Trabajo Práctico N°3 - Ciencia de Datos

Clasificación de pobres en la EPH

Grupo 10

Francisca Cetra, Mariano Ripoll y Justina Rivero Ayerza

Profesora Magistral

María Noelia Romero

Profesor Tutorial

Tomas Enrique Buscaglia

Semestre y año de presentación

2do Semestre 2025

Link al repositorio de Github:

https://github.com/JustinaRiveroA/Ciencia_Datos_TP3_Grupo10

A. Enfoque de validación

1. y 2.

Al estratificar el split, la prevalencia de pobre=1 se mantiene igual en train y test (train: 0.4137; test: 0.4140), aproximadamente un 41%, por lo que las métricas de evaluación no quedan distorsionadas por diferencias en la mezcla de clases. Trabajamos solamente con los datos del año 2025.

Tabla 1. Diferencias de medias entre entrenamiento y testeo (datos 2025)

Split	Test (30%)	Train (70%)	Diferencia (Train - Test)	P-valor (Test T)
Educación	9.932	10.189	0.256	0.104
Edad	37.629	37.236	-0.393	0.606
Edad 2	1951.602	1877.964	-73.638	0.255
Sexo	1.519	1.526	0.007	0.669
Es jefe[1]	0.370	0.396	0.026	0.102
Categoría ocupacional	1.227	1.342	0.115	0.014
Horas trabajadas	14.168	15.127	0.959	0.167
Adultos equivalentes	3.444	3.452	0.009	0.877
Ratio Dependencia	0.690	0.667	-0.023	0.324
Ocupados Share	0.439	0.453	0.014	0.175
Educación jefe	11.650	11.568	-0.082	0.483
Jefe mujer[2]	0.429	0.453	0.024	0.141
Estado civil: Separado/ Divorciado	0.067	0.071	0.004	0.622
Estado civil: Soltero	0.483	0.477	-0.006	0.713
Estado civil: Unido	0.176	0.201	0.025	0.056
Estado civil: Viudo	0.061	0.050	-0.011	0.157
Cobertura médica: No Paga	0.271	0.292	0.021	0.157

[1] Variable que toma valor 1 si es jefe de hogar, 0 si no lo es.
[2] Variable que vale 1 si el jefe de hogar es mujer, 0 si no lo es.

Cobertura médica: Ns/Nr	0.002	0.001	-0.001	0.506
Cobertura médica: Obra Social	0.642	0.633	0.009	0.586
Cobertura médica: Prepaga/ Mutual	0.061	0.058	-0.003	0.739
Cobertura médica: Publico	0.000	0.001	0.001	0.083
Estado: Inactivo	0.422	0.388	-0.033	0.042
Estado: Menor de 10	0.114	0.109	-0.005	0.639
Estado: Ocupado	0.426	0.459	0.032	0.051

Las diferencias de medias entre train y test son en general muy pequeñas y no significativas. El split quedó bien balanceado: solo se observan señales puntuales en la variable de Categoría Ocupacional (diferencia de aproximadamente 0.115, $p \approx 0.014$) y en la variable categórica Estado, el valor de "Inactivo" muestra una diferencia de $\approx -0{,}033$ ($p\approx 0.042$), el valor de "Ocupado" (diferencia de aproximadamente 0.032, $p\approx 0.051$) y Estado Civil: Unido ($\Delta\approx 0.025$, $p \approx 0.056$). La magnitud de estos desvíos es baja y no altera la composición de covariables, por lo que no esperamos sesgos por la partición.

