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

Taller de Programación

Maestría en Economía Aplicada

Tercer trimestre 2025

Trabajo Práctico N° 4

Link GitHub https://github.com/jorge-ux272/Big-Data-UBA--Grupo-8

Docente: Noelia Romero

Alumnos:
Jorge Eduardo Bolaños Gamarra
Mario Antonio Valdivia Reyes
Héctor Sebastián San Martin

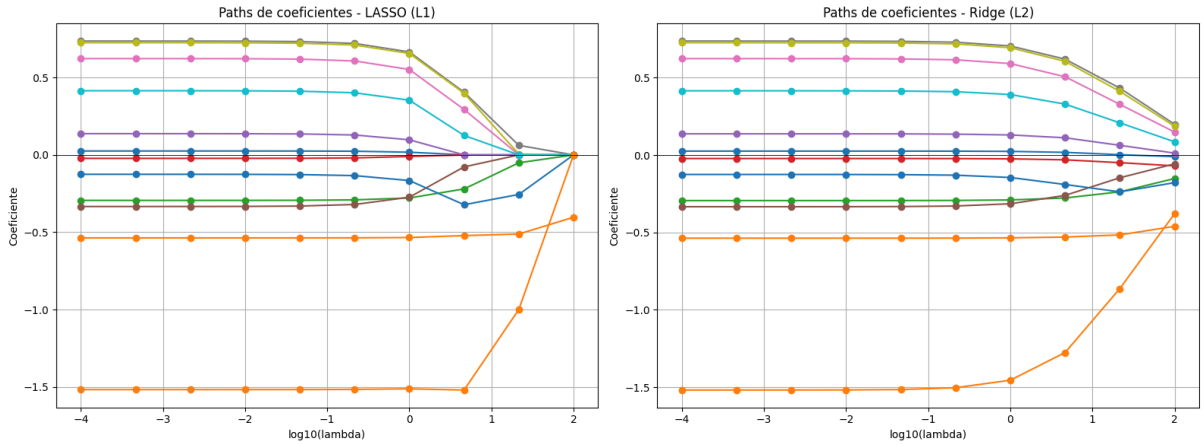
Introducción

El presente trabajo compara el desempeño predictivo de diferentes algoritmos de machine learning para clasificar la condición de pobreza en hogares, utilizando datos de la Encuesta Permanente de Hogares (EPH). El análisis desarrollado se enfoca en la implementación y evaluación de cinco modelos: Regresión Logística (sin regularización), Regresión Logística con regularización L1 (LASSO) y L2 (Ridge), Árboles de Decisión y K-Nearest Neighbors (KNN) (k vecinos más cercanos).

