Universidad de Buenos Aires
Facultad de Ciencias Económicas
Escuela de Estudios de Posgrado

Maestría en Economía Aplicada

Materia:
Taller de Programación

Docente:
Noelia Romero

Informe – Trabajo Práctico # 3

Clasificando pobres con la EPH

Grupo # 5:

Cammisi, Andrés
Porco, Matías
Pineda, David

Link GitHub: https://github.com/DavidPT0902/Grupo_5_UBA_2025.git

Noviembre, 2025

Predicción de Pobreza con Modelos de Clasificación: Un Enfoque Empírico con Logit y KNN

Introducción

La pobreza es un fenómeno multidimensional cuya correcta identificación condiciona la eficacia de políticas públicas sensibles, especialmente aquellas basadas en la asignación de recursos escasos, como los programas alimentarios. Contar con un modelo predictivo que permita anticipar la vulnerabilidad de los hogares, particularmente cuando el ingreso no está disponible, resulta fundamental para una intervención focalizada. Este trabajo utiliza datos de la EPH (2005–2025) y combina la partición de la base respondieron con diferentes clasificadores supervisados para evaluar su capacidad de predicción sobre la condición de pobreza. Luego, se aplica el mejor método en la base norespondieron para inferir pobreza en hogares donde el ingreso no fue declarado.

La estructura sigue las etapas del proceso de modelado: validación del set de entrenamiento, ajuste de un modelo Logit, estimación mediante K Vecinos Cercanos (KNN), comparación fuera de muestra y, finalmente, aplicación del clasificador seleccionado. El análisis se presenta en forma integrada, pero acompañado de los números y procedimientos correspondientes.

A. Enfoque de validación

1) Tabla de diferencia de medias, selección de variables y discusión

Las variables señaladas capturan dimensiones socioeconómicas relacionadas con el riesgo de estar en condición de pobreza, ya que son aproximaciones al tipo de inserción laboral, la acumulación de capital humano, la estructura de roles dentro del hogar y su tamaño, entre otras. Todas fueron previamente recodificadas en los trabajos prácticos anteriores, asegurando consistencia entre años y bases. A continuación, damos razones para la elección de cada una de estas variables.

- *Parentesco:* La posición dentro del hogar puede influir en el acceso a ingresos y en la exposición a la vulnerabilidad (por ejemplo, jefatura de hogar, cónyuge, hijo/a).
- *Sexo:* La evidencia empírica muestra brechas de ingreso y de participación laboral según género, lo que afecta directamente el riesgo de pobreza.
- *Edad:* La edad se correlaciona con el ciclo de vida laboral, la acumulación de capital humano y la probabilidad de dependencia económica.
- *Estado civil:* El estado civil afecta la estructura de ingresos y responsabilidades del hogar (por ejemplo, monoparentalidad), factores relevantes para la pobreza monetaria.

- *Condición de actividad:* La participación laboral es un determinante directo del ingreso y, por lo tanto, parecería ser un predictor natural para clasificar la situación de pobreza.
- *Nivel educativo:* Es uno de los determinantes más del ingreso y de las posibilidades de conseguir un empleo.
- *Categoría ocupacional:* Diferencias entre asalariados, cuentapropistas o empleadores se asocian a distintos niveles de estabilidad laboral y remuneración.
- *Categoría de inactividad:* Identifica grupos con riesgo elevado (estudiantes, jubilados, inactivos no definidos). Permite discriminar entre inactividad asociada a inversión en capital humano vs. inactividad estructural.
- *Miembros del hogar:* El tamaño del hogar y la relación entre perceptores de ingresos y dependientes es fundamental para medir necesidades equivalentes y exposición a privaciones.

El split 70/30 con semilla 444 produjo un conjunto de entrenamiento de 6.538 observaciones y uno de testeo de 2.802. La tasa de pobreza quedó prácticamente igual entre ambos (46,2% vs 46,3%).

