FCE-UBA
UNIVERSIDAD DE BUENOS AIRES
FACULTAD DE CIENCIAS ECONÓMICAS

BIG DATA Y MACHINE LEARNING

TRABAJO PRÁCTICO N° 4

"CLASIFICANDO DE POBRES CON LA EPH"

(Encuesta Permanente de Hogares)

GRUPO 3

Rafael Pablo Pinto Chambi - Reg. 908586

Javier Rodolfo Aguirre - Reg. 819698

Lautaro Manuel Bogado – Reg. 899596

Enlace de repositorio: https://github.com/RafaelPCh/BigDataUBA-Grupo3.git

Docente a cargo: María Noelia Romero

Segundo Cuatrimestre de 2025

A.

1

Para la estimación de la variable objetivo pobre, se seleccionó un conjunto de variables de la EPH que cubren dimensiones socioeconómicas clave. Sin embargo, el set inicial de predictores presentaba problemas severos de separación perfecta, lo cual impedía la convergencia del modelo y la obtención de coeficientes interpretables. La modificación metodológica más significativa para resolver esta inestabilidad fue la exclusión de las variables CAT_OCUP (Categoría ocupacional) y CAT_INAC (Categoría de inactividad). En su lugar, la dimensión del mercado laboral se capturó mediante las variables ESTADO (Condición de actividad), horastrab (Horas trabajadas) y V14 (Ingresos no laborales), que ofrecieron una alternativa robusta.

Además de esta exclusión, el modelo requirió el agrupamiento de categorías en varios predictores claves que también generaban separación debido a la escasez de observaciones en algunos de sus niveles. Específicamente, las categorías extremas de IX_TOT (Miembros del hogar) e IV2 (Habitaciones) fueron consolidadas para crear grupos más grandes. Del mismo modo, fue necesario agrupar los niveles con pocas observaciones en CH03 (Parentesco) y CH15 (Completitud educativa). Este preprocesamiento fue un paso esencial para asegurar la estabilidad numérica y permitir una estimación robusta de los coeficientes.

El modelo final se complementó con determinantes fundamentales que no presentaban problemas de estimación. Se mantuvo EDUC (Años de escolaridad) como medida principal de capital humano, junto con los controles demográficos de CH06 (Edad), su término cuadrático (EDAD_2) y CH04 (Sexo). Asimismo, se incluyeron indicadores clave de NBI y hábitat como IV1 (Tipo de vivienda), IV6 (Acceso al agua), IV8 (Desagüe), II8 (Electricidad) y V2 (Servicio doméstico). Finalmente, se retuvo la variable AGLOMERADO para controlar por las heterogeneidades regionales pertinentes.

Al analizar las tablas de diferencia (Tablas 1 ,2,3,4) de medias entre los conjuntos de entrenamiento y testeo, se concluye que la partición del dataset es, en general, adecuada y balanceada.

Se observa que las diferencias en las medias para la mayoría de las variables son de pequeña magnitud. Esta observación fue corroborada al realizar las pruebas estadísticas pertinentes (test t de diferencia de medias), las cuales no arrojaron resultados estadísticamente significativos en casi todos los casos, sugiriendo que ambos subconjuntos son representativos de la misma población.

La única excepción relevante detectada es la variable EDUC en la base de datos correspondiente al año 2005, para la cual sí se encontró una diferencia de medias estadísticamente significativa entre los conjuntos de entrenamiento y testeo.

B

3 (tabla 5)

Al analizar los determinantes de la pobreza, los resultados del modelo logístico final (Tabla 1) son robustos y consistentes con la teoría económica. El factor más determinante es, con diferencia, la estructura demográfica del hogar (IX_TOT). El impacto es escalonado y severo: en comparación con un hogar unipersonal, uno de tres miembros (IX_TOT_3.0) ya multiplica las chances (odds) de ser pobre por 3.83. Este efecto se dispara exponencialmente, llegando a un Odds Ratio de 15.32 para hogares de siete o más miembros, lo que evidencia cómo la tasa de dependencia y la dilución del ingreso son factores críticos de vulnerabilidad.

Inmediatamente después, emergen los indicadores de pobreza estructural. La calidad de la vivienda (IV1_2.0) se revela como un predictor clave; habitar en una vivienda de menor calidad (Tipo B) multiplica las chances de ser pobre por 3.68 en comparación con una vivienda "Tipo A". De forma similar, el acceso a servicios básicos (IV6_2.0) también muestra un fuerte impacto, con un Odds Ratio de 3.15, subrayando que las condiciones del hábitat son un componente fundamental de la pobreza.

Finalmente, entre los factores protectores, el capital humano (EDUC) es el más significativo. Cada año adicional de escolaridad reduce las chances de ser pobre en un 36.1% (OR = 0.639), confirmando su rol como principal herramienta de movilidad social. Le sigue la intensidad laboral (horas trabajadas): por cada hora adicional trabajada por semana, las chances de caer en la pobreza se reducen en un 32% (OR = 0.680), destacando la importancia no solo de estar ocupado, sino de evitar la subocupación. En conjunto, estas variables delinean un escenario donde la pobreza en Argentina está fuertemente ligada a la demografía del hogar, las condiciones de la vivienda y el nivel educativo y de inserción laboral de sus miembros.

Punto 4:

Para ilustrar la probabilidad predicha por el modelo, se seleccionó la variable Edad (utilizando los datos de x_train_2025). Esta elección se justifica por dos motivos principales: primero, la edad representa el ciclo de vida del individuo, un factor sociodemográfico fundamental en el análisis de la pobreza; segundo, en el modelo general fue una de las variables más significativas.

El gráfico ilustra la relación entre la Edad del individuo (eje X) y la probabilidad predicha de ser pobre (eje Y), según el modelo logístico. La curva azul representa esta probabilidad, mientras que los puntos verdes (Y obs.) muestran los valores observados reales (0 o 1) y los puntos rojos (Y predicho) la clasificación binaria del modelo. Se observa una relación no lineal y decreciente a lo largo de casi todo el ciclo de vida: la probabilidad de ser pobre es máxima en las edades más tempranas, comenzando por encima del 60% (0.6), y desciende de manera sostenida a medida que la edad aumenta, estabilizándose cerca del 40% (0.4) en la vejez.

C

Punto 5

En KNN, un K pequeño genera fronteras de decisión muy irregulares, lo que implica bajo sesgo y alta varianza, con riesgo de sobreajuste. Al aumentar K, el modelo promedia más vecinos, suaviza la frontera y reduce la varianza, pero a costa de elevar el sesgo (tendencia al subajuste). El K óptimo equilibra esta compensación (trade-off), como sugiere el máximo de precisión observado en el Gráfico 2, localizado alrededor de K≈7–9.

