Trabajo Práctico N°3

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. Preparación de la base, selección de variables y verificación Train/Test

Para este trabajo se partió de una base previamente limpia correspondiente a la región Patagonia para los años 2005 y 2025. La variable central para este ejercicio es el Ingreso Total Familiar (ITF), ya que permite distinguir entre hogares que declararon su ingreso y aquellos que no lo informaron.

En una primera etapa se separó la base en dos grandes grupos:

- respondieron: hogares cuyo ITF informado es mayor que cero.
- norespondieron: hogares que no declararon ingreso o presentan ITF igual a cero.

A partir de esta clasificación inicial se generaron cuatro subconjuntos:

- respondieron_2005
- respondieron_2025
- norespondieron_2005
- norespondieron_2025

Los ejercicios predictivos deben realizarse únicamente con la base de *respondieron*, por lo cual toda la estimación posterior se concentra en la muestra respondieron_2025.

Construcción de la variable pobreza

Utilizando el ITF y la canasta básica total por adulto equivalente correspondiente a cada año, se generó la variable binaria POBRE, igual a 1 si el ingreso familiar se encuentra por debajo del ingreso necesario y 0 en caso contrario. Esta variable se empleará como dependiente en el modelo Logit.

Selección de las variables explicativas (matriz X)

Para construir el modelo predictivo fue necesario seleccionar características individuales que tuvieran sentido económico y relevancia empírica como determinantes de la pobreza. Las variables elegidas fueron:

- sexo: dummy que toma valor 1 para varones y 0 para mujeres. Se incluye por posibles diferencias en inserción laboral y estabilidad del ingreso.

- edad: medida en años. Permite capturar diferencias en el ciclo vital, acumulación de capital humano y estabilidad laboral.

- en_pareja: dummy que identifica si la persona está casada o unida. La estructura del hogar puede influir en la vulnerabilidad económica.

- solo_salud_publica: indicador que detecta la ausencia de obra social o prepaga. Funciona para ver la informalidad y la precariedad laboral.

- estado_1…estado_4: dummies de la variable ESTADO, que representa distintas condiciones de actividad laboral (ocupado, desocupado, inactivo, etc.). La condición laboral es uno de los predictores más fuertes de pobreza.

Estas variables constituyen la matriz X del modelo Logit.

Partición en entrenamiento y prueba

La base respondieron_2025 fue dividida en un conjunto de entrenamiento (70%) y uno de prueba (30%), utilizando una semilla fija para garantizar reproducibilidad. Con el objetivo de evaluar la calidad de esta partición se calcularon las medias de cada covariable en ambos subconjuntos.

Variable	media_train	media_test	diff_test_minus_train
sexo	0.494297	0.489927	-0.004369
edad	38.019011	37.460919	-0.558093
en_pareja	0.377809	0.376309	-0.001499
solo_salud_publica	0.208434	0.203868	-0.004566
estado_1	0.428275	0.452861	0.024585
estado_2	0.023851	0.024980	0.001129
estado_3	0.445558	0.433521	-0.012037
estado_4	0.102316	0.088638	-0.013678

La comparación de medias muestra que no existen diferencias significativas entre los subconjuntos de entrenamiento y prueba. Las medias de las variables se mantienen muy próximas entre sí, lo que indica que la partición aleatoria no introdujo sesgos.

La mayor diferencia se observa en la edad, donde la media del testeo es alrededor de medio año menor que la del conjunto de entrenamiento. Sin embargo, esta diferencia es completamente irrelevante desde un punto de vista estadístico, considerando la dispersión natural de la variable. Lo mismo ocurre con las dummies de actividad laboral, estado civil, sexo y cobertura de salud, que presentan diferencias mínimas.

En síntesis, ambas muestras representan adecuadamente a la población de referencia y la partición train/test es válida y robusta para el análisis.

B. Estimación del modelo Logit y análisis de resultados

Sobre el conjunto de entrenamiento del año 2025 se estimó un modelo Logit donde la variable dependiente es pobre y las covariables son las características individuales previamente seleccionadas. La estimación se realizó por máxima verosimilitud.