La comparación de medias muestra diferencias muy pequeñas, como se observa a continuación

Tabla A.1 — Diferencia de medias entre Train y Test (2025)

Variable	Media Train	Media Test	Diferencia	Diferencia %
Parentesco	2.254	2.225	0.029	1.29%
Sexo	1.527	1.510	0.017	1.10%
Edad	36.894	37.475	-0.581	-1.58%
Estado civil	3.427	3.420	0.007	0.21%
Condición de actividad	2.240	2.254	-0.014	-0.61%
Nivel educativo	3.716	3.686	0.030	0.80%
Categoría ocupacional	1.230	1.200	0.030	2.42%
Categoría de inactividad	1.732	1.748	-0.016	-0.91%
Miembros del hogar	3.445	3.443	0.002	0.05%

En promedio, alrededor del 1% y ninguna por encima del 2,5%. Incluso en la variable con mayor diferencia relativa, cat_ocup, la discrepancia fue apenas 2,42%. Esta estabilidad indica que la partición no generó desbalance sistemático y que el

conjunto de testeo resulta adecuado para evaluar el desempeño real de los clasificadores sin sesgos de composición.

2) Separación por año

Se generaron las cuatro bases requeridas: respondieron_2005 (14.481 casos), respondieron_2025 (9.340), norespondieron_2005 (170) y norespondieron_2025 (4.463). La estimación y evaluación del modelo se concentrará en 2025, aplicándose posteriormente a quienes no declararon su ingreso ese año.

B. Modelo de Regresión Logística

1) Estimación de Efectos Marginales

Se estimó un modelo de regresión logística para el año 2025, donde la variable dependiente es pobre y el vector de regresores estaba conformado por las variables listadas previamente. El modelo se ajustó sobre x_train de la base "respondieron_2025", e incluye un intercepto para capturar la probabilidad de ser pobre cuando todas las variables toman su categoría base. La tabla B.1 contiene los coeficientes, errores estándar, valores p y odds ratios para cada regresor.

Tabla B.1 — Resultados del modelo Logit (2025)
(Variable dependiente: pobre = 1)

Variable	Coeficiente	Error Estándar	P-value	Odds Ratio
Constante	-1.7161	0.3043	0.0000	0.1798
Parentesco	-0.0791	0.0243	0.0012	0.9239
Sexo	0.0941	0.0578	0.1037	1.0986
Edad	-0.0021	0.0018	0.2504	0.9979
Estado civil	0.0160	0.0225	0.4788	1.0161
Condición de actividad	0.0633	0.0748	0.3974	1.0653
Nivel educativo	-0.2308	0.0169	0.0000	0.7939
Categoría ocupacional	0.0731	0.0558	0.1902	1.0758
Categoría de inactividad	0.2192	0.0283	0.0000	1.2450
Miembros del hogar	0.5434	0.0234	0.0000	1.7218

Los resultados permiten extraer algunos patrones importantes. Primero, el coeficiente asociado a nivel educativo (-0,2308) es negativo y altamente significativo ($p<0,01$), con odds ratio de 0,7939, lo que indica que, manteniendo las demás características constantes, un aumento de una categoría en el nivel educativo (por ejemplo, pasar de primaria completa a secundaria, o de secundaria a superior), reduce aproximadamente 20% las probabilidades relativas de ser pobre. Este

resultado es consistente con la literatura que destaca el papel del capital humano como mecanismo de protección frente a la pobreza. En segundo lugar, la cantidad de miembros del hogar muestra un coeficiente positivo (0,5434) y signficativo ($p<0,01$), con un odd ratio de 1,7218; por lo tanto, un miembro adicional en el hogar incrementa aproximadamente 72% las chances de que la persona sea calificada como pobre, manteniendo todo lo demás constante. Esto refleja que, ante una cantidad dada de recursos, los hogares más numerosos enfrentan una mayor presión para satisfacer las necesidades básicas de todos los miembros, elevando con ello el riesgo de pobreza.