En cuanto a la matriz X, incluimos 24 variables sin ingresos ni identificadores (evita leakage) que capturan capital humano, demografía/ciclo de vida, inserción laboral y estructura del hogar: variable de educación, edad y edad al cuadrado (efecto no lineal de edad), Sexo, una variable binaria que indica si es Jefe de hogar o no, otra variable categórica que indica el Estado civil (Soltero, Unido, Separado/Divorciado, Viudo), la Cobertura de salud médica (Obra Social, Prepaga/Mutual, Público, No paga, NsNr), otra variable que indica si el individuo es Ocupado, Inactivo, o Menor de 10 años. También incluimos una variable de Categoría Ocupacional, las Horas trabajadas Asimismo, las variables "Adultos equivalentes", "Ratio Dependencia" y "Ocupados Share" fueron creadas con el objetivo de actuar como variables de "Estructura y Composición del Hogar"; en sí son tres indicadores agregados. Primero, la variable Adultos equivalentes, mide las necesidades económicas del hogar ajustadas por su composición. Se calcula sumando los factores de "adulto equivalente" de cada miembro, asignados según su edad (CH06) y sexo (CH04), permitiendo así una comparación más precisa que el simple conteo de personas. En segundo lugar, se creó Ratio Dependencia, que refleja la carga económica que soportan los miembros en edad productiva con respecto a quienes tenderían a ser más

dependientes e improductivos por edad. Este ratio se calcula como el cociente entre las personas dependientes (menores de 15 y mayores de 65) y las personas en edad de trabajar (según CH06), proveyendo un indicador claro de la presión sobre los generadores de ingreso. Finalmente, la variable Ocupados Share mide la proporción de miembros del hogar que se encuentran efectivamente trabajando (identificados a través de la variable ESTADO), lo que funciona como un proxy directo de la capacidad del hogar para generar ingresos. Por último, una variable que indica el nivel educativo del jefe de hogar y otra binaria que toma el valor de 1 si el jefe es mujer y 0 si no lo es. El conjunto de estas variables resume habilidades y recursos de las personas, su capacidad de generar ingresos (inserción e intensidad laboral), necesidades y composición del hogar. Son todas variables que teóricamente pueden justificarse como relevantes para entender o explicar la pobreza.

B. Modelo de regresión logística

3.

Tabla 2. Modelo de Logit reducido. Coeficientes, Errores Estándar, Odds-Ratios (con datos de gente que respondió ITF en 2025).

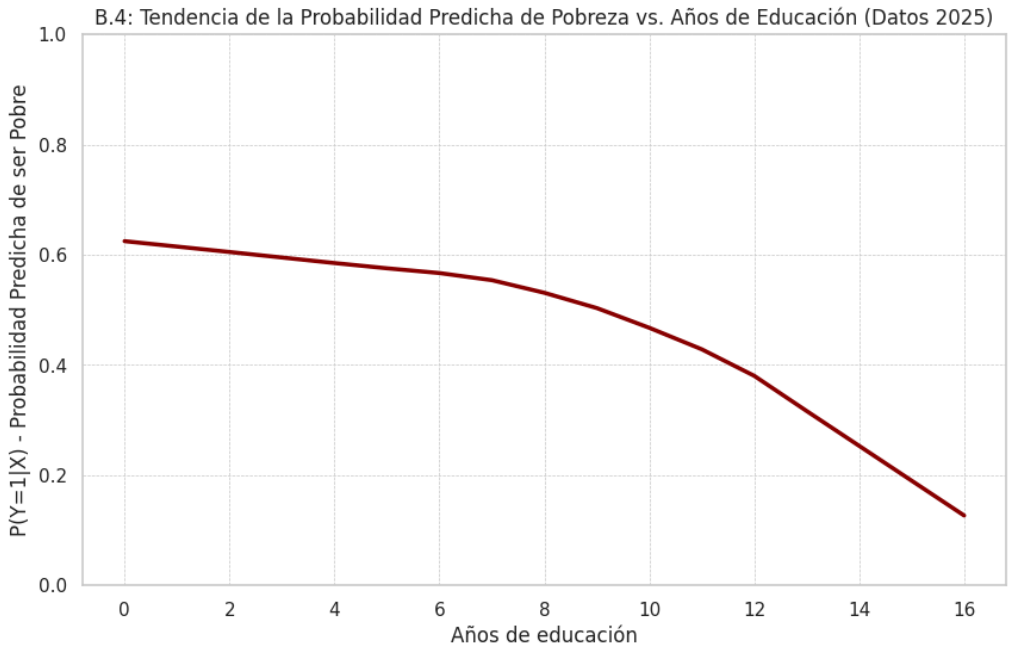
	Coeficiente (Log-Odds)	Error Estándar (Std.Err.)	Odds-Ratio (OR)	P-valor (P>|z|)
Intercept	=+2.2568	0.5499	9.5529	0.000
Educación	-0.0696	0.0187	0.9328	0.000
Edad 2	-0.0002	0.0000	0.9998	0.000
Sexo	=+0.2406	0.1038	1.2720	0.021
Es jefe	=+0.4885	0.1304	1.6299	0.000
Categoría ocupacional	-0.3452	0.1258	0.7081	0.006
Adultos equivalentes	=+0.1970	0.0377	1.2177	0.000
ocupados_share	-4.0267	0.2495	0.0178	0.000
Educación jefe	-0.1167	0.0175	0.8898	0.000
Jefe mujer	=+0.6262	0.1048	1.8705	0.000
Estado Civil: Viudo	-1.1282	0.2600	0.3236	0.000
Cobertura médica: No Paga	=+2.4453	0.2547	11.5342	0.000

