Predicción de *No-shows* a citas médicas con aprendizaje de máquinas: un estudio con datos reales de Brasil

Carlos Castillo

11 de diciembre de 2025

Resumen

Se presenta un estudio de predicción de *no-shows* en atención ambulatoria utilizando modelos tabulares (regresión logística y Random Forest). Empleamos un conjunto de datos real de 110,527 citas (Vitória, Brasil), con prevalencia de *no-show* de 20.190 %. Mostramos resultados de validación cruzada y en prueba (ROC-AUC, PR-AUC) y, para evaluación orientada a operación, reportamos P@10 % y P@20 %. Los modelos se implementan en un pipeline reproducible de scikit-learn. También esbozamos un protocolo de datos sintéticos para futuras pruebas controladas.

1. Introducción

Las ausencias a cita (*no-shows*) constituyen un shock operativo persistente en la atención ambulatoria: reducen la productividad, generan capacidad ociosa difícil de reasignar y deterioran la continuidad del cuidado, con consecuencias clínicas y económicas no triviales (Hasvold & Wootton, 2011, Salazar et al., 2022). Además, el ausentismo no es aleatorio: suele correlacionarse con variables observables registradas al agendamiento (edad, barrio, franja horaria, tiempo de espera entre programación y atención), lo que abre la puerta a modelos predictivos útiles en la práctica (Salazar et al., 2022).

Dos cuerpos de evidencia motivan este estudio. Por un lado, la literatura de recordatorios muestra que los mensajes de texto (SMS) incrementan la asistencia frente a no recordar y, a menudo, con costos menores que las llamadas telefónicas, aunque con heterogeneidad entre contextos y calidad de evidencia baja–moderada (Gurol-Urganci et al., 2013, Hasvold & Wootton, 2011). Por otro lado, trabajos recientes en aprendizaje de máquinas (ML) documentan que, con un preprocesamiento cuidadoso y métricas de evaluación adecuadas al desbalance, los modelos tabulares logran desempeños competitivos para priorizar pacientes con alto riesgo de *no-show* en entornos reales (Liu et al., 2022, Salazar et al., 2022).

Este artículo se enfoca en la componente predictiva con datos reales. Usando un conjunto público de 110,527 citas ambulatorias de Vitória (Brasil), construimos un pipeline reproducible con dos clasificadores de referencia: regresión logística y Random Forest, y un preprocesamiento explícito (codificación one-hot en categóricas y paso directo en numéricas). Se reportan métricas globales (ROC-AUC, PR-AUC) y métricas de priorización operativa (Precision@k), que capturan cuántos *no-shows* se concentran en el top-k de riesgo cuando la capacidad de intervención es limitada (Vickers & Elkin, 2006).

Deliberadamente, **no** se abordan cuestiones causales (ex, sesgo de selección en el envío histórico de SMS) ni se proponen reglas de asignación, esas preguntas cruciales para traducir *scores* en decisiones requieren diseños experimentales o métodos causales específicos y se dejan para trabajo futuro. El

objetivo en esta etapa es fijar una línea base clara, reproducible y evaluada fuera de muestra que pueda integrarse, sin fricción, a flujos institucionales.

2. Datos

2.1. Fuente, limpieza y variables

Se usa el conjunto público *Medical Appointment No Shows* (110,527 filas, 14 columnas) y, tras la limpieza mínima (estandarización de encabezados y tipos, derivación de `wait_days`, `appt_weekday`, `sched_hour`, eliminación de 1 edad negativa), se trabaja con 110,526 registros. La etiqueta es `no_show` (1 = no asistió, 0 = asistió). El balance final se muestra en la Tabla 1 (prevalencia $\approx 20.190\,\%$).

Cuadro 1: Balance de clases (muestra final).

Clase	**Conteo**	**Proporción**
Asistió (y=0)	88,207	0.7981
No-show (y=1)	22,319	0.2019

Variables usadas: Numéricas: `age`, `handcap` (ordinal 0–4), `wait_days`, `sched_hour`. Dummies: `scholarship`, `hipertension`, `diabetes`, `alcoholism`, `sms_received`. Categóricas: `gender` (F/M), `neighbourhood` ($\sim$ 81 niveles), `appt_weekday` (en el archivo aparecen 6 días, no hay domingo).

2.2. Evidencia descriptiva y exploración visual

La Figura 5 resume las distribuciones marginales de todas las variables del conjunto de datos ya depurado. En `age` se observa el patrón típico de atención ambulatoria: una masa principal entre $\sim$20 y 70 años, presencia de niños y adultos mayores, y algunos valores extremos altos (≥ 100). En `wait_days` la distribución es marcadamente asimétrica a la derecha: la mayoría de las citas se programan con esperas cortas y existe una cola de esperas largas (decenas de días), compatible con agendas por cupos. `sched_hour` exhibe la naturaleza discreta de la programación (horas enteras entre 6:00 y 20:00, con picos en la mañana), lo que anticipa bandas horizontales en relaciones bivariadas. `handcap` es una variable ordinal fuertemente cero–inflada (casi todo en 0, pocos casos en 1–4), para modelos lineales conviene considerar codificación ordinal explícita o una versión binaria (≥ 1). Entre las categóricas, `gender` está desbalanceada a favor de mujeres (fenómeno habitual en APS) y `appt_weekday` muestra seis niveles (lunes–sábado), sin programaciones en domingo en esta base.`neighbourhood` tiene cardinalidad moderada ($\approx$ 81 niveles). Aunque es manejable con *one-hot*, usamos `handle_unknown=ignore"` para garantizar robustez ante categorías no vistas en evaluación/despliegue. Las dummies clínicas (`hipertension`, `diabetes`, `alcoholism`) son poco prevalentes coherente con registros poblacionales y previsiblemente aportarán señal en interacción con edad más que de forma marginal. Finalmente, `sms_received` aparece escasa y desbalanceada hacia 0.

La matriz de correlaciones de Spearman en la Figura 6 confirma baja colinealidad global y tres asociaciones de magnitud relevante: (i) `wait_days`–`sms_received` $\rho \approx 0.57$, consistente con que esperas largas activan más recordatorios, (ii) `age`–`hipertension` $\rho \approx 0.50$, y (iii) `age`–`diabetes` $\rho \approx 0.29$, patrones clínicamente plausibles. El resto de correlaciones es pequeño en valor absoluto ($|\rho| \lesssim 0.13$). En términos de modelado, esto sugiere que gran parte de la señal potencial provendrá

de no linealidades e interacciones (ex, edad × comorbilidades o espera × día/hora), más que de efectos lineales marginales.
Las relaciones bivariadas de la Figura 7, trazadas sobre una submuestra estratificada para hacer visible la clase minoritaria, refuerzan la intuición anterior. En *wait_days vs. age* la nube es triangular: predominan esperas cortas en todo el rango etario, los *no-show* (rojo) se mezclan con los asistidos (azul) sin fronteras lineales claras, con un leve aumento de densidad de *no-shows* cuando la espera se alarga. En *sched_hour vs. age* se aprecian bandas horizontales por horas, con mezcla prácticamente uniforme por clase indicio de efecto marginal débil de la hora y, en su caso, no monótono. En *sched_hour vs. wait_days* predominan nuevamente esperas cortas para casi todas las horas, cualquier señal útil parece emerger más de combinaciones (hora × día, hora × tipo de paciente) que de cada variable por separado. Estos patrones justifican evaluar, junto a una base lineal (logística), modelos capaces de capturar interacciones y no linealidades (ex, Random Forest).
Desagregando el resultado por calendario, la Figura 8 muestra una heterogeneidad moderada por día de la semana: la tasa de *no-show* se mantiene cercana al 20 % de lunes a jueves (20.6 %, 20.1 %, 19.7 %, 19.4 %), sube el viernes (21.2 %) y alcanza un máximo el sábado (23.1 %). La amplitud total es ∼3.7 p.p. (sábado vs. jueves). Este patrón es compatible con mezcla de casos y prácticas de programación (mayor proporción de citas opcionales/reprogramadas hacia fin de semana), por lo que no se interpreta causalmente, sí justifica incluir `appt_weekday` (one–hot ordenado) y explorar interacciones con `sched_hour` y `wait_days`.
Por último, la Figura 9 compara tasas brutas según recordatorio por SMS: $\hat{p}_{\text{no SMS}} \approx 16.7\,\%$ frente a $\hat{p}_{\text{SMS}} \approx 27.6\,\%$, diferencia de ∼10.9 p.p. (relativa ∼+65 %). Con tamaños muestrales muy distintos (aprox. 75k sin SMS y 35k con SMS), los errores estándar son pequeños ($\text{se}(\hat{p}) = \sqrt{\hat{p}(1-\hat{p})/n} \approx$ 0.13–0.24 p.p.). Dado que `sms_received` está fuertemente correlacionada con `wait_days` ($\rho \approx 0.57$), este contraste refleja *targeting* operativo y diferencias de mezcla temporal, no un efecto causal del mensaje. En la parte predictiva, se trata `sms_received` como una característica informativa del proceso, cualquier inferencia causal requeriría un diseño o identificación específicos.

