
Big Data y Machine Learning (UBA) - 2025

Trabajo práctico 4: Clasificando pobres con la EPH

Grupo 1

Francisco Ariel Gonzalez Häberle 903012

David Jimenez Jaldin 909323

Lourdes Sofia Beltramo 888198

Link github: https://github.com/davidjjj14082000-png/BigDataUBA-Grupo-1

## A) Enfoque de validación

### Ejercicio 1

| | Media_Train | Media_Test | Diferencia |
|---|---|---|---|
| const | 1.000 | 1.000 | 0.000 |
| ANO4 | 2011.587 | 2011.352 | 0.234 |
| CH04 | 1.527 | 1.520 | 0.007 |
| CH06 | 35.880 | 36.033 | -0.153 |
| CH07 | 3.463 | 3.474 | -0.011 |
| CH08 | 2.311 | 2.300 | 0.011 |
| NIVEL_ED | 3.540 | 3.517 | 0.023 |
| CAT_INAC | 1.644 | 1.640 | 0.004 |
| ESTADO | 2.201 | 2.201 | -0.000 |

Los promedios de todas las variables son muy parecidos en los conjuntos de entrenamiento y testeo, como se puede observar en la tabla de diferencias de medias. Las diferencias detectadas son insignificantes, todas por debajo de 0.25 unidades, lo que señala que no hay discrepancias relevantes entre las medias de las variables en los dos subconjuntos de datos. Esto indica que la división original se llevó a cabo de forma aleatoria, lo que asegura que las muestras de entrenamiento y de testeo son representativas del total de la población. En consecuencia, se espera que el modelo de predicción no presente sesgos asociados a una mala separación de los datos y que su desempeño en testeo sea consistente con el observado en entrenamiento.

## B) Modelo de regresión logística

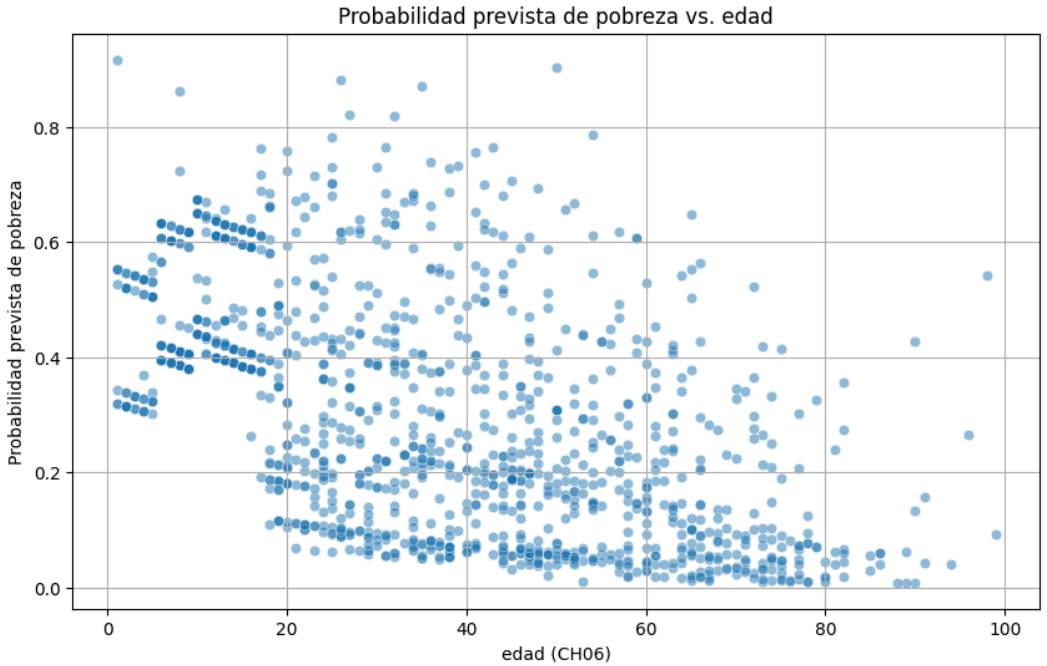
### Ejercicio 3: estimación y efectos marginales

```
Optimization terminated successfully.
         Current function value: 0.496474
         Iterations 7
```

