

Big Data y Machine Learning para Economía Aplicada

Taller 3, 2023-02

Link del repositorio en Github: <https://github.com/afdz27/Taller-3.git>

Equipo de competencia en Kaggle: *Equipo Rocket*

1. Introducción

De acuerdo con la ONU (2023), en el mundo hay más de 700 millones de personas que viven en situación de extrema pobreza, comprendida por la incapacidad de satisfacer las necesidades básicas, como son la salud, la educación, el acceso al agua, entre otras. Uno de los parámetros para identificar a las personas que viven en estas condiciones es la pobreza monetaria, que es estimado indirectamente a través de la comparación del ingreso per cápita con el costo monetario de adquirir una canasta de bienes (Cepeda et al., 2019). En Colombia, desde el 2002 la tendencia de la pobreza monetaria ha sido decreciente, sin embargo, además de que la tasa de reducción se ha disminuido, el valor al 2018 era del 27% (Cepeda et al., 2019). Según el Banco Mundial (2020), las zonas con mayor pobreza en Colombia a 2015 eran el Choco, La Guajira y el Cauca con valores de 26.1%, 20.9% y 12.4%, respectivamente.

La predicción de la pobreza es un objetivo crítico para lograr la implementación efectiva de políticas, que permita direccionar los recursos eficientemente a aquellos hogares que están en una situación de pobreza o que están a punto de llegar a esa situación (Li et al., 2022). Dado que los censos nacionales son costosos y complejos logísticamente y que los modelos económicos utilizados hasta el momento dependen de varias suposiciones, los métodos de machine learning se ven como una opción prometedora, especialmente cuando la cantidad de información ha aumentado considerablemente en los últimos años (Pathways for Prosperity Commission, 2023).

El objetivo de este trabajo es clasificar una muestra de hogares en Colombia de si son pobres o no, se presentan algunos métodos de machine learning utilizados para resolver problemas de clasificación y los resultados obtenidos, tomando como fuente de datos la base del DANE correspondiente a la Medición de Pobreza Monetaria y Desigualdad 2018 en Colombia. Estos datos corresponden a encuestas realizadas a hogares del país en la Gran Encuesta Integrada de Hogares - GEIH 2018.

Para este ejercicio se utilizaron cuatro bases de datos, dos correspondientes a muestras de los hogares, una de entrenamiento y otra de prueba, con variables físicas del hogar, del arriendo, de los ingresos por unidad de gasto y de la clasificación realizada por el DANE de la pobreza monetaria (estas dos últimas variables solamente están en la de entrenamiento), y las otras dos correspondientes a individuos dentro de estos hogares, igualmente una de entrenamiento y otra de prueba, con variables sociodemográficas, de aspectos laborales (tipo de ocupación, recibo de subsidios o de bonificaciones, entre otras) y de varios tipos de ingreso (estos últimos solamente en la base de entrenamiento).

Para este problema Set se realizaron en total 43 estimaciones, de las cuales se centró la atención en modelos de predicción de ingreso a través de redes neuronales. El mejor modelo estimado se obtuvo con el enfoque de predicción del ingreso mediante el uso de redes neuronales. Sin embargo, el resultado es muy similar al obtenido con otras especificaciones y modelos, lo que indica que se pueden incorporar otras variables que permitan tener una mejor predicción. En cuanto a los modelos de clasificación de pobres y no pobres el mejor resultado fue obtenido con el modelo Lasso, que obtuvo un accuracy de 0,2123 y un score en kaggle de 0.34. Los demás modelos implementados no superaron el 0.25 como puntaje final. Al comparar los resultados obtenidos por los modelos de clasificación de pobres y no pobres, y la predicción del ingreso se puede observar que los segundos presentaron un mejor resultado tanto a nivel de F train como del score obtenido en la competencia

Data y análisis descriptivo de las variables

Para la limpieza y realizar el correspondiente análisis de datos se realizaron las siguientes acciones: En la base de hogares, se hizo un filtro de las variables buscando aquellas que pudieran aportar a la predicción del ingreso eliminando aquellas que fueran redundantes unas con otras, por ejemplo, en bonificaciones y subsidios se preguntaba si tenía y luego nuevamente se reafirmaba si tenía o no; luego se realizó una limpieza de datos y valores atípicos aplicando el rango

intercuartil de 1,5, dejando los valores de test; en las dos bases las columnas de amortización y arriendo estimado, como tienen incidencia sobre el ingreso de la unidad de gasto, se ajustaron mediante resta para mantener la variable de arriendo estimado que es la que realmente afecta al ingreso corregido por unidad de gasto para mejorar la predicción.

