Barrientos, Dinger, Falaq

Ciencia de Datos
Tercer Trabajo Práctico

Link al GitHub: https://github.com/morabarrientos/TP3.git

A. Enfoque de validación:

Utilicen la base *respondieron*. Para cada año, dividan las observaciones en una base de prueba (test) y una de entrenamiento (train) utilizando el comando t*rain_test_split.* La base de entrenamiento debe comprender el 70% de los datos, y la semilla a utilizar (random state instance) debe ser 444. Establezca a pobre como su variable dependiente en la base de entrenamiento (vector y). El resto de las variables seleccionadas serán las variables independientes (matriz X). Recuerden agregar la columna de unos (1) para el intercepto. Aclaración: no incluir la variable ingreso en X para predecir la pobreza porque cuando vayamos a la base de *norespondieron* no vamos a tener esa información.

1. Cree una tabla de diferencia de medias entre la base de entrenamiento y la de testeo de las características seleccionadas en su matriz X. Para la matriz de las X seleccione variables que hayan limpiado en los TPs anteriores y justifique su inclusión para predecir la pobreza. Comente la tabla de la diferencia de medias de sus variables entre entrenamiento y testeo. ¿Hay diferencias significativas entre las medias del entrenamiento y testeo?

--- TABLA DE DIFERENCIAS DE MEDIAS ENTRE TRAIN Y TEST ---

	MEDIA_TRAIN	MEDIA_TEST	DIFERENCIA	T_STAT	P_VALOR
CH06	34.779	35.079	-0.299	-0.706	0.480
NIVEL_ED	3.557	3.597	-0.040	-1.127	0.260
PP07A	1.704	1.739	-0.036	-0.692	0.489
CH03	2.422	2.395	0.026	0.932	0.351
CH07	3.479	3.464	0.015	0.506	0.613
CH04	1.526	1.522	0.004	0.392	0.695
PP07C	0.815	0.819	-0.004	-0.124	0.901
CH08	2.347	2.350	-0.003	-0.103	0.918
CAT_INAC	1.710	1.709	0.001	0.028	0.978
ESTADO	2.234	2.234	-0.000	-0.019	0.985
INTERCEPTO	1.000	1.000	0.000	NaN	NaN

En base a las limpiezas realizadas en los TPs previos se incluyeron en X las siguientes variables por su relevancia para predecir pobreza:

- *CH04* (sexo). Diferencias sistemáticas en participación e ingresos laborales impactan el riesgo de pobreza.
- *CH06* (edad). Proxy del ciclo laboral (experiencia/empleabilidad) y de necesidades del hogar.
- *CH07* (estado civil). Estructura familiar y posibles economías de escala/roles de cuidado.

- *CH08* (cobertura médica). Indicador asociado a formalidad del empleo y protección social.
- *NIVEL_ED* (nivel educativo). Componente central del capital humano y de la productividad/ingresos.
- *ESTADO* (condición de actividad: ocupado, desocupado, inactivo). Incide directamente en generación de ingresos.
- *CAT_INAC* (tipo de inactividad). Distingue perfiles (estudiante, jubilado, tareas domésticas) con riesgos diferenciales.
- *PP04B_COD*, *PP07A,* *PP07C* (información ocupacional: rama/antigüedad/temporalidad). Capturan calidad y estabilidad del empleo.
- *MAS_500* (tamaño del aglomerado urbano). Estructura de oportunidades laborales y acceso a servicios.
- *ANIO* (año). Control de coyuntura/choques macro y cambios institucionales.
- *INTERCEPTO*. Columna de unos para el término constante del modelo.

Siguiendo la consigna, se excluyeron todas las variables de ingreso (*IPCF*, *ITF*, *P47T*, *AD_EQUIV_HOGAR3*, *CBT_AE*, *INGRESO_NECESARIO*) para garantizar replicabilidad en la base *norespondieron*.