Cobertura médica: Obra Social	=+0.7095	0.2401	2.0329	0.003
Cobertura médica: Publico	=+2.6165	1.2643	13.6884	0.038
Estado: Inactivo	-1.2112	0.3625	0.2978	0.001
Estado: Menor de 10	-1.7505	0.4346	0.1737	0.000

Estimamos un modelo Logit sobre el set de entrenamiento (2025) con selección backward. En la Tabla 2 reportamos coeficientes en log-odds, errores estándar, odds-ratios (OR) y p-valores. Los signos y los OR confirman intuiciones: más educación (individuo y jefe) reduce la probabilidad de pobreza (OR<1), mayor proporción de ocupados en el hogar reduce fuertemente la pobreza (OR≈0.02), y hogares con jefa mujer presentan mayor riesgo (OR≈1.87). Las variables no significativas fueron eliminadas.

4.

Gráfico 1. Visualización $\widehat{P}(Y = 1|X)$ y variable de Educación (en años).



El gráfico muestra la relación entre los años de educación (eje X) y la probabilidad predicha de ser pobre (eje Y). Se puede observar una relación monótonamente decreciente: a medida que suben los años de educación, la probabilidad predicha de ser pobre cae. La curva (LOWESS) sugiere además no linealidad: entre 0–10 años la pendiente es suave; alrededor de secundario completo (~12 años) empieza a caer más rápido y el descenso se

acentúa entre 12–16 años de educación (que puede ser el ingreso a, o la terminación de terciario-universitario). Esto es consistente con el modelo: el Odds-Ratio de Educación < 1 ves decir, cada año adicional de educación reduce la probabilidad en ~0,9 p.p. (ceteris paribus). Elegimos usar la variable Educación porque resume capital humano y es un determinante central de ingresos y posibilidad de tener un empleo. Además, su interpretación es directa y la curva deja ver los quiebres en los niveles de finalización de cada año. El gráfico es una aproximación de dependencia parcial basada en las predicciones del Logit. Ilustra el patrón promedio manteniendo el resto de covariables según se observan en la muestra.

C. Método de vecinos cercanos (KNN)

5.

Tabla 3. Métricas, KNN K=1

	Accuracy	Precision	Recall	F1-Score
Train	0.9950	0.9951	0.9927	0.9939
Test	0.7339	0.6721	0.6974	0.6845

Con K=1 el rendimiento en train es casi perfecto (Accuracy≈0.995; F1≈0.994) pero cae fuerte en test (Accuracy≈0.734; F1≈0.685): evidencia clara de overfitting (varianza alta, sesgo bajo).

Tabla 4. Métricas KNN, K=5

	Accuracy	Precision	Recall	F1-Score
Train	0.8428	0.8154	0.8016	0.8085
Test	0.7471	0.7018	0.6767	0.6890

Al pasar a K=5, disminuye la varianza: las métricas de train y test se acercan (Accuracy test≈0.747; F1 test≈0.689).

