Universidad de los Andes
Big Data and Machine Learning for Applied Economics
Taller 2

Integrantes:
David Santiago Caraballo Candela[1]
Sergio David Pinilla Padilla[2]
Juan Diego Valencia Romero[3]

30 de septiembre de 2022

GitHub: Repositorio Taller 2

1 Introducción

El problema de la pobreza en Colombia es casi inherente a la historia del país. A pesar de 130 años de crecimiento económico sostenido iniciado en 1890, de la expansión educativa, la urbanización y la "modernización" de los sectores productivos, en Colombia la pobreza y la violencia siguen siendo altas aunque se hayan logrado reducir en los años recientes, particularmente en la última década. Algunos autores se han encargado de estudiar las causas y los orígenes de este flagelo que aún agobia a cerca del 27% de la población total (DANE, 2019). Robinson (2016) sostiene que el alcance y la persistencia de la pobreza y la violencia en Colombia son una consecuencia de las facetas extractivas de las instituciones políticas y económicas nacidas desde la independencia e, incluso, la conquista. Por su parte, Herrera (2017) cree que el problema colapsa a uno de riesgo moral entre pobres y ricos así como de formales e informales, pues ninguno tiene incentivos económicos sostenibles a moverse al lado "rival". Incluso, Ocampo & Gómez-Arteaga (2017) consideran que la tardía expansión de los sistemas de protección social habría contribuido más a la reducción de la pobreza que el crecimiento del Producto Interno Bruto (PIB), lo que hoy en día sí hacen. La medición de la pobreza es difícil, larga y costosa. Al construir mejores modelos, es posible diseñar encuestas con menos preguntas y más especificidad que midan de forma rápida y económica la eficacia de nuevas políticas públicas, intervenciones y programas sociales.

Este trabajo constituye una aplicación práctica de los tópicos aprendidos en el curso de *Big Data and Machine Learning for Applied Economics* sobre clasificación y predicción. Los datos utilizados en este ejercicio provienen de la GEIH (Gran Encuesta Integrada de Hogares), una encuesta mediante la cual se solicita información sobre las condiciones de empleo de las personas además de las características generales de la población a nivel nacional, cabecera-resto, regional, departamental, y para cada una de las capitales de los departamentos. Esta encuesta, que utilizaba el marco geoestadístico del Censo Nacional de Población y Vivienda (CNPV) de 2005, a la cual se refiere como GEIH-M05, se actualizó para estar acorde con el marco y los resultados del CNPV de 2018. Tal actualización permitió tener en cuenta los cambios en la evolución demográfica, organización del territorio y distribución de la población que ha experimentado el país entre las dos operaciones censales. Dichos cambios aumentaron la visibilidad estadística de otros grupos marginales en la muestra (e.g. minorías, habitantes en vía terciaria, entre otros). Luego, esta información constituye un insumo robusto para analizar el comportamiento de los ingresos de los hogares colombianos a fin de estimarlos y asignar un estatus de pobreza, además, vía clasificación.

Nuestra aproximación creó una medida ponderada *W-MIR* de la tasa de falsos positivos con falsos negativos a razón 1:3 en penalización, respectivamente. Los resultados de regresión sugieren un modelo cuantílico (Q25) para predecir el ingreso de los hogares y comparar con la línea de pobreza ya que tiene una sensitividad del 97,77% y su medida ponderada entre el resto de especificaciones es la mínima con 16,52%. Por su parte, la estrategia de clasificación con mayor poder predictivo es *QDA* con una tasa de falsos negativos del 13,1% y *W-MIR* del 20,84%. Identificamos que los grandes predictores de la pobreza en los hogares están fuertemente correlacionados con el estatus socio-económico de(l)/la jefe(a) del hogar y su cónyuge, el grado educativo promedio alcanzando, su participación en los regímenes de salud y las características generales de la vivienda.

El documento se organiza de la siguiente manera: la sección 2 contiene una descripción detallada de los datos a utilizar así como su procesamiento; en la sección 3 se construyen dos aproximaciones empíricas; la primera, de estimación y, la segunda, de clasificación, ambas con el objetivo de determinar el estatus de pobreza de los hogares en la muestra. La sección 4 aborda los resultados encontrados en términos de las métricas de evaluación del modelo. Las consideraciones finales y recomendaciones de política son presentadas en la sección 5.

[1]Código: 201813007.
[2]Código: 201814755.
[3]Código: 201815561.

