UNIVERSIDAD DE BUENOS AIRES - 2025

Big Data y Machine Learning

Trabajo Práctico N° 4

Clasificación de pobreza usando la EPH

Fecha de entrega: 11 de noviembre a las 13:00 hs.

Link al repositorio del grupo:
https://github.com/StefanoPistoia/Big-Data-Grupo-2/tree/main

Parte A - Enfoque de validación

Tabla 1. Diferencia entre medias para las variables seleccionadas

Año	Variable	Media_Train	Media_Test	P Value
2005	**horastrab**	13,31	13,68	0,60
2005	**educ**	8,39	8,15	0,10
2005	**edad**	29,07	28,63	0,47
2005	**edad2**	1269,65	1238,69	0,50
2005	**cobertura_medica**	2,64	2,73	0,09
2005	**sexo**	1,52	1,51	0,73
2005	**estado_laboral**	2,47	2,45	0,63
2005	**estado_civil**	3,74	3,75	0,92
2025	**horastrab**	36,74	35,93	0,46
2025	**educ**	9,76	9,86	0,58
2025	**edad**	35,85	34,92	0,25
2025	**edad2**	1755,19	1677,01	0,24
2025	**cobertura_medica**	2,22	2,26	0,78
2025	**sexo**	1,53	1,53	0,98
2025	**estado_laboral**	2,34	2,32	0,68
2025	**estado_civil**	3,54	3,60	0,33

En esta tabla incluimos variables que creemos relevantes para predecir el ingreso. La primera de ellas es *horas trabajadas*[1] que intuimos tiene una correlación positiva con el ingreso total. Por una lógica similar incluimos el estado *laboral*. Después tenemos *edad*[2], que puede ser un proxy para experiencia laboral y nos indica una correlación positiva pero no lineal por eso la incluimos al cuadrado también. Es lógico que una c*obertura médica* sólo pueda ser costeada por gente con mayores ingresos así que añadimos esa variable. Con respecto al *sexo*, la literatura sobre brecha salarial por género es vasta asi que nos pareció una inclusión obligatoria. Y finalmente incluimos *estado civil* porque existe una diferencia, los hombres casados ganan más que los solteros y las mujeres menos[3].

En cuanto a los tests de medias no observamos diferencias significativas al 5% para ninguna variable por lo que concluimos que ambos grupos están balanceados.

[1] Work More, Make Much More? The Relationship between Lifetime Hours Worked and Lifetime Earnings (2024, Federal Reserve Bank of St. Louis)
[2] Gasparini, Marchionni & Sosa Escudero (2005), CEDLAS – "Distribución del ingreso en América Latina"
[3] An analysis of income differentials by marital status (2008, R. Madalozzo)

