Universidad de Buenos Aires
Facultad de Ciencias Económicas
Escuela de Estudios de Posgrado

Histogramas, Kernels y Métodos No Supervisados usando la EPH

Trabajo Práctico N°3 - Grupo 15

DOCENTE: NOELIA ROMERO

ASIGNATURA: TALLER DE PROGRAMACIÓN

Repositorio GitHub - Grupo 15

INTEGRANTES: CHRISTIAN CAMPOS, ALEJANDRO ALCOCER, JULIÁN DELGADILLO MARÍN
POSGRADO: MAESTRÍA EN ECONOMÍA APLICADA

Fecha de Entrega: Viernes 14 de Noviembre de 2025

Resumen

El presente trabajo desarrolla un ejercicio de modelización predictiva orientado a la identificación de hogares en situación de pobreza utilizando microdatos de la Encuesta Permanente de Hogares (EPH) para los años 2005 y 2025. A partir de un conjunto de variables sociodemográficas y económicas, se construyeron y compararon distintos modelos de clasificación, entre ellos regresión logística y el algoritmo K-Nearest Neighbors (KNN).

El análisis incluyó la estimación de matrices de confusión, curvas ROC y métricas de desempeño como Accuracy, Recall, Precision y AUC, permitiendo evaluar la capacidad de los modelos para discriminar entre hogares pobres y no pobres. Los resultados muestran que el modelo Logit presenta un mejor desempeño general, mientras que el modelo KNN logra una mayor sensibilidad en la identificación de hogares pobres.

Adicionalmente, se aplicó el modelo seleccionado a la población de no respondientes en 2025, estimando la proporción de individuos potencialmente pobres dentro de este grupo. Este ejercicio permite discutir las implicancias de los errores de clasificación en el contexto de políticas públicas y la asignación eficiente de recursos escasos.

Palabras clave: pobreza, modelos de clasificación, regresión logística, KNN, curva ROC, políticas públicas

Cuadro 1. Tamaño de las bases de datos utilizadas

Base de datos	Filas (observaciones)	Columnas (variables)
Respondieron 2005	7043	174
Respondieron 2025	4309	174
No respondieron 2005	101	169
No respondieron 2025	2872	169

Cuadro 2. Distribución de la variable pobreza por año

Categoría	2005 (cantidad)	2025 (cantidad)
No pobre (0)	4833	2975
Pobre (1)	2210	1334
Total	7043	4309

El análisis del Cuadro 1 evidencia cambios importantes en la estructura de las bases de datos utilizadas. Entre 2005 y 2025 se observa una disminución significativa en el número de hogares respondientes, pasando de 7.043 a 4.309 observaciones, lo que podría reducir el nivel de representatividad y capacidad de generalización de los modelos estimados para 2025. Paralelamente, el número de no respondientes se incrementa de forma considerable, de 101 en 2005 a 2.872 en 2025, lo que convierte al tratamiento del sesgo por no respuesta en un elemento crucial. A su vez, las bases utilizadas difieren en número de variables: las observaciones respondientes poseen 174 variables frente a las 169 de la base de no respondientes, lo que sugiere evaluar si las cinco variables ausentes tienen relevancia predictiva para el modelo aplicado.

Por su parte, el Cuadro 2 confirma que los totales de hogares analizados (7.043 en 2005 y 4.309 en 2025) coinciden con las observaciones respondientes del Cuadro 1, reforzando que estas corresponden a las muestras efectivamente utilizadas para el entrenamiento y validación de los modelos. La proporción de pobreza se mantiene relativamente estable entre ambos años, siendo de aproximadamente 31,38 % en 2005 y 30,96 % en 2025. Esta estabilidad y el hecho de que la clase minoritaria represente cerca del 31 % en ambos casos muestran que la variable de pobreza se encuentra adecuadamente balanceada, evitando la necesidad de métodos adicionales de oversampling o undersampling y garantizando interpretaciones confiables de métricas como la precisión. En conjunto, los resultados enfatizan la importancia de considerar el aumento drástico de no respondientes en 2025 y su potencial impacto en la calidad de las estimaciones de pobreza si no se aborda adecuadamente.

Cuadro 3. Diferencia de medias entre conjuntos de entrenamiento y prueba (2005)

Variable	Media Train	Media Test	Diferencia
ch04	1.5203	1.5182	0.0021
ch06	32.7156	32.7405	-0.0249
ad_equiv_hogar	3.5518	3.5961	-0.0443
cbt_equiv	205.0700	205.0700	0.0000

Cuadro 4. Diferencia de medias entre conjuntos de entrenamiento y prueba (2025)

Variable	Media Train	Media Test	Diferencia
ch04	1.5249	1.5205	0.0044
ch06	36.9768	37.4323	-0.4555
nivel_ed	3.8034	3.9080	-0.1046
region	1.0000	1.0000	0.0000
aglomerado	32.7908	32.7695	0.0213
cat_ocup	1.2779	1.3503	-0.0725
cat_inac	1.6651	1.5886	0.0766
ad_equiv_hogar	2.7864	2.7806	0.0058
cbt_equiv	365177.0000	365177.0000	0.0000

El análisis presentado en los Cuadros 3 y 4 tiene como propósito verificar si los conjuntos de entrenamiento y prueba provienen de la misma distribución de probabilidad. Este contraste es fundamental, ya que diferencias de medias pequeñas sugieren que la partición de los datos fue aleatoria y adecuada, evitando problemas de sobreajuste y permitiendo que el desempeño observado sobre el conjunto de prueba sea un estimador confiable del rendimiento del modelo sobre datos no utilizados en la estimación.