A. Modelo de Regresión Logística con Regularización: Ridge y LASSO

1. Visualización

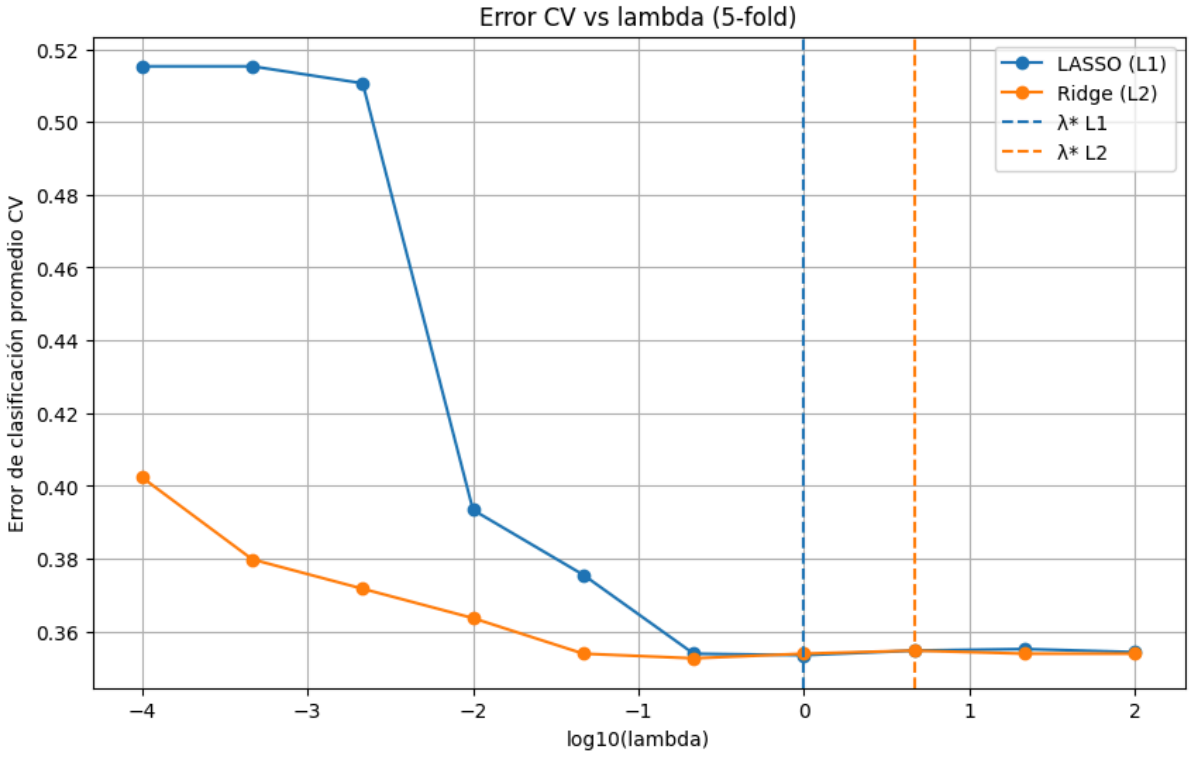
Figura 1. Path de coeficientes modelos Lasso y Ridge



Los paths de coeficientes muestran que a medida que aumenta λ, LASSO (L1) va llevando varios coeficientes a cero, para la selección automática de variables, mientras que Ridge (L2) solo reduce la magnitud de los coeficientes sin eliminarlos, manteniendo todas las variables en el modelo. Ambos métodos coinciden en que la variable más relevante es edad transformada, siendo la última en reducirse al incrementar la penalización. Por tanto, los gráficos muestran que LASSO genera modelos más interpretables, mientras que Ridge es útil si se asume que todas las variables aportan información.

2. Penalidad óptima por Cross-validation y visualización

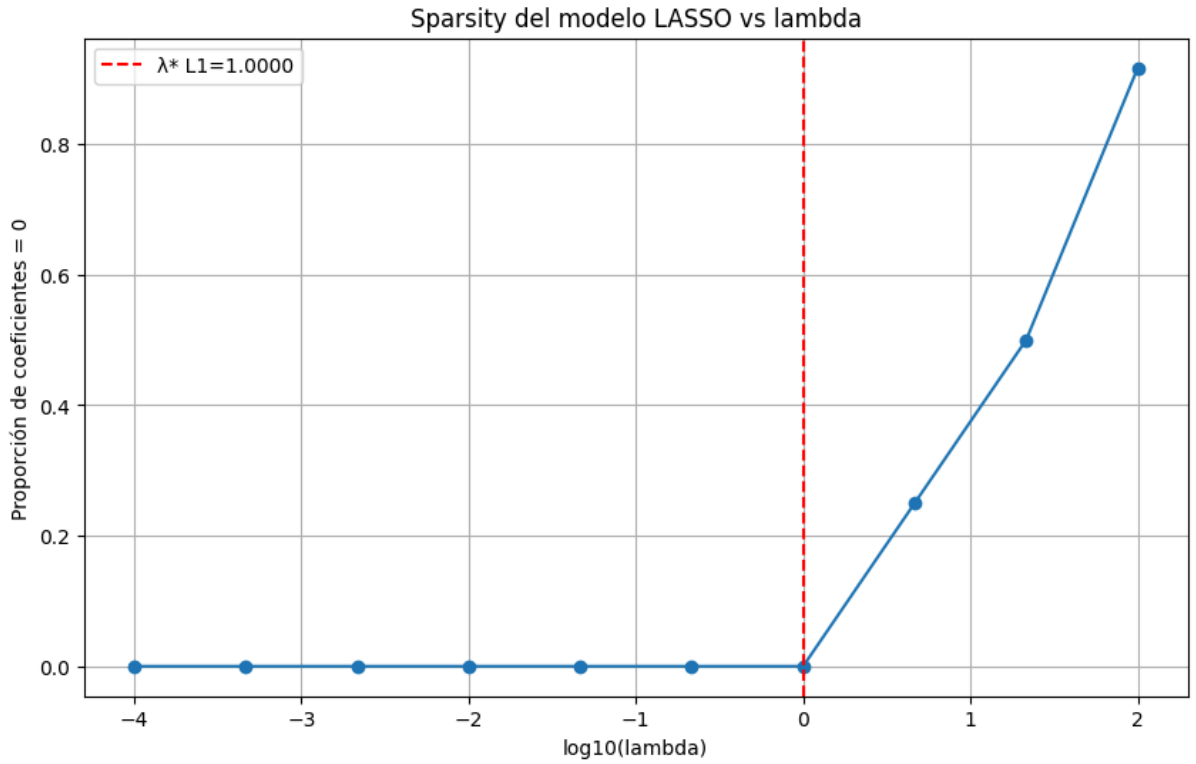
Figura 2. Error de clasificación promedio versus lambda