| | Coeficiente | Error Estándar | Odd Ratio |
|---|---|---|---|
| const | -0.440340 | 3.217262e-01 | 0.643817 |
| CH04 | 0.104810 | 9.489611e-02 | 1.110500 |
| CH06 | -0.021302 | 4.076088e-03 | 0.978923 |
| CH07 | -0.102714 | 3.725284e-02 | 0.902385 |
| CH08 | 0.286905 | 2.333934e-02 | 1.332298 |
| NIVEL_ED_2 | 0.277801 | 2.318519e-01 | 1.320223 |
| NIVEL_ED_3 | -0.122726 | 1.856327e-01 | 0.884506 |
| NIVEL_ED_4 | -0.347999 | 2.113986e-01 | 0.706100 |
| NIVEL_ED_5 | -1.155016 | 2.242692e-01 | 0.315053 |
| NIVEL_ED_6 | -1.608418 | 2.584352e-01 | 0.200204 |
| NIVEL_ED_7 | -0.175317 | 4.449437e-01 | 0.839191 |
| CAT_INAC_1 | -1.011259 | 9.911510e+06 | 0.363761 |
| CAT_INAC_2 | -1.847685 | 9.911510e+06 | 0.157602 |
| CAT_INAC_3 | 0.350005 | 9.911510e+06 | 1.419074 |
| CAT_INAC_4 | 0.563601 | 9.911510e+06 | 1.756987 |
| CAT_INAC_5 | 0.085360 | 9.911510e+06 | 1.089109 |
| CAT_INAC_6 | 1.047111 | 9.911510e+06 | 2.849407 |
| CAT_INAC_7 | 0.918952 | 9.911510e+06 | 2.506663 |
| ESTADO_2 | 1.114876 | 2.131200e-01 | 3.049190 |
| ESTADO_3 | 0.186904 | 9.911510e+06 | 1.205512 |
| ESTADO_4 | -0.080820 | 9.911510e+06 | 0.922360 |

La estimación de la regresión logística realizada con la base de entrenamiento muestra el signo, magnitud y significancia de los coeficientes. Los coeficientes negativos indican que un aumento en esa variable disminuye la probabilidad de pobreza, mientras que los coeficientes positivos la incrementan. En este caso, las variables CH06 (edad), CH07 (nivel educativo del jefe de hogar), NIVEL_ED y ESTADO presentan coeficientes negativos y *odds ratios* menores a 1, lo que sugiere que a mayor edad, mayor nivel educativo y determinadas condiciones regionales, disminuye la probabilidad de encontrarse en situación de pobreza.

Ejercicio 4: visualización



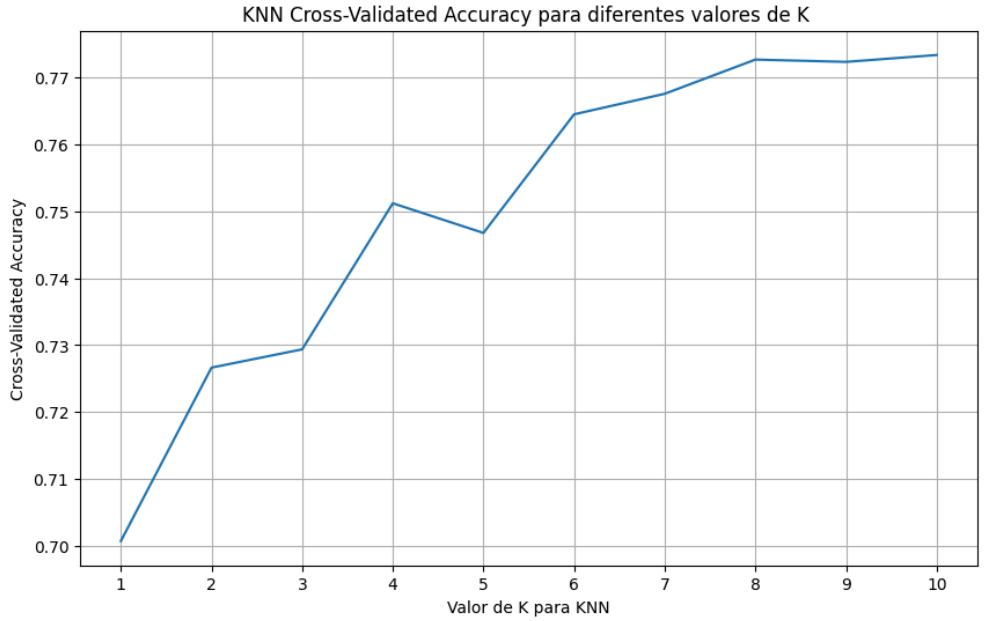

La probabilidad estimada de ser pobre en función de la edad (CH06) se muestra en este gráfico de dispersión, que está basado en el modelo de regresión logística calculado. Cada punto en el gráfico simboliza a una persona dentro del conjunto de prueba.
Al analizar el gráfico, podemos observar la tendencia general de cómo cambia la probabilidad de pobreza a medida que se envejece. Una dispersión de puntos agrupados señala edades en las que la probabilidad predicha es parecida, mientras que una dispersión más extensa indica una variabilidad más alta en el riesgo de pobreza para esas edades.

