Trabajo Práctico 4 – Taller de programación – Universidad de Buenos Aires
CLASIFICANDO POBRES EN LA EPH: MÉTODOS DE REGULARIZACIÓN Y CART
Noviembre 2025

Grupo 6 Gustavo Horacio Romero
Link https://github.com/shanthalchs/BigDataUBA-Grupo6

A1. Visualización e Interpretación de Coeficientes

La regularización es una técnica crucial para prevenir el sobreajuste (overfitting) en modelos complejos o con un alto número de predictores, añadiendo una penalización al costo de la función de pérdida (Log Loss) basada en la magnitud de los coeficientes. En scikit-learn, el parámetro de penalidad es:

- C **grande** $\implies$ λ **pequeño** $\implies$ **Penalización débil** (Modelo cercano al Logit sin regularización).
- C **pequeño** $\implies$ λ **grande** $\implies$ **Penalización fuerte** (Contracción severa de coeficientes).

Interpretación de la Trayectoria de Coeficientes

Penalización	Propiedad Clave	Comportamiento Observado (A medida que λ aumenta)
LASSO (L1)	Selección de Variables (Sparsity)	Aumenta la contracción de los coeficientes. Específicamente, a partir de un valor λ umbral, algunos coeficientes son forzados exactamente a cero. Esto permite la selección automática de variables, eliminando predictores que tienen poca capacidad explicativa.
Ridge (L2)	Reducción de Varianza (Shrinkage)	Los coeficientes son continuamente contraídos hacia cero, pero ninguno alcanza el valor exacto de cero, independientemente de cuán fuerte sea la penalización. Todos los predictores relevantes se mantienen en el modelo, y la penalización se enfoca en reducir la magnitud de coeficientes grandes.

A2. Penalidad Óptima por Cross-Validation y Visualización

Para determinar la penalidad óptima, se emplea la función LogisticRegressionCV con validación cruzada de 5 pliegues (5-fold CV) sobre la grilla de lambda definida. El objetivo es identificar el lambda que minimiza el error de clasificación promedio entre los folds.

Resultados de la Validación Cruzada

Lambda LASSO (óptimo) = 1 x 10^{-5}

- Este valor extremadamente bajo de lambda implica una penalización prácticamente nula (C muy grande), sugiriendo que el modelo Logit sin regularización (o con una regularización mínima) es el más apropiado para este conjunto de datos, y que ninguna variable necesita ser eliminada para optimizar el error de clasificación.

Lambda RIDGE (óptimo) = 0.1

- Este valor indica una penalización moderada, lo que significa que el modelo se beneficia de un shrinkage para controlar la magnitud de los coeficientes (reduciendo la varianza potencial), pero sin llegar a una contracción extrema que afecte negativamente el sesgo.