El gráfico muestra que el modelo Ridge (L2) alcanza un menor error de clasificación promedio que LASSO (L1), lo que sugiere que conservar todos los predictores penalizados

suavemente ofrece mejor desempeño que eliminar algunos. Sin embargo, LASSO sigue siendo útil si se busca un modelo más interpretable y esparso.

Figura 3. Esparcidad del modelo



A medida que aumenta el parámetro de regularización λ, el modelo LASSO elimina más coeficientes, con lo cual sube la sparsity. En λ=1.0 se alcanza un punto de equilibrio donde el modelo conserva las variables más informativas y descarta las menos relevantes, lo que permite tener interpretabilidad sin sacrificar demasiado rendimiento.

3. Estimación con λ^cv y comparación de coeficientes

Tabla 1. Comparación de coeficientes estimados según el método de regularización aplicado

Variable	Logit sin penalidad	LASSO (L1)	Ridge (L2)
AEDUC_cat_Superior_completo_Postgrado	-1.518191	-1.512861	-1.275223
AEDUC_cat_Primaria_completa	0.736556	0.666004	0.618378
AEDUC_cat_Secundaria_incompleta	0.725944	0.656206	0.605078
AEDUC_cat_Primaria_incompleta	0.623145	0.552583	0.505162
CH06_scaled	-0.536889	-0.534499	-0.528934

