Universidad de los Andes

Maestría Economía Aplicada: *Big Data and Machine Learning for Applied Economics*

**Grupo:** Laura Natalia Capacho, Sebastián David Beltrán y Yurani Gonzalez[[1]](#endnote-1)

**GitHub URL:** <https://github.com/sbeltro/G10_PS2>

***Problem Set 2: Predicting Poverty[[2]](#footnote-1)***

**1. Introducción:**

En Colombia, el índice de pobreza monetaria fue de 39,3% en 2021 y la pobreza multidimensional de 16,0% (DANE, 2022). La ONU (2022) señala la erradicación de la pobreza como un imperativo ético, social, político y económico a nivel mundial, por esto, identificar correctamente los hogares que deben priorizarse es fundamental. El objetivo de este trabajo es construir un modelo predictivo de pobreza en hogares colombianos. Se caracterizó la pobreza desde dos enfoques: un problema de clasificación, donde encontramos que 25,7% de hogares en la base de prueba son pobres, y un modelo de predicción de ingreso comparado con la línea de pobreza, que indica que 32,1% de dichos hogares son pobres.

Destacamos que durante el estudio encontramos que las variables disponibles en la fuente de datos dispuesta para el ejercicio[[3]](#endnote-2), no permiten una aproximación sobresaliente de pobreza, pues las características del hogar que contiene son limitadas, y no se encuentran variables correspondientes a características de vivienda o condiciones de vida, comúnmente usadas para identificar pobreza (Kambuya, 2020). Para lograr una mejor caracterización de los hogares y enfocar de forma óptima las intervenciones es necesario ampliar las preguntas que contiene la encuesta para capturar otros factores, como lo hace la ELCA[[4]](#endnote-3).

**2. Datos**

En el estudio se utilizaron 4 bases[[5]](#endnote-4) de la GEIH 2018. En la tabla 1 se presentan las principales estadísticas descriptivas de las variables de interés con las que se entrenaron los modelos. Tenemos una muestra de 164,960 hogares, 33,024 clasificados como pobres y 131,936 no pobres. Se observa que en todas sus características existe una diferencia significativa entre los grupos: una mayor proporción de hogares pobres recibe subsidios (82,2% frente al 46% de no pobres), y ayudas monetarias de hogares nacionales e instituciones (29,6% y 31,6% respectivamente, frente al 20,6% y 11,0% en no pobres). Además, vemos que los hogares pobres duermen en promedio 2 personas por cuarto, el doble que en los no pobres.

En la tabla 2 se caracterizan los hogares que buscamos clasificar (*test).* La muestra se compone de 66,168 hogares, de los cuales, el 56,2% recibe subsidios, 13,2% subsidio familiar y tan solo 0,2% educativo. Asimismo, encontramos que el 23,2% recibió ayuda de hogares nacionales y 15,7% de instituciones, y vemos que en esta muestra duermen cerca de 2 personas en cada cuarto. (En la Gráfica 1 se observan las características de las dos muestras).

**3. Modelos y resultados**

**3.1 Modelos de clasificación**

La primera aproximación de un modelo predictivo de pobreza se abordó a través de un problema de clasificación, donde utilizando un clasificador bayesiano[[6]](#endnote-5) se asignan probabilidades a que un hogar sea pobre con base en sus características. En concreto, un hogar se clasificó: , si la variable de pobreza estimada es mayor a la regla “r”, entonces I es igual a 1, indicando que el hogar es pobre.

Estimación de pobreza: revisamos la base disponible y seleccionamos las características que permitirían identificar si el hogar es pobre, dada la escasa disponibilidad de información a nivel del hogar, se construyeron características agregadas a partir de variables individuales. Construimos un conjunto de 12 variables para hacer la estimación (Anexo 1). Posteriormente, hicimos una exploración con el método de *Best Subset Selection* para identificar aquellas que tendrían mayor relevancia y se evaluaron 6 modelos[[7]](#endnote-6). En las estimaciones se priorizaron la sensibilidad (buscando un alto nivel en el indicador) y el ratio de falsos negativos (buscando un bajo nivel), y se ajustó una regla como punto de corte.

Modelo seleccionado: luego de analizar los modelos y sus resultados en la matriz de confusión[[8]](#endnote-7) (Tabla 4), seleccionamos el modelo *logit* pues obtuvo los mejores resultados. Este modelo considera 4 variables explicativas: (i) *Subsidiado\_hg* que toma el valor de 1 si al menos un miembro de hogar (MdH) pertenece al régimen subsidiado de salud; (ii) *ayudaInstituciones\_hg* que indica si al menos un MdH recibió ayuda monetaria de una institución nacional o extranjera (=1), (iii) *personaxCuarto\_hg* que indica la ratio entre MdH y cuartos donde duermen; y (iv) *educ\_hg* que es el promedio de años de escolaridad de los MdH. Las variables fueron construidas en la base de personas de entrenamiento y se agruparon por hogar (*id) (*Tabla 3 contiene los resultados de la estimación)[[9]](#endnote-8).