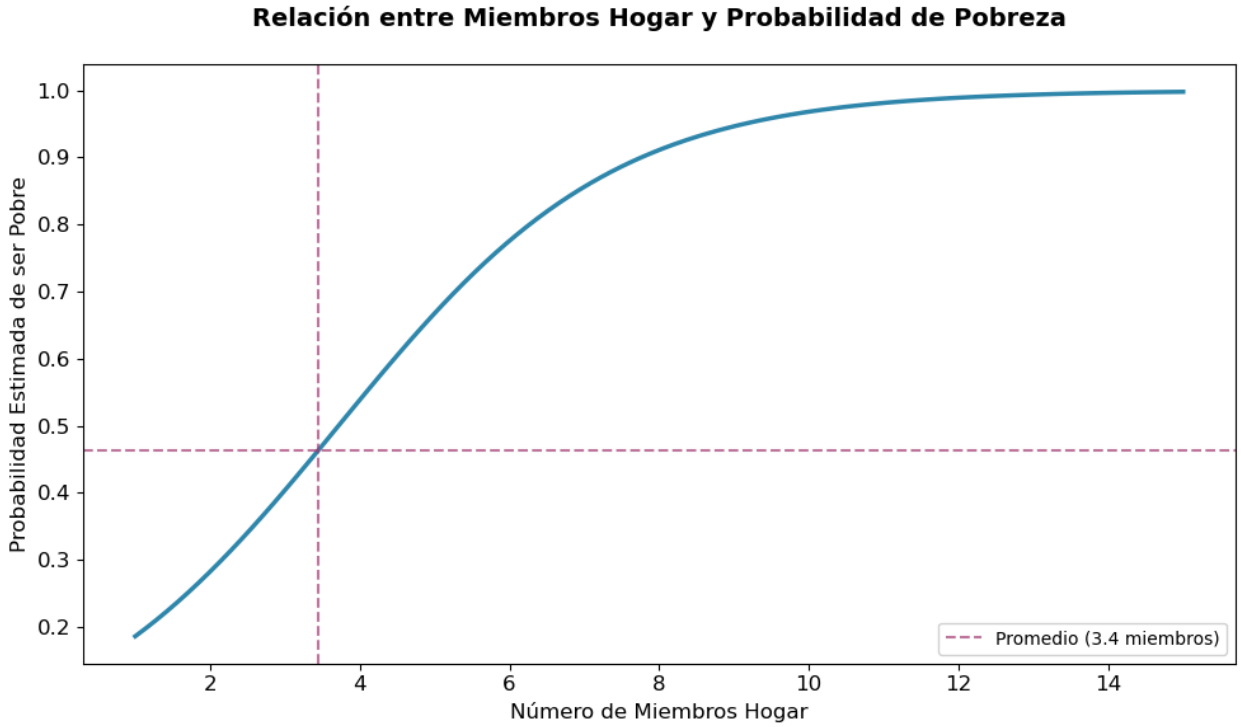
Por su parte, la categoría de inactividad también aparece asociada significativamente con la pobreza, su coeficiente es 0,2192 y su odd ratio es de 1,245, sugiriendo que algunos inactivos (por ejemplo, algunas personas dedicadas a labores domésticas o inactivos por problemas de salud) tienen mayor probabilidad de encontrarse por debajo de la línea de pobreza que otros grupos, controlando las demás variables.

Finalmente, variables como sexo, edad, estado civil, condición de actividad y categoría ocupacional no resultan estadísticamente signficativas al 5%, aunque presentan signos acordes con algunos supuestos económicos; por ejemplo, mayor pobreza asociada a ciertos estados civiles o posiciones ocupacionales más precarias. En síntesis, los factores más importantes en nuestro modelo son el nivel educativo, el tamaño del hogar y el tipo de inactividad, todos altamente significativos. Otras variables (sexo, edad, estado civil, ocupación, condición de actividad) no son significativas porque posiblemente sus efectos hayan quedado subsumidos por variables más estructurales, como educación e inactividad.

