Modelo de clasificación de trabajo infantil para Colombia usando aprendizaje de máquinas

Camila Alejandra Velasco Contreras

Jairo Alexander Torres Preciado

Agosto 2 de 2022

Resumen

Las aplicaciones empíricas del aprendizaje de máquinas en las ciencias sociales han venido en aumento en el último tiempo, especialmente en temas de economía. En esa línea, este trabajo presenta un modelo de clasificación para predecir trabajo infantil en Colombia. A partir de datos de la Encuesta Longitudinal Colombiana de la Universidad de los Andes (ELCA) en su versión de 2016, se obtuvieron variables individuales del menor, así como variables sobre sus padres y su hogar que explican el contexto en el que el niño vive. Se entrenaron múltiples modelos de clasificación para predecir si un menor trabaja o no, con el fin de aportar una herramienta que permita a las autoridades la identificación temprana de niños en esta situación, así como de aquellos que están en condiciones de vulnerabilidad y podrían iniciar a trabajar en cualquier momento. Los resultados muestran que el modelo que mejor predice en la muestra de prueba es el logit con cutoff alternativo, el cual arrojó una sensibilidad o recall de 0.667, la más alta entre todos los algoritmos probados.

Palabras clave: Trabajo de menores, aprendizaje de máquinas, modelo de clasificación, Colombia.

JEL: O15, J21, C1, C52

1. Introducción

Se considera que existe trabajo infantil en Colombia cuando niños y adolescentes de menos de 15 años ejercen cualquier actividad laboral, cuando se dedican a oficios del hogar con una intensidad superior a 15 horas semanales, y cuando jóvenes entre 15 y 18 años efectúan trabajos considerados peligrosos o nocivos (ICBF, 2013). Las más recientes cifras del DANE[1] muestran que la tasa de trabajo infantil, durante el último trimestre de 2021, fue de 4.8% en todo el territorio nacional. También se evidencia que los niños tienen mayor participación dentro del mercado laboral en comparación con las niñas, mostrando tasas de 6.4% y 3.2% respectivamente. Así mismo, las actividades en las que más trabajan los menores son agricultura, ganadería, caza, silvicultura y pesca, seguida por comercio y reparación de vehículos e industrias manufactureras.

Si bien la tasa de trabajo infantil ha mostrado una tendencia decreciente en la última década, pasando de 10.3% en 2012 a 4.8% en 2021, esta sigue siendo una problemática de especial relevancia en Colombia por las implicaciones y consecuencias que trae consigo. Se trata de un fenómeno que obliga a hacer sacrificios en el bienestar futuro del menor a cambio de beneficios inmediatos para él y otras personas, lo cual implica que hay una separación entre los beneficios presentes y los costos futuros, al tiempo que los beneficios son asumidos por distintos individuos (Bernal y Cárdenas, 2006). En ese sentido, el trabajo infantil afecta a los niños, toda vez que impide que satisfagan sus necesidades de recreación, trae repercusiones en su salud que pueden permanecer hasta su edad adulta y obstaculiza su educación, lo cual afecta su formación de capital humano y tiene un efecto sobre sus ingresos futuros (Pedraza y Ribero 2006).

Por lo anterior, en este trabajo se estimaron varios modelos de clasificación, usando métodos de aprendizaje de máquinas, que permiten clasificar a los niños en dos categorías, si trabajan o no, con base en variables determinantes del trabajo infantil relacionadas con características del menor y de su hogar, así como del entorno en el cual vive. El modelo que arrojó mejores resultados fue el logit con cutoff alternativo, el cual presentó la métrica más alta de sensibilidad (0.667), es decir, tuvo mayor acierto para identificar a los verdaderos menores que están trabajando.

Esta investigación contribuye a la escasa literatura de aprendizaje de máquinas aplicada al trabajo infantil, en la que se destacan los trabajos de Libaque-Saenz et al. (2018) para Perú y Chocobar (2021) para Argentina y, además, se constituye en una novedad en cuanto a su aplicación en Colombia. Así mismo, es un aporte importante para predecir el trabajo infantil a partir de ciertas variables de los menores, sus hogares y el contexto en el que crecen, lo cual permitirá desarrollar estrategias de política pública tendientes a identificar niños en situación de vulnerabilidad que están trabajando o que podrían verse forzados a hacerlo en el futuro y así, adoptar medidas para evitar que trabajen y abandonen los estudios.