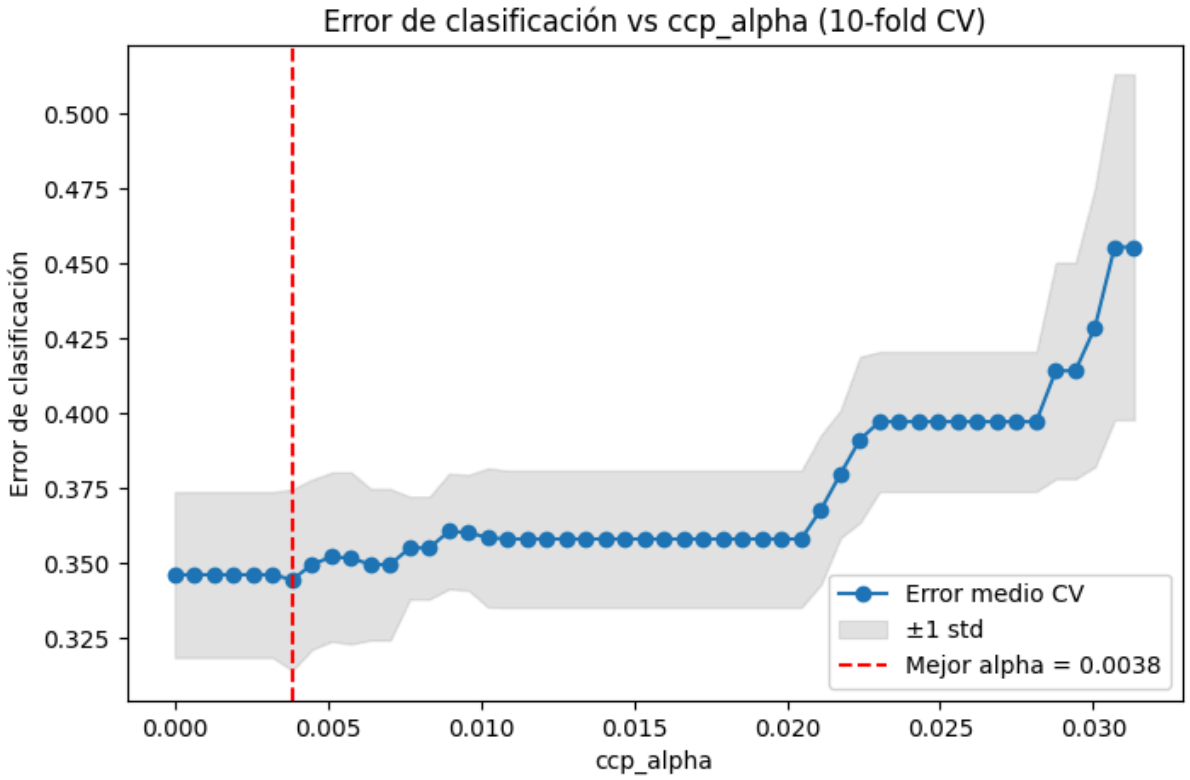
AEDUC_cat_Secundaria_completa	0.415506	0.353836	0.327990
Viudo	-0.332432	-0.273680	-0.268675
HORASTRAB_fulltime	-0.294094	-0.278665	-0.278329
SeparadoDivorciado	0.134654	0.097704	0.108886
AEDUC_cat_Superior_incompleto	-0.124790	-0.165808	-0.191857
CH04	0.025181	0.017547	0.016456
Casado	-0.022509	-0.010069	-0.032897

Al comparar los coeficientes del logit sin penalidad junto con LASSO (L1) y Ridge (L2) usando λ_cv, se observa que la regularización reduce la magnitud de los coeficientes sin cambiar su signo. LASSO tiende a acercar algunos coeficientes a cero y Ridge los contrae suavemente, pero en este caso ninguna variable se elimina, dado que el modelo considera que todas las variables tienen información relevante.