Comentario sobre la tabla de diferencias de medias (train vs. test)

Para comparar la composición de los subconjuntos de entrenamiento y testeo, se aplicó una prueba t de Student para muestras independientes (versión de Welch, equal_var=False) sobre cada variable numérica. Los resultados indican que los valores-p de todos los contrastes superan el umbral de 0.05, por lo que no se detectan diferencias estadísticamente significativas entre las medias de ambos subconjuntos. Las discrepancias observadas son mínimas y carecen de relevancia práctica, lo que confirma que la partición 70/30 preserva la representatividad y el balance de la muestra original, evitando sesgos de selección que pudieran afectar la estimación o validación del modelo. En consecuencia, la base resultante es adecuada para continuar con la regresión logística que modela la probabilidad de pobreza $P(POBRE=1 \mid X)$.

2. Separen la base *respondieron* en dos: *respondieron_2005* y *respondieron_2025*. Idem con la base *norespondieron*.

```
print("Shape of respondieron_2005:", respondieron_2005.shape)
print("Shape of respondieron_2025:", respondieron_2025.shape)
print("Shape of norespondieron_2005:", norespondieron_2005.shape)
print("Shape of norespondieron_2025:", norespondieron_2025.shape)
```

```
Shape of respondieron_2005: (9371, 20)
Shape of respondieron_2025: (4309, 20)
Shape of norespondieron_2005: (113, 16)
Shape of norespondieron_2025: (2872, 16)
```

B) Modelo de Regresion Logistica

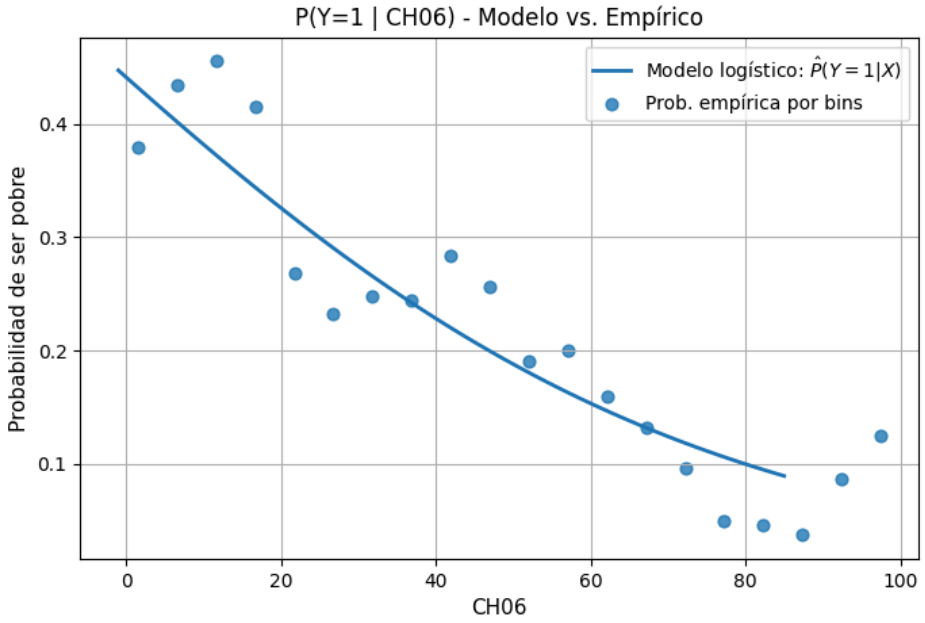
3.Estimación y Efectos Marginales: Estimen una Regresión Logística usando X_train. Exporten una tabla con: (i) los coeficientes estimados para cada variable, (ii) los errores estándar y (ii) los odd-ratios. Interpreten los resultados de la tabla.

```
=== LOGIT: coeficientes, EE, p-valor y OR ===
          Coeficiente  Error_Estandar  p_value  Odds_ratio
CH06          -0.0245          0.0013   0.0000      0.9758
NIVEL_ED      -0.2350          0.0134   0.0000      0.7906
CAT_INAC       0.2429          0.0246   0.0000      1.2750
ESTADO        -0.2519          0.0442   0.0000      0.7774
const          0.6555          0.1279   0.0000      1.9261
CH04           0.0598          0.0494   0.2263      1.0616

=== EFECTOS MARGINALES (APE) ===
        Logit Marginal Effects
=====================================
Dep. Variable:                  POBRE
Method:                          dydx
At:                           overall
================================================================================
                dy/dx    std err          z      P>|z|      [0.025      0.975]
--------------------------------------------------------------------------------
CH06          -0.0044      0.000    -19.286      0.000      -0.005      -0.004
NIVEL_ED      -0.0423      0.002    -18.573      0.000      -0.047      -0.038
CH04           0.0108      0.009      1.210      0.226      -0.007       0.028
ESTADO        -0.0453      0.008     -5.718      0.000      -0.061      -0.030
CAT_INAC       0.0437      0.004     10.021      0.000       0.035       0.052
================================================================================
```