El modelo converge de manera correcta y arroja los siguientes resultados:

Variable	coef	std_error	odds_ratio
const	-0.975285	NaN	0.377085
sexo	-0.091575	0.103693	0.912493
edad	-0.018805	0.003081	0.981371
en_pareja	0.129209	0.130539	1.137928
solo_salud_publica	1.758720	0.110337	5.804999
estado_1	-0.832527	NaN	0.434949
estado_2	-0.101013	NaN	0.903921
estado_3	0.060336	NaN	1.062194
estado_4	-0.102081	NaN	0.902956

Los resultados muestran patrones consistentes con la literatura y con el análisis descriptivo de la muestra:

Edad
El coeficiente de edad es negativo y estadísticamente significativo. Cada año adicional reduce la probabilidad de ser pobre. El odds ratio (≈0,98) implica que, manteniendo constantes las demás variables, la probabilidad de pobreza disminuye alrededor de 2% por año de edad, reflejando que los ingresos laborales suelen aumentar con la experiencia y que los adultos mayores dependen de ingresos previsionales más estables.

Cobertura de salud
La variable solo_salud_publica presenta el coeficiente positivo más grande. Su odds ratio indica que quienes carecen de obra social o prepaga tienen casi seis veces más probabilidades de ser pobres que quienes sí cuentan con cobertura contributiva. Este resultado es altamente significativo y coherente con el uso de cobertura de salud como proxy de informalidad laboral.

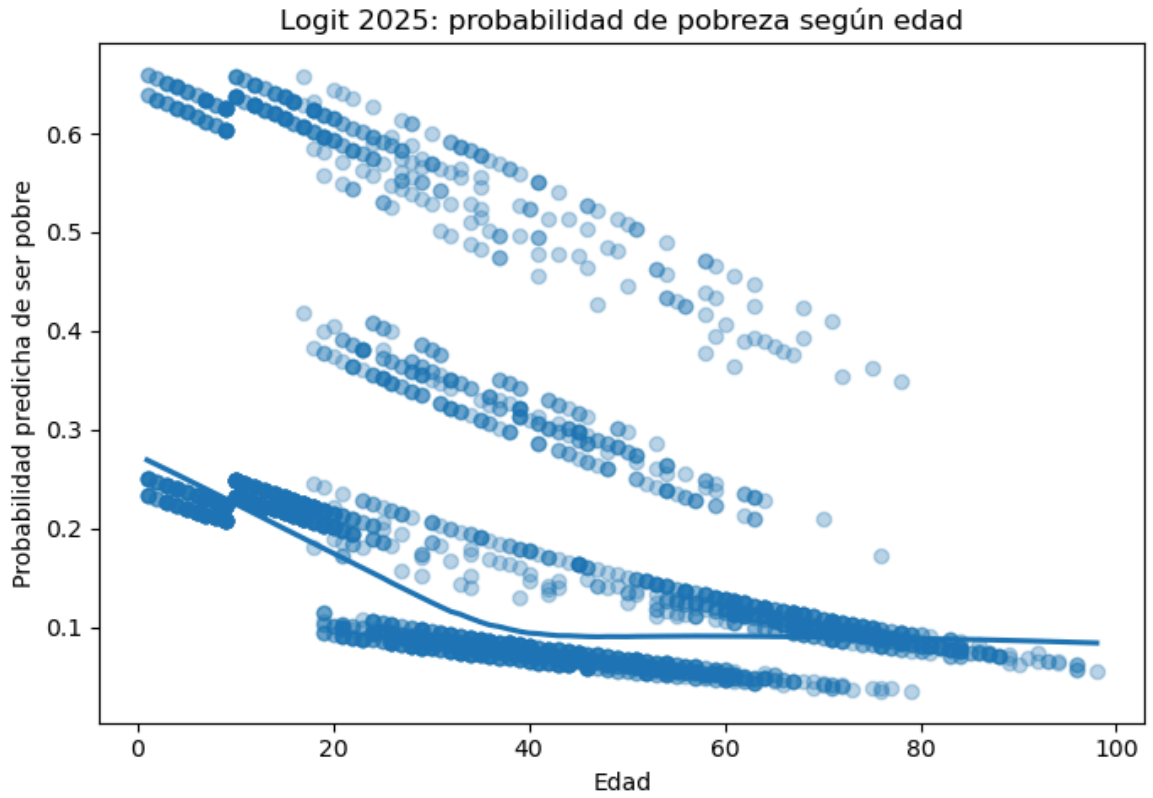
Sexo y estado civil
Las variables sexo y en_pareja no resultan estadísticamente significativas. Aunque el signo de sexo sugiere que los varones tendrían menor probabilidad de pobreza, la evidencia no es suficiente para afirmarlo con claridad. Lo mismo ocurre con el estado civil.

