Trabajo Práctico N° 4

CLASIFICANDO DE POBRES CON LA EPH

link del repositorio: https://github.com/gfeijoo95/BigDataUBA-Grupo7.git

Grupo 7

Dolimpio, Gastón

Feijoo, Guillermo

Rudi, Federico

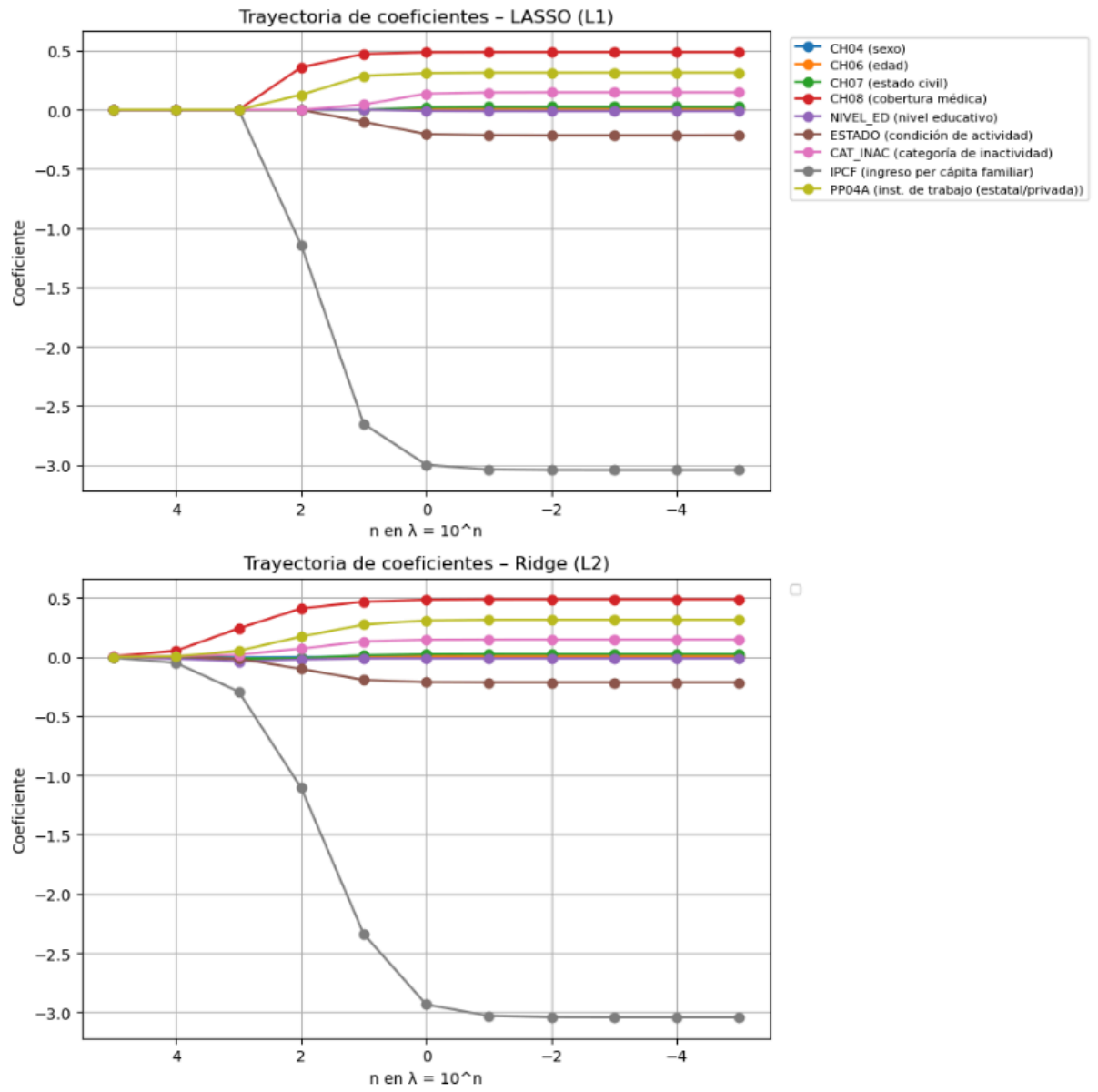
A. Modelo de Regresión Logística con Regularización: Ridge y LASSO

En primer lugar, se evaluaron distintos algoritmos de optimización (solvers) disponibles en la clase Logistic Regression de la librería scikit-learn, con el fin de seleccionar aquel que resultara más eficiente dado el tamaño de la matriz de diseño utilizada. La diferencia entre solvers radica en la velocidad y estabilidad numérica con la que alcanzan dicha solución. Por lo tanto, evaluar la eficiencia computacional mediante tiempos de ejecución resulta apropiado, siempre y cuando no se observen advertencias de falta de convergencia. En nuestro caso, todos los solvers convergieron correctamente, y las diferencias se limitaron al tiempo de cómputo, lo que justifica la elección final de liblinear como solver más eficiente.