B. Árboles de Decisión

4. Estimen un árbol de decisión podado (CART)

Figura 4. Error de clasificación versus CCP del árbol de decisión

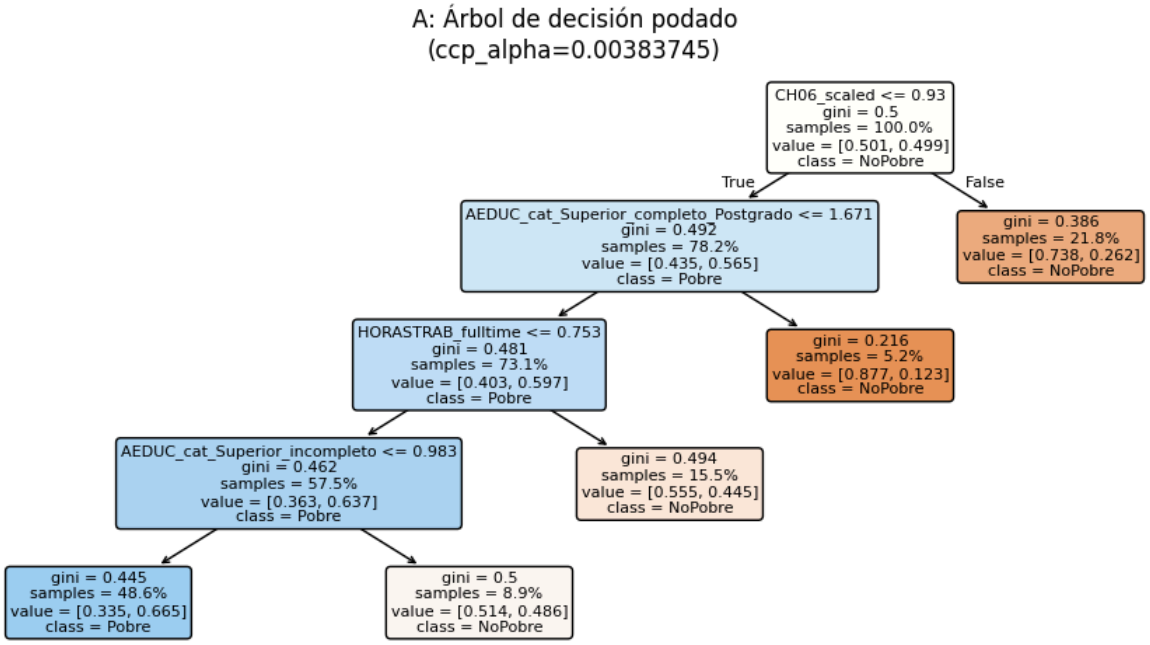


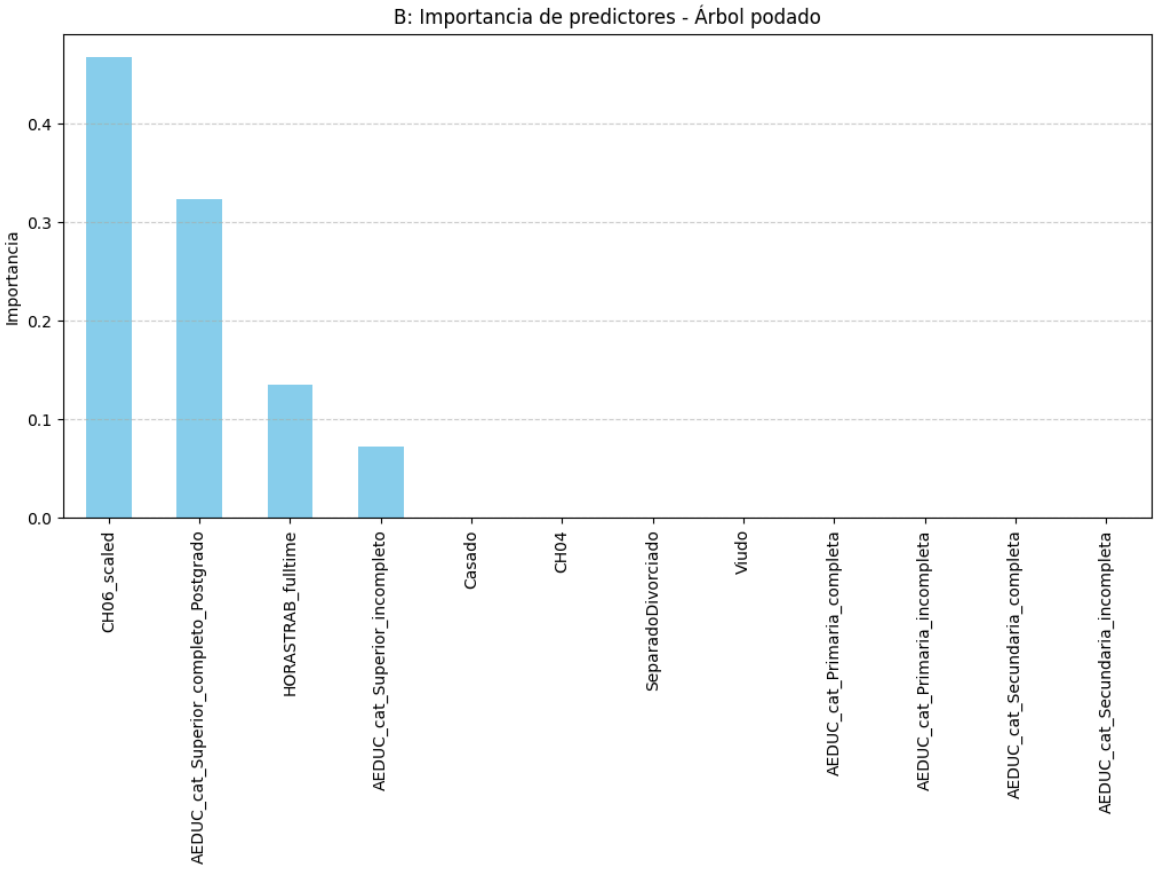
El gráfico del error de clasificación promedio frente a ccp_alpha muestra valores bajos de ccp_alpha (menores a 0.0038), siendo que el error se mantiene prácticamente constante en