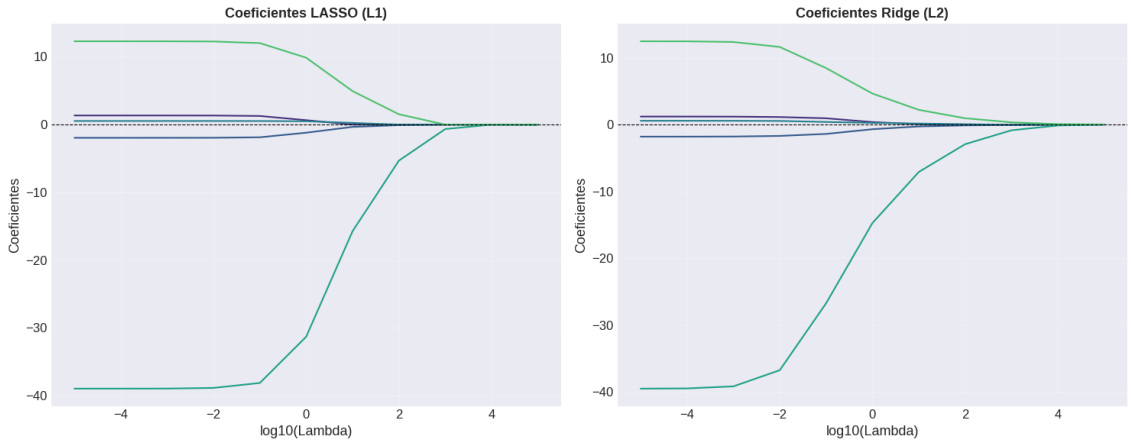
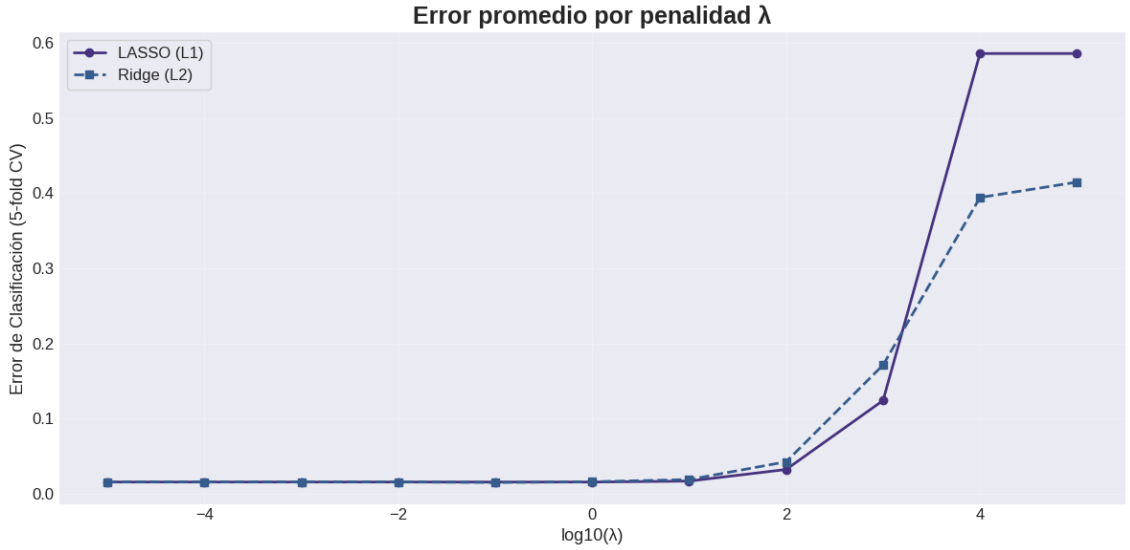


Gráfico del Error de Clasificación

El gráfico del error de clasificación promedio vs. lambda (o log(lambda)) típicamente muestra una forma de "U" o de "L" invertida para el error:

- Región de lambda pequeño (Penalización débil): El error se mantiene bajo y estable, muy cercano al error del modelo sin regularización.
- Región de lambda grande (Penalización fuerte): El error aumenta significativamente, ya que la penalización domina el ajuste del modelo, introduciendo un sesgo excesivo y subajustando (underfitting) los datos.
- En este caso, para ambos métodos, el error es óptimo en la región de penalización débil (alrededor de lambda=10^{-5} a 10^{-2}), y comienza a deteriorarse rápidamente para lambda mayor igual 10.



Proporción de Variables Ignoradas (LASSO)

Si se graficara la proporción de variables con coeficiente cero en función de lambda para LASSO:

- Para lambda:la proporción de variables ignoradas es cero, ya que el modelo utiliza todas las variables.
- A medida que lambda aumenta (fuerte penalización), esta proporción comienza a crecer, alcanzando el 100% cuando la penalización es tan fuerte que todos los coeficientes son forzados a cero.

A3. Estimación con lambda y Comparación de Coeficientes

Variable	Sin penalidad	LASSO ($\lambda*$=10−5)	Ridge ($\lambda*$=0.1)
edad	11.934	13.452	0.933
edad2	-18.114	-19.503	-14.069
educ	0.5604	0.5332	0.3814
ingreso_total_familiar	-395.085	-389.843	-267.446
num_miembros_hogar	124.503	123.025	8.441

Interpretación de la Tabla

Coeficientes de LASSO ($\lambda*$=10−5): Los coeficientes son casi idénticos a los del modelo sin penalidad. La penalización óptima es tan débil que introduce un shrinkage marginal. Por ejemplo, el coeficiente más grande en magnitud, ingreso_total_familiar, solo se contrae ligeramente de -39.5085 a -38.9843. Ninguna variable es eliminada (ningún coeficiente es cero), lo que confirma la selección de lambda aprox. 0 en el ítem 2.