Tabla 5. Métricas KNN, K=10

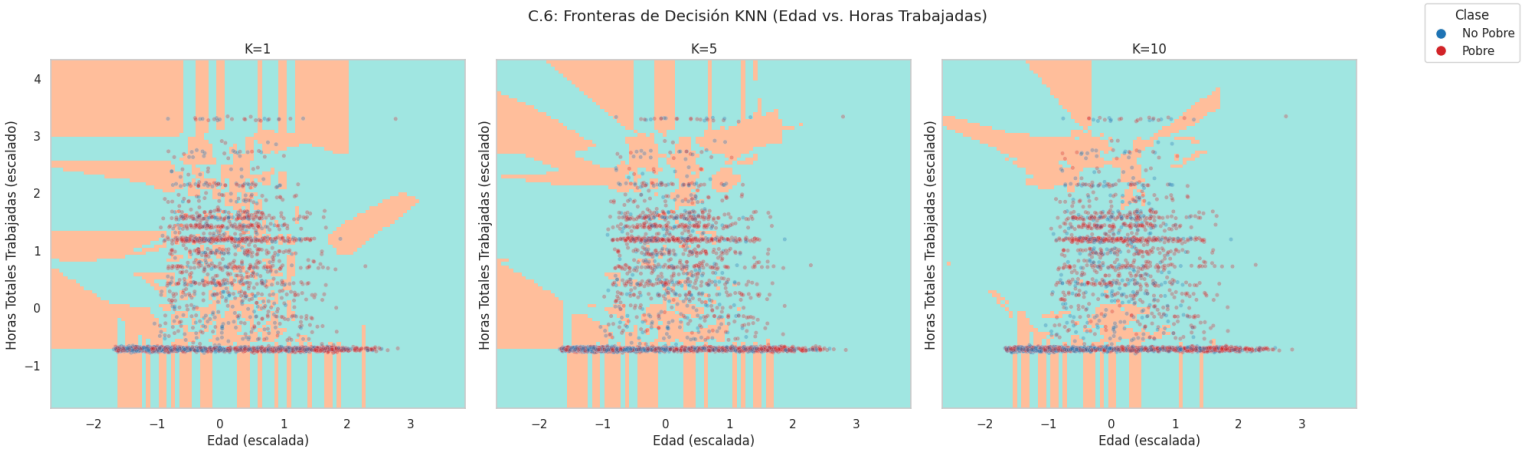
	Accuracy	Precision	Recall	F1-Score
Train	0.8075	0.8178	0.6879	0.7473
Test	0.7595	0.7611	0.6109	0.6778

Con K=10 el modelo se suaviza aún más: Accuracy test sube levemente (≈0.760) pero F1 test baja un poco (≈0.678), lo que sugiere más sesgo y menor varianza.

Vemos el trade-off sesgo–varianza: K chico sobre-ajusta; K más grande generaliza mejor pero puede perder detalle en la frontera de decisión. En nuestros datos, entre K=5 y K=10 hay un compromiso razonable (mejor Accuracy con K=10, mejor F1 con K=5).

6.

Gráfico 2. Visualización fronteras de decisión de Vecinos Cercanos (KNN).



El Gráfico 2 muestra las fronteras de decisión en el plano Edad (“CH06”) vs Horas trabajadas (“PP3E_TOT”). Con ambas variables estandarizadas para K=1, K= 5 y K=10.

