las variables originales seleccionadas previamente, sin aplicar recodificaciones ni construir índices adicionales. El objetivo fue mantener la mayor cantidad de información bruta disponible, aunque con el costo potencial de mayor ruido o redundancia en el análisis.

- **Versión 3 — Dataset híbrido:** Esta versión combinó los enfoques anteriores, incorporando tanto las variables recodificadas como los índices, junto con las variables originales utilizadas en su construcción. A su vez, se priorizó el uso de recodificaciones realizadas oficialmente por el equipo técnico de la ENFR, descartando las variables crudas en los casos en que ya existía una transformación validada. Esta estrategia buscó un equilibrio entre riqueza informativa y claridad analítica.

Para evaluar el desempeño de cada versión y definir cuál resultaba más adecuada como insumo para la modelización, se entrenaron tres modelos de *Random Forest*, uno por cada dataset, utilizando un proceso de optimización de hiperparámetros automatizado con `Optuna` [28]. El *Random Forest* es un algoritmo de aprendizaje automático basado en árboles de decisión que construye múltiples árboles a partir de subconjuntos aleatorios de los datos y promedia sus predicciones. Esta estrategia le permite reducir el sobreajuste y mejorar la capacidad de generalización frente a modelos individuales. Su versatilidad, interpretabilidad relativa y robustez ante valores atípicos o relaciones no lineales lo hacen especialmente adecuado para tareas de clasificación en contextos sociales [29]. Por otro lado, `Optuna` es una biblioteca de código abierto para la optimización eficiente de hiperparámetros que utiliza estrategias de búsqueda inteligentes —como búsqueda bayesiana

o TPE (Tree-structured Parzen Estimator)— para encontrar combinaciones óptimas de parámetros con menor costo computacional que una búsqueda exhaustiva. La combinación de *Random Forest* con `Optuna` permitió ajustar automáticamente parámetros clave (como la profundidad de los árboles o la cantidad de estimadores), maximizando así el rendimiento de cada versión. Las métricas utilizadas para la comparación fueron la exactitud (accuracy), el F1-score y el área bajo la curva ROC (AUCROC).

Una vez seleccionado el dataset, se llevó a cabo una evaluación adicional para verificar si las variables recodificadas por este trabajo aportaban efectivamente valor al modelo en comparación con las versiones originales provistas por la ENFR. Para ello, se reemplazó cada variable recodificada de forma individual por su versión original correspondiente, manteniendo el resto del dataset sin modificaciones.

3.5. Modelado

Una vez seleccionado el dataset final, se avanzó con la etapa de modelado, dividida en dos fases complementarias. En ambas se trabajó con tres algoritmos de aprendizaje supervisado del tipo ensemble: *Random Forest* [29], *XGBoost* [30] y *HistGradientBoosting* [31], cuyos hiperparámetros fueron optimizados utilizando la biblioteca `Optuna` [28], al igual que en la etapa de evaluación comparativa entre datasets.

Random Forest, ya introducido en la sección anterior, fue complementado con dos variantes basadas en *Gradient Boosting*:

- *XGBoost* (Extreme Gradient Boosting) es un algoritmo que construye árboles secuencialmente, donde cada uno corrige los errores del anterior, optimizando una función objetivo diferenciable. Su eficiencia y capacidad de regularización lo han convertido en una de las herramientas más populares en la ciencia de datos contemporánea.
- *HistGradientBoosting*, por su parte, es una implementación eficiente de *Gradient Boosting* incluida en `scikit-learn`, una biblioteca de aprendizaje automático de código abierto ampliamente utilizada en Python [32]. Este modelo emplea histogramas para acelerar el entrenamiento y manejar de forma eficaz variables numéricas continuas. Además de su buen rendimiento predictivo, presenta una integración nativa con otras herramientas del ecosistema `scikit-learn`, lo que facilita su evaluación y visualización dentro de pipelines reproducibles.

3.5.1. Selección del Mejor Modelo Completo

En una primera instancia, se entrenaron los tres modelos mencionados sobre el dataset final completo —*Random Forest*, *XGBoost* y *HistGradientBoosting*— con hiperparámetros optimizados mediante `Optuna`. La evaluación se realizó utilizando tres métricas estándar en problemas de clasificación supervisada:

- **Accuracy:** mide la proporción de observaciones correctamente clasificadas sobre el total de casos. Es una métrica general útil, aunque puede resultar engañosa en contextos con clases desbalanceadas.
- **F1-score:** representa el promedio armónico entre la precisión (proporción de verdaderos positivos entre los positivos predichos) y la sensibilidad o recall (proporción

de verdaderos positivos entre los positivos reales). Es especialmente relevante cuando los falsos positivos y los falsos negativos tienen consecuencias importantes, como suele ocurrir en problemas de salud pública.

- **AUC-ROC (Área Bajo la Curva ROC):** indica la capacidad del modelo para distinguir entre las clases positivas y negativas. A diferencia del accuracy, no depende de un umbral de clasificación específico y es menos sensible al desbalance de clases, lo que la convierte en una métrica robusta para problemas donde una clase es más frecuente que la otra.

Dado el contexto del problema —la predicción de la realización de una mamografía como práctica preventiva de salud—, se priorizó el *AUC-ROC* como métrica principal para la comparación entre modelos. Esta métrica permite evaluar la capacidad del modelo para discriminar entre mujeres que se realizaron o no la mamografía, sin depender de un umbral específico de clasificación. Por su naturaleza integral y su robustez frente a distintos escenarios, el *AUC-ROC* resulta especialmente útil como criterio de selección en tareas de clasificación sanitaria.

Finalmente, se seleccionó aquel modelo que mostró el mejor desempeño para ser utilizado en las etapas posteriores del análisis, incluidas la detección de grupos de riesgo y la evaluación de variables predictoras clave.

3.5.2. Identificación de Variables Relevantes

En una segunda etapa de modelado, se retuvieron principalmente variables relacionadas con condiciones sociodemográficas, cobertura sanitaria, percepciones de salud y características del hogar. En cambio, se excluyeron aquellas vinculadas a la realización de otras prácticas preventivas —como el test de Papanicolaou (PAP), controles de glucemia, presión arterial o colesterol—, dado que se esperaba que estuvieran fuertemente asociadas con la realización de mamografías. El objetivo de esta decisión fue identificar predictores específicos de la conducta de tamizaje mamográfico, más allá del hábito general de control médico.

Se volvieron a entrenar los tres modelos ajustados y, para cada uno, se identificaron las 40 variables con mayor importancia relativa en la predicción. Luego, se seleccionaron aquellas variables que aparecían consistentemente entre las más importantes en los tres modelos, lo cual permitió obtener un conjunto robusto de predictores clave.

A partir de este subconjunto, se buscó validar estadísticamente la relevancia de las variables predictoras identificadas por los modelos de *machine learning*. El objetivo de esta fase no fue construir un nuevo modelo de predicción, sino evaluar la significancia estadística de cada variable y la magnitud de su efecto sobre la probabilidad de realización de la mamografía, a través de la estimación de los *odds ratios* (OR) ajustados. Para ello, se adoptó una estrategia en dos etapas.

En primer lugar, se evaluó la presencia de multicolinealidad entre las variables independientes mediante el cálculo del *Factor de Inflación de la Varianza* (VIF, por sus siglas en inglés). Este indicador cuantifica cuánto se incrementa la varianza de los coeficientes estimados en un modelo de regresión debido a la correlación lineal entre predictores. Valores de VIF superiores a 7 o 10 suelen indicar problemas potenciales de colinealidad que pueden afectar la estabilidad de las estimaciones y significación estadística [33].

En segundo lugar, se estimó un modelo de regresión logística para explicar la probabilidad de realización de una mamografía. Esta técnica es adecuada cuando la variable dependiente es dicotómica y permite interpretar la contribución de cada predictor mediante la estimación de los *odds ratios* (OR), que reflejan la magnitud y dirección del efecto ajustado de cada variable sobre la variable respuesta. A partir del modelo, se obtuvieron los *p-valores* asociados a cada coeficiente, que indican su significancia estadística.

Este enfoque permitió identificar no solo las variables más relevantes desde el punto de vista predictivo, sino también aquellas que presentan asociaciones estadísticamente significativas y sustantivamente interpretables con la realización de la mamografía.

3.6. Grupos de Riesgo

Para identificar grupos de riesgo en relación con la no realización de mamografías, se utilizó el modelo previamente seleccionado para predecir la probabilidad de pertenecer a la clase negativa (`control_mamografia` $= 0$). A partir de las probabilidades generadas sobre el conjunto de testeo, se definió como grupo de riesgo al subconjunto de mujeres con una probabilidad igual o superior a 0,8 de no haberse realizado esta práctica preventiva.

Este análisis se enfocó en las variables categóricas más relevantes identificadas en la etapa anterior, consideradas factores de riesgo robustos para la exclusión preventiva. Se comparó la frecuencia relativa de cada categoría en el grupo de riesgo respecto del total del dataset, calculando la diferencia porcentual relativa para detectar aquellas características sobrerrepresentadas en el grupo.

Asimismo, se exploraron combinaciones de 4, 5 y 6 variables con mayor sobrerrepresentación conjunta, con el objetivo de identificar perfiles multivariados de alto riesgo. Para cada combinación, se calculó la proporción de personas del grupo de riesgo que presentaban simultáneamente todas las características definidas, y se la comparó con su frecuencia en la muestra total. Este procedimiento permitió identificar configuraciones específicas de factores que, en conjunto, se asocian con una probabilidad significativamente mayor de no haberse realizado una mamografía, aportando evidencia útil para el diseño de intervenciones focalizadas.

De esta forma, se buscó caracterizar integralmente al grupo de personas con alta probabilidad de no realizarse una mamografía, a partir de los factores de riesgo ya identificados por el modelo.

4. RESULTADOS

4.1. Análisis Exploratorio de Muestra Analítica

La muestra analizada está compuesta por 4.693 mujeres residentes en Argentina de entre 50 y 69 años, provenientes de la Encuesta Nacional de Factores de Riesgo (ENFR) 2018. En la Figura 4.1 se muestra la distribución de la variable dependiente del estudio: la realización de una mamografía en los últimos dos años. Se observa que el 62,6 % de las mujeres de la muestra reportó haberse realizado la práctica preventiva en ese período, mientras que el 36,9 % no lo hizo.

