Trabajo Práctico 3

Ciencia de Datos

Alumnas

Juana Cisneros, Francesca Lo Tartaro , Lucía Martin

Profesores

María Noelia Romero,Ignacio Anchorena

Asignatura

Ciencia de Datos

Teórica 1 - Tutorial 2.

Semestre y año de presentación

2do Semestre 2025

Cisneros, Lo Tartaro, Martin

Link al repositorio: https://github.com/lucia1martin/Trabajo-Practico-3

A. Enfoque de validación

1. **Tabla 1:** *Comparación de medias entre conjuntos de entrenamiento y testeo*

Variables	Media (Train)	Media (Test)	Diferencia	p-valor	¿Diferencia significativa?
Edad	52.603	53.108	−0.505	0.0807	No
Estado conyugal: Viudo/a	0.122	0.132	−0.010	0.0856	No
Estado conyugal: Divorciada/o	0.143	0.133	0.009	0.1231	No
Estado conyugal: Casado/a	0.286	0.281	0.005	0.5125	No
Estado conyugal: Soltero/a	0.231	0.233	−0.002	0.7407	No
Estado conyugal: Unido/a	0.218	0.220	−0.002	0.7958	No
Sexo: Varón	0.473	0.477	−0.004	0.6609	No
Horas trabajadas semanales	36.763	36.282	0.481	0.3173	No
Obra social en el trabajo	0.322	0.320	0.002	0.8223	No
Nivel educativo (N.E): Superior	0.226	0.216	0.010	0.1510	No
N.E.: Primario completo	0.176	0.183	−0.007	0.2782	No
N.E.: Superior Incompleto	0.115	0.123	−0.008	0.1350	No
N.E.: Secundario incompleto	0.151	0.145	0.006	0.2982	No
N.E.: Secundario completo	0.271	0.270	0.001	0.9461	No
N. E.: Sin instrucción	0.006	0.006	0.000	0.7916	No
N.E.: Primario incompleto	0.056	0.057	−0.001	0.7724	No

La partición 70/30 con random_state = 444 quedó balanceada: en la tabla de diferencias de medias ninguna variable muestra diferencias estadísticamente significativas entre entrenamiento y test (todos los p-values > 0,05), y los desvíos son pequeños. Los mayores son Edad (−0,505 años; p=0,0807) y Horas trabajadas semanales (0,481; p=0,3173), ambos no significativos, por lo que no se observa sesgo de muestreo relevante. La matriz X incluye dummies de nivel educativo (capital humano que reduce probabilidad de pobreza), estado conyugal (estructura del hogar y red de apoyo), obra social (señal de empleo formal/acceso a salud), sexo y edad (ciclo de vida laboral), además de horas trabajadas(intensidad de inserción laboral); todas son variables limpiadas en TPs previos y disponibles/replicables en *norespondieron*, cumpliendo la restricción de excluir ingreso para evitar fuga de información.

2. **Tabla 2:** *Tasa de respuesta y no respuesta: 2005 vs 2025*

Bases	Respondieron	No respondieron
2005	9278, 16	70,16

2025	8695, 16	1047, 16

En la tabla se puede observar que en *respondieron* hay buena cantidad de datos para poder estimar modelos (9278 casos en 2005; 8695 en 2025). Mientras que en *no respondieron* el contraste es alto (70 casos en 2005; 1047 en 2025). Esta falta de datos deja en desventaja a *no respondieron* para el año 2005 por su escasa cantidad de observaciones.
Con estos tamaños, el split 70/30 con random_state = 444 es adecuado para los casos que respondieron en 2005 y 2025. Luego, se aplica cada modelo a su respectivo grupo de no respondieron. Mientras que en 2025 los resultados serán más estables, en 2005 la precisión es limitada por el tamaño de muestra tan pequeño.