Entonces, se tiene las siguientes implicaciones para el modelado: el conjunto de evidencias sugiere: (i) tratar `wait_days` y `sched_hour` con codificaciones/transformaciones acordes a su forma discreta y asimétrica, (ii) controlar la cardinalidad de `neighbourhood` para evitar una expansión excesiva de dummies, (iii) considerar codificación ordinal/binarización de `handcap`, y (iv) incorporar pesos de clase en el entrenamiento dada la prevalencia de *no-show* (≈ 20.2 %). Sobre esta base se implementa una línea baseline (logística con *one–hot*) y un modelo no lineal (Random Forest) para capturar interacciones de calendario y perfil clínico.

3. Metodología

3.1. Split estratificado y protocolo fuera de muestra

Se parte de X (covariables) y $y =$ `no_show` $\in \{0, 1\}$. Se realiza un split estratificado entrenamiento/prueba de 80/20 con `random_state=42`. El conjunto final tiene $N = 110{,}526$ observaciones: $N_{\text{train}} = 88{,}420$ y $N_{\text{test}} = 22{,}106$, manteniendo prevalencia prácticamente idéntica en ambos subconjuntos ($\Pr(y{=}1) \approx 0.2019$).

3.2. Preprocesamiento y espacio de características

Se usa un `ColumnTransformer` con: (i) *one-hot encoding* para `gender`, `neighbourhood` y `appt_weekday`, con `handle_unknown=ïgnore"`, y (ii) paso directo para el resto de variables numéricas/binarias. La

salida del preprocesamiento se construye en formato denso. Tras la expansión categórica, el número total de características es $p = 98$ (89 dummies categóricas + 9 variables pasadas directamente).

3.3. Modelos

Sea $\{(x_i, y_i)\}_{i=1}^n$ una muestra i.i.d. donde $x_i \in \mathcal{X}$ es el vector de covariables observadas al agendamiento y $y_i \in \{0, 1\}$ indica ausentismo ($y_i = 1$ si el paciente no asiste). Denote por $\phi : \mathcal{X} \to \mathbb{R}^p$ el mapeo de preprocesamiento (expansión *one-hot* de categóricas y paso directo de numéricas/binarias), y por $z_i = \phi(x_i)$ el vector de características resultante. Ambos modelos producen un *score* $\hat{s}(x) \in [0, 1]$ interpretable como probabilidad estimada de *no-show*, y se estiman dentro de un Pipeline que encapsula $\phi(\cdot)$ y el estimador, de modo que ϕ se ajusta únicamente con datos de entrenamiento en cada partición, evitando data leakage.

Regresión logística La regresión logística especifica un modelo paramétrico para la probabilidad condicional:

$$\Pr(y_i = 1 \mid x_i) = \Pr(y_i = 1 \mid z_i) = \sigma(\beta_0 + \beta^\top z_i), \qquad \sigma(t) = \frac{1}{1 + e^{-t}}. \tag{1}$$

Los parámetros (β_0, β) se estiman por máxima verosimilitud penalizada (*ridge*), resolviendo

$$\min_{\beta_0, \beta} \sum_{i=1}^{n} w_{y_i} \log\Big(1 + \exp\big(-(2y_i - 1)(\beta_0 + \beta^\top z_i)\big)\Big) + \frac{\lambda}{2}\|\beta\|_2^2, \tag{2}$$

donde w_y son pesos por clase y $\lambda > 0$ controla la regularización. En esta implementación, `class_weight="balanced"` fija w_1 y w_0 inversamente proporcionales a las frecuencias de clase en entrenamiento (equivalente a dar mayor costo a errores sobre la clase minoritaria). El parámetro λ corresponde a la penalización ℓ_2 por defecto en `scikit-learn` y `max_iter=1000` garantiza convergencia numérica del solver.

Random Forest. El Random Forest es un ensamble no paramétrico de árboles de clasificación. Cada árbol $b \in \{1, \ldots, B\}$ se entrena sobre una muestra *bootstrap* de los datos y, en cada nodo, la partición se elige considerando un subconjunto aleatorio de características. Cada árbol induce una regla de probabilidad $\hat{s}_b(x) = \hat{s}_b(z) \in [0, 1]$ (proporción de clase 1 en la hoja terminal). El bosque promedia estas probabilidades:

$$\hat{s}_{\mathrm{RF}}(x) = \frac{1}{B} \sum_{b=1}^{B} \hat{s}_b(x). \tag{3}$$

En la práctica fije $B = 300$ (`n_estimators=300`) y se controla complejidad imponiendo un tamaño mínimo de hoja `min_samples_leaf=20`, lo cual reduce varianza al evitar particiones demasiado finas. Para manejar el desbalance, `class_weight="balanced_subsample"` aplica pesos por clase dentro de cada muestra *bootstrap*. La aleatoriedad se controla con `random_state=42` y el entrenamiento se paraleliza con `n_jobs=-1`.

3.4. Validación cruzada en entrenamiento y métricas

La selección de desempeño se realiza en el conjunto de entrenamiento usando validación cruzada estratificada de 5 pliegues (`StratifiedKFold`, `shuffle=True`, `random_state=42`). Se reporta ROC-AUC y PR-AUC (Average Precision, AP).
En el conjunto de prueba reservado se reporta además P@10 % y P@20 %, definidas como la proporción de verdaderos *no-shows* dentro del 10 % (20 %) de pacientes con mayor riesgo predicho.

Cuadro 2: Desempeño en validación cruzada (5-fold) sobre `train`.

Modelo	**AUC (mean)**	**AUC (sd)**	**AP (mean)**	**AP (sd)**
Regresión logística	0.670	0.005	0.306	0.004
Random Forest	0.736	0.005	0.370	0.004

Cuadro 3: Desempeño en prueba reservada (`test`).

Modelo	**ROC-AUC**	**PR-AUC** (AP)	P@10 %	P@20 %
Regresión logística	0.674	0.310	0.356	0.345
Random Forest	0.738	0.372	0.418	0.385

4. Resultados

4.1. Desempeño en validación cruzada (entrenamiento)

La Tabla 2 resume el desempeño promedio (y desviación estándar) en validación cruzada estratificada de 5 pliegues sobre el conjunto de entrenamiento. El Random Forest domina a la logística tanto en ROC-AUC como en PR-AUC, con variabilidad pequeña entre pliegues, lo que sugiere estabilidad fuera de muestra dentro del mismo régimen de datos.

4.2. Desempeño fuera de muestra (prueba reservada)

La Tabla 3 reporta el desempeño en el conjunto de prueba reservado ($N_{\text{test}} = 22{,}106$). El patrón se mantiene: el Random Forest obtiene AUC = 0.738 y AP = 0.372, frente a AUC = 0.674 y AP = 0.310 de la logística.