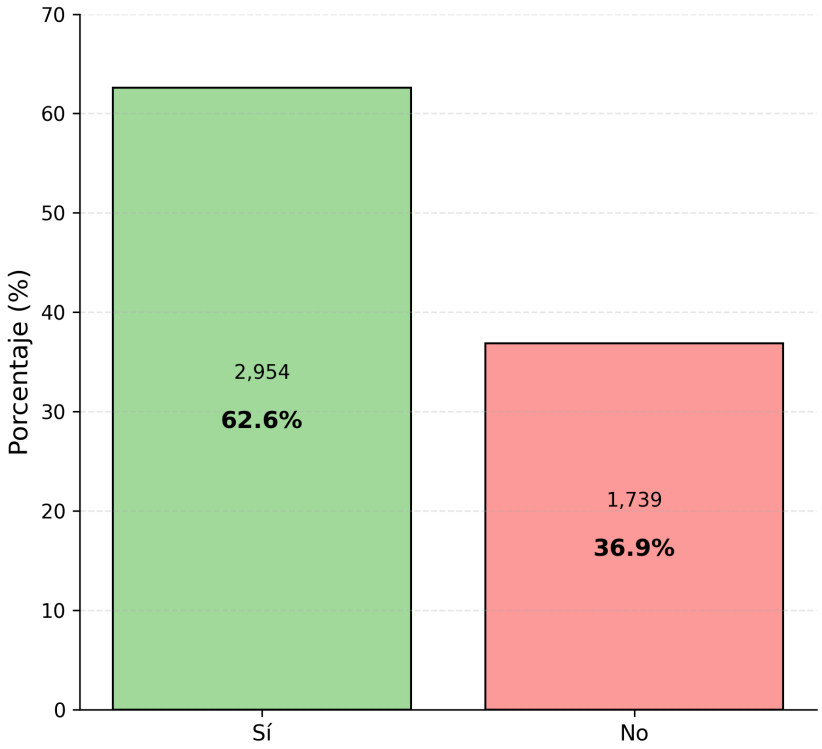


Figura 4.1: Distribución de mujeres de 50 a 69 años en la Argentina según realización de mamografía en los últimos dos años. Se excluyen los registros con Ns/Nc.

Una vez descripta la distribución general de la variable dependiente `control_mamografia`, se exploraron sus asociaciones con las variables independientes en el marco de un análisis exploratorio inicial. En esta sección se presentan los resultados de algunos de estos cruces, con el objetivo de detectar patrones preliminares de acceso desigual a la mamografía entre distintos grupos de mujeres.

Se seleccionaron aquellos que resultaron más relevantes por la magnitud de las diferencias observadas y por su capacidad para ilustrar desigualdades estructurales que atraviesan el acceso a esta práctica preventiva.

Una de las variables más ilustrativas fue la provincia de residencia. En la Figura 4.2, se presenta un mapa de calor que muestra el porcentaje de mujeres de entre 50 y 69 años en la muestra que accedieron a una mamografía en los últimos dos años, desagregado por provincia. Estos valores se calculan a partir de los datos observados en la muestra analítica.

En este mapa se puede observar un patrón de desigualdad territorial: mientras que CABA y La Pampa exhiben los valores más altos —con más del 82 % de cobertura—, provincias del norte como Formosa, Chaco y Corrientes registraron porcentajes considerablemente más bajos, por debajo del 50 %, incluso llegando a 37 %, como Santiago del Estero. Estas diferencias sugieren la existencia de factores estructurales que limitan el acceso a los servicios de salud preventiva en ciertas regiones del país.

Otra variable que mostró una asociación marcada con la realización de mamografías fue el nivel educativo del jefe o jefa de hogar. Tal como se observa en la Figura 4.3, existe una relación positiva entre el nivel educativo alcanzado y la proporción de mujeres que accedieron a esta práctica preventiva.

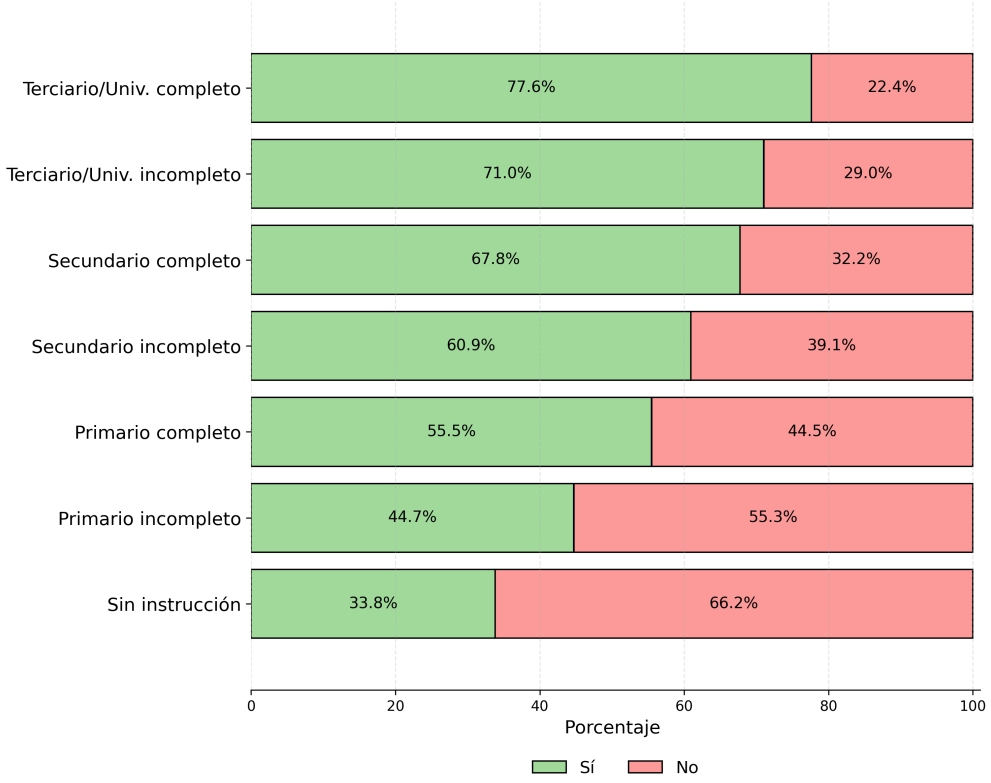


Figura 4.3: Realización de mamografía en mujeres de 50 a 69 años en los últimos dos años según nivel educativo del jefe/a de hogar.

Entre los hogares con jefatura sin instrucción formal, apenas el 33,8 % de las mujeres se realizaron una mamografía en los últimos dos años. En cambio, cuando el jefe o jefa de hogar tiene estudios universitarios completos, ese porcentaje asciende al 77,6 %. Esta tendencia es consistente a lo largo de los distintos niveles educativos y sugiere que la educación —como indicador de capital cultural y acceso a recursos— cumple un rol importante en la utilización de servicios preventivos de salud.

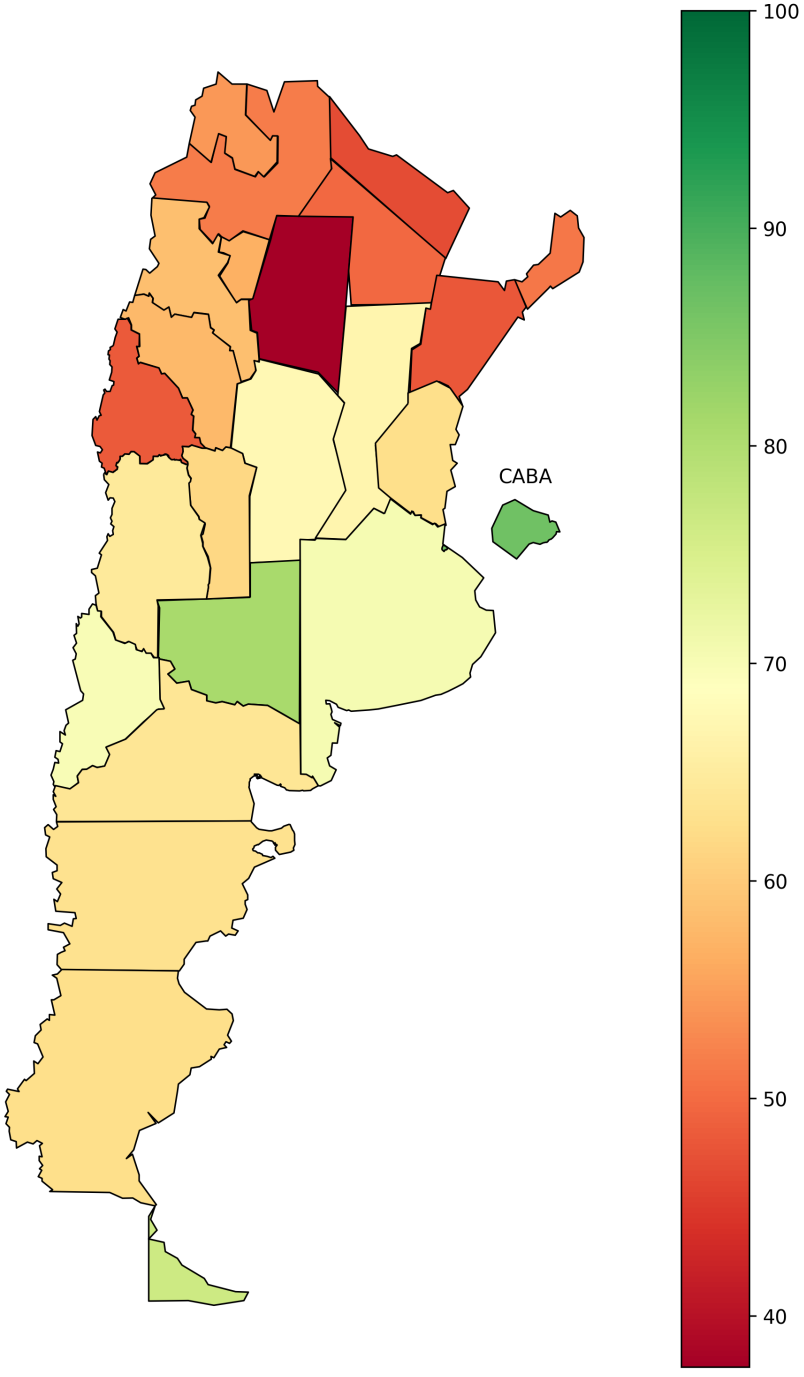


Figura 4.2: Porcentaje de mujeres de 50 a 69 años que se realizaron una mamografía en los últimos dos años, por provincia. Cálculos realizados sobre la muestra analítica.

El nivel socioeconómico del hogar también mostró una asociación clara con la realización de mamografías. Esta variable es la que fue construida a partir de los ingresos del hogar disponibles en la encuesta, agrupada en cinco categorías ordinales (ABC1, C2, C3, D1 y D2E).

En la Figura 4.4 se observa que el porcentaje de mujeres que accedieron a esta práctica preventiva disminuye sistemáticamente a medida que se desciende en la escala de nivel socioeconómico.