2) Visualización

Para ilustrar la forma funcional, se graficó la probabilidad estimada de ser pobre según el número de miembros del hogar, manteniendo el resto de las variables en su valor medio.

Figura B.2 – Relación entre miembros del hogar y probabilidad de pobreza



Como puede observarse, a medida que aumenta el número de personas que integran el hogar, la probabilidad estimada de pobreza crece de forma sustancial. Hogares pequeños (1–2 miembros) presentan probabilidades inferiores al 30%. A partir de 4–5 miembros, la probabilidad supera el 50%. En hogares numerosos (8 o más miembros), la probabilidad se acerca a 1 (es decir, casi certeza de ser pobres). La forma en "S" de la relación refleja las características de la función logística: al inicio, el incremento en la probabilidad es moderado. Entre 3 y 7 miembros, el crecimiento es muy acelerado. Finalmente, se estabiliza cerca del límite superior (100%). La línea vertical punteada indica el tamaño promedio de los hogares de la muestra: 3,4 miembros. La línea horizontal muestra que, en ese tamaño promedio, la probabilidad de ser pobre se ubica alrededor del 45–50%. Es decir, un hogar "típico" ya presenta un riesgo relativamente alto de pobreza.

C. Método de Vecinos Cercanos (KNN)

1) Clasificación con $K = \{1, 5, 10\}$ y trade-off

El algoritmo KNN fue implementado sobre las mismas variables, luego de estandarizar las numéricas. Con K=1 la precisión en entrenamiento fue 0,895, reflejando un ajuste muy fino a los datos; con K=5 y K=10 esta precisión baja a 0,773 y 0,754 respectivamente. Esto resume el trade-off fundamental: K pequeño implica baja sesgo pero alta varianza (sobreajuste), mientras que K grande suaviza

la frontera de decisión y reduce la varianza, al precio de introducir algo de sesgo. Elegir un K intermedio permite captar patrones sin sobre-ajustarse al ruido local.

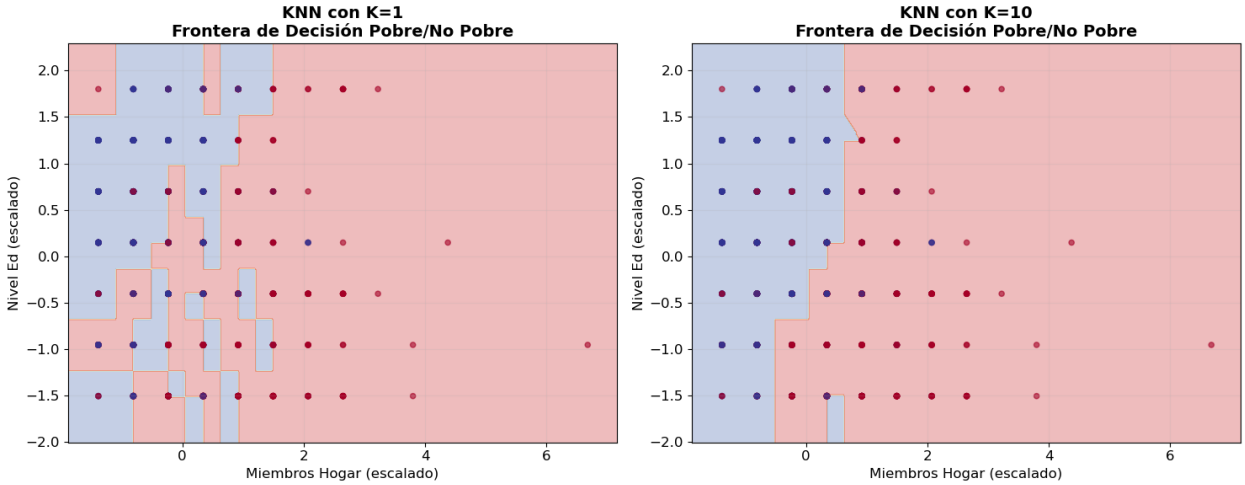
Tabla C.1 — Resumen de Precisión en entrenamiento por K

K	Precisión Train
1	0.895
5	0.773
10	0.754

2) Fronteras de decisión

Se graficaron las fronteras de decisión en el plano generado por miembros_hogar y nivel_ed para K=1 y K=10.