torno al 35-36%, lo que indica que el árbol es complejo y presenta poco sesgo pero riesgo de sobreajuste. A partir del ccp_alpha óptimo (0.0038, marcado con la línea roja vertical), el error comienza a aumentar progresivamente a medida que la poda se vuelve más agresiva, reduciendo el tamaño del árbol. Esto representa el óptimo entre sesgo y varianza según la validación cruzada, logrando el menor error de clasificación (aproximadamente de 35.5%). El aumento del error para ccp_alpha mayores a 0.025 confirma la necesidad del pruning para evitar complejidad.

5. Visualización del árbol podado por cross-validation

Figura 5. Gráfico del árbol de decisión e importancia de predictores





El árbol podado final es simple e interpretable, ya que utiliza tres predictores. El nodo raíz es AEDUC_cat_Superior_completo_Postgrado ≤ 0.5 (no tener estudios superiores completos es un factor de riesgo de pobreza), seguido por CH06_scaled (edad estandarizada) y HORASTRAB_fulltime (trabajar jornada completa o no). Esto coincide con la importancia de variables que muestra el panel B, donde esas tres variables dominan y el resto tiene importancia casi nula. En ese sentido, este resultado no coincide con la selección de variables realizada por Lasso, ya que éste mantuvo prácticamente todas las variables.

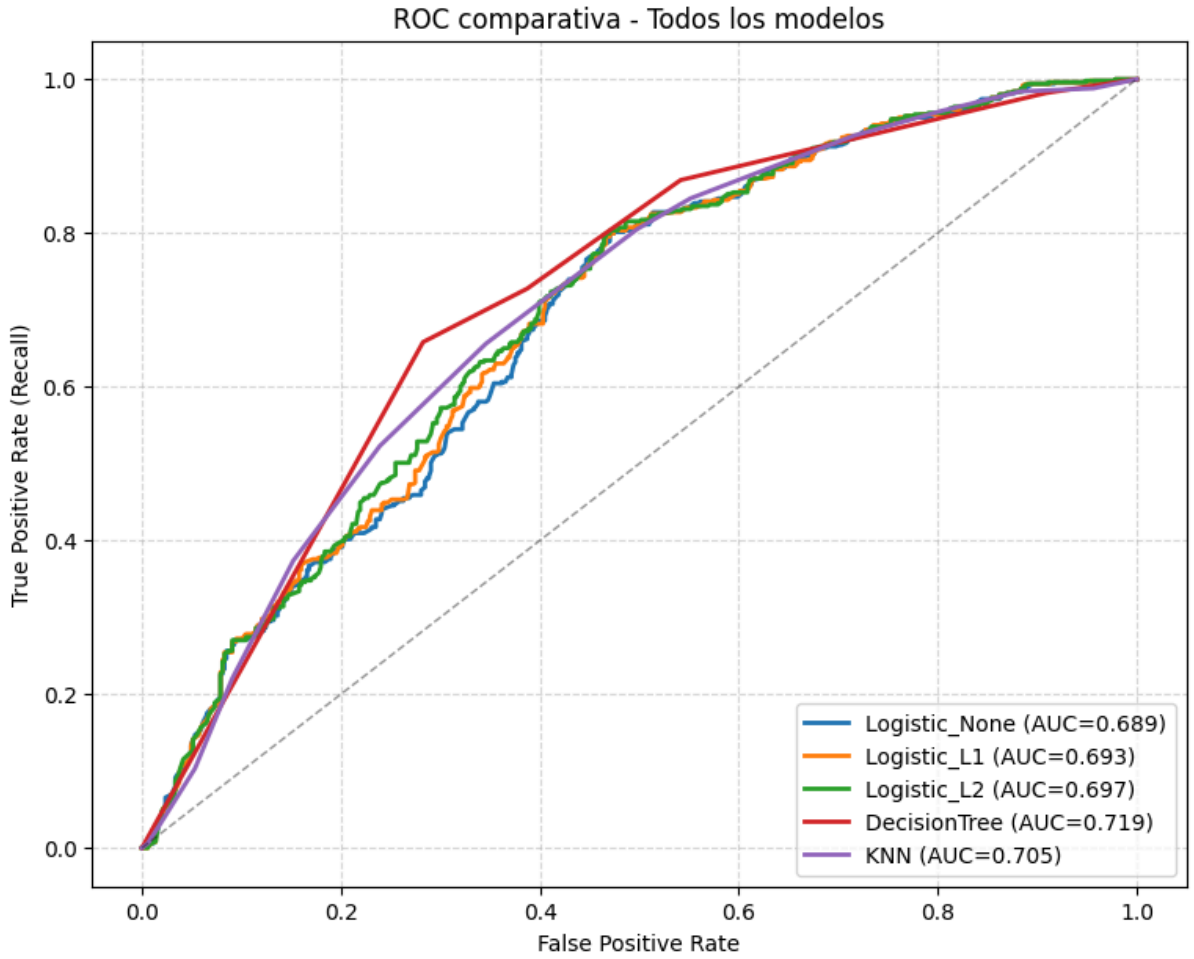
6. Matriz de confusión (p>0,5), la curva ROC y las dos métricas

