www.ecuadorencifras.gob.ec

LIBRO
METODOLÓGICO

IMPUTACIÓN TALLA MENORES DE 2 AÑOS

ENSANUT 2018

INEC | Buenas cifras, mejores vidas

Metodología de imputación de talla para menores de 2 años en ENSANUT 2018

En el marco de la implementación de la primera Encuesta Nacional sobre Desnutrición Infantil (ENDI), el Instituto Nacional de Estadística y Censos (INEC) organizó mesas técnicas con expertos de diferentes agencias internacionales. Estas mesas trataron temas como comparabilidad entre las operaciones estadísticas en temas de desnutrición, la calidad de las cifras y mediciones antropométricas y el diagnóstico de las mediciones antropométricas de la información levantada por el INEC en encuestas desde 2006 hasta 2023. Como resultado de estas mesas técnicas se concluyó que la información de tallas en el año 2018 no cumpliría con las recomendaciones emitidas en 2019 sobre todo en aquellos menores de dos años, al existir un porcentaje no despreciable de poblaciones por encima del estándar de crecimiento de Organización Mundial de la Salud (OMS); y que posiblemente la cifra de Desnutrición Crónica Infantil (DCI) para este grupo etario en el 2018 se encuentra sobreestimada. A partir de estos resultados, el INEC se propuso desarrollar un ajuste a nivel de tasa con la asistencia técnica de expertos. Un primer estudio agregado, utilizando información histórica, informó que el valor más probable de DCI para menores de dos años en 2018 es de 23,6%. Así, la finalidad de este estudio fue explorar diversas técnicas de imputación para generar una nueva variable de talla en la base de datos en menores de dos años para el año 2018. Para esto, se exploraron métodos de regresión, imputación múltiple y aprendizaje automático. Los resultados obtenidos corroboran el análisis agregado determinando un valor más probable de prevalencia de la DCI en el año 2018 es de 23,6% para menores de 2 años. De los métodos explorados el modelo de aprendizaje automático XGBoost evidenció las mejores propiedades en cuanto a sus predicciones, evidenciando validez interna y externa al predecir niveles de DCI.

Nota técnica imputación de talla para menores de 2 años ENSAUT 2018

Coordinación
Coordinación General Técnica de Innovación en Métricas y Análisis de la Información

Dirección/Departamento
Dirección de Innovación en Métricas y Metodologías

Elaborado por:
Danilo Vera
Santiago Valdivieso
Galo Egas G.

Revisado por:
Galo Egas G.
Director de Innovación en Métricas y Metodologías

Aprobado por:
Darío Velez
Coordinador General de Innovación en Métricas y Análisis de la Información

Contenido

Índice de Tablas

Índice de Figuras

Índice de Anexos

Agradecimientos

El Instituto Nacional de Estadística y Censos (INEC) extiende su gratitud al Grupo del Banco Mundial por su invaluable aporte en el desarrollo e implementación de metodologías de inferencia estadística y predicción, que han sido fundamentales tanto en el desarrollo conceptual de las metodologías como en la disponibilidad de recursos tecnológicos para el entrenamiento de modelos. Estos esfuerzos han conducido a la imputación de tallas para menores de dos años en ENSANUT 2018 y consecuente la estimación de la prevalencia de la Desnutrición Crónica Infantil en el marco de la coherencia y consistencia interna e histórica del fenómeno.

Igualmente, el INEC expresa su agradecimiento al Fondo de las Naciones Unidas para la Infancia (UNICEF) y a la Organización Panamericana de la Salud (OPS) por su invaluable apoyo en brindar asistencia técnica especializada. Esta colaboración fue fundamental en la actualización de la metodología empleada en la evaluación de la calidad de las mediciones antropométricas de talla con base en los nuevos estándares establecidos en el año 2019.

Por último, los autores desean expresar su agradecimiento a Mónica Pozo, quien ha realizado valiosas aportaciones y contribuciones tanto a la definición de los parámetros de análisis como a la estructuración del presente estudio.

Ratificando su compromiso con la producción de estadísticas de la más alta calidad, que permitan a los responsables de políticas y a la ciudadanía contar con información confiable y oportuna, el INEC continuará trabajando estrechamente con el Grupo Conjunto de Estimaciones de Desnutrición (JME, por sus siglas en inglés) para la profundización, el desarrollo y la actualización de los estándares de calidad en encuestas de nutrición a nivel mundial.

1. Introducción

En 2019, la Organización Mundial de la Salud y el Fondo de las Naciones Unidas para la Infancia OMS (OMS; UNICEF, 2019) publicaron un conjunto de recomendaciones sobre el levantamiento, evaluación e interpretación de datos antropométricos. Entre otros aportes, estas ofrecen estándares de calidad con los que las operaciones estadísticas pueden contrastar sus resultados y determinar si reflejan correctamente la realidad que desean medir.

En reunión del 10 de marzo de 2023, el Gobierno del Ecuador, el Sistema de Naciones Unidas y el Banco Mundial acordaron organizar mesas técnicas con la participación del INEC y de expertos de las agencias internacionales para la revisión del diseño muestral, análisis de los datos y comparabilidad de las estimaciones de la Encuesta de Desnutrición Infantil (ENDI).

Una primera mesa técnica revisó el diseño muestral de la ENDI y analizó su comparabilidad con la de la Encuesta Nacional de Salud y Nutrición del 2018 (ENSANUT). La misión integrada por CEPAL, UNICEF y Banco Mundial concluyó que el diseño muestral de la ENDI no presenta ningún tipo de sesgo y se recomendó replicar el diseño muestral de ENDI en ENSANUT para asegurar comparabilidad. Al realizar este ejercicio, el INEC concluyó que el efecto del diseño muestral de ENSANUT es mínimo; por lo que, no afecta la comparabilidad de las series (INEC, 2023a).

Por otro lado, una segunda mesa técnica revisó la calidad de los datos antropométricos de las encuestas de nutrición en el Ecuador sobre la base de los nuevos estándares de calidad de 2019 y el análisis de determinantes de la desnutrición crónica infantil (DCI) que realizó el INEC. Para tal efecto, especialistas de OPS y UNICEF revisaron toda la documentación relacionada con ENDI y ENSANUT 18, metodologías de cálculo y resultados obtenidos por el INEC.

Los resultados de esta misión concluyeron que la calidad de las mediciones antropométricas de la ENDI es de alta calidad, no presenta problemas en cuanto a casos faltantes, distribución etaria, del dígito decimal y de número entero de pesos y estaturas, de puntajes Z o de desviaciones estándar. Mientras que, los indicadores de calidad apuntarían a que la prevalencia de la DCI en grupos etarios de 0 a 23 meses en la ENSANUT 2018 estaría sobrestimada y se sugirió utilizar las recomendaciones de OMS y UNICEF (2019) para el análisis de las mediciones antropométricas de las encuestas.

En los meses de agosto y septiembre de 2023, el INEC realizó un diagnóstico de calidad de los datos antropométricos de talla en el que concluyó que, en concordancia con UNICEF, la ENSANUT 2018 presenta indicadores elevados en menores de dos años. Particularmente, el exceso de probabilidad en la cola derecha observado es implausible, pues implicaría un crecimiento por encima de una población adecuadamente nutrida. En este sentido, se desarrolló un ajuste a nivel de tasa con la asistencia técnica del Banco Mundial, utilizando información histórica y se estimó que el valor más probable de DCI para menores de dos años en 2018 es de 23,6% (INEC, 2023b).

No obstante, existe la necesidad de conocer la tasa de DCI a niveles subnacionales por propósitos de planificación nacional, desarrollo de política pública y con fines de investigación. En este sentido, se contó nuevamente con la asistencia técnica del Banco Mundial en el estudio de diversas técnicas de imputación para generar

una nueva variable de talla en la base de datos en menores de dos años. Se exploraron métodos de regresión con variantes bayesianas, cuantílicas y jerárquicas, métodos de imputación múltiple con regresión bayesiana y Predictive Mean Matching (PMM), así como un método de aprendizaje automático (XGBoost). De todos los métodos explorados, este último se seleccionó para generar las imputaciones de talla en virtud de que evidenció las mejores propiedades en sus estimaciones.

2. Tratamiento de los datos

Para la estimación de modelos que puedan predecir adecuadamente la talla y especialmente considerando los métodos de aprendizaje supervisado, se consideró pertinente construir un pool de datos con la mayor cantidad de información disponible. En la presente sección se describe el procedimiento realizado desde la identificación de las fuentes de información, homologación, hasta contar con una base de datos que pueda ser trabajada en el proceso de modelado.

2.1. Diagnóstico de las fuentes de información y homologación de variables

La fuente de datos para el proceso de modelado consolida la información de seis operaciones estadísticas: ECV-2006, ENSANUT-2012, ECV-2014, ENSANUT-2018 y ENDI-2022. Estas contemplan la temática de la nutrición, medidas antropométricas, condiciones de vida y permiten generar estimaciones de desnutrición crónica infantil en menores de 5 años. Contar con información suficiente para el entrenamiento de los distintos modelos consolidará su capacidad predictiva y evitará que se incurra en problemas de dimensionalidad. La Tabla 1 resume la cantidad de observaciones por encuesta:

Tabla 1. Cantidad de observaciones por fuente

Encuesta	Observaciones	
	Menores de 2 años	Menores de 5 años
ECV-2006	2312	6068
ENSANUT-2012	4029	8653
ECV-2014	4136	11231
ENSANUT-2018	7034*	18714
ENDI-2023	7993	21531
Total	**25504**	**66197**

Fuentes: ECV 2006, ENSANUT 2012, ECV 2014, ENSANUT 2018, ENDI 2023.
* El grupo objetivo consta de 7034 menores de 2 años sobre los que se va a realizar la imputación de la talla, luego del proceso de modelado.
Elaboración: DINME - INEC

El proceso de diagnóstico inicia con la definición de las variables que son relevantes para el fenómeno de estudio. Al respecto, UNICEF en el 2011, menciona lo siguiente:

> La desnutrición infantil es el resultado de la ingesta insuficiente de alimentos (en cantidad y calidad), la falta de una atención adecuada y la aparición de enfermedades infecciosas. Detrás de estas causas inmediatas, hay otras subyacentes como son la falta de acceso a los alimentos, la falta de atención sanitaria, la utilización de sistemas de agua y saneamiento insalubres, y las prácticas deficientes de cuidado y alimentación. En el origen de todo ello están las causas básicas que incluyen factores sociales, económicos y políticos como la pobreza, la desigualdad o una escasa educación de las madres. Un niño que sufre desnutrición crónica presenta un retraso en su crecimiento. Indica una carencia de los nutrientes necesarios durante un tiempo prolongado, por lo que aumenta el riesgo de que contraiga enfermedades y afecta al desarrollo físico e intelectual del niño. (UNICEF, 2011)

En este sentido, resulta importante explorar diversos grupos de variables que se relacionan con el problema de la DCI de acuerdo con las causas inmediatas, subyacentes y básicas. No obstante, la necesidad de construir un pool de datos con una cantidad suficiente de individuos para evitar problemas de dimensionalidad implica que se dejen de lado algunas variables que pueden considerarse explicativas y predictivas de la DCI, pero, que no están disponibles en todas las

fuentes de información o no son comparables entre fuentes. Para el ejercicio se consideran los siguientes componentes[1]:

- La DCI puede verse influenciada por la ubicación geográfica de los menores; es así como, la prevalencia varía significativamente entre provincias o regiones (Aguaysa, 2023) y, esta es mayor en los niños que se encuentran en al área rural (Albuja, 2022). Se consideran las **variables geográficas** como: área, región y provincia.
- La presencia de DCI es estructural y se relaciona con las condiciones de pobreza (aunque no son predictores perfectos), por tanto, se incluyen **características del hogar/vivienda** como: tipo de vivienda, vía de acceso principal, ocupación de la vivienda, material del piso, paredes y techo; agua, de donde proviene y como la consume; energía eléctrica; servicio de recolección de basura; combustible para cocinar; número de cuartos/dormitorios; e ingreso del hogar e ingreso per cápita. Además, a nivel macro los determinantes incluyen la pobreza estructural (Aguaysa, 2023), por lo cual, se consideran las **Dimensiones NBI**[2]: dependencia económica, acceso a escolaridad, materiales de vivienda deficientes, servicios básicos inadecuados y hacinamiento, para lo cual es necesario conocer el tamaño del hogar.
- En lo referente a las características de la madre o jefe de hogar, se espera que un perfil educativo, socio-cultural o hereditario permitan comprender el comportamiento de la DCI en menores de 5 años. Por esta razón fueron consideradas **variables del jefe de hogar y madre del menor** como: edad, pea, escolaridad, título de educación superior, Autoidentificación étnica, seguro, ocupación y estado civil; adicional en el caso del jefe de hogar el sexo y en el caso particular de las madres se obtiene su estado de embarazo.
- Como características que se asocian directamente a la unidad de análisis se encuentran las **variables del niño/a**, en donde se identifica si el padre o madre viven en el hogar, sexo, edad, etnia[3], lactancia exclusiva, vacunas (bcg, pentavalente, opv, srp), carné o libreta integral, registro del peso al nacer y variables antropométricas como: peso, talla, zlen, dci.

Como paso siguiente, se revisaron los formularios de todas las encuestas para identificar la similitud en preguntas, opciones de respuesta y flujos. Una vez que se han identificado las variables, se ha realizado tanto en formulación como en las opciones de respuesta; juntamente con aquellas que requerirían algún proceso de tratamiento o comprensión del flujo lógico para llegar a ser comparables. Adicionalmente, para nutrir el set de variables disponible se calculó algunos indicadores como: dimensiones NBI, ingresos o escolaridad.