La variable arriendo tenía muchos missing values, porque en la encuesta preguntaba sobre el valor del arriendo, pero para quienes por ejemplo estaban pagando un crédito hipotecario, no tomaba este gasto como un arriendo. Es decir, la variable arriendo estimado buscó corregir este desfase.

Finalmente, sobre la base de individuos, tenía muchos vacíos en variables discretas, que en este caso se imputaron los datos con la moda.

Las bases de datos utilizadas para el ejercicio contienen varias variables de interés que están identificadas con la nomenclatura utilizada por la DIAN. Luego de identificar el tipo de variable y la información contenida, se renombraron aquellas que se consideraron como importantes para la predicción correspondiente. Dado que el enfoque de este ejercicio requiere predecir si un hogar es pobre o no mediante dos caminos, el primero siendo la predicción directa de la variable Pobre y el segundo la predicción del ingreso del hogar, la selección de estas variables debía contemplar aquellas características de los hogares que pudieran influir en esta clasificación, por ejemplo, el lugar en el que están ubicados, el número de habitaciones, tanto habitadas como no habitadas, el tipo de propiedad (arrendada, propia, etc.), las variables proxy del gasto realizado que demanda la propiedad y la cantidad de personas que viven allí; por otro lado, para los individuos, dado que la clasificación de los hogares en términos de pobreza depende del ingreso que conjuntamente se genera en el hogar por los individuos que viven allí, es fundamental mantener las variables que se consideran necesarias para predecir adecuadamente el ingreso de la persona.

A partir de la teoría económica y del primer taller, para las bases de datos de los individuos, se consideraron la educación, el sexo y la edad; adicional a estas, teniendo en consideración que el tipo de ocupación, la formalidad del trabajo, el hecho de recibir subsidios y bonificaciones, el estado laboral, tienen influencia sobre los ingresos, se incluyeron dentro del análisis.

Por otro lado, teniendo en cuenta que desde las bases de datos de los hogares se debe predecir la condición de pobreza y que esta depende del ingreso de los individuos, deben crearse nuevas variables que agrupen características relevantes para la determinación del ingreso; por ejemplo, el sexo de la jefe de hogar o el porcentaje de personas inactivas dentro del hogar, entre otros, pueden tener relevancia en que el hogar pueda tener más o menos ingresos.

En cuanto a las estadísticas descriptivas se tiene lo siguiente:

Las unidades de gasto observadas, están conformadas en su mayor parte por hogares de entre 2 a 4 personas, luego por los hogares de 5 personas y los hogares unipersonales en el train. Para el test, los hogares con más de 6 personas en adelante son poco usuales. Proporcionalmente se encontró que cerca del 22,5% de las unidades de gasto corresponde a los conformados por 3 personas, seguido de los de 2 personas con el 21% de la muestra, y en tercer lugar por los integrados por 4 personas que componen el 20% de la observación.

Es importante resaltar que en el test las unidades de gasto unipersonales son el 15% de la observación, mientras que los que están conformados por 5 integrantes apenas supera el 10%. Las unidades de gasto de más de 6 personas están por debajo del 5% de la muestra.

La variable de número de cuartos, se encontró que para el test el número de cuartos, el 42% tienen 2 cuartos, el 39% 1 cuarto. En el caso del train, los hogares con 2 cuartos corresponden al 40%, con 1 cuarto el 38%











	Unique (#)	Missing (%)	Mean	SD	Min	Median	Max	
Clase	1	0	1.0	0.0	1.0	1.0	1.0	
cuartos	4	0	3.4	0.9	2.0	3.0	5.0	
cuartosHab	5	0	2.0	0.8	1.0	2.0	5.0	
Propiedad	6	0	2.5	1.2	1.0	3.0	6.0	
ArriendoEst	274	0	199196.4	239210.6	0.0	70000.0	1000000.0	
Arriendo	341	0	164569.6	217848.0	0.0	0.0	850000.0	
Nper	7	0	3.3	1.5	1.0	3.0	7.0	
Npersug	7	0	3.3	1.5	1.0	3.0	7.0	
Ing totug	33080	0	1557518.5	1031153.2	0.0	1300000.0	5050000.0	
Ing totugarr	34165	0	1741986.9	1086299.5	0.0	1500000.0	5461044.3	

Ilustración 1 Estadísticas descriptivas de la base de train de hogares











	Unique (#)	Missing (%)	Mean	SD	Min	Median	Max	
Clase	2	0	1.1	0.3	1.0	1.0	2.0	
cuartos	13	0	3.4	1.2	1.0	3.0	18.0	
cuartosHab	10	0	2.0	0.9	1.0	2.0	10.0	
Propiedad	6	0	2.5	1.3	1.0	3.0	6.0	
ArriendoEst	286	0	289392.8	4700373.8	0.0	100000.0	800000000.0	
Arriendo	417	0	160637.3	660063.0	0.0	0.0	150000000.0	
Nper	21	0	3.3	1.8	1.0	3.0	21.0	
Npersug	21	0	3.3	1.8	1.0	3.0	21.0	
Li	288	0	120192.3	7423.5	99544.8	121460.3	131125.6	
Lp	288	0	270984.5	34849.5	167222.5	280645.6	303816.7	