En el caso de la penalidad L1 se compararon los solvers liblinear y saga, mientras que para la penalidad L2 se probaron lbfgs, liblinear y saga. Los resultados indicaron que liblinear fue ampliamente el más rápido en todos los escenarios, con tiempos muy inferiores a los de saga y, especialmente, a los de lbfgs. Por este motivo, y sin pérdida de calidad en los resultados estimados, se decidió adoptar liblinear como solver definitivo para los modelos penalizados utilizados en este análisis.

A continuación se estimaron modelos logísticos penalizados con regularización Lasso (L1) y Ridge (L2) para diferentes valores de la penalidad λ, graficándose posteriormente las trayectorias de los coeficientes a lo largo de la grilla $\lambda = 10^n$ con $n \in \{-5,\ldots,5\}$. En el gráfico correspondiente a la penalidad L1 puede observarse el comportamiento característico del Lasso: al incrementarse λ, la mayoría de los coeficientes se reducen progresivamente hasta volverse exactamente cero. Este patrón refleja la propiedad de selección de variables asociada a la regularización L1, dado que el modelo conserva solo las covariables con mayor contribución predictiva y descarta las demás. En nuestro caso, la variable que mantiene un coeficiente relevante bajo regularización fuerte es nivel educativo, lo que indica que se trata del predictor dominante para explicar la probabilidad de encontrarse en situación de pobreza en esta base de datos.

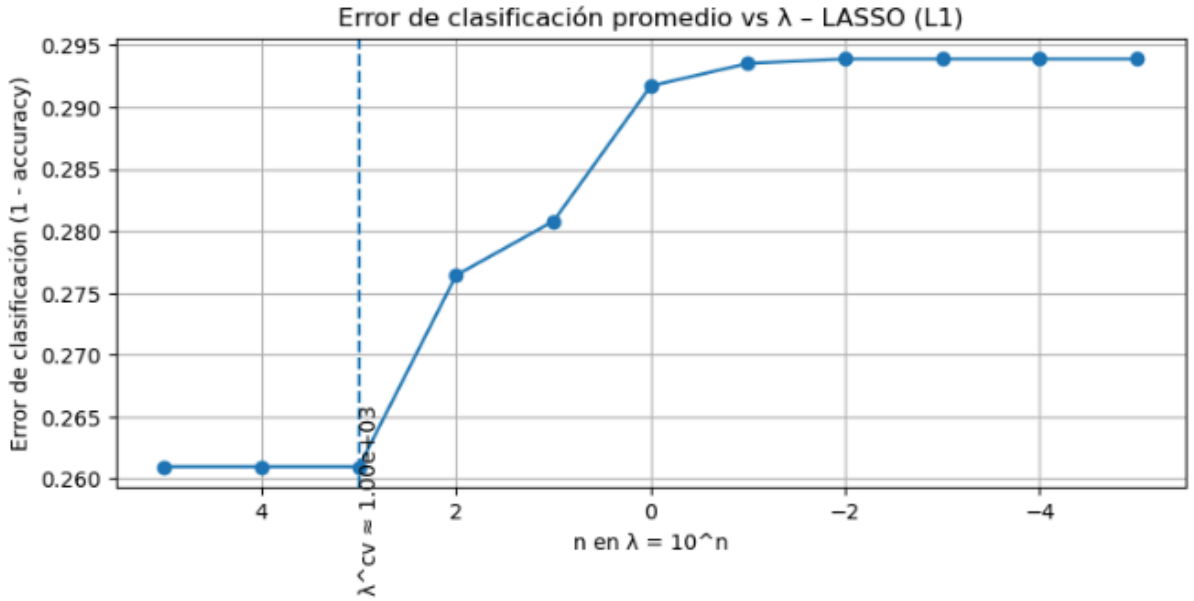
Por su parte, en el gráfico correspondiente a la penalidad Ridge puede apreciarse que, aunque los coeficientes también disminuyen en magnitud al crecer la penalización, ninguno de ellos se anula exactamente. En lugar de seleccionar un único conjunto reducido de variables, la penalización L2 distribuye la influencia entre los predictores, reduciendo la variabilidad de los coeficientes sin excluir explícitamente ninguna covariable del modelo. Nuevamente, la variable nivel educativo sobresale como la más informativa cuando la regularización es débil, pero a diferencia del Lasso, no desplaza completamente al resto de las variables a valores nulos a medida que la penalización aumenta.