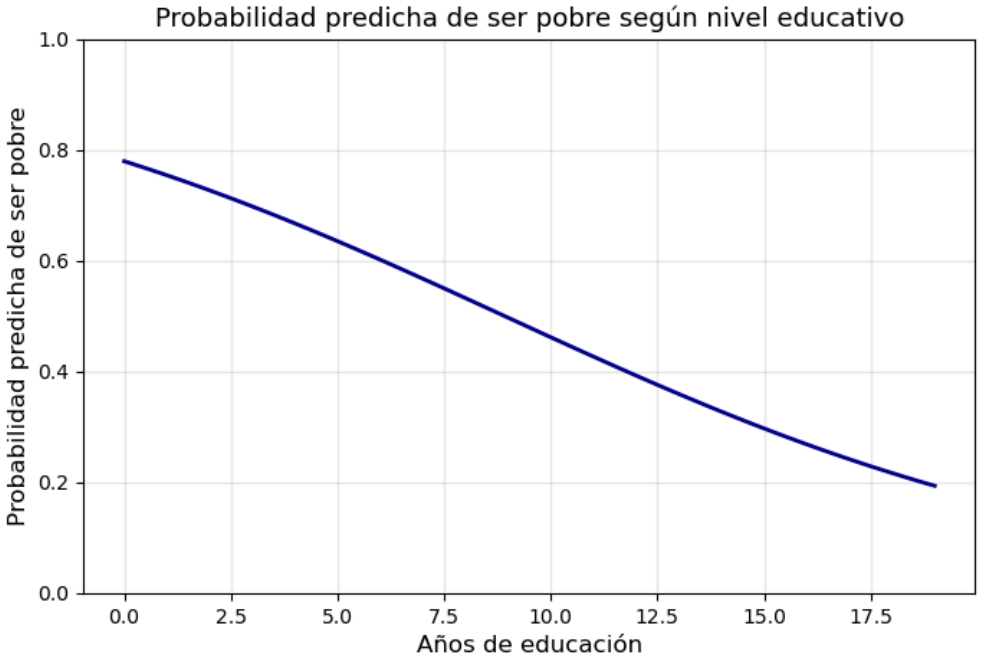
Parte B - Modelo de Regresión Logística

Tabla 2. Coeficientes de regresión logística

Variable	Coef. (β)	exp(β)	Error Std.	p-value
const	0.87	2.38	0.43	0.04
horastrab	-0.01	0.99	0.0	0.0
educ	-0.14	0.87	0.02	0.0
edad	0.06	1.06	0.02	0.0
edad2	-0.0	1.0	0.0	0.0
coberturamedica_mutual_prepaga_emergencia (base: obra social)	-0.47	0.63	0.69	0.5
cobertura_medica_plan_seguro_publico	1.41	4.09	0.21	0.0
cobertura_medica_ninguna	1.39	4.03	0.11	0.0
cobertura_medica_obra_social_y_mutual_prepaga_emergencia	0.59	1.8	1.21	0.63
sexo_femenino	-0.1	0.91	0.1	0.36
estado_laboral_desocupado (base: ocupado)	0.97	2.63	0.43	0.02
estado_laboral_inactivo	0.15	1.16	0.2	0.47
estado_laboral_menor_de_10	-0.56	0.57	0.34	0.1
estado_civil_casado (base: unido)	0.32	1.37	0.17	0.06
estado_civil_separado	0.33	1.39	0.26	0.21
estado_civil_viudo	0.28	1.33	0.31	0.37
estado_civil_soltero	-0.5	0.61	0.16	0.0

Observando los coeficientes signficativos al 5% o más, vemos que por cada hora trabajada la probabilidad de ser pobre disminuye en un 1%, cada año de educación adicional reduce en un 13% la probabilidad, la edad sin elevar al cuadrado resulta relevante, con un 6% más de probabilidad de ser pobre por cada año extra. Luego vemos que con respecto a tener obra social, tener un plan o seguro público aumenta por 4 tus probabilidades de ser pobre, mismo caso que no tener ninguna cobertura médica. Sorprendentemente no encontramos efectos por género. Estar desocupado multiplica las odds por 2.6 respecto a estar ocupado. Y finalmente el estado civil.

Gráfico 1.

