Trabajo práctico Nº3

Ciencia de Datos

Grupo Nº4

Alumnos

Juan Augusto Alvarez Simonassi , Camila Gioja , María Delfina González Elosú

Profesores

María Noelia Romero

Ignacio Anchorena

Tomas Enrique Buscaglia

Parte A: Enfoque de validación

Para comenzar con el preparado de la base, utilizamos la base de datos respondieron, donde se almacenaron datos de personas que sí respondieron su ingreso (*ITF*), y agregamos variables creadas anteriormente como *EDAD2*, *EDUC* (referida a años de educación, una variable discreta), *horastrab* y *adulto_equiv* (referida al peso relativo individual en terminos de edad y sexo). Por otro lado, para elegir las variables de interés, primero convertimos en dummies algunas variables como por ejemplo *CH06* (referido a sexo), que la convertimos en *MUJER* (y 0 en caso contrario), *CH07* (referido a estado civil) que lo convertimos en *PAREJA* (para personas unidas o casadas, y otros casos como 0). También descompusimos en dummies variables como *CH08* (referido a la cobertura de salud) en *OBRA_SOCIAL*, *PREPAGA, PLAN_PUBLICO* y *SIN_COBERTURA,* separamos *ESTADO* (referido al estado de ocupación) en *OCUPADO, DESOCUPADO,* e *INACTIVO,* separamos *CAT_OCUP* (categoría de ocupación) en *PATRON, CUENTA_PROPIA, OBRERO,* y *FAMILIAR_SNREM,* y por último, separamos *CAT_INAC* (categoría de inactividad) en *ESTUDIANTE, AMA_CASA* y *JUBILADO*. Decidimos utilizar esas variables porque representan factores sociodemográficos, educativos y laborales que influyen directamente en la probabilidad de que un hogar se encuentre en situación de pobreza. No incluímos variables íntimamente relacionadas con el *ITF,* y tuvimos la precaución al utilizar variables relacionadas entre sí para evitar problemas de colinealidad. Se trabajó con una base de datos de *respondieron* para la base de datos del 2005 y otra para la del 2025.

Posteriormente, cada base se dividió en un conjunto de entrenamiento (70%) y un conjunto de prueba (30%) utilizando el método *train_test_split* con una semilla fija (*random_state = 444)* para asegurar reproducibilidad. El entrenamiento se realizó de forma estratificada, de modo que la proporción de personas pobres y no pobres se mantuviera constante en ambos subconjuntos. Luego, se elaboraron dos tablas de diferencias de medias entre las variables del conjunto de entrenamiento y el de prueba **(Tabla 1; Tabla 2)**. Esta tabla permitió verificar que no existieran diferencias significativas entre ambos grupos, garantizando así la representatividad de la muestra utilizada para el entrenamiento. Los resultados mostraron que las medias son muy similares entre *train* y *test*, y que ninguna variable presenta diferencias estadísticamente significativas (*p-value >0.05*), con la excepción de la variable *PREPAGA* para el 2005 (con una diferencia significativa leve). Esto indica que la partición de los datos fue adecuada tanto para 2005 como para 2025; los conjuntos de entrenamientos y testeo están balanceadas y los conjuntos son comparables.

Parte B: Modelo de Regresión Logística

Se estimaron dos modelos de regresión logística binaria para los años 2005 y 2025, con el objetivo de analizar los determinantes de la probabilidad de ser pobre. En ambos casos, la estimación se realizó utilizando el conjunto de entrenamiento (*x_train)* y se incluyeron las variables sociodemográficas relevantes mencionadas anteriormente. Los modelos finales conservaron únicamente aquellas variables significativas *($p < 0.05$)* y que no presentaran colinealidad, verificada mediante el factor de inflación de la varianza (VIF). Este indicador permitió detectar cuando dos o más variables explicativas estaban correlacionadas entre sí, asegurando independencia.