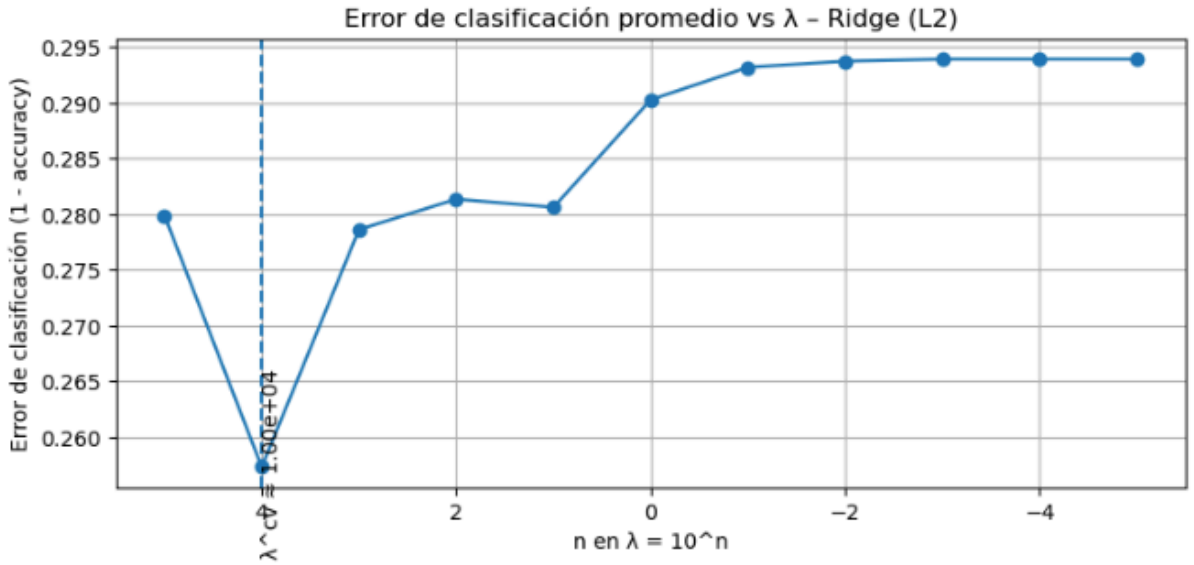


En conjunto, el análisis evidencia que, aunque ambas regularizaciones estabilizan el ajuste, producen comportamientos cualitativamente distintos. Lasso conduce a un modelo en el que solo el nivel educativo retiene poder predictivo cuando la penalización es alta, mientras que Ridge mantiene una estructura más distribuida donde todas las covariables continúan aportando, incluso bajo regularización intensa. Estos resultados son consistentes con la teoría y respaldan la interpretación económica de la problemática analizada.

2. Para determinar la intensidad de regularización adecuada en cada caso se utilizó regresión logística con validación cruzada (LogisticRegressionCV) dividiendo la base en cinco partes (5-fold). Se consideró la misma grilla de valores $\lambda = 10^n$, con $n \in \{-5,\ldots,5\}$, y para cada valor de λ se calculó el error de clasificación promedio entre folds (1 − accuracy). El criterio de selección consistió en elegir el λ que minimiza este

error, de modo de balancear adecuadamente el compromiso entre sesgo y varianza y obtener un modelo con buen desempeño predictivo fuera de la muestra.