2 Datos

Los datos aquí empleados, provenientes de la GEIH, se organizaban en cuatro (4) bases distintas, cada una con un nivel observacional y/o con muestra distinta. Por un lado, los archivos nombrados `train_personas.Rds` y `test_personas.Rds` corresponden a muestras de observaciones de la GEIH al nivel del individuo, que fueron clasificadas *ex-ante* entre un conjunto de la muestra que debería ser utilizado para entrenar los modelos (542.941) y otro que debía ser utilizado para evaluarlos (219.812). Mientras que, por otro lado, los archivos `train_hogares.Rds` y `test_hogares.Rds` corresponden a muestras de observaciones de la GEIH a nivel hogar, que también están clasificadas entre muestra para *train* (164.960) y para *test* (66.168), respectivamente.

En este ejercicio, cada una de las bases de datos cumple una función distinta. Las bases de datos *ex-ante* identificadas como *test* no poseen información acerca del ingreso de los individuos ni el hogar y, por lo tanto, se busca predecir el nivel de ingreso (y la situación de pobreza) de estos hogares ubicados en el *test*. Para lograr esto, en principio se utilizaron las bases de datos *train*, que sí poseían información acerca del ingreso de los individuos y del hogar, como insumo para entrenar y evaluar la efectividad de los modelos de predicción esbozados en la sección 3.2, mediante el uso de variables socio-económicas comunes entre *test* y *train*. Posteriormente, el modelo entrenado y revisado con los datos *train* se implementó sobre las variables socio-económicas presentes en la base *test* para realizar la predicción del ingreso del hogar.

La Figura 1 (Anexos) muestra que la distribución del ingreso per cápita del hogar en la muestra *train* esta sesgado hacia la izquierda, con valores atípicos en la cola derecha. Además, la Tabla 1 (Anexos) concluye que el hogar colombiano promedio dispone de $\approx$ COP \$860.000 por persona para subsistir (i.e. comer, vestir, vivienda, entre otros), mientras que el individuo promedio devenga al mes $\approx$ COP 775.000. La edad promedio es de 33,5 años, el 47% de la muestra son mujeres, el 25% de los individuos son pobres, y la mayoría están concentrados entre los estratos 1 al 3 -cerca de un 68%-, y poseen algún tipo de educación (principalmente de primaria o bachillerato -$\approx$ 47%-).

Debido que, como objetivo central se busca desarrollar un modelo que efectivamente prediga el ingreso de los hogares de la base *test*, se seleccionó dentro de la muestra *train* a una sub-muestra que emule las características de la muestra objetivo. Para ello se utilizó la metodología de *Propensity Score Matching* (*PSM*) que permitió calcular la probabilidad de que una observación de *train* al nivel del individuo perteneciera a *test*, condicional a las variables observables entre ambas bases. Gracias a estos resultados (Figura 2, Anexos), se logró identificar dentro de *train* a una sub-muestra de individuos que probablemente podrían hacer parte de *test* ($PSM \geq 0,5$), y que alcanzaban a corresponder al 18,3% del mismo *train*. Esta información se guardó en la variable `dummy_psm`.

A partir de aquí, y para seguir explotando el gran volumen de los datos, se realizó un ejercicio de pre-procesamiento en el que se agrupó la información contenida de los individuos (60 variables) a nivel hogar de la siguiente manera: i) para las variables de edad y `dummy_psm` se calculó la media; ii) para las variables continuas se hizo una suma vertical y; iii) para las variables categóricas y *dummies* se seleccionó a la categoría del jefe del hogar como característica del hogar en su totalidad. Lo anterior permitió lograr dos objetivos específicos: i) incrementar la cantidad de información contenida a nivel hogar, sin necesidad de aumentar el espacio muestral y; ii) con la media de `dummy_psm`, se logró establecer un criterio de selección ($\mathbb{E}[$`dummy_psm` $> 0,5]$) para identificar a una sub-muestra de 33.631 hogares provenientes de *train* (20,4%) que probablemente podrían hacer parte de la muestra objetivo (*test*). Entonces, con una partición 80:20, las observaciones de esta sub-muestra de hogares constituirán una nueva muestra de evaluación (*test 2.0* -20,4%-), y el otro 79,6% restante de *train* (131.329) integrará una nueva muestra de entrenamiento (*train 2.0*).

Por último, debido al fenómeno de conformabilidad de matrices por el alto número de variables (>250) se decidió realizar una selección previa de *features* para reducir la variabilidad y los costos computacionales. En la siguiente sección se profundiza más sobre la manipulación de las variables.