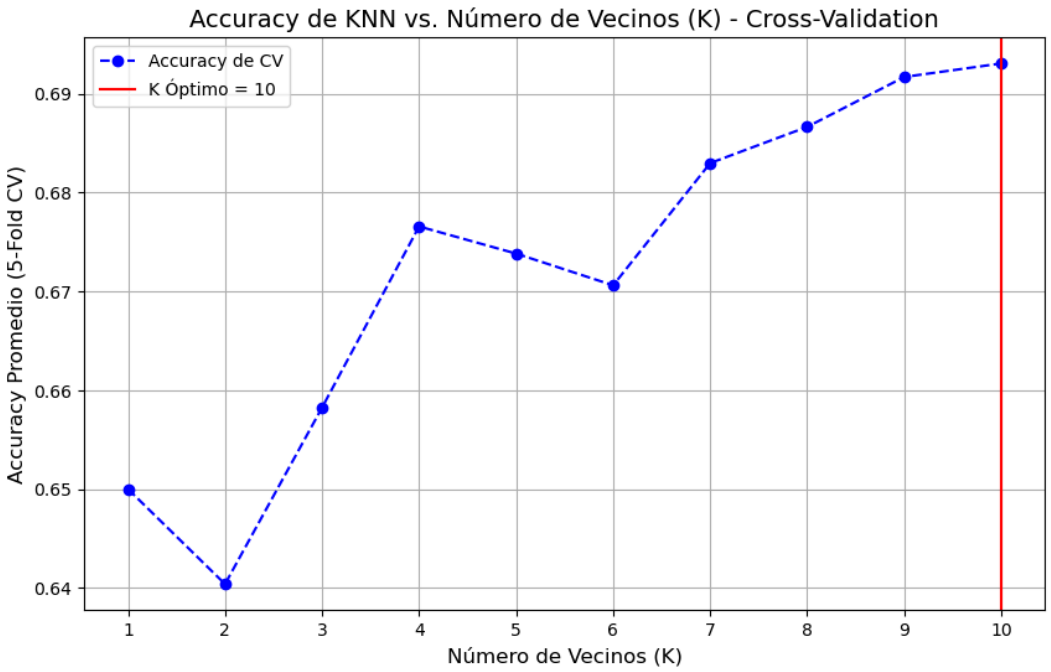


Lo que veíamos en la tabla de coeficientes lo confirmamos en el gráfico, es notable ver la diferencia entre una persona sin educación y una persona con educación superior, 60% puntos porcentuales en la probabilidad de ser pobre, realmente ayuda a tomar dimensión de la importancia de la educación.

Parte C - Método de Vecinos Cercanos (KNN)

La elección del K óptimo en este modelo ocurre en el punto donde se encuentra el balance entre el trade off de sesgo y varianza. En este caso vemos que para K=1, estamos ante un sobreajuste y el ruido de los datos puede llevar a una muy alta varianza. Para K=10 en cambio tenemos la mejor precisión entre las 3 opciones de parámetros, si se continuara incrementando el K es probable que lleve a un modelo muy simple y un aumento del sesgo que no minimice el error de clasificación

