Departamento Administrativo
Nacional de Estadística

DANE
INFORMACIÓN PARA TODOS

Dirección General

Imputación de la condición de informalidad de los ocupados en Colombia para marzo y abril de 2020

Agosto 2020

Contenido

Equipo de trabajo:

Departamento Administrativo Nacional de Estadística DANE
Andrés García Suaza-Asesor externo, profesor Universidad EIA
Anderson Leal Vélez
Angélica Joana Suárez
Cristhyan Leonardo Naranjo
Jaime Sebastián Lobo Tovar
José de Jesús Lobo Camargo
Juan Sebastián Ordóñez Herrera

1. INTRODUCCIÓN

El DANE tiene como misión implementar procesos de producción y comunicación de información estadística que cumplan con estándares internacionales para la toma de decisiones (Decreto 262 de 2004). En la pandemia del Coronavirus COVID-19, estos procesos de producción han enfrentado retos importantes como consecuencia de las políticas de confinamiento que afectan los operativos de campo.

En el caso del mercado laboral, la Gran Encuesta Integrada de Hogares (GEIH), el instrumento de caracterización debió responder a los retos de operación en medio del confinamiento, principalmente durante los meses de marzo y abril. Lo anterior, supuso realizar modificaciones al proceso de recolección a fin de garantizar un conjunto de indicadores que permitieran mantener el seguimiento del comportamiento del mercado laboral. Estas modificaciones han tenido lugar en un número importante de países y siguen las sugerencias realizadas por la OIT (2020).

Los indicadores de seguimiento al mercado de trabajo como tasa desempleo, tasa de participación y tasa de ocupación, así como la caracterización socioeconómica de la población mantuvieron su calidad y reporte. Sin embargo, el indicador de informalidad, el cual se compone de variables como el tipo de ocupación, el tamaño de las empresas y el oficio, presenta incompletitud para los meses mencionados.

En este sentido, como parte del desarrollo de estadísticas exploratorias que está llevando a cabo el DANE, en el cual se integran registros administrativos, se propone imputar la condición de informalidad a nivel de microdatos para los ocupados encuestados en la GEIH utilizando la Planilla Integrada de Liquidación de Aportes (PILA) como fuente de información adicional para estimar la probabilidad de que un ocupado desempeñe su actividad económica en el sector informal.

Para ello, se utilizan algoritmos de aprendizaje de máquina, en particular *Random Forest*, sobre el conjunto de datos de los ocupados de la GEIH para el año 2019 y el primer semestre de 2020, considerando además de un conjunto de variables socioeconómicas, un indicador de emparejamiento entre la GEIH y la PILA. Este ejercicio de imputación hace posible estimar la tasa de informalidad para los meses de marzo y abril de 2020.

2. OBJETIVO

Imputar la condición de informalidad del módulo de ocupados de la GEIH a través de la implementación de algoritmos de aprendizaje de máquina y la integración de registros PILA que permita completar la serie de informalidad registrada con periodicidad mensual.

3. REFERENTES

La Comisión Económica para Europa de las Naciones Unidas (UNECE) resalta la importancia del uso de las metodologías de aprendizaje de máquina en la producción de estadísticas oficiales. En particular, la UNECE señala que las Oficinas Nacionales de Estadísticas pueden hacer uso de estos métodos con el objetivo de: (i) realizar inferencia, (ii) corregir la unidad de no respuesta, (iii) imputar la no respuesta, (iv) medir el error de medición de un modelo, y (iv) realizar predicciones en un futuro cercano (UNECE, 2018). En particular, la metodología *Random Forest (RF)* es un algoritmo de aprendizaje que utiliza las predicciones de múltiples arboles de decisión para construir un pronóstico agregado. Entre las ventajas de esta metodología se encuentra que es computacionalmente eficiente y no requiere el cambio frecuente de los parámetros del modelo (IMF, 2020).

Esta metodología ha sido empleada por Eurostat con el objetivo de obtener una estimación no paramétrica del crecimiento del Producto Interno Bruto (PIB) y analizar las variables que permiten explicar las fluctuaciones de corto plazo. Este pronóstico utiliza datos armonizados de encuestas consumidores y empresas. Asimismo, en UNECE (2018) se propone hacer uso de un RF en dos etapas para la edición e imputación de la variable ausencia parcial del trabajo, desempleo, trabajo temporal, horas extras, haciendo uso de los datos de la Encuesta de Fuerza Laboral de Noruega.

4. METODOLOGÍA

4.1. Variables

4.1.1. Variable dependiente y muestra objeto de estudio

El objetivo del modelo propuesto consiste en realizar un ejercicio de pronóstico de la condición de informalidad de lo ocupados, la cual se define como una variable dicotómica que toma el valor de 1 en caso de que el ocupado sea informal y 0 en caso de que este sea formal.

Para efectos de estadísticas oficiales, los ocupados informales son las personas que durante el período de referencia se encontraban en una de las siguientes situaciones:

1. Los empleados particulares y los obreros que laboran en establecimientos, negocios o empresas que ocupen hasta cinco personas en todas sus agencias y sucursales, incluyendo al patrono y/o socio.
2. Los trabajadores familiares sin remuneración en empresas de cinco trabajadores o menos.
3. Los trabajadores sin remuneración en empresas o negocios de otros hogares.
4. Los empleados domésticos en empresas de cinco trabajadores o menos.
5. Los jornaleros o peones en empresas de cinco trabajadores o menos.
6. Los trabajadores por cuenta propia que laboran en establecimientos hasta de cinco personas, excepto los independientes profesionales.
7. Los patrones o empleadores en empresas de cinco trabajadores o menos.
8. Se excluyen los obreros o empleados del gobierno.

Esta variable se calcula para el periodo comprendido entre enero de 2019 y febrero de 2020, además de los meses de mayo y junio de 2020, en los cuales la operación estadística de la GEIH retomó los módulos completos. La inclusión de estos meses es crucial ya que el modelo estimado tendrá información sobre los cambios en los patrones de empleo que pudieron derivarse de la pandemia del Coronavirus COVID-19.

Los datos considerados se dividen en tres grupos: i. entrenamiento, correspondiente a un 70% de la muestra; ii. testeo, el 30% restante; y iii. la base objeto de imputación correspondiente a los ocupados de los meses de marzo y abril de 2020. De acuerdo con la definición de informalidad, existen grupos de ocupados en los que se conoce su condición de informalidad con la información disponible. En particular, los obreros o empleados del gobierno, así como los trabajadores por cuenta propia profesionales, por lo que se excluyen de la muestra, y se asigna el valor de 0 en la variable dicotómica. De esta forma, la muestra consiste en 276.419, correspondientes a las principales 23 ciudades o áreas metropolitanas.

4.2.2. Variables independientes

Con el fin de ejecutar el entrenamiento de un algoritmo que permita imputar la condición de informalidad a los ocupados, se considera un conjunto de variables disponibles en la GEIH que permiten capturar características socioeconómicas y de empleo de los individuos. En particular, se consideran las variables sexo, edad, ciudad, oficio, actividad económica, posición ocupacional y el registro de novedades como vacaciones o suspensiones. Además, se incluyen el mes y el año de reporte de la encuesta para controlar patrones de estacionalidad en el mercado de trabajo.

Esta información se complementa con datos provenientes de la PILA, que permiten tener variación adicional importante para imputar la condición de informalidad. En particular, es usual considerar que los registros de PILA corresponden al componente formal del empleo, por lo que un ocupado en GEIH que se encuentre registrado en PILA puede tener un empleo en condiciones de formalidad con alta probabilidad. De hecho, la tabla 1 muestra que, para los periodos de enero y abril de 2019 y enero de 2020, un gran porcentaje de los ocupados que se emparejan entre GEIH y PILA corresponden a ocupados formales. Lo cual coincide con un alto porcentaje de informales, casi 80%, entre aquellos que no se encuentran en el emparejamiento entre estas fuentes de información. Esto es un indicio importante de que la información proveniente de los registros administrativos puede ser relevante para determinar la condición de informalidad.

Tabla 1. Porcentaje de informalidad del emparejamiento GEIH PILA

	ene-19		abr-19		ene-20	
	No PILA	PILA	No PILA	PILA	No PILA	PILA
Formal	22,14	77,36	21,66	78,63	20,75	78,51
Informal	77,86	22,64	78,34	21,37	79,25	21,49

Fuente: DANE, GEIH-PILA

De este modo, entre las características individuales se considera una variable dicotómica que indica si el registro de GEIH se encuentra entre los registros de PILA. Este indicador, que como se verá es importante en el ejercicio de clasificación, constituye una importante fuente de aprovechamiento de registros administrativos como insumo para complementar la operación estadística. Además de la variable que representa la unión entre estos conjuntos de datos, desde la PILA, se incluye el tipo de cotizante y desde GEIH se incluye un indicador que valida la calidad en la identificación de los encuestados. Finalmente, teniendo en cuenta el impacto del COVID-19 en los patrones de empleo, se incluye una variable dicotómica que

toma el valor de 1 durante el periodo de pandemia[1]. Para implementar el algoritmo es necesario efectuar un procedimiento de codificación para las variables cualitativas.

4.2. MODELO

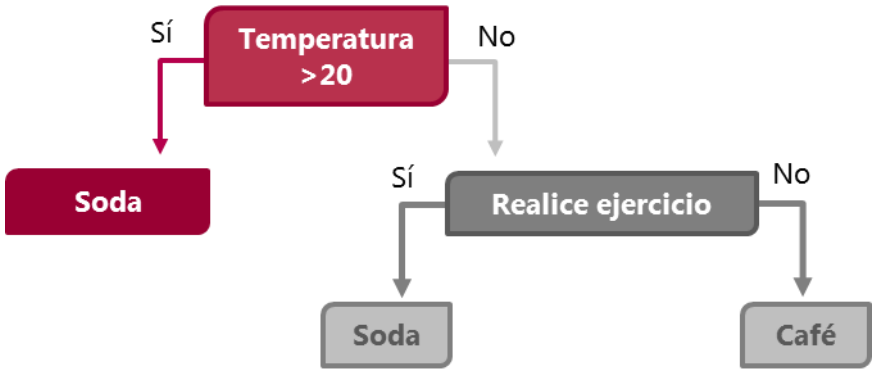
Teniendo en cuenta que el objetivo es construir modelos que permitan inferir la condición más probable de los ocupados en términos de informalidad a partir de sus características, se propone implementar algoritmos de aprendizaje de máquina, los cuales se vienen implementando cada vez más en el análisis económico y de mercado laboral (Gerunov 2014 y Athey e Imbens, 2019). Dadas las características del ejercicio, se propone la aplicación del algoritmo denominado *Random Forest (RF),* un refinamiento del algoritmo denominado árboles de decisión (ver James et al, 2013 y Lantz 2015). El algoritmo de árbol de decisión basa su proceso de aprendizaje en la separación registros en subconjuntos con mayor nivel de homogeneidad, lo cual se puede medir a través de un índice de entropía[2]. De esta forma el algoritmo construye una serie de decisiones simples, semejante a las sentencias de la forma *si….entonces…* en un algoritmo computacional.

Para ejemplificar el funcionamiento del algoritmo, suponga que desea decidir qué bebida tomar al desayuno. Dicha decisión puede depender de un conjunto de factores, entre ellos la temperatura y las actividades que realiza durante la mañana. De esta forma, si la temperatura es alta o decide realizar ejercicio preferirá una bebida fría, por ejemplo, soda, en otros casos, café. De esta forma, la elección de la bebida tiene como primer criterio la temperatura, y en segundo la realización de actividad física. Esta secuencia de decisiones puede representarte en un árbol como lo muestra la Figura 1. Esta misma lógica puede implementarse sobre las características observables de los individuos para determinar si un ocupado es formal o informal.

Figura 1. Ejemplo del mecanismo de aprendizaje del árbol de decisiones

[1] Los meses de marzo a junio de 2020.

[2] Minimizar este índice suele ser el criterio estadístico para determinar la efectividad de las separaciones con el objetivo de realizar clasificar objetos, en este caso ocupados entre formales e informales.



Bajo esta misma perspectiva el algoritmo de árboles de decisión es entrenado a partir de un conjunto de registros que permiten inferir patrones y establecer reglas de clasificación de ellos ocupados informales. Este mismo principio se aplica al RF que hace parte de los modelos de ensamble de árboles de decisión. En particular, un RF es un conjunto de árboles de decisión que hace más robusta la tarea de clasificación a través de la elección aleatoria de conjuntos de datos y variables (ver Breiman, 2001, y Lantz, 2015). De esta forma el ejercicio de clasificación no depende solo de la construcción de un árbol, sino que permite generalizar patrones sobre los datos sin caer fácilmente en la memorización de estos por parte del algoritmo.

Para el entrenamiento del RF se considera un 70% de la muestra, mientras que el 30% restante se utiliza para analizar el poder predictivo del modelo estimado. Para evaluar el desempeño del algoritmo se obtiene métricas como el F1 *score*[3] y el nivel de precisión.

[3] El Valor-F en estadística es la medida de precisión como un promedio ponderado de la precisión (proporción de los clasificados como informales correctamente clasificados) y el *recall* (proporción de informales clasificados correctamente), donde un puntaje F1 alcanza su mejor valor en 1 y el peor puntaje en 0.

5. RESULTADOS

La implementación del algoritmo RF requiere la selección de parámetros que determinan el proceso de aprendizaje de máquina. A priori, la selección de dichos parámetros es arbitraria, sin embargo, a partir de la comparación de escenarios (o proceso de validación cruzada) es posible determinar la configuración de parámetros que genera el mejor desempeño. En este caso estos parámetros hacen referencia a el porcentaje máximo de variables que entrena cada árbol y el número de clasificadores o árboles.

El algoritmo se implementó a diferentes configuraciones de datos, que diferían entre los meses incluidos en la muestra y la inclusión de variables de cruce de PILA y de periodo COVID-19. Es decir, se realizan diferentes combinaciones en cuanto la inclusión de estas variables y de los datos correspondientes a febrero y mayo. La intención de estos ejercicios es identificar si las variables de cruce son relevantes, así como la importancia de incluir datos del periodo COVID-19, como es el caso de los meses de mayo y junio.

De este ejercicio resulta que la mejor configuración corresponde a la inclusión de toda la información disponible. Implementando la técnica de validación cruzada se obtiene que la selección sugerida de parámetros corresponde a la estimación de 134 árboles de decisión utilizando un 26,8% de las variables disponibles en cada uno de estos. La Tabla 2 presenta el error de clasificación obtenido en la muestra de entrenamiento y la de testeo, donde se puede observar que el modelo tiene un buen nivel de desempeño.

Tabla 2. Error de entrenamiento y de prueba del algoritmo RF.

Error de entrenamiento	11,52%
Error de prueba	11,47%

Fuente: elaboración propia

Un análisis más profundo del desempeño del algoritmo puede realizarse a través de la matriz de confusión presentada en la Tabla 3. Esta matriz muestra el nivel de acierto del modelo y además permite observar cómo se distribuyen los errores de clasificación dentro del conjunto de datos de testeo o prueba. Es importante señalar que el algoritmo posiblemente no presenta un sesgo en los errores de clasificación, es decir los errores tienden a distribuirse de manera relativamente uniforme entre ocupados formales que son clasificados como informales, y viceversa.

Tabla 3. Matriz de confusión del conjunto de datos prueba del algoritmo RF

			Valores Predicción	
Valores observados	**Categorías**		**Formal**	**Informal**
			0	**1**
	Formal	**0**	32.360	5.174
	Informal	**1**	5.545	50.221

Fuente: elaboración propia

Al observar los resultados, se puede evidenciar un buen desempeño del modelo. En particular, el modelo clasificó de manera correcta 32.360 ocupados como formales y 50.221 como informales, de un total de 93.300 registros. Métricas como *Precision, Recall y F1 score*[4], permiten validar los resultados observados en la matriz de confusión, en un algoritmo con buen desempeño estas métricas son cercanas a uno[5]. Por su parte, el F1 score promedia las métricas anteriores, y por tanto permite medir el desempeño global del modelo. Los resultados obtenidos en estas métricas (ver Tabla 4) muestran un desempeño satisfactorio del algoritmo.

Tabla 4. Métricas asociadas al algoritmo RF

Precision	0,8537

[4] La Precisión hace referencia al porcentaje de registros clasificados como informales que efectivamente pertenecen a este grupo, mientras que el *Recall* mide el porcentaje de informales que son correctamente clasificados.

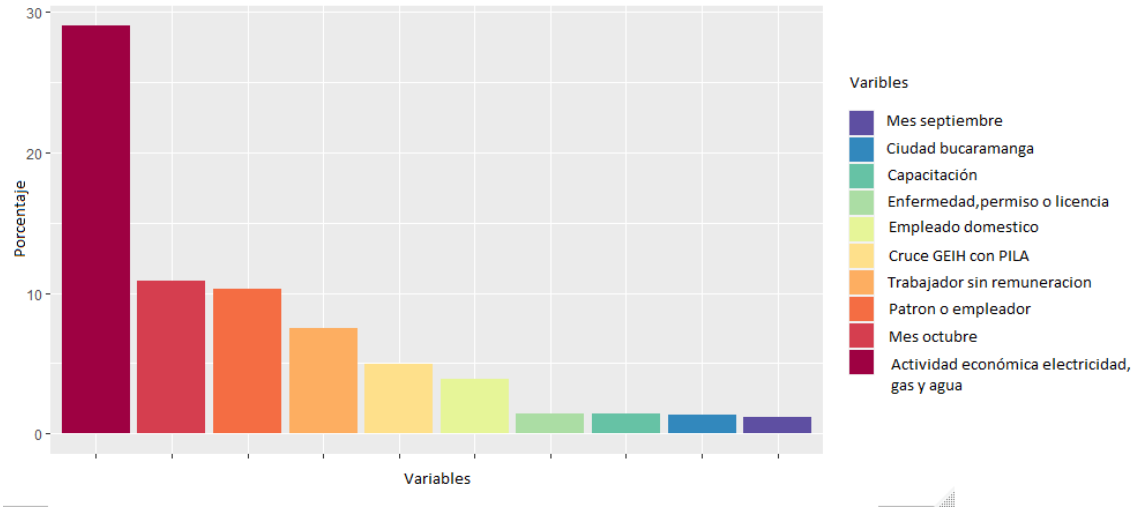
[5] Esto corresponde al hecho de observar valores pequeños en la matriz de confusión por fuera de la diagonal.

Recall	0,8622
F1 score	0,8579

Fuente: elaboración propia

Un atributo interesante de este algoritmo, no generalizable a otros algoritmos de aprendizaje de máquina, es que ofrece la posibilidad de inferir la contribución de cada una de las variables en el proceso de clasificación. Los resultados, presentados en la Figura 2, indican que las cinco variables con mayor incidencia en la clasificación de los ocupados entre formal e informal son la actividad económica electricidad, gas y agua; mes de octubre, posiciones ocupacionales de empleador y trabajador familiar sin remuneración, y la variable de cruce PILA-GEIH. Esta última muestra la utilidad en el uso de registros administrativos para el propósito de la imputación.

Figura 2. Importancia relativa de las características



Fuente: elaboración propia

El RF obtenido se aplica a los meses de marzo y abril de donde se imputa la condición de informalidad de los ocupados. Esto permite no solo realizar los análisis a nivel de microdatos y correlacionar la informalidad con características de los individuos, sino también recuperar la serie de la tasa de informalidad (ver tabla 5).

Tabla 5. Resultado de la imputación de informalidad

		mar-19	abr-19	mar-20 (e)	abr-20 (e)
13 ciudades y áreas metropolitanas	Formales	5.761.320	5.563.060	5.490.990	4.327.579
	Informales	4.999.454	5.068.967	4.317.637	3.335.724
	Informalidad (%)	46,46%	47,68%	44,02%	43,53%
23 ciudades y áreas metropolitanas	Formales	6.236.682	6.061.929	5.926.923	4.678.564
	Informales	5.693.664	5.744.017	4.903.411	3.724.109
	Informalidad (%)	47,72%	48,65%	45,27%	44,32%

Fuente: elaboración propia GEIH

6. BIBLIOGRAFÍA

- Athey, S., & Imbens, G. W. (2019). Machine learning methods economists should know about, arxiv.

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

- Gerunov, A. (2014). Big data approaches to modeling the labor market.

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

- Lantz, B. (2015). Machine learning with R. Packt Publishing Ltd.

- OIT (2020). COVID-19: Orientaciones para la recolección de datos de las estadísticas del trabajo. https://ilo.org/wcmsp5/groups/public/---dgreports/---stat/documents/publication /wcms_745104.pdf

- UNECE (2018). The use of machine learning in official statistics

- UNECE (2018). Two-phase and double machine learning for data editing and imputation