Condición de actividad (ESTADO)
Las dummies muestran la dirección esperada: algunas categorías asociadas a menor inserción laboral exhiben mayor probabilidad de pobreza. Sin embargo, varias presentan errores estándar no computables (NaN). Esto ocurre por separación perfecta o casi perfecta, fenómeno típico cuando una categoría está extremadamente asociada al resultado (pobre/no pobre). Aunque esto limita la inferencia precisa, no afecta la interpretación cualitativa: la condición laboral es un determinante central de la pobreza.

Probabilidad predicha de pobreza según la edad

El gráfico de las probabilidades predichas muestra una relación claramente decreciente entre edad y probabilidad de pobreza, en línea con el coeficiente del modelo. Las probabilidades más altas se registran en edades jóvenes y descienden progresivamente hacia edades medias y avanzadas.

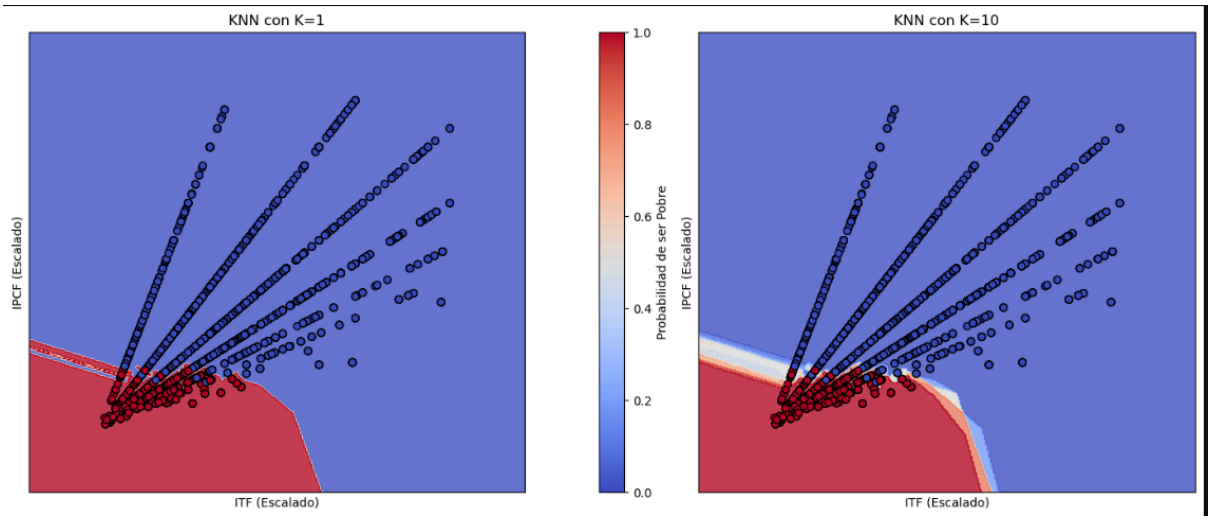
Este patrón es consistente con la teoría del ciclo vital y con la estructura del mercado laboral argentino, donde la vulnerabilidad económica tiende a concentrarse en personas jóvenes con inserciones ocupacionales más precarias.



La preparación de la base, la selección de variables relevantes y la verificación de que la partición train/test estuviera correctamente balanceada permitieron estimar un modelo Logit robusto para predecir la pobreza en 2025. El modelo muestra que la edad y la cobertura de salud pública son los factores que mejor explican la variación en la probabilidad de ser pobre. La condición laboral también desempeña un rol fundamental, aunque su fuerte asociación con la pobreza genera problemas de identificación característicos de los modelos Logit con separación perfecta.