En el modelo del 2005 (**Tabla 3**), las variables *IX_TOT*, *SIN_COBERTURA*, y *ESTUDIANTE* presentan coeficientes positivos y significativos, indicando que una mayor cantidad de integrantes del

hogar, no tener cobertura médica o ser estudiante aumenta la probabilidad de ser pobre. Por el contrario, *EDUC,* y *horastrab* muestran coeficientes negativos y significativos,lo que sugiere que mayor nivel educativo y más horas trabajadas reducen la probabilidad de pobreza. La variable *EDAD* se incorporó junto con su término cuadrático (*EDAD2),* ya que ambos resultaron significativos. La variable *EDAD* mostró un coeficiente positivo y *EDAD2* un coeficiente negativo, evidenciando una relación no lineal con la pobreza; la probabilidad de ser pobre aumenta con la edad hasta cierto punto, pero luego disminuye. Decidimos mantener ambas variables en el modelo por su relevancia analítica, más allá de su alta correlación. Finalmente, las categorías laborales *INACTIVO*, *CUENTA_PROPIA* y *OBRERO* también mostraron asociaciones positivas, reflejando la mayor vulnerabilidad de los trabajadores informales o fuera del mercado laboral formal.

Para el 2025 **(Tabla 4)** vemos un modelo coherente y consistente con lo esperado en términos socioeconómicos. Las variables *IX_TOT* y *adulto_equiv* conservan coeficientes positivos y significativos, confirmando que los hogares más grandes o con mayor carga demográfica tienen más probabilidades de ser pobres. Las variables *EDUC* y *horastrab* mantienen efectos negativos, reafirmando el rol de la educación y del trabajo en la reducción de la pobreza. La variable *MUJER* presenta un coeficiente positivo, lo que indica que las mujeres tienen una probabilidad ligeramente mayor de ser pobres que los hombres, en línea con brechas persistentes de género. *SIN_COBERTURA* continúa siendo uno de los factores más relevantes, y *JUBILADO* muestra un efecto negativo, sugiriendo que los adultos mayores que perciben jubilación tienen menor riesgo de pobreza. La variable *PREPAGA* fue finalmente excluida del modelo por su escasa relevancia teórica y su posible solapamiento con el indicador de cobertura médica.

Los odds ratios permiten cuantificar estos efectos. En 2005, una persona sin cobertura médica tenía 6,8 veces más probabilidades de ser pobre que una con cobertura, mientras que cada año adicional de educación reducía la probabilidad relativa de pobreza en aproximadamente un 13%. En 2025, la magnitud de los efectos se mantiene coherente, aunque con menor intensidad: los individuos sin cobertura médica tienen 4,6 veces más probabilidades de ser pobres, mientras que los jubilados presentan una probabilidad 70% menor de encontrarse en esa situación. En ambos modelos, el Pseudo R^2 (~0.28) refleja un nivel de ajuste moderado pero apropiado para modelos sociales, indicando que la educación, el trabajo y el acceso a cobertura médica siguen siendo los principales mecanismos protectores frente a la pobreza, mientras que la estructura del hogar continúa siendo un determinante estructural clave. Sin embargo, hay algunas variables que dejaron de ser tan relevantes en la medición de la pobreza, como por ejemplo la EDAD, que dejó de ser significativa en el modelo de 2025.

Posteriormente se graficó la probabilidad predicha de ser pobre en función de las horas de trabajo por semana, y manteniendo el resto de las variables predictoras fijas (ceteris paribus). Elegimos hacer una ilustración para cada año para poder hacer una comparación entre ambos modelos. En ambos casos, la relación entre horas trabajadas y probabilidad de pobreza resulta ser negativa y monotónica; a medida que aumentan las horas trabajadas, disminuye la probabilidad de ser pobre. Sin embargo, como el rango de valores observados de la variable *horastrab* es relativamente acotado, la porción visible de la curva corresponde solo al tramo decreciente de la función logística. En síntesis, para ambos casos los gráficos muestran que las personas que trabajan menos horas (o no trabajan) tienen una probabilidad sustancialmente mayor de ser pobres, mientras que aquellos con jornadas laborales más extensas tienen una probabilidad menor.

Parte C: Método de Vecinos Cercanos (KNN)

En esta sección se estimaron tres modelos *K-Nearest Neighbors* (KNN) con K = {1, 5, 10} para los años 2005 y 2025, utilizando la matriz de entrenamiento (X_train) y la variable objetivo (y_train). Las variables predictoras fueron previamente estandarizadas mediante StandardScaler(), asegurando que todas tuvieran la misma escala en el cálculo de distancias.

