Universidad de San Andrés
Departamento de Economía
Maestría en Economía

Prediciendo trabajo infantil: comparación de técnicas de econometría tradicional y machine learning

Emilia Belén CHOCOBAR
DNI: 39.782.674
Mentor: Walter SOSA ESCUDERO

Buenos Aires, Argentina
16 de Diciembre, 2021

Tesis Maestría en Economía de

Emilia Belén CHOCOBAR autor

Prediciendo trabajo infantil: comparación de técnicas de econometría tradicional y machine learning

Resumen

La literatura actual sobre trabajo infantil es amplia, y en su gran mayoría, los autores buscan determinar sus causas y determinantes. No obstante, estos trabajos solamente utilizan modelos de la econometría tradicional. Si el interés consiste en predecir trabajo infantil generalmente se usan modelos como logit o probit. Existen otros modelos de la literatura de machine learning que también son útiles para predecir, sin embargo el uso de estas técnicas en este campo permanece casi sin explorar.

Este estudio busca comparar el desempeño de distintos modelos para predecir trabajo infantil fuera de la muestra de entrenamiento. Los modelos utilizados son logit, random forest, gradient boosted model, extrem grandient boost, ridge y lasso. A su vez, se puede obtener información sobre cuáles son las variables más relevantes a la hora de predecir si un niño, niña o adolescente trabaja. Saber cuáles son las variables más relevantes y cuál modelo cumple mejor el objetivo de predecir trabajo infantil es de suma importancia y utilidad, esto permitirá orientar de manera más efectiva los recursos escasos de los responsables de políticas y de los distintos organismos internacionales.

Palabras claves: machine learning, modelos econométricos, predecir, trabajo infantil.

Códigos JEL: C5, C13, J13.

Predicting child labor: comparison of traditional econometric and machine learning techniques

Abstract

Child labor literature is extensive, and most authors seek to determine it's causes and determinants. However, these papers only use traditional econometrics models. If the interest consists in predicting child labor, models such as logit or probit are generally used. There are other models from the machine learning literature that are also useful for prediction, however the use of these techniques in this field remains largely unexplored.

I compare the performance of different models to predict child labor outside the training sample. The models used are logit, random forest, gradient boosted model, extrem grandient boost, ridge, and lasso. I can know which are the most relevant variables when predicting whether a child or adolescent works, also we can know which model best predictis child labor. This is extremely important and useful as it will allow the scarce resources of policy makers and different international organizations to be directed more effectively.

Keywords: machine learning, econometric models, predict, child labor.

JEL codes: C5, C13, J13.

1. Introducción

El trabajo infantil es un problema crítico, tanto en los países desarrollados como en aquellos en vía de desarrollo. El motivo principal de esta preocupación es su efecto negativo sobre el capital humano y el desarrollo económico de un país (Hanushek, 2013).

Existe una extensa literatura que busca identificar cuáles son sus causas y sus principales determinantes. La explicación más común es la que asume que el trabajo infantil es inevitable en hogares de bajo nivel económico, ya que no pueden sobrevivir sin el aporte salarial de los niños/jóvenes y es por esto que la educación y otras actividades son consideradas como bienes de lujo (Basu y Van, 1998; Basu y Tzannatos, 2003). Por su parte, Emerson y de Souza (2000) encuentran que el trabajo infantil persiste a lo largo de las generaciones, esto quiere decir que los padres que trabajaron en algún momento de su niñez tienen mayor probabilidad de hacer trabajar a sus hijos. Edmonds y Turk (2003), encontraron que hogares vietnamitas con negocio propio son más propensos a tener niños en el mercado laboral. Bhalotra y Heady (2003) afirman que el tamaño de la propiedad (también considerado como un indicativo de riqueza) está negativamente relacionado con el trabajo infantil, al menos para los casos de Ghana y Pakistán.

Otra corriente de investigación hace hincapié en las políticas públicas empleadas por los gobiernos. Hiraoka (1997) y Post (2002) descubrieron que las diferencias entre las tasas de escolaridad y las tasas de trabajo infantil de América Latina y Asia son consecuencia de las diferencias en las políticas educativas y las leyes nacionales. Weiner (1991) encuentra que las diferencias de las tasas de trabajo infantil están vinculadas a las creencias sobre educación que tienen los sectores económicos más altos y las coaliciones políticas. Por su parte, Ranjan (2001), Jafarey y Lahiri (1999) encuentran que el trabajo infantil es consecuencia de las restricciones al crédito.

Para Argentina, Torre (2008) usa datos de la Encuesta de Niños, Niñas y Adolescentes (EANNA) para identificar las principales variables que determinan el trabajo infantil. Utiliza el modelo logit multinomial y las variables que incluye son la edad, género y educación, zona geográfica, características de la vivienda, acceso a servicios y tamaño del hogar. Todas las variables mencionadas resultaron significativas en el modelo y tienen el signo esperado de acuerdo a la teoría, con excepción de acceso al agua, la única variable que no fue estadísticamente significativa. Waisgrais (2007) utiliza la misma base de datos y el mismo modelo. El autor encuentra una relación negativa entre el trabajo infantil y el ingreso familiar. Otro hallazgo importante es que no hay diferencias de trabajo infantil entre aquellos hogares que se encuentran bajo la línea de pobreza y aquellos que no. Este trabajo también verifica que se cumple el axioma del lujo en Argentina, ya que los datos sugieren que los hogares recurren al trabajo infantil con el fin de suavizar el ingreso familiar.

Estos y tantos otros trabajos más analizan los efectos causales y los determinantes del

trabajo infantil con técnicas de la econometría tradicional. Estas técnicas permiten alcanzar eficazmente los objetivos propuestos en cada una de las investigaciones. Sin embargo, si lo que nos interesa es la capacidad predictiva, en economía se suele hacer referencia exclusivamente a modelos tradicionales como logit o probit. Los modelos de la literatura de machine learning juegan un rol muy importante como alternativa viable, ya que pueden mejorar altamente la capacidad predictiva al explotar las relaciones no lineales entre las variables (Mullainathan y Spiess, 2017), sobre todo cuando se cuenta con una extensa base de datos. Sin embargo, el uso de estas técnicas en este campo permanece casi inexplorado. Se puede nombrar dos trabajos relevantes para la Latinoamérica, como es el de Lazo et al. (2017), quien utiliza datos de Perú para comparar la capacidad predictiva del modelo logit y redes neuronales, y en este caso los resultados sugieren que redes neuronales tienen mejor performance. Por otra parte, Rodrigues, Prata y Silva (2015) mediante técnicas de minería de datos analiza base de datos sociales de Brasil para descubrir patrones válidos para estudiar trabajo infantil. Ambos autores coinciden que en los factores mas relevantes son los relacionados con ingresos, sexo, transferencias que recibe el hogar, etc.

Poder predecir los niños, niñas y adolescentes (NNyA) con mayores probabilidades de caer en el trabajo infantil es de suma importancia para los gobiernos, hacedores de políticas y demás organismos internacionales. Contar con información sobre cuáles son las características más relevantes puede orientar sus recursos de una manera mucho más efectiva y eficiente en pos de su erradicación. Por lo tanto, el objetivo de este trabajo es comparar distintos modelos que permiten predecir fuera de la muestra de entrenamiento a los NNyA que trabajan, determinar las ventajas y desventajas de su aplicación y conocer las variables más relevantes incluidas en cada uno.

Siguiendo la literatura argentina sobre trabajo infantil, este trabajo hace uso de la Encuesta de Actividades de Niños, Niñas y Adolescentes (EANNA) realizada en 2016-2017. Esta base de datos proporciona información valiosa acerca de las distintas actividades y características socio económicas de la población objetivo, como así también del hogar y miembros del hogar con los que habitan. Debido a la numerosa cantidad de variables y a la multicolinealidad presente en los datos, esta base constituye un marco adecuado para la aplicación de distintos métodos predictivos.

El trabajo se estructura de la siguiente manera, en la Sección 2 se describe la base de datos que se utiliza en este estudio (EANNA 2016-2017), resaltando tanto las limitaciones como las ventajas de utilizarla. Además, se incluye una breve caracterización de la situación socioeconómica de los NNyA. En la Sección 3, se da a conocer las distintas técnicas empleadas con el objetivo de predecir trabajo infantil y además se comparan las ventajas y desventajas de su utilización. En la sección siguiente se muestran los resultados y el criterio tomado para poder comparar la performance de las técnicas. Finalmente, en la Sección 5 se presentan las

principales conclusiones y lecciones aprendidas en el trabajo.

2. Datos

Los datos que se utilizan para predecir el trabajo infantil provienen de la Encuesta de Actividades de Niños, Niñas y Adolescentes (EANNA) 2016-2017 realizada en Argentina por muestreo probabilístico, tanto en zonas urbanas como rurales.

La encuesta está conformada por tres cuestionarios:

- Cuestionario 1: registra las características estructurales de la vivienda, como por ejemplo materiales con las cuales fue construida la vivienda, acceso a distintos servicios, como luz, agua, etc.

- Cuestionario 2: recoge información sobre las características sociodemográficas de todos los miembros del hogar, por ejemplo, nivel educativo, categoría ocupacional, ingresos. También incluye información sobre si los miembros son beneficiarios de programas de transferencias de ingresos, como por ejemplo la Asignación Universal por Hijo para la protección social (AUH).

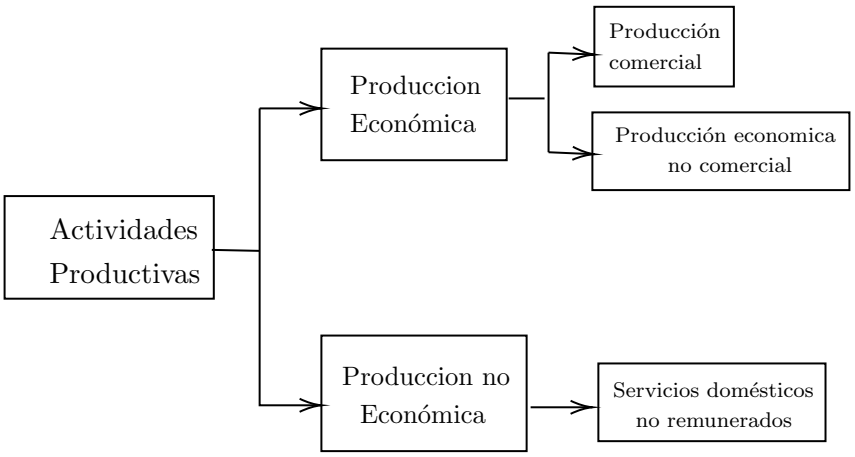
- Cuestionario 3: condensa toda la información sobre los NNyA de 5 a 17 años miembros del hogar. Se pregunta sobre la asistencia a la escuela, las trayectorias educativas, las actividades que realizan en el tiempo libre, y también aquellas de carácter productivo.