## C. Método de Vecinos Cercanos (KNN)

### Ejercicio 5: Estimación

La selección de K en el modelo KNN afecta directamente la compensación entre varianza y sesgo. Un modelo con un valor de K chico (por ejemplo, K=1) es de bajo sesgo pero alta varianza, porque la predicción se basa solo en el vecino más próximo, lo que lo hace susceptible al ruido presente en los datos de entrenamiento. Un modelo con un K grande resulta en un alto sesgo y baja varianza, dado que se promedia la predicción sobre más vecinos, lo que suaviza el límite de decisión, pero puede pasar por alto patrones locales relevantes. Hallando un K óptimo se logra equilibrar esta relación para conseguir un rendimiento óptimo.

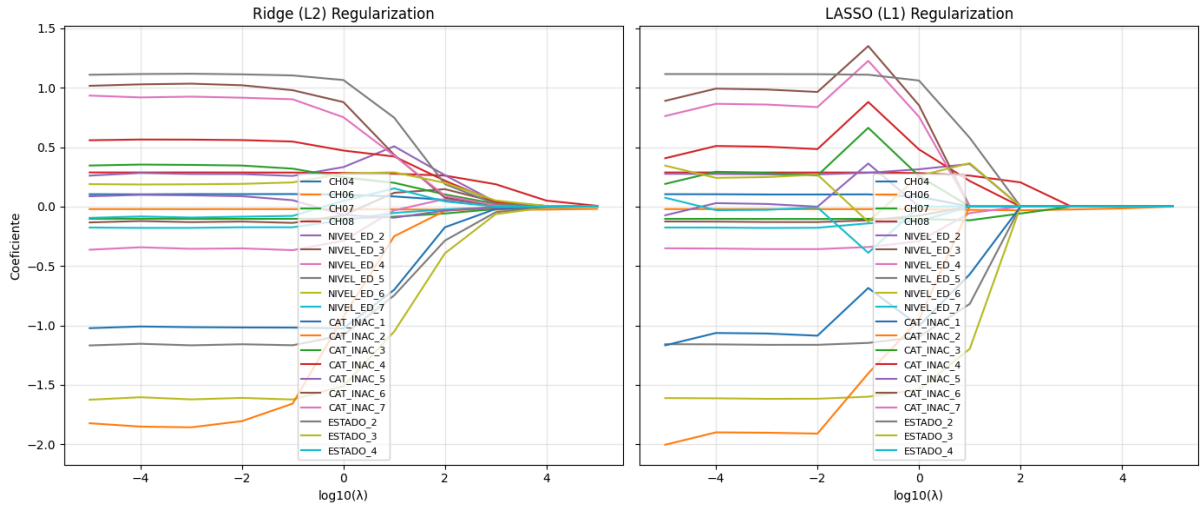
### Ejercicio 7: K óptimo por CV




El gráfico muestra la precisión promedio obtenida a través de validación cruzada para diferentes valores de K en el modelo KNN. El eje horizontal representa el número de vecinos (K), y el eje vertical la precisión promedio de la validación cruzada.
Observando el gráfico, buscamos el valor de K que maximiza la precisión promedio, en este caso el k óptimo es 10. Este valor de K se considera el óptimo porque ofrece el mejor equilibrio entre sesgo y varianza para este conjunto de datos, lo que resulta en el mejor rendimiento de generalización en datos no vistos durante la validación cruzada.

