Lic. en Ciencias del Comportamiento

Trabajo Práctico N° 4

Alumnas
Agustina Kiessner
Isabela Nicola
Clara Mollón

Profesores
María Noelia Romero
Tomas Enrique Buscaglia
Ignacio Anchorena

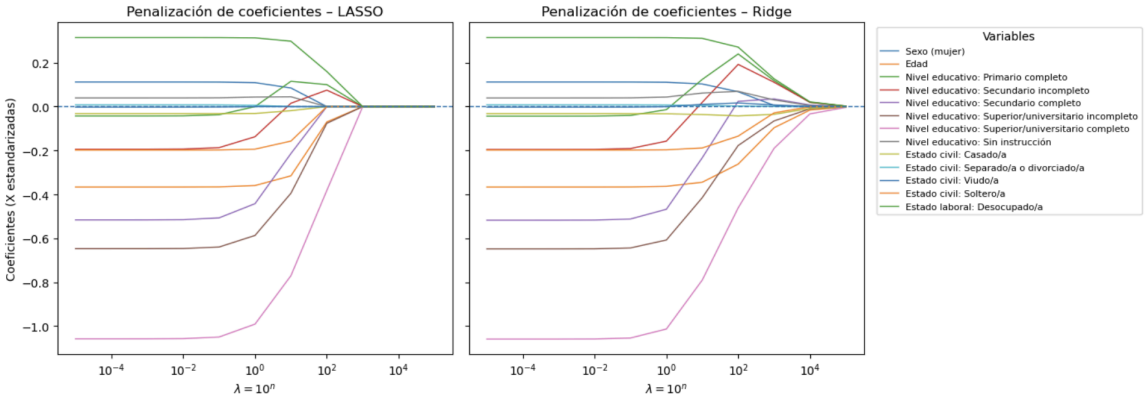
Asignatura
Ciencia de datos - Tutorial 3

Fecha de presentación
28 de Noviembre de 2025

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

1.

Figura 1



Nota. Se utilizó el *solver liblinear*, ya que es compatible con LASSO y Ridge. Como el modelo no presentó problemas de convergencia, no fue necesario utilizar otros *solvers*.

En la Figura 1 se muestra cómo los coeficientes estimados de la regresión logística penalizada varían a medida que aumenta la penalización, bajo LASSO y Ridge, utilizando $\lambda = 10^n$ con $n \in [-5, 5]$. Todas las variables explicativas fueron previamente estandarizadas y se removió el intercepto para garantizar comparabilidad entre coeficientes. Los predictores incluidos corresponden a sexo, edad, nivel educativo (*dummies*), estado civil (*dummies*) y condición laboral.

En ambos modelos se observa el patrón esperado de *shrinkage*, al incrementarse λ, los coeficientes se reducen de manera progresiva y convergen hacia cero, lo que refleja cómo la penalización restringe la magnitud de los parámetros. Este encogimiento reduce la varianza del estimador y ayuda a controlar el sobreajuste, a costa de introducir un mayor sesgo, mostrando el *trade-off* sesgo-varianza.

La diferencia central entre penalidades es clara en la figura 1. LASSO tiende a llevar varios coeficientes exactamente a cero cuando λ crece, realizando una selección explícita de variables. En la figura 1 se aprecia que diversas categorías de nivel educativo y ciertas categorías de estado civil pierden rápidamente su contribución al modelo, mientras que predictores como sexo (mujer) y edad muestran comportamientos más estables, lo que sugiere una contribución relativamente más robusta al modelo incluso bajo regularización fuerte. Por el contrario, Ridge contrae todos los coeficientes de forma gradual pero sin anular ninguno, de modo que todas las curvas descienden suavemente hacia cero sin llegar a eliminar predictores, manteniendo todas las variables en el modelo.