El cuestionario 3 contiene información valiosa para poder identificar a aquellos NNyA que trabajan. Sin embargo, es necesario aclarar qué definición se adopta para clasificar a NNyA que trabajan. Cierto es que la conceptualización del término trabajo infantil es sumamente compleja de abordar y esto se debe principalmente a la falta de unicidad en su definición y por la heterogeneidad y multiplicidad que lo caracteriza.

Para la Organización Internacional del Trabajo (OIT) el trabajo infantil es "toda forma de trabajo que priva a los niños/jóvenes de su niñez, de su potencial y de su dignidad, es decir como un término englobante de cualquier tipo de actividad laboral cuya realización es nociva para el desarrollo físico y mental de un infante". Para el INDEC trabajo infantil refiere al conjunto de actividades productivas desarrolladas por los NNyA de 5 a 17 años, aunque la ley argentina establezca que personas de 16 y 17 años pueden trabajar en condiciones particulares. Se entiende por actividades productivas a todo tipo de trabajo en el mercado, actividades de autoconsumo y la ocupación intensiva de tareas domésticas.

Teniendo en cuenta la Figura 1, el INDEC considera como actividades productivas por parte de los NNyA a actividades tanto de producción económica, por ejemplo aquellas con fin de comercialización en el mercado y la elaboración de bienes de uso propio; como a

Figura 1: Clasificación de las actividades NNyA.



Fuente: INDEC

las actividades no económicas que son por ejemplo, las tareas domésticas y personales no remuneradas realizadas de manera intensiva para el consumo del hogar.

En base a esta clasificación, la EANNA 2016-2017, incluye cuatro variables binarias que permiten predecir trabajo infantil:

- Mercado: actividades laborales que producen bienes y/o servicios con valor en el mercado. También incluye toda actividad realizada para terceros a cambio de algún tipo de beneficio, ya sea monetario o material.
- Autoconsumo: comprende las actividades de producción de bienes primarios para el consumo del hogar.
- Actividad doméstica intensiva: abarca todas aquellas tareas intensivas en el hogar que imposibilita la asistencia o rendimiento adecuado en la escuela.
- Trabaja: adopta valor 1 si realiza al menos una de las actividades mencionadas anteriormente, de lo contrario, 0.

Teniendo en cuenta todo lo descrito, los microdatos disponibles son ideales para este tipo de trabajo. Puntualmente, la variable "Trabaja" es una variable binaria que representa 0 si el NNyA no trabajó en el último tiempo, y 1 en caso contrario. Por lo tanto, esta es la variable dependiente que se utilizará dentro de los distintos modelos para predecir trabajo infantil, misma variable empleada por Torre (2008) y Waisgrais (2007). Las variables independientes incorporadas al modelo siguen el mismo criterio adoptado por dichos autores y corresponden a características estructurales del hogar, información socio económicas del jefe y del NNyA, características escolares y tareas extra curriculares que realiza el NNyA. No se utilizaron

todas las variables porque muchas de ellas presentan gran cantidad de valores faltantes. La selección de variables a incluir en el modelo se realiza en base a cantidad de valores faltantes, posibilidad de imputar valores y cuestionarios respondidos por los NNyA. Finalmente, los modelos incorporan 12 variables características de la vivienda, 28 variables referentes al jefe del hogar y 8 sobre las características del NNyA, con un total de 16.115 observaciones[1].

En este trabajo sólo se emplean los datos de las zonas urbanas. Esta decisión se toma en base a que existen diferencias estructurales entre zonas urbanas y rurales. Por ejemplo, es mucho más común observar a NNyA realizar actividades productivas y de autoconsumo en las zonas rurales por costumbre o tradición familiar, mientras que en las zonas urbanas no es tan frecuente. Esto lo demostraron Emerson y de Souza (2000), quienes observan que el trabajo infantil se perpetua a lo largo de las generaciones. Y dado, que la base de datos no cuenta con información sobre si el NNyA trabaja por costumbre o tradición, no se puede aislar su relevancia, y al unir ambas bases, rural y urbana, en una sola y predecir con el mismo modelo, reduciría la eficacia predictiva en la submuestra urbana de testeo.

Es necesario agregar que la proporción de niños trabajando en las zonas urbanas es bajo. Entre los NyN de 5 a 15 años, el 8,4 % realiza al menos una actividad productiva. El 2,6 % solamente realiza actividades para el mercado, 1,3 % actividades para el autoconsumo y 3,5 % actividad doméstica intensa (Gráfico.9). Si bien, en el total de NyN, las mujeres y varones se distribuyen de manera casi equitativa (49 % de niñas y 51 % de niños), al interior de cada una de las actividades se observan diferencias tanto en las actividades dirigidas al mercado como al autoconsumo, se evidencia la mayor participación de los niños (62,8 % y 82,6 % de los casos, respectivamente) en detrimento de la colaboración en tareas domésticas (42,8 %) (Gráfico.10). Se observa que en el NOA se registran los niveles de trabajo infantil más elevados (11,5 %), seguido por GBA (9,4 %) y el NEA (9,3 %). En la Patagonia, en cambio, se verifica la incidencia más baja, con una tasa que alcanza el 6,0 % (Gráfico.11).

En cuanto a los adolescentes de 15 y 17 años, el 30 % realiza al menos una actividad productiva, el 13,3 % solo actividad para el mercado, 2,6 % solo actividades de autoconsumo y 9 % actividades domésticas intensivas (Gráfico.12). Al igual que lo señalado para los NyN, tanto en las actividades dirigidas al mercado como al autoconsumo, se evidencia una mayor presencia relativa de los adolescentes varones (65,2 % y 73,9 % de los casos, respectivamente), en detrimento de su participación en tareas domésticas (35,0 %)(Gráfico.13).

Otros datos de interés (y que serán útiles al observar la sección de resultados) son los referentes al ingreso de la ocupación principal, edad del jefe del hogar y del NNyA, condición del hogar y nivel educativo del hogar. Todos estos datos pueden visualizarse en Cuadro.1,

[1]En el Anexo se describen las variables utilizadas. El cuestionario 3 incluye un set de preguntas sobre actividades que realiza el NNyA, en ese caso esas variables fueron útiles para determinar si trabaja o no, pero no fueron incorporadas en el modelo

Gráfico.2 y Gráfico.3.

Cuadro 1: Variables de interés

	No trabaja				Trabaja			
	min	max	promedio	std	min	max	promedio	std
ingreso ocupacion principal	140	250001	11804	9939	100	250000	9023	10627
edad jefe del hogar	15	98	44	12	15	92	45	12
edad del NNyA	1	17	11	4	5	17	14	3

Figura 2: Condiciones del hogar. Total urbano

		Total hogares con NNyA	Hogares con al menos un NNyA que trabaja	Hogares con NNyA que no trabajan
			%	
Paredes	**Total**	**100,0**	**100,0**	**100,0**
	Deficiente	3,8	5,1	3,5
	No deficiente	96,2	94,9	96,5
Pisos	**Total**	**100,0**	**100,0**	**100,0**
	Deficiente	0,8	1,7	0,6
	No deficiente	99,2	98,3	99,4

Fuente: EANNA Urbana (2016/2017), INDEC y MTEySS.

Figura 3: clima educativo. Total urbano

Clima educativo	Total hogares con NNyA	Hogares con al menos un NNyA que trabaja	Hogares con NNyA que no trabajan
		%	
Total	**100,0**	**100,0**	**100,0**
Bajo	51,7	68,2	48,7
Medio	31,8	23,0	33,4
Alto	16,5	8,8	17,9

Fuente: EANNA Urbana (2016/2017), INDEC y MTEySS.

3. Metodología

Dado que el fin de este estudio es predecir el trabajo infantil fuera de la muestra de entrenamiento se utilizan distintos modelos que permiten predecir variables binarias. De la literatura econométrica tradicional, se emplea el modelo logit, de la literatura de machine

learning, ridge y lasso, y explotando las relaciones entre las distintas variables de la extensa base de datos se emplean tres tipos de árboles, random forest y gradient boosted model (GBM) y extreme gradient boost (XGBoost). En un principio nada indicaría que un determinado modelo pueda ser mejor que otro. Finalmente, se compararán la performance de los distintos modelos con el objetivo de saber cuál es el que predice mejor en la pequeña submuestra de NNyA que trabajan.

3.1. Logit

Los modelos logit son de respuesta binaria y se utilizan generalmente para calcular la probabilidad de respuesta, dado un conjunto de variables predictoras. El resultado del modelo es la estimación de la probabilidad que tiene un nuevo individuo de pertenecer a un grupo o a otro. Además, al tratarse de un análisis de clasificación, también es posible identificar las variables más importantes que explican las diferencias entre grupos. Este modelo, además, evita las limitaciones del modelo de probabilidad lineal.

$$p_i = P(y = 1/x) = F(\beta_0 + \beta_1 x_1 + \beta_2 x_2 + ... + \beta_k x_k) = F(\boldsymbol{x\beta}) \tag{1}$$

donde F es una función que asume valores entre cero y uno, lo que asegura que las probabilidades de respuesta estimada estén dentro de dicho rango. La función F es la función logística, cuya formula es la siguiente:

$$F(x\beta) = \Lambda(z) = \frac{e^{x\beta}}{1 + e^{x\beta}} \tag{2}$$

que está entre cero y uno para todos los números reales. Esta es la función de distribución acumulada (FDA) para una variable aleatoria logística estándar. Esta función se comporta de la siguiente manera: F(z) $\to$ 0 a medida que z $\to -\infty$, y F(z)$\to$1 a medida que z$\to \infty$.

El estimador $\hat{\beta}$ se estima por máxima verosimilitud y la predicción es:

$$\hat{p_i} = P(y = 1/x) = \frac{e^{x\hat{\beta}}}{1 + e^{x\hat{\beta}}} \tag{3}$$

De acuerdo al clasificador de Bayes se elige el estado más probable minimiza el riesgo esperado.

$$\hat{Y_i} = 1[\hat{p_i} \geq 0,5] \tag{4}$$

Es decir, con un modelo logit se puede encontrar la probabilidad de ocurrencia del trabajo infantil y a partir del clasificador de Bayes se puede determinar si el evento sucedería o no. También es posible determinar cuáles son las variables estadísticamente significativas para el modelo, pero esto no implica que el modelo seleccione e incorpore solamente algunas variables, como si lo hace lasso.

3.2. Lasso y ridge

Logit es el método comúnmente utilizado cuando se tiene una variable dependiente binaria. Sin embargo, existen ciertas desventajas cuando se cuenta con una gran cantidad de variables independientes correlacionadas. El estimador es consistente pero con alta varianza, es decir, modelos más complejos tienden a ser menos sesgados, pero con mayor varianza, lo cual afecta al error de predicción. El desafío es saber "cuánto" sesgo tolerar para bajar considerablemente la varianza. Las estimaciones ridge (Hoerl y Kennard, 1970; Le Cessie y Houwelingen, 1992; Tibshirani, 1996) y lasso (Park y Casella, 2008; Tibshirani, 1996) son útiles, ya que regularizan los coeficientes compensando un pequeño aumento en el sesgo con una mayor reducción en la varianza de la predicción.