Punto 6 Grafico 3

Usamos la educación y la edad en un modelo KNN de pobreza porque consideramos que estas variables son demográficas y socioeconómicas fundamentales que definen el espacio de las características de la vulnerabilidad. Sabemos que la educación es un fuerte predictor del nivel de ingresos y oportunidades al determinar el acceso a empleos de calidad, mientras que la edad se relaciona intrínsecamente con la etapa del ciclo de vida y la acumulación de capital humano. Al utilizar KNN, buscamos agrupar a los individuos con niveles educativos y edades similares, ya que asumimos que la proximidad en estas dimensiones implica una vulnerabilidad económica comparable.

Punto 7

En el Gráfico 3 se presenta el accuracy promedio del modelo KNN, obtenido mediante validación cruzada (5-fold CV), en función del número de vecinos (K) de 1 a 10. Al analizar la curva, observamos que el rendimiento fluctúa, alcanzando un mínimo en K=2 y mostrando una tendencia general a mejorar a medida que K aumenta. Basándonos en esta visualización, el número óptimo de vecinos para este modelo es K=10, ya que este valor registra el accuracy promedio más alto de la serie, situándose por encima de 0.68.

D

Punto 8

El Gráfico 5 ilustra cómo las penalidades L1 (LASSO) y L2 (Ridge) afectan a los coeficientes del modelo a medida que varía la fuerza de la regularización, representada por Lambda (λ). En ambos gráficos, un Lambda bajo (izquierda) se asemeja a un modelo sin regularizar, mientras que un Lambda alto (derecha) impone una penalización fuerte.

La regularización LASSO (L1) se caracteriza por su capacidad de realizar selección de variables. Como se observa en el gráfico de la izquierda, a medida que Lambda aumenta, los coeficientes no solo se reducen, sino que muchos de ellos son forzados a ser exactamente cero. Esto resulta en un modelo más simple que utiliza solo un

subconjunto de los predictores originales, eliminando los que considera menos relevantes.

En contraste, la regularización Ridge (L2), mostrada en el gráfico de la derecha, también reduce la magnitud de los coeficientes a medida que Lambda aumenta, pero lo hace de manera diferente. La penalidad L2 encoge los coeficientes acercándolos a cero, pero sin forzarlos a ser exactamente cero (a menos que ya lo fueran). Por lo tanto, Ridge mantiene todas las variables en el modelo, aunque reduce la influencia de aquellas con coeficientes grandes, siendo especialmente efectivo para mitigar problemas de multicolineidad y reducir la varianza general del modelo.

Punto 9

El análisis de regularización mediante validación cruzada, observado en el Gráfico 6, permitió identificar los hiperparámetros óptimos para los modelos LASSO y Ridge. Para el modelo LASSO, se determinó un valor de lambda (λ) óptimo de 0.1000, el cual minimiza el error de predicción medio, situándolo aproximadamente en 0.28. En comparación, el modelo Ridge alcanzó su rendimiento óptimo con un lambda de 0.0100, registrando un error de predicción medio levemente superior, cercano a 0.285. Ambos gráficos de error muestran la curva de error medio y su variabilidad (± 1 desviación estándar), confirmando la robustez de los lambdas seleccionados en el punto de menor error.

Un aspecto fundamental del análisis LASSO es su capacidad para la selección de variables, ilustrada en el gráfico inferior. Este muestra la proporción de variables cuyos coeficientes son forzados a cero a medida que aumenta la penalización lambda. Es notable que en el valor de λ óptimo (0.1000), la proporción de variables ignoradas es prácticamente nula. Esto indica que, para alcanzar el máximo poder predictivo, el modelo LASSO requiere retener casi todas las variables disponibles, y que una mayor simplificación del modelo (un lambda más alto) resultaría en un incremento del error de predicción.

Punto 10

La Tabla 6 compara los coeficientes (la importancia y dirección de cada variable) para un modelo logístico estimado bajo tres condiciones: sin penalización (regresión logística estándar) y con dos tipos de regularización, L1 (LASSO) y L2 (Ridge). El objetivo de estas penalizaciones es prevenir el sobreajuste y manejar la multicolineidad, y la tabla muestra cómo cada una "castiga" o modifica los coeficientes del modelo original.

Al comparar las columnas, se observa claramente el efecto de cada penalización. El modelo "Sin Penalidad" sirve como línea base, mostrando los coeficientes en su magnitud máxima (ej. IX_TOT_7,0 con 2.74). Ambas regularizaciones, L1 y L2, "encogen" (shrink) sistemáticamente estos coeficientes hacia cero. Se advierte que L1 (LASSO) es una penalización más agresiva, ya que reduce los coeficientes de forma más drástica que L2 (ver CH03_5,0, que pasa de -1.40 a -1.11 con L1, pero solo a -1.33 con L2). Esta capacidad de L1 para llevar coeficientes irrelevantes a cero (o muy cerca, como en AGLOMERADO) la hace útil para la selección de variables, mientras que L2

(Ridge) simplemente modera todos los pesos de manera más suave, reteniendo todas las variables.

E

Punto 11

La evaluación del rendimiento predictivo en la base de prueba de 2025 revela una clara superioridad de los modelos basados en Regresión Logística frente al modelo de K Vecinos Más Cercanos (KNN). El Logit estándar, Logit LASSO y Logit Ridge ofrecen un desempeño de discriminación casi idéntico (Tabla 7), evidenciado por un Área Bajo la Curva (AUC) consistente alrededor de 0.80 (Gráfico 8). Este valor indica una capacidad robusta y similar entre las tres variantes de Logit para diferenciar entre individuos "Pobres" y "No Pobres". Por otro lado, el modelo KNN (K=10) se queda notablemente rezagado, con un AUC de 0.7645, sugiriendo que es un clasificador significativamente menos potente para esta tarea.

La matriz de confusión del modelo Logit con un umbral de 0.5 muestra que el modelo clasificó correctamente a 4254 individuos (2337 No Pobres y 1917 Pobres). Sin embargo, cometió un total de 1586 errores, siendo el Falso Positivo (800 casos de No Pobres clasificados como Pobres) ligeramente mayor que el Falso Negativo (786 casos de Pobres clasificados como No Pobres).

Las métricas de clasificación en un umbral fijo confirman la superioridad de los Logit y exponen la debilidad crítica del KNN. Los modelos Logit alcanzan una Precision de 0.706 y un Recall (Sensibilidad) de aproximadamente 0.71, lo que significa que aciertan en el 70.6% de sus predicciones de pobreza y logran identificar al 71% de la población realmente pobre. En contraste, el KNN, aunque mantiene una Precision aceptable (0.703), tiene un Recall muy bajo, de solo 0.57. Este bajo nivel de sensibilidad es un defecto importante en la predicción de pobreza, ya que implica que el modelo KNN está fallando en identificar al 43% de los casos positivos reales, lo que resultaría en una alta tasa de Falsos Negativos y dejaría a una parte significativa de la población vulnerable sin clasificar correctamente.