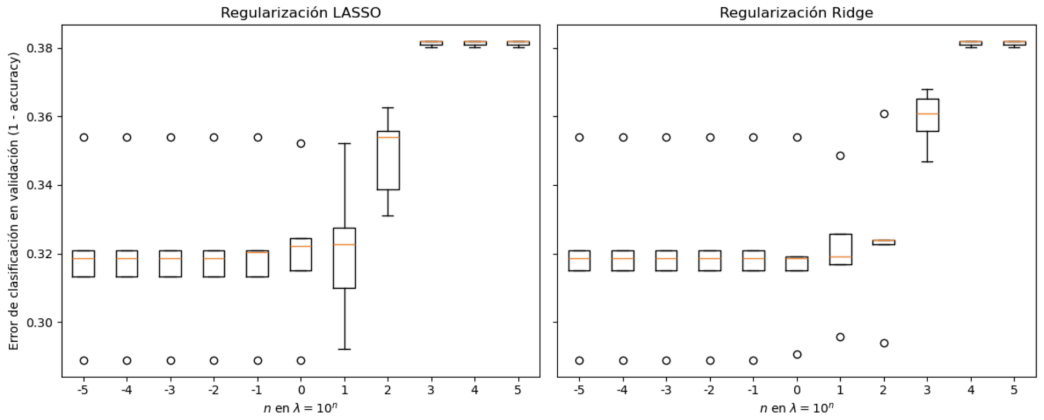
2.

La selección de la penalidad óptima para los modelos logísticos regularizados se realizó mediante validación cruzada estratificada de 5 particiones. La búsqueda se llevó a cabo sobre la grilla definida previamente, con valores de lambda expresados como 10^n para n entre -5 y 5. En este estimador, la penalización se controla a través del parámetro C, que es igual al inverso de lambda ($C = 1/\lambda$). Los valores óptimos seleccionados para cada modelo se resumen en la Tabla 1.

Tabla 1

Modelo	λ^{cv} óptimo	C = 1/λ	Error de clasificación (CV)
LASSO	10^{-5}	100000	0.3191
Ridge	1.0	1.0	0.3195

Figura 2



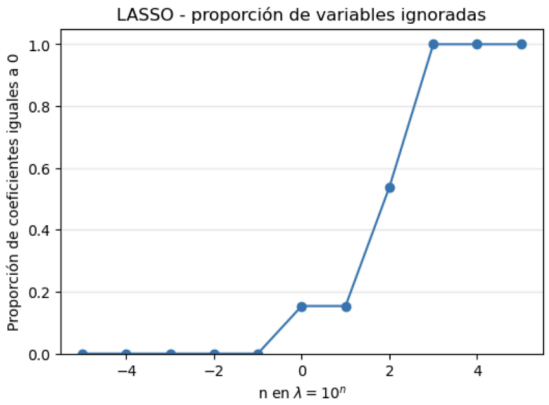
Nota. Dado que trabajamos con regresión logística y la variable objetivo es binaria, utilizamos como métrica de desempeño el error de clasificación, definido como $1-$*accuracy*, que representa la proporción de observaciones mal clasificadas en cada *fold*.

Para cada valor de lambda, el modelo se entrenó sobre cuatro particiones y se evaluó en la restante, obteniendo cinco medidas de desempeño por cada nivel de penalización. La selección del parámetro óptimo se basó en el error de clasificación promedio, es decir, la proporción de observaciones mal predichas en cada fold. Esta métrica resulta adecuada para modelos logísticos, dado que la variable dependiente es binaria y el objetivo es distinguir correctamente entre ambas categorías.

En el caso de LASSO, la validación cruzada seleccionó una penalidad extremadamente chica, $\lambda^{cv} = 10^{-5}$ (equivalente a C = 100000). El error de clasificación promedio asociado a este valor fue 0.3191, el más bajo dentro de la grilla considerada. Los boxplots muestran que, para valores de lambda muy pequeños (penalización débil), el modelo mantiene un desempeño estable y favorable. Sin embargo, a medida que λ aumenta, el error de clasificación crece de forma pronunciada, indicando un deterioro en la capacidad predictiva del modelo debido a una regularización excesiva.

En el caso de Ridge, la validación cruzada seleccionó una penalidad intermedia $\lambda^{cv} = 1$ (C = 1). El error de clasificación promedio correspondiente fue 0,3195, muy similar al obtenido con LASSO. Los boxplots de la Figura 2 muestran un patrón coherente, para valores moderados de λ el error se mantiene relativamente bajo y estable, mientras que penalizaciones más fuertes incrementan el error de clasificación, sugiriendo que el modelo comienza a subajustar. Ambos modelos alcanzan desempeños comparables en validación, aunque LASSO ofrece la ventaja adicional de poder anular coeficientes y de facilitar la selección de variables.

Figura 3



Nota. En el eje vertical se muestra la proporción de coeficientes exactamente iguales a cero. Un valor igual a 1 implica que LASSO anuló todos los coeficientes asociados a las variables explicativas.