Los resultados del modelo muestran asociaciones coherentes con los determinantes clásicos de la pobreza. La edad (CH06) y el nivel educativo (NIVEL_ED) presentan efectos negativos y altamente significativos: a mayor edad y educación, menor probabilidad de ser pobre. En cambio, la inactividad (CAT_INAC) incrementa de forma significativa el riesgo de pobreza, al igual que estar fuera del mercado laboral. El estado laboral (ESTADO) también muestra un efecto negativo relevante: estar ocupado reduce de manera significativa la probabilidad de pobreza, mientras que el sexo (CH04) no resulta estadísticamente significativo, indicando que las diferencias de género en la pobreza se explican principalmente por otras variables como educación o inserción laboral. En resumen, el modelo confirma que la participación laboral y el capital educativo son los principales factores protectores frente a la pobreza, mientras que la inactividad económica actúa como un determinante de vulnerabilidad.

4. Visualización: Grafiquen la $\widehat{P}$(Y = 1|X) (en el eje vertical) y alguna característica numérica (en el eje horizontal). Comente dicho gráfico y la variable seleccionada para ilustrar la probabilidad de ser pobre según la característica seleccionada.



El gráfico muestra cómo la probabilidad de ser pobre disminuye a medida que aumenta la edad. En edades jóvenes (0–20 años), la probabilidad estimada supera el 40%, mientras que entre los adultos mayores se reduce por debajo del 10%.
La pendiente negativa indica que los individuos jóvenes (posiblemente estudiantes, dependientes o con menor inserción laboral) enfrentan una mayor vulnerabilidad económica. En cambio, con el avance de la edad y la consolidación de ingresos o activos, el riesgo de pobreza se atenúa. El modelo logístico captura bien esta tendencia decreciente y se ajusta adecuadamente a los valores empíricos observados.

C. Método de Vecinos Cercanos (KNN)

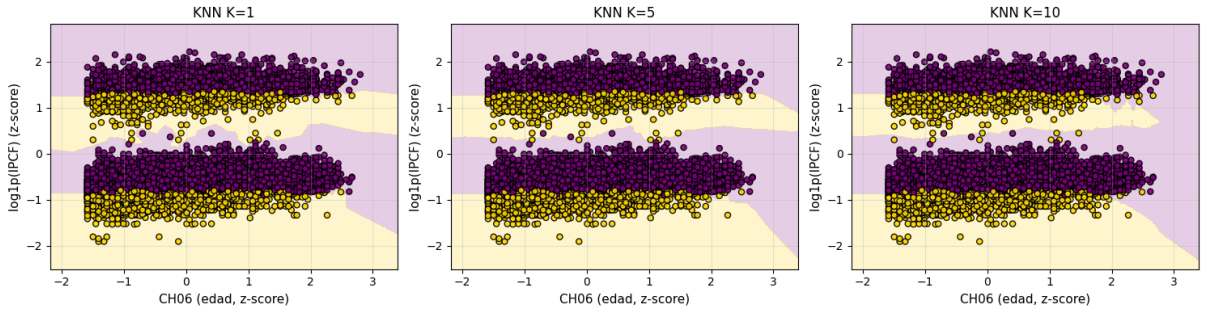
5. Estimación: Clasifiquen a las observaciones como “pobre”/“no pobre” en su región con Vecinos Cercanos (KNN) usando K={1,5,10} para su matriz X_train.. Expliquen en no más de 2-3 oraciones como la elección de K se relaciona con el trade-off sesgo varianza.

