Universidad de Buenos Aires

Facultad de Ciencias Exactas y Naturales

Ciencia de datos aplicada al estudio de las desigualdades socioeconómicas y territoriales en la realización de Papanicolaou en Argentina

Tesis de Licenciatura en Ciencias de Datos

Valentina Camila Durán

Directora: María Soledad Férnandez

Codirectora: Luciana Bruno

Facultad de Ciencias Exactas y Naturales de la UBA, 2025

CIENCIA DE DATOS APLICADA AL ESTUDIO DE LAS DESIGUALDADES SOCIOECONÓMICAS Y TERRITORIALES EN LA REALIZACIÓN DE PAPANICOLAOU EN ARGENTINA

El cáncer de cuello de útero es una enfermedad prevenible, pero aún causa alrededor de 2.200 muertes anuales en Argentina. La prueba de Papanicolaou (PAP) es uno de los métodos más efectivos para su detección temprana y forma parte de las estrategias recomendadas de prevención. Sin embargo, no todas las mujeres acceden por igual a esta práctica: su realización está asociada a factores socioeconómicos y territoriales, siendo significativamente menor en los grupos en situación de mayor vulnerabilidad socioeconómica.

El objetivo de esta tesis es identificar desigualdades en la realización del PAP en Argentina, caracterizando los grupos de mujeres con menor probabilidad de acceso. Para ello se utilizó la Encuesta Nacional de Factores de Riesgo (ENFR) 2018, aplicando análisis exploratorios, que incluyó la construcción de indicadores socioeconómicos y habitacionales, y distintos enfoques de modelado: regresión logística, regularización LASSO y algoritmos de Machine Learning (Random Forest y LightGBM).

Los resultados de los modelos fueron consistentes en cuanto a la identificación de variables relacionadas con la realización de PAPs: mujeres de hogares con mayores niveles de precariedad, sin cobertura de salud y con menor instrucción se relacionaron con menor realización de la práctica preventiva y por ende conforman un grupo con mayor riesgo. Por otro lado, la jurisdicción de residencia fue una variable relevante, mostrando la existencia de desigualdades territoriales en la realización de esta práctica. El desempeño de todos los modelos ajustados en cuanto a su capacidad de predecir nuevos datos fue similar.

Estos hallazgos refuerzan la evidencia sobre las desigualdades en salud y ofrecen insumos para el diseño de políticas públicas focalizadas, que promuevan el acceso equitativo a prácticas preventivas como lo es el PAP. Además, el análisis comparado de modelos sugiere que, en contextos como la salud pública, los enfoques explicativos —como la regresión logística— siguen siendo herramientas valiosas, tanto por su capacidad predictiva como por su mayor interpretabilidad para informar decisiones basadas en evidencia.

Palabras claves: Realización de Papanicolaou, desigualdades en salud, regresión logística, LASSO, Random Forest, LightGBM.

DATA SCIENCE MODELS APPLIED TO THE ANALYSIS OF SOCIOECONOMIC AND TERRITORIAL INEQUALITIES IN PAP TEST IN ARGENTINA

Cervical cancer is a preventable disease, yet it still causes around 2,200 deaths per year in Argentina. The Papanicolaou test (PAP) is one of the most effective methods for early detection and is part of the recommended prevention strategies. However, not all women have equal access to this practice: its uptake is associated with socioeconomic and territorial factors and is significantly lower among socioeconomically vulnerable groups.

The aim of this thesis is to identify inequalities in PAP uptake in Argentina by characterizing groups of women with a lower likelihood of access. To this end, data from the 2018 National Risk Factor Survey (ENFR) was used, applying exploratory analyses—including the construction of socioeconomic and housing indicators—and different modeling approaches: logistic regression, LASSO regularization, and machine learning algorithms (Random Forest and LightGBM).

The models consistently identified variables associated with lower PAP uptake: women living in more precarious households, without health coverage, and with lower education levels were less likely to undergo this preventive test, forming a higher-risk group. Additionally, province of residence proved to be a relevant factor, highlighting territorial disparities in PAP uptake. All fitted models showed similar performance in predicting new data.

These findings reinforce existing evidence on health inequalities and provide inputs for the design of targeted public policies that promote equitable access to preventive practices such as the PAP. Furthermore, the comparison of modeling approaches suggests that, in public health contexts, explanatory models—such as logistic regression—remain valuable tools due to both their predictive power and greater interpretability for evidence-based decision-making.

Keywords: Papanicolaou uptake, health inequalities, logistic regression, LASSO, Random Forest, LightGBM.

AGRADECIMIENTOS

Quiero agradecer a mi mamá, Claudia, a mi papá, Willy, y a mis hermanos, Santi y Manu, que me acompañaron en todo este proceso, desde la elección de la carrera hasta cada examen. Incluso, me ayudaron con sus propios conocimientos en los momentos de estudio, siendo un pilar fundamental en todo mi recorrido. Gracias por estar siempre a mi lado, con su apoyo y amor incondicional.

A mis amigas de la secundaria, que me acompañaron y apoyaron incluso cuando tuve que quedarme días enteros estudiando, deseándome suerte y fuerzas en cada examen. Gracias por su paciencia y amor durante todos estos años.

A todos mis amigos de la facultad. Sin ellos, la carrera no hubiera sido fácil de transitar. Gracias por las cursadas compartidas, los mates, las risas y también los llantos. En especial gracias a Luz y Sol, que me acompañaron desde la primera materia, sin ellas, cada cursada, cada estudio para los exámenes y cada trabajo práctico no hubieran sido lo mismo. Me motivaron a seguir cada día. Son amistades que empezaron en esta carrera y que me llevo para siempre.

A mis directoras, Sole y Luciana, por su guía y todo el tiempo que le dedicaron. Gracias por ayudarme a repensar mis ideas y darles forma, por sus sugerencias y por acompañarme con tanta dedicación y compromiso en este proceso.

A la Facultad de Ciencias Exactas y Naturales de la Universidad de Buenos Aires, que fue mucho más que el lugar donde cursé la carrera, fue una segunda casa para mí. Donde aprendí, me reí, lloré y crecí personalmente. Me enorgullece haberme formado en esta facultad que prioriza la formación científica, fomentando la educación pública y la participación de las mujeres en la ciencia, que es un poco de lo que trata esta tesis.

Y gracias a todas las personas que, de una forma u otra, me acompañaron y alentaron en este camino. Gracias a toda esta ayuda y amor, hoy pude llegar hasta acá.

A mi familia y amigos.

Índice general

1. INTRODUCCIÓN

En Argentina, las políticas públicas de salud promueven la realización periódica de prácticas preventivas como herramienta fundamental para la detección temprana y el control de enfermedades. Entre ellas, el estudio de Papanicolaou (PAP) se destaca como el principal método para prevenir el cáncer de cuello de útero, una enfermedad causada principalmente por el Virus del Papiloma Humano (VPH), que afecta a las células del cuello uterino. Aunque es altamente prevenible, sigue siendo el tercer cáncer más frecuente en mujeres y la quinta causa de muerte por cáncer en el país, con más de 4.600 casos nuevos y aproximadamente 2.200 muertes anuales [1, 2].

El PAP es una prueba sencilla y efectiva que permite detectar lesiones precancerosas en estadios iniciales [3]. Según las recomendaciones del Programa Nacional de Prevención del Cáncer Cervicouterino (PNPCC), todas las personas con cuello uterino de entre 25 y 64 años deben realizarse este estudio al menos cada tres años, luego de dos resultados normales consecutivos con un año de diferencia [1]. Esta práctica se encuentra disponible de forma gratuita en los centros de salud públicos de todo el país [4].

A pesar de su importancia, el acceso al PAP no es equitativo. Numerosos estudios han documentado desigualdades en la realización de esta práctica, asociadas a factores socioeconómicos y territoriales: las mujeres en situación de vulnerabilidad —como aquellas con menor nivel educativo, sin cobertura de salud o con bajos ingresos— presentan menores tasas de realización del PAP [5, 6, 7, 8]. También se observan disparidades geográficas: la provincia y el aglomerado urbano de residencia se asocian con las probabilidades de realizarse este control preventivo [5, 6, 7]. Además de estas dimensiones estructurales, factores individuales y conductuales como la edad, la percepción de la propia salud, el consumo de tabaco o la actividad física también se relacionan con la decisión o posibilidad de realizarse el PAP [5, 6].

La Encuesta Nacional de Factores de Riesgo (ENFR) [9] es una herramienta pública, representativa y anonimizada que releva datos sobre factores de riesgo (como el consumo de tabaco, alcohol, alimentación y actividad física), procesos de atención en el sistema de salud y datos relacionados con enfermedades no transmisibles (hipertensión, diabetes, colesterol elevado y otras). Consta de cuatro ediciones realizadas en los años 2005, 2009, 2013 y 2018, y se constituye como una herramienta clave para describir la situación sanitaria y social de la población adulta en Argentina. La ENFR es un relevamiento de gran importancia para el sistema de salud argentino, ya que forma parte de la estrategia nacional para la vigilancia y control de las enfermedades no transmisibles (ENT). Es llevada a cabo de manera conjunta por el Ministerio de Salud de la Nación y el Instituto Nacional de Estadística y Censos (INDEC) [10]. En este trabajo se utilizó la última edición (2018), que brinda información válida y confiable para analizar la relación entre factores socioeconómicos, demográficos y de salud. En particular, esta base de datos posibilita el análisis de las desigualdades en la realización de PAPs y su relación con múltiples factores a diferentes escalas: individuales, del hogar y territoriales.

El objetivo central de esta tesis es identificar desigualdades socioeconómicas y terri-

toriales en la realización del PAP en Argentina y caracterizar los grupos de mujeres con menor probabilidad de acceso a este examen preventivo. Para ello, se implementan métodos estadísticos explicativos como la regresión logística, técnicas de regularización (LASSO) y modelos predictivos basados en Machine Learning (árboles de decisión, random forest y LGBM, entre otros). Esta integración metodológica busca no solo describir patrones y desigualdades, sino también aportar insumos para la formulación de políticas públicas orientadas a reducir las brechas en la prevención del cáncer de cuello de útero. La articulación con herramientas de Ciencia de Datos y el acceso a información actualizada son claves para identificar áreas prioritarias y monitorear intervenciones de salud pública [11].

Existen trabajos previos que abordan el estudio de la realización de PAP en Argentina a partir de los datos de la ENFR [5, 6, 7], pero en esta tesis se propone la incorporación de herramientas de Machine Learning en el proceso de identificación de desigualdades en la realización de PAP, que es un enfoque no abordado previamente. La articulación con herramientas de Ciencia de Datos y el acceso a información actualizada son claves para identificar áreas prioritarias y monitorear intervenciones de salud pública con enfoque territorial [11].

2. OBJETIVOS

El objetivo general de esta tesis es identificar desigualdades socioeconómicas y territoriales en la realización del estudio de Papanicolaou (PAP) en Argentina. Se busca analizar cómo distintos factores, que abarcan dimensiones individuales, del hogar y del territorio, se relacionan con la práctica de este examen preventivo. La intención final es aportar evidencia empírica que permita caracterizar los grupos de mujeres con menor acceso al PAP, con el propósito de contribuir al diseño de políticas de salud más equitativas y focalizadas, que puedan fortalecer la prevención del cáncer de cuello de útero y mejorar la equidad en el acceso a los controles de salud.

2.1. Objetivos específicos

1. **Realizar un análisis exploratorio y limpieza de datos**, con el fin de comprender la calidad y características de la base de datos de la Encuesta Nacional de Factores de Riesgo (ENFR, 2018).

2. **Construcción de variables socioeconómicas**, integrando información de diversas variables —como educación, cobertura de salud, situación laboral e ingresos del hogar— para construir indicadores más robustos y consistentes.

3. **Identificar características socioeconómicas y territoriales** relacionadas con la realización del PAP en Argentina, mediante el ajuste de modelos (regresión logística, LASSO, random forest y LGBM).

4. **Identificar variables que permitan describir los grupos de mujeres donde se requiere mayor promoción de la práctica del PAP**, contribuyendo a focalizar estrategias de salud pública y a desarrollar intervenciones más efectivas en la prevención del cáncer cervicouterino.

3. METODOLOGÍA

La metodología empleada en esta tesis combina técnicas de análisis exploratorio, construcción de variables y modelos explicativos y predictivos.

3.1. Fuente de datos y variables

La fuente principal de información de esta tesis es la Encuesta Nacional de Factores de Riesgo (ENFR), [9], en su cuarta edición, correspondiente al año 2018.