En el Cuadro 3, correspondiente a 2005, se observa que las diferencias entre las medias de las variables en los conjuntos de entrenamiento y prueba son extremadamente reducidas, siendo la mayor de ellas de apenas $-0{,}0443$. Esto confirma que la partición realizada fue adecuada y que ambos conjuntos son estadísticamente similares, lo cual asegura una validación sólida y representativa del proceso de modelización para dicho año. Por su parte, el Cuadro 4, correspondiente a 2025, muestra un comportamiento similar, aunque destaca la variable `ch06`, para la cual se registra una diferencia de medias cercana a $-0{,}4555$, lo que indica una desviación moderada en la media —posiblemente de edad— entre los conjuntos. Si bien esta diferencia no invalida la partición, sí sugiere que podría influir en caso de observar discrepancias relevantes en el rendimiento del modelo de 2025 entre entrenamiento y prueba. En conjunto, ambas comparaciones demuestran que la división de datos es estadísticamente consistente en los dos períodos, lo que brinda confianza en la posterior evaluación del desempeño predictivo de los modelos estimados.

Cuadro 5. Tamaño de las matrices de entrenamiento y prueba

Año	Conjunto	Observaciones	Variables
2005	Entrenamiento (Train)	4930	10
2005	Prueba (Test)	2113	10
2025	Entrenamiento (Train)	3016	10
2025	Prueba (Test)	1293	10

El Cuadro 5 muestra que, aunque las bases originales contenían 174 variables, el proceso de modelización utilizó únicamente 10, lo que implica una selección de características que mejora la interpretabilidad del modelo y reduce el riesgo de sobreajuste. Asimismo, los tamaños de muestra en ambos años (7.043 en 2005 y 4.309 en 2025) coinciden plenamente con los hogares respondientes del Cuadro 1, confirmando que se empleó la totalidad de la información disponible. Finalmente, la división aproximada 70/30 entre entrenamiento y prueba en ambos casos es consistente con las buenas prácticas en aprendizaje automático y garantiza una validación adecuada del desempeño de los modelos.

El Cuadro 6 muestra los resultados del modelo Logit mediante el uso de Odds Ratios (OR), los cuales permiten identificar los factores que incrementan o reducen la probabilidad de que un hogar sea pobre. En términos generales, características como el tamaño del hogar (medido por `ad_equiv_hogar`) y el sexo del jefe de hogar presentan OR mayores que 1, indicando un aumento del riesgo de pobreza, mientras que variables como la edad y el nivel educativo presentan OR menores que 1, actuando como factores protectores. Asimismo, las categorías laborales e inactivas presentan OR significativamente inferiores a la categoría de referencia, sugiriendo que el grupo base es el que concentra las mayores probabilidades de pobreza. En conjunto, el modelo identifica adecuadamente los principales determinantes del riesgo de pobreza, coherentes con patrones socioeconómicos observados en la literatura.

Cuadro 6. Resultados del modelo Logit (coeficientes y odds ratio)

Variable	Coeficiente	Error estándar	Odds Ratio
Intercepto (const)	1.1981	0.5053	3.3139
ch04 (Sexo)	0.1135	0.0460	1.1202
ch06 (Edad)	-0.3474	0.0872	0.7065
ad_equiv_hogar	0.5468	0.0519	1.7277
aglomerado_33	0.6887	0.1310	1.9911
cat_ocup_1	-3.6362	0.7844	0.0264
cat_ocup_2	-1.6232	0.4929	0.1973
cat_ocup_3	-2.4812	0.4833	0.0836
cat_ocup_4	-7.6929	26.0473	0.0005
cat_inac_1	-3.0937	0.5449	0.0453
cat_inac_2	-2.6750	1.1819	0.0689
cat_inac_3	-2.4269	0.4962	0.0883
cat_inac_4	-1.5988	0.4979	0.2021
cat_inac_5	-2.2846	0.6630	0.1018
cat_inac_6	-1.7157	0.5869	0.1798
cat_inac_7	-1.8220	0.5652	0.1617
nivel_ed_2	-0.0556	0.2116	0.9459
nivel_ed_3	-0.1113	0.1443	0.8947
nivel_ed_4	-0.3256	0.1806	0.7221
nivel_ed_5	-1.0472	0.1891	0.3509
nivel_ed_6	-1.5780	0.2333	0.2064
nivel_ed_7	-0.2623	0.4390	0.7693

La Figura 1 muestra la relación entre el número de adultos equivalentes y la probabilidad predicha de pobreza, observándose una asociación claramente positiva: a medida que aumenta el tamaño del hogar, también lo hace la probabilidad estimada de pobreza. Este comportamiento es consistente con los resultados del modelo Logit presentados en el Cuadro **??**, donde `ad_equiv_hogar` presenta un Odds Ratio mayor que 1. Si bien se aprecia una concentración de hogares entre uno y cuatro adultos equivalentes, la dispersión en la probabilidad predicha muestra que esta variable, aunque importante, no es determinante por sí sola, requiriendo el aporte explicativo de otras características sociodemográficas incluidas en el modelo.