## D. Modelo de Regresión Logística con Regularización: Ridg y LASSO

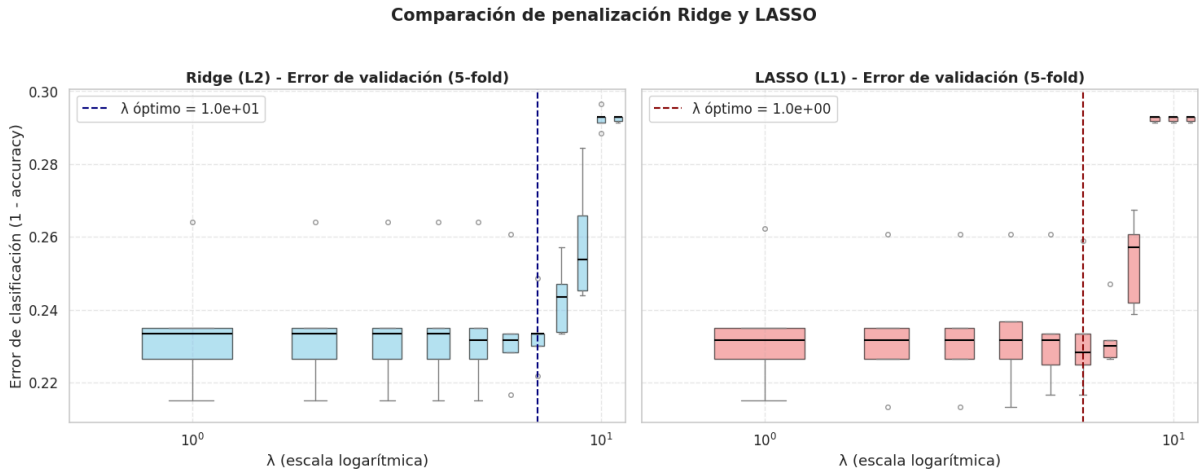
### Ejercicio 8



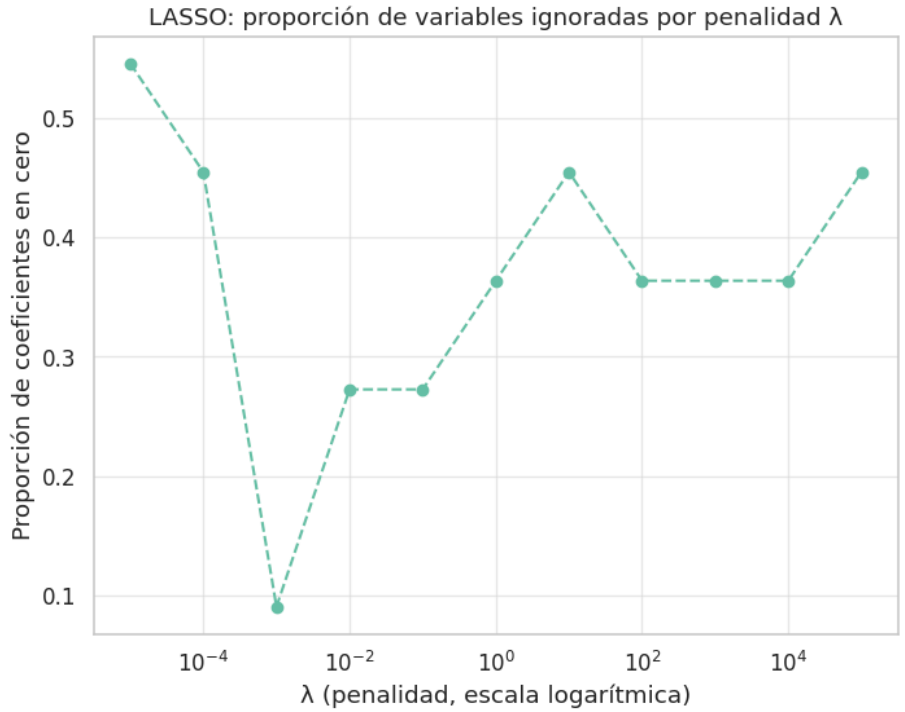

Estas gráficas ilustran cómo la regularización afecta los coeficientes del modelo al variar $\lambda$. En Ridge (L2), los coeficientes se reducen progresivamente a medida que aumenta la penalización, pero no llegan a ser exactamente cero, lo que implica que todas las variables

permanecen en el modelo aunque con menor influencia. En cambio, en LASSO (L1), el aumento de $\lambda$ reduce la magnitud de los coeficientes y, también lleva algunos exactamente a cero, realizando una selección automática de variables. En conjunto, Ridge controla la magnitud de los parámetros, mientras que LASSO promueve modelos más simples y parcimoniosos.

## Ejercicio 9




Los boxplots muestran el error de validación (1 – accuracy) para distintos valores de λen modelos Ridge y LASSO utilizando validación cruzada (5-fold).
El λcv óptimo fue 1.0e+01 para Ridge y 1.0e+00 para LASSO. En ambos casos, el error se mantiene bajo en un rango intermedio de penalización, aumentando para valores extremos. Ridge penaliza suavemente todos los coeficientes, mientras que LASSO simplifica el modelo al forzar algunos coeficientes a cero.