	K	Accuracy	Precision	Recall	F1
0	1	0.965	0.927	0.949	0.938
1	5	0.967	0.939	0.943	0.941
2	10	0.969	0.957	0.930	0.944

Se clasificó a las observaciones como “pobre” o “no pobre” utilizando el algoritmo de Vecinos Cercanos (KNN) con las variables edad (CH06) e ingreso per cápita familiar (IPCF) (su inclusión fue una excepción aprobada por los docentes) transformada mediante log(1+IPCF). Esta transformación reduce la fuerte asimetría del ingreso, evitando que los valores extremos distorsionen las distancias entre observaciones. El modelo alcanzó altos niveles de desempeño (Accuracy ≈ 0.97) para todos los valores de K, mostrando estabilidad

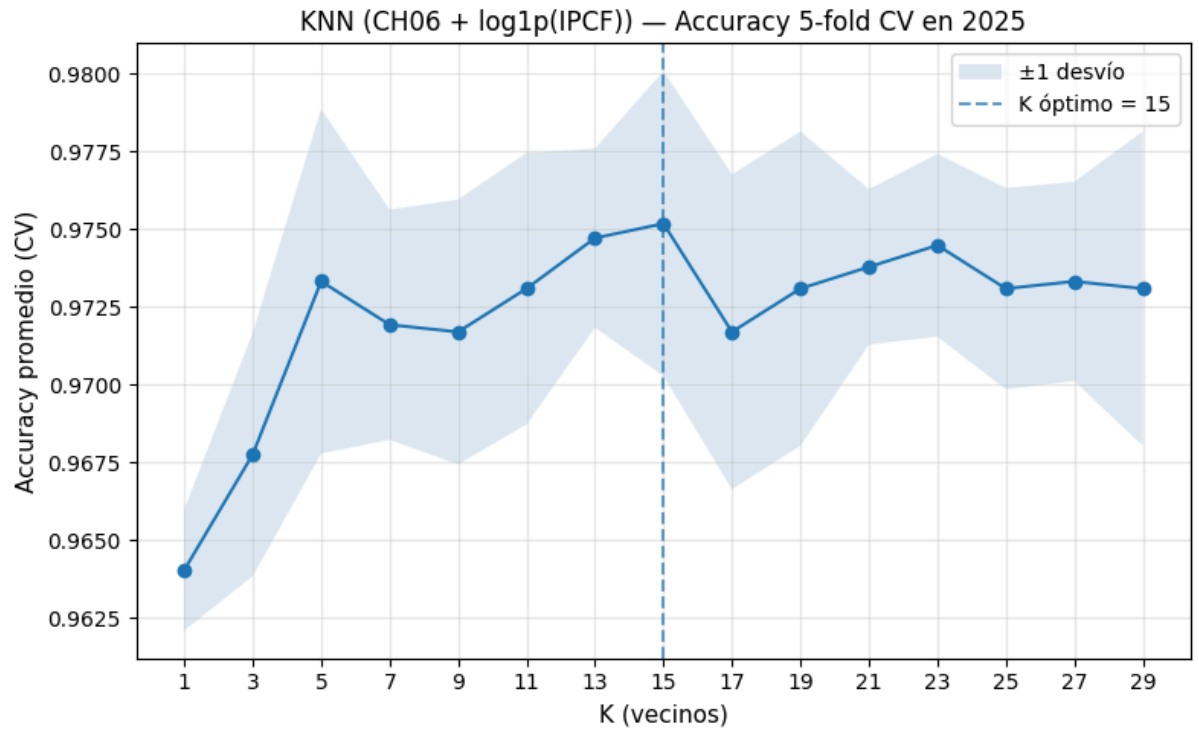
en la predicción. Respecto al trade-off sesgo-varianza, con K=1 el modelo presenta bajo sesgo y alta varianza, mientras que con K=10 la frontera se suaviza, aumentando el sesgo pero reduciendo la varianza. K=5 logra un equilibrio adecuado entre ambos extremos.