Los modelos se entrenaron y evaluaron dentro del conjunto de entrenamiento, con el objetivo de observar cómo varía su desempeño a medida que cambia el número de vecinos (*K)*. En el caso de la base de 2005 **(Figura 3)**, el accuracy fue de 0.654 con K = 1, 0.703 con K = 5 y 0.711 con K = 10. Para la base de 2025 (**Figura 4**), los resultados fueron ligeramente superiores: 0.680 con K = 1, 0.723 con K = 5 y 0.730 con K = 10. Se observa, en ambos períodos, una mejora progresiva en el desempeño del modelo a medida que aumenta el número de vecinos, alcanzando su valor máximo con K = 10.

Para visualizar este comportamiento, se graficaron (**Figura 3 y 4**) las fronteras de decisión del modelo KNN para las variables *EDUC* (nivel educativo) y horastrab (horas trabajadas). En el caso de K=1, la frontera es muy irregular y fragmentada, ya que el modelo clasifica cada observación en función de su vecino más cercano, lo que genera sobreajuste y una representación visual "dura" o poco suave. A medida que K aumenta (K=5 y K=10), las fronteras se vuelven más suaves y estables, representando una mejor capacidad de generalización. Esto es lo que ilustra el trade-off sesgo-varianza. En K = 1 hay poco sesgo y mucha varianza, y a medida que vamos aumentando el número de vecinos, la varianza disminuye y el sesgo aumenta. En términos interpretativos, las regiones con menor nivel educativo y menor cantidad de horas trabajadas concentran una mayor proporción de observaciones clasificadas como pobres, mientras que los individuos con mayor educación y más horas trabajadas presentan una menor probabilidad de pobreza. Si bien visualmente los gráficos pueden parecer irregulares, esto se debe principalmente a la naturaleza de las variables elegidas; ambas toman valores discretos por lo que al graficar la probabilidad predicha o las fronteras de decisión, los puntos tienden a superponerse o agruparse en líneas verticales, lo que da lugar a una visualización con bloques o zonas densas.

Luego de analizar el comportamiento del modelo con distintos valores de K, se buscó determinar el número óptimo de vecinos que maximiza el desempeño del clasificador. Para ello, se implementó una validación cruzada de 5 particiones (CV=5) sobre la base de entrenamiento, probando valores de K entre 1 y 30. En cada iteración, el modelo se construyó mediante un *Pipeline* que incluye el escalado de variables (*StandardScaler*) y el clasificador *KNeighborsClassifier*, evaluando el *accuracy* promedio obtenido en los distintos folds de validación (**Figura 5**). El gráfico muestra cómo varía el *accuracy* a medida que aumenta el número de vecinos. Se observa que el desempeño mejora rápidamente para valores bajos de K y luego se estabilizó. El valor óptimo de K fue K = 25, donde el accuracy promedio alcanzó 0.76. Con este valor se entrenó el modelo final KNN con K óptimo (K-CV), que se utilizó en las etapas siguientes del trabajo para poder evaluar su desempeño en la base de test y predecir pobreza en la base de *norespondieron_2025*. Este procedimiento permite reducir el riesgo de sobreajuste y seleccionar el modelo que logra el mejor equilibrio entre sesgo y varianza, basándose únicamente en los datos de entrenamiento.

Parte D: Desempeño de modelos, elección y predicción por fuera de la muestra

Con el objetivo de evaluar el desempeño del modelo logístico, se implementó un análisis basado en la métrica AUC (Área Bajo la Curva ROC) **(Figura 6)**. Se utilizaron las funciones del módulo