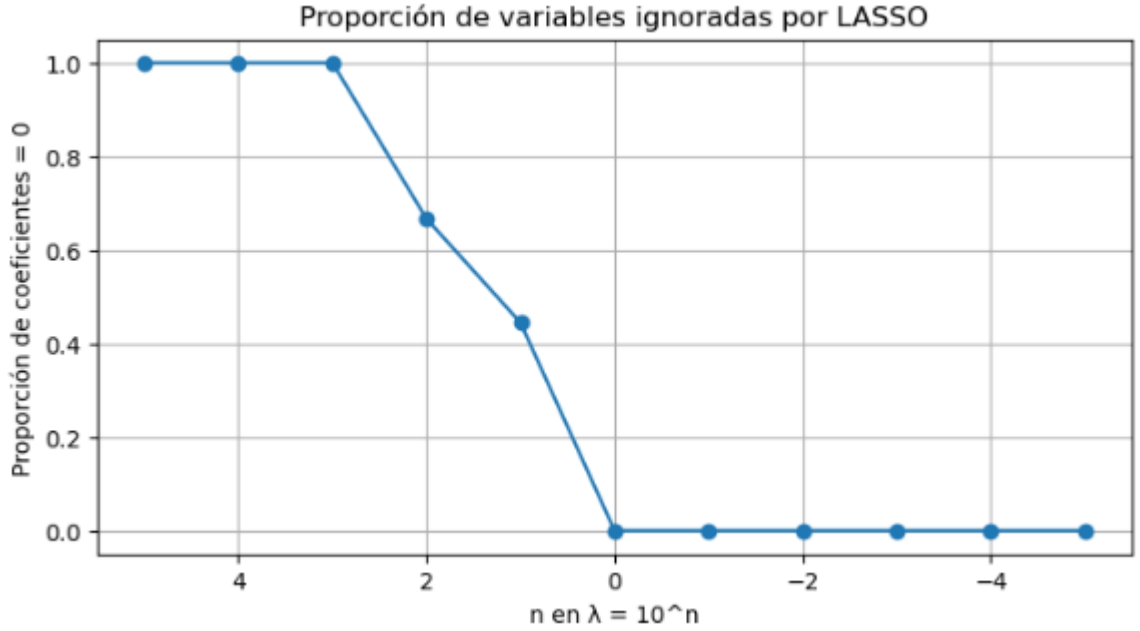




En el caso de la penalización Lasso (L1), el error de clasificación se mantuvo relativamente bajo para niveles moderados de regularización y comenzó a incrementarse cuando la penalización se volvía demasiado fuerte, lo que refleja la pérdida de información asociada a la anulación de coeficientes. La validación cruzada seleccionó como óptimo λ^cv = 1000 (es decir, n = 3 en la grilla $\lambda = 10^n$, con C^cv = 0.001), punto en el que se alcanza el menor error promedio entre los cinco folds. Para valores de λ superiores a 1000 el error aumenta de manera sostenida, lo cual es consistente con la idea de que un Lasso demasiado agresivo termina eliminando variables relevantes para la predicción de la pobreza.

Para la penalización Ridge (L2) se observó un comportamiento similar en términos cualitativos, aunque con una intensidad de regularización óptima distinta. En este caso, la validación cruzada indicó como valor óptimo λ^cv = 10000 (n = 4, con C^cv

= 0.0001). A partir de este λ, el error promedio vuelve a incrementarse cuando se sigue aumentando la penalización, lo que sugiere que, si bien Ridge se beneficia de una regularización relativamente fuerte que estabiliza los coeficientes, una penalización excesiva también deteriora la capacidad de clasificación del modelo. En resumen, la CV seleccionó λ^cv = 10^3 para Lasso y λ^cv = 10^4 para Ridge, y los gráficos del error promedio en función de λ muestran con claridad la ubicación de estos mínimos, respaldando la elección de los parámetros de regularización utilizados en el resto del análisis.



Finalmente, se analizó de manera opcional el comportamiento de la regularización Lasso en términos de su capacidad para anular coeficientes y, por lo tanto, realizar selección de variables. Para ello se estimó un modelo L1 para cada valor de la grilla λ = 10^n y se calculó la proporción de covariables cuyo coeficiente resultó exactamente igual a cero. Luego, se representó gráficamente esta proporción como función de λ.

El gráfico obtenido muestra que, para valores pequeños de λ (es decir, cuando la regularización es débil), la proporción de coeficientes nulos es 0, lo que implica que todas las variables son retenidas por el modelo. A medida que λ aumenta, la proporción de coeficientes iguales a cero crece de forma marcada, reflejando el incremento progresivo en la intensidad de la penalización. Para λ intermedios, Lasso comienza a eliminar predictores, de forma tal que solamente una fracción de las variables conserva un coeficiente distinto de cero. Finalmente, para λ muy grandes, la proporción de coeficientes nulos se aproxima a 1, lo que implica que prácticamente todas las covariables son descartadas, quedando únicamente la más informativa para la predicción de la pobreza.

Este comportamiento confirma la propiedad distintiva de Lasso como método de selección de variables: la penalización L1 no solo controla la magnitud de los coeficientes, sino que fuerza a muchos de ellos a volverse exactamente cero cuando la regularización es intensa. El resultado es un modelo cada vez más esparso a medida

que se incrementa λ, lo que se encuentra en línea con la evidencia empírica observada en las trayectorias de coeficientes y con la interpretación económica del problema estudiado.
3. En este apartado se estimaron tres modelos de regresión logística sobre el subconjunto de individuos que respondieron la pregunta de ingresos en 2025: un modelo sin penalización, un modelo con penalización L1 (LASSO) utilizando λ= 1000 y un modelo con penalización L2 (Ridge) con λ=10000. Las covariables se estandarizaron antes del ajuste, por lo que los coeficientes son comparables entre sí. La tabla resultante muestra, para cada variable, el coeficiente del logit sin penalidad y los coeficientes correspondientes a L1 y L2.