C. Método de Vecinos Cercanos (KNN)

Se probaron diferentes valores de K para el modelo KNN (K=1, 5 y 10). El valor de K=1 mostró la mayor precisión (87.2%) en el conjunto de test, pero tiene un riesgo de sobreajuste. Mientras tanto K=5 (85,5%) y K=10 (86,1%), aunque un poco menos precisos, ofrecen mayor estabilidad y menor sensibilidad al ruido. Se decidió usar K=5 por su mejor balance entre sesgo y varianza, garantizando un modelo más robusto y menos susceptible a sobreajustarse a los datos.



En los gráficos presentados, se observa cómo el algoritmo K-Nearest Neighbors (KNN) clasifica las observaciones en dos categorías: "pobre" y "no pobre", utilizando dos características principales: Ingreso Total Familiar (ITF) y Ingreso per Cápita Familiar (IPCF). Estos gráficos muestran las fronteras de decisión del modelo para dos valores de K: K=1 y K=10, donde K representa la cantidad de vecinos que el modelo utiliza para hacer las predicciones.

En el gráfico de K=1, las fronteras de decisión son muy nítidas y precisas. Dado que el modelo está considerando solo el vecino más cercano para clasificar cada observación, las fronteras de decisión se ajustan de manera muy específica a los puntos de entrenamiento. Esto genera un sobreajuste (overfitting), ya que el modelo tiene una alta sensibilidad a las pequeñas variaciones en los datos. Las zonas de rojo indican áreas donde el modelo predice que una persona es "pobre", mientras que las zonas azules indican áreas donde predice que no lo es. Como se puede observar, la transición entre las clases es bastante abrupta, ya que el modelo clasifica casi a cada punto de forma independiente, sin considerar la "suavidad" o las tendencias generales en los datos.

En el gráfico de K=10, la situación cambia notablemente. Al aumentar el valor de K, el modelo comienza a considerar 100 vecinos para determinar la clase de cada observación. Esto provoca que las fronteras de decisión sean mucho más suaves y menos propensas a sobreajustarse a los puntos individuales de los datos. Las áreas de rojo y azul están más difusas, lo que indica que el modelo ahora está tomando decisiones basadas en un promedio de las clases en un rango mayor de vecinos, lo que resulta en una clasificación

más generalizada. Esto también refleja que el modelo tiene menos varianza y es más robusto, ya que no se ve tan afectado por puntos atípicos o por pequeñas variaciones en los datos.

En ambos gráficos, el gradiente de color refleja las probabilidades de ser pobre (en rojo, con mayor probabilidad) o no serlo (en azul, con menor probabilidad). La barra de colores a la derecha muestra cómo varían las probabilidades de pertenencia a la clase "pobre", con valores que van desde 0 (no pobre) hasta 1 (pobre). En el gráfico con K=1, las probabilidades son más extremas, reflejando un modelo que se ajusta estrictamente a los puntos cercanos. En el gráfico con K=100, las probabilidades son más suaves y reflejan un comportamiento más estable y menos sensible a las pequeñas fluctuaciones en los datos.