4.3. Lectura operativa: P@k, PR y ganancias acumuladas

Curva Precision–Recall (PR). Para un umbral t, defina $\hat{y}(t) = \mathbf{1}\{\hat{s}(x) \geq t\}$. La precisión y el recall son

$$\text{Precision}(t) = \frac{\text{TP}(t)}{\text{TP}(t) + \text{FP}(t)}, \qquad \text{Recall}(t) = \frac{\text{TP}(t)}{\text{TP}(t) + \text{FN}(t)}.$$

La Figura 1 muestra que el Random Forest mantiene mayor precisión para un mismo nivel de recall, coherente con $\text{AP}_{\text{test}} = 0.372$ frente a 0.310 en logística (Tabla 3). En un problema desbalanceado, la referencia natural es la precisión base, igual a la prevalencia $\Pr(y = 1) \approx 0.202$.

Priorización y P@k. Si la institución solo puede intervenir a una fracción k de pacientes, se ordena por score $\hat{s}_i$ y defina $\mathcal{T}_k$ como el conjunto top-k de tamaño $|\mathcal{T}_k| = \lceil kN_{\text{test}} \rceil$. Entonces

$$\text{P@k} = \frac{1}{|\mathcal{T}_k|} \sum_{i \in \mathcal{T}_k} \mathbf{1}\{y_i = 1\}, \qquad \text{Lift@}k = \frac{\text{P@k}}{\Pr(y = 1)}.$$

En prueba, con $\Pr(y = 1) \approx 0.202$, la logística obtiene P@10 % = 0.356 (Lift@10 % $\approx 1.76\times$) y el Random Forest P@10 % = 0.418 (Lift@10 % $\approx 2.07\times$). Para $k = 20\,\%$, los valores son 0.345 vs. 0.385 (lift $\approx 1.71\times$ vs. $1.91\times$).

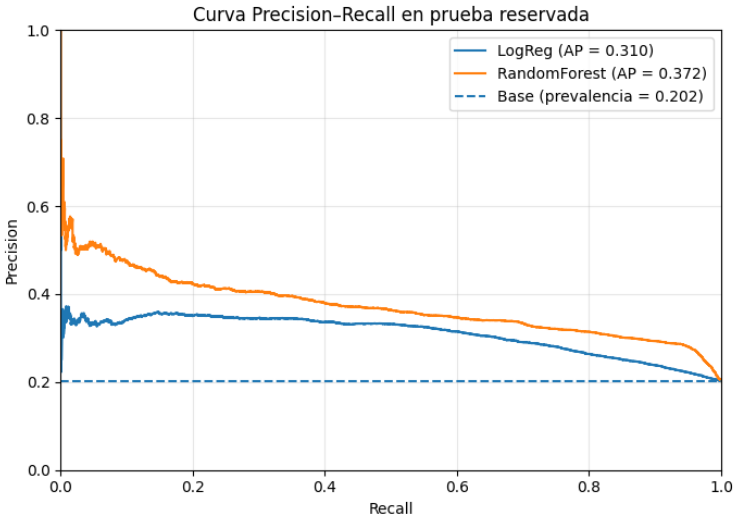

Figura 1: Curvas Precision–Recall (PR) en prueba reservada. La línea horizontal punteada corresponde a la precisión base, igual a la prevalencia de *no-show* en el conjunto de prueba ($\Pr(y = 1) \approx 0.202$).

Curva de ganancias acumuladas (gains). La Figura 2 reporta

$$G(k) = \frac{\sum_{i \in \mathcal{T}_k} \mathbf{1}\{y_i = 1\}}{\sum_{i=1}^{N_{\text{test}}} \mathbf{1}\{y_i = 1\}},$$

que mide la fracción de todos los *no-shows* capturados al priorizar una fracción k de pacientes. Bajo selección aleatoria, $G(k) = k$. En nuestros resultados, priorizar el top-10 % captura $G(0.10) = 0.177$ con logística y $G(0.10) = 0.207$ con Random Forest, para top-20 %, $G(0.20) = 0.341$ y 0.382, respectivamente. Esto muestra que el score del Random Forest concentra más riesgo y, por tanto, es superior como herramienta de focalización cuando la capacidad de intervención es limitada, sin interpretar causalmente covariables operativas (ex, SMS).

4.4. Endogeneidad del SMS y riesgo contrafactual para política

En la base histórica, el recordatorio por SMS no se asigna al azar. Denote: $D \in \{0, 1\}$ indicador de SMS ($D = 1$ si recibió SMS), $Y \in \{0, 1\}$ indicador de *no-show*, y X covariables pre-tratamiento observables (edad, espera, día/hora, barrio, etc.). Bajo asignación endógena, entrenar un único modelo con D como covariable puede recuperar el riesgo bajo la política histórica (mezcla):

$$\Pr(Y = 1 \mid X) = e(X)\, p_1(X) + (1 - e(X))\, p_0(X),$$

donde $e(X) = \Pr(D = 1 \mid X)$, $p_1(X) = \Pr(Y = 1 \mid X, D = 1)$ y $p_0(X) = \Pr(Y = 1 \mid X, D = 0)$. Sin embargo, para decidir *ex-ante* si enviar SMS se necesita precisamente $p_0(X)$, el riesgo sin tratamiento. En particular, entrenar sobre la mezcla y luego "poner $D = 0$" al aplicar una política no produce un score interpretable como riesgo contrafactual sin intervención.

Enfoque propuesto (dos modelos). Separa explícitamente: (i) un modelo descriptivo que predice bajo la política histórica (puede incluir D como covariable), útil para EDA y para documentar targeting operacional, y (ii) un modelo de política ex-ante que estima

$$p_0(X) = \Pr(Y = 1 \mid X, D = 0),$$

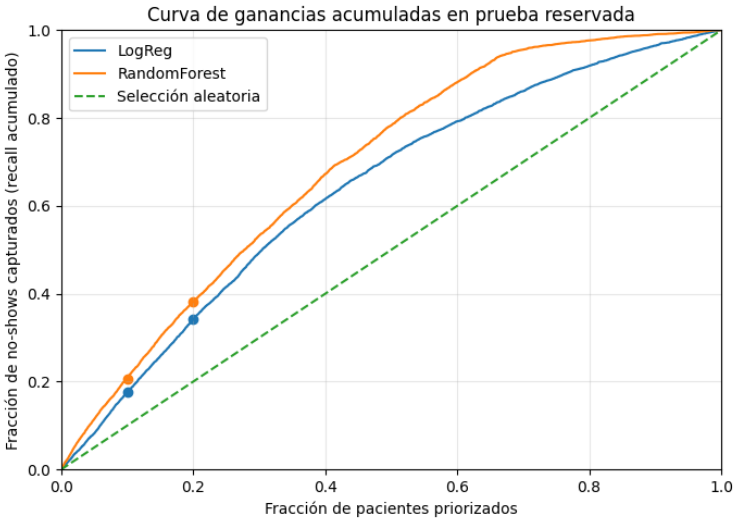

Figura 2: Curvas de ganancias acumuladas en prueba reservada. En el eje x se muestra la fracción priorizada (top-k por score) y en el eje y la fracción de *no-shows* capturados (recall acumulado). La diagonal representa selección aleatoria ($G(k) = k$).

entrenado solo en el subconjunto $D = 0$ y usando únicamente covariables pre-tratamiento. Para corregir el sesgo de selección inducido por quién cae en $D = 0$, se estima $e(X) = \Pr(D = 1 \mid X)$ y se repondera las observaciones con $D = 0$ mediante

$$w(X) = \frac{1}{1 - e(X)},$$

bajo el supuesto de ignorabilidad condicional en X y soporte común. El score de política es entonces $\widehat{p}_0(X)$.