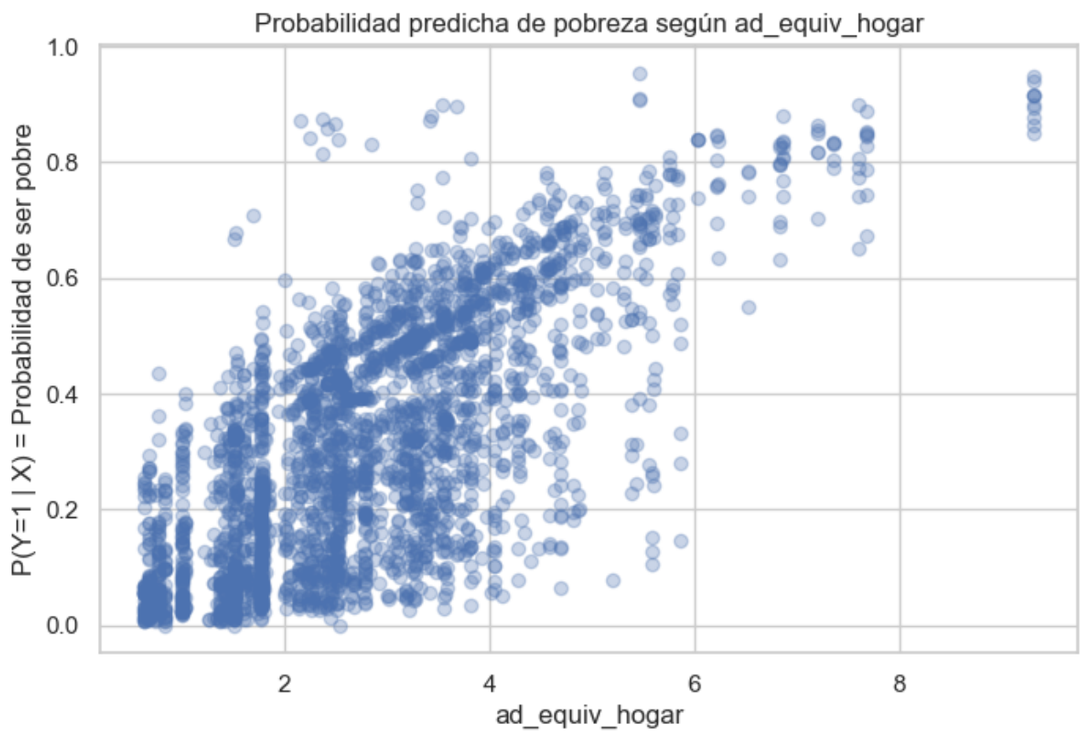

Figura 1. Probabilidad predicha de pobreza según adultos equivalentes del hogar

La Figura 2 muestra una relación negativa entre la edad del jefe de hogar y la probabilidad predicha de pobreza, donde a mayor edad la probabilidad estimada tiende a disminuir y la dispersión de los valores se reduce. Este patrón es coherente con los resultados del modelo Logit en el Cuadro **??**, donde la variable `ch06` presenta un Odds Ratio menor que 1, indicando que la edad actúa como un factor protector frente a la pobreza. Asimismo, se observa que los hogares encabezados por personas jóvenes presentan el mayor riesgo y variabilidad en la predicción, mientras que en edades avanzadas la probabilidad de pobreza se concentra en valores bajos, sugiriendo mayor estabilidad socioeconómica en estos grupos.