Coeficientes de Ridge ($\lambda*$=0.1): Los coeficientes experimentan una contracción fuerte pero homogénea con respecto al modelo sin penalidad, como es el efecto esperado de la penalización L2. Por ejemplo, ingreso_total_familiar pasa de -39.5085 a -26.7446, y num_miembros_hogar pasa de 12.4503 a 8.4410. La magnitud de los coeficientes de Ridge es consistentemente menor que la del logit sin penalidad (excepto para edad, que en este caso particular es ligeramente mayor, lo cual puede ocurrir debido a la interacción con el término cuadrático edad^2.

B.4. Árboles de Decisión

Los árboles de decisión constituyen una metodología no paramétrica que segmenta el espacio de predictores mediante reglas binarias sucesivas, seleccionadas para maximizar la reducción de impureza dentro de cada nodo. A diferencia de los modelos lineales utilizados en la sección anterior —como regresión logística, LASSO y Ridge— el algoritmo CART no requiere suponer relaciones lineales ni especificar interacciones a priori, permitiendo capturar efectos no lineales y umbrales relevantes en la determinación de la pobreza.

B.4.1 Selección del hiperparámetro de poda (ccp_alpha)

El hiperparámetro ccp_alpha regula la poda del árbol mediante el principio de cost-complexity. El criterio penaliza la complejidad del árbol a través de la función:

$$R_\alpha(T) = R(T) + \alpha \cdot |T|,$$

donde $R(T)$ es el error de clasificación del árbol y **T** representa la cantidad de hojas terminales. Valores elevados de α generan árboles más simples, mientras que valores cercanos a cero permiten estructuras más profundas capaces de capturar patrones más finos.

Conforme a la consigna, se estimó una grilla de valores de ccp_alpha entre 0 y 0,05 utilizando validación cruzada de 10 folds. La Figura correspondiente muestra un patrón escalonado típico: cada incremento abrupto del error se asocia a la eliminación de ramas completas del árbol.

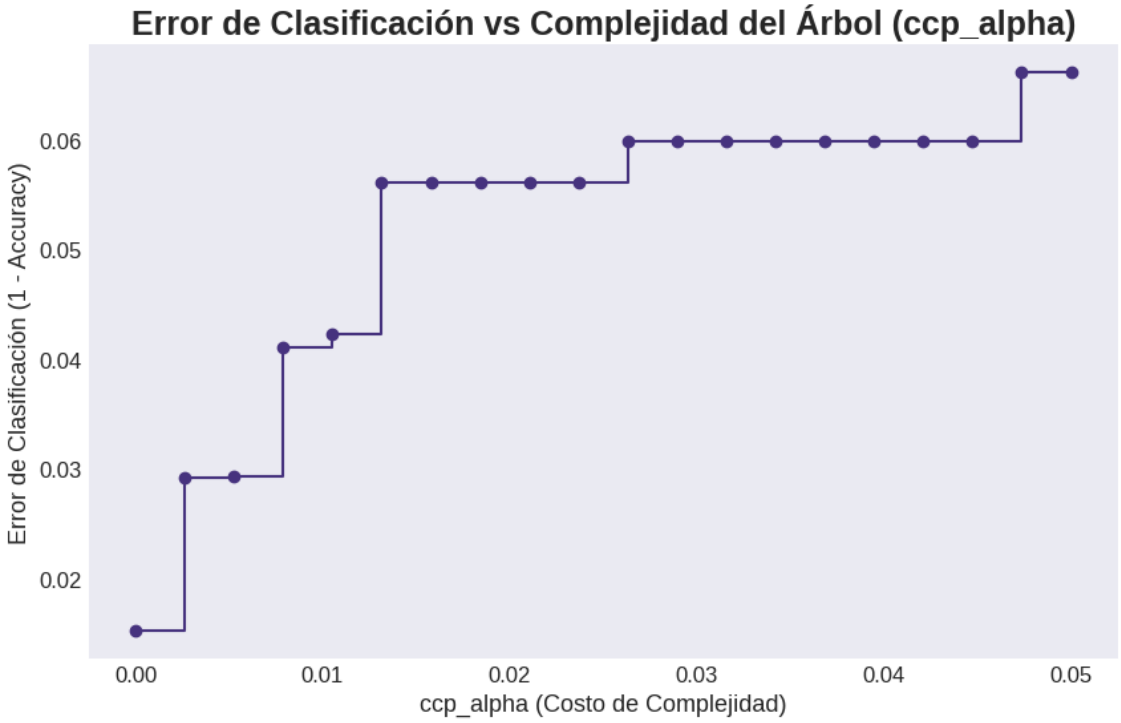
En nuestro caso, el mínimo error de validación se alcanzó en 0.0