Tomando en consideración que la unidad de análisis para el presente estudio son los menores de 5 años, las variables identificadas debido a su naturaleza fueron agrupadas en 6 grupos de variables (Anexo 1), donde además se incluyen variables adicionales necesarias para la identificación de los individuos en la base apilada y declaración del diseño muestral como: unidad primaria de muestreo (upm), estrato, factor de expansión (fexp) e identificador del individuo dentro de la base (id).

[1] Algunas variables importantes como: bajo peso al nacer, talla de la madre, anemia no forman parte del presente estudio debido a diferentes motivos: cantidad importante de valores perdidos (bajo peso al nacer), no disponibilidad de información en una encuesta completa (talla de la madre, anemia), dificultad en la homologación, entre otras.

[2] Pobreza por Necesidades Básicas Insatisfechas: https://www.ecuadorencifras.gob.ec/pobreza-por-necesidades-basicas-insatisfechas/

[3] En el caso de los menores de la base ECV-2006 se imputa la etnia del jefe de hogar.

Una vez que se han identificado las variables que comparten características similares entre las operaciones estadísticas el paso siguiente consiste en realizar el proceso de homologación sobre estas variables para que puedan ser almacenadas en una sola base de datos apilada.

En este sentido, el proceso de homologación de variables garantiza la comparabilidad entre las preguntas de las distintas fuentes de información en cuanto a categorías y población objetivo. Para ello, la formulación de la pregunta debe ser afín, el flujo lógico no debe excluir individuos o poblaciones diferentes entre las encuestas y las opciones de respuesta deben ser equivalentes. Para este proceso se tomó como guía o base pívot la base ENDI 2023; por lo tanto, las variables y categorías antes de este periodo se ajustaron a esta base de referencia. Se hace especial énfasis en las categorías de las preguntas dado que la mayor parte de las variables son categóricas (ordinales o nominales). La cantidad reducida de variables numéricas (discretas o continuas) guardan las mismas definiciones en cuanto a unidades de medida para las distintas encuestas.

Un punto fundamental en el procesamiento de cualquier base de datos es el tratamiento de valores perdidos (NAs), los cuales, en función a la naturaleza de la pregunta pueden presentarse por pérdida de información al momento del levantamiento o por el flujo lógico de llenado del cuestionario (Medina & Galván, 2007). En el primer caso, se prefiere hacer uso de registros completos[4]; es decir, se omiten los registros que no cuentan con información en las variables homologadas, principalmente en aquellas que se relacionan directamente con los menores de 5 años: talla, peso, edad, entre otras. En el segundo caso, se considera el NA como una categoría adicional.

Asimismo, la construcción de indicadores permite generar nuevas variables en función de la información disponible con la finalidad de simplificar la información provista por las variables o representar componentes adicionales para contextos de interés. Ejemplos de indicadores son las dimensiones de la pobreza por NBI, ingresos, pea o la escolaridad.

Finalmente, se considera como primer insumo la base apilada obtenida luego del proceso de identificación de preguntas, construcción de indicadores (NBI, pea, ingresos), homologación de variables y tratamiento de valores perdidos. Este proceso da como resultado una base de 63046 observaciones y 64 variables sin NAs[5] para el proceso de modelado y 10 variables adicionales (objetivo, de identificación y diseño muestral).

2.2. Transformación de la base de datos (factores)

Los diferentes modelos requieren capturar información adicional de las variables predictoras, especialmente en el modelo de aprendizaje automático. Por ello, se realizan transformaciones sobre las 64 variables presentes en el **primer insumo** las cuales dependen del tipo de variable: dicotómica, categórica o numérica.

- En las variables dicotómicas no se realiza transformaciones, solamente se verifica que tomen valores de 0 o 1.

[4] En este caso del total de 7034 menores de 2 años solamente se considera 6773 menores que cuentan con información completa en las variables homologadas y a quienes es factible imputar una talla. Adicionalmente, un análisis de datos atípicos descarta 34 observaciones (Anexo 4), por lo que se limita la imputación de la talla a un total de 6739 menores de 2 años.

[5] A excepción de los NAs en las variables dcronica_5 y dcronica_2, los cuales se presentan cuando |zlen| > 6 y en el caso específico de la dcronica_2 para los niños que tienen 2 o más años.

- En las variables categóricas se hace uso de binarización, donde a cada categoría se representa mediante una variable dummy (0 o 1).
- En las variables numéricas se realizan transformaciones polinómicas (grados 2 y 3) y transformaciones sinusoidales (seno y coseno).

Adicional a las transformaciones realizadas, y con la finalidad de aportar más información a los modelos de aprendizaje automático (XGBoost o ADA-ENET), se realiza una combinación por parejas de un subconjunto[6] priorizado de variables dicotómicas y categóricas: *region, v_tipo_vivienda, v_techo, v_pared, v_piso, v_agua_recibe, v_agua_sum, v_servicio_hog, v_energia, v_ocupacion, year_c, mad_parto, mad_lugar_parto, mad_etnia, v_via_acceso, area, jefe_sexo, jefe_pea, n_sexo, nbi_pob, n_bcg, n_pentavalente, n_opv, n_srp, mad_pea*.

Consecuentemente, se genera un segundo insumo dado por la base resultante de la transformación y combinación de variables. Esta base consta de 63046 observaciones y 4106 variable para el proceso de modelado, la cual además contienen variables de identificación, diseño muestral y variables objetivo (ver Anexos 1 y 4).

[6] Se prioriza un subconjunto de variables para generar una base de datos de una dimensión adecuada para el proceso de entrenamiento esta priorización se la realiza con el apoyo de expertos.

3. Proceso de modelado

3.1. Estrategia metodológica

Existen diversas formas de imputación según la literatura, estas varían en cuanto a sus supuestos, sofisticación y su empleo puede depender de la razón detrás de la necesidad de completar la información. Es así como, habitualmente las operaciones estadísticas pueden contener información faltante debido a la no respuesta por fatiga, desconocimiento de la información solicitada o rechazo de participar en la investigación por parte del informante (Medina & Galván, 2007); en estos casos y suponiendo que se cumple el supuesto de que los datos omitidos se encuentran distribuidos de manera aleatoria (Moneta, et al., 2022), la falta de información es parcial por lo que se emplean comúnmente métodos de análisis de datos completos (listwise), análisis de datos disponibles (pairwise), imputación por medias no condicionadas, imputación por medias condicionadas por métodos de regresión, etc.

Por otro lado, del diagnóstico de calidad de los datos antropométricos en el que se analizó, entre otras encuestas, a la ENSANUT (INEC, 2023b), se evidenció que los datos de talla han sido observados con un exceso de varianza, por lo que, todas las mediciones individuales no corresponden a la realidad, a pesar de que distribucionalmente están centrados alrededor del verdadero parámetro poblacional.

Adicionalmente, considerando que no se cuenta con información del mismo periodo o fuente[7] para hacer una imputación parcial, los métodos a utilizarse girarán alrededor de la construcción de uno o varios modelos que permitan inferir una variable de interés sobre un conjunto de variables explicativas y no alrededor de niveles dados o vecinos cercanos intra-periodo.

Se toma como punto de partida el resultado del ejercicio previo en el que se estimó que el valor más probable de DCI para menores de dos años en 2018 es de 23.6%, y se explorarán los siguientes métodos: (1) imputación única con métodos de regresión lineal, (2) imputación múltiple y (3) predicción con métodos de aprendizaje automático.

Para la selección de variables a incluir en los primeros dos métodos se especificó un modelo lineal con una red elástica adaptativa (ADA-ENET, por su nombre en inglés). Con él, se identificaron las variables que mejor explican la varianza de la talla para cada uno de los grupos etarios de estudio. De entre estas últimas, destaca el peso, que en todos los modelos tiene el mayor poder predictivo. Un listado de las variables utilizadas en cada en estos dos primeros métodos se expone en el Anexo 1.

Por otro lado, para el método de predicción con aprendizaje automático, en un primer momento se incluyeron todas las variables de la base descrita en el apartado anterior. Como ejercicio complementario, se mantuvieron solamente las cien variables con mayor poder explicativo (un listado de estas se expone en el Anexo 1). En este caso, el peso también fue la variable con mayor influencia en la determinación de la talla.

Cabe señalar que en todos los métodos se utilizó la base apilada que resultó del proceso descrito en el apartado anterior. Para obtener los pesos de las variables

[7] Se asume que todas las mediciones de ENSANUT 2018 se observaron con ruido en los tres grupos etarios por lo que la imputación de realiza sobre la totalidad del conjunto de tallas.

independientes, se utilizaron las observaciones de la ECV 2006, ENSANUT 2012, ECV 2014 y ENDI 2023. Los datos de la talla de la ENSANUT 2018 se consideraron como perdidos.

En el proceso de modelado existieron dos limitaciones comunes a todos los modelos. La primera refiere a la temporalidad. Dado que no hubo disponibilidad de bases con información cercana a 2018, no fue posible entrenar los modelos con información de ese periodo. Esto obliga a asumir que las relaciones encontradas entre 2006 y 2023 se mantuvieron en 2018. No obstante, cabe mencionar que, dado el número de observaciones de las bases de 2023 y 2014 (las más cercanas al año de interés), éstas tuvieron mayor peso.

La segunda se refiere a la relación funcional entre el peso y la talla. Esta primera variable fue la que tuvo mayor impacto sobre las predicciones de todos los modelos, encontrándose una relación positiva con la talla. No obstante, los modelos parecen ser incapaces de identificar una forma funcional útil para identificar casos de sobrepeso y obesidad, que se miden como una relación entre el peso y la talla. Esto ya que, a este tipo de casos, los modelos les asignan una talla lo suficientemente alta como para escapar la asignación de estas condiciones. Por esta razón, la talla que se publica puede no conservar los patrones reales de los niños con estos cuadros de malnutrición.

3.2. Métodos de Imputación

En la presente sección se describen los métodos explorados para generar una imputación de talla para ENSANUT 2018 en menores de dos años.

3.2.1. Métodos de Regresión

Predice la variable dependiente como función lineal de un conjunto de variables explicativas. En este documento se utilizó el método de mínimos cuadrados ordinarios para definir la línea de la regresión. Este método consiste en minimizar la distancia entre los valores predichos por la regresión y los observados. Existen varios tipos de regresiones lineales que difieren en términos de la complejidad de las relaciones que identifican entre la variable dependiente y las explicativas. A continuación, se indican los tipos de regresión lineal que se exploraron para el presente ejercicio:

- **Estándar:** se imputa la media condicional de la distribución de la variable dependiente dadas las variables independientes. Además, asume que la relación entre la variable dependiente y las independientes es la misma para distintos grupos poblacionales.
- **Jerárquica (modelo mixto):** sigue un proceso similar a la estándar, pero permite que la relación lineal entre la variable dependiente e independiente difiera en términos de intercepto y pendiente para distintos grupos poblacionales. En nuestro caso, se especificó que la relación entre la talla y el peso pueda diferir entre áreas (urbana/rural), etnias y niveles de instrucción de las madres.
- **Cuantílica:** imputa los cuantiles condicionales de la distribución de la variable dependiente dadas las variables independientes. Para este ejercicio, se realizó predicciones para los cuantiles 25, 50 y 75 de la talla.

3.2.2. Imputación múltiple con ecuaciones encadenadas

Una de las limitaciones de los modelos de regresión es que no logran capturar la incertidumbre que involucra el proceso de imputación, fallando en replicar la variabilidad observada en los datos no imputados. Una técnica comúnmente utilizada para solventar este problema es la imputación múltiple con ecuaciones encadenadas, que consiste en imputar un valor varias veces a la misma observación introduciendo un componente aleatorio en el proceso. Esto resulta en varias bases de datos que difieren en los valores que toma la variable imputada. Posteriormente, se calcula de la desnutrición crónica infantil con cada una de las bases y luego se promedia los resultados obtenidos.

Existen varios métodos de imputación que discrepan en la forma en que se calculan/eligen los valores que se imputan. En este trabajo se exploraron dos de los métodos más utilizados en el campo:

- **Regresión lineal:** utiliza un método similar a la regresión lineal estándar, pero en lugar de imputar la media condicional de la distribución condicionada, imputa varias veces los valores obtenidos aleatoriamente de esta distribución.
- **Predictive Mean Matching**[8] **(PMM):** luego de calcular los valores predichos de una regresión lineal estándar, de la misma forma que en el método anterior, se encuentran las "k" observaciones más parecidas a la que se quiere imputar. En la literatura, a estas observaciones se las llama "vecinos más cercanos". A estos últimos los selecciona calculando la diferencia entre la predicción de las observaciones no imputadas y la predicción de la observación a imputarse.
  Posteriormente, de entre los vecinos más cercanos, se elige aleatoriamente uno de ellos y se imputa el valor de la variable de interés que se predijo para esta observación. En este documento se muestran las imputaciones que resultan de utilizar 5 vecinos más cercanos. No obstante, los resultados son robustos a un cambio de estos parámetros (se obtienen resultados muy similares si se utiliza un k igual a 3 o a 10).