La edición 2018 de la ENFR es relevante por ser la más reciente disponible y por la riqueza de la información relevada. Se realizó durante el último trimestre de dicho año en localidades urbanas de 5.000 habitantes o más, logrando más de 29.000 entrevistas completas.

El cuestionario de la ENFR se estructura en dos bloques principales:

- **El Bloque Hogar**, que releva información sobre las características de la vivienda, las condiciones del hogar y la situación socioeconómica de sus miembros. Según la ENFR 2018, se considera hogar a la persona o grupo de personas, parientes o no, que habitan bajo el mismo techo y comparten los gastos de alimentación [12]. Este bloque incluye datos sobre ingresos del hogar, cobertura de salud y situación laboral del jefe o jefa de hogar.
- **Bloque Individual** se aplica a una persona adulta seleccionada aleatoriamente dentro del hogar y aborda dimensiones clave como máximo nivel educativo alcanzado, salud general, hábitos de consumo, actividad física, alimentación y prácticas preventivas. Es en este bloque donde se incluye la pregunta sobre la realización del estudio de Papanicolaou (PAP) en los últimos dos años, variable que constituye el foco de esta investigación.

La distribución geográfica de los puntos de muestreo de la ENFR 2018 se presenta en la Figura 3.1, la cual ilustra la cobertura de este relevamiento en localidades urbanas de todo el país. Los datos y la documentación de la ENFR son de acceso público y se pueden consultar en el portal del INDEC, [9].

3.1.1. Definición de la muestra

La muestra utilizada en este trabajo se extrajo a partir de la base completa de la ENFR 2018, que cuenta con 29.224 casos. Para el análisis de la realización del estudio de Papanicolaou (PAP), se consideraron exclusivamente los datos de mujeres de 25 años o más.

La muestra analítica incluye 14.805 mujeres de 25 años o más, lo que representa aproximadamente el 50 % de la población total relevada por la encuesta (ver Figura 3.2). La

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)

decisión de este recorte etario responde a las recomendaciones generales del Ministerio de Salud para la realización de esta práctica preventiva [4].

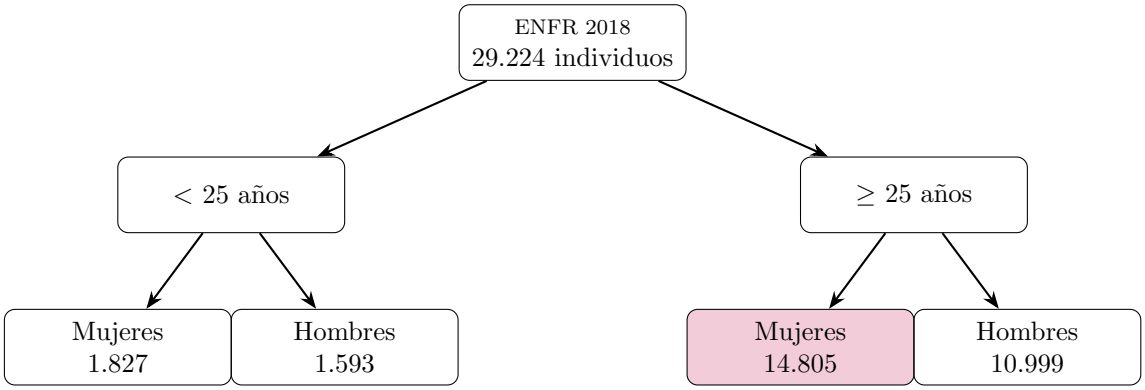

Figura 3.2: Diagrama de flujo de los datos de la ENFR 2018 que conforman la muestra analitica, se destaca en rojo la conformada por mujeres mayores a 25 años utilizada en este estudio.

3.2. Análisis exploratorio

En esta sección se exploró tanto la variable de realización del estudio de Papanicolaou (PAP) bajo análisis (ver Sección 3.2.1), como las potenciales variables predictoras, que describen características individuales, del hogar y del entorno de las encuestadas (ver Sección 3.2.2). Estas variables buscan capturar distintas dimensiones del nivel socioeconómico y territorial.

Se realizó primero una revisión detallada de los valores faltantes y registros inconsistentes tanto para la variable respuesta como para las potenciales predictoras, así como la identificación de categorías con baja variabilidad en relación a la realización de PAP o gran desbalance. En caso de ser necesario, se agruparon o recodificaron categorías para asegurar un análisis más robusto.

La ENFR 2018 posee **variables categóricas** (como el tipo de vivienda, la región geográfica o la cobertura de salud) y algunas **variables continuas** (por ejemplo, la edad, la cantidad de personas en el hogar o los ingresos del hogar). Para las variables categóricas, se examinaron las frecuencias absolutas y relativas de cada categoría, así como su distribución en la muestra analítica. En el caso de las variables continuas, se describieron algunas de sus medidas estadísticas como la media, la mediana, la desviación estándar y los percentiles. Además, se construyeron histogramas y diagramas de densidad para detectar posibles asimetrías y la presencia de valores atípicos.

A lo largo de esta etapa, se exploraron tanto variables a escala de la persona encuestada (edad, condición de actividad, nivel educativo, etc.) como sociodemográficas a escala del hogar (características de la vivienda, hacinamiento, tipo de hogar) y variables relacionadas con hábitos de salud de la persona (percepción de salud general, prácticas preventivas de salud, realización de actividad física, entre otras). En cada caso, se describió cada posible variable predictora en relación a la realización del PAP en los últimos dos años. Además, se llevaron a cabo comparaciones entre diferentes variables de regionalización para detectar desigualdades geográficas.

Estos análisis exploratorios permitieron no solo tener un panorama detallado de la muestra analítica y de las variables relevantes, sino también identificar patrones preliminares de desigualdad y posibles factores de riesgo que serán retomados y profundizados en las siguientes secciones mediante modelos explicativos y predictivos.

3.2.1. Variable respuesta

La variable respuesta bajo análisis es `control_pap`, es una variable dicotómica informada en la ENFR que da cuenta de si la mujer se realizó una prueba de Papanicolau en los últimos dos años: toma el valor 1 si la mujer realizó el PAP en los últimos dos años, 2 si no lo realizó, y 99 en caso de respuesta Ns/Nc. Esta variable, si bien es ofrecida como tal en los datos disponibles por la ENFR, fue construida a partir de dos preguntas de la ENFR 2018:

bipp03 "*¿Alguna vez se hizo un Papanicolaou?*"
Opciones: 1. Sí;
2. No;
99. Ns/Nc.

bipp04 "*¿Cuándo fue la última vez que se hizo un Papanicolaou?*"
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.

Durante el análisis exploratorio, se observó en ambas variables que existían valores faltantes y categorías "No sabe/No contesta" (Ns/Nc). En `bipp03` (¿Alguna vez se hizo un Papanicolaou?), 1.822 mujeres respondieron que nunca se hicieron un PAP, y 68 mujeres respondieron Ns/Nc. Estos casos se correspondieron en la codificación a los valores faltantes (NAs) en `bipp04`, ya que esta última pregunta sólo es respondida por quienes declararon haberse hecho un PAP alguna vez.

La elección de `control_pap` como variable de respuesta se debe a que, según el Ministerio de Salud de la Nación [4], el PAP debe hacerse al menos una vez al año y, si durante dos años consecutivos los resultados son negativos, el intervalo puede extenderse. Por ello, se definió una ventana de dos años en esta tesis, considerando que refleja tanto la recomendación general de control anual como la flexibilidad que existe para espaciar los controles, permitiendo capturar la práctica real de la población.

3.2.2. Variables predictoras

A partir del total de 287 variables relevadas por la ENFR, se seleccionó un conjunto específico de 36 variables para el análisis de desigualdades en la realización del Papanicolaou (PAP). Estas variables representan dimensiones socioeconómicas y territoriales —a nivel individual, del hogar y del entorno— que son el núcleo de esta tesis.

En la Tabla 3.1, se presenta un resumen de estas variables seleccionadas que conforman la base para los modelos explicativos y predictivos que se desarrollarán en las siguientes secciones. Se preservan los nombres de las variables tal cual figuran en la encuesta para facilitar la reproducibilidad del estudio.

Tabla 3.1: Principales variables predictoras seleccionadas

Categoría	Variable	Descripción
Territoriales	cod_provincia	Código de provincia de residencia
	region	Región geográfica
	tamanio_aglomerado	Tamaño del aglomerado urbano
	aglomerado	Aglomerado urbano específico
	localidades_150	Localidad con $\geq$150.000 hab.
Características del hogar	bhcv01	Tipo de vivienda
	bhcv02	Número de ambientes
	bhcv03	Material del piso
	bhcv04	Material del techo
	bhcv05	Presencia de cielorraso
	bhcv06	Medio principal para cocinar
	bhcv07	Disponibilidad de agua por cañería
	bhcv08	Fuente de obtención de agua
	bhcv09	Presencia de baño o letrina
	bhcv10	Tipo de inodoro o letrina
	bhcv11	Tipo de desagüe del inodoro
	bhho01	Uso exclusivo del baño
	bhho02	Número de ambientes exclusivos del hogar
	cant_componentes	Número total de miembros del hogar
	tipo_hogar	Tipo de hogar (estructura familiar)
	quintil_uc	Quintil de ingresos
	bhih03	Obtuvo AUH u otro plan social
	bhch03_j	Sexo del jefe/a de hogar
	rango_edad_j	Rango de edad del jefe/a de hogar
	bhch05_j	Estado civil del jefe/a
	nivel_instruccion_j	Nivel educativo del jefe/a
	cobertura_salud_j	Cobertura de salud del jefe/a
	bhsl06	Horas semanales trabajadas
	condicion_actividad_j	Situación laboral del jefe/a
Individual (encuestada)	bhch02	Parentesco con el jefe/a
	rango_edad	Rango de edad
	bhch05	Estado civil
	nivel_instruccion	Nivel educativo
	cobertura_salud	Cobertura de salud
	bisl06	Horas semanales trabajadas
	condicion_actividad	Situación laboral

3.3. Recodificación y construcción de variables

Recodificación de variables

En esta etapa del trabajo se revisaron las variables socioeconómicas y territoriales para mejorar su utilidad en los modelos explicativos y predictivos. Se identificaron aquellas variables con un elevado número de categorías, categorías con frecuencias muy bajas o que presentaban redundancias conceptuales. Para resolverlo, se agruparon o recodificaron estas variables en categorías más sintéticas, facilitando su interpretación y reduciendo la complejidad de los modelos.

El criterio general fue preservar la relevancia conceptual de cada variable —por ejemplo, manteniendo diferencias importantes como el nivel educativo o el tipo de vivienda— pero consolidando categorías minoritarias o poco representadas que podrían dificultar el análisis. Así, se recodificaron variables como el `tipo de vivienda`, diferenciando entre casas, departamentos y una categoría "otros" que agrupa casillas, piezas de inquilinato y demás. Además, las variables `material del piso` y `material del techo` se reagruparon en niveles de calidad alta, media o baja.

Se simplificaron las categorías de `nivel de instrucción` tanto para el/la jefe/a de hogar como para el respondente individual, estableciendo tres grandes grupos: bajo nivel educativo (hasta secundario incompleto), nivel medio (hasta secundario completo o terciario incompleto) y nivel alto (terciario, universitario completo o superior). También se recodificaron las variables sobre la estructura del hogar (`tipo de hogar`), distinguiendo entre hogares unipersonales, conyugales y no conyugales, para capturar diferencias relevantes en la convivencia y organización familiar.

Sin embargo, no se descartaron las variables originales para el estudio. Se compararon las variables en su codificación original con las variables recodificadas en relación a la relevancia de las mismas en los modelos.

Construcción de variables

Además de las recodificaciones realizadas, se desarrollaron diversos índices para capturar de manera más integral las condiciones socioeconómicas y habitacionales de los hogares. Estos índices no estaban directamente disponibles en la base de datos, por lo que fue necesario construirlos a partir de variables de la ENFR 2018 siguiendo criterios metodológicos consolidados.