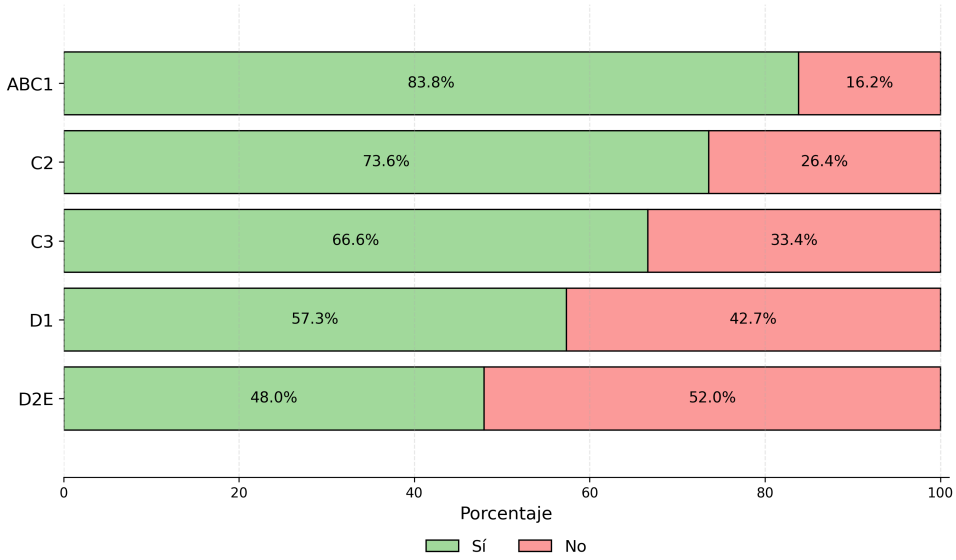


Figura 4.4: Realización de mamografía en mujeres de 50 a 69 años en los últimos dos años según nivel socioeconómico del hogar.

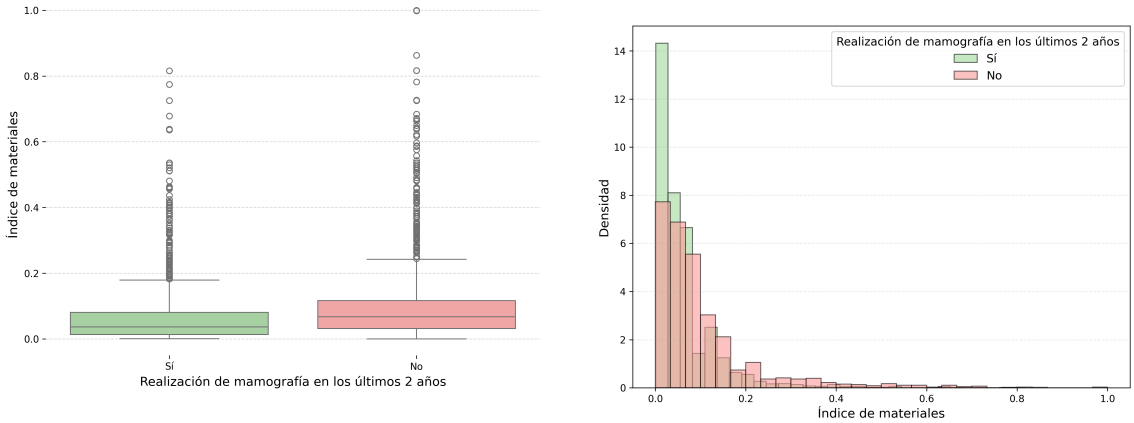
Entre las mujeres pertenecientes al estrato ABC1, el 83,8 % se realizó una mamografía en los últimos dos años. Esta proporción cae a 73,6 % en el nivel C2, 66,6 % en C3, 57,3 % en D1 y apenas 48,0 % en el nivel D2E.

Por último, se exploró la asociación entre la realización de mamografías y el índice material, una variable continua construida a partir de un Análisis de Correspondencias Múltiples sobre condiciones habitacionales relevadas por la ENFR 2018 (ver sección Metodología). Esta versión continua del índice permite captar gradientes más sutiles en las condiciones materiales, cuanto mayor es el índice, mayor la precariedad de los materiales de la vivienda.

La Figura 4.5 presenta dos representaciones complementarias de esta relación. El boxplot (a) muestra que las mujeres que no se realizaron una mamografía en los últimos dos años se caracterizan por pertenecer mayoritariamente a hogares con mayor precariedad: el valor mediano del índice es más alto y la distribución es más dispersa en ese grupo. A su vez, el histograma (b) refuerza esta observación al evidenciar una mayor concentración de mujeres sin mamografía en los tramos superiores del índice, es decir, en situaciones materiales más desfavorables.

Si bien la mayoría de las mujeres —tanto con como sin mamografía realizada en los últimos dos años— se concentran en niveles bajos del índice, el contraste se vuelve más

claro en los sectores más vulnerables: en los valores altos del índice, la proporción de mujeres que no accedieron al estudio es visiblemente mayor. Esto sugiere una posible relación entre las condiciones materiales más extremas y el menor acceso al control. Esta tendencia también se refleja en el histograma (b), donde se observa que la mayor densidad de casos —especialmente entre quienes sí realizaron la mamografía— se concentra en los valores más bajos del índice, es decir, entre quienes viven en condiciones materiales relativamente menos precarias.



(a) Boxplot del índice según realización de mamografía

(b) Distribución del índice por grupo

Figura 4.5: Relación entre precariedad material y realización de mamografía (versión continua del índice). A mayor valor del índice, mayor es la precariedad del hogar.

4.2. Evaluación Comparativa de Versiones de Dataset

A continuación se presentan los resultados obtenidos para las tres versiones de dataset evaluadas. En la Tabla 4.1 se muestran las métricas de *accuracy*, *F1-score* y *AUC-ROC* correspondientes a cada modelo de *Random Forest*.

Tabla 4.1: Desempeño de modelos de Random Forest según versión de dataset.

Versión de dataset	**Accuracy**	**F1-score**	**AUC-ROC**
Versión 2 — Originales sin transformar	0.8552	0.8838	**0.9051**
Versión 1 — Recodificadas e índices	0.8520	0.8821	0.8998
Versión 3 — Híbrido (seleccionada)	0.8498	0.8806	0.8955

Si bien la Versión 3 no obtuvo el mejor resultado absoluto en *AUC-ROC*, sus métricas de desempeño fueron altamente competitivas y las diferencias con las otras versiones resultaron marginales (menores a 0,01 en todas las métricas). Dado que los rendimientos fueron prácticamente equivalentes, se optó por seleccionar la Versión 3 por su mayor simplicidad interpretativa y coherencia con los objetivos del estudio. Por ello, la decisión final se basó en un criterio más amplio, que combinó aspectos técnicos, interpretativos y estratégicos.

La Versión 3 fue diseñada como un dataset híbrido que integra las ventajas de las variables originales y las transformaciones realizadas, priorizando aquellas recodificaciones validadas por el equipo técnico de la ENFR y descartando variables crudas cuando ya

existía una versión más robusta. Esta versión favorece la interpretación de los resultados, un aspecto central para los objetivos de este trabajo.

Ajustes Exploratorios sobre la Versión Seleccionada

Se exploraron ajustes puntuales sobre la estructura de la Versión 3 con el objetivo de evaluar posibles mejoras. Estas pruebas mantuvieron fijos los hiperparámetros previamente optimizados y buscaron detectar variables que pudieran enriquecer el modelo sin comprometer su simplicidad.

- **Inclusión de `quintil_uc`:** Esta variable —indicador del ingreso del hogar— mostró una leve mejora en *AUC-ROC* (hasta 0,8963) y aportó valor interpretativo relevante, por lo que se decidió incorporarla, a pesar de que el modelo ya incluía la variable de Nivel Socioeconómico (NSE) construida específicamente para este estudio.
- **Reemplazo de `control_colesterol` por `bico01` y `bico02`:** La variable `control_colesterol` indica si la mujer se realizó un control de colesterol en los últimos dos años, mientras que `bico01` señala si alguna vez se lo realizó y `bico02` detalla hace cuánto. Si bien estas variables adicionales permiten una descomposición más fina del comportamiento, su incorporación no mejoró las métricas del modelo y podría introducir ruido o redundancia. Por ello, se optó por mantener la versión sintética original.

A partir de este análisis, se realizó a la Versión 3 del dataset la incorporación de `quintil_uc` como variable adicional.

Ninguno de los reemplazos de la evaluación adicional de variables recodificadas produjo una mejora en el *AUC-ROC* del modelo, que fue la métrica prioritaria para la elección de la versión final. Las diferencias en F1-score y accuracy también fueron mínimas en todos los casos. La única excepción fue la variable de nivel educativo del respondente, cuya versión original obtuvo un valor levemente superior en *AUC-ROC*. No obstante, se decidió mantener la versión recodificada, ya que sintetiza los niveles educativos en una estructura ordinal más compacta y coherente con los objetivos del estudio, lo que facilita su interpretación y su uso en los modelos.

En síntesis, tras comparar distintas versiones del dataset, aplicar ajustes específicos sobre la estructura seleccionada e inspeccionar el aporte real de las recodificaciones implementadas, se definió una base analítica robusta y coherente con los objetivos del estudio. Esta versión final —basada en una estrategia híbrida de selección de variables— logra un equilibrio entre rendimiento predictivo, claridad interpretativa y viabilidad metodológica. Sobre esta base se construyeron y evaluaron los modelos que se presentan en la próxima sección.

4.3. Ajuste Supervisado con Dataset Completo

En esta etapa inicial se entrenaron tres algoritmos de aprendizaje supervisado para predecir la realización de mamografías en mujeres de entre 50 y 69 años. Se utilizó el conjunto completo de variables disponibles, incluyendo tanto factores estructurales como

otras prácticas preventivas de salud (variables de control). Los modelos considerados fueron *Random Forest*, *XGBoost* y *HistGradientBoosting*, con hiperparámetros optimizados mediante la biblioteca `Optuna`.

4.3.1. Importancia de Variables

En los tres modelos ajustados se observaron patrones consistentes en cuanto a la importancia relativa de las variables predictoras. A pesar de que cada algoritmo emplea criterios distintos para estimar dicha relevancia, se identificó un conjunto estable de predictores destacados que se repitieron en los tres enfoques.

Como se muestra en la Tabla 4.2, muchas de las variables con mayor importancia relativa corresponden a controles de otras prácticas preventivas de salud, como el Papanicolaou (`control_pap`), estudios de colon (`control_colon`) y colesterol (`control_colesterol`), junto con controles de hipertensión, glucemia y hábitos saludables generales.

Tabla 4.2: Variables con mayor importancia relativa en los tres modelos supervisados.

Variable	**Descripción**
`control_pap`	Realización del test de Papanicolaou (PAP) en los últimos dos años.
`control_colon`	Realización de algún estudio preventivo de cáncer de colon alguna vez.
`control_colesterol`	Control de colesterol en los últimos dos años.
`indice_material`	Índice sintético construido a partir de condiciones habitacionales.
`bidi07_1.0`	Última vez que se midió la glucemia o el azúcar en sangre: hace menos de un año.
`biaf10_04`	Prácticas de actividad física en la última semana: realizadas con fines deportivos o de mejora física.
`imc_categorias`	Índice de masa corporal (IMC) del respondente, categorizado por rangos.
`cobertura_salud`	Tipo de cobertura médica: con obra social o prepaga.
`nivel_instruccion_recod_1.0`	Nivel educativo alcanzado por el respondente: nivel bajo educativo.
`nivel_instruccion_j_recod_1.0`	Nivel educativo del jefe o jefa del hogar: nivel bajo educativo.
`bhcv06_recod_1`	Tipo de combustible utilizado para cocinar en el hogar: gas de red.
`promedio_fv_diario`	Promedio diario de consumo de frutas y verduras (en porciones).