3 Modelos & Resultados

La presente sección desarrolla dos metodologías para determinar el estatus de pobreza de los hogares. Una a través de la estimación de siete (7) modelos de regresión que predicen el ingreso del hogar y esto se compara con un umbral de línea de pobreza y otra vía clasificación con seis (6) técnicas. Las especificaciones utilizan un total de 26 predictores repartidos en tres categorías. Las variables continuas (j) fueron estandarizadas con media 0 ($\mu_j = 0$) y desviación estándar igual a 1 ($\sigma_j^2 = 1$). Además, se truncó el ingreso per cápita del hogar a su percentil 95 (COP \$7,2 millones). Igualmente, para resolver el problema de falta de información en ciertas variables explicativas (i.e. dicótomas), se realizó un ejercicio de imputación de valores con aquel más repetido (la moda).

Primero, se inicia concentrándose en las variables socio-económicas y demográficas tradicionales del hogar (i.e. edad promedio, ciudad de ubicación y si está en zona rural, sexo y máximo nivel educativo del jefe del hogar)[4]. Luego, en variables sobre la condición del hogar y su vivienda (i.e. número de miembros, si la vivienda es propia, valor de las cuotas de amortización, valor del uso mensual del predio,[5] y si hay hacinamiento)[6]. Tercero, variables del mercado laboral para cada hogar (i.e. si el jefe del hogar esta afiliado a seguridad social, si hace parte del régimen subsidiado, si esta trabajando o buscando trabajo, el tiempo que lleva trabajando en su empresa, si el mes pasado recibió primas en el trabajo, si ha recibido pagos en especie por su trabajo, si se desplaza en transporte de la empresa, si la empresa es pequeña,[7] si esta cotizando pensión, si el mes pasado recibió ingresos laborales, por arriendo o por dividendos, y, por ultimo, el numero total de horas trabajadas a la semana en el hogar). Estas 23 variables (algunas continuas, otras categóricas y otras dicótomas) corresponderán al espacio de regresores que se le aplicará inicialmente a los modelos de regresión y clasificación especificados. Sin embargo, existe la posibilidad de que ciertas especificaciones terminen seleccionando algunos de estos *features*, lo cual, potencialmente puede ayudar a reducir la varianza mediante una disminución del espacio de regresores.

3.1 Regresión

Primero se abordó el problema de predicción de la pobreza en los hogares de manera indirecta. Se especificaron de siete (7) modelos de regresión que buscaban predecir el ingreso de los hogares, y, al comparar el nivel predicho de ingreso con la linea de pobreza del hogar, se pudo clasificar a cada hogar del *test* como pobre o no. Para poder evaluar cuál de las especificaciones predice mejor por fuera de muestra, se va a utilizar una métrica de evaluación asimétrica personalizada (denominada únicamente para este documento como *Weighted Misidentification Ratio* o *W-MIR*), en la cual clasificar a un hogar pobre como no pobre -falso negativo- tendrá un costo tres veces superior a clasificar a un hogar no pobre como pobre -falso positivo-[8].

Las primeras tres especificaciones para predecir el ingreso corresponderán a modelos de *baseline* que poseen una función de pérdida cuadrática: Lasso, Ridge y *Elastic Net (EN)*. Luego de realizar un entrenamiento inicial de cada modelo mediante el uso de *10-Fold Cross-Validation* para encontrar los parámetros óptimos: Lasso fue optimizado con un $\lambda_L^* = 0,0139$ y no se presentó selección; Ridge fue optimizado con un $\lambda_R^* = 4,04$; y, por ultimo, *Elastic Net* fue optimizado con $\lambda_{EN}^* = 1$[9] y $\alpha_{EN}^* = 0,0139$, lo cual implica que se presentó una solución de esquina y la *Elastic Net* seleccionó a Lasso ($\lambda_L^* = 0,0139$) como su especificación óptima. En este orden de ideas, si *EN* selecciona un Lasso como su especificación óptima, y si Lasso reduce ligeramente los coeficientes pero no excluye a ningún *feature* de la especificación. Entonces, intuitivamente esto podría estar indicando que todos los predictores previamente pre-seleccionados poseen información relevante para estimar el ingreso de los hogares de la muestra.

Sin embargo, dado que la función de pérdida requerida para este ejercicio es lineal asimétrica y no cuadrática, se debe tener en cuenta que Christoffersen & Diebolt (1997) demuestran que la mejor predicción para este tipo de costos asimétricos proviene de regresiones cuantílicas, más no de proyecciones lineales de la media condicional (MCO). En especial, debido que el costo de sobrestimar el ingreso es tres veces superior al de subestimarlo, se debería trabajar con regresiones cuantílicas al percentil 25.[10] Por lo tanto, se reconfiguraron las funciones de pérdida para las tres *baseline functions* utilizando una métrica D^2 *Pinball (P25)*[11] y esto derivó en la estimación sobre la base de datos *train 2.0* de tres modelos similares a regresiones cuantílicas para el percentil 25 con extensiones Lasso, Ridge y *EN*[12].

