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

Taller de Programación

Maestría en Economía Aplicada

Tercer trimestre 2025

Trabajo Práctico N° 3

Link GitHub https://github.com/jorge-ux272/Big-Data-UBA--Grupo-8

Docente: Noelia Romero

Alumnos:
Jorge Eduardo Bolaños Gamarra
Mario Antonio Valdivia Reyes
Héctor Sebastián San Martin

Diferencia de medias entre las bases de entrenamiento y de testeo

Para este punto, se obtuvieron las medias de las variables entre los conjuntos de entrenamiento y prueba, junto con la diferencia, el estadístico t y el p-valor de una prueba de diferencia de medias. El objetivo es verificar si la partición de datos generó subconjuntos estadísticamente equivalentes.

La diferencia representa el cambio entre la media del conjunto de entrenamiento y la del conjunto de prueba. El estadístico t mide la magnitud relativa de esa diferencia, y el p-valor indica la probabilidad de observar una diferencia igual o mayor si las medias fueran realmente iguales. La hipótesis nula plantea que las medias son iguales en ambos subconjuntos, mientras que la hipótesis alternativa sostiene que son diferentes.

La mayoría de las variables presentan p-valores mayores a 0,05, lo que indica que no hay evidencia estadística suficiente para afirmar que las medias difieren significativamente entre los conjuntos de entrenamiento y prueba.

Algunas variables, como CH06, CH07_5, CH07_3 y NIVEL_ED_5, muestran diferencias moderadas, aunque susp-valores (superiores a 0,17) no alcanzan significancia estadística. En el caso de HORASTRAB y AEDUC no se calcularon el estadístico t ni el p-valor, posiblemente debido a falta de variabilidad o a la presencia de valores extremos.

En conjunto, la partición de datos puede considerarse balanceada, ya que no se observan diferencias significativas en la distribución de las variables entre ambos subconjuntos. Esto sugiere que el modelo entrenado no estará sesgado por una segmentación inadecuada de los datos.

Tabla 1. Tabla de diferencia de medias

Variable	Nombre	Media_Train	Media_Test	Diferencia	t-stat	p-value
AD_EQUIV_HOGAR	**Adulto equivalente del hogar**	0.579354	0.573187	0.006166	0.278670	0.780504
AEDUC	**Variable transformada de educación**	9.858156	9.825254	0.032902	-	-
CH06	**Edad en años cumplidos**	33.556361	34.052419	-0.496057	-1.366470	0.171822
CH07_1	**Estado civil: Soltero/a**	0.121867	0.122678	-0.000811	-0.152685	0.878650
CH07_2	**Estado civil: Casado/a o conviviente**	0.272978	0.274968	-0.001990	-0.275082	0.783259
CH07_3	**Estado civil: Separado/a o divorciado/a**	0.042172	0.046717	-0.004544	-1.347321	0.177908
CH07_4	**Estado civil: Viudo/a**	0.056283	0.059592	-0.003309	-0.869140	0.384791
CH07_5	**Estado civil: Unión libre / otro**	0.506700	0.495862	0.010839	1.337272	0.181164
HORASTRAB	**Horas trabajadas en la semana**	16.681482	15.691448	0.990035	-	-
NIVEL_ED_1	**Nivel educativo: Sin instrucción**	0.187451	0.190362	-0.002912	-0.458307	0.646742
NIVEL_ED_2	**Nivel educativo: Primaria incompleta**	0.179962	0.184293	-0.004331	-0.690910	0.489638
NIVEL_ED_3	**Nivel educativo: Primaria completa**	0.170266	0.171602	-0.001336	-0.218714	0.826877
NIVEL_ED_4	**Nivel educativo: Secundaria incompleta**	0.159231	0.155601	0.003630	0.616038	0.537883
NIVEL_ED_5	**Nivel educativo: Secundaria completa**	0.111777	0.105389	0.006388	1.273210	0.202972
NIVEL_ED_6	**Nivel educativo: Superior / universitario**	0.098691	0.102078	-0.003387	-0.693171	0.488218

Las características seleccionadas presentan valores promedio muy similares en ambas muestras, lo que sugiere que se tiene una partición adecuada y sin sesgo sistemático. En la mayoría de los casos, las diferencias de medias son pequeñas y los valores p son elevados, lo que indica que no existen diferencias estadísticamente significativas. Esto respalda la consistencia entre ambos subconjuntos y sugiere que el modelo podrá generalizar al no enfrentar cambios en la distribución de las variables explicativas. Cabe aclarar que las variables AEDUC y HORASTRAB generan un t-stat nulo debido a que su varianza es baja, tienen distribución casi idéntica en train y test, y, en el caso de HORASTRAB, hay gran concentración en el valor 0.