Como análisis complementario, se examinó la proporción de variables con coeficiente exactamente igual a cero para cada nivel de penalización. El patrón observado es el esperado para LASSO, con penalizaciones bajas prácticamente no se eliminan predictores, mientras que para valores mayores (por encima de 10^2) la proporción de coeficientes nulos crece abruptamente hasta alcanzar el 100%, indicando que el modelo anula todos los coeficientes asociados a las variables explicativas.

3.

Tabla 2

	Coeficientes sin penalidad	Coeficientes Lasso	Coeficientes Ridge
Sexo (mujer)	0.1121	0.1121	0.1110
Edad	-0.3662	-0.3662	-0.3634
Nivel educativo: Primario completo	-0.0434	-0.0428	-0.0140
Nivel educativo: Secundario incompleto	-0.1950	-0.1941	-0.1577
Nivel educativo: Secundario completo	-0.5173	-0.5161	-0.4681
Nivel educativo: Superior/universitario incompleto	-0.6481	-0.6472	-0.6083
Nivel educativo: Superior/universitario completo	-1.0589	-1.0578	-1.0141
Nivel educativo: Sin	0.0394	0.0395	0.0434

instrucción			
Estado civil: Casado/a	-0.0322	-0.0322	-0.0328
Estado civil: Separado/a o divorciado/a	0.0078	0.0078	0.0073
Estado civil: Viudo/a	-0.0024	-0.0024	-0.0002
Estado civil: Soltero/a	-0.1979	-0.1979	-0.1969
Estado laboral: Desocupado/a	0.3146	0.3146	0.3143

Estimamos tres modelos logísticos utilizando la matriz X_train: un modelo sin penalidad, un modelo con penalización por LASSO y otro con penalización por Ridge, empleando en ambos casos los valores óptimos de lambda obtenidos por validación cruzada en el ejercicio anterior. A partir de estas estimaciones construimos una tabla comparativa que muestra, para cada variable, los coeficientes sin penalizar y los coeficientes ajustados bajo LASSO y Ridge.

En todos los predictores observamos el efecto esperado de la regularización, tanto en LASSO como en Ridge se reduce la magnitud de los coeficientes en relación con el modelo sin penalidad. En este caso, la penalización de LASSO no llevó ningún coeficiente exactamente a cero, lo cual es coherente con que el lambda óptimo seleccionado para LASSO fue extremadamente pequeño. Por su parte, la penalización Ridge produjo un encogimiento suave y sistemático en todos los coeficientes, pero sin eliminar variables del modelo (ningún coeficiente se vuelve exactamente cero).

La regularización atenúa la fuerza de las asociaciones estimadas, aunque sin modificar su signo ni su interpretación sustantiva, y la penalización Ridge, a diferencia de LASSO, no anula predictores sino que únicamente reduce su magnitud.

4.

En esta consigna, se estimó un árbol de decisión CART, eligiendo el hiperparámetro de costo de complejidad del árbol (ccp_aplha) como criterio de poda. Se generó una grilla completa de valores posibles generada por el árbol inicial sin podar, y para cada uno se evaluó su desempeño mediante 10-fold *cross validation*.

En la Figura 4 se muestra el accuracy por medio de validación cruzada en función de ccp_alpha. Se puede ver que para valores muy chicos de ccp_alpha, el modelo mantiene una performance relativamente alta y estable con *accuracy* alrededor de 0.66 y 0.67, lo que es habitual cuando el árbol completo tiene muchas divisiones pequeñas, ya que los valores chicos producen podas muy leves y el desempeño apenas varía. Por esta razón, pueden verse muchos puntos acumulados en la zona izquierda del gráfico. Cuando aumenta el valor de ccp_alpha, el árbol empieza a podarse más, eliminando ramas que aportan información útil (baja el accuracy). El ccp_alpha seleccionado por CV fue de 0.00123 (Ver Tabla 3). Con ese valor, se entrenó el árbol podado, el cual obtuvo un *accuracy* en train de 0.7 y en *test* de 0.682 (Ver Tabla 3). La cercanía entre estos valores sugieren que el árbol podado generaliza

mejor que el árbol completo y reduce el sobreajuste atípico de un CART sin poda.