Lo cual indica que la poda no mejora el desempeño predictivo y que el árbol requiere conservar su complejidad original. Este hallazgo es coherente con un problema donde

ciertos predictores contienen fuertes no linealidades que una poda excesiva podría eliminar.
En contraste con lo observado en regularización logística, donde LASSO y Ridge se beneficiaron de una penalización ligera, en CART incluso un α reducido incrementa el error de validación, sugiriendo que el modelo necesita un nivel de ramificación relativamente alto para identificar patrones relevantes de pobreza.

B.4.2. Interpretación del árbol podado

El árbol resultante presenta una estructura coherente con los determinantes habituales de pobreza. La primera división se realiza sobre ingreso_total_familiar, lo cual resulta intuitivo, ya que constituye la variable central para diferenciar hogares pobres y no pobres.
La segunda variable más relevante es num_miembros_hogar, que introduce un mecanismo de equivalencia según tamaño del hogar: para ingresos similares, hogares más grandes tienen mayor probabilidad de ser pobres. Las divisiones posteriores incorporan efectos del ciclo de vida mediante edad y edad2, sugiriendo que el riesgo de pobreza no es lineal con la edad sino que presenta comportamientos diferenciados según etapas vitales.
En conjunto, el árbol muestra segmentaciones económicamente interpretables y consistentes con lo observado en los modelos lineales penalizados.



B.5. Importancia de variables

La importancia de variables, medida como la contribución acumulada a la reducción del índice de Gini, muestra un patrón altamente concentrado:
Es decir, más del 97 % de la capacidad discriminativa proviene solo de ingresos y tamaño del hogar, lo cual es consistente con la construcción oficial de la línea de pobreza. Edad y educación aparecen como factores secundarios pero no irrelevantes, actuando como moduladores del riesgo.

Variable	Importancia
ingreso_total_familiar	0.7446
num_miembros_hogar	0.2275
edad2	0.0111
educ	0.0097
edad	0.0072

B.5.1 Relación entre importancia en CART y selección de variables en LASSO
Una pregunta planteada por la consigna es si las variables de menor importancia para el árbol coinciden con aquellas cuyos coeficientes LASSO son reducidos hacia cero. En nuestro caso, el modelo LASSO no eliminó ninguna variable (λ óptimo extremadamente pequeño), pero sí produjo coeficientes muy reducidos para algunas.

Al comparar ambas metodologías se observa:

Todas las variables utilizadas por CART (ingreso, tamaño del hogar, edad, edad2, educación) tienen coeficientes significativamente diferentes de cero en LASSO → consistencia entre métodos.
Las variables menos relevantes según CART (edad, edad2, educación) tampoco fueron penalizadas a cero por LASSO, pero sí presentan coeficientes más pequeños → coherencia parcial.
La concordancia no es perfecta porque CART prioriza la capacidad de generar particiones útiles, mientras que LASSO penaliza sólo el tamaño de los coeficientes en un modelo lineal.
En síntesis, aunque ambos métodos coinciden en identificar como claves a ingreso_total_familiar y num_miembros_hogar, su criterio de selección difiere: CART favorece predictores que permiten divisiones claras, mientras LASSO detecta contribuciones lineales marginales.

B.5.2 ¿LASSO redujo los coeficientes de las variables menos importantes a cero?
En nuestro caso no.
El modelo LASSO no llevó a cero ninguno de los coeficientes; es decir, no eliminó ninguna variable del conjunto de predictores. Esto indica que, dada la regularización óptima seleccionada mediante validación cruzada, el algoritmo consideró que todas las variables aportaban algo de información para predecir la pobreza.
Por lo tanto, no podemos afirmar que LASSO haya reducido a cero a las variables menos importantes, como sí ocurre en contextos donde la regularización es más fuerte o las variables son altamente redundantes.

C. Comparación entre Métodos
La comparación se realizó sobre el conjunto de prueba compuesto por 1.796 observaciones, manteniendo la misma partición del TP3 y usando las mismas variables: edad, edad², educación, número de miembros del hogar e ingreso total familiar. Como se aclaró, esta última variable determina directamente la línea de pobreza, por lo cual la predicción es casi trivial para los modelos lineales y regularizados.