Figura C.2 – Fronteras para K=1 y K=10



Las figuras exhiben dos comportamientos distintos: con K=1 la frontera es muy irregular y responde a cada punto de entrenamiento, mientras que con K=10 las regiones se tornan más homogéneas, lo que refleja una generalización más estable aunque menos precisa en áreas de alta variabilidad.

3) Selección del K óptimo mediante validación cruzada

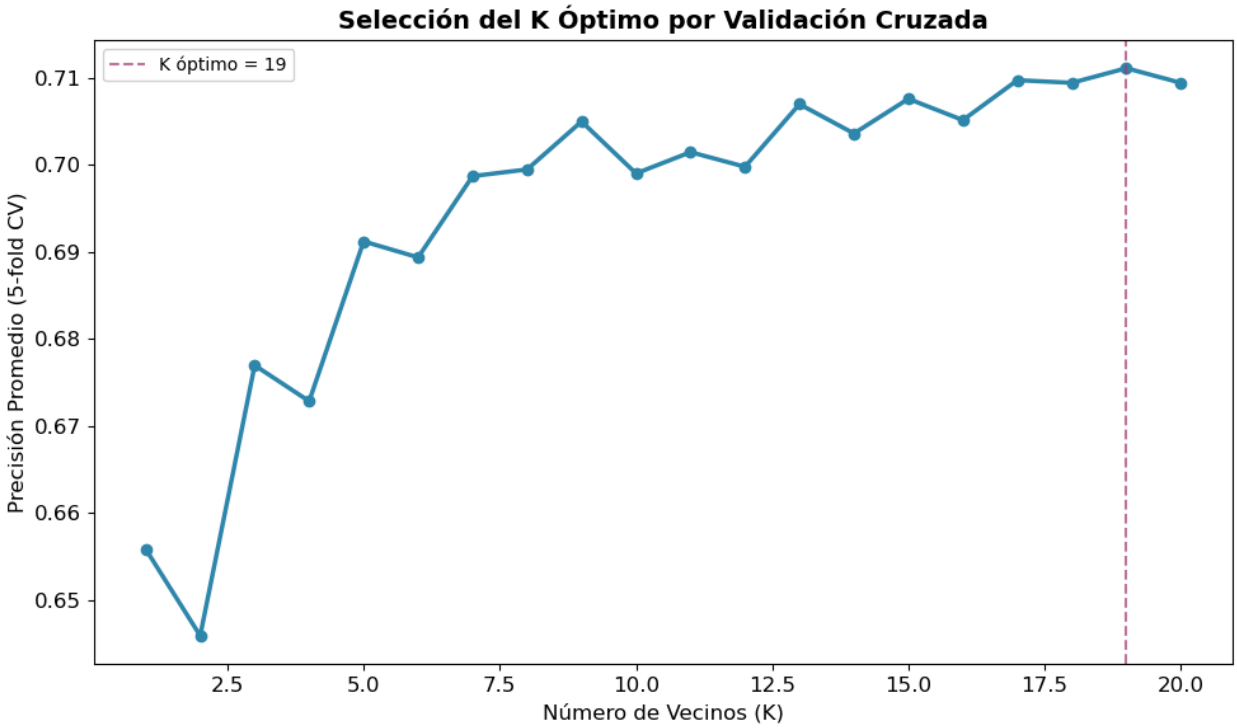
Se realizó una validación cruzada 5-fold.