Regla económica (umbral). Si el objetivo es decidir un contacto (SMS o SMS+llamada) con costo C y beneficio neto por cita recuperada b, y si el tratamiento reduce el riesgo en proporción θ, una regla operativa natural es:

$$\widehat{p}_0(X)\,\theta\,b \geq C \quad \Longleftrightarrow \quad \widehat{p}_0(X) \geq p^* = \frac{C}{b\theta}.$$

En este trabajo no identificamos causalmente θ con estos datos, para una etapa aplicada, θ debe venir de evidencia experimental o de una estrategia causal adicional.

Alternativa (variables instrumentales). Si existiera variación exógena en el envío de SMS (ex, shocks del proveedor, cambios discretos de política, reglas por cupos), podría utilizarse un instrumento Z para identificar un efecto local del tratamiento y así estimar θ relevante para política.

4.5. Score de política bajo asignación endógena de SMS

Dado que el envío de SMS (D) responde a decisiones operativas históricas, el score relevante para una política ex-ante es $\widehat{p}_0(X) \approx \Pr(Y = 1 \mid X, D = 0)$, es decir, el riesgo contrafactual sin intervención. Se estima $\widehat{p}_0$ entrenando modelos únicamente en el subconjunto $D = 0$ y reponderando por $w(X) = 1/(1 - \hat{e}(X))$, donde $\hat{e}(X) = \Pr(D = 1 \mid X)$ es el *propensity score*. Para estabilizar la

Cuadro 4: Desempeño del *score de política* $\widehat{p}_0(X)$ en prueba restringida a $D = 0$ (con pesos IPW capados al p99).

Modelo	**ROC-AUC**	**PR-AUC** (AP)	P@10 %	P@20 %
Policy-LogReg (p_0)	0.641	0.271	0.357	0.296
Policy-RF (p_0)	0.783	0.372	0.425	0.376

estimación ante solapamiento imperfecto (pesos extremos), se çapan"los pesos al percentil 99. La Tabla 4 resume el desempeño de estos modelos de política en la muestra de prueba restringida a $D = 0$.

Curva PR en $D = 0$. La Figura 3 muestra las curvas Precision–Recall en prueba restringida a $D = 0$. La línea base de precisión es la prevalencia en $D = 0$, $\Pr(Y = 1 \mid D = 0) \approx 0.164$. El Random Forest domina a la logística y obtiene mayor área bajo la curva: AP = 0.372 frente a AP = 0.271.

Ganancias acumuladas y focalización. Sea $\mathcal{T}_k$ el conjunto top-k por $\widehat{p}_0(X)$ (de tamaño $|\mathcal{T}_k| = \lceil kN_{D=0} \rceil$). Defina la ganancia acumulada:

$$G(k) = \frac{\sum_{i \in \mathcal{T}_k} \mathbf{1}\{Y_i = 1\}}{\sum_{i:D_i=0} \mathbf{1}\{Y_i = 1\}}.$$

Bajo selección aleatoria, $G(k) = k$. En nuestros resultados, para $k = 10\,\%$ la logística captura $G(0.10) = 0.217$ y el Random Forest $G(0.10) = 0.259$, para $k = 20\,\%$, las capturas son $G(0.20) = 0.361$ y $G(0.20) = 0.457$, respectivamente.

Con $N_{D=0} = 14{,}918$ y $\Pr(Y = 1 \mid D = 0) \approx 0.164$, esto implica que priorizar el top-10 % ($\approx 1{,}492$ pacientes) captura en torno a $0.217 \times 2{,}452 \approx 532$ *no-shows* con logística y $0.259 \times 2{,}452 \approx 635$ con Random Forest, frente a ≈ 245 bajo priorización aleatoria. En suma, el score $\widehat{p}_0(X)$ permite focalizar riesgo de manera consistente con una política ex-ante, y el Random Forest es sustancialmente superior para este propósito.

5. Discusión, limitaciones y trabajo futuro

5.1. Qué predicen los modelos (y qué no)

En este trabajo se estima *scores* predictivos con fines de priorización operativa. El modelo estándar entrenado sobre toda la muestra (mezcla de $D = 0$ y $D = 1$) aproxima la probabilidad de *no-show* bajo el histórico de asignación de SMS, es decir, una combinación de $p_0(X) = \Pr(Y = 1 \mid X, D = 0)$ y $p_1(X) = \Pr(Y = 1 \mid X, D = 1)$ ponderada por el *propensity score* $e(X) = \Pr(D = 1 \mid X)$. Su score es útil para anticipar ausentismo dado el proceso actual, pero no debe leerse como riesgo contrafactual bajo una política alternativa.
El *score de política* $\widehat{p}_0(X) \approx \Pr(Y = 1 \mid X, D = 0)$, en cambio, está diseñado para decisiones ex-ante: intenta aproximar el riesgo sin intervención (sin SMS) usando solo covariables pre-tratamiento y reponderación IPW en el subconjunto $D = 0$. En ningún caso identificamos el efecto causal del SMS, la asociación bruta entre `sms_received` y *no-show* refleja targeting operativo y mezcla temporal (correlación con `wait_days`), no necesariamente impacto del mensaje.

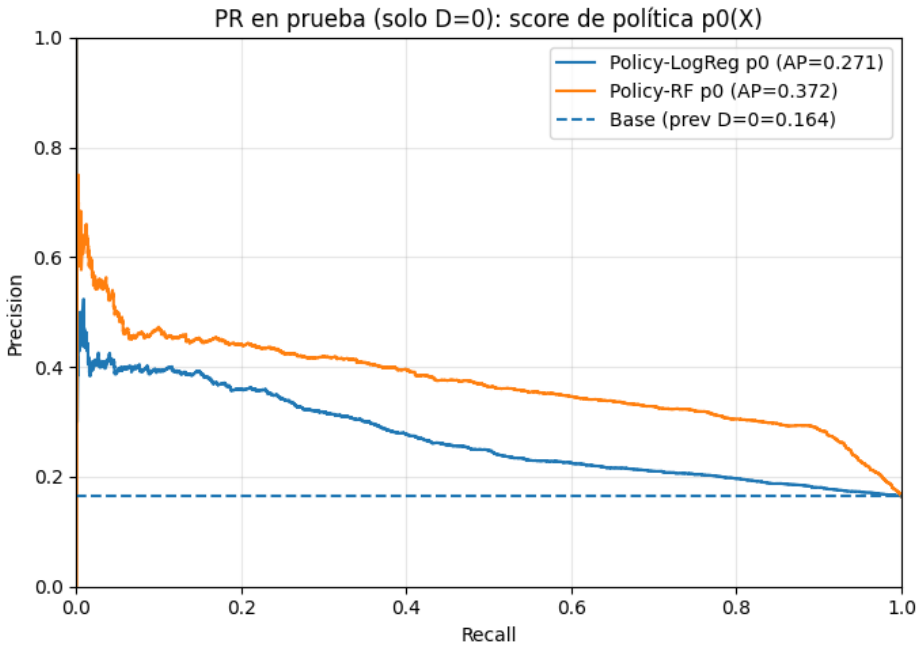

Figura 3: Curvas Precision–Recall en prueba restringida a $D = 0$ (no se envió SMS). Scores de política $\widehat{p}_0(X) = \Pr(Y = 1 \mid X, D = 0)$. La línea horizontal punteada corresponde a la prevalencia en $D = 0$: $\Pr(Y = 1 \mid D = 0) \approx 0.164$.