Ilustración 2 Estadísticas descriptivas de la base de test de hogares

Modelo y resultados

Predicción por clasificación de pobres y no pobres

Los modelos de clasificación consisten en usar datos de características de individuos y situaciones para entrenar un modelo predefinido y poder asignar una categoría específica de acuerdo a la información de entrada asociada a los elementos de la muestra. Para este caso, el objetivo es identificar cuando un hogar es pobre o no pobre, de acuerdo a la definición de la pobreza, es decir, el modelo planteado tiene como variable regresada la característica asociada a la pobreza, la cual se

constituye como una variable categorica dummy. Aunque la clasificación se puede hacer para varias categorías, en este caso el ejercicio de predicción se concentra en la variable dicótoma Pobre, para la cual los hogares serán pobres (1) o no pobres (0). A diferencia de la predicción de ingreso, en este ejercicio se usa la información de los hogares para la clasificación.

Dado que para el ejercicio se tienen bases de datos de los individuos que componen el hogar e información de los mismos hogares, para lograr maximizar el beneficio de la información, se seleccionaron variables de las dos bases para crear nuevas características asociadas a cada hogar, partiendo de la información individual consignada en las bases de datos de train y test suministradas por el problema set. En este caso, se identificaron 20 variables adicionales que se podían extraer y que eran relevantes para la estimación de la pobreza (como se puede ver en el archivo llamado Diccionario de variables en la carpeta stores del repositorio de Github). De esas variables, se seleccionaron 10 y se crearon en los dataframes correspondientes, cómo se puede ver en la tabla 1:

Tabla 1 Variables adicionales para modelos de clasificación

No.	Variable	Tipo de variable	Descripción	Justificación
1	jefe_mujer	Categorica dicótoma	Describe si el jefe del hogar es mujer (1) o no lo es (0)	Dado que existe una brecha salarial entre hombres y mujeres y es posible que si el jefe de hogar es mujer los ingresos del mismo sean más bajos, por lo que hay mayor posibilidad de que este hogar se encuentre en pobreza. Además del efecto negativo sobre la experiencia y la educación que experimentan las mujeres el dedicarse al hogar y el cuidado de los hijos
2	edad_jefe_hogar	Numérica discreta	Describe la edad de la persona identificada como jefe de hogar	La edad está asociada a la experiencia y la capacidad productiva de las personas, por lo cual puede influir en el ingreso que perciben los hogares. Si hay una mayor cantidad de personas en la unidad de gasto en edad laboral es más factible que el mismo no se encuentre en la pobreza
3	menores_edad	Numérica discreta	Cantidad de personas menores de 18 años en la unidad de gasto	A mayor número de menores de edad, menor posibilidad de generación de ingresos acumulados e incremento de gastos del hogar como aseo, educación, alimentación, entre otros. Según lo anterior, a mayor número de menores de edad, el hogar se ve obligado a destinar su ingreso total a necesidades básicas, obligando a dejar de lado otro tipo de actividades. Así mismo, limita las posibilidades de destinar una mayor cantidad de recursos a educación de calidad, y por ende al acceso de oportunidades derivadas de la misma
4	ocupación_jefe_hogar	Categorica nominal	Ocupación de la persona identificada como jefe de hogar	Las ocupaciones tienen diferentes remuneraciones, por lo que pueden incidir en el ingreso. Por lo anterior se incluye la variable
5	Educación_jefe_hogar	Categorica ordinal	Máximo nivel educativo alcanzado por el jefe de hogar,	El nivel educativo en Colombia muchas veces permite el acceso a diferentes puestos de trabajo, que tienen remuneraciones diferentes. Por lo anterior, y porque la ocupación es