3.2.3. Métodos de aprendizaje automático

Dentro de la gama de algoritmos de aprendizaje automático (machine-learning) existentes destaca el modelo XGBoost (Chen & Guestrin, 2016) (eXtreme Gradient Boosting) el cual es un algoritmo que ha despertado gran interés, pues, aunque es relativamente reciente es considerado actualmente el estado del arte en algoritmos de aprendizaje automático por sus resultados (Espinosa, 2020). XGBoost es una implementación de árboles de decisión para el aumento de gradiente (gradient-boosting) que tiene tres componentes principales:

- **Función de pérdida:** la función de la función de pérdida es estimar cuál es la mejor manera de hacer predicciones del modelo con los datos proporcionados.
- **Aprendiz débil:** El alumno débil es aquel que clasifica los datos de manera tan deficiente en comparación con las conjeturas aleatorias. Los alumnos débiles son en su mayoría árboles de decisión.
- **Modelo aditivo:** es un proceso iterativo y secuencial en el que se agregan árboles de decisión paso a paso. Cada iteración debería reducir el valor de

[8] Predictive Mean Matching (PMM) es un método de imputación estadística ampliamente utilizado para valores faltantes, propuesto por primera vez por Donald B. Rubin en 1986 y R. J. A. Little en 1988.

la función de pérdida. Se agrega una cantidad fija de árboles o el entrenamiento se detiene una vez que la pérdida alcanza un nivel aceptable o ya no mejora en un conjunto de datos de validación externa.

En el boosting, los árboles se construyen secuencialmente de modo que cada árbol posterior tenga como objetivo reducir los errores del árbol anterior. Cada árbol aprende de sus predecesores y actualiza los errores residuales. Por lo tanto, el árbol que crezca a continuación en la secuencia aprenderá de una versión actualizada de los residuos.

La técnica del gradient-boosting consta de tres pasos:

1. Se define un modelo inicial F_0 para predecir la variable objetivo y. Este modelo estará asociado a un residual $(y - F_0)$
2. Se ajusta un nuevo modelo h_1 a los residuos del paso anterior.
3. Luego, F_0 y h_1 se combinan para obtener F_1, la versión mejorada de F_0. El error cuadrático medio de F_1 será menor que el de F_0:

$$F_1(x) < -F_0(x) + h_1(x)$$

Para mejorar el rendimiento de F_1, se modela los residuos de F_1 dando lugar a F_2:

$$F_2(x) < -F_1(x) + h_2(x)$$

Repitiendo este proceso m-veces, se logra minimizar los residuos tanto como sea posible:

$$F_m(x) < -F_{m-1}(x) + h_m(x)$$

Los modelos que forman el conjunto, se conocen como aprendices base, pueden provenir del mismo algoritmo de aprendizaje o de diferentes algoritmos de aprendizaje. En el boosting los árboles de decisión son considerados aprendices débiles los cuales individualmente tienen un sesgo alto y poder predictivo bajo, pero aporta información vital para la predicción final, creando en conjunto un aprendiz fuerte el cual reduce el sesgo y la varianza. Además, los aprendices (aditivos) no alteran las funciones creadas en los pasos anteriores si no que imparten información propia para reducir los errores (Analytics Vidhya, 2018).

Algunas de las características del modelo XGBoost se resumen en Analytics Vidhya (2018) y Shiksha (2023) que parten del trabajo de Chen y Guestrin, (2016):

- Manejo de datos dispersos: los valores faltantes o la transformación de variables como la binarización o interacciones hacen que los datos sean escasos, para lo cual XGBoost incorpora un algoritmo de búsqueda de división para manejar diferentes tipos de patrones.
- Estructura de bloques para aprendizaje en paralelo: para una computación más rápida, XGBoost puede utilizar varios núcleos en la CPU. Los datos se clasifican y almacenan en unidades de memoria llamadas bloques. A diferencia de otros algoritmos, esto permite reutilizar el diseño de los datos en iteraciones posteriores, en lugar de volver a calcularlos.
- Poda de árboles: XGBoost utiliza el parámetro max_depth según se especifica el criterio de parada para la división de la rama y comienza a podar los árboles hacia atrás. Este enfoque de profundidad mejora significativamente el rendimiento computacional.

- Validación cruzada: la implementación de XGBoost viene con un método de validación cruzada incorporado. Esto ayuda al algoritmo a evitar el sobreajuste cuando el conjunto de datos no es tan grande.
- Conciencia de caché y computación fuera del núcleo: XGBoost se ha diseñado teniendo en cuenta el uso óptimo del hardware. Debido a esta propiedad, el algoritmo funciona asignando memorias intermedias internas en cada paso y, por lo tanto, utiliza la caché de la manera más eficiente.
- No linealidad: XGBoost puede detectar y aprender de patrones de datos no lineales.

Un algoritmo complejo de aprendizaje automático como el XGBoost viene con una cantidad considerable de parámetros, por lo que los alcances del ajuste de parámetros también son altos (Shiksha, 2023). En este sentido, es necesario dimensionar apropiadamente los parámetros que guiarán a los árboles de decisión en cada paso. En este caso se contemplan 4 parámetros[9]: la tasa de aprendizaje, la profundidad del árbol de decisión, el submuestreo de registros y el submuestreo de columnas:

- **eta:** (Tasa de aprendizaje). El rango es de 0 a 1. Un valor bajo de eta se traduce en un modelo robusto al sobreajuste. En el caso del presente análisis de consideran tasas de aprendizaje de 0,08 hasta 0,12 con un salto de 0,01.
- **max_depth:** (Profundidad del árbol). El número máximo de nodos de bifurcación de los árboles de decisión usados en el entrenamiento. Una mayor profundidad puede devolver mejores resultados, pero puede resultar en sobreajuste. Se consideran los siguientes valores: 6 hasta 18 con un salto de 2.
- **subsample:** (Submuestra de registros). Proporción de submuestra de la instancia de entrenamiento. Establecerlo en 0,5 significa que XGBoost recopiló aleatoriamente la mitad de las instancias de datos para hacer crecer árboles y esto evitará el sobreajuste. El rango es de 0 a 1. Se consideran los siguientes valores: 0,2 hasta 0,5 con un salto de 0,1.
- **colsample_bytree:** (Submuestra de columnas) Proporción de submuestra de columnas al construir cada árbol. El rango es de 0 a 1. Se consideran los valores de 0,7 hasta 1,0 con un salto de 0,1.

Para realizar el recorrido de los parámetros se construye una grilla con todas las combinaciones posibles una vez que se han fijado los valores para cada parámetro. La grilla resultante de los parámetros establecidos indica una cantidad de 560 posibilidades.

Una vez que se ha descrito el planteamiento de los modelos XGBoost y se conoce que, para encontrar un modelo óptimo es necesario realizar el recorrido de una grilla de parámetros resulta imperante comprender cómo se aplicó este modelo a los insumos disponibles para obtener un resultado concluyente. Así, el procedimiento realizado en el proceso de modelado, parte de la base apilada y transformada (segundo insumo), donde se definen los escenarios para la ejecución de los modelos. La ejecución de los modelos consiste en recorrer la grilla de parámetros para luego escoger el modelo que ofrece mejores resultados.

[9] Existen más parámetros que pueden ser analizados, pero ello requiere un esfuerzo más grande al momento de recorrer la grilla, por lo cual se analizan los 4 parámetros más importantes. Adicionalmente, es necesario indicar parámetros propios de la tarea de aprendizaje, que está dada por los árboles de decisión, como el objective que indica el tipo de tarea de puntuación a realizar (reg:linear) y el nround: el número de iteraciones que se realizarán antes de detener el proceso de ajuste (se establece en 50).

Los escenarios comprenden la estrategia utilizada para la ejecución del modelo[10] a partir de la información disponible en la base de datos aplicada y transformada. Se contemplan 6 escenarios, descritos en la Tabla 2:

Tabla 2. Escenarios proceso de modelado XGBoost

Escenario	Set de entrenamiento	Set de aplicación
Escenario 1	Toda la información disponible: - ECV 2006, ENSANUT 2012, ECV 2014, ENDI 2023 (grupos 1-6) - ENSANUT 2018 (grupos 4-6)	ENSANUT 2018 (grupos 1-3)
Escenario 2	Información niños/as menores de 2 años: - ECV 2006, ENSANUT 2012, ECV 2014, ENDI 2023 (grupos 1-3)	ENSANUT 2018 (grupos 1-3)
Escenario 3	Información niños/as menores de 1 año: - ECV 2006, ENSANUT 2012, ECV 2014, ENDI 2023 (grupos 1 y 2)	ENSANUT 2018 (grupos 1 y 2)
Escenario 4	Información niños/as de 12 - 23 meses: - ECV 2006, ENSANUT 2012, ECV 2014, ENDI 2023 (grupo 3)	ENSANUT 2018 (grupo 3)
Escenario 5	Información niños/as de 0 - 5 meses: - ECV 2006, ENSANUT 2012, ECV 2014, ENDI 2023 (grupo 1)	ENSANUT 2018 (grupo 1)
Escenario 6	Información niños/as de 0 - 5 meses: - ECV 2006, ENSANUT 2012, ECV 2014, ENDI 2023 (grupo 2)	ENSANUT 2018 (grupo 2)

Nota: Los grupos de edad corresponde a los siguientes rangos: (1) 0-5 meses, (2) 6-11 meses, (3) 12-23 meses, (4) 24-35 meses, (5) 36-47 meses y (6) 48-59 meses.
Elaboración: DINME - INEC

Las variantes realizadas a estos escenarios consisten en realizar el entrenamiento separando a los menores de acuerdo con el sexo, principalmente en los escenarios 4-6 donde se distingue a cada grupo etario.

Luego del proceso de ejecución se revisan los resultados obtenidos en cada uno de los escenarios y se contrastan métricas tradicionales como rmse y accuracy. No obstante, es necesario considerar que la imputación a realizarse corresponde a individuos que forman parte de un diseño muestral, por tanto, deben cumplir con estimaciones desagregadas adherentes a este diseño. En este sentido, la elección de los mejores modelos debe atender criterios de consistencia que aseguren una correcta selección de las nuevas tallas en los menores de 2 años de la ENSANUT 2018, y eso se complementa con un criterio experto en la selección inteligente del mejor modelo. Una estrategia de selección de modelos permitiría combinar resultados de los diferentes escenarios sobre todo de los escenarios que entrena cada grupo por separado (escenarios 4, 5 y 6); además esto puede realizarse en los modelos resultantes de la variación de los escenarios por sexo.

Los resultados del modelo XGBoost condicionados a la combinación de parámetros y combinación de resultados evidencias que uno de los parámetros que ofrece un mayor poder predictivo es la tasa de aprendizaje (parámetro eta), dado que

[10] El proceso de modelado mediante XGBoost incorpora un método de validación cruzada que ayuda al algoritmo a evitar el sobreajuste. En ese caso es necesario definir como conjuntos de entrenamiento (train) y prueba (test), para este estudio se definen un 20% de los datos para el test.

genera resultados diferenciados en los diferentes valores. Por tal motivo este parámetro rigió el análisis de resultados y la selección de los mejores modelos.

El modelo XGBoost al ser un modelo de aprendizaje automático fue implementado con la mayor cantidad de variables (factores) posible[11] con el propósito de capturar la información no observable que los modelos de regresión posiblemente no son capaces de capturar. Entre las variables que aportan más información predictiva la talla fueron: sexo[12], peso, edad en días (del menor), edad de la madre, ingreso per cápita, escolaridad de la madre, escolaridad del jefe de hogar, edad del jefe de hogar e ingresos del hogar. Considerando que el peso, la edad en días y el sexo constituyen la base para la estimación de las curvas de crecimiento, resulta razonable esperar que una predicción condicionada al comportamiento de estas variables no se ajuste a una tendencia histórico en cuanto a la distribución de los datos, si no que se encuentre alineada a un comportamiento teórico; lo cual se ve reflejado en la cantidad de excesos en la cola derecha que predice estos modelos (0,05%).

Como parte del proceso de selección de modelos se realiza un proceso de selección de variables en función de la ganancia que representa cada variable en los mejores modelos. La ganancia implica la contribución relativa de la variable correspondiente al modelo calculado tomando la contribución de cada característica para cada árbol en el modelo. Un valor más alto de esta métrica en comparación con otra variable implica que es más importante para generar una predicción. Se selecciona aproximadamente 130 variables que representan una ganancia acumulada entre 75% y 95 %. Se vuelve a entrenar los modelos y generar predicciones, pero ahora con un modelo mucho más simple garantizando predicciones consistentes.

La naturaleza aleatoria y los métodos numéricos detrás del modelo XGBoost oculta en los algoritmos y funciones que permiten entrenar un determinado modelo hace que, aproximarse a un único resultado considerado óptimo puede convertirse en una tarea imposible ya que esto dependerá de la forma en que se exploren las posibles combinaciones. El recorrido de una grilla de parámetros es un método válido para encontrar parámetros óptimos[13] y seleccionar modelos. Si bien, es posible que exista algún modelo que no haya sido explorado y que no sólo tenga mejores capacidades predictivas, sino que cumpla con criterios históricos de distribución de tamaños; el presente ejercicio cumple su propósito al explorar una grilla amplia que explora de manera general la topología del problema, y por tanto, la predicción encontrada es adecuada para predecir las tasas de dci en los grupos siempre que se cumpla con criterios de consistencia histórica, por categorías de desagregaciones, y distribucionales.

[11] La cantidad de variables se encuentra condicionada a la capacidad tecnológica (Anexo 4).
[12] El sexo se incluye implícitamente como un predictor dado que la separación entre hombres y mujeres al momento de entrenar los modelos produjo mejores resultados.
[13] Un enfoque alternativo puede ser el de Optimización Bayesiana.

