


UBA FCE
Universidad de Buenos Aires
Facultad de Ciencias Económicas

Taller de Programación – Trabajo Práctico N.º 3 (2025)

Tema: Métodos No Supervisados y Visualización

Grupo 4

Ángel Enrique Zapata Barros

David Andrés Robalino Chica

Federico Walter Kisza

Docente: María Noelia Romero

Repositorio GitHub del Grupo:

https://github.com/Analisis1983/Grupo4_UBA_2025

Parte A: Enfoque de validación

1. Para verificar si cada uno de los predictores tienen distribuciones similares en el grupo de entrenamiento y en el grupo de testeo se analiza la diferencia de medias para cada variable. Se usa un muestreo estratificado porque al ser la variable dependiente binaria (pobre/no pobre), existe el riesgo de que las muestras queden desbalanceadas. De esta manera nos aseguramos que en ambas muestras se mantenga la misma proporción de pobres y no pobres. A continuación se muestran los resultados para cada predictor.

Tabla de diferencias de medias y estadísticas descriptivas

Región Patagónica

Variable	N entren.	Media entren.	Desvío entren.	N testeo	Media testeo	Desvío testeo	t-testeo	p-valor
Edad	5005	34,7	21,5	2145	34,5	21,6	0,4	0,7
Varón	5005	0,5	0,5	2145	0,5	0,5	-0,2	0,8
Años de educación	5005	9,5	5,7	2145	9,5	5,6	0,3	0,7
Ocupado	5005	0,4	0,5	2145	0,4	0,5	0,3	0,8
Tiene cobertura médica	5005	0,8	0,4	2145	0,8	0,4	-0,6	0,5

Fuente: Elaboración propia en base a INDEC

Se incluyeron las siguientes variables como predictores:

- Sexo: generalmente se observan diferencias en los niveles de pobreza entre hombres y mujeres, vinculadas al acceso al trabajo o a puestos jerárquicos (techo de cristal). Variable dummy
- Edad: las personas muy jóvenes suelen presentar mayor inestabilidad laboral o ingresos más bajos, mientras que los adultos en edades intermedias concentran las tasas más altas de empleo y mejores salarios. Por el contrario, las personas mayores pueden enfrentar mayores dificultades para reinsertarse laboralmente o depender de ingresos no laborales. Variable cuantitativa
- Años de educación: variable cuantitativa importante para definir la capacidad de acceso al trabajo en mejores condiciones y, por ende, a mayores ingresos.
- Cobertura médica: se creo una dummy que define si tiene o no a partir de todas las categorías que vienen en la EPH. Refleja de manera simplificada formalidad laboral, capacidad de pago o vulnerabilidad familiar.
- Se decidió omitir las dummies de cat_ocupacional a pesar de ser relevantes por simplicidad y para evitar multicolinealidad, además de que en algunas de las categorías (como patrón) hay muy pocas

observaciones lo que hacía que las diferencias entre medias sean significativas. Por otro lado, el objetivo es predecir pobreza y no grados de pobreza.

No se encuentran diferencias significativas entre las medias de entrenamiento y de testeo (p-valor mayor a 0,05 en todos los casos)

Parte B: Modelo de regresión logística

3.

Coeficientes, errores estándar y odd-ratios de modelo de regresión logística

Región patagónica

	Coeficiente	Error estándar	Odds ratio
Edad	-0,01	0,00	0,99
Varón	-0,27	0,10	0,76
Años de educación	-0,05	0,01	0,95
Ocupado	-0,66	0,12	0,52
Tiene cobertura médica	-1,68	0,11	0,19