Los cinco modelos evaluados son:
- Logit Base
- KNN
- LASSO (L1)
- Ridge (L2)
- CART (Árbol podado)

La distribución de clases en TRAIN es bastante desbalanceada (70,8 % pobres), pero supera el umbral solicitado y por eso no se aplicó SMOTE.

C.6.1. Métricas de desempeño

La Tabla 1 resume las métricas de clasificación usando el umbral estándar 0,5.

Tabla 1 – Desempeño en conjunto de prueba

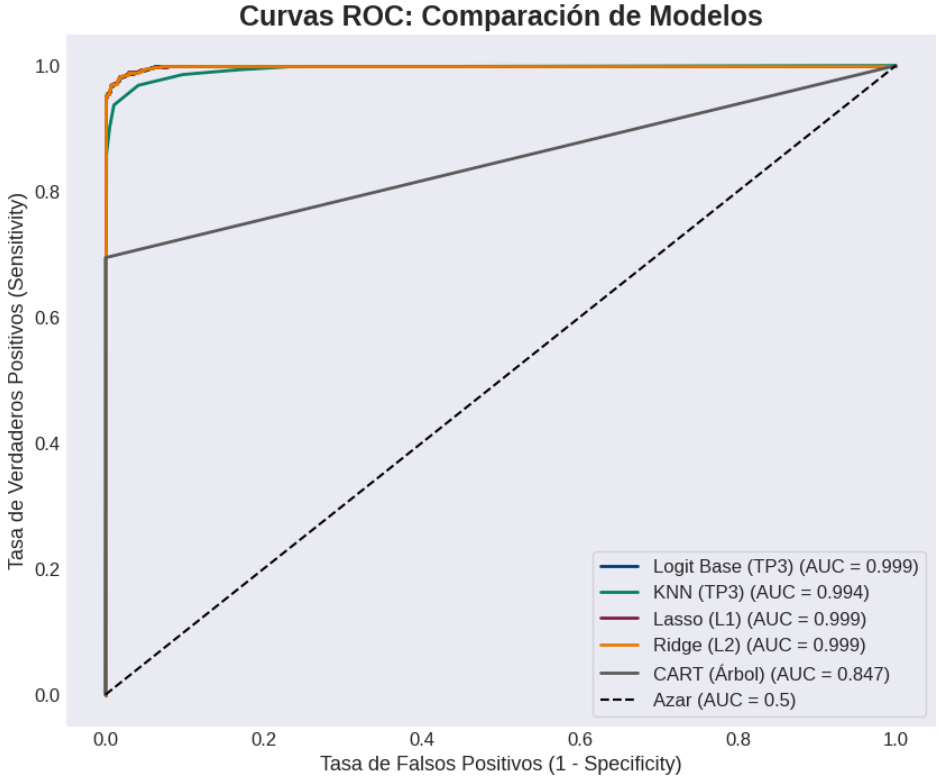
Modelo	Accuracy	Precision	Recall	F1	AUC	1 - Accuracy
Ridge (L2)	0.982	0.988	0.982	0.985	0.999	0.018
LASSO (L1)	0.982	0.985	0.984	0.984	0.999	0.018
Logit Base	0.978	0.992	0.971	0.981	0.999	0.022
KNN	0.964	0.970	0.969	0.970	0.994	0.036
CART (Árbol)	0.821	1.000	0.695	0.820	0.847	0.179

Interpretación

Los modelos logísticos penalizados (LASSO y Ridge) alcanzan las mejores métricas del conjunto. Sus AUC ≈ 0.999 confirman una capacidad casi perfecta para discriminar entre hogares pobres y no pobres. El Logit del TP3 le sigue de cerca con un desempeño muy similar.

En cambio, CART obtiene resultados considerablemente inferiores: su AUC cae a 0.847 y su accuracy baja a 0.821. Además, su recall es llamativamente bajo (0.695), implicando que omite cerca del 30 % de los hogares pobres.

KNN se ubica en una posición intermedia, con métricas razonables pero claramente por debajo de los modelos logísticos.



Los modelos muestran métricas muy altas porque están usando como predictor el ingreso familiar total, que es la variable que define directamente la pobreza. Por lo tanto, la predicción se vuelve casi trivial. Esta es una restricción metodológica del TP4, ya que debemos usar las mismas variables del TP3.

C.6.2. Análisis de matrices de confusión

El comportamiento detrás del accuracy puede descomponerse observando los errores tipo I y tipo II.