4. Discusión y selección del modelo óptimo

Para seleccionar el modelo más pertinente, es necesario evaluar tres aspectos de los datos: (1) su cercanía frente a las estimaciones realizadas a nivel agregado por el "estándar de oro", (2) el grado de ajuste de la distribución resultante de cada modelo con los estándares de la OMS y (3) la consistencia de las desagregaciones de la DCI respecto a la tendencia histórica.

Respecto al primer criterio, el "estándar de oro" es el método que se utilizó para predecir la prevalencia de DCI a nivel agregado (INEC, 2023b). En este último, se realizó un ajuste a los niveles de DCI por grupo etario utilizando una simulación de la distribución que se habría visto en la ENSANUT 2018 de no existir problemas de medición, así como un promedio con la tendencia histórica. Los supuestos detrás de estos métodos son acordes con la literatura y la evidencia empírica del mundo y del Ecuador. Las tasas obtenidas con este método se muestran en la Tabla 3.

Tabla 3. Resultados del estándar de oro

Grupo edad	DCI predicha	Límite Inferior	Límite Superior
Menores de 2	23,6%	20,9%	25,9%
0-5 meses	13,7%	11,1%	16,3%
6-11 meses	18,1%	15,4%	20,8%
12-23 meses	29,8%	27,1%	31,8%

Elaboración: DINME - INEC

Mientras mayor sea la cercanía de los resultados obtenidos por los modelos de ajuste a nivel de microdato con este estándar de oro, se juzgará que estos son de mayor calidad.

En cuanto al segundo criterio, los estándares de la OMS recomiendan que la distribución de la talla para la edad (ZLEN, por sus siglas en inglés) obtenida de la recolección de datos se parezca lo más posible a una distribución normal estándar con un sesgo de la media hacia la izquierda que refleja la DCI de un país. Así, se esperaría que la distribución obtenida tenga una desviación estándar cercana a 1 y que la concentración de las observaciones en la cola derecha de la distribución no sea superior a 2,3%. Es razonable esperar que una buena imputación de calidad genere una distribución con las mismas propiedades.

Finalmente, el tercer criterio utiliza las tendencias y niveles históricos para calibrar las estimaciones y asegurar que sean confiables. Este tipo de análisis es posible dada la naturaleza estructural del fenómeno de la DCI.

Para evaluar los modelos a la luz de los primeros dos criterios, en la Tabla 4 se exponen las tasas de DCI de los distintos métodos para los grupos etarios de menores de 2 años, 0-5 meses, 6-11 meses y 12-23 meses y su distancia frente al estándar de oro. También se exponen los excesos en la cola derecha de la distribución y la desviación estándar del z-score de la talla para la edad (ambos indicadores que se utilizan para evaluar la calidad de las mediciones de talla).

Analizando la distancia frente al estándar de oro, los modelos de XGBoost serían los mejores, existiendo diferencias cercanas a 0,1% por grupos y sin diferencia en la tasa agregada de menores de dos años. Los modelos de imputación múltiple con los métodos bayesiano y PMM le siguen con una subestimación de 1 p.p. y 2,1 p.p., respectivamente. No obstante, cuando se observan las diferencias a nivel de grupos

etarios, se aprecian diferencias que superan los 4,5 p.p. Vale mencionar que el caso de la imputación múltiple con regresión lineal bayesiana, a pesar de la cercanía en los menores de 2, está dada por una sobreestimación de los primeros dos grupos etarios y una subestimación del tercero. Los modelos de regresión lineal (imputación única), por su parte, otorgan subestimaciones especialmente altas que van desde 5,5 p.p. hasta 7,8 p.p. en menores de dos. Estas son aún más elevadas en algunos grupos etarios.

Tabla 4. Resultados de los métodos explorados

Técnica	Método	Grupo edad	DCI predicha	Distancia estándar de oro	Exceso +2DE z-score	Desviación estándar
Imputación única con métodos de regresión lineal	Estándar	Menores de 2	15,8%	-7,80%	1,5%	1,19
		0-5 meses	9,5%	-4,20%	3,3%	1,29
		6-11 meses	10,1%	-8,00%	0,0%	1,20
		12-23 meses	20,7%	-9,10%	1,5%	1,10
	Jerárquica	Menores de 2	16,9%	-6,70%	1,6%	1,37
		0-5 meses	10,7%	-3,00%	3,6%	1,71
		6-11 meses	12,4%	-5,70%	0,6%	1,17
		12-23 meses	21,2%	-8,60%	1,3%	1,30
	Cuantil	Menores de 2	18,1%	-5,50%	7,6%	1,86
		0-5 meses	11,6%	-2,10%	11,3%	1,88
		6-11 meses	14,9%	-3,20%	9,0%	1,78
		12-23 meses	21,9%	-7,90%	5,6%	1,83
Imputación múltiple	Regresión lineal bayesiana	Menores de 2	22,6%	-1,00%	4,2%	1,66
		0-5 meses	18,3%	4,60%	6,8%	1,85
		6-11 meses	20,2%	2,10%	2,7%	1,54
		12-23 meses	25,3%	-4,50%	3,3%	1,57
	Predictive Mean Matching	Menores de 2	20,5%	-3,10%	3,3%	1,61
		0-5 meses	15.4%	1,70%	5,7%	1,79
		6-11 meses	16.5%	-1,60%	2,5%	1,53
		12-23 meses	24,2%	-5,60%	2,9%	1,53
Aprendizaje automático	XGBoost	Menores de 2	23,6%	0,00%	0,1%	0,96
		0-5 meses	13,6%	-0,10%	0,3%	1,07
		6-11 meses	17,8%	-0,30%	0,0%	0,89
		12-23 meses	29,9%	0,10%	0,0%	0,88

Elaboración: DINME - INEC

Adicionalmente, cabe señalar que las predicciones de todos los modelos excepto el XGBoost implican un ajuste mayor en el grupo etario de 12-23 frente a los demás. Esto es inconsistente con el hecho de que los mayores problemas de medición habrían ocurrido en los grupos etarios más jóvenes, como se muestra en el estudio de la Calidad de los datos en las estimaciones de retraso en talla de las encuestas de nutrición infantil 2006 – 2023 (INEC, 2023b), especialmente en la ENSANUT 2018.

Respecto a la concentración de casos en la cola derecha de la distribución, todos los modelos, excepto el XGBoost, presentan excesos respecto al estándar de la OMS (porcentaje superior a 2,3% en el área de +2DE) en algún grupo etario. No obstante, excluyendo a la regresión cuantílica y los grupos de 0-5 meses de los modelos de imputación múltiple, estos excesos parecen aceptables con relación a lo que se ha observado en la tendencia histórica.

En el caso de las desviaciones estándar, el modelo XGBoost es que el más se acerca a 1. Sin embargo, es el único método que tiene un estadístico menor a uno. Este último hecho es inconsistente con la tendencia histórica de encuestas similares. La

desviación estándar promedio entre 2006 y 2014 es de 1,37 de 0 a 5 meses, 1,40 de 6-11 meses y de 1,57 de 12-23 meses. Considerando esto, los modelos de regresión lineal estándar y jerárquica son los que entregarían un estadístico más creíble, dadas las características del levantamiento del 2018. El resto de los modelos otorgan una desviación estándar demasiado alta en todos los grupos etarios.

Finalmente, con relación a la consistencia con la tendencia histórica de las desagregaciones geográficas y sociodemográficas, los modelos de XGBoost también presentan los mejores resultados. En cambio, los modelos basados en regresiones presentan el peor desempeño. En los modelos problemáticos, la totalidad de los casos atípicos ocurren por debajo del mínimo histórico. Así, para que el lector pueda observar esta problemática, en la tabla que se presenta en el Anexo 3 se comparan las desagregaciones es posible encontrar prevalencias atípicas frente al mínimo histórico de los años 2006-2023 (excluyendo 2018), lo cual corrobora el correcto comportamiento de las estimaciones del modelo XGBoost.

En resumen, el modelo de XGBoost es el que tiene mejor desempeño conjunto en los criterios 1 y 3. Es el que tiene mayor cercanía frente a los resultados del estándar de oro y, en términos de su consistencia de sus desagregaciones con la serie histórica, empata con el PMM, en tanto ninguno de los dos presenta datos atípicos. En cuanto a la concentración de observaciones en +2DE (parte del criterio 2), el XGBoost también presenta el mejor escenario (Anexo 3). No obstante, es importante considerar que su desviación estándar es particularmente baja en relación con encuestas similares. A pesar de esta última limitación, por el resto de las bondades que ofrece se considera que este modelo es el que mejor desempeño tiene de forma general, por lo que es el que se utilizará para la imputación oficial. A continuación, se exponen los resultados de este modelo con los respectivos intervalos de confianza.

Tabla 5. Tasas de DCI modelo XGBoost por grupo etario e intervalos de confianza

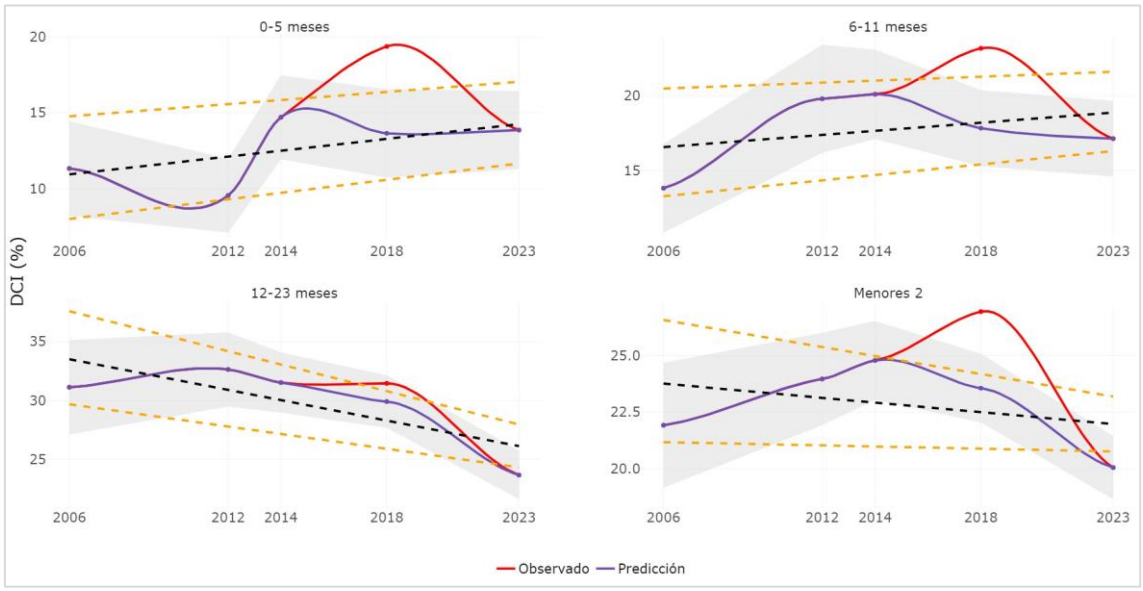
Método	Grupo edad	DCI predicha	Límite Inferior	Límite Superior
Estándar	Menores de 2	23,6%	21,8%	25,3%
	0-5 meses	13,6%	10,6%	16,7%
	6-11 meses	17,8%	15,0%	20,7%
	12-23 meses	29,9%	27,7%	32,1%

Elaboración: DINME - INEC

Adicionalmente, se incluye un gráfico que ilustra el cambio frente a la estimación original de las tasas de DCI en la ENSANUT 2018.

La Figura 1, representa cómo el mejor ajuste provisto por el modelo XGBoost guarda concordancia con la tendencia histórica con sus intervalos de confianza en los tres grupos etarios y en el agregado en menores de 2 años. Esta condición es más notoria en los grupos 1 y 2 (0-5 meses y 6-11 meses respectivamente), en donde se ve claramente como la estimación observada de la ENSANUT 2018 se encuentra por fuera del comportamiento esperable de la tendencia de la DCI. A pesar de que el impacto es menos evidente en el grupo 3 (12-23 meses) y que el intervalo de confianza de la estimación XGboost contiene a la estimación original, es pertinente observar que esta última no cae dentro del intervalo de confianza de la tendencia, lo que sugiere que la nueva tasa es un valor más confiable de DCI en este grupo etario.