Índice de hacinamiento

Se calculó como el cociente entre la cantidad total de personas en el hogar y la cantidad de ambientes exclusivos, es decir, aquellos utilizados únicamente por las personas del hogar, excluyendo baño, cocina, pasillos, lavadero y garage, según los criterios definidos por el INDEC[13].

$$\text{Índice de hacinamiento} = \frac{\text{Cantidad de personas en el hogar}}{\text{Cantidad de ambientes exclusivos}}$$

A partir de este índice, se construyó una variable categórica con cuatro niveles para capturar distintos grados de ocupación del espacio: hogar holgado, sin hacinamiento, hacinamiento moderado y hacinamiento crítico. El umbral para identificar *hacinamiento crítico* se tomó de la definición oficial del INDEC [13], que considera que un hogar se encuentra en esta situación cuando hay más de tres personas por ambiente. Las demás categorías fueron definidas en este trabajo para reflejar matices adicionales en las condiciones habitacionales.

Tabla 3.2: Clasificación del índice de hacinamiento adaptada de los criterios oficiales del INDEC.

Categoría	**Índice de hacinamiento (personas por ambiente)**	**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.

Índice de Necesidades Básicas Insatisfechas (NBI)

El **NBI** es una herramienta oficial desarrollada por el INDEC para identificar hogares en situación de vulnerabilidad estructural. Según el INDEC [14], un hogar presenta **NBI** si tiene al menos una de las siguientes carencias:

- NBI 1 — Vivienda inadecuada: Son los hogares que viven en habitaciones de inquilinato, hotel o pensión, viviendas no destinadas a fines habitacionales, viviendas precarias y otro tipo de vivienda. Se excluye a las viviendas tipo casa, departamento y rancho.
- NBI 2 — Condiciones sanitarias: Hogares que no cuentan con baño o letrina.
- NBI 3 — Hacinamiento crítico: Hogares con más de tres personas por cada ambiente exclusivo disponible.
- NBI 4 — Asistencia escolar incompleta: Hogares con al menos un niño o niña de entre 6 y 12 años que no asiste a la escuela.
- NBI 5 — Capacidad de subsistencia reducida: Hogares donde hay cuatro o más personas por miembro ocupado y el jefe o jefa de hogar no completó el tercer grado de la primaria.

En este trabajo, se lograron reconstruir las tres primeras características a partir de las variables disponibles en la **ENFR 2018**, respetando la metodología oficial del INDEC.

NBI 1: Vivienda — Se identificaron como viviendas inadecuadas los hogares que residen en habitaciones de inquilinato, hoteles/pensiones, viviendas no destinadas a habitación u "otros" (categorías 4, 5, 6, 7 de `bhcv01`).

NBI 2: Condiciones sanitarias — Hogares que declararon no tener baño ni letrina, a partir de la variable `bhcv09`.

NBI 3: Hacinamiento crítico — Se clasificó como crítico si se superan las 3 personas por ambiente.

Sin embargo, dado que la ENFR no incluye datos sobre asistencia escolar ni sobre la ocupación de todos los miembros del hogar, se construyó un *proxy* alternativo que captura situaciones similares a las características 4 y 5:

Proxy de NBI 4 y 5 — Hogares cuyo jefe no completó la primaria o no tiene instrucción (`nivel_instruccion_j` en categorías 1 y 2). Esto refleja trayectorias de desventaja estructural vinculadas a menor escolarización de los hijos y a menores ingresos familiares [5, 6].

Construccion del indicador de NBI — Un hogar fue clasificado como **con NBI** si presentaba al menos una de las situaciones de NBI 1. 2, 3 o proxy de 4 y 5 descritas anteriormente.

Índice de Nivel Socioeconómico (NSE)

El **NSE** es un indicador ampliamente utilizado para segmentar la estructura social de los hogares [15]. Dado que la ENFR no permite calcular el algoritmo completo oficial (que considera educación, ocupación y tenencia de bienes), se empleó el ingreso mensual del hogar como aproximación.

A partir de la distribución de ingresos y siguiendo proporciones reportadas por la Sociedad Argentina de Investigadores de Marketing y Opinión (SAIMO) [16] en el año 2018, se asignaron los siguientes estratos:

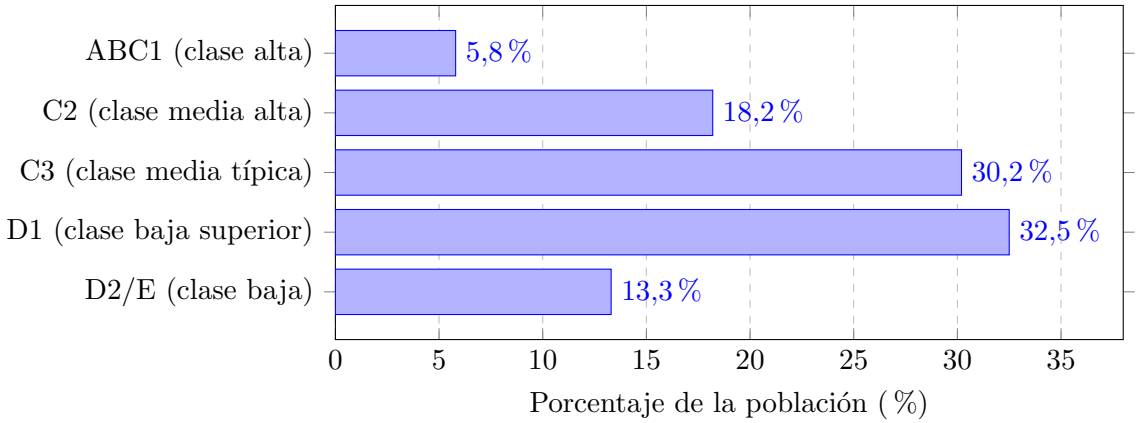

Figura 3.3: Distribución proporcional de la población por nivel socioeconómico según SAIMO (2018).

De esta manera, se particionaron los registros de ingresos del hogar de la ENFR 2018 para que la distribución de niveles socioeconómicos refleje las proporciones publicadas por SAIMO en 2018. Este procedimiento permitió asignar a cada hogar de la encuesta una clase socioeconómica (ABC1, C2, C3, D1 o D2/E), reproduciendo la estructura de ingresos de la población argentina en términos relativos.

Índice material

Se construyó un **índice material** para sintetizar las condiciones estructurales de las viviendas a partir de varias variables del módulo *Características de la Vivienda* de la ENFR 2018. Se incluyeron las siguientes dimensiones: material del piso (`bhcv03`), material del techo (`bhcv04`), presencia de cielorraso (`bhcv05`), tipo de combustible para cocinar (`bhcv06`), ubicación del acceso al agua (`bhcv07`), fuente de provisión de agua (`bhcv08`), tipo de inodoro (`bhcv10`) y destino del desagüe del inodoro (`bhcv11`).

Para combinar estas variables cualitativas se utilizó la técnica de *Análisis de Correspondencias Múltiples* (MCA) [17], implementada con la librería `prince` de Python [18]. El MCA es una técnica estadística que resume la información de varias variables categóricas en componentes numéricos.

En este caso, fue interpretado como un **gradiente de calidad material**: valores más altos del índice indican viviendas con mayor precariedad (materiales de baja calidad o servicios básicos deficientes), mientras que valores más bajos corresponden a viviendas con mejores características constructivas y de acceso a infraestructura.

Por otro lado, se construyó una versión categórica *Índice material terciles* dividiendo los valores del índice material en terciles mediante la función `qcut` dentro de la librería `Pandas` de Python [19]:

- **Alto:** viviendas con mejores condiciones materiales.
- **Medio:** viviendas con condiciones materiales intermedias.
- **Bajo:** viviendas con mayor precariedad material.

En su conjunto, los índices construidos permiten complementar la descripción socioeconómica de los hogares, capturando no sólo el nivel de ingresos, sino también aspectos habitacionales y estructurales que han sido descritas en la literatura en la capacidad de acceder a prácticas preventivas de salud, como el estudio de PAP [7, 6, 5].

3.4. Modelos

En esta sección se presentaron los tres enfoques de modelado empleados para abordar la problemática de la no realización del Papanicolaou (PAP). En primer lugar, se utilizó la **regresión logística** con selección de variables mediante la técnica de stepwise como modelo explicativo principal, lo que permitió identificar y cuantificar la magnitud del efecto de los factores de riesgo asociados a la menor realización del PAP. En segundo lugar, se ajustó un **modelo de regularización LASSO** como técnica de selección automática de variables, evaluando su capacidad predictiva y la importancia relativa de cada variable. Por último, se implementaron **modelos de aprendizaje automático supervisado** para comparar la capacidad predictiva y explorar la relevancia de los distintos predictores.

3.4.1. Regresión logística

La construcción de la regresión logística inició con un **pre-procesamiento de las variables** para ser incluidos en los modelos que consistió con la limpieza y codificación de las variables del conjunto de datos. Las variables categóricas con valores faltantes se completaron con la categoría "Ns/Nc" para evitar la pérdida de registros. Las variables continuas se completaron con su mediana. Además, se homogenizó la codificación de variables binarias a valores 0/1, y las categóricas se convirtieron al tipo `category` en Python [19] para un tratamiento consistente durante el modelado. En particular, la variable `control_pap`, que indica la realización del PAP en los últimos dos años, se codificó como 1 para "no se realizó" y 0 para "sí se realizó".

Selección inicial de variables

La selección de variables predictoras comenzó con el cálculo del coeficiente de asociación de **Cramér's V** [20] entre cada variable categórica y la respuesta binaria `control_pap`. Este coeficiente mide la intensidad de la asociación entre variables categóricas (0 indica ausencia de asociación, mientras que 1 representa asociación perfecta), permitiendo identificar predictores con mayor relación con la variable objetivo.

A partir de este análisis, se realizó un **primer filtro manual**, priorizando la selección de:

- Las **versiones recodificadas** de variables que en sus versiones originales presentaban desbalence o interpretabilidad limitada.
- Evitar el **solapamiento conceptual** entre variables, eligiendo una única representante a partir del resultado del coeficiente de Cramér's V.

Control de multicolinealidad

Dado que la estimación de los coeficientes en la regresión logística se ve afectada por la presencia de colinealidad entre predictoras, se evaluó la multicolinealidad mediante el cálculo de **Factores de Inflación de la Varianza (VIF)**. Para ello, se generaron variables dummy —es decir, variables binarias que representan cada categoría de una variable cualitativa— utilizando `drop_first=True` para evitar redundancias. Luego, se calcularon los VIF y se eliminaron aquellas variables con valores elevados que podían afectar la estabilidad del modelo, priorizando aquellas con mayor relevancia para el análisis. Se considera que un valor de VIF menor a 10 es aceptable para no presentar multicolinealidad preocupante en el modelo [21].

Modelo múltiple

Se ajustó un modelo múltiple considerando todas las predictoras seleccionadas luego del proceso inicial (univariado) y del control de la multicolinealidad. Para evaluar la

contribución de cada predictora en el modelo múltiple, se aplicó un **Test de Verosimilitud** [22]. Este test compara la log-verosimilitud de un **modelo completo** (con todas las variables) con la de modelos **reducidos** (donde se elimina una variable por vez). Si la disminución de la log-verosimilitud al excluir una variable es estadísticamente significativa —es decir, si el **p-valor** asociado al test es menor a 0,05—, se concluye que dicha variable aporta información relevante a la realización del PAP. El estadístico de verosimilitud (LR statistic) cuantifica esta diferencia: valores grandes indican que la variable eliminada contribuye significativamente al ajuste, ya que el modelo completo mejora notablemente frente al reducido.

Performance del modelo seleccionado

La muestra se dividió en un conjunto de entrenamiento (80 %) y otro de evaluación (20 %), utilizando la función `train_test_split` de la librería `sklearn` [23] con estratificación según la respuesta. Esto aseguró la misma proporción de clases en ambos conjuntos, controlando el desbalanceo de la variable objetivo.

Con la selección de variables predictoras a incorporar en el modelo múltiple final (regresión logística) se realizó un ajuste del modelo con la función `smf.logit` de `statsmodels` [24], incorporando las variables categóricas mediante la notación `C()` en la fórmula. Este modelo permitió estimar los **odds ratios (OR)** para cada predictor, junto con sus respectivos intervalos de confianza (IC) del 95 %.