No.	Variable	Tipo de variable	Descripción	Justificación
			según la clasificación de la base de datos: desde primaria hasta educación superior	determinante del ingreso se creó esta variable. Además, es de especial relevancia la del jefe de hogar, dado que es principalmente el quien responde por las necesidades del hogar.
6	Porcentaje_edad_trabajo	Numérica continua	Cantidad de personas en la unidad de gasto en edad de trabajar (se asume que pueden trabajar los mayores de 18 años)	Indica la posibilidad del hogar para generar ingresos
7	ocupados	Numérica discreta	Cantidad de personas en la unidad de gasto mayores a 18 años que cuentan con trabajo o con fuente de ingreso derivada de la prestación de servicios laborales	Indica la fuerza laboral del hogar y el efecto de la contribución agregada de los miembros de la unidad de gasto para generar ingresos
8	Porcentaje_ocupados	Numérica continua	Porcentaje de personas en edad de trabajar ocupados	Ratio que permite verificar el uso de la totalidad del recurso de la familia para generar ingresos. Un menor número de ocupados significa menores ingresos para la familia
9	maxEducLevel_hogar	Categorica ordinal	Máximo nivel educativo de toda la unidad de gasto	Aunque el jefe de hogar es principalmente quien responde por las necesidades de la familia, puede ser que los hijos tengan un nivel educativo más alto, por lo que esto puede representar mayores ingresos. Un nivel educativo más alto también significa acceso a otros servicios que no pudo tener el jefe de hogar, por lo que sería un indicador de superación de pobreza y mejoramiento de oportunidades
10	personasxhab	Numérica continua	Número de personas por habitación	El número de personas por habitación es un indicador de que tan grande es la residencia del hogar y por ende la capacidad del mismo de acceso a lugares amplios mediante el pago de arriendo o compra de vivienda. A mayor número de personas por habitación para dormir, mayor es la probabilidad de que una familia se encuentre en la pobreza.

Fuente: Elaboración propia

Las variables de la tabla XXX se obtuvieron mediante la iteración de código en R de la base de personas, usando como llave de merge el id de cada hogar.

Una vez finalizada la creación de las variables y se seleccionaron las definitivas, se crearon tres modelos para iniciar el proceso de clasificación. Se seleccionó el modelo #3 debido a que presentaba los mejores resultados, este fue:

$$\text{Pobre} = \beta_0 + \beta_1 \text{personasxhab} + \beta_2 \text{ArriendoEst} + \beta_3 \text{Propiedad} + \beta_4 \text{jefe_mujer} + \beta_5 \text{edad_jefe_hogar} + \beta_6 \text{maxEducLevel_hogar} + \beta_7 \text{ocupacion_jefe_hogar} + \beta_8 \text{menores_edad}$$

En el anterior modelo se agregaron las variables ArriendoEst y Propiedad las cuales reflejan una aproximación al arriendo que se estima puede pagar el hogar y el tipo de vivienda que habitan (arrendada, propia, otro) respectivamente de las bases de datos de hogar. Este modelo consta de 8 variables finales que fueron seleccionadas de acuerdo al desempeño de los modelos corridos. Dentro del modelo hay 4 variables numéricas y 4 categóricas.

Para la clasificación, se usaron modelos logit, lasso, ElasticNet y Lasso con criterio de selección ROC. Sin embargo, teniendo en cuenta que la variable a predecir es dicotoma (1 si es pobre y 0 si es no pobre), la predicción se hace usando la probabilidad de que cada observación sea clasificada como pobre. En todos los métodos se usó el tipo de predicción de probabilidad, en el que para valores mayores a 0.5 se asignaba la categoría de pobre, y para menores a ese valor se asignaba la categoría no pobre (Clasificador de Bayes $c = 0.5$), a excepción de un caso en el que se intentó con un valor más alto, dado el desbalanceo de los clasificados como pobres en la base de train. Se probaron por primera vez todos los métodos, siendo el modelo de Lasso el mejor (modelos Lasso 2, 4 y 5). Los resultados obtenidos se ven en la tabla 2:

Tabla 2. Comparación de modelos de clasificación de pobres y no pobres

Nombre	Hiperparámetro	lambda	TP	FP	TN	FN	Accurac y	Resultado F1 score train	Resultado F1 score test
logit1	c=0.5	-	277 3	179 5	817	1619 1	0,1663	0,23567	0,23
logit2_clasificación	c=0.5		301 4	155 4	803	1620 5	0,1763	0,25341	0,23
lasso_1_clasificación	alfa = 0.01 length=200 c= 0.8	1,0233	586	398 2	774	9264 4	0,3860	0,08128	0,07
lasso_2_clasificación	alfa = 0.01 length=200 c=0.5	1,0232	458 6	0	0	1700 8	0,2123	0,35034	0,34
lasso3_ROC_clasificación	alfa = 0.01 length=200 c=0.5	0,0136	106 6	350 2	436 0	1264 8	0,2514	0,11661	0,22
elastic_net_1	alfa = 0.01 mixture=0.5 c=0.5	0,0002	286 9	169 9	834	1617 1	0,1716	0,24305	0,21
lasso_5_clasificación	alfa = 0.001 length=500 c=0.5	1.0023	456 7	1	0	1700 8	0,2116	0,34938	0,34
RedesClas_01	-	-	-	-	-	-	0,8455	0.5494	No sometido

Fuente: Elaboración propia

Como resultado de los modelos cargados en Kaggle se obtuvo que el mejor modelo fue el lasso_2_clasificación con un score de 0.34. Debido a que Lasso arrojó el mejor resultado, se corrieron más modelos cambiando los parámetros. Se cambiaron parámetros en la grilla, usando un length de 70, 200 y 500 con combinaciones de λ de 0,01 y 0,001. El resultado no varió y el máximo score obtenido fue el del modelo 2. Al comparar con todos los modelos de predicción de ingreso,

estos modelos presentaron mejores scores, por lo que se optó por continuar implementando estos, como se puede ver en la siguiente sección.