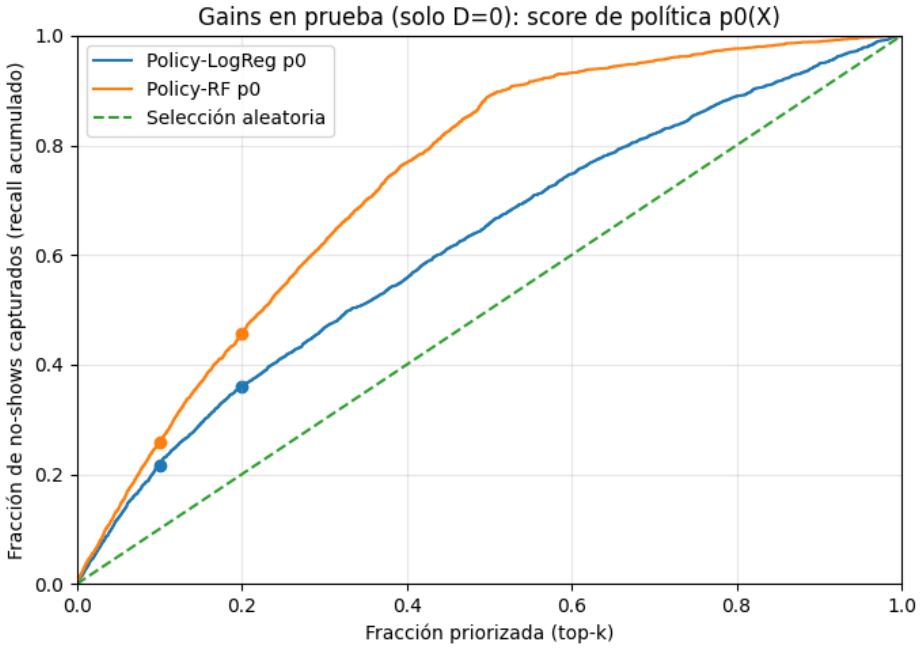

Figura 4: Curvas de ganancias acumuladas en prueba restringida a $D = 0$. En el eje x se muestra la fracción priorizada (top-k) por $\widehat{p}_0(X)$ y en el eje y la fracción de *no-shows* capturados (recall acumulado). La diagonal representa selección aleatoria ($G(k) = k$).

5.2. Lectura económica y operativa

Para un problema desbalanceado ($\Pr(Y = 1) \approx 0.20$), los resultados muestran que un modelo no lineal (Random Forest) concentra más riesgo que una base lineal logística, tanto en métricas globales (ROC-AUC, PR-AUC) como en métricas de priorización (P@k, ganancias acumuladas $G(k)$). Si la institución solo puede intervenir sobre una fracción k de pacientes, el ranking por score permite capturar una fracción de *no-shows* claramente superior a la selección aleatoria. Operativamente, esto se traduce en focalizar recordatorios, reprogramaciones proactivas o confirmaciones telefónicas en el top-k de riesgo, respetando restricciones de capacidad.

En el módulo de política, el score $\widehat{p}_0(X)$ mantiene esta capacidad de concentración dentro del subconjunto $D = 0$, lo que sugiere que, aun corrigiendo por asignación endógena, es posible construir reglas de priorización ex-ante consistentes con riesgo sin intervención.

5.3. Limitaciones principales

Este es un ejercicio predictivo, con varias limitaciones relevantes:
Asignación endógena de SMS: `sms_received` refleja decisiones operativas históricas, no asignación aleatoria. La construcción de $\widehat{p}_0(X)$ descansa en supuestos de ignorabilidad condicional y soporte común, si existen determinantes no observados (ex, severidad clínica, historial de adherencia, restricciones de transporte) que afectan a la vez D y Y, el score de política puede estar sesgado.
Soporte común imperfecto (overlap) y pesos extremos: las distribuciones de $e(X)$ para $D = 0$ y $D = 1$ difieren de forma apreciable, generando pesos IPW muy grandes en algunas regiones de X. Sin capping, el tamaño efectivo de muestra (ESS) para $D = 0$ es reducido, lo que aumenta varianza. Capar pesos (p95, p99 o valor fijo) mejora el ESS a costa de introducir algo de sesgo, por eso reporta diagnósticos de overlap, distribución de pesos y sensibilidad del desempeño a la regla de capping.
Calibración de probabilidades: aunque la calidad de ranking es buena (particularmente en Random Forest), las curvas de calibración muestran que los modelos tienden a sobreestimar el riesgo en los deciles altos. Para reglas basadas en umbrales monetarios p^*, la calibración se vuelve central, aquí solo se da un primer diagnóstico con curvas de confiabilidad y Brier score.
Validez externa y dataset shift: los datos provienen de una sola ciudad y de un periodo concreto, y el split 80/20 es aleatorio. Cambios en la agenda, en la política de recordatorios o en la composición de pacientes pueden alterar la relación entre covariables y ausentismo. En despliegue real se requiere monitoreo de drift y, posiblemente, reentrenamiento periódico.

5.4. Trabajo futuro

A partir de estos resultados, se ven varias extensiones:
Calibración post-estimación: aplicar métodos isotónicos o de Platt en un conjunto de validación (tanto para el modelo estándar como para el score de política) y evaluar métricas como Brier y Expected Calibration Error (ECE), además de P@k (fue literatura que encontré pero no entendí bien).
Evaluación temporal y robustez a drift: reemplazar el split aleatorio por un split temporal (entrenar en meses iniciales y evaluar en meses posteriores) y monitorear la estabilidad de AUC, AP y P@k a lo largo del tiempo.
De scores a decisiones: combinar el score $\widehat{p}_0(X)$ con información sobre costos C, beneficios b y una estimación causal del efecto θ de SMS/llamada (idealmente proveniente de un RCT, una discontinuidad de política o un instrumento válido) para derivar reglas de intervención basadas en net benefit, por ejemplo vía Decision Curve Analysis.

Modelos adicionales y comparación: explorar variantes gradiente (GBDT, XGBoost/LightGBM) y otros modelos tabulares competitivos, manteniendo el mismo protocolo de evaluación fuera de muestra y los mismos diagnósticos de overlap, pesos y calibración.

5.5. Reproducibilidad

Todos los modelos se implementan en `scikit-learn` mediante Pipeline, con semilla fija (random_state=42) y preprocesamiento (*one-hot* con `handle_unknown=ïgnore"`) ajustado exclusivamente sobre entrenamiento en cada pliegue de validación cruzada, evitando data leakage. El código proporciona los scripts necesarios para regenerar tablas, figuras y experimentos, de modo que los resultados puedan auditarse y extenderse sobre una base reproducible.

6. Conclusiones

Usando un conjunto público de 110,526 citas ambulatorias en Brasil (prevalencia de *no-show* $\approx 20.2\,\%$), construimos un pipeline reproducible con regresión logística y Random Forest para predecir ausentismo. En evaluación fuera de muestra, el Random Forest domina consistentemente en métricas globales (AUC y AP) y en métricas operativas de focalización (P@k y ganancias acumuladas), lo que indica que captura no linealidades e interacciones relevantes (calendario, espera, perfil del paciente) que un modelo lineal no recoge completamente.

Adicionalmente, se abroda el problema operativo de endogeneidad en el envío histórico de SMS separando un score descriptivo (política histórica) de un score de política ex-ante $\widehat{p}_0(X) = \Pr(Y = 1 \mid X, D = 0)$, entrenado en $D = 0$ y reponderado con IPW. Los diagnósticos de overlap, distribución de pesos y ESS muestran que la reponderación puede ser inestable sin estabilización, al capar pesos (p95/p99) obtenemos desempeños sustantivamente mejores y robustos. En conjunto, los resultados sugieren que un enfoque predictivo simple puede ser un insumo útil para priorización bajo restricciones de capacidad, mientras que la traducción del score a una regla óptima de intervención requiere un componente causal adicional para identificar la efectividad del recordatorio.

A. Exploración descriptiva adicional (EDA)

Esta sección recopila las figuras completas de EDA que sustentan la discusión de la Sección A. El objetivo es documentar la estructura básica de las variables (forma de las distribuciones, colinealidad y relaciones bivariadas) para facilitar la auditoría del pipeline.