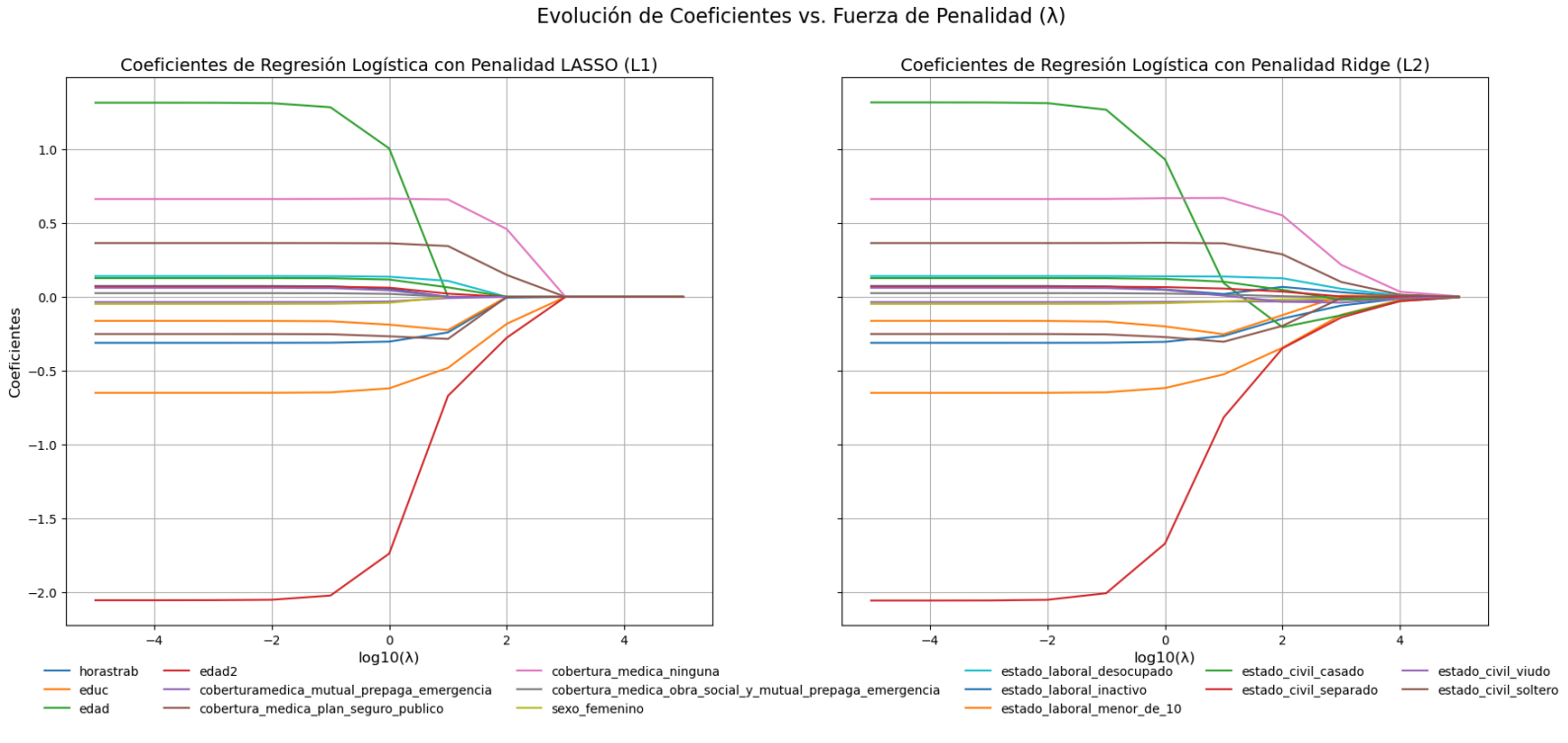
.Gráfico 2.



En el gráfico con Cross Validation confirmamos que efectivamente el número óptimo de vecinos es 10.

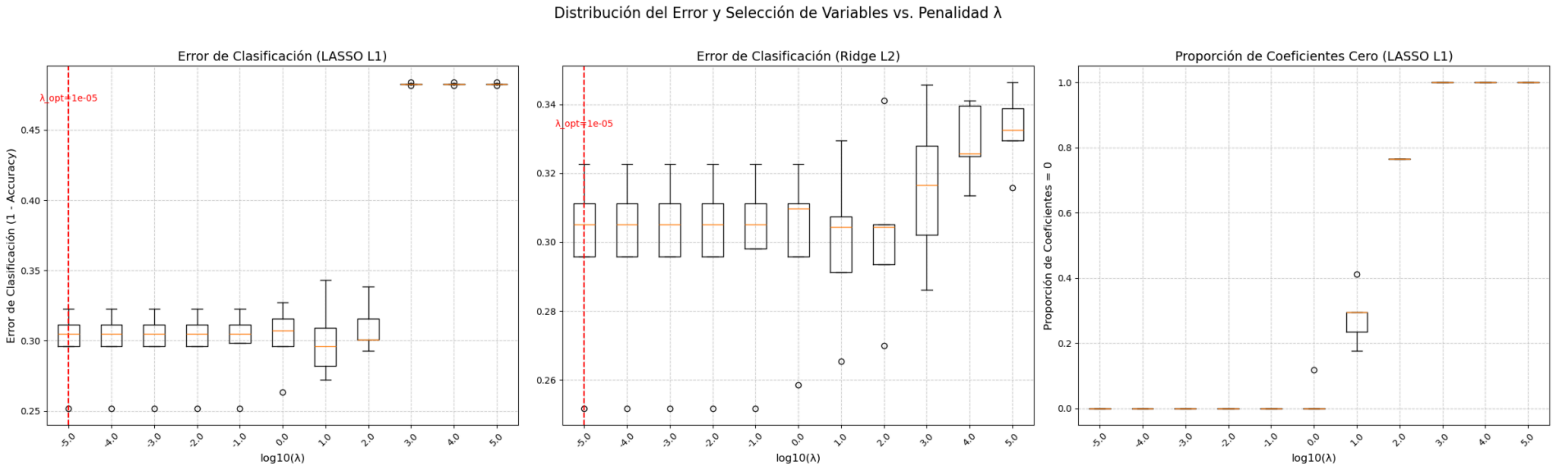
Parte D - Modelo de Regresión Logística con Regularización: Ridge y LASSO