Los **odds ratios (OR)** expresan la razón entre las probabilidades (razón de odds) de que ocurra un evento (en nuestro caso definimos *no realizarse el PAP*) en una categoría de una variable, comparadas con la categoría de referencia [21]:

$$\text{OR} = \frac{\text{odds de no realizarse el PAP en el grupo}}{\text{odds de no realizarse el PAP en el grupo de referencia}}$$

Los **odds**, a su vez, se definen como el cociente entre la probabilidad de que ocurra el evento y la de que no ocurra [22]:

$$\text{odds} = \frac{P(\text{No realizarse el PAP})}{P(\text{Realizarse el PAP})}$$

Un **OR mayor a 1** indica que el grupo tiene mayor probabilidad de no realizarse el control en comparación con el grupo de referencia. Por el contrario, un **OR menor a 1** sugiere menor probabilidad. Los **intervalos de confianza** del 95 % (IC 95 %) permiten evaluar la precisión de esta estimación: si el IC no incluye el valor 1, se considera que la asociación observada es estadísticamente significativa.

Adicionalmente se evaluó el desempeño del modelo en el conjunto de prueba para asegurarse de que conserve cierto poder predictivo. Para ello, se utilizó la métrica del **Área Bajo la Curva ROC (AUROC)** [25], que mide la capacidad del modelo para discriminar correctamente entre mujeres que no se realizaron el PAP y aquellas que sí lo hicieron. El AUROC tiene la ventaja de considerar todos los posibles umbrales de decisión, por lo que un valor cercano a 1 indica excelente capacidad de discriminación, mientras que un valor cercano a 0.5 refleja un desempeño no mejor que el azar.

Además, para las jurisdicciones, se calcularon las **probabilidades promedio** predichas por el modelo de no realización del PAP. Estas predicciones se obtuvieron directamente mediante la función de Python `predict` [23] a partir de los coeficientes del modelo final. En este caso, se promedió la probabilidad predicha para cada jurisdicción, considerando la heterogeneidad real de las demás características de la población en el conjunto de datos. Así, se obtuvo un valor promedio de probabilidad para cada jurisdicción, que permitió realizar una comparación relativa entre ellas considerando la variabilidad de las demás características.

3.4.2. Modelo de regularización (LASSO)

Como estrategia de modelado y selección de variables alternativa se empleó un modelo de regresión logística penalizada mediante **LASSO (Least Absolute Shrinkage and Selection Operator)** [26].

LASSO es una técnica de regularización que modifica la función de costo de la regresión logística agregando una penalización proporcional a la suma de los valores absolutos de los coeficientes:

$$\text{RSS} + \lambda \sum_{j=1}^{p} |\beta_j|$$

donde:

- **RSS** es la suma de los errores al cuadrado.
- λ es un hiperparámetro que controla la fuerza de la penalización.
- $\sum |\beta_j|$ penaliza los valores absolutos de los coeficientes.

Este término de penalización fuerza a que muchos coeficientes se reduzcan exactamente a cero cuando λ es suficientemente grande, logrando de forma automática la **selección de variables**. A diferencia del enfoque empleado en la estrategia de modelado de la regresión logística descrito en la sección anterior —que elimina variables en función de la significancia estadística de sus coeficientes—, LASSO considera el aporte global de cada variable al error total, descartando aquellas que no resultan relevantes en la predicción. La estrategia adoptada fue mantener todas las variables en el modelo para que LASSO identificara cuáles realmente contribuyen, sin necesidad de evitar incluir variables correlacionadas en el modelo: la penalización controla el efecto de las variables correlacionadas.

Para entrenar el modelo LASSO, se realizaron los siguientes pasos de **preprocesamiento de las variables**:

1. **Limpieza e imputación**: los valores faltantes se completaron con la categoría "Ns/Nc" en categoricas y -1 en continuas, una categoría que LASSO puede utilizar como señal informativa [25].
2. **División en conjunto de entrenamiento y prueba**: Se separó el conjunto en entrenamiento (80 %) y prueba (20 %) mediante `train_test_split` con estratificación por la respuesta.

3. **Codificación de variables**: dado que LASSO no trabaja con variables categóricas directamente, se convirtieron todas las variables categóricas en variables dummy (*one-hot encoding*, es decir, variables binarias que indican la pertenencia a cada categoría).

4. **Estandarización**: se aplicó la transformación `StandardScaler` [27] únicamente a las variables continuas, para garantizar que todas las variables numéricas tuvieran media cero y varianza uno. Se hizo después de la división de datos para no filtrar información del conjunto de prueba en el conjunto de entrenamiento, un paso fundamental para evitar el *data leakage* [25].

Para el **entrenamiento del modelo LASSO** se utilizó la clase `LogisticRegression` de `scikit-learn` con penalización L1 (`penalty='l1'`) [23].

Para la **evaluación del modelo** ajustado según las variables cuyos coeficientes no fueron reducidos a cero, se evaluó el desempeño del modelo LASSO en el conjunto de prueba mediante la métrica **AUROC**. Esto permitió comprobar la capacidad discriminativa y la robustez del modelo ajustado, y compararla con la regresión logística.

3.4.3. Modelos predictivos

Como paso complementario a los modelos construidos con fines explicativos, se implementaron distintos modelos de **Machine Learning supervisado** para evaluar la capacidad predictiva de los modelos sobre nuevos datos y comparar la importancia de las variables identificadas por los diferentes modelos ajustados.

Se inició con la limpieza y transformación de las variables, utilizando el mismo **preprocesamiento** descrito en la Sección `3.4.1`, que incluyó la imputación de valores faltantes, la división de los datos en entrenamiento y prueba, la conversión de las variables categóricas en variables dummy (es decir, binarias), ya que los modelos de aprendizaje automático utilizados no pueden procesar directamente datos categóricos, y la estandarización de las variables continuas, dado que algunos modelos (como SVM o KNN) son sensibles a las escalas de las variables.

Entrenamiento y comparación de modelos

Se ajustaron los siguientes algoritmos [25]:

- **Árbol de decisión**
- **K-Nearest Neighbors (KNN)**
- **Support Vector Machine (SVM)**
- **Naive Bayes**
- **Random Forest**
- **LightGBM (LGBM)**

Para cada modelo, se implementó una búsqueda aleatoria de hiperparámetros ("RandomizedSearchCV" [28]) con validación cruzada estratificada de 5 pliegues.

Para los modelos con mejor desempeño —ver sección 4.2.3- se analizaron las **importancias de las variables**. En estos modelos, la importancia de cada variable se calcula como la cantidad de veces que la variable se utiliza para dividir los nodos en los árboles de decisión y el grado en que esa división mejora la pureza o la reducción de la impureza (en términos de ganancia de información o disminución de la entropía, dependiendo del modelo) [25].

Este proceso permitió identificar las variables que más contribuyen a la predicción de la no realización del PAP. Luego, se construyeron **modelos reducidos** entrenados solo con estas variables más importantes, seleccionadas de acuerdo con la **importancia acumulada hasta el 80 % del total**. Esta estrategia buscó mantener un equilibrio entre la interpretabilidad y la capacidad predictiva de los modelos, reduciendo el número de variables sin perder gran parte de la información relevante.

Por último, se evaluó la capacidad predictiva de los modelos ajustados y reducidos en el conjunto de prueba mediante AUROC.

4. RESULTADOS

4.1. Descripción de la muestra

En la Figura 4.1 se detalla la distribución porcentual del control de PAP por cada categoría de las variables consideradas, junto con el número total de observaciones para cada grupo.

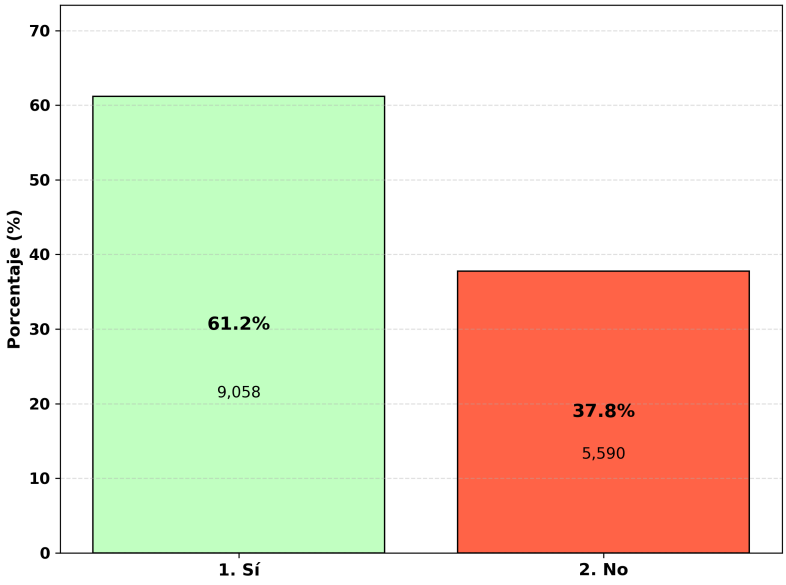

Figura 4.1: Distribución de mujeres de 25 años o más según la realización de un PAP (`control_pap`) en los últimos dos años. Se excluyen los registron con Ns/Nc (157 registros).

Los resultados de la distribución porcentual de la realización o no del PAP para cada variable predictora utilizada, discriminada por catagoría se presentan en la Tabla 4.1. Por motivos de extensión y claridad, se decidió no mostrar en detalle todas las variables categóricas usadas, sino un subconjunto representativo que incluye las variables recodificadas, las que describen al individuo y aquellas construidas o recodificadas que sintetizan la información sobre las características del hogar.

Tabla 4.1: Porcentajes de realización del PAP según variables seleccionadas. El color celeste indica las categorías correspondientes al nivel `Territorial`, el verde al nivel `Hogar` y el rosa al `Individual`

Categoría	**control_pap**		**Total**
	NO	**SI**	
region (1 - Metropolitana)	29.2	70.8	1807

Categoría	control_pap		Total
	NO	SI	
region (2 - Pampeana)	34.3	65.7	4753
region (3 - Noroeste)	46.3	53.7	2661
region (4 - Noreste)	43.8	56.2	2074
region (5 - Cuyo)	43.0	57.0	1320
region (6 - Patagónica)	35.8	64.2	2033
tamanio_aglomerado (1 - >1.500.000)	29.2	70.8	1807
tamanio_aglomerado (2 - entre 500.001 y 1.500.000)	36.1	63.9	1484
tamanio_aglomerado (3 - entre 150.001 y 500.000)	37.9	62.1	2856
tamanio_aglomerado (4 - <150.000)	40.5	59.5	8501
bhcv01_recod (1 - Casa)	39.6	60.4	12491
bhcv01_recod (2 - Departamento)	27.9	72.1	1952
bhcv01_recod (3 - Otros)	47.3	52.7	205
bhho01 (0 - Sin baño exclusivo del hogar)	48.4	51.6	153
bhho01 (1 - Baño exclusivo del hogar)	38.0	62.0	14446
bhho01 (Ns/Nc - No sabe/no contesta)	51.0	49.0	49
nivel_hacinamiento (Adecuado - Sin hacinamiento)	35.8	64.2	6658
nivel_hacinamiento (Crítico)	47.2	52.8	197
nivel_hacinamiento (Moderado)	47.9	52.1	526
nivel_hacinamiento (Holgado)	39.4	60.6	7267
tipo_hogar_recod (1 - Unipersonal)	44.7	55.3	3493
tipo_hogar_recod (2 - Conyugal completo)	31.0	69.0	6346
tipo_hogar_recod (3 - Conyugal completo con otros miembros)	43.3	56.7	827
tipo_hogar_recod (4 - Conyugal incompleto)	40.1	59.9	3327
tipo_hogar_recod (5 - No conyugal)	55.9	44.1	655
quintil_uc (1 - Más bajo)	47.7	52.3	2827
quintil_uc (2 - Bajo)	45.4	54.6	3040
quintil_uc (3 - Medio)	38.5	61.5	2876
quintil_uc (4 - Alto)	34.0	66.0	3034
quintil_uc (5 - Más alto)	25.2	74.8	2871
bhih03 (0 - No recibió por AUH)	37.2	62.8	11557
bhih03 (1 - Recibió por AUH)	42.3	57.7	2936
bhih03 (Ns/Nc)	29.7	70.3	155
nbi (0 - Sin necesidades básicas insatisfechas)	38.0	62.0	14303
nbi (1 - Con necesidades básicas insatisfechas)	46.4	53.6	345
nse_2018_empirico (ABC1 - Alto)	22.1	77.9	774
nse_2018_empirico (C2 - Medio alto)	25.9	74.1	2602
nse_2018_empirico (C3 - Medio)	34.9	65.1	4417
nse_2018_empirico (D1 - Bajo)	46.0	54.0	4912
nse_2018_empirico (D2E - Muy bajo)	48.8	51.2	1943
indice_material_terciles (Alto)	33.3	66.7	6583

Categoría	control_pap		Total
	NO	SI	
indice_material_terciles (Medio)	37.2	62.8	3561
indice_material_terciles (Bajo)	46.0	54.0	4504
rango_edad (2 - 25 a 34 años)	31.4	68.6	3190
rango_edad (3 - 35 a 44 años)	28.4	71.6	4505
rango_edad (4 - 45 a 54 años)	35.8	64.2	3563
rango_edad (5 - 55 años o más)	59.9	40.1	3390
cobertura_salud (0 - Pública)	43.9	56.1	3802
cobertura_salud (1 - Privada)	36.1	63.9	10846
nivel_instruccion_recod (1.0 - Hasta secundario incompleto)	51.0	49.0	6217
nivel_instruccion_recod (2.0 - Secundaria completa o universitario incompleto)	32.9	67.1	5032
nivel_instruccion_recod (3.0 - Universitaria)	22.3	77.7	3384
nivel_instruccion_recod (Ns/Nc - No sabe/no contesta)	60.0	40.0	15
condicion_actividad (1 - Ocupado)	29.3	70.7	7813
condicion_actividad (2 - Desocupado)	34.7	65.3	590
condicion_actividad (3 - Inactivo)	49.5	50.5	6245

4.2. Modelos

En esta sección se presentaron los resultados obtenidos para los tres enfoques de modelado aplicados: la **regresión logística**, el **modelo LASSO** y los **modelos predictivos de aprendizaje automático**.