2. Datos

La fuente de información de la que se obtuvieron los datos fue la Encuesta Longitudinal Colombiana de la Universidad de los Andes (ELCA) en su versión de 2016. En particular, se utilizaron los paneles de niños de sector urbano y rural. De aquí se obtuvieron variables referentes a cada uno de los menores tales como “trabajo”, que es la variable de clasificación, y la edad del menor. Además, se construyeron las variables “padre presente” y “madre presente” que señalan si el niño vive con cada padre o no.

[1] GEIH - trabajo infantil (dane.gov.co)

Esta base se complementó con variables de los paneles de personas de sector urbano y rural. De esta última se obtuvieron predictores adicionales relacionados con los menores como el sexo, si estudia o no, si ha migrado en los últimos tres años, y si dejó de estudiar en algún momento durante los últimos tres años.

Adicionalmente, la base se enriqueció con variables tomadas de los paneles de hogares urbanos y rurales que, en general, dan muestra de la situación socioeconómica en la que vive cada niño. Específicamente, se agregaron predictores como la región en donde viven, número de personas que vive en el hogar, el estrato de la vivienda, tipo de vivienda (casa, apartamento o cuarto), así como predictores que dan cuenta de la asistencia que reciben por parte del Estado en forma de ayudas desde entidades como el Instituto Colombiano de Bienestar Familiar y el SENA, o programas de transferencias monetarias como Familias en Acción, Programa Adulto Mayor (actualmente Colombia Mayor), Jóvenes en Acción y ayuda a desplazados por conflicto armado. Por último, se construyó la variable de ingreso total del hogar (ing_tot_kn) a partir de la suma de los ingresos por trabajo agrario y no agrario, pensiones y arriendos.

Luego de ajustar el panel con cada uno de los predictores antes mencionados, borrando observaciones que no tenían información completa e imputando aquellas que tenían algunos datos faltantes, se obtuvo un panel definitivo con 2351 observaciones, el cual se utilizó para entrenar y probar los modelos. La Tabla 1 muestra las estadísticas descriptivas de la base de datos. Se observa que la proporción de niños que trabaja en la muestra es de 4%, cifra similar al dato más reciente que presentó el DANE de 4.8% para todo el país. Así mismo, se evidencia que los niños trabajan en mayor proporción que las niñas, que los menores que no estudian son minoría, apenas 26, que corresponden a 1.1% y que la mayor parte de ellos no ha dejado de estudiar ni ha migrado en los últimos tres años.

En cuanto a los hogares, el promedio de personas que vive en cada uno es 5 y la mayor parte de ellos vive en casas estrato 1, 2 o 3. El ingreso promedio de cada hogar es 1,393,079 pesos. En su mayoría, los integrantes de estos hogares son beneficiarios del programa Familias en Acción, pero no lo son de los programas Adulto Mayor y Jóvenes en Acción, no han recibido ayuda a desplazados ni han participado en programas de formación del SENA o programas de asistencia del ICBF para menores.

Tabla 1. Estadísticas descriptivas

Característica	N = 2,351[1]
Niño trabaja	
Si	95 (4.0%)
No	2,256 (96%)
Edad Niño	7.61 (1.13)
Sexo niño	
Hombre	1,202 (51%)
Mujer	1,149 (49%)
Niño estudia	
Sí	2,325 (99%)
No	26 (1.1%)

Migró en los ult. 3 años	
Sí	1 (<0.1%)
No	2,350 (100%)
Niño dejó de estudiar	
Sí	99 (4.2%)
No	2,252 (96%)
No. personas en el hogar	5.34 (2.91)
Tipo vivienda	
Casa	1,883 (80%)
Apartamento	424 (18%)
Cuarto	44 (1.9%)
Estrato	
1	29 (1.2%)
2	1,140 (48%)
3	919 (39%)
4	231 (9.8%)
5	28 (1.2%)
6	4 (0.2%)
Familias en acción (ult. año)	
Si	1,265 (54%)
No	1,086 (46%)
Programa adulto mayor	
Sí	254 (11%)
No	2,097 (89%)
Hogar beneficiario ICBF (ult.año)	
Sí	323 (14%)
No	2,028 (86%)
Hogar beneficiario Sena (ult.año)	
Sí	122 (5.2%)
No	2,229 (95%)
Jóvenes en acción (ult. año)	
Si	23 (1.0%)