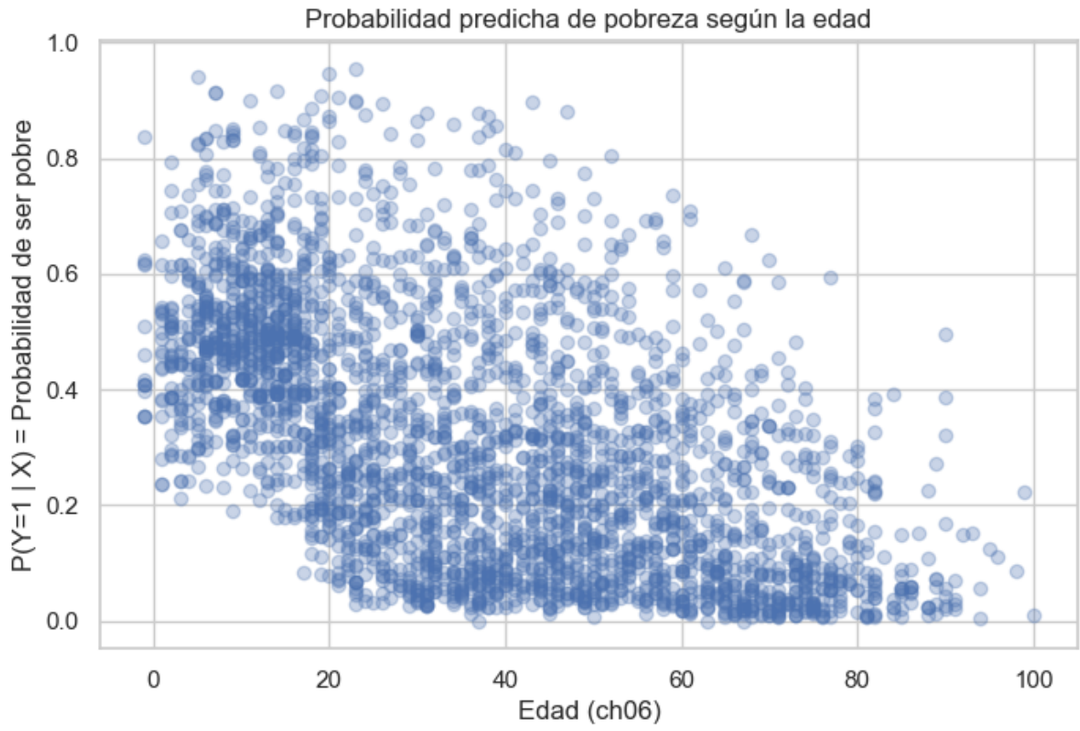

Figura 2. Probabilidad predicha de pobreza según la edad

La Figura 3 evidencia que el nivel educativo del jefe de hogar actúa como un fuerte factor protector frente a la pobreza, ya que conforme aumenta el nivel formativo la probabilidad predicha de ser pobre disminuye de forma clara. Este patrón coincide con los resultados del modelo Logit del Cuadro **??**, donde todas las categorías educativas presentan Odds Ratios inferiores a 1. Mientras los niveles educativos más bajos muestran mayores probabilidades y dispersión, los niveles altos concentran la mayoría de las predicciones en valores cercanos a cero, confirmando que la educación permite discriminar eficazmente entre hogares con mayor y menor riesgo socioeconómico.

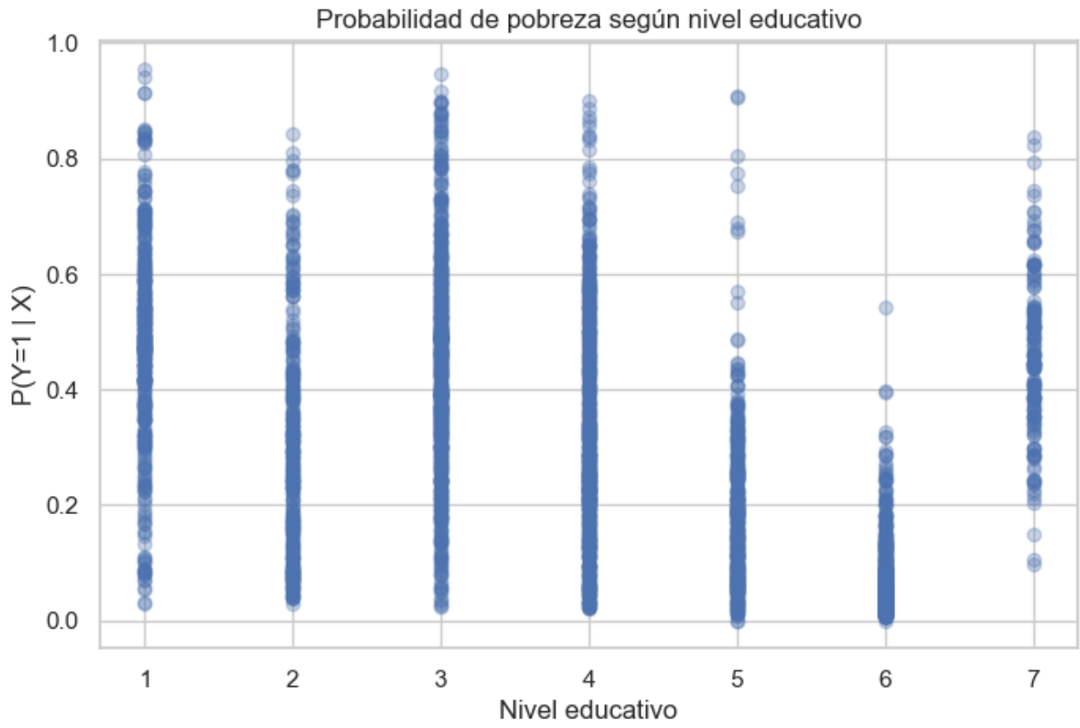

Figura 3. Probabilidad de pobreza según nivel educativo

El Cuadro 7 muestra el desempeño del modelo KNN para distintos valores de K, observándose que el valor $K = 1$ arroja un Accuracy de 0.99, lo cual constituye una señal clara de sobreajuste, dado que el modelo memoriza excesivamente los datos de entrenamiento y no generaliza adecuadamente. A medida que K aumenta a 5 y 10, el Accuracy cae a 0.80 y 0.77 respectivamente, reflejando un comportamiento más realista y estable. En consecuencia, valores intermedios como $K = 5$ resultan más adecuados para la comparación con el modelo Logit, al ofrecer un mejor equilibrio entre variabilidad y capacidad predictiva.

Cuadro 7. Desempeño del modelo KNN para distintos valores de K (base 2025)

Valor de K	**Accuracy**
1	0.99
5	0.80
10	0.77

La Figura 4 ilustra de forma clara por qué el modelo KNN con $K = 1$ tiende al sobreajuste. La frontera de decisión aparece altamente irregular y fragmentada, lo que indica que el algoritmo intenta clasificar de manera perfecta cada observación del conjunto de entrenamiento, en lugar de generar una separación general y estable entre las clases. Esto es coherente con el resultado del Cuadro **??**, donde el Accuracy alcanza valores cercanos a 1 para $K = 1$, señal de una memorización excesiva de los datos. Además, esta complejidad se traduce en una baja capacidad de generalización frente a datos nuevos, por lo que valores de K mayores resultan más adecuados para lograr un desempeño más equilibrado del modelo.