Tabla C.2 — Mejor K por validación cruzada 5-fold

Métrica	Valor
K óptimo	19
Precisión promedio	0.711

Cuando se restringe K al rango pedido por la consigna (1 a 10), el mejor desempeño aparece cerca de K=9–10. Al extender la búsqueda a 1–20, el valor óptimo resulta ser K=19, con precisión CV aproximada de 0,711.

Figura C.3 – Selección de K por CV



Este resultado muestra que la estructura de los datos favorece fronteras relativamente suaves.

D. Desempeño fuera de la muestra y análisis para políticas públicas

1) Comparación entre Logit y KNN-CV

Al evaluar en X_test_2025, el Logit con umbral 0,5, se observa la siguiente matriz de confusión:

Tabla D.1 — Matriz de Confusión Logit (2025)

	Pred No Pobre	Pred Pobre
Real No Pobre	1134	372
Real Pobre	499	797

La matriz de confusión indica que el Logit identifica correctamente a la mayoría de los no pobres, pero comete un número importante de falsos negativos (pobres clasificados como no pobres). En términos globales, el área bajo la curva ROC también favorece al KNN (0,779 contra 0,765).

Figura D.2 – Curva ROC Logit vs KNN

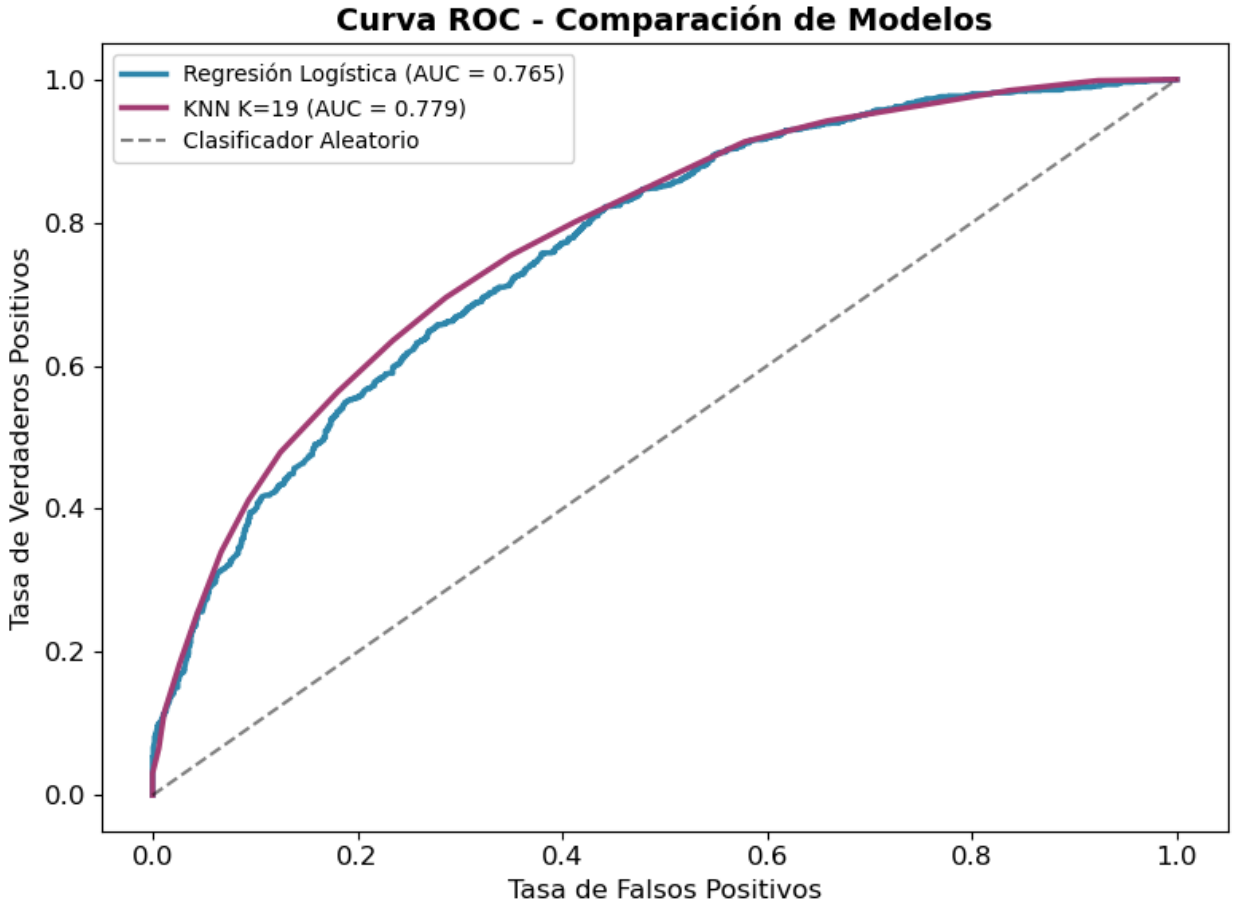


Tabla D.2 — Métricas de Clasificación

	Logit	KNN
Accuracy	0.689151	0.705211
Precisión (Pobre)	0.681779	0.69983
Sensibilidad (Pobre)	0.614969	0.635031
F1-Score (Pobre)	0.646653	0.665858

La comparación indica que ambos modelos separan razonablemente bien pobres y no pobres, pero el KNN-CV obtiene un desempeño levemente superior en métricas clave, especialmente en recall y F1, lo que resulta relevante para aplicaciones de política.

2) Elección del modelo para asignación de alimentos

Suponiendo que el Ministerio enfrenta recursos escasos y desea minimizar el error de dejar hogares pobres fuera del programa (error tipo II), la métrica central es el recall de la clase pobre. En este análisis, el KNN seleccionado logra recuperar una mayor proporción de pobres sin sacrificar de modo sustantivo la precisión. Por ello,