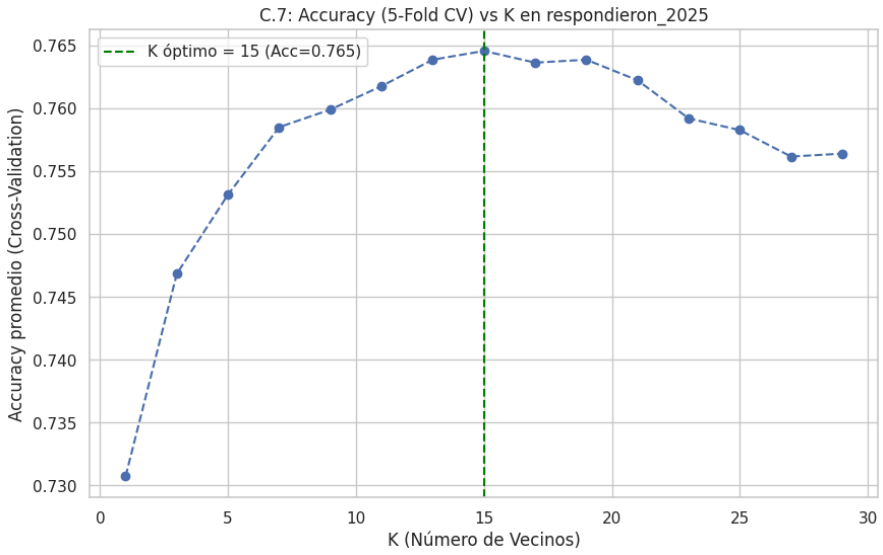
Con K=1 la frontera es muy dentada y fragmentada (sensibilidad extrema al ruido local). Con K=5 y K=10 las regiones “Pobre/No Pobre” se vuelven (un poco) más suaves y estables, coherente con la reducción de varianza al aumentar K. El patrón general indica que, manteniendo horas trabajadas bajas y edades extremas, aumenta la probabilidad de ser pobre, mientras que más horas trabajadas tienden a asociarse con “No Pobre”; la suavización de la frontera con K altos refleja esa relación promedio. Igualmente, esos 3 gráficos no son muy claros para observar la frontera de decisión de KNN. Los resultados no son muy interpretables debido a la distribución no “relativamente homogenea” ni genuinamente continua de valores (solo enteros) que toman las variables, incluso tenerlos “scaled” no ayuda.

Además respecto a la distribución no homogénea, como puede verse, forzamos a cero varios NaNs de horas trabajadas dado que los valores nulos en esta variable no son errores, sino una consecuencia estructural del estado no laboral de los individuos (inactivos, desocupados, estudiantes, etc), se procedió a su imputación forzosa a cero. Este enfoque es

lógicamente coherente, ya que representa con precisión la cantidad de horas trabajadas por quienes no tienen una ocupación. Fundamentalmente, esta decisión evitó la eliminación masiva de filas que habría ocurrido con otros métodos, en sí previniendo así un severo sesgo de selección y garantizando que las predicciones del modelo sean generalizables a toda la población en estudio, en lugar de limitarse a un subgrupo sesgado de individuos ocupados. Además, permite esto evitar achicar la muestra de norespondieron2025, teniendo una muestra más amplia sobre la que predecir y permitiendo que los datos de la variable sean mejores predictores que si la muestra estuviera sesgada y no tuviera tantos ceros; sesgo que creemos es peor que el que puede llegar a introducir imputar a cero algunos pocos casos de personas que genuinamente no quisieran responder.

7.

Gráfico 3. K-óptimo con Cross Validation



Hicimos 5-Fold Cross Validation en la base *respondieron_2025* probando K=1, 3,…, 29. El K óptimo según Accuracy promedio fue K=15 (Accuracy CV≈0.765), como se ve en el Gráfico 3. Esto es consistente con las tablas 3, 4 y 5. Los valores de K intermedios-altos reducen la varianza sin perder (e incluso mejorando) Accuracy respecto de K=5/10.

A medida que aumenta K crece el sesgo y cae la varianza: K=1 sobre-ajusta (alto rendimiento en *train*, caída en *test*), mientras que K=10–20 ofrece un compromiso más estable. Esta tendencia se confirma con la validación cruzada, que ubica el óptimo alrededor de K=15.

D. Desempeño de modelos, elección y predicción afuera de la muestra

8. Se compararon tres clasificadores en el conjunto de testeo, con umbral 0.5: el modelo Logit (reducido), KNN con validación cruzada (CV) con un K óptimo = 15, y Logit con regularización Ridge (L2). Las métricas reportadas fueron Accuracy, AUC-ROC (capacidad de discriminación), Precision, Recall y F1 para la clase positiva Pobre.