Segmentación según “respondieron” y “no respondieron”

El procedimiento consiste en dividir o segmentar las bases de datos originales en dos grupos temporales distintos: por un lado, se crea la base respondieron_2005 y la base norespondieron_2005 que contienen los registros correspondientes al año 2005; por otro lado, se generan las bases respondieron_2025 y norespondieron_2025 con los registros del año 2025. Esta separación permite analizar por separado la información de quienes respondieron y quienes no lo hicieron para cada uno

de los dos años, facilitando así comparaciones temporales y el estudio de la evolución de las características o comportamientos de cada grupo a lo largo del tiempo.

Tabla 2. Separación de la base respondieron

Categoría	2005 (obs.)	2025 (obs.)
Respondieron	14,760	3,363
No respondieron	198	1,406

La separación de la base entre los años 2005 y 2025 muestra una diferencia en el volumen de observaciones para ambos grupos. Esto refleja un cambio en el tamaño o cobertura de la muestra a lo largo del tiempo. La tabla evidencia variaciones en la composición temporal de las bases, lo cual es importante considerar en el análisis posterior para evitar sesgos derivados.

Estimación y efectos marginales de "respondieron" 2025

Se realizó la estimación de una Regresión Logística usando X_train de respondieron_2025, a partir de lo cual se exportó una tabla con los coeficientes estimados para cada variable, los errores estándar y los odd-ratios.

Tabla 3. Estimación y Efectos Marginales

Variable	Coeficiente	Error estándar	Odds Ratio (exp(beta))
Constante	1.7710	0.4443	5.8768
CH06 (Edad en años cumplidos)	-0.0186	0.0053	0.9815
HORASTRAB (Horas trabajadas en la semana)	-0.0095	0.0036	0.9906
AD_EQUIV_HOGAR (Adulto equivalente del hogar)	0.5200	0.0519	1.6820
AEDUC (Variable transformada de educación)	-0.2020	0.0208	0.8171

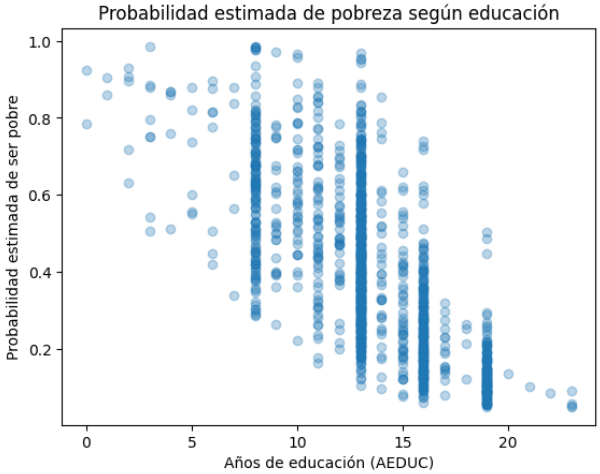
El modelo logit muestra que, si se mantienen constantes las demás variables, la probabilidad de ser pobre disminuye ligeramente a medida que aumentan la edad (OR = 0.98) y las horas trabajadas por semana (OR = 0.99), lo que sugiere que personas mayores y con mayor inserción laboral tienen un riesgo menor de pobreza. Por otro lado, el tamaño del hogar medido por adultos equivalentes incrementa la probabilidad de pobreza (OR = 1.68), indicando que hogares más grandes enfrentan mayor vulnerabilidad económica. Finalmente, se observa que la educación reduce la probabilidad de

ser pobre (OR = 0.82), mostrando que mayores niveles educativos actúan como un factor protector frente a la pobreza.

Relación entre años de educación y probabilidad de ser pobre

El Gráfico 1 muestra la relación estimada entre los años de educación (AEDUC) y la probabilidad de ser pobre, según el modelo logístico ajustado. Se observa una clara tendencia decreciente: a medida que aumentan los años de educación (AEDUC), la probabilidad estimada de encontrarse en situación de pobreza tiende a descender. Esto es consistente con la evidencia empírica habitual, donde mayores niveles educativos suelen asociarse con mejores oportunidades laborales y mayores ingresos. Si bien existe dispersión en los valores individuales (reflejada por los puntos del gráfico) la tendencia global indica que la educación actúa como un factor protector frente a la pobreza, reduciendo de manera progresiva la probabilidad prevista por el modelo.

Gráfico 1: Probabilidad estimada de ser pobre respecto a los años de educación.