y dado que también domina en F1 y AUC, se concluye que el modelo KNN-CV es más adecuado como herramienta de focalización, siempre sujeto a una calibración del umbral acorde al presupuesto disponible.

3) Predicción de pobreza en norespondieron_2025

Con el KNN-CV final, de las 4.463 personas que no declararon su ingreso, 1.836 son clasificadas como pobres, lo que corresponde a 41,1% del total. Esta estimación constituye una primera aproximación operativa para la identificación de hogares potencialmente vulnerables entre quienes no aportan información de ingresos y otorga una base para priorizar intervenciones.

Conclusión

A partir de un enfoque riguroso de validación y comparación de modelos, se construyó un clasificador capaz de anticipar la condición de pobreza a partir de información sociodemográfica básica. La regresión logística ofreció interpretaciones claras y efectos sustantivos coherentes —como la fuerte asociación entre tamaño del hogar e inactividad con mayores riesgos de pobreza y el rol protector de la educación—, mientras que el KNN, tras seleccionar el número óptimo de vecinos mediante validación cruzada, mostró un desempeño levemente superior en las métricas relevantes para política, en particular en la capacidad de identificar correctamente a los hogares pobres.

En contextos donde los errores de exclusión son especialmente costosos, como la asignación de programas alimentarios, priorizar el recall y el F1 resulta más adecuado que centrarse exclusivamente en la exactitud global. Bajo estos criterios, el KNN-CV surge como la alternativa preferible. La predicción sobre los individuos sin declaración de ingreso sugiere que cerca del 41% podría encontrarse en situación de pobreza, una cifra útil como referencia inicial para la focalización del programa. En conjunto, los resultados del presente análisis evidencian cómo los métodos de clasificación, cuando son acompañados por una validación cuidadosa y una interpretación sustantiva, se convierten en herramientas valiosas para la toma de decisiones basada en datos.