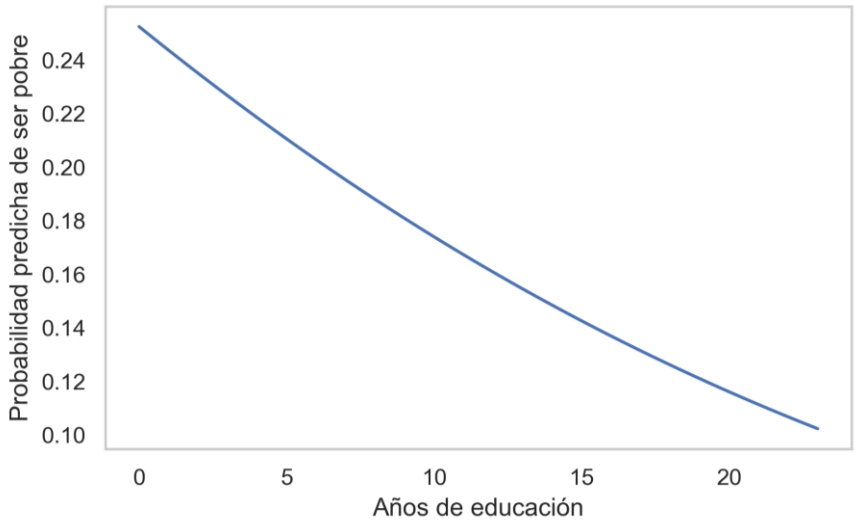
Fuente: Elaboración propia en base a INDEC

- Todos los coeficientes son negativos, lo que implica un efecto negativo de cada uno de ellos sobre la pobreza. Cabe destacar que los coeficientes miden el logaritmo de los odds
- Manteniendo todo lo demás constante, aumentar en un año la edad reduce en un 1% la probabilidad de ser pobre
- Ser varón, en promedio, reduce las chances de ser pobre en un 24% respecto a las mujeres.
- Cada año adicional de educación reduce las chances de ser pobre en un 5%
- Estar ocupado reduce la probabilidad de ser pobre en un 48%
- Tener cobertura médica las reduce en un 81%, siendo claramente el más significativo para predecir las probabilidades de ser pobre

4. Visualización:

Probabilidad estimada de ser pobre en función de la edad

Región Patagónica



Fuente: Elaboración propia en base a INDEC

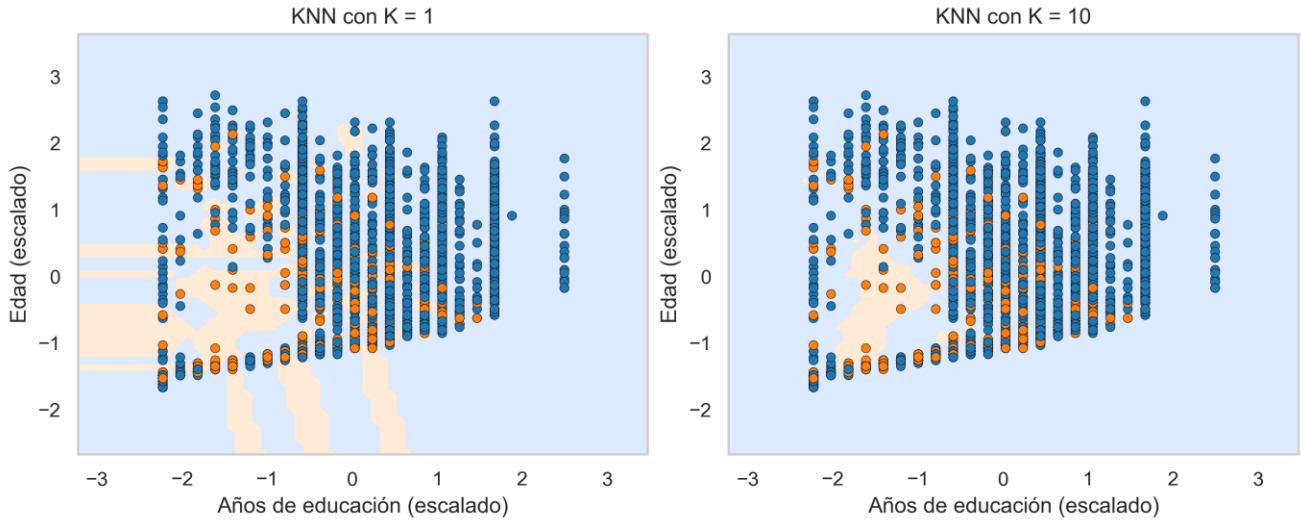
El gráfico refleja una relación inversa entre los años de educación y la probabilidad estimada de pobreza. Mientras que las con bajos estudios tienen un 20% o más de se pobres, en personas con estudios superiores baja a en torno al 10%.