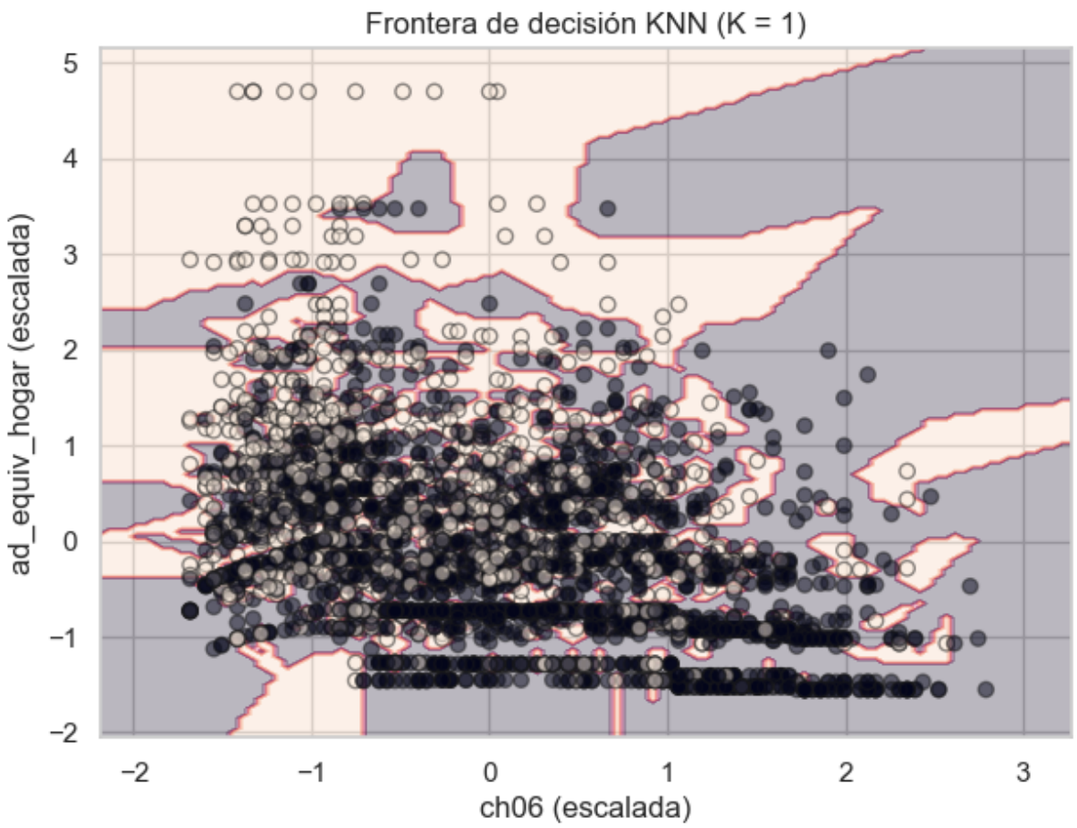

Figura 4. Frontera de decisión del modelo KNN con $K = 1$

La Figura 5 muestra la frontera de decisión del modelo KNN con $K = 10$, la cual se observa mucho más suave y menos fragmentada en comparación con la Figura 4 correspondiente a $K = 1$. Al considerar 10 vecinos, el modelo generaliza mejor y se vuelve menos sensible al ruido y a los outliers, reflejando un Accuracy más realista de 0.77 según el Cuadro 7. La clasificación identifica adecuadamente las zonas de riesgo de pobreza, asociadas a hogares con mayor número de adultos equivalentes y edad baja, y las áreas de protección, correspondientes a hogares con mayor edad y menos adultos equivalentes. Este análisis confirma que $K = 10$ es un valor de hiperparámetro robusto para la comparación con el modelo Logit.