Figura 4

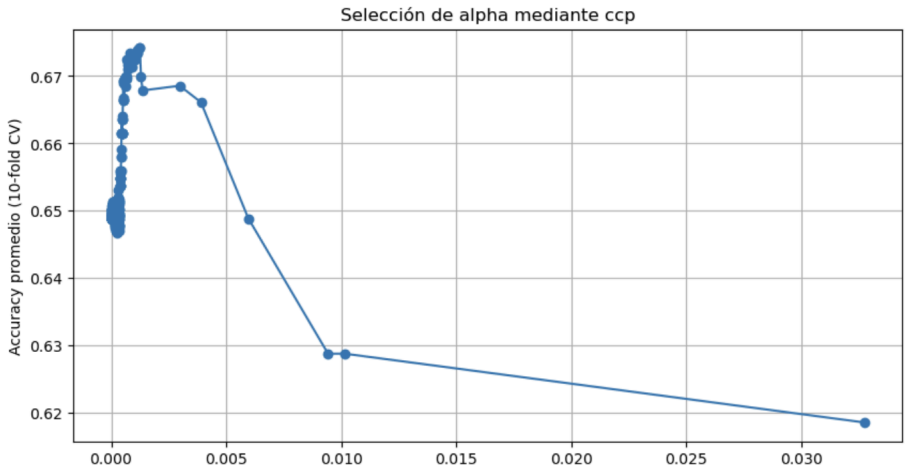


Tabla 3

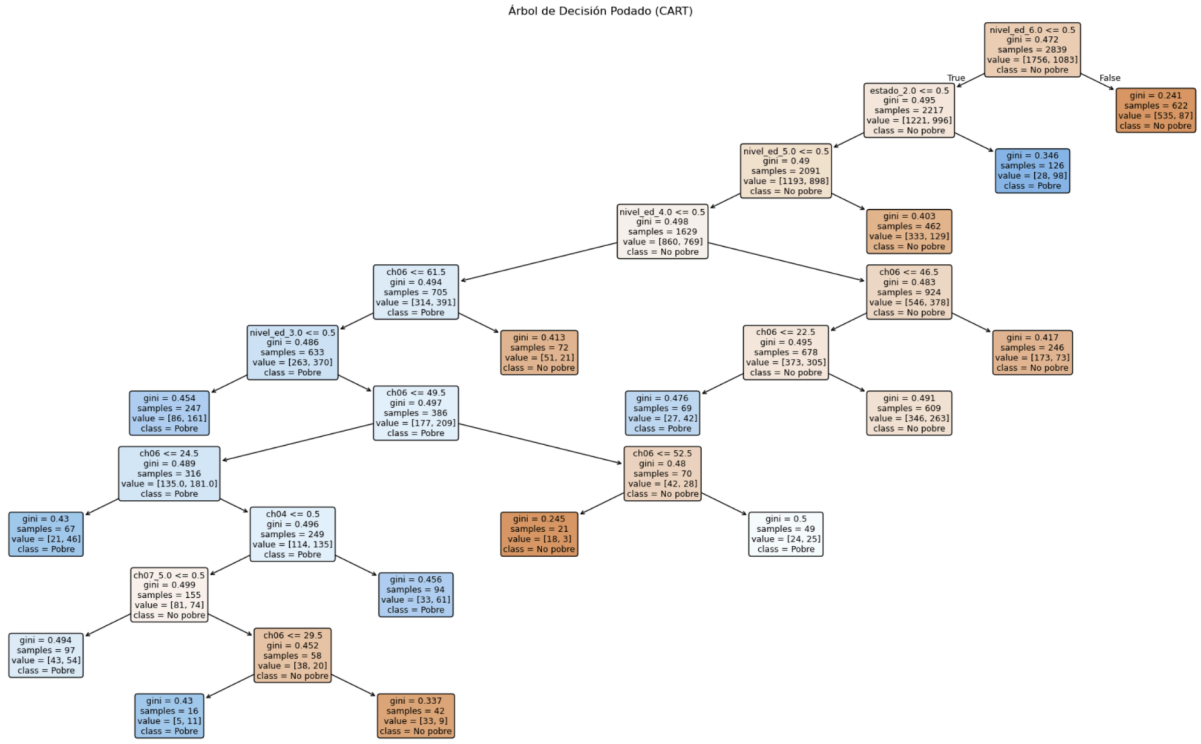
Métrica	Valor
Cantidad de valores de ccp_alpha	312
Mejor ccp_alpha (10-fold CV)	0.001239
Accuracy en entrenamiento	0.700
Accuracy en test	0.682

Nota. En la Tabla 3 se presentan los resultados del árbol de decisión podado mediante validación cruzada.