---

[4] Creado a partir de dos (2) variables dicótomas; i) `univ[1=(P6210=6)]` que hace referencia a grado universitario y ii) `bajaeduc[1=(P6210=1|2|3)]` que hace referencia a baja educación (ver estadísticas descriptivas en Anexos).

[5] Esta constituye la unión del valor de los arriendos de los que hogares que, efectivamente, pagan y las disposiciones a pagar por el uso de la vivienda donde habitan; quienes no lo hacen (son dueños).

[6] Se define hacinamiento según la ELCA si habitan entre 2,5 o más personas por cuarto para dormir. De manera que la variable se construyó como `haci`$= \frac{\texttt{Nper}}{\texttt{p5010}}$ donde `Nper` corresponde al número de personas habitando el hogar y `p5010` al número de cuartos donde duermen las personas del hogar incluyendo sala-comedor.

[7] Como variable dicótoma que toma el valor de 1 si la empresa posee menos de 50 empleados y 0 de lo contrario.

[8] La métrica de evaluación implementada corresponde a un promedio ponderado entre la *False Negative Rate (FNR)* y la *False Positive Rate (FPR)*: $W\text{-}MIR = 0,75 \times (FNR) + 0,25 \times (FPR) = 0,75 \times (\frac{FN}{TP+FN}) + 0,25 \times (\frac{FP}{TN+FP})$

[9] $\lambda_{EN} = \frac{\lambda_L}{\lambda_L + \lambda_R}$

[10] Se utiliza cuando hay una función de pérdidas lineal a ambos lados pero asimétrica: $L(\varepsilon = y - \hat{y}) = a|y-\hat{y}|, y-\hat{y} > 0 \wedge b|y-\hat{y}|, y-\hat{y} < 0$. En estos casos, el pronostico óptimo de la variable a estimar corresponderá al percentil $\frac{a}{a+b}$ de la distribución predicha. Entonces, para este ejercicio en particular, con $a = 1$ y $b = 3$, el pronostico óptimo del ingreso debería corresponder al percentil 25 de la distribución.

[11] Definido como una función de puntuación de la regresión, es una métrica análoga al R^2 pero aplicable para regresiones cuantílicas, dado que retorna la fracción de la pérdida de *Pinball*(P25), $dev(y, y_{P25})$, explicada por la estimación de $\hat{y}$. Se define como $D^2(y, \hat{y}) = 1 - \frac{dev(y,\hat{y})}{dev(y,y_{P25})}$

[12] Estas tres especificaciones no provienen directamente de una regresión cuantílica, pero con la función de pérdida D^2 *Pinball (P25)* las predicciones deberían asemejarse a las de una regresión cuantílica al percentil 25 con su correspondiente técnica de reducción/selección.

Luego de utilizar *10-Fold Cross-Validation* para el entrenamiento, se encontró que: para la regresión *P25* con Lasso, $\lambda^*_{L,25} = 1,00 \times 10^{-9}$ lo cual es consistente con el hecho de que no se presentó selección y se utilizaron todas las variables iniciales; para *P25* con Ridge, $\lambda^*_{R,25} = 1,00 \times 10^{-9}$; y para *P25* con *Elastic Net*, $\lambda^*_{EN,25} = 0$ y $\alpha^*_{EN,25} = 0$, la especificación toma la forma de un Ridge, pero no solo se queda con todas las variables del *train 2.0* (no hay selección), sino que la penalización converge a 0 (no hay reducción). Esto puede implicar que, intuitivamente, el uso de regresiones con funciones de pérdida que penalizan más la sobrestimación del ingreso puede estar reduciendo de forma *ex-ante* los coeficientes óptimos utilizados por el modelo entrenado. Consecuentemente, como los coeficientes ya están reducidos para evitar sobre-estimaciones, la labor del Lasso y del Ridge de penalizar por la magnitud de los estimadores se vuelve redundante.

Adicionalmente, teniendo en cuenta la literatura previamente referenciada, se consideró relevante el ejercicio de estimar una especificación adicional que tomara directamente la forma de una regresión cuantílica al percentil 25 (sin incluir procedimientos de selección de variables o reducción de parámetros). Esto, con la intención de analizar si el uso de las otras metodologías de predicción previamente propuestas si pueden mejorar los resultados en comparación con el óptimo teórico.

Tabla 2. Métricas de evaluación: modelos de regresión.