Predicción del ingreso

En esta sección la predicción de la clasificación del hogar en pobre o no se realiza mediante una primera predicción del ingreso por individuo i . Este ingreso se agrupa por hogar j , se ajusta por la cantidad de personas que hay en la unidad de gasto y por un arriendo estimado cuando no se está pagando arriendo; luego, se compara la línea de pobreza –estimada de acuerdo con la valoración de la canasta de bienes necesaria para satisfacer ciertas necesidades–, como se muestra a continuación:

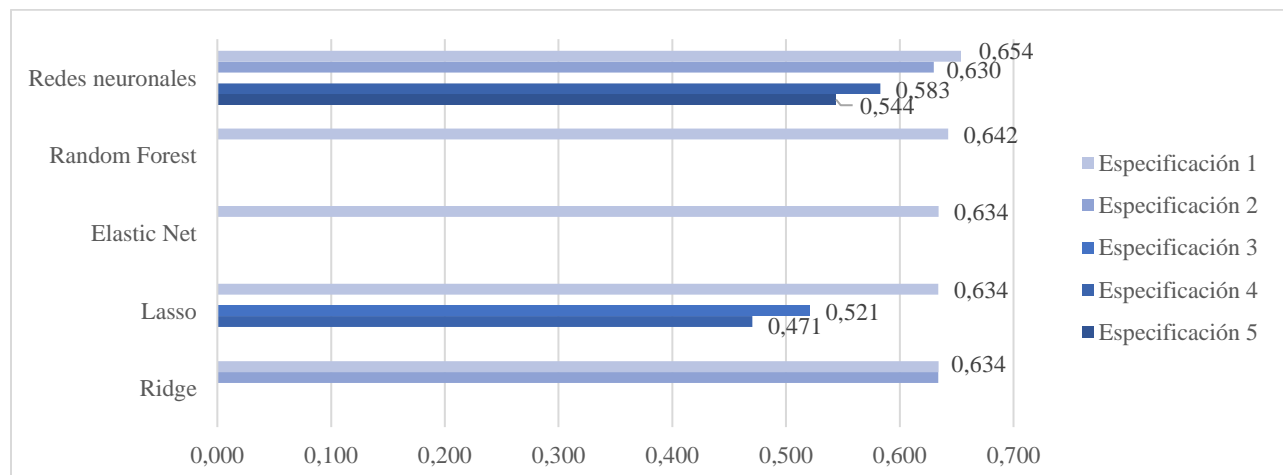
$$\begin{aligned}
 \text{Pobre} &= 1, \text{ si } \frac{\sum_{i=1}^{n \text{ del hogar}} \text{Ingreso}_i + \text{Arriendo Estimado}_j}{N_persug_j} \geq \text{línea de pobreza}_j \\
 \text{Pobre} &= 0, \text{ si } \frac{\sum_{i=1}^{n \text{ del hogar}} \text{Ingreso}_i + \text{Arriendo Estimado}_j}{N_persug_j} < \text{línea de pobreza}_j
 \end{aligned}$$

La predicción del ingreso se realiza utilizando diferentes modelos y especificaciones. Respecto a modelos, se utilizaron modelos de regularización, de *random forest* y de redes; en cuanto a especificaciones se utilizaron las siguientes variantes:

1. Especificación 1: $\text{ingreso} = f(X)$, siendo X todas las variables seleccionadas y descritas en la sección de datos, que son características de los individuos.
2. Especificación 2: $\text{ingreso} = \text{Edad} + \text{Edad}^2 + f(X_1)$, siendo X_1 todas las variables características de los individuos distintas de las descritas explícitamente en la ecuación.
3. Especificación 3: $\log(\text{ingreso}) = f(X)$, siendo X todas las variables seleccionadas y descritas en la sección de datos, que son características de los individuos.
4. Especificación 4: $\log(\text{ingreso}) = \text{Edad} + \text{Edad}^2 + f(X_1)$, siendo X_1 todas las variables características de los individuos distintas de las descritas explícitamente en la ecuación.
5. Especificación 5: $\log(\text{ingreso}) = \text{Edad} + \text{Edad}^2 + \text{sexo} * \text{Edad} + \text{sexo} + f(X_2)$, siendo X_2 todas las variables características de los individuos distintas de las descritas explícitamente en la ecuación.