En conclusión, la consistencia en el AUC y en las métricas de Precision y Recall entre las tres regresiones logísticas sugiere que las penalizaciones LASSO y Ridge, aunque influyeron en la selección y encogimiento de coeficientes (como se vio en el análisis anterior), no mejoraron la capacidad predictiva final sobre el conjunto de prueba. Por lo tanto, cualquiera de los modelos Logit es preferible al KNN. Dependiendo de si la prioridad es la simplicidad y transparencia (Logit estándar) o la identificación de las variables más influyentes (Logit LASSO), estas variantes son las opciones más adecuadas para predecir la pobreza en 2025.

Punto 12

Cuando el Ministerio de Capital Humano tiene la tarea crítica de dirigir recursos escasos, como un programa de alimentos, a la población más vulnerable, la selección del modelo de clasificación debe priorizar la minimización del Error Tipo II (Falso

Negativo). Este error, donde una persona que es realmente pobre es clasificada erróamente como no pobre, representa el costo social más alto para la política pública, ya que implica dejar sin ayuda a quienes más la necesitan. Por esta razón, el modelo "mejor" para este objetivo no es el que tiene la mayor precisión general, sino aquel que maximiza la métrica de Recall (Sensibilidad), que mide la proporción de casos positivos reales que el modelo logra identificar.

Al examinar los resultados, los tres modelos basados en Regresión Logística superan dramáticamente al modelo KNN. Si bien las diferencias son mínimas entre ellos, el modelo Logit Ridge registra el Recall más alto con un valor de 0.71, ligeramente superior al 0.709 del Logit estándar y el Logit LASSO. Este margen, aunque pequeño, indica que el Logit Ridge es el más efectivo para capturar y, por lo tanto, incluir, al 71% de la población pobre en el programa.

En consecuencia, el modelo Logit Ridge es la opción recomendada para la asignación de recursos. Su superioridad en Recall garantiza que, de todos los modelos probados, es el que mejor cumple con la prioridad de inclusión, mitigando el riesgo de que los grupos vulnerables queden fuera del programa. Aunque esto podría conllevar una ligera penalización en términos de eficiencia (un Recall alto a veces se asocia con un mayor número de Falsos Positivos, o asignación a no pobres), el imperativo ético y social de un programa de alimentos exige priorizar la identificación de la necesidad sobre la minimización absoluta del gasto ineficiente.

Punto 13

El Gráfico 9 ilustra el resultado de aplicar el modelo Logit Ridge a la base de datos de individuos que no respondieron a la encuesta. La predicción indica que el 45.9% de este grupo podría ser clasificado como pobre. Si bien esta proporción puede parecer elevada, su interpretación requiere considerar dos factores cruciales: primero, el cálculo oficial de la pobreza mediante la EPH utiliza factores de ponderación individual que no están aplicados en esta clasificación, por lo que la extrapolación directa al total de la población debe hacerse con cautela. Segundo, el elevado porcentaje es coherente con el contexto geográfico de la muestra (región del NOA), la cual históricamente presenta características socioeconómicas más vulnerables en comparación con los principales centros urbanos de Argentina

Apéndice:

Tabla 1 2005

Variable	Media Train	Media Test	Diferencia (Train - Test)
AGLOMERADO	23,2816	23,2891	-0,0076
CH03	2,7608	2,7816	-0,0208
CH04	0,488	0,4835	0,0045
CH06	29,6372	29,6249	0,0123
CH07	3,7478	3,7777	-0,0299
CH08	0,5441	0,5482	-0,0042
CH09	1,1396	1,1332	0,0064
CH15	1,3804	1,3733	0,0071
es_estudiante	0,2911	0,2893	0,0018
ESTADO	2,4662	2,4235	0,0427
CAT_OCUP	1,0404	1,0821	-0,0416
CAT_INAC	2,0832	2,024	0,0591
IV1	1,1042	1,1136	-0,0094
IV2	3,4693	3,4987	-0,0294
IV6	1,1984	1,1775	0,0209
IV8	1,0183	1,0191	-0,0008
II8	0,4865	0,4966	-0,0101
IX_TOT	4,8665	4,8534	0,0131
V2	1,7131	1,7082	0,0048
V14	1,8668	1,8643	0,0025
EDUC	7,7779	8,0173	-0,2394
horastrab	14,1305	14,1585	-0,0281
EDAD_2	1324,6345	1322,6259	2,0086

Tabla 2 2025

Variable	Media Train	Media Test	Diferencia (Train - Test)
AGLOMERADO	22,8078	22,8122	-0,0044
CH03	2,6119	2,5991	0,0128
CH04	0,4764	0,4863	-0,0099
CH06	35,1175	34,7764	0,3411
CH07	3,7574	3,7435	0,0139
CH08	0,6543	0,6387	0,0156
CH09	1,0655	1,0709	-0,0053
CH15	1,271	1,2812	-0,0102
es_estudiante	0,2734	0,2709	0,0025
ESTADO	2,2382	2,2622	-0,0239
CAT_OCUP	1,209	1,183	0,026
CAT_INAC	1,807	1,8668	-0,0598
IV1	1,1311	1,1392	-0,0081

IV2	3,4257	3,3932	0,0325
IV6	1,0308	1,0324	-0,0016
IV8	1,0036	1,0034	0,0002
II8	0,49	0,4914	-0,0014
IX_TOT	4,038	4,0685	-0,0305
V2	1,5931	1,595	-0,0019
V14	1,8353	1,8337	0,0016
EDUC	9,8401	9,693	0,1471
horastrab	15,8653	15,1942	0,6711
EDAD_2	1686,6323	1679,9373	6,695