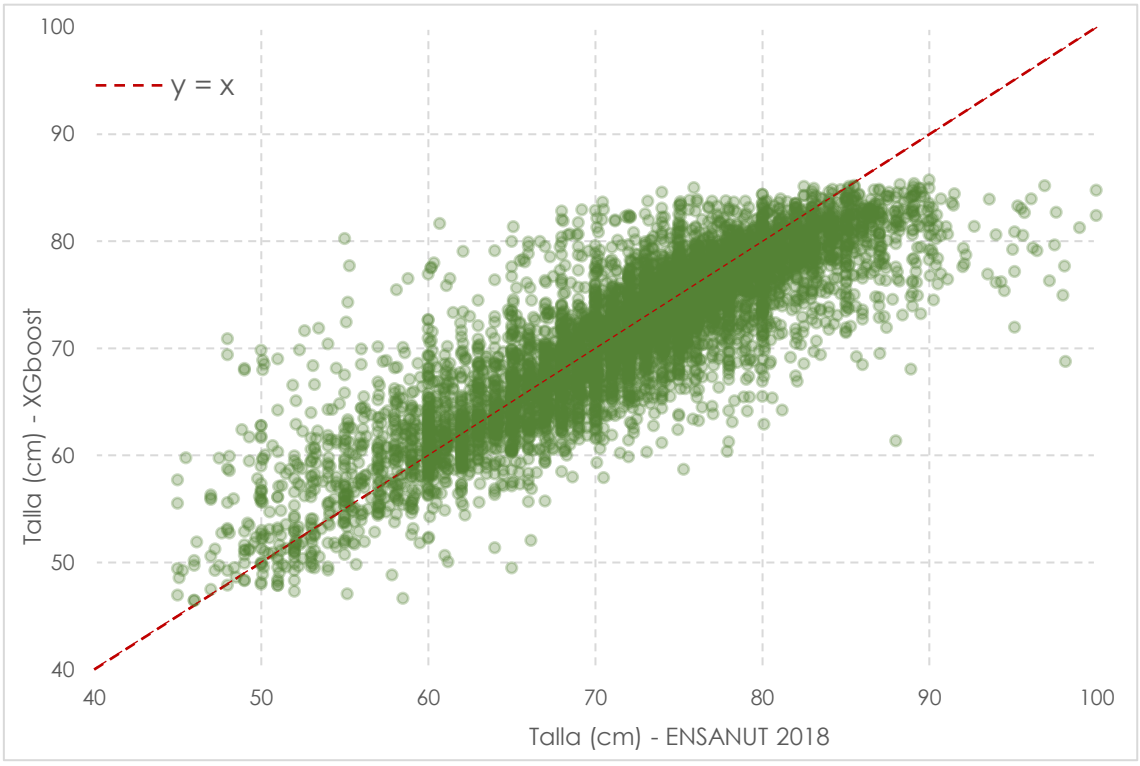
Figura 1. Tendencia 2006-2023, regresión lineal bayesiana, predicción XGBoost



Fuente: ECV 2006, ENSANUT 2012, ECV 2014, ENSANUT 2018, ENDI 2023, resultados XGBoost
Elaboración: DINME – INEC
Nota: El área sombreada representa el intervalo de confianza de la seria con la predicción XGBoost.

Asimismo, al realizar una asociación de las tallas anteriores vs las nuevas (XGBoost) mediante un diagrama de dispersión entre las dos tallas (Figura 2), se observa que el modelo XGboost muestra un comportamiento similar al de la talla observada con menor dispersión, esto subiere que el ajuste del modelo es adecuado y razonable para emular el comportamiento de la medición de talla a nivel de microdato.

Figura 2. Gráfico de dispersión talla ENSANUT 2018 vs predicción XGBoost



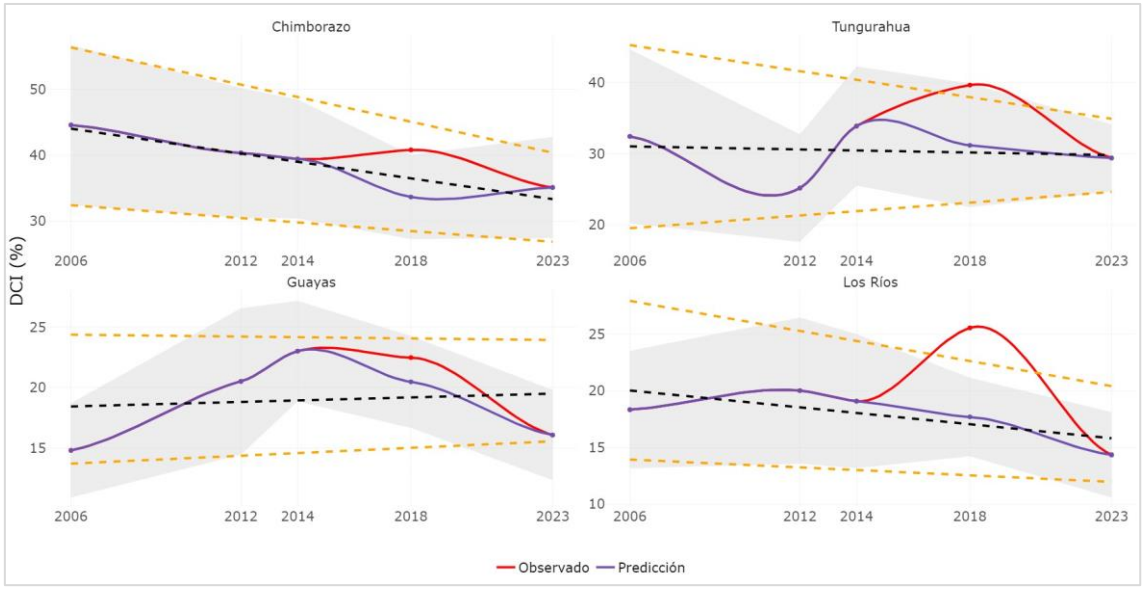
Fuente: ENSANUT 2018, resultados XGBoost
Elaboración: DINME - INEC

La naturaleza de un modelo implica una posible distorsión entre lo predicho vs lo observado, no obstante, como se puede observar en la Figura 2, el modelo no presenta un sesgo sistemático dentro del grupo de estudio. Esto pone en evidencia las bondades del método XGBoost además de las propiedades ya descritas en cuanto a distribución, desagregaciones y tendencia histórica en las cuales sobresale entre los métodos estudiados.

En la figura 3 se presentan algunos resultados a nivel de provincia que permiten ampliar la discusión sobre los resultados del modelo XGboost. En particular, para la provincia de Chimborazo, una de las provincias que históricamente presenta mayores niveles de DCI, el modelo predice una tasa de 33,7% que es estadísticamente indistinta que el nivel sugerido por la tendencia y el nivel de 2023, lo que sugiere que la DCI en Chimborazo no sufrió cambios entre el 2018 y 2013, conclusión similar a la obtenida con el dato observado.

Por su parte, para las provincias de Tungurahua y Los Ríos las estimaciones XGBoost predicen una tasa más razonable y cercana a la tendencia histórica, De igual manera, la estimación de DCI para Guayas mediante XGboost se acerca más a la media tendencial, por lo que se mejora la precisión de la estimación, esto a pesar de que no existe diferencia estadísticamente significativa entre la estimación original y la imputada.

Figura 3. DCI histórica provincias Chimborazo, Tungurahua, Guayas, Los Ríos



Fuente: ECV 2006, ENSANUT 2012, ECV 2014, ENSANUT 2018, ENDI 2023, resultados XGBoost
Elaboración: DINME – INEC
Nota: El área sombreada representa el intervalo de confianza de la seria con la predicción XGBoost.

Además, la robustez del modelo XGBoost se evidenció mediante un estudio de combinaciones y agregaciones de los distintos escenarios estudiados, donde pudo comprobar que la distribución de los datos entre el modelo elegido y el modelo agregado es estadísticamente igual (Anexo 5).

Finalmente, los distintos estudios sobre las técnicas utilizadas en la imputación de la talla, así como en los resultados obtenidos, han permitido determinar que el modelo XGBoost brinda un resultado consistente, no solo como técnica de imputación, si no como un modelo que replica de forma adecuada una tasa de DCI si cuenta con

información de buena calidad. En este sentido, si la calidad de la información que ingresa no es buena, el resultado normalmente tampoco es bueno. Esto se ha corroborado con información de la ENDI 2023 y ENSANUT 2018 mostrando un buen promedio de ajuste en tasa para modelos seleccionados en la ENDI mayor al que se observa en la ENSANUT 2018 (ver Anexo 6).

5. Conclusiones y Recomendaciones

5.1. Conclusiones

- Se utilizaron métodos de regresión con variantes bayesianas, cuantílicas y jerárquicas, se utilizó imputación múltiple con regresiones bayesianas y Predictive Mean Matching; y, métodos de aprendizaje automático a través de algoritmos de XGBoost. Los resultados corroboran el análisis realizado utilizando métodos de tendencia lineal y ajuste normal determinando un valor más probable de prevalencia de la DCI en la ENSANUT 2018, en este caso se ratifica que esta se encuentra en 23,6% para menores de 2 años.
- De los métodos que se exploraron el modelo XGBoost evidenció las mejores propiedades en cuanto a sus predicciones, evidenciando validez interna y externa al predecir niveles de DCI cercanos a lo observado en otros grupos etarios, desagregaciones y validez histórica en cuanto a tendencias.
- La talla obtenida por medio del método elegido presenta una debilidad al considerar el cuadro antropométrico completo que incluye el peso, por este motivo, se recomienda a los interesados tratar con cuidado el uso de indicadores distintos a la DCI, como desnutrición aguda o doble carga, en donde el modelo no logra replicar un comportamiento apegado al historial observado.

5.2. Recomendaciones

- Con la finalidad de evitar sesgos en la información que informe de manera erra la toma de decisiones, se recomienda utilizar la nueva talla imputada para la estimación de la DCI en los menores de 2 años para ENSANUT 2018 y utilizar los métodos estándar (recomendados en el documento de metodología y diseño muestral de la encuesta) para el cálculo de errores de estimación.
- Dependiendo de las necesidades de investigación y estimación, se recomienda a los usuarios tomar en consideración las tendencias y estimaciones de la DCI de los grupos de 2 a 5 años en donde no se han realizado modificaciones.

Bibliografía

Aguaysa, M. (2023). Desnutrición Crónica Infantil. Revisión Bibliográfica Sistemática. *Maestría en Nutrición y Dietética*. Quito: UDLA. Obtenido de https://dspace.udla.edu.ec/bitstream/33000/15442/1/UDLA-EC-TMND-2023-120.pdf

Albuja, W. (2022). Determinantes socioeconómicos de la desnutrición crónica en menores de cinco años: evidencia desde Ecuador. *Inter disciplina, 10*(28), 591-611. doi:10.22201/ceiich.24485705e.2022.28.83314.

Analytics Vidhya. (08 de Septiembre de 2018). *Algorithms*. Obtenido de XGBoost: Introduction to XGBoost Algorithm in Machine Learning: https://www.analyticsvidhya.com/blog/2018/09/an-end-to-end-guide-to-understand-the-math-behind-xgboost/

Chen, T., & Guestrin, C. (2016). XGBoost: A Scalable Tree Boosting System. *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data*, 785-794. Obtenido de https://dl.acm.org/doi/pdf/10.1145/2939672.2939785

Espinosa, J. (2020). Aplicación de algoritmos Random Forest y XGBoost en una base de solicitudes de tarjetas de crédito. *Ingeniería, investigación y tecnología, 21*(3), 1-16. Obtenido de https://www.scielo.org.mx/pdf/iit/v21n3/1405-7743-iit-21-03-00002.pdf

INEC. (2023a). *La Desnutrición Infantil: Cálculo del factor de expansión con base a la ENDI 2022 - ENSANUT 2018.* Quito-Ecuador.

INEC. (2023b). Calidad de los datos en las estimaciones de retraso en talla de las encuestas de nutrición infantil 2006-2023. (G. Egas, S. Valdivieso, & M. Pozo, Edits.) Quito: INEC. Obtenido de https://www.ecuadorencifras.gob.ec/documentos/web-inec/ENDI/Documento_calidad_mediciones_DCI_2006-2023.pdf

Medina, F., & Galván, M. (Julio de 2007). Imputación de datos: teoría y práctica. *Estudios estadísticos y prospectivos*. Santiago de Chile: CEPAL. Obtenido de https://repositorio.cepal.org/server/api/core/bitstreams/02dd479f-fae2-43c4-b5ec-5419fa7f6190/content

Moneta, A. M., Juárez, M. A., Camilo Caro, L., & Allub, M., (2022, Mayo). Una revisión del supuesto MAR en datos faltantes de la EPH. *Serie Documentos de Trabajo de Investigación de la Facultad de Ciencias Económicas (DTI-FCE)*. Universidad Nacional de Córdoba. Retrieved from https://revistas.unc.edu.ar/index.php/DTI/article/view/37746/37746

OMS; UNICEF. (2019). *Recomendaciones para la obtención de datos, el análisis y la elaboración de informes sobre indicadores antropométricos en niños.* Ginebra: Printed in Switzerland.