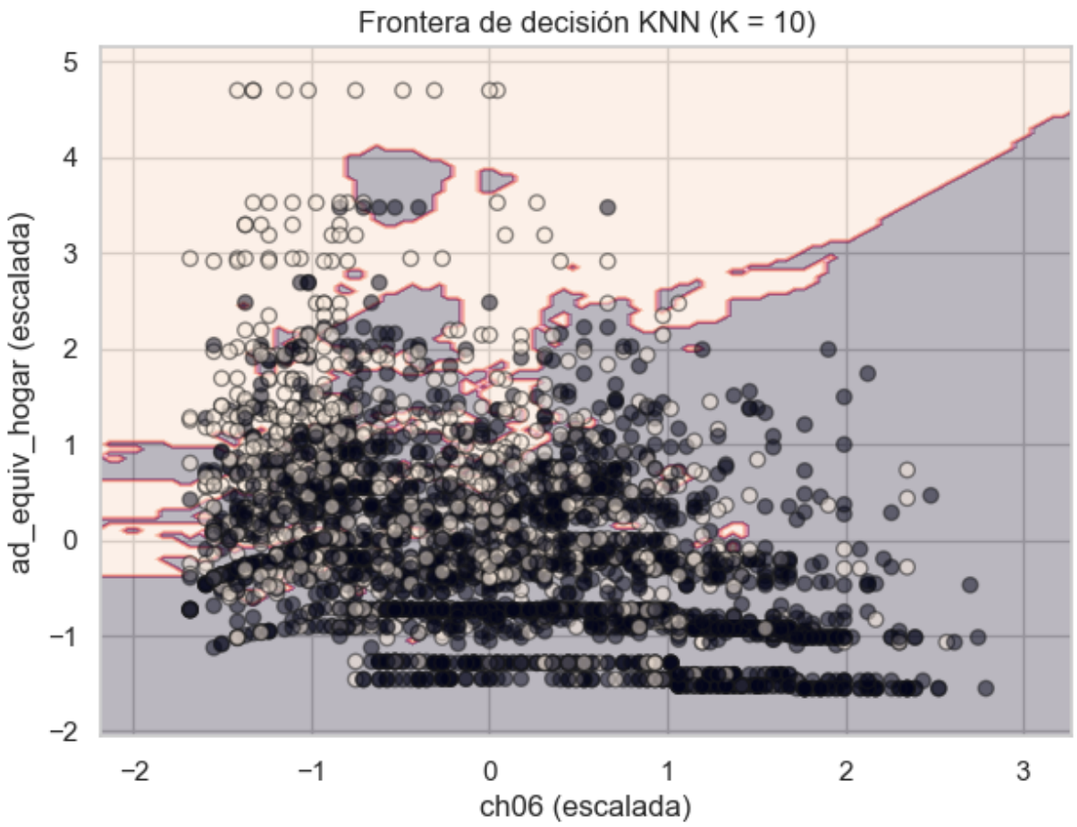

Figura 5. Frontera de decisión del modelo KNN con $K = 10$

El Cuadro 8 presenta los resultados de la validación cruzada (5-fold CV) para el modelo KNN sobre la base de 2025, proporcionando una estimación más realista y generalizable de su desempeño. Se observa que el Accuracy de $K = 1$ cae de 0.99 a 0.66 en la validación cruzada, confirmando el sobreajuste identificado previamente. A partir de $K = 2$ el Accuracy se estabiliza alrededor de 0.70, alcanzando un máximo de 0.71 para $K \geq 6$, lo que indica un desempeño robusto y consistente. Por lo tanto, el modelo KNN seleccionado con K entre 6 y 10 ofrece un 71 % de clasificaciones correctas y constituye una referencia fiable para comparaciones con el modelo Logit.

Cuadro 8. Resultados de Cross-Validation (5-fold) para KNN (Año 2025)

Número de Vecinos (K)	**Accuracy promedio (5-fold CV)**
1	0.66
2	0.70
3	0.70
4	0.70
5	0.70
6	0.71
7	0.70
8	0.71
9	0.71
10	0.71

La Figura 6 ilustra la evolución del Accuracy promedio (5-fold CV) del modelo KNN a medida que se incrementa el número de vecinos (K). Se observa un aumento rápido desde 0.66 en $K = 1$ hasta 0.70 en $K = 2$, seguido de una meseta entre $K = 2$ y $K = 5$, y alcanzando un máximo de aproximadamente 0.712 en $K = 6$. Esta visualización confirma que $K = 1$ está sobreajustado y que el valor óptimo de K es 6, asegurando un modelo KNN robusto y preciso con un desempeño de clasificación cercano al 71 %, listo para compararse con el modelo Logit.