Métrica/Metodología	Lasso	Ridge	*EN*	Lasso (p25)	Ridge (p25)	*EN* (p25)	Q25
W-MIR	0,2443	0,2180	0,2443	0,2200	0,2181	0,2200	0,1652
FNR	0,2532	0,2047	0,2532	0,2027	0,2046	0,2027	0,0223
FPR	0,2176	0,2582	0,2176	0,2719	0,2586	0,2719	0,5941

Ahora, dado que en estas especificaciones no se presentó una selección automática de variables relevantes (no hubo *"Lasso's free lunch"*), pero que igualmente un problema de sobre-especificación no identificado podría estar aumentando la varianza de las predicciones, se decidió realizar una prueba de robustez sobre los dos mejores modelos (así como se evidencia más adelante en la Tabla 3, Ridge y regresión cuantílica (Q25) fueron los modelos que presentaron mejor *W-MIR*). Utilizando el resultado de una matriz de correlaciones entre la variable objetivo (ingreso del hogar) y los *features*, se seleccionaron a las cinco variables que presentaban una mayor correlación con la variable de interés para constituir una nueva sub-muestra de entrenamiento (*train-selec*) pero con menor cantidad de predictores, estas variables fueron: i) si el jefe del hogar pertenece a régimen de salud subsidiado; ii) el mayor nivel educativo del jefe del hogar; iii) si el jefe del hogar trabaja en una empresa pequeña; iv) si la vivienda del hogar es propia y; v) el nivel de hacinamiento en la vivienda.

Tabla 3. Métricas de evaluación: modelos de regresión adicionales.

Métrica/Metodología	Ridge	Q25	Ridge-TS	Q25-TS
W-MIR	0,2180	0,1652	0,2974	0,2054
FNR	0,2047	0,0223	0,3006	0,0695
FPR	0,2582	0,5941	0,2880	0,6130

Al realizar estas dos nuevas estimaciones, la Tabla 3 esboza que Ridge con *train-selec* (TS) requiere de una mayor penalidad para optimizar sus resultados de predicción, $\lambda^*_{R,TS} = 32,32 > \lambda^*_R = 4,04$. No obstante, como se muestra en la Tabla 2, al momento de evaluar estos modelos dentro del *test-selec*, esta mayor reducción en los coeficientes no es suficiente para compensar por la pérdida de información de reducir el número de predictores. Este mismo comportamiento se evidencia para la regresión Q25, la cual tampoco obtuvo ganancias en poder explicativo cuando se redujo el número de predictores (Q25-TS). Por lo tanto, para los modelos de regresión, se puede llegar a dos conclusiones generales: primero, el hecho de que Lasso no eliminara ninguno de los predictores implica que todos proveían información relevante para pronosticar si un hogar va a ser pobre o no (según su nivel de ingreso), lo cual se comprobó al contrastar los resultados iniciales de Ridge y Q25 con los resultados de sus modelos con menores variables; y segundo, se corroboran empíricamente los resultados teóricos de Christoffersen & Diebolt (1997), debido a que el mejor modelo de regresión fue el Q25, y a que en la mayoría de los casos los modelos con función de perdida D^2 *Pinball (P25)* se comportaron mejor que sus contrapartes con perdida cuadrática.

3.2 Clasificación

Ahora bien, aquí se ocupa del problema de predicción de la pobreza en los hogares de forma directa, asignando valores de 1 si el hogar el pobre y 0 de lo contrario. Son siete (7) las técnicas empleadas y, al igual que en los modelos de regresión, la métrica de evaluación sobre la clasificación de falsos positivos *vs.* falsos negativos es la *W-MIR*. Dos metodologías corresponden a *Linear Discriminant Analysis (LDA)* con *Singular Value Decomposition (SDV)*[13] y *LDA* por MCO.

[13]Es una factorización de una matriz real o compleja. Generaliza la descomposición propia de una matriz normal cuadrada con una base propia orto-normal a cualquier matriz de tamaño $m \times n$.

Asimismo, la sofisticación de las anteriores responde a *Quadratic Discrimininant Analysis (QDA)* que, a diferencia de *LDA*, asume que cada clase (de las observaciones) tiene su propia matriz de covarianza. Luego, un modelo Logit con distribución logística de los errores. Las siguientes dos son Bayesianas; particularmente, el clasificador *Naive Bayes (NV)* con una distribución de Bernoulli para las variables dicótomas y el clasificador *NB* con distribución Gaussiana (normal) para las variables continuas. Finalmente, *K-Nearest-Neighbors (KNN)* con $K^* = 5$, iterado mediante *10-Fold Cross-Validation* sobre una grilla de hiperparámetros de forma $\mathcal{G} = \{5, 10, 15, 20\}$.