sklearn.metrics para calcular la curva ROC. El modelo obtuvo un AUC de 0.823, lo cual indica que tiene una buena capacidad de discriminación entre positivos y negativos. En términos prácticos, esto significa que el modelo tiene una probabilidad del 82.3% de asignar una probabilidad de pobreza mayor a un hogar pobre que a uno no pobre, lo que refleja un desempeño predictivo muy satisfactorio. La curva ROC se mantiene claramente por encima de la diagonal aleatoria, mostrando un buen equilibrio entre sensibilidad y especificidad. Cuando observamos específicamente la matriz de desempeño **(Tabla 5)**, se observó la precisión promedio (≈ 0.77) y el F1-score (≈ 0.74) que indican que el desempeño general fue aceptable. El modelo identifica bien la clase negativa (*recall* = 0.619) aunque tiene menor sensibilidad para la clase positiva (*recall* = 0.619). La exactitud global es de 77.2% lo que sugiere un rendimiento consistente, pero con un margen de mejora en la detección de positivos. Con respecto a la matriz de confusión **(Tabla 6)**, se puede observar que hay más Falsos Negativos que Falsos Positivos, aunque esta relación está bien balanceada.

Para el año 2025, se implementó también un modelo de vecinos cercanos con el objetivo de evaluar el desempeño de predicción de la condición socioeconómica de los individuos (pobre/no pobre). La curva ROC **(Figura 7)** alcanzó un 0.811, lo que indica que tiene una buena capacidad discriminatoria. Esto significa que la probabilidad que el modelo asigne una puntuación de probabilidad más alta a un caso positivo que a un caso negativo. Hay casi la misma cantidad de falsos positivos y falsos negativos. En la matriz de confusión **(Figura 8)** se observa un 74% de exactitud global, lo que indica que clasifica correctamente tres de cada cuatro observaciones. El F1-score promedio de aprox 0.73 respalda la consistencia del modelo **(Matriz de confusión - Figura 8)**. Comparando generalmente los modelos, se puede ver que el modelo logístico tiene un AUC ligeramente superior (0.823 vs. 0.811), lo que indica una mejor capacidad para diferenciar entre clases. La diferencia no es grande, ambos tienen un buen nivel de discriminación. A su vez, el logit logra un 3.2% más de exactitud, mostrando un rendimiento algo superior en el conjunto de prueba. El modelo de regresión logística presenta mayor precisión en la clase "No pobre" (0.809) y menor en la clase "Pobre" (0.691). El KNN tiene un mayor balance entre las clases, igualando la precisión y mejorando el recall para la clase "Pobre" (0.69 frente a 0.619). KNN es entonces más sensible a los casos positivos, aunque sacrifica algo de precisión general.

Se puede decir que si el objetivo principal es maximizar la precisión y la estabilidad global, el modelo Logit es preferible, pero si el objetivo es identificar mayor sensibilidad a la población en situación de pobreza, el modelo de NKK tiene un mejor equilibrio entre sensibilidad y precisión. Con respecto al punto 9, dado que el propósito del Ministerio es minimizar los falsos negativos y asegurar que las personas pobres sean detectadas, el KNN con K óptimo (K-CV) puede considerarse más adecuado, ya que presenta mayor recall (0.69), identificando un porcentaje más alto de hogares vulnerables, aún a costa de una leve pérdida de exactitud general.

Finalmente, con el método seleccionado (KNN-CV) como el más adecuado, se aplicó el modelo sobre la base norespondieron25 con el objetivo de predecir la probabilidad de pobreza entre quienes no declararon su ingreso. El modelo determinó como óptimo un número de vecinos K = 25 (con *weights = uniform*), alcanzando un *accuracy* promedio de 0.76 durante la validación cruzada. Al aplicarlo sobre los individuos de la base, el modelo clasificó como pobres al 43.6% de las personas (474 de 1088 observaciones). Este resultado indica que a 4 de cada 10 personas que no declararon su ingreso serían consideradas pobres según el modelo, un valor coherente con la tasa observada entre quienes sí

respondieron. Esto sugiere que el modelo mantiene una consistencia razonable fuera de la muestra, y que puede servir como una herramienta útil para identificar grupos vulnerables en contextos de información incompleta.

Anexo

Tabla 1.