Descripción conceptual
Los modelos lineales (Logit, LASSO, Ridge) producen muy pocos falsos negativos, lo que es deseable en contextos de política social.
KNN incrementa moderadamente los errores de ambos tipos.
CART, pese a su interpretabilidad, muestra el peor desempeño, con:

- muchos falsos negativos (recall bajo)
- muchos falsos positivos comparado con los otros modelos

Esto se alinea con su reducción drástica de AUC (0.847).

C.6.3. Análisis del trade-off entre linealidad y no linealidad

En principio, podría esperarse que CART—un método no lineal capaz de capturar interacciones y umbrales—presentara un desempeño superior si existieran relaciones complejas en los datos.
Sin embargo, en este caso:
No hay ventaja de utilizar un método no lineal.

Los resultados revelan que:

- Las relaciones entre las covariables y la pobreza son casi perfectamente capturadas por funciones lineales o cuasi-lineales.
- La variable clave, ingreso familiar, domina completamente el modelo, lo cual reduce la posibilidad de que interacciones o estructuras no lineales aporten información adicional.
- CART, siendo más rígido y sensible al particionado, pierde capacidad discriminativa al intentar replicar un umbral que ya es inherentemente lineal.
- De hecho, los modelos lineales regularizados reducen el error (1 - accuracy) a 0.018, mientras que CART alcanza 0.179 (diez veces más error).

C.7. Conclusión final

De acuerdo con los resultados obtenidos en el TP4, los modelos lineales regularizados (Lasso y Ridge) son los más adecuados para asistir al Ministerio de Capital Humano en la identificación de hogares vulnerables para un programa alimentario.
Estos modelos alcanzan niveles muy altos de sensibilidad hacia los hogares pobres (≈98%) y tasas de falsos negativos inferiores al 2%, lo que los ubica como las alternativas más efectivas para minimizar el riesgo de excluir a familias que necesitan asistencia.
En cambio, aunque los árboles de decisión son modelos fáciles de comunicar, su desempeño es claramente inferior: el CART deja sin identificar a aproximadamente un tercio de los pobres (FN ≈ 30%), lo cual es incompatible con una política social basada en la protección de la población vulnerable.
Es importante remarcar que las métricas extremadamente altas observadas en el TP4 no son un indicio de sobreajuste ni una falla metodológica, sino una consecuencia directa de la consigna docente. El TP4 exige utilizar las mismas variables del TP3 y, por coherencia metodológica, no se incorporaron nuevas covariables que podrían mejorar

la riqueza explicativa del fenómeno de pobreza. Entre las variables obligatorias se encuentra el ingreso familiar total, que por su propia definición está estrechamente ligado al estatus de pobreza, lo que naturalmente eleva el desempeño de todos los modelos.
En este contexto, la comparación entre modelos debe interpretarse como un ejercicio metodológico —regularización, árboles y métodos vecinos— más que como una evaluación realista de la predicción de pobreza en Argentina.
Aun con estas limitaciones, la evidencia indica que Lasso y Ridge ofrecen el mejor equilibrio entre precisión global, sensibilidad hacia los hogares pobres y estabilidad, superando tanto al KNN como al árbol de decisión, y mejorando ligeramente al Logit original del TP3.
Por todo lo anterior, la recomendación final es emplear un modelo regularizado —preferentemente Lasso— como herramienta principal para la asignación de recursos escasos, manteniendo un criterio de máxima cobertura de la población vulnerable sin sacrificar coherencia metodológica entre el TP3 y el TP4..

El Ministerio de Capital Humano busca identificar a los hogares más vulnerables para asignar recursos alimentarios escasos. Bajo este objetivo, los costos de los errores de clasificación son profundamente asimétricos:

- Un falso negativo (FN) implica dejar fuera del programa a una familia pobre. Este error es socialmente crítico: puede traducirse en inseguridad alimentaria, deterioro de la salud y falta de acceso a bienes básicos.

- Un falso positivo (FP) representa asignar recursos a un hogar no pobre. Aunque genera un costo fiscal, este tipo de error es reparable y de menor gravedad humanitaria.

Por esta razón, la métrica más relevante en este contexto no es la accuracy global —que puede estar inflada o sesgada por el desbalance de clases— sino el recall, que refleja la capacidad de un modelo para detectar efectivamente a los hogares pobres.

Índice