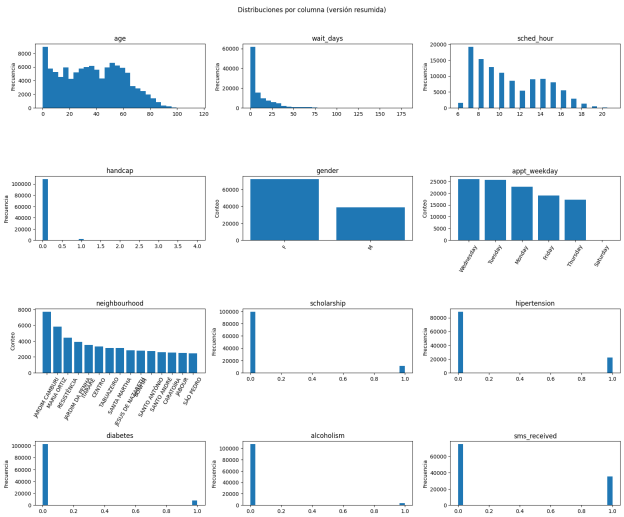

Figura 5: Distribuciones marginales de las variables sobre la muestra depurada.

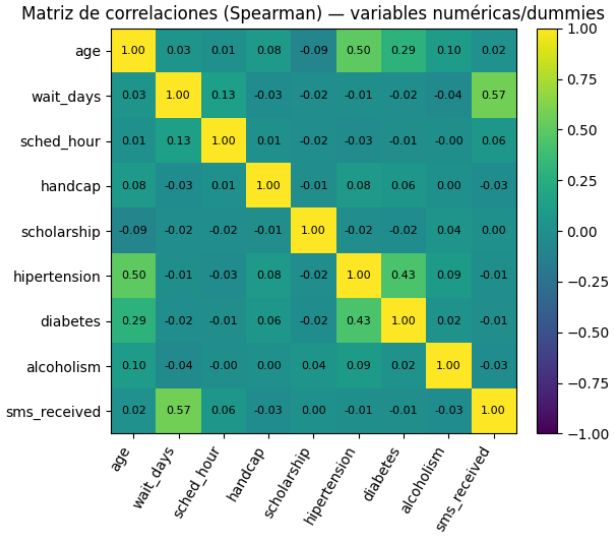

Figura 6: Correlaciones de Spearman en variables numéricas y dummies. Se observa baja colinealidad global, con asociaciones destacadas `age`–`hipertension`, `age`–`diabetes` y `wait_days`–`sms_received`.

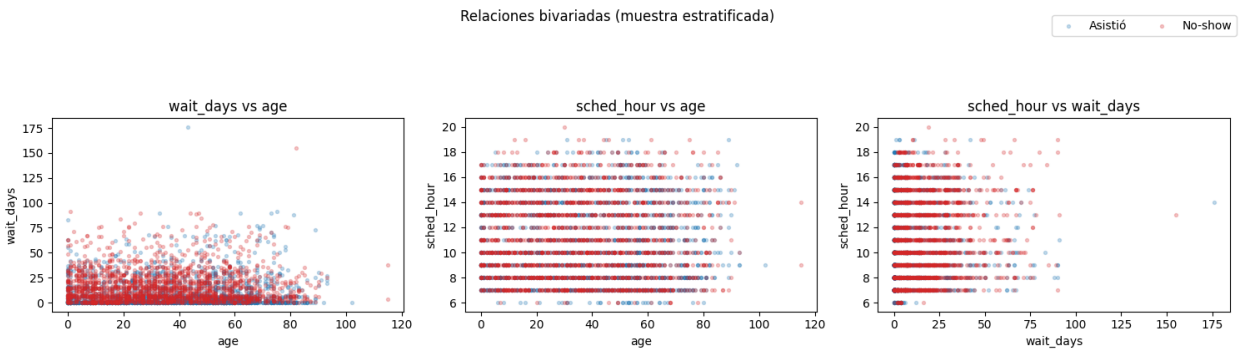

Figura 7: Relaciones bivariadas sobre submuestra estratificada (azul: asistió, rojo: *no-show*). Las nubes confirman la ausencia de fronteras lineales claras, especialmente en `wait_days` y `sched_hour`, y motivan el uso de modelos no lineales.

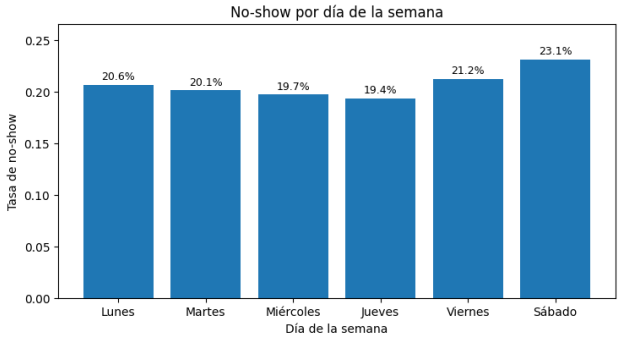

Figura 8: Tasa de no-show por día de la semana (lunes–sábado). El ausentismo varía entre ∼19.4 % y 23.1 %, con máximo en sábado.

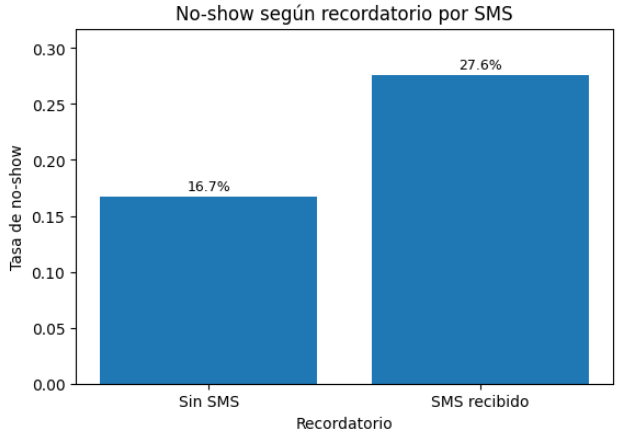

Figura 9: Tasa de no-show según recordatorio por SMS (bruta, observacional). La mayor tasa en el grupo con SMS refleja *targeting* y mezcla temporal, no necesariamente un efecto causal del mensaje.

B. Diagnósticos de solapamiento e IPW para el score de política

Nota sobre fuentes y uso de IA. Los conceptos de *propensity score*, tamaño efectivo de muestra (ESS) y winsorización de pesos IPW que se utilizan en este apéndice provienen de la literatura estándar de inferencia causal y no se trabajaron en detalle en el curso. Decidí incorporarlos como un ejercicio exploratorio para entender mejor las limitaciones de la reponderación y documentar diagnósticos básicos de solapamiento y estabilidad de los pesos.
Para estudiar estos temas y redactar la sección utilicé un asistente de inteligencia artificial (Gemini gratis) como apoyo en el código, en la explicación conceptual y en la propuesta de los gráficos/tablas de robustez. El procesamiento de los datos, la elección de los escenarios de sensibilidad (sin cap, p95, p99 y cap fijo) sí fueron realizados por mí. Esta sección debe entenderse como material complementario que estoy aprendiendo a usar, más que como parte central evaluada del curso:
Definimos el *propensity score* $\hat{e}(X) = \Pr(D = 1 \mid X)$ y los pesos para $D = 0$ como

$$w(X) = \frac{1}{1 - \hat{e}(X)}.$$

El tamaño efectivo de muestra (ESS) de los no tratados ponderados se calcula como

$$\text{ESS} = \frac{\left(\sum_{i:D_i=0} w_i\right)^2}{\sum_{i:D_i=0} w_i^2}.$$

En nuestra base, el subsample de entrenamiento con $D = 0$ tiene $N_{D=0}^{\text{train}} = 60{,}126$ observaciones. Sin winsorización, los pesos presentan colas muy pesadas: el ESS cae a ESS ≈ 920 (alrededor del 1.5 % de la muestra efectiva), lo que implica alta varianza. Al capar en el percentil 99, $w_i^{\text{cap}} = \text{mín}\{w_i, q_{0.99}(w)\}$, el ESS aumenta a $\text{ESS}^{\text{cap}} \approx 10{,}485$ (17 % de la muestra), con pérdida de información acotada y mejoras sustanciales en estabilidad.