Shiksha. (9 de Agosto de 2023). *Shiksha Online*. Obtenido de XGBoost Algorithm in Machine Learning: https://www.shiksha.com/online-courses/articles/xgboost-algorithm-in-machine-learning/

UNICEF. (2011). *La Desnutrición Infantil. Causas, consecuencias y estrategias para su prevención y tratamiento.* Madrid: UNICEF España. Obtenido de https://www.salud.gob.ec/wp-content/uploads/2016/09/Dossierdesnutricion.pdf

Anexos

Anexo 1: Definición de variables

Tipo de variable	Variable	Descripción	C1	C2	C3	C4	C5	CH
Variables Dicotómicas	area	Área	2	2	2	2	2	2
	v_cocina	Combustible o energía para cocinar	4	6	4	5	6	2
	jefe_sexo	Sexo del jefe de hogar	2	2	2	2	2	2
	jefe_pea	PEA del jefe de hogar	2	2	2	2	2	2
	mad_pea	PEA de la madre del menor	2	2	2	2	2	2
	n_sexo	Sexo del menor de 5 años	2	2	2	2	2	2
	n_padre	El padre del menor vive en el hogar	2	2	2	2	2	2
	n_madre	La madre del menor vive en el hogar	2	2	2	2	2	2
	n_parentesco	Parentesco del niño con el jefe del hogar	13	9	13	9	10	2
	n_carne	Tiene carne o libreta integral de salud	2	3	2	3	3	2
	n_bcg	Tiene dosis BCG	2	2	2	2	2	2
	n_pentavalente	Tiene dosis PENTAVALENTE	2	2	2	2	2	2
	n_opv	Tiene dosis OPV	2	2	2	2	2	2
	n_srp	Tiene dosis SRP	2	2	2	2	2	2
	nbi_pob	Pobreza por Necesidad Básicas insatisfechas	2	2	2	2	2	2
	nbi_expob	Pobreza extrema por Necesidad Básicas insatisfechas	2	2	2	2	2	2
	nbi_depec	Dependencia económica	2	2	2	2	2	2
	nbi_acc_esco	Sin acceso a escolaridad	2	2	2	2	2	2
	nbi_matviv_def	Materiales deficientes de la vivienda	2	2	2	2	2	2
	nbi_ser_viv	Servicios básicos inadecuado	2	2	2	2	2	2
	nbi_hcam	Hacinamiento	2	2	2	2	2	2
Variables Categóricas	year_c	Año de levantamiento (categórica)	5	5	5	5	5	5
	prov	Provincia	21	24	25	25	25	25
	region	Región	3	4	4	4	3	3
	v_via_acceso	Tipo de vía	5	6	6	6	8	3
	v_tipo_vivienda	Tipo de vivienda	6	8	8	8	8	6
	v_techo	Material predominante del techo o cubierta de la vivienda	6	6	7	6	6	6
	v_pared	Material predominante de las paredes exteriores de la vivienda	7	7	8	7	8	7
	v_piso	Material predominante de piso de la vivienda	8	7	8	8	8	7
	v_agua_recibe	Agua que recibe la vivienda es	3	4	4	4	4	4
	v_agua_sum	De dónde recibe el agua principalmente este hogar	7	5	6	7	5	5
	v_servicio_hog	Servicio higiénico de la vivienda	5	5	5	5	7	4
	v_energia	Dispone la vivienda de la energía proviene de	5	5	5	4	2 + 5	3
	v_basura	Como elimina la basura de la vivienda en este hogar	5	6	6	6	7	4
	v_ocupación	Tenencia de la vivienda	7	7	6	7	6	5
	jefe_etnia	Etnia referente al jefe de hogar	6	8	8	8	8	5

	jefe_estado_civil	Estado civil/conyugal del jefe de hogar	6	6	6	7	6	6
	jefe_seguro	Aporta actualmente el jefe de hogar	6	4 + 6	7	6	7	6
	jefe_ocupacion	Ocupación del jefe de hogar	15	9	17	9	9	8
	jefe_titulo	Título de educación superior del jefe de hogar	3	3	3	3	3	3
	jefe_inst	Nivel de instrucción del jefe de hogar	9	10	11	10	13	3
	mad_inst	Nivel de instrucción de la madre	9	10	11	10	13	3
	mad_titulo	Título de educación superior de la madre	3	3	3	3	3	3
	mad_etnia	Etnia referente a la madre del menor	6	8	8	8	8	5
	mad_estado_civil	Estado civil/conyugal de la madre del menor	6	6	6	7	6	6
	mad_seguro	Aporta actualmente la madre del menor	6	4 + 6	7	6	7	6
	mad_ocupacion	Ocupación de la madre del menor	15	9	17	9	9	8
	mad_lugar_parto	Lugar del parto	12	15	14	10	10	5
	mad_profesional	Persona o profesional que atendió el parto	7	7	8	9	9	5
	mad_parto	Tipo de parto	3	3	3	2	2	2
	mad_embarazada	Embarazo reportado por la madre	3	3	3	3	3	3
	mad_peso_nacer	Registro del peso al nacer del menor	3	3	3	3	3	3
	n_etnia	Etnia del menor de 5 años	6	8	8	8	8	5
Variables Numéricas	v_cuartos	Números de cuartos que dispone el hogar						
	v_dormitorio	Números de cuartos para dormir						
	jefe_edad	Edad en años cumplidos del jefe de hogar						
	ingresoh	Ingreso del hogar						
	ingreso_pc	Ingreso per-cápita						
	mad_edad	Edad en años cumplidos de la madre del menor						
	n_edaddias_nin	Edad en días de los niños/as de la encuesta						
	peso	Peso del menor en Kg						
	mad_escol	Escolaridad de la madre						
	jefe_escol	Escolaridad del jefe de hogar						
	numh	Tamaño del hogar						
Identificación y diseño muestral	id_base	Identificador de la persona						
	id_upm	Identificador de la upm						
	estrato	Identificador del estrato						
	fexp	Factor de expansión						
Variables objetivo y adicionales	talla	Talla en cm						
	zlen	Indicador antropométrico						
	dcronica_5	DCI en menores de 5 años						
	dcronica_2	DCI en menores de 2 años						
	n_grupo_edad_nin	Grupo de edad						
	year	Año de levantamiento						
C1: Categorías ECV 2006; C2: Categorías ENSANUT 2012; C3: Categorías ECV 2014; C4: Categorías ENSANUT 2018; C5: Categorías ENDI 2023; CH: Categorías HOMOLOGACIÓN								

Anexo 2: Escenarios modelos XGBoost

Los escenarios de análisis se establecen en función de la información disponible y la descomposición de los datos en conjuntos de entrenamiento y prueba. Estos escenarios brindan un contexto general de la aplicación tanto en los métodos de aprendizaje automático como en los modelos de regresión. Los escenarios a estudiar se resumen en la siguiente tabla:

Escenarios	Base de entrenamiento	Aplicación
Escenario 1	Todo: - ECV06 - ENSANUT12 - ECV14 - ENSANUT18 (grupos 4-6) - ENDI23	- ENSANUT18 grupos 1-3 - ENSANUT18 grupos 1-2 - ENSANUT18 grupo 3
Escenario 2	Solo edades 0-2 años: - ECV06 (grupos 1-3) - ENSANUT12 (grupos 1-3) - ECV14 (grupos 1-3) - ENDI23 (grupos 1-3)	- ENSANUT18 grupos 1-3 - ENSANUT18 grupos 1-2 - ENSANUT18 grupo 3
Escenario 3	Solo edades 0-1 año: - ECV06 (grupos 1-2) - ENSANUT12 (grupos 1-2) - ECV14 (grupos 1-2) - ENDI23 (grupos 1-2)	- ENSANUT18 grupos 1-2
Escenario 4	Solo edades 1-2 años: - ECV06 (grupo 3) - ENSANUT12 (grupo 3) - ECV14 (grupo 3) - ENDI23 (grupos 3)	- ENSANUT18 grupo 3
Escenario 5	Solo edades 0-5 meses: - ECV06 (grupo 1) - ENSANUT12 (grupo 1) - ECV14 (grupo 1) - ENDI23 (grupos 1)	- ENSANUT18 grupo 1
Escenario 6	Solo edades 1-2 años: - ECV06 (grupo 2) - ENSANUT12 (grupo 2) - ECV14 (grupo 2) - ENDI23 (grupos 2)	- ENSANUT18 grupo 2

Los primeros dos escenarios toman como información de entrenamiento bases de dimensiones considerables por tal motivo estas requieren un mayor tiempo de ejecución. En el caso de los escenarios 3 al 6 estos se restringen a entrenar únicamente en los grupos de edad correspondiente al grupo de aplicación que se desea puntuar.

Anexo 3: Desagregaciones resultados de los modelos

Desagregación	Categoría	Etiqueta	2006	2012	2014	2018_obs	2023	Pred1	Pred2	Pred3	Pred4	Pred5	Pred6
menores_2	1	Menores de 2	21,9%	24,0%	24,8%	27,0%	20,1%	15,8%	16,9%	18,1%	12,1%	20,5%	23,6%
gedad	1	0-5 meses	11,3%	9,5%	14,7%	19,4%	13,9%	9,5%	10,7%	11,6%	6,8%	15,4%	13,6%
	2	6-11 meses	13,8%	19,8%	20,1%	23,2%	17,1%	10,1%	12,4%	14,9%	8,7%	16,5%	17,8%
	3	12-23 meses	31,1%	32,6%	31,5%	31,5%	23,7%	20,7%	21,2%	21,9%	15,7%	24,2%	29,9%
area	1	Urbana	17,1%	21,4%	22,3%	25,4%	18,9%	13,9%	14,9%	16,0%	10,3%	18,9%	21,0%
	2	Rural	30,1%	29,0%	30,1%	30,3%	21,9%	19,7%	20,9%	22,2%	15,9%	23,6%	28,9%
region	1	Sierra	28,5%	27,7%	27,3%	28,6%	23,9%	19,2%	19,7%	19,4%	14,5%	23,0%	26,0%
	2	Costa	16,3%	20,7%	22,3%	25,3%	17,3%	12,9%	14,2%	16,4%	9,9%	18,3%	21,1%
	3	Amazonia	22,8%	19,4%	28,4%	30,8%	19,6%	19,7%	21,4%	23,3%	16,3%	22,8%	28,9%
region_ar	1	Sierra - urbana	20,8%	24,5%	23,5%	26,5%	21,3%	18,0%	19,2%	16,6%	12,4%	21,0%	22,0%
	2	Sierra - rural	38,8%	33,0%	33,7%	32,4%	27,7%	21,3%	20,7%	24,4%	18,4%	26,5%	33,1%
	3	Costa - urbana	14,5%	19,0%	21,5%	24,6%	18,0%	11,7%	12,7%	15,8%	9,2%	17,9%	20,6%
	4	Costa - rural	20,6%	26,0%	25,0%	27,3%	15,6%	16,2%	18,7%	18,2%	11,7%	19,6%	22,6%
	5	Amazonía - urbana	17,0%	14,6%	21,0%	27,5%	10,4%	8,6%	8,0%	14,2%	6,4%	14,0%	17,0%
	6	Amazonía - rural	25,1%	21,8%	31,8%	32,4%	22,8%	25,2%	28,1%	27,7%	21,3%	27,1%	34,7%
sexo	1	Hombre	25,7%	26,8%	28,1%	30,7%	23,5%	20,8%	21,6%	19,5%	15,1%	23,3%	28,5%
	2	Mujer	17,5%	21,1%	20,9%	22,8%	16,5%	10,2%	11,5%	16,4%	8,8%	17,3%	18,1%
etnia_ind	0	No indígena		23,1%	23,2%	26,2%	18,8%	14,8%	15,7%	17,1%	11,0%	19,8%	22,5%
	1	Indígena		32,4%	40,8%	36,5%	33,4%	28,0%	31,3%	29,2%	25,3%	29,0%	36,0%
escol_mom	1	Ninguno/Básica	25,1%	26,8%	30,1%	30,4%	26,3%	20,6%	21,3%	22,3%	16,1%	23,9%	26,8%
	2	Media/Bachillerato	18,5%	21,4%	21,3%	26,3%	18,8%	14,0%	14,9%	17,3%	10,9%	19,8%	23,4%
	3	Superior	12,0%	16,7%	18,2%	23,3%	12,7%	10,3%	12,5%	11,4%	7,0%	15,1%	17,5%
nbi_pob	0	No Pobre NBI	16,0%	21,9%	20,3%	25,4%	17,9%	13,9%	14,8%	15,7%	10,6%	19,0%	21,9%
	1	Pobre NBI	25,0%	25,9%	29,3%	28,9%	23,0%	18,1%	19,3%	21,0%	14,0%	22,3%	25,7%
nbi_depec	0	No dependencia económica	21,8%	24,1%	24,6%	27,1%	20,0%	15,8%	16,8%	18,1%	12,1%	20,4%	23,6%
	1	Dependencia económica	27,6%	9,1%	40,6%	22,6%	27,1%	17,0%	19,9%	16,7%	12,2%	21,3%	22,0%
nbi_matviv_def	0	Calidad materiales	20,0%	22,7%	23,8%	26,5%	19,7%	14,6%	15,8%	17,2%	11,5%	20,0%	23,0%
	1	Sin calidad materiales	29,2%	36,5%	32,3%	32,2%	23,7%	28,3%	28,4%	26,9%	18,8%	25,4%	30,2%
nbi_ser_viv	0	Calidad servicios	17,1%	22,2%	22,3%	25,5%	19,8%	14,9%	15,7%	17,0%	11,4%	19,7%	22,6%
	1	Sin calidad de servicios	26,9%	26,2%	30,0%	31,8%	20,8%	18,7%	20,8%	21,5%	14,6%	23,0%	26,8%
nbi_hcam	0	No hacinamiento	17,8%	22,7%	22,4%	26,0%	18,0%	14,3%	15,4%	16,2%	10,9%	19,5%	22,8%
	1	Hacinamiento	26,5%	30,3%	29,9%	29,4%	25,7%	19,4%	20,4%	22,6%	15,2%	22,8%	25,5%
nbi_acc_esco	0		21,3%	24,0%	24,6%	27,0%	20,0%	15,7%	16,8%	18,0%	12,0%	20,4%	23,5%
	1		38,8%	22,2%	46,1%	23,2%	21,7%	29,3%	28,5%	24,0%	22,1%	31,7%	34,3%

Pred1: Reg. lineal estándar; Pred2: Reg. lineal jerárquica; Pred3: Reg. lineal cuantil; Pred4: Imputación múltiple lineal; Pred5: Predictive Mean Matching; Pred6: XGBoost

Anexo 4: Resumen de observaciones y variables.