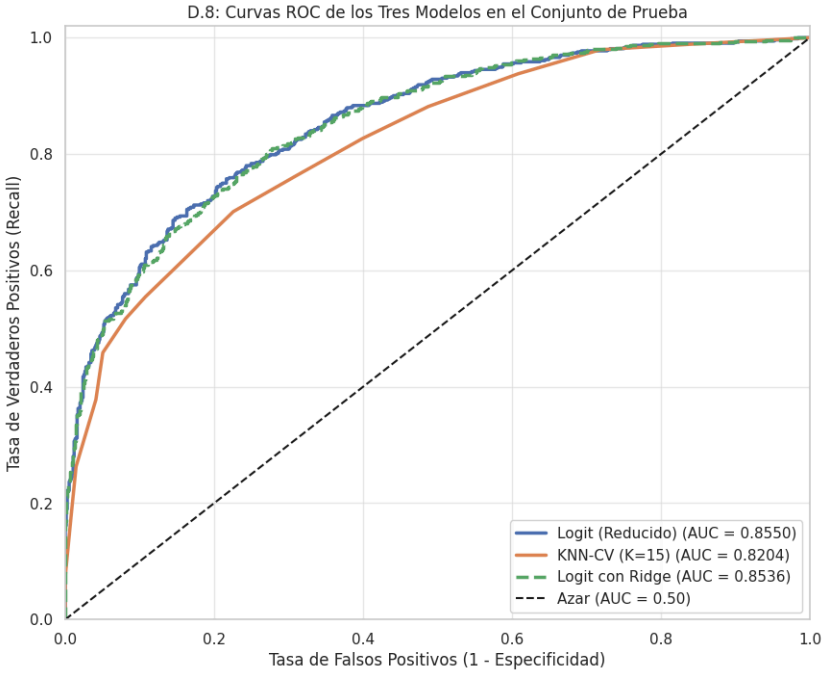
Para comparar los modelos sobre el conjunto de prueba utilizamos, en primer lugar, las matrices de confusión con umbral de clasificación 0.50. En el caso del Logit reducido, la matriz es [[650, 103], [182, 350]]. Interpretando en el orden estándar [NoPobre, Pobre], esto implica 650 verdaderos negativos (TN), 103 falsos positivos (FP), 182 falsos negativos (FN) y 350 (TP) verdaderos positivos. En términos de métricas para la clase Pobre, esto se traduce en una Precision cercana a 0.77 y un recall de aproximadamente 0.66, valores que se reflejan también en el reporte de clasificación (ver Tabla 6). Para KNN-CV (K=15), la matriz es [[640, 113], [209, 323]], con 323 TP y 209 FN. El Recall cae a aproximadamente 0.61 y la medida de Precision ronda 0.74. En síntesis, con el mismo umbral, el Logit recupera más pobres (menos falsos negativos) que KNN, al costo de un leve aumento de falsos positivos.

Tabla 6. Comparación de métricas en testeo

	Accuracy	AUC-ROC	Precision (Pobre)	Recall (Pobre)	F1-Score (Pobre)
Logit (Reducido)	0.7782	0.8550	0.7726	0.6579	0.7107
KNN-CV (K=15)	0.7494	0.8204	0.7408	0.6071	0.6674
Logit con Ridge	0.7774	0.7774	0.7746	0.6523	0.7082

Como se puede observar en la Tabla 6, el Logit alcanza un valor de Accuracy = 0.778, un índice AUC-ROC= 0.885, Precisión = 0.773, Recall = 0658 y F1-Score = 0.711. Mientras que KNN-CV obtuvo un Accuracy de 0.749, índice de AUC-ROC = 0.820, Precision = 0.741, Recall = 0.607 y F1 = 0.667. Vemos que el modelo de Logit supera al KNN con K óptimo encontrado con CV (K=15) de manera consistente.