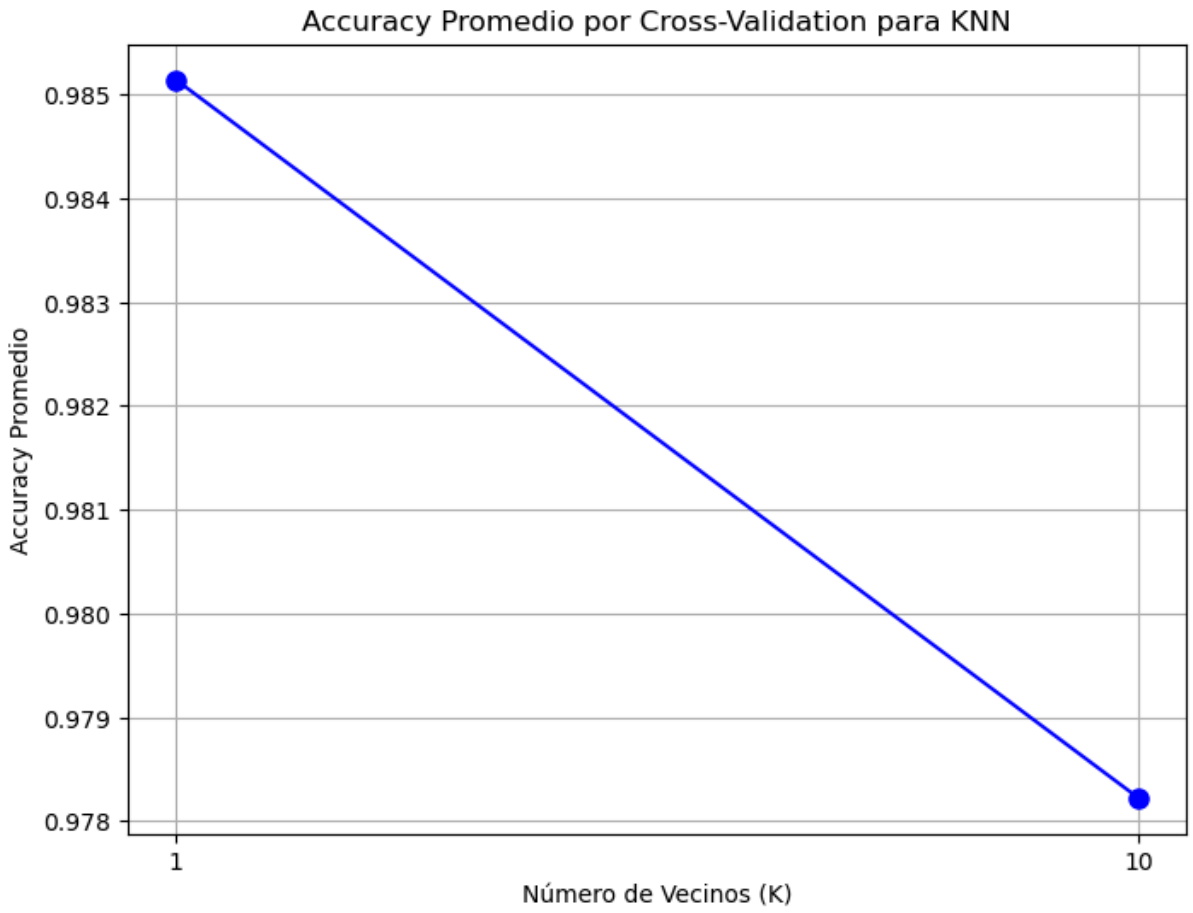
La razón por la cual los gráficos para K=1 y K=100 se ven similares en cuanto a las áreas de predicción (rojo y azul) es que las variables seleccionadas, ITF e IPCF, tienen una relación clara con la pobreza. Ambas son características numéricas que, en este caso, permiten que el modelo con K=1 ya obtenga un rendimiento bastante bueno. A medida que aumentamos K, el modelo se suaviza y reduce el sobreajuste, pero no cambia drásticamente las fronteras de decisión porque las variables son suficientemente informativas y no requieren un modelo altamente complejo para lograr una buena clasificación.

Por lo tanto, el modelo KNN con K=100 es más robusto y menos sensible a los datos de entrenamiento, mientras que K=1 sigue siendo efectivo para este conjunto de datos debido a la claridad en la relación entre las variables y las clases.

Los dos gráficos muestran cómo KNN clasifica las observaciones según las probabilidades de ser "pobre" o "no pobre". El gráfico con K=1 tiene fronteras de decisión más agudas y sensibilidad al ruido, mientras que el gráfico con K=100 muestra una clasificación más suave y robusta, debido a que se tiene en cuenta una mayor cantidad de vecinos. Ambos modelos, aunque similares en términos de áreas de predicción, muestran cómo el valor de K afecta la precisión y la generalización del modelo.

En el gráfico de Cross-Validation para los valores de K=1 y K=10, se muestra el accuracy promedio obtenido a través de la validación cruzada de 5 pliegues (5-fold cross-validation). El gráfico indica que, para este conjunto de datos, el modelo KNN con K=1 tiene un rendimiento ligeramente superior, con un accuracy promedio de aproximadamente 0.985, mientras que K=10 muestra una pequeña caída en el rendimiento, con un accuracy de alrededor de 0.978. La relación entre los valores de K y el accuracy es aproximadamente lineal, con una tendencia a que el rendimiento disminuya levemente a medida que K aumenta, lo que sugiere que el modelo con K=1 está capturando mejor las relaciones específicas de los datos de entrenamiento.