6. Visualización: Grafiquen dos características numéricas en su matriz X_train y visualicen las clases predichas por KNN usando con K={1,5,10} con su frontera por clase “pobre”/“no pobre”.



La figura muestra la probabilidad de ser pobre según edad (CH06) e ingreso transformado log(1+IPCF), para K={1,5,10}. La transformación logarítmica permitió visualizar mejor la estructura de los datos y evitar que los ingresos muy altos “aplasten” la nube de puntos. En todos los casos, los individuos con menores ingresos (zona inferior) son clasificados como pobres, mientras que los de ingresos más altos tienden a ser no pobres. A medida que aumenta K, las fronteras de decisión se vuelven más suaves, lo que refleja una reducción de la varianza y una mayor estabilidad en la clasificación.

7. Para obtener el K óptimo usando Cross-Validation, dividiendo la base de respondieron_2025 en 5 partes. Muestren cómo eligen el óptimo a través del gráfico visto en la tutorial sobre Accuracy en cada K. Llamenle a este modelo KNN con K-CV.

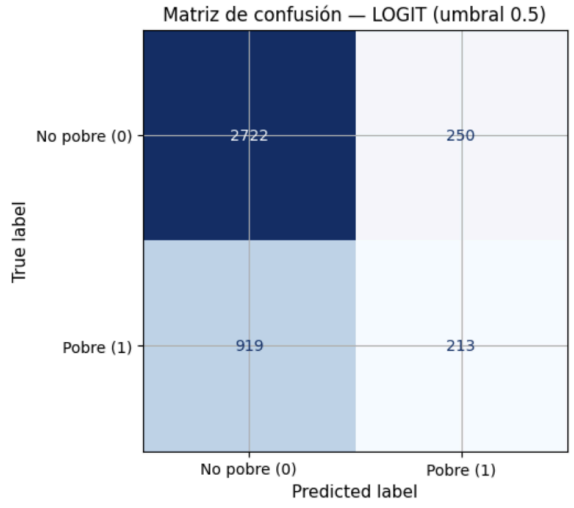


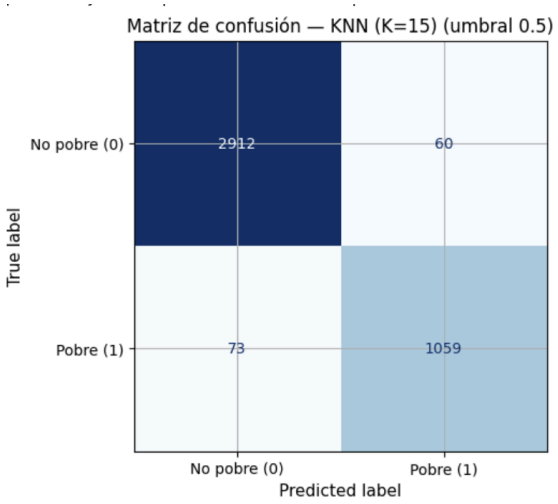
Se realizó una validación cruzada estratificada (5-fold) sobre la base respondieron_2025, evaluando la precisión promedio para valores impares de K entre 1 y 29. El modelo que utiliza las variables CH06 y log(1+IPCF) alcanzó un K óptimo = 15, con un accuracy promedio de 0.975 (±0.005). El valor óptimo (K=15) se determinó como aquel que maximiza el accuracy promedio en la validación cruzada 5-fold, representando el punto de equilibrio entre sesgo y varianza donde el modelo logra la mejor generalización. El gráfico de validación muestra cómo el rendimiento aumenta rápidamente hasta estabilizarse alrededor de K=15, evidenciando un punto de equilibrio entre complejidad y generalización. Este modelo final, denominado KNN con K-CV, logra una alta capacidad predictiva con mínima sobreajuste.