B. Modelo de Regresión Logística.

3.Estimación y Efectos Marginales

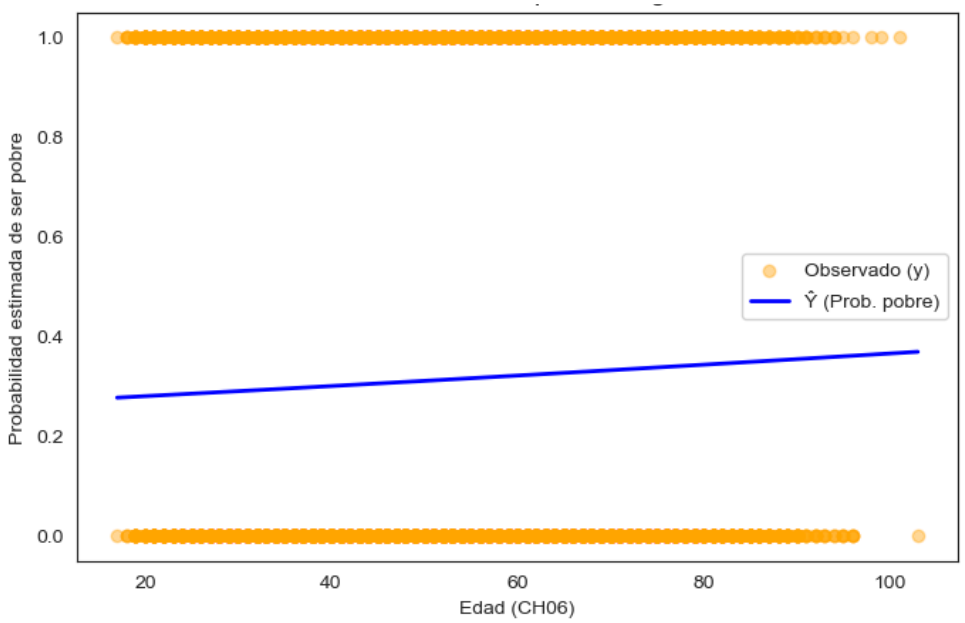
Tabla 3: *Regresión logística (X_train) sobre probabilidad de pobreza*

Variable	Coeficiente	Error Estándar	Odds Ratio
Intercepto	-0.6621	0.1323	0.5157
Edad	-0.0024	0.0016	0.9977
Horas trabajadas semanales	0.0015	0.0006	1.0015
Sexo: Varón	0.0196	0.0453	1.0198
Obra social en el trabajo	-0.5505	0.0508	0.5766
Estado conyugal: Unido/a	0.1731	0.0640	1.1890
Estado conyugal: Casado/a	0.1731	3.576335e+06	0.8946
Estado conyugal: Divorciada/o	0.3328	0.0742	1.3949
Estado conyugal: Viudo/a	0.2332	0.0869	1.2626
Estado conyugal: Soltero/a	0.2839	0.0639	1.3283
N.E. : Primario completo	0.0047	0.0964	1.0047
N.E. : Secundario incompleto	-0.1016	0.1008	0.9034
N.E. : Secundario completo	-0.1103	0.0952	0.8955
N.E. : Superior completo	-0.1409	0.0979	0.8685
N.E: Sin instrucción	-0.0618	0.2724	0.9401
N.E. : Superior incompleto	-0.1399	0.1083	0.8695

Estimamos una regresión logística con la matriz X Train (intercepto incluido y sin ingreso) y reportamos coeficientes, errores estándar y odds ratios. El resultado de los parámetros sugiere: obra social se asocia con menor probabilidad de ser pobre (OR = 0,576), es decir, las chances caen 42% manteniendo el resto constante; los estados conyugales (salvo “casado/a” (OR = 0,894)) elevan el riesgo de ser pobre; edad y horas trabajadas tienen efectos pequeños; y los niveles educativos apuntan en la dirección esperada, bajan el riesgo de ser pobre pero la evidencia es débil.