Comparación de medias entre conjuntos train y test (2005)

Variable	*MEAN train*	*MEAN test*	Diferencia	*p-valor*
EDAD2	1570.862795	1546.610119	24.252676	0.656106
EDAD	33.910019	33.617560	0.292460	0.660933
horastrab	16.273772	16.162202	0.111569	0.886268
EDUC	10.664327	10.601190	0.063136	0.618703
PREPAGA	0.063816	0.081101	-0.017285	0.045425
PAREJA	0.421825	0.435268	-0.013443	0.405328
OBRERO	0.350989	0.340030	0.010959	0.479187
SIN_COBERTURA	0.429164	0.418899	0.010265	0.523988
INACTIVO	0.421187	0.430804	-0.009617	0.551269
JUBILADO	0.080408	0.072173	0.008236	0.336696
AMA_CASA	0.115826	0.122768	-0.006941	0.513593
IX_TOT	4.597320	4.604167	-0.006847	0.917285
ESTUDIANTE	0.278239	0.283482	-0.005243	0.720884
PATRON	0.017549	0.021577	-0.004028	0.382027
CUENTA_PROPIA	0.089981	0.093006	-0.003025	0.748423
adulto_equiv	0.826305	0.829152	-0.002847	0.520184
OCUPADO	0.434269	0.431548	0.002722	0.866248
OBRA_SOCIAL	0.534142	0.536458	-0.002317	0.886757

Nota: No se observan diferencias significativas salvo en la variable PREPAG (p = 0.045), lo que indica una adecuada aleatorización del muestreo

Tabla 2

Comparación de medias entre conjuntos train y test (2025)

Variable	*MEAN train*	*MEAN test*	Diferencia	*p-valor*
EDAD2	1886.255397	1912.592593	-26.337195	0.705490
EDAD	37.382077	37.671415	-0.289338	0.723421

EDUC	10.843992	10.814815	0.029177	0.847049
OBRA_SOCIAL	0.611813	0.588794	0.023019	0.203140
OBRERO	0.340122	0.323837	0.016286	0.346884
CUENTA_PROPIA	0.109165	0.125356	-0.016191	0.177190
SIN_COBERTURA	0.364155	0.379867	-0.015712	0.378580
AMA_CASA	0.123422	0.113010	0.010411	0.377942
PLAN_PUBLICO	0.008554	0.016144	-0.007590	0.078243
MUJER	0.529124	0.536562	-0.007438	0.685792
ESTUDIANTE	0.246436	0.240266	0.006170	0.695946
INACTIVO	0.439511	0.445394	-0.005883	0.747997
IX_TOT	3.897352	3.891738	0.005614	0.936105
OCUPADO	0.439919	0.444444	-0.004526	0.804751
PATRON	0.011405	0.014245	-0.002840	0.502690
DESOCUPADO	0.026477	0.024691	0.001785	0.757410
JUBILADO	0.098982	0.100665	-0.001683	0.879098
horastrab	15.703870	15.702754	0.001116	0.998891

Nota: No se observan diferencias estadísticamente significativas en ninguna variable

Tabla 3

Coeficientes del modelo LOGIT 2005

Variable	Coeficiente	Error estándar	Odds Ratio
const	-2.7214	0.2910	0.065785
EDAD	0.0382	0.0145	1.038907
EDAD2	-0.0005	0.0002	0.999473
IX_TOT	0.3505	0.0260	1.419770
EDUC	-0.1378	0.0148	0.871246
SIN_COBERTURA	1.9192	0.0967	6.815440
horastrab	-0.0130	0.0028	0.987118
INACTIVO	0.4003	0.1746	1.492225
CUENTA_PROPIA	0.8281	0.2532	2.288946
OBRERO	0.4814	0.2130	1.618286
ESTUDIANTE	0.5771	0.1660	1.780836

Tabla 4

Coeficientes del modelo LOGIT 2025

Variable	Coeficiente	Error estándar	Odds Ratio

const	-4.624846	0.559377	0.009805
IX_TOT	0.501122	0.032574	1.650573
adulto_equiv	3.139798	0.683838	23.099196
EDUC	-0.077722	0.016350	0.925222
MUJER	0.640131	0.165149	1.896729
PREPAGA	1.341555	0.339576	3.824986
SIN_COBERTURA	1.521949	0.105964	4.581147
horastrab	-0.018243	0.002666	0.981922
JUBILADO	-1.237526	0.265853	0.290101