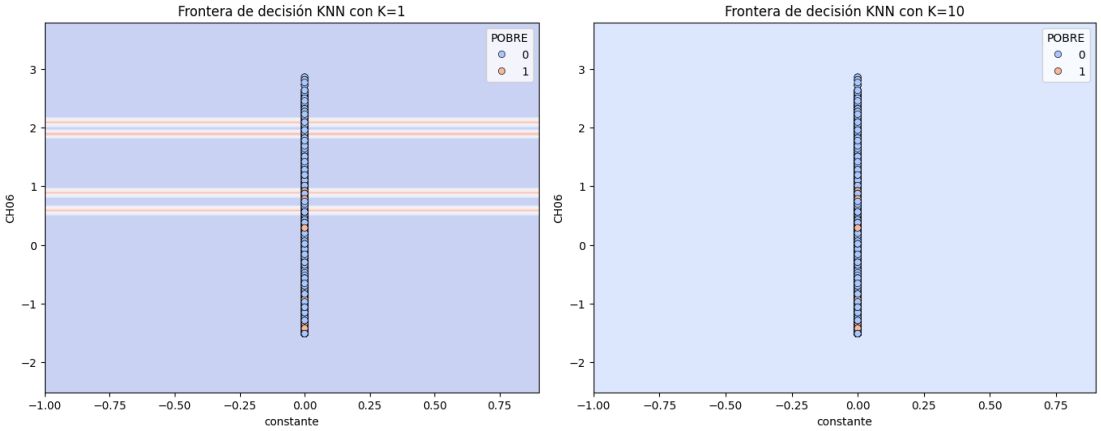


Estimación por métodos de vecinos cercanos

La elección de K en el clasificador KNN refleja el clásico trade-off sesgo–varianza. Con valores pequeños (como K=1), el modelo se ajusta muy de cerca a los datos de entrenamiento, lo que implica bajo sesgo pero alta varianza, aumentando el riesgo de sobreajuste. A medida que K aumenta, las predicciones se suavizan al promediar más vecinos, lo que reduce la varianza pero introduce mayor sesgo, pudiendo perder detalle en patrones locales. En el caso de valores K=5 y K=10, estos ofrecen un equilibrio más estable con desempeños prácticamente idénticos y ligeramente superiores al de K=1.

Punto 6: Visualización

Gráfico 2: Visualización de las características Edad (CH06) y constante.

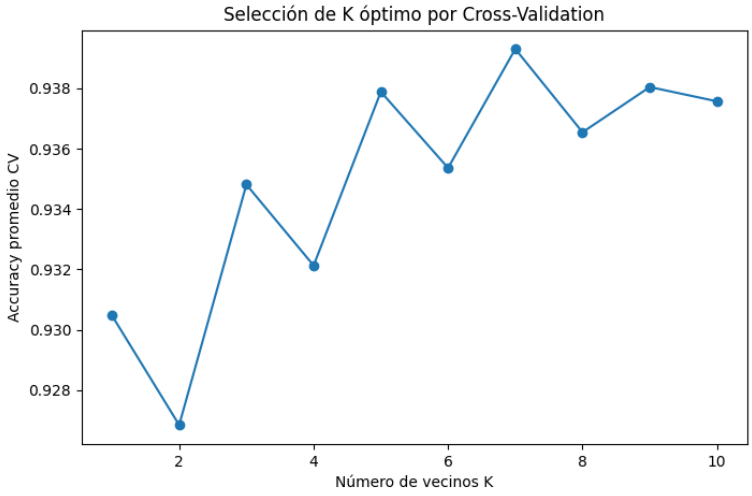


Los gráficos muestran la frontera de decisión del clasificador KNN utilizando las variables constante y edad para los casos de K=1 y K=10. Como era de esperar, la variable constante no aporta capacidad de separación, por lo que la clasificación depende casi exclusivamente de Edad. Con K=1, la frontera es altamente irregular y segmentada, reflejando una fuerte sensibilidad a valores puntuales del conjunto de entrenamiento y, por lo tanto, un riesgo elevado de sobreajuste (alta varianza). En cambio, con K=10 la frontera se vuelve mucho más suave y estable, ya que el modelo promedia la información de más vecinos. Esto reduce la varianza y genera una clasificación más robusta, aunque con un sesgo levemente mayor.

Determinación de K-óptimo por Cross-validation

Se hizo la división de la base X_train de “respondieron_2025” en 5 partes (5-fold) para obtener el K óptimo por Cross-Validation con K=(1,10).

Gráfico 3: Selección K óptimo por Cross-Validation



El gráfico muestra la evolución del accuracy promedio obtenido mediante validación cruzada para distintos valores del hiperparámetro *K* del modelo KNN. El desempeño presenta variaciones moderadas entre K=2 y K=10, alcanzando su valor máximo en K=7. Esto indica que dicho valor logra el mejor equilibrio entre sesgo y varianza, evitando tanto el sobreajuste asociado a valores muy bajos de K como el subajuste que puede aparecer con valores demasiado altos. En consecuencia, seleccionar K=7 como hiperparámetro óptimo favorece una mejor capacidad de generalización del modelo y un menor riesgo de errores sistemáticos en la predicción.