Parte C: Método de Vecinos Cercanos (KNN)

5. Un valor con un K bajo genera un modelo con bajo sesgo pero alta varianza, ya que se ajusta casi perfectamente a los datos de entrenamiento, aumentando los errores cuando se intenta predecir valores por fuera de la muestra. En cambio, a medida que K aumenta, el modelo se vuelve más estable y generaliza mejor, pero a costa de aumentar el sesgo, ya que se suavizan las fronteras de decisión.

6. **Visualización**

Fronteras de decisión del modelo KNN para distintos valores de K



Fuente: Elaboración propia en base a INDEC

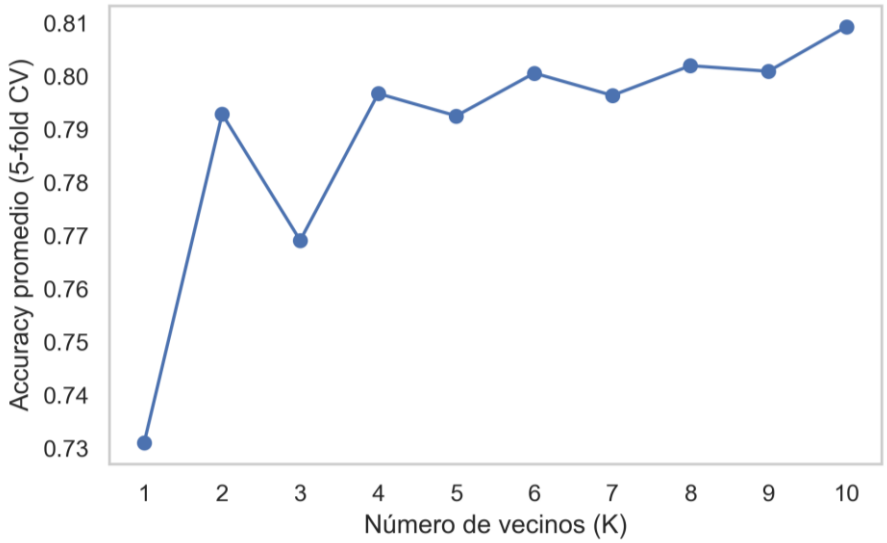
Con K=1 la frontera de decisión es muy irregular: el modelo sigue demasiado de cerca cada punto y termina sobreajustando, lo que genera clasificaciones caóticas en zonas donde las clases se mezclan. Con K=10 la frontera se vuelve más suave y estable, concentrando la probabilidad de pobreza en un área bien definida; aunque aumenta el sesgo y se pierden algunos casos en subgrupos pequeños. En ambos gráficos se observa lo señalado previamente: menores niveles de educación y edad se asocian con una mayor probabilidad de pobreza, concentrándose los puntos de pobres en el cuadrante inferior izquierdo.

7. K. óptimo por Cross Validation.

   El número óptimo de vecinos cercanos según el *accuracy* promedio es 10 y con él trabajaremos. De todas formas, podría elegirse el 6 en pos de mantener un buen equilibrio entre el sesgo y la varianza, ya que la ganancia no es tan marcada en los números de K siguientes.

Selección de K óptimo por Cross-Validation

Región patagónica



Fuente: Elaboración propia en base a INDEC

Parte D: Desempeño de modelos afuera de la muerta, métricas y políticas públicas.