4.3.2. Comparación de Modelos y Selección Final

La Tabla 4.3 presenta las métricas de desempeño obtenidas por cada modelo. Todos alcanzaron niveles altos de rendimiento general, aunque *XGBoost* se destacó en las tres métricas evaluadas. En particular, obtuvo el mayor valor de accuracy (85,73 %), el mayor

F1-score (88,66 %) y, especialmente, el mayor *AUC-ROC* (0,9069), métrica prioritaria en este estudio.

Tabla 4.3: Métricas de desempeño de los modelos predictivos incluyendo todo el conjunto de predictoras seleccionadas.

Modelo	Accuracy	F1-score	AUC-ROC
Random Forest	0,8445	0,8821	0,9052
HistGradientBoosting	0,8552	0,8853	0,9022
XGBoost	**0,8573**	**0,8866**	**0,9069**

Por este motivo, se seleccionó *XGBoost* como modelo base para los análisis posteriores orientados a la detección de grupos de riesgo.

4.4. Ajuste de Modelos sin Variables de Conducta Sanitaria

El objetivo del ajuste de variables predictoras fue identificar con mayor claridad qué variables sociales, económicas, materiales o geográficas explican la desigualdad en el acceso sin depender de conductas preventivas previas.

Cabe destacar que no se excluyeron variables vinculadas a percepciones personales o conductas generales de salud —como la autoevaluación del estado de salud o acciones para controlar el peso—, ya que estas no refieren a controles médicos específicos, sino a disposiciones generales frente al cuidado personal.

La Tabla 4.4 muestra las variables eliminadas en relación al set completo utilizado en la sección anterior.

Tabla 4.4: Variables no incluídas en el modelo por representar prácticas preventivas ya realizadas.

Variable	Descripción
`control_hipertension`	Control de presión arterial en los últimos dos años.
`bipc01`	En el último año un médico, un enfermero u otro profesional de la salud le ha dicho que tiene que bajar de peso.
`control_colesterol`	Control de colesterol alguna vez por autorreporte.
`control_diabetes`	Control de glucemia o azúcar en sangre alguna vez por autorreporte.
`control_pap`	Realización del test de Papanicolaou (PAP) en los últimos dos años.
`control_colon`	Realización de algún estudio preventivo de cáncer de colon alguna vez.
`bidi07`	Última vez que se midió la glucemia o el azúcar en sangre.
`bicc03_recod`	Última vez que se realizó un estudio preventivo de cáncer de colon.
`prevalencia_hipertension`	Prevalencia de presión arterial elevada por autorreporte.
`prevalencia_colesterol`	Prevalencia de colesterol elevado por autorreporte.
`prevalencia_diabetes`	Prevalencia de glucemia elevada/diabetes por autorreporte.

Como puede observarse, los modelos continúan alcanzando niveles satisfactorios de las métricas analizadas, aunque algo menores que los obtenidos con el dataset completo (ver Tabla 4.5). Esto sugiere que los factores sociales, materiales y demográficos conservan un valor predictivo importante aún sin la presencia de variables de conducta preventiva.

Tabla 4.5: Métricas de desempeño de los modelos sin variables de control sanitario.

Modelo	**Accuracy**	**F1-score**	**AUC-ROC**
Random Forest	0,6731	0,7744	0,7351
XGBoost	0,6869	0,7707	0,7364
HistGradientBoosting	**0,6922**	**0,7758**	**0,7391**

Una vez confirmado que los modelos conservan un nivel de desempeño razonable incluso sin las variables de control, se procedió a comparar nuevamente las variables con mayor importancia relativa en cada uno de ellos. El objetivo fue identificar qué factores estructurales, sociales, económicos o contextuales se mantienen como predictivos en ausencia de información directa sobre conductas sanitarias previas.

La Tabla 4.6 que se muestra en la siguiente página presenta las variables que aparecieron de forma consistente como relevantes en los tres modelos ajustados en esta etapa.

Como se puede observar, muchas de estas variables remiten a condiciones estructurales del hogar (como el índice material o el hacinamiento), a dimensiones socioeconómicas (nivel educativo del respondente y del jefe del hogar, tipo de cobertura de salud), así como a características contextuales (tamaño del aglomerado urbano, provincia de residencia). También emergen como relevantes ciertos indicadores de percepción subjetiva de la salud de cada uno (como la autoevaluación del estado general o la presencia de dolor/malestar al momento de la realización de la encuesta), junto con algunos hábitos vinculados al cuidado cotidiano, como el uso del cinturón de seguridad, la práctica de actividad física con fines saludables y el consumo de frutas y verduras. Además, se puede ver como relevante la situación conyugal de la respondente, particularmente que está casada.

Tabla 4.6: Variables con alta importancia relativa en los tres modelos sin controles sanitarios.

Variable	**Descripción**
promedio_fv_diario	Promedio diario de consumo de frutas y verduras (en porciones).
indice_material	Índice sintético construido a partir de condiciones habitacionales.
hacinamiento	Nivel de hacinamiento del hogar.
nse_2018_empirico	Nivel socioeconómico, construido a partir de los ingresos del hogar.
bipc03	En estos momentos está haciendo algo para mantener controlado su peso
cobertura_salud	Tipo de cobertura médica del respondente.
cobertura_salud_j	Tipo de cobertura médica del jefe/a del hogar.
bisg05	Nivel de dolor o malestar autorreportado al momento de la entrevista.
bile03_1	Si maneja o viaja en auto usa siempre cinturón de seguridad
nivel_instruccion_j_recod_1	Nivel educativo del jefe o jefa del hogar: nivel bajo educativo.
nivel_instruccion_recod_1	Nivel educativo alcanzado por el respondente: nivel bajo educativo.
tamanio_aglomerado_4	Tamaño de aglomerado: Menos de 150.000 habitantes.
bisg02	Nivel de dificultad para caminar autorreportado al momento de la entrevista.
consumo_tabaco_100_3	Condición de fumador: No fumador.
bhcv06_recod_1	Para cocinar utiliza principalmente: gas de red.
barreras_fyv_recod_4	Tiene barreras económicas para el consumo de frutas y verduras.
bhsl06_2.0	Trabaja habitualmente entre 35 y 45 horas semanales.
barreras_fyv_recod_2	No presenta barreras en el consumo de frutas y verduras: considera que consume una cantidad adecuada.
cod_provincia_6	Vive en la provincia de Buenos Aires.
rango_edad_j_5	El jefe del hogar más de 65 años.
bile03_98	No viaja en auto.
biaf10_04	Las actividades físicas que realizó en la última semana fueron para mejorar su condición física/hacer deporte
nivel_instruccion_recod_3	Nivel educativo alcanzado por el respondente: nivel alto educativo.
bhch05_2	La respondete es casada.
bisg01	Autopercepción del estado de salud
cod_provincia_2	Vive en la Ciudad Autónoma de Buenos Aires.
tamanio_aglomerado_1	Tamaño de aglomerado: Más de 1.500.000 habitantes.

4.4.1. Modelo Explicativo: Variables Significativas

En esta etapa, se trabajó sobre el conjunto de variables identificadas como más relevantes en los modelos supervisados sin controles sanitarios (ver Tabla 4.6).

Previo a la estimación del modelo, se evaluó la presencia de colinealidad entre los predictores mediante el cálculo del *Variance Inflation Factor* (VIF). Todos los valores resultaron por debajo del umbral crítico de 7, por lo que no fue necesario excluir ninguna variable por colinealidad.

A continuación se presentan los resultados del modelo de regresión logística. La Tabla 4.7 muestra las variables que resultaron estadísticamente significativas (p-valor $< 0{,}05$), ordenadas por su nivel de significancia.

Tabla 4.7: Variables con significancia estadística (p-valor $< 0{,}05$).

Variable (descripción)	**p-valor**
`bile03_1`: Siempre usa cinturón de seguridad al viajar en auto	<0,000001
`bhch05_2`: Situación conyugal del respondente: casada	<0,000001
`biaf10_04_1`: Las actividades físicas realizadas en la última semana fueron para mejorar su condición física/hacer deporte	<0,000001
`cod_provincia_6`: Provincia de residencia: Buenos Aires	0,000003
`rango_edad_j_5`: Edad del jefe/a de hogar: mayor a 64 años	0,000004
`cod_provincia_2`: Provincia de residencia: CABA	0,000014
`bhsl06_2`: Cantidad de horas semanales trabajadas: Entre 35 y 45 horas	0,000034
`tamanio_aglomerado_4`: Residencia en aglomerado de menos de 150.000 habitantes	0,000257
`cobertura_salud`: Cobertura de salud del respondente: obra social o prepaga	0,000450
`bisg02_3`: Percepción de salud del respondente: buena	0,000802
`cod_provincia_86`: Provincia de residencia: Santiago del Estero	0,001939
`bisg05_2`: Tiene dolor/malestar moderado el día de hoy	0,002471
`hacinamiento`: Nivel de hacinamiento del hogar	0,005682
`indice_material`: Índice de condiciones materiales del hogar	0,014376
`bhcv06_recod_1`: Para cocinar utiliza gas de red	0,020673
`barreras_fyv_recod_4`: No tiene barreras en el acceso a frutas y verduras	0,030314
`nse_2018_empirico_D2E`: Nivel socioeconómico D2E	0,042506

Se pueden observar múltiples variables con asociación estadísticamente significativa con la no realización de mamografías. Entre ellas, se incluyen factores vinculados al entorno físico y social —como el nivel de hacinamiento y el índice de condiciones materiales del hogar—, así como variables territoriales (residencia en determinadas provincias o en aglomerados pequeños). También se identifican variables relacionadas con el acceso al sistema de salud y el estado percibido por la persona encuestada, tales como la percepción de salud, la cobertura sanitaria, el dolor/malestar actual, y ciertas prácticas preventivas o de autocuidado. Asimismo, algunas características del jefe o jefa de hogar —como su edad o situación conyugal— resultan significativas en el modelo.

Estas asociaciones fueron obtenidas luego de ajustar por el resto de las variables inclui-

das en el modelo, lo que permite identificar su contribución específica al riesgo estimado de no haberse realizado una mamografía en los últimos dos años.

Además del análisis de significancia estadística mediante p-valores, se calcularon los *odds ratios* (OR) para aquellas variables que resultaron estadísticamente significativas en el modelo (p-valor $< 0{,}05$), con el objetivo de estimar la magnitud del efecto asociado a cada factor.

Los *odds ratios* permiten interpretar la fuerza de asociación entre cada predictor y la variable dependiente, en este caso, la probabilidad de haberse realizado una mamografía en los últimos dos años. Un OR mayor a 1 indica mayores chances de realización, mientras que un OR menor a 1 indica menores chances, es decir, un posible factor de riesgo.