La regresión ridge logística, se obtiene al maximizar la función de verosimilitud con un parámetro que penaliza todos los coeficientes excepto a la constante.

Partiendo del modelo logit:

$$p_i = \pi_i = \frac{e^{x_i\beta}}{1 + e^{x_i\beta}} \tag{5}$$

donde x_i es la i-ésima fila de una matriz de n observaciones con p variables independientes y una columna de 1s para la constante, y β es el vector de columna de los coeficientes de regresión. Las estimaciones de los parámetros se obtienen a través de maximizar la función log-likelihood (donde y_i es la variable independiente):

$$\begin{aligned} l(B) &= \sum_{n}^{i=1} [y_i log(\pi_i) + (1 - y_i) log(1 - \pi_i)] \\ &= \sum_{n}^{i=1} [y_i log(\frac{\pi_i}{1 - \pi_i}) + log(1 - \pi_i)] \\ &= \sum_{n}^{i=1} [y_i x_i \beta - log(1 + e^{x_i\beta})] \end{aligned} \tag{6}$$

Los estimadores de ridge logístico depende del valor que asume un parámetro de ajuste $\lambda \geq 0$. Este se selecciona por separado. Para obtener el valor de los coeficientes se maximiza la siguiente función, que a diferencia de la Ecuación 6, se agrega a la función L_2 una penalización de ridge (Duffy y Santner, 1989; Cessie y Houwelingen, 1992). Por lo tanto, la ecuación de maximización restringida es :

$$l_\lambda^R(\beta) = \sum_{i=1}^{n} [y_i x_i \beta - log(1 + e^{x_i\beta})] - \lambda \sum_{j=i}^{p} \beta_j^2 \tag{7}$$

A medida que aumenta la penalización λ, los coeficientes ridge se acercarán a cero, pero ninguno de ellos será exactamente cero. Cuando $\lambda = 0$ el término de regularización no tiene efecto y los estimadores serán iguales a los del modelo logit. La desventaja de usar este modelo es que no selecciona variables, e incluye todas en el modelo final (James et al., 2013).

Lasso logit es otra alternativa de regularización que supera la desventaja de ridge, ya que reduce el número de predictores en el modelo final. La versión penalizada de la función log-likelihood a maximizar es ahora (Hastie et al., 2009):

$$l_{\lambda}^{R}(\beta) = \sum_{i=1}^{n}[y_i x_i \beta - log(1 + e^{x_i \beta})] - \lambda \sum_{j=i}^{p} |\beta_j| \tag{8}$$

Lasso penaliza L_1 en lugar de L_2. Esta penalización se usa tanto para la selección de predictores como para la selección de ellos, ya que cuando el λ es suficientemente grande, fuerza a que algunos coeficientes sean iguales a cero (James et al., 2013). La ventaja de lasso es que solo incluye un subconjunto de variables predictoras en el modelo final, lo cual mejora la interpretabilidad del modelo.

Tanto para el caso de ridge como para el de lasso los valores de λ juegan un rol fundamental. Para elegir el mejor valor λ para el modelo se usa *k-fold cross-validation* (Hastie et al., 2009). Los datos se dividen en k subconjuntos de aproximadamente el mismo tamaño y uno de los subconjuntos se convierte en el conjunto de validación. Los subconjuntos k-1 restantes se utilizan como datos de entrenamiento. Este procedimiento se repite k veces, cada vez con un conjunto de validación diferente, y el valor óptimo de λ se estima de manera que se maximice la función *cross-validated log-likelihood* (Goeman, 2010)

Entonces, ¿cuáles son los beneficios de estos modelos? Ambos modelos parten de logit, por lo tanto ambos son capaces de clasificar a los NNyA que trabajan y que no trabajan. El aporte de estos modelos, o la distinción de ellos, se centra en introducir cierta "cantidad" de sesgo para reducir la varianza, de tal manera que las predicciones sean mucho más precisas que en logit. A su vez, lasso se diferencia de ridge porque ataca un problema adicional, la reducción de dimensionalidad. Lasso reduce a cero las variables que no son importantes para el modelo.

3.3. Árboles

Los árboles de decisión también son modelos predictivos formados por reglas binarias con las que se logra dividir las observaciones en función de sus características y predecir el valor de la variable de respuesta. La principal diferencia con los modelos anteriores, es que los árboles permiten que las distintas variables interactúen entre sí y condicionen la predicción. Por ejemplo, si Y_i es la variable dependiente, y tiene dos predictores X_1 y X_2, el

algoritmo lo que hace es elegir una variable y particionar el espacio, a su vez el algoritmo decide que punto de esa variable particionar, la media, la moda, etc. Este algoritmo se repite recursivamente, dependiendo el modelo de árbol que se utilice, se decide el momento en el cual deja de actuar el algoritmo. De esta manera, y a modo ilustrativo, la forma en que se relacionan las variables son semejantes a las ramificaciones de los árboles.

Usar la técnica de árboles para predecir tiene muchas ventajas, no parten de ningún supuesto estadístico sobre la forma de distribución de los datos, pueden usar tanto predictores numéricos como categóricos y no son muy influenciados por valores atípicos. Los árboles son fáciles de interpretar, ya que se los pueden presentar con esquemas que muestren las relaciones entre las distintas variables, además es posible identificar las variables más importantes del modelo (con excepción de random forest). A su vez, son capaces de seleccionar predictores de forma automática. La técnica de árboles puede aplicarse tanto a problemas de regresión como de clasificación. Y es por esto, que en este trabajo se utilizan para predecir una clasificación, si el NNyA trabaja o no.

Sin embargo presentan ciertas desventajas, por ejemplo, son sensibles a datos de entrenamiento desbalanceados, es decir, una categoría se diferencia notoriamente por sobre las demás. Un claro ejemplo es la predicción de trabajo infantil, ya que solo el 12 % de la muestra de NNyA trabaja. El sobre ajuste también es una desventaja, esto implica que el modelo dentro de la muestra de entrenamiento predice correctamente, pero fuera de ella, no. Sin embargo, las técnicas de árboles seleccionadas en este trabajo tratan de solucionar dicho problema.

Dentro de la familia de árboles, random forest es uno de los modelos más conocidos. Está formado por un conjunto de árboles de decisión, cada uno entrenado con una muestra distinta (bootstrapping). Esta aleatoriedad en la muestras utilizadas busca disminuir la correlación entre los árboles. La predicción de una nueva observación se obtiene agregando las predicciones de todos los árboles individuales que forman el modelo.

Otra variante son los modelos gradient boosting (GBM). Estos modelos están compuestos por un conjunto de árboles individuales entrenados secuencialmente. La idea fundamental es que cada nuevo árbol mejora los errores de los árboles anteriores. Para predecir se agregan las predicciones de todos los árboles individuales que forman el modelo. Por otra parte, existe una variante de GBM, XGBoost que son las siglas de Extreme gradient boosting. Este modelo es una implementación específica del GBM, ya que utiliza aproximaciones más precisas para encontrar el mejor modelo de árbol y evita de mejor manera el sobre ajuste. A diferencia de todas la técnicas antes mencionadas, tanto GBM como XGBoost permiten modelar la data con valores faltantes, de esta manera tenemos más variables/observaciones con las cuales trabajar.

3.4. Comparación de métodos

Todos los modelos descriptos presentan las características mínimas necesarias para predecir una variable binaria como lo es el trabajo infantil. En primera instancia, ninguno parecería ser superior a otro ya que todos los modelos presentan desventajas y ventajas.

Logit estima la probabilidad de ocurrencia de que un NNyA trabaje o no, y en base al clasificador de Bayes se predice si trabaja o no. Logit permite diferenciar las variables estadísticamente significativas de aquellas que no lo son, como así también permite saber como se relacionan las variables independientes con la variable dependiente, en sentido y magnitud. Sin embargo este modelo presenta ciertas dificultades que son sobrellevadas por ridge y lasso. Estos métodos incorporan el *trade-off* entre sesgo y varianza al modelo, de tal manera que al permitir cierto sesgo en el modelo, se reduce drásticamente la varianza y así se obtienen predicciones mucho más precisas. Cierto es que el tema de la reducción de la dimensionalidad es un tema distinto al hecho de predecir correctamente, pero el único modelo que logra hacerlo es lasso, ya que lleva los coeficientes a cero, mientras que todos los otros modelos no lo hacen. Ridge y lasso parten de logit, por lo tanto también es posible identificar el sentido de las relaciones entre la variable dependientes y las independientes. Por otro lado, los distintos modelos de árboles son alternativas totalmente distintas, en primer lugar, la principal diferencia es la forma en que trabaja el algoritmo, como ya se explicó, las variables interactúan entre sí para poder predecir, por lo tanto las variables dependen del valor que asumen otras, no existe una relación directa con la variable dependiente como en los casos anteriores. Otra ventaja de usar árboles, específicamente GBM y XGBoost es que ambos modelos permiten trabajar con valores faltantes, lo cual aumenta la cantidad de observaciones en la base de datos. Sin embargo, una gran desventaja de estos modelos es que el algoritmo no clasifica bien en caso de muestras desbalanceadas, como es el caso de trabajo infantil.

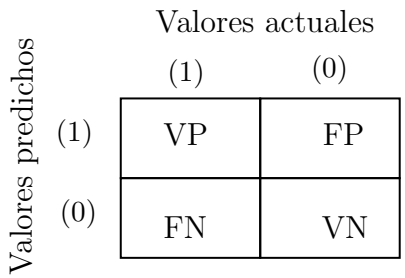
Todos estos métodos resultan propuestas interesantes, todos tienen ventajas y desventajas pero todos son capaces de predecir variables binarias e identificar las variables relevantes para el modelo. Ahora bien, una vez identificadas las variables más importantes interesa saber qué tan bien predice cada modelo, qué tan correcta es dicha predicción.

La matriz de confusión es una herramienta que permite evaluar el desempeño de cualquier algoritmo de clasificación. Permite tener una mejor aproximación respecto a que tan bien o mal está clasificando el modelo, para ello se cuentan los aciertos y errores de cada una de las categorías en clasificación.

La Figura 4 muestra la matriz de confusión para un clasificador en dos clases, como es el caso de todos los modelos que estamos usando para predecir trabajo infantil.

VP (verdadero positivo) es el número de predicciones correctas de que un caso es positivo. FP (falso positivo) es es el número de predicciones incorrectas de que un caso es positivo,

Figura 4: Matriz de Confusión



o sea la predicción es positiva cuando realmente el valor tendría que ser negativo. A estos casos también se les denomina errores de tipo I. FN (falso negativo) la predicción es negativa cuando realmente el valor tendría que ser positivo. A estos casos también se les denomina errores de tipo II. VN (verdadero negativo) es el número de predicciones correctas de que un caso es negativo.