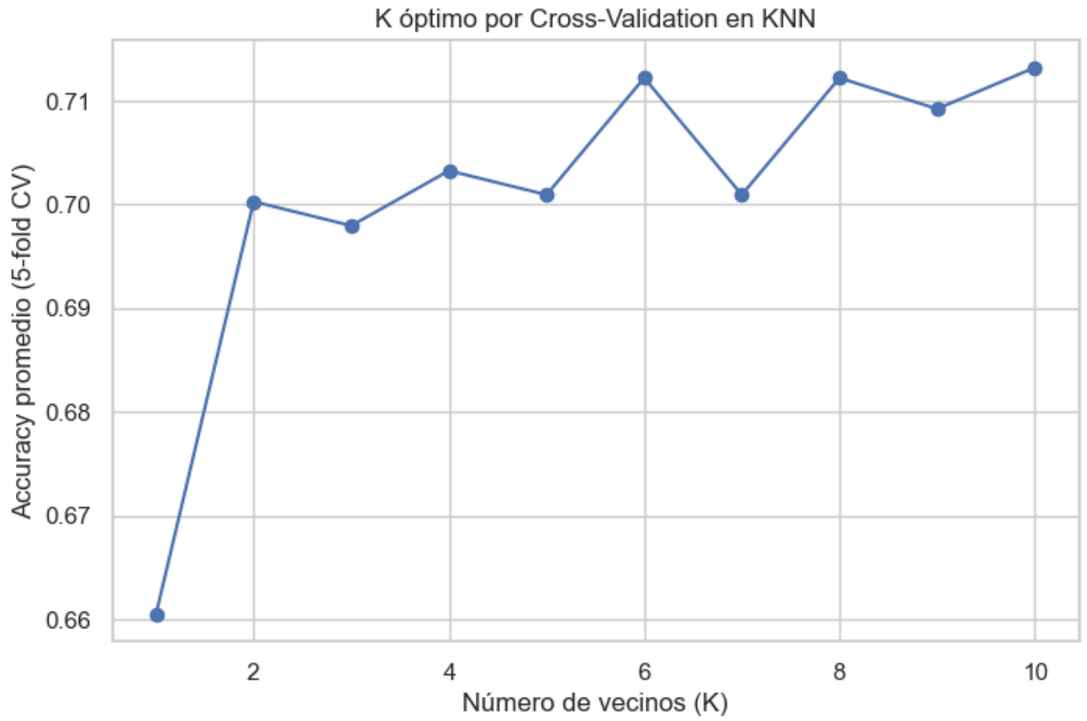

Figura 6. Selección del número óptimo de vecinos (K) por validación cruzada

La Figura 7 presenta la matriz de confusión del modelo Logit para el conjunto de prueba de 2025, donde la clase Pobre (1) es la positiva. El modelo clasifica correctamente 807 hogares No Pobres (VN) y 161 hogares Pobres (VP), mientras que existen 96 Falsos Positivos (FP) y 229 Falsos Negativos (FN). Esto da un total de 1293 observaciones, coincidiendo con el Cuadro 5. Las métricas calculadas muestran un Accuracy de 0.7486, indicando que el modelo acierta en casi el 75 % de los casos, una Precision de 0.6265 y un Recall de 0.4128. En consecuencia, aunque la precisión general es buena, la capacidad para identificar correctamente a los hogares pobres es limitada, dejando un alto número de Falsos Negativos que podrían implicar exclusión de beneficiarios en políticas sociales.

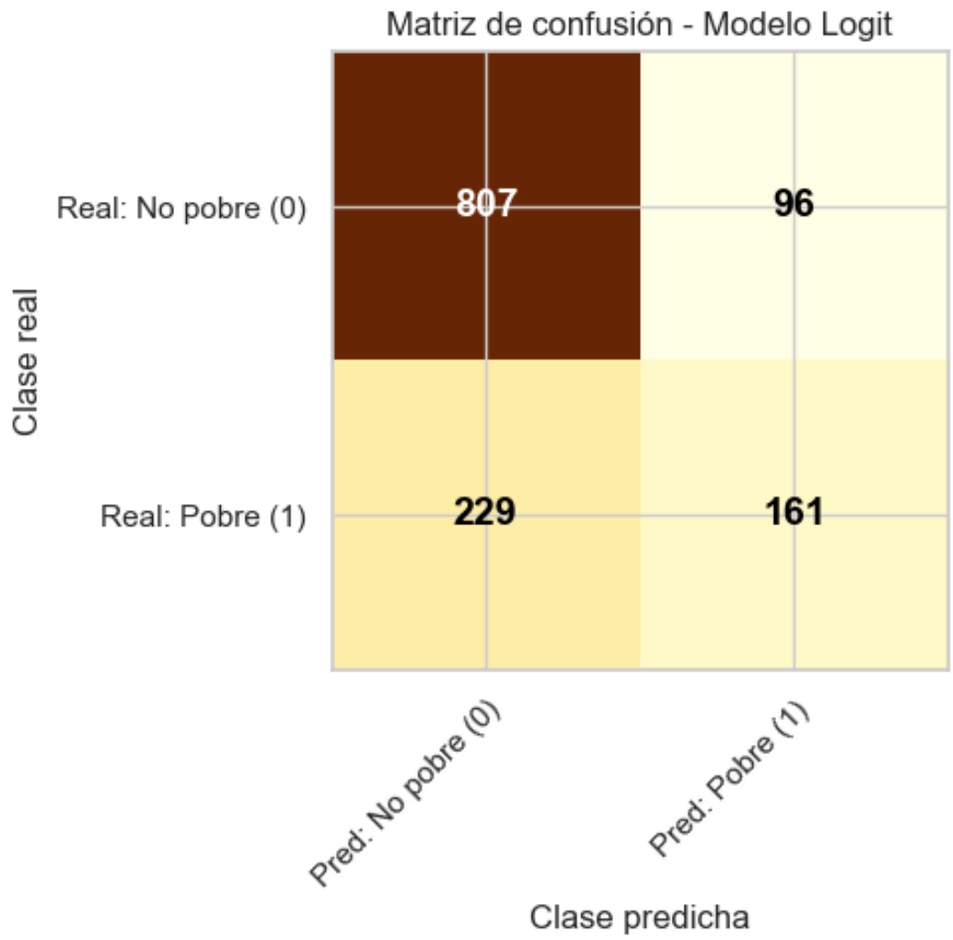

Figura 7. Matriz de confusión del modelo Logit

La Curva ROC compara la capacidad de discriminación de los modelos Logit y KNN (K=1). El modelo Logit presenta una curva muy por encima de la línea diagonal, con un AUC de 0.802, lo que indica que hay un 80.2 % de

probabilidad de clasificar correctamente un hogar pobre y uno no pobre. En contraste, el modelo KNN sobreajustado con $K = 1$ tiene un AUC de 0.647, reflejando un desempeño inferior y menor capacidad de discriminación en la mayoría de los umbrales. Esta comparación muestra que, en términos de AUC, el modelo Logit es superior, ofreciendo una combinación de buena discriminación y alta interpretabilidad, ya que permite identificar claramente los factores de riesgo y protección, como el tamaño del hogar, la edad y el nivel educativo. A pesar de su bajo Recall, el Logit se posiciona como el modelo más confiable para fines de política pública.