4.2.1. Regresión Logística

La Figura 4.2 muestra los resultados del valor del índice de Cramér's V, con las variables ordenadas de mayor asociación a menor asociación con la variable dependiente.

A partir del primer proceso de seleccion en base a Cramér's V obtuvimos los siguientes resultados:

- **Variables recodificadas vs. originales:** Se priorizaron las versiones recodificadas por cuestiones de balance y significado, excepto en `tipo_hogar`, donde la versión original tenía mayor sentido interpretativo.
- **Variables conceptualmente similares:** Se seleccionó una de ellas en base al coeficiente de Cramér's V:
  - Entre `nse` y `quintil`, se mantuvo `nse`.
  - Entre `cod_provincia`, `aglomerado`, `region`, `tamano_aglomerado` y `localidades_150`, se eligió `cod_provincia` como proxy geográfico.

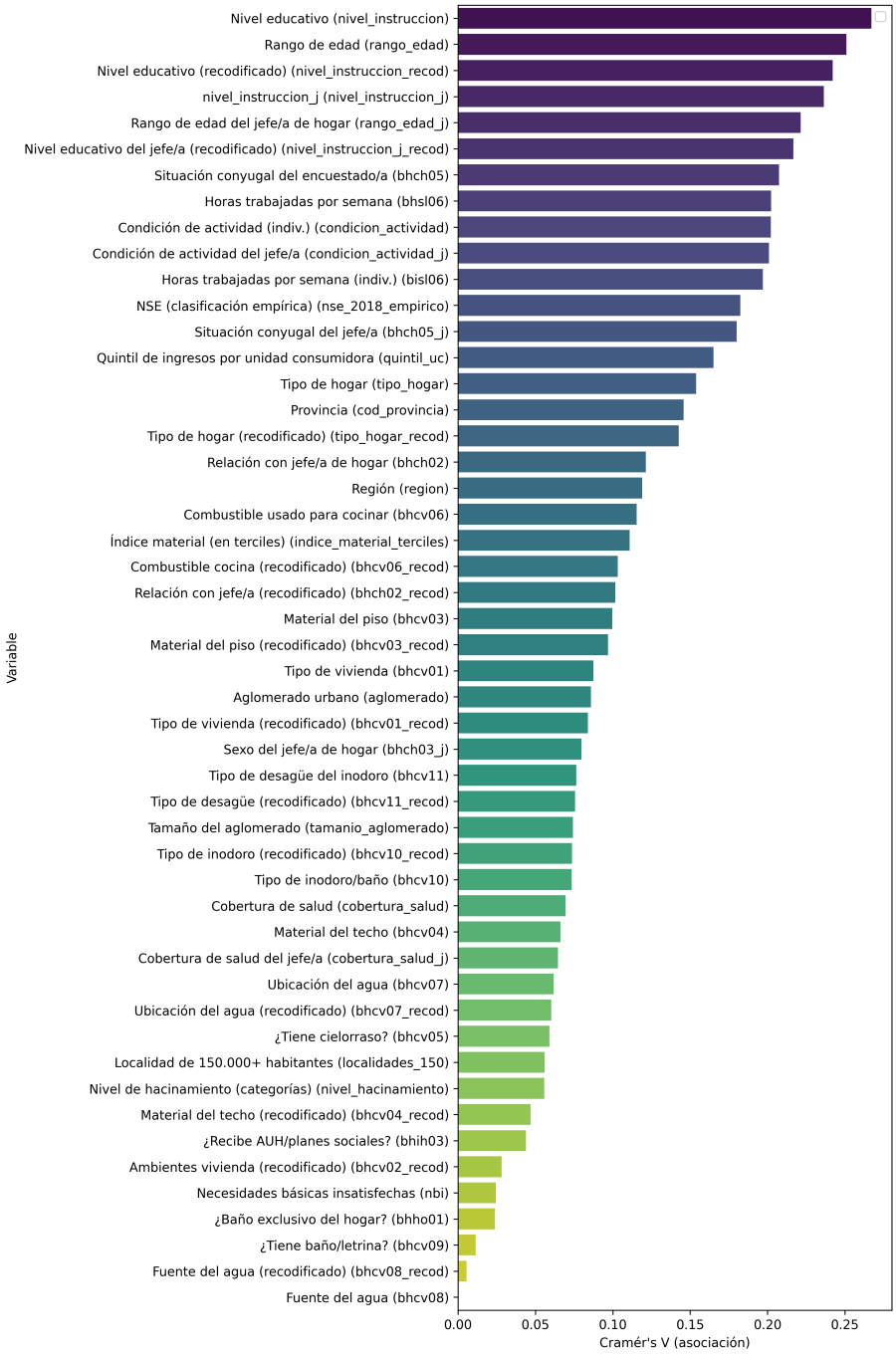

Figura 4.2: Coeficiente de Cramér's V entre variables categóricas y la variable bajo análisis (`control_pap`).

- Como el `índice material` mostró una mayor asociación en Cramér's V que los indicadores individuales que lo componen, se priorizó el índice y se eliminaron las variables componentes (y sus versiones recodificadas).

En base al cálculo del *Variance Inflation Factor* (VIF) se encontró que algunos VIF eran extremadamente elevados. Estos resultados se explicaron por la alta colinealidad con las variables relacionadas a la jefatura del hogar, ya que casi el 60 % de las mujeres en la muestra son jefas de hogar. Por ello, se decidió descartar todas las variables relacionadas a la jefatura.

Luego, se identificó que persistían valores elevados de VIF entre grupos de variables que representaban constructos similares. En particular, se evaluaron las siguientes alternativas:

- **Índice material**: se comparó la versión continua y la categorizada en terciles.
- **Hacinamiento**: se evaluó la variable continua frente a la version categóricas (nivel de hacinamiento) y las variables originales (cantidad de componentes y cantidad de personas en el hogar).

El criterio general para la selección fue priorizar el p-valor en los modelos multivariados, eligiendo aquellas variables con efectos significativos en la relación con la realización del PAP. Por esto, se mantuvo la variable continua de `hacinamiento`. Sin embargo, en el caso del `índice material`, se priorizó la versión categorizada en terciles por su mayor valor comunicativo e interpretativo, aun cuando la versión continua también resultara significativa.

Con estas decisiones, se recalculó el VIF de las variables en un modelo reducido, encontrándose valores aceptables (es decir, valores menores que 10 [21]).

Los resultados de la prueba de verosimilitud se resumen a continuación:

Variables con p-valor < 0.05:

- cod_provincia, bhch05, indice_material_terciles, nivel_instruccion_recod, rango_edad, cobertura_salud, condicion_actividad, hacinamiento, tipo_hogar, nse_2018_empirico, bhcv01_recod.

Variables con p-valor ≥ 0.5:

- nbi, bhcv02_recod.

De este modo, se definió el conjunto final de variables incluidas en el modelo de regresión logística (ver Tabla 4.2).

Tabla 4.2: Variables finales incluidas en el modelo explicativo ordenadas de menor a mayor según el p-valor. El color celeste indica las categorías correspondientes al nivel `Territorial`, el verde al nivel `Hogar` y el rosa al `Individual`.

Variable	**Descripción**	**p-valor**
cod_provincia	Provincia de residencia (proxy geográfico)	0.0000
bhch05	Situación conyugal de la mujer	0.0000
indice_material_terciles	Índice material de la vivienda (en terciles)	0.0000
nivel_instruccion_recod	Nivel educativo alcanzado (recodificado)	0.0000
rango_edad	Rango etario de la mujer	0.0000
cobertura_salud	Presencia o no de cobertura de salud	0.0000
condicion_actividad	Condición de actividad (ocupación)	0.0002
hacinamiento	Hacinamiento en la vivienda	0.0003
tipo_hogar	Tipo de hogar (estructura familiar)	0.0004
nse_2018_empirico	Nivel socioeconómico (clasificación empírica)	0.0010
bhcv01_recod	Tipo de vivienda (recodificado)	0.0112

En el modelo final, los VIF tomaron valores menores a 7.

La Figura 4.3 presenta los *odds ratios* e intervalos de confianza del modelo de regresión logística ajustado final (no se incluyen los OR de las provincias).

Para las provincias, se muestran los valores predichos por el modelo (probabilidades ajustadas, Figura 4.4).

En cuanto a la evaluación del desempeño del modelo, el área bajo la curva ROC (AUROC) fue de 0.72. Este valor confirma que el modelo tiene una capacidad adecuada para discriminar entre mujeres que realizan y no realizan el control de PAP, ofreciendo un punto de partida sólido para la interpretación de los resultados y las conclusiones.

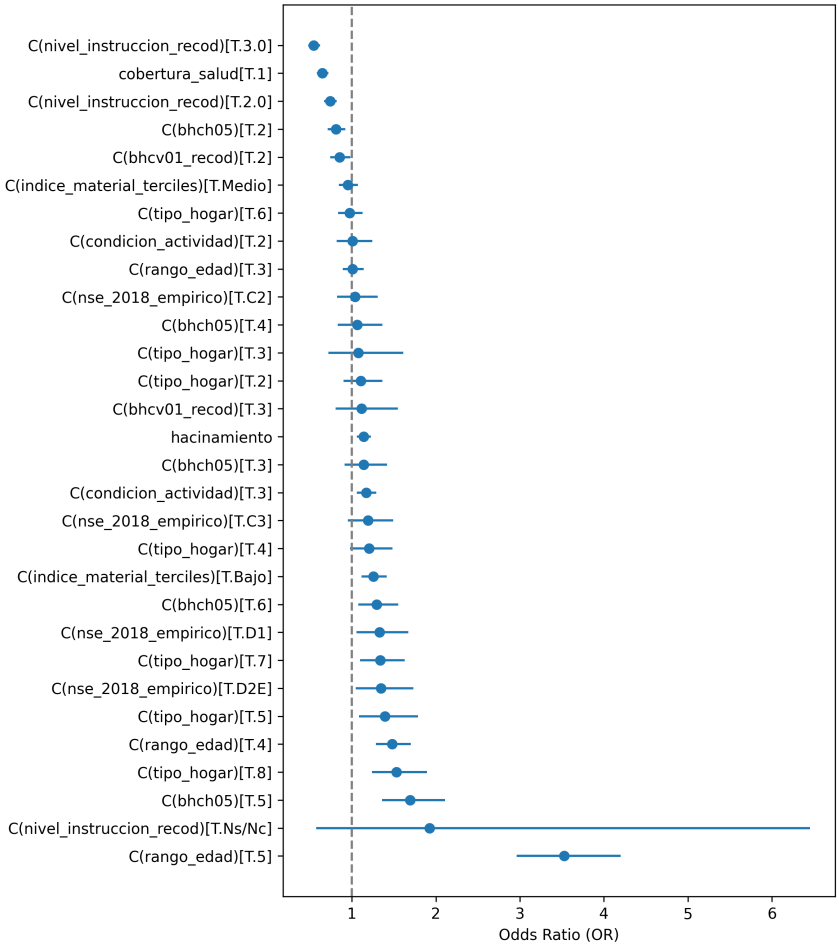

Figura 4.3: Forest plot del modelo final (sin provincias). Se muestran los valores de OR (círculos) y sus respectivos intervalos de confianza al 95 % (líneas horizontales). La línea punteada indica el valor de referencia OR = 1, correspondiente a la ausencia de efecto.