La evaluación de los modelos se realizó mediante el cálculo del *F1-score* en el conjunto de validación, a partir de una matriz de confusión que se construía con la predicción y con los valores reales del conjunto en cuestión. En la siguiente gráfica se muestran los mejores resultados del *F1-score* para los modelos evaluados:

Figura 1. Mejores resultados F1- score de predicción del parámetro de Pobre con la predicción del ingreso



Fuente: Elaboración propia

Para el detalle de los modelos que se simularon, véase el anexo de este reporte. Como puede verse, se exploraron diferentes valores de hiperparámetros para lograr ver cómo esto afectaba la predicción. Muchos de los modelos no fueron sometidos, teniendo en cuenta que tenían resultados inferiores o similares a algunos que sí fueron cargados a Kaggle.

El mejor modelo para este enfoque fue el de redes neuronales, con la especificación 1. Sin embargo, este resultado es muy similar al obtenido con otras especificaciones y modelos. Así mismo, en Kaggle, no superó el valor de 0.56.

En redes neuronales se desarrollaron una cantidad mayor de variantes que en los otros modelos, teniendo en cuenta la cantidad de hiperparámetros que este tenía. Se identificó que el mejor modelo tenía dos capas intermedias, con tres su desempeño bajaba.

Otro aspecto importante es que, si bien el modelo de *Random Forest* tenía uno de los mejores resultados en F1-score, su valor en Kaggle fue de 0.44. Esto puede deberse a que su valor de *Recall* era muy inferior (0.55) a los de los modelos que tuvieron un valor en Kaggle más alto (entre 0.7 y 0.8).

Tabla 3. Modelos de predicción pobres mediante estimación de ingreso estimados

ID de modelo	Especificación	Modelo	Hiperparámetros	Accuracy	Precision	Recall	F1-score - Validation Data	F1-score - Test Data
1	Especificación 1	Lasso	alfa = 1×10^{-5}	0.85	0.68	0.60	0.63	0.55
2	Especificación 3	Lasso	alfa = 1×10^{-5}	0.71	0.40	0.75	0.52	0.4
3	Especificación 4	Lasso	alfa = 1×10^{-5}	0.65	0.34	0.74	0.47	No sometido
4	Especificación 1	Ridge	alfa = 1×10^{-5}	0.86	0.69	0.59	0.63	0.55
5	Especificación 1	Ridge	alfa = 1×10^{-10}	0.86	0.69	0.59	0.63	No sometido
6	Especificación 1	Elastic Net	alfa = 1×10^{-10} ; mixture = 0.5	0.85	0.68	0.60	0.63	No sometido
7	Especificación 1	Elastic Net	alfa = 1×10^{-10} ; mixture = 0.75	0.85	0.68	0.60	0.63	0.55
8	Especificación 1	Random Forest	mtry = 7; min.node.size = 5	0.87	0.76	0.56	0.64	0.43
9	Especificación 1	Random Forest	mtry = 5; min.node.size = 5	0.86	0.77	0.51	0.61	No sometido
10	Especificación 1	Random Forest	mtry = 5; min.node.size = 5; Imputación de ceros (0) en EdadTrabajo = 0	0.86	0.77	0.51	0.62	No sometido
11	Especificación 1	Redes neuronales	Epocas: 500; Bache: 5112; Capa densa 1: 8; Activación 1: ReLu; Capa densa 2: 0; Activación 2: N/A; Capa densa 3: 0; Activación 3: N/A; Dropout: 0.5	0.84	0.61	0.67	0.64	No sometido
12	Especificación 1	Redes neuronales	Epocas: 500; Bache: 5112; Capa densa 1: 16; Activación 1: ReLu; Capa densa 2: 0; Activación 2: N/A; Capa densa 3: 0; Activación 3: N/A; Dropout: 0.5	0.79	0.21	0.00	0.00	No sometido
13	Especificación 1	Redes neuronales	Epocas: 500; Bache: 5112; Capa densa 1: 8; Activación 1: ReLu; Capa densa 2: 8; Activación 2: ReLu; Capa densa 3: 0; Activación 3: N/A; Dropout: 0.5	0.83	0.59	0.70	0.64	0.56
14	Especificación 1	Redes neuronales	Epocas: 250; Bache: 5112; Capa densa 1: 16; Activación 1: ReLu; Capa densa 2: 8; Activación 2: ReLu; Capa densa 3: 0; Activación 3: N/A; Dropout: 0.5	0.84	0.60	0.71	0.65	0.56