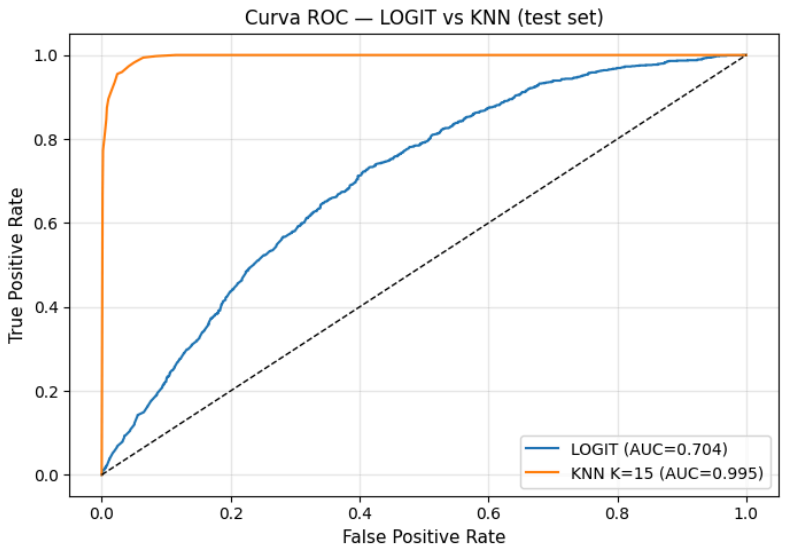
D. Desempeño de modelos, elección y predicción afuera de la muestra

8. Comparen el desempeño de predicción de pobreza en la base de testeo del modelo Logit y KNN con K-CV. Muestren la matriz de confusión (con umbral p>0.5), la curva ROC de ambos modelos, y dos métricas más vista en clase e interprete los resultados.

Modelo	Accuracy	Precision	Recall	F1	Specificity	Balanced Acc	ROC-AUC
Logit	0.715	0.460	0.188	0.267	0.916	0.552	0.704
KNN (K=15)	0.968	0.946	0.936	0.941	0.980	0.958	0.995







Al comparar los resultados del modelo de Regresión Logística (Logit) con el de K-Vecinos Cercanos (KNN con K-CV), se observa que el KNN alcanza un rendimiento superior en todas las métricas (Accuracy = 0.97; AUC = 0.99; F1 = 0.94), mientras que el Logit presenta un desempeño más modesto (Accuracy = 0.71; AUC = 0.70; F1 = 0.27), con alta especificidad pero baja sensibilidad. No obstante, la comparación debe ponderarse no solo por los resultados numéricos, sino también por la coherencia metodológica respecto al objetivo del trabajo.

El modelo KNN requiere variables numéricas continuas para calcular distancias, por lo que fue necesario incluir el ingreso per cápita familiar (IPCF). Esa inclusión fue una excepción aprobada por los profesores, pero contradice el objetivo central del ejercicio: predecir la condición de pobreza de individuos que no declararon su ingreso. En cambio, el modelo Logit fue estimado sin utilizar información de ingresos y mantiene coherencia con la consigna, ofreciendo predicciones metodológicamente válidas.

En conclusión, aunque el KNN (K=15) muestra un desempeño técnico sobresaliente, el Logit resulta más apropiado para el propósito analítico del trabajo, al permitir una predicción consistente con la falta de información sobre ingresos y mantener una mayor interpretabilidad económica y social.