Tabla 4. Métricas de evaluación: modelos de clasificación.

Métrica/Metodología	*LDA-SVD*	*LDA*-MCO	Logit	*QDA*	*NB-Ber*	*NB-Gau*	*KNN*
W-MIR	0,4281	0,2102	0,2879	0,2084	0,2916	0,4585	0,2211
FNR	0,5482	0,2101	0,3461	0,1309	0,2674	0,4263	0,1762
FPR	0,0675	0,2105	0,1133	0,4407	0,3641	0,5551	0,3554

La Tabla 4 exhibe los valores calculados de *W-MIR, FNR* y *FPR* para los modelos de clasificación. Análogo a los resultados encontrados acerca de los modelos de regresión con Q25, *QDA* domina sobre el resto en la métrica de evaluación personalizada y la tasa de falsos negativos con valores de 0,2084 (20,84%) y 0,1309 (13,09%), respectivamente. Por su parte, *LDA-SVD* es significativamente menor frente a las demás metodologías pues su tasa de falsos positivos es, al menos, un 40% menor que el *"second best"* (*LDA*-MCO) con 0,0675 (6,75%). Nuevamente, no existió una preselección de variables presumiblemente relevantes, pero sigue la sospecha que algunas de ellas puedan estar generado ruido o sobre-identifiquen a hogares que en realidad no son pobres. Por ello se repite la prueba de robustez hecha en la subsección anterior dada una muestra *train-selec*. Sin embargo, las variables circunscritas dentro de *train-selec* son diferentes, en particular; i) si el jefe del hogar está ocupado; ii) si el jefe del hogar pertenece a régimen de salud subsidiado; iii) el mayor nivel educativo del jefe del hogar; iv) si el jefe del hogar trabaja en una empresa pequeña y; v) la suma de las horas trabajadas a la semana en el hogar.

Tabla 5. Métricas de evaluación: modelos de clasificación adicionales.

Métrica/Metodología	*LDA-MCO*	*QDA*	*LDA-MCO*-TS	*QDA*-TS
W-MIR	0,2102	0,2084	0.3061	0.2942
FNR	0,2101	0,1309	0.2720	0.3370
FPR	0,2105	0,4407	0.4082	0.1658

Luego de correr estas dos nuevas estimaciones, es posible dilucidar en la Tabla 5 que *QDA* mantiene su ventaja comparativa respecto a *W-MIR* como *FNR*, pero con menor número de regresores. Este resultado no se sostiene para la *FPR* pues allí si se pone en relieve que la reducción de variables hizo disminuir la tasa en poco más del 50% de *QDA*, pero no lo suficiente como para acercarse a *LDA-MCO*. Esto lleva a resultados cualitativamente idénticos a los encontrados en la subsección anterior pues no se excluyen variables relevantes que arrojen aquel método de clasificación con las menores tasas de error y vuelve a cumplirse lo demostrado en Christoffersen & Diebolt (1997). Por lo que ni en el análisis de regresión ni en clasificación soporta sub-seleccionar los "mejores" regresores.

3.3 Veredicto

De los catorce (14) modelos originales mas los cuatro (4) modelos adicionales especificados e implementados para intentar predecir (tanto de forma directa como indirecta) si un hogar es pobre o no. Se encontró que, bajo un supuesto de perdida asimétrica en donde pesan tres veces mas los falsos negativos a los falsos positivos (3:1), los mejores predictores son aquellos modelos que: i) así como lo indica la literatura internalizan mejor la función de perdida requerida, como el Q25; ii) toman en cuenta la posible existencia de heteroscedasticidad dentro de las observaciones de la muestra, como el *QDA* y; iii) buscan incluir a todos los posibles predictores relevantes. Además, a partir de los resultados, también se puede evidenciar que no necesariamente las estimaciones directas de una variable dicótoma, mediante modelos de clasificación, van a presentar un desempeño predictivo estrictamente superior a las estimaciones indirectas de esta variable dicótoma, mediante modelos de regresión. Tanto así que, en esta ocasión, el modelo con mejor desempeño en las métricas mas relevantes (*W-MIR* y *FNR*) correspondió a un modelo de regresión.

En especifico, el modelo de regresión cuantílica al percentil 25 fue el ganador indiscutido, ostentando un *W-MIR* de 0,1652 (16,52%) y un *FNR* de 0,0223 (2,23%), lo cual implica que este modelo es altamente efectivo a la hora de no clasificar erróneamente a los pobres como no pobres (el objetivo principal de este ejercicio). Pero, esto a costa de presentar el *FPR* mas alto de todas las estimaciones, 0,5941 (59,41%), lo que implica que este modelo fácilmente termina identificando a personas no pobres como pobres.

Por el lado de los modelos de clasificación, el modelo *QDA* fue mejor que sus pares tanto en *W-MIR* como en *FNR*, pero fue sustancialmente menor efectivo que el Q25. Ahora, si hay que resaltar que su *FPR* no fue ni tan alto como el de Q25 ni fue el mas alto dentro de los modelos de clasificación, lo cual podría indicar que este modelo maneja mejor el *trade-off* entre *FP* y *FN* dentro de sus predicciones. Por lo tanto, a partir de la evidencia empírica recopilada, seria razonable esperar que el modelo Q25 sea el mejor a la hora de predecir la incidencia de la pobreza, sin clasificar mal a los hogares pobres, dentro de la muestra objetivo *test*.

4 Conclusiones

En este documento, se buscó desarrollar una metodología de predicción que permitiera mejorar la identificación de hogares pobres a partir de sus características sociales, demográficas y económicas, sin la necesidad de tener datos explícitos acerca de sus niveles de ingreso. Mas sin embargo, prestando especial cuidado en no cometer el error de clasificar erróneamente (como no pobre) a hogares vulnerables que sí deberían ser clasificados como tal. Luego de implementar hasta dieciocho (18) especificaciones diferentes, tanto de regresión como de clasificación (contando las pruebas de robustez), se encontró que las mejores son aquellas que son consistentes con los resultados teóricos de Christoffersen & Diebolt (1997), según los cuales, ante funciones de perdidas asimétricas el pronostico óptimo corresponde a una predicción cuantílica. No obstante, estimaciones alternativas de clasificación que tomen en cuenta la presencia de heteroscedasticidad, también pueden ser efectivas para cumplir con este objetivo de predicción.

Con base en lo planteado anteriormente, es altamente recomendable la aplicación práctica de *approaches* como los que captura el documento para impactar sobre las variables identificadas como relevantes en los modelos y así lograr combatir la pobreza, sin dejar a ninguno de los hogares verdaderamente vulnerables sin la cobertura que pueden estar requiriendo para superar trampas de pobreza. De hecho, cuando este objetivo (mejor identificación) se logra resulta más sencillo notar que en los frentes como: i) transferencias monetarias; donde entre el 20% y el 30% de los ingresos de los hogares pobres y vulnerables en el promedio nacional corresponden a las ayudas directas que el Gobierno otorga para subsanar necesidades como alimentación, educación y salud. Es necesario focalizar aún más el destino de los recursos a hogares que presenten jefes de hogar con características de bajo ingreso, poco nivel educativo y ausencia de régimen de salud, toda vez que se hace una supervisión estricta de su manejo. También que; ii) la inclusión productiva debe asegurar la entrega a los más vulnerables de las herramientas básicas para que, realmente, puedan alcanzar las oportunidades en igualdad de condiciones. Entre estas la provisión de mejores bienes y servicios públicos, la mejora integral de vías e infraestructura y el acceso y la conectividad a las redes de comunicación (i.e. internet).

Finalmente que una; iii) tributación equilibrada y responsabilidad fiscal: equilibra "la cancha" entre el sector formal e informal (ver estadísticas descriptivas en Anexos) así como la tasa impositiva a las personas jurídicas *vs.* las naturales para lograr mayor captación y destinar aún más recursos a los programas sociales focalizados. Esto con un correcto manejo de los impuestos por su buena captación y prudencia en el endeudamiento, se convierten en un círculo virtuoso con alto potencial de reducir la brecha entre los ricos y pobres vía distribución del ingreso.

Bibliografía

Christoffersen, P., & Diebolt, F. (1997). Optimal Prediction Under Asymetric Loss. *Econometric Theory*, 13, pp. 808-817.

DANE. (2019). Pobreza monetaria en Colombia. *Boletín Técnico*. (Disponible en: https://www.dane.gov.co/files/investigaciones/condiciones_vida/pobreza/2018/bt_pobreza_monetaria_18.pdf)

Herrera, C. (2017). *Pobreza y Prejuicio* (1st ed.). Colombia: Editorial Planeta.

Ocampo, J., & Gómez-Arteaga, N. (2017). Los sistemas de protección social, la redistribución y el crecimiento en América Latina. *Repositorio CEPAL*. (Disponible en: https://repositorio.cepal.org/bitstream/handle/11362/42030/1/RVE122_Ocampo.pdf)

Robinson, J. (2016). La miseria en Colombia. *Revista Desarrollo y Sociedad*, Número 76, pp. 9-90.

Anexos

Figuras

Figura 1. Ingreso per cápita (hogar).

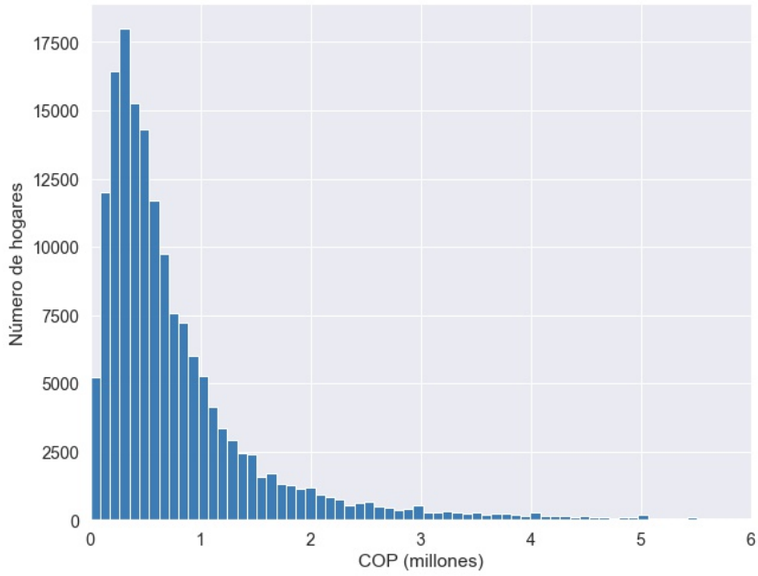
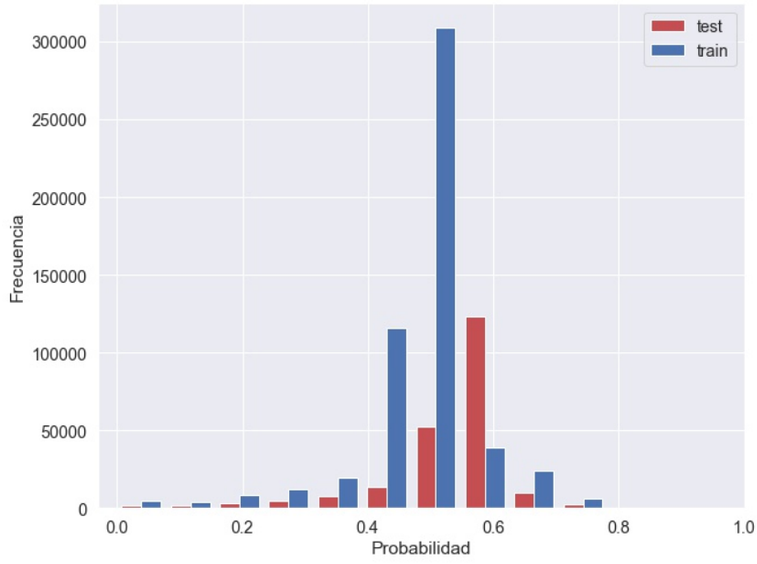


Figura 2. *PSM*: probabilidad de pertenencia a *test*.



Tablas

Tabla 1. Estadísticas descriptivas (sin agrupar).

VARIABLE	Obs.	Media	Desviación	Mínimo	Mediana	Máximo
Ingreso per cápita (hogar)	164960	864194.36	1225467.34	0	541666.67	88833333.33
Ingreso total (persona)	447512	775591.56	1380447.03	0	430000	85833333.33
Personas en el hogar	164960	3.29	1.77	1	3	22
Edad	542941	33.55	21.64	0	31	110
Pobre	542941	0.25	0.43	0	0	1
Sexo [1=Mujer]	542941	0.47	0.50	0	0	1
Estrato 1	542941	0.22	0.41	0	0	1
Estrato 2	542941	0.26	0.44	0	0	1
Estrato 3	542941	0.22	0.42	0	0	1
Estrato 4	542941	0.11	0.32	0	0	1
Estrato 5	542941	0.10	0.30	0	0	1
Estrato 6	542941	0.09	0.28	0	0	1
Sin estudios	542941	0.05	0.22	0	0	1
Preescolar	542941	0.03	0.17	0	0	1
Básica primaria	542941	0.25	0.43	0	0	1
Básica secundaria	542941	0.17	0.38	0	0	1
Bachillerato	542941	0.22	0.41	0	0	1
Superior	542941	0.24	0.42	0	0	1
Subsidio	115975	0.21	0.41	0	0	1
Ocupados	542941	0.46	0.50	0	0	1