Tabla 3 2025

Variable	Mean train	sd train	Mean test	sd test	t-test	p-value
AGLOMERADO	22,807754	3,6853737	22,812158	3,7348068	-0,049808	0,9602771
CH03	2,6119105	1,5434744	2,5991438	1,5183186	0,3478628	0,7279589
CH04	0,4764189	0,4995435	0,4863014	0,4998551	-0,827804	0,4078233
CH06	35,117506	21,297285	34,77637	21,693824	0,6666291	0,505041
CH07	3,7573941	1,5971459	3,7434932	1,6034169	0,3638334	0,7159986
CH08	0,6542766	0,4756986	0,6386986	0,4804188	1,3664644	0,1718576
CH09	1,0655476	0,2750815	1,0708904	0,2910586	-0,798689	0,4245089
CH15	1,2709832	0,6511066	1,2811644	0,6655376	-0,650071	0,515677
es_estudiante	0,2733813	0,445784	0,2708904	0,4444577	0,2340622	0,8149468
ESTADO	2,2382094	1,1261898	2,2621575	1,1272548	-0,889716	0,3736637
CAT_OCUP	1,2090328	1,3993881	1,1830479	1,392818	0,7782294	0,4364729
CAT_INAC	1,8069544	1,9518145	1,8667808	1,9801358	-1,277244	0,2015781
IV1	1,1310951	0,3480688	1,1392123	0,3612077	-0,964947	0,3346189
IV2	3,4256595	1,2853047	3,3931507	1,2928278	1,056687	0,2907081
IV6	1,0307754	0,1839527	1,032363	0,1854824	-0,360308	0,7186326
IV8	1,0035971	0,05988	1,0034247	0,0584253	0,1214178	0,9033654
II8	0,490008	0,5000001	0,4914384	0,4999695	-0,119729	0,904703
IX_TOT	4,0379696	1,6589925	4,0684932	1,6877607	-0,766018	0,4437034
V2	1,5931255	0,4913493	1,5950342	0,4909274	-0,162624	0,8708214
V14	1,8353317	0,370955	1,8337329	0,3723521	0,1801825	0,857017
EDUC	9,8401279	4,5376408	9,6929795	4,6186386	1,3499176	0,177106
horastrab	15,865308	21,345838	15,194178	20,761838	1,3266365	0,1846946
EDAD_2	1686,6323	1738,9291	1679,9373	1793,6266	0,1596097	0,8731952

Tabla 4 2005

Variable	Mean train	sd train	Mean test	sd test	t-test	p-value
AGLOMERADO	23,2815719	4,15805114	23,2891468	4,10543365	-0,07833147	0,93756759
CH03	2,76077833	1,52398233	2,78162798	1,55821348	-0,58208674	0,5605342
CH04	0,48798169	0,49995092	0,48349134	0,49976824	0,38477578	0,70042006
CH06	29,6371614	21,1292094	29,6248774	21,0965695	0,0249152	0,98012361
CH07	3,74780618	1,56905912	3,77770513	1,55071632	-0,81911171	0,41276249
CH08	0,54406715	0,49814934	0,54821837	0,49771024	-0,35705785	0,72106372
CH09	1,13964136	0,39601607	1,13321347	0,39583226	0,695382	0,48684856
CH15	1,38038916	0,73768567	1,37332462	0,7263605	0,41211072	0,68027635
es_estudiante	0,29111026	0,45436092	0,28931023	0,45347925	0,16980055	0,86517394
ESTADO	2,46623426	1,15815425	2,42350441	1,1517023	1,58305914	0,11347211
CAT_OCUP	1,04044258	1,37114666	1,08205296	1,39000904	-1,29457858	0,19552521
CAT_INAC	2,08317436	1,96785891	2,02402746	1,97910732	1,28528675	0,19875185
IV1	1,10415872	0,36264753	1,11359922	0,39118298	-1,08872487	0,27632457
IV2	3,46928653	1,38844434	3,49869238	1,41082769	-0,90283033	0,36665919
IV6	1,19839756	0,42839719	1,17750899	0,41730298	2,10486736	0,03535427
IV8	1,01831362	0,13410852	1,0191239	0,13697168	-0,25715292	0,7970712
II8	0,48645555	0,49991189	0,49656751	0,50002908	-0,86640007	0,38631274
IX_TOT	4,86646318	1,74295333	4,85338346	1,70275551	0,32367568	0,74619753
V2	1,71308661	0,4524071	1,70823799	0,45461052	0,45841759	0,64667247
V14	1,86684472	0,33980731	1,86433475	0,34246074	0,31566308	0,75227145
EDUC	7,77794735	4,86280341	8,01732592	4,8621367	-2,10874371	0,03501676
horastrab	14,1304845	22,6650972	14,1585485	22,6611469	-0,05304226	0,95770037
EDAD_2	1324,63449	1618,30026	1322,62586	1636,01748	0,05299314	0,95773949

Tabla 5 punto 3

Variable	Coeficiente	Error estándar	Odds Ratio
CH06	0,626149996	0,413524935	1,87039567
EDUC	-0,447724682	0,082313206	0,63908061
horastrab	-0,385272428	0,100084209	0,680265287
EDAD_2	-1,026051418	0,373315763	0,358419419
AGLOMERADO_19.0	-0,480263761	0,174987882	0,618620202
AGLOMERADO_22.0	-0,016053578	0,175403913	0,984074594
AGLOMERADO_23.0	-0,19680346	0,173026516	0,821352046
AGLOMERADO_25.0	0,084645485	0,189600173	1,088331169
AGLOMERADO_29.0	-0,089007208	0,178260437	0,914838979
CH03_2.0	-0,243838343	0,181547744	0,783614301
CH03_3.0	-1,210015243	0,235991743	0,298192734
CH03_4.0	-1,358038597	0,49293287	0,257164686
CH03_5.0	-1,545281706	0,338863613	0,213251788
CH03_6.0	-1,023355109	0,692799156	0,359387132
CH03_7.0	-1,403819946	0,30116327	0,245656774
CH03_10.0	0,505670971	0,895043234	1,658097682
CH04_1	0,243555484	0,103206593	1,275777102
CH07_2.0	-0,453724627	0,187161299	0,635257642
CH07_3.0	0,050450567	0,266367783	1,051744871
CH07_4.0	-0,35564342	0,33863748	0,700722439
CH07_5.0	0,145263993	0,202693574	1,156344797
CH08_1	-1,161905851	0,110902204	0,312889292
CH09_2.0	0,022841016	0,301507841	1,023103869
CH09_3.0	0,675362918	0,681025894	1,964745888
CH15_2.0	-0,12656008	0,199447983	0,88112121
CH15_3.0	-0,145324362	0,185455195	0,864741758
CH15_4.0	-0,420649122	0,543118167	0,656620455
es_estudiante_1	0,178695676	0,193503128	1,195656821

ESTADO_2.0	0,199281872	0,360953794	1,220525949
ESTADO_3.0	0,070224614	0,222044265	1,072749109
ESTADO_4.0	-0,67485879	0,347311523	0,509228324
IV1_2.0	-0,042826517	0,165190355	0,958077586
IV1_3.0	1,330513754	1,012229434	3,782986411
IV2_2.0	0,279118663	0,271178285	1,321964203
IV2_3.0	0,079651078	0,275466568	1,082909151
IV2_4.0	-0,255606232	0,287161795	0,774446861
IV2_5.0	0,363767405	0,317452283	1,438739531
IV2_6.0	-0,4325491	0,361932087	0,648852994
IV2_7.0	0,221611079	0,416209032	1,248085876
IV6_2.0	1,150375985	0,39365837	3,159380567
IV6_3.0	-0,999320082	1,201912708	0,368129654
IV8_2.0	-0,803803633	0,945326662	0,447623128
II8_1.0	-0,388554612	0,108046426	0,678036192
IX_TOT_2.0	0,647845375	0,285328528	1,911417999
IX_TOT_3.0	1,534585373	0,27848131	4,639401508
IX_TOT_4.0	1,891399131	0,282468062	6,628636529
IX_TOT_5.0	2,342303228	0,295667557	10,40517448
IX_TOT_6.0	2,520009329	0,310948765	12,42871261
IX_TOT_7.0	2,729248322	0,323675775	15,32136596
V2_2.0	0,113447501	0,126166784	1,120133082
V14_2.0	-0,494942145	0,132264374	0,609606175