Tabla 2. Matriz de confusión y reporte de clasificación para cada modelo

Métrica	Logística	Logística L1	Logística L2	Decision Tree	KNN
TN	299	297	299	363	331
FP	207	209	207	143	175
FN	147	140	142	172	173
TP	356	363	361	331	330
Accuracy	0.649	0.654	0.654	0.688	0.655

1-Accuracy	0.351	0.346	0.346	0.312	0.345
F1-score	0.668	0.675	0.674	0.678	0.655
Precision	0.632	0.635	0.636	0.698	0.653
Recall	0.708	0.722	0.718	0.658	0.656
AUC	0.689	0.693	0.697	0.719	0.705

Figura 6. Curva ROC comparativa



La evaluación muestra que el árbol de decisión es el mejor modelo, al tener un mayor accuracy (0.688), menor tasa de error (0.312), mejor precisión (0.698) y mayor AUC (0.719). Los modelos logísticos, con o sin regularización, presentan desempeño similar (accuracy de alrededor de 0.65, y AUC 0.69), mientras que KNN queda ligeramente detrás (accuracy 0.655, AUC 0.705). El árbol reduce en un 11 % la tasa de error respecto al mejor logit y comete menos falsos positivos, lo cual respalda la ventaja de usar un método no lineal para capturar interacciones y efectos no lineales.

Además, el árbol tiene como ventaja que tiene alta interpretabilidad y también precisión, ya que usa pocas variables y tiene reglas claras, lo que permite superar el trade-off entre comunicación y rendimiento. En tal sentido, es una opción adecuada para clasificar

pobreza en la EPH, aunque se debe reconocer que en general los accuracies y AUC se mantienen por debajo de un umbral de 0.75.

7. ¿Cambió su respuesta con respecto a cuál es el "mejor" modelo para asignar recursos escasos a los más necesitados?

Considerando la situación del Ministerio de Capital Humano, el modelo más adecuado para asignar recursos escasos a los hogares pobres es la Regresión Logística con penalización LASSO o Ridge, ya que minimiza los falsos negativos y maximiza el recall de la clase pobre (detectando 140–142 de los hogares pobres frente a 172 del árbol podado). Aunque el árbol es más preciso y comunicable, deja fuera a más beneficiarios reales, lo cual es un factor importante para un objetivo de inclusión social. La recomendación por tanto es usar el logit regularizado como modelo principal, calibrando el umbral para alcanzar un recall ≥ 0.80, y emplear el árbol podado como herramienta de comunicación complementaria para explicar los factores de vulnerabilidad.