Comparación de desempeño: Logit vs. KNN (K=7)

Se evaluó el desempeño de dos modelos de clasificación para predecir pobreza sobre la base de test del año 2025. Los modelos comparados fueron regresión logística (Logit) y con K=7 (KNN). Se utilizaron tres métricas: Accuracy, F1-score y AUC (Área bajo la curva ROC).

La Curva ROC y la matriz de confusión brindan evidencia para concluir que el modelo KNN con K=7 (AUC=0.845) tiene un eficaz desempeño predictivo (AUC=0.742), ya que su curva se acerca más a la esquina superior izquierda, lo cual confirma que se tiene capacidad para discriminar entre las clases. A su vez, la matriz de confusión revela una buena precisión al predecir la clase "No pobre" (1120 aciertos) y a los "Pobres" (1171 aciertos o verdaderos positivos). Esto es consistente con un AUC de 0.742, lo cual confirma que, aunque el modelo es viable, el clasificador KNN sería preferible para esta tarea debido a su mayor poder de clasificación general.

Gráfico 4. Matriz de confusión y Curva ROC

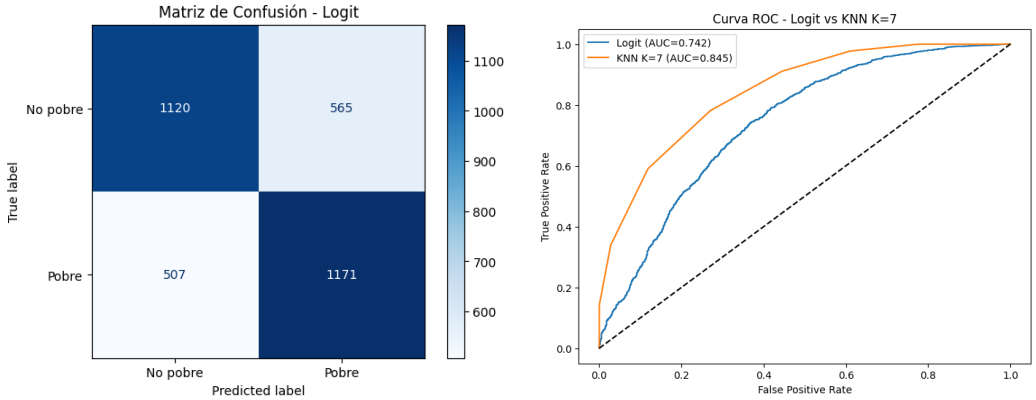


Tabla 4: Comparación de métricas modelos Logit vs KNN

#	MODELO	ACCURACY	F1-SCORE	AUC
0	LOGIT	0.681237	0.685999	0.742169
1	KNN K=7	0.755575	0.761463	0.844706

La matriz de confusión del modelo Logit (con umbral p > 0.5) muestra un desempeño moderado, con una sensibilidad de aproximadamente 70% para la clase "Pobre", aunque con una cantidad considerable de falsos positivos. Por otro lado, la curva ROC evidencia que el modelo KNN tiene mejor capacidad de discriminación entre clases, con un AUC superior (0.845 vs 0.742). En conjunto, el KNN supera al Logit en todas las métricas evaluadas, lo que indica una mayor eficacia para clasificar correctamente a los individuos pobres y no pobres. Si bien el modelo Logit ofrece ventajas interpretativas, el modelo KNN resulta más adecuado para tareas de predicción en este contexto.

Modelo de clasificación más adecuado para predecir pobres y asignar recursos

En el contexto del Ministerio de Capital Humano, el objetivo principal del programa es identificar correctamente a los grupos vulnerables para asignar recursos alimentarios escasos. Por ello, el modelo de clasificación “mejor” no es necesariamente el que obtiene mayor exactitud global (accuracy), sino aquel que minimiza el riesgo de cometer errores tipo II, es decir, clasificar como “no pobre” a un hogar que sí lo es. Este tipo de error tiene un costo social elevado, ya que implica dejar sin asistencia a quienes más la necesitan.

En este sentido, aunque el modelo KNN presenta un desempeño ligeramente superior en términos de accuracy, el modelo Logit resulta más adecuado para fines de política pública, ya que permite

interpretar las probabilidades estimadas y ajustar el umbral de decisión para priorizar la identificación de los hogares pobres (mayor recall). De este modo, se favorece una asignación más inclusiva y socialmente eficiente de los recursos del programa.

Predicción de personas pobres dentro de la base "no respondieron" 2025

Utilizando el modelo seleccionado (KNN con k=7), se aplicó el procedimiento de predicción sobre la base "*norespondieron_2025"*, replicando exactamente el mismo procedimiento usado en el entrenamiento: imputación por medias, estandarización y el mismo conjunto de variables explicativas. El modelo clasificó como pobres al 52,35% de las personas que no respondieron la encuesta en el año 2025.