Gráfico 3.



En el Gráfico 3 tenemos el eje x que representa el valor de la penalización del método de regularización, a medida que aumentamos nuestro parámetro se reduce el valor de los coeficientes, en el caso de LASSO algunos se reducen a 0 eliminando la variable del modelo, y en el caso de Ridge son llevados a valores muy pequeños. En LASSO se interpretan claramente las variables que resisten la penalización, años de educación, no tener cobertura médica,tener un seguro público y la edad al cuadrado.

Gráfico 4.



Como se observa en el Gráfico 4 la penalización óptima para LASSO y Ridge es 0.00001 en ambos casos, un valor extremadamente bajo. Corriendo los modelos con este λ obtenemos la siguiente tabla y vemos que los coeficientes son exactamente iguales excepto por variaciones infinitesimales, lo que nos está diciendo que el modelo que minimiza el error de clasificación es el logit sin penalización y no hay ruido fuerte o colinealidad excesiva.

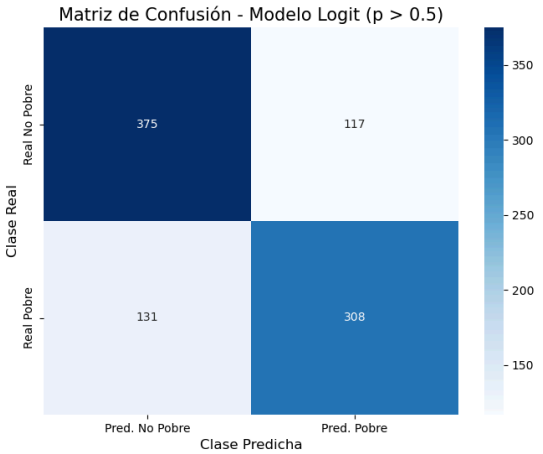
Tabla 3. Coeficientes de regresión logística con penalización

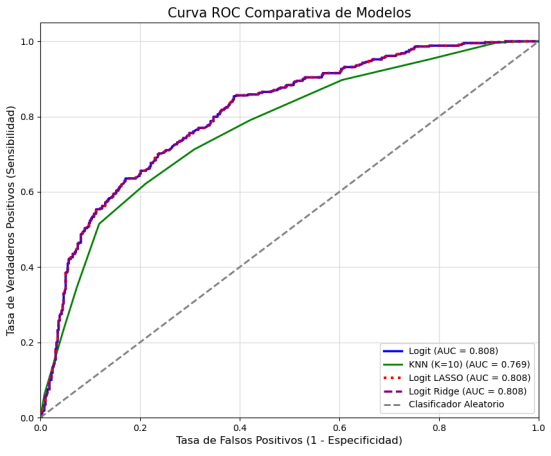
Variable	Sin penalizar	L1	L2
const	0.86740	0.86910	0.86780
horastrab	-0.01450	-0.01450	-0.01450
educ	-0.14160	-0.14160	-0.14160
edad	0.06190	0.06180	0.06190
edad2	-0.00110	-0.00110	-0.00110
coberturamedica_mutual_prepaga_emergencia (base: obra social)	-0.46920	-0.46910	-0.46910
cobertura_medica_plan_seguro_publico	140.910	140.920	140.910
cobertura_medica_ninguna	139.320	139.330	139.310
cobertura_medica_obra_social_y_mutual_prepaga_emergencia	0.58740	0.58700	0.58780