Observaciones

Encuesta	Observaciones Base Homologada		Observaciones luego del tratamiento de valores perdidos	
	< 2 años	< 5 años	< 2 años	< 5 años
ECV 2006	2312	6068	2242	5725
ENSANUT 2012	4029	8653	3850	8137
ECV 2014	4136	11231	4058	10789
ENSANUT 2018	7034	18714	6773	17630
ENDI 2023	7993	21531	7881	20765
Total	**25504**	**66197**	**24804**	**63046**

La base de datos homologada cuenta con un total de 66197 (observaciones) niños/as menores de 5 años que tienen un registro no vacío de talla. Es importante en este caso considerar menores que cuenten con un valor en la variable talla dado que las técnicas empleadas (regresión, imputación múltiple y XGBoost) son supervisadas y no se aplican a individuos que no cuentan con información en la variable objetivo (talla). Un subconjunto de 25504 corresponde a niños/as menores de 2 años, de los cuales 7034 pertenecen a la ENSANUT 2018. Este grupo es sobre el cual se requiere generar la nueva talla por tanto no forma parte de los entrenamientos de los distintos modelos.

Luego del tratamiento de valores perdidos solamente 6773 de los 7034 menores de 2 años de la ENSANUT 2018 tienen información completa en las variables homologadas (explicativas) entonces es solamente sobre este grupo sobre el cual es posible generar una imputación de la talla.

Adicionalmente, de los 6773 menores de 2 años, una vez que se ha elegido el mejor escenario para la imputación de la talla, mismo que cumple con condiciones de distribución y desagregaciones, se hace un estudio posterior de valores atípicos en las estimaciones. Este estudio a posteriori no pretendió validar el comportamiento de valores atípicos en las estimaciones del modelo, dado que se asume que la talla inicial no es válida si no que pretende analizar casos extremos en la condición de DCI. El estudio hace uso de un análisis de regresión y análisis de componentes principales para determinar estos casos extremos y se determina que 34 observaciones presentan una diferencia muy alta entre los valores predichos y los reales. Por tanto, de los 6773, solamente 6739 cuentan con un valor de talla.

Variables

Tipo de variable	Base1	Transformaciones	Interacciones	Base2
Dicotómicas	21	-	3855	4046
Categóricas	32	170		
Numéricas	11	44	-	55
Adicionales	10	-	-	5
Total	74			4106

Del total de variables inicialmente homologadas se selecciona aquellas que constituyen variables explicativas para el estudio de la talla en contextos de DCI. Se tiene un total de 21 variables dicotómicas, 32 variables categóricas, 11 numéricas y 10 variables adicionales (identificadores y diseño muestral). Para generar la base de

datos para el entrenamiento de los modelos, en especial el modelo XGBoost, se genera variables como transformaciones de las iniciales. En el caso de las variables numéricas se obtiene el cuadrado, cubo, seno y coseno de las variables. En las variables categóricas se realiza un proceso de binarización para generar tantas variables dummies como categorías tenga cada una de las variables. Adicionalmente, se generan interacciones entre variables dicotómicas y categóricas y posteriormente se binariza cada una de las interacciones. El primer paso para realizar las interacciones realizar combinaciones de pares de variables entre los grupos de variables dicotómicas y categóricas:

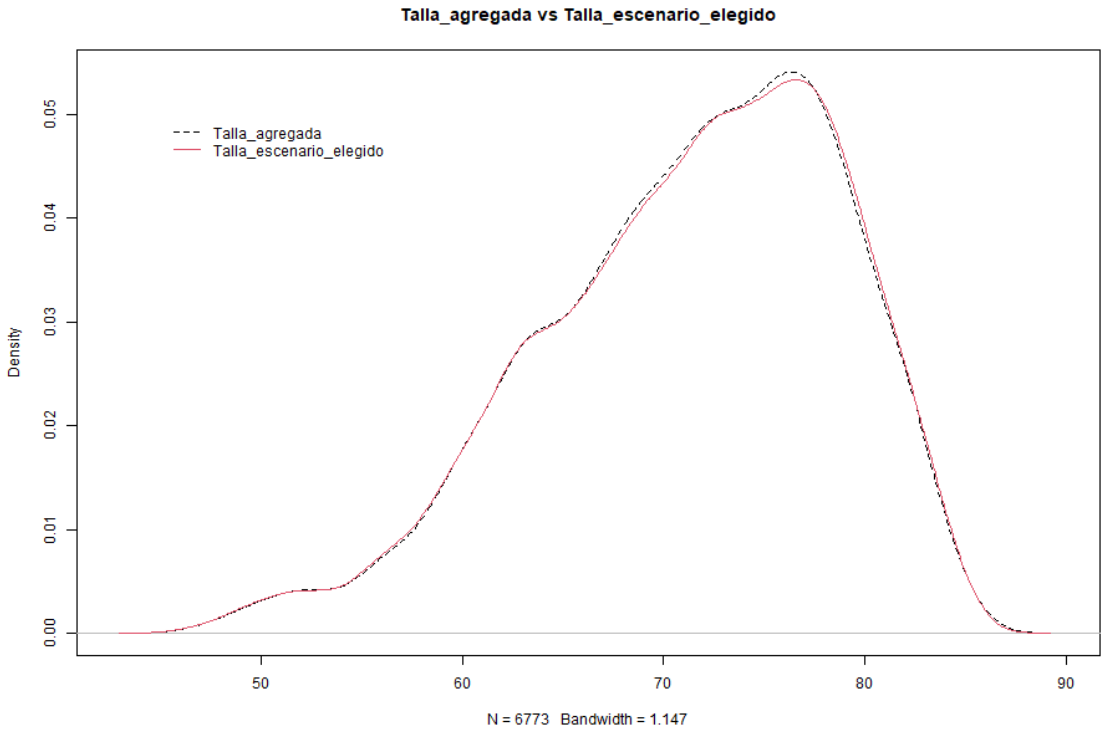
$$\binom{21+32}{2} = \binom{53}{2} = \frac{53!}{(53-2)! \cdot 2!} = \frac{53 \cdot 52 \cdot 51!}{51! \cdot 2} = 53 \cdot 26 = 1378$$

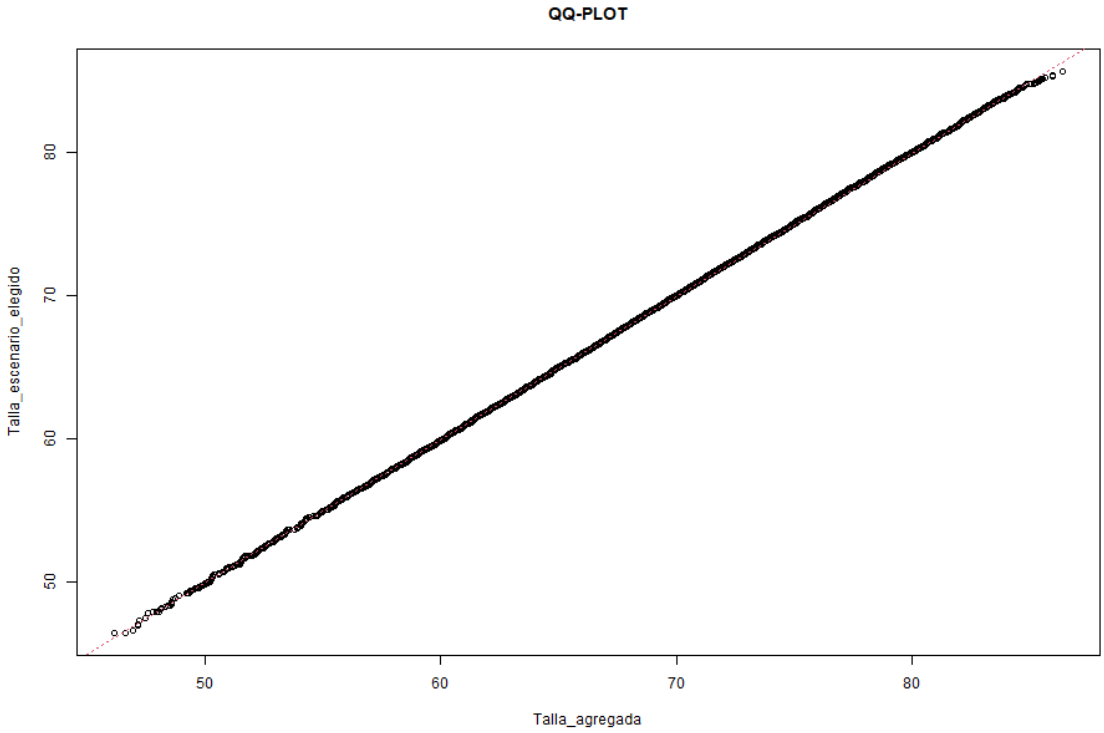
Estas 1378 combinaciones que luego pasan por un proceso de binarización dan lugar a 21376 variables lo cual limita la capacidad tecnológica para manejar el entrenamiento de los modelos. Por este motivo, se selecciona un subconjunto de las variables sobre las cuales realizar este procedimiento, 15 dicotómicas y 10 categóricas:

$$\binom{15+10}{2} = \binom{25}{2} = \frac{25!}{(25-2)! \cdot 2!} = \frac{25 \cdot 24 \cdot 23!}{23! \cdot 2} = 25 \cdot 12 = 300$$

Estas 300 combinaciones dan lugar a las 3855 variables binarias, las cuales sumadas a las 21 variables dicotómicas, las 170 variables binarizadas a partir de las variables categóricas, las 11 variables numéricas y sus 44 transformaciones suman un total de 4101 variables que forman parte del proceso de entrenamiento del modelo XGBoost, principalmente. La base de datos final tiene 4106 variables dado que se incluyen 5 variables adicionales como id, talla, dcronica, zlen y grupo.

Anexo 5: Comparación talla agregada vs talla del escenario/modelo elegido





Anexo 6: Calidad de estimaciones con información ENDI 2023 y ENSANUT 2018

ENDI 2023

grupo_sexo	col_ssample	sub_sample	max_depth	eta	dci_train	dci_test	dci_xgb	endi_train	endi_test	dci_endi	dif_train	dif_test	dif
g1_h	0,7	0,2	6	0,08	0,19	0,15	0,18	0,15	0,20	0,16	0,04	0,05	0,02
g1_h	0,8	0,2	6	0,08	0,18	0,15	0,17	0,15	0,20	0,16	0,03	0,05	0,02
g1_h	0,9	0,2	6	0,08	0,18	0,13	0,17	0,15	0,20	0,16	0,03	0,07	0,02
g1_h	0,7	0,2	6	0,09	0,14	0,08	0,13	0,15	0,20	0,16	0,01	0,12	0,03
g1_h	0,8	0,2	6	0,09	0,14	0,09	0,13	0,15	0,20	0,16	0,01	0,11	0,03
g1_h	0,9	0,2	6	0,09	0,14	0,09	0,13	0,15	0,20	0,16	0,01	0,11	0,03
g1_h	0,7	0,2	6	0,10	0,11	0,08	0,10	0,15	0,20	0,16	0,04	0,12	0,05
g1_h	0,8	0,2	6	0,10	0,12	0,08	0,11	0,15	0,20	0,16	0,03	0,12	0,04
g1_h	0,9	0,2	6	0,10	0,12	0,08	0,11	0,15	0,20	0,16	0,03	0,12	0,05
g1_h	0,7	0,2	6	0,11	0,11	0,08	0,11	0,15	0,20	0,16	0,03	0,12	0,05
g1_h	0,8	0,2	6	0,11	0,11	0,08	0,11	0,15	0,20	0,16	0,03	0,12	0,05
g1_h	0,9	0,2	6	0,11	0,11	0,08	0,10	0,15	0,20	0,16	0,04	0,12	0,05
g1_h	0,7	0,2	6	0,12	0,11	0,07	0,10	0,15	0,20	0,16	0,04	0,13	0,05
g1_h	0,8	0,2	6	0,12	0,10	0,08	0,10	0,15	0,20	0,16	0,05	0,12	0,06
g1_h	0,9	0,2	6	0,12	0,11	0,07	0,10	0,15	0,20	0,16	0,04	0,13	0,05
g1_m	0,7	0,2	6	0,08	0,08	0,11	0,09	0,11	0,09	0,12	0,03	0,02	0,03
g1_m	0,8	0,2	6	0,08	0,11	0,10	0,11	0,11	0,09	0,12	0,00	0,01	0,01
g1_m	0,9	0,2	6	0,08	0,11	0,11	0,11	0,11	0,09	0,12	0,00	0,01	0,01
g1_m	0,7	0,2	6	0,09	0,07	0,05	0,07	0,11	0,09	0,12	0,04	0,04	0,05
g1_m	0,8	0,2	6	0,09	0,07	0,07	0,07	0,11	0,09	0,12	0,04	0,02	0,05
g1_m	0,9	0,2	6	0,09	0,07	0,08	0,07	0,11	0,09	0,12	0,04	0,01	0,05
g1_m	0,7	0,2	6	0,10	0,06	0,04	0,06	0,11	0,09	0,12	0,05	0,06	0,06
g1_m	0,8	0,2	6	0,10	0,06	0,04	0,06	0,11	0,09	0,12	0,05	0,06	0,06
g1_m	0,9	0,2	6	0,10	0,06	0,05	0,06	0,11	0,09	0,12	0,05	0,05	0,06
g1_m	0,7	0,2	6	0,11	0,05	0,04	0,05	0,11	0,09	0,12	0,06	0,06	0,07
g1_m	0,8	0,2	6	0,11	0,06	0,04	0,05	0,11	0,09	0,12	0,05	0,06	0,06
g1_m	0,9	0,2	6	0,11	0,06	0,04	0,06	0,11	0,09	0,12	0,05	0,06	0,06
g1_m	0,7	0,2	6	0,12	0,06	0,03	0,05	0,11	0,09	0,12	0,05	0,06	0,07
g1_m	0,8	0,2	6	0,12	0,06	0,03	0,05	0,11	0,09	0,12	0,05	0,06	0,07
g1_m	0,9	0,2	6	0,12	0,06	0,03	0,05	0,11	0,09	0,12	0,05	0,06	0,07
g2_h	0,7	0,2	6	0,08	0,34	0,26	0,33	0,23	0,14	0,21	0,11	0,13	0,11
g2_h	0,8	0,2	6	0,08	0,34	0,28	0,33	0,23	0,14	0,21	0,11	0,14	0,12
g2_h	0,9	0,2	6	0,08	0,33	0,25	0,32	0,23	0,14	0,21	0,11	0,11	0,11
g2_h	0,7	0,2	6	0,09	0,26	0,18	0,24	0,23	0,14	0,21	0,03	0,05	0,03
g2_h	0,8	0,2	6	0,09	0,25	0,20	0,24	0,23	0,14	0,21	0,02	0,06	0,02
g2_h	0,9	0,2	6	0,09	0,26	0,18	0,24	0,23	0,14	0,21	0,03	0,05	0,03
g2_h	0,7	0,2	6	0,10	0,20	0,15	0,19	0,23	0,14	0,21	0,03	0,01	0,02
g2_h	0,8	0,2	6	0,10	0,21	0,16	0,20	0,23	0,14	0,21	0,02	0,02	0,01
g2_h	0,9	0,2	6	0,10	0,20	0,17	0,20	0,23	0,14	0,21	0,03	0,04	0,02
g2_h	0,7	0,2	6	0,11	0,16	0,13	0,16	0,23	0,14	0,21	0,07	0,01	0,06
g2_h	0,8	0,2	6	0,11	0,18	0,14	0,17	0,23	0,14	0,21	0,05	0,00	0,04
g2_h	0,9	0,2	6	0,11	0,16	0,13	0,15	0,23	0,14	0,21	0,07	0,01	0,06