La Figura 4.6 presenta un gráfico tipo *forest plot* con los OR e intervalos de confianza al 95 % para las variables que resultaron estadísticamente significativas (p-valor $< 0{,}05$). Dado que la variable dependiente toma el valor 1 si la persona se realizó una mamografía, se puede interpretar que las variables con OR menor a 1 presentan asociación con menor probabilidad de realización, mientras que los valores mayores a 1 se vinculan con una mayor probabilidad.

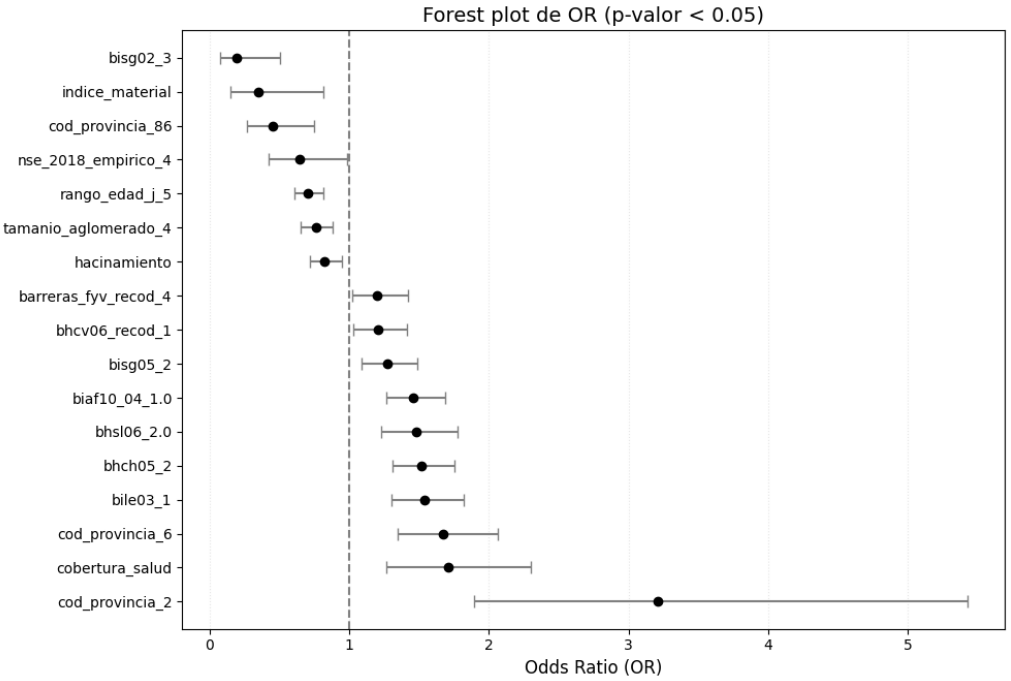


Figura 4.6: Forest plot de los *odds ratios* (OR) para la realización de mamografía en los ultimos 2 años con intervalos de confianza al 95 % para las variables estadísticamente significativas (p-valor < 0.05).

Entre los factores con *odds ratios* (OR) menores a 1 —es decir, asociados a menores chances de realización de mamografías— se encuentran la percepción regular o mala del estado de salud, el índice material del hogar y un nivel socioeconómico bajo. Por el contrario, variables con OR mayores a 1, como contar con cobertura de salud o residir en la Ciudad Autónoma de Buenos Aires o en la provincia de Buenos Aires, se asocian con

mayores chances de acceso al estudio, actuando como factores de protección.

4.5. Grupos de Riesgo

En esta sección del trabajo se presentan los resultados del análisis descriptivo del grupo de riesgo, comenzando por las variables categóricas individuales. Se identificaron aquellas categorías sobrerrepresentadas en el grupo en comparación con su frecuencia en la muestra total, lo que permite visualizar qué factores están más estrechamente asociados con una probabilidad elevada de exclusión preventiva.

En primer lugar, se analizó la variable `cod_provincia`, la cual indica el código de la provincia en donde se encuestó a cada respondente. La Tabla 4.8 presenta las provincias ordenadas según su sobrerrepresentación en el grupo de riesgo. Se reporta la frecuencia relativa de cada provincia dentro del grupo de riesgo y en la muestra total, ambas expresadas en porcentaje. También se incluye la diferencia absoluta (en puntos porcentuales) y la diferencia relativa porcentual entre ambas proporciones. Los valores positivos indican una mayor presencia en el grupo de riesgo en comparación con la muestra general.

Tabla 4.8: Participación diferencial en grupo de mayor riesgo de no realización de mamografía según provincia.

Provincia	**Grupo de riesgo (%)**	**Total muestra (%)**	**Dif. Absoluta (pp)**	**Dif. relativa (%)**
Formosa	6,28	2,56	+3,73	+145,8
Chaco	6,81	3,30	+3,50	+106,2
San Juan	6,81	3,30	+3,50	+106,2
Corrientes	8,90	4,69	+4,21	+89,9
Santiago del Estero	2,62	1,38	+1,23	+89,1
Misiones	5,76	3,30	+2,46	+74,4
Entre Ríos	7,85	4,79	+3,06	+63,9
Catamarca	3,66	2,45	+1,22	+49,6
La Rioja	4,19	3,19	+0,99	+31,1
Tucumán	3,14	2,56	+0,59	+22,9
Salta	4,71	3,94	+0,77	+19,6
Chubut	3,66	3,94	-0,28	-7,0
Neuquén	2,09	2,56	-0,46	-18,1
Jujuy	2,09	2,66	-0,57	-21,3
Córdoba	5,24	6,82	-1,58	-23,2
Buenos Aires	13,61	19,06	-5,45	-28,6
Santa Cruz	1,05	1,49	-0,44	-29,8
Mendoza	1,57	2,66	-1,09	-41,0
Río Negro	3,14	5,64	-2,50	-44,3
Tierra del Fuego	0,52	0,96	-0,43	-45,4
Santa Fe	2,62	6,28	-3,67	-58,3
CABA	2,09	5,22	-3,12	-59,9
San Luis	1,57	4,26	-2,69	-63,1
La Pampa	0,00	2,98	-2,98	-100,0

Entre todas las provincias, algunas presentan una sobrerrepresentación especialmente marcada dentro del grupo de riesgo en comparación con su peso en el total de la muestra. Por ejemplo, Corrientes (8,9 % en el grupo de riesgo frente a 4,7 % en la muestra total), Formosa (6,3 % vs. 2,6 %), Chaco y San Juan (ambas con 6,8 % vs. 3,3 %) duplican o incluso más su presencia relativa en este subconjunto.

En el extremo opuesto, varias provincias muestran una subrepresentación significativa en el grupo de riesgo, es decir, su proporción dentro del grupo es sustancialmente menor a la esperada por su peso en la muestra general. Entre ellas se destacan Buenos Aires (13,6 % vs. 19,1 %), Santa Fe (2,6 % vs. 6,3 %), CABA (2,1 % vs. 5,2 %) y La Pampa, que directamente no presenta casos con probabilidad mayor o igual a 0,8 en el modelo, pese a representar casi el 3 % de la muestra.

Luego del análisis territorial por provincia, se exploró la distribución del grupo de riesgo según el tamaño del aglomerado urbano de residencia, cuya variable en la ENFR está definida como `tamanio_aglomerado`. La Tabla 4.9 presenta la proporción de mujeres en cada categoría de tamaño de aglomerado dentro del grupo de riesgo y en la muestra total, expresadas en porcentaje, asi como la diferencia absoluta entre ambas proporciones y la diferencia relativa.

Tabla 4.9: Participación diferencial en grupo de mayor riesgo de no realización de mamografía según tamaño del aglomerado urbano.

Tamaño del aglomerado	**Grupo de riesgo (%)**	**Total muestra (%)**	**Dif. absoluta (pp)**	**Dif. relativa (%)**
Menos de 150.000 habitantes	65,45	60,38	+5,06	+8,4
150.001 a 500.000 habitantes	17,28	17,15	+0,13	+0,8
500.001 a 1.500.000 habitantes	8,38	10,12	–1,74	–17,2
Más de 1.500.000 habitantes	8,90	12,35	–3,45	–28,0

Se observa que el 65,5 % del grupo de riesgo reside en localidades con menos de 150.000 habitantes, mientras que ese estrato representa el 60,4 % de la muestra total, lo que da un 8,4 % de aumento. Los aglomerados intermedios (150.001 a 500.000 habitantes) presentan proporciones similares en ambos grupos. En cambio, los aglomerados de mayor tamaño muestran una menor representación en el grupo de riesgo respecto de su peso en la muestra: aquellos entre 500.001 y 1.500.000 habitantes disminuyen su porcentaje por el 17,2 % en el grupo de riesgo y los de más de 1.500.000 habitantes un 28 %.

Otra dimensión relevante para caracterizar al grupo de riesgo es la estructura familiar del hogar (Tabla 4.10). Los hogares no conyugales representan el 8,9 % del grupo de riesgo, frente al 5,8 % en la muestra total. Los hogares conyugales completos con otros miembros también presentan una mayor proporción en el grupo de riesgo, al igual que los hogares conyugales incompletos.

Tabla 4.10: Participación diferencial en grupo de mayor riesgo de no realización de mamografía según tipo de estructura familiar del hogar.

Tipo de hogar	Grupo de riesgo (%)	Total muestra (%)	Dif. absoluta (pp)	Dif. relativa (%)
Hogar no conyugal	8,90	5,75	+3,15	+54,8
Hogar conyugal completo con otros miembros	10,47	7,35	+3,12	+42,5
Hogar conyugal incompleto	25,65	18,32	+7,34	+40,1
Hogar unipersonal	30,37	31,63	–1,26	–4,0
Hogar conyugal completo (con o sin hijos)	24,61	36,95	–12,35	–33,4

Por su parte, los hogares unipersonales presentan proporciones similares en ambos grupos (30,4 % en el grupo de riesgo y 31,6 % en la muestra). En cambio, los hogares conyugales completos (con o sin hijos) muestran una menor presencia relativa en el grupo de riesgo (24,6 %) respecto del total de la muestra (37,0 %), lo que equivale a un 33,4 % menos.

La Tabla 4.11 presenta la distribución del grupo de riesgo, muestra y diferencia según nivel socioeconómico (NSE), basado en la variable construida `nse_empirico_2018` (ver sección Metodología). Los niveles más bajos muestran mayores proporciones en el grupo de riesgo en comparación con la muestra total: el 43,9 % pertenece al nivel D1 (vs. 36,9 %), y el 23,0 % al nivel D2E (vs. 12,8 %). En particular, D2E presenta una diferencia relativa del +80,3 %.

Tabla 4.11: Participación diferencial en grupo de mayor riesgo de no realización de mamografía según nivel socioeconómico (NSE).