No	2,328 (99%)
Recibió ayudas desplazados (ult. año)	
Sí	82 (3.5%)
No	2,269 (97%)
Padre vive en el hogar	
Sí	819 (35%)
No	1,532 (65%)
Madre vive en el hogar	
Sí	172 (7.3%)
No	2,179 (93%)
Ingreso total del hogar	1,393,079 (1,482,425)

[1]Promedio (Est.Desv)/ Frecuencia(%)

3. Modelo

El modelo seleccionado fue un logit, teniendo como variables explicativas el sexo y la edad del menor, si el niño estudia o no, el número de personas en el hogar, el tipo de vivienda en la que vive la familia, el estrato, si el padre y la madre viven en el hogar, el ingreso total del hogar, la región del país en la que está ubicado y si la familia fue beneficiaria en el último año de familias en acción, jóvenes en acción, ICBF, SENA, programa Adulto Mayor o ayudas a desplazados. Pese a que la literatura encuentra gran relación entre la educación de los padres y el riesgo de trabajo infantil, esta variable no pudo ser incluida debido a las pocas observaciones que se tenían (menos del 10% de la muestra). Dado que la muestra está altamente desbalanceada y sólo el 4% de los casos el niño trabaja, se estableció un punto de corte alternativo de probabilidades predichas que determinan la predicción del evento. El punto de corte es 0.033, es decir, que los niños con probabilidad por encima de este umbral se clasifican como si trabajaran. Este nuevo punto de corte es aquel que minimiza la distancia al modelo perfecto según la curva de ROC (100% de sensibilidad y especificidad).

Se escogió este modelo por encima de otros 6 modelos, dado que arrojó la sensibilidad más alta, del 67%. En este caso, es de mayor interés la correcta predicción de los niños que trabajan respecto a los no trabajan, por lo cual se considera que la sensibilidad es la métrica apropiada para comparar los distintos modelos. Los otros modelos estimados fueron:

- Logit con muestra balanceada (up-sample)
- Elastic net con punto de corte alternativo (0.036). Alfa=0.1 y lambda= 0.0053, escogidos con validación cruzada.
- Elastic net con muestra balanceada (up-sample). Alfa=0.55 y lambda=0.0263, escogidos con validación cruzada.
- Árbol simple: 8 nodos terminales. Variables seleccionadas por el árbol: Región, padre presente, sexo del niño, edad del niño e ingreso total del hogar.
- Árbol simple con up-sample: 16 nodos terminales. Variables seleccionadas por el árbol: Región, programa de adulto mayor, sexo del niño, ingreso total del hogar, edad del niño, estrato, personas en el hogar, ayuda desplazados e icbf,

- Árbol con up-sample y prune: Selección de número óptimo de nodos terminales con validación cruzada. 14 nodos terminales. Mismas variables del anterior árbol.
- Random Forest: Muestra balanceada. Variables más importantes: Region, ingreso total y sexo.

A continuación, se muestra la Tabla 2 con las principales métricas de comparación entre modelos:

Tabla 2. Métricas de comparación de modelos

Modelo	Precisión	Sensibilidad	Especificidad
Logit con cutoff alternativo	0,641	**0,667**	0,640
Elastic net cut-off alternativo	0,652	0,611	0,653
Logit con upsample	0,720	0,500	0,729
Elastic net con upsample	0,722	0,611	0,727
Árbol simple	0,953	NA	0,953
Árbol con upsample	0,524	0,065	0,970
Árbol con upsample y prune	0,549	0,068	0,971
Random Forest	0,906	0,095	0,956

4. Resultados

Con el modelo seleccionado se obtuvo la siguiente matriz de confusión, la cual evidencia que este algoritmo predice correctamente a los verdaderos positivos (niños que sí trabajan)

Tabla 3. Matriz de confusión Modelo Logit

		Referencia	
		Sí trabaja	**No trabaja**
Predicción	**Sí trabaja**	12	162
	No trabaja	6	288

En el Anexo 1 se muestran los resultados del modelo escogido,

5. Conclusiones y recomendaciones

En este estudio se utilizaron datos de la ELCA 2016 para entrenar distintos modelos que permitieran predecir trabajo infantil. Esta problemática es de especial interés en Colombia, ya que, si bien las cifras de trabajo en menores se han reducido, aún son muchos los niños que se ven obligados a trabajar para sobrevivir y ayudar a sus familias. Después de estimar siete modelos, se encontró que el que mejor predice la variable de trabajo infantil es el logit con muestra balanceada y punto de corte de 0.033, el cual arrojó una sensibilidad de 0.667. Esta métrica de evaluación del modelo es la más relevante en este caso, puesto que indica el número de niños trabajadores que fueron correctamente identificados (verdaderos positivos), lo cual es el objetivo del modelo de clasificación.

Aunque una de las limitaciones de esta investigación es el pequeño tamaño de la muestra, se puede concluir que el modelo escogido funciona bien, toda vez que predice correctamente la mayoría de los niños que trabajan. Esto sugiere que se puede entrenar el modelo con muestras más grandes que tengan variables similares relacionadas con el niño, su hogar y el contexto en el que vive, tales como el censo general, o la base de datos del SISBEN. A su vez, los resultados de las predicciones servirán para que las autoridades competentes atiendan a los menores en esta situación y les garanticen todos sus derechos o para prevenir que los niños en situación de vulnerabilidad caigan en el trabajo infantil.

6. Código y base de datos

Código y base de datos disponibles en: camivel/Trabajo-Final (github.com)

Referencias

Bernal, R. y Cárdenas, M. (2006). Trabajo Infantil en Colombia. Fedesarrollo.

Chcocobar, B. (2021). Prediciendo trabajo infantil: comparación de técnicas de econometría tradicional y machine learning. [Tesis de maestría]. Universidad de San Andrés. Buenos Aires.

ICBF (2013). Una doble mirada al trabajo infantil en Colombia. Observatorio del Bienestar de la Niñez. Boletín Especial No. 10.

Libaque-Saenz, C.F., Lazo, J., Lopez-Yucra, K.G., Bravo, E.R. (2018). Could Machine Learning Improve the Prediction of Child Labor in Peru? In: Lossio-Ventura, J., Alatrista-Salas, H. (eds) Information Management and Big Data. SIMBig 2017. Communications in Computer and Information Science, vol 795. Springer, Cham.

Pedraza, A. y Ribero, R. (2006). El trabajo infantil y juvenil en Colombia y algunas de sus consecuencias claves. Revista Latinoamericana de Ciencias Sociales, Niñez y Juventud. Vol. 4, Nº 1.

Anexos

Anexo 1. Modelo Final Logit

Characteristic	log(OR)[1]	95% CI[1]	p-value
Edad_niño	-0.26	-0.52, 0.00	0.054
Sexo mujer	0.29	0.03, 0.55	0.032
Estudia No	1.8	-289, 293	>0.9
Migró_ult3Sí	0.07	-321, 321	>0.9
dejo_estudiarSí	0.20	0.00, 0.40	0.052
RegionLbOriental	-0.10	-0.56, 0.36	0.7
RegionLbCentral	-0.24	-0.62, 0.15	0.2
RegionLbPacífica	-0.07	-0.52, 0.37	0.7
RegionLbBogotá	-0.27	-0.66, 0.12	0.2
Personas Hogar	0.60	0.09, 1.1	0.021
tipo_viviendaApartamento	0.60	0.12, 1.1	0.014
tipo_viviendaCuarto	-0.06	-0.27, 0.14	0.5
Estrato 1	0.21	-0.85, 1.3	0.7
Estrato 2	0.00	-1.0, 1.0	>0.9
Estrato 3	0.16	-0.57, 0.90	0.7
Estrato 4	1.7	-285, 289	>0.9
Estrato5	0.52	-314, 315	>0.9
Familias_Accion NO	0.02	-0.27, 0.32	0.9
prg_adultomayor NO	-0.13	-0.46, 0.20	0.4
Icbf NO	0.24	0.00, 0.47	0.046
Sena NO	-0.09	-0.42, 0.24	0.6
jovenes_accion NO	-1.3	-273, 270	>0.9
ayu_desplazados NO	-2.7	-295, 290	>0.9
padre_presente No	-0.34	-0.66, -0.03	0.034
madre_presente No	-0.09	-0.47, 0.29	0.6
Ingreso total	-0.15	-0.38, 0.09	0.2

[1]OR = Odds Ratio, CI = Confidence Interval