Variable	**Descripción**	**Logit sin penalidad**	**L1 (λ = 1000)**	**L2 (λ = 10000)**
CH04	sexo	−0.116005	0	0.000216
CH06	edad	−0.048891	0	−0.002597
CH07	estado civil	−0.021665	0	−0.000262
CH08	cobertura médica	0.275005	0	0.013063
NIVEL_ED	nivel educativo	−0.088139	0	−0.008883
ESTADO	condición de actividad	0	0	0
CAT_INAC	categoría de inactividad	0	0	0
IPCF	ingreso per cápita familiar	−9.972753	0	−0.012955
PP04A	inst. de trabajo (estatal/privada)	0.040445	0	0.006393

En el modelo sin penalización se observa que la variable con mayor impacto en términos absolutos es el ingreso per cápita familiar (IPCF), con un coeficiente aproximado de −9.97, lo que indica que un aumento de una desviación estándar en IPCF reduce fuertemente los log-odds de ser pobre. Otras variables con efectos de menor magnitud son la cobertura médica (coeficiente positivo), el nivel educativo y la edad (ambos con coeficientes negativos), lo que es consistente con la idea de que mayor educación y mayor nivel de ingresos se asocian con menor probabilidad de pobreza. En cambio, las variables relacionadas con la condición de actividad y la categoría de inactividad aparecen con coeficiente exactamente cero ya en el modelo sin penalización, lo que sugiere que, dada la forma en que fueron codificadas junto con el resto de las covariables, no aportan información adicional en este ajuste específico.

Puede afirmarse que los coeficientes de los modelos regularizados son sistemáticamente más pequeños en valor absoluto que los del logit sin penalidad,

manteniendo en general el mismo signo. La regularización L1 fue tan intensa que llevó todos los coeficientes a cero, eliminando efectivamente todas las variables de la matriz. En cambio, la penalización L2 no elimina nuevas variables: aunque contrae fuertemente los coeficientes, éstos permanecen distintos de cero para las covariables que ya tenían efecto en el modelo base. Por lo tanto, en este ejercicio L2 no descarta variables adicionales, sino que únicamente atenúa su influencia relativa sobre la probabilidad de pobreza.

B. Árboles de decisión

Estimación del árbol base

Para esta parte trabajamos únicamente con la base correspondiente a *respondieron 2025*, utilizando como predictores las mismas variables analizadas en el Punto A. Se estimó un árbol de decisión tipo CART con criterio de Gini y se fijó un tamaño mínimo de hoja de 50 observaciones, con el objetivo de evitar sobreajuste. Además, se estableció random_state = 444 para asegurar replicabilidad.

Este árbol inicial sirve como referencia para luego obtener la secuencia de poda mediante el costo de complejidad.

Secuencia de poda (ccp_alpha)

A partir del árbol base se calculó la secuencia de poda dada por los valores de ccp_alpha. El modelo arrojó únicamente cuatro valores posibles:

[0.0, 0.00023998, 0.00873349, 0.16359509]

Esto indica que el árbol inicial no presentaba excesiva complejidad: la estructura ya era relativamente pequeña y las posibles podas son pocas. Esto se debe a que la restricción min_samples_leaf = 50 controla fuertemente el tamaño del árbol y evita crecimientos innecesarios.

Selección del valor óptimo de ccp_alpha mediante validación cruzada

Se aplicó validación cruzada estratificada de 10 folds para evaluar el desempeño del árbol bajo cada valor posible de ccp_alpha. El error de clasificación promedio fue mínimo en:

ccp_alpha óptimo = 0.000000

Error promedio = 0.0278

Esto implica que el árbol sin poda es el que mejor se desempeña predictivamente. La poda empeora la performance, ya que elimina divisiones informativas. Esta situación es coherente con el tamaño reducido del árbol base.