C. Sensibilidad al cap de pesos IPW

Para cuantificar el *trade-off* sesgo–varianza, repetimos el entrenamiento del score de política variando la regla de winsorización de pesos (sin cap, p95, p99 y cap fijo en 20). La Tabla 5 reporta el desempeño en prueba restringida a $D = 0$.

Cuadro 5: Sensibilidad del score de política a la winsorización de pesos IPW (prueba restringida a $D = 0$).

Modelo	**Cap**	**ROC-AUC**	**PR-AUC**	P@10 %	P@20 %
Policy-LogReg	none	0.554	0.194	0.227	0.207
Policy-LogReg	p95	0.688	0.299	0.384	0.341
Policy-LogReg	p99	0.641	0.271	0.357	0.296
Policy-LogReg	20	0.664	0.288	0.383	0.322
Policy-RF	none	0.778	0.361	0.418	0.369
Policy-RF	p95	0.785	0.377	0.425	0.378
Policy-RF	p99	0.783	0.372	0.425	0.376
Policy-RF	20	0.784	0.377	0.424	0.383

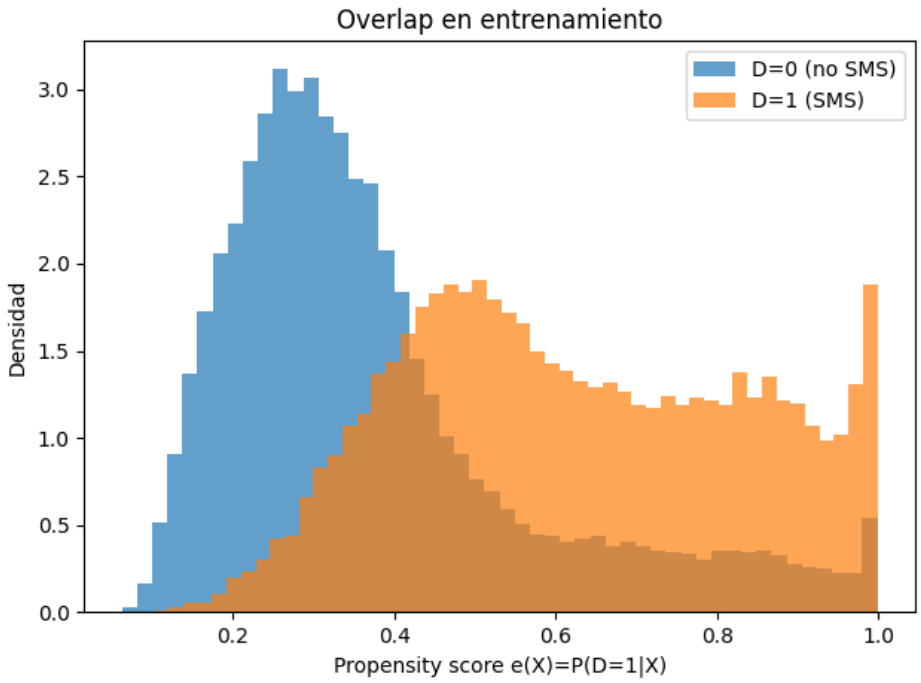

Figura 10: Distribución del propensity score $\hat{e}(X)$ en entrenamiento por grupo D (diagnóstico de soporte común/overlap). El grupo $D = 0$ se concentra en valores intermedios (~ 0.2–0.4) mientras que $D = 1$ acumula probabilidad en rangos altos ($\hat{e}(X) \gtrsim 0.5$), con zona de solapamiento razonable en torno a 0.3–0.6 pero masas no triviales cerca de los extremos.

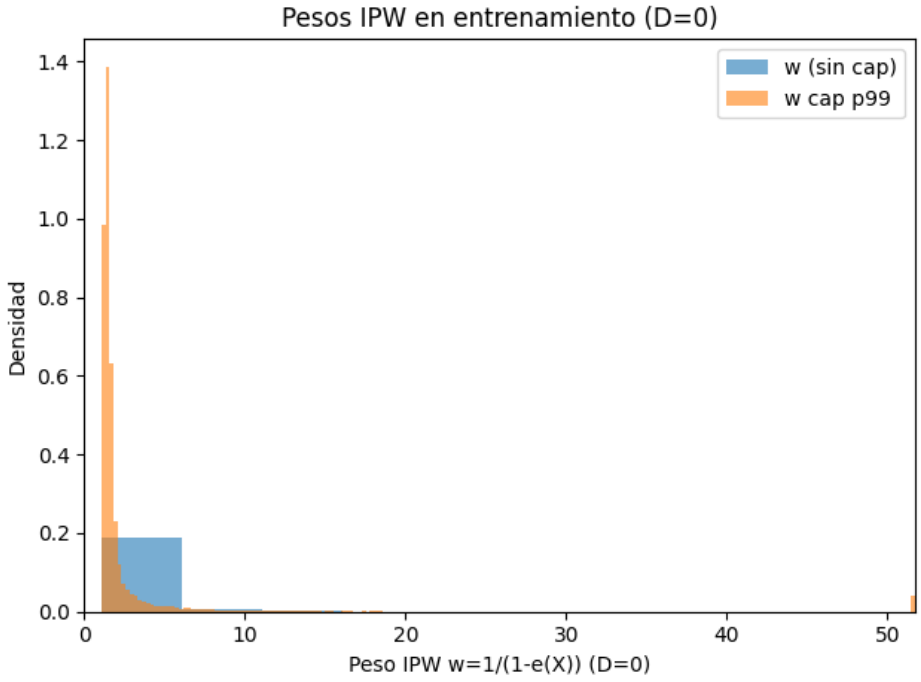

Figura 11: Distribución de pesos IPW $w(X)$ para $D = 0$ (sin cap vs. cap p99). Sin cap aparecen pesos extremos (hasta valores superiores a 50), responsables del ESS bajo. El cap p99 recorta la cola derecha y concentra la masa en el rango $[1, 10]$, aumentando el ESS y estabilizando la estimación.

Dado que la prevalencia en $D = 0$ es $\Pr(Y = 1 \mid D = 0) \approx 0.164$, las precisiones del Random Forest en el top-10 % (0.418–0.425) implican *lifts* cercanos a 2.6 veces la línea base, prácticamente invariantes al cap. En cambio, para la logística el desempeño sin cap es pobre (AUC 0.554, AP 0.194) y mejora notablemente con cualquier forma de winsorización, con valores algo superiores para p95 y cap 20. En el cuerpo del artículo tomamos p99 como especificación de referencia por su buen compromiso entre ESS y desempeño, y dejamos el resto de reglas como robustez adicional.

D. Calibración (curva de confiabilidad y Brier score)

La calibración evalúa si las probabilidades predichas $\hat{s}(x)$ corresponden a frecuencias empíricas. En las Figuras 12 y 13 reportamos curvas de confiabilidad por *bins* para los modelos principales (logística y RF en el conjunto de prueba completo) y para los modelos de política (Policy-LogReg y Policy-RF en prueba restringida a $D = 0$).