Figura 1

Función logística de la probabilidad de pobreza según horas de trabajo - 2005

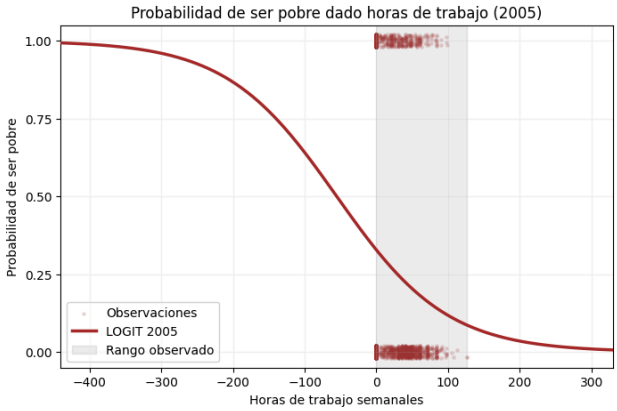


Figura 2

Función logística de la probabilidad de pobreza según horas de trabajo - 2025

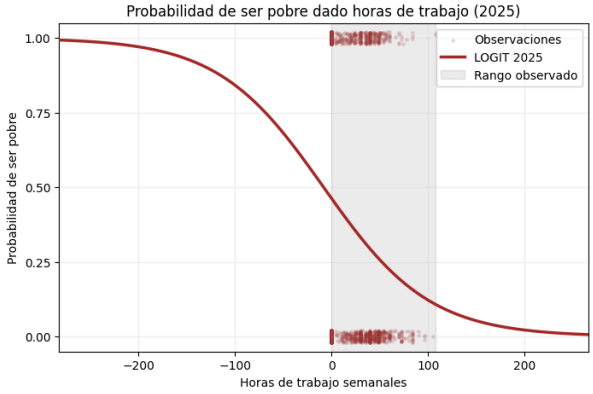


Figura 3

Frontera de decisión en vecinos cercanos (KNN) - 2005 para las variables horastrab y EDUC

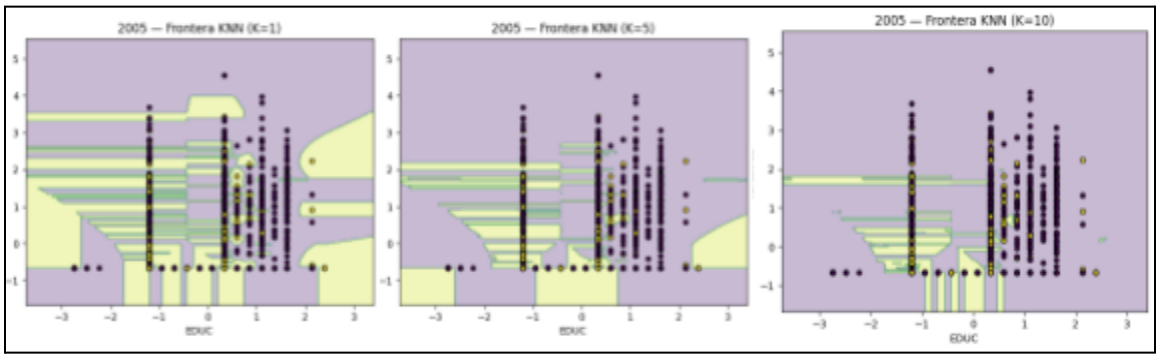
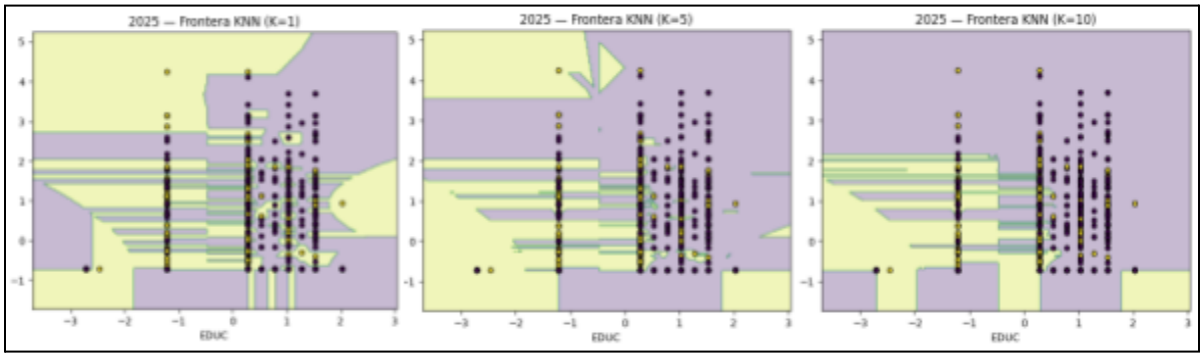


