Problem Set 2: Predicting Poverty

Big Data y Machine Learning para Economía Aplicada

Gustavo Adolfo Castillo Álvarez (201812166),
Alexander Almeida Ramírez (202225165),
Jorge Luis Congacha Yunda (201920042) y
Jaime Orlando Buitrago González (200612390)

14-04-2024

1 Introducción

La lucha contra la pobreza ha sido un tema central en la agenda global durante décadas, debido a que afecta a millones de personas en el mundo. En este sentido organismos como el Banco Mundial han desempeñado un papel importante en la recopilación de datos y generación de análisis que ayudan a comprender la magnitud y las tendencias de la pobreza a nivel mundial. Sin embargo, este esfuerzo enfrenta desafíos especiales en países o regiones con limitaciones de datos, recursos, crisis y cambios constantes.

En este contexto, la predicción de la pobreza se vuelve fundamental para analizar como fenómenos como la pandemia de COVID-19 pueden estar afectando este aspecto, especialmente la pobreza extrema en diversas regiones (Yonzan y Lakner 2023). Se ha observado un avance en la generación de predicciones de la pobreza por parte de entidades como el Banco Mundial (Castaneda Aguilar et al. 2024), así como otras instituciones nacionales e internacionales, con el propósito de proporcionar información relevante para la formulación de políticas públicas.

En Colombia, también se han realizado avances en la medición de la pobreza a través de organismos como el DANE y universidades. No obstante, continúa siendo un desafío importante, complejo y costoso que requiere cierto tiempo. Por lo tanto, la predicción de la pobreza mediante modelos de aprendizaje automático puede resultar eficiente y accesible para los territorios, ofreciendo una oportunidad para agilizar y mejorar la evaluación de la pobreza. Esto permitiría una mayor focalización de recursos y políticas, siendo aún más efectivo al incorporar datos específicos de cada territorio, como conflictos armados, condiciones de vida, estructura y actividad productiva.

En este contexto, como una alternativa a las limitaciones de recursos económicos y de tiempo, este estudio propone desarrollar un modelo de clasificación y predicción de la pobreza a nivel de hogares en Colombia, usando datos del Departamento Administrativo Nacional de Estadística - DANE y la Misión para el Empalme de las Series de Empleo, Pobreza y Desigualdad - MESE) del año 2018, que incluyen información a nivel de individuo y hogar. Estos datos se dividen en dos bloques: una base de entrenamiento y una base de prueba, ambas a nivel de persona y hogar. La diferencia entre estas bases radica en que la de entrenamiento contiene todas las variables necesarias para el cálculo de la pobreza, lo que sirve como conjunto de entrenamiento para nuestra predicción, mientras que la de prueba carece de variables como ingresos, necesarias para este cálculo.

Ahora, dado que estas bases contienen más de 200 variables, uno de los primeros pasos fue seleccionar las más relevantes para nuestro propósito. En este sentido, se optó por seleccionar XXX variables, teniendo en cuenta el esfuerzo y la capacidad computacional requerida para ejecutar modelos con una gran cantidad de variables y observaciones. Estas variables, seleccionadas a partir de una revisión de literatura, incluyen características socioeconómicas como edad, género, nivel educativo y tipo de vivienda, entre otras. Después de plantear y probar varios modelos de regresión y clasificación de pobreza, como Logit, Bagging, Boosting, Random-Forest, Logit-Carret, QDA e IDA, se determinó que el modelo con mejores resultados era el QDA, con una precisión del 0.56. Es importante destacar que la precisión de este modelo no difiere mucho de otros, como el Logit-Lasso y Bagging, cuyos puntajes según Kaggle son 0.56 y 0.55, respectivamente.

En conclusión, se evidencia que es posible realizar una predicción adecuada de la pobreza con un conjunto reducido de variables y herramientas adecuadas de Big Data y Machine Learning. Esta información constituye un insumo importante para la formulación de políticas públicas dirigidas a combatir la pobreza. Sin embargo, se reconoce la necesidad de incorporar otras variables socioeconómicas y de vivienda para lograr una estimación más precisa de los ingresos del hogar y, por ende, una mejor estimación de la pobreza. También se recomienda explorar otras especificaciones del modelo, como aumentar el grado polinomial en las variables y realizar interacciones entre variables.

2 Datos

Para el Problem Set se utilizó el módulo de pobreza monetaria de la Gran Encuesta Integrada de Hogares (GEIH) de 2018. Mensualmente, el DANE recolecta información sobre los ingresos y el mercado laboral de una muestra representativa de la población colombiana, así la muestra de cada mes es representativa para Colombia y anualmente para 23 departamentos y Bogotá, sus capitales y áreas metropolitanas, y otros dominios (rural y urbano) (DANE 2019). La operación estadística tiene aproximadamente 750 mil observaciones o personas, 230 mil hogares y 30 mil viviendas.

Con la información de los ingresos el DANE calcula anualmente la pobreza monetaria. El Comité de Expertos emite los conceptos técnicos con los que se definen las líneas de pobreza y pobreza extrema con las que se clasifican a los hogares [1] como pobres. Si el ingreso per cápita es inferior a las líneas definidas, el hogar se considera pobre, pues no cuenta con los suficientes ingresos para cubrir los requerimientos nutricionales mínimos (pobreza extrema) o los bienes y servicios básicos (CONPES 2012). Para cada uno de los miembros del hogar, se tienen en cuenta los ingresos por salarios, ganancias u honorarios, ingresos en especie, otras fuentes, para definir el ingreso total. Adicionalmente, se realizan imputaciones y correcciones a las bases de datos (CONPES 2012), de tal manera que se obtiene información oportuna y de calidad.

Para resolver la pregunta asociada al Problem Set se utilizó una base de datos previamente preparada para realizar predicciones sobre la pobreza monetaria e ingreso de la población Colombia en 2018. Esta base de datos descargada de *Kaggle* (Sarmiento-Barbieri 2024), cuenta con 762.753 observaciones a nivel de personas y 231.128 observaciones para hogares. La primera está divida en 543.109 observaciones en una base de entrenamiento (train) y 219.644 de prueba (test), y la segunda en 164.960 observaciones para entrenamiento y 66.168 de test.

La división de las bases se realizó para efectos pedagógico, eliminando algunas de las variables de base de datos train. Por tal motivo, las predicciones se realizaron sólo con las variables compartidas

[1] En estricto sentido se denominan unidades de gasto, pero para simplificar se utilizará el términos hogares

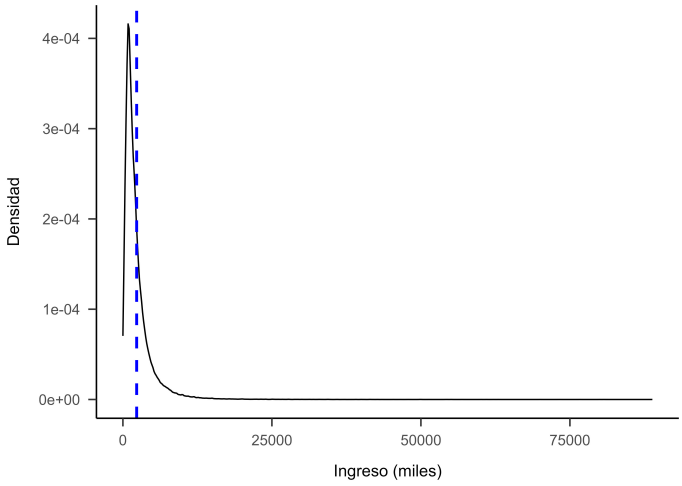
por las bases train y test. Se manipularon las variables categóricas, creando *dummies* para cada categoría, y para continúas reemplazando con 0 valores perdidos. Como la predicción se realizó por hogar, se agregaron las variables como promedios y sumas de personas. Las estadísticas descriptivas de las variables transformadas utilizadas se observan a continuación:

Statistic	N	Mean	St. Dev.	Min	Max
edad	164,960	37.44	16.88	5.67	102.00
edad_2	164,960	1,890.50	1,486.86	77.25	10,404.00
Genero	164,960	0.53	0.28	0.00	1.00
estudiante	164,960	0.12	0.19	0.00	1.00
busca_trabajo	164,960	0.03	0.10	0.00	1.00
amo_casa	164,960	0.20	0.24	0.00	1.00
hijos_hogar	164,960	0.20	0.24	0.00	1.00
primaria	164,960	0.05	0.16	0.00	1.00
secundaria	164,960	0.16	0.24	0.00	1.00
media	164,960	0.22	0.28	0.00	1.00
superior	164,960	0.26	0.33	0.00	1.00
Ingtot	164,960	2,102.59	2,532.55	0.00	85,833.33
Ingtotugarr	164,960	2,307.86	2,628.93	0.00	88,833.33
exp_trab_actual	148,947	65.80	91.44	0.00	948.00
horas_trab_usual	164,960	30.39	18.05	0.00	130.00
Pobre	164,960	0.20	0.40	0	1
Nper	164,960	3.29	1.77	1	28
num_menores	164,960	0.75	1.01	0	14
num_adulto	164,960	0.32	0.61	0	6
eps	164,959	0.93	0.21	0.00	1.00
rural	164,960	0.91	0.29	0	1
num_cuartos	164,960	3.39	1.24	1	98
num_cuartos_dormir	164,960	1.99	0.90	1	15
Npersug	164,960	3.28	1.77	1	28
vivienda_arriendo	164,960	0.39	0.49	0	1
vivienda_propia	164,960	0.00	0.00	0	0
contributivo	164,960	0.40	0.41	0.00	1.00
Desempleado	164,960	0.05	0.15	0.00	1.00
Inactivo	164,960	0.31	0.31	0.00	1.00
Ocupado	164,960	0.50	0.32	0.00	1.00

Cuadro 1: Estadísticas descriptivas

Como se observa en la tabla anterior, en la base de entrenamiento predominan las personas con formación hasta media y superior (20% y 26%, respectivamente), la mayoría tiene hijos menores (75%), la mitad tiene población ocupada y es mujer (50%). Tomando en cuenta que el ingreso promedio es de 2,1 millones y la cantidad de personas promedio es de 3.2 personas, el ingreso per cápita es de $641.032,3, lo cual supera las líneas de pobreza e indigencia. No obstante, también se observa una desviación estándar alta, incluso superior al ingreso promedio, esto explica que el 20% de los hogares sean pobres.

Figura 1: Distribución del ingreso. Base train



De hecho, la distribución del ingreso permite ver una gran concentración de hogares por debajo del ingreso promedio. Es por tanto, que es apremiante buscar alternativas para estimar la población pobre por otros medios, incluidos el Machine Learning.

3 Modelos y resultados

3.1 Modelos de Clasificación

Para clasificar los hogares se buscó directamente predecir la clase $k \in \{Pobre, NoPobre\}$, a la que pertenecen los hogares, en este sentido se trató de un problema de clasificación binaria pues solo hay dos clases posibles. Dado que un hogar es clasficado como pobre directamente si su ingreso es menor a una línea de pobreza, se buscó estimar la categoría de un hogar a partir de características socioeconómicas del hogar: proporción de mujeres, edad promedio de los integrantes del hogar, número de menores, etc. Así, el objetivo era aproximar la probabilidad condicional de que un hogar con cierto vector de características $\boldsymbol{X}$ fuera pobre o no, i.e. $\mathbb{P}[y_i = k|X_i = x] = p(x)$. La expresión analítica de $p(x)$ depende del modelo estimado. En las dos regresiones logísticas, con y sin regularización lasso, se preprocesaron estandarizandolas. En todos los modelos realizados se usaron 15 variables independientes: edad media del hogar, edad media al cuadrado, proporción de mujeres, proporción de estudiantes, proporción de personas buscando trabajo, proporción de personas dedicadas a oficios del hogar, proporción de personas con primaria, prop. de personas con secundaria, prop. de personas con educación superior, número de cuartos, prop. de personas en régimen contributivo, número de adultos mayores, dicotómica de hogar rural, y dicotómica de si la vivienta es arrendada.

Utilizamos 6 diferentes algoritmos de clasificación: logit, QDA, LDA, *random forest*, *bagging* y *GBM*. Dado que cada uno de estos modelos "aprende" la frontera de decisión utilizando diferentes enfoques sus ecuaciones respectivas varían significativamente. El único modelo discriminativo que implementamos, el logit, directamente estima la probabilidad condicional mediante una función logística (1).

$$\mathbb{P}[y_i|X_i] = \frac{\exp(\boldsymbol{X}\boldsymbol{\beta})}{1+\exp(\boldsymbol{X}\boldsymbol{\beta})} \tag{1}$$

$$\mathbb{P}[y_i = k|X_i = x] = \frac{\pi_k f_k(x)}{\pi_1 f_1(x) + \pi_2 f_2(x)}, k \in \{0, 1\} \tag{2}$$

Los dos modelos generativos como el QDA y LDA en cambio descomponen $p(x)$ como una distribución posterior y mediante Bayes y haciendo supuestos sobre la densidad $f_k(x)$ condicional a la clase k la estiman de diferentes formas, en ambos casos suponiendo que surgen de distribuciones Gausianas mutivariadas. En el caso de LDA supone una varianza común entre las clases, mientras que QDA supone una varianza y media específica a cada clase.

Los últimos modelos basados en árboles posibilita unas relaciones no lineales entre las características y la variable dicotómica de Pobre. Tanto el *Random Forest* como el *bagging* agregan varios árboles de forma independiente usando bootstrap en cada nuevo árbol, utilizando el primero un subconjunto alreatorio de los predictores en cada iteración.

3.2 Modelos de Predicción de Ingreso

Los modelos de regresión tenían como objetivo predecir el ingreso del hogar, y a partir de este escoger el umbral de línea de pobreza a partir del cuál obtuviéramos el mejor desempeño fuera de muestra. Para este ejercicio utilizamos también los mismos algoritmos que aquellos usados previamente. Se supuso una relación lineal entre los predictores y el ingreso y se predijo el ingreso mediante los modelos de regresión tradicionales.

3.3 Modelos Finales (QDA)

Para el modelo de clasificación, se estimó la siguiente especificación:

$$Pobre = edad + edad^2 + mujer + estud + trab + ofi + prim + secun + media + super + cuartos + contri + adul + rur + arr$$

La siguiente tabla presenta la descripción de cada una de estas variables: edad(debido a la relación de u invertida entre el ingreso y la edad), genero (existe evidencia de que las mujeres tienen menor ingresos que los hombres), variables que identifican la ocupación (un hogar con mayor cantidad de personas que estudian, buscan trabajo o se dedican a los oficios del hogar por lo general tendrán menores ingresos), variables que identifican la educación (existe una correlación entre educación e ingresos), número de cuartos en la vivienda(un hogar con mayor cantidad de cuartos es más probable que no sea pobre), porcentaje de personas que están en el régimen contributivo(una familia con régimen contributivo implica que tienen mayores ingresos o empleo formal), número de adultos(los hogares con mayor cantidad de adultos implica que está conformado con personas con capacidad de trabajar), si la vivienda es rural(existe evidencia de que las personas que viven en zona rural tienen menos ingreso) y si la vivienda es arrendada (las personas que viven en arriendo pueden tener menos ingresos que personas que tienen casa propia).

Predictor	Descripción
Edad	Edad promedio de las personas dentro de un hogar.
Edad_2	El cuadrado de la variable edad.
estud	Número de personas dentro del hogar que se dedicaron mayor parte del tiempo a estudiar.
mujer	Proporción de personas dentro del hogar que son mujeres.
trab	Número de personas dentro del hogar que se dedicaron mayor parte del tiempo a buscar trabajo.
ofi	Número de personas dentro del hogar que se dedicaron mayor parte del tiempo a oficios del hogar.
prim	Número de personas dentro del hogar con educación primaria como mayor nivel educativo.
secun	Número de personas dentro del hogar con educación secundaria como mayor nivel educativo.
media	Número de personas dentro del hogar con educación media como mayor nivel educativo.
super	Número de personas dentro del hogar con educación superior(Técnico o tecnológico, Universitario o Postgrado) como mayor nivel educativo.
cuartos	Número de cuartos dentro del hogar.
adul	Número de adultos dentro del hogar.
rur	=1 si el hogar está en la zona rural y 0 si está ubicado en la zona urbana.
arri	=1 si la vivienda ocupada por el hogar es en arriendo o subarriendo
contri	Promedio de personas dentro del hogar que están afiliados al régimen contributivo de seguridad social en salud.

Para el enfoque de clasificación, se realizaron cinco estimaciones: Logit, QDA, LDA, logit-ridge y random Forest. Todas las estimaciones de estas metodologías se realizaron con las mismas variables. Para todos los modelos, escogimos los hiperparámetros usando cross validación. Este proceso consiste en dividir los datos en k subconjuntos y entrenarlos k veces, utilizando un subconjunto de prueba y los restantes de entrenamiento en cada entrenamiento. En el caso de los modelos que se estimaron en este proyecto, se realizó 5 folds o particiones en las que se dividieron los datos en la validación cruzada. Se realizó este proceso debido a que permite evaluar el rendimiento de los modelos de forma más robusta, ya que permite evaluar las predicciones y garantizar que son independientes de la partición entre datos de entramiento y prueba. Con esto, se garantiza que no existan problemas de sobreajuste(por ejemplo, que tenga una precisión de 1 en el conjunto train, pero que tenga una precisión muy baja en el conjunto de datos fuera de la muestra).

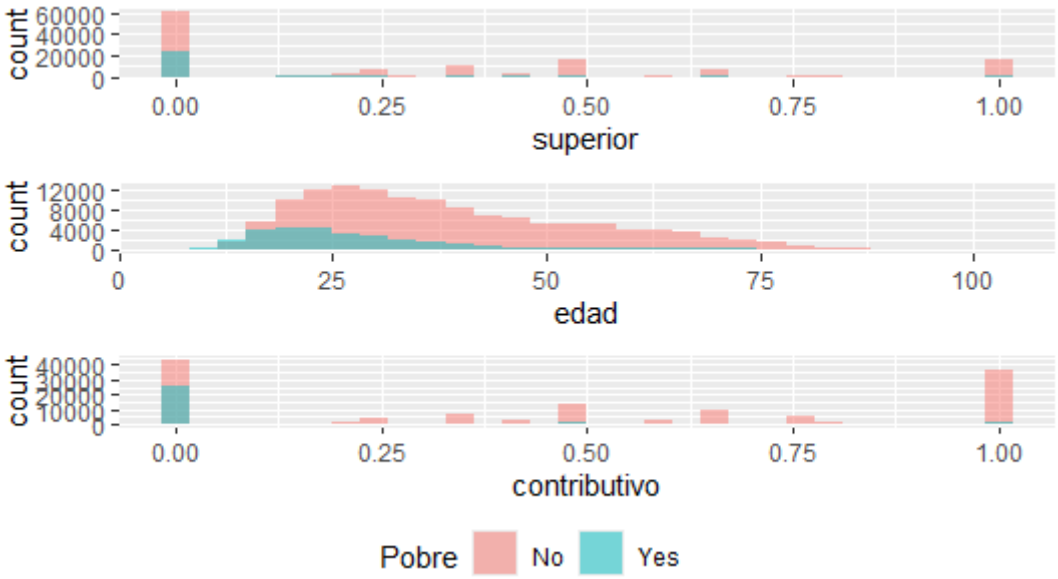
Al momento de realizar las predicciones, el modelo con mejor Accuracy en el test set fue el Análisis Discriminante Cuadrático (QDA). Este es un algoritmo de clasificación supervisada se utiliza para separar las clases diferentes basadas en características observadas. Este algoritmo asume que cada observación de cada clase tiene su propia distribución gaussiana y resulta de introducir estimaciones de los parámetros en el teorema de Bayes para realizar la predicción. Además, supone que cada clase tiene su propia matriz de covarianza.

En la siguiente tabla se puede observar la tabla con la correlación entre la variable pobre(transformada en dicotómica) y las demás variables. Se puede apreciar que las variables que más contribuyen a la predicción son la edad promedio del hogar, el porcentaje de personas que están en el régimen contributivo y el número de personas del hogar con educación superior, debido a que tienen una

correlación alta con la variable que clasifica a la población en dos grupos. Podemos observar también, en la gráfica x, la distribución de estas variables según la clase a la que pertenece(pobre o no pobre). Este gráfico muestra que estas mismas variables son las que tienen distribuciones que menos se solapan entre los grupos.

variable	correlación
edad	-0.20
edad	-0.15
mujer	0.04
estud	0.09
trab	0.08
ofi	0.09
prim	0.14
secun	0.09
media	-0.06
super	-0.25
cuartos	-0.14
contri	-0.37
adul	-0.04
rur	-0.08
arr	0.05

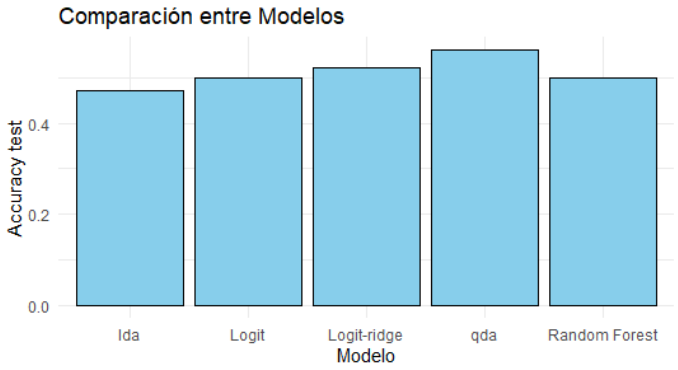
Figura 2: Distribución de las variables más importantes entre grupos



Las métricas que se utilizaron para seleccionar el mejor modelo fue la exactitud (Accuracy) en los datos de test. Esta métrica consiste en el cociente entre los valores verdaderos predichos correctamente más los valores falsos predichos correctamente sobre el total de las observaciones ((TP+TN)/(P+N)). Este valor en resumen mide la cantidad de aciertos que tuvimos al predecir los si una persona era pobre o no.

En la gráfica x se puede observar la exactitud en la base test para cada uno de los modelos que se estimaron. El modelo con mejor rendimiento fue el qda(con un F1 de 0.56), seguido por el Logit-ridge(0.52). Los demás modelos, tuvieron rendimientos iguales o menores a 0.5 (random forest y logit del 0.5 y lda de 0.47). Una posible explicación a que el modelo QDA sea más preciso que el logit, el lda y el logit-ridge es porque el QDA es más flexible en términos de asumir diferentes matrices de covarianza para cada clase, lo que brinda una mayor capacidad de ajuste y permite adapatarse mejor al conjunto de datos. En cuanto al por qué el QDA tiene un mejor rendimiento que el random forest puede ser explicado porque al utilizar 15 variables, el modelo se estaría sobreajustando a los datos de entrenamiento, por lo que la predicción fuera de la muestra es baja.

Figura 3: Comparación precisión



4 Referencias bibliográficas

Castaneda Aguilar, R. A., C. Diaz-Bonilla, T. Fujs, C. Lakner, M. C. Nguyen, M. Viveros, y S. K. T. Baah. 2024. «March 2024 global poverty update from the World Bank: first estimates of global poverty until 2022 from survey data». World Bank Blogs. https://blogs.worldbank.org/en/opendata/march-2024-global-poverty-update-from-the-world-bank--first-esti.

CONPES. 2012. «CONPES 150. Metodologías oficiales y arreglos institucionales para la medición de la pobreza en Colombia». https://colaboracion.dnp.gov.co/CDT/Conpes/Social/150.pdf.

DANE. 2019. «Medición de Pobreza Monetaria y Desigualdad 2018». [Base de datos].`https://microdatos.dane.gov.co/index.php/catalog/608`

Sarmiento-Barbieri, I. 2024. «Dataset Pronlem Set 2». [Base de datos].`https://www.kaggle.com/t/59ef509497064da4b855e6c0148af484`

Yonzan, D. G., N. Mahler, y C. Lakner. 2023. «Poverty is back to pre-COVID levels globally, but not for low-income countries». World Bank Blogs. https://blogs.worldbank.org/en/opendata/poverty-back-pre-covid-levels-globally-not-low-income-countries.