El modelo seleccionado no tiene hiperparámetros alfa o lambda que busquen maximizar alguna métrica. No obstante, 3 de los 6 modelos estimados sí contaban con dichos hiperparámetros, que corresponde a estimaciones *Lasso*, tomó valores y buscó maximizar la sensibilidad y la curva *ROC*. De igual forma, como se trató de un modelo *logit* convencional, no se aplicó un método de *class imbalances*, sin embargo, 2 de los modelos considerados utilizaron los métodos de *Up-sampling* y *Down-sampling[[10]](#endnote-9)*.

Desempeño del modelo: El modelo *logit* arrojó una sensibilidad (*Sensitivity)[[11]](#endnote-10),* es decir, una proporción de hogares pobres clasificados correctamente de 57.8% y una tasa de 11.1% de falsos negativos[[12]](#endnote-11), que mide la proporción de verdaderos hogares pobres que el modelo clasificó erróneamente. Además, exhibe un tasa de falsos positivos[[13]](#endnote-12), que indica la proporción de “falsos pobres” sobre el total de hogares que el modelo clasifica como pobres, de 52.91%, 0.80 en el área bajo la curva (*AUC[[14]](#endnote-13)*, en inglés), y se observa que su curva ROC[[15]](#endnote-14) refleja el mejor desempeño entre los modelos analizados (Gráfica 2).

**3.2 Modelos de regresión de ingresos**

Una segunda aproximación para construir un modelo predictivo de pobreza de los hogares es a través de su ingreso. Estimamos el ingreso del hogar a partir de la agregación de una estimación individual del ingreso de cada MdH[[16]](#endnote-15). Luego, utilizando la línea de pobreza (Lp) definida por el DANE, se hace una comparación y clasifica el hogar, . Si el ingreso estimado es menor a la Lp se clasifica como pobre, y la variable .

Estimación del ingreso: Reconocemos que los ingresos, sobre todo en hogares pobres, pueden provenir de diversas fuentes, que además de salario se componen de otros ingresos[[17]](#endnote-16). Sin embargo, dentro de las variables disponibles no contamos con información suficiente para estimar apropiadamente dichos ingresos. Por lo anterior, el ingreso predicho corresponde al ingreso laboral del individuo que logramos aproximar de mejor manera.

Para la estimación se usaron 2 métodos de selección de modelos: *Best Subset Selection* (*BSuS)* y *Backward Stepwise Selection (BSwS)[[18]](#endnote-17)*, que permiten encontrar las variables de interés (Anexo 2) que mejor ajustan el ingreso. Además, se estimaron 6 modelos usando el método de regularización *Elastic Net[[19]](#endnote-18)*, que fuerza a que los coeficientes del modelo tiendan a cero, minimizado el riesgo de sobreajuste (*overfitting)*, reduciendo la varianza, mitigando el efecto de correlaciones entre predictores y mejorando la estimación.

Después de realizar las estimaciones, se calculó el error cuadrático medio (*MSE* en inglés) de cada modelo (Tabla 5). En la Gráfica 3 se observa que 5 de los modelos estimados tienen un nivel de ajuste similar (*BSwS y* 4 de *Elastic Net*), mientras el modelo de *BSuS*, y 2 del otro método reportan un ajuste más deficiente frente a los demás. En el modelo (3) es donde encontramos el mejor ajuste, en la medida que presenta el menor error cuadrático medio.

Modelo seleccionado: El modelo (3) seleccionado corresponde a la estimación utilizando *Elastic Net*, y validación cruzada en 5 conjuntos[[20]](#endnote-19), de un modelo donde la variable dependiente es , que denota el ingreso laboral del individuo y las variables explicativas son: , una variable categórica que indica si la persona trabaja en microempresa (=1), , que determina si la persona se encuentra ocupada (=1), , es el nivel educativo medido en años y , denota la ocupación del individuo.

En la estimación las variables fueron centradas y escaladas con media 0 y desviación estándar 1, para asegurar que la penalización que impone el modelo se aplique por igual sobre cada coeficiente. La estimación utilizó los hiperparámetros[[21]](#endnote-20) óptimos y , que minimizan el *MSE*, valores que se encontraron empleando validación cruzada en 5 conjuntos.