Gráfico 4. Curvas ROC de los tres modelos en *testeo*



Complementariamente, comparamos el poder discriminatorio con la curva ROC. El Gráfico 9 muestra que la curva del Logit reducido se ubica por encima de la de KNN-CV, con un AUC-ROC ≈ 0.855 frente a 0.820 de KNN. El Logit con regularización Ridge (AUC ≈ 0.854) prácticamente se superpone con el Logit reducido, por lo que ambos superan a KNN en esta métrica. Esta evidencia se refuerza en la Tabla 6, donde el Logit reducido alcanza valores mayores que KNN. En conjunto, la matriz de confusión, las métricas resumidas y el análisis de las curvas ROC indican que el Logit reducido ofrece un mejor equilibrio para detectar pobreza que KNN-CV bajo el umbral estándar de 0.5.

9. Bajo un criterio de política pública, el costo de un Falso Negativo (clasificar como "No pobre" a quien sí lo es) es considerablemente mayor que el de un Falso Positivo, ya que implica negarle asistencia a un hogar vulnerable. Por ello, el objetivo principal es maximizar el Recall de la clase "Pobre", manteniendo una precisión razonable para no diluir la efectividad del programa. Basado en el análisis comparativo del punto anterior (Tabla 6), el modelo Logit reducido es el que presenta el mejor desempeño general y, por tanto, se elige como base para esta optimización. La siguiente tabla explora cómo varían las métricas de este modelo al ajustar el umbral de clasificación para priorizar el Recall.

Tabla 7. Rendimiento del Modelo logit a diferentes umbrales

Umbral	Accuracy	Precision (Pobre)	Recall (Pobre)	F1-Score (Pobre)	Falsos Negativos	Falsos Positivos
0.50	0.7782	0.7726	0.6579	0.7107	182	103
0.45	0.7805	0.7490	0.7068	0.7273	156	126
0.40	0.7696	0.7100	0.7500	0.7294	133	163
0.35	0.7549	0.6719	0.7970	0.7291	108	207
0.30	0.7377	0.6395	0.8402	0.7262	85	252
0.25	0.7206	0.6128	0.8835	0.7236	62	297
0.20	0.6840	0.5739	0.9192	0.7066	43	363
0.15	0.6482	0.5432	0.9455	0.6900	29	423
0.10	0.5984	0.5079	0.9662	0.6658	18	498
0.05	0.5424	0.4747	0.9868	0.6410	7	581

Como se observa en la Tabla 7, reducir el umbral de probabilidad impacta directamente en el trade-off entre Precisión y Recall. Al bajar el punto de corte de 0.50 a 0.35, el Recall (Pobre) aumenta significativamente de 0.658 a 0.797, y el número de Falsos Negativos se reduce drásticamente de 182 a 108. Este cambio logra que el programa alcance a una mayor proporción de la población objetivo. Si bien la Precisión disminuye (de 0.773 a 0.672), este es un compromiso aceptable y deliberado para minimizar el riesgo de exclusión social. Por lo tanto, se elige como modelo final y operativo el Logit reducido con un umbral de clasificación optimizado en 0.35.

10. Finalmente, aplicamos el Logit reducido con umbral optimizado = 0.35 a la base *norespondieron_2025* (previa preparación de X consistente con el set de entrenamiento). Reportamos, de manera comparativa, la proporción de observaciones clasificadas como pobres bajo el umbral por defecto (0.50) y bajo el umbral elegido (0.35). Con $p > 0.50$, la proporción estimada de pobres es 34.63%. Al ajustar el umbral a 0.35 para priorizar el Recall, la proporción sube a 45.7%, lo que representa un incremento de 11.07 puntos porcentuales en la identificación de potenciales beneficiarios del plan. Este resultado es coherente con el objetivo de política: el umbral más bajo captura a una fracción mayor de hogares vulnerables, reduciendo falsos negativos en un subconjunto poblacional donde la no respuesta puede asociarse a condiciones de mayor precariedad. Naturalmente, esta

decisión implica admitir más falsos positivos; sin embargo, ese costo se considera aceptable frente al riesgo de exclusión de personas pobres.