Universidad de San Andrés

Departamento de Economía

Maestría en Economía

Pobreza en Argentina: un análisis predictivo utilizando herramientas de Machine Learning

Autor: Andrés DABÚS

DNI: 38.919.616

Mentor: Walter SOSA ESCUDERO

Bahía Blanca

21 de diciembre, 2020

Tesis de Maestría en Economía de

Andrés DABÚS

"Pobreza en Argentina: un análisis predictivo utilizando herramientas de Machine Learning"

Resumen

Las mediciones de los niveles de pobreza a partir del ingreso monetario pueden estar afectadas por inconvenientes prácticos como la no respuesta y la subdeclaración del mismo en las encuestas de hogares. Además, muchas críticas sugieren que esta variable puede no representar de manera confiable el nivel de vida y las privaciones de los individuos. Por estas razones, en este trabajo se analiza la posibilidad de identificar a las familias pobres y no pobres a partir de variables alternativas al ingreso, las cuales representan otras dimensiones del bienestar distintas a la monetaria. Implementando diferentes algoritmos de la literatura de Machine Learning y utilizando las bases de datos de la Encuesta Permanente de Hogares (EPH), se encontró que se pudo clasificar correctamente al 84,25% de las observaciones. Asimismo, se muestra que es posible reducir el espacio original de predictores mediante LASSO, aunque no considerablemente, y que, a partir de conditional random forest, las variables más relevantes para determinar el status de pobreza son la cantidad total de miembros del hogar, el tipo de cobertura médica que tiene el jefe de hogar, la cantidad de miembros del hogar con edad mayor o igual a 10 años cumplidos, la edad del jefe de hogar, la categoría de inactividad del jefe de hogar y la cantidad de miembros del hogar menores de 10 años.

Palabras clave: pobreza, algoritmos de aprendizaje supervisado, clasificación, predictores no monetarios

“Poverty in Argentina: a predictive analysis using Machine Learning tools”

Abstract

Measurements of poverty levels based on monetary income can be affected by practical inconveniences such as non-response and underreporting in household surveys. Furthermore, many critics suggest that this variable may not reliably represent the standard of living and deprivations of individuals. For these reasons, this work analyzes the possibility of identifying poor and non-poor families using alternative variables to income, which represent other dimensions of well-being different from monetary. By implementing several algorithms from Machine Learning literature and using databases from the Permanent Household Survey (EPH), it was found that 84,25% of the observations could be correctly classified. Likewise, it is shown that it is possible to reduce the original space of predictors through LASSO, although not considerably, and that, according to Conditional Random Forest, the most relevant variables to determine poverty status are the total number of household members, the type of medical coverage that the head of household has, the number of household members over 10 years old, the age of the head of household, the category of inactivity of the head of household and the number of household members under 10 years old.

Keywords: poverty, supervised learning algorithms, classification, non-monetary predictors

Códigos JEL: C38, C53, I32

Índice

1. Introducción

Existen importantes razones que justifican el estudio y el análisis de la pobreza. Siguiendo a Gasparini et al.(2012), en primer lugar es un fenómeno percibido como un mal en sí mismo, ya que implica una carencia de oportunidades como la imposibilidad de satisfacer necesidades humanas elementales. En segundo lugar, por los potenciales efectos que puede llegar a tener sobre otras variables sociales y económicas relevantes. Por estos motivos, su reducción forma parte de la agenda de los gobiernos la región y de los países en vías en desarrollo en general. Para ello, la correcta identificación de los hogares pobres resulta primordial a la hora de llevar a cabo políticas públicas que busquen combatir esta problemática.

En Argentina el Instituto Nacional de Estadísticas y Censos (INDEC) elabora semestralmente indicadores de pobreza e indigencia utilizando una metodología indirecta también conocida como "de línea". Esta estrategia de medición consiste en establecer un umbral de ingresos a partir del cual se considera que un hogar es capaz de adquirir un conjunto de bienes y servicios (canasta básica total) que le permita satisfacer sus necesidades básicas. De esta forma, se consideran pobres a aquellos hogares en donde el ingreso total familiar no supera el valor de dicha canasta. Es decir, la pobreza se mide sobre la distribución del ingreso. No obstante, este enfoque monetario tiene ciertos problemas prácticos. Fitzpatrick et al. (2018) sostienen que las encuestas nacionales de hogares que recolectan datos de consumo o de ingreso en pos de generar indicadores de pobreza suelen ser complejas lo que, a su vez, puede provocar que la elaboración de este tipo de mediciones sociales se lleve a cabo con poca frecuencia. McBride y Nichols (2018) también sugieren que las encuestas de ingresos y gastos para identificar individuos en condición de pobreza suelen ser más costosas y que su realización demanda mucho tiempo.

Otro problema asociado a este tipo de encuestas es la no respuesta. Esto se refiere al hecho de que algunas personas o bien deciden no contestar algún ítem del cuestionario[1] (no respuesta

[1]En el caso de la no respuesta parcial, los encuestados, típicamente, se rehúsan a contestar preguntas relacionadas con el ingreso (Gasparini et al., 2012).

parcial), o deciden no contestarlo en su totalidad (no respuesta total).[2] En principio, esto no representaría un problema relevante siempre y cuando se dé de manera aleatoria. Sin embargo, este fenómeno suele estar correlacionado con variables socioeconómicas relevantes (Groves y Couper, 1998). Salvia y Donza (1999) sugieren que en la Encuesta Permanente de Hogares (EPH) las personas perceptoras de ingresos altos tienden a no responder y, por ende, a estar subrepresentadas en la muestra. Asimismo, esto se corresponde con que en la práctica los encuestadores encuentran que las personas de mayores recursos tienden a estar menos predispuestas y a ser más temerosas de colaborar con el encuestador (Gasparini et al., 2012). Naturalmente, si no se realiza ningún tipo de corrección, los indicadores resultantes sobreestimarían la pobreza.

Estas mediciones también suelen estar afectadas por el hecho de que las personas son propensas a subdeclarar sus niveles de ingreso y consumo en las encuestas. Son dos las principales causas de este fenómeno: la dificultad de las encuestas para abarcar todas las fuentes de ingreso y la subdeclaración deliberada de los niveles de ingreso por parte de los encuestados.[3] A diferencia de la no respuesta, esta cuestión suele ser más delicada debido a que no es posible identificar quién incurrió en tal hecho. En cambio, en la no respuesta se sabe con exactitud qué individuos se negaron a colaborar. Asimismo, se estima que el grado de subdeclaración está positivamente correlacionado con el nivel de ingreso (Gasparini et al., 2012), lo que afectaría su valor medio, y por tanto el nivel de pobreza estimada. De este modo, dado que este fenómeno se mide sobre la distribución del ingreso, la subdeclaración provocaría estimaciones sesgadas de pobreza.

Si bien existen ciertos procedimientos que intentan aliviar estos problemas, no brindan soluciones concluyentes. Para la no respuesta parcial se suelen utilizar factores de expansión que le asignan un peso mayor a aquellas observaciones con menor probabilidad de respuesta en pos de corregir el sesgo muestral. Otros métodos alternativos son los de imputación, mediante los cuales se busca estimar el ingreso de los individuos que no lo reportaron, pero

[2]El INDEC da cuenta del "creciente deterioro de la respuesta de los hogares sobre sus ingresos monetarios"durante el reciente período 2007-2015 en el siguiente documento técnico `https://www.indec.gob.ar/ftp/cuadros/sociedad/nota_EPH_ingresos_06_17.pdf`.

[3]La dificultad de abarcar todas las fuentes de ingreso se da, principalmente, porque los individuos no las suelen recordar con exactitud (sobre todo, cuando se trata de ingresos por trabajos esporádicos y de ingresos por rentas de capital).

que sí respondieron el resto de las preguntas de la encuesta, a partir del ingreso de personas parecidas que respondieron todo el cuestionario (Gasparini et al., 2012). No obstante, el problema con este procedimiento es que el ingreso depende de factores no observables, de modo que se podría ver afectada la calidad y confiabilidad de dichas estimaciones. Por otro lado, las soluciones para la no respuesta total están menos exploradas.

Los procedimientos para afrontar el problema de la subdeclaración son mucho menos claros y no hay consenso en cuanto a si es conveniente aplicarlos. El ajuste más conocido consiste en comparar y emparejar el total de los ingresos por cada fuente de las Cuentas Nacionales con un agregado computado a partir de la encuesta de hogares. Sin embargo, el problema radica en que usualmente el ajuste considera que la subdeclaración es uniforme en cada fuente de ingreso y, dado que existe una relación positiva entre la magnitud de la subdeclaración y el nivel de ingresos, dicho ajuste podría provocar que se subestime la pobreza.

Por otra parte, existen ciertas críticas que sugieren que el ingreso puede no reflejar de manera confiable el nivel de bienestar y las privaciones de los individuos. Una de ellas tiene que ver con su capacidad para captar el nivel de vida de las personas mayores retiradas del mercado laboral (Gasparini et al., 2012). Muchas veces estas suelen tener ingresos bajos, pero mantienen un nivel de vida más alto a través de la liquidación de ahorros. Es decir, mediante el enfoque monetario de la pobreza estas personas podrían llegar a ser consideradas pobres cuando en realidad gozan de un bienestar mayor. Otra cuestión está asociada a la alta volatilidad que suele tener esta variable. Si está muy influida por factores estacionales, es posible que al momento de ser reportada algunas personas sean clasificadas como pobres cuando en realidad no tienen carencias en términos de educación, vivienda y otras variables representativas del nivel de bienestar, ya que sus ingresos de largo plazo pueden ser más altos.

Para sortear estos problemas asociados a dicha variable (el ingreso), esta tesina tiene como objetivo, en primer lugar, analizar en qué medida es posible clasificar a los hogares pobres y no pobres a partir de variables alternativas que captan dimensiones del bienestar distintas a la monetaria. En otras palabras, si se puede lograr en gran parte una correcta y precisa identificación de aquellos hogares en condición de pobreza utilizando predictores no monetarios, esto permitiría en cierta medida prescindir del ingreso y así sortear los problemas de la

subdeclaración y de la no respuesta.[4] En particular, el conjunto de predictores utilizados reflejan distintos aspectos del nivel de vida como las condiciones habitacionales y de vivienda, el acceso a servicios públicos, características sociodemográficas, educativas y diferentes cuestiones que refieren al estado de actividad de los individuos y a la formalidad laboral. Es importante aclarar que estos son, naturalmente, menos propensos a sufrir los problemas de la no respuesta y de la subdeclaración, ya que, por un lado, no representan información tan sensible como el ingreso. Por ejemplo, no es difícil notar que es probable que resulte menos incómodo o comprometedor para el encuestado responder una pregunta acerca de una condición de la vivienda, como el material del cual están hechos los pisos, que contestar acerca del monto de ingresos que percibe. Por otro lado, es información que resulta más difícil de olvidar para los individuos y, por lo tanto, es menos probable que subdeclaren por este motivo. A modo de ilustración, es más factible que el respondiente sepa contestar con exactitud cuál es el número de ambientes que tiene su vivienda que recordar cuáles fueron sus ingresos percibidos por trabajos esporádicos o por rentas de capital.

En segundo lugar, se llevan a cabo dos ejercicios que complementan al primer objetivo. Por un lado, se busca establecer si se puede reducir el espacio original de características a un subgrupo de variables que contengan los datos verdaderamente relevantes para realizar dicha clasificación. En términos del propósito de este trabajo esto resulta importante, ya que, de ser posible, permitiría desarrollar modelos más parsimoniosos e interpretables sin perder información valiosa. Asimismo, se realiza un ordenamiento de la importancia de los predictores.

A partir de las finalidades mencionadas en los dos párrafos anteriores, este trabajo contribuye al desarrollo de medidas de pobreza para Argentina que no solamente sortean los mencionados problemas de la metodología tradicional basada en el ingreso, sino que son más abarcativas al tener en cuenta aspectos multidimensionales que hacen al bienestar humano. A su vez, también contribuye a determinar cuáles de estas variables son relevantes para la elaboración de dichas mediciones.

[4] Si no se utiliza el ingreso para determinar qué hogares son pobres y cuáles no, se evitarían, necesariamente, los inconvenientes prácticos y metodológicos expuestos anteriormente.

La metodología utilizada para tales fines es un conjunto de herramientas brindadas por la literatura de machine learning, ya que presentan grandes ventajas cuando se trata de tareas de tipo clasificatorias. Estas consisten, particularmente, en algoritmos de aprendizaje supervisado. Para el primer objetivo se analiza y se compara el poder predictivo de dos modelos logísticos regularizados por ridge y LASSO, de *random forest*, *conditional random forest* y *support vector machines*. Para reducir la cantidad de variables se presentan los predictores seleccionados por LASSO, mientras que la importancia de las variables se analiza a partir de random forest y de conditional random forest. Además, los datos utilizados corresponden a las bases de microdatos de la EPH del INDEC para el período 2016-2019.

El trabajo se organiza de la siguiente manera. En la siguiente sección se realiza una revisión de la literatura existente sobre distintas aplicaciones de aprendizaje estadístico para predecir la condición de pobreza. La sección 3 explica la metodología general a seguir en cada modelo, mientras que en la sección 4 se detalla cómo funcionan los algoritmos implementados. La sección 5 explica cuáles son los datos utilizados, en la sección 6 se exponen los principales resultados obtenidos y, por último, en el apartado 7 se presentan las reflexiones finales.

2. Revisión de la literatura

Si bien las técnicas de aprendizaje automático han sido extensamente aplicadas en otras disciplinas, esto no ha sido así en el caso de la economía, en donde la literatura es relativamente reciente y escasa. Típicamente, en los problemas predictivos o clasificatorios las herramientas tradicionales de la econometría involucran algún tipo de modelo de regresión lineal o logístico en los cuales se prioriza el ajuste adentro de la muestra. En cambio, machine learning se enfoca en mejorar la capacidad predictiva afuera de la muestra y se destaca por capturar relaciones no lineales entre los predictores y por identificar patrones generales entre los datos (Varian, 2014). En este sentido, se detallan a continuación algunos trabajos que han aplicado estas técnicas para estudiar la problemática de la pobreza en distintos contextos.

Sohnesen y Stender (2017) utilizan datos de gastos de consumo de Malawi, Ruanda, Etiopía, Uganda, Albania y Tanzania para predecir la pobreza rural y urbana a través de random

forest y de una variante de la regresión lineal. Estos encuentran que el primer método tiene un error de predicción menor para ambos tipos de pobreza. Thoplan (2014) también implementa este algoritmo con el objetivo de mejorar la precisión de la clasificación de aquellas personas en situación de pobreza en Mauritania. Utilizando datos censales de este país para el año 2000, encuentra que las principales variables que predicen la pobreza son las horas trabajadas por semana, la edad y el máximo nivel educativo alcanzado. Asimismo, descubre una brecha con respecto al género, ya que las mujeres tienen más chances de ser clasificadas como pobres en comparación con los hombres.

Por su parte, McBride y Nichols (2018) muestran que las técnicas de machine learning, en particular random forest y random forest por cuantiles, pueden mejorar el desempeño predictivo afuera de la muestra de los *proxy mean tests* (PMT). Kambuya (2020) también busca mejorar la eficiencia y la precisión de los PMT en términos de selección de variables y de error de predicción afuera de la muestra. Utilizando datos de la encuesta de hogares de Tailandia para el año 2016, aplica random forest, LASSO (abreviación en inglés de *Least Absolute Shrinkage and Selection Operator*) y el método de regresión *stepwise*. Como resultado, encuentra un trade-off entre el error de exclusión (hogares pobres clasificados como no pobres) y de inclusión (hogares no pobres clasificados como pobres) según el método que se utilice para seleccionar variables. Mientras random forest presenta una tasa menor del primer tipo de error, las otras técnicas presentan una tasa más del último. En igual sentido, el trabajo de Otok y Seftiana (2014) tiene como objetivo aumentar la eficacia y la eficiencia de los programas de asistencia del gobierno de Jombang, Indonesia. Utilizando CART (en inglés Classification and Regression Trees) y random forest encuentran que este último método es más preciso a la hora de clasificar hogares pobres y hogares pobres de manera crónica. Asimismo, Kshirsagar et al.(2017) muestran a través de una regresión logística regularizada (elastic net logistic regression) que se puede mejorar la predicción afuera de la muestra de los PMT para el caso de Zambia.

Por otra parte, Afzal, Hersh, y Newhouse (2015) comparan el desempeño predictivo afuera de la muestra de modelos construidos manualmente contra aquellos en los que las variables fueron seleccionadas algorítmicamente para Pakistán y Sri Lanka. En particular, en el primer caso se utilizaron modelos construidos de antemano a partir de cierta información previa

acerca del proceso generador de datos. En el segundo caso se construyeron modelos a partir de las variables seleccionadas por el algoritmo LASSO y por el algoritmo forward stepwise. Los autores también evalúan si mejora la precisión de los modelos al añadir información obtenida de imágenes satelitales. A partir de estas obtuvieron indicadores de la proporción de tierra utilizada y ocupada, la elevación del terreno, la densidad de la población, de la actividad lumínica de noche, entre otras. Como principal resultado encontraron que el estimador LASSO produce mejoras sustanciales en el error predictivo en comparación con los otros métodos utilizados. Además, sugieren que el uso de imágenes satelitales puede mejorar las predicciones cuando los datos son más escasos, como es el caso de Sri Lanka. Rincón (2019) utiliza técnicas de machine learning con el objetivo de mejorar la frecuencia de las estimaciones de pobreza en México. Para ello, evalúa y compara el desempeño y el poder predictivo de la regresión logística tradicional contra algoritmos como random forest, análisis discriminante lineal (ADL) y support vector machines. Este encuentra que el método que menor error de predicción presenta es random forest, mientras que el que peor desempeño tiene es el ADL. Además, sugiere que las variables que más inciden en la pobreza son el ingreso laboral, el número de miembros por hogar y la ubicación rural geográfica, mientras que los predictores menos importantes son el género, el nivel educativo y el acceso a un plan de cobertura médica.