A partir de los datos de la matriz de confusión se definieron distintos términos estándar para medir el desempeño de un clasificador:

- Precisión: total de predicciones correctas sobre el total de predicciones realizadas.
- Sensibilidad: La "tasa positiva verdadera" - el porcentaje de individuos que el modelo predijo correctamente que ocurriría. $Sensibilidad = \frac{VP}{VP+FN}$
- Especificidad: La "tasa negativa verdadera" - el porcentaje de individuos que el modelo predijo correctamente no incurriría. $Especificidad = \frac{VN}{FP+VN}$
- Tasa total de clasificación errónea: el porcentaje del total de clasificaciones incorrectas realizadas por el modelo.

En este estudio se hace principal énfasis en la *tasa de sensibilidad*, ya que este valor es el que verdaderamente estaría dando información sobre que tan bien predice cada modelo a la pequeña submuestra de NNyA que trabajan. Cierto es que existen otros indicadores, como por ejemplo el *valor F1* que se calcula haciendo la media armónica entre la sensibilidad y la precisión. Este indicador implica que al investigador le importa por igual la sensibilidad y la precisión, y en este estudio no es así, nos importa principalmente la *sensibilidad*. Otra métrica posible de evaluar es el área por debajo de la curva ROC, que se obtiene al trazar la *tasa sensibilidad* frente a *1 - especificidad*. Esta área solo sería un adicional gráfico, pero no aportaría mas información sobre la clasificación particular de los NNyA que trabajan. Existen otras métricas que permiten evaluar y comparar la capacidad predictiva de los modelos, como son AIC, BIC, pérdida logarítmica, índice Jaccard, etc, pero todas ellas se centran en evaluar que tan precisas son las predicciones 0 y 1, estas métricas no diferencias o no tienen fin práctico cuando se esta trabajando con muestras desbalanceadas, como es este caso.

4. Resultados

Se probaron seis modelos distintos para predecir trabajo infantil: ridge, lasso, logit, random forest, gradient boosted bodel(GBM) y extrem gradient boost (XGBoost). La variable dependiente en todos los modelos es "Trabaja", extraída de la base de datos EANNA, donde 0 indica que el NNyA no trabaja y 1 caso contrario. El conjunto de variables predictoras, tal como se dijo en la sección 2, corresponden a variables referentes a características socioeconómicas del hogar, del jefe de familia y del niño (ver anexo 1), además se incluyen características de infraestructura del hogar y los servicios a los que tienen acceso.

Primer paso, previo a obtener los resultados de los distintos modelos, se realizó un análisis de la base de datos y se decidió tratar a los valores faltantes de ciertas variables. A saber, en todas las variables referentes a ingresos, edad de jefe de hogar, y cantidad de repeticiones por parte del NNyA se imputó el valor promedio de cada una de las variables mencionadas. Y en cuanto a las variables que dependen de la edad, como por ejemplos, si Falta/faltaba a menudo a la escuela Primaria EGB o a la escuela secundaria o polimodal, fueron utilizadas para crear una nueva variable que combina a ambas, es decir se crea una variable que indica si falta/faltaba, independientemente del nivel educativo al que asiste el NNyA. Lo mismo sucede con las variables repitió alguna vez un grado de primaria/año de EGB y secundaria o polimodal, ambas se convierten en la variable repitió". Finalmente, la base de datos (total, es decir, muestra de entrenamiento y testeo) que se utiliza para los modelos ridge, lasso, logit y random forest contiene 12.823, donde el 13 % del total de NNyA de la muestra reportan trabajar. Mientras que para el caso de GBM y XGBoost, se utilizan 16.115 observaciones, ya que estos permiten modelar también los datos faltantes (el 10 % de los NNyA trabajan).

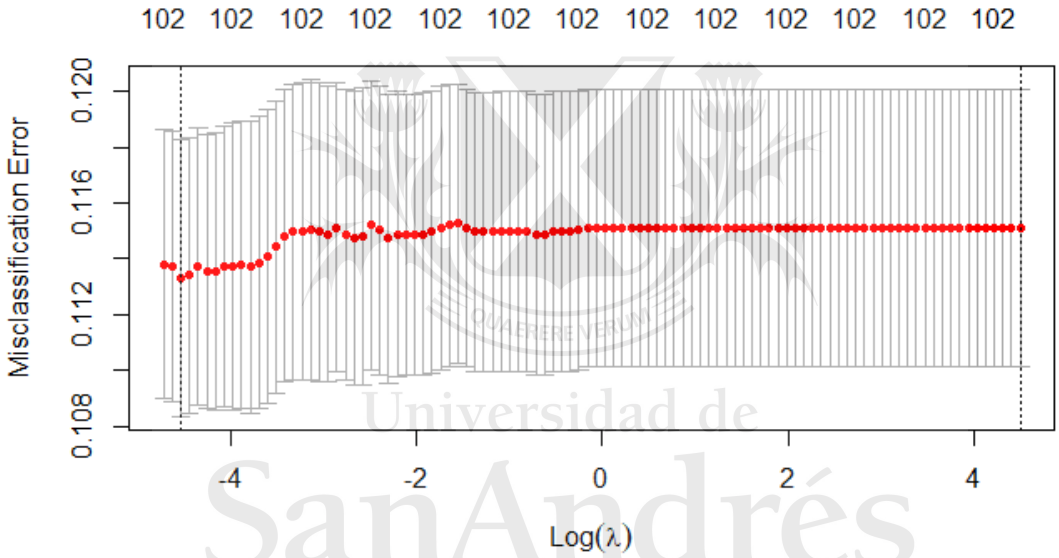
Todos los modelos son entrenados con el 80 % de la muestra y con el 20 % restante se evalúa la capacidad predictiva. Al dividir la muestra se mantiene en cada subsample un 10 % de NNyA que trabajan en cada una[2]. Al dividir la muestra de esta forma es posible conocer que tan bien se predice a los NNyA que trabajan ya que se conocen cuales son las observaciones que forman parte de la muestra de testeo y se pueden calcular las medidas de performance descriptas en al apartado anterior. En todos los casos para ajustar los hiperparámetros de cada modelo se utiliza la librería *scikit-learn*, con la función *GridSearchCV*. Este algoritmo prueba todas las combinaciones posibles de parámetros que se desean probar en el modelo. Cada uno de esos parámetros se prueba en una serie de iteraciones de validación cruzada.

[2]La división fue realizada de manera aleatoria una sola vez, es decir todos los modelos usaron las mismas particiones de data

4.1. Ridge

El primer modelo que se evalúa es ridge y es implementado en *RStudio* con el paquete *glmnet*. Como se explicó en la sección 3, el resultado del proceso de clasificación depende del hiperparámetro λ que determina el grado de penalización. Para poder elegir el mejor λ para nuestro modelo se utiliza la Figura 5 que muestra el error cuadrático medio obtenido por validación cruzada para cada valor del logaritmo de λ junto con la barra de error correspondiente. En este caso, el λ con el que consigue el menor error es 0,0104 ($log(\lambda) \approx -4{,}56$)

Figura 5: Error Cuadrático Medio - ridge



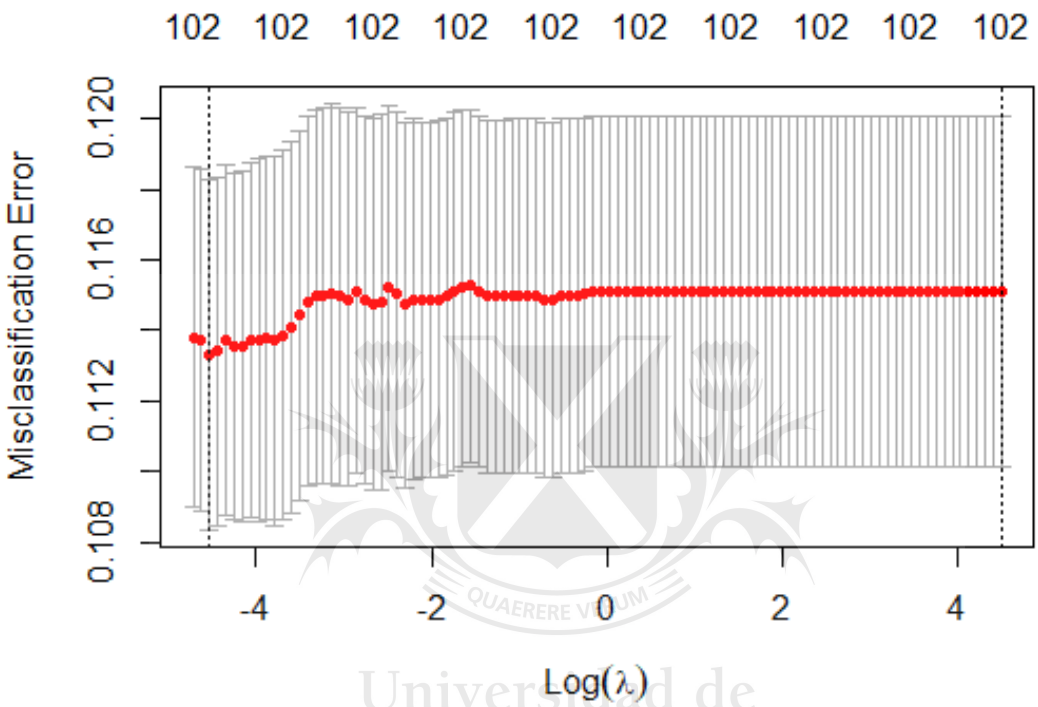
Los coeficientes con valores más altos son aquellos correspondientes, en su gran mayoría al sexo y edad del jefe del hogar y del NNyA, nivel educativo del niño y del grupo familiar, ingreso y horas destinadas a la ocupación principal y composición familiar, es decir cantidad de de miembros y cantidad de NNyA. La región a la que pertenecen y los deseos que tienen los NNyA para su mayoría de edad también son variables importantes a tener en cuenta dentro del modelo.

4.2. Lasso

El proceso para estimar un modelo lasso y la identificación del mejor valor de λ es muy parecido al caso de ridge. En este caso, tambien se corre el modelo en *RStudio* con el paquete

glmnet. Al igual que en ridge, para elegir el mejor λ se observa la Figura 6. En este caso, el λ con el que consigue el menor error es 0,0134 ($log(\lambda) \approx -4{,}31$) .

Figura 6: Error Cuadrático Medio - lasso



A diferencia del modelo ridge, lasso sí selecciona las variables a incorporar dentro del modelo, y ellas son: sexo y edad del NNyA, nivel educativo al que asiste, si falta o no a la escuela, cantidad de veces que repitió algún curso, deseos para su mayoría de edad; en cuanto al jefe del hogar las variables seleccionadas son sexo y edad, ocupación principal, horarios destinado e ingreso percibido; también selecciona variables referentes al hogar, cantidad de miembros y nivel educativo del hogar. Lasso no selecciona ninguna variable relacionada con la estructura edilicia del hogar ni los servicios a los que tienen acceso.

4.3. Logit

Por su parte el modelo logit es estimado con la función glm de *RStudio* e indica que solo algunas variables son estadísticamente significativas. El Cuadro 1 muestra solamente las variable significativas, sin embargo el modelo a la hora de predecir utiliza todas las variables.

En este caso, las variables coinciden con las mencionadas en ridge y lasso, edad y sexo del NNyA y del jefe del hogar, deseos para la mayoría de edad, composición familiar, etc.;

Cuadro 2: Coeficientes logit

Variable	Coef
Región NOA	-3.474e+00***(0.334)
Región Cuyo	-.0331***(1.149)
Región Pampeana	-.249***(0.321)
Región Patagonia	-0.347**(0.130)
Edad NNyA	2.59***(0.146)
Sexo NNyA	1.773**(0.156)
Asiste a la escuela	0.680**(0.159)
Deseos p/18años	3.179***(0.747)
Material del piso	4.498*(0.632)
Cant de pob 0-17 años	1.723***(0.759)
Clima Edu Hogar	-2.07***(0.418)
Edad jefe	-8.878**(0.347)
Sexo jefe	6.190**(0.308)
Situación Conyugal	0.394***(0.185)
Horas en la ocup ppl	7.17*(0.003)
Ayuda del estado	0.447*(0.261)
Observations	12,893
Log Likelihood	−2,907.699
Akaike Inf. Crit.	6,063.398
Note:	*p<0.1; **p<0.05; ***p<0.01

sin embargo, hay dos variables más a tener en cuenta, material del piso y ayuda por parte del estado. En el caso de ridge los coeficientes que acompañan a estas dos ultimas variables son muy bajos y en el caso de lasso, los coeficientes son cero. La mayoría de estos coeficientes también coinciden con los resultados vistos en Torre (2008) y Waisgrais (2007). Agua, material de paredes y desagüe si resultaron estadísticamente significativos en el trabajo de los autores mencionados, mientras que en este trabajo no.[3]

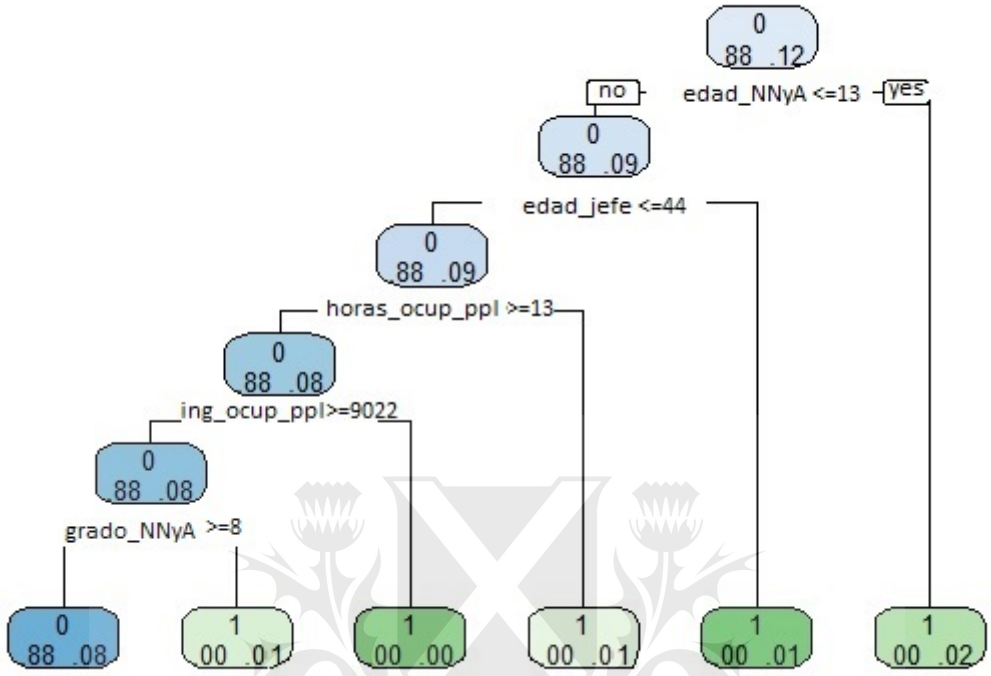
[3]La coincidencia se da en el signo del coeficiente, no exactamente en la magnitud. Las diferencias son consecuencia de las variables incorporadas en los modelos.

4.4. Random Forest

Por otra parte, al usar las distintas especificaciones de árboles, se encuentra que las variables importantes para predecir trabajo infantil, siguen siendo las mismas. Este modelo también es implementado en *Rstudio* con el paquete *randomForest*.Las primeras quince variables más importantes que incorpora random forest al modelo son (ordenadas de forma decreciente de importancia de acuerdo al Mean Decrease Gini que es una medida de importancia variable basada en el índice de impurezas de Gini utilizado para el cálculo de divisiones en árboles.) edad del NNyA y del jefe del hogar, horas destinadas a la ocupación principal, ingreso percibido, nivel educativo del NNyA, region a la que pertenece, situación conyugal, deseos del NNyA para la mayoría de edad, tamaño del hogar, cantidad de menores edad que conviven en el hogar, a quien le pertenece la vivienda, categoría ocupacional, cantidad de veces que repitió la escuela el NNyA, y el clima educativo del hogar.

La Figura 7 muestra como random forest relaciona las variables para poder predecir trabajo infantil. En esta figura, a modo explicativo, solo se representan las relaciones entre cinco variables, sin embargo, el modelo incluye y relaciona a todas las variables de la base de datos. El modelo para clasificar si un NNyA trabaja parte de su edad, aquellos que tienen más de 13 años son clasificados como NNyA que trabajan, mientras que aquellos que son menores de 13 años dependen de la edad del jefe del hogar. Por lo tanto, si el NNyA tiene hasta 12 años y el jefe es mayor de 44 años se clasificará como NNyA que trabaja. A su vez, si el NNyA vive en un hogar en el cual el jefe es menor de 44, el modelo random forest indica que hay que tener en cuenta la cantidad de horas que dedica a la ocupación principal, si el jefe le dedica menos de 13 horas, el NNyA será clasificado como trabajador. Si el jefe trabaja más de 13 horas, pero su sueldo es menor a $9.022, el modelo clasificará al NNyA cargo como trabajador. Si el sueldo es mayor, la clasificación dependerá del nivel educativo del NNyA, si tiene al menos 8 años de educación, el modelo lo clasificará como no trabajador. El modelo continua y el valor predicho depende de como interactúan las variables. Los modelos de árboles permiten estudiar cómo se relacionan las variables explicativas para poder predecir. No todas se relacionan de manera lineal con la variable dependiente, mas bien dependen del valor que puede llegar a tomar otra variable y así sucesivamente considerando a todas las variables incluidas en el modelo.

Figura 7: Árbol random forest



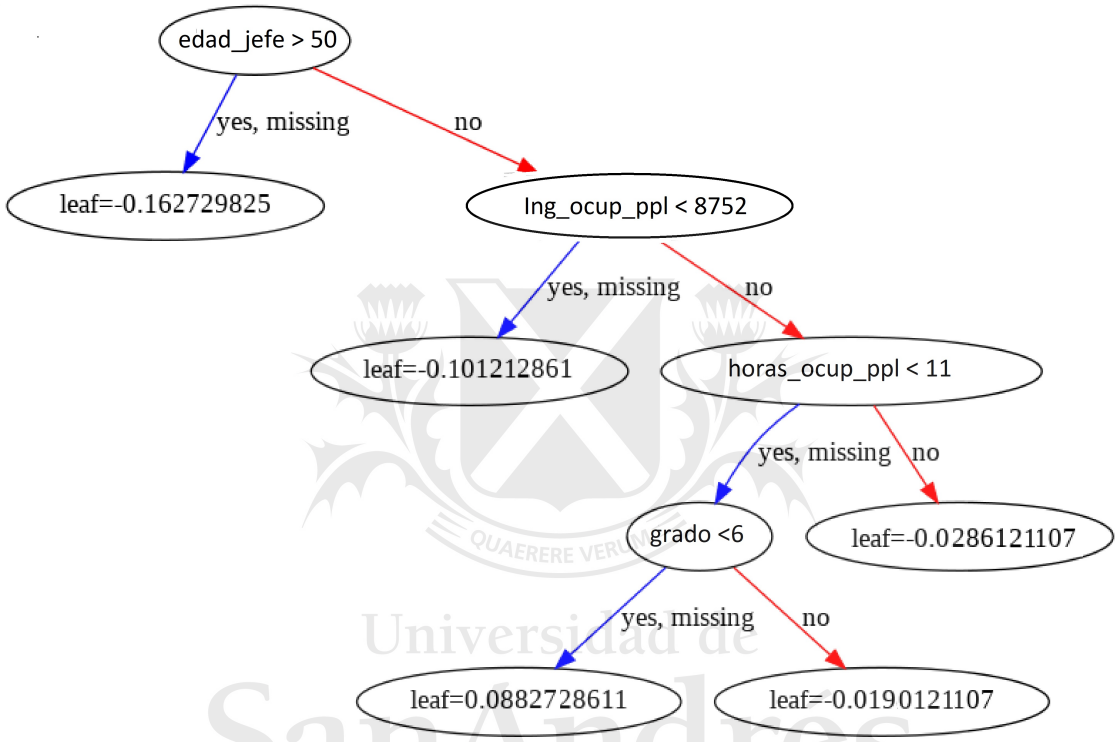
4.5. Gradient Boosted Model

GBM es otro modelo dentro de la familia de árboles de machine learning. Este modelo es implementados en *Python* con el paquete *scikit-learn* y la función *GradientBoosting-Classifier*. Los resultados siguen la misma lógica que los modelos anteriormente descriptos, las variables más importantes (ordenadas de acuerdo al Mean Decrease Gini) siguen siendo las mismas. La edad y sexo del NNyA y del jefe del hogar, características de la ocupación principal, como los horas destinadas e ingreso percibido, composición del hogar, etc. siguen siendo fundamentales para predecir el trabajo infantil. GBM es un modelo similar a random forest, las primeras quince variables más influyentes y su orden de importancia coinciden en su gran mayoría, la única diferencia es que GBM incorpora si el hogar recibe AUH, mientras que random forest no la considera como variable principal.

La Figura 8 muestra como se relacionan dichas variables. Para predecir si el NNyA trabaja o no GBM parte de la edad del jefe (al igual que random forest). Aquellos NNyA cuyo jefe de hogar tenga más de 50 años serán predichos como trabajadores, caso contrario su clasificación dependerá del ingreso de la ocupación principal. Si el ingreso del jefe del hogar es menor a $8.752 (monto similar al de random forest) el NNyA será predicho como trabajador. En caso de que el ingreso sea mayor, random forest indica que se ha de tener en

cuenta la cantidad de horas destinadas a la ocupación principal, si la persona trabaja menos de 11 horas, el NNyA a cargo será clasificado como trabajador. El árbol completo continúa con más ramificaciones que corresponden a las demás variables incluidas dentro del modelo. La Figura 8 muestra la lógica con la cual interactúan las variables, misma lógica con la que trabaja random forest.

Figura 8: Árbol GBM



4.6. XGBoost

XGBoost tiene la ventaja de poder modelar incluso valores faltantes (al igual que GBM). El algoritmo trabaja con una "dirección predeterminada" (Chen y Guestrin, 2016) a la hora de decidir como clasificar a los valores faltantes, la dirección predeterminada puede ser 1 o 0, y se aprende en el proceso de construcción del árbol para elegir la mejor dirección que optimice la pérdida de bondad de ajuste durante el entrenamiento. Es por esto que en este modelo se incluyeron más variables que en todos los anteriores ya que muchas variables eliminadas para los modelos anteriores se debieron a gran cantidad de valores faltantes en ellas (ver Anexo 1). Este modelo es implementados en *Python* con el paquete *XGBoost* y la función *XGBoostClassifier*.

Las variables más importantes para el modelo XGBoost (de acuerdo al Mean Decrease Gini) son la edad del jefe del hogar y del NNyA, el ingreso de la ocupación principal, nivel

educativo al que asiste el NNyA, tamaño del hogar, si percibe AUH, sexo del NNyA y del jefe, situación conyugal, deseos para la mayoría de edad. El gráfico de cómo se relacionan las variables es muy similar al mostrado en GBM, al menos para las primeras variables más importantes.

4.7. Comparación de modelos

Este trabajo busca comparar los distintos modelos que permiten pronosticar trabajo infantil, para ello se evalúan distintos factores, por ejemplo, las variables utilizadas para el pronóstico, las relaciones entre la variable dependiente e independientes, relaciones entre variables independientes (en el caso de árboles), si el modelo selecciona o no variables y las distintas medidas de performance que indican que tan bien predice cada modelo a la pequeña muestra de NNyA que trabajan.

Primero y principal, cabe resaltar que en todos los modelos es posible identificar cuáles son las variables más importantes. Todas ellas resultan estar en linea con los resultados obtenidos en trabajos anteriores como los de Torre (2008) y Waisgrais (2007). Para todos los modelos las variables más importantes o estadísticamente significativas se concentran en: edad y sexo del NNyA y del jefe del hogar, región a la que pertenece, deseos para la mayoría de edad, cantidad de menores de edad en el hogar, horas destinadas e ingreso percibido por ocupación principal y situación conyugal. Existen otras variables que sólo fueron relevantes en determinados modelos, como por ejemplo, si el hogar recibe alguna ayuda del estado y el material con el que esta construido el piso del hogar solo resultaron estadísticamente significativas en logit. Como ya se explicó, lasso, a diferencia de todos los modelos, es el único capaz de seleccionar variables a incorporar dentro del modelo, en este caso las únicas variables incorporadas en el modelo son las que se repiten en todos los modelos conjuntamente con el ingreso de la ocupación principal. Todos estos modelos permiten solamente relaciones directas entre las variables independientes y dependientes, esto quiere decir, que es posible saber en qué sentido y en qué magnitud afecta cada variable predictora a la variable "Trabaja". Los árboles no pueden ser interpretados de la misma forma, en estos casos las variables se relacionan entre sí y dependen unas de otras para lograr predecir. En los tres tipos de árboles, las principales variables son las mismas que se mencionaron para los modelos logit, ridge y lasso, a saber, edad, sexo, ingresos y horas destinadas a la ocupación principal, el nivel educativo al que asiste el niño, deseos para la mayoría de edad, composición familiar, etc. Random forest, a diferencia del resto incorpora tenencia de la vivienda y si el NNyA repitió o no la escuela. GBM solo agrega la variable AUH (si recibe o no la Asignación Universal por Hijo). ¿Cómo se relacionan estas variables? Los modelos de árboles permiten visualizar que la variable ingreso de ocupación principal sólo importa para predecir trabajo infantil si el jefe del hogar es mayor a determinada edad, y a su vez el ingreso depende del tiempo que

le dedica, y a su vez al curso al que asiste el NNyA. Por lo tanto, estos modelos permiten conocer que no solo importan las variables, sino también que determinado grupo dentro de cada variable importa más que otros para predecir trabajo infantil.

Es necesario aclarar que hay variables que no son importantes para ningún modelo. En su gran mayoría, las variables relacionadas con características de la vivienda, como el material de construcción, acceso a servicios y tenencia de la vivienda, no son significativas o son de muy poca relevancia en comparación de las demás variables. Otras variables pertenecientes al cuestionario 3, como por ejemplo, si el NNyA llega tarde, falta o repite el colegio no son muy relevantes. Si el hogar recibe algún tipo de ayuda como ropa, medicamentos, dinero, o transferencias por parte del estado, tampoco resultaron importante.

Ahora bien, ya se saben cuales son las variables relevantes para predecir trabajo infantil, pero, ¿qué tan precisas son la predicciones de los modelos? Como se puede observar en el Cuadro 3, todos los modelos logran predecir correctamente fuera y dentro de la muestra. Pero dado que no tenemos gran porcentaje de NNyA que trabajan, nos interesa, saber qué modelo predice mejor a este pequeño sector que si trabaja. Por lo tanto, la mejor idea para poder comparar el desempeño de los modelos fuera de la muestra de entrenamiento, es usar los datos de la matriz de confusión de cada uno de los modelos.

Las medidas construidas a partir de la matriz son:

Cuadro 3: Evaluación de performance

Modelo	**Acc - train**	**Acc-test**	**Sensibilidad**	**Especificidad**	**Clasificación errónea**
ridge	89 %	88 %	7 %	99 %	11 %
lasso	93 %	90 %	0 %	100 %	10 %
logit	91 %	88 %	50 %	89 %	15 %
RF	92 %	88 %	53 %	99 %	5 %
GBM	93 %	91 %	54 %	99 %	4 %
XGBoost	92 %	89 %	53 %	99 %	5 %

Dado que 0 es el valor que asume la variable Trabaja cuando el NNyA no participa del mercado laboral, *especificidad* siempre será un porcentaje muy alto (ya que el 90 % de los niños no trabajan). Por lo tanto, *sensibilidad* es el indicador que realmente nos interesa.

En base a los resultados obtenidos, los distintos modelos construidos en base a métodos de árboles predicen mejor el pequeño subconjunto de niños trabajando. En los tres casos, más del 50 % de los NNyA que trabajan en la muestra de entrenamiento son predichos correctamente. Particularmente, en base al análisis realizado, GBM es capaz de clasificar mejor a los NNyA que trabajan, ya que es el modelo con mayor porcentaje de *sensibilidad* y menor *tasa de clasificación errónea*.

5. Conclusión

Existe una extensa literatura internacional y nacional sobre trabajo infantil, sin embargo todos estos trabajos se interesan por las causas y determinantes de este fenómeno y para ello utilizan modelos de la econometría tradicional. Es por esto que este trabajo se enfoca en comparar y evaluar la capacidad predictiva de distintas técnicas de econometría tradicional como logit y técnicas de machine learning como lasso, ridge y las diferentes especificaciones de árboles.

Una vez analizadas las particularidades de cada modelo ninguno parece ser superior a otro. Todos son capaces de predecir variables binarias como es el caso de trabajo infantil, todos logran identificar las variables relevantes, todos presentan ventajas y desventajas.

En cuanto a las variables, logit incorpora todas ellas dentro del modelo y permite identificar las variables estadísticamente significativas. Ridge y lasso, son variantes de logit que incorporan el trade-off de sesgo y varianza, es decir, tolerando cierto sesgo es posible reducir la varianza y así obtener predicciones más precisas, a su vez lasso, a diferencia de ridge, selecciona variables. Los modelos de árboles permiten visualizar interacciones entre las variables y observar cómo la clasificación de si un NNyA es trabajador depende de cómo se relacionan las variables y a su vez dependen de ciertos grupos dentro de las variables.

Al igual que Torre (2008) y Waisgrais (2007) este trabajo utiliza los datos de la EANNA 2016-2017. Los coeficientes de los modelos logit propuestos por estos autores coinciden, en su gran mayoría, con las variables más importantes o estadísticamente significativas de los modelos propuestos en este trabajo, edad y sexo del NNyA y del jefe del hogar, región a la que pertenece, deseos para la mayoría de edad, cantidad de menores de edad en el hogar, horas destinadas a la ocupación principal y situación conyugal. En contraposición con los resultados mostrados en Torre (2008) y Waisgrais (2007) agua, material de paredes y desagüe si fueron estadísticamente significativas, mientras que todos los modelos de este trabajo coinciden en que estas variables no aportan información para predecir trabajo infantil [4]. Tampoco aportan información aquellas variables relacionadas con la estructura edilicia de la casa, los distintos servicios a los que tienen acceso y las respuestas del niño en la escuela, asistencia, llegadas tardes, cantidad de repeticiones, etc..

Para saber cuan precisas son las predicciones se evalúan distintos indicadores de performance construidos en base a la matriz de confusión de cada modelo. Todos logran predecir adecuadamente trabajo infantil en la muestra de entrenamiento y en la de testeo, pero cabe recordar que la muestra con la que se trabaja es desbalanceada, 13 % del total de NNyA trabajan, entonces hay que hacer énfasis en predecir correctamente a aquellos NNyA de la

[4] Cabe resaltar que el objetivo de la investigaciones de estos autores se centran en analizar los determinantes de trabajo infantil, mientras que en este trabajo el objetivo principal es comparar el desempeño de las distintas técnicas predictivas.

muestra de testeo que trabajan. En este caso, los resultados no son realmente buenos. El indicador de sensibilidad es extremadamente bajo para ridge y lasso. Logit por su parte logra predecir correctamente sólo el 50 % del total de NNyA que trabajan de la muestra de testeo, es decir que la mitad de NNyA que trabajan no están siendo predichos como tales. Respecto a los modelos de árboles los resultados no son muy distintos los de logit, la tasa de sensibilidad también ronda el 50 %.

Entonces, se sabe cuáles son las ventajas y desventajas que tiene cada modelo para predecir, además se sabe que todas las variables relevantes y no relevantes son prácticamente similares en todos los modelos. La principal diferencia reside en cuan bien se predice la submuestra de NNyA que trabajan. Lasso y ridge no son alternativas muy prometedoras ya que el indicador de sensibilidad es demasiado bajo, logit y las demás especificaciones de árboles resultan ser alternativas que predicen mejor a los NNyA que trabajan, al menos al 50 %. Ha de tenerse en cuenta que se trabaja con una muestra desbalanceada, aun así, estos modelos lograron predecir el 50 % correctamente. Quedará para futuras investigaciones estudiar la implementación de técnicas de re-balanceo, como por ejemplo submuestrear, es decir quitar observaciones de la clase mayoritaria o contrariamente, sobremuestrear, incrementar la cantidad de observaciones de la clase minoritaria. También existen algoritmos, como *SMOTE* que son capaces de generar muestras sintéticas a partir de la muestra minoritaria que ya se tienen.

De esta manera este trabajo realiza un gran aporte a la literatura nacional e internacional sobre trabajo infantil ya que muestra que no sólo las técnicas de logit o probit (utilizadas en toda la literatura vigente) sino también las técnicas de árboles de machine learning son alternativas viables para predecir y conocer los determinantes de trabajo infantil. Por lo tanto, este trabajo refleja el gran potencial de las técnicas de machine learning para enfrentar el estudio de problemas sociales. A su vez, un subproducto de este trabajo es haber encontrado un subconjunto de la muestra que logra identificar a NNyA que trabajan (al menos el 50 % de ellos), lo cual resulta de sumo interés para los distintos organismos internacionales y hacedores de políticas públicas. Estos modelos pueden ser empleados por distintos organismos con información sobre las características de NNyA para así predecir quienes de ellos posiblemente se encuentren activos laboralmente (al menos el 50 %) y tener mayor foco en cada uno de ellos. Al saber que todos los modelos coinciden en cuales son las variables más relevantes hace más fácil concentrar los esfuerzos en monitorear principalmente estas características y así evitar que la cantidad de NNyA que trabajan aumente.

Cabe también recordar que esta literatura que vincula temas sociales, como lo es el trabajo infantil, con la aplicación de técnicas de machine learning es aún nueva y muy poco explorada. Hay grandes posibilidades de seguir investigando y es por esto que quedarán para futuras investigaciones efectuar pruebas de robustez con otros modelos como support

vector machine (lineal o no lineal), redes neuronales, componentes principales (que reduce la dimensionalidad). vecinos cercanos, técnicas no paramétricas, elastic net (que combina las técnicas utilizadas de lasso y ridge). A su vez, resultaría sumamente interesante replicar este estudio en base a grupos etarios de los NNyA. El campo de investigación es sumamente amplio ya que la literatura de machine learning está constantemente incorporando nuevos modelos predictivos. También es importante reconocer que este trabajo, como así también los lineamientos que se sugieren para futuras investigaciones, no son limitantes a la literatura de trabajo infantil, sino también para estudiar otros tipos de problemáticas sociales.

6. Bibliografía

Basu, K., y Tzannatos, Z. (2003). The global child labor problem: what do we know and what can we do?. The world bank economic review, 17(2), 147-173.

Basu, K., y Van, P. H. (1998). The economics of child labor. American economic review, 412-427.

Bhalotra, S., y Heady, C. (2003). Child farm labor: The wealth paradox. The World Bank Economic Review, 17(2), 197-227.

Chen, T., Guestrin, C. (2016, August). Xgboost: A scalable tree boosting system. In Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining (pp. 785-794).

Duffy, D. E., y Santner, T. J. (1989). On the small sample properties of norm-restricted maximum likelihood estimators for logistic regression models. Communications in Statistics-Theory and Methods, 18(3), 959-980.

Edmonds, E. V., y Turk, C. (2004). Child labor in transition in Vietnam (Vol. 2774). World Bank Publications.

Goeman J. J., (2010). L1 and L2 Penalized Regression Models. Technical document for CRAN, http://cran.r-project.org/web/packages/penalized/penalized.pdf.

Hastie, T., Tibshirani, R., Friedman, J., (2009). The Elements of Statistical Learning: Data Mining, Inference, and Prediction. Second Edition. Springer.

Hiraoka, R. (1997). Whither child labour in India? Myth of its relation to poverty and economic development. Cornell University. Hoerl, A. E., y Kennard, R. W. (1970). ridge regression: applications to nonorthogonal problems. Technometrics, 12(1), 69-82.

Hoerl, A. E., Kennard, R. W. (1970). Ridge regression: Biased estimation for nonorthogonal problems. Technometrics, 12(1), 55-67.

Instituto Nacional de Estadística y Censos - I.N.D.E.C. (2018) Encuesta de Actividades de Niños, Niñas y Adolescentes 2016-2017. - 1a ed . -

Jafarey, S., y Lahiri, S. (1999). Will trade sanctions reduce child labour?: The role of credit markets. Journal of Development Economics, 68(1), 137-156.

James, G., Witten, D., Hastie, T., Tibshirani, R. (2013). An introduction to statistical learning (Vol. 112, p. 18). New York: springer.

Lazo, C. F. L. S. J. (2017). Predicting child labor in Peru: A comparison of logistic regression and neural networks techniques.

Le Cessie, S., y Van Houwelingen, J. C. (1992). ridge estimators in logistic regression. Journal of the Royal Statistical Society: Series C (Applied Statistics), 41(1), 191-201.

Mullainathan, S., Spiess, J. (2017). Machine learning: an applied econometric approach. Journal of Economic Perspectives, 31(2), 87-106.

Park, T., y Casella, G. (2008). The bayesian lasso. Journal of the American Statistical Association, 103(482), 681-686.

Post, D. (2002). Children's work, schooling, and welfare in Latin America. Boulder, CO.

Rodrigues, D. C., Prata, D. N., Silva, M. A. (2015). Exploring social data to understand child labor. International Journal of Social Science and Humanity, 5(1), 29.

Ranjan, P. (2001). Credit constraints and the phenomenon of child labor. Journal of development economics, 64(1), 81-102.

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. Journal of the Royal Statistical Society: Series B (Methodological), 58(1), 267-288.

Torre, J. (2008). Determinantes del trabajo infantil en Argentina. Anales, Asociación Argentina de Economía Política, 43.

Waisgrais, S. (2007). El trabajo de niñas, niños y adolescentes. Concepto, metodología y resultados. El trabajo infantil en la Argentina. Análisis y desafíos para la política pública, Oficina Internacional del Trabajo y Ministerio de Trabajo, Empleo y Seguridad Social, Argentina, 99-128.

Weiner, M. (1991). The child and the state in India: Child labor and education policy in comparative perspective. Princeton University Press.

7. Anexo I

A continuación se presentan las variables utilizadas en los modelos. Aquellas variables con * son las que además se incorporaron en el modelo XGBoost.

Características del NNyA

- Región
- Edad del niño
- Sexo del niño
- Asiste a la escuela
- Grado al que asiste
- Tipo de jornada*
- Tipo de gestión del establecimiento al que asiste*
- Llega/llegaba tarde a la escuela a menudo*
- Falta/faltaba a menudo a la escuela Primaria EGB*
- Repitió alguna vez un grado de primaria/año de EGB*
- Falta/faltaba a menudo a la escuela Secundario o polimodal
- Repitió alguna vez un grado de secundaria/ polimodal
- La semana pasada ayudó en un negocio, por dinero
- El año pasado ayudó en un negocio, por dinero
- La semana pasada cuidó personas fuera de su hogar por dinero o propina
- El año pasado cuidó a personas fuera de su hogar por dinero o propina
- La semana pasada repartió volantes, entradas, etc. por dinero o propina
- El año pasado repartió volantes, entradas, etc. por dinero o propina
- La semana pasada vendió algo en la feria, calle, tren, colectivo
- El año pasado vendió algo en la feria, calle, tren, colectivo
- La semana pasada cortó el pasto para ganar algún dinero o propina

- El año pasado cortó el pasto para ganar algún dinero o propina
- La semana pasada limpió parabrisas, malabares o cuidar autos por propina
- El año pasado limpió parabrisas, malabares o cuidar autos por propina
- La semana pasada hizo mandados a alguien fuera de su hogar para ganar dinero
- El año pasado hizo mandados a alguien fuera de su hogar para ganar dinero
- La semana pasada paseó perros por dinero
- El año pasado paseó perros por dinero
- La semana pasada hizo reparto de comida, transportó mercaderías o cargas
- El año pasado hizo reparto de comida, transportó mercaderías o cargas
- La semana pasada limpió casas o negocios, lavó o planchó ropa para afuera
- El año pasado limpió casas o negocios, lavó o planchó ropa para afuera
- La semana pasada juntó en la calle papeles, cartones para vender
- El año pasado juntó en la calle papeles, cartones para vender
- La semana pasada hizo pan, empanadas, dulces para vender
- El año pasado hizo pan, empanadas, dulces para vender
- La semana pasada hizo tejidos, costuras, artesanías u otros productos para vender
- El año pasado hizo tejidos, costuras, artesanías u otros productos para vender
- La semana pasada ayudó en la construcción o reparación de otra vivienda
- El año pasado ayudó en la construcción o reparación de otra vivienda
- La semana pasada participó en desfile de modelos, castings de TV, etc
- El año pasado participó en desfile de modelos, castings de TV, etc
- La semana pasada hizo alguna otra actividad por plata
- El año pasado hizo alguna otra actividad por plata
- La semana pasada ayudó a alguien en una actividad para ganar dinero

- El año pasado ayudó a alguien en una actividad para ganar dinero
- Deseos para cuando tenga 18 años
- Realizan al menos una actividad productiva
- Asistencia a la escuela
- Repetición de grado o año en la escuela
- Cantidad de repitencias en grado o año de la escuela

Características del hogar

- Tipo de vivienda
- Material de las paredes exteriores
- Material de los pisos
- Disponibilidad del agua
- Tenencia de la vivienda
- Identifica deficiencia en el material de las paredes
- Identifica deficiencia en el material de los pisos
- Cantidad de NNyA de 5 a 17 en el hogar
- Cantidad de NNyA de 0 a 17 en el hogar
- Clima educativo del hogar (de los mayores de 18 años)
- Identifica hogares con algún miembro perceptor de AUH
- Identifica hogares con percepción de asistencia social

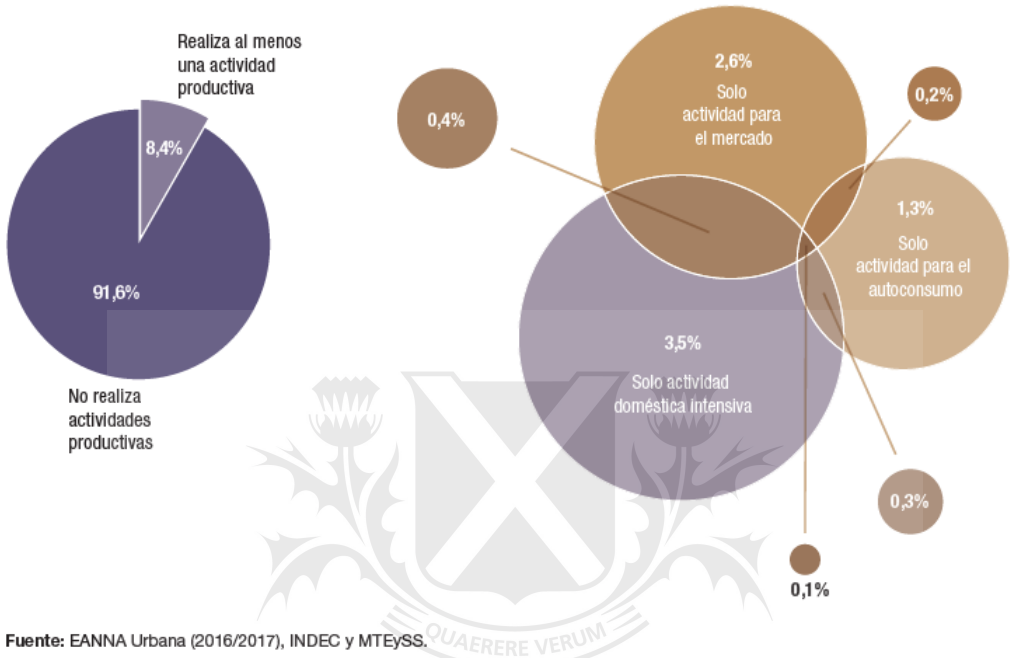
Características del jefe del hogar

- Edad
- Sexo
- Situación conyugal
- Sabe leer y escribir

- Asistencia escolar
- Último grado o año aprobado
- Tipo de primer trabajo
- Hizo changas la semana pasada
- Disponibilidad para trabajar
- Búsqueda de trabajo en el mes
- Cantidad de ocupaciones*
- Horas de trabajo ocupación principal
- Horas de trabajo ocupaciones secundarias
- Descuento jubilatorio*
- El mes pasado, recibió del Estado dinero en vales para comprar alimentos
- El mes pasado, recibió del Estado dinero en efectivo
- El mes pasado, no recibió dinero del Estado
- Ayuda en alimentos
- Ayuda en ropa/zapatos
- Ayuda en medicamentos
- Ayuda en dinero
- Ayuda de otro tipo
- Ninguna ayuda
- ¿Se reconoce como descendiente o perteneciente a un pueblo indígena?
- Condición de actividad de los adultos
- Categoría ocupacional de los adultos
- Ingreso de la ocupación principal (en pesos)
- Cantidad de personas en el hogar

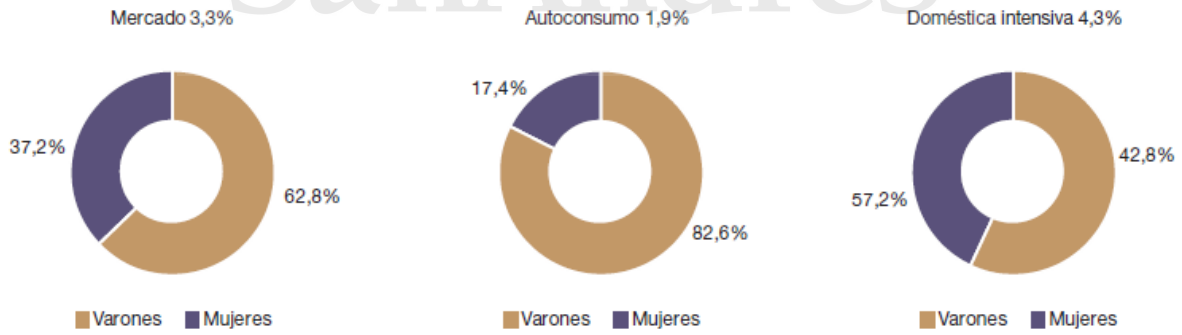
8. Anexo II: gráficos

Figura 9: Participación en actividades productivas. Niños y niñas de 5 a 15 años. Total urbano



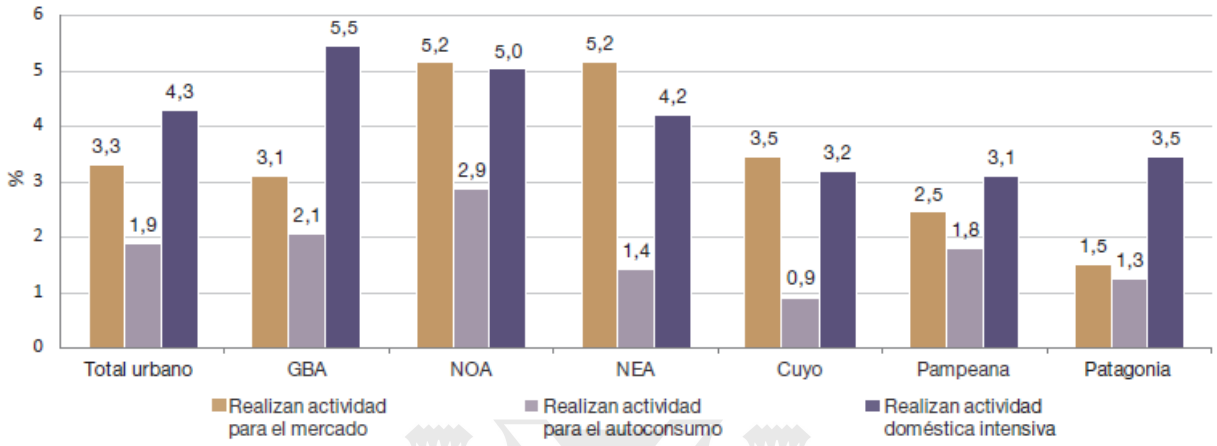
Fuente: EANNA Urbana (2016/2017), INDEC y MTEySS.

Figura 10: Distribución por sexo en actividades productivas. Niños y niñas de 5 a 15 años. Total urbano



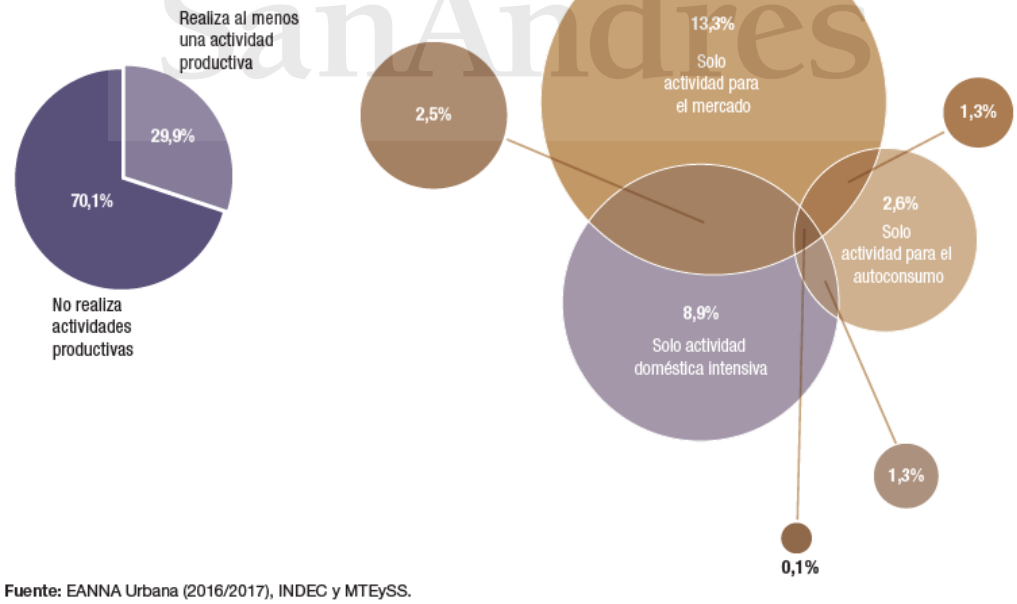
Fuente: EANNA Urbana (2016/2017), INDEC y MTEySS.

Figura 11: Incidencia de las actividades productivas según región. Niños y niñas de 5 a 15 años. Total urbano



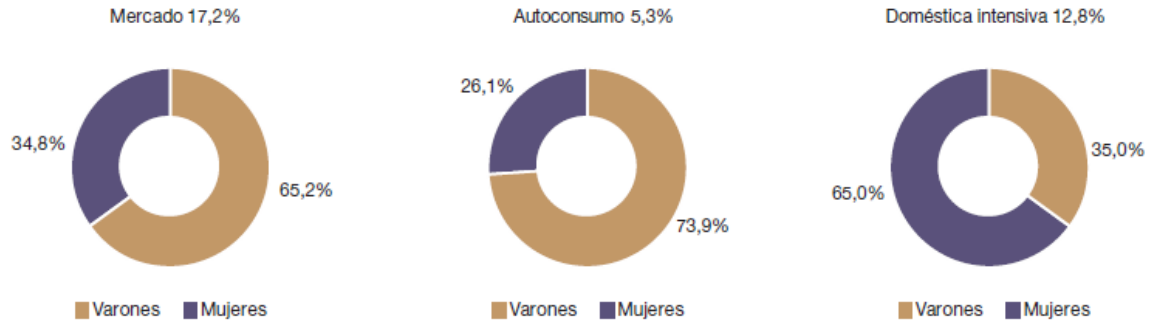
Fuente: EANNA Urbana (2016/2017), INDEC y MTEySS.

Figura 12: Participación en actividades productivas. Adolescentes de 16 y 17 años. Total urbano



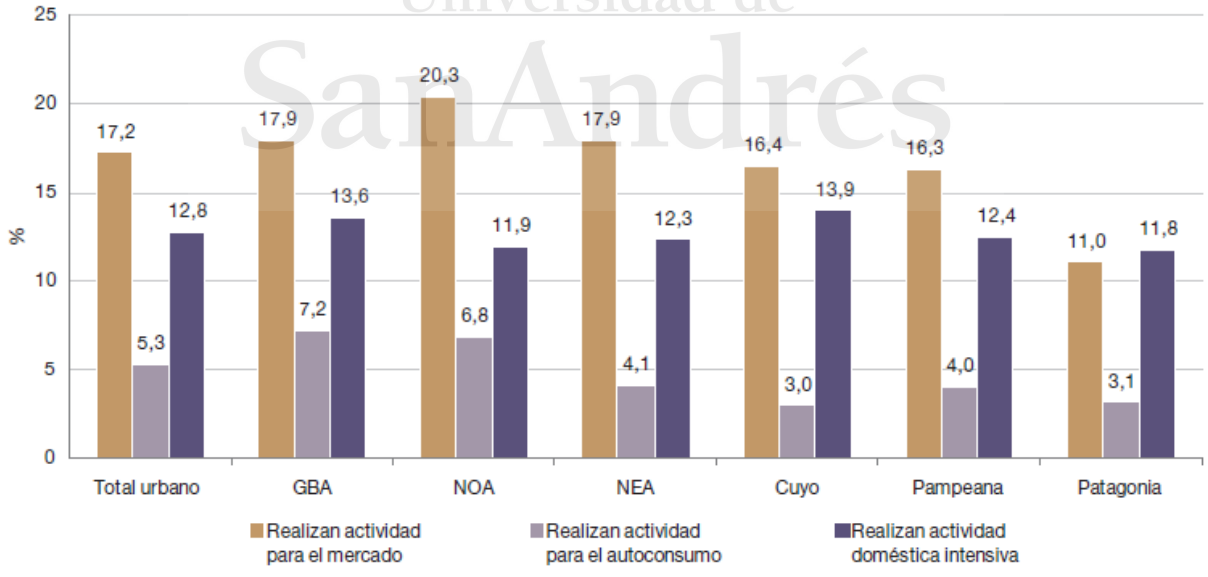
Fuente: EANNA Urbana (2016/2017), INDEC y MTEySS.

Figura 13: Distribución por sexo en actividades productivas. Adolescentes de 16 y 17 años. Total urbano



Fuente: EANNA Urbana (2016/2017), INDEC y MTEySS.

Figura 14: Incidencia de las actividades productivas según región. Adolescentes de 16 y 17 años. Total urbano



Fuente: EANNA Urbana (2016/2017), INDEC y MTEySS.