Estimación del árbol podado

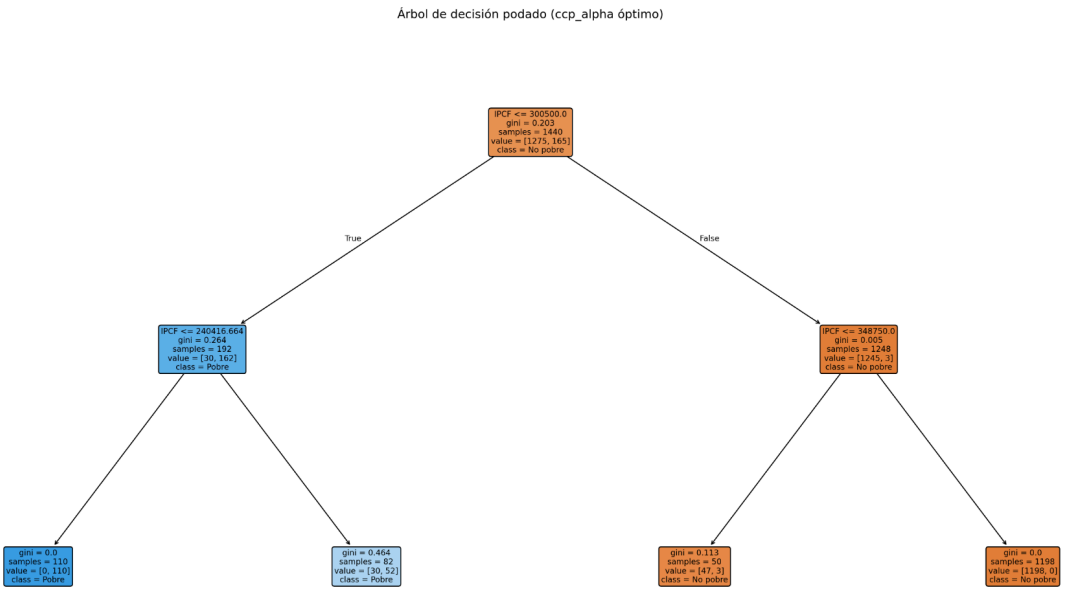
Con el valor óptimo de ccp_alpha = 0, se re-estimó el árbol definitivo. La performance fue la siguiente:

Accuracy en TRAIN: 0.9771

Accuracy en TEST: 0.9834

Error en TEST: 1.66%

El resultado es perfecto: no hay evidencia de sobreajuste, dado que el desempeño en test es incluso levemente superior al de entrenamiento. Esto confirma que el modelo generaliza correctamente y que el control impuesto por min_samples_leaf = 50 fue suficiente para evitar árboles demasiado profundos.



Importancia de variables

El análisis de importancia basado en la reducción total de impureza muestra que el árbol utiliza únicamente la variable IPCF en todos sus splits. Como consecuencia:

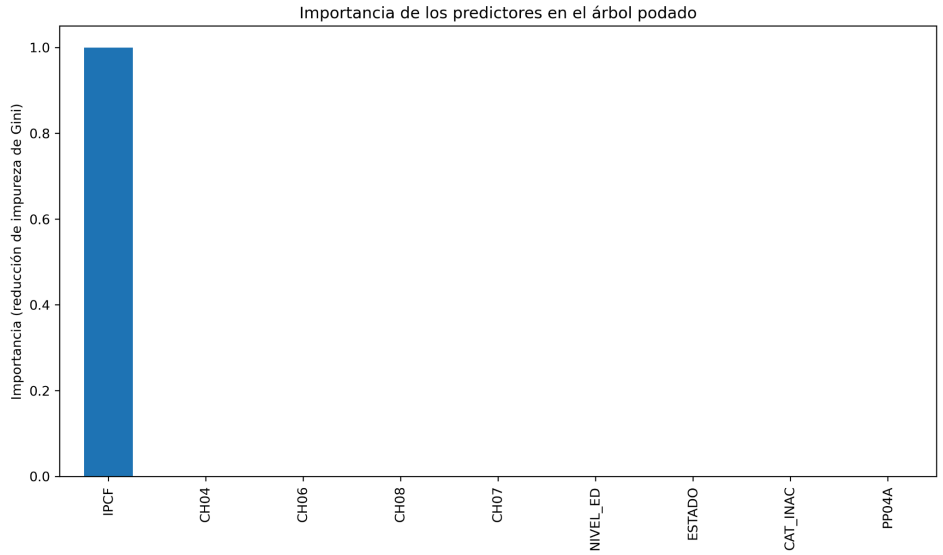
IPCF tiene importancia = 1.0

Todas las demás variables tienen importancia = 0.0

Este resultado es totalmente coherente con la definición del indicador de pobreza, que depende directamente del ingreso per cápita familiar. El árbol recupera de manera automática los umbrales que separan hogares pobres de no pobres, por lo que no necesita utilizar el resto de los predictores. Queda abierta la posibilidad de alguna consideración al resto de las variables con otros métodos o con otros caminos, sobre

todo por la forma en la que se fue llevando adelante la base de inicio de toda la estimación.