4.Visualización

Gráfico 1:*Probabilidad estimada de pobreza según la edad*



El gráfico muestra en el eje horizontal la variable Edad mientras que en el eje vertical la probabilidad estimada de ser pobre. Los puntos naranjas son los valores observados (0,1) mientras que la recta azul representa la predicción del modelo. La pendiente es levemente ascendente (acorde con un odds ratio cercano a 1), lo que resalta que la edad por sí sola tiene poco poder explicativo y que la categorización de ser o no pobre es explicada por otras variables .

C. Método de Vecinos Cercanos (KNN)

5.Estimación:

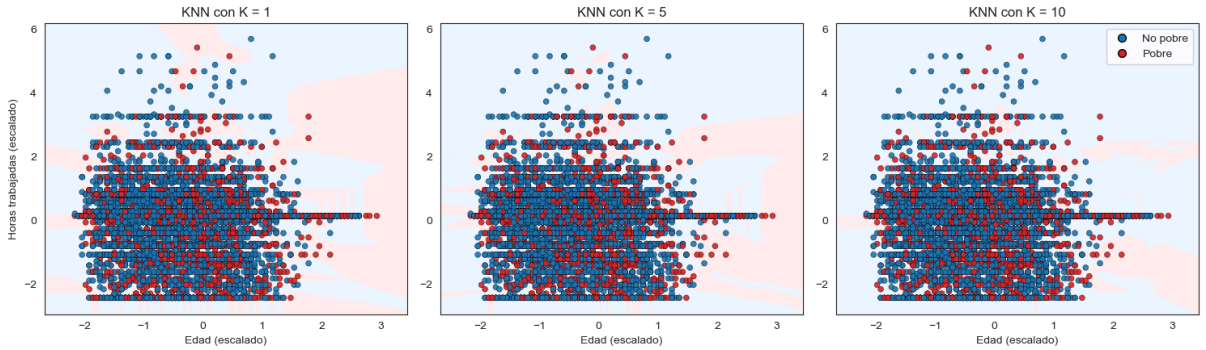
Tabla 4: *KNN para clasificación de pobreza por región (K = 1, 5, 10) con X_train*

K	Exactitud	Exactitud balanceada	Precisión 1	Recall 1	F1
1	0,595	0,527	0,352	0,346	0,349
5	0.628	0,518	0,353	0,222	0,272
10	0.664	0,513	0,374	0,108	0,167

A medida que K aumenta, la precisión global del modelo sube de 0,595 a 0,664, mientras que la capacidad del modelo para identificar correctamente los casos positivos cae de 0,346 a 0,108. Esto muestra que valores bajos de K capturan patrones locales, reduciendo el sesgo pero aumentando la varianza (*overfitting*), mientras que K altos hacen al modelo más general y estable, pero menos capaz de identificar la clase positiva (*underfitting*), reflejando el clásico trade-off sesgo-varianza.

6. Visualización:

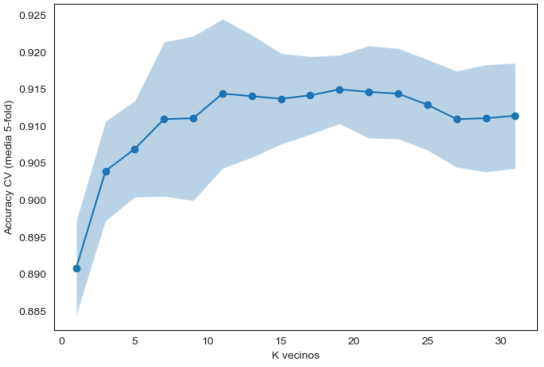
Gráfico 2: *Fronteras de decisión KNN (K = 1, 5 y 10) con Edad y Horas trabajadas*



El gráfico muestra cómo, para distintos valores de K, KNN pinta el plano como "pobre" (rojo) o "no pobre" (azul). Con K=1 el modelo tiene bajo sesgo y alta varianza: sigue mucho el dato local, detecta más pobres pero es más ruidoso. Con K=5 y K=10 la frontera se suaviza: sube el sesgo y baja la varianza, el modelo generaliza más y suele favorecer a la clase mayoritaria. En resumen, al aumentar K ganás estabilidad y algo de exactitud global, pero perdés sensibilidad para pobres. Si el objetivo es encontrar pobres, conviene un K más bajo.