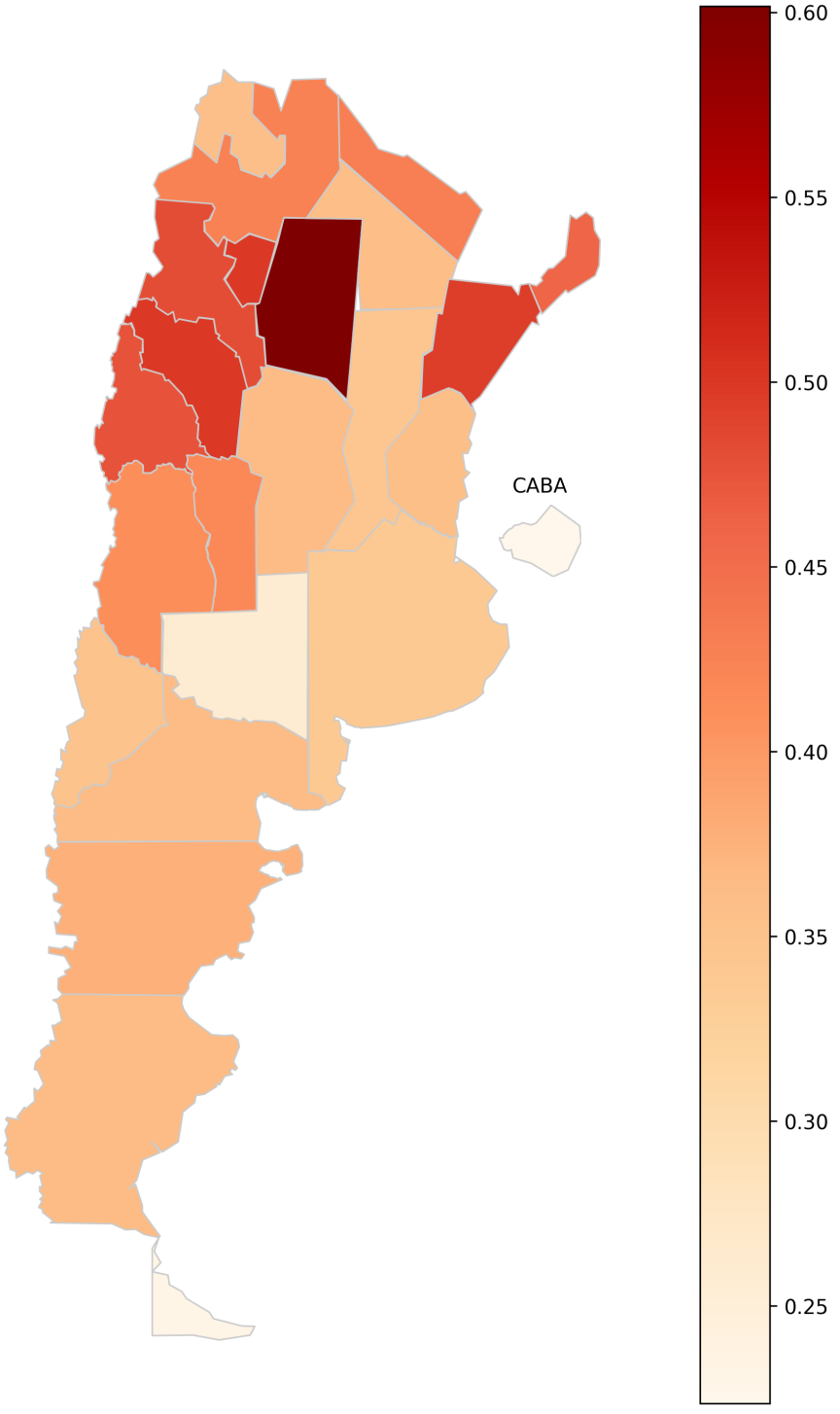

Figura 4.4: Valores ajustados predichos por el modelo de la **no** realización de PAP por provincia.

4.2.2. Modelo de regularización (LASSO)

En la Figura 4.5 se presentan los *odds ratios* (OR) estimados por el modelo LASSO para las variables seleccionadas (sin incluir las provincias).

Para mejorar la legibilidad del gráfico, se excluyeron las variables cuyos OR se encuentran cercanos a 1 (entre 0.9 y 1.1). Esto permite destacar las asociaciones más relevantes y facilita la comparación entre las variables.

En cuanto a los valores predichos por el modelo para las provincias, se verificó que las probabilidades promedio de no realización del PAP son similares a las obtenidas en el modelo explicativo, por lo que no se vuelve a incluir el mapa de provincias en esta sección.

Finalmente, se evaluó el rendimiento de este modelo Lasso en el conjunto de test, obteniéndose un área bajo la curva ROC (AUROC) de 0.72. Esto confirma que el modelo Lasso logra una capacidad de discriminación comparable al modelo de regresión logística.

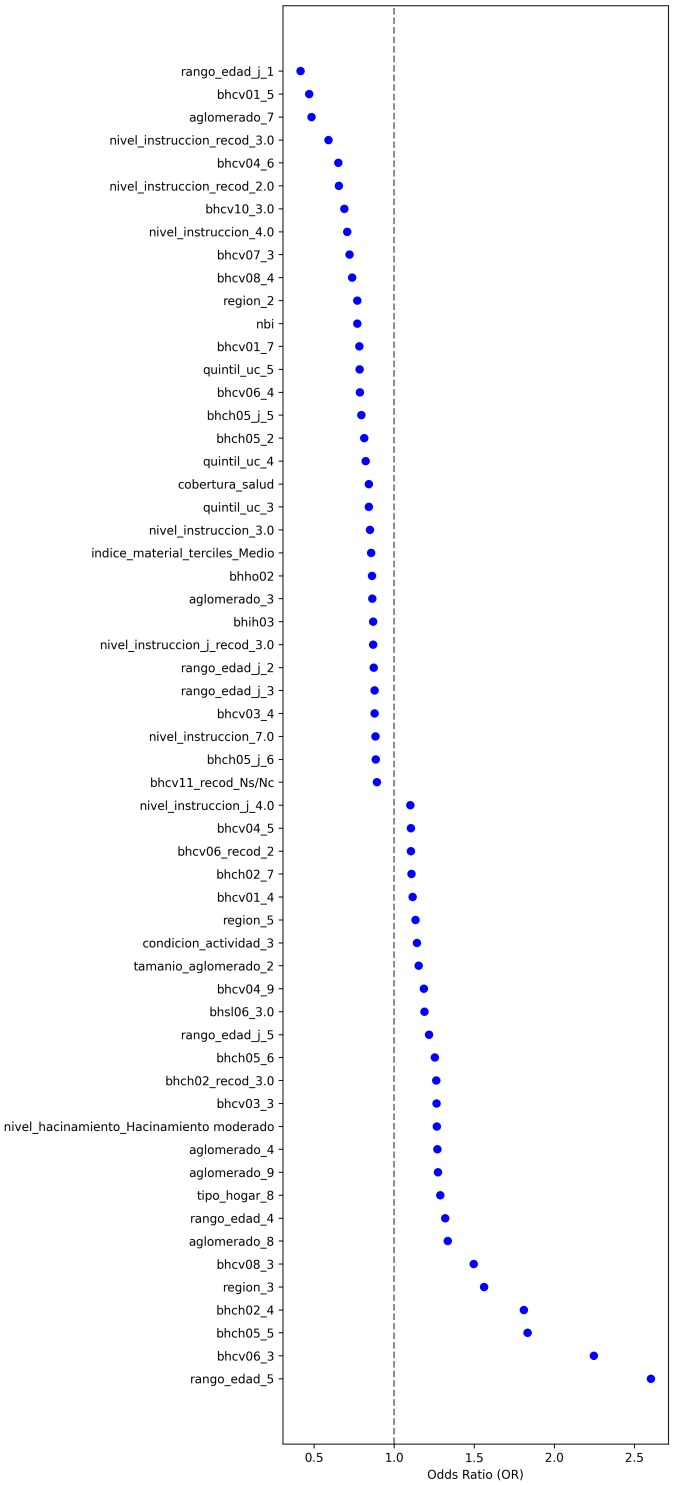

Figura 4.5: Forest plot de los *odds ratios* estimados por Lasso (sin provincias, mostrando solo OR entre 0.9 y 1.1).

4.2.3. Modelos predictivos

A continuación se presentan los resultados de los modelos de aprendizaje automático evaluados. La Tabla 4.3 muestra los valores de AUROC obtenidos mediante validación cruzada para cada modelo. **LGBM** y **Random Forest** obtuvieron el mejor desempeño, por lo que fueron seleccionados para el análisis posterior.

Tabla 4.3: Modelos predictivos según AUROC

Modelo	AUROC promedio
LGBM	**0.7143**
Random Forest	**0.7129**
SVM	0.7000
Naive Bayes	0.6850
Árbol de decisión	0.6840
KNN	0.6780

En relación a la **importancia acumulada de las variables**, seleccionando aquellas que explicaban aproximadamente el 80 % de la importancia total. La Figura 4.6 muestra las 30 variables más relevantes para LGBM y Random Forest.

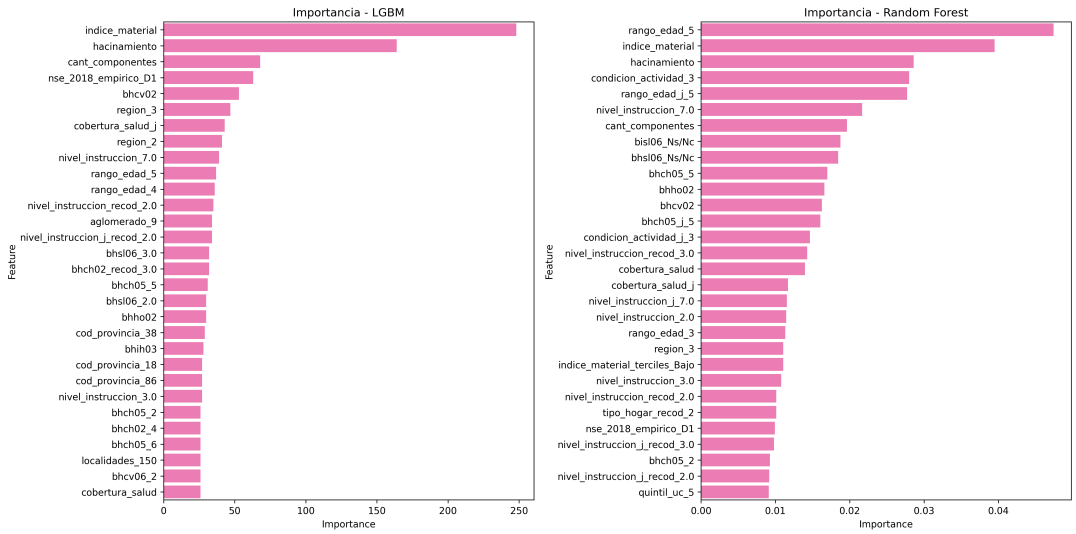

Figura 4.6: Importancia de las 30 variables principales en los modelos LGBM (izquierda) y Random Forest (derecha).

Los resultados de la evaluación de ambos modelos reducidos a las variables que representaba el 80 % de la importancia acumulada fueron los siguientes:

- **LGBM reducido**: AUROC = 0.7134
- **Random Forest reducido**: AUROC = 0.7115

Estos resultados muestran que los modelos predictivos de aprendizaje automático ofre-

cieron un nivel de discriminación similar (AUROC ≈ 0.71-0.72) al de los modelos explicativos.