sexo_femenino	-0.09500	-0.09490	-0.09500
estado_laboral_desocupado (base: ocupado)	0.96870	0.96860	0.96900
estado_laboral_inactivo	0.14730	0.14700	0.14730
estado_laboral_menor_de_10	-0.56350	-0.56440	-0.56390
estado_civil_casado (base: unido)	0.31740	0.31740	0.31730
estado_civil_separado	0.32830	0.32820	0.32830
estado_civil_viudo	0.28310	0.28280	0.28290
estado_civil_soltero	-0.50230	-0.50260	-0.50240

Parte E - Desempeño de modelos afuera de la muestra, métricas y políticas públicas

La matriz de confusión resultante aplicar el modelo Logit a la base de testeo de 2025 y la curva ROC para los 4 modelos (Logit, KNN, Logit + LASSO, Logit + Ridge) son las siguientes:





Y elegimos las siguientes métricas de performance:

Modelo	Accuracy	F1-Score	AUC
Logit	0.7336	0.713	0.8083
KNN (K=10)	0.71	0.6691	0.7689
Logit LASSO	0.7336	0.713	0.8083
Logit Ridge	0.7336	0.713	0.8083

A excepción de KNN, que tiene una performance inferior, cualquiera de los otros tres modelos tiene el mismo desempeño. Nuestra interpretación adjudica este resultado a que los modelos *penalizadores* sobre Logit, que son LASSO y Ridge, al no penalizar por irrelevancia a ninguno de los predictores (de hecho el lambda óptimo obtenido antes es 0), resultan ser equivalentes a Logit.

Para un método de clasificación orientado a una recomendación de política de distribución de alimentos, consideramos que el problema de optimización debería ser la minimización de falsos negativos (error tipo 2), es decir, reducir al mínimo el predecir como 'no pobres' a aquellos que en realidad lo son. Nuestro argumento se basa en que el fenómeno que se está tratando, la indigencia, es prioritario combatirla en el marco de un Estado de Bienestar, y en que los costos sociales (o externalidades) de este error tipo 2 excederían ampliamente a los del tipo 1. Para este problema, nuestro orden de preferencia de los métodos es: Elastic Net, Ridge, LASSO (de *mejor* a *peor*).

En primer lugar, porque siendo la *especificidad* la prioridad (minimizar a los falsos negativos a costa de asumir falsos positivos), la performance de MSE_{test} en LASSO no es la mejor debido a que la multicolinealidad de ciertos predictores podría generar una inestabilidad no deseada en la predicción.

Segundo, nuestra inclinación a aceptar cierto sesgo para tener un método estable (con baja varianza) nos lleva a preferir un menor grado de complejidad (en cantidad y grado del espacio de predictores). En este sentido, preferimos Elastic Net por sobre Ridge ya que el primero conserva una virtud de LASSO que hace a la estabilidad: la selectividad de las variables explicativas. En última instancia, una manipulación criteriosa tanto del umbral de clasificación como de los ponderadores $l1$ y $l2$ para Elastic Net permitirán adecuar los resultados del método a los objetivos del programa y su restricción presupuestaria, que nunca puede desestimarse.

Adicionalmente, creemos que la elección de un método menos complejo tiene beneficios derivados en torno a la comunicabilidad, siendo que se trata de una cuestión de política pública, donde la rendición de cuentas es tan importante como la acción propiamente dicha que se está llevando a cabo y la intención que se tiene con ella.

Con el método elegido, se identificó como pobres a un 46% de las personas encuestadas que no reportaron su ingreso total familiar.