Tabla 6

Variable	Sin Penalidad	L1 (LASSO)	L2 (Ridge)
IX_TOT_7,0	2,748988277	2,677952502	2,737234851
IX_TOT_6,0	2,536932553	2,469480579	2,527678161
IX_TOT_5,0	2,356128089	2,293252669	2,349059313
IX_TOT_4,0	1,907376179	1,843594061	1,899069104
IX_TOT_3,0	1,551278486	1,489137486	1,542425161
CH03_5,0	-1,545845589	-1,51115264	-1,536110309
CH03_7,0	-1,404558533	-1,37123752	-1,399613661
CH03_4,0	-1,31694003	-1,32497433	-1,345705336
CH03_3,0	-1,208138621	-1,18100926	-1,204852065

IV1_3,0	1,191958002	1,218373707	1,328631333
IV6_2,0	1,182683804	1,14120365	1,157503057
CH08_1	-1,160155556	-1,15835238	-1,15953422
CH03_6,0	-1,024440126	-0,96767851	-1,017195906
EDAD_2	-1,012029245	-1,0028347	-1,03569406
IV8_2,0	-0,956828676	-0,77504712	-0,795817191
IV6_3,0	-0,742739698	-0,87615814	-0,96899995
CH09_3,0	0,722887677	0,612801436	0,673818645
ESTADO_4,0	-0,673249712	-0,66436096	-0,662042092
IX_TOT_2,0	0,663302349	0,605897602	0,656462959
CH03_10,0	0,627569206	0,443352769	0,500617084
CH06	0,61625195	0,599879335	0,638881864
V14_2,0	-0,485098254	-0,48604503	-0,490186701
AGLOMERADO_19,0	-0,478313976	-0,46537907	-0,474594306
CH07_2,0	-0,450121913	-0,44569469	-0,450003058
EDUC	-0,445801428	-0,44501602	-0,447711751
CH15_4,0	-0,441193047	-0,38306577	-0,416226774
IV2_6,0	-0,400518436	-0,41940526	-0,411296898
IV2_5,0	0,394827181	0,368082134	0,38374819
II8_1,0	-0,386966068	-0,38711577	-0,387886608
horastrab	-0,379510555	-0,38064127	-0,381859869
CH07_4,0	-0,353060717	-0,34299165	-0,345741916
IV2_2,0	0,310200146	0,286517841	0,300099245
IV2_7,0	0,244106312	0,220461747	0,24261509
CH04_1	0,243031689	0,242101556	0,24542661
CH03_2,0	-0,235010318	-0,23281423	-0,237904502
IV2_4,0	-0,224603806	-0,24509474	-0,234093449
ESTADO_2,0	0,215411679	0,195098451	0,205383919
AGLOMERADO_23,0	-0,195498369	-0,1856409	-0,19229071
es_estudiante_1	0,182083635	0,164326861	0,176626438
CH07_5,0	0,15809577	0,135869811	0,154511272
CH15_3,0	-0,141494075	-0,13875776	-0,142314823
CH15_2,0	-0,123814257	-0,12051536	-

			0,123627611
V2_2,0	0,119009356	0,117310465	0,119505169
IV2_3,0	0,111683935	0,088545588	0,10058951
AGLOMERADO_29,0	-0,086687131	-0,07490015	-0,082628815
AGLOMERADO_25,0	0,085770432	0,093018863	0,09154717
ESTADO_3,0	0,080026187	0,07724327	0,080063214
CH07_3,0	0,064944754	0,050063725	0,060422001
IV1_2,0	-0,040374451	-0,04215324	-0,040958114
CH09_2,0	0,036764434	0,001503199	0,02331137
AGLOMERADO_22,0	-0,01423836	-0,00311495	-0,009147984

Tabla 7:

Modelo	Precision	Recall
Logit	0.706	0.709
Logit LASSO	0.706	0.709
Logit Ridge	0.706	0.71
KNN (K=10)	0.703	0.57

Gráfico 1

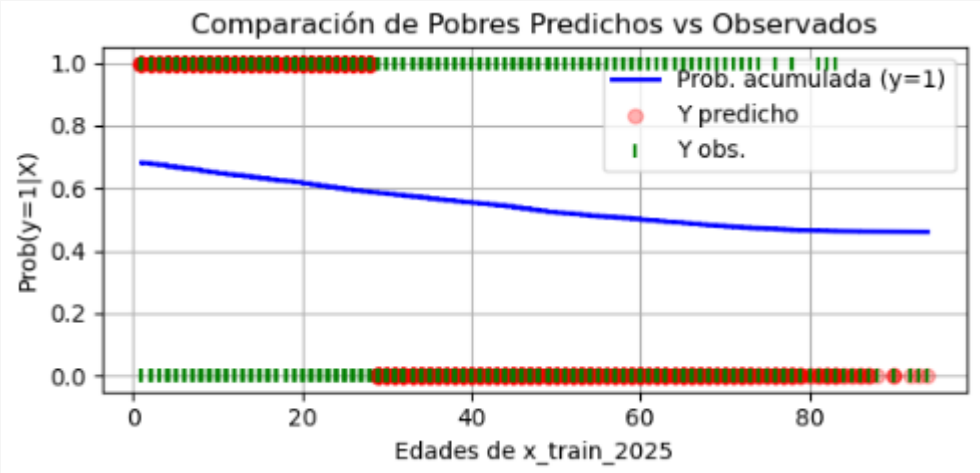

Gráfico 2:

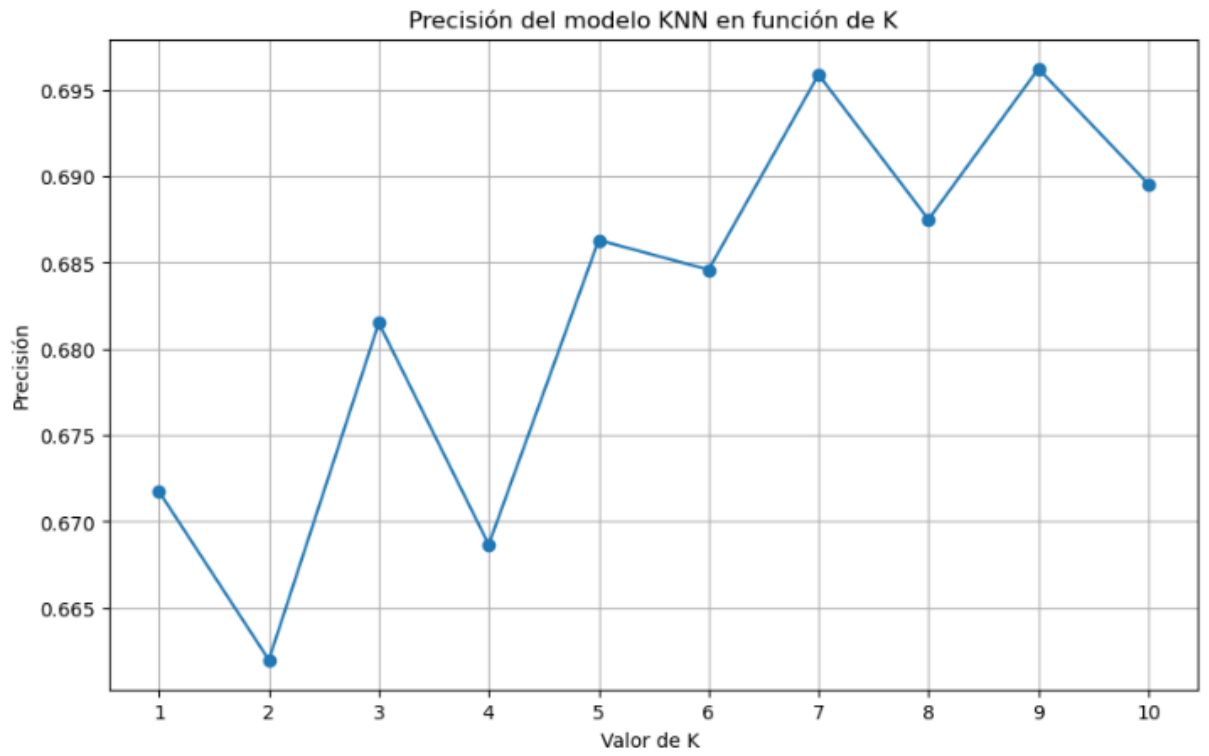

Gráfico 3:

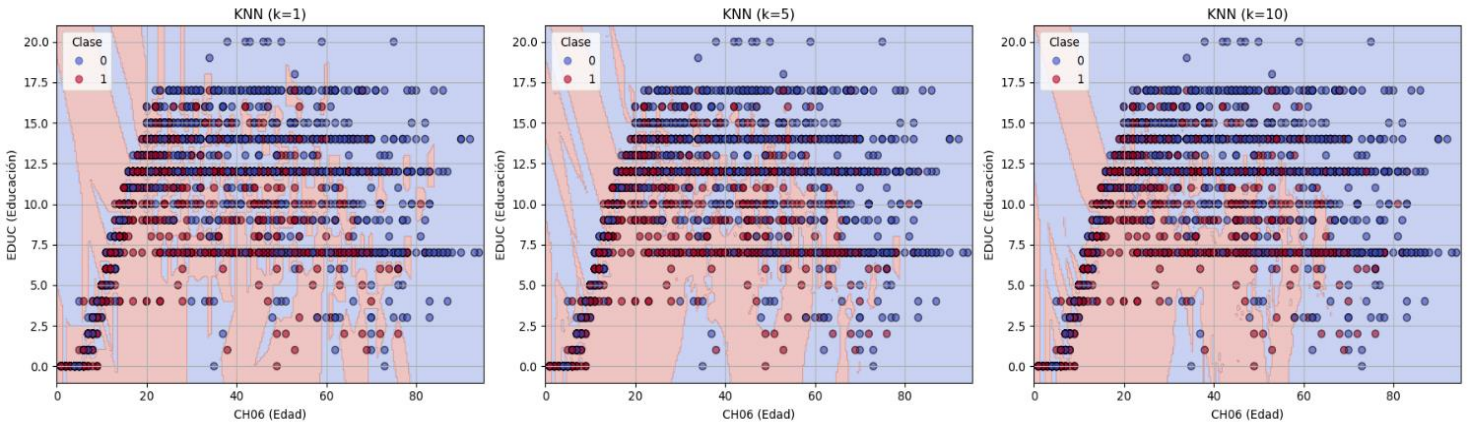

Gráfico 4:

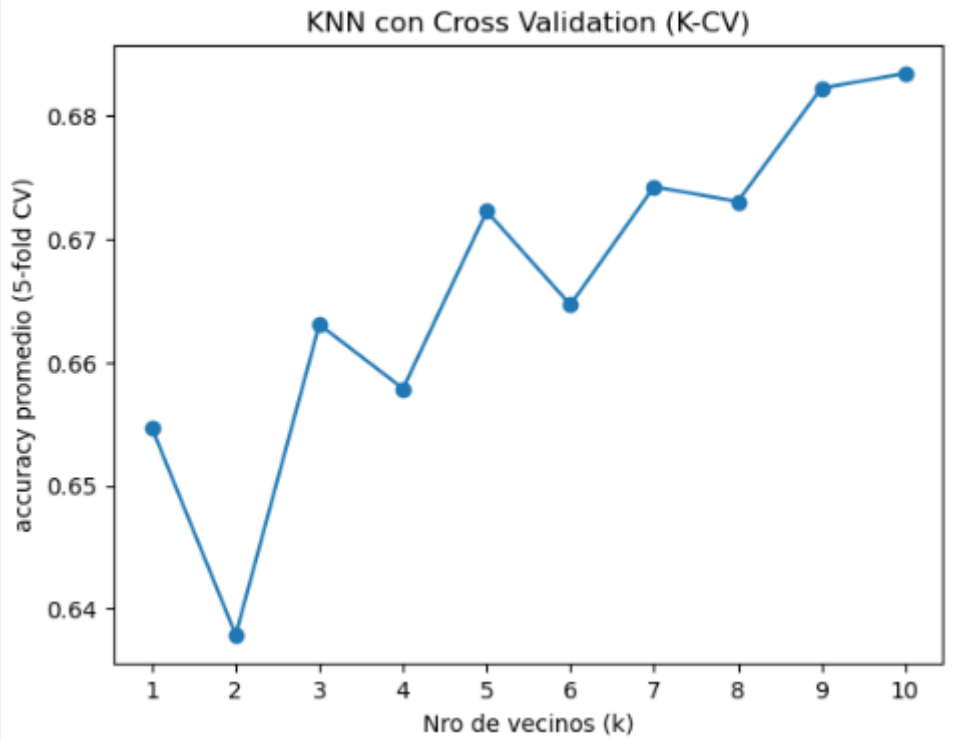

Gráfico 5.