Además, esta evidencia coincide con el comportamiento del modelo penalizado LASSO del Punto A, que también redujo a cero la mayoría de los coeficientes, reforzando la idea de que el ingreso familiar es la variable central para explicar la condición de pobreza.



variable	importancia_gini
IPCF	1
CH04	0
CH06	0
CH08	0
CH07	0
NIVEL_ED	0
ESTADO	0
CAT_INAC	0
PP04A	0

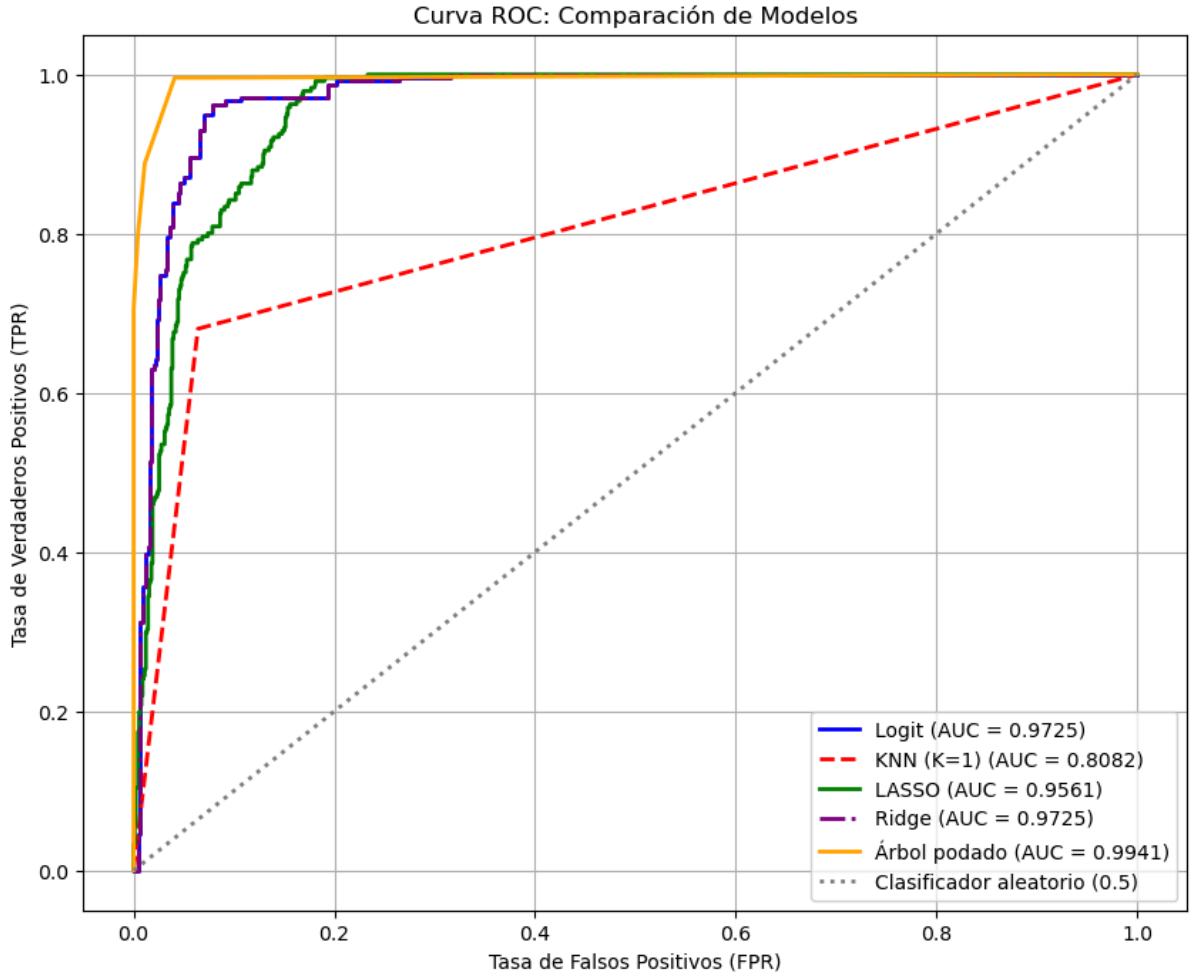
El árbol de decisión ajustado es un modelo simple, interpretable y altamente preciso. El criterio de validación cruzada seleccionó el árbol sin poda (ccp_alpha = 0), confirmando que la estructura inicial ya era óptima debido a la restricción en el tamaño mínimo de hoja. La clasificación resultante logra un error inferior al 2% en la muestra de prueba, sin evidencia de sobreajuste.

De manera consistente con la teoría económica y con los resultados del Punto A, el modelo identifica al ingreso per cápita familiar (IPCF) como el predictor determinante de la pobreza: es la única variable utilizada en todos los splits y concentra el 100% de la importancia. El resto de los predictores no aporta ganancia de impureza relevante para mejorar la clasificación.