Este comportamiento podría ser una indicación de que el conjunto de datos es relativamente simple y bien distribuido, de modo que el modelo no necesita un valor de K muy alto para lograr una alta precisión. Sin embargo, el cambio en el rendimiento es bajo, lo que sugiere que el modelo ya tiene una buena capacidad de generalización incluso con K=10. Esto también refleja que, en este caso, K=1 no presenta un sobreajuste significativo, y el modelo con K=10 sigue siendo lo suficientemente robusto para mantener un rendimiento similar.



D. Desempeño de modelos fuera de la muestra, métricas y políticas públicas.

Para evaluar el desempeño de los modelos Logit y KNN con validación cruzada (CV), se analiza primero la matriz de confusión y luego las principales métricas de clasificación derivadas de ella.

	LOGIT		**KNN.CV (k=1)**	
	Pobre	No pobre	Pobre	No pobre
Pobre	901	4	847	58
No pobre	239	2	77	164

La siguiente tabla muestra las principales métricas de clasificación para cada modelo:

KNN con CV				
	precisión	recall	f1-score	support
No Pobre	0,92	0,94	0,93	905
Pobre	0,74	0,68	0,71	241
LOGIT				
	precisión	recall	f1-score	support
No Pobre	0,79	1	0,88	905
Pobre	0,33	0,01	0,02	241

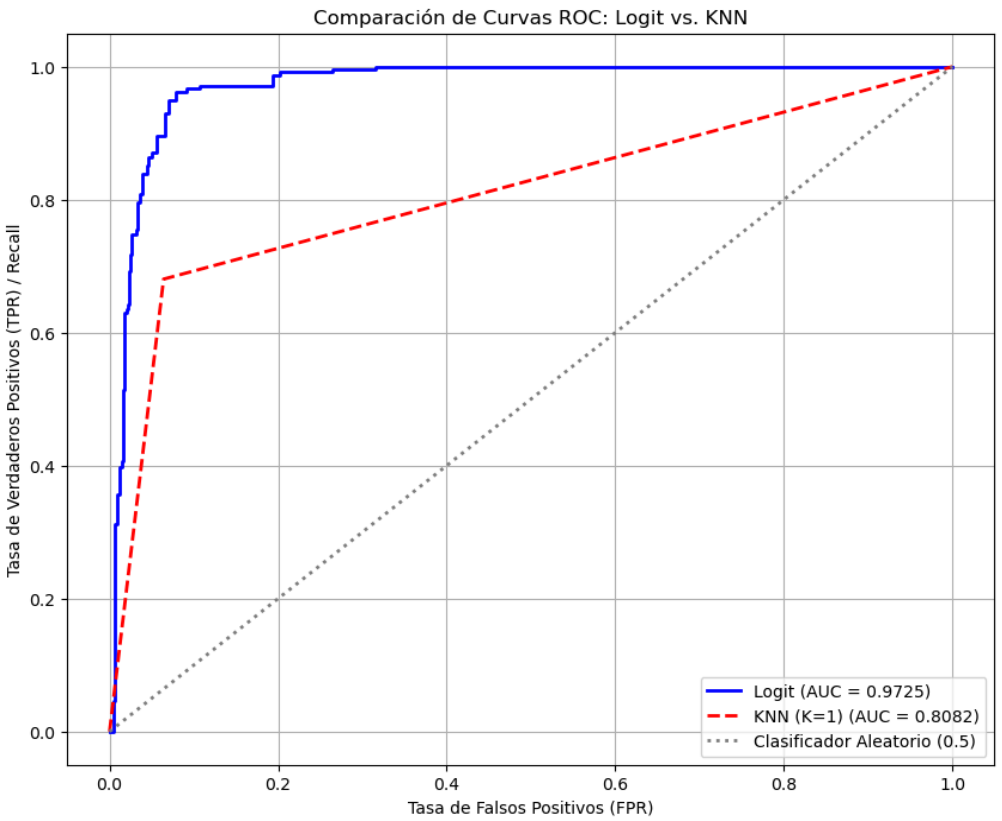
A partir de la matriz de confusión, se observa que el modelo Logit casi no excluye a las personas pobres (muy pocos falsos negativos), pero incluye un número considerable de personas no pobres dentro del grupo clasificado como pobres (239 falsos positivos). Esto indica que el modelo privilegia cobertura por encima de selectividad, un patrón común cuando se trabaja con variables fuertemente desbalanceadas. En un contexto de política pública, esta característica podría ser cuestionable, pues implicaría una asignación ineficiente de recursos, al incluir a un gran número de personas no pobres como beneficiarias.