NSE	Grupo de riesgo (%)	Total muestra (%)	Dif. absoluta (pp)	Dif. relativa (%)
D2E	23,04	12,78	+10,26	+80,3
D1	43,98	36,85	+7,13	+19,4
C3	26,70	28,86	–2,16	–7,5
C2	5,24	16,40	–11,16	–68,1
ABC1	1,05	5,11	–4,06	–79,5

En contraste, los niveles socioeconómicos medios y altos se encuentran subrepresentados. En particular, el estrato ABC1 concentra apenas el 1,1 % del grupo de riesgo, frente al 5,1 % de la muestra.

La Tabla 4.12 presenta las principales combinaciones multivariadas ordenadas por número de variables y diferencia relativa porcentual.

Tabla 4.12: Principales combinaciones de variables sobrerrepresentadas en el grupo de riesgo.

Nº var.	**Combinación de variables**	**Total (%)**	**Riesgo (%)**	**Dif. relativa (%)**
4	bhch05_2 + cobertura_salud_0 + nivel_instruccion_recod_1 + tamanio_aglomerado_4	3,0	8,9	+196,7
4	cobertura_salud_0 + nivel_instruccion_recod_1 + tamanio_aglomerado_4 + tipo_hogar_recod_2	2,9	7,9	+172,4
4	bhch05_5 + nivel_instruccion_recod_1 + nse_2018_empirico_D1 + tamanio_aglomerado_4	3,9	8,4	+115,4
5	cobertura_salud_0 + cobertura_salud_j_0 + condicion_actividad_3 + nse_2018_empirico_D2E + tamanio_aglomerado_4	1,1	2,6	+136,4
5	bhch05_5 + condicion_actividad_3 + nse_2018_empirico_D1 + tamanio_aglomerado_4 + tipo_hogar_recod_1	1,7	3,1	+82,4
5	bhch05_2 + condicion_actividad_3 + nivel_instruccion_recod_1 + nse_2018_empirico_D1 + tamanio_aglomerado_4	3,0	4,7	+56,7
6	cobertura_salud_0 + cobertura_salud_j_0 + condicion_actividad_3 + nivel_instruccion_recod_1 + nse_2018_empirico_D2E + tipo_hogar_recod_2	1,6	5,8	+262,5
6	aglomerado_9 + bhih03 + condicion_actividad_3 + nivel_instruccion_recod_1 + nse_2018_empirico_D1 + tipo_hogar_recod_1	1,2	3,1	+158,3
6	aglomerado_1 + cobertura_salud_0 + cobertura_salud_j_0 + condicion_actividad_3 + nivel_instruccion_recod_1 + tamanio_aglomerado_1	0,9	2,1	+133,3

Entre las combinaciones con mayor sobrerrepresentación se destacan aquellas que incluyen factores vinculados al nivel educativo, la falta de cobertura de salud (tanto del respondente como del jefe/a de hogar), el tamaño del aglomerado de residencia y la composición del hogar (que además tiene que ver con la situación conyugal de la respondente). Estas variables aparecen recurrentemente en los perfiles identificados, lo que sugiere que su presencia simultánea está asociada a una mayor probabilidad de no haberse realizado una mamografía. Por su parte, también hace su primera aparición la variable `bhih03`, que indica si la respondente percibió algún ingreso en dinero o en especie en los últimos 6 meses por Asignación Universal por Hijo, planes sociales u otras transferencias estatales.

5. DISCUSIONES Y CONCLUSIONES

Este trabajo tuvo como objetivo principal identificar los factores socioeconómicos a escala individual y hogar asociados con la probabilidad de mujeres de entre 50 y 69 años en Argentina de realizarse una mamografía al menos cada dos años, utilizando para ello herramientas de machine learning aplicadas a los datos de la Encuesta Nacional de Factores de Riesgo (ENFR) 2018. A su vez, se buscó identificar grupos de mujeres con menor acceso a esta práctica, con el fin de generar evidencia útil para orientar políticas públicas para la prevención del cáncer de mama.

5.1. Brechas en la Realización de Mamografías

Del análisis exploratorio descriptivo y del modelado predictivo se sugiere la presencia de brechas importantes en la cobertura de mamografías entre distintos subgrupos de la población. A nivel nacional, aproximadamente dos de cada tres mujeres de 50 a 69 años reportaron haberse realizado una mamografía de control en los últimos dos años, lo que deja a cerca de un tercio sin adherencia al tamizaje recomendado. Sin embargo, esta proporción nacional oculta disparidades significativas asociadas a factores socioeconómicos y regionales, que fueron observadas tanto en la exploración inicial como en los resultados de los modelos.

Como era esperable, las mujeres que habían realizado otras prácticas preventivas en el pasado, tales como el papanicolau, controles de presión arterial o glucemia, presentaron una mayor asociación con haber accedido a una mamografía en los últimos dos años. Esto sugiere la existencia de un patrón de comportamiento preventivo consistente, donde quienes están más integradas al sistema de salud y acostumbran a realizar controles periódicos tienden también a cumplir con el tamizaje mamográfico requerido [16]. Este resultado fue observado tanto en el análisis exploratorio como en los modelos predictivos, donde las variables relacionadas al uso previo del sistema sanitario contribuyeron de forma positiva a la predicción de realización de mamografías.

Se observó un claro gradiente de realización de mamografía según el nivel educativo del jefe o jefa del hogar: las mujeres provenientes de hogares con menor nivel de escolaridad —especialmente aquellos con primaria incompleta o solo primaria— presentaron los niveles más bajos de acceso a mamografías. Esta asociación, que los modelos predictivos confirmaron como una de las más relevantes, sugiere que la educación opera como indicador de otros recursos como el conocimiento sanitario y el nivel socioeconómico. Este patrón coincide con lo hallado por Lamfre y Hasdeu [15], quienes identificaron a la baja escolaridad como un fuerte predictor de menor uso del tamizaje, y con Nuche-Berenguer y Sakellariou [16], quienes señalaron que las mujeres con menor educación eran significativamente menos propensas a realizarse estudios preventivos. A su vez, la tenencia de cobertura médica (obra social o prepaga) fue otro determinante central: su ausencia se asoció sistemáticamente con menor acceso a mamografías, tanto en los modelos predictivos como en la literatura previa, que también la señala como una de las principales barreras estructurales para el tamizaje [16].

Más allá del nivel educativo del jefe o jefa, los indicadores de posición socioeconómica

familiar y las condiciones materiales del hogar —como el índice material, el nivel socioeconómico y el hacinamiento— mostraron asociación con la realización de mamografías. Las mujeres pertenecientes a hogares con mejores condiciones de vida presentaron niveles más altos de realización del estudio, mientras que aquellas en contextos de mayor vulnerabilidad socioeconómica mostraron menor probabilidad de realización. Estos hallazgos se enmarcan en el enfoque propuesto por la Comisión sobre Determinantes Sociales de la Salud de la Organización Mundial de la Salud, que sostiene que las condiciones sociales estructurales —como el entorno material, los ingresos o el acceso a servicios— influyen de manera directa en las oportunidades de alcanzar y sostener una buena salud [22]. En este caso, la capacidad de una mujer para realizarse controles preventivos no depende exclusivamente de su voluntad individual, sino que está parcialmente condicionada por los recursos de su entorno familiar y sus condiciones de vida. Las variables consideradas como *proxies* del nivel socioeconómico en los modelos predictivos contribuyeron de forma notable a explicar la variación de la variable respuesta, complementando el efecto ya observado de la educación y la cobertura médica.

Por otro lado, el análisis realizado reveló diferencias importantes en el acceso a mamografías según la provincia de residencia. Algunas jurisdicciones, especialmente del norte del país, como Santiago del Estero y Formosa, mostraron proporciones notablemente más bajas de realización del estudio, muy por debajo del promedio nacional. En cambio, otras como la Ciudad Autónoma de Buenos Aires (CABA), con alta concentración de infraestructura sanitaria, o La Pampa, con políticas activas de prevención —como mamografías sin turno ni orden médica y unidades móviles para zonas alejadas [34]— presentaron niveles de cobertura considerablemente más altos. Estas diferencias territoriales reflejan desigualdades estructurales en la distribución y disponibilidad de servicios de salud. Como también se observó en la tesis de Damiani [11], algunas provincias presentan un rezago persistente en la cobertura de mamografías que no puede explicarse únicamente por las características individuales de las mujeres. Al igual que en su estudio, donde se emplearon modelos eco-epidemiológicos para identificar regiones con patrones de bajo acceso y seguimiento de esta práctica preventiva, los resultados de esta tesis muestran que tanto la provincia de residencia como el *tamaño de aglomerado* muestran asociación con su realización. Las mujeres que residen en aglomerados de menor tamaño presentaron menor realización de mamografías, incluso controlando por otros factores. Desde un enfoque predictivo, estos hallazgos refuerzan la necesidad de diseñar políticas de prevención que contemplen activamente la dimensión geográfica y las diferencias estructurales entre contextos locales.

Además de los factores estructurales, se identificaron características relacionales y conductuales asociadas al acceso a mamografías. La situación conyugal mostró un patrón claro: estar casada se vinculó con mayor probabilidad de realización del estudio. Este patrón fue consistente con el tipo de hogar: los hogares no conyugales y los hogares conyugales incompletos presentaron niveles de cobertura inferiores, lo que sugeriría que el acompañamiento familiar puede funcionar como facilitador, ya sea por soporte emocional, informativo o logístico. En línea con estos hallazgos, Nuche-Berenguer y Sakellariou [16] también encontraron una asociación significativa entre estado civil y uso de servicios preventivos, destacando el peso de los vínculos familiares en las desigualdades de acceso.

También se observaron diferencias vinculadas a las condiciones materiales del entorno y a los estilos de autocuidado. Las mujeres con dificultades económicas para acceder a una alimentación saludable o que declararon sentirse mal de salud al momento de la encuesta

presentaron menor probabilidad de realizarse una mamografía, posiblemente por priorizar otros problemas o por percibir la prevención como secundaria. En contraste, prácticas como usar cinturón de seguridad, no fumar o hacer ejercicio regularmente se asociaron con mayor propensión al tamizaje, actuando como señales de estilos de vida orientados al cuidado personal. Estas dimensiones actitudinales, escasamente abordadas en trabajos previos, fueron incluidas aquí a través de indicadores concretos. Si bien el estudio de Nuche-Berenguer y Sakellariou [16] consideró la autopercepción de salud como variable de control, esta tesis se incluyeron diferentes variables que describen los comportamientos observables, lo cual permite describir patrones más amplios y detallados de cuidado. En conjunto, los hallazgos sugieren que la prevención depende tanto de factores estructurales como de disposiciones subjetivas hacia la salud.

