Evaluación de modelos de *nowcasting* de la tasa de POBREZA MONETARIA

LUIS EDUARDO CASTILLO* Y SAMANTHA GUILLÉN**

Contar con información frecuente de la pobreza es un elemento clave para evaluar el bienestar de los hogares peruanos. Sin embargo, en Perú, la tasa de pobreza oficial de un año determinado es publicada en el segundo trimestre del año siguiente, por lo que cuenta con cuatro meses de rezago. Con el fin de realizar estimaciones con mayor frecuencia, en este artículo se proponen modelos de *machine learning* y microsimulaciones. En particular, el método de *machine learning* permite utilizar información adelantada de las ENAHO trimestrales para calcular tasas de pobreza trimestrales. Por su parte, el método de microsimulaciones permite estimar una tasa de pobreza anual durante el año en curso, empleando información económica ejecutada y proyectada. Los resultados indican que ambos métodos podrían ser útiles para dicho objetivo.

* **Jefe, Departamento de Políticas Sociales y Regionales del BCRP**
luiseduardo.castillo@bcrp.gob.pe

** **Especialista, Departamento de Políticas Sociales y Regionales del BCRP**
samantha.guillen@bcrp.gob.pe

En Perú, la tasa de pobreza monetaria oficial es publicada de forma anual por el Instituto Nacional de Estadística e Informática (INEI) con base en las mediciones de la Encuesta Nacional de Hogares (ENAHO). Esta encuesta se ejecuta de forma continua a lo largo del año, pero los procesos de imputación y estimación que sirven para el cálculo de la tasa de pobreza recién se realizan con la encuesta anual. Todo esto conlleva a que la tasa de pobreza de un año determinado recién sea publicada en el segundo trimestre del siguiente año. Este rezago en la publicación de la cifra genera el interés por predecir de forma adelanta la tasa de pobreza del año en curso.

En este contexto, este artículo explora el uso de algoritmos de *machine learning* (ML) y de microsimulaciones para predecir la tasa de pobreza, con la finalidad de aportar al debate sobre la estimación más frecuente de este indicador clave de bienestar. Estas técnicas de *nowcasting* de la pobreza se alimentan de indicadores económicos agregados, como el PBI o la inflación, pero también de la información de ingresos, gastos y características de los hogares que se encuentran en los adelantos trimestrales de la ENAHO.

> El *machine learning* (aprendizaje automático) es una técnica de inteligencia artificial que permite que **los algoritmos predigan y reconozcan patrones basados en un conjunto de datos.**

MEDICIÓN DE LA POBREZA Y LA ENAHO

En Perú, la pobreza monetaria se calcula al comparar el gasto per cápita mensual de un hogar con el valor de una canasta básica de consumo, también denominada línea de pobreza. De esta manera, se considera que un hogar es pobre si tiene un gasto per cápita inferior a dicha línea. Esta metodología requiere, por ende, información del gasto total de los hogares, así como del valor actualizado de la línea de pobreza, los cuales únicamente están disponibles en la versión anual de la ENAHO.

$$Gasto\ per\ cápita\ mensual < línea\ de\ pobreza \rightarrow Hogar\ es\ pobre$$

Al respecto, los avances trimestrales de la ENAHO no reportan directamente el gasto total del hogar ni la línea de pobreza. Si bien algunos módulos de la ENAHO trimestral, como los de educación, salud o vivienda, incluyen preguntas sobre gastos en estos rubros, la información es parcial y no permite reconstruir la totalidad del gasto reportado en la ENAHO anual. Por ello, la pobreza no puede ser estimada directamente de manera trimestral con la referida encuesta.

CUADRO 1 Diferencias entre la ENAHO anual y trimestral

	ENAHO anual	ENAHO trimestral
Pobreza y gasto del hogar	✓	× Información parcial de gastos del hogar
Empleo e ingresos	✓	×
Educación	✓	✓
Salud	✓	✓
Vivienda	✓	✓
Características sociodemográficas	✓	✓

Frente a esta limitación, se exploran dos alternativas. En primer lugar, el uso de modelos de *machine learning* de clasificación permitiría identificar la condición de pobreza de un hogar, aun sin información completa sobre su gasto o línea de pobreza. Para ello, los algoritmos utilizarían información de otras características sociodemográficas y del contexto macroeconómico de un hogar para "aprender" a clasificarlo como pobre o no pobre. De esta manera, sería posible aprovechar los datos disponibles en la ENAHO trimestral, así como variables macroeconómicas para generar estimaciones de pobreza.

No obstante, el método de *machine learning* requiere información adelantada de la ENAHO trimestral, tal que es útil únicamente cuando estas bases están disponibles. En su defecto, un método más flexible es el de microsimulaciones. Este método estima un gasto per cápita contrafactual con base en cambios esperados en el ingreso, empleo y nivel de precios que experimentan los hogares durante el periodo de análisis. Para ello, se parte de los hogares observados en la última publicación anual de la ENAHO, tal que no se requiere de los adelantos de la ENAHO trimestral.

MODELOS DE *MACHINE LEARNING*

El *machine learning* (aprendizaje automático) es una técnica de inteligencia artificial que permite que los algoritmos predigan y reconozcan patrones basados en un conjunto de datos. En este artículo, se utilizan algoritmos de ML para clasificar a un hogar en una de dos categorías: pobre o no pobre. Para ello, se

utilizan modelos de clasificación con aprendizaje supervisado, que consisten en entrenar a los modelos con información histórica para clasificar a cada hogar en una categoría.

En específico, se entrenaron cuatro modelos de ML: regresión logística, *decision tree* (árboles de decisión), *random forest* (bosque aleatorio) y *K-nearest neighbors* (K vecinos más cercanos).

- **Regresión logística:** Este modelo estima la probabilidad condicional de que un hogar sea pobre o no mediante una regresión logística. Con base en un punto de corte en la probabilidad, clasifica a un hogar como pobre o no pobre.
- ***Decision tree*:** Este modelo clasifica a los hogares según reglas sobre las respuestas a las variables predictoras incluidas en la estimación. En el ejemplo ilustrativo, se puede observar cómo funciona un modelo *decision tree* para clasificar a un hogar como pobre. Si el ingreso es "alto", pero el hogar no tiene servicios básicos y el jefe tiene un nivel de educación "bajo", entonces se considera al hogar como pobre. Cabe destacar que este es un ejemplo para explicar el algoritmo y no representa al modelo final usado en este documento.
- ***Random forest*:** Este modelo utiliza diferentes árboles de decisión para generar predicciones (pobre o no pobre), usando diferentes subconjuntos de datos y características de manera aleatoria. La predicción final consiste en elegir por mayoría la clase más votada entre todos los árboles de decisión.
- ***K-nearest neighbors* (KNN):** El modelo asigna una clasificación (pobre o no pobre) a una observación basándose en las clases de sus K vecinos más cercanos en el espacio de características. Así, KNN es un modelo supervisado dado que utiliza datos etiquetados de la ENAHO anual durante el entrenamiento para realizar predicciones sobre nuevas observaciones.

Al ejecutar modelos de ML supervisado, los datos usualmente se dividen en dos subconjuntos: entrenamiento y prueba. El conjunto de entrenamiento se utiliza para ajustar los parámetros de los modelos y representa el 20 por ciento de la muestra total. Esto sigue la división usualmente empleada en ML. Similarmente, el conjunto de prueba sirve para evaluar el rendimiento.

Dado que la proporción de hogares pobres encuestados es significativamente menor que la de no pobres, se enfrenta un potencial problema de desbalance en los datos que perjudicaría la estimación. De forma intuitiva, si el algoritmo busca aprender de los hogares pobres, pero encuentra poco de estos en la muestra, fallará en el proceso de clasificación. Por ello, se realiza una técnica de *oversampling* (sobremuestreo), que consiste en generar nuevas muestras sintéticas que replican las observaciones de la categoría minoritaria (en este caso, la de los hogares pobres). Esto permite que el modelo aprenda mejor de

GRÁFICO 1 Ejemplo ilustrativo de un modelo *decision tree*

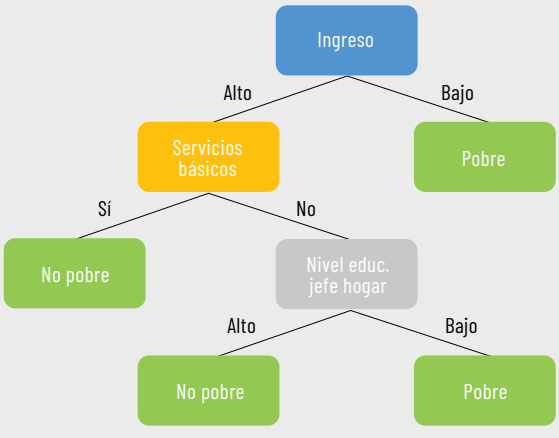


GRÁFICO 2 Ejemplo ilustrativo de un modelo *random forest*

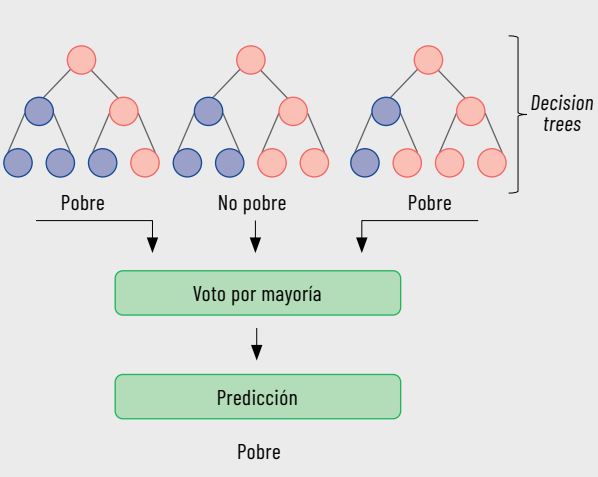
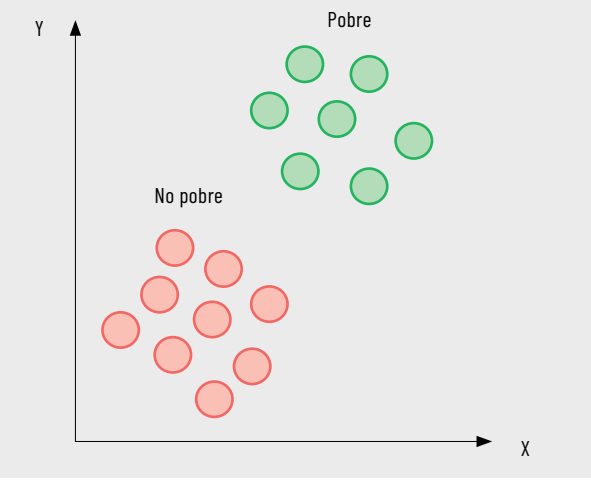


GRÁFICO 3 Ejemplo ilustrativo de un modelo KNN



CUADRO 2 **Número de hogares según tipo de ENAHO**

Periodo	Hogares presentes en la base anual y trimestral	Hogares presentes solo en la base trimestral	Hogares presentes solo en la base anual
2020	32 272	220	2 218
2021	33 006	120	1 239
2022	32 668	177	1 545
2023	32 314	249	1 572

FUENTE: INEI-ENAHO.

los patrones de hogares en situación de pobreza.

Para entrenar los modelos, se utilizó información de la ENAHO de 2020 a 2022. En específico, se enlazan las bases trimestrales —que no identifican la situación de pobreza de un hogar— con el identificador de hogar pobre de las bases anuales. Con ello, se utiliza información de los hogares presentes en las bases anuales y trimestrales. Este paso es clave para que los algoritmos se entrenen para clasificar a un hogar como pobre o no. Cabe mencionar que a las variables de estas bases se les realizó un tratamiento de valores faltantes imputándolos mediante un algoritmo de vecinos cercanos (*K-nearest neighbors*)[1]. Los datos de 2023 se utilizan para evaluar el rendimiento de los modelos fuera de los años de entrenamiento (2020 a 2022).

Para entrenar a los algoritmos, se emplearon variables relacionadas a las características de la vivienda del hogar y del jefe del hogar, así como a las condiciones macroeconómicas que enfrentaron en el momento de su entrevista. Cabe mencionar que también se considera el agregado de los gastos presentes en las bases trimestrales ("gasto básico") y una línea de pobreza actualizada según la inflación del año.

RESULTADOS

Los modelos de predicción son evaluados según las siguientes métricas:

- ***Accuracy* (exactitud):** proporción de predicciones correctas sobre el total de predicciones. Mide qué tan bien el modelo clasifica en general.
- ***Precision* (precisión):** proporción de verdaderos positivos (pobres clasificados correctamente) entre todas las predicciones positivas (predicciones de pobres). Mide cuán preciso es el modelo cuando predice una clase positiva, es decir, cuántas de las predicciones positivas son realmente correctas.
- ***Recall* (sensibilidad):** proporción de verdaderos positivos entre todos los casos que realmente son

CUADRO 3 **Variables predictoras en los modelos de clasificación**

Rubro	Variable	Descripción
Variables temporales	Año	Año de la encuesta
	Trimestre	Trimestre de la encuesta
	Pandemia	Control por año 2020
Características de la vivienda	Área geográfica	Urbano o rural
	Región	Costa, sierra o selva, y Lima Metropolitana
	Número de miembros del hogar	
	Vivienda inadecuada	Si tiene pared, techo o combustible inadecuados, o piso tierra
	Servicios básicos	Acceso a electricidad, desagüe y agua
Características del jefe del hogar	Nivel educativo	Primaria, secundaria, superior no universitaria o superior universitaria
	Edad	
	Sexo	
Gasto básico y pobreza	Gasto básico real per cápita total	Aproximación al gasto total de los hogares
	Decil de gasto básico real per cápita	
	Otros rubros de gasto	Gasto real per cápita en alimentos, alojamiento, salud, transporte y restaurantes
	Línea de pobreza	Línea de pobreza actualizada con la inflación
Condiciones macroeconómicas	Inflación	Variación del deflactor de ingresos de la ENAHO
	Crecimiento de PBI regional	Por trimestres, variación interanual
	Crecimiento del empleo nacional	Por trimestres, variación interanual

1 La imputación mediante KNN consiste en llenar valores faltantes de una variable usando la moda de los valores de los k-vecinos más cercanos considerando que los vecinos comparten patrones similares. Sin esta imputación, se pierden alrededor de 3 mil observaciones de un total de 98 mil observaciones.

positivos (pobres). Mide cuántos de los casos positivos reales el modelo fue capaz de identificar correctamente.

- **AUC (área bajo la curva ROC[2]):** evalúa el rendimiento del modelo en diferentes umbrales de clasificación. Mide la capacidad del modelo para distinguir entre clases.

Los resultados indican que los modelos con mejor ajuste serían *random forest* y regresión logística. El primero genera alrededor de 90,7 por ciento de predicciones correctas, y clasifica correctamente al 80,8 por ciento de hogares pobres. Asimismo, según el AUC, el modelo tendría un buen rendimiento en su clasificación. Cabe señalar que estas métricas se basan en la muestra de prueba.

Dado que el objetivo consiste en predecir la tasa de pobreza trimestral, se puede evaluar el desempeño de cada modelo con la raíz del error cuadrático medio (ECM). Bajo este criterio, los modelos de regresión logística y *random forest* son los que generan mejores predicciones, incluso en años fuera de la muestra (2023).

Con el objetivo de continuar mejorando la predicción, se entrenan los modelos de *random forest* y regresión logística —que son los que mostraron mejor rendimiento— por área geográfica. Esto busca que los algoritmos se enfoquen en las características específicas de la pobreza rural y urbana. Al entrenar los modelos por separado, se logra mejorar la capacidad predictiva. En la regresión logística, la raíz del ECM de 2023 se mantiene en 1,1 p.p., mientras que en *random forest* disminuye de 3,3 p.p. a 1,5 p.p.

CUADRO 4 Métricas de evaluación

	Accuracy	*Precision*	*Recall*	AUC
Regresión logística	90,5	75,7	79,6	95,6
Decision tree	90,1	74,3	80,0	94,9
Random forest	90,7	75,9	80,8	96,1
K-nearest neighbors	88,7	68,1	85,6	94,4

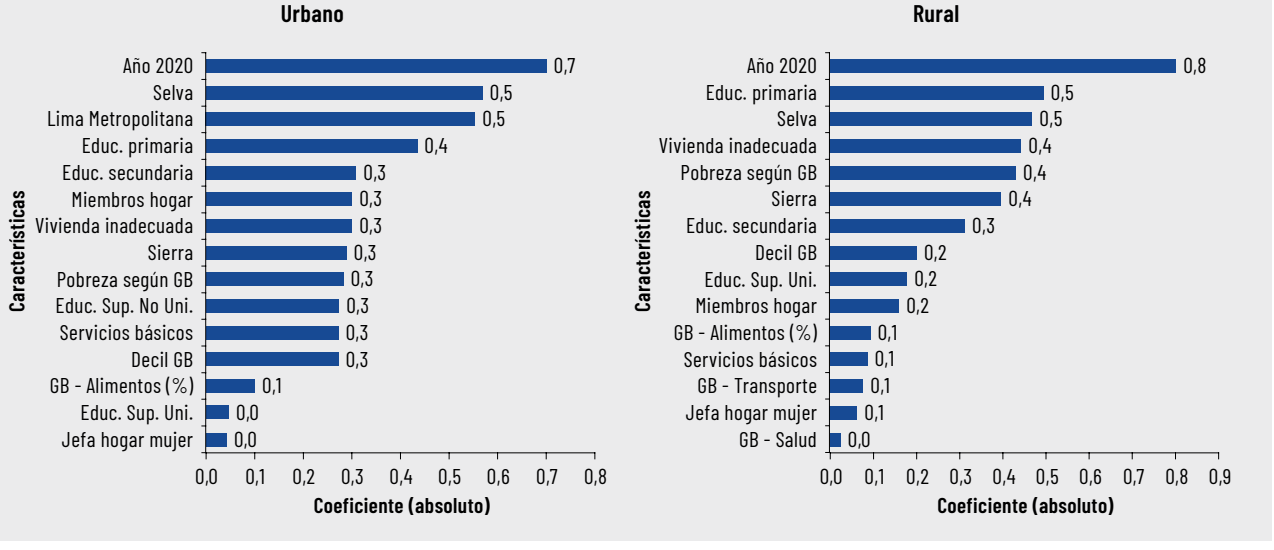
CUADRO 5 Raíz del error cuadrático medio, 2020-2023

	2020-2022	2023
Regresión logística	1,9	1,1
Decision tree	2,5	4,0
Random forest	2,5	3,3
K-nearest neighbors	6,0	7,6

CUADRO 6 Métricas de evaluación

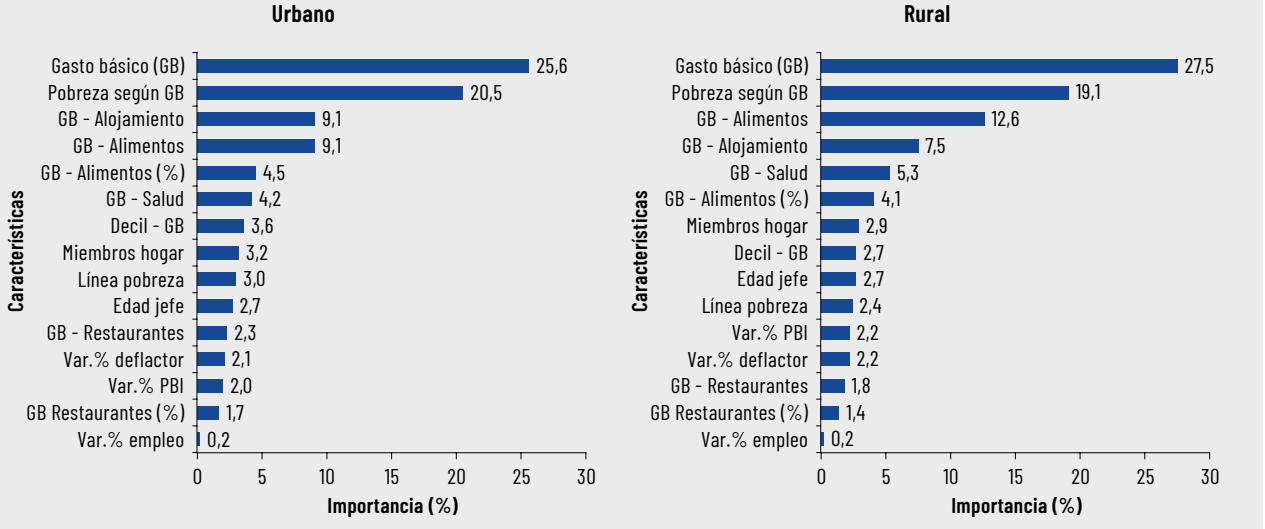
	Accuracy	*Precision*	*Recall*	AUC
Regresión logística				
Rural	87,2	80,6	78,1	94,2
Urbano	92,0	75,1	67,8	96,4
Random forest				
Rural	87,4	79,8	80,4	94,4
Urbano	92,3	74,3	72,3	96,6

GRÁFICO 4 Regresión logística: coeficientes absolutos



2 La curva ROC (*receiver operating characteristic*) es una representación gráfica utilizada para evaluar la capacidad de un modelo de clasificación binaria. Muestra la relación entre la tasa de verdaderos positivos (sensibilidad) y la tasa de falsos positivos para diferentes umbrales de decisión del modelo. El área bajo la curva ROC (AUC) es una métrica común que indica qué tan bien el modelo distingue entre las clases, siendo 1 un modelo perfecto y 0.5 un modelo sin capacidad predictiva (similar a una predicción al azar).

GRÁFICO 5 ***Random forest*: importancia de las variables predictoras**
(En porcentaje)



Al desglosar por la importancia de cada variable en el modelo *random forest*, se puede observar que el gasto básico y la aproximación de pobreza juegan un rol importante en la capacidad predictiva del modelo.

En el modelo de regresión logística, la educación del jefe del hogar, la región y área geográfica contribuirían a explicar la probabilidad de que un hogar sea pobre o no.

> La pobreza monetaria es uno de los indicadores de bienestar más importantes, al representar la capacidad para satisfacer necesidades básicas. Como tal, **a pesar de ser necesario para una estimación adecuadamente técnica y robusta, el rezago en la publicación de las cifras dificulta el monitoreo continuo de las condiciones de vida de la población.**

MÉTODO DE MICROSIMULACIONES

El método de microsimulaciones consiste en simular las variaciones en los ingresos laborales y no laborales, en el estado de empleo y en el costo de la canasta básica (i.e., línea de pobreza monetaria) para el grupo de hogares observados en la última ENAHO anual publicada. Con ello, se genera un gasto básico contrafactual que permite reclasificar a los hogares como pobres o no pobres. Por ejemplo, para estimar la tasa de pobreza de 2024, se partiría de los hogares observados en la ENAHO de 2023. Luego, se le insumen los cambios previstos en ingresos, empleo y línea de pobreza durante el año, dada la información económica disponible. En términos sencillos, el método respondería la siguiente pregunta: ¿cuál sería la tasa de pobreza si los hogares de 2023 reciben los choques económicos esperados para 2024?

Este método entonces no requiere de adelantos de la ENAHO trimestral, sino que parte de la última publicación anual para recuperar la tasa de pobreza anual. Es decir, el método permite adelantar una tasa de pobreza anual utilizando información económica ejecutada.

Por la naturaleza del método, la estimación de la tasa de pobreza depende de los supuestos sobre la evolución de los ingresos laborales, la transición en el empleo y el costo de la canasta básica. Los supuestos principales planteados para el ejercicio son:

- **La línea de pobreza crecería según los datos ejecutados de la inflación promedio.** En específico, se utiliza la inflación del rubro de alimentos y bebidas para actualizar el componente alimentario de la línea de pobreza. El componente no alimentario de la línea de pobreza se actualiza con

CUADRO 7 **Transiciones entre estados de empleo, 2011-2019**

Situación inicial	Situación final							
	Dependiente formal		Dependiente informal		Independiente		Desempleado	
	Prob.	Premio/ penalidad	Prob.	Premio/ penalidad	Prob.	Premio/ penalidad	Prob.	Premio/ penalidad
Dependiente formal	84%	-	11%	0,4	5%	0,4	0%	1
Dependiente informal	14%	2,3	64%	-	22%	0,9	0%	1
Independiente	3%	2,5	12%	1,1	85%	-	0%	1
Desempleado	10%	A	24%	B	11%	C	55%	-

NOTA: EN EL CASO DEL PASO DEL DESEMPLEO A UNA SITUACIÓN DE EMPLEO, SE COLOCAN LOS INGRESOS DEL PRIMER CUARTIL DE LOS TRABAJADORES DEPENDIENTE FORMAL (A), DEPENDIENTE INFORMAL (B) E INDEPENDIENTE (C).
FUENTE: INEI-ENAHO PANEL 2011-2019.

la inflación sin alimentos y bebidas. La nueva línea de pobreza es la suma de ambos componentes.

- **La variación en el ingreso laboral real de los trabajadores** equivaldría al crecimiento del PBI del sector en que labora dicho trabajador, multiplicado por la elasticidad ingreso laboral–PBI de cada tipo de trabajador[3].
- **Asimismo, se consideran transiciones entre tipos de empleo que podrían generar pérdidas o ganancias de ingresos para cada trabajador.** Con base en la información de empleo, se estima una matriz de transición entre trabajador dependiente formal, dependiente informal, independiente y desempleo.

  Se considera que el cambio en el estado laboral viene acompañado, según sea el caso, de una penalidad/bonificación a los ingresos, que se estima a partir de las remuneraciones promedios de los diferentes tipos de empleos. Por ejemplo, si un trabajador transita de dependiente formal a independiente, la penalidad sería equivalente a la diferencia porcentual entre el ingreso promedio de los trabajadores independientes y dependientes formales.
- **La propensión marginal por consumir de los hogares se mantendría constante.** Este supuesto permite recrear la distribución de gasto de las familias a partir de las nuevas series de ingresos. Con ello, se compara el gasto per cápita con la línea de pobreza actualizada para determinar si un hogar es pobre.
- Es importante mencionar que se pueden incorporar cambios en las políticas de transferencias del gobierno. Si se observa un incremento en el monto de pago a los usuarios de programas sociales, un aumento significativo en el número de beneficiarios, o alguna transferencia extraordinaria por parte del gobierno (por ejemplo, los bonos de la pandemia), esto se podría agregar dentro del ingreso total del hogar.

Dado que las transiciones entre estados de empleo se computan de forma aleatoria, se repite el ejercicio contrafactual 100 veces para converger a un estimado promedio de pobreza. En las simulaciones, asumir que un individuo en particular pase o no al desempleo o a la informalidad/formalidad modifica el salario final. Por ello, es necesario repetir los cálculos y promediar los resultados.

Finalmente, se puede agregar que este método puede ser utilizado para proyecciones, y no solo para técnicas de *nowcasting*. En ese caso, el método no estaría condicionado a los datos ejecutados de PBI por sector o de inflación por rubros, sino a las proyecciones de tales variables.

RESULTADOS

Para evaluar el desempeño del modelo, se calcula la raíz del error cuadrático medio para el periodo 2011 a 2023, utilizando como referencia las tasas de pobreza anuales. Es decir, el método proyecta dichas tasas, las cuales se utilizan como referencia para el cálculo del error cuadrático medio. Con ello, la raíz del error cuadrático medio del método de microsimulaciones es 2,8 p.p.

COMENTARIOS FINALES

Este artículo presenta dos métodos diferentes para la estimación de la tasa de pobreza monetaria ante la ausencia de datos oficiales. La pobreza monetaria es uno de los indicadores de bienestar más importantes, al representar la capacidad para satisfacer necesidades básicas. Como tal, a pesar de ser necesario para una estimación adecuadamente técnica y robusta, el rezago en la publicación de las cifras dificulta el monitoreo continuo de las condiciones de vida de la población. Por ello, los modelos de *machine learning* y el método de microsimulaciones podrían ser utilizados para proveer información oportuna a los analistas económicos y hacedores de política.

3 Las elasticidades se estiman mediante una regresión lineal con datos trimestrales de ingreso laboral real y PBI entre 2012 y 2023, controlado por la ocurrencia de la pandemia (2020 y 2021), para cada tipo de trabajador. Particularmente, se usa la siguiente especificación para cada tipo de trabajador: $\text{var}\ ing = \beta + \beta_1 \text{var}\, PBI + \beta_2\, pandemia$. Asimismo, al pasar a ingresos nominales, se asume una tasa de inflación promedio.