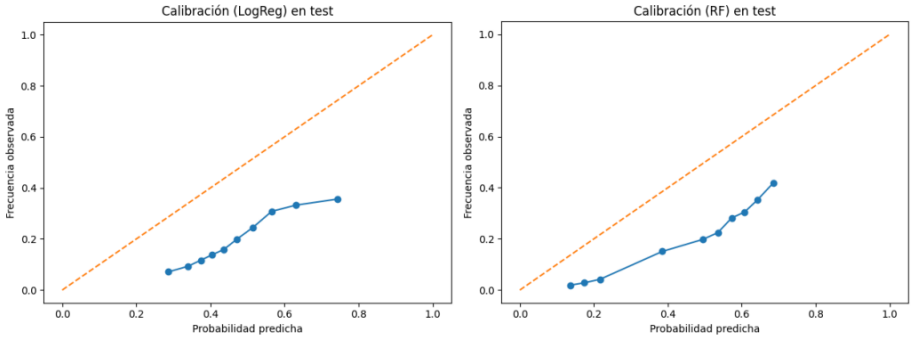

Figura 12: Curvas de confiabilidad en prueba (modelos principales logística y RF). En ambos casos los puntos se sitúan por debajo de la diagonal, indicando cierta sobreestimación del riesgo, algo menor en el RF.

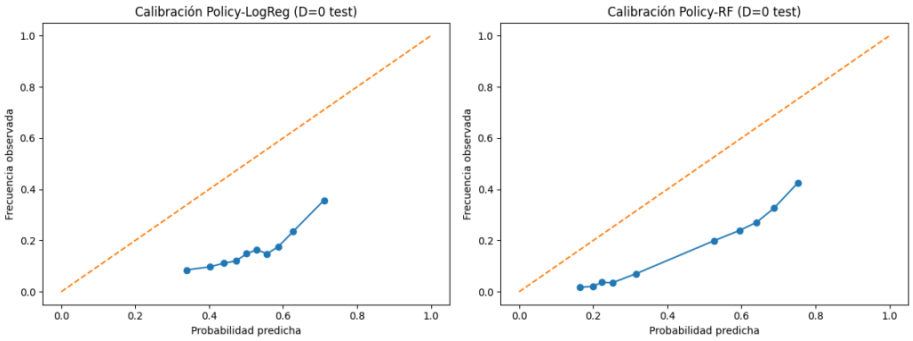

Figura 13: Curvas de confiabilidad en prueba restringida a $D = 0$ (Policy-LogReg y Policy-RF). El modelo de política basado en RF está mejor calibrado que el logístico, aunque ambos tienden a sobreestimar el riesgo en los deciles altos.

Para los modelos de política, el Brier score (error cuadrático medio de probabilidad) es Brier(Policy-LogReg) $\approx$ 0.258 y Brier(Policy-RF) ≈ 0.198, coherente con la mejor combinación de ranking y calibración del RF. En conjunto, los diagnósticos sugieren que los scores son útiles como herramientas de ranking (P@k, ganancias), pero que una etapa de calibración post-hoc (Platt o isotónica) sería recomendable

si se desea interpretar las probabilidades de forma absoluta o aplicar reglas basadas en umbrales monetarios p^*.

E. Interpretabilidad: importancia de variables (permutation importance)

Como diagnóstico de interpretabilidad global, calculamos *permutation importance* en prueba para el modelo de política (Policy-RF), medida como la caída media en AP al permutar cada covariable. La Figura 14 muestra el ranking visual y la Tabla 6 reporta las caídas medias y su desviación estándar.

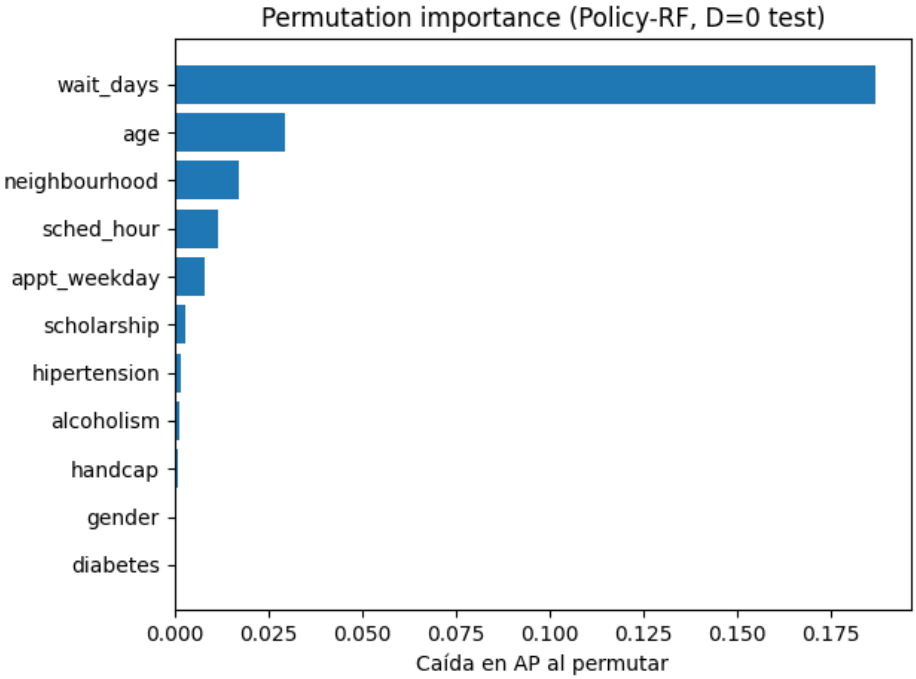

Figura 14: Permutation importance (principales covariables) para el score de política (Policy-RF) en $D = 0$ test, medida como caída en AP al permutar cada variable.

La importancia dominante de `wait_days` es consistente con la intuición clínica y con la EDA: la espera prolongada aumenta el riesgo de ausentismo. Le siguen `age` y `neighbourhood`, lo que sugiere heterogeneidad por ciclo de vida y por contexto geográfico. Día y hora de la cita aportan señal adicional pero de menor magnitud, mientras que las variables clínicas y demográficas básicas (`gender`, `diabetes`, `handcap`) tienen contribuciones marginales a la capacidad predictiva del modelo de política.

F. Robustez temporal (si se dispone de fechas)

Finalmente, un chequeo más exigente de generalización consiste en introducir una partición temporal usando `AppointmentDay`: entrenar en meses iniciales y evaluar en meses posteriores. Esto reduce el riesgo de optimismo asociado a un split aleatorio donde tren y prueba comparten la misma mezcla temporal. En un despliegue real, este tipo de evaluación temporal, combinada con monitoreo en

Cuadro 6: Permutation importance para Policy-RF en prueba restringida a $D = 0$.

Variable	Caída media en AP	Desv. estándar
`wait_days`	0.187	0.003
`age`	0.030	0.004
`neighbourhood`	0.017	0.003
`sched_hour`	0.012	0.002
`appt_weekday`	0.008	0.003
`scholarship`	0.003	0.001
`hipertension`	0.002	0.002
`alcoholism`	0.001	0.000
`handcap`	0.001	0.001
`gender`	0.000	0.001
`diabetes`	0.000	0.000

línea de métricas como AUC, AP y P@k, es clave para detectar *drift* y decidir cuándo reentrenar el modelo.

Referencias

[1] J. Aroba. *Medical Appointment No Shows (Kaggle dataset)*. Disponible en: `https://www.kaggle.com/datasets/joniarroba/noshowappointments`.

[2] A. J. Vickers y E. B. Elkin. *Decision curve analysis: A novel method for evaluating prediction models. Medical Decision Making*, 26(6):565–574, 2006.

[3] I. Gurol-Urganci, T. de Jongh, V. Vodopivec-Jamsek, R. Atun y J. Car. *Mobile phone messaging reminders for attendance at healthcare appointments. Cochrane Database of Systematic Reviews*, (12):CD007458, 2013. doi: 10.1002/14651858.CD007458.pub3.

[4] P. E. Hasvold y R. Wootton. *Use of telephone and text-message reminders to improve attendance at hospital appointments: A systematic review. Journal of Telemedicine and Telecare*, 17(7):358–364, 2011.

[5] D. Liu et al. *Machine learning approaches to predicting medical appointment* no-shows *in pediatric care. npj Digital Medicine*, 5:48, 2022.

[6] L. H. A. Salazar et al. *No-show in medical appointments with machine learning techniques: A systematic literature review. Information*, 13(11):507, 2022.