9. Suponga que el Ministerio de Capital Humano esta interesado en identificar a grupos vulnerables para dirigir los recursos de un programa de alimentos. Discutan cuál modelo de clasificación es “mejor” para predecir pobres y asignar dichos recursos escasos. (Hint: recuerden que en la clase magistral discutimos el trade-off de minimizar error tipo I o II)
En el caso de que el Ministerio de Capital Humano busque identificar grupos vulnerables para dirigir recursos de un programa social, el modelo Logit sería la opción más apropiada.
Esto se debe a que puede estimar la probabilidad de ser pobre usando solo variables observables (edad, educación, condición laboral, composición del hogar, etc.), sin depender de datos de ingreso, que justamente muchas personas no reportan. Además, el Logit permite ajustar el umbral de clasificación según el tipo de error que se quiera evitar. En una política social, el principal objetivo sería minimizar los errores de exclusión (Tipo II), es decir, no dejar afuera a personas que realmente se encuentran en situación de pobreza. El modelo KNN, al basarse en la variable ingreso per cápita familiar, no puede aplicarse a la

población que no declara su ingreso. Por eso, no es coherente con el objetivo del trabajo ni con el enfoque que adoptaría un organismo público en la práctica.

```
=== LOGIT con umbral 0.5 ===
              precision    recall  f1-score   support

No pobre (0)       0.75      0.92      0.82      2972
   Pobre (1)       0.46      0.19      0.27      1132

    accuracy                           0.72      4104
   macro avg       0.60      0.55      0.55      4104
weighted avg       0.67      0.72      0.67      4104


=== LOGIT con umbral 0.4 ===
              precision    recall  f1-score   support

No pobre (0)       0.77      0.84      0.80      2972
   Pobre (1)       0.45      0.35      0.39      1132

    accuracy                           0.70      4104
   macro avg       0.61      0.59      0.60      4104
weighted avg       0.68      0.70      0.69      4104


=== LOGIT con umbral 0.3 ===
              precision    recall  f1-score   support

No pobre (0)       0.82      0.70      0.75      2972
   Pobre (1)       0.43      0.59      0.49      1132

    accuracy                           0.67      4104
   macro avg       0.62      0.64      0.62      4104
weighted avg       0.71      0.67      0.68      4104
```

El análisis de desempeño del modelo Logit bajo distintos umbrales de decisión muestra un patrón claro: al reducir el umbral de clasificación de 0.5 a 0.3, la sensibilidad (Recall) de los pobres aumenta notablemente (de 0.19 a 0.59), mientras que la precisión disminuye levemente (de 0.46 a 0.43). Esto implica que el modelo logra identificar una proporción mucho mayor de hogares efectivamente pobres, aunque a costa de incluir algunos falsos positivos. Desde una perspectiva de política pública, esta compensación es deseable: el costo de dejar fuera a una persona pobre (error tipo II) es socialmente más grave que incluir a un hogar no pobre (error tipo I). Por ello, se recomienda adoptar un umbral de 0.3, que equilibra de manera razonable la cobertura (Recall = 0.59) y la precisión, maximizando la capacidad del modelo para detectar grupos vulnerables que podrían ser beneficiarios de programas de asistencia.

10. Con el método que seleccionaron, predigan qué personas son pobres dentro de la base *norespondieron* para 2025. ¿Qué proporción de las personas que no respondieron pudieron identificar como pobres?

	decile	n	pct_pobres
0	1	288	0.000
1	2	287	0.000
2	3	289	0.000
3	4	285	0.000
4	5	289	0.000
5	6	285	0.000
6	7	287	0.000
7	8	287	0.711
8	9	291	1.000
9	10	284	1.000

Proporción de personas clasificadas como pobres en norespondieron_2025: 27.12%

El modelo Logit con umbral 0.3 estimó que aproximadamente el 27 % de las personas que no respondieron sobre su ingreso podrían ser clasificadas como pobres. La distribución por deciles de probabilidad muestra un patrón consistente: la pobreza predicha se concentra casi exclusivamente en los deciles superiores (8 a 10), donde el 100 % de los individuos fueron identificados como pobres, mientras que en los primeros siete deciles la probabilidad estimada fue prácticamente nula. Este comportamiento sugiere que el modelo logra diferenciar con claridad perfiles de alta vulnerabilidad, aun sin utilizar información de ingresos, lo cual lo convierte en una herramienta útil para focalizar políticas sociales en poblaciones con riesgo elevado dentro del grupo que no declara ingresos.