ID de modelo	Especificación	Modelo	Hiperparámetros	Accuracy	Precision	Recall	F1-score - Validation Data	F1-score - Test Data
15	Especificación 1	Redes neuronales	Epocas: 500; Bache: 5112; Capa densa 1: 16; Activación 1: ReLu; Capa densa 2: 8; Activación 2: ReLu; Capa densa 3: 0; Activación 3: N/A; Dropout: 0.5	0.84	0.61	0.68	0.64	No sometido
16	Especificación 1	Redes neuronales	Epocas: 250; Bache: 5112; Capa densa 1: 16; Activación 1: ReLu; Capa densa 2: 16; Activación 2: ReLu; Capa densa 3: 0; Activación 3: N/A; Dropout: 0.5	0.83	0.59	0.67	0.63	No sometido
17	Especificación 1	Redes neuronales	Epocas: 200; Bache: 5112; Capa densa 1: 64; Activación 1: ReLu; Capa densa 2: 32; Activación 2: ReLu; Capa densa 3: 0; Activación 3: N/A; Dropout: 0.5	0.84	0.63	0.63	0.63	No sometido
18	Especificación 1	Redes neuronales	Epocas: 200; Bache: 5112; Capa densa 1: 16; Activación 1: ReLu; Capa densa 2: 8; Activación 2: ReLu; Capa densa 3: 4; Activación 3: ReLu; Dropout: 0.5	0.81	0.53	0.80	0.64	No sometido
19	Especificación 1	Redes neuronales	Epocas: 500; Bache: 5112; Capa densa 1: 16; Activación 1: ReLu; Capa densa 2: 8; Activación 2: ReLu; Capa densa 3: 0; Activación 3: N/A; Dropout: 0.5	0.83	0.58	0.72	0.64	No sometido
20	Especificación 1	Redes neuronales	Epocas: 500; Bache: 10224; Capa densa 1: 16; Activación 1: ReLu; Capa densa 2: 8; Activación 2: ReLu; Capa densa 3: 0; Activación 3: N/A; Dropout: 0.5	0.84	0.60	0.72	0.65	No sometido
21	Especificación 1	Redes neuronales	Epocas: 750; Bache: 10224; Capa densa 1: 16; Activación 1: ReLu; Capa densa 2: 8; Activación 2: ReLu; Capa densa 3: 0; Activación 3: N/A; Dropout: 0.5	0.84	0.60	0.69	0.64	No sometido
22	Especificación 1	Redes neuronales	Epocas: 500; Bache: 5112; Capa densa 1: 16; Activación 1: ReLu; Capa densa 2: 8; Activación 2: ReLu; Capa densa 3: 0; Activación 3: N/A; Dropout: 0.25	0.84	0.63	0.63	0.63	No sometido
23	Especificación 1	Redes neuronales	Epocas: 250; Bache: 2556; Capa densa 1: 16; Activación 1: ReLu; Capa densa 2: 8; Activación 2: ReLu; Capa densa 3: 0; Activación 3: N/A; Dropout: 0.25	0.84	0.62	0.64	0.63	No sometido
24	Especificación 5	Redes neuronales	Epocas: 500; Bache: 5112; Capa densa 1: 16; Activación 1: ReLu; Capa densa 2: 8; Activación 2: ReLu; Capa densa 3: 0; Activación 3: N/A; Dropout: 0.5	0.32	0.24	1.00	0.38	No sometido
25	Especificación 5	Redes neuronales	Epocas: 100; Bache: 5112; Capa densa 1: 64; Activación 1: ReLu; Capa densa 2: 32; Activación 2: ReLu; Capa densa 3: 0; Activación 3: N/A; Dropout: 0.5	0.66	0.38	0.97	0.54	No sometido
26	Especificación 5	Redes neuronales	Epocas: 50; Bache: 5112; Capa densa 1: 64; Activación 1: ReLu; Capa densa 2: 32; Activación 2: ReLu; Capa densa 3: 16; Activación 3: ReLu; Dropout: 0.5	0.47	0.29	0.99	0.44	No sometido
27	Especificación 4	Redes neuronales	Epocas: 50; Bache: 5112; Capa densa 1: 64; Activación 1: ReLu; Capa densa 2: 32; Activación 2: ReLu; Capa densa 3: 0; Activación 3: N/A; Dropout: 0.5	0.66	0.38	0.96	0.55	No sometido
28	Especificación 4	Redes neuronales	Epocas: 50; Bache: 5112; Capa densa 1: 64; Activación 1: Sigmoid; Capa densa 2: 32; Activación 2: ReLu; Capa densa 3: 0; Activación 3: N/A; Dropout: 0.5	0.73	0.43	0.90	0.58	No sometido
29	Especificación 4	Redes neuronales	Epocas: 50; Bache: 5112; Capa densa 1: 64; Activación 1: Sigmoid; Capa densa 2: 32; Activación 2: Sigmoid; Capa densa 3: 0; Activación 3: N/A; Dropout: 0.5	0.65	0.37	0.95	0.53	No sometido
30	Especificación 4	Redes neuronales	Epocas: 50; Bache: 5112; Capa densa 1: 64; Activación 1: ReLu; Capa densa 2: 32; Activación 2: Sigmoid; Capa densa 3: 0; Activación 3: N/A; Dropout: 0.5	0.58	0.32	0.92	0.48	No sometido
31	Especificación 4	Redes neuronales	Epocas: 50; Bache: 5112; Capa densa 1: 64; Activación 1: tanh; Capa densa 2: 32; Activación 2: ReLu; Capa densa 3: 0; Activación 3: N/A; Dropout: 0.5	0.65	0.37	0.95	0.54	No sometido
32	Especificación 2	Redes neuronales	Epocas: 500; Bache: 5112; Capa densa 1: 16; Activación 1: ReLu; Capa densa 2: 8; Activación 2: ReLu; Capa densa 3: 0; Activación 3: N/A; Dropout: 0.5	0.84	0.61	0.69	0.65	0.56
33	Especificación 2	Ridge	alfa = 1×10^{-10}	0.86	0.69	0.59	0.63	No sometido