LASSO descarta más variables a medida que λ crece, cuando las penalizaciones son  chicas, casi todas siguen activas, pero con valores altos, se demuestra cómo la regularización LASSO (penalización L1) lleva cada vez más coeficientes a cero a medida que aumenta el parámetro de penalización (lambda), realizando así una selección de características.

Ejercicio 10 *Estimación con  y comparación de coeficientes*

Visualización de la tabla de resultados:

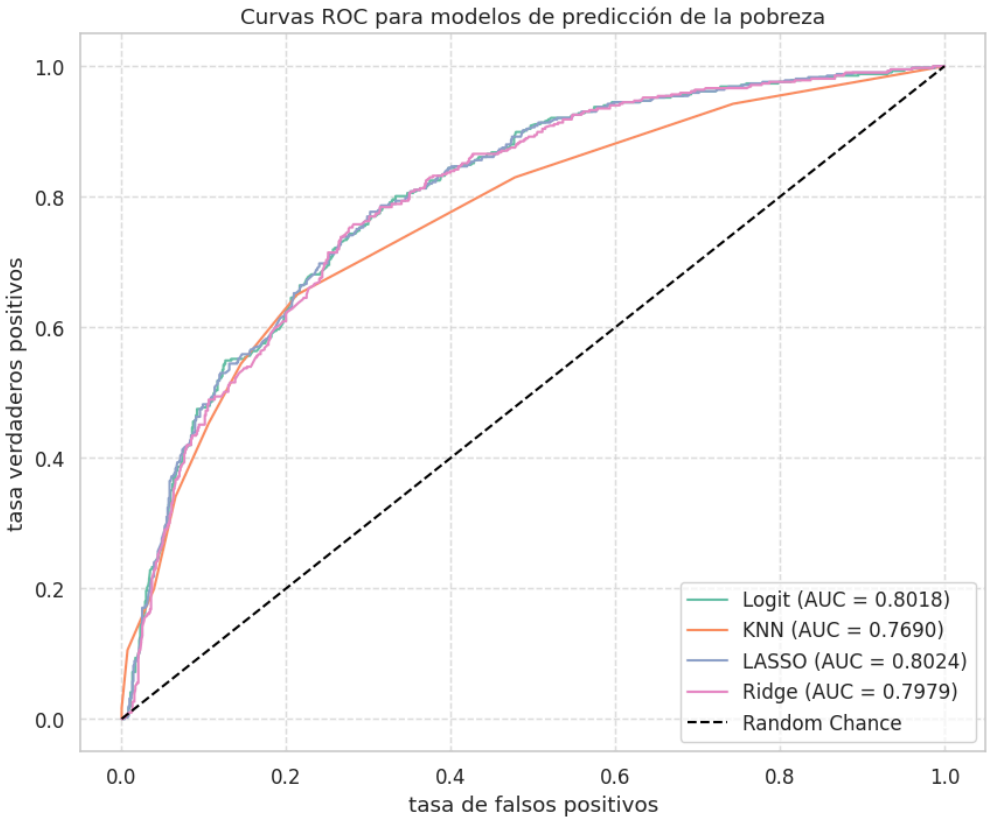
| | **Sin penalidad** | **L1 (Lasso)** | **L2 (Ridge)** |
|---|---|---|---|
| **Variables** | | | |
| **CH04** | 0.0526 | 0.0492 | 0.0497 |
| **CH06** | -0.4801 | -0.4581 | -0.4307 |
| **CH07** | -0.1701 | -0.1651 | -0.1563 |
| **CH08** | 0.5442 | 0.5407 | 0.5303 |
| **NIVEL_ED_Primario incompleto (incluye educación especial)** | -0.0912 | -0.0730 | -0.0472 |
| **NIVEL_ED_Secundario completo** | -0.2551 | -0.2301 | -0.1927 |
| **NIVEL_ED_Secundario incompleto** | -0.1642 | -0.1405 | -0.1041 |
| **NIVEL_ED_Sin instrucción** | -0.0962 | -0.0851 | -0.0697 |
| **NIVEL_ED_Superior/universitario completo** | -0.7009 | -0.6762 | -0.6262 |
| **NIVEL_ED_Superior/universitario incompleto** | -0.4903 | -0.4681 | -0.4300 |
| **ESTADO_Entrevista no realizada** | 0.0000 | 0.0000 | 0.0000 |
| **ESTADO_Inactivo** | -0.0539 | 0.0000 | -0.0394 |
| **ESTADO_Menor de 10 años** | -0.1187 | -0.0807 | -0.0990 |
| **ESTADO_Ocupado** | -0.5538 | -0.5347 | -0.4946 |
| **CAT_INAC_Discapacitado** | 0.0546 | 0.0531 | 0.0563 |
| **CAT_INAC_Estudiante** | -0.0899 | -0.0813 | -0.0674 |
| **CAT_INAC_Individuo Ocupado** | 0.1269 | 0.1668 | 0.1004 |
| **CAT_INAC_Jubilado/pensionado** | -0.4782 | -0.4746 | -0.4555 |