7.

Gráfico 3: *Selección de K óptimo en KNN con K-CV (*5 particiones*)*



K óptimo:

El valor óptimo de K se determinó mediante validación cruzada de 5 particiones, evaluando el accuracy promedio para distintos valores de K. El K=19 fue seleccionado por presentar el mejor desempeño medio (accuracy CV = 0.915), reflejando el equilibrio entre sesgo y varianza del modelo.

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

8.

Gráfico 4: *Predicción de pobreza Logit y KNN con K-CV*

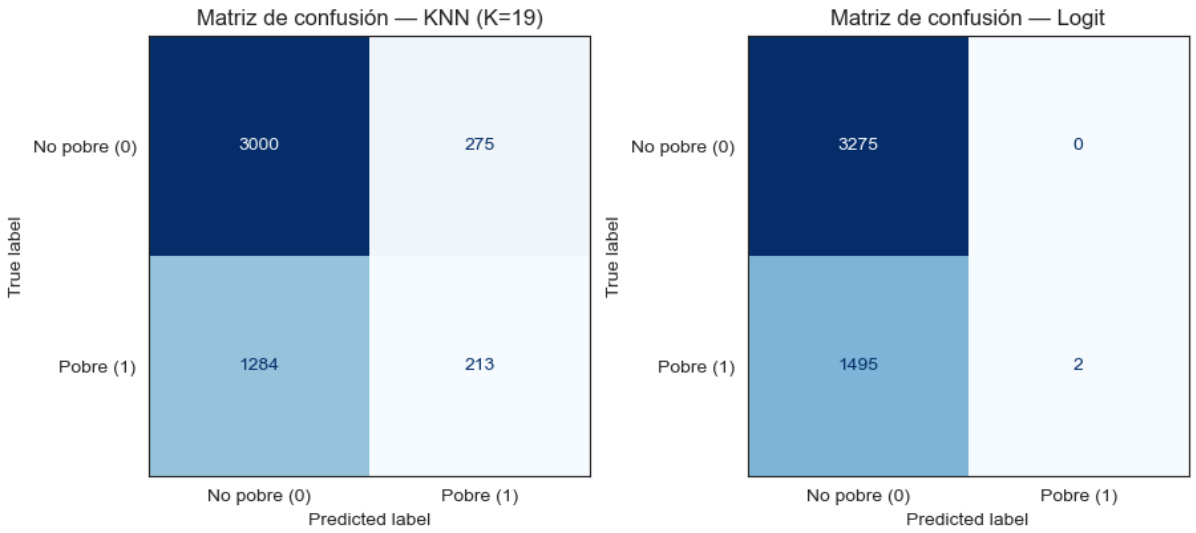
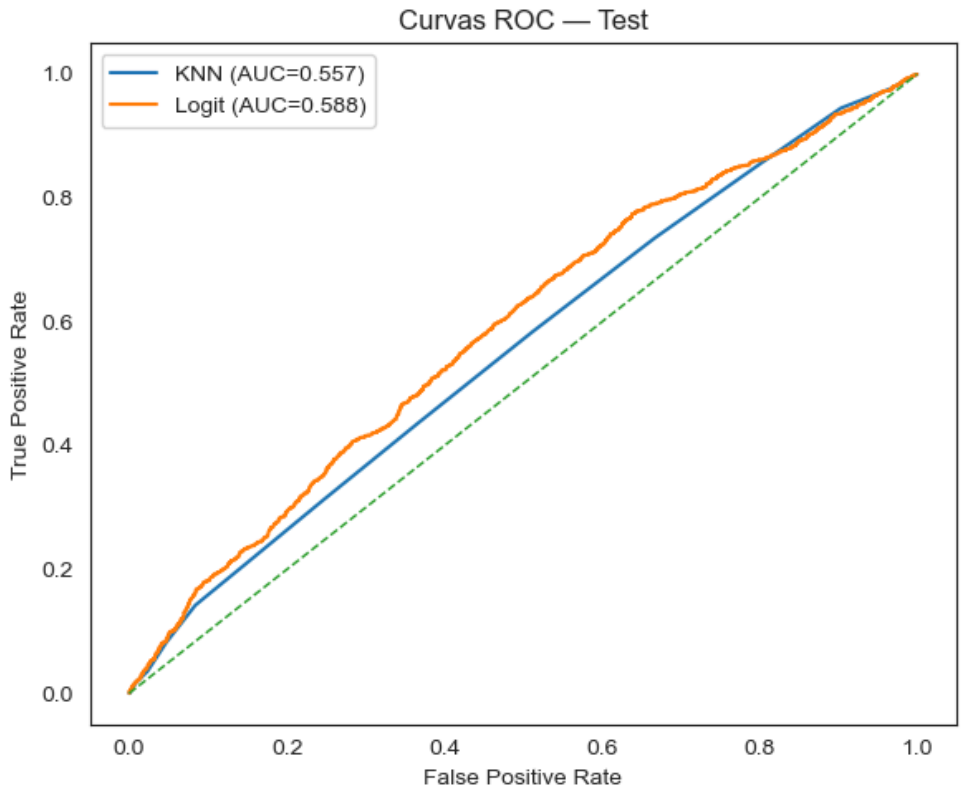