5.

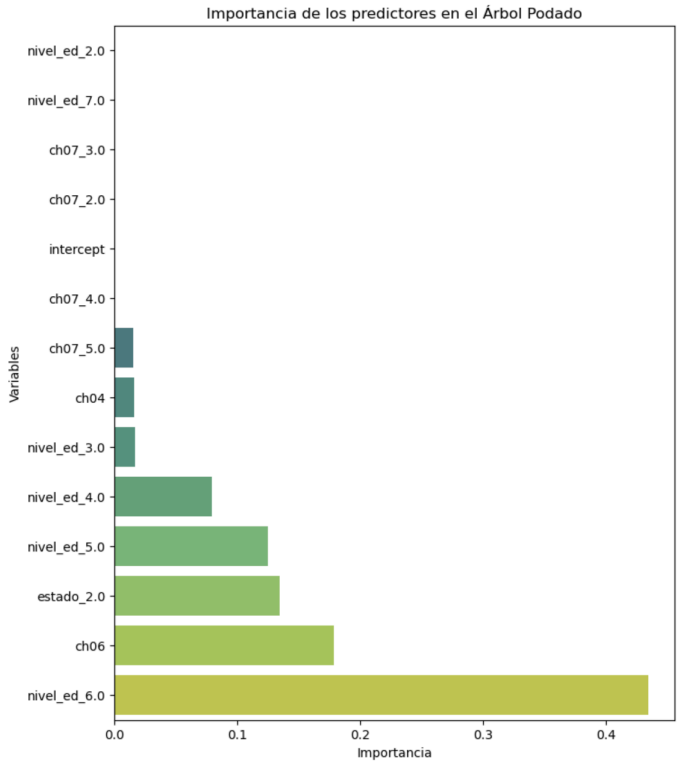
En el panel A (Figura 5), puede verse el árbol de decisión final obtenido tras aplicar *cost-complexity* pruning utilizando el valor óptimo de ccp_alpha seleccionado mediante 10-*fold cross validation*. Este árbol es más simple que el árbol completo crecido, lo que evita el sobreajuste y mejora la capacidad de generalización. Presenta una profundidad moderada (de 5-6 niveles), con una división inicial basada en la variable de nivel educativo, lo que indica que es la más informativa para discriminar entre pobreza y no pobreza. Las divisiones que le siguen son sobre edad y estado laboral, que son variables fuertemente asociadas con la condición de pobreza. La poda del árbol permitió reducir ramas irrelevantes, evitar *splits* con muy pocos casos y mantener un modelo simple, interpretable y estable.

Figura 5



En el Panel B (Figura 6) se presenta el gráfico de importancia de predictores del árbol podado. La variable con mayor importancia es una dummy de nivel educativo (nivel_ed_6 = Superior/Universitario incompleto), seguida por ch06 (edad), la variable *dummy* de estado (estado_2 = Desocupado), y algunas categorías intermedias de nivel educativo. Estas variables concentran la mayor reducción de impureza dentro del modelo y, por lo tanto, son las que más influyen en la clasificación final. Por otro lado, varias *dummies* de nivel educativo (nivel_ed_3= primario completo y nivel_ed_7= universitario completo) y de ch07 (variable de ocupación) muestran importancias muy bajas. Esto es consistente con lo observado en el modelo LASSO, donde los coeficientes de esas mismas variables fueron penalizados hacia valores cercanos a cero, que indica una contribución predictiva débil y redundante. Ambos métodos descartan predictores poco informativos, reforzando la robustez de las conclusiones del análisis.

Figura 6



C. Comparación entre métodos

6.

A continuación comparamos el desempeño de los modelos estimados en el TP3 (logit sin penalidad y KNN con validación cruzada) y en el TP4 (LASSO, Ridge y el árbol podado), incorporando además la métrica 1–*accuracy* como medida del error total de clasificación.