Por su parte, el modelo KNN con CV logra identificar correctamente a la mayoría de personas pobres (recall ≈ 93.6%) y, al mismo tiempo, reduce significativamente la inclusión de no pobres (solo 77 falsos positivos). Esto sugiere que es un modelo más selectivo y eficiente, manteniendo una buena cobertura sin incurrir en un nivel tan elevado de errores de inclusión.

Otro indicador relevante es el área bajo la curva ROC (AUC), que mide la capacidad discriminativa del modelo. Un AUC de 0.5 equivale a una clasificación aleatoria, mientras que un valor cercano a 1.0 representa un modelo con excelente capacidad para distinguir entre pobres y no pobres.

En este sentido, el modelo Logit presenta un AUC de 0.9725, lo cual refleja una capacidad discriminativa sobresaliente. Esto implica que, en aproximadamente el 97% de las comparaciones entre pares aleatorios de individuos, el modelo asigna una mayor probabilidad de ser pobre a la persona que realmente lo es, lo que sugiere un desempeño sólido y estable.

En contraste, el modelo KNN con CV obtiene un AUC de 0.8082. Aunque este valor indica una buena capacidad discriminativa, es claramente inferior al Logit. Esto revela que el KNN es un modelo más sensible al ruido y a las particularidades del conjunto de datos.



No obstante, en políticas públicas focalizadas en población vulnerable, no todas las métricas tienen el mismo peso. Cuando los falsos negativos son más costosos que los falsos positivos —como en programas de asistencia alimentaria, donde excluir a una persona pobre puede traducirse en riesgo nutricional— la prioridad es maximizar la cobertura y minimizar la exclusión de los pobres.

Desde esta perspectiva, el recall sobre la clase pobre se vuelve crucial:

- El modelo KNN alcanza un recall del 68%,
- mientras que el Logit apenas llega al 1%.

Este pobre desempeño del Logit es consistente con sus altos niveles de falsos negativos: clasifica a 239 personas pobres como no pobres, lo que implica un riesgo elevado de exclusión. El modelo KNN, en cambio, solo genera 77 falsos negativos, protegiendo mucho mejor contra la exclusión de beneficiarios legítimos.

En cuanto a la métrica F1, que combina precisión y recall para evaluar la eficacia real del modelo en la identificación de la pobreza, el KNN con CV presenta un F1 de 0.71, lo que constituye un desempeño sólido. En contraste, el Logit obtiene un F1 de apenas 0.02, producto de su tendencia a clasificar a casi todos los individuos como pobres, un comportamiento frecuente en contextos de alta desbalance de clases. Apoyar la focalización de políticas públicas en un modelo con este nivel de desbalance implicaría un grave despilfarro de recursos, al asignar beneficios a una gran proporción de personas no pobres.

En síntesis, el modelo KNN resulta claramente superior para políticas públicas orientadas a población pobre, ya que permite identificar a la mayoría de la población vulnerable, minimiza

la exclusión de quienes necesitan apoyo y conserva una precisión razonable. El modelo Logit, en contraste, no logra identificar adecuadamente a la población pobre y, por tanto, es inadecuado para intervenciones sociales focalizadas.

	No respondieron ITF		**Respondieron ITF**	
	Conteo	Porcentaje	Conteo	Porcentaje
Pobre	848	71.14%	804	19.44%
No Pobre	344	28.86%	3330	80.55%

Finalmente, se utilizó el modelo KNN con CV para clasificar a las personas que no respondieron la pregunta sobre el ingreso total familiar (ITF). Los resultados indican que aproximadamente el 71.1% de quienes omitieron esta información (848 personas) se encuentran en condición de pobreza. Este porcentaje es considerablemente superior al 19.4% observado entre quienes sí proporcionaron su ingreso. Una explicación plausible es que muchas de las personas que no declaran su ingreso lo hacen porque en realidad no cuentan con ingresos suficientes, lo cual refuerza el patrón de vulnerabilidad detectado.