8. Comparación de desempeño de predicción entre modelos Logit y KNN con K-CV

Matriz de confusión

Modelo Logit con umbral p > 0,5

	No pobre (Pred)	Pobre (Pred)
No pobre (Real)	934	44
Pobre (Real)	189	58

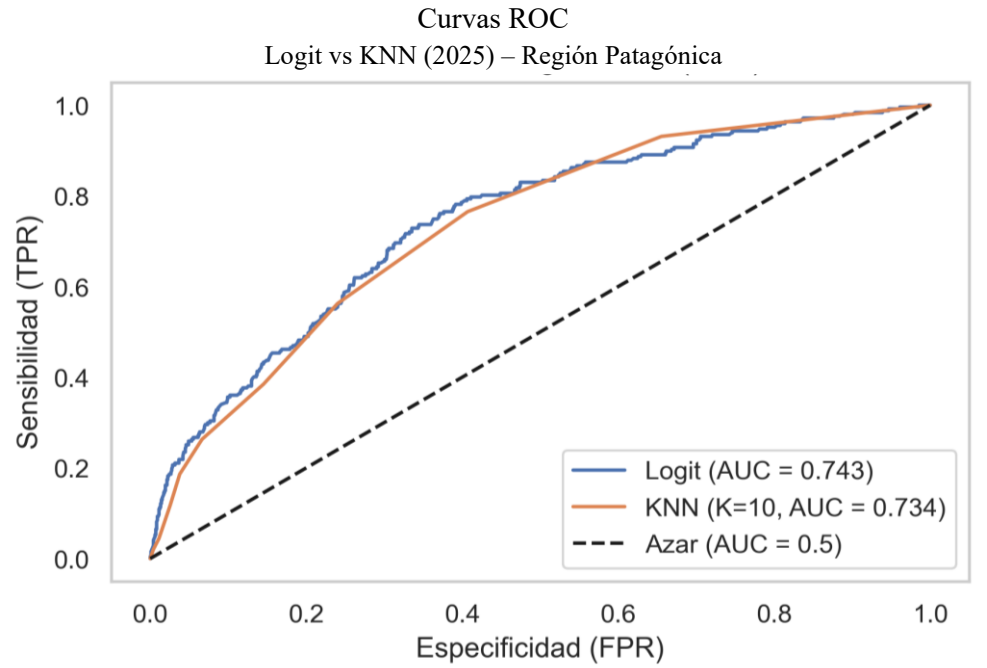
Modelo KKN

	No pobre (Pred)	Pobre (Pred)
No pobre (Real)	941	37
Pobre (Real)	201	46

Fuente: Elaboración propia en base a INDEC

El Logit logra identificar más pobres reales (TP = 58) que el KNN (TP = 46), por lo que comete menos falsos negativos (189 vs. 201). Esto es relevante porque el falso negativo representa a una persona pobre clasificada como no pobre. Por el contrario, el KNN presenta menos falsos positivos (37 vs. 44), es decir, clasifica a menos personas no pobres como pobres.

Curva ROC de ambos modelos



Fuente: Elaboración propia en base a INDEC

El modelo Logit presenta un área bajo la curva (AUC) apenas superior (0,743 vs. 0,734), lo que indica una leve mejor capacidad para distinguir entre pobres y no pobres en distintos umbrales de decisión. Ambos modelos superan ampliamente el desempeño del azar (AUC = 0,5), aunque la ventaja del Logit es consistente a lo largo de casi toda la curva.

Tabla de métricas de clasificación

Modelo	Accuracy	Recall (pobre=1)
Logit	0,810	0,235
KNN (K=10)	0,806	0,186

Fuente: Elaboración propia en base a INDEC

El desempeño de ambos modelos es similar en términos de *accuracy* (alrededor del 81%), pero existen diferencias importantes en la capacidad para identificar correctamente a las personas pobres, que es el objetivo principal del ejercicio. El modelo Logit presenta un recall (TP/TP+FN) mayor (0,235) que el KNN (0,186), lo que implica que detecta a una mayor proporción de pobres reales y comete menos errores tipo II (falsos negativos)

9. El hacedor de política pública debe minimizar el error de tipo 2 en este caso, es decir, dejar a la menor cantidad de pobres sin alimentos (la menor cantidad de falsos negativos). En este caso, el modelo que mejor se comporta es el Logit, ya que tiene un mayor recall (TP/TP+FN) y la tasa de falsos negativos (1 - recall) es más baja.
10. Un 16,7% del total de personas de la EPH de 2005 y 2025 que no respondieron la muestra fueron identificadas como pobres usando el modelo Logit construido en los incisos anteriores.