En conjunto, los resultados muestran que la realización de la mamografía no puede entenderse como una decisión puramente individual, sino como una práctica situada, condicionada por factores estructurales —como la educación, la cobertura médica, los ingresos, el entorno territorial— y también por disposiciones subjetivas hacia el cuidado de la salud, expresadas en prácticas cotidianas y contextos relacionales. La convivencia, los vínculos familiares y los estilos de vida emergieron como elementos que facilitan o limitan la prevención, reforzando la idea del cuidado como fenómeno socialmente mediado. En este marco, la aplicación de técnicas de ciencia de datos constituyó un aporte central de esta tesis, al permitir estimar riesgos individuales, jerarquizar variables y detectar combinaciones de factores que incrementan la exclusión, superando los enfoques tradicionales y ofreciendo herramientas más precisas para orientar políticas públicas equitativas. Reconocer esta doble dimensión —estructural y actitudinal— resulta clave para diseñar estrategias preventivas más eficaces e inclusivas.

5.2. Condiciones Acumuladas de Riesgo

Además del análisis univariado, los resultados mostraron que muchas de estas condiciones se combinan entre sí y se refuerzan. No se trata de factores aislados, sino de situaciones que suelen combinarse: por ejemplo, las mujeres con bajo nivel educativo también tienden a no tener cobertura médica, vivir en hogares con menos recursos y estar fuera del mercado laboral, lo que resulta coherente con patrones estructurales de desigualdad observados en otros ámbitos sociales. Esta acumulación de desventajas —ya señalada también en trabajos como los de Lamfre y Hasdeu (2019) [15] y Nuche-Berenguer y Sakellariou (2021) [16]— genera perfiles persistentes de exclusión del sistema preventivo, donde múltiples barreras se potencian entre sí.

Este enfoque más integral permite entender mejor cómo se configuran las desigualdades en salud y por qué algunas mujeres quedan sistemáticamente por fuera de las prácticas preventivas. Reconocer la existencia de estos grupos de riesgo —no definidos por una única característica, sino por combinaciones de condiciones adversas— es fundamental para orientar políticas públicas más precisas. Identificar quiénes son, dónde están y qué barreras enfrentan es el primer paso para diseñar intervenciones más efectivas y equitativas en materia de prevención.

5.3. Limitaciones del Estudio

Si bien los hallazgos de esta investigación aportan evidencia relevante, es importante reconocer algunas limitaciones que deben ser consideradas al interpretar los resultados y que abren oportunidades para investigaciones futuras. En primer lugar, al basarse en datos de una encuesta transversal (ENFR 2018), no se permite establecer relaciones causales ni descartar la influencia de variables no observadas. Por ejemplo, la asociación entre educación y prácticas preventivas podría estar mediada por otros factores estructurales no captados.

En segundo lugar, si bien la ENFR ofrece un panorama amplio, no contempla todos los factores que podrían incidir en el comportamiento preventivo particular de las mamografías. Por ejemplo, no se relevan datos sobre conocimientos o percepciones sobre el cáncer de mama, antecedentes familiares, interacciones médico-paciente específicas o la disponibilidad local de mamógrafos, lo que limita el alcance del análisis. Además, al haber trabajado con información auto-reportada, algunas respuestas pueden estar afectadas por errores de recuerdo o sesgos de deseabilidad social, particularmente en preguntas sensibles como la realización o no de controles médicos.

Una tercera limitación tiene que ver con el universo poblacional relevado por la ENFR, que se restringe a la población urbana del país. Esta restricción podría excluir dinámicas específicas de zonas totalmente rurales, donde el acceso a prácticas preventivas podría estar condicionado por factores distintos. Sin embargo, según el Censo Nacional de 2022, aproximadamente el 92 % de la población argentina reside en áreas urbanas, por lo que los resultados mantienen un alto nivel de representatividad a nivel nacional [35].

Por último, los resultados reflejan la situación sanitaria del país en el año 2018, y podrían haber cambiado debido a factores externos como nuevas políticas públicas, campañas de sensibilización o el impacto de la pandemia de COVID-19 sobre los controles preventivos. Por esta razón, toda generalización a años posteriores debe realizarse con precaución, y se recomienda actualizar este análisis con fuentes más recientes, con el fin de evaluar la persistencia o evolución de las desigualdades observadas.

5.4. Recomendaciones para Futuras Investigaciones y Políticas Públicas

Los resultados de esta tesis permiten delinear algunas recomendaciones relevantes para el diseño de políticas públicas y futuras líneas de investigación en el campo. En primer lugar, es importante continuar monitoreando la evolución del acceso a mamografías a través de futuras ediciones de la ENFR y otras fuentes, integrando herramientas de ciencia de datos al sistema de vigilancia sanitaria. La combinación de encuestas nacionales con registros administrativos o datos de programas de cáncer permitiría un seguimiento más detallado y una detección temprana de cambios en las brechas de acceso.

Por otra parte, resulta clave complementar el análisis cuantitativo con estudios cualitativos centrados en los grupos de mujeres que no acceden al tamizaje. Entrevistas o grupos focales podrían revelar barreras culturales, percepciones de riesgo, miedos u obstáculos logísticos que los modelos no pueden captar. Esta dimensión permitiría diseñar intervenciones más sensibles al contexto social de las mujeres en situación de vulnerabilidad.

Por último, en cuanto a la acción estatal, los hallazgos sugieren priorizar la focalización de políticas preventivas en los grupos más rezagados: mujeres sin cobertura médica, de

bajo nivel educativo, fuera del mercado laboral y residentes en provincias del norte del país o en aglomerados pequeños. Algunas estrategias posibles pueden incluir operativos móviles de mamografía, como en la provincia de La Pampa [34], la búsqueda activa a través de promotoras de salud, recordatorios personalizados enfocados en los grupos de riesgo y articulación con la atención primaria para facilitar turnos y seguimiento.

En conclusión, esta tesis ha intentado aportar evidencia útil para comprender las desigualdades en el acceso a la prevención del cáncer de mama en Argentina. El enfoque basado en ciencia de datos, articulado con una mirada social, permite no solo describir el problema, sino orientar intervenciones concretas. Avanzar hacia una mayor equidad en salud requiere actuar tanto sobre las condiciones estructurales que limitan el acceso como sobre las prácticas culturales y los mecanismos institucionales que pueden facilitarlo.

Apéndice

.1. Otras Recodificaciones de Variables

En esta sección del apéndice se detallan las recodificaciones aplicadas a otras variables utilizadas en el análisis, pero que no fueron incluidas en el cuerpo principal del trabajo por no resultar centrales para la discusión.

.1.1. Fuente de agua del hogar

La variable original `bhcv08`, que indica el modo de acceso al agua en el hogar, fue recodificada en una variable binaria para simplificar su tratamiento en los modelos predictivos. La categoría 1 (agua proveniente de red pública) representa más del 95 % de los casos en la muestra, mientras que las demás categorías —perforación con bomba a motor o manual, aljibe, pozo u otras fuentes— presentan una frecuencia considerablemente menor. Si bien estas alternativas reflejan distintas condiciones de acceso, todas comparten la característica de no provenir de un sistema centralizado de provisión de agua. Por lo tanto, se unificaron en una única categoría que indica acceso limitado o no formal al recurso.

La recodificación adoptada fue la siguiente:

Tabla .1: Recategorización de la variable `bhcv08` (fuente de agua del hogar).

Nueva categoría	**Descripción**
1	Agua de red pública (agua corriente)
0	Agua de pozo, perforación, aljibe u otra fuente

.1.2. Cantidad de ambientes

La variable original `bhcv02` registra la cantidad total de ambientes del hogar, excluyendo baño, cocina, pasillos, lavadero y garage. Dado que se trata de una variable numérica discreta con un rango amplio, se procedió a su recategorización con el objetivo de facilitar su interpretación y mitigar posibles efectos espúrios por valores extremos con muy baja frecuencia.

Se agruparon los valores en tres categorías principales, considerando que las diferencias entre hogares con 7, 8 o más ambientes resultan marginales en términos habitacionales, mientras que la distinción entre viviendas pequeñas, medias y amplias sí refleja umbrales significativos desde una perspectiva social y material. Esta decisión busca capturar adecuadamente el tamaño relativo de la vivienda y su potencial asociación con condiciones de bienestar y espacio residencial disponible.

La recodificación utilizada fue la siguiente:

Tabla .2: Recategorización de la variable `bhcv02` (cantidad de ambientes en la vivienda).

Categoría	**Descripción**
1	1 o 2 ambientes (viviendas pequeñas)
2	3 ambientes (vivienda media típica)
3	4 o más ambientes (viviendas amplias)

.1.3. Relación de parentesco con el jefe/a de hogar

La variable original `bhch02` indica el vínculo de parentesco de cada integrante del hogar con el jefe o la jefa del mismo. Su codificación contempla diez categorías, que incluyen tanto relaciones familiares directas como vínculos más lejanos o no familiares:

- 1: Jefe/a de hogar
- 2: Cónyuge o pareja
- 3: Hijo/a o hijastro/a
- 4: Padre/madre
- 5: Hermano/a
- 6: Suegro/a
- 7: Yerno
- 8: Nieto/a
- 9: Otro familiar
- 10: Otro no familiar

Dado que muchas de estas categorías presentan frecuencias muy bajas y representan roles secundarios en la estructura del hogar, se procedió a una recategorización orientada a simplificar su uso en el análisis, lo que se puede observar en la Tabla .3.

Tabla .3: Recategorización de la variable `bhch02` (relación con el jefe/a de hogar).

Nueva categoría	Descripción	Códigos originales incluidos
1	Jefe/a de hogar	1
2	Cónyuge o pareja del jefe/a	2
3	Otros familiares o no familiares	3, 4, 5, 6, 7, 8, 9, 10

La nueva variable distingue tres grandes grupos: jefe/a de hogar, cónyuge o pareja, y otros familiares o no familiares. Esta decisión responde, por un lado, a la concentración de casos en las dos primeras categorías y, por otro, a la necesidad de reducir la fragmentación de clases para evitar problemas de sobreajuste y mejorar la estabilidad de los modelos predictivos. Además, permite conservar la distinción funcional más relevante dentro del hogar, vinculada a la toma de decisiones y al acceso a recursos.

.2. Intento Preliminar de Clasificación Socioeconómica Nominal

Como se mencionó en la sección Metodología, en una primera etapa de construcción de la variable indicadora de nivel socioeconómico se intentó aplicar directamente la segmentación de niveles socioeconómicos difundida por el consultor Guillermo Oliveto (Consultora W), publicada en el diario *La Nación* el 7 de abril de 2025 [36]. Esta clasificación retoma la estructura de cinco niveles propuesta por SAIMO y CEIM, pero la expresa en términos de ingresos netos mensuales del hogar, acompañados del porcentaje estimado de la población que pertenece a cada estrato.