Naturalmente, por la heterogeneidad de las variables empleadas en este trabajo las cuales representan distintas dimensiones del bienestar, existe una contundente relación con la literatura de la pobreza multidimensional. Caruso, Sosa Escudero, y Svarc (2015) aplican el método de *clusters* para establecer cuáles son las dimensiones más relevantes que determinan la condición de pobreza. Utilizando los datos de la encuesta mundial de Gallup (*Gallup World Poll*), en primera instancia logran identificar a las personas pobres utilizando todo el espacio de características. Luego, siguiendo la estrategia de Fraiman et al. (2008) reducen la dimensionalidad del mismo seleccionando el mínimo número de variables que reproduzcan la clasificación hecha en el primer paso lo más precisamente posible. Los autores encuentran que las tres variables más relevantes para clasificar a las personas pobres y no pobres son el ingreso mensual de los hogares, no haber dispuesto de dinero suficiente para comprar alimentos al menos en tres ocasiones distintas durante el último año y la presencia de una computadora en el hogar. Gasparini et al.(2013) también utilizan los datos de la encuesta de Gallup

para estudiar las dimensiones más importantes de la pobreza en Latinoamérica y el Caribe. Estos autores encuentran que las más relevantes son tres: el ingreso monetario, el bienestar subjetivo y las necesidades básicas (por ejemplo, agua y electricidad). Asimismo, establecen que dichos resultados sugieren que el ingreso es relevante cuando se busca determinar el nivel de vida, pero que, no obstante, el bienestar es un fenómeno multidimensional que no puede ser captado completamente por esta única variable. Alkire y Santos (2010), a partir de datos de encuestas de hogares, estiman un nuevo índice de pobreza multidimensional para 104 países en vías de desarrollo, el cual está compuesto por 10 indicadores que se corresponden con las tres dimensiones que componen al índice de desarrollo humano: la salud, la educación y el estándar de vida. Por su parte, Battiston et al. (2013) estudian la pobreza para cinco países de Latinoamérica a partir de seis dimensiones distintas del bienestar representadas por el ingreso, la asistencia escolar de los niños, la educación del jefe de hogar, la higiene, el acceso al agua y el albergue de los individuos.

Cardinale Lagomarsino et al. (2016) aplican random forest para Argentina usando datos de la EPH pertenecientes al período 2003-2015 con el objetivo de mejorar la capacidad predictiva y de determinar cuáles son los principales predictores no monetarios de la pobreza para dicho país. Estos encuentran que son la cantidad de miembros del hogar, la edad del principal sostén del hogar, la cobertura médica y el nivel de educación. Fitzpatrick et al. (2018) realizaron un reporte para el Banco Mundial en donde predicen la pobreza para Malawi e Indonesia utilizando y comparando distintos algoritmos. Por un lado, utilizan algunos más convencionales como random forest, support vector machines, regresión logística, análisis discriminante lineal y k- vecinos cercanos. Por otro lado, también aplican otras técnicas más avanzadas a esta tarea predictiva como los algoritmos *crowd-sourced*, algoritmos combinatorios, *deep-learning* y *automated machine learning*. En línea con el anterior trabajo, procuran utilizar solamente como predictores variables categóricas simples de recolectar, dejando de lado los predictores monetarios. Estos sugieren que, si bien no se destaca ningún algoritmo en particular, las técnicas de machine learning tienen una mayor precisión a la hora de clasificar e identificar a los hogares en condición de pobreza en comparación con el análisis de regresión tradicional.

Por último, es interesante mencionar que a principios del 2018 el Banco Mundial llevó a cabo

una competencia llamada "Pover-t Test: Predicting Poverty". El objetivo de la misma era que científicos de datos de todo el mundo desarrollen algoritmos para predecir el status pobreza utilizando datos de consumo a nivel hogar e individual para tres países en vías de desarrollo en pos de dotar de estas herramientas a dicha institución para combatir esta problemática.[5]

El presente trabajo busca hacer una contribución a esta literatura reciente. Si bien el enfoque es similar al de Cardinale Lagomarsino et al. (2016), existen ciertas diferencias con respecto a esta investigación. En primer lugar, se explora el potencial predictivo de otros algoritmos, aparte de random forest. En segundo lugar, como bien se mencionó, los autores emplearon random forest para analizar la importancia de las variables. En cambio, en este trabajo también se emplea conditional random forest, ya que, como se explica más adelante, resulta ser un método más robusto para esta tarea particular. Por último, también se presentan las variables seleccionadas por LASSO para analizar en qué medida es posible reducir el espacio original de estas a un conjunto más chico.

3. Metodología

3.1. Procedimiento general

Los procedimientos a aplicar son los que brinda la literatura de machine learning. Son varias las diferencias que este paradigma guarda con respecto al enfoque tradicional de la econometría frecuentista. Por un lado, este último se concentra en estimar un modelo, usualmente representado por $y_i = x_i'\beta + \mu_i$, exógenamente dado en dónde la relación entre la variable dependiente y_i y el vector regresores x_i' está determinada por una teoría o una "estructura" (Sosa Escudero, 2018). Es decir, se busca *estimar* de la mejor manera posible a los coeficientes β en donde la calidad del estimador suele estar asociada con ciertas propiedades deseables. Generalmente, suele existir una preferencia lexicográfica por la insesgadez, mientras que la eficiencia del estimador suele tener una importancia secundaria. En cambio, el paradigma del aprendizaje estadístico o automático tienen como objetivo *predecir*

[5] https://www.drivendata.org/competitions/50/ worldbank-poverty-prediction/page/99/

y en base a x, donde el modelo en sí mismo no tiene un papel relevante. En particular, se busca predecir correctamente afuera de la muestra, es decir, evaluar la capacidad predictiva en observaciones que no fueron utilizadas para construir el modelo. Dado que el desempeño predictivo puede mejorar sustancialmente al tolerar cierto sesgo haciendo que disminuya la varianza, la preferencia por métodos sesgados suele ser otra diferencia crucial con respecto al enfoque de la econometría clásica (Sosa Escudero, 2018). Esto se refleja en el grado de complejidad del modelo: modelos más complejos tienden a ser menos sesgados, pero a tener una mayor varianza y a ser más erráticos, mientras que una menor complejidad permite bajar la varianza a costa de un mayor sesgo. Típicamente, suele existir un hiperparámetro que controla dicha complejidad de manera que se maximice la precisión de las predicciones de acuerdo a una función de pérdida. Asimismo, machine learning también resulta útil cuando se busca entender o encontrar patrones generales entre los datos. Siguiendo a Plulikova (2016), existen dos tipos de algoritmos de aprendizaje estadístico que generalmente se utilizan para este tipo de tareas.

1. **Algoritmos de aprendizaje no supervisado**: buscan encontrar algún tipo de estructura entre los datos sin presumir de antemano que exista. Trabajan solamente a partir de un espacio de características sin ninguna variable de salida o de respuesta.
2. **Algoritmos de aprendizaje supervisado**: presume cierto patrón inicial entre los datos y, típicamente, trabajan con variables predictoras o de características y con variables de respuesta. Suelen ser utilizados para tareas de tipo predictivas.

El problema de clasificación binaria del presente trabajo se encuentra dentro del grupo de técnicas de aprendizaje supervisado. Para ello, se siguen los siguientes pasos generales para cada algoritmo utilizado.

1. Se entrenan a los modelos con los datos pertenecientes al período 2016-2018 y se elige el valor óptimo de sus respectivos parámetros a través de validación cruzada. De esta

manera, se optimiza la complejidad de los mismos evitando el sobreajuste.[6]

2. Una vez entrenados, se utilizan los datos del año 2019 como base de prueba. De esta forma, se evalúa en qué medida estos modelos pueden clasificar correctamente a este conjunto de observaciones que no fue utilizado para su entrenamiento.[7]

3. Posteriormente, se computan una serie de métricas tradicionales indicativas de la capacidad predictiva de cada método detalladas a continuación.

Cabe resaltar que para estos procedimientos se utilizaron distintas funciones del software de uso gratuito R.

3.2. Medidas de desempeño predictivo

Existen distintas formas de evaluar y comparar el desempeño clasificatorio de este tipo de modelos. Dado que la variable de respuesta es binaria, existen cuatro escenarios posibles: que el modelo prediga que un hogar es pobre cuando lo es (verdadero positivo), que prediga que es pobre cuando no lo es (falso positivo), que prediga que no es pobre cuando no lo es (verdadero negativo) y que prediga que no es pobre cuando el hogar lo es (falso negativo)[8]. Equivalentemente, los casos de falsos positivos y falsos negativos representan el error de tipo 1 y de tipo 2 respectivamente. Estos escenarios conforman lo que se conoce como matriz de confusión. Siguiendo a Fitzpatrck, algunas de las métricas de desempeño predictivo más comunes son las siguientes.

- **Tasa de aciertos:** mide la proporción del total de las observaciones que fueron correc-

[6]Los umbrales de decisión para los respectivos modelos también pueden ser pensados como hiperparámetros adicionales a optimizar. Por ello, estos fueron elegidos de manera tal que maximicen el desempeño predictivo de los modelos (en particular, la tasa de aciertos tal como se explica en la subsección 3.2). Además, si bien los únicos modelos probabilísticos son las regresiones logísticas regularizadas, es sencillo obtener las probabilidades predichas cuando se trata de clasificadores discretos como en el caso de los modelos de árboles (Fawcett, 2006).

[7]La base de entrenamiento contiene aproximadamente al 70 % de los datos mientras que en la de prueba se encuentra el 30 % restante.

[8]Respectivamente, estos casos se representan mediante sus abreviaciones: VP, FP, VN y FN.

tamente predichas. Se calcula como

$$\frac{VP + VN}{VP + FP + VN + FN}$$

- **Sensibilidad o tasa de verdaderos positivos:** es la proporción de hogares pobres clasificados correctamente en relación con el total de hogares que pertenecen a esta categoría.

$$\frac{VP}{VP + FN}$$

- **Especificidad o tasa de verdaderos negativos:** mide la proporción de hogares no pobres clasificados correctamente en relación con el total de observaciones que pertenecen a esta categoría.

$$\frac{VN}{VN + FP}$$

- **Precisión:** de todas las observaciones clasificadas como pobres, mide que proporción fue clasificada correctamente.

$$\frac{VP}{VP + FP}$$

- **F1:** esta medida comúnmente se utiliza cuando se calculan la precisión y la sensibilidad. Tiene rango $[0, 1]$ y es la media armónica entre estas métricas. Intuitivamente, puede pensarse como una media aritmética, pero con la diferencia de que siempre va a tomar valores más cercanos al componente más chico. Es decir, si la precisión es alta pero la sensibilidad es baja, entonces el F1 va a ser bajo y lo mismo para el caso contrario. En cambio, cuando ambos componentes tienen valores más parecidos, esta medida suele parecerse más a una media aritmética.

$$\frac{2VP}{2VP + FP + FN}$$

- **Área bajo la curva ROC:** la curva ROC (abreviación en inglés de "*receiver operating characteristic*") o curva de características operativas del receptor, es una representación gráfica de la tasa de verdaderos positivos frente a la tasa de falsos positivos para

distintos umbrales de clasificación. Intuitivamente, esta curva mide la capacidad de un modelo para distinguir correctamente casos ciertos de "falsas alarmas". El área bajo esta curva, con rango $[0, 1]$ es indicativo del grado de dicha separabilidad. Por ende, mientras mayor sea el área de esta región, mejor clasifica el modelo.

En un escenario ideal el número de falsos positivos y falsos negativos sería igual a cero. No obstante, dado que esto es imposible de lograr, en la práctica se suele priorizar la minimización del error de tipo 1 o de tipo 2 dependiendo del contexto y de cada problema en particular. De esta preferencia también va a resultar conveniente prestar más atención a unas u otras medidas. Si el costo de cometer errores de tipo 1 es alto, el modelo preferido será aquel cuyas medidas de especificidad y precisión sean más altas. Por el contrario, si resulta más importante minimizar la tasa de falsos negativos, el mejor modelo va a ser el que tenga una medida de sensibilidad más alta. Una cuestión a tener en cuenta en estos casos es que, lógicamente, disminuir un tipo de error implica necesariamente aumentar el otro. Por ejemplo, es fácil ver que un modelo que clasifique a todas las observaciones como pertenecientes a la categoría 1 tendría una sensibilidad igual a 1 pero una baja precisión y una especificidad nula.

Por otro lado, cuando no existe una diferencia grande entre el costo que se le asigna a uno u otro tipo de error y, por lo tanto, interesa clasificar correctamente a ambas clases se suele prestar más atención a la tasa de aciertos y al índice F1. A diferencia de las métricas anteriores que dependen de los falsos positivos o de los falsos negativos, estas medidas están afectadas por ambos tipos de error. Sin embargo, a la hora de utilizarlas hay que tener en cuenta ciertas consideraciones. En primer lugar, la tasa de aciertos es sensible al desbalanceo muestral y sus valores pueden resultar engañosos a la hora de evaluar un modelo en estos casos. Por lo tanto, este indicador debería utilizarse únicamente cuando las clases están balanceadas. En segundo lugar, el índice F1 no tiene este problema, pero le asigna igual peso a las dos medidas que componen esta media armónica. En caso de que se busque darle más importancia a la precisión o a la sensibilidad, debería computarse un índice F1 ponderado.

Dado que en este trabajo interesa clasificar correctamente a los hogares pobres y no pobres y que, en principio, no hay una preferencia por minimizar un tipo de error específico, serían apropiados tanto el índice F1 como la tasa de aciertos. Sin embargo, se opta por utilizar

esta última como principal medida de desempeño predictivo dada su fácil interpretabilidad. Aunque la muestra original presenta cierto desbalanceo entre las clases, en la sección 5 se detalla cómo se aborda este problema para que los valores de dicha medida sean confiables.

4. Algoritmos

4.1. Regresión logística

En primer lugar, se explica cómo se lleva a cabo una tarea de clasificación utilizando esta herramienta más convencional. La regresión logística suele ser utilizada para clasificar observaciones en dos categorías. Específicamente, es un modelo para la probabilidad condicional de ocurrencia de Y dado un conjunto de predictores X:

$$p \equiv Pr(Y = 1|X)$$

Donde el modelo no lineal para p tiene la siguiente forma

$$p = F(\mathbf{x}'\beta) = \frac{e^{\mathbf{x}'\beta}}{1 + e^{\mathbf{x}'\beta}}$$

β es el vector de coeficientes y x' representa a los predictores. Típicamente, β es estimado a través del método de máxima verosimilitud. Este consiste en estimar los valores de dichos parámetros de manera tal que la probabilidad predicha para cada observación, $\hat{p}(x_i)$, a través de la función logística se asemeje lo más posible al valor real observado de los mismos. En otras palabras, buscamos que las estimaciones de los parámetros $\hat{\beta}$ sean tales que al introducirlos en el modelo dado en la ecuación (2), arroje una probabilidad cercana a 1 para aquellas observaciones pertenecientes a la clase representada por este valor, y un valor cercano a cero para aquellas observaciones que pertenecen a la otra clase (James et al., 2013). Formalmente, en este caso implicaría elegir los valores de $\hat{\beta}$ tal que maximicen la siguiente función de

verosimilitud:

$$L(\beta) = \prod_{i:y_i=1} p_i \prod_{i:y_i=0} (1-p_i) = \prod_{i=1}^{n} p_i^{y_i}(1-p_i)^{1-y_i} \tag{1.1}$$

Para obtener dichos valores de β, se aplica logaritmos a la función de verosimilitud.

$$l(\beta) = \sum_{i=1}^{n} \{y_i \ln p_i + (1-y_i) \ln (1-p_i)\} \tag{1.2}$$

$$l(\beta) = \sum_{i=1}^{n} \left\{ y_i x_i' \beta - \ln (1 + e^{x_i' \beta}) \right\} \tag{1.3}$$

Luego, se deriva con respecto a β y se iguala a cero.

$$\frac{\partial l(\beta)}{\partial \beta} = \sum_{i=1}^{n} x_i (y_i - p_i) = 0 \tag{1.4}$$

No obstante, este resulta ser un sistema de K ecuaciones no lineales en β, por lo que no se se puede obtener una solución explícita. Por ello, para resolverlo sería necesario aplicar el método de Newton-Raphson. [9]

Una vez obtenidas las predicciones de las probabilidades individuales, se utiliza un umbral de decisión μ a partir del cual se determina si una observación pertenece a una clase o a la otra [10]. De esta forma, si un individuo tiene una probabilidad predicha mayor a μ, es clasificada como pobre. En cambio, si es menor, es clasificada como no pobre:

$$\hat{Y}_i = 1[\hat{p}_i > \mu]$$

Es menester destacar que, a diferencia de los otros dos algoritmos (*random forest*, *conditional*

[9] Ver Hastie, Tibshirani, y Friedman (2009), p.120.

[10] Tal como se ha mencionado en la nota al pie 6, μ se elige de forma tal que se maximice la tasa de aciertos.

random forest y *support vector machines*) este modelo logístico es lineal en el sentido de que no incorpora términos de interacción, términos polinómicos o algún tipo de transformación de los predictores. En cambio, los otros dos métodos tienen la capacidad de capturar este tipo de relaciones no lineales de manera automática (Mullainathan y Spiess, 2017).

4.1.1. Métodos de regularización

La complejidad de este modelo viene dada por la cantidad de predictores que utiliza. Incorporar todos ellos llevaría a que el mismo sobreajuste a los datos y que, por ende, tenga un desempeño predictivo afuera de la muestra relativamente pobre. Por lo general, existen dos formas de controlar el grado de complejidad de manera que mejore la capacidad predictiva y la intepretabilidad del mismo. Una de estas alternativas consiste en restringir el número de variables eligiendo un subconjunto de los p predictores originales a través del método *best subset selection* o *stepwise*. Sin embargo, estas técnicas suelen tener ciertas desventajas. El primero de estos métodos suele ser inviable computacionalmente, ya que el algoritmo tiene que estimar todos lo subconjuntos de modelos posibles y elegir el que minimice el error de predicción afuera de la muestra por validación cruzada. Esto implicaría que para p predictores se tengan que estimar 2^p modelos. [11] Por su parte, los métodos *stepwise* no tienen asociada esta desventaja computacional, ya que este algoritmo tiene que estimar un número de modelos más reducido. Específicamente, para p predictores tiene que estimar $\frac{p(p+1)}{2}$ modelos.[12] No obstante, no garantiza que el modelo óptimo sea encontrado.

La otra alternativa para controlar el grado de complejidad consiste en implementar métodos de regularización. Dadas las desventajas asociadas a las técnicas de selección de variables recién mencionadas, este trabajo aplica este segundo enfoque. Particularmente, se emplean los métodos de ridge y LASSO detallados a continuación.

Ridge: por lo expuesto anteriormente, sabemos que los coeficientes de la regresión logística se obtienen maximizando el logaritmo de la función de verosimilitud expresada en la fórmula

[11] Por ejemplo, para 17 predictores se tienen que estimar 131072 modelos
[12] Por ejemplo, para 17 predictores *stepwise* tiene que estimar 153 modelos

(1.2). Los coeficientes de Ridge en este caso se obtienen de maximizar una función parecida que es la siguiente:

$$l^R(\beta) = \sum_{i=1}^{n} \left\{ y_i x_i' \beta - \ln\left(1 + e^{x_i' \beta}\right) \right\} - \lambda \sum_{j=2}^{p} \beta_j^2 \tag{1.5}$$

Donde el segundo término es conocido como "penalidad de contracción" (en inglés *shrinkage penalty*) o término de regularización. Esto resulta ser un problema de maximización restringida donde el primer término penaliza la falta de ajuste adentro de la muestra y el segundo penaliza la cantidad de predictores utilizados. λ es el hiperparámetro que regula el peso relativo de estas penalidades y la complejidad del modelo cuyo valor óptimo, típicamente, se selecciona mediante validación cruzada. Es fácil ver que cuando $\lambda \to 0$, los coeficientes de Ridge que maximizan esta función tienden a ser los mismos que los de la regresión logística. En cambio, cuando $\lambda \to \infty$, los coeficientes tienden a cero. aunque por la forma que tiene el segundo término estos nunca llegan a ser exactamente iguales a cero. Por lo tanto, a diferencia de *best subset selection*, *stepwise regression* y LASSO, ridge siempre va a incluir a los p predictores originales. Cabe resaltar que este "encogimiento" hacia cero de los coeficientes no se aplica al término intercepto.

La validación cruzada se llevó a cabo con $k = 10$ particiones aleatorias mediante la función *cv.glmnet* y, luego, el modelo fue entrenado con el lambda óptimo a través de la función *glmnet*. Ambas se encuentran en el paquete *glmnet*.

Group LASSO (abreviación en inglés de *Least Absolute Shrinkage and Selection Operator*): en primer, lugar se explica en que consiste la regularización por LASSO. Este es un método de regularización alternativo que, a diferencia de ridge, es considerado una manera formal y algorítmica de seleccionar variables. Sus coeficientes se obtienen resolviendo un problema de maximización restringida diferente:

$$l^L(\beta) = \sum_{i=1}^{n} \left\{ y_i x_i' \beta - \ln\left(1 + e^{x_i' \beta}\right) \right\} - \lambda \sum_{j=2}^{p} \mid \beta_j \mid \tag{1.6}$$

Al igual que en el caso anterior, el primer término penaliza la falta de ajuste adentro de la muestra, el segundo penaliza la incorporación de predictores y λ es el parámetro que regula la complejidad del modelo. Cuando $\lambda \to 0$, los coeficientes obtenidos son los mismos que el modelo logístico original y cuando $\lambda \to \infty$ los coeficientes tienden a cero. Sin embargo, para un λ suficientemente grande algunos coeficientes pueden efectivamente llegar a ser iguales a cero. En el caso extremo el modelo solamente incluiría el término intercepto, ya que este encogimiento no aplica a este coeficiente. Intuitivamente, LASSO elimina las variables que guardan poca relación con la variable de respuesta. Por lo tanto, el número de predictores incluidos en el modelo guarda una relación inversa con el valor que tome λ.

No obstante, en la práctica cuando el conjunto de datos contiene variables categóricas con varias clases, para estimar este modelo es necesario reexpresar estos predictores en forma de dummies. Si bien en principio no hay inconvenientes, puede afectar la interpretabilidad de los resultados ya que para la misma variable es posible que el método seleccione algunas categorías y no otras. Por ello, en pos de un mejor entendimiento a la hora de seleccionar predictores se utiliza una extensión de este método conocido como *group LASSO*. Desarrollada por Yuan y Lin (2006), esta alternativa permite que el algoritmo seleccione grupos de variables. De esta manera, todas las clases de una variable categórica son seleccionadas o excluidas. La lógica es la misma que el método tradicional de LASSO sólo que el problema de optimización es el siguiente.

$$l^{GL}(\beta) = \sum_{i=1}^{n} \left\{ y_i x_i' \beta - \ln\left(1 + e^{x_i'\beta}\right) \right\} - \lambda \sum_{g=1}^{G} || \beta_g || \tag{1.7}$$

En donde G representa a los grupos de variables. De la misma manera que ridge, el valor de este hiperparámetro se selecciona por validación cruzada con $k = 10$ particiones aleatorias. A su vez, la función para llevar a cabo esto fue *cv.grpreg* y el modelo fue entrenado con la función *grpreg*. Ambas pertenecen al paquete *grpreg*.

4.2. Random forest

Este algoritmo fue introducido por Breiman (2001). Como se ha mencionado anteriormente, permite capturar y modelar de manera automática las relaciones no lineales entre los predictores. De esta forma, si la relación subyacente entre los predictores y la variable de interés es más bien de este tipo, es esperable que este algoritmo tenga un mejor desempeño que un modelo lineal y viceversa.

Para entender cómo funciona es conveniente comenzar explicando en qué consisten los árboles de clasificación y regresión, ya que constituyen la unidad más básica de este algoritmo más complejo. En términos generales, CART (abreviatura en inglés de Classification and Regression Trees) consiste en una serie de particiones binarias y recursivas de los datos en función de la variable de respuesta de manera que estos queden agrupados los más homogéneamente posible en términos de los predictores (Cardinale lagomarsino et al., 2016). Para realizar estas particiones, el algoritmo determina cuál va a ser la variable y el valor de la misma a partir de la cual dividirá los datos. Este procedimiento se repite para luego clasificar a las observaciones en la clase más frecuente del grupo a la que pertenecen. Esto se puede resumir en los siguientes pasos:

1. Se divide el espacio de los X_p predictores en j regiones, $R_1, R_2, ..., Rj$.
2. A cada observación que cae en la región R_j se la clasifica en la categoría k que resulta más frecuente en dicha región.

Entonces, dada la variable j y el punto de partición s, se pueden definir los siguientes semiplanos.

$$R_1(j,s) = \{X|X_j \leq s\}, R_2(j,s) = \{X|X_j > s\} \tag{2.1}$$

De esta forma, la variable y el punto de partición son elegidas de manera que minimicen la impureza en cada semiplano, lo que resulta equivalente a maximizar la homogeneidad. En este

caso, como medida de impureza se utilizará el índice de Gini. Sea $\hat{p}_{mk} \equiv 1/n_m \cdot \sum_{i:x_i \in R_m} I(y_i = k)$ la proporción de observaciones pertenecientes a la clase k en la región o nodo m, el índice de Gini se define de la siguiente forma

$$GINI_m = \sum_{k=1}^{K} \hat{p}_{mk}(1 - \hat{p}_{mk}) \tag{2.2}$$

Intuitivamente, esta medida de impureza nos indica cual es la probabilidad de clasificar mal una observación. Se puede observar que cuantas más son las observaciones que pertenecen a una determinada categoría, menor es el índice de Gini. Por ende, cuanto menor es el valor de esta medida, menor es la impureza en el nodo o región. Entonces, el par (X_j, s) en donde la partición es óptima resuelve

$$\min_{j,s} \{n_1 GINI_1 + n_2 GINI_2\} \tag{2.3}$$

De esta forma, todas las observaciones que caigan en la región m son clasificadas de acuerdo a la clase k correspondiente a la mayoría de las observaciones que allí se encuentran. [13] Estas particiones binarias recursivas, en el límite, dan como resultado un árbol de clasificación que tiene un nodo por cada observación. Dado que la complejidad de este modelo viene dada por la cantidad de nodos terminales que tiene el árbol, esto implicaría un problema de sobreajuste[14].

Para elegir la complejidad óptima de este modelo se suele utilizar el método de poda (en inglés *Weakest Link Pruning*). No obstante, aun aplicando este mecanismo de búsqueda del subárbol óptimo, CART presenta ciertas dificultades. Específicamente, es un método poco robusto a los datos debido a que un pequeño cambio en los mismos puede resultar en un árbol estimado muy distinto. La razón de dicha inestabilidad es la naturaleza jerárquica de los predictores inherentes al método de árboles. En otras palabras, es un método que presenta

[13] Formalmente se expresa $k(m) = argmax_k \hat{p}_{mk}$

[14] Un árbol que presente un nodo terminal por observación presentaría un sesgo bajo y una varianza alta. En cambio, cuando se reducen la cantidad de nodos terminales y, por ende, aumenta el número de observaciones en cada uno de ellos la varianza se reduce mientras que el sesgo aumenta

alta varianza.

Una forma de solucionar este problema es mediante *bagging* o *bootstrap aggregation*. Este procedimiento fue introducido también por Breiman (1996) y consiste en tomar B muestras aleatorias de tamaño n con reemplazo de la muestra original. Luego, para cada una de estas muestras se estima un árbol y se guardan sus predicciones individuales. Posteriormente, se realiza la clasificación a través de un voto por mayoría. Es decir, la estimación de *Bagging* da como resultado un vector $\hat{f}_{bag(x)}$ de tamaño K $[p_1(x), p_2(x), ..., p_K(x)]$ en donde $p_K(x)$ representa la proporción de árboles que predicen la clase k para una determinada observación. De esta forma, *bagging* clasifica a las observaciones a partir de la clase con más votos de los B árboles, $\hat{G}_{bag}(x) = argmax_K \hat{f}_{bag}(x)$. Para el caso de los árboles de regresión se toma como predicción al promedio de las predicciones de cada árbol. Intuitivamente, la idea es que la varianza del promedio es menor que la de un sólo árbol, por lo que la predicción conjunta de todos los árboles es más robusta.

Un potencial problema que puede surgir con *bagging* es cuando hay predictores "fuertes", es decir, variables que reducen mucho el índice de Gini cuando se hace una partición en base a ellas. Si los árboles entrenados están independiente e idénticamente distribuidos, la varianza del promedio de las predicciones viene dada por $\frac{1}{B}\sigma^2$. Sin embargo, la presencia de tal tipo de predictores provocaría que muchos de los árboles entrenados a través de bootstrap utilicen las mismas variables para hacer las primeras particiones. De esta forma, los árboles se parecerían mucho entre sí, lo que es equivalente a decir que estarían correlacionados. Entonces, si los B árboles están idénticamente distribuidos pero no son independientes entre sí, la varianza de las predicciones vendría dada por:

$$\rho\sigma^2 + \frac{1-\rho}{B}\sigma^2$$

En donde ρ representa el coeficiente de correlación por pares de árboles. A medida que aumenta B, el segundo término tiende a cero mientras que el primero se mantiene constante. Por ello, una forma de disminuir dicha correlación y, por ende, la varianza es restringiendo y aleatorizando los predictores utilizados a la hora de realizar las particiones de manera

que no todos árboles sean construidos a partir de los mismos predictores. El algoritmo que descorrelaciona a los árboles a través de este procedimiento es *random forest*. En este caso, el *trade off* sesgo varianza viene regulado por la cantidad de m predictores aleatorios utilizados para realizar las particiones binarias: cuanto más grande es m el modelo va a tener un menor sesgo y una varianza más grande y cuando más chico es, la varianza va a ser menor a costa de un mayor sesgo. Por lo tanto, de los p predictores originales, este método sólo utiliza $m < p$ elegidos al azar donde, típicamente, $m = \sqrt{p}$. Resumiendo, para tareas de clasificación, *random forest* sigue los siguientes pasos (Friedman et al., 2009).

1. Entrena B árboles a partir de muestras obtenidas por *bootstrap* de la muestra original. Para ello, repite recursivamente los siguientes pasos:

   a) Selecciona aleatoriamente m variables de los p predictores originales, donde generalmente $m = \sqrt{p}$.

   b) A partir de los m predictores seleccionados, realiza la partición óptima en la variable y el punto tal que resuelvan (2.3).

2. Clasifica a las observaciones a través del voto mayoritario de los árboles entrenados. Sea $\hat{C}_b(x)$ la clase predicha por el árbol b, entonces $\hat{C}^B_{rf}(x) = votomayoritario\left\{\hat{C}_b(x)\right\}_1^B$

La función utilizada para entrenar este modelo fue *randomForest* que provee el paquete llamado con el mismo nombre. Siguiendo la regla general, se utilizaron $m = \sqrt{p}$ predictores y el número de árboles entrenados fue de 500.

4.3. Conditional random forest

Este método, inicialmente desarrollado por Hothorn et al. (2006), es una implementación de random forest con la diferencia de que las unidades básicas que lo componen no son árboles de clasificación y regresión (CART), sino árboles de inferencia condicional. Estos realizan las particiones binarias recursivas de manera diferente siguiendo los siguientes pasos.

1. Testea la hipótesis nula global de que la variable de respuesta Y es independiente de todos los predictores X_j. La misma se formula en términos de la independencia parcial para cada predictor donde en términos formales queda expresada $H_0 = \cap_{j=1}^{m} H_0^j$ donde $H_0^j : D(\mathbf{Y}|X_j) = D(\mathbf{Y})$ y donde $D(\mathbf{Y}|X_j)$ representa la distribución de Y condicional en X_j. Esto se lleva a cabo a través de tests de permutación, a partir de los cuales es posible obtener le distribución del estadístico de la hipótesis nula así, como los p valores para cada predictor.

2. Si H_0 no es rechazada, el proceso se detiene automáticamente. Si es rechazada, se elige para realizar la partición el predictor que esté más fuertemente asociado con la variable de respuesta, que será aquel cuyo p valor sera más chico.

3. Una vez elegida la variable, se elige el punto de partición óptimo que maximiza el grado de discrepancia entre las dos submuestras resultantes a partir del estadístico computado en el primer paso. (Hothorn et al., 2006).

4. Luego repite recursivamente los pasos anteriores y el árbol condicional dejará de crecer cuando la hipótesis global nula no pueda ser rechazada en ninguna partición hasta el momento hecha.

Este algoritmo tiene dos importantes diferencias en comparación con random forest. En primer lugar, los árboles de inferencia condicional no necesitan de weakest link pruning para evitar el sobreajuste, ya que el árbol alcanza el tamaño óptimo cuando la hipótesis global de independencia no es rechazada. La segunda diferencia es que cuando se cuentan con predictores de distinto tipo (continuos y categóricos) y de diferente escala, CART y, por ende, random forest tienen un sesgo a elegir variables que tienen muchos potenciales puntos de corte a pesar de que estas puedan no tener una relación fuerte con la variable de respuesta. Strobl et al. (2007) señalan que son dos las fuentes de dicho sesgo en estos algoritmos. [15] Por un lado, la primera viene dada por el criterio del índice de Gini para realizar particiones. En aquellos predictores con más puntos de cortes es probable que el algoritmo encuentre un punto de partición que produzca mayores ganancias de acuerdo a

[15] Este problema también ha sido estudiado en trabajos como White y Liu (1994) y Shih (2004).

dicho criterio de manera casual. La segunda fuente de dicho sesgo es inducida cuando se entrena a cada árbol con muestras extraídas mediante *bootstrap*. Los autores muestran que para un conjunto de predictores independientes de una variable de respuesta Y esta técnica de resampleo aleatorio con reemplazo induce a una asociación artificial que se acentúa con el número de categorías y puntos de corte.

En este sentido, conditional random forest soluciona la primera fuente de sesgo dado que no utiliza el criterio del índice de gini para realizar particiones, sino los p-valores obtenidos para cada variable en el primer paso descrito. Sin embargo, para poder solucionar la segunda fuente de sesgo es necesario entrenar a los árboles condicionales con muestras aleatorias sin reemplazo.

Por ello, para entrenar este modelo se utilizó la función *cforest* del paquete *partykit*. Tanto el número de predictores aleatorios para realizar las particiones y la cantidad de árboles entrenados fueron los mismos que en el caso de random forest. Asimismo, para entrenar cada árbol las muestras aleatorias sin reemplazo utilizadas fueron de una fracción equivalente al 60 % de las observaciones originales.

4.4. Support vector machines

Esta herramienta fue desarrollada inicialmente por Cortes y Vapnik (1995). A grandes rasgos, consiste en establecer un límite de clasificación no lineal mediante la construcción de límites lineales de decisión, o hiperplanos separadores, en un espacio transformado y aumentado del conjunto de las características o predictores originales (Rincón, 2019). Para entender mejor cómo funciona este algoritmo, resulta conveniente explicar previamente en qué consisten el clasificicador de máximo márgen y *support vector classifier* (SVC).

En primer lugar, es menester introducir el concepto de hiperplano. En un espacio de p dimensiones, un hiperplano es un subespacio plano de $p-1$ dimensiones. Por ejemplo, en un espacio de dos dimensiones, un hiperplano es una línea; en un espacio de 3 dimensiones, el hiperplano sería un plano (James, Witten, Hastie, y Tibshirani, 2013). Por ende, en un espacio de p dimensiones, un hiperplano es definido por la siguiente ecuación

$$\beta_0 + \beta_1 x_1 + \beta_2 x_2 + ... + \beta_p x_p = \beta_0 + \boldsymbol{\beta}' x = 0 \tag{3.1}$$

En donde $\boldsymbol{\beta}'$ representa al vector de coeficientes. De esta forma, si un punto $x = (x_1, x_2, ..., x_p)^T$ en un espacio de p dimensiones satisface la ecuación 12, se encuentra sobre el semiplano. Si no se cumple de manera que

$$\beta_0 + \boldsymbol{\beta}' x > 0 \tag{3.2}$$

El punto se encontraría de un lado del hiperplano. En cambio, si se da que

$$\beta_0 + \boldsymbol{\beta}' x < 0 \tag{3.3}$$

El punto se encontraría del otro lado del hiperplano. Es decir, en este contexto, un hiperplano divide el espacio en dos partes. Ahora bien, consideremos un conjunto de puntos u observaciones x_i en un espacio de p características o dimensiones donde a cada uno le corresponde un valor de la variable de respuesta y_i $\{(x_1, y_1), ..., (x_i, y_i)\}$ tal que $y_i \in \{-1, 1\}$ en lugar de tomar los valores 0 y 1. Además, supongamos que existe un hiperplano que puede separar perfectamente a estas dos clases de observaciones en donde el lado positivo corresponde a las observaciones de la clase $y_i = 1$ y el lado negativa a las de clase $y_i = -1$. Entonces, la perfecta separabilidad implica que

$$y_i(\beta_0 + \boldsymbol{\beta}' x_i) > 0 \tag{3.4}$$

No obstante, resulta que son infinitos los hiperplanos que pueden realizar esta separación. Una solución natural sería elegir el hiperplano separador que maximice la distancia a los puntos más cercanos de cada clase (Vapnik, 1996). Tales puntos son llamados vectores soporte (*support vectors*) y son las únicas observaciones que afectan o determinan la solución a este problema. La distancia perpendicular de estos a dicho hiperplano se define como margen y el

hiperplano resultante es conocido como el clasificador de máximo margen. También llamado hiperplano separador óptimo, surge del siguiente problema de optimización restringida.

$$\begin{aligned} &\max_{\beta_0,\boldsymbol{\beta}} M \\ &\text{s.a. } \|\boldsymbol{\beta}\| = \sum_{j=1}^{p} \beta_j^2 = 1 \\ &\qquad y_i(x_i'\boldsymbol{\beta} + \beta_0) \geq M, \, \forall i = 1, ..., n \end{aligned} \tag{3.5}$$

En donde M es el margen. La primera restricción sirve para poder asegurar que la distancia perpendicular de un punto x_i al hiperplano viene dada por $y_i(x_i'\boldsymbol{\beta}+\beta_0)$. La segunda restricción implica que, condicional a que M sea positivo, las observaciones se encuentran del lado correcto. Esto asegura que las observaciones se encuentran al menos a una distancia igual a M del hiperplano. De esta forma, los parámetros β_0 y $\boldsymbol{\beta}$ son elegidos de manera que esa distancia se maximice. Este problema puede ser expresado de una forma más conveniente. En primer lugar, podemos deshacernos de la primera restricción. Como $\|\boldsymbol{\beta}\| = 1$, podemos reexpresar la segunda restricción

$$\frac{1}{\|\boldsymbol{\beta}\|} y_i(x_i'\boldsymbol{\beta} + \beta_0) \geq M, \forall i = 1, ..., n \tag{3.6}$$

$$y_i(x_i'\boldsymbol{\beta} + \beta_0) \geq M\|\boldsymbol{\beta}\| \tag{3.7}$$

Ya que para todo valor de $\boldsymbol{\beta}$ y β_0 que satisfaga dicha desigualdad también lo hará cualquier valor positivo reescalado, podemos establecer que $M = \frac{1}{\|\boldsymbol{\beta}\|}$ (Hastie et al., 2009). De esta manera, el problema de maximización en (3.5) es equivalente a

$$\begin{aligned} &\min_{\beta_0,\boldsymbol{\beta}} \|\boldsymbol{\beta}\| \\ &\text{s.a. } y_i(x_i'\boldsymbol{\beta} + \beta_0) \geq 1, \, \forall i = 1, ..., n \end{aligned} \tag{3.8}$$

Como resultado, el hiperplano separador óptimo produce una función $\hat{f}(x) = x'\hat{\boldsymbol{\beta}} + \beta_0$ y la regla de clasificación puede expresarse como

$$\hat{y}_i(x) = \begin{cases} +1, & \text{si } \hat{f}(x) = x_i'\hat{\boldsymbol{\beta}} + \hat{\beta}_0 > 0 \\ -1, & \text{si } \hat{f}(x) = x_i'\hat{\boldsymbol{\beta}} + \hat{\beta}_0 < 0 \end{cases} \tag{3.9}$$

Equivalentemente, el clasificador se puede reexpresar como $\hat{G}(x) = signo(\hat{f}(x))$. Sin embargo, raras son las veces en las que las observaciones pertenecientes a la muestra de entrenamiento son perfectamente separables linealmente. Además, el clasificador de máximo margen suele ser sensible o poco robusto a los datos, lo que es un indicio de sobreajuste y, por ende, de peores predicciones afuera de la muestra. Una forma de enfrentar este problema consiste en maximizar el margen, pero, al mismo, tiempo permitir que algunas observaciones se encuentren dentro del mismo o incluso en el lado incorrecto del hiperplano. El problema es similar al anterior, pero con una nueva restricción.

$$\begin{aligned} &\min_{\beta_0, \boldsymbol{\beta}} \|\boldsymbol{\beta}\| \\ &\text{s.a. } y_i(x_i'\boldsymbol{\beta} + \beta_0) \geq 1 - \epsilon_i, \forall i = 1, ..., n \\ &\epsilon_i \geq 0, \sum_{i=1}^{n} \epsilon_i \leq C \end{aligned} \tag{3.10}$$

Donde ϵ_i representa la distancia o la proporción en la que las observaciones se encuentran del lado incorrecto del margen. Si $\epsilon_i = 0$, entonces la observación i se encuentra del lado correcto del margen y lo contrario ocurre cuando $\epsilon_i > 0$. Cuando $\epsilon_i > 1$, la observación no solo viola el margen, sino que también se encuentra del lado incorrecto del hiperplano, es decir que fue mal clasificada. En este caso, son estas observaciones las que determinan cuál es el hiperplano separador óptimo. Esto resulta equivalente a afirmar que las observaciones que se encuentran en el lado correcto del margen no afecta a la solución de *support vector classifier*. En este sentido, C es el parámetro que regula el número y la intensidad máxima de observaciones del lado incorrecto del margen que se está dispuesto tolerar. Cuanto más grande es C más observaciones se permiten que estén de lado correcto y, de esta forma, más vectores soporte

habría. Cuanto más chico es dicho valor, ocurre lo contrario. De esto se desprende que C regula el *trade-off* entre el sesgo y la varianza del modelo. En el primer caso mencionado, aumenta el sesgo, pero disminuye la varianza. En cambio, cuando $C = 0$, el modelo se vuelve insesgado y ajusta bien adentro de la muestra pero al mismo tiempo se vuelve muy errático. Típicamente, el valor óptimo de este hiperparámetro se elige por validación cruzada. Luego, la regla clasificatoria es la misma que en el caso anterior dado por (3.9).

Si bien support vector classifier es un clasificador más flexible, no deja de tener un límite o borde de decisión lineal y muchas veces las observaciones no pueden separarse de esta forma. En estos casos, no es difícil notar que cualquier clasificador lineal no sería una herramienta adecuada. Una posible solución a este problema sería agrandar el espacio de las características de los predictores añadiendo términos polinómicos, de interacción y otro tipo de funciones de los predictores, aunque esto en muchas ocasiones resulta computacionalmente ineficiente y prohibitivo. En este sentido, support vector machines es una alternativa superadora, ya que permite ajustar un límite de decisión no lineal utilizando *kernels*, lo que implica no tener que añadir un gran número de términos adicionales. Intuitivamente, esto permite llevar el espacio original de características a una dimensión superior de forma que sea posible separar linealmente a las observaciones en este espacio aumentado. Para explicar esto, es conveniente reexpresar el problema de maximización (3.8) de la siguiente forma.

$$\begin{aligned} &\min_{\beta_0,\boldsymbol{\beta}} \ \frac{1}{2}\|\boldsymbol{\beta}\|^2 \\ &\text{s.a. } y_i(x_i'\boldsymbol{\beta} + \beta_0) \geq 1, \forall i = 1, ..., n \end{aligned} \tag{3.11}$$

Donde el problema primal de esta optimización se puede escribir como

$$L(\beta_0, \boldsymbol{\beta}, \lambda) = \frac{1}{2}\|\boldsymbol{\beta}\|^2 - \sum_{i=1}^{n} \lambda_i \left[y_i(x_i'\boldsymbol{\beta} + \beta_0) - 1\right] \tag{3.12}$$

Asimismo, para obtener el problema dual primero debemos obtener las condiciones de primer orden con respecto a β_0 y $\boldsymbol{\beta}$, expresadas a continuación.

$$\boldsymbol{\beta} = \sum_{i=1}^{n} \lambda_i y_i x_i \tag{3.13}$$

$$0 = \sum_{i=1}^{n} \lambda_i y_i \tag{3.14}$$

Luego, reemplazamos en el primal y obtenemos finalmente el problema dual.

$$L_D = \sum_{i=1}^{n} \lambda_i - \frac{1}{2} \sum_{i=1}^{n} \sum_{j=1}^{n} \lambda_i \lambda_j y_i y_j x_i' x_j \tag{3.15}$$

Este problema depende exclusivamente del producto interno del vector de características de las observaciones $\langle x_i \cdot x_j \rangle$. En los casos anteriores del clasificador de máximo margen y de *support vector classifier* este término estaba "linealizado" de manera que

$$\langle x_i', x_j \rangle = \sum^{p} x_{ip}' x_{jp} \tag{3.16}$$

Además, si reemplazamos (3.13) en (3.9) podemos reexpresar los anteriores clasificadores lineales.

$$\hat{G}(x) = signo(\boldsymbol{\beta} \sum_{i=1}^{n} \lambda_i y_i x_i' x_j + \hat{\beta}_0) \tag{3.17}$$

Como se ha mencionado, cuando las observaciones no son linealmente separables en el espacio original de características la idea es llevar a este mismo a una dimensión superior en el cual puedan separarse mediante un hiperplano. Esto requiere modificar el producto interno de las características de las observaciones. En forma genérica, este producto transformado quedaría expresado $\langle h(x_i') \cdot h(x_j) \rangle$. La manera en que este método realiza esta transformación es a través de una función kernel $K(x_i, x_j)$. [16] No obstante, para esto es condición necesaria que

[16]Un kernel es una función simétrica ($K(x_i, x_j) = K(x_j, x_i)$) y no negativa ($K(x_i, x_j \geq 0)$) que en este contexto puede ser interpretada como una medida de similitud entre dos observaciones.

la matriz de Gram del kernel sea positiva definida (Murphy, 2012). [17] Si se cumple esta condición, el teorema de Mercer demuestra que existe una función $h(.)$ tal que:

$$K(x_i, x_j) = h(x_i)' \cdot h(x_j) \tag{3.18}$$

Este resultado muestra que, bajo dicha condición, el kernel es equivalente al producto interno de las observaciones en un espacio ampliado o transformado de características. De esta forma, es posible generalizar el producto interno de las observaciones y el problema dual sería el siguiente.

$$L_D = \sum_{i=1}^{n} \lambda_i - \frac{1}{2} \sum_{i=1}^{n} \sum_{j=1}^{n} \lambda_i \lambda_j y_i y_j K(x_i, x_j) \tag{3.19}$$

Donde el clasificador de support vector machines puede expresarse como.

$$\hat{G}(x) = signo(\boldsymbol{\beta} \sum_{i=1}^{n} \lambda_i y_i K(x_i, x_j) + \hat{\beta}_0) \tag{3.20}$$

La gran ventaja de este procedimiento, conocido como "truco kernel" (en inglés *kernel trick*), es que permite "no linealizar" el producto interno sin necesidad de conocer la forma explícita de $h(.)$, solamente es necesario conocer la función kernel. Algunas de las alternativas más conocidas son las siguientes.[18]

- **Kernel lineal**: $K(x_i, x_j) = \langle x_i \cdot x_j \rangle$
- **Kernel polinómico de grado p:** $K(x_i, x_j) = (1 + \langle x_i \cdot x_j \rangle)^p$
- **Kernel radial:** $K(x_i, x_j) = exp(-\gamma ||x_i - x_j||)^2$
- **Kernel de red neuronal:** $tanh(\kappa_1 \langle x_i \cdot x_j \rangle + \kappa_2)$

[17] Cuando esta condición se cumple al kernel se lo llama de Mercer.

[18] Aplicar el kernel lineal resulta equivalente a *support vector classifier*, ya que el espacio original de características no es transformado.

Una vez ampliado el espacio de características, el problema resulta igual a los casos anteriores: encontrar el hiperplano separador óptimo. Al igual que en *support vector classifier*, el parámetro que regula el *trade off* sesgo-varianza es C.

En la práctica, un inconveniente que suele presentar este algoritmo es que su entrenamiento puede demandar mucho tiempo cuando la base de entrenamiento es relativamente grande (Lantz, 2015). Por ello, siguiendo a Rincón (2019), se implementa una versión modificada de este algoritmo desarrollada por Steinwart y Thomann (2017). Esta alternativa consiste en dividir al espacio de características en subespacios más pequeños llamados "células" donde el modelo es entrenado de manera local en cada una de estas. De esta forma, como en cada célula hay un número menor de observaciones, el tiempo de entrenamiento se reduce considerablemente. La validación cruzada se lleva a cabo en cada subespacio de manera separada de manera que termina habiendo igual cantidad de parámetros óptimos como de subespacios. Por último, para clasificar una observación determinada es utilizado el modelo que se entrenó en el subespacio al que la observación pertenece. No obstante, actualmente esta alternativa puede implementarse utilizando solamente un kernel radial por lo que en este trabajo se aplica esta especificación.

Como se expuso anteriormente, este kernel tiene la forma $K(x_i, x_j) = exp(-\gamma||x_i - x_j||)^2$ donde el término $||x_i - x_j||$ representa la distancia euclídea entre una observación de entrenamiento y una de testeo. Cuanto más grande es esta distancia, menor es el valor que toma la función kernel entendida en este caso como la similitud entre los datos. Es decir, las observaciones de entrenamiento que se encuentran más alejadas en términos euclídeos de las de testeo tienen un peso menor a la hora de clasificar a estas últimas. Esto significa que el kernel tiene un comportamiento "local", ya que solo las observaciones cercanas influyen en la clase que se le atribuye a las observaciones afuera de la muestra (James et al., 2013). Los hiperparámetros a calibrar son C y γ donde necesariamente habrá un par de estos por cada célula. El parámetro γ regula el grado en el que la distancia entre los datos influye en la función kernel. Cuanto más chico es, las observaciones más lejanas tienen mayor influencia en la clasificación. Inversamente, cuanto más grande es γ las observaciones más lejanas tendrán un peso menor. Intuitivamente, regula que tan "local" es la clasificación. En términos del trade off sesgo-varianza, cuanto mayor es γ, más insesgado y errático es el modelo y cuanto menor es, aumenta el sesgo y

disminuye la varianza.

Al igual que en los modelos anteriores, los valores óptimos de C y γ se determinan por validación cruzada utilizando $k = 10$ particiones donde la función implementada para entrenar este modelo es *mcSVM* que se encuentra en el paquete *liquidSVM*.

5. Datos

Los datos utilizados corresponden a las bases trimestrales de microdatos de la EPH a nivel hogar e individual para el período que abarca entre el segundo trimestre de 2016 y el último trimestre de 2019.[19] Siguiendo la metodología del INDEC, se calcularon las líneas de pobreza de cada hogar a partir de la unidad de referencia de adulto equivalente y de los coeficientes de las unidades consumidoras, las cuales a su vez varían por regiones.[20] Luego, para determinar la condición de pobreza de un hogar, se comparó este umbral con el ingreso total familiar. De esta forma, se creó la variable de respuesta que toma valor 1 si este último es estrictamente menor que la línea de pobreza del hogar y toma valor 0 en el caso contrario.

Dado que en este trabajo se busca identificar a los hogares pobres y no pobres con variables no monetarias, se eliminaron todas aquellas relacionadas con el ingreso. A su vez, cada hogar fue identificado a partir del jefe/a del mismo. La incorporación del conjunto total de los 132 predictores que se emplean para realizar las predicciones de la condición de pobreza en este trabajo se debe a que son indicadores fundamentales del nivel de vida de los individuos, así como permiten incluir un conjunto heterogéneo de distintas dimensiones que hacen al bienestar. De esta forma, esto permite contar con información heterogénea como las características habitacionales y de vivienda, el acceso a servicios públicos, la ubicación geográfica y el número de integrantes del hogar, como también con cualidades individuales del jefe/a en cuestión como el estado de actividad, el nivel educativo, características sociodemográficas y el grado

[19] Luego de su intervención en el año 2007, el INDEC reanudó la elaboración de indicadores de pobreza e indigencia a partir del año 2016. Es por ello que se presume que no hay estadísticas oficiales disponibles para el primer trimestre de dicho año.

[20] Ver `https://www.indec.gob.ar/ftp/cuadros/sociedad/EPH_metodologia_22_pobreza.pdf`

de formalidad. En total se cuenta con 132 predictores.[21]

Las variables PP04B_COD y PP11B_COD que refieren a la rama de actividad de la ocupación principal para los ocupados y de la última ocupación principal para los desocupados, respectivamente, fueron reemplazadas por predictores que indican a cuál de los 24 sectores generales del Clasificador de Actividades para Encuestas Sociodemográficas (CAES) pertenecen dichas actividades.[22] Por otro lado, las variables PP04D_COD y PP11D_COD las cuales se refieren al tipo de ocupación los ocupados y al último tipo de ocupación de los desocupados, de acuerdo al Clasificador Nacional de Ocupaciones (CNO), también fueron reemplazadas por variables que indican el carácter ocupacional, la jerarquía ocupacional, la tecnología ocupacional y la calificación ocupacional de ese empleo según dicho clasificador.[23]

Otra cuestión relevante tiene que ver con que los datos presentan un significativo desbalanceo muestral. Esto se debe al hecho de que la mayoría de los hogares son no pobres y es una fracción menor, pero no insignificante, la de aquellos que se encuentran en condición de pobreza. En la muestra, aproximadamente, un 23 % de los hogares pertenecen a esta última categoría. Esto puede representar un problema, ya que un algoritmo que clasifique a todos los hogares como no pobres clasificaría correctamente casi al 80 % de las observaciones. De esta forma, mientras más desbalanceada está la muestra aumentaría artificialmente la precisión de las estimaciones. Siguiendo a Somasundaram y Reddy (2016), existen dos maneras de lidiar con este problema. La primera implica modificar los algoritmos de manera que le den una ponderación mayor a las observaciones de la categoría minoritaria, mientras que la segunda consiste en métodos de remuestreo aleatorio que equilibren la proporción de observaciones de las categorías de la variable de respuesta. Las técnicas más utilizadas dentro de este segundo grupo son:

- **Sobremuestreo:** consiste en incrementar aleatoriamente el número de observaciones

[21] Para más detalles, las tablas B.1 y B.2 del anexo presentan un listado y una descripción de las variables utilizadas.

[22] Estas variables son *sec_ ocup* y *sec_ desocup* para los ocupados y los desocupados respectivamente. A su vez, a las observaciones a las que no les correspondía responder se las codificó con el número "444".

[23] Estas son *car_ ocup*, *jer_ ocup*, *tec_ ocup* y *calif_ ocup* para los ocupados y para la última ocupación de los desocupados son *car_ desocup*, *jer_ desocup*, *tec_ desocup* y *cal_ desocup*. A su vez, a las observaciones a las que no les correspondía responder se las codificó con el número "444".

de la clase subrepresentada de manera que ambas categorías queden equilibradas.

- **Submuestreo** reduce aleatoriamente el número de observaciones de la clase mayoritaria con el mismo fin.
- **Muestro híbrido:** combina las dos técnicas anteriores para obtener una muestra más balanceada.

Particularmente, en este trabajo se adopta la técnica de submuestreo aleatorio de forma que todos los modelos son entrenados con clases balanceadas. Tal como señalan Fernández et al. (2018), a diferencia del sobremuestreo aleatorio, esta alternativa no incrementa las probabilidades de incurrir en un sobreajuste, ya que no crea réplicas exactas de la clase minoritaria.

Por último, para el caso de ridge, LASSO y support vector machines los predictores fueron normalizados de forma tal que los mismos tengan un rango de $[0, 1]$. Para dichos métodos, particularmente, esto es una práctica común debido a que sus resultados pueden verse afectados por la escala de las variables.

6. Resultados

6.1. Habilidad predictiva de los algoritmos

En esta sección se presentan los principales resultados obtenidos. La tabla 1 resume los valores de las medidas de desempeño predictivo para cada modelo. [24]

[24]Los umbrales de decisión o *"cut off"* óptimos para los modelos son 0.6724 (ridge), 0.6784 (group LASSO), 0.577 (random forest), 0.6312 (cond. random forest) y 0.8396 (SVM)

Tabla 1: **Medidas de desempeño predictivo afuera de la muestra**

	Aciertos	Sensibilidad	Especificidad	Precisión	F1	ABROC
Ridge	0.8351	0.6126	0.9158	0.7251	0.6641	0.888
Group LASSO	0.8357	0.6092	0.9179	0.7290	0.6637	0.889
Random forest	0.8425	0.6787	0.902	0.7152	0.6964	0.8966
Cond. random forest	0.8296	0.6239	0.9042	0.7025	0.6608	0.8843
SVM (kernel radial)	0.8265	0.5733	0.9183	0.7179	0.6375	0.8748

Nota: la presente tabla muestra las métricas de desempeño predictivo afuera de la muestra obtenidas para todos los modelos entrenados.

Como bien se ha mencionado, la métrica más adecuada para evaluar los distintos métodos en este trabajo es la tasa de aciertos. Se puede observar que esta no difiere mucho entre los modelos. El que presenta el indicador más alto es random forest siendo este de 84,25 %. Luego, siguen los modelos logísticos regularizados de los cuales group LASSO resulta ser marginalmente mejor que ridge con tasas del 83,57 % y 83,51 %, respectivamente. Por su parte, conditional random forest y support vector machines tiene un desempeño un poco más bajo, ya que sus correspondientes tasas de aciertos fueron del 82.96 % y 82,65 %. Es decir, solamente teniendo en cuenta variables que representan dimensiones del bienestar distintas de la monetaria, estos métodos pueden clasificar correctamente a una gran parte de las observaciones. Esto se ve reflejado en la figura A.4 del anexo que compara las tasas de pobreza estimadas mediante random forest (el modelo que mejor predijo) y a partir del método de "línea" empleado por el INDEC. Es interesante notar que, para todo el período de entrenamiento, el algoritmo de árboles arroja una tasa de pobreza mayor que la metodología de líneas en todos los trimestres, siendo esta diferencia de 3,4 puntos porcentuales, en promedio. Sin embargo, esta discrepancia se revierte en el período de testeo, ya que random forest subestima a la tasa de pobreza de acuerdo a la metodología del INDEC, en promedio, por 1,34 puntos porcentuales. Una posible explicación para dicho fenómeno se centra en los hogares que tienen un ingreso marginalmente superior al valor de su línea de pobreza, los cuales serían clasificados como no pobres de acuerdo a la metodología del INDEC cuando en el resto de las dimensiones del bienestar capturadas en las variables utilizadas pueden presentar significativas carencias. Por lo tanto, es razonable que la tasa de pobreza estimada por el mencionado algoritmo durante todo el período de entrenamiento sea sistemáticamente mayor.

Por otro lado, es posible que las tendencias y el signo de las discrepancias entre las tasas de ambas metodologías se reviertan durante el período de testeo por algún potencial evento que haya afectado a la pobreza por ingresos que, además, el algoritmo no capta. La tendencia creciente de la tasa de pobreza medida por líneas que se observa en los primeros dos trimestres del 2018 (y que continúa hasta el último trimestre del 2019) podría estar respondiendo a los episodios de fuertes devaluaciones y de crisis cambiarias que ocurrieron en Argentina durante ese período de tiempo. Estos impactan directamente sobre el costo de vida, provocando una caída en el ingreso real y, por lo tanto, mayores niveles de pobreza. Sin embargo, esto no se observa tanto en la tendencia de la tasa de pobreza estimada por random forest, ya que como bien se ha mencionado, no se utilizó ninguna variable monetaria para entrenar dicho algoritmo ni para realizar las predicciones. Por ende, los mencionados acontecimientos pueden no afectar directamente a los predictores utilizados, por lo que esto no impactaría tanto en la tendencia de la tasa predicha por el método de árboles.[25]

Si bien ambas metodologías presentadas son de distinta naturaleza y presentan diferencias en sus estimaciones como bien se explicó en el párrafo anterior, pueden también ser pensadas como complementos mutuos a la hora de determinar la condición de pobreza. No hay que perder de vista que el ingreso monetario no deja de ser una dimensión relevante del bienestar, ya que resume mucha información acerca de la capacidad de consumo de los individuos. Por ello, si se buscara conciliar ambos enfoques sería necesario entrenar los algoritmos y realizar las predicciones no solo a partir las dimensiones del bienestar representadas por el conjunto de variables empleadas en esta tesina, sino también a partir de predictores monetarios elegidos mediante algún criterio de forma tal que se minimicen los problemas asociados a estos que motivan este trabajo.

Por último, es interesante ver que, si bien el resto de las medidas tienen una relevancia menor en este contexto, el ordenamiento cambia en algunas de ellas. Si se buscara minimizar el tipo de error 1, el modelo preferido podría ser group LASSO o support vector machines, ya que presentan las mayores tasas de precisión y especificidad, respectivamente. En cambio,

[25]Para cada trimestre del 2019 (es decir, los utilizados para evaluar los modelos) las tasas de pobreza obtenidas por random forest fueron 25,21 %, 25,43 %, 25,44 % y 24,94 %, respectivamente. Por otro lado, las obtenidas mediante las líneas de pobreza del INDEC fueron 25,25 %, 27,82 %, 24,84 % y 28,45 %.

si los falsos negativos tuvieran un costo relativo mayor, el modelo preferido seguiría siendo conditional random forest debido a que su tasa de sensibilidad del 67,87 % es la más alta.

6.2. Reducción del espacio de predictores e importancia de las variables

Hasta ahora se ha determinado que utilizando solamente predictores no monetarios es posible clasificar correctamente a una gran proporción de las observaciones que no conformaron parte de la muestra de entrenamiento, sorteando así los problemas que tiene el ingreso para establecer la condición de pobreza. Sin embargo, existen dos cuestiones adicionales que complementan este resultado.

En primer lugar, en términos de los objetivos de este trabajo, resulta importante determinar si es posible reducir el espacio original de variables a un subconjunto que capte la "esencia" de los datos (Murphy, 2012). Esto permitiría no solo contar con modelos predictivos más parsimoniosos, sino también reducir el costo de la recolección de la información. Para ello se utilizó LASSO que, como se ha mencionado anteriormente, es un método que sirve para seleccionar predictores de manera formal y algorítmica. Como puede observarse en la tabla B.3 del anexo, de los 132 predictores originales, este método seleccionó a 80. Esto quiere decir que el espacio original de características puede reducirse, aunque no sustancialmente, ya que la cantidad de variables elegidas lejos está de ser un reducido grupo. A nivel hogar fueron elegidas algunas que refieren a las condiciones materiales de la vivienda como el tipo de la misma (IV1), el material del cual están hechos los pisos, el techo interior y el techo exterior (IV3, IV4 e IV5 respectivamente). También fueron seleccionados predictores que describen las condiciones habitacionales como la cantidad de ambientes utilizados para su uso exclusivo (II1), si la vivienda tiene lavadero (II4_2), si tiene garage (II4_3), el régimen de tenencia de la propiedad (II7) y el tipo de combustible utilizado para cocinar (II8). Otras variables elegidas también fueron la cantidad total de miembros de hogar (IX_TOT), la cantidad de miembros con 10 años cumplidos o más (IX_MAYEQ10) y la región geográfica donde está ubicado el hogar.

Por otra parte, en lo que respecta a las características individuales del jefe de hogar fueron escogidas variables sociodemográficas como el sexo (CH04) y la edad (CH06). También son importantes predictores respectivos a su educación que indican si sabe leer y escribir (CH09), si el establecimiento educativo donde asiste o asistió, en caso de haber asistido, es público o privado (CH11) y el máximo nivel educativo alcanzado (NIVEL_ED). Además, fueron seleccionadas la condición de actividad (ESTADO), la categoría de inactividad (CAT_INAC) y las distintas formas en las cuales estuvieron buscando trabajo (PP02C1, PP02C2, PP02C6 y PP02C7). Para los que están ocupados son relevantes la cantidad de ocupaciones (PP03D), la cantidad de horas trabajadas en la ocupación principal como en otras ocupaciones (PP3E_TOT y PP3F_TOT), el sector económico donde trabaja (sec_ocup), el carácter ocupacional (car_ocup), la jerarquía ocupacional (jer_ocup), entre otras. Para los desocupados importa la cantidad de tiempo que hace que está buscando trabajo (PP10A), si en ese período realizó algún trabajo ocasional (PP10C), si su última ocupación principal era en una institución de caracter público, privado o de otro tipo (PP11A), el número total de personas que trabajaban en ese lugar (PP11C), si era un trabajo transitorio o permanente (PP11L1) así como su jerarquía (jer_desocup), tecnología (tec_desocup) y calificación ocupacional (cal_desocup) de acuerdo al Clasificador Nacional de Ocupaciones. Por último, también se escogieron variables que refieren al grado de formalidad de las ocupaciones como el tipo de comprobante que le dan al momento de cobrar (PP07K), si tiene vacaciones pagas (PP07G1) o si tiene algún otro tipo de beneficio laboral (PP07F4).

El segundo punto consiste en analizar el ordenamiento de los predictores de acuerdo a su importancia. La intención en esta instancia también es llevar a cabo el ejercicio de robustez propuesto por Cardinale Lagomarsino et al. (2016) en pos de solucionar el sesgo de random forest hacia predictores continuos y con muchas categorías mediante conditional random forest. En primer lugar, resulta útil explicar cuáles son las dos medidas de importancia que random forest permite computar. Una de ellas es el decrecimiento promedio en la precisión (en inglés *Mean Decrease Accuracy*). Para calcular esta medida, por un lado, se tienen en cuenta los errores de clasificación afuera de la bolsa de cada árbol (out of bag) una vez que el modelo es entrenado. Luego se permutan aleatoriamente los valores de un predictor X_j, lo que "simularía" la eliminación de la variable en cuestión. Posteriormente, se vuelve a estimar

el error de predicción de cada árbol y se calcula la diferencia entre estos (antes y después de la permutación) para luego promediarla por la cantidad de árboles. Intuitivamente, esta medida indica cuánto se reduce la precisión del modelo cuando se "elimina" dicha variable X_j.[26] La segunda medida es el decrecimiento medio de Gini (en inglés, *Mean Decrease Gini*). Esta mide qué tanto disminuye la impureza cuando el algoritmo realiza una partición en un predictor X_j promediado por la cantidad de árboles. Naturalmente, las fuentes de sesgo expuestas en la subsección 4.3 provocarían que estas medidas también sean sesgadas a favor de las variables con muchos puntos de corte.

La medida de importancia computada a partir conditional random forest entrenado con muestras aleatorias sin reemplazo sería insesgada en este sentido. Esta se calcula a partir del mismo principio de permutación que *mean decrease accuracy*. En las figuras A.5 y A.6 del anexo se pueden observar los resultados para los respectivos métodos. Si se presta atención a las 6 variables más importantes elegidas por random forest mediante su medida de permutación, se observa a la región donde se encuentra el hogar, el sector o industria económica en la cual trabaja el jefe de hogar, la cantidad total de personas que trabajan en ese lugar, el régimen de tenencia de la propiedad, la edad del jefe de hogar y el estado civil del mismo. A través del criterio de Gini el método seleccionó a la cantidad total de miembros del hogar, la edad del jefe de hogar, el tipo de cobertura médica del jefe de hogar, su nivel de educación, la cantidad de miembros con 10 años cumplidos o más como así los que tienen menos de 10 años. Por su parte, conditional random forest seleccionó insesgadamente a la cantidad total de miembros del hogar, al tipo de cobertura médica que tiene el jefe de hogar, la cantidad de miembros del hogar con edad mayor o igual a 10 años cumplidos, la edad del jefe de hogar, a la categoría de inactividad del jefe de hogar y a la cantidad de miembros del hogar menores de 10 años como las seis variables más relevantes. Si bien estas últimas son en su mayoría las mismas que seleccionó random forest mediante su medida de Gini, el ordenamiento es distinto y la categoría de inactividad cobra más importancia que el nivel de educación.

Es necesario tener en cuenta una limitación que puede tener este resultado. Strobl et al. (2008) señalan que las medidas de permutación pueden estar afectadas cuando los predictores

[26] Cuanto más impreciso se vuelve el modelo al permutar dicho predictor, más importante es este último.

están muy correlacionados entre sí, ya que puede existir un sesgo a favor de estas variables. Como solución desarrollan una medida de importancia condicional que reflejaría mejor la relevancia de los predictores. Sin embargo, por restricciones computacionales, esta medida no se pudo calcular y se llevó a cabo de manera no condicional. Por ende, en este trabajo se pudo sortear el sesgo hacia predictores con muchos puntos de corte, pero podría estar latente el sesgo hacia predictores con un grado alto de correlación.

Por último, en pos de establecer una mayor claridad acerca de las diferencias de los resultados encontrados en comparación con la literatura anterior y, específicamente, con respecto a Cardinale Lagomarsino et al. (2016), no está de más mencionar que la principal distinción con respecto a este trabajo radica en que conditional random forest seleccionó como uno de los predictores más importante a la categoría de inactividad del jefe de hogar y no a la educación del mismo (esta última variable si fue encontrada en la investigación de los autores). Si bien el resto de los predictores más importantes coinciden, esta tesina asegura de cierta forma que dicha importancia no está sesgada por los potenciales puntos de corte de las variables. En cuanto al poder predictivo, la tasa de aciertos de este trabajo fue 0,75 puntos porcentuales menor que la de los mencionados autores, ya que dicha métrica correspondiente al modelo entrenado por estos fue del 85 %. No obstante, es menester tener en cuenta que los períodos de tiempo evaluados y, por ende las muestras utilizadas, son diferentes. Asimismo, al igual que trabajos que comparan distintos algoritmos como el de Sohnesen y Stender (2017), Otok y Seftiana (2014) y Rincón (2019), en esta tesina también el mejor modelo resultó ser random forest. Es decir, parte de la evidencia sugiere que este método tiene un mejor desempeño relativo a la hora de identificar la condición de pobreza en distintos contextos. En cuanto a los predictores más importantes, esta tesina también encuentra algunas similitudes y diferencias con respecto a la literatura. Al igual que Thoplan (2014), la edad resulta ser una variable relevante, aunque no las horas trabajadas, ni el máximo nivel educativo alcanzado ni tampoco el género. En comparación con Rincón (2019), este trabajo también encuentra que el número de miembros del hogar es relevante, pero a diferencia de dicho autor, también lo es el tipo de cobertura médica. En cuanto al desempeño predictivo que esta literatura obtiene suele variar de acuerdo a cada trabajo particular. Por ejemplo, McBride y Nichols (2018) tienen una tasa de aciertos que ronda el 80 %, el modelo de Rincón (2019) es en promedio de

93 % aproximadamente, mientras que Otok y Seftiana alcanzan una tasa de aciertos de 98 %. Fitzpatrick et al. (2018) consiguen métricas que van desde el 70 % hasta el 91 %, dependiendo del algoritmo implementado y de la especificación de los mismos.

7. Reflexiones finales

Las mediciones de pobreza a partir del ingreso pueden verse afectadas por inconvenientes prácticos como la no respuesta y la subdeclaración del mismo cuando se llevan a cabo las encuestas de hogares. Además, muchas críticas sugieren que esta variable puede no ser un reflejo confiable del nivel de bienestar de los individuos. Para sortear estos problemas en este trabajo se estudió la posibilidad de clasificar entre hogares pobres y no pobres a partir de predictores no monetarios que reflejan distintos aspectos de las condiciones de vida a nivel hogar e individual. Para este propósito, se utilizaron distintos métodos provenientes de la literatura de machine learning debido a su gran potencial cuando se trata de tareas de tipo predictivo.

Los principales resultados pueden resumirse en los siguientes tres puntos. En primer lugar, si bien no hay dudas de que el ingreso puede representar una dimensión importante del bienestar, se pudo clasificar correctamente a una gran proporción de las observaciones sin tener en cuenta a dicha variable. Entre los modelos implementados, los cuales varían en su naturaleza y en su capacidad para captar relaciones lineales y no lineales, el que mejor desempeño predictivo tuvo fue random forest. Este pudo predecir acertadamente la condición de pobreza y de no pobreza del 84,25 % del total de los hogares y, a su vez, sus estimaciones de la tasa trimestral de pobreza tuvieron una discrepancia promedio de 1,34 puntos porcentuales con respecto a la calculada por la metodología de "líneas" en el paríodo de testeo. No obstante, no existió una diferencia muy grande con respecto al desempeño predictivo del resto de los modelos (ridge, LASSO, conditional random forest y support vector machines), ya que en el peor de los casos la tasa de aciertos fue del 82 %. Esto muestra que, en gran medida, podría llegar a ser factible prescindir del ingreso para establecer dicha condición de pobreza y evitar los problemas expuestos. En segundo lugar, fue posible reducir el espacio original de características, pero

no considerablemente. De las 132 variables utilizadas, LASSO redujo a cero solamente los coeficientes de 52 de ellas, por lo que el número de predictores seleccionados sigue siendo bastante alto. Por último, se utilizó conditional random forest para determinar la importancia de las variables. Evitando posibles sesgos por la escala y el número de potenciales puntos de partición, este método seleccionó como predictores más relevantes a la cantidad total de miembros del hogar, el tipo de cobertura médica del jefe de hogar, la cantidad de miembros del hogar mayores a 10 años, la edad del jefe de hogar, la categoría de inactividad del mismo y la cantidad de miembros con menos de 10 años cumplidos,

En lo que respecta a futuros trabajos, sería relevante estudiar el potencial de otro tipo de algoritmos clasificatorios como los pertenecientes a la literatura de deep learning, como también analizar si el desempeño predictivo de support vector machines mejora utilizando otros tipos de kernels. A su vez, si bien el conjunto de predictores utilizados abarca cuestiones importantes acerca de la calidad habitacional y de vivienda, de educación, de trabajo y de acceso a los servicios públicos, hay dimensiones no monetarias del bienestar que no se reflejan en los datos utilizados como los niveles de seguridad, de libertad e incluso percepciones subjetivas. Incluir este tipo de información sería relevante no sólo para lograr potenciales mejoras en términos predictivos sino también para analizar si cambia el ordenamiento de la importancia de las variables. Además, en el caso de futuras versiones que busquen reconciliar los enfoques de la pobreza monetaria y no monetaria empleando los métodos de aprendizaje automático, sería conveniente incluir variables que reflejen la dimensión del ingreso para realizar dichas predicciones. Para sortear el problema del sesgo hacia predictores correlacionados sin recurrir a medidas de permutación condicional podrían entrenarse especificaciones de conditional random forest que contengan la mayor cantidad de predictores independientes entre sí de forma que cada uno represente una dimensión o una cuestión del bienestar distinta. Por ejemplo, para el caso de las variables que refieren a la cantidad de miembros del hogar las cuales tienen un grado de correlación alto (IX_TOT, IX_MAYEQ10 e IX_MEN10) podría tenerse en cuenta solamente a la variable IX_TOT, ya que contemplaría también a las otras dos. Otras alternativas que se podrían considerar son las presentadas por Hooker y Mentch (2019). A modo de ejemplo, sugieren una medida de importancia alternativa denominada *dropped importance variable* la cual consiste en calcular la diferencia del error de entrenamiento

cuando se excluye un predictor X_j. Por último, sería enriquecedor también hacer un análisis a nivel desagregado para determinar si los predictores más importantes varía según la región geográfica que se estudie dada la heterogeneidad de la pobreza en Argentina.

Referencias

Afzal, M., Hersh, J., y Newhouse, D. (2015). Building a better model: Variable selection to predict poverty in pakistan and sri lanka.

Alkire, S., y Santos, M. E. (2010). Acute multidimensional poverty: A new index for developing countries. *United Nations development programme human development report office background paper*(2010/11).

Battiston, D., Cruces, G., Lopez-Calva, L. F., Lugo, M. A., y Santos, M. E. (2013). Income and beyond: Multidimensional poverty in six latin american countries. *Social indicators research*, *112*(2), 291–314.

Breiman, L. (1996). Bagging predictors. *Machine learning*, *24*(2), 123–140.

Breiman, L. (2001). Random forests. *Machine learning*, *45*(1), 5–32.

Cardinale Lagomarsino, B., Chagalj, C., y Romero, N. (2016). Predicción de la pobreza en argentina usando random forest. *LI Reunión Anual de la Asociación Argentina de Economía Política*.

Caruso, G., Sosa Escudero, W., y Svarc, M. (2015). Deprivation and the dimensionality of welfare: A variable-selection cluster-analysis approach. *Review of Income and Wealth*, *61*(4), 702-722. Descargado de `https://onlinelibrary.wiley.com/doi/abs/10.1111/roiw.12127` doi: 10.1111/roiw.12127

Cortes, C., y Vapnik, V. (1995). Support-vector networks. *Machine learning*, *20*(3), 273–297.

Fawcett, T. (2006). An introduction to roc analysis. *Pattern recognition letters*, *27*(8), 861–874.

Fernández, A., García, S., Galar, M., Prati, R. C., Krawczyk, B., y Herrera, F. (2018). *Learning from imbalanced data sets*. Springer.

Fitzpatrick, C., Bull, P., y Dupriez, O. (2018). Machine learning for poverty prediction: A comparative assessment of classification algorithms. *Wired at: www. github. com*.

Fraiman, R., Justel, A., y Svarc, M. (2008). Selection of variables for cluster analysis and classification rules. *Journal of the American Statistical Association*, *103*(483), 1294–1303.

Gasparini, L., Cicowiez, M., y Sosa Escudero, W. (2012). *Pobreza y desigualdad en américa*

latina. Temas Grupo Editorial.

Gasparini, L., Sosa-Escudero, W., Marchionni, M., y Olivieri, S. (2013). Multidimensional poverty in latin america and the caribbean: new evidence from the gallup world poll. *The Journal of Economic Inequality*, *11*(2), 195–214.

Groves, R. M., y Couper, M. P. (1998). *Nonresponse in household interview surveys*. John Wiley & Sons.

Hastie, T., Tibshirani, R., y Friedman, J. (2009). *The elements of statistical learning: data mining, inference, and prediction*. Springer Science & Business Media.

Hooker, G., y Mentch, L. (2019). Please stop permuting features: An explanation and alternatives. *arXiv preprint arXiv:1905.03151*.

Hothorn, T., Hornik, K., y Zeileis, A. (2006). Unbiased recursive partitioning: A conditional inference framework. *Journal of Computational and Graphical statistics*, *15*(3), 651–674.

James, G., Witten, D., Hastie, T., y Tibshirani, R. (2013). *An introduction to statistical learning* (Vol. 112). Springer.

Kambuya, P. (2020). Better model selection for poverty targeting through machine learning: A case study in thailand. *Thailand and The World Economy*, *38*(1), 91–116.

Kshirsagar, V., Wieczorek, J., Ramanathan, S., y Wells, R. (2017). Household poverty classification in data-scarce environments: a machine learning approach. *arXiv preprint arXiv:1711.06813*.

Lantz, B. (2015). Machine learning with r. packt publishing. *Birmingham Mumbai*.

McBride, L., y Nichols, A. (2018). Retooling poverty targeting using out-of-sample validation and machine learning. *The World Bank Economic Review*, *32*(3), 531–550.

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

Murphy, K. P. (2012). *Machine learning: a probabilistic perspective*. MIT press.

Otok, B. W., y Seftiana, D. (2014). The classification of poor households in jombang with random forest classification and regression trees (rf-cart) approach as the solution in achieving the 2015 indonesian mdgs' targets. *International Journal of Science and Research (IJSR) Volume*, *3*.

Plulikova, N. (2016). Poverty analysis using machine learning methods. *Bachelor's in Mathematics Thesis, Comenius University in Bratislava*.

Rincón, R. (2019). Quarterly multidimensional poverty predictions in mexico using machine learning algorithms. *Dirección de Investigación Económica, Banco de México*.

Salvia, A., y Donza, E. (1999). Problemas de medición y sesgos de estimación derivados derivados de la no respuesta a preguntas de ingresos en la eph (1990-1998). *Asociación Argentina de Especialistas de Estudios del Trabajo/ASET*(18), 93–120.

Shih, Y.-S. (2004). A note on split selection bias in classification trees. *Computational statistics & data analysis*, *45*(3), 457–466.

Sohnesen, T. P., y Stender, N. (2017). Is random forest a superior methodology for predicting poverty? an empirical assessment. *Poverty & Public Policy*, *9*(1), 118–133.

Somasundaram, A., y Reddy, U. S. (2016). Data imbalance: effects and solutions for classification of large and highly imbalanced data. En *International conference on research in engineering, computers and technology (icrect 2016)* (pp. 1–16).

Sosa Escudero, W. (2018). *Big data y aprendizaje automático: Ideas y desafíos para economistas.* en Una nueva econometría: Automatización, big data, econometría espacial y estructural. Universidad Nacional del Sur.

Steinwart, I., y Thomann, P. (2017). liquidsvm: A fast and versatile svm package. *arXiv preprint arXiv:1702.06899*.

Strobl, C., Boulesteix, A.-L., Kneib, T., Augustin, T., y Zeileis, A. (2008). Conditional variable importance for random forests. *BMC bioinformatics*, *9*(1), 307.

Thoplan, R. (2014). Random forests for poverty classification. *International Journal of Sciences: Basic and Applied Research (IJSBAR), North America*, *17*.

Vapnik, V. (1996). *The nature of statistical learning theory.* Springer science & business media.

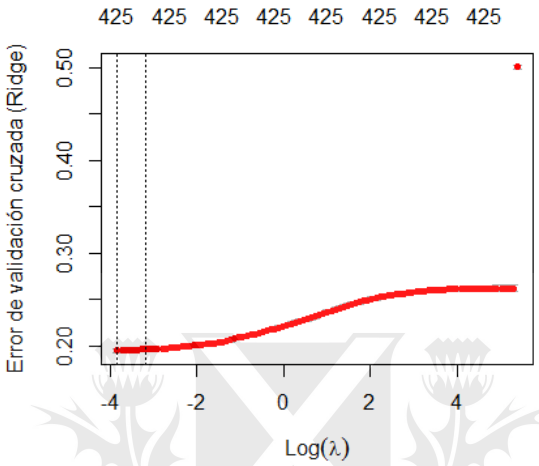
Varian, H. R. (2014). Big data: New tricks for econometrics. *Journal of Economic Perspectives*, *28*(2), 3–28.

White, A. P., y Liu, W. Z. (1994). Bias in information-based measures in decision tree induction. *Machine Learning*, *15*(3), 321–329.

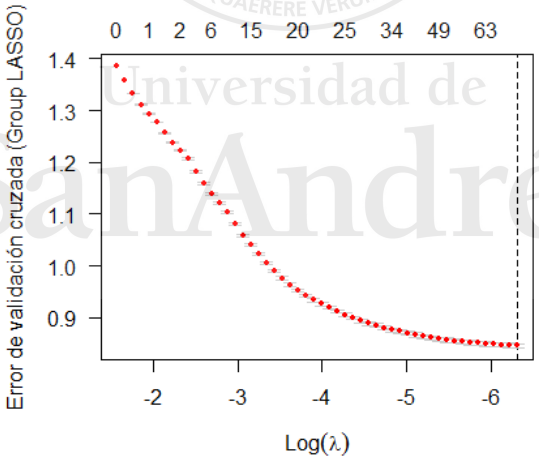
Yuan, M., y Lin, Y. (2006). Model selection and estimation in regression with grouped variables. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, *68*(1), 49–67.

Anexos

A. Gráficos



(a)



(b)

Figura A.1: Las subfiguras (a) y (b) muestran los valores del logaritmo del hiperparámetro λ que minimiza el error de validación cruzada para los modelos logísticos regularizados por Ridge y group LASSO respectivamente. Los mismos son 0,02177 y 0,0018 para los respectivos modelos. Asimismo, group LASSO seleccionó 80 predictores de los 132 originales.

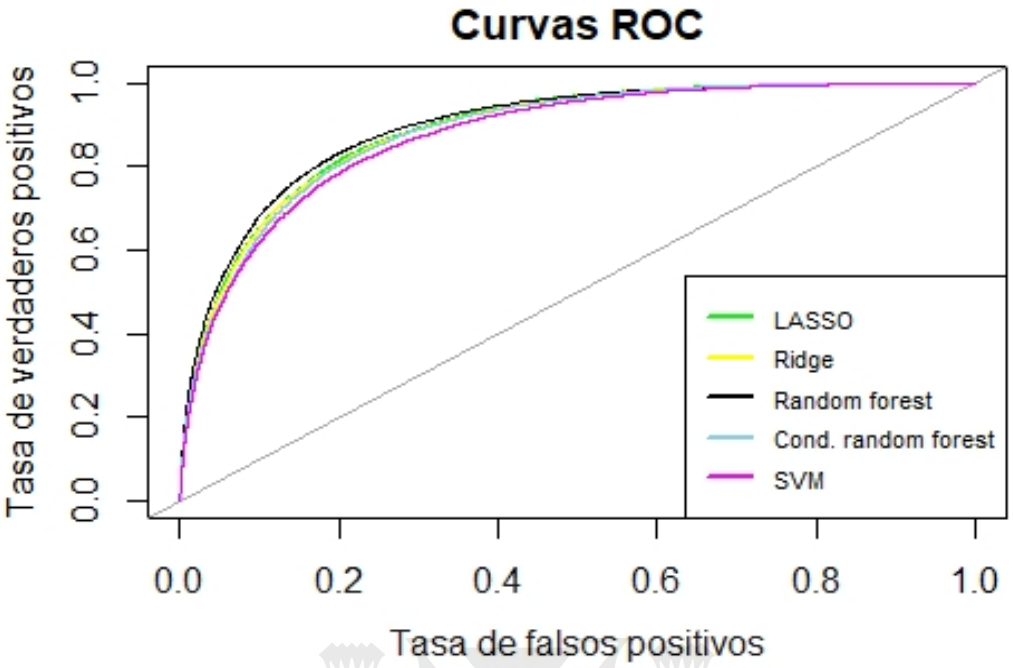


Figura A.2: En este gráfico se presentan las curvas ROC obtenidas para cada modelo. Como se puede observar, no existen diferencias significativas entre estas, lo cual se corresponde con que el área abajo de cada curva es parecida tal como se muestra en la sección de resultados. Solamente la curva negra de random forest está un poco por encima del resto.

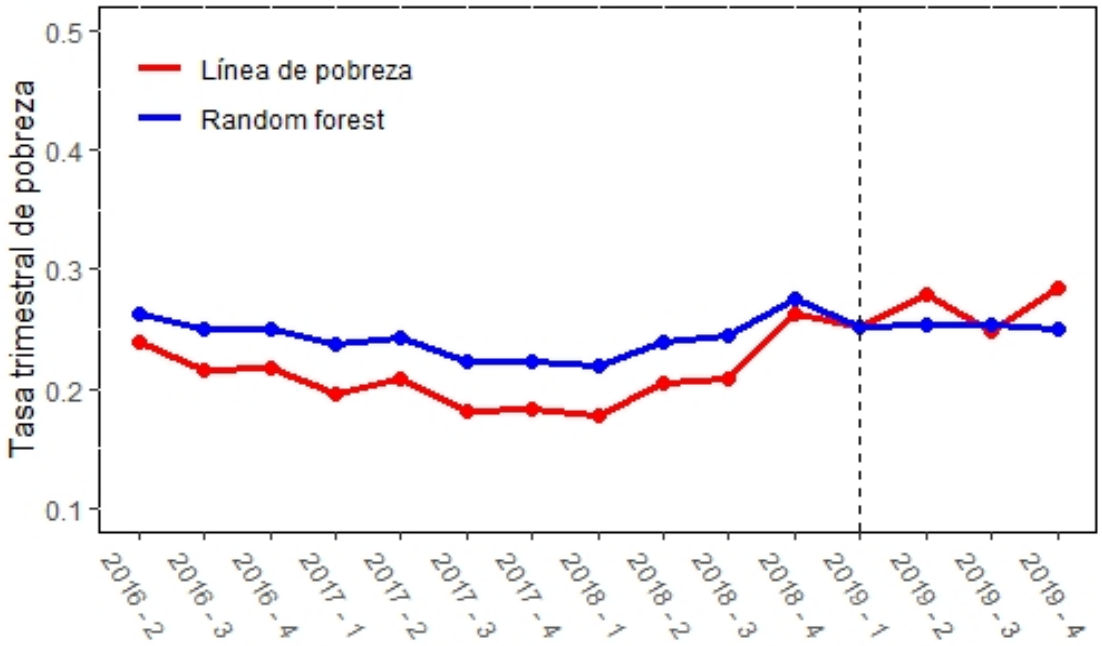


Figura A.3: Este gráfico compara las tasas trimestrales de pobreza obtenidas a partir de la metodología de línea del INDEC y de random forest (el método que mejor desempeño predictivo tuvo). No obstante, el período de mayor interés es el que se encuentra a la derecha de la línea punteada, ya que los datos pertenecientes al mismo fueron utilizados para evaluar el modelo. En este lapso, se puede observar que para el primer y el tercer trimestre del año 2019 las tasas estimadas por ambos métodos son prácticamente iguales. Sin embargo, hay mayores discrepancias para los 2 trimestres restantes ya que en ambos períodos de tiempo, random forest subestima las mediciones de la línea de pobreza. Por último, en promedio para todo el período de testeo, dicha subestimación es de 1,34 puntos porcentuales.

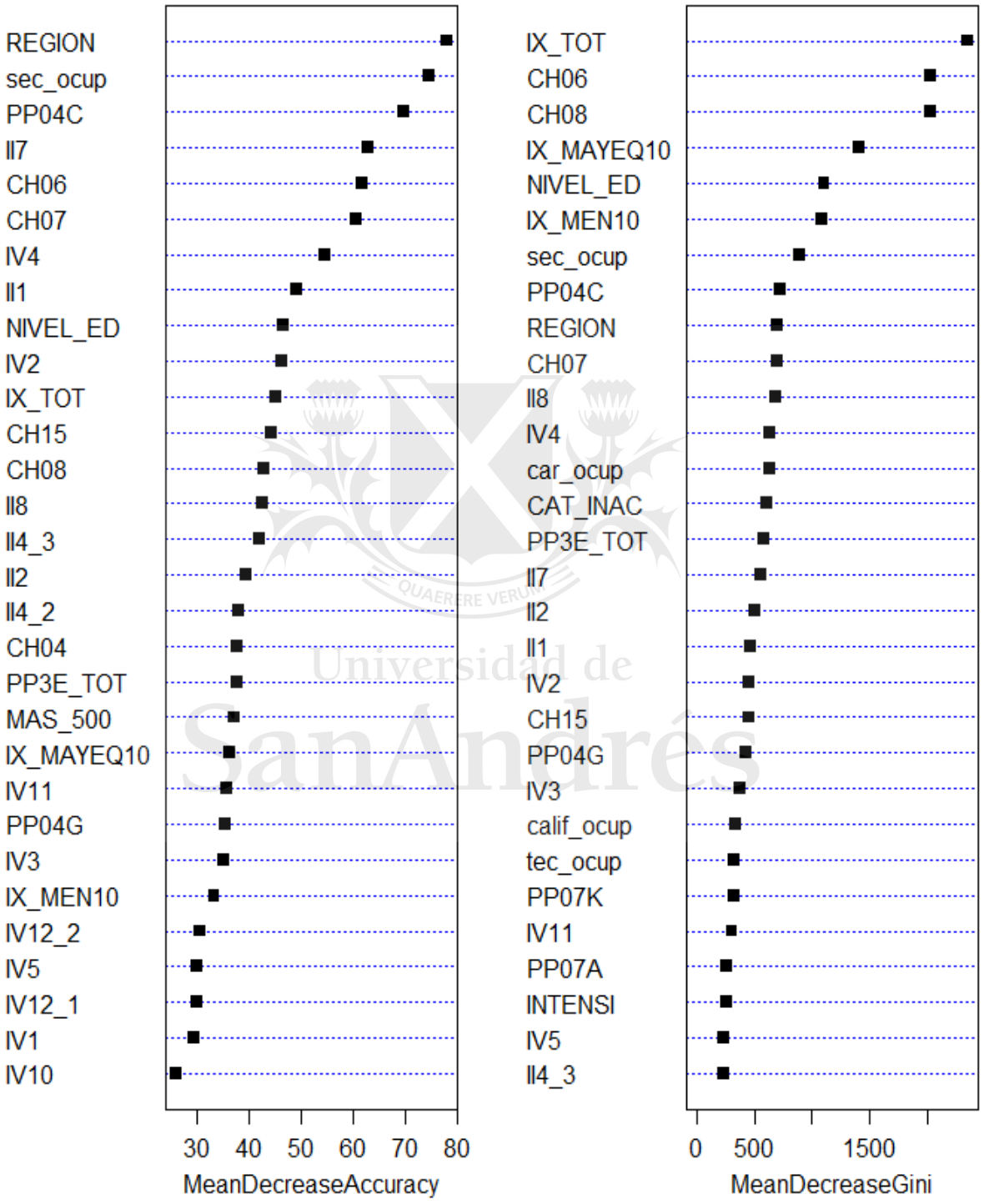


Figura A.4: Esta figura muestra las 30 variables más relevantes elegidas por random forest de acuerdo en base al *mean decrease accuracy* y al *mean decrease gini*.

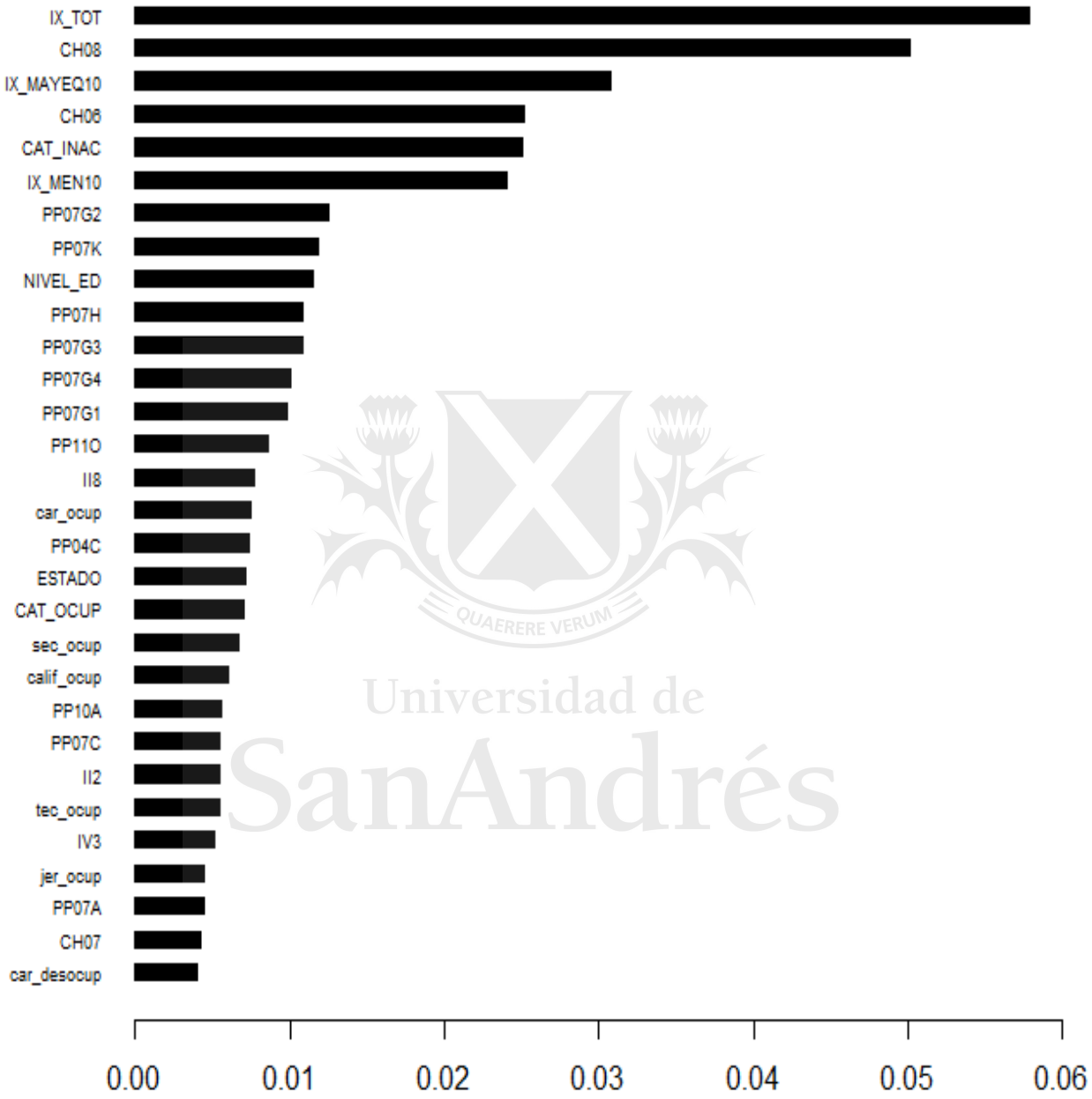


Figura A.5: Esta figura presenta las 30 variables más importantes elegidas por conditional random forest mediante el mismo principio de permutación utilizado para computar *mean decrease accuracy*.

B. Tablas

Variables a nivel hogar	
Categóricas	
REGIÓN	Región geográfica donde se encuentra el hogar (GBA, NOA, NEA, Patagonia, Pampeana).
MAS_500	Si el hogar se encuentra en un aglomerado de más de 500 mil habitantes.
IV1	Tipo de vivienda (casa, departamento, pieza de inquilinato, pieza en hotel/pensión o local no construido para habitación).
IV3	Material del que están hechos los pisos interiores de la vivienda (mosaico/baldosa/madera/cerámica/alfombra, cemento/ladrillo fijo, ladrillo suelto/tierra).
IV4	Material del que está recubierto el techo exterior de la vivienda (membrana/cubierta asfáltica, baldosa/losa sin cubierta, pizarra/teja, chapa de metal sin cubierta, chapa de fibrocemento/plástico, chapa de cartón, caña/tabla/paja con barro/paja sola o departamento en propiedad horizontal).
IV5	Si el techo de la vivienda tiene cielorraso o revestimiento interior.
IV6	Si la vivienda tiene agua por cañería dentro de la vivienda, fuera de la vivienda, pero dentro del terreno, o fuera del terreno.
IV7	Si el agua es de red pública (agua corriente), perforación con bomba a motor o perforación con bomba manual.
IV8	Si la vivienda tiene baño o letrina.
IV9	Si el baño o letrina se encuentra dentro de la vivienda, fuera de la vivienda, pero dentro del terreno, o fuera del terreno.
IV10	Si el baño tiene inodoro con botón/mochila/de agua, inodoro sin botón/cadena y con arrastre de agua (a balde) o letrina (sin arrastre de agua).
IV11	El desagüe del baño es a la red pública (cloaca), a un cámara séptica y pozo ciego, sólo a pozo ciego o a un hoyo/excavación en la tierra
IV12_1	Si la vivienda está ubicada cerca de basural/es (3 cuadras o menos).
IV12_2	Si la vivienda está ubicada en zona inundable.
IV12_3	Si la vivienda está ubicada en una villa de emergencia.
II3	Si en la vivienda hay alguna habitación utilizada exclusivamente como lugar de trabajo.
II4_1	Si la vivienda tiene cuarto de cocina.
II4_2	Si la vivienda tiene lavadero.
II4_3	Si la vivienda tiene garage.
II5	Si alguna de las habitaciones referidas en II4_1, II4_2 y II4_3 es utilizada para dormir.
II6	Si alguna de las habitaciones referidas en II4_1, II4_2 y II4_3 es utilizada como lugar de trabajo.

II7	Régimen de tenencia de la vivienda (propietario de la vivienda y del terreno, propietario solamente de la vivienda, inquilino, ocupante por pago de impuestos/expensas, ocupante en relación de dependencia, ocupante gratuito con permiso, ocupante de hecho sin permiso, o si la vivienda está en sucesión).
II8	Combustible utilizado para cocinar (gas de red, gas de tubo/garrafa o kerosene/leña/carbón).
II9	Si el baño es de uso exclusivo del hogar, si es compartido con otro/s hogar/es de la misma vivienda, si es compartido con otra/s vivienda/s o si no tiene baño.
Numéricas	
IV2	Cantidad total de ambientes que tiene la vivienda, sin contar baño/s, cocina, pasillo/s, lavadero y garage.
II1	Cantidad de ambientes que tiene el hogar para su uso exclusivo.
II2	Cantidad de ambientes de II1 que se utilizan habitualmente para dormir.
II3_1	Cantidad de ambientes que se utilizan exclusivamente como lugar de trabajo, (si II3 es afirmativa).
II5_1	Cantidad de ambientes de las preguntas II4_1, II4_2 y II4_3 que se utilizan para dormir (si II5 es afirmativa).
II6_1	Cantidad de ambientes de las preguntas II4_1, II4_2 y II4_3 que se utilizan exclusivamente como lugar de trabajo (si II6 es afirmativa).
IX_TOT	Cantidad de miembros del hogar.
IX_MEN10	Cantidad de miembros del hogar menores de 10 años.
IX_MAYEQ10	Cantidad de miembros del hogar con 10 o más años cumplidos.

Tabla B.1

Variables a nivel individual del jefe de hogar	
Categóricas	
CH04	Sexo (varón o mujer).
CH07	Si actualmente está unido, casado, separado/a (o divorciado/a), viudo/a o soltero/a.
CH08	Tipo de cobertura médica (obra social; mutual/prepaga/servicio de emergencia; planes y seguros públicos; no paga ni le descuentan; obra social y mutual/prepaga/ servicio de emergencia; obra social y planes y seguros públicos; mutual/prepaga/servicio de emergencia/ planes y seguros públicos; obra social, mutual/prepaga/servicio de emergencia y planes y seguros públicos.
CH09	Si sabe leer y escribir.
CH10	Si asiste o asistió a algún establecimiento educativo.
CH11	Si ese establecimiento es público o privado (si CH10 es afirmativa).
NIVEL_ED	Máximo nivel de educación alcanzado (primaria incompleta, primaria completa, secundaria incompleta, secundaria completa, superior universitaria incompleta, superior universitaria completa o sin instrucción).
CH15	Lugar de nacimiento (en la localidad donde vive, en otra localidad de la misma provincia, en otra provincia, en un país limítrofe o en otro país).
CH16	Lugar donde vivía hace 5 años (en la localidad donde vive, en otra localidad de la misma provincia, en otra provincia, en un país limítrofe o en otro país).
ESTADO	Condición de actividad (ocupado, desocupado o inactivo)
CAT_OCUP	Categoría ocupacional para ocupados y desocupados con ocupación anterior (patrón, cuenta propista, obrero/empleado o trabajador familiar sin remuneración).
CAT_INAC	Categoría de inactividad (jubilado/pensionado, rentista, estudiante, ama de casa, menor de 6 años, discapacitado).
PP02C1	Si estuvo buscando trabajo haciendo contactos y entrevistas.
PP02C2	Si estuvo buscando trabajo enviando currículums y contestando avisos.
PP02C3	Si estuvo buscando trabajo presentándose en establecimientos.
PP02C4	Si estuvo buscando trabajar haciendo algo por su cuenta.
PP02C5	Si estuvo buscando trabajo poniendo carteles en negocios o preguntando por su barrio.
PP02C6	Si estuvo buscando trabajo consultando a parientes y amigos.
PP02C7	Si estuvo buscando trabajo anotándose en bolsas, listas o planes de empleo, en agencias/contratistas o si algún tercero le está buscando trabajo.
PP02C8	Si estuvo buscando trabajo de otra forma activa.

PP02E	Razón por la cual durante los últimos 30 días no buscó trabajo, en caso de que así sea (está suspendido; ya tiene trabajo asegurado; se cansó de buscar trabajo; hay poco trabajo en esta época del año; otras razones).
PP02H	Si en los últimos 12 meses buscó trabajo en algún momento.
PP02I	Si en los últimos 12 meses trabajó en algún momento
PP03C	Si la semana pasada tenía un sólo empleo/ocupación/actividad o más (para los ocupados que trabajaron la semana de referencia).
PP03G	Si la semana pasada quería trabajar más horas o no (para los ocupados que trabajaron la semana de referencia).
PP03H	Si hubiera conseguido más horas: las hubiera podido trabajar esa semana, las hubiera podido empezar a trabajar en dos semanas a más tardar o si no hubiera podido trabajar más horas (para los ocupados que trabajaron la semana de referencia).
PP03I	Si en los últimos 30 días buscó trabajar más horas (para todos los ocupados).
PP03J	Si aparte de este/os trabajo/s estuvo buscando algún otro empleo/ocupación/actividad (para todos los ocupados).
INTENSI	Si es subocupado por insuficiencia horaria, ocupado pleno, sobreocupado u ocupado que no trabajó en la semana (para todos los ocupados).
PP04A	Si el negocio/empresa/institución/actividad en donde lleva a cabo su ocupación principal (la que más horas semanales le dedica) es estatal, privada o de otro tipo.
PP04B1	Si presta servicio doméstico en hogares.
PP04C	Número de personas que trabajan allí en total (1, 2, 3, 4, 5, entre 6 y 10, entre 11 y 25, entre 26 y 40, entre 41 y 100, entre 101 y 200, entre 201 y 500 o más de 500).
PP04G	Lugar donde realiza principalmente sus tareas (en un local/oficina/establecimiento/negocio/taller/chacra/finca; en puesto ó kiosco fijo callejero; en vehículos: bicicleta/moto/autos/barcos/botes sin incluir servicios de transporte; en vehículo para transporte de personas y mercaderías ya sean aéreos, marítimos o terrestres; en obras en construcción, de infraestructura, minería ó similares; en esa vivienda sin lugar exclusivo; en la vivienda del socio ó del patrón; en el domicilio/local de los clientes; en la calle/espacios públicos/ambulante/de casa en casa/puesto móvil callejero; en otro lugar).
PP05C_1	Si en ese negocio/empresa/actividad tiene maquinarias/equipos propios, prestados/alquilados o si no tiene (para los trabajadores independientes).
PP05C_2	Si ese negocio/empresa/actividad tiene un local propio, prestado/alquilado o si no tiene (para los trabajadores independientes).
PP05C_3	Si ese negocio/empresa/actividad tiene un vehículo propio, prestado/alquilado o si no tiene (para los trabajadores independientes).
PP05E	Si para la actividad del negocio, en los últimos 3 meses, tuvo que gastar en la compra de materias primas, productos, pagar servicios u otros gastos (para los trabajadores independientes).

PP05F	Si ese negocio/empresa/actividad trabaja habitualmente para un sólo cliente o para varios clientes (para los trabajadores independientes).
PP05H	Tiempo que los trabajadores independientes han estado trabajando en ese empleo en forma continua con interrupciones laborales no mayores a 15 días (menos de 1 mes; de 1 a 3 meses; más de 3 meses a 6 meses; más de 6 meses a 1 año; más de 1 año a 5 años; más de 5 años).
PP06A	Si en ese negocio/empresa/actividad tiene socios o familiares asociados (para los trabajadores independientes).
PP06E	Si ese negocio/empresa/actividad es una sociedad jurídicamente constituida, si es una sociedad de otra forma legal o si es una sociedad convenida de palabra (para los trabajadores independientes).
PP06H	Si es una actividad o negocio familiar (para los trabajadores independientes).
PP07A	Tiempo que los trabajadores asalariados están trabajando en ese empleo en forma continua sin interrupciones de la relación laboral en la misma empresa/negocio/institución (menos de 1 mes; de 1 a 3 meses; más de 3 meses a 6 meses; más de 6 meses a 1 año; más de 1 año a 5 años; más de 5 años).
PP07C	Si ese empleo tiene tiempo de finalización por ser una changa/trabajo transitorio o si no lo tiene por ser un trabajo permanente (para los trabajadores asalariados).
PP07D	Si PP07C es afirmativa, la duración de ese trabajo transitorio (sólo fue esa vez/sólo cuando lo llaman; hasta 3 meses; más de 3 meses a 6 meses; más de 6 meses a 12 meses; más de 1 año).
PP07E	Si PP07C es afirmativa, el tipo de trabajo transitorio (un plan de empleo; un período de prueba; una beca/pasantía/aprendizaje; otro)
PP07F1	Si en el trabajo le dan de comer gratis (para la ocupación principal de los trabajadores asalariados).
PP07F2	Si en el trabajo le dan vivienda (para la ocupación principal de los trabajadores asalariados).
PP07F3	Si en el trabajo le dan algún producto o mercadería (para la ocupación principal de los trabajadores asalariados).
PP07F4	Si en el trabajo le dan algún otro beneficio como un automóvil o un teléfono celular (para la ocupación principal de los trabajadores asalariados).
PP07F5	Si en el trabajo no recibe ningún beneficio (para la ocupación principal de los trabajadores asalariados).
PP07G1	Si en el trabajo tiene vacaciones pagas (para la ocupación principal de los trabajadores asalariados).
PP07G2	Si en el trabajo le dan aguinaldo (para la ocupación principal de los trabajadores asalariados).
PP07G3	Si en el trabajo tiene días pagos por enfermedad (para la ocupación principal de los trabajadores asalariados).
PP07G4	Si en el trabajo tiene obra social (para la ocupación principal de los asalariados).

PP07G_59	Si en el trabajo no tiene ningún beneficio contemplado en PP07G1, PP07G2, PP07G3 y PP07G4 (para la ocupación principal de los trabajadores asalariados).
PP07H	Si por ese trabajo tiene descuento jubilatorio (para la ocupación principal de los trabajadores asalariados).
PP07I	Si aporta por sí mismo a algún sistema jubilatorio (para los trabajadores asalariados).
PP07J	Si el turno habitual de trabajo es de mañana/tarde, de noche o de otro tipo (para la ocupación principal de los trabajadores asalariados).
PP07K	Si cuando cobra le dan un recibo con sello/membrete/firma del empleador, le dan un papel/recibo sin nada, si no le entregan nada o si no cobra por ser trabajador sin pago/ad-honorem (para la ocupación principal de los trabajadores asalariados).
PP09A	Si en su ocupación principal trabaja en la Ciudad de Buenos Aires, en partidos del Gran Buenos Aires, en ambos o en otro lugar (para los ocupados de la región GBA).
PP09B	Si en su ocupación principal trabaja en la ciudad donde reside (para los ocupados de Posadas, Formosa, Corrientes, Resistencia, Santa Fe, Paraná y Neuquén).
PP09C	Si en su ocupación principal no trabaja en la ciudad donde reside, lo hace en: otro lugar de esa provincia; otra provincia; otro país (para los ocupados de Posadas, Formosa, Corrientes, Resistencia, Santa Fe, Paraná y Neuquén).
PP10A	Para los desocupados, cantidad de tiempo que hace que están buscando trabajo (menos de 1 mes; de 1 a 3 meses; más de 3 meses a 6 meses; más de 6 meses a 12 meses; más de 1 año).
PP10C	Si durante ese tiempo realizó algún trabajo/changa (para los desocupados).
PP10D	Si ha trabajado alguna vez (para los desocupados).
PP10E	Para los desocupados, cantidad de tiempo que hace que terminaron su último trabajo/changa (menos de 1 mes; de 1 a 3 meses; más de 3 meses a 6 meses; más de 6 meses a 12 meses; más de 1 año a 3 años; más de 3 años).
PP11A	Si el negocio/empresa/institución/actividad en la que trabajaba era estatal, privada o de otro tipo (para los desocupados con empleo anterior).
PP11B1	Si prestaba servicios domésticos en hogares particulares (para los desocupados con empleo anterior).
PP11C	Para los desocupados con empleo anterior, el número de personas que trabajaban allí en total (1, 2, 3, 4, 5, de 6 a 10, de 11 a 25, de 26 a 40, de 41 a 100, de 101 a 200, de 201 a 500 o más de 500).
PP11L	Para los desocupados con empleo anterior, razón principal por la que dejaron esa actividad (falta de clientes/clientes que no pagan; falta de capital/equipamiento; trabajo estacional; tenía gastos demasiado altos; jubilación/retiro; causas personales; otras causas laborales).
PP11L1	Si ese trabajo era transitorio o permanente (para los desocupados con empleo anterior).
PP11M	Si ese trabajo era un plan de empleo, un período de prueba u otro tipo de trabajo (para los desocupados con empleo anterior).

PP11N	Si en ese trabajo le hacían descuentos jubilatorios (para los desocupados con empleo anterior).
PP11O	Para los desocupados con empleo anterior, razón principal por la que dejó ese trabajo (despido/cierre; por retiro voluntario del sector público; por jubilación; fin de trabajo temporario/ estacional; le pagaban poco/no le pagaban; malas relaciones laborales/malas condiciones de trabajo; renuncia obligada/pactada; otras causas laborales; por razones personales).
PP11P	Si cerró la empresa (para los desocupados con empleo anterior).
PP11Q	Si fue la única persona que se quedó sin trabajo (para los desocupados con empleo anterior).
PP11R	Si le enviaron telegrama (para los desocupados con empleo anterior).
PP11S	Si le pagaron indemnización (para los desocupados con empleo anterior).
PP11T	Si está cobrando seguro de desempleo (para los desocupados con empleo anterior).
sec_ocup	Para los ocupados, la rama de actividad económica a la que pertenece el negocio/empresa/institución en la que trabaja de acuerdo a las 24 secciones generales de la Clasificación de Actividades Económicas para Encuestas Sociodemográficas del Mercosur (CAES).
car_ocup	Para los ocupados, carácter ocupacional de acuerdo a los 10 grandes grupos ocupacionales del Clasificador Nacional de Ocupaciones (CNO).
jer_ocup	Para los ocupados, jerarquía ocupacional de acuerdo al CNO.
tec_ocup	Para los ocupados, tecnología ocupacional de acuerdo al CNO.
calif_ocup	Para los ocupados, calificación ocupacional de acuerdo al CNO.
sec_desocup	Para los desocupados, la rama de actividad económica a la que pertenecía o pertenece el negocio/empresa/institución en la que trabajaba de acuerdo a las 24 secciones generales de la CAES.
car_desocup	Para los desocupados, carácter ocupacional de su empleo anterior de acuerdo a los 10 grandes grupos ocupacionales del CNO.
jer_desocup	Para los desocupados, jerarquía ocupacional de su empleo anterior de acuerdo al CNO.
tec_desocup	Para los desocupados, tecnología ocupacional de su empleo anterior de acuerdo al CNO.
cal_desocup	Para los desocupados, calificación ocupacional de su empleo anterior de acuerdo al CNO.
Numéricas	
CH06	Años cumplidos.
PP03D	Cantidad de ocupaciones (para los ocupados que trabajaron en la semana de referencia.)
PP3E_TOT	Total de horas que trabajó en la semana en la ocupación principal (para los ocupados que trabajaron en la semana de referencia).
PP3F_TOT	Total de horas que trabajó en la semana en otras ocupaciones (para los ocupados que trabajaron en la semana de referencia).

PP04B2	Cantidad de casas en las que trabaja (para los ocupados que prestan servicios domésticos).
antig_asal	Antigüedad laboral en días de los trabajadores asalariados.
antig_indep	Antigüedad laboral en días de los trabajadores independientes.
antig_desocup	Para los desocupados, cantidad de tiempo en días que trabajaron en su último empleo.

Tabla B.2

Variables seleccionadas por LASSO (80 de 132)	
Nivel hogar	
REGIÓN	Región geográfica donde se encuentra el hogar (GBA, NOA, NEA, Patagonia, Pampeana).
IV1	Tipo de vivienda (casa, departamento, pieza de inquilinato, pieza en hotel/pensión o local no construido para habitación).
IV3	Material del que están hechos los pisos interiores de la vivienda (mosaico/baldosa/madera/cerámica/alfombra, cemento/ladrillo fijo, ladrillo suelto/tierra).
IV4	Material del que está recubierto el techo exterior de la vivienda (membrana/cubierta asfáltica, baldosa/losa sin cubierta, pizarra/teja, chapa de metal sin cubierta, chapa de fibrocemento/plástico, chapa de cartón, caña/tabla/paja con barro/paja sola o departamento en propiedad horizontal).
IV5	Si el techo de la vivienda tiene cielorraso o revestimiento interior.
IV7	Si el agua es de red pública (agua corriente), perforación con bomba a motor o perforación con bomba manual.
IV10	Si el baño tiene inodoro con botón/mochila/de agua, inodoro sin botón/cadena y con arrastre de agua (a balde) o letrina (sin arrastre de agua).
IV11	El desagüe del baño es a la red pública (cloaca), a un cámara séptica y pozo ciego, sólo a pozo ciego o a un hoyo/excavación en la tierra.
IV12_1	Si la vivienda está ubicada cerca de basural/es (3 cuadras o menos).
IV12_2	Si la vivienda está ubicada en zona inundable.
II1	Cantidad de ambientes que tiene el hogar para su uso exclusivo.
II2	Cantidad de ambientes de II1 que se utilizan habitualmente para dormir.
II3_1	Cantidad de ambientes que se utilizan exclusivamente como lugar de trabajo (si II3 es afirmativa).
II4_2	Si la vivienda tiene lavadero.
II4_3	Si la vivienda tiene garage.
II6_1	Cantidad de ambientes de las preguntas II4_1, II4_2 y II4_3 que se utilizan exclusivamente como lugar de trabajo (si II6 es afirmativa).
II7	Régimen de tenencia de la vivienda (propietario de la vivienda y del terreno, propietario solamente de la vivienda, inquilino, ocupante por pago de impuestos/expensas, ocupante en relación de dependencia, ocupante gratuito con permiso, ocupante de hecho sin permiso, o si la vivienda está en sucesión).
II8	Combustible utilizado para cocinar (gas de red, gas de tubo/garrafa o kerosene/leña/carbón).
II9	Si el baño es de uso exclusivo del hogar, si es compartido con otro/s hogar/es de la misma vivienda, si es compartido con otra/s vivienda/s o si no tiene baño.

IX_TOT	Cantidad de miembros del hogar.
IX_MAYEQ10	Cantidad de miembros del hogar con 10 o más años cumplidos.
Nivel individual	
CH04	Sexo (varón o mujer).
CH07	Si actualmente está unido, casado, separado/a (o divorciado/a), viudo/a o soltero/a.
CH08	Tipo de cobertura médica (obra social; mutual / prepaga / servicio de emergencia; planes y seguros públicos; no paga ni le descuentan; obra social y mutual/prepaga/ servicio de emergencia; obra social y planes y seguros públicos; mutual/prepaga/ servicio de emergencia/ planes y seguros públicos; obra social, mutual/prepaga/servicio de emergencia y planes y seguros públicos.
CH09	Si sabe leer y escribir.
CH11	Si ese establecimiento es público o privado (si CH10 es afirmativa).
NIVEL_ED	Máximo nivel de educación alcanzado (primaria incompleta, primaria completa, secundaria incompleta, secundaria completa, superior universitaria incompleta, superior universitaria completa o sin instrucción).
CH15	Lugar de nacimiento (en la localidad donde vive, en otra localidad de la misma provincia, en otra provincia, en un país limítrofe o en otro país).
CH16	Lugar donde vivía hace 5 años (en la localidad donde vive, en otra localidad de la misma provincia, en otra provincia, en un país limítrofe o en otro país).
ESTADO	Condición actual de actividad (ocupado, desocupado o inactivo)
CAT_INAC	Categoría de inactividad (jubilado/ pensionado, rentista, estudiante, ama de casa, menor de 6 años, discapacitado).
PP02C1	Si estuvo buscando trabajo haciendo contactos y entrevistas.
PP02C2	Si estuvo buscando trabajo enviando currículums y contestando avisos.
PP02C6	Si estuvo buscando trabajo consultando a parientes y amigos.
PP02C7	Si estuvo buscando trabajo anotándose en bolsas, listas o planes de empleo, en agencias/contratistas o si algún tercero le está buscando trabajo.
PP02E	Razón por la cual durante los últimos 30 días no buscó trabajo, en caso de que así sea (está suspendido; ya tiene trabajo asegurado; se cansó de buscar trabajo; hay poco trabajo en esta época del año; otras razones).
PP02H	Si en los últimos 12 meses buscó trabajo en algún momento.
PP02I	Si en los últimos 12 meses trabajó en algún momento
PP03G	Si la semana pasada quería trabajar más horas o no (para los ocupados que trabajaron la semana de referencia).

PP03H	Si hubiera conseguido más horas las hubiera podido trabajar esa semana, las hubiera podido empezar a trabajar en dos semanas a más tardar o si no hubiera podido trabajar más horas (para los ocupados que trabajaron la semana de referencia).
PP03I	Si en los últimos 30 días buscó trabajar más horas (para todos los ocupados).
PP03J	Si aparte de este/os trabajo/s estuvo buscando algún otro empleo/ocupación/actividad (para todos los ocupados).
INTENSI	Si es subocupado por insuficiencia horaria, ocupado pleno, sobreocupado u ocupado que no trabajó en la semana (para todos los ocupados).
PP04B1	Si presta servicio doméstico en hogares.
PP04C	Número de personas que trabajan allí en total, incluído (1, 2, 3, 4, 5, entre 6 y 10, entre 11 y 25, entre 26 y 40, entre 41 y 100, entre 101 y 200, entre 201 y 500 o más de 500).
PP04G	Lugar donde realiza principalmente sus tareas (en un local/ oficina/establecimiento/negocio/taller/chacra/finca; en puesto ó kiosco fijo callejero; en vehículos: bicicleta/moto/autos/barcos/botes sin incluir servicios de transporte; en vehículo para transporte de personas y mercaderías ya sean aéreos, marítimos o terrestres; en obras en construcción, de infraestructura, minería ó similares; en esa vivienda sin lugar exclusivo; en la vivienda del socio ó del patrón; en el domicilio/local de los clientes; en la calle/espacios públicos/ambulante/de casa en casa/puesto móvil callejero; en otro lugar).
PP05C_1	Si en ese negocio/empresa/actividad tiene maquinarias/equipos propios, prestados/alquilados o si no tiene (para los trabajadores independientes).
PP05C_3	Si ese negocio/empresa/actividad tiene un vehículo propio, prestado/alquilado o si no tiene (para los trabajadores independientes).
PP05E	Si para la actividad del negocio, en los últimos 3 meses, tuvo que gastar en la compra de materias primas, productos, pagar servicios u otros gastos (para los trabajadores independientes).
PP05H	Tiempo que los trabajadores independientes han estado trabajando en ese empleo en forma continua con interrupciones laborales no mayores a 15 días (menos de 1 mes; de 1 a 3 meses; más de 3 meses a 6 meses; más de 6 meses a 1 año; más de 1 año a 5 años; más de 5 años).
PP06E	Si ese negocio/empresa/actividad es una sociedad jurídicamente constituída, si es una sociedad de otra forma legal si es una sociedad convenida de palabra (para los trabajadores independientes).
PP07A	Tiempo que los trabajadores asalariados están trabajando en ese empleo en forma continua sin interrupciones de la relación laboral en la misma empresa/negocio/institución (menos de 1 mes; de 1 a 3 meses; más de 3 meses a 6 meses; más de 6 meses a 1 año; más de 1 año a 5 años; más de 5 años).
PP07C	Si ese empleo tiene tiempo de finalización por ser una changa/trabajo transitorio o si no lo tiene por ser un trabajo permanente (para los trabajadores asalariados).
PP07E	Si PP07C es afirmativa, el tipo de trabajo transitorio (un plan de empleo; un período de prueba; una beca/pasantía/aprendizaje; otro)

PP07F1	Si en el trabajo le dan de comer gratis (para la ocupación principal de los trabajadores asalariados).
PP07F4	Si en el trabajo le dan algún otro beneficio como un automóvil o un teléfono celular (para la ocupación principal de los trabajadores asalariados).
PP07G1	Si en el trabajo tiene vacaciones pagas (para la ocupación principal de los trabajadores asalariados).
PP07J	Si el turno habitual de trabajo es de mañana/tarde, de noche o de otro tipo (para la ocupación principal de los trabajadores asalariados).
PP07K	Si cuando cobra le dan un recibo con sello/membrete/firma del empleador, le dan un papel/recibo sin nada, si no le entregan nada o si no cobra por ser trabajador sin pago/ad-honorem (para la ocupación principal de los trabajadores asalariados).
PP09A	Si en su ocupación principal trabaja en la Ciudad de Buenos Aires, en partidos del Gran Buenos Aires, en ambos o en otro lugar (para los ocupados de la región GBA).
PP09B	Si en su ocupación principal trabaja en esa ciudad (para los ocupados de Posadas, Formosa, Corrientes, Resistencia, Santa fe, Paraná y Neuquén).
PP10A	Para los desocupados, tiempo que hace que están buscando trabajo (menos de 1 mes; de 1 a 3 meses; más de 3 meses a 6 meses; más de 6 meses a 12 meses; más de 1 año).
PP10C	Si durante ese tiempo realizó algún trabajo/changa (para los desocupados).
PP11A	Si el negocio/empresa/institución/actividad en la que trabajaba era estatal, privada o de otro tipo (para los desocupados con empleo anterior).
PP11C	Para los desocupados con empleo anterior, número de personas que trabajaban allí en total, incluído (1, 2, 3, 4, 5, de 6 a 10, de 11 a 25, de 26 a 40, de 41 a 100, de 101 a 200,de 201 a 500 o más de 500).
PP11L1	Si ese trabajo era transitorio o permanente (para los desocupados con empleo anterior).
PP11Q	Si fue la única persona que se quedó sin trabajo (para los desocupados con empleo anterior).
sec_ocup	Para los ocupados, la rama de actividad económica a la que pertenece el negocio/empresa/institución en la que trabaja de acuerdo a las 24 secciones generales de la Clasificación de Actividades Económicas para Encuestas Sociodemográficas del Mercosur (CAES).
car_ocup	Para los ocupados, carácter ocupacional de acuerdo a los 10 grandes grupos ocupacionales del Clasificador Nacional de Ocupaciones (CNO).
jer_ocup	Para los ocupados, jerarquía ocupacional de acuerdo al CNO.
tec_ocup	Para los ocupados, tecnología ocupacional de acuerdo al CNO.
calif_ocup	Para los ocupados, calificación ocupacional de acuerdo al CNO.
jer_desocup	Para los desocupados, que jerarquía ocupacional tenían en su empleo anterior de acuerdo al CNO.
tec_desocup	Para los desocupados, que tecnología ocupacional tenían en su empleo anterior de acuerdo al CNO.

cal_desocup	Para los desocupados, que calificación ocupacional tenían en su empleo anterior de acuerdo al CNO.
CH06	Años cumplidos.
PP03D	Cantidad de ocupaciones (para los ocupados que trabajaron en la semana de referencia.)
PP3E_TOT	Total de horas que trabajó en la semana en la ocupación principal (para los ocupados que trabajaron en la semana de referencia).
PP3F_TOT	Total de horas que trabajó en la semana en otras ocupaciones (para los ocupados que trabajaron en la seamana de referencia).
PP04B2	Cantidad de casas en las que trabaja (para los ocupados que prestan servicios domésticos.)

Tabla B.3