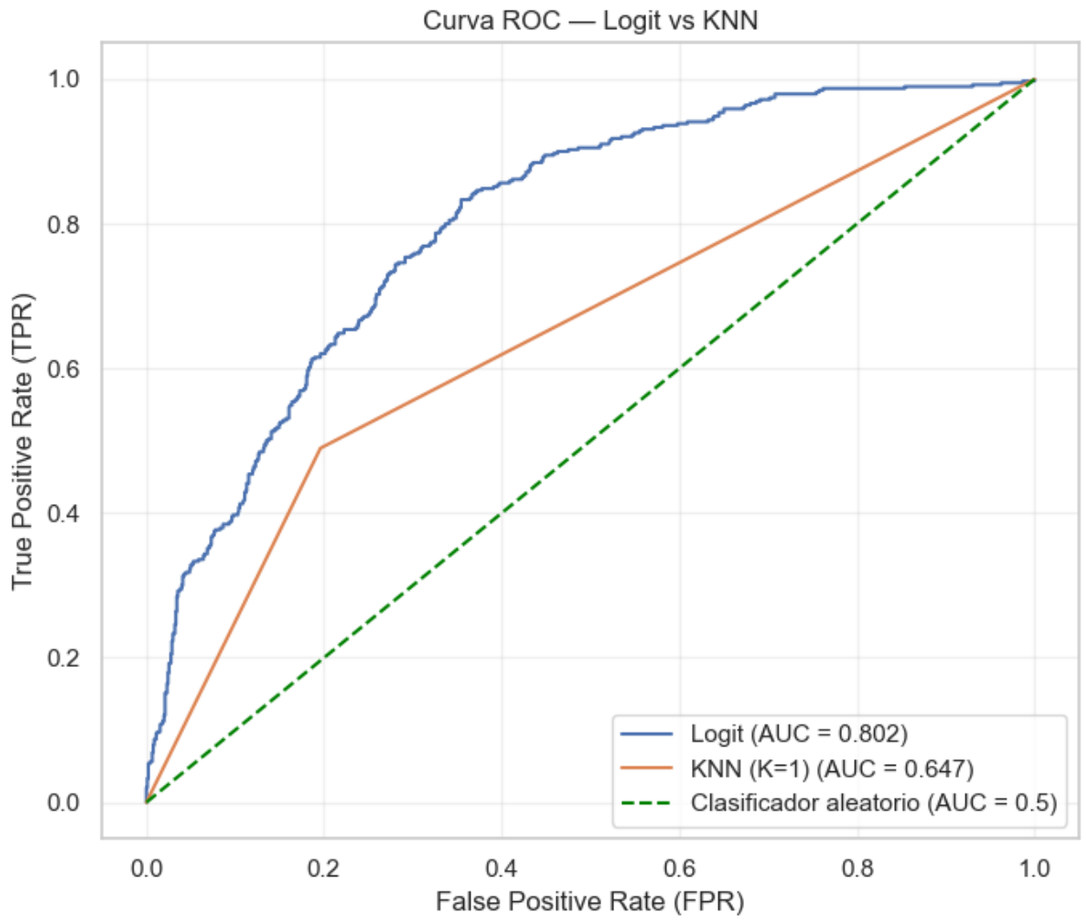

Figura 8. Curva ROC comparativa entre los modelos Logit y KNN

El modelo Logit presenta un mejor desempeño general frente al KNN ($K = 1$), con un Accuracy de 0.75 frente a 0.71 y un AUC de 0.80 frente a 0.65, lo que indica mayor capacidad de discriminación y confiabilidad en la clasificación de hogares pobres y no pobres. Ambos modelos muestran un Balanced Accuracy de 0.65, lo que sugiere que clasifican las clases minoritarias de manera similar.

En términos de asignación de recursos (Precision), Logit es más eficiente (0.63 vs. 0.52), minimizando los falsos positivos, mientras que KNN tiene mayor Recall (0.49 vs. 0.41), identificando un mayor porcentaje de hogares realmente pobres y reduciendo los falsos negativos. El F1-score es idéntico en ambos modelos (0.50), reflejando un compromiso similar entre eficiencia y cobertura.

Si la prioridad es la eficiencia en la asignación de recursos y la interpretabilidad de los factores de riesgo, el Logit es la opción recomendada debido a su mayor AUC y Precision. Existe un trade-off entre Precision (Logit) y Recall (KNN), por lo que la elección final dependerá de si la política pública prioriza cobertura o eficiencia. Se debe notar que el KNN utilizado en la comparación no es el óptimo ($K = 6$), cuyo desempeño habría sido ligeramente mejor, aunque probablemente Logit mantendría la ventaja en AUC. El siguiente paso es aplicar el modelo seleccionado a la población de no respondientes de 2025.

Modelo	Accuracy	Precisión	Recall	F1-score	AUC	Balanced Accuracy
Logit	0.75	0.63	0.41	0.50	0.80	0.65
KNN (K=1)	0.71	0.52	0.49	0.50	0.65	0.65

Cuadro 9. Comparación de métricas de desempeño entre los modelos Logit y KNN (año 2025).