La Figura 7 presenta las matrices de confusión del TP3 y la Figura 8 las del TP4. La regresión logística muestra un comportamiento conservador donde clasifica correctamente a la mayoría de los casos negativos pero produce una cantidad elevada de falsos negativos, lo que refleja una sensibilidad reducida y un (1 – *accuracy*) elevado. El modelo KNN del TP3 mejora esta dimensión, aumentando los verdaderos positivos a costa de incrementar los falsos positivos, lo que es coherente con su frontera de decisión más flexible. Por su parte, los modelos logísticos penalizados (LASSO y Ridge) generan matrices prácticamente idénticas entre sí y muy similares a la del logit sin penalidad, lo que confirma que las penalizaciones no modifican sustancialmente la clasificación dadas las características del diseño de predictores. Esto era esperable, ya que el valor óptimo de penalización seleccionado para LASSO fue extremadamente pequeño, lo que implica que, en la práctica, el modelo funciona casi igual que un logit sin regularización. El árbol podado, en cambio, muestra un leve aumento en verdaderos positivos y negativos, sugiriendo una capacidad algo mayor para capturar relaciones no lineales.

Figura 7

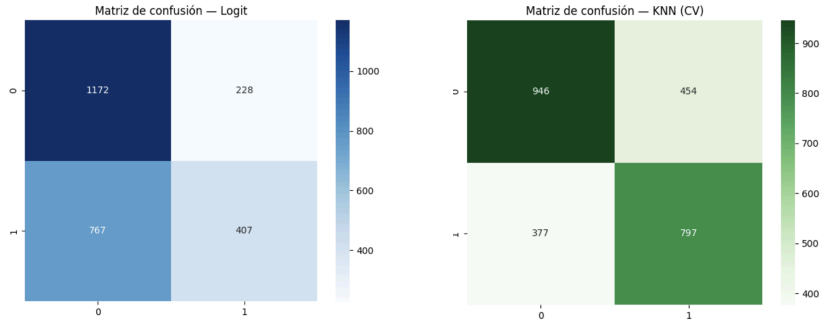
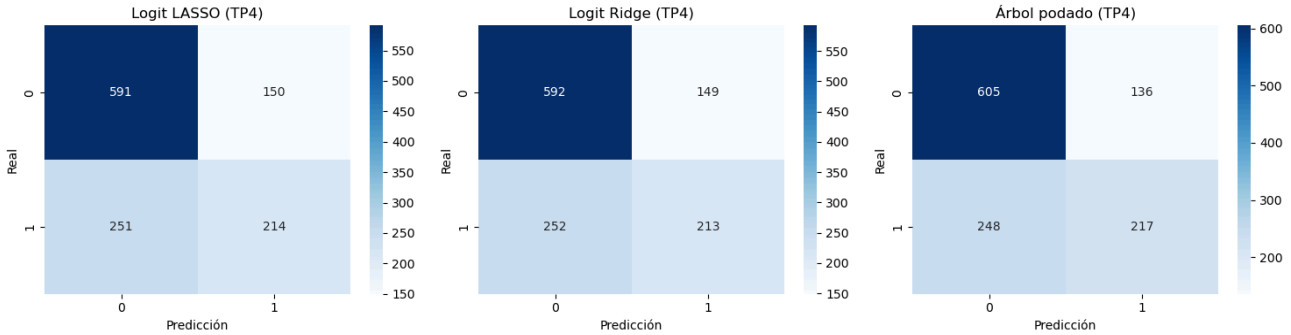
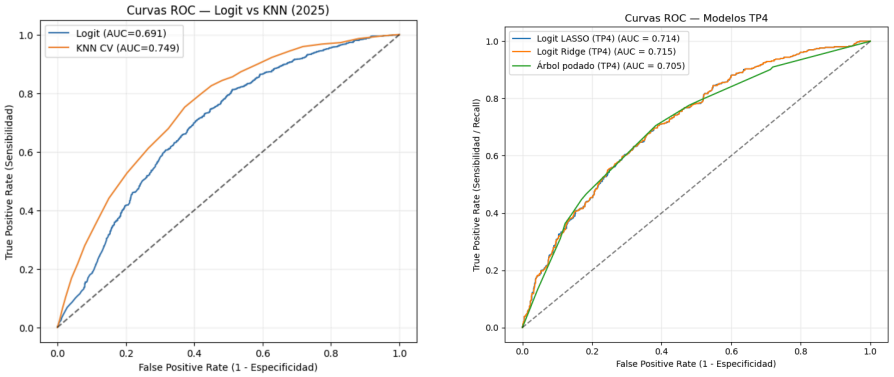


Figura 8



Las curvas ROC (Figura 9) refuerzan este patrón. En la comparación TP3, la curva del KNN domina claramente a la del logit, evidenciando un AUC superior y, por ende, una mayor capacidad discriminatoria. Entre los modelos del TP4, las curvas de LASSO, Ridge y logit son prácticamente superpuestas, con AUC cercanos a 0.71, lo que indica que las penalizaciones no aportan mejoras significativas. El árbol podado presenta un AUC apenas inferior, aunque igualmente cercano, lo cual sugiere que capta cierta no linealidad sin lograr un salto sustantivo en la capacidad predictiva. Por lo tanto, las curvas ROC muestran que los modelos lineales del TP4 tienen un desempeño moderado y muy homogéneo, mientras que el KNN del TP3 continúa siendo el modelo con mejor discriminación.