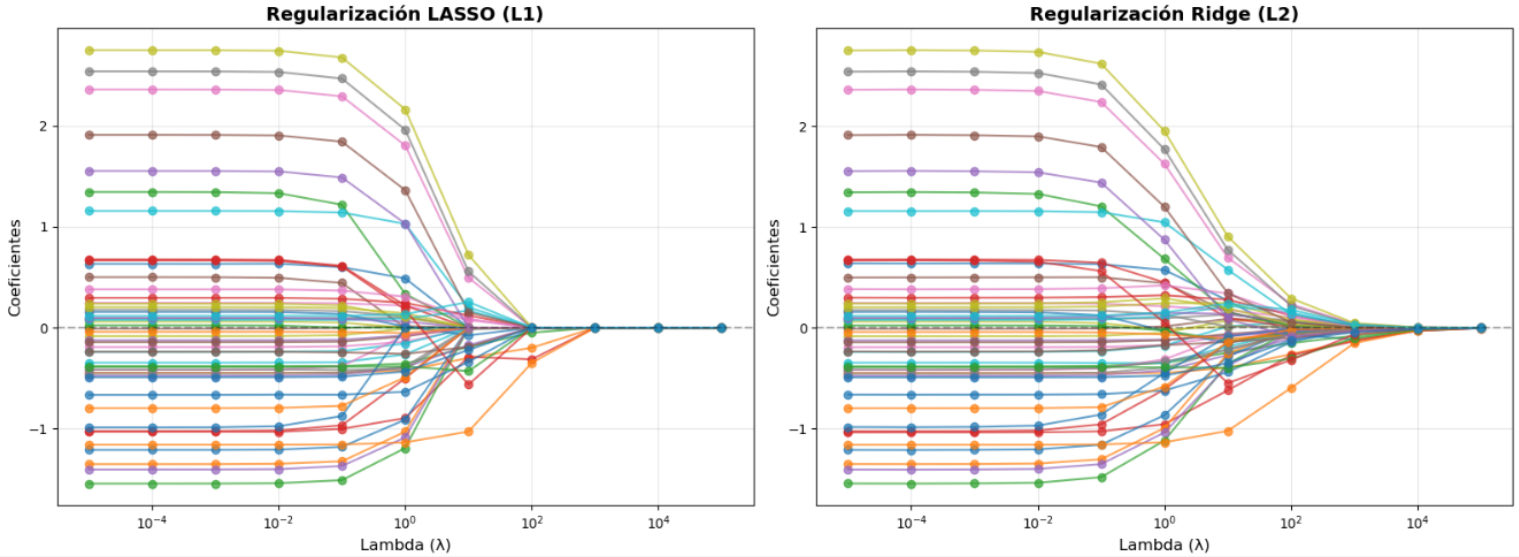

Gráfico 6

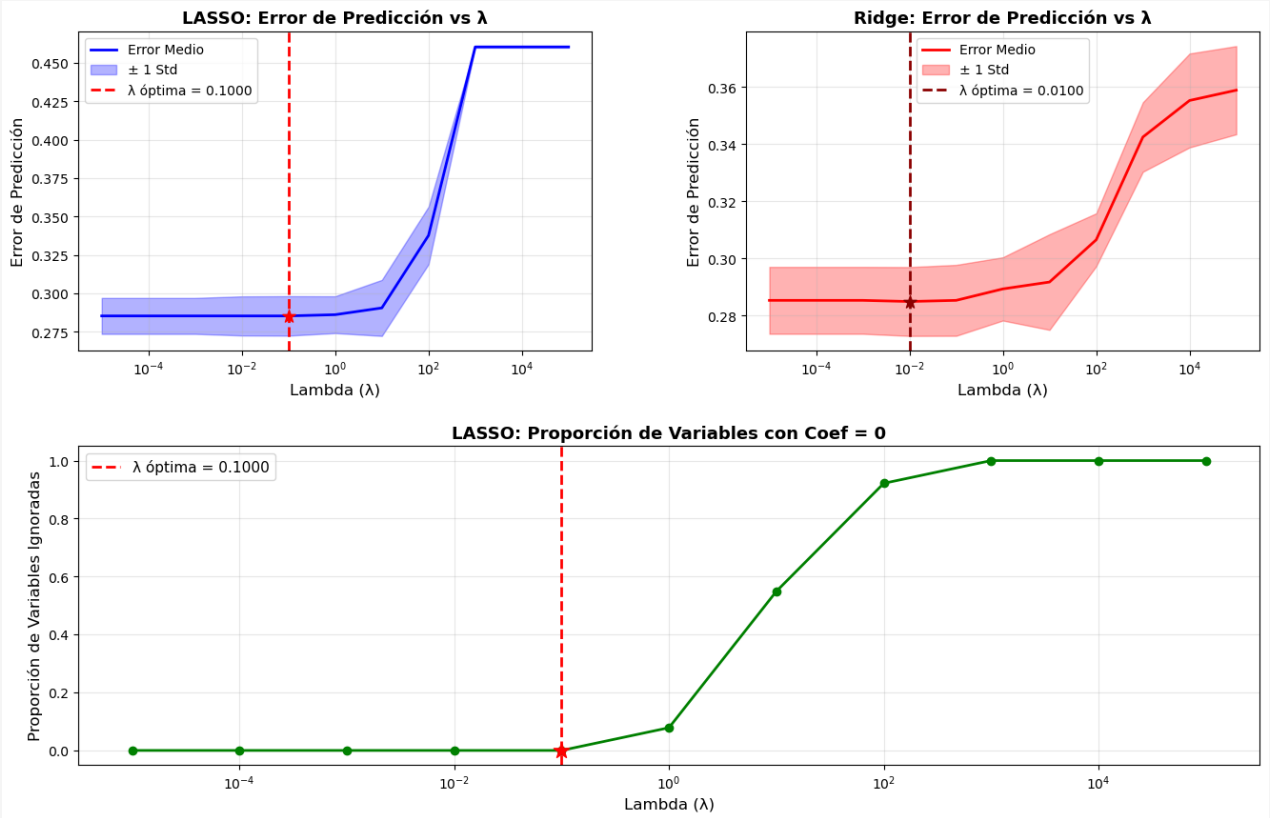

Gráfico 7

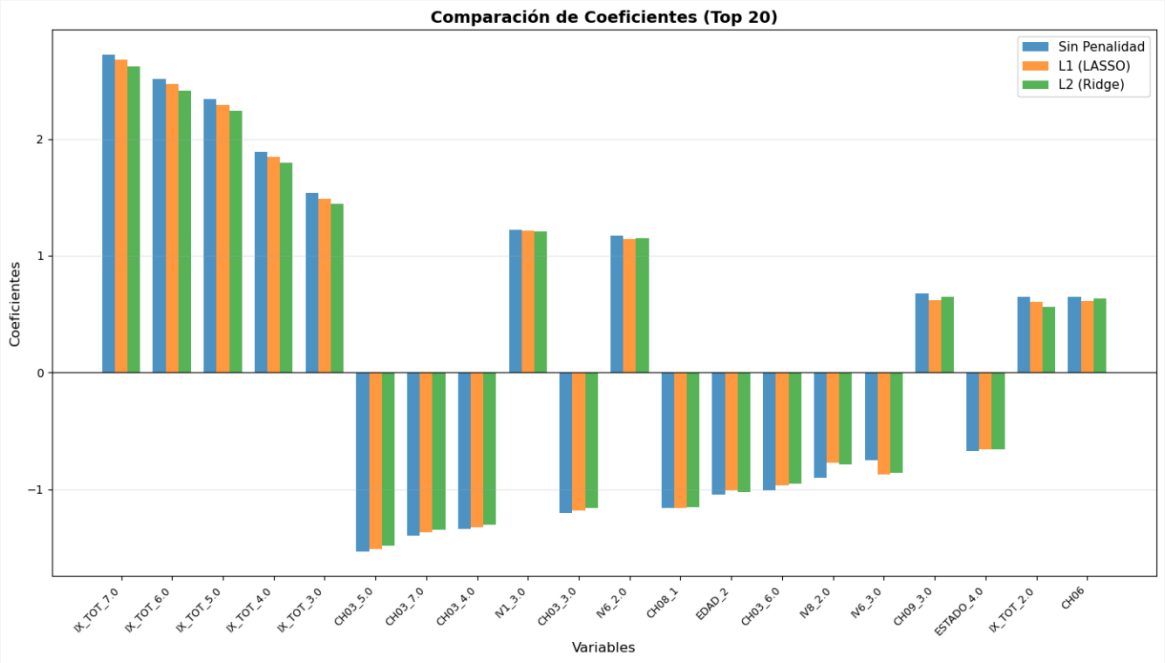