Figura 4

Frontera de decisión en vecinos cercanos (KNN) - 2025 para las variables horastrab y EDUC



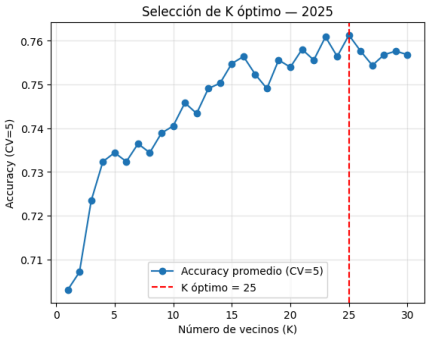


Figura 5

Selección de K óptimo - 2025

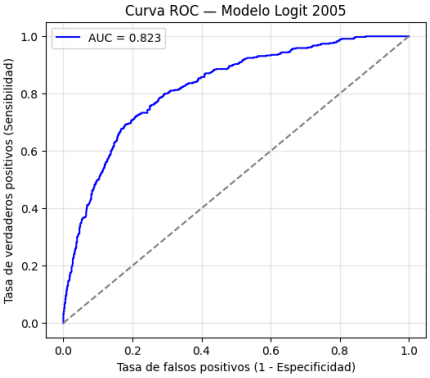


Figura 6

Curva ROC - Modelo Logit 2005

Tabla 5

Matriz de desempeño de modelo logístico 2005

Clase	Precisión	Recall	F1-score	Support
0	0.809	0.853	0.831	879
1	0.691	0.619	0.653	465
Accuracy			0.772	1344
Macro avg	0.750	0.736	0.742	1344
Weighted	0.768	0.772	0.769	1344

Tabla 6

Matriz de confusión del modelo logístico 2005

	Predicho 0	Predicho 1
Real 0	750	129
Real 1	177	288

Figura 7

Curva ROC - Modelo KNN 2025

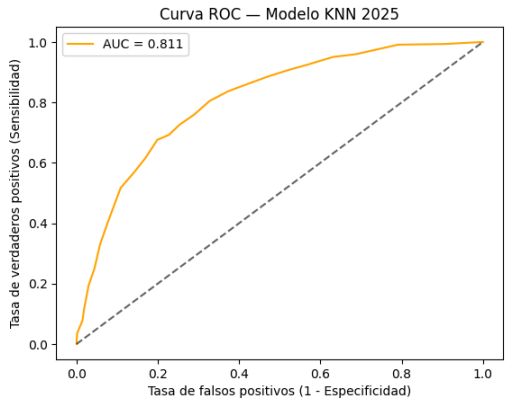


Tabla 7

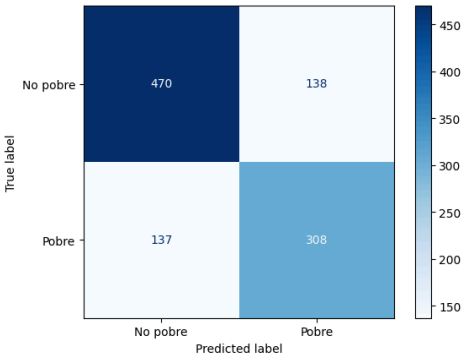
Matriz de desempeño de KNN

Clase	Precisión	Recall	F1-score	Support
0	0.77	0.77	0.77	608
1	0.69	0.69	0.69	445
Accuracy			0.74	1053

Macro avg	0.73	0.73	0.73	1053
Weighted	0.74	0.74	0.74	1053

Figura 8

Matriz de confusión de KNN 2025



Github: