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)

2024-04-14

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, Mahler 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

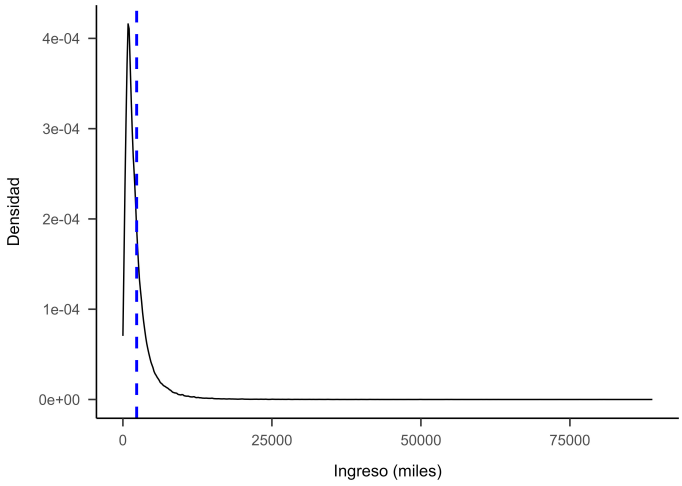
por las bases train y test. Se manipularon las variables categóricas, creando *dummies* para cada categoría, y para continúas reemplazando con 0 valores perdidos. Como la predicción se realizó por hogar, se agregaron las variables como promedios y sumas de personas. Las estadísticas descriptivas de las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

Statistic	N	Mean	St. Dev.	Min	Median	Max
edad	164,960	37.44	16.88	5.67	33.50	102.00
edad_2	164,960	1,890.50	1,486.86	77.25	1,400.50	10,404.00
Genero	164,960	0.53	0.28	0.00	0.50	1.00
estudiante	164,960	0.12	0.19	0.00	0.00	1.00
busca_trabajo	164,960	0.03	0.10	0.00	0.00	1.00
amo_casa	164,960	0.20	0.24	0.00	0.17	1.00
hijos_hogar	164,960	0.20	0.24	0.00	0.17	1.00
primaria	164,960	0.05	0.16	0.00	0.00	1.00
secundaria	164,960	0.16	0.24	0.00	0.00	1.00
media	164,960	0.22	0.28	0.00	0.07	1.00
superior	164,960	0.26	0.33	0.00	0.00	1.00
Ingtot	164,960	2,102,586.00	2,532,552.00	0.00	1,400,000.00	85,833,333.00
Ingtotugarr	164,960	2,307,865.00	2,628,933.00	0.00	1,581,242.00	88,833,333.00
exp_trab_actual	148,947	65.80	91.44	0.00	31.50	948.00
horas_trab_usual	164,960	30.39	18.05	0.00	31.20	130.00
Pobre	164,960	0.20	0.40	0	0	1
Nper	164,960	3.29	1.77	1	3	28
num_menores	164,960	0.75	1.01	0	0	14
num_adulto	164,960	0.32	0.61	0	0	6
eps	164,959	0.93	0.21	0.00	1.00	1.00
rural	164,960	0.91	0.29	0	1	1
num_cuartos	164,960	3.39	1.24	1	3	98
num_cuartos_dormir	164,960	1.99	0.90	1	2	15
Npersug	164,960	3.28	1.77	1	3	28
vivienda_arriendo	164,960	0.39	0.49	0	0	1
vivienda_propia	164,960	0.00	0.00	0	0	0
contributivo	164,960	0.40	0.41	0.00	0.33	1.00
Desempleado	164,960	0.05	0.15	0.00	0.00	1.00
Inactivo	164,960	0.31	0.31	0.00	0.25	1.00
Ocupado	164,960	0.50	0.32	0.00	0.50	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.

2.1 Procesamiento de Datos

2.1.1 Consolidación de base "clean"

2.1.2 Construcción de variables (feature engineering)

2.1.3 Ejercicio con variables DANE

2.2 Análisis Descriptivo

3 Modelos y resultados

3.1 Modelos de Clasificación

3.2 Modelos de Predicción de Ingreso

3.3 Modelos Finales (QDA)

4 Conclusiones

5 Referencias bibliográficas

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