Fuente: Elaboración propia

Conclusiones

La pobreza es un indicador usado para poder clasificar a personas y hogares y así determinar su dificultad al acceso de bienes y servicios para su diario vivir, poniendo en riesgo sus vidas. Está representa las dificultades que tiene una sociedad de garantizar accesos a servicios básicos para todos sus habitantes y refleja la brecha entre los ingresos que pueden tener todos los habitantes de un determinado territorio. Sin embargo, para todos los países es costosos revisar dicha clasificación

año año, por lo que su predicción es de especial importancia para la formulación de políticas públicas que puedan mejorar la condición de vida de estas personas.

En este problema se evidenció que no es sencillo realizar una estimación demasiado cercana a la realidad. En el desarrollo de este documento se encontró que el mejor modelo estimado se obtuvo con el enfoque de predicción del ingreso mediante el uso de redes neuronales, con la especificación 1. Sin embargo, el resultado obtenido es muy similar al obtenido con otras especificaciones y modelos, lo que indica que se pueden incorporar otras variables que permitan tener una mejor predicción. Al cambiar los parámetros de todos los modelos implementados, se evidenció un aumento en el tiempo de cómputo. Sin embargo, el resultado fue similar, lo cual se comprobó con la estimación del F train y con el resultado obtenido en la aplicación de kaggle.

En cuanto a los modelos de clasificación de pobres y no pobres el mejor resultado fue obtenido con el modelo Lasso, que tuvo punto de corte de probabilidad $c=0.5$, un Alpha de 0,01 y un length para la grilla de 200 (estos valores se incrementaron sin obtener una mejora en el resultado final). Con este modelo se obtuvo un accuracy de 0,2123 y un score en kaggle de 0.34. Los demás modelos implementados no superaron el 0.25 como puntaje final.

Al comparar los resultados obtenidos por los modelos de clasificación de pobres y no pobres, y la predicción del ingreso se puede observar que los segundos presentaron un mejor resultado tanto a nivel de F train como del score obtenido en la competencia, lo cual se puede deber a la selección de las variables, los modelos usados y la relación de pobres de la base de train.

Bibliografía

Cepeda, L., Rivas, G., Álvarez, S., Katherine Rodríguez, R., & Sánchez, W. (2019). *POBREZA MONETARIA Y POBREZA MULTIDIMENSIONAL DEPARTAMENTO NACIONAL DE PLANEACIÓN*.

Li, Q., Yu, S., Échevin, D., & Fan, M. (2022). Is poverty predictable with machine learning? A study of DHS data from Kyrgyzstan. *Socio-Economic Planning Sciences*, 81. <https://doi.org/10.1016/j.seps.2021.101195>

ONU. (2023). *Objetivo 1: Poner fin a la pobreza en todas sus formas en todo el mundo*. <https://www.un.org/sustainabledevelopment/es/poverty/>

Pathways for Prosperity Commission. (2023). *Can machine learning predict poverty?* <https://pathwayscommission.bsg.ox.ac.uk/blog/can-machine-learning-predict-poverty/>

World Bank. (2020). *Global Subnational Atlas of Poverty (GSAP)*. <https://pipmaps.worldbank.org/en/data/datatopics/poverty-portal/poverty-geospatial?dataset=PovertyRate2.15-gsap&zoomLevel=5&lat=20.16336578378857&lng=-83.45214843750001>