El modelo de predicción del ingreso individual se entrenó con la base *train*, y luego se agregaron los ingresos predichos de los MdH para obtener un estimado del ingreso agregado del hogar, que comparamos frente a la Lp[[22]](#endnote-21) y clasificamos los hogares en “pobre” y “no pobre”. Encontramos que, de un total de 164,960 hogares, 52,035 son clasificados como pobres, mientras 112,925 son clasificados como no pobres.

1. **Notas**

   Los códigos de los estudiantes del grupo son 202121025, 202121021 y 201212100, respectivamente. [↑](#endnote-ref-1)
2. \* Todas las tablas y gráficas mencionadas en el documento se encuentran en la sección de Anexos al final. [↑](#footnote-ref-1)
3. (Gran Encuesta Integrada de Hogares 2018 del DANE). [↑](#endnote-ref-2)
4. Encuesta Longitudinal Colombiana de la Universidad de Los Andes. [↑](#endnote-ref-3)
5. Las bases se encuentran divididas en dos muestras, una de entrenamiento (*train*) y otra de prueba (*test*), de cada una se contó con una base de datos a nivel de personas y otra a nivel hogar. Dado que la estimación que nos interesa es pobreza del hogar, se exponen las características en esta agregación. [↑](#endnote-ref-4)
6. El clasificador se entrenó en una base de entrenamiento donde conocemos la “clasificación correcta” de los hogares. [↑](#endnote-ref-5)
7. (1) *logit*; (2) *logit* con un control de *Two-Class Summary* y *5-fold Cross Validation (CV)*; (3) *logit* con un control de *5 Stats* y *5-fold CV*; (4) *logit* con control de *5 Stats* y *5-fold CV*, maximizando con *Lasso* la *ROC*; (5) *logit* con control de *5 Stats* y *5-fold CV*, maximizando con *Lasso* la sensibilidad y usando *Up-sampling*; y (6) *logit* con control de *5 Stats* y *5-fold CV*, maximizando con *Lasso* la sensibilidad y usando *Down-sampling.* [↑](#endnote-ref-6)
8. La matriz de confusión brinda información sobre la sensibilidad, ratio de falsos negativos y positivos y demás indicadores [↑](#endnote-ref-7)
9. Se observa que todas las variables resultaron significativas a un nivel del 1%, y además los coeficientes tienen los signos esperados. [↑](#endnote-ref-8)
10. *Up-sampling* es un método que simula puntos adicionales de la clase minoría (pobre) para balancear entre dos clases y *Down-sampling* reduce de manera aleatoria la clase mayoritaria (no pobre) para balancear las clases. [↑](#endnote-ref-9)
11. *Sensitivity* es la ratio entre el total de hogares clasificados como pobres y el número real de dichos hogares. [↑](#endnote-ref-10)
12. La tasa de falsos negativos es la ratio entre el número falsos negativos (hogares mal clasificados como no pobres) sobre el total de hogares clasificados como no pobres por el modelo. [↑](#endnote-ref-11)
13. La tasa de falsos positivos es la ratio entre el número falsos positivos (hogares mal clasificados como pobres) sobre el total de hogares clasificados como pobres por el modelo. [↑](#endnote-ref-12)
14. El *AUC* toma un valor entre 0 y 1, donde un valor cercano a 1 indica que el modelo está seleccionado a los verdaderos positivos y tiene un ratio bajo de falsos positivos. [↑](#endnote-ref-13)
15. La *Receiver Operating Characteristic Curve (ROC)* es una curva que mide la predicción del modelo frente al ratio de verdaderos positivos y falsos positivos. [↑](#endnote-ref-14)
16. , donde N es el número de miembros del hogar e es una función del ingreso individual. [↑](#endnote-ref-15)
17. Otros ingresos como subsidios, auxilios y ayudas del Estado, instituciones y otros hogares. [↑](#endnote-ref-16)
18. *Best Subset Selection* combina los predictores disponibles y *Backward Stepwise Selection* (introduce todas las variables en la ecuación y excluye secuencialmente una tras otra) [↑](#endnote-ref-17)
19. Este método es una combinación de las penalizaciones que imponen sobre los coeficientes lasso y ridge, dos métodos de regularización. [↑](#endnote-ref-18)
20. *5-fold Cross Validation* en inglés [↑](#endnote-ref-19)
21. En el modelo existen dos hiperparámetros: alfa (α)que controla el grado en que influye cada penalización y toma valores [0,1], así, si se aplica *Lasso* y si se aplica *Ridge;* y lambda (λ)que es el hiperparámetro de regularización. [↑](#endnote-ref-20)
22. Se multiplicó la línea de pobreza (Lp) por el número de personas en la unidad de gasto, puesto que la Lp está en términos per cápita, y el ingreso estimado es agregado del hogar. [↑](#endnote-ref-21)