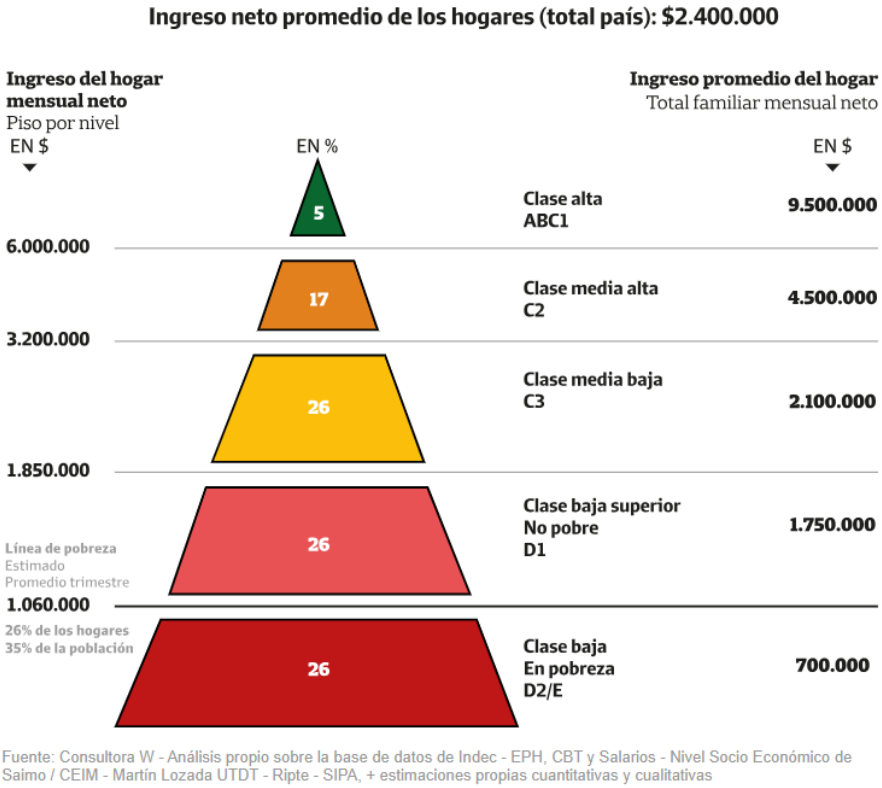


Figura .1: Pirámide de niveles socioeconómicos según ingresos mensuales del hogar (Consultora W, 2025). Fuente: *La Nación*, 07/04/2025.[36]

Para adaptar estos tramos de ingreso a los datos de la ENFR 2018, se intentó ajustar los valores nominales de 2025 utilizando la serie del Índice de Precios al Consumidor (IPC) del INDEC. Sin embargo, la aplicación de estos cortes a la variable de ingreso de la encuesta resultó insatisfactoria: más del 70 % de los hogares quedaron clasificados en la categoría más baja (D2E), lo que evidenció una desalineación entre la estructura actualizada y la distribución real de ingresos en la base de 2018. Por este motivo, se descartó esta opción como estrategia principal de segmentación.

Bibliografía

[1] Organización Mundial de la Salud. Cáncer de mama. Página web, 2024. Información oficial sobre el cáncer de mama publicada por la OMS.

[2] Y. S. Sun, Z. Zhao, Z. N. Yang, F. Xu, H. J. Lu, Z. Y. Zhu, W. Shi, J. Jiang, P. P. Yao, and H. P. Zhu. Risk factors and preventions of breast cancer. *International Journal of Biological Sciences*, 13(11):1387–1397, 2017. Artículo revisado por pares sobre factores de riesgo y prevención del cáncer de mama.

[3] International Agency for Research on Cancer. Cancer today – breast cancer fact sheet (globocan 2022). `https://gco.iarc.who.int/today/en/fact-sheets-cancers`, 2024. Accedido en marzo de 2025.

[4] International Agency for Research on Cancer. Breast cancer cases and deaths are projected to rise globally – press release no. 361. `https://gco.iarc.who.int/en`, 2025. Publicado el 24 de febrero de 2025. Accedido en marzo de 2025.

[5] United Nations Development Programme. Human development report 2023/24: Breaking the gridlock. `https://hdr.undp.org/`, 2024. Accedido en marzo de 2025.

[6] Organización Mundial de la Salud. Marco de aplicación de la iniciativa mundial contra el cáncer de mama, 2022. Accedido en marzo de 2025.

[7] Instituto Nacional del Cáncer. Mortalidad por cáncer de mama en mujeres. estadísticas 2022. `https://www.argentina.gob.ar/salud/instituto-nacional-del-cancer/estadisticas/mortalidad-cm`, 2023. Accedido en marzo de 2025.

[8] Ministerio de Salud de la República Argentina. Programa nacional de control de cáncer de mama. Página web, 2024. Información oficial sobre el Programa Nacional de Control de Cáncer de Mama en Argentina.

[9] Ministerio de Salud de la República Argentina. Tamizaje. Página web, 2024. Información oficial sobre tamizaje de cáncer de mama en Argentina.

[10] Instituto Nacional del Cáncer Ministerio de Salud de la Nación. El cáncer de mama en 8 palabras: una guía breve para atención primaria de la salud. *Ministerio de Salud de la Nación, Presidencia de la Nación*, 2015. Guía dirigida a equipos de atención primaria para la detección, diagnóstico y tratamiento del cáncer de mama.

[11] Magdalena Damiani Quiroz. Análisis eco-epidemiológico de las variaciones espacio-temporales en la realización de mamografías como práctica preventiva de salud en argentina y su relación con características del entorno. Tesis académica, 2024. Tesis sobre análisis eco-epidemiológico en la realización de mamografías en Argentina.

[12] R. Mottram, W. L. Knerr, D. Gallacher, H. Fraser, L. Al-Khudairy, A. Ayorinde, S. Williamson, C. Nduka, O. A. Uthman, S. Johnson, A. Tsertsvadze, C. Stinton, S. Taylor-Phillips, and A. Clarke. Factors associated with attendance at screening for

breast cancer: a systematic review and meta-analysis. *BMJ Open*, 11:e046660, 2021. Revisión sistemática y meta-análisis sobre asistencia al tamizaje de cáncer de mama.

[13] Instituto Nacional de Estadística y Censos de la República Argentina. Encuesta nacional de factores de riesgo. Página web, s.f. Base de datos sobre factores de riesgo en Argentina.

[14] Instituto Nacional de Estadística y Censos (INDEC). Sitio web del instituto nacional de estadística y censos (indec). Página web, 2025. Portal oficial con acceso a estadísticas y publicaciones sobre Argentina.

[15] Laura Lamfre and Santiago Hasdeu. Desigualdades sociales en salud y prácticas preventivas de cáncer en mujeres. *Cuadernos de Investigación. Serie Economía*, 8:68–96, 2019. Análisis del impacto de las desigualdades sociales en la prevención del cáncer en mujeres basado en la Encuesta Nacional de Factores de Riesgo 2018.

[16] Bernardo Nuche-Berenguer and Dikaios Sakellariou. Socioeconomic determinants of participation in cancer screening in argentina: A cross-sectional study. *Frontiers in Public Health*, 9, 2021. Estudio transversal sobre los determinantes socioeconómicos en la participación del cribado de cáncer en Argentina.

[17] Instituto Nacional de Estadística y Censos (INDEC) and Secretaría de Gobierno de Salud de la Nación. *4° Encuesta Nacional de Factores de Riesgo. Resultados definitivos*. INDEC y Secretaría de Gobierno de Salud, Buenos Aires, Argentina, 2018. Accedido en marzo de 2025.

[18] Ministerio de Salud de la Nación and Instituto Nacional de Estadística y Censos (INDEC). Encuesta nacional de factores de riesgo 2018 – nota técnica, 2019. Accedido en marzo de 2025.

[19] Harald Cramér. *Mathematical Methods of Statistics*. Princeton University Press, 1946.

[20] Instituto Nacional de Estadística y Censos (INDEC). Glosario de términos estadísticos. Página web, 2023. Definiciones oficiales de indicadores habitacionales y sociales.

[21] Instituto Nacional de Estadística y Censos (INDEC). Indicadores de condiciones de vida de los hogares. Página web, 2023. Informe técnico sobre indicadores habitacionales basados en la EPH.

[22] Commission on Social Determinants of Health. *Closing the gap in a generation: Health equity through action on the social determinants of health*. World Health Organization, Geneva, 2008. Informe emblemático de la OMS sobre determinantes sociales de la salud.

[23] Instituto Nacional de Estadística y Censos (INDEC). Censo nacional 2022. resultados preliminares sobre educación. Página web, 2022. Informe oficial sobre niveles educativos de la población argentina.

[24] Naciones Unidas. International standard classification of education (isced). Documento técnico, 2009. Clasificación internacional de niveles educativos utilizada por UNESCO y organismos internacionales.

[25] Instituto Nacional de Estadística y Censos (INDEC). Necesidades básicas insatisfechas (nbi). Página web, 2025. Sección del sitio oficial del INDEC dedicada a indicadores de necesidades básicas insatisfechas.

[26] Sociedad Argentina de Investigadores de Marketing y Opinión (SAIMO). Sitio oficial de saimo. Página web, 2024. Información institucional y metodológica sobre niveles socioeconómicos en Argentina.

[27] Oscar Muraro. Evolución del nivel socioeconómico en argentina. actualización a 2023. Presentación en PDF, 2024. Informe técnico elaborado a partir de datos de EPH, INDEC y el algoritmo SAIMO/CEIM. Disponible en: `https://saimo.org.ar`.

[28] Takuya Akiba, Shotaro Sano, Toshihiko Yanase, Takeru Ohta, and Masanori Koyama. Optuna: A next-generation hyperparameter optimization framework. *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, pages 2623–2631, 2019.

[29] Leo Breiman. Random forests. *Machine learning*, 45(1):5–32, 2001.

[30] Tianqi Chen and Carlos Guestrin. Xgboost: A scalable tree boosting system. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 785–794. ACM, 2016.

[31] Fabian Pedregosa, Gael Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, Jake Vanderplas, Alexandre Passos, David Cournapeau, Matthieu Brucher, Matthieu Perrot, and Edouard Duchesnay. Scikit-learn: Machine learning in python. `https://scikit-learn.org/stable/modules/generated/sklearn.ensemble.HistGradientBoostingClassifier.html`, 2011.

[32] Fabian Pedregosa, Gael Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, Jake Vanderplas, Alexandre Passos, David Cournapeau, Matthieu Brucher, Matthieu Perrot, and Edouard Duchesnay. Scikit-learn: Machine learning in python. *Journal of Machine Learning Research*, 12:2825–2830, 2011.

[33] David W Hosmer, Stanley Lemeshow, and Rodney X Sturdivant. *Applied Logistic Regression*. Wiley, Hoboken, NJ, 3rd edition, 2013.

[34] Ministerio de Salud de La Pampa. Mamografías en mujeres de 50 a 69 años sin turno ni orden médica, 2024. Consultado el 3 de junio de 2025.

[35] INDEC. Censo nacional de población, hogares y viviendas 2022. resultados provisionales. `https://www.indec.gob.ar/indec/web/Nivel4-Tema-2-41-135`, 2023. Consultado en junio de 2025.

[36] Guillermo Oliveto. Clase media: entre la mutación genética y la esperanza realista. `https://www.lanacion.com.ar/economia/clase-media-entre-la-mutacion-genetica-y-la-esperanza-realista-nid07042025/`, 2025. Entrevista publicada en La Nación que presenta la segmentación actualizada del NSE según la Consultora W.