| | | | |
|---|---|---|---|
| **CAT_INAC_Menor de 6 años** | -0.1022 | -0.0955 | -0.0862 |
| **CAT_INAC_Otros** | 0.0470 | 0.0471 | 0.0525 |
| **CAT_INAC_Rentista** | -0.1661 | -0.1601 | -0.1564 |

Interpretacion: Observamos que para ambos metodos de regularizacion la magnitud los coeficientes es menor en todos los casos, salvo para aquellos con entrevista no realizada ya que sus coeficientes son 0 al no poseer información relevante en la categoría, respecto al modelo sin penalidad, lo cual cumple con la función de sesgar los coeficientes para obtener una menor varianza en los mismos. Respecto a la penalidad de Lasso observamos que vuelve 0 al coeficiente de ESTADO_Inactivo, cumpliendo su funcion de seleccionar variables al excluir aquellas que poseen un elevado grado de multicolinealidad, como es en el caso de la dummy de inactivo con las categorias de inactividad. Por su parte, Ridge penaliza la magnitud de los coeficientes acercandolos hacia cero, lo cual se ve reflejado para todos los coeficientes del modelo (a excepcion de ESTADO_Inactivo). Por último destacamos que el modelo con penalidad L1 mantiene el signo de todos los coeficientes del modelo base (a excepcion de ESTADO_Inactivo, mientras que la penalidad L2 altera el signo de 11 coeficientes.

## E. Desempeño de modelos afuera de la muestra, métricas y políticas públicas

### Ejercicio 11



Los valores AUC  confirman esta observación: LASSO presenta el valor AUC más alto (0,8024), seguido de cerca por LOGIT (0,8018) y ridge (0,7979), y finalmente KNN (0,7690). Un valor AUC más alto indica una mejor capacidad discriminatoria general.
El modelo LASSO presenta la mayor precisión, lo que indica que clasifica correctamente más instancias que los demás modelos.
La puntuación F1 es la media armónica de la precisión y la exhaustividad, lo que proporciona una medida equilibrada del rendimiento de un modelo, especialmente útil cuando hay desequilibrio de clases. El modelo KNN  alcanza la puntuación F1 más alta (0,5439), lo que indica un mejor equilibrio entre la correcta identificación de casos positivos (exhaustividad) y la precisión (no etiquetar erróneamente los casos negativos como positivos). Seguido por LASSO y luego logit y ridge.
En conjunto, los resultados muestran que no existe un único modelo claramente superior, sino que cada uno presenta fortalezas distintas. El modelo LASSO exhibe la mayor capacidad discriminatoria global (AUC = 0,8024) y la mejor precisión general, lo que lo convierte en una buena opción cuando se busca maximizar la exactitud total de las predicciones. Sin embargo, el modelo KNN alcanza el mayor F1-Score (0,5439), mostrando un mejor equilibrio entre la detección de hogares pobres y la reducción de falsos positivos.

## Ejercicio 12

Discusión:
Considerando el trade-off entre los errores tipo I y II, en un programa social como la asignación de alimentos el costo de un error tipo II (no identificar a un pobre) es mayor que el de un error tipo I (dar ayuda a un no pobre).
Por lo tanto, conviene elegir un modelo que priorice la sensibilidad (recall), aun a costa de menor precisión.
En este caso, una regresión logística con un umbral de clasificación ajustado (<0.5) permitiría identificar un mayor número de hogares pobres, mientras que un random forest podría lograr mayor exactitud general pero menor transparencia y control sobre este equilibrio.
En consecuencia, la regresión logística sería más adecuada si se busca un modelo interpretable, transparente y socialmente sensible al error tipo II, mientras que el random forest podría usarse en etapas exploratorias donde se priorice la eficiencia predictiva.