Grafico 8

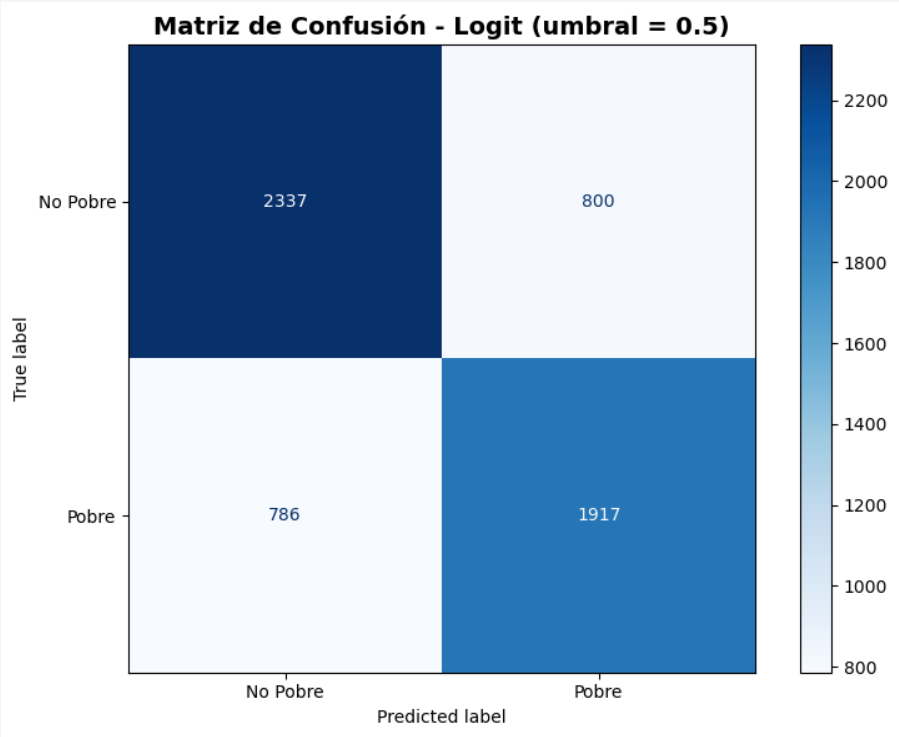

Gráfico 9

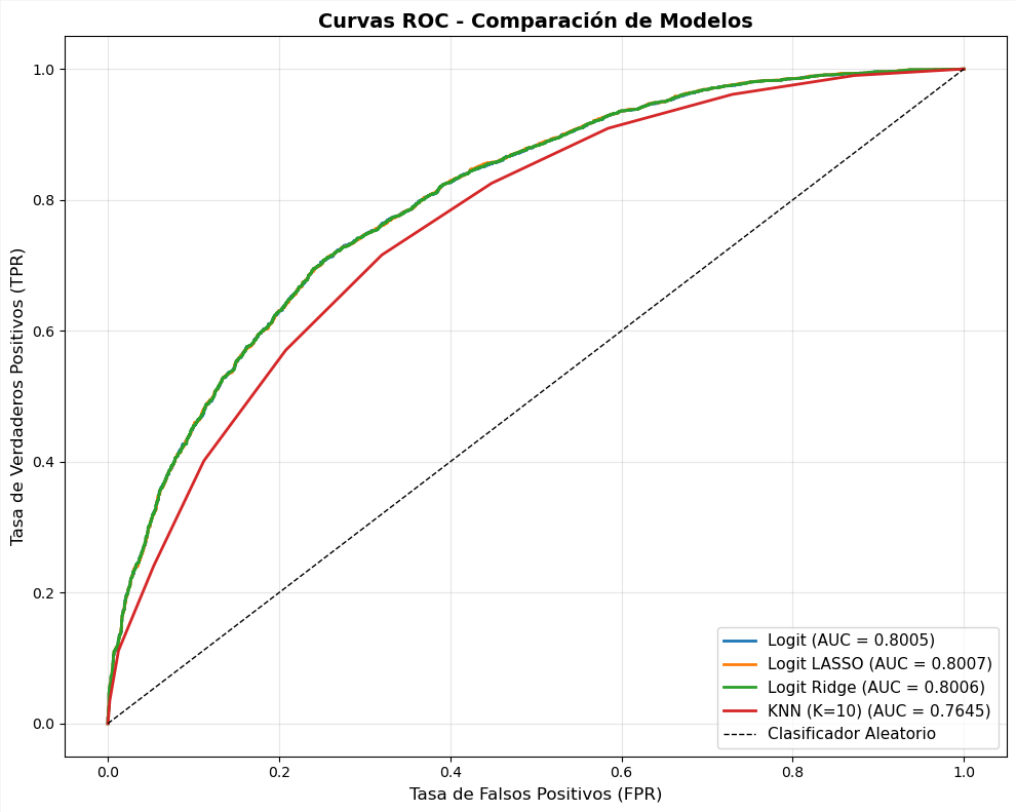

Grafico 10: