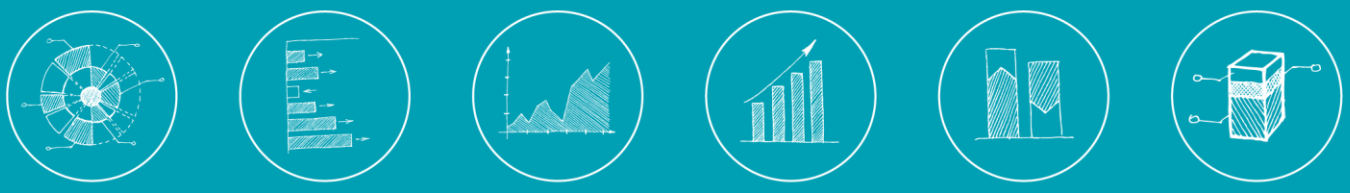
Cuaderno de Trabajo
No. 15 – Mayo 2023

Imputación de Información en Encuestas de Hogares a través de técnicas de Machine Learning: Análisis del caso ecuatoriano

2023



Elaboración técnica:

Buenas cifras,
mejores vidas

Autoridades:

Roberto Castillo A.
Director Ejecutivo

Jorge García-Guerrero
Subdirector General

Lorena Moreno E.
Coordinadora General Técnica de Innovación en Métricas y Análisis de la Información

Cristhian Rosales C.
Director de Estudios y Análisis de la Información

Revisión Institucional:

Cristhian Rosales C.

Autores:

Diego Del Pozo V.
Instituto Nacional de Estadística y Censos, Ecuador

Andrés Villacís M.
Instituto Nacional de Estadística y Censos, Ecuador

Elizabeth Feijoó S.
Instituto Nacional de Estadística y Censos, Ecuador

Los Cuadernos de Trabajo Temáticos son documentos que presentan análisis de fenómenos sociales, económicos y ambientales con el objetivo de promover la investigación e incentivar el debate.

Las interpretaciones y opiniones expresadas en este documento pertenecen a los autores y no reflejan el punto de vista oficial del Instituto Nacional de Estadística y Censos (INEC). El INEC ha realizado una revisión del documento, no obstante, no garantiza la exactitud de los datos que figuran en el documento.

Expresamos nuestro agradecimiento a Cristhian Rosales por su invaluable contribución y orientación, los cuales han sido fundamentales para la realización exitosa de este estudio.

Imputación de Información en Encuestas de Hogares a través de técnicas de Machine Learning: Análisis del caso ecuatoriano

Diego Del Pozo V.; Andrés Villacís M.; Elizabeth Feijoó S.

Resumen

Este estudio propone una solución para la imputación de información perdida o incompleta en la Encuesta Nacional de Empleo, Desempleo y Subempleo (ENEMDU) anual del 2021. Específicamente, se enfoca en las preguntas relacionadas con la tenencia de Registro Único de Contribuyentes (RUC) del lugar de trabajo e ingreso laboral, utilizando técnicas de *machine learning*. A lo largo del documento se presentan y evalúan distintos modelos, como son *ensemble learning* (*random forest* y XGBoost), *neural networks* (*multilayer perceptron*) y métodos supervisados simples (*support vector machine*), para luego definir como la técnica de imputación más óptima (XGboost) y realizar la imputación de las variables. Además, se analiza la consistencia de los resultados obtenidos reconstruyendo los principales indicadores laborales a partir de las variables imputadas y contrastando con las cifras oficiales.

Palabras clave: *machine learning*, imputación, datos perdidos, encuestas.

Introducción

El manejo de datos perdidos (*missing*), faltantes o incompletos es un problema habitual en los análisis estadísticos, lo que puede afectar la precisión de las cifras oficiales obtenidas a partir de encuestas de hogares, censos poblacionales, registros administrativos o cualquier análisis cuantitativo que se base en grandes conjuntos de datos generado por institutos de estadística (Rosati, 2021; Donza, 2013).

En particular, en el caso de las encuestas de hogares, es común encontrar información faltante o incompleta, que puede ser causada por diversos motivos técnicos, como: i) la no respuesta de toda la unidad de análisis, es decir, cuando el hogar no desea participar en la encuesta o el hogar no es ubicado por los encuestadores; ii) la no respuesta a ítems determinados de la encuesta, lo cual sucede cuando la familia o el individuo no desea responder a las preguntas realizadas por el encuestador, ya sea por dificultad en la comprensión de la pregunta, falta de conocimiento de la respuesta o la resistencia del informante a otorgar cierta información; iii) errores humanos durante el pre-procesamiento de la información (Thamelo et al., 2021; Restrepo y Marín, 2012).

Por lo tanto, los institutos de estadística han tenido que buscar herramientas para subsanar la existencia de datos perdidos de manera continua y permanente (Rosati, 2021; Restrepo y Marín, 2012). Entre las estrategias técnicas abordas para resolver este problema se encuentra la imputación de la información faltante utilizando técnicas estadísticas, como el reemplazo de los valores faltantes por la media/mediana/moda de las unidades observadas en la variable, la duplicación de valores conocidos pertenecientes a individuos con características similares, entre otras (Lin y Tsai, 2019; Donza, 2013; CEPAL, 2007). Además, se han implementado técnicas basadas en modelos de *machine learning*, que permiten la estimación de la información faltante

para diferentes fuentes de información y volumen de datos (UNECE, 2022; Rosati, 2021; Tlamelo et al., 2021). Esta última técnica ha experimentado un crecimiento constante en los últimos años debido a su versatilidad en la estimación de información perdida para variables cuantitativas y cualitativas de diferentes fuentes de información y por presentar una precisión en las estimaciones de los valores perdidos superior a la obtenida mediante la aplicación de técnicas de imputación estadísticas[1] (Rosati, 2021; Tlamelo et al., 2021).

Este estudio investiga la capacidad de diversas técnicas de *machine learning* para imputar información perdida en la Encuesta Nacional de Empleo, Desempleo y Subempleo (ENEMDU) anual de 2021, llevada a cabo por el Instituto Nacional de Estadística y Censos (INEC). Específicamente, el estudio se centra en las preguntas relacionadas con la tenencia de Registro Único de Contribuyentes (RUC)[2] del lugar de trabajo (p49) y la variable de ingreso laboral (ila). En el caso de la variable ila, se construyó una nueva variable, la cual se denominó new_ila. La variable mantuvo los criterios metodológicos oficiales[3], pero en el caso de los ingresos negativos no los asignó como valores faltantes, sino que preservó su valor. Además, se categorizó a la variable en tres grupos: 1.Ingreso mayor o igual al salario básico unificado - SBU[4], 2. Ingreso menor o igual al SBU, y 3.Ingresos negativos.

En el estudio se probaron distintas técnicas de *machine learning*, como aprendizaje por conjuntos (*Random Forest* - RF y XGBoost), *neural networks* (*Multilayer Perceptron* - MLP) y métodos supervisados simples (*Support Vector Machine* - SVM). El estudio evaluó el desempeño de estas técnicas en la imputación de las variables p49 y new_ila en la ENEMDU de 2021. Se utilizó la métrica de precisión (*accuracy*) de cada modelo para seleccionar la técnica más efectiva en la imputación definitiva de la información perdida. Los resultados mostraron que el modelo XGBoost tuvo un rendimiento superior, con una precisión del 88,7% para la variable p49 y del 83,8% para la variable new_ila, en comparación con las técnicas RF (precisión del 88,1% para p49 y 83,3% para new_ila), MLP (precisión del 88,4% para p49 y 83,3% para new_ila) y SVM (precisión del 88,4% para p49 y 83,2% para new_ila). Por lo tanto, se seleccionó XGBoost para realizar la imputación definitiva.

Una vez determinada la técnica con mejor rendimiento predictivo (XGBoost), se imputó la información perdida en las preguntas p49 y new_ila. Luego, con la información completa, se generaron las variables de sectorización del empleo (que incluye la variable p49 en su construcción) y clasificación de la condición de actividad de los empleados (que incluye la variable new_ila en su construcción).

Como resultado de la imputación, se observaron cambios en la distribución de las variables. En el caso de la variable p49, se encontró que de las 424.904 personas empleadas que respondieron "no sabe" en la pregunta, 307.525 personas fueron

[1] Dependiendo de la técnica estadística implementada para la imputación de la información faltante se puede atenuar sesgos de información, así como reducir la veracidad y precisión de los indicadores estimados (CEPAL, 2007).
[2] El Registro Único de Contribuyentes (RUC) es el documento que identifica e individualiza a los contribuyentes, personas físicas o jurídicas, para fines tributarios (SRI, 2022).
[3] La metodología oficial del INEC para la construcción del ila se encuentra en el siguiente link: https://www.ecuadorencifras.gob.ec/documentos/web-inec/Bibliotecas/Revista_Estadistica/Aspectos_metodologicos_sobre_la_medicion_de_la_pobreza_por_ingresos_en_el_Ecuador.pdf
[4] En el Ecuador el salario básico unificado (SBU) para el año 2021 fue de $400 dólares en el territorio continental y de $720 dólares en la región insular de Galápagos.

categorizadas como "sí" y 117.378 personas como "no". Por otro lado, al imputar la variable new_ila, se logró predecir los ingresos para un total de 66.443 empleados que inicialmente no proporcionaron información sobre sus ingresos laborales. Estos empleados se distribuyeron de la siguiente manera: 52.828 personas fueron clasificadas como 1.Ingreso mayor o igual al SBU, 13.605 personas como 2.Ingreso menor al SBU y 10 personas como 3.Ingreso negativo.

Una vez observados los cambios en la distribución de las variables imputadas, se procedió a comparar los resultados de los indicadores generados a partir de estas variables, específicamente el sector del empleo y la condición de actividad, antes y después de la imputación.

El análisis de los indicadores estimados basados en la variable de sectorización del empleo, tanto antes como después de la imputación, reveló un aumento significativo en la tasa de empleo en el sector formal, pasando del 42,9% al 46,8%. Además, se observó un cambio en la tasa de empleo en el sector informal, que pasó del 49,6% al 51,0%, también con una diferencia estadísticamente significativa. Es importante destacar que, aunque las diferencias en los estimadores mencionados resultaron ser estadísticamente significativas, no se encontraron cambios en la dispersión en comparación con las cifras oficiales.

En cuanto a los indicadores derivados de la variable condición de actividad de los empleados, se observó que la tasa de empleo adecuado experimentó un ligero incremento del 32,5% al 33,1%, aunque esta diferencia no fue estadísticamente significativa. Del mismo modo, la tasa de otro empleo inadecuado también experimentó un leve aumento del 27,2% al 27,3%, sin alcanzar la significancia estadística. Los resultados obtenidos no revelaron cambios en la dispersión.

Estos hallazgos indican que la imputación de la información perdida mediante técnicas de *machine learning* no presentó un impacto significativo en la estimación de estos indicadores de empleo; sino que por el contrario, se mantuvieron las cifras consistentes.

El documento se encuentra organizado de la siguiente manera: la primera sección muestra la revisión de literatura de las experiencias en la imputación de encuestas y el manejo de datos perdidos; la segunda sección profundiza las fuentes de información y metodologías implementadas en el estudio; la tercera sección ilustra los resultados alcanzados; y por último, la cuarta sección recaba las conclusiones desprendidas del estudio.

1. Revisión de literatura

1.1. Imputación de encuestas de hogares

Las encuestas de hogares son el resultado de un conjunto de procesos estandarizados y homologados en un modelo de producción estadística[5], cuyo objetivo es garantizar

[5] En el INEC el modelo de producción estadística consta de ocho fases: planificación, diseño, construcción, recolección, procesamiento, análisis, difusión y evaluación. De las cuales las etapas de recolección (contempla la recopilación de toda la información necesaria a partir del uso de diferentes métodos de recolección, para su posterior almacenamiento en un ambiente apropiado y seguro) y procesamiento (comprende la depuración de datos, generación de resultados estadísticos y preparación para el análisis y difusión) son del interés para fines del documento (Castillo y Puebla, 2016).

la calidad de la información obtenida (Castillo y Puebla, 2016). A pesar de esto, estas encuestas pueden presentar errores de medida. Por esta razón, se han desarrollado dos conceptos para conocer las fortalezas y debilidades de las mediciones efectuadas en encuestas de hogares: i) el primero es la completitud, que se refiere a la medida en que se obtienen respuestas a la pregunta indagada, es decir, la proporción de encuestados que proveen una respuesta válida a la pregunta; ii) el segundo es la estrategia de indagación, que se refiere a los abordajes completos y las preguntas empleadas para captar diferentes tipos de información (Castillo y Puebla, 2016; Beccaria y Glusman, 2013). A partir de estos dos conceptos, las oficinas nacionales de estadística pueden analizar la calidad de sus operaciones estadísticas y proponer mejoras para superar las limitaciones de las fases de recolección (información perdida, incompleta o sub-reportada) y de procesamiento de la información (información atípica, incorrecta, no fiable), que pueden decantar en la ausencia de información (Castillo y Puebla, 2016).

Este documento se enfoca en la estrategia de indagación y en la problemática asociada a la ausencia total o parcial de información (valores perdidos o sin respuesta) que se presenta regularmente en encuestas de hogares. Según la literatura teórica, la existencia de valores perdidos se relaciona con errores de medición, ignorancia o negativa de los informantes, problemas en el muestreo[6], y fallas en los equipos de recolección y procesamiento de la información. Por consiguiente, esta situación tiene un efecto inmediato en el análisis de los datos y en la generación de resultados (Raja y Thangavel, 2019; Restrepo y Marín, 2012; CEPAL, 2007).

En la literatura empírica se detallan diferentes técnicas de imputación para tratar los valores perdidos en encuestas y otras operaciones estadísticas. En el pasado, se han utilizado estrategias poco recomendadas como son la eliminación de los casos con valores perdidos (*listwise o case deletion*[7]*, pairwise deletion*[8] y ajuste de ponderadores[9]), la cual puede resultar en la perdida de información relevante y provocar resultados sesgados, inferencias estadísticas ineficientes o conclusiones erróneas (Lin y Tsai, 2019; Raja y Thangavel, 2019; Aydilek y Arslan, 201; Baraldi y Enders, 2019; Donza, 2013; CEPAL, 2007). En contraste, se han desarrollado técnicas robustas de imputación estadística[10], tales como imputación por la media/mediana/moda[11], media condicional de datos agrupados[12], *hot deck*[13], *cold deck*[14], regresión[15], máxima verosimilitud[16] o imputación

[6] Corresponde a la diferencia de las estimaciones entre valores muestrales y poblacionales (Arce, Cárdenas, Canales y Lehmann, 2019).
[7] Eliminación o descarte de los datos faltantes conocido como análisis con datos completos (Baraldi y Enders; 2019, Donza; 2013; CEPAL, 2007).
[8] Consiste en tomar para el análisis todos los datos de que se dispone para cada variable analizada; mientras que los casos incompletos se eliminan análisis por análisis, de tal forma que cualquier caso dado puede contribuir a algunos análisis pero no a otros (Baraldi y Enders;2019, Donza; 2013; CEPAL,2007).
[9] Corresponde a la generación de nuevos factores de ponderación que suplan la no respuesta (Baraldi y Enders; 2019, Donza; 2013; CEPAL, 2007).
[10] Las técnicas estadísticas de imputación se basan en el reemplazo de los valores faltantes a partir de la estimación de modelos de regresión o el uso de estadísticos (media, moda o mediana) (Donza, 2013; CEPAL, 2007).
[11] Se basa en el reemplazo de los valores faltantes de una variable por la media/mediana/moda de las unidades observadas en la variable (Lin y Tsai, 2019; Donza, 2013; CEPAL, 2007).
[12] Forma categorías a partir de covariables que se encuentran correlacionadas con la variable de interés y posteriormente imputa los datos faltantes con observaciones que corresponden a las submuestrras que tienen características comunes (Donza, 2013; CEPAL, 2007).
[13] *Hot deck*: es una técnica de duplicación de valores conocidos, para lo cual se asigna al registro que no posee valor el dato correspondiente a un registro con valor conocido (Donza, 2013; CEPAL, 2007).
[14] *Cold deck*: es una técnica que asigna valores a los datos faltantes en función de información conocida de otras fuentes de información (Donza, 2013; CEPAL, 2007).
[15] Consiste en la aplicación de modelos de regresión para imputar la información en la variable Y, a partir de covariables Xi correlacionadas con la variable dependiente (Donza, 2013; CEPAL, 2007).
[16] Comprende la estimación de los parámetros del modelo con los datos completos con la función de máxima verosimilitud, para luego usar los parámetros estimados para la predicción de los valores perdidos y consecuentemente completarlos (CEPAL, 2007).

múltiple[17](Lin y Tsai, 2019 Donza, 2013; CEPAL, 2007) y técnicas de aprendizaje automático o *machine learning* - ML[18] como son *neural networks* (Hasan et al., 2021; Zhu et al., 2011; Richman et al., 2009); *clustering* y decision trees (Hasan et al., 2021; Loh et al., 2019); *k nearest neighbors* - KNN (Hasan et al., 2021; Zhang, 2012); MLP(Hasan et al., 2021; Rosati, 2021; Jerez et al., 2010); *RF* (Rosati, 2021; Jerez et al., 2010); y SVM (Tlamelo et al., 2021; Yang y Shami, 2020).

Dentro de las técnicas de imputación de valores perdidos en encuestas empleadas por los institutos de estadística (IES) se tiene el caso de Estados Unidos, donde el *Bureau of The Census*, *Bureau of Labor Statistics* y el *Bureau of Economic Analysis*, utilizan metodologías como las técnicas de ponderación (modificación del peso de la probabilidad de un determinado individuo al ajustar las faltas de respuesta) e imputación (reemplazo de valor perdido específico por un valor numérico)[19] para ajustar la no respuesta (Eltinge, Kozlow y Luery, s.f). Durante la pandemia del COVID-19, el Reino Unido implementó la técnica de imputación de vecino más cercano para las variables de empleo, desempleo, actividad económica y horas trabajadas de la encuesta de fuerza laboral – *Labour Force Survey*- (Office for National Statistics, 2021). En la encuesta de ingresos – *Canadian Income Survey* – 2020 (CIS) de Canadá, los datos faltantes de las variables relacionadas con ingresos, trabajo, asistencia escolar, seguridad alimentaria, vivienda y costos de servicios públicos se imputaron a través del enfoque de vecino más cercano y *cold deck* (contempla el Censo de 2021 como datos donantes) (Canada Statistics, 2020). En la encuesta de gastos del hogar – *Household Expenditure Survey* de Australia, la imputación de la no respuesta se realizó a partir de la metodología *hot deck*, en la que se reemplaza la información faltante por la de individuos con características similares (región, sexo, edad, estatus laboral e ingreso) que respondieron completamente a la encuesta (conocidos como donantes) (Australian Bureau of Statistics, 2017).

A nivel regional, se han realizado diversos ejercicios de imputación en encuestas. En Argentina, Rosati (2021) utilizó técnicas de RF, XGBoost y MLP para imputar la variable de ingreso de la Encuesta Permanente de Hogares entre 2004 y 2016, mientras que Donza (2013) aplicó el método estadístico de máxima verosimilitud para imputar la no respuesta en las preguntas de ingreso laboral de la Encuesta Permanente de Hogares del Gran Buenos Aires. En Colombia, Dane (2020) aplicó el algoritmo de RF para imputar la condición de informalidad[20]de los ocupados en la Gran Encuesta Integrada de Hogares (GEIH)[21]durante marzo y abril de 2020, lo que permitió estimar la tasa de

[17] El método de imputación múltiple se compone de tres etapas, en la primera cada valor perdido es reemplazado por un conjunto m>1 valores generados por simulación y consecuentemente se generan m conjuntos de datos completos; la segunda etapa consiste en aplicar a cada uno de los conjuntos de datos m el método de análisis deseado; y, por último los resultados obtenidos son combinados para producir una estimación global (Donza, 2013; CEPAL, 2007).

[18] Las técnicas de machine learning consisten en crear un modelo predictivo para estimar valores que sustituirán los valores perdidos en la base de datos de análisis (Jerez et al.,2010).

[19] Imputación determinística: el valor sustituido es obtenido a partir de un proceso determinístico (media de la muestra o valor pronosticado a partir de una regresión) (Eltinge et al., s.f).
Imputación estocástica: el valor de reemplazo es seleccionado al azar de las observaciones disponibles de una celda específica o cluster (Eltinge et al., s.f).

[20] Con la finalidad de manejar estadísticas oficiales, se contemplan como ocupados informales a aquellos individuos que durante el periodo de referencia se encontraron en alguna de las siguientes situaciones: empleados particulares y obreros que laboran en establecimientos, negocios o empresas de hasta cinco empleados en todas sus agencias o sucursales; trabajadores familiares sin remuneración en empresas de cinco trabajadores o menos; trabajadores sin remuneración en empresas o negocios de otros hogares; empleados domésticos en empresas de cinco trabajadores o menos; jornaleros o peones en empresas de cinco trabajadores o menos; trabajadores por cuenta propia que laboran en establecimientos hasta cinco personas, a excepción de los profesionales independientes; patronos o empleadores en empresas de cinco trabajadores o menos; se excluyen obreros o empleados del gobierno (DANE, 2020).

[21] Como fuente adicional se utiliza la planilla integrada de liquidación de aportes (PILA) para la determinación de la probabilidad de que un ocupado realice sus actividades productivas en el sector informal (DANE, 2020).

informalidad para ese periodo. Restrepo y Marín (2012) evaluaron diferentes metodologías de imputación (media no condicionada, regresión estocástica, *hot deck* con y sin regresión, imputación múltiple normal multivariada e imputación múltiple con ecuaciones encadenadas) para la variable de ingresos en la GEIH para el 2010, obteniendo resultados similares para cada modelo. En Chile, Arce et al. (2019) aplicaron diferentes técnicas de imputación en la VIII Encuesta de Presupuestos Familiares (EPF) para gastos diarios (las técnicas de imputación fueron factor de no respuesta[22], ajuste por peso diario, media condicionada y *hot deck*) e ingresos asociados al trabajo principal y jubilaciones (se usaron métodos estadísticos de imputación por regresión de Heckman[23], *hot deck*, media condicional e imputación múltiple). Alfaro y Fuezalida (2009) realizan la imputación del ingreso laboral, activos financieros y deudas de los entrevistados en la Encuesta de Protección Social (EPS) 2004, para lo cual aplican la metodología de imputación múltiple a partir de información únicamente para jefes de hogar. Finalmente, en México, Rodríguez y López (2015) imputan las observaciones con ingresos faltantes de la Encuesta Nacional de Ocupaciones y Empleo (ENOE) a partir de la metodología *hot deck*.

1.2. Mecanismos para el manejo de datos perdidos

Los datos perdidos y la no respuesta en encuestas pueden ocurrir, según la teoría de los datos faltantes de Rubin (1976), de acuerdo a tres mecanismos teóricos: *missing completely at random*, *missing at random* y *not missing at random*.

Para definir la falta de datos, Thamelo et al. (2021) define a Y como un conjunto de datos, la cual se descompone en Y_0 (valores observados)[24] y Y_m (valores perdidos)[25]; y, se define una matriz S de valores perdidos en la que:

$$S:=\begin{cases} 0, si\ Y\ es\ observado \\ 1, si\ Y\ es\ perdido \end{cases}$$

También se asume "q" como un vector de valores que indica la relación entre la omisión en S y el conjunto de datos Y. En ese sentido, los mecanismos de los valores perdidos son definidos de acuerdo a la probabilidad de si un valor es observado o perdido, como se señala a continuación:

***Missing Completely at Random* (MCAR):** los valores perdidos son independientes de las medidas observados y no observadas. La probabilidad MCAR se define así:

$$p(S \mid q).$$

***Missing at Random* (MAR)**: los valores perdidos solo se relacionan con los valores observados. La probabilidad MAR se define como:

$$p(S \mid Y_0, q).$$

[22] Consiste en reponderar la información de gastos existente, lo cual solventa el problema de perdidos.
[23] Completa los datos perdidos a partir de una regresión en dos etapas de los datos completos (Arce el al., 2019).
[24] Se refiere al resto de variables del conjunto de datos.
[25] Se refiere a la propia variable analizada con valores perdidos.

***Not Missing at Random* (MNAR):** los valores perdidos dependen de igual forma de los valores perdidos como de los observados. La probabilidad MNAR se define como:

$$p(S \mid Y_0, Y_m, q).$$

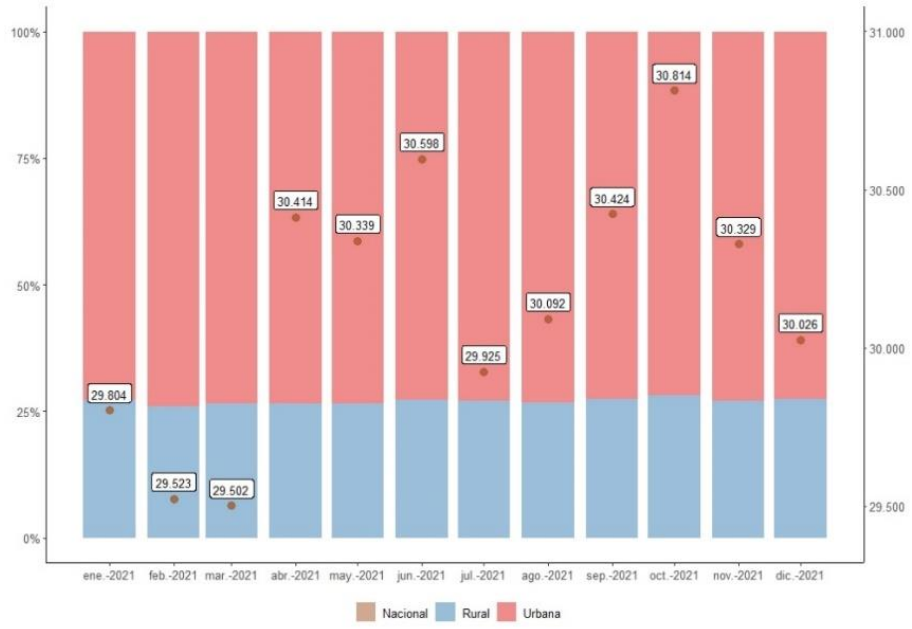
2. Datos y metodología

2.1. Datos

Esta investigación tiene como objetivo el imputar información en preguntas de la encuesta de Empleo, Desempleo y Subempleo (ENEMDU) anual del 2021 del Ecuador, que reflejan valores perdidos o incompletos. La ENEMDU es una encuesta de hogares recogida mensualmente que se dirige a individuos de 5 años y más de edad, y se realiza en aproximadamente 9.000 viviendas, lo que equivale a alrededor de 30.000 individuos en todo el territorio nacional. El diseño muestral[26] de la encuesta es de tipo probabilístico bietápico de elementos, con estratificación geográfica por dominios de estudio y con representatividad nacional, urbano y rural. Además, la encuesta utiliza el factor de expansión como ponderador para llevar las cifras muestrales a poblacionales, según el INEC (2021).

Asimismo, el INEC cuenta con la encuesta anual, la cual se compone de las encuestas mensuales levantadas durante el año y abarca alrededor de 108.192 viviendas (361.790 individuos). Esta encuesta es representativa a nivel nacional, urbano, rural, cinco ciudades autorepresentadas (Quito, Guayaquil, Cuenca, Machala y Ambato) y 24 provincias del país, como se indica en la Figura 1 que muestra el comportamiento de las observaciones en el periodo de análisis (INEC, 2022).

Figura 1. Comportamiento de registros ENEMDU (Datos muestrales)



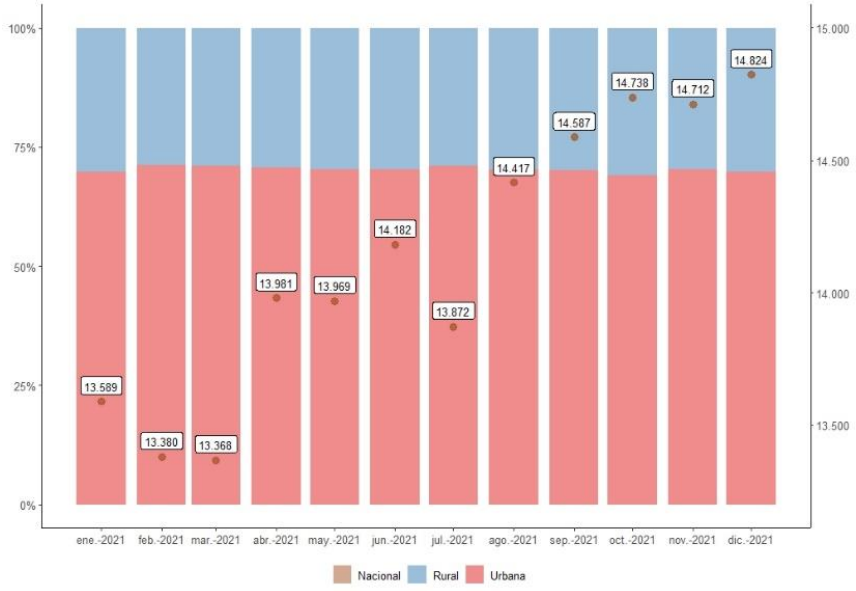
Fuente: ENEMDU 2021

[26] La selección de la muestra es realizada de forma aleatoria, en dos etapas, la primera basada en la selección de Unidades Primarias de Muestreo (UPM) por estrato socioeconómico (alto, medio, bajo) y la segunda enfocada en la selección de viviendas ocupadas dentro de las UPM seleccionadas en la primera etapa (INEC, 2021).

La encuesta contempla un esquema de levantamiento de información a través de un formulario dividido en varias secciones, y que a su vez sigue un esquema de flujos controlado[27], lo que permite el direccionamiento de los informantes de acuerdo a sus características y situación laboral. En ese sentido, al levantarse cómputos de las respuestas obtenidas en las distintas preguntas que conforman la encuesta, se observan valores perdidos o *missing* en preguntas abiertas, o dentro de la categoría "no sabe". Esto se asocia con la negativa o desconocimiento de los informantes para proveer información.

Con esto en consideración, se realizará un análisis de la calidad de la información de los empleados (Véase en la Figura 2 la distribución de empleados por periodo) en las preguntas asociadas con registros contables (pregunta 48 – p48), tenencia de RUC del lugar de trabajo (pregunta 49 – p49), afiliación a la seguridad social (pregunta 61b1 – p61b1) e ingresos laborales (ila), las cuales inciden en la obtención de varios indicadores laborales. Si es necesario, se imputará aquellas preguntas que cuentan con la categoría "no sabe" o que reflejan información perdida pese a la aplicación del control de flujo del formulario.

Figura 2. Comportamiento de registros de empleados ENEMDU (Datos muestrales)



Fuente: ENEMDU 2021

2.1.1. Análisis de datos faltantes (perdidos (*missing*) o "no sabe") como potenciales variables a imputar

La selección de las variables a ser imputadas se basa en el análisis de las preguntas que pueden tener potenciales valores perdidos o que tienen dentro de la pregunta la categoría "no sabe". Por tanto, este apartado analiza a las preguntas p48, p49 y p61b1

[27] El cumplimiento de los flujos del formulario es controlado a partir de mallas de validación, implementados en los sistemas de captación (el INEC de Ecuador denomina al sistema de captación de información como Sistema Integrado de Producción Estadística - SIPE).

y la variable de ingreso laboral (ila) de la ENEMDU[28]. Como se mencionó anteriormente, estas variables son fundamentales para calcular varios indicadores laborales.

2.1.1.1. Manejo de registros contables (p48) y RUC del negocio (p49)

En este apartado, se analizan las preguntas p48 y p49, las cuales se relacionan con la tenencia de registros contables y RUC del lugar de trabajo del empleado, respectivamente. La Tabla 1 muestra el número de registros de cada una de las categorías de la pregunta p48 de la encuesta para los meses que conforman el 2021. Es posible observar que existen 9.223 registros (6,7 % en promedio mensual) que responden a la pregunta como "no sabe", lo que la convierte en candidata para ser imputada.

Tabla 1. Distribución y tasa de no respuesta - no sabe- de la pregunta 48 (Datos muestrales)

p48 (¿El estableciemiento o lugar donde trababaja lleva:...?		ene-21	feb-21	mar-21	abr-21	may-21	jun-21	jul-21	ago-21	sep-21	oct-21	nov-21	dic-21
Frecuencia	Registros contables	1.451	1.391	1.365	1.325	1.288	1.242	1.271	1.299	1.357	1.295	1.372	1.333
	Cuaderno de cuentas	1.835	1.847	1.997	2.062	1.891	2.079	1.972	2.209	2.094	2.413	2.458	2.368
	No lleva contabilidad	6.693	6.565	6.602	6.983	7.275	7.204	6.987	7.217	7.465	7.561	7.280	7.545
	No sabe	848	906	750	804	829	876	828	840	725	581	738	568
Porcentaje	Registros contables	13,0%	12,6%	12,4%	11,5%	11,1%	10,6%	11,2%	10,9%	11,4%	10,7%	11,3%	11,0%
	Cuaderno de cuentas	16,5%	16,8%	18,2%	17,9%	16,3%	17,8%	17,3%	18,6%	17,5%	19,9%	20,2%	19,5%
	No lleva contabilidad	60,2%	59,7%	60,1%	60,7%	62,8%	61,7%	61,5%	60,8%	62,5%	62,3%	59,8%	62,2%
	No sabe	7,6%	8,2%	6,8%	7,0%	7,2%	7,5%	7,3%	7,1%	6,1%	4,8%	6,1%	4,7%

Nota: La muestra corresponde a los individuos empleados
Fuente: ENEMDU 2021

En la Tabla 2 se presenta el número de observaciones correspondientes a cada categoría de la pregunta p49. Se observa que un total de 8.153 observaciones (5,9% en promedio mensual) indicaron "no sabe" como respuesta. Es importante recuperar la información clasificada en la categoría "no sabe" debido a su relevancia en la generación de indicadores de sectorización laboral.

Tabla 2. Distribución y tasa de no respuesta -no sabe- de la pregunta 49 (Datos muestrales)

p49 (¿El estableciemiento o lugar donde trababaja tiene RUC?		ene-21	feb-21	mar-21	abr-21	may-21	jun-21	jul-21	ago-21	sep-21	oct-21	nov-21	dic-21
Frecuencia	Si	4.310	4.325	4.323	4.369	4.422	4.506	4.360	4.590	4.761	4.781	4.948	4.947
	No	5.792	5.553	5.739	6.111	6.101	6.154	5.945	6.206	6.265	6.542	6.269	6.412
	No sabe	725	831	652	694	760	741	753	769	615	527	631	455
Porcentaje	Si	38,8%	39,3%	39,4%	38,0%	38,2%	38,6%	38,4%	38,7%	39,9%	39,4%	40,7%	40,8%
	No	52,1%	50,5%	52,3%	53,1%	52,6%	52,7%	52,3%	52,3%	52,5%	53,9%	51,5%	52,9%
	No sabe	6,5%	7,6%	5,9%	6,0%	6,6%	6,3%	6,6%	6,5%	5,2%	4,3%	5,2%	3,8%

Nota: La muestra corresponde a los individuos empleados
Fuente: ENEMDU 2021

2.1.1.2. Afiliación a la seguridad social (pregunta 61b1)

Otra de las preguntas en las que existe la categoría "no sabe" y que podría ser imputada es la pregunta p61b1, la cual hace referencia a la afiliación a la seguridad social. Sin embargo, como se observa en la Tabla 3, el número de observaciones con respuesta

[28] El formulario de la encuesta ENEMDU se encuentra disponible en el siguiente link: https://www.ecuadorencifras.gob.ec/documentos/web-inec/EMPLEO/2023/Febrero/202302_Formulario_ENEMDU.pdf

en el campo "no sabe" es un total de 46 (0,0% en promedio mensual), por lo que resulta infructuoso incluir esta variable dentro del ejercicio de imputación.

Tabla 3. Distribución y tasa de no respuesta -no sabe- de la pregunta 61b1 (Datos muestrales)

p61b1 (A cúal seguro aporta actualmente?)		ene-21	feb-21	mar-21	abr-21	may-21	jun-21	jul-21	ago-21	sep-21	oct-21	nov-21	dic-21
Frecuencia	IESS general	3.865	3.788	3.786	3.919	3.941	3.951	3.800	4.073	4.164	3.834	4.004	4.005
	IESS voluntario	614	666	670	681	640	710	673	722	758	765	764	776
	Seguro campesino	572	516	536	584	559	536	576	529	510	577	515	525
	Seguro del issfa o isspol	126	109	110	134	111	95	154	93	93	135	96	86
	No aporta	17.510	17.480	17.468	17.898	18.010	17.988	17.742	17.732	17.775	18.348	18.057	17.790
	No sabe	2	3	2	1	1	1	4	3	3	2	10	14
Porcentaje	IESS general	17,0%	16,8%	16,8%	16,9%	16,9%	17,0%	16,6%	17,6%	17,9%	16,2%	17,1%	17,3%
	IESS voluntario	2,7%	3,0%	3,0%	2,9%	2,8%	3,0%	2,9%	3,1%	3,3%	3,2%	3,3%	3,3%
	Seguro campesino	2,5%	2,3%	2,4%	2,5%	2,4%	2,3%	2,5%	2,3%	2,2%	2,4%	2,2%	2,3%
	Seguro del issfa o isspol	0,6%	0,5%	0,5%	0,6%	0,5%	0,4%	0,7%	0,4%	0,4%	0,6%	0,4%	0,4%
	No aporta	77,2%	77,5%	77,4%	77,1%	77,4%	77,3%	77,3%	76,6%	76,3%	77,5%	77,0%	76,7%
	No sabe	0,0%	0,0%	0,0%	0,0%	0,0%	0,0%	0,0%	0,0%	0,0%	0,0%	0,0%	0,1%

Nota: La muestra corresponde a todos los individuos
Fuente: ENEMDU 2021

2.1.1.3. Ingreso laboral (ila)

Por último, en esta investigación se plantea la imputación de los valores *missing* de la variable de ingreso laboral (ila), que es necesaria para la generación de la variable de condición de actividad de los empleados. El análisis descriptivo de la variable se presenta en la Tabla 4, donde se observa que hay 3.782 observaciones (2,6% en promedio mensual) con valores *missing* que pueden ser imputados.

Tabla 4. Distribución de la información del ila oficial (Datos muestrales)

ila (Ingreso laboral)		ene-21	feb-21	mar-21	abr-21	may-21	jun-21	jul-21	ago-21	sep-21	oct-21	nov-21	dic-21
Frecuencia	Valores en ingreso	11.196	11.156	11.144	11.590	11.574	11.822	11.715	12.077	12.363	12.312	12.400	12.540
	Valores perdidos en ingreso	344	365	305	346	361	285	291	303	264	290	378	250
Porcentaje	Valores en ingreso	97,0%	96,8%	97,3%	97,1%	97,0%	97,6%	97,6%	97,6%	97,9%	97,7%	97,0%	98,0%
	Valores perdidos en ingreso	3,0%	3,2%	2,7%	2,9%	3,0%	2,4%	2,4%	2,4%	2,1%	2,3%	3,0%	2,0%

Nota: La muestra corresponde a los individuos empleados y que no se autoidentifican como trabajadores no remunerados en la pregunta p42
Fuente: ENEMDU 2021

No obstante, en la metodología oficial para el cálculo del ingreso laboral (ila), los ingresos negativos[29] o aquellos que presentan inconsistencias[30] en las variables utilizadas para su construcción se consideran como valores *missing*. Por lo tanto, se propone generar una nueva variable del ila (denominada new_ila) que permita mejorar la precisión de la predicción, al construirse bajo las mismas condiciones que la variable original, pero sin asignar como *missing* a los valores negativos.

[29] Este particular ocurre principalmente en los trabajadores independientes (patronos y cuenta propia), quienes por el giro de negocio pueden obtener pérdidas en el periodo de referencia. Específicamente, un trabajador independiente podría reportar gastos mayores a sus ingresos brutos, pero esto no significa que su ingreso final del periodo sea negativo; es posible que haga uso de sus ahorros o adquiera una deuda. En ENEMDU el ingreso negativo resulta de la diferencia entre los ingresos brutos (p63 y p64) y los gastos de funcionamiento del negocio (p65) (Castillo y Puebla, 2016).

[30] De acuerdo con Castillo y Puebla (2016) este particular ocurre bajo cuatro condiciones: i) En alguna de las preguntas que conforma el ingreso laboral existen códigos 9´s a tres, cuatro o cinco dígitos; ii) Cuando un trabajador dependiente en su actividad principal, no informa o no conoce sus ingresos monetarios (p66) o, si un trabajador independiente en su actividad principal, no informa o no conoce sus ingresos brutos; iii) cuando un trabajador reporta ingresos en su actividad principal, manifiesta contar con una segunda actividad, pero no informa o no conoce sobre los mismos y no reporta ingresos no laborales; iv) cuando un trabajador reporta ingresos en su actividad principal, manifiesta tener una segunda actividad, pero no informa o no conoce acerca de los mismos, pero sí reporta ingresos no laborales.

La nueva variable del ila (new_ila) se compone de la variable ila original, y de las variables de ingreso de la ocupación principal (ila 1) e ingreso de la ocupación secundaria (ila 2). Para su construcción, se contemplan los siguientes aspectos: i) Se agrega ila 1 e ila 2, para generar la variable new_ila; ii) Si ila 2 es *missing*, se toma el ingreso del ila 1 para new_ila; iii) Si ila 1 es igual a cero, ila 2 es *missing* y la variable original del ila es *missing* (lo que indica un ingreso incoherente según la construcción original de la variable), entonces new_ila se coloca como *missing*.

En la Tabla 5 se muestra la distribución de la variable new_ila y los valores potenciales a imputar (valores *missing*). Se puede observar que la variable presenta 1.681 observaciones con valores *missing* (1,2% en promedio mensual), lo que significa 2.101 observaciones menos que el ila bajo la metodología oficial (ila original). Esto se debe principalmente a la modificación del manejo de los ingresos negativos en la metodología de construcción de la variable new_ila, y que en la variable ila original se encontraban como *missing*.

La recuperación de la información que se encuentra como *missing* permitirá afinar la clasificación de la condición de actividad de los individuos con empleo, especialmente aquellos categorizados inicialmente como no clasificados o empleo inadecuado debido la falta de información del ila.

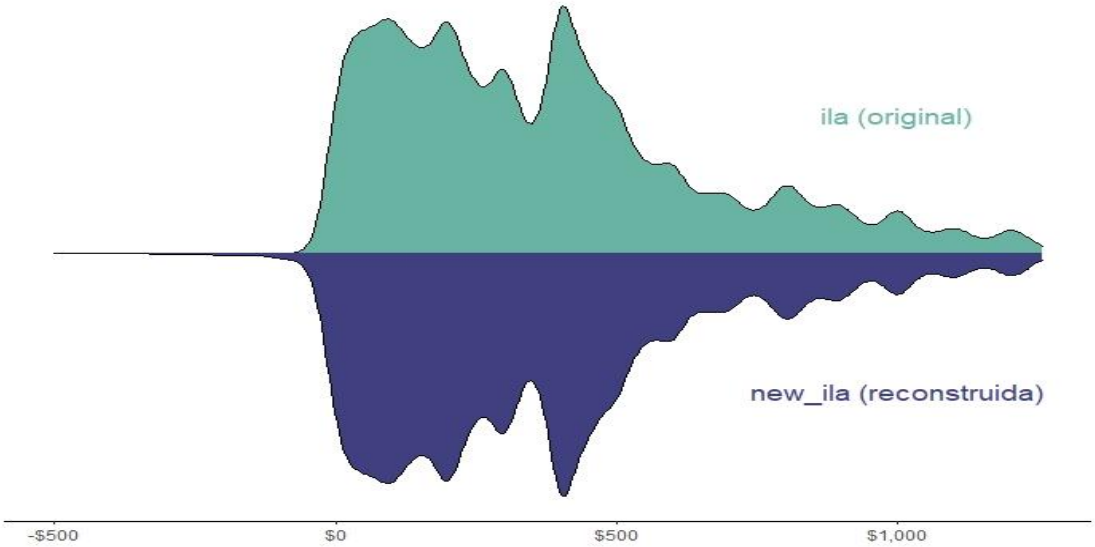
Tabla 5. Distribución de la información d la nueva variable ila (Datos muestrales)

ila (Ingreso laboral)		ene-21	feb-21	mar-21	abr-21	may-21	jun-21	jul-21	ago-21	sep-21	oct-21	nov-21	dic-21
Frecuencia	Valores en ingreso	11.380	11.330	11.305	11.761	11.750	11.988	11.882	12.251	12.516	12.494	12.630	12.703
	Valores perdidos en ingreso	160	191	144	175	185	119	124	129	111	108	148	87
Porcentaje	Valores en ingreso	98,6%	98,3%	98,7%	98,5%	98,4%	99,0%	99,0%	99,0%	99,1%	99,1%	98,8%	99,3%
	Valores perdidos en ingreso	1,4%	1,7%	1,3%	1,5%	1,6%	1,0%	1,0%	1,0%	0,9%	0,9%	1,2%	0,7%

Nota: La muestra corresponde a los individuos empleados y que no se autoidentifican como trabajadores no remunerados en la pregunta p42
Fuente: ENEMDU 2021

Adicionalmente, complementando el análisis anterior, la Figura 3 muestra la distribución de las variables ila_original y new_ila reconstruida, y se observa la existencia de una diferencia mínima en su distribución, la cual se debe específicamente a los ingresos laborales negativos como ya se mencionó anteriormente.

Figura 3. Distribución del agregado del Ingreso laboral (ila original vs. new_ila)



Notas:
1) Se eliminan datos atípicos (outliers) para una mejor visualización de la distribución.
2) Solo se consideran valores no perdidos.
Fuente: Encuesta Nacional de Empleo, Desempleo y Subempleo 2021

Por último, con la finalidad de tener una lógica común en la generación de los modelos de ML (clasificación), se convierte la variable new_ila en categórica. Para lo cual se propusieron tres categorías: 1.Ingreso mayor o igual a SBU; 2.Ingreso menor o igual al SBU; y 3.Ingresos negativos. La Tabla 6 muestra la distribución de la variable new_ila en cada una de estas categorías, así como las observaciones a ser imputadas.

Tabla 6. Distribución de la información de la variable new_ila categórica (Datos muestrales)

Ila (Ingreso laboral) reconstruido categórico		ene-21	feb-21	mar-21	abr-21	may-21	jun-21	jul-21	ago-21	sep-21	oct-21	nov-21	dic-21
Frecuencia	Ingreso mayor o igual a SBU	5.506	5.302	5.254	5.302	5.339	5.454	5.574	5.804	5.942	5.814	6.026	6.223
	Ingreso menor a SBU	5.690	5.854	5.890	6.288	6.235	6.368	6.141	6.273	6.421	6.498	6.374	6.317
	Ingreso negativo	184	174	161	171	176	166	167	174	153	182	230	163
	Valores perdidos en ingreso	160	191	144	175	185	119	124	129	111	108	148	87
Procentaje	Ingreso mayor o igual a SBU	47,7%	46,0%	45,9%	44,4%	44,7%	45,0%	46,4%	46,9%	47,1%	46,1%	47,2%	48,7%
	Ingreso menor a SBU	49,3%	50,8%	51,4%	52,7%	52,2%	52,6%	51,1%	50,7%	50,9%	51,6%	49,9%	49,4%
	Ingreso negativo	1,6%	1,5%	1,4%	1,4%	1,5%	1,4%	1,4%	1,4%	1,2%	1,4%	1,8%	1,3%
	Valores perdidos en ingreso	1,4%	1,7%	1,3%	1,5%	1,6%	1,0%	1,0%	1,0%	0,9%	0,9%	1,2%	0,7%

Nota: La muestra corresponde a los individuos empleados y que no se autoidentifican como trabajadores no remunerados en la pregunta p42
Fuente: ENEMDU 2021

2.1.2. Construcción de indicadores laborales

Una vez han sido analizadas las preguntas que serán imputadas en este ejercicio (p49 y new_ila), este apartado tiene como un objetivo servir como preámbulo metodológico al destacar las principales consideraciones técnicas para la construcción de los indicadores laborales (sectorización laboral y de clasificación de los empleados de acuerdo a su condición de actividad), que se derivan de las variables a ser imputadas.

2.1.2.1. Sectorización del empleo[31]

La sectorización se refiere a la agrupación de unidades de producción similares, que comparten características comunes en términos de funciones y comportamientos (Molina et al., 2015). En el contexto del empleo, la sectorización se basa en dos aspectos principales: i) el número de personas que trabajan en el lugar de trabajo (pregunta 47a - p47a)[32]; y ii) la tenencia de RUC del lugar de trabajo del empleado (p49).

Según Molina et al. (2015), los empleados se clasifican en cuatro categorías según su sector: formal, que incluye a las personas que trabajan en establecimientos con 100 o más empleados (pregunta 47a); informal, que se refiere a las personas que trabajan en unidades productivas con menos de 100 empleados (p47a) y que no tienen RUC (p49); empleo doméstico, que abarca a las personas cuya ocupación es la de empleado/a doméstico/a; y, no clasificados, que comprende a las personas que no proporcionan información sobre el RUC de la empresa en la que trabajan (p49).

2.1.2.2. Condición de actividad[33]

La metodología para clasificar a la población empleada según su condición de actividad se basa en la identificación de factores relevantes como el ingreso laboral (ila), el tiempo de trabajo[34] y el deseo disponibilidad de trabajar más horas[35](Castillo, 2015).

Castillo (2015) propone una clasificación de los ocupados por su condición de actividad en tres grupos mutuamente excluyentes[36]: empleo adecuado, empleo inadecuado y empleo no clasificado. Dentro del empleo inadecuado, se incluyen tres subcategorías: subempleo, otro empleo inadecuado y empleo inadecuado no remunerado.

Para clasificar el empleo adecuado, se toman en cuenta el ingreso laboral o beneficio mensual y la jornada laboral, estableciendo que un empleado es considerado como adecuado si su ingreso y el tiempo de trabajo cumplen con la normativa vigente[37], independientemente de su deseo o disponibilidad para trabajar horas adicionales.

El empleo inadecuado mientras tanto considera a los trabajadores con deficiencias en términos de ingresos laborales y horas de trabajo. Dentro de esta categoría, las

[31] Para mayor detalle de la construcción de la sectorización del empleo diríjase al siguiente link: https://www.ecuadorencifras.gob.ec/documentos/web-inec/Estudios%20e%20Investigaciones/Trabajo_empleo/4.%20REM-Actualizacion_metodologica_empleo_informal.pdf
[32] La pregunta 47 en el formulario menciona lo siguiente: ¿Cuántas personas trabajan usualmente en la empresa o negocio?; la pregunta 47a corresponde al valor numérico de la respuesta.
[33] Para mayor detalle de la construcción de la condición de actividad de las personas empleadas diríjase al siguiente link: https://www.ecuadorencifras.gob.ec/wp-content/uploads/downloads/2015/02/Empleo-y-condici%C3%B3n-de-actividad-en-Ecuador.pdf
[34] En Ecuador, de acuerdo con el Código del Trabajo, el umbral laboral semanal es de 40 horas (Castillo, 2015).
[35] De acuerdo al diseño de la encuesta ENEMDU, para reflejar el deseo de trabajar horas adicionales se incluye el factor de disponibilidad de trabajar horas adicionales (Castillo, 2015).
[36] La clasificación en empleo adecuado y empleo inadecuado se encuentran en concordancia con los derechos y garantías de los trabajadores estipulados en la Constitución de la República del Ecuador y del Código del Trabajo (Castillo, 2015).
[37] La remuneración o beneficio mensual percibido por el trabajador es contemplado adecuado si cubre con las necesidades básicas del trabajador y de su familia. Mientras que una jornada laboral es adecuada, si el trabajador destina al menos el tiempo fijado en un umbral legal (Castillo, 2015).

subcategorías (subempleo[38], otro empleo inadecuado[39] y empleo inadecuado no remunerado[40]) se determinan según el deseo y la disponibilidad del trabajador para trabajar horas adicionales, así como la percepción de remuneraciones y beneficios. Por último, la categoría de empleo no clasificado incluye a los trabajadores que no proporcionan información completa sobre ingresos laborales y horas de trabajo, lo que no permite una clasificación clara en ninguna de las categorías anteriores.

2.2. Metodología

2.2.1. Técnicas de machine learning usadas para imputación en encuestas de hogares

Los métodos de imputación que se utilizarán en este estudio se basan en técnicas de *machine learning* como son *ensemble learning* (RF y XGBoost)[41], *neural networks*[42] (MLP) y métodos supervisados simples[43](SVM). A continuación, se detalla cada uno de ellos:

RF: es un método de ensamble tipo *bagging*[44] el cual genera variabilidad a nivel de unidades de análisis y de atributos; para cada iteración no se utiliza el total de predictores, sino una muestra de dicho conjunto, es decir, en lugar de dividir el espacio en funcion del total de predictores, para cada árbol se utiliza su submuestra. El algoritmo de construcción de cada árbol se basa en la definición del total de árboles o iteraciones a entrenar; para luego determinar el total de casos de entrenamiento (n) y número de variables clasificadoras (K), extraer una muestra *bootstrap*[45] de los casos de entrenamiento[46], realizar la predicción y repetir el proceso de forma iterativa a través de todos los árboles definidos en el ensamble[47] (Rosati, 2021).

XGBoost: es una técnica de árboles de decisión de tipo *boosting*[48] optimizada a partir del algoritmo *Gradient Boosting*, la cual consiste en el entrenamiento de los errores de

[38] El subempleo es entendido como una situación inadecuada de empleo, donde el trabajador denota su deseo y disponibilidad de trabajar horas adicionales. Así, un trabajador puede ser subempleo por horas, por ingresos o por ambas simultáneamente. Dentro del subempleo se clasifica al subempleo por insuficiencia de tiempo de trabajo, el cual engloba a las personas en condición de insuficiencia de horas y de ingresos y horas, y que desean y se encuentran disponibles para trabajar más horas; mientras que los individuos que presentan insuficiencia de ingresos y tienen deseo y disponibilidad de trabajar horas adicionales, se clasificarán como subempleo por insuficiencia de ingresos. Aquellos trabajadores con ingresos inferiores al salario básico unificado y sin deseo y disponibilidad de trabajar más horas adicionales se incluirán dentro de la categoría de otro empleo inadecuado (Castillo, 2015).

[39] Los trabajadores que por la naturaleza del giro de su negocio presenten ingresos negativos o pérdidas, son clasificados como empleados inadecuados, ya que no cuentan con un ingreso adecuado. En estricto sentido, una parte de estos trabajadores se destinará al subempleo, si presenta el deseo y la disponibilidad de trabajar horas adicionales, caso contrario se ubicará en la categoría de otro empleo inadecuado (Castillo, 2015).

[40] Las personas que no reciben ningún tipo de compensación, al menos no monetaria por su trabajo, son clasificados como trabajadores inadecuados no remunerados; dentro de este grupo se encuentran pequeñas unidades familiares rurales dedicadas a la agricultura (Castillo, 2015).

[41] *Ensemble learning* es conocida como ensamble de modelos o sistemas de clasificadores múltiples, es una técnica que tiende a incrementar la capacidad predictiva de los clasificadores base (regresiones, *decisión trees o neural networks*) a partir de submuestras de los datos originales y la estimación para cada submuestra de un modelo (Rosati, 2021).

[42] *Neural Networks* son un tipo de sistema de modelado predictivo a partir del ajuste iterativo de parámetros (Kaiser J. , 2014).

[43] Los modelos supervisados se basan en la generación de información a partir de los datos disponibles con casos objetivo (etiquetados), de tal manera que el algoritmo pueda ser utilizado para predecir nuevos casos (no etiquetados) (Berry et al., 2020).

[44] *Bagging* es una técnica combinada, en la que cada ensamble o clasificación es entrenado usando diferentes muestras de entrenamiento a partir de la muestra original, se seleccionan N elementos uniformemente aleatorios con reemplazos (Tlamelo et al., 2021).

[45] *Bootstrap* es un método que se basa en la extracción de remuestras sucesivas con reposición de una muestra original, este método es usado en la estimación de distribuciones muestrales de distintos estimadores (medidas de tendencia central y dispersión) (Rosati, 2021).

[46] Para cada nodo del árbol se muestrea k predictores de K y se calcula la mejor partición a partir de k variables de la muestra de entrenamiento. Cada árbol es construido por completo evitando algún tipo de poda.

[47] Cada caso es ubicado en la clase que ha sido clasificado la mayor cantidad de veces a lo largo de los árboles generados (votación - *voting*).

[48] *Boosting* es una técnica de ensamble iterativo basada en remuestreo, en la que el algoritmo corrige los errores cometidos en las iteraciones previas (Rosati, 2021).

clasificación a partir de un nuevo árbol de decisión (Ergul y Kamisli, 2021). Rosati (2021) manifiesta que operativamente el algoritmo se basa en que, para cada iteración, se debe entrenar un modelo base (*decision tree*), calcular los errores del modelo, entrenar un nuevo modelo sobre los errores del modelo, agregar el modelo al ensamble, y, por último, agregar los resultados.

MLP: esta técnica es conocida como *Feed Forward Neural Network* y se basa en múltiples unidades computacionales interconectadas, en donde cada capa se encuentra conectada con las neuronas de la siguiente capa (Jeres et al., 2010). Concretamente, el método plantea que un MLP se compone de una capa de *input*, otra de *output* y dos o más capas de cómputo conocidas como unidades de umbral lineal; el proceso consiste en que cada unidad de umbral lineal toma como input un vector y realiza transformaciones lineales y no lineales sobre el mismo, para que luego el valor resultante (*ouput*) de la transformación sea enviado a la siguiente capa. Este proceso se desarrolla secuencialmente, hasta que se tiene una última capa con el *output* deseado (Rosati, 2021).

SVM: es un método supervisado simple que identifica puntos de datos desde un espacio de baja dimensión a un espacio de alta dimensión para hacerlos linealmente separables y luego generar un hiperplano como límite de clasificación para dividir los puntos de datos (Yang y Shami, 2020; Stewart et al., s.f). Operativamente, la técnica identifica para una muestra de entrenamiento un hiperplano de separación óptimo que maximice la distancia desde el mismo hacia los puntos más cercanos (Tlamelo et al., 2021).

2.2.2. Hiperparámetros y métricas de evaluación de las técnicas de *machine learning*

Una vez que se han seleccionado las técnicas de *machine learning* para llevar a cabo la imputación, es crucial definir los parámetros que controlarán el comportamiento, precisión y efectividad de cada algoritmo, así como las métricas de evaluación de la predicción realizada por cada uno.

Existen dos tipos de parámetros en los modelos de *machine learning*: los parámetros y los hiperparámetros[49]. Los parámetros se desarrollan durante el proceso de aprendizaje basado en los datos, como los pesos de las capas en redes neuronales. Por otro lado, los hiperparámetros deben ser definidos previamente a la aplicación de cada modelo como parte de su arquitectura y configuración. Estos hiperparámetros incluyen parámetros de penalización en modelos SVM y tasas de aprendizaje en *neural networks* (Yang y Shami, 2020).

En el caso de los algoritmos de ensamble basados en *decision trees* (RF y XGBoost), los hiperparámetros incluyen el número de árboles a entrenar, la configuración de cada árbol y el grado de aprendizaje de cada uno. Para los algoritmos MLP, los hiperparámetros se refieren al número de capas y neuronas en cada capa (Rosati, 2021). En los modelos SVM, los hiperparámetros relevantes son el tipo de función de

[49]Los hiperparámetros son un número de valores que deben ser establecidos para determinar el comportamiento del algoritmo y controlar el problema de sobreajuste (*overfitting*) de los modelos de ML (Rosati, 2021).
Sobreajuste u *overfitting* ocurre cuando un modelo de ML se encuentra sobreentrenado y no es generalizable; es decir, el modelo no capta verdaderamente a los datos, y los confunde con ruidos y errores aleatorios de los propios datos (Rosati, 2021).

kernel (lineal, radial, polinomial y sigmoidal), que mide la similitud entre dos puntos de los datos (Yang y Shami, 2020).

En cuanto a las métricas de evaluación de la predicción, es importante distinguir su aplicación en problemas de clasificación y regresión en *machine learning*[50]. Los problemas de regresión están relacionados con variables continuas, mientras que los problemas de clasificación involucran variables discretas. Para los problemas de clasificación, se consideran estadísticas derivadas de la matriz de confusión, como la *sensitivity* (VPR), *specificity* (VNR), *false positive ratio* (FPR), *false negative ratio* (FNR), *precision* (PPV), *negative predective value* (NPV), *false discovery rate* (FDR), *false omission rate*(FOR), *Cohen´s Kappa coefficient*, *F1 - score* y *accuracy*, entre las más comunes. En cambio, para los problemas de regresión se suelen utilizar medidas como *mean absolute error* (MAE), *mean absolute percentage error* (MAPE) y *root mean squared error* (RMSE) (Hasan et al., 2021).

2.2.3. Estimación de modelos de *machine learning*

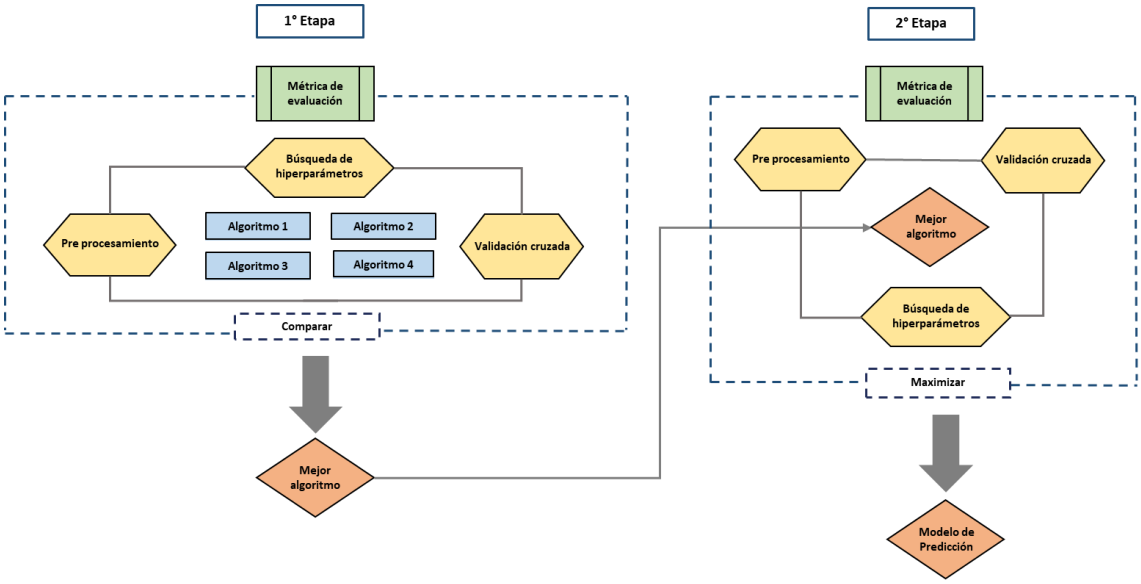
La estimación de modelos con alto poder predictivo a partir de algoritmos de ML es una tarea exhaustiva, en gran parte por lo específico y demandante (computacionalmente) que puede llegar a ser el proceso de selección, preparación, entrenamiento y evaluación de los mismos. La literatura relacionada con técnicas de imputación a través de ML muestra un extenso uso de algoritmos de distinta índole (supervisado, no supervisado, ensamble y *neural networks*) cada uno con sus ventajas, desventajas y requerimientos individuales.

Bajo estos antecedentes, este estudio plantea la imputación de las variables p49 y new_ila[51] de la encuesta ENEMDU, asumiendo por la presencia de flujos de respuesta que el patrón que siguen los datos perdidos son de tipo MAR[52]; es decir, la probabilidad de datos perdidos está relacionado únicamente con variables observadas del conjunto de datos.

En cuanto al proceso de imputación, se busca la obtención de un modelo de predicción adecuado, que sea preciso y cuente con alto poder predictivo, para este fin, se establece un proceso dividido en dos etapas: la primera busca la elección del mejor algoritmo de predicción de un conjunto de algoritmos de ML, mientras la segunda maximiza el poder predictivo del algoritmo elegido, obteniendo el modelo de predicción óptimo a utilizar para imputar definitivamente cada pregunta de interés (la Figura 4 ilustra el proceso descrito).

[50] El Anexo 1 describe las métricas de evaluación de la predicción para problemas de clasificación y regresión.
[51] Dado que la variable p48 no tiene influencia sobre los indicadores relacionados con el mercado laboral, se ha excluido la variable de esta sección. No obstante, los resultados de la imputación de la variable p48 al igual que para la variable new_ila continuo se encuentran en el Anexo 3.
[52] Graham (2009, 2012) expone que, no es posible determinar de forma inequívoca a los datos perdidos dentro de uno de los tres patrones conocidos en la literatura (MCAR, MAR, MNAR), como resultado la elección del patrón de datos perdidos puede ser asumida.

Figura 4. Proceso de imputación



Con respecto a la primera etapa, se toma como referencia cuatro tipos de algoritmos de ML comúnmente utilizados en la literatura (Rosati, 2021; DANE, 2020; Lin y Tsai, 2019), dos algoritmos de tipo ensamble (RF y XGBoost), un algoritmo de *neural networks* (MLP) y un algoritmo de aprendizaje supervisado simple (SVM).

La elección del mejor algoritmo se realiza a partir de la métrica de exactitud de la predicción *accuracy*[53] dado que se trata de un problema de clasificación; además, con la intención de garantizar la comparabilidad entre modelos y minimizar las divergencias técnicas se propone utilizar el método de búsqueda aleatoria con validación cruzada (k[54] = 5) para la definición de hiperparámetros. Posteriormente, se definen las variables que entraran a cada modelo (véase la Tabla 7) en función de los flujos de respuesta establecidos dentro del formulario de la encuesta.

Tabla 7. Variables para modelado por pregunta de interés (1° etapa)

Variables	Descripción	Tipo de variable	p49	New_IIa
area	Área	Categórica	-	x
dominio	Dominio	Categórica	x	x
p02	Sexo	Categórica	-	x
p03	Edad	Numérica	x	x
p15	Como se considera	Categórica	x	-
p20	Trabajó la semana pasada	Categórica	x	-
p24	Horas de trabajo en la semana anterior	Numérica	x	-
p42	Categoría de ocupación	Categórica	x	x
p45	Cuántos años trabaja	Numérica	x	x
p46	Sitio de trabajo	Categórica	x	-
p47b	Número personas trabajan en el establecimiento	Numérica	-	-
p48	Manejo de registros contables en la empresa que trabaja	Categórica	-	-

[53] Una exploración de la literatura muestra que el *accuracy, sensitivity* y *specificity* son las métricas de evaluación usadas más frecuentemente. Estas son empleadas por el 39,1%, 11,0% y 14,1% de la literatura respectivamente (Hasan et al., 2021). La precisión (*accuracy)* define la calidad del modelo a través del contraste entre las predicciones correctas contra el total de predicciones; para conocer su método de cálculo diríjase al Anexo 1.

[54] k o *fold* se refiere a una división de los datos de entrenamiento en subconjuntos disjuntos y no solapados, donde, su valor indica el número de divisiones que serán realizadas sobre el conjunto de datos (Hastie et al., 2009). La elección de un número optimo depende de distintos factores asociados al objetivo de investigación; sin embargo, la literatura recomienda lo valores cinco o diez. (Breiman y Spector,1992; Kohavi,1995)

p49	El establecimiento tiene RUC	Categórica	-	-
p61b1	Afiliación a la seguridad social	Categórica	-	x
niv_ins	Nivel de instrucción	Categórica	x	x
nnivins		Categórica	-	-
grupo1	Grupo de Ocupación CIUO8 (población empleada de 15 años y más de edad)	Categórica	-	x
crama_ciiu4	Rama de actividad ciiu4 a 1 dígito - agrupación ENEMDU	Categórica	x	x
horas	Horas de trabajo semanal	Numérica	-	x
secemp	Sectores de los Empleados (clasificación para >= 15 años de edad)	Categórica	-	x
Total de variables para modelado			**10**	**12**

Nota: La elección de las variables se realizó en base a los flujos de las preguntas a ser imputadas y del conocimiento de la encuesta de los investigadores
Fuente: ENEMDU 2021

Posterior a la selección de variables, se aplica un pre-procesamiento a los algoritmos SVM y MLP, que incluye la recodificación de variables categóricas mediante el método *one hot encoding*[55] y estandarización de variables numéricas[56]. En contraste, los algoritmos RF y XGBoost no requieren de un pre-procesamiento específico, ya que cuentan con uno dentro de su funcionalidad. Los resultados del *accuracy* para cada modelo y variable imputada se presenta en la Tabla 8, a partir de la cual se llega a la conclusión que el mejor modelo a ser implementado es de ensamble XGBoost para ambas variables.

Tabla 8. Resultados obtenidos por cada algoritmo considerado y pregunta de imputación

Accuracy				
Pregunta	***Support Vector Machine***	***XGBoost***	***Random Forest***	***Multilayer Perceptron***
P49	88,4%	**88,7%**	88,1%	88,4%
New_ila	83,2%	**83,8%**	83,3%	83,3%

Una vez se definió el modelo óptimo (XGBoost[57]) para la imputación en la primera etapa, el siguiente paso, es realizar un análisis exploratorio de variables (EDA por sus siglas en inglés) más exhaustivo. Durante este análisis, se realizaron diferentes procedimientos, como la revisión de valores faltantes, la identificación y eliminación de información redundante o poco informativa, entre otros[58]. A diferencia de la selección realizada en la primera etapa, este enfoque consideró no solo los flujos de respuesta, sino también las características de las variables (tipo de variable, existencia absoluta de valores únicos, información recodificada, entre otros). Como resultado, se seleccionaron aquellas variables que presentan una relación importante con la variable que será imputada, al tiempo que proporcionan información completa, generalizada y una mayor variabilidad en los datos. La Tabla 9 muestra las variables que finalmente ingresaron al modelo en la segunda etapa para cada variable de interés.

Tabla 9. Variables para modelado por pregunta de interés (2° etapa)

Variables	**Descripción**	**Tipo de variable**	**p49**	**new_ila**
area	Área	Categórica	x	x

[55] Consiste en la generación de variables dummy por categoría.
[56] El valor reescalado se desprende de la formula $\frac{X-\mu}{\sigma}$ donde μ corresponde a la media y σ a la desviación estándar de la variable.
[57] La generación de los modelos finales en el software estadístico R se realiza a partir del conjunto de librerías "*tidymodels*", esto debido a que ofrecen más opciones tanto para procesamiento de variables como métodos de búsqueda de hiperparámetros, además permiten la revisión y presentación de resultados de forma sencilla.
[58] El Anexo 2 muestra el proceso completo del análisis exploratorio de variables, es decir, evidencia los criterios de eliminación de variables y el detalle las variables resultantes consideradas en los modelos.

p02	Sexo	Categórica	x	x
p03	Edad	Numérica	x	x
p04	Relación de Parentesco	Categórica	x	x
p06	Estado civil	Categórica	x	x
p10a	Nivel de instrucción	Categórica	x	x
p15	Como se considera	Categórica	x	x
p24	Horas de trabajo en la semana anterior	Numérica	x	x
p42	Categoría de ocupación	Categórica	x	x
p45	Cuántos años trabaja	Numérica	x	x
p46	Sitio de trabajo	Categórica	x	x
p47a	Tamaño del establecimiento	Categórica	-	x
p47b	Número personas trabajan en el establecimiento	Numérica	x	-
p48	Manejo de registros contables en la empresa que trabaja	Categórica	x	-
p49	El establecimiento tiene RUC	Categórica	-	-
p50	Número de trabajos	Categórica	x	x
p51a	Horas de trabajo principal	Numérica	x	x
p61b1	Afiliación a la seguridad social	Categórica	x	x
nnivins	Nivel de instrucción	Categórica	x	x
grupo1	Grupo de Ocupación CIUO8 (población ocupada de 15 años y más de edad)	Categórica	x	x
rama_ciiu4	Rama de actividad ciiu4 a 1 dígito (población ocupada de 15 años y más de edad)	Categórica	x	x
prov_n	Provincia	Categórica	x	x
ila	Ingreso laboral	Numérica	x	-
hsize	Tamaño del hogar	Numérica	x	x
horas	Horas de trabajo semanal	Numérica	x	x
escol	Años promedio de escolaridad	Numérica	x	x
secemp	Sectores de los empleados (clasificación para >= 15 años de edad)	Categórica	-	x
Total de variables para modelado			**24**	**23**

Fuente: ENEMDU 2021

Luego de haber seleccionada a la técnica de predicción y las variables para modelado, se inicia la segunda etapa centrada en maximizar el poder predictivo del algoritmo elegido (XGBoost), a través del pre-procesamiento de las variables de entrada y la definición de hiperparámetros[59]. El pre-procesamiento de variables consistió en la eliminación de variables con predominancia de valores únicos (varianza cercana a cero) y la agrupación de categorías con baja presencia[60]; lo que a su vez contribuyó a reducir la demanda de recursos computacionales.

En relación con la búsqueda de hiperparámetros, la literatura señala la existencia de tres enfoques distintos. En primer lugar, se encuentra la definición de valores a través de procesos aleatorios, que fue utilizado en la primera etapa. El segundo enfoque implica la selección de valores dentro de rangos predefinidos para cada parámetro de interés. El tercer enfoque contempla las técnicas automáticas que utilizan modelación, las cuales han sido respaldadas como el enfoque más eficaz entre los tres presentados (Thornton et al., 2013; Bergstra et al., 2011). Por lo tanto, en esta etapa se opta por utilizar el enfoque de métodos automáticos, específicamente la optimización bayesiana[61], que ha demostrado tener un alto grado de eficacia en la definición de hiperparámetros

[59]El algoritmo XGBoost posee una amplia variedad de hiperparámetros que pueden ser tuneados (la guía del algoritmo se encuentra disponible en: https://xgboost.readthedocs.io/en/stable/parameter.html). No obstante, para los modelos desarrollados en este estudio se consideran únicamente los hiperparámetros habilitados en *tidymodels* para algoritmos de tipo *boosted tree* (disponibles en: https://parsnip.tidymodels.org/reference/details_boost_tree_xgboost.html) y los parámetros alfa y lambda (regularizaciones).

[60] Categorías con poca representatividad en la variable (≤ 5%) son agregadas en una nueva categoría denominada "otro".

[61]Método que cuya finalidad es buscar el valor máximo o mínimo global de una función objetivo. Se basa en técnicas probabilísticas y modelos bayesianos para encontrar la mejor solución posible.

para algoritmos basados en árboles (Yang y Shami, 2020; Putatunda y Rama, 2018). Asimismo, se minimiza la posibilidad de sobreentrenamiento (*overfitting*) con la generación de particiones a partir del conjunto de entrenamiento (validación cruzada - k = 10) estratificados, para lo cual se usa las variables de interés (p49 y new_ila) como estrato[62]. Con el ajuste del modelo de predicción (el cual implica encontrar los mejores valores de los hiperparámetros, es decir, aquellos que maximizan el rendimiento del modelo en términos de exactitud y capacidad predictiva), se procede a corroborar la capacidad de generalización del modelo en el conjunto de validación, y se evalúa el mismo mediante el uso de diferentes métricas obtenidas por cada modelo.

De forma específica, se utilizan tres métricas: *accuracy*, que representa la proporción de predicciones correctas en relación con el total de predicciones realizadas, *Cohen´s Kappa coefficient* o kappa que evalúa la concordancia entre las predicciones y las clases reales, considerando el acuerdo que se podría esperar por el azar; y F1-score que combina la precision (capacidad de hacer predicciones correctas) y *recall* (capacidad de encontrar todas las instancias relevantes) en una única medida. La Tabla 10 muestra las métricas obtenidas para la variable p49, lo que evidencia el alto poder predictivo del modelo. Se observa que el modelo tiene la capacidad de realizar un gran número de predicciones correctas, como se refleja en un accuracy del 89,4%. Además, muestra un equilibrio entre la capacidad de identificar la categoría de interés y predecir su valor real, con un F1-score de 88,3%. Sin embargo, la métrica Kappa refleja un valor ligeramente inferior a las dos anteriores, con un 78,6%. Estos resultados indican que el modelo está obteniendo buenos resultados en la clasificación general, pero existe una pequeña discrepancia en la concordancia específica entre las predicciones y las clases reales, que no puede atribuirse únicamente al azar. A pesar de esta discrepancia, se considera que el modelo sigue siendo efectivo y presenta un rendimiento aceptable en términos de precisión y capacidad predictiva.

Tabla 10. Métricas de evaluación obtenidas para la tenencia de RUC

Pregunta	Tipo clasificación	*Accuracy*[63]	Kappa	*F1 score*
p49	Binaria	89,4%	78,6%	88,3%

Nota: No se presentan resultados por categorías individuales, ya que las métricas presentadas se derivan directamente de la matriz de confusión, que representa el contraste entre dos categorías. En este caso, los valores de las métricas por categoría serían los mismos.

Por otro lado, en la Tabla 11 se presentan los resultados del modelo asociado a la pregunta new_ila. En primera instancia, se observa un buen poder predictivo, como se refleja en su precisión del 84,1%. Sin embargo, al analizar las métricas más robustas, como el kappa (68,6%) y el *F1-score* (57,3%), se aprecia que el valor de precisión podría no reflejar con precisión la capacidad de predicción verdadera del modelo. De hecho, de acuerdo con las métricas kappa y *F1-score*, se evidencia que el modelo posee una capacidad de predicción mucho menor en comparación con el *accuracy* y carece de la habilidad de identificar con precisión la clase de interés y su valor real.

Tabla 11. Métricas de evaluación obtenidas para el agregado del ingreso

Pregunta	Tipo clasificación	*Accuracy*	Kappa	*F1 score*
new_ila	Múltiple	84,1%	68,6%	57,3%

[62] Permite mantener la distribución de la variable a imputar (categorías) en los conjuntos de entrenamiento y prueba, así como en cada partición (*fold*) de la validación cruzada; además, minimiza la posibilidad de sub-representación en el caso de la variable menos predominante (Forman y Scholz, 2010).

[63] El Anexo 6 refleja el *accuracy* para los modelos implementados en las ENEMDU mensual de octubre 2021 y IV trimestre de 2021.

Sin embargo, es fundamental tener en cuenta que en problemas de clasificación múltiple, el rendimiento del modelo se calcula mediante el promedio de las métricas individuales de cada categoría. Esto significa que el desempeño de una categoría en particular puede tener un impacto significativo en las métricas globales. Por lo tanto, aunque el modelo pueda mostrar un buen rendimiento en la mayoría de las categorías, su efectividad en una categoría específica podría verse afectada y reflejarse en el promedio de las métricas del modelo, como es el caso de la variable new_ila. La Tabla 12 detalla las métricas obtenidas para cada categoría del agregado del ingreso.

Tabla 12. Métricas de evaluación obtenidas por categoría del agregado del ingreso

Pregunta	Categoría	*Accuracy*	Kappa	*F1 score*
New_ila	Ingreso mayor o igual a SBU	85,5%	70,8%	84,1%
	Ingreso menor a SBU	84,3%	64,6%	85,4%
	Ingreso negativo	98,4%	2,3%	2,4%

Al examinar las métricas de evaluación de cada categoría, se pueden identificar comportamientos particulares en cada categoría. En primer lugar, se nota que todas las categorías (1.Ingreso mayor o igual a SBU, 2.Ingreso menor a SBU y 3.Ingreso negativo) muestran un alto poder predictivo en términos de precisión (*accuracy*), con valores de 85,5%, 84,3% y 98,4% respectivamente. Especialmente para la categoría 3. Ingreso negativo, la cual muestra una predicción casi perfecta.

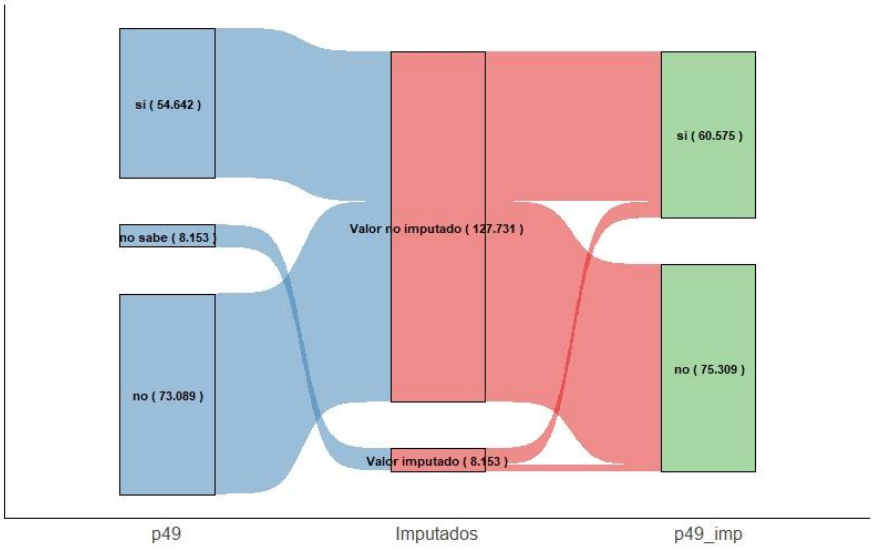
A pesar de esto, al analizar la métrica Kappa, se observan resultados diferentes. Aunque en general se nota una reducción de los valores en comparación con los de precisión (*accuracy*), las escalas varían en cada categoría. Las categorías 1.Ingreso mayor o igual a SBU y 2.Ingreso menor a SBU presentan valores altos de Kappa (70,8% y 64,6% respectivamente), los cuales no pueden compararse con el valor obtenido por la categoría 3.Ingreso negativo, que es solo del 2,3%. Este comportamiento se mantiene en la métrica *F1-score*, donde también se observa un valor bajo de 2,4% para la categoría 3.Ingreso negativo, a pesar de que las categorías 1.Ingreso mayor o igual a SBU y 2.Ingreso menor a SBU presentan valores similares a las métricas de precisión (*accuracy*). Esta discrepancia entre las métricas de predicción podría atribuirse a la predominancia de una o varias categorías sobre otras en la fuente de información (clases desbalanceadas), lo cual incide en las métricas. Como resultado, el modelo de predicción asignará la predicción de nuevos registros a una de las dos categorías con mayor representación (98,5% está representado por las categorías 1.Ingreso mayor o igual a SBU y 2.Ingreso menor a SBU). En caso de que tome el valor de 3.Ingreso negativo, esta predicción podría considerarse casi como una asignación aleatoria.

De este modo, y considerando lo mencionado por Japkowicz & Shan (2011), es importante señalar que los valores óptimos para cada métrica no siguen un estándar, sino que dependen del contexto y las características de la fuente de información. En el presente trabajo, se concluye que los modelos asociados tanto a la pregunta p49 como a new_ila son buenos modelos de predicción. Esta conclusión se cumple especialmente en la predicción de la variable new_ila, ya que, aunque una categoría presenta valores bajos en las métricas de evaluación, esta categoría no es de particular interés en el contexto del estudio. Por lo tanto, se destaca el desempeño general satisfactorio de los modelos en las categorías de mayor relevancia.

3. Resultados

Esta sección tiene como objetivo mostrar los cambios en la distribución de las variables p49 y new_ila como resultado de la imputación en comparación con las variables originales. Además, presenta las transiciones laborales que se observaron a partir de las variables de sectorización del empleo y clasificación de la población según su condición de actividad después de la imputación. Para comenzar, se realiza un análisis muestral de la imputación de las variables p49 y new_ila, utilizando un gráfico de Sankey[64] que permite visualizar la distribución de las observaciones imputadas en las diferentes categorías de las variables. La Figura 5 muestra los resultados de la imputación para la variable p49. De las 8.553 observaciones imputadas (muestrales), que originalmente correspondían a la categoría "no sabe", se distribuyeron de la siguiente manera después de la imputación: 5.933 observaciones (72,8%) fueron clasificadas como "si" y 2.220 observaciones (27,2%) se clasificaron como "no".

Figura 5. Flujo de valores imputados para la pregunta p49 (Datos muéstrales)

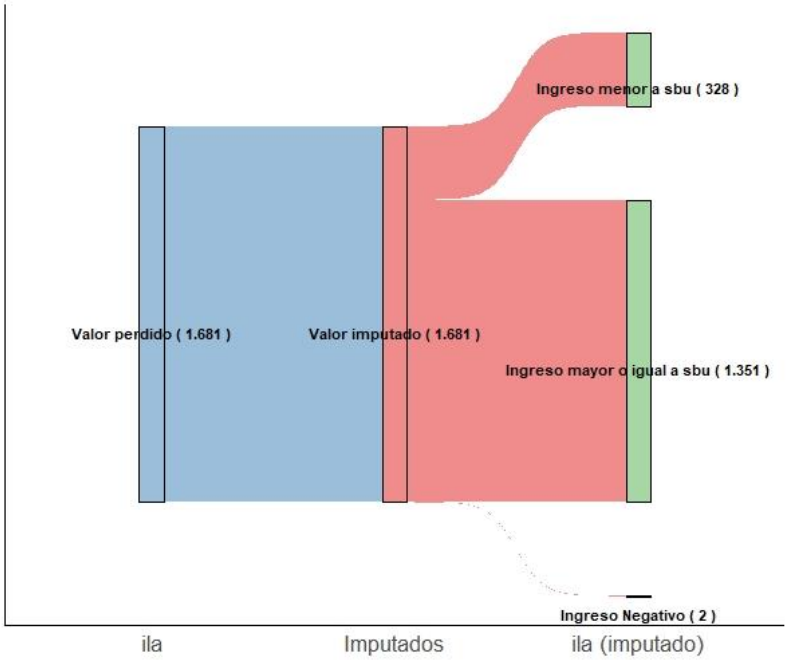


Fuente: ENEMDU 2021

La Figura 6, por su parte, representa el flujo de registros que fueron sometidos a imputación en la variable new_ila. Se observa que de los 1.681 valores imputados, lo cual representa el 0,8%[65] del total de valores perdidos en la muestra de la variable, 1.351 valores (80,4%) fueron clasificados como 1.Ingreso mayor o igual al SBU, 328 valores (19,5%) fueron clasificados como 2.Ingreso menor al SBU, y solo 2 observaciones (0,1%) fueron predichas como 3.Ingreso negativo.

[64] El gráfico de Sankey es una técnica de visualización que permite la representación e identificación de flujos entre variables de interés, para más información dirigirse al url: https://www.data-to-viz.com/graph/sankey.html
[65] Las demás observaciones que en la variable new_ila tienen valores *missing* corresponden a empleados no remunerados e individuos pertenecientes a la población económicamente inactiva.

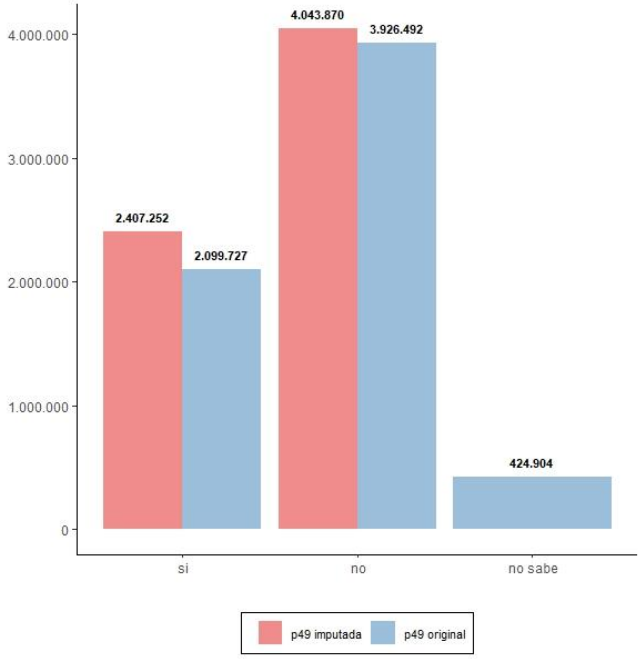
Figura 6. Flujo de valores imputados para el agregado del ingreso (new_ila) (Datos muéstrales)



Fuente: ENEMDU 2021

Una vez se ha visualizado el flujo de la muestra imputada a través de las categorías de las variables p49 y new_ila, se presentan los cambios en la distribución de las categorías a nivel poblacional. En particular, en relación a la pregunta p49, se observa un cambio significativo como resultado de la imputación. La Figura 7 ilustra que 307.525 personas que originalmente se encontraban en la categoría "no sabe" han sido reasignadas a la categoría "si", mientras que 117.378 personas han sido reasignadas a la categoría "no".

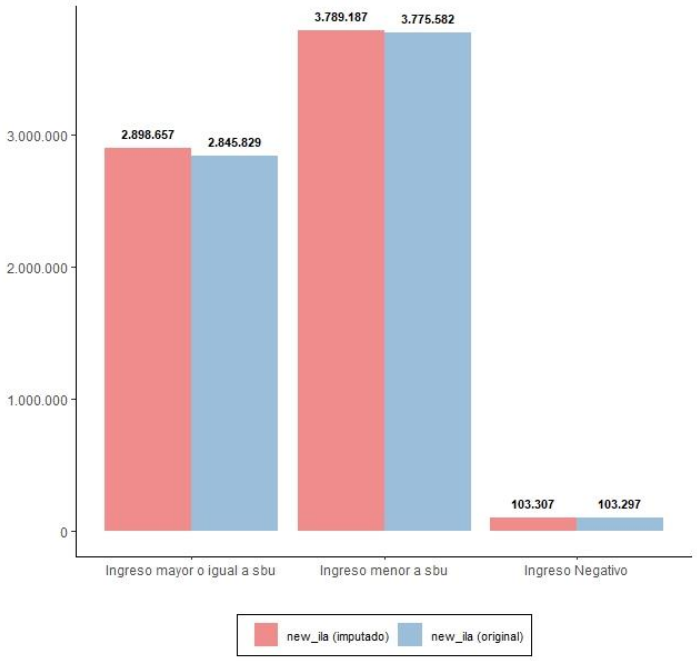
Figura 7. Distribución poblacional de la variable p49 vs. p49 imputada



Fuente: ENEMDU 2021

La imputación de la variable new_ila, al igual que en el caso de la pregunta p49, también resultó en un cambio en la distribución poblacional. Específicamente, se observó que un total de 66.443 personas experimentaron una reasignación en las categorías de esta variable. De este modo, se encontró que 52.828 personas pasaron a clasificarse en la categoría 1.Ingreso mayor o igual al SBU, mientras que 13.605 personas se ubicaron en la categoría 2.Ingreso menor al SBU. Además, un pequeño grupo de 10 personas se movió a la categoría 3.Ingreso negativo.

Figura 8. Distribución poblacional de la variable new_ila vs. new_ila imputada



Fuente: ENEMDU 2021

Luego de analizar los cambios resultantes de la imputación en las preguntas, a continuación, se construyen las variables de sectorización del empleo y condición de actividad. A través de estas variables, se explorarán las transiciones de la población entre diferentes grupos poblacionales. Este análisis permitirá comprender cómo se han modificado la clasificación del empleo y la condición laboral de las personas como resultado de la imputación de datos.

3.1. Sectorización del empleo

La variable de sectorización del empleo[66] se reconstruyó utilizando la variable p49 imputada (consulte la sección 2.1.2.1 para obtener más detalles sobre su construcción). En esta sección, se presenta el comportamiento de los distintos grupos poblacionales que conforman la variable.

En este sentido, en la Tabla 13 se observa que un total de 424.904 personas (equivalente al 5,4% del empleo total) que se encontraban como no clasificado por sector, se redistribuyeron, de tal forma que, 307.525 personas (equivalente al 3,8% del empleo total) pasaron a ubicarse en el sector formal y 117.378 personas (equivalente al 1,5% del

[66] Las categorías que comprenden la variable son: 1.Sector formal; 2.Sector informal; 3.Empleo doméstico; 4.No clasificado por sector.

empleo total) al sector informal. Así, al analizar la proporción del empleo en el sector formal, se observa un cambio del 42,9% al 46,8% como resultado de la redistribución del grupo originalmente clasificado como no clasificado por sector. En cuanto al empleo en el sector informal, el cambio en la proporción es menor, pasando del 49,5% al 51,0%

Tabla 13. Transiciones poblacionales de la sectorización del empleo

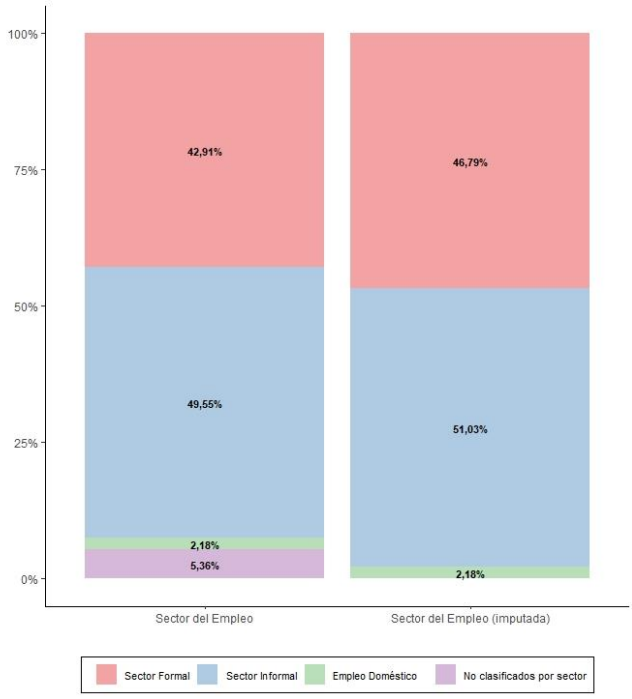
Sector del empleo		Sector del empleo (imputado)				Total
		Sector Formal	Sector Informal	Empleo Doméstico	No clasificado por sector	
Sector del empleo (original)	Sector Formal	3.400.196 (84.650)	-	-	-	3.400.196 (84.650)
		42,9%	-	-	-	42,9%
	Sector Informal	-	3.926.492 (73.089)	-	-	3.926.492 (74.089)
		-	49,5%	-	-	49,5%
	Empleo Doméstico	-	-	173.005 (3.727)	-	173.005 (3.727)
		-	-	2,2%	-	2,2%
	No clasificado por sector	307.525 (5.933)	117.378 (2.220)	-	-	424.904 (8.153)
		3,8%	1,5%	-	-	5,4%
Total		3.707.721 (90.583)	4.043.870 (75.309)	173.005 (3.727)	-	7.924.595 (169.619)
		46,8%	51,0%	2,2%	-	100%

Nota:
1) Los valores entre paréntesis corresponden al total de personas a nivel muestral
2) Las celdas resaltadas corresponden a las poblaciones que cambian como resultado de la imputación de la pregunta p49
Fuente: ENEMDU 2021 (imputada)

Al calcular los indicadores laborales basados en la variable de sectorización del empleo antes y después de la imputación, se observan cambios significativos como resultado de la reclasificación de la población que originalmente se encontraba en la categoría no clasificado por sector. Estos cambios se reflejan en los indicadores, tal como se muestra en la Figura 9. En particular, se observa que el indicador de empleo en el sector formal aumentó del 42,9% al 46,8%, lo cual representa un cambio estadísticamente significativo.

Del mismo modo, el empleo en el sector informal experimentó un aumento del 49,6% al 51,0%, también con una diferencia estadísticamente significativa. Además de los cambios en los indicadores, se analizaron los errores estándar y los intervalos de confianza asociados. Se pudo observar que no existen cambios significativos en la dispersión de los indicadores antes y después de la imputación. Esto indica que la precisión de los indicadores se mantiene constante, lo cual es un aspecto positivo en términos de confiabilidad de las estimaciones (para mayor detalle diríjase a los Anexos 4 y 5).

Figura 9. Comparación de los indicadores de sectorización laboral antes y después de la imputación de la pregunta p49



Nota:
1) El denominador de los indicadores el empleo total
2) El aumento de la proporción de empleados en el sector formal e informal representa un cambio estadísticamente significativo al 95% de nivel confianza.
Fuente: ENEMDU 2021 (imputada)

3.2. Condición de actividad de los empleados

La variable de condición de actividad de los empleados también fue reconstruida a partir de la variable imputada new_ila, siguiendo los procedimientos detallados en la sección 2.1.2.2. Esta reconstrucción implicó una reclasificación de las categorías de la variable en comparación con la variable original.

En la Tabla 14 se muestra la fluctuación de las categorías poblacionales de los empleados tras la imputación del new_ila. Se observa que el cambio se encuentra en las personas que inicialmente se clasificaron en la categoría de empleo no clasificado. De las 61.131 personas (0,8% del empleo total) que se encontraban en esta categoría, se produjo una transición donde 51.598 personas (0,7% del empleo total) pasaron a ser clasificadas como empleo adecuado, 8.923 personas (0,0%) se clasificaron como otro empleo inadecuado, y 610 personas (0,0% del empleo total) se clasificaron como subempleo por insuficiencia de ingresos. En tanto que, 5.312 personas (0,0% del empleo total) que carecían de ingreso pertenecientes a la categoría de subempleo por insuficiencia de horas se mantuvieron en la misma categoría pese a la imputación, debido a que prima en la construcción del indicador las horas trabajadas respecto de los ingresos laborales.

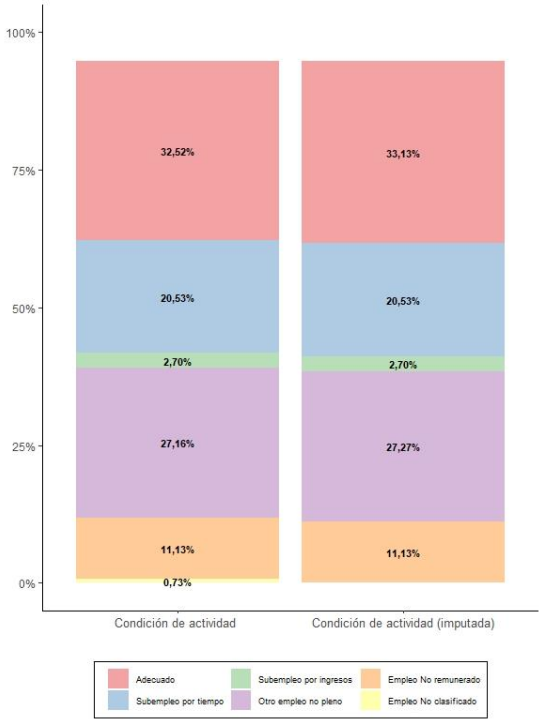
Tabla 14. Transiciones poblacionales de la condición de actividad

Condición de actividad del empleo		Condición de actividad imputada						Total
		Adecuado	Subempleo por insuficiencia de tiempo	Subempleo por insuficiencia de ingresos	Otro empleo inadecuado	Empleo No remunerado	Empleo No clasificado	
Condición de actividad original	Adecuado	2.719.099 (64.596)	-	-	-	-	-	2.719.099 (64.596)
		34,3%						34,3%
	Subempleo por insuficiencia de tiempo	-	1.717.003 (34.649)	-	-	-	-	1.717.003 (34.649)
			21,7%					21,7%
	Subempleo por insuficiencia de ingresos	-	-	225.496 (4.843)	-	-	-	225.496 (4.843)
				2,8%				2,8%
	Otro empleo inadecuado	-	-	-	2.271.234 (45.185)	-	-	2.271.234 (45.185)
					28,7%			28,7%
	Empleo No remunerado	-	-	-	-	930.633 (18.796)	-	930.633 (18.796)
						11,7%		11,7%
	Empleo No clasificado	51.598 (1.320)	-	610 (13)	8.923 (217)	-	-	61.131 (1.550)
		0,7%		0,0%	0,0%			0,8%
	Total	2.770.697 (65.916)	1.717.003 (34.649)	226.105 (4.856)	2.280.157 (45.402)	930.633 (18.796)	-	7.924.595 (169.619)
		35,0%	21,7%	2,8%	28,8%	11,7%	-	100%

Nota:
1) Los valores entre paréntesis corresponden al total de personas a nivel muestral
2) Las celdas resaltadas corresponden a las poblaciones que cambian como resultado de la imputación del ila
Fuente: ENEMDU 2021 (imputada)

Por último, la Figura 10 muestra los cambios en los indicadores laborales derivados de la condición de actividad antes y después de la imputación de la variable new_ila. Se observa que el empleo no clasificado, que originalmente representaba el 0,1% de la población económicamente activa (PEA), desaparece. Asimismo, se observa un ligero aumento en la tasa de empleo adecuado, que pasa del 32,5% de la PEA al 33,1% de la PEA, y el otro empleo inadecuado cambia del 27,2% de la PEA al 27,3% de la PEA. Sin embargo, las diferencias en los indicadores derivados de la imputación no resultaron ser estadísticamente significativas (para mayor detalle el diríjase al Anexo 4 y 5). En cuanto a la dispersión de los indicadores, no se observaron cambios significativos y la precisión se mantuvo. Esto sugiere que la imputación de la variable new_ila no tuvo un impacto sustancial en la estimación de los indicadores laborales y que los resultados se mantuvieron consistentes.

Figura 10. Condición de actividad antes y después de la imputación



Nota:
1) El denominador de los indicadores es la PEA
2) No se observan diferencias estadísticamente significativas
Fuente: ENEMDU 2021 (imputada)

4. Conclusiones

Este documento presenta una propuesta metodológica para la imputación de datos perdidos o sin respuesta en encuestas de hogares mediante el uso de técnicas de ML, tales como *ensamble learning* (RF y XGBoost), *neural networks* (MLP) y métodos supervisados simples (SVM). Se discute además cómo este proceso impacta en las estimaciones de indicadores.

Esta metodología fue utilizada para imputar las variables correspondientes a la tenencia de RUC del lugar de trabajo de los empleados (p49) y al ingreso laboral (ila) de la ENEMDU anual del 2021 realizada por el INEC. Con el fin de mejorar la predicción de la variable ila se construyó una nueva variable (new_ila) que, a diferencia de la variable original, mantiene los ingresos laborales negativos en lugar de asignarlos como valores *missing*. Además, para mantener la misma categorización de predicción (clasificación) para ambas variables, se transformó de numérica a categórica, definiendo así tres categorías: 1.Ingreso mayor o igual al SBU; 2.Ingreso menor o igual al SBU; y 3.Ingresos negativos.

El proceso de imputación comenzó con la evaluación del mejor modelo de predicción utilizando criterios estándar como la búsqueda aleatoria de parámetros y la validación cruzada con *5 folds* para la información faltante de ambas variables (lo que permitió la comparación de los modelos), utilizando la métrica *accuracy*.

Los resultados obtenidos mostraron que los modelos de ensamble RF y XGBoost presentaron la mejor performance de predicción, siendo el modelo XGBoost el que

obtuvo el *accuracy* más alto, con un 88,7% para la variable p49 y un 83,2% para la variable new_ila. Estos resultados son consistentes con los hallazgos de Rosati (2021) y Jerez et al. (2010). Una vez seleccionada la técnica XGBoost[67] como la de mejor desempeño, se utilizó esta metodología para la imputación definitiva de ambas variables, pero previamente se optimizó el algortimo a través de un análisis exploratorio de variables y la optimización bayesiana para la búsqueda de hiperparámetros.

La imputación de ambas variables mostró un *accuracy* en la predicción de la variable p49 de 89,4% y new_ila de 84,1%, lo cual demostró la acertada selección de variables y la optimización de hiperparámetros, y una mejora en la precisión de la predicción de la imputación en comparación con los modelos preliminares.

Como resultado de la imputación, se observaron cambios en la distribución de las variables. En el caso de la variable p49, se encontró que la población empleada que inicialmente respondió "no sabe" en la pregunta se distribuyó en las categorías "si" y "no" con 307.525 y 117.378 individuos, respectivamente, después de la imputación. Por otro lado, la imputación de la variable new_ila permitió predecir los ingresos para un total de 66.443 empleados que inicialmente no reportaron información sobre los ingresos laborales y no presentaron ingresos incoherentes. Los cambios en la distribución de los empleados se reflejaron en 52.828 personas que fueron clasificadas como ingreso mayor o igual al SBU, 13.605 personas como ingreso menor al SBU y 10 personas como ingreso negativo.

Una vez que se imputaron ambas variables, se procedió al cálculo de los indicadores laborales de sectorización laboral y condición de actividad del empleo, y se compararon los resultados con los estimadores oficiales. Las estimaciones realizadas para la variable de sectorización del empleo, basada en la pregunta p49 imputada, revelaron una reclasificación completa de los individuos que inicialmente no estaban clasificados por sector, asignándolos a las categorías de empleo formal e informal. Específicamente, de los 424.904 individuos clasificados en la categoría no clasificado por sector, se reasignaron 307.525 al sector formal y 117.378 al sector informal. Este resultado tuvo un impacto significativo en las tasas de empleo en ambos sectores. La tasa de empleo en el sector formal aumentó del 42,9% al 46,8%, mientras que la tasa de empleo en el sector informal experimentó un cambio del 49,6% al 51,0%. Ambas diferencias resultaron ser estadísticamente significativas y la precisión del indicador se mantuvo inalterada al analizar los errores estándar de los indicadores antes y después de la imputación.

Por otro lado, la variable de condición de actividad, construida en base a la variable new_ila, reveló cambios en la distribución de los empleados. Inicialmente, de las 61.130 personas clasificadas como empleo no clasificado, se reasignaron 51.598 al empleo adecuado, 8.923 a otro empleo inadecuado y 610 al subempleo por insuficiencia de ingresos.

[67] Con el objetivo de evaluar la solidez del enfoque utilizado, se utilizó la metodología basada en modelos XGBoost para predecir las variables p49 y new_ila en encuestas adicionales, como la ENEMDU mensual de octubre de 2021 y la ENEMDU trimestral del cuarto trimestre de 2021. Los resultados mostraron que en algunos casos se logró obtener un *accuracy* aún mayor que la alcanzada en la base de datos anual (con una *accuracy* de 84,2% para new_ila mensual y anual), lo cual confirma la capacidad predictiva de los modelos desarrollados. Este análisis adicional refuerza la confianza en la robustez del enfoque utilizado.

Estos cambios tuvieron efectos en las estimaciones de los indicadores correspondientes. La tasa de empleo adecuado aumentó ligeramente del 32,5% de la PEA al 33,1% de la PEA, mientras que la tasa de otro empleo inadecuado también experimentó un incremento mínimo, del 27,2% de la PEA al 27,3% de la PEA. Sin embargo, ninguna de estas diferencias resultó ser estadísticamente significativa. Además, al analizar la dispersión a través de los errores estándar, no se evidenciaron cambios significativos en la precisión de los indicadores antes y después de la imputación.

Así, este estudio resaltó la importancia de la imputación de información perdida o faltante en bases de datos, especialmente en encuestas de empleo. La presencia de valores faltantes puede afectar la calidad y confiabilidad de los resultados obtenidos. Es por eso que el uso de técnicas de *machine learning* se vuelve fundamental en este proceso. Las técnicas de *machine learning* ofrecen una poderosa herramienta para la imputación de datos perdidos. Al utilizar algoritmos avanzados y modelos predictivos, estas técnicas permiten inferir los valores faltantes a partir de la información disponible en los datos. Esto contribuye a completar la base de datos de manera precisa y confiable, evitando así la distorsión de los análisis y estimaciones.

Además, el estudio demostró la eficacia de las técnicas de *machine learning*, específicamente en la imputación de información perdida en encuestas de empleo. Al utilizar algoritmos como XGBoost, se logró obtener resultados con un alto nivel de precisión y poder predictivo, lo que proporciona una base sólida para la generación de estimaciones confiables.

Por último, se recomienda que los modelos analizados en futuros estudios combinen distintas técnicas de búsqueda de hiperparámetros y realicen un análisis exploratorio de variables más exhaustivo. Esto permitirá mejorar la precisión y el rendimiento de los modelos de imputación utilizados. Además, es importante explorar otras técnicas de *machine learning* que sean más eficientes computacionalmente en comparación con las utilizadas en este estudio. Algunas opciones a considerar son los algoritmos genéticos, la eliminación recursiva de variables y la optimización bayesiana. Estas técnicas pueden ofrecer resultados aún más precisos y eficientes en términos de tiempo de procesamiento.

Se sugiere también que en ejercicios futuros se realice un contraste entre la imputación mediante técnicas de *machine learning* y las técnicas de imputación tradicionales, como *hot deck, cold deck* o vecino más cercano. Este contraste permitirá evaluar y comparar el desempeño de diferentes enfoques de imputación y determinar cuál es el más adecuado para cada situación.

Bibliografía

Alfaro, R., & Fuezalida, M. (2009). Imputación múltiple en encuestas microeconómicas . *Cuaderno de economía*, 273 - 288.

Arce, A., Cárdenas, A., Canales, V., & Lehmann, K. (2019). Métodos de imputación VIII EPF: Gastos diarios e ingresos de la actividad laboral y jubilaciones. *Instituto Nacional de Estadística - Chile*, 1-114.

Australian Bureau of Statistics. (2017). *Household Expenditure Survey, Australia: Summary of Results methodology*. Obtenido de Household Expenditure Survey, Australia:

Summary of Results methodology: https://www.abs.gov.au/methodologies/household-expenditure-survey-australia-summary-results-methodology/2015-16

Aydilek, I., & Arslan, A. (2013). A hybrid method for imputation of missing values using optimized fuzzy c-means with support vector regression and genetic algorithm. *Information Sciences*, 25 - 35.

Baraldi, A., & Enders, C. (2010). An introduction to modern missing data analyses. *Journal of School Psychology 48*, 5-37.

Beccaria, & Gluzmann. (2013). Medición de los Ingresos y la Pobreza Oficial en América Latina y el Caribe. *CEDLAS*, 1 - 142.

Bergstra, J., Bardenet, R., Bengio, Y., & Kégl, B. (2011). Algorithms for Hyper-Parameter Optimization. *Advances in neural information processing systems, 24*.

Berry, M., Mohamed, A., & Wah Yap, B. (2020). Supervised and Unsupervised Learning for Data Science. Switzerland : Springer Nature .

Breiman, L., & Spector, P. (1992). Submodel Selection and Evaluation in Regression. The X-Random Case. *International Statistical Review / Revue Internationale de Statistique, 60*(3), 291 - 319. doi:10.2307/1403680

Canada, S. (2020). *Canadian Income Survey - 2020 (CIS)*. Obtenido de Canadian Income Survey - 2020 (CIS): https://www23.statcan.gc.ca/imdb/p2SV.pl?Function=getSurvey&SDDS=5200

Castillo y Rosero. (2015). Empleo y condición de actividad en el Ecuador. *Revista de Estadística y Metodologías - Volumen 1*, 30 - 53.

Castillo, & Puebla. (2016). Aspectos metodológicos sobre la medición de la pobreza por ingresos en el Ecuador. *Revista de Estadística y Metodologías - Número 2*, 23 - 25.

Castillo, R. (2015). *Empleo y condición de actividad en Ecuador.*

CEPAL. (2007). *Imputación de datos: teoría y práctica.*

Deng, Y., & Lumley, T. (2021). MULTIPLE IMPUTATION THROUGH XGBOOST. 1 -14.

Departamento Administrativo Nacional de Estadística (DANE). (2020). Imputación de la condición de informalidad de los ocupados en Colombia para marzo y abril de 2020. 1-14.

Donza, E. (2013). Método de imputación de la no respuesta en las preguntas de ingresos en la Encuesta Permanente de Hogares. Gran Buenos Aires 1990 - 2010. *X Jornadas de Sociología. Facultad de Ciencias Sociales, Universidad de Buenos Aires*, (págs. 1 - 25). Buenos Aires.

Eltinge, J., Kozlow, R., & Luery, D. (s.f.). Imputation in Three Federal Statistical Agencies. 1-23.

Ergul, Z., & Kamisli, Z. (2021). Performance Analysis of XGBoost Classifier with Missing Data. *Conference Paper*, 1-5.

Forman, G., & Scholz, M. (2010). Apples-to-Apples in Cross-Validation Studies: Pitfalls in Classifier Performance Measurement. *Acm Sigkdd Explorations Newsletter*, 49-57.

Graham, J. W. (2009). Missing data analysis: Making it work in the real world. *Annual review of psychology, 60*, 549-576.

Graham, J. W. (2012). *Missing data: Analysis and design.* Springer Science & Business Media.

Hapfelmeier, & Ulm. (2014). Variable selection by Random Forests using data with missing values. *Computational Statistics and Data Analysis*, 129-139.

Hasan, Ahraful, Shidhartho, Aishwariya, Tasnim, & Sunanda. (2021). Missing value imputation affects the performance of machine learning: A review and analysis of the literature (2010–2021). *Informatics in Medicine Unlocked 27*, 1-23.

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

Instituto Nacional de Estadística y Censos (INEC). (2021). *Diseño muestral de la Encuesta Nacional de Empleo, Desempleo y Subempleo (ENEMDU)*.

Instituto Nacional de Estadística y Censos (INEC). (2022). *Diseño muestral - Encuesta Nacional de Empleo, Desempleo y Subempleo - ENEMDU Anual.*

Jail, A. (2010). Data clustering: 50 years beyond K-means. *Pattern Recognition Letters*, 651 - 666.

Japkowicz, N., & Shan, M. (2011). Performance Measures I. En *Evaluating Learning Algorithms: A Classification Perspective* (págs. 74 - 110). Cambridge: Cambridge University Press. doi:10.1017/CBO9780511921803.004

Jerez, J., Molina, I., García-Laencina, P., Alba, E., Ribelles, N., Martín, M., & Franco, L. (2010). Missing data imputation using statistical and machine learning methods in a real breast cancer problem. *Artificial Intelligence in Medicine*, 105 - 115.

Kaiser, & Jiří. (s.f.). Algorithm for Missing Values Imputation in Categorical Data with Use of Association Rules . *ACEEE*, 1-4.

Kaiser, J. (2014). Dealing with Missing Values in Data. *Journal of systems integration*, 42-51.

Khun, M., & Jhonson, K. (2013). Classification Trees and Rule-Based Models. En M. Kuhn, & K. Johnson, *Applied predictive modeling* (Vol. 16, págs. 369 - 413). New York: Springer. doi:10.1007/978-1-4614-6849-3

Kohavi, R. (1995). A Study of Cross-Validation and Bootstrap for Accuracy Estimation and Model Selection. *Ijcai, 14*(2), 1137-1145.

Lin, & Tsai. (2019). Missing value imputation: a review and analysis of the literature (2006 - 2017). *Artificial Intelligence Review*, 1-23.

Lobato, F., Sales, C., Tadaiesky, V., Dias, L., Ramos, L., & Santana, A. (2015). Multi-Objective Genetic Algorithm For Missing Data Imputation. *Pattern Recognition Letters*, 1-9.

Loh, Wei-Lin, Etinge, J., Cho, M. J., & Li, Y. (2019). CLASSIFICATION AND REGRESSION TREES AND FORESTS FOR INCOMPLETE DATA FROM SAMPLE SURVEYS. *Statistica Sinica*, 431-453.

Molina, Rivadeneira, & Rosero. (2015). *Actualización metodológica: el empleo en el sector informal.*

Office for National Statistics. (2021). Labour Force Survey: alternative imputation during the coronavirus (COVID-19) pandem,ic. 1-11.

Putatunda, S., & Rama, K. (Noviembre de 2018). A comparative analysis of hyperopt as against other approaches for hyper-parameter optimization of XGBoost. *Proceedings of the 2018 international conference on signal processing and machine learning*, 6-10.

Raja, & Thangavel. (2019). Missing value imputation using unsupervised machine learning techniques. *METHODOLOGIES AND APPLICATION*, 2 - 32.

Rajoub, B. (2020). Biomedical Signal Processing and Artificial Intelligence in Healthcare. Dubai, United Arab Emirates.

Restrepo, M., & Marín, J. (2012). Imputación de ingresos en la Gran Encuesta Integrada de Hogares (geih) de 2010. *Revista Desarrollo y Sociedad*, 219-243.

Richman, M., Trafalis, T., & Adrianto, I. (2009). Missing Data Imputation Through Machine Learning Algorithms. *Artificial Intelligence Methods in the Environmental Sciences*, 153 - 169.

Rodríguez, E., & López, B. (2015). Imputación de ingresos laborales. Una aplicación con encuestas de empleo en México. *EL TRIMESTRE ECONÓMICO* , 117-146.

Rosati, G. (2021). Métodos de Machine Learning como alternativa para la imputación de datos perdidos. Un ejercicio en base a la Encuesta Permanente de Hogares. *Estudios del trabajo*, 1-24.

Rubin, D. (1976). Inference and missing data. *Biometrika*, 581 - 592.

Servicio de Rentas Internas (SRI). (2022). *Servicio de Rentas Internas (SRI)*. Obtenido de https://www.sri.gob.ec/ruc-personas-naturales#%C2%BFqu%C3%A9-es

Snoek, J., Larochelle, H., & Adams, R. P. (2012). Practical bayesian optimization of machine learning algorithms. *Advances in neural information processing systems, 25*.

Statistics, A. B. (2017). *Hosehold Expenditure Survey, Australia: Summary of Results methodology*.

Stewart, T., Zeng, D., & Wu, M. (s.f.). Constructing support vector machines with missing data. 1-27.

Thornton, C., Hutter, F., Hoos, H. H., & Leyton-Brown, K. (Agosto de 2013). Auto-WEKA: Combined selection and hyperparameter optimization of classification algorithms. *Proceedings of the 19th ACM SIGKDD international conference on Knowledge discovery and data mining*, 847-855.

Tlamelo, E., Maupong, T., Mpoeleng, D., Semong, T., Mphago, B., & Tabona, O. (2021). A survey on missing data in machine learning. *Journal of Big Data*, 1-37.

UNECE. (2022). Machine Learning Imputation for Social Surveys: Random Forest imputation of ONS' Household Financial Survey. *MODERNSTATS*, 1-12.

Yang, L., & Shami, A. (2020). On hyperparameter optimization of machine learning algorithms: Theory and practice. *Neurocomputing*, 295-316.

Zhanga, S. (2012). Nearest neighbor selection for iteratively kNN imputation. *The Journal of Systems and Software*, 2541 - 2552.

Zhu, X., Zhang, S., Jin, Z., Zhang, Z., Xu, & Zhuoming. (2022). Missing Value Estimation Mixed-Attribute Data Sets. *IEEE TRANSACTIONS ON KNOWLEDGE AND DATA ENGINEERING*, 110 - 121.

Anexos

Anexo 1: Estadísticos de evaluación de clasificación

Tabla A1.1. Métricas de evaluación de la clasificación

	Condición predicha		
Condición real	**Total poblacional=P+N**	**Condición predicha positiva (PP)**	**Condición predicha negativa (PN)**
	Condición real Positivo (P)	Verdaderos positivos (VP) (Predicción correcta)	Falsos negativos (FN) (Error tipo 2)
	Condición real negativo (N)	Falsos positivos (FP) (Error tipo 1)	Verdaderos negativos (VN) (Predicción correcta)

Fuente: Hasan et al. (2021)

Tabla A1.2. Métricas de evaluación de un problema de clasificación

Estadístico	**Definición**
$VPR = \frac{VP}{VP + FN}$	Sensitivity: mide la proporción de reales positivos correctamente identificados como positivos sobre el total de condiciones reales positivas.
$VNR = \frac{VN}{VN + FP}$	Specificity: mide la proporción de los reales negativos que son identificados correctamente como negativos, sobre el total de condiciones reales negativas.
$FPR = \frac{FP}{FP + VN}$	False positive ratio: mide la proporción falsos positivos, sobre el total de condiciones reales negativas, probabilidad de rechazar falsamente la hipótesis de error tipo 1.
$FNR = \frac{FN}{FN + VP}$	False negative ratio: mide la proporción de falsos negativos sobre el total de condiciones reales positivas; prueba la hipótesis de error tipo 2.
$PPV = \frac{VP}{VP + FP}$	Precision: mide la proporción de observaciones reales positivas sobre el total de condiciones reales positivas.
$NPV = \frac{VN}{VN + FN}$	Negative predictive value: mide la proporción de observaciones reales negativas sobre el total de condiciones predichas negativas.
$FDR = \frac{FP}{FP + VP}$	False discovery rate: mide la proporción de observaciones reales negativas sobre el total de condiciones predichas positivas
$FOR = \frac{FN}{FN + VN}$	False omission rate: mide la proporción de falsos negativos sobre el total de predichos negativos, considerando la probabilidad de error tipo 2.
$ACC = \frac{VP + VN}{P + N}$	Accuracy: mide que tan bien una prueba de clasificación binaria identifica o excluye una condición.
$F1\ score = 2 * \frac{PPV * VPR}{PPV + VPR}$	F1 score: corresponde a la media armónica de precisión y recuperación.
$kappa = \frac{N * (VP + VN) - \left(((VP + FN) * (VP + FP)) + ((FN + VN) * (FP * VN))\right)}{N^2 - \left(((VP + FN) * (VP + FP)) + ((FN + VN) * (FP * VN))\right)}$	Cohen´s Kappa coefficient: mide el grado de acuerdo o desacuerdo entre dos evaluadores (valore real vs. valor predicho).

Nota:
(1) Error tipo 1: Falso positivo.
(2) Error tipo 2: Falso negativo.
Fuente: Hasan et al. (2021)

Tabla A1.3. Métricas de evaluación de un problema de regresión

Métrica	**Definición**	**Utilidad**
Media absoluta del error (MAE)*	$MAE = \frac{1}{n}\sum_1^n \lvert x_i - \hat{x}_i \rvert$, donde x_i=valores actuales de las n muestras x_i=valor imputado de las n muestras	Estima la media del error de los valores predichos y valores reales y evalúa la imputación de valores continuos.

Error porcentual absoluto medio (MAPE)*	$MAPE = \frac{100}{n}\sum_i^n \lvert\frac{x_i - \bar{x}_i}{x_i}\rvert$, donde x_i=valores actuales de las n muestras $\hat{x}_i$=valor imputado de las n muestras	Mide el error porcentual absoluto medio que ha sido imputado en las posiciones que faltan para los atributos continuos.
Raíz cuadrada del error medio (RMSE)*	$RMSE = \sqrt{\frac{1}{n}\sum_i^n (x_i - \bar{x}_i)^2}$, donde x_i=valores actuales de las n muestras $\hat{x}_i$=valor imputado de las n muestras	Mide el error cuadrático medio de las variables continuas pronosticadas con respecto a las variables reales.

Nota:
*variables continua.
Fuente: Hasan et al. (2021)

Anexo 2: Análisis exploratorio (EDA) de la Encuesta Nacional de Empleo, Desempleo y Subempleo 2021

Estructura general de la fuente de información

Antes de generar modelos de clasificación, es importante realizar una exploración de los datos disponibles para obtener un conocimiento profundo de su estructura e identificar posibles relaciones entre variables. Esta exploración, conocida como Análisis Exploratorio de Datos (EDA), también ayuda en la selección de variables a incorporar en los modelos, alerta al investigador sobre la necesidad de aplicar transformaciones o pre-procesamiento a las variables, y en general mejora la eficiencia del modelo. En la Tabla A2.1 se presenta una primera revisión realizada sobre el conjunto total de observaciones, donde se analizan diferentes aspectos y características de los datos.

Tabla A2.1. Estructura de la Encuesta Nacional de Empleo, Desempleo y Subempleo 2021

Descripción		Valores
No. de registros (filas)		361.790
No. de variables (columnas)		262
No. de variables de tipo numérico (numérico - entero)		130
No. de variables de tipo categóricas (factor)		132
%. de variables con casos completos		22,9% (60)
%. de variables que contienen casos perdidos	> 0% y < 50%	18,7% (49)
	≥ 50% y < 90%	31,3% (83)
	≥ 90%	27,1% (70)

Fuente: ENEMDU 2021

La base de datos cuenta con un alto número de observaciones y variables, lo cual también se acompaña de un elevado número de valores perdidos (48.236.985 en toda la base de datos) distribuidos en todas las variables de la base de datos (un total de 202 variables tienen valores perdidos). Es importante destacar que, en el contexto de ENEMDU, estos valores perdidos no se deben a la falta de información o respuestas, sino que están relacionados con los flujos establecidos para garantizar el correcto direccionamiento de la encuesta.

Registros disponibles para cada pregunta de imputación (p48, p49, new_ila[68])

La ENEMDU 2021, posee una alta cantidad de observaciones (361.790), a pesar de esto la presencia de criterios como: i) flujos internos dentro de la encuesta; ii) población

[68] La variable new_ila mencionada hace referencia a la variable numérica del ingreso laboral reconstruida en el marco de la investigación y a la variable categórica derivada.

objetivo (15 años o más; iii) registros sujetos a imputación ("no sabe"), reducen el número de registros que efectivamente pueden ser utilizados para la generación de modelos de ML. Bajo estas consideraciones, la Tabla A2.2 presenta los registros disponibles así como el porcentaje de información para cada variable de interés.

Tabla A2.2. Total de registros para cada pregunta de imputación

Pregunta	Total registros ENEMDU	Total registros para modelado	% Registros disponibles para modelado (% respecto del total de registros ENEMDU)
P48	361.790	100.389	27,7%
P49		101.594	28,1%
new_lla**		144.847	40,0%

** La fila new_ila representa el número de registros disponibles tanto para la variable new_ila numérica como categórico.
Fuente: ENEMDU 2021

Depuración de variables para modelado

Una vez que se ha establecido el universo para cada pregunta de imputación, es necesario realizar un proceso de depuración debido al alto número de variables disponibles (262). En este proceso, se prioriza la no reducción del número de registros y se enfoca en seleccionar las variables con mayor poder predictivo[69].

Variables con valores perdidos

En primer lugar, eliminan las variables que presentan valores perdidos. Esta acción se lleva a cabo por dos razones principales: i) la existencia de flujos internos en la encuesta que excluyen a individuos específicos, y ii) la presencia alta o absoluta de valores perdidos en múltiples variables. En la Tabla A2.3 se muestra el número total de variables eliminadas para cada pregunta de imputación.

Tabla A2.3. Variables eliminadas por presencia de valores perdidos

Descripción		P48	p49	new_ila*
Descomposición variables con valores perdidos	> 0% y < 50%	27	27	41
	≥ 50% y < 90%	49	49	39
	≥ 90%	67	67	67
No. de variables eliminadas		**143**	**143**	**147**

Nota:
* La columna new_ila representa el número de variables eliminadas tanto para la variable new_ila numérico como categórico.
Fuente: ENEMDU 2021

Variables identificadoras y del diseño muestral

Luego, se procede a eliminar las variables identificadoras (id_upm, id_viv, id_hogar, id_per, upm, conglomerado, vivienda, hogar, persona, dominio, zona de planificación, y las variables relacionadas con el diseño muestral (estrato anual, factor de expansión anual, panel de muestreo). Estas variables se eliminan ya que su función principal es establecer la unicidad de las observaciones y permitir la aplicación de técnicas

[69] El proceso de depuración es individual; sin embargo, las acciones realizadas fueron aplicadas a cada pregunta de imputación (p48, p49 y new_ila) razón por la que se presentan de forma generalizada. De existir acciones concretas para una pregunta específica se menciona como un apartado individual.

estadísticas para obtener resultados a nivel poblacional. Los detalles de esta eliminación se presentan en la Tabla A2.4.

Tabla A2.4. Variables eliminadas por ser identificadoras o del diseño muestral

Variable	Descripción
conglomerado	Conglomerado
cod_inf	Código del informante
plan_muestreo	Estrato anual
fexp	Factor de expansión anual
upm	Unidad primaria de muestreo
panelm	Panel de muestreo
vivienda	Vivienda
hogar	Hogar
id_upm	Identificador de la unidad primaria de muestreo
id_viv	Identificador de la vivienda
id_hogar	Identificador del hogar
id_per	Identificador de la persona
zdp	Zonas de planificación
dominio	Dominio de análisis
Total de variables eliminadas	**14**

Fuente: ENEMDU 2021

Variables con varianza cero (único valor - constante)

Una vez eliminadas las variables innecesarias, se realiza un análisis para identificar aquellas columnas que contienen un valor constante. Este proceso tiene como objetivo eliminar las variables que no presentan variabilidad (varianza cero) y, por lo tanto, no aportan poder predictivo. En la Tabla A2.5 se detallan las variables que han sido depuradas en cada una de las preguntas de interés.

Tabla A2.5. Variables eliminadas por ser de varianza cero por pregunta de imputación

Variable	Descripción	P48	p49	New_lla*
p47a	Tamaño del establecimiento	x	x	-
Ano	Año	x	x	x
__000004	Temporal de apoyo	x	x	x
__000005	Temporal de apoyo	x	x	x
Pet	Población en Edad de Trabajar	x	x	x
Pei	Población Económicamente Inactiva	x	x	x
Pea	Población Económicamente Activa	x	x	x
Empleo	Población con Empleo	x	x	x
des_ab	Desempleo abierto	x	x	x
des_oc	Desempleo oculto	x	x	x
Desempleo	Población sin Empleo	x	x	x
n_r	Empleo no remunerado	x	x	x
n_c	Empleo no clasificado	x	x	x
den_tba_basica	Tasa bruta de asistencia a Educación Básica (ratio den)	x	x	x
den_tba_prepa	Tasa bruta de asistencia a E. Básica Preparatoria y Elemental (ratio den)	x	x	x
den_tba_bmedia	Tasa bruta de asistencia a E. Básica Media (ratio den)	x	x	x

den_tba_bsuperior	Tasa bruta de asistencia a E. Básica Superior (ratio den)	x	x	x
den_tba_primaria	Tasa bruta de asistencia a Educación Primaria (ratio den)	x	x	x
Vinc	Empleado Público-Privado	x	x	-
Total de variables eliminadas		**19**	**19**	**17**

Nota:
* La columna new_ila representa el número de variables disponibles tanto para el ila numérico como categórico
Fuente: ENEMDU 2021

Variables de indicadores educativos

Los indicadores educativos generados en la base de datos son eliminados debido a su carácter restrictivo al abarcar solo ciertos grupos de edad. En su lugar, se prioriza el uso de variables más generales, como el nivel de instrucción. Las variables educativas eliminadas se detallan en la Tabla A2.6.

Tabla A2.6. Variables eliminadas por ser indicadores educativos

Variable	Descripción
num_tba_basica	Tasa bruta de asistencia a Educación Básica (ratio num)
num_tba_bach	Tasa bruta de asistencia a Bachillerato (ratio num)
num_tba_superior	Tasa bruta de asistencia a Educación Superior (ratio num)
num_tba_prepa	Tasa bruta de asistencia a E. Básica Preparatoria y Elemental (ratio num)
num_tba_bmedia	Tasa bruta de asistencia a E. Básica Media (ratio num)
num_tba_bsuperior	Tasa bruta de asistencia a E. Básica Superior (ratio num)
num_tba_primaria	Tasa bruta de asistencia a Educación Primaria (ratio num)
num_tba_secundaria	Tasa bruta de asistencia a Educación Secundaria (ratio num)
den_tba_bach	Tasa bruta de asistencia a Bachillerato (ratio den)
den_tba_superior	Tasa bruta de asistencia a Educación Superior (ratio den)
den_tba_secundaria	Tasa bruta de asistencia a Educación Secundaria (ratio den)
Total de variables eliminadas	**11**

Fuente: ENEMDU 2021

Variables poco informativas

En la Tabla A2.7 se muestran las variables que se ha determinado que no contribuyen al proceso de predicción en el contexto de la investigación. Estas variables incluyen aquellas que son de apoyo en la clasificación de la condición de actividad de la población, o que son referenciales para los periodos de análisis, entre otras.

Tabla A2.7. Variables eliminadas por ser poco informativas

Variable	Descripción
Adec	Empleo adecuado
Analfa	Tasa de analfabetismo (15 años o más)
ced01a	Tiene cédula de identidad
condact	Condición de actividad agregada
condact_n	Nueva condición de actividad
condact_n1	Recodificación de la nueva condición de actividad
d_d	Deseo y disponibilidad de trabajar horas adicionales
fecha	Fecha
fecha1	Fecha1
Ipc	Índice de precios al consumidor
lindigencia	Línea de extrema pobreza

lpobreza	Línea de pobreza
mes	Mes
otro_i	Otro empleo inadecuado
p01	Código de persona
p07	Asiste a clases
p15aa	Donde nació
p20	Trabajó la semana pasada
p71a	Recibió ingresos derivados del capital
p72a	Recibe jubilación o pensiones
p73a	Recibió regalos o donaciones
p74a	Recibió dinero del exterior
p75	Recibió el bono de desarrollo humano
p77	Recibió el bono por discapacidad
periodo	Periodo
rn	Región natural
sbu	salario básico unificado
sub_h	Subempleo por insuficiencia de tiempo de trabajo
sub_w	Subempleo por insuficiencia de ingresos
t	Umbral para las horas trabajadas
Total de variables eliminadas	**30**

Fuente: ENEMDU 2021

Variables con información redundante

Se depuran aquellas variables con información similar, tomando como criterio de selección el nivel de desagregación y cardinalidad (número de categorias) de la variable evaluada. Se mantienen las variables con alta desagregación y con un número de categorías maximo de 21, las variables eliminadas se presentan en la Tabla A2.8.

Tabla A2.8. Variables eliminadas por tener información redundante

Variable	Descripción
afiliados	Tipo de seguro
c_ocupa	Recodificación categoría de ocupación
ciudad	Ciudad
crama_ciiu4	Rama de actividad
etnia	Recodificación p15 (cómo se considera)
g_edad	Grupos de edad
ingrl	Ingreso laboral
niv_inst	Nivel de instrucción
p05a	Seguro social - alternativa 1
p05b	Seguro social - alternativa 2
p40	Rama de actividad
p41	Grupo de ocupación
prov	Provincias
rama1d_ciiu4	Rama de actividad ciiu4 a 1 dígito
rama2d_ciiu4	Rama de actividad ciiu4 a 2 dígitos
rama3d_ciiu4	Rama de actividad ciiu4 a 3 dígitos
rama4d_ciiu4	Rama de actividad ciiu4 a 4 dígitos
seguro	Seguridad social
Total de variables eliminadas	**18**

Fuente: ENEMDU 2021

Depuraciones adicionales

Se eliminan también a aquellas variables construidas a partir de las variables que serán imputadas, ya que pueden afectar a las predicciones realizadas, la Tabla A2.9 las resume.

Tabla A2.9. Variables eliminadas por depuraciones adicionales

Variable	Descripción	P48	p49	ila - cat	ila-con
secemp	Sectorización del empleo		x		
vinc	Condición de actividad de los empleado (Públicos-Privados)			x	x
W	Umbral de ingresos laborales	x	x		

Fuente: ENEMDU 2021

Variables que entran a modelamiento

Finalizado el proceso de depuración, de las 262 variables disponibles en la ENEMDU anual 2021 efectivas para modelo, bajo lo expuesto en este Anexo, se mantienen alrededor de 28 variables. La Tabla A2.10 resume las variables resultantes del proceso de depuración.

Tabla A2.10. Variables resultantes del proceso de depuración

Variables	Descripción	Tipo de variable	p48	p49	new_ila*
area	Área	Categórica	x	x	x
p02	Sexo	Categórica	x	x	x
p03	Edad	Numérica	x	x	x
p04	Relación de parentesco	Categórica	x	x	x
p06	Estado civil	Categórica	x	x	x
p10a	Nivel de instrucción	Categórica	x	x	x
p15	Etnia	Categórica	x	x	x
p24	Horas efectivas de trabajo en la ocupación principal	Numérica	x	x	x
p42	Categoría de ocupación	Categórica	x	x	x
p45	Experiencia	Numérica	x	x	x
p46	Sitio de trabajo	Categórica	x	x	x
p47a	Tamaño del establecimiento	Categórica	-	-	x
p47b	Número personas trabajan en el establecimiento	Numérica	x	x	-
p48	Tenencia de libros contables en el lugar de trabajo	Categórica	-	x	-
p49	Tenencia de RUC en el lugar de trabajo	Categórica	x	-	-
p50	Número de trabajos	Categórica	x	x	x
p51a	Horas habituales de trabajo principal	Numérica	x	x	x
p61b1	Afiliación a la seguridad social	Categórica	x	x	x
nnivins	Nivel de instrucción	Categórica	x	x	x
grupo1	Grupo de ocupación ciuo8 (población ocupada de 15 años y más)	Categórica	x	x	x
rama_ciiu4	Rama de actividad ciiu4 a 1 dígito (población ocupada de 15 años y más)	Categórica	x	x	x
prov_n	Provincia	Categórica	x	x	x
ila	Ingreso laboral	Numérica	x	x	-
hsize	Tamaño del hogar	Numérica	x	x	x
horas	Horas de trabajo semanal	Numérica	x	x	x
escol	Años promedio de escolaridad	Numérica	x	x	x

secemp	Sectorización del empleo	Categórica	x	-	x
Total de variables para modelado			**25**	**24**	**23**

* La columna new_ila representa el número de variables disponibles tanto para el las variables new_ila numérico como categórico

Fuente: ENEMDU 2021

Anexo 3: Proceso de imputación ingreso laboral – new_ila (numérico) y p48

Inicialmente, en esta investigación se consideraron cuatro variables como candidatas para la imputación (p48, p49, p61b1 e ila). Sin embargo, al revisar el número de observaciones sujetas a imputación, se determinó que la variable p61b1 tenía un número muy reducido de observaciones, por lo que fue descartada. Además, se determinó que la variable p48 no tenía una influencia significativa en los indicadores laborales, por lo que también se descartó.

En cuanto a la variable del ingreso agregado, se optó por utilizar la variable new_ila y se transformó en una variable categórica. Sin embargo, en este anexo se presenta el proceso de imputación para las variables p48 y new_ila (numérica), las cuales no fueron incluidas en la sección de metodología, pero que tenían valores que podían ser objeto de imputación.

Generación de modelos de predicción a partir de algoritmos de ML

En primer lugar, es importante mencionar que tanto los problemas de regresión (variables numéricas) como de clasificación (variables categóricas) utilizan algoritmos de predicción que comparten similitudes. La principal diferencia entre ellos radica en las métricas de evaluación utilizadas para medir la calidad de las predicciones obtenidas. En este sentido, se utilizaron dos métricas específicas para evaluar la calidad de los modelos generados. Para la variable categórica de interés, p48, se empleará la métrica *accuracy*. Por otro lado, para el agregado del ingreso numérico (new_ila), se utilizará la métrica de la raíz del error cuadrático medio (RSME).

Una vez definidas las métricas de evaluación, se procedió a generar los primeros modelos de predicción siguiendo el proceso de dos etapas propuesto en la metodología. En primer lugar, se seleccionó el mejor algoritmo entre cuatro candidatos (RF, XGBoost, MLP y SVM), los cuales fueron sometidos a pre-procesamiento, validación cruzada y búsqueda de hiperparámetros. Luego, se extrajeron las principales métricas de evaluación y se realizaron comparaciones. Los resultados se presentan en la Tabla A3.1.

Tabla A3.1. Métricas de evaluación obtenidas por cada algoritmo considerado y para cada pregunta de imputación

Pregunta	Métrica	SVM	XGBoost	RF	MLP
new_lla numérico	RMSE	$1.140	$923	$924	$1.369
p48	Accuracy	82,19%	82,4%	81,0%	81,4%

Al igual que en la sección de metodología, se concluye que el algoritmo XGBoost es el más idóneo para la generación de predicciones (menor RMSE en new_ila y mayor *accuracy* en p48). En ese sentido, el modelo final utiliza los lineamientos mencionados en la metodología para pre-procesamiento, se agrupan a las categorías poco

representativas y se eliminan las variables con varianza cercana a cero. Adicionalmente, se establece la validación cruzada estratificada (k = 10) y se seleccionan los hiperparámetros (optimización bayesiana). Las variables que entraran a modelado son el resultado de un análisis exploratorio y depuración de la fuente de información.

Tabla A3.2. Métricas de evaluación[70] obtenidas.

Pregunta	Métricas	Valores
p48	*Accuracy*	81,7%
	Kappa	64,1%
	F1-score	75,5%
new_ila numérico	RMSE	$ 943.09
	MAE	$209,79

La Tabla A3.3 complementa lo presentado para la variable p48, al desagregar las métricas de evaluación por categoría.

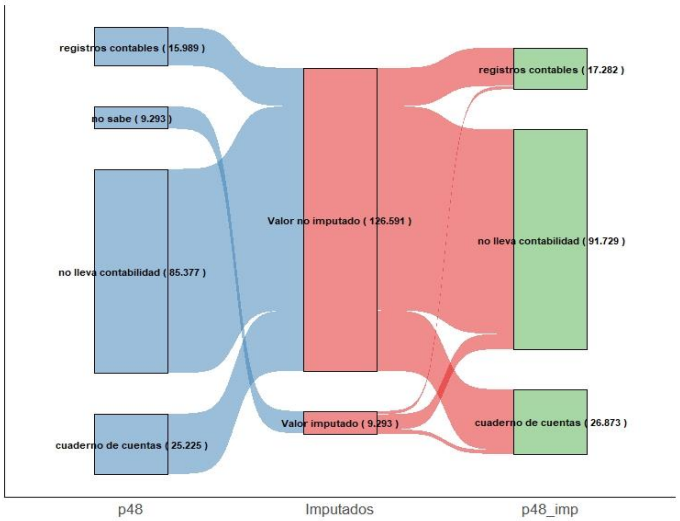
Tabla A3.3. Métricas de evaluación obtenidas por cada categoría de la variable p48

Pregunta	Categoría	*Accuracy*	Kappa	*F1 - score*
p48	registros contables	94,2%	76,9%	80,3%
	cuaderno de cuentas	82,7%	45,6%	56,4%
	no lleva contabilidad	28,5%	23,2%	1,6%

Una vez evaluados y definidos los modelos, se procede a la generación de las predicciones (imputaciones) sobre las variables de interés. En la Figura A3.1 se presenta un gráfico de Sankey que ilustra las observaciones imputadas de la variable p48. Se observa que de las 9.293 observaciones inicialmente clasificadas en la categoría "no sabe", después de la imputación, 6.352 observaciones (68,4%) se clasificaron como "no lleva contabilidad"; 1.648 observaciones (17,7%) se clasificaron como "cuaderno de cuentas"; y 1.293 observaciones (13,9%) se clasificaron como "lleva contabilidad".

[70] Las métricas RMSE y MAE representan la diferencia entre el valor predicho y real (valores más pequeños se asocian a mejores resultados). Por otro lado, el R cuadrado hace alusión el ajuste del modelo con respecto a la variable de interés (ila) y puede tomar valores entre 0 a 1, se prefiere valores cercanos a 1.

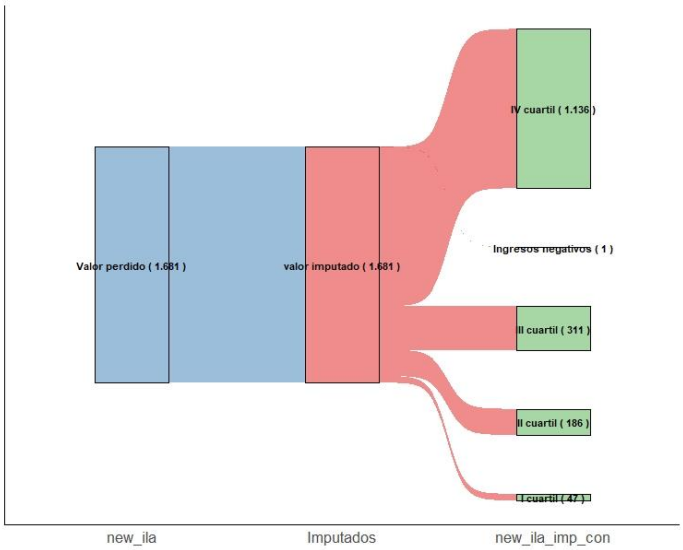
Figura A3.1. Flujo de valores imputados de la pregunta sobre registros contables (datos muestrales)



Fuente: ENEMDU 2021

Por último, la variable new_ila numérica presenta un comportamiento similar al obtenido por su contraparte categórica, como se aprecia en la Figura A3.2[71].

Figura A3.2. Flujo de valores imputados para la variable new_ila numérica (datos muestrales)



Nota: Para facilidad de visualización se categorizó a la variable ila en cuartiles
Fuente: ENEMDU 2021

La mayoría de las observaciones imputadas corresponden al cuarto cuartil de ingresos, con valores que van desde más de $594 hasta $113,100, totalizando 1.136 observaciones. De las observaciones restantes, 311 tienen valores en el tercer cuartil, que van desde más de $350 hasta $594, mientras que 186 se encuentran en el segundo cuartil, con valores que van desde más de $150 hasta $350. Además, 47 observaciones están en el primer cuartil, con valores que van desde $0 hasta $150, y solo una observación tiene un valor negativo (menos de $0).

[71] Se agrupa la información en cuatro cuartiles, Ingresos negativos y valores perdidos para una mejor visualización de la transición de las observaciones imputados.

Anexo 4. Estimaciones de los datos originales y datos imputados de los indicadores del sector en el empleo

Tabla A4.1. Estadísticos resultantes de los indicadores de sectorización del empleo (original – imputada)

Sector del empleo		Datos						
		Tasa	error	CV	LI	LS	Deff	Rango limites
Formal (original)	NACIONAL*	42,9%	0,0	1,13	42,0%	43,9%	15,4	1,9%
	Urbano*	53,5%	0,0	0,9	52,5%	54,5%	10,6	2,0%
	Rural*	23,8%	0,0	3,2	22,3%	25,3%	18,1	3,0%
Formal (imputada)	NACIONAL*	46,8%	0,0	1,1	45,8%	47,8%	15,9	1,9%
	Urbano*	57,5%	0,0	0,9	56,6%	58,5%	10,2	1,9%
	Rural*	27,5%	0,0	3,0	25,9%	29,1%	19,5	3,2%
Informal (original)	NACIONAL*	49,6%	0,0	1,0	48,6%	50,5%	16,3	2,0%
	Urbano	38,7%	0,0	1,3	37,8%	39,7%	10,6	1,9%
	Rural	69,0%	0,0	1,2	67,4%	70,7%	19,9	3,4%
Informal (imputada)	NACIONAL*	51,0%	0,0	1,0	50,0%	52,0%	16,4	2,0%
	Urbano	39,8%	0,0	1,2	38,8%	40,7%	10,5	1,9%
	Rural	71,3%	0,0	1,2	69,7%	73,0%	19,8	3,3%
Doméstico (original)	NACIONAL	2,2%	0,0	3,6	2,0%	2,3%	4,6	0,3%
	Urbano	2,7%	0,0	3,8	2,5%	2,9%	4,2	0,4%
	Rural	1,2%	0,0	9,	1,0%	1,4%	5,8	0,4%
Doméstico (imputada)	NACIONAL	2,2%	0,0	3,6	2,0%	2,3%	4,6	0,3%
	Urbano	2,7%	0,0	3,8	2,5%	2,9%	4,2	0,4%
	Rural	1,2%	0,0	9,1	1,0%	1,4%	5,8	0,4%
No clasificado en el sector (original)	NACIONAL*	5,4%	0,0	2,8	5,1%	5,7%	7,2	0,6%
	Urbano*	5,0%	0,0	3,3	4,7%	5,4%	5,9	0,7%
	Rural*	6,0%	0,0	5,0	5,4%	6,5%	9,1	1,2%
No clasificado en el sector (imputada)	NACIONAL*	0,0%	0,0	.	.	.	0,0	.
	Urbano*	0,0%	0,0	.	.	.	0,0	.
	Rural*	0,0%	0,0	.	.	.	0,0	.

Nota:
Tasa: indicador estimado
Error: error estándar del indicador
CV: coeficiente de variación del indicador, el cual debe ser ≤ a 15 para que una cifra sea representativa
LI y LS: intervalos de confianza inferior y superior del indicador
Deff: Efecto de diseño de la muestra, que comprende un ajuste utilizado definir el tamaño de la muestra de la encuesta
Rango límites: Diferencia entre el LS y el LI
***** Diferencia estadísticamente significativa
Fuente: ENEMDU 2021

Tabla A4.2. Prueba de hipótesis de la diferencia de los indicadores de sectorización del empleo antes y después de imputación

Sector del empleo		Diferencia Tasas	Prueba de hipótesis (p-value)
Formal	NACIONAL*	3,9%	0,0
	Urbano*	4,0%	0,0
	Rural*	3,7%	0,0
Informal	NACIONAL*	1,5%	0,0
	Urbano	1,1%	0,1
	Rural	2,3%	0,1
Doméstico	NACIONAL	0,0%	1,0
	Urbano	0,0%	1,0

	Rural	0,0%	1,0
No clasificado en el sector	**NACIONAL***	-5,4%	0,0
	Urbano*	-5,0%	0,0
	Rural*	-6,0%	0,0

Nota:
***** Diferencia estadísticamente significativa
Fuente: ENEMDU 2021

Anexo 5. Estimaciones de los indicadores desprendidos de la condición de actividad de las personas con empleo antes y después de la imputación

Tabla A5.1. Estadísticos resultantes de los indicadores de laborales (previo a imputación)

Condición de actividad del empleo		**Datos**						
		Tasa	**error**	**CV**	**LI**	**LS**	**Deff**	**Rango limites**
Adecuado (original)	**NACIONAL**	32,5%	0,0	1,2	31,8%	33,3%	10,8	1,5%
	Urbano	39,8%	0,0	1,0	39,0%	40,6%	7,7	1,6%
	Rural	18,7%	0,0	3,1	17,6%	19,8%	12,9	2,3%
Adecuado (imputado)	**NACIONAL**	33,1%	0,0	1,1	32,4%	33,9%	11,1	1,5%
	Urbano	40,7%	0,0	1,0	39,9%	41,6%	8,0	1,6%
	Rural	18,8%	0,0	3,1	17,7%	20,0%	13,0	2,3%
Inadecuado (original)	**NACIONAL**	61,5%	0,0	0,7	60,7%	62,4%	13,0	1,7%
	Urbano	52,3%	0,0	0,9	51,4%	53,2%	9,5	1,8%
	Rural	78,9%	0,0	0,8	77,6%	80,1%	13,4	2,4%
Inadecuado (imputado)	**NACIONAL**	61,6%	0,0	0,7	60,8%	62,5%	12,9	1,7%
	Urbano	52,5%	0,0	0,9	51,6%	53,4%	9,4	1,8%
	Rural	78,9%	0,0	0,9	77,7%	80,1%	13,4	2,4%
Subempleo (original)	**NACIONAL**	23,2%	0,0	1,2	22,7%	23,8%	7,3	1,1%
	Urbano	22,8%	0,0	1,3	22,3%	23,4%	5,6	1,2%
	Rural	24,0%	0,0	2,4	22,9%	25,1%	10,3	2,2%
Subempleo (imputado)	**NACIONAL**	23,2%	0,0	1,2	22,7%	23,8%	7,3	1,1%
	Urbano	22,8%	0,0	1,3	22,3%	23,4%	5,6	1,2%
	Rural	24,0%	0,0	2,4	22,9%	25,1%	10,3	2,2%
Subempleo por horas (original)	**NACIONAL**	20,5%	0,0	1,3	20,0%	21,1%	7,2	1,0%
	Urbano	20,4%	0,0	1,3	19,9%	21,0%	5,1	1,1%
	Rural	20,8%	0,0	2,7	19,7%	21,8%	12,0	2,2%
Subempleo por horas (imputado)	**NACIONAL**	20,5%	0,0	1,3	20,0%	21,1%	7,2	1,0%
	Urbano	20,4%	0,0	1,3	19,9%	21,0%	5,1	1,1%
	Rural	20,8%	0,0	2,7	19,7%	21,8%	11,1	2,2%
Subempleo por ingresos (original)	**NACIONAL**	2,7%	0,0	3,3	2,5%	2,9%	5,2	0,4%
	Urbano	2,4%	0,0	4,1	2,2%	2,6%	4,5	0,4%
	Rural	3,2%	0,0	5,4	2,9%	3,6%	5,8	0,7%
Subempleo por ingresos (imputado)	**NACIONAL**	2,7%	0,0	3,3	2,5%	2,9%	5,2	0,4%
	Urbano	2,4%	0,0	4,1	2,2%	2,6%	4,5	0,4%
	Rural	3,2%	0,0	5,4	2,9%	3,6%	5,8	0,7%
Otro empleo no pleno (original)	**NACIONAL**	27,2%	0,0	1,0	26,6%	27,7%	6,6	1,1%
	Urbano	24,0%	0,0	1,5	23,3%	24,7%	8,4	1,5%
	Rural	33,1%	0,0	1,2	32,3%	33,9%	4,0	1,5%
Otro empleo no pleno (imputado)	**NACIONAL**	27,3%	0,0	1,0	26,7%	27,8%	6,6	1,1%
	Urbano	24,2%	0,0	1,5	23,4%	24,9%	8,3	1,5%
	Rural	33,1%	0,0	1,2	32,4%	33,9%	4,0	1,5%
Empleo no remunerado (original)	**NACIONAL**	11,1%	0,0	2,5	10,6%	11,7%	13,7	1,1%
	Urbano	5,5%	0,0	3,2	5,1%	5,8%	6,7	0,7%
	Rural	21,8%	0,0	2,8	20,6%	23,0%	12,6	2,4%

Empleo no remunerado (imputado)	**NACIONAL**	11,1%	0,0	2,5	10,6%	11,7%	13,7	1,1%
	Urbano	5,5%	0,0	3,2	5,1%	5,8%	6,7	0,7%
	Rural	21,8%	0,0	2,8	20,6%	23,0%	12,6	2,4%
Empleo No clasificado (imputado)	**NACIONAL***	0,5%	0,0	5,9	0,4%	0,5%	4,3	0,1%
	Urbano*	0,7%	0,0	6,2	0,6%	0,7%	4,5	0,2%
	Rural*	0,1%	0,0	19,3	0,1%	0,1%	2,8	0,1%
Empleo No clasificado (imputado)	**NACIONAL***	0,0%	0,0	.	.	.	0,0	.
	Urbano*	0,0%	0,0	.	.	.	0,0	.
	Rural*	0,0%	0,0	.	.	.	0,0	.

Nota:
Tasa: indicador estimado
Error: error estándar del indicador
CV: coeficiente de variación del indicador, el cual debe ser ≤ a 15 para que una cifra sea representativa
LI y LS: intervalos de confianza inferior y superior del indicador
Deff: Efecto de diseño de la muestra, que comprende un ajuste utilizado definir el tamaño de la muestra de la encuesta
Rango límites: Diferencia entre el LS y el LI
***** Diferencia estadísticamente significativa
Fuente: ENEMDU 2021

Tabla A5.2. Prueba de hipótesis de la diferencia de los indicadores de laborales antes y después de imputación

Clasificación del empleo		**Diferencia**	**Prueba de hipótesis (p-value)**
Adecuado	**NACIONAL**	0,6%	0,3
	Urbano	0,9%	0,1
	Rural	0,1%	0,9
Inadecuado	**NACIONAL**	0,1%	0,9
	Urbano	0,2%	0,8
	Rural	0,1%	1,0
Subempleo	**NACIONAL**	0,0%	1,0
	Urbano	0,0%	1,0
	Rural	0,0%	1,0
Subempleo por horas	**NACIONAL**	0,0%	1,0
	Urbano	0,0%	1,0
	Rural	0,0%	1,0
Subempleo por ingresos	**NACIONAL**	0,0%	1,0
	Urbano	0,0%	1,0
	Rural	0,0%	1,0
Otro empleo inadecuado	**NACIONAL**	0,1%	0,8
	Urbano	0,1%	0,8
	Rural	0,0%	1,0
Empleo no remunerado	**NACIONAL**	0,0%	1,0
	Urbano	0,0%	1,0
	Rural	0,0%	1,0
Empleo No clasificado	**NACIONAL***	-0,5%	0,0
	Urbano*	-0,7%	0,0
	Rural*	-0,1%	0,0

Nota:
***** Diferencia estadísticamente significativa
Fuente: ENEMDU 2021

Anexo 6. Modelos adicionales

Tabla A6.1. *Accuracy* ENEMDU mensual (octubre) 2021 y IV trimestre 2021

ENEMDU	Periodo	Variable	*Accuracy*
Mensual	Octubre 2021	p48	84,5%
		p49	89,4%
		new_ila categórico	84,2%
Trimestral	IV trimestre 2021	p48	84,2%
		p49	88,9%
		new_ila categórico	84,2%

Fuente: ENEMDU octubre 2021 y IV trimestre 2021

INEC | Buenas cifras, **mejores vidas**

@ecuadorencifras

@ecuadorencifras

@InecEcuador

t.me/ecuadorencifras

INEC/Ecuador

INECEcuador

Administración Central (Quito)
Juan Larrea N15-36 y José Riofrío,
Teléfonos: (02) 2544 326 - 2544 561 Fax: (02) 2509 836
Código postal: 170410
correo-e: inec@inec.gob.ec

www.ecuadorencifras.gob.ec