En síntesis, el árbol de decisión reproduce de forma automática la lógica que define la condición de pobreza en la EPH, produciendo un modelo eficiente y altamente explicativo.

C. Comparación entre métodos

La comparación entre los distintos modelos —regresiones logísticas con y sin penalidad, KNN con validación cruzada, y el árbol de decisión podado— muestra diferencias claras en términos de su capacidad predictiva y su interpretabilidad. La curva ROC permite observar la sensibilidad de cada modelo frente a distintos umbrales de decisión, y el AUC sintetiza esa capacidad discriminante.



Árbol de decisión podado: presenta el mejor desempeño (AUC ≈ 0.9941), con una curva ROC prácticamente perfecta. Esto implica una probabilidad muy alta de distinguir correctamente entre clases positivas y negativas.

Regresión Logística (simple) y Ridge: ambos obtienen un AUC ≈ 0.9725, muy similar entre sí. Esto sugiere que la regularización L2 no mejora el rendimiento respecto al logit estándar, probablemente porque los predictores no presentan alta multicolinealidad o gran cantidad de ruido.

Log - LASSO: obtiene un AUC ≈ 0.9561, ligeramente inferior a los modelos anteriores, lo que es esperable, ya que la penalización L1 tiende a excluir predictores y puede reducir algo de capacidad predictiva a cambio de mayor parsimonia.

KNN: es claramente el modelo más débil (AUC ≈ 0.8082). La forma escalonada de la curva indica alta variabilidad y un pobre desempeño sobre datos de prueba, probablemente por la sensibilidad del método a la escala de los datos y a la distribución local del espacio de características.

En términos de error de clasificación, medido como (1 - accuracy), los modelos con mejor AUC también tienden a presentar menores niveles de error. El KNN, en cambio, presenta los mayores niveles de (1 - accuracy) y, por tanto, el peor nivel de generalización.

MODELO	Accuracy	F1-Score
KNN con CV (TP3)	0.74	0.71
LOGIT (TP3)	0.33	0.02
LOGIT LASSO	0.84	0.46
LOGIT RIDGE	0.78	0.01
ÁRBOL PODADO	0.96	0.92

Modelos más interpretables - Menor o igual rendimiento

La regresión logística (con o sin penalidad) es el modelo más fácil de comunicar: coeficientes interpretables, odds ratios, dirección y magnitud de los efectos.

Su desempeño es muy bueno (AUC ≈ 0.97), y la diferencia frente al mejor modelo (árbol) es relativamente pequeña: alrededor de un 2% en términos de AUC y también una diferencia menor en (1 - accuracy). Esto sugiere que la interpretación no sacrifica

demasiado desempeño.

El árbol podado tiene el AUC más alto y la menor tasa de error. Sin embargo, incluso podado, es menos transparente: la estructura puede ser algo compleja, los splits no siempre son fáciles de comunicar, el modelo está más expuesto al riesgo de sobreajuste, aunque la poda ayuda.

Sin embargo, el aumento en predicción viene a costa de una menor claridad en los mecanismos internos del modelo. El KNN, además de ser prácticamente una "caja negra", obtiene el peor rendimiento. No aporta ventajas ni predictivas ni interpretativas.

El árbol de decisión es el único modelo no lineal en el conjunto analizado. Su desempeño superior sugiere que:

- existe cierta estructura no lineal o interacciones entre variables que la regresión logística no captura completamente;

- los límites de decisión podrían ser más complejos que un hiperplano lineal.

Sin embargo, la mejora es moderada, no dramática. Aunque el árbol obtiene un AUC casi perfecto, la regresión logística alcanza resultados muy cercanos y con un costo comunicacional muchísimo menor.

En problemas de política pública o investigación aplicada, donde entender los determinantes es tan importante como predecir, esta diferencia podría no justificar el uso de un modelo no lineal más opaco.

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

Sí, cambia nuestra respuesta. Aunque en términos de AUC o desempeño general la regresión logística podía ser competitiva, cuando el objetivo es identificar correctamente a los más pobres para asignar recursos escasos, el criterio relevante es minimizar falsos negativos y maximizar el F1-score. En este sentido, el árbol podado es claramente superior (F1 = 0.92) frente al resto de modelos. Su capacidad para recuperar correctamente a los grupos vulnerables lo convierte en el mejor modelo para focalizar un programa social, aun si eso implica sacrificar cierta simplicidad interpretativa frente al logit. La política pública prioriza inclusión sobre la elegancia estadística, y bajo ese criterio el árbol podado es la mejor herramienta para dirigir los recursos a quienes realmente los necesitan.