Gráfico 5: *Curva ROC: KNN & Logit*



Dos métricas más vistas en clase:

Tabla 5: *Desempeño fuera de muestra (test): comparación entre Logit y KNN (K=19)*

Modelo	Exactitud	AUC ROC
KNN (K = 19)	0,673	0,557
Logit	0,687	0,588

Ambos modelos logran un desempeño aceptable, con resultados similares en términos de precisión general. El Logit muestra una clasificación más conservadora, mientras que el KNN identifica más hogares pobres, aunque con algunos errores adicionales. Las curvas ROC reflejan un poder discriminatorio moderado, adecuado para un modelo basado en variables no monetarias.

9. Dado el objetivo del Ministerio de Capital Humano de identificar hogares vulnerables para dirigir recursos alimentarios, el costo de los errores no es simétrico: es más grave dejar fuera a un hogar pobre (error tipo II) que incluir por error a un hogar no pobre (error tipo I). En este contexto, aunque el modelo Logit presenta una precisión global ligeramente mayor, el modelo KNN (K=19) resulta más adecuado, ya que logra identificar una mayor proporción de hogares pobres aun con un leve aumento en falsos positivos. Aunque el Logit obtiene mejores valores de AUC y Accuracy, el KNN identifica una cantidad significativamente mayor de pobres reales (161 frente a 2), ofreciendo en consecuencia un mejor equilibrio entre errores tipo I y tipo II. En este tipo de política social, es preferible minimizar el error tipo II, es decir, evitar excluir a hogares pobres, aun cuando ello implique una ligera pérdida de eficiencia en la asignación de recursos.

10.

Tabla 6: *Predicción de pobreza en norespondieron 2025 (KNN, K=19)*

Conjunto	N casos	Pobres predichos (n)	Pobres predichos (%)	Proporción ponderada (%)
No respondieron 2025	1.047	123	11,75	10,86

Usando el modelo **KNN (K = 19)** seleccionado en el punto anterior, se aplicó la predicción de pobreza sobre la base *norespondieron 2025*. El modelo identificó aproximadamente **11,8 %** de los individuos como pobres, mientras que la proporción ponderada por los factores de expansión de la EPH asciende a **10,9 %**.

Estos resultados demuestran que, aun considerando a quienes no respondieron la encuesta, hay un grupo potencialmente vulnerable que podría ser considerado dentro de políticas de asistencia. El hecho de que las proporciones simples y ponderadas sean parecidas sugiere que la falta de respuesta no afecta demasiado la estimación de pobreza en este grupo.