Figura 9



La Tabla 4 sintetiza las métricas de desempeño de cada modelo. La regresión logística presenta la menor sensibilidad y el mayor (1 – *accuracy*), reflejando una alta proporción de falsos negativos. LASSO y Ridge mejoran parcialmente estos valores, pero su desempeño global es prácticamente idéntico entre sí y no supera al modelo base. El árbol podado obtiene el mayor *accuracy* dentro de los modelos y un nivel de sensibilidad mayor que los modelos lineales, reduciendo el error total a 0.318. Finalmente, el KNN presenta la mayor sensibilidad del conjunto y un AUC superior, posicionándose como el modelo más eficaz en la identificación de casos positivos, aunque con un incremento en falsos positivos reflejado en su (1 – *accuracy*).

En conclusión, los resultados muestran un *trade-off* claro entre interpretabilidad y performance predictiva. Los modelos lineales ofrecen mayor estabilidad y son más fáciles de comunicar, pero presentan limitaciones importantes en sensibilidad. Los métodos no lineales, especialmente el KNN, mejoran la detección de casos positivos y poseen mejor capacidad discriminatoria, aunque incurren en un mayor número de errores globales. Por lo tanto, existe una ventaja relativa en utilizar un modelo no lineal cuando se prioriza la identificación de positivos, mientras que los modelos lineales resultan preferibles cuando el énfasis se coloca en la interpretabilidad y en la comunicación clara de los resultados.

Tabla 4

Modelo	**Accuracy**	**Sensibilidad**	**1 - Accuracy**
Logit	0.613	0.347	0.387
KNN (CV)	0.677	0.679	0.333
LASSO	0.667	0.460	0.333
Ridge	0.667	0.458	0.333
Árbol podado	0.682	0.467	0.318

7.

En el TP3 habíamos concluido que, desde la perspectiva del Ministerio de Capital Humano, el modelo más adecuado era el KNN con validación cruzada, dado que maximiza la sensibilidad y minimizaba el error Tipo II (falsos negativos), que es el error más costoso en una política social focalizada. Aunque el KNN cometía una gran cantidad de falsos positivos, su capacidad para detectar a la mayoría de los hogares vulnerables justificaba su elección.

Al incorporar los resultados del TP4, esta conclusión no cambia. Los modelos lineales estimados con LASSO y Ridge, así como el árbol podado, muestran mejoras acotadas y no superan al KNN en términos de sensibilidad ni en su capacidad para identificar correctamente los casos positivos. Si bien el árbol podado ofrece un desempeño algo más equilibrado que el logit y reduce parcialmente el error total, su sensibilidad sigue siendo considerablemente inferior a la del KNN. Del mismo modo, las penalizaciones aplicadas al logit no producen mejoras suficientes como para alterar esta jerarquía.

Por lo tanto, dado que el objetivo del Ministerio continúa siendo minimizar los falsos negativos, es decir evitar excluir a un hogar vulnerable del programa alimentario, el modelo que prioriza la sensibilidad sigue siendo la opción más adecuada. En este sentido, el KNN con validación cruzada continúa siendo el mejor modelo para asignar recursos escasos a los más necesitados, aun cuando implique aceptar un mayor número de falsos positivos. En una política social de este tipo, este intercambio resulta razonable, ya que el costo de omitir a un hogar pobre es significativamente mayor que el costo de asistir a un hogar no pobre.

Apéndice
Link al GitHub: https://github.com/isanicola/CC408-Grupo-T3-8/tree/main/TP4

Diccionario para la base de datos 2005
https://www.indec.gob.ar/ftp/cuadros/menusuperior/eph/EPH_disenoreg_T1_2005.pdf

Diccionario para la base de datos 2025
https://www.indec.gob.ar/ftp/cuadros/menusuperior/eph/EPH_registro_1T2025.pdf