4.2.4. Comparación de modelos

A continuación, se presenta la tabla de las variables seleccionadas por cada uno de los modelos considerados. El color celeste indica las categorías correspondientes al nivel `Territorial`, el verde al nivel `Hogar` y el rosa al `Individual`

Variable	Reg Logística	LASSO	LGBM	Random Forest
cod_provincia	✓	✓	✓	✓
region		✓	✓	✓
tamanio_aglomerado		✓	✓	✓
aglomerado		✓	✓	✓
localidades_150		✓	✓	✓
indice_material_terciles	✓	✓	✓	✓
tipo_hogar	✓	✓	✓	✓
nse_2018_empirico	✓	✓	✓	✓
bhcv01_recod	✓		✓	✓
hacinamiento	✓		✓	✓
bhcv01		✓	✓	✓
bhcv02_recod		✓	✓	✓
bhcv05		✓	✓	✓
bhcv06		✓	✓	✓
bhho02		✓	✓	✓
cant_componentes		✓	✓	✓
tipo_hogar_recod		✓	✓	✓
quintil_uc		✓	✓	✓
bhch03_j		✓	✓	✓
rango_edad_j		✓	✓	✓
nivel_instruccion_j		✓	✓	✓
nivel_instruccion_j_recod		✓	✓	✓
cobertura_salud_j		✓	✓	✓
bhsl06		✓	✓	✓
condicion_actividad_j		✓	✓	✓
bhcv03_recod		✓		✓
bhcv04		✓		✓
bhcv04_recod		✓		✓
bhcv06_recod		✓		✓
bhcv11		✓		✓
bhcv11_recod		✓		✓
nivel_hacinamiento		✓		✓
bhch05_j		✓		✓
bhcv02			✓	✓
indice_material			✓	✓
bhcv03		✓		
bhcv07		✓		

Variable	Reg Logística	LASSO	LGBM	Random Forest
bhcv08		✓		
bhcv09		✓		
bhcv10		✓		
bhho01		✓		
nbi		✓		
rango_edad	✓	✓	✓	✓
bhch05	✓	✓	✓	✓
nivel_instruccion_recod	✓	✓	✓	✓
cobertura_salud	✓	✓	✓	✓
condicion_actividad	✓	✓	✓	✓
bhih03		✓	✓	✓
nivel_instruccion		✓	✓	✓
bhch02		✓	✓	✓
bhch02_recod		✓	✓	✓
bisl06		✓		✓

Los cuatro modelos seleccionaron variables a diferentes escalas (territorial, hogar e individual), siendo las variables seleccionadas por todos los modelos: *cod_provincia* (jurisdicción de residencia), *indice_material_terciles* (nivel de material del hogar), *tipo_hogar* (estructura del hogar), *nse_2018_empirico* (nivel socioeconómico estimado), *rango_edad*, *bhch05* (situación conyugal), *nivel_instruccion_recod* (nivel de instrucción recategorizado), *cobertura_salud* (acceso a cobertura de salud) y *condicion_actividad* (condición de actividad económica).

5. DISCUSIÓN

Consistencias entre modelos

El análisis realizado muestra un patrón claro de desigualdades socioeconómicas y territoriales en la realización del estudio de Papanicolaou (PAP) en Argentina. A pesar de las diferencias metodológicas, los modelos utilizados —la regresión logística, el LASSO y los modelos de machine learning (LGBM y Random Forest)— coincidieron en resaltar un conjunto común de factores de riesgo.

En primer lugar, se destaca la **dimensión socioeconómica**: los niveles económicos más bajos —especialmente `nse_2018_empirico_D1` y `nse_2018_empirico_D2E`— están asociados a un menor acceso al PAP. Asimismo, las **condiciones materiales del hogar**, medidas por el `índice material` y el `hacinamiento`, emergen de forma transversal como determinantes clave. La **falta de cobertura de salud** o la **inactividad laboral** también contribuyen a reforzar estas desigualdades.

Por otro lado, los modelos muestran una clara relevancia del **nivel de instrucción**: aunque varía la categoría exacta destacada según el enfoque, existe acuerdo entre los modelos en que niveles educativos más altos se vinculan con una mayor probabilidad de realizarse el estudio.

Estos resultados están en línea con lo reportado por estudios previos, como *De Maio et al. (2012)* [5], *Nuche-Berenguer y Sakellariou (2021)* [6] y *Lamfre y Hasdeu (2019)* [7], donde afirman que las desigualdades estructurales y las condiciones socioeconómicas desfavorables constituyen barreras significativas para el acceso al PAP. Dichos estudios destacan cómo los bajos niveles de ingreso, la menor educación y las condiciones materiales precarias del hogar limitan el acceso a estas prácticas preventivas. Por ejemplo, *De Maio (2012)* muestra gradientes sociales marcados en la realización del PAP, mientras que *Nuche-Berenguer (2021)* enfatiza la relevancia del nivel educativo y de ingresos como determinantes. *Lamfre (2019)*, por su parte, refuerza la importancia de las condiciones del hogar y materiales de la vivienda, además del nivel educativo y de ingresos, para explicar estas desigualdades.

En segundo lugar, los modelos identifican **diferencias jurisdiccionales** significativas: jurisdicciones como CABA, Tierra del Fuego y La Pampa presentan menor riesgo de no realización, mientras que Santiago del Estero, muestra un riesgo significativamente mayor. Estas disparidades reflejan la importancia de las dinámicas territoriales en el acceso a la prevención y la necesidad de políticas públicas adaptadas a estas realidades. Estos hallazgos coinciden con lo reportado por *De Maio et al. (2012)* [5], que también destaca la existencia de desigualdades en el acceso al PAP vinculadas a factores territoriales en Argentina.

Por otro lado, la **edad** aparece como un factor de control que resulta relevante en todos los modelos: las mujeres de 50 años o más presentan sistemáticamente menor probabilidad de realización del pap, señalando la necesidad de estrategias de salud específicas para este

grupo. También el **estado civil** (`viudas` y `solteras`) se asocia a un mayor riesgo de no realización del PAP.

Diferencias entre modelos

Si bien existe consenso en las variables estructurales clave, el **modelo LASSO** y los modelos **LGBM y Random Forest** incorporaron otras variables y categorías que no aparecen en la regresión logística. Esto no implica que dichas variables no sean relevantes, sino que en la regresión logística la estimación de sus coeficientes puede verse alterada frente a la presencia de variables altamente correlacionadas [29]. Por ejemplo, LASSO capturó la importancia de casi todas las variables del `índice material` de forma individual, en contraste con los otros modelos que tienden a considerarlas solo en conjunto. Estas diferencias no significan que la regresión no reconozca la importancia de estos factores, sino que sus efectos se ven absorbidos por otras variables correlacionadas, algo que el LASSO y los modelos de árboles manejan mejor [26, 25].

Por otro lado, tanto LASSO como la regresión logística destacaron el rol de ciertos tipos de hogar, como el tipo 5 (*hogar multipersonal conyugal completo con hijos y con otros miembros*) y el tipo 8 (*hogar multipersonal no conyugal*), que no fueron tan relevantes en los modelos de Machine Learning. Esto podría evidenciar que estos hogares suelen presentar dinámicas más complejas —como mayores responsabilidades de cuidado y convivencia con más personas— que pueden dificultar la priorización de la salud individual y la organización de controles preventivos como el PAP.

Además, los modelos basados en árboles (LGBM y Random Forest) destacan la importancia de `cant_componentes` (cantidad de miembros del hogar) y `nivel_instruccion_recod_2.0` (educación secundaria incompleta). Esto refuerza la evidencia previa sobre cómo las condiciones del hogar y el nivel educativo son determinantes clave en la participación en controles preventivos como el PAP [5, 7, 6]. A diferencia de los modelos lineales, los métodos basados en árboles pueden capturar relaciones no lineales y manejar de forma más flexible la colinealidad entre variables. Por este motivo, estas variables se retienen en los modelos de árboles incluso cuando sus efectos están parcialmente solapados con otros determinantes [25].

Finalmente, se observa que en la regresión logística algunas categorías pierden significancia (con intervalos de confianza que cruzan el valor 1 en los OR), como las categorías intermedias de nivel socioeconómico (C2 y C3). Esto sugiere que su importancia puede depender del enfoque metodológico: LASSO y los modelos de árboles captan señales más difusas o correlacionadas, mientras que el modelo de regresión logística con selección de variables stepwise exige mayor robustez estadística para identificar asociaciones sólidas [29].

Aportes y limitaciones de los modelos

La combinación de modelos explicativos y predictivos permitió un análisis robusto. La **regresión logística** facilitó la identificación de factores de riesgo e informar un OR con sus respectivos IC y la identificación de efectos estadísticamente significativos. Por otro lado, el **modelo LASSO** sirvió como punto de comparación, resaltando la relevancia

de muchas más variables (incluso las del índice material de forma individual), aunque a costa de complejidad interpretativa. Y por último, los **modelos de Machine Learning** (LGBM y Random Forest) aportaron flexibilidad, simplicidad y la capacidad de capturar relaciones no lineales o interacciones complejas.

Cabe destacar que, a pesar de sus diferencias, todos los modelos presentaron capacidad predictiva (medida mediante AUROC) similar. Esto coincide con el estudio reportado por *Christodoulou et al. (2019)* [30], que concluye que no existe una diferencia significativa en el rendimiento predictivo entre algoritmos de aprendizaje automático y la regresión logística en modelos de predicción clínica. En el caso de esta tesis, esto podría indicar que el poder predictivo no depende tanto de la complejidad del modelo, sino de las variables estructurales y territoriales subyacentes en los datos. Esta idea es particularmente valiosa para quienes provienen del áreas de salud y están más familiarizados con modelos tradicionales como la regresión logística, ya que mostró que siguen siendo herramientas explicativas y predictivas confiables.

Sin embargo, los modelos tradicionales presentan limitaciones importantes. La regresión logística, a menos que se incluyan explícitamente términos de interacción o transformaciones en la fórmula del modelo, no puede detectar relaciones no lineales ni interacciones relevantes. Además, requiere un trabajo previo de limpieza, selección de variables y control de la colinealidad. Esto, por un lado, es positivo porque fomenta un análisis más cuidadoso y comprensible, pero también puede resultar tedioso y demandante en términos de tiempo y recursos, además de que quedan fuera del modelo por colinealidad variables que podrían ser relevantes, lo cual debe ser reportado. El modelo LASSO, aunque útil para la selección automática de variables, tiende a incluir muchas categorías (como las del índice material), lo que complica la interpretación directa de los resultados. Los modelos de Machine Learning, por su parte, requieren un manejo cuidadoso de la gran cantidad de variables dummificadas, lo que puede derivar en interpretaciones más engorrosas. Sin embargo, estos modelos son más eficientes y suelen requerir menos trabajo manual en la preparación de datos, y tienen la ventaja de captar automáticamente relaciones no lineales e interacciones complejas.

Desde la perspectiva de ciencia de datos, si bien existen modelos mucho más complejos (desde los árboles hasta modelos como los transformers o redes neuronales profundas), en temas de salud pública donde la interpretación es clave, modelos como la regresión logística continúan siendo una herramienta valiosa. Además, en el caso analizado de esta tesis, se observó que el desempeño predictivo del modelo de regresión logística fue comparable al resto de los modelos ajustados, pudiendo combinar capacidad explicativa con buen desempeño predictivo. Estos hallazgos se refuerzan los planteos de *Rudin (2019)* [31] y *Doshi-Velez y Kim (2017)* [32], quienes argumentan que, en contextos de alto impacto social como la salud pública, el uso de modelos interpretables es esencial para garantizar transparencia, confianza y la comprensión directa de los determinantes de riesgo. Aunque los modelos de aprendizaje automático tienen la ventaja de captar automáticamente relaciones no lineales e interacciones complejas, esto no se traduce necesariamente en un mejor rendimiento predictivo en datos estructurados como los utilizados aquí. Por ende, la regresión logística sigue siendo una herramienta de referencia que combina explicabilidad y confiabilidad para la toma de decisiones en estos contextos.

Factores de riesgo identificados

El análisis conjunto de los modelos permitió identificar variables que permiten caracterizar **grupos de riesgo consistentes** para la no realización del estudio de Papanicolaou (PAP) en Argentina. Estas variables relevantes describen distintas dimensiones del nivel socioeconómico (ingresos, educación, condiciones materiales), tanto a escala individual como del hogar. En particular, los niveles más bajos de NSE —especialmente `nse_2018_empirico_D1` y `nse_2018_empirico_D2E`— muestran consistentemente menor cobertura. Las **condiciones materiales y de hacinamiento** —medidas por el `índice material` y la variable de `hacinamiento`— muestran que las mujeres pertenecientes a hogares con mayor precariedad material y hacinamiento severo enfrentan mayores dificultades para acceder al PAP, reflejando desigualdades en las condiciones de vida [5, 7, 6]. La **falta de cobertura de salud** y la **inactividad laboral** también surgen como variables importantes, mientras que trabajar más de 45 horas semanales (`bhsl06=3`) se asocia a menor probabilidad de control, evidenciando cómo las exigencias laborales pueden obstaculizar el acceso a la salud preventiva [33]. Además, existe consenso entre los modelos en que los niveles educativos más altos se asocian a una mayor probabilidad de realizarse el estudio, destacando la relevancia del **nivel de instrucción**. Esto también, se condice con lo reportado por la bibliografía, que documenta cómo la escolaridad más alta aumenta la participación en prácticas preventivas como el PAP en Argentina [5, 7, 6]. En cuanto a la estructura del hogar, destacan los hogares tipo 5 (*hogar multipersonal conyugal completo con hijos y con otros miembros*) y tipo 8 (*hogar multipersonal no conyugal*) como factores de riesgo en la regresión logística y LASSO, probablemente por las múltiples responsabilidades y la convivencia con varios miembros que dificultan la priorización de la salud individual. Además, pertenecer a un hogar donde se es *otro familiar que no sea cónyuge/pareja o no familiar* del jefe o jefa de hogar surge también como una variable significativa, lo que podría sugerir que la posición de menor centralidad en el hogar podría limitar el acceso o la prioridad para realizar controles preventivos.

Los modelos también identifican **diferencias territoriales significativas**: jurisdicciones como CABA, Tierra del Fuego y La Pampa presentan menor riesgo de no realización, mientras que otras, como Santiago del Estero, Corrientes y La Rioja muestran un riesgo significativamente mayor, reflejando la importancia de las dinámicas territoriales en el acceso a la prevención [5]. Vivir en aglomerados más pequeños —como Gran Salta o localidades con menos de 150.000 habitantes— se observa en este análisis como un posible factor asociado a menor acceso al control, lo que sugiere que las dinámicas poblacionales y el acceso a servicios varían según el contexto territorial. Finalmente, variables como la **edad** —en particular, las mujeres de 50 años o más— se destacan como un factor de riesgo transversal respaldado por la bibliografía [7], mientras que el **estado civil** (`viudas` y `solteras`) surgió en este análisis como otra posible variable asociada a mayor riesgo, lo que subraya la necesidad de estrategias de salud específicas para estos grupos [6].

En conjunto, los factores de riesgo identificados en este trabajo refuerzan la necesidad e importancia de implementar políticas públicas focalizadas con el objetivo de mejorar la cobertura y la equidad en la salud preventiva en Argentina. La consistencia de los modelos en señalar **determinantes estructurales** como la precariedad material, la falta de cobertura de salud y el hacinamiento deja en claro que la equidad en el acceso a la prevención no puede lograrse sin abordar estas desigualdades de base.

Además, los hallazgos muestran que tanto **factores socioeconómicos** como **territoriales** —como la provincia de residencia, el tamaño del aglomerado o la carga laboral— se relacionan fuertemente con la probabilidad de realizarse el PAP. Esto subraya la necesidad de políticas **descentralizadas** que no sólo garanticen la provisión de recursos, sino que también reconozcan las barreras culturales, económicas y sociales que enfrentan las mujeres en situación de mayor vulnerabilidad.

En conjunto, estos resultados ponen de manifiesto que la **equidad en salud** requiere un enfoque que considere la diversidad de contextos y necesidades, y que priorice a los grupos más afectados por estas desigualdades estructurales.

5.1. Conclusiones finales y líneas futuras

Los resultados obtenidos confirman que las desigualdades socioeconómicas y territoriales siguen siendo determinantes clave en la no realización del estudio de Papanicolaou (PAP) en Argentina. La consistencia entre modelos y la identificación de factores de riesgo consistentes entre distintas metodologías empleadas refuerzan la urgencia de políticas públicas **focalizadas, adaptadas a las realidades locales y sensibles al territorio**, con el objetivo de reducir estas brechas en el acceso a la prevención. La metodología aplicada, que combina diferentes estrategias de modelado, permitió un análisis integral que combina interpretabilidad y capacidad de capturar relaciones complejas, lo que resulta fundamental para comprender la realización del PAP en Argentina.

A futuro, un paso relevante sería comparar estos resultados con los de futuras ediciones de la Encuesta Nacional de Factores de Riesgo (ENFR), ya que la edición prevista para 2022 fue suspendida por la pandemia de COVID-19 y la ENFR 2018 continúa siendo la fuente más actualizada y completa para este análisis. Por otro lado, variables que en este trabajo no fueron identificadas como relevantes, como `bhih03` (Si recibió algún ingreso por AUH u otro plan social), podrían recuperarse en estudios futuros. Finalmente, también sería valioso refinar el uso del modelo LASSO, explorando no sólo la presencia de coeficientes distintos de cero, sino también su magnitud y relevancia. Esto permitiría construir un conjunto de variables todavía más refinado.

En conjunto, esta tesis contribuye a la comprensión de las desigualdades en el acceso a una práctica preventiva de cáncer cervicouterino en Argentina y ofrece un marco metodológico replicable para continuar estudiando tanto el acceso a esta práctica como la posibilidad de extenderlo al estudio de otras prácticas preventivas.

Bibliografía

[1] Instituto Nacional del Cáncer - Programa Nacional de Prevención del Cáncer Cervicouterino (PNPCC). Programa nacional de prevención del cáncer cervicouterino, 2024. Accedido el 18 de marzo de 2025.

[2] J. Ferlay, M. Ervik, F. Lam, M. Colombet, L. Mery, M. Piñeros, A. Znaor, I. Soerjomataram, and F. Bray. Global cancer observatory: Cancer today. 2024. Disponible en `https://gco.iarc.fr/today`.

[3] S. Dasgupta. The efficiency of cervical pap and comparison of conventional pap smear and liquid-based cytology: A review. *Cureus*, 15(11):e48343, 2023. https://doi.org/10.7759/cureus.48343.

[4] Ministerio de Salud de la Nación. Cáncer de cuello de útero y el PAP. `https://www.argentina.gob.ar/salud/cancer/tipos/cancer-de-cuello-de-utero-ccu#:~:text=cuello%20de%20%C3%BAtero.-,El%20PAP,p%C3%BAblicos%20de%20todo%20el%20pa%C3%ADs.`, 2025. Accedido el 4 de junio de 2025.

[5] F.G. De Maio, B. Linetzky, and D. Ferrante. Changes in the social gradients for pap smears and mammograms in argentina: Evidence from the 2005 and 2009 national risk factor surveys. *Public Health*, 126:821–826, 2012.

[6] Bernardo Nuche-Berenguer and Dikaios Sakellariou. Socioeconomic determinants of participation in cancer screening in argentina: A cross-sectional study. *Frontiers in Public Health*, 9:699108, 2021.

[7] 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.

[8] S. Arrossi, M. Paolino, and R. Sankaranarayanan. Determinants of women's participation in cervical cancer screening in a high mortality setting in argentina. *Cancer Detection and Prevention*, 32:295–303, 2008.

[9] Instituto Nacional de Estadística y Censos de la República Argentina. Encuesta nacional de factores de riesgo. `https://www.indec.gob.ar/indec/web/Nivel4-Tema-4-32-68`, s.f. Base de datos sobre factores de riesgo en Argentina.

[10] Instituto Nacional de Estadística y Censos (INDEC). Sitio web del instituto nacional de estadística y censos (indec). `https://www.indec.gob.ar/`, 2025. Portal oficial con acceso a estadísticas y publicaciones sobre Argentina.

[11] Ahmad Reza Hosseinpoor, Nicole Bergen, Aluisio J.D. Barros, Kerry L.M. Wong, Ties Boerma, and Cesar G. Victora. Monitoring subnational regional inequalities in health: measurement approaches and challenges. *International Journal for Equity in Health*, 15:18, 2016.

[12] INDEC. Encuesta nacional de factores de riesgo 2018 - cuestionario. https://www.indec.gob.ar/ftp/cuadros/sociedad/cuestionario_enfr_2018.pdf, 2018. Disponible en línea. Consultado el 16 de junio de 2025.

[13] Instituto Nacional de Estadística y Censos (INDEC). Indicadores de condiciones de vida de los hogares en 31 aglomerados urbanos. 4to trimestre de 2019, 2019. [Accedido: junio 2025].

[14] 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.

[15] Cámara de Empresas de Investigación Social y de Mercado (CEIM), Asociación Argentina de Marketing (AAM), and Sociedad Argentina de Investigadores de Marketing y Opinión (SAIMO). ¿Qué es el NSE? https://ceim.org.ar/wp-content/uploads/2022/06/Proyecto-2022-NSE-Fichas.pdf, 2022. Accedido el 4 de junio de 2025.

[16] 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.

[17] Michael Greenacre. *Correspondence Analysis in Practice*. Chapman and Hall/CRC, 2007.

[18] Max Halford. prince: Python library for correspondence analysis and related techniques. https://github.com/MaxHalford/prince, 2019. [Accedido: junio 2025].

[19] The pandas development team. pandas: powerful Python data analysis toolkit, 2024. [Accedido: junio 2025].

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

[21] Gareth James, Daniela Witten, Trevor Hastie, and Robert Tibshirani. *An Introduction to Statistical Learning with Applications in R*. Springer, 2013.

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

[23] Pedregosa, F. and Varoquaux, G. and Gramfort, A. and Michel, V. and Thirion, B. and Grisel, O. and Blondel, M. and Prettenhofer, P. and Weiss, R. and Dubourg, V. and Vanderplas, J. and Passos, A. and Cournapeau, D. and Brucher, M. and Perrot, M. and Duchesnay, É. Scikit-learn: Machine Learning in Python, 2011. [Accedido: junio 2025].

[24] Seabold, Skipper and Perktold, Josef and others. statsmodels: Statistical models in Python. https://www.statsmodels.org/stable/generated/statsmodels.formula.api.logit.html, 2010. [Accedido: junio 2025].

[25] Andreas C. Müller and Sarah Guido. *Introduction to Machine Learning with Python: A Guide for Data Scientists*. O'Reilly Media, Sebastopol, CA, 2016. Libro introductorio sobre aprendizaje automático con Python y scikit-learn.

[26] Robert Tibshirani. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society: Series B (Methodological)*, 58(1):267–288, 1996.

[27] Pedregosa, F. and Varoquaux, G. and Gramfort, A. and Michel, V. and Thirion, B. and Grisel, O. and Blondel, M. and Prettenhofer, P. and Weiss, R. and Dubourg, V. and Vanderplas, J. and Passos, A. and Cournapeau, D. and Brucher, M. and Perrot, M. and Duchesnay, É. Scikit-learn: Machine Learning in Python – Preprocessing data, 2011. [Accedido: junio 2025].

[28] Pedregosa, F. and Varoquaux, G. and Gramfort, A. and Michel, V. and Thirion, B. and Grisel, O. and Blondel, M. and Prettenhofer, P. and Weiss, R. and Dubourg, V. and Vanderplas, J. and Passos, A. and Cournapeau, D. and Brucher, M. and Perrot, M. and Duchesnay, É. Scikit-learn: Machine Learning in Python – RandomizedSearchCV, 2011. [Accedido: junio 2025].

[29] Carsten F. Dormann, Jane Elith, Sven Bacher, Carsten Buchmann, Gudrun Carl, Clémentine Carré, Jaime R. García Marquéz, Bernd Gruber, Bernard Lafourcade, Pedro J. Leitão, Tamara Münkemüller, Colin McClean, Patrick E. Osborne, Björn Reineking, Boris Schröder, Andrew K. Skidmore, Damaris Zurell, and Sven Lautenbach. Collinearity: a review of methods to deal with it and a simulation study evaluating their performance. *Ecography*, 36(1):27–46, 2013.

[30] Evangelos Christodoulou, Junyan Ma, Gary S. Collins, Ewout W. Steyerberg, Jan Y. Verbakel, and Ben Van Calster. A systematic review shows no performance benefit of machine learning over logistic regression for clinical prediction models. *Journal of Clinical Epidemiology*, 110:12–22, 2019.

[31] Cynthia Rudin. Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nature Machine Intelligence*, 1:206–215, 2019.

[32] Finale Doshi-Velez and Been Kim. Towards a rigorous science of interpretable machine learning. *arXiv preprint arXiv:1702.08608*, 2017.

[33] Organización Panamericana de la Salud (OPS). Las funciones esenciales de la salud pública en las américas: una renovación para el siglo xxi, 2017. Disponible en: `https://iris.paho.org/handle/10665.2/34015` [Accedido: junio de 2025].