g2_h	0,7	0,2	6	0,12	0,16	0,14	0,16	0,23	0,14	0,21	0,07	0,01	0,06
g2_h	0,8	0,2	6	0,12	0,16	0,12	0,15	0,23	0,14	0,21	0,07	0,01	0,06
g2_h	0,9	0,2	6	0,12	0,16	0,15	0,16	0,23	0,14	0,21	0,06	0,02	0,05
g2_m	0,7	0,2	6	0,08	0,16	0,17	0,16	0,11	0,17	0,13	0,04	0,01	0,03
g2_m	0,8	0,2	6	0,08	0,15	0,19	0,16	0,11	0,17	0,13	0,04	0,03	0,03
g2_m	0,9	0,2	6	0,08	0,16	0,18	0,16	0,11	0,17	0,13	0,04	0,01	0,04
g2_m	0,7	0,2	6	0,09	0,10	0,12	0,10	0,11	0,17	0,13	0,02	0,04	0,02
g2_m	0,8	0,2	6	0,09	0,10	0,13	0,10	0,11	0,17	0,13	0,02	0,04	0,02
g2_m	0,9	0,2	6	0,09	0,10	0,13	0,11	0,11	0,17	0,13	0,01	0,04	0,02
g2_m	0,7	0,2	6	0,10	0,09	0,10	0,09	0,11	0,17	0,13	0,02	0,07	0,04
g2_m	0,8	0,2	6	0,10	0,09	0,09	0,09	0,11	0,17	0,13	0,03	0,07	0,04
g2_m	0,9	0,2	6	0,10	0,09	0,10	0,09	0,11	0,17	0,13	0,02	0,07	0,04
g2_m	0,7	0,2	6	0,11	0,08	0,08	0,08	0,11	0,17	0,13	0,04	0,09	0,05
g2_m	0,8	0,2	6	0,11	0,07	0,06	0,07	0,11	0,17	0,13	0,04	0,11	0,06
g2_m	0,9	0,2	6	0,11	0,08	0,07	0,08	0,11	0,17	0,13	0,03	0,10	0,05
g2_m	0,7	0,2	6	0,12	0,07	0,08	0,07	0,11	0,17	0,13	0,04	0,08	0,05
g2_m	0,8	0,2	6	0,12	0,07	0,08	0,07	0,11	0,17	0,13	0,04	0,09	0,05
g2_m	0,9	0,2	6	0,12	0,07	0,07	0,07	0,11	0,17	0,13	0,04	0,10	0,05
g3_h	0,7	0,2	6	0,08	0,40	0,44	0,41	0,26	0,34	0,27	0,15	0,11	0,13
g3_h	0,8	0,2	6	0,08	0,41	0,43	0,41	0,26	0,34	0,27	0,15	0,10	0,14
g3_h	0,9	0,2	6	0,08	0,40	0,44	0,41	0,26	0,34	0,27	0,14	0,10	0,13
g3_h	0,7	0,2	6	0,09	0,29	0,34	0,30	0,26	0,34	0,27	0,03	0,01	0,03
g3_h	0,8	0,2	6	0,09	0,29	0,33	0,30	0,26	0,34	0,27	0,04	0,01	0,03
g3_h	0,9	0,2	6	0,09	0,30	0,32	0,30	0,26	0,34	0,27	0,04	0,01	0,03
g3_h	0,7	0,2	6	0,10	0,25	0,30	0,26	0,26	0,34	0,27	0,01	0,04	0,01
g3_h	0,8	0,2	6	0,10	0,24	0,27	0,24	0,26	0,34	0,27	0,02	0,07	0,03
g3_h	0,9	0,2	6	0,10	0,24	0,28	0,24	0,26	0,34	0,27	0,02	0,06	0,03
g3_h	0,7	0,2	6	0,11	0,21	0,23	0,22	0,26	0,34	0,27	0,04	0,11	0,06
g3_h	0,8	0,2	6	0,11	0,22	0,23	0,22	0,26	0,34	0,27	0,04	0,10	0,05
g3_h	0,9	0,2	6	0,11	0,22	0,23	0,22	0,26	0,34	0,27	0,04	0,11	0,06
g3_h	0,7	0,2	6	0,12	0,20	0,21	0,20	0,26	0,34	0,27	0,05	0,12	0,07
g3_h	0,8	0,2	6	0,12	0,20	0,21	0,20	0,26	0,34	0,27	0,05	0,13	0,07
g3_h	0,9	0,2	6	0,12	0,21	0,20	0,20	0,26	0,34	0,27	0,05	0,14	0,07
g3_m	0,7	0,2	6	0,08	0,30	0,34	0,30	0,20	0,21	0,20	0,10	0,13	0,10
g3_m	0,8	0,2	6	0,08	0,29	0,34	0,30	0,20	0,21	0,20	0,09	0,13	0,10
g3_m	0,9	0,2	6	0,08	0,30	0,34	0,31	0,20	0,21	0,20	0,10	0,14	0,11
g3_m	0,7	0,2	6	0,09	0,22	0,23	0,23	0,20	0,21	0,20	0,02	0,02	0,03
g3_m	0,8	0,2	6	0,09	0,22	0,24	0,22	0,20	0,21	0,20	0,02	0,03	0,02
g3_m	0,9	0,2	6	0,09	0,23	0,23	0,23	0,20	0,21	0,20	0,03	0,02	0,03
g3_m	0,7	0,2	6	0,1	0,20	0,18	0,19	0,20	0,21	0,20	0,00	0,02	0,00
g3_m	0,8	0,2	6	0,1	0,20	0,19	0,20	0,20	0,21	0,20	0,00	0,02	0,00
g3_m	0,9	0,2	6	0,1	0,19	0,19	0,19	0,20	0,21	0,20	0,01	0,02	0,01
g3_m	0,7	0,2	6	0,11	0,16	0,16	0,16	0,20	0,21	0,20	0,04	0,04	0,03
g3_m	0,8	0,2	6	0,11	0,17	0,18	0,17	0,20	0,21	0,20	0,03	0,03	0,03
g3_m	0,9	0,2	6	0,11	0,17	0,17	0,17	0,20	0,21	0,20	0,03	0,04	0,03
g3_m	0,7	0,2	6	0,12	0,16	0,17	0,16	0,20	0,21	0,20	0,04	0,04	0,04
g3_m	0,8	0,2	6	0,12	0,15	0,15	0,15	0,20	0,21	0,20	0,05	0,06	0,05
g3_m	0,9	0,2	6	0,12	0,15	0,17	0,16	0,20	0,21	0,20	0,05	0,04	0,04

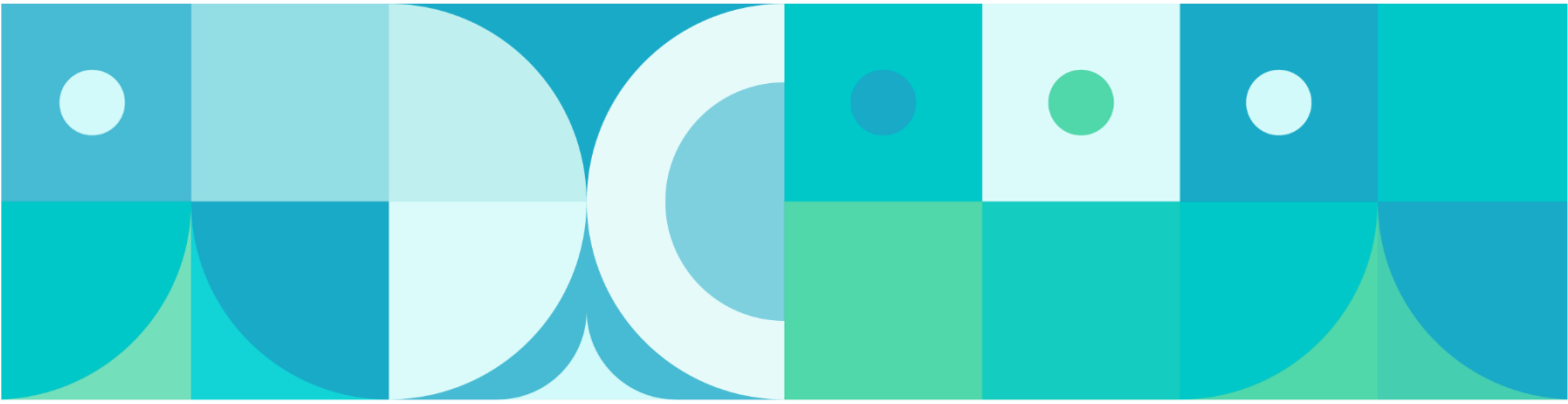
ENSANUT 2018

grupo_sexo	col_sample	sub_sample	max_depth	eta	dci_train	dci_test	dci_xgb	ens_train	ens_test	dci_ens	dif_train	dif_test	dif
g1_h	0,7	0,2	6	0,08	0,25	0,32	0,26	0,23	0,19	0,22	0,02	0,13	0,04
g1_h	0,8	0,2	6	0,08	0,26	0,34	0,27	0,23	0,19	0,22	0,03	0,15	0,05
g1_h	0,9	0,2	6	0,08	0,26	0,36	0,28	0,23	0,19	0,22	0,03	0,17	0,06
g1_h	0,7	0,2	6	0,09	0,22	0,28	0,23	0,23	0,19	0,22	0,01	0,09	0,01
g1_h	0,8	0,2	6	0,09	0,21	0,26	0,22	0,23	0,19	0,22	0,02	0,07	0,01
g1_h	0,9	0,2	6	0,09	0,22	0,28	0,24	0,23	0,19	0,22	0,01	0,09	0,01
g1_h	0,7	0,2	6	0,10	0,21	0,25	0,22	0,23	0,19	0,22	0,02	0,06	0,01
g1_h	0,8	0,2	6	0,10	0,20	0,27	0,22	0,23	0,19	0,22	0,03	0,08	0,01
g1_h	0,9	0,2	6	0,10	0,19	0,24	0,20	0,23	0,19	0,22	0,04	0,04	0,02
g1_h	0,7	0,2	6	0,11	0,18	0,24	0,19	0,23	0,19	0,22	0,06	0,05	0,04
g1_h	0,8	0,2	6	0,11	0,20	0,25	0,21	0,23	0,19	0,22	0,04	0,06	0,02
g1_h	0,9	0,2	6	0,11	0,20	0,26	0,21	0,23	0,19	0,22	0,03	0,07	0,01
g1_h	0,7	0,2	6	0,12	0,20	0,19	0,20	0,23	0,19	0,22	0,03	0,01	0,03
g1_h	0,8	0,2	6	0,12	0,19	0,22	0,19	0,23	0,19	0,22	0,04	0,03	0,03
g1_h	0,9	0,2	6	0,12	0,19	0,15	0,18	0,23	0,19	0,22	0,05	0,04	0,05
g1_m	0,7	0,2	6	0,08	0,11	0,10	0,11	0,16	0,18	0,17	0,05	0,08	0,06
g1_m	0,8	0,2	6	0,08	0,13	0,08	0,12	0,16	0,18	0,17	0,03	0,10	0,05
g1_m	0,9	0,2	6	0,08	0,12	0,09	0,11	0,16	0,18	0,17	0,04	0,08	0,05
g1_m	0,7	0,2	6	0,09	0,09	0,08	0,09	0,16	0,18	0,17	0,06	0,09	0,08
g1_m	0,8	0,2	6	0,09	0,08	0,08	0,08	0,16	0,18	0,17	0,08	0,10	0,09
g1_m	0,9	0,2	6	0,09	0,09	0,07	0,09	0,16	0,18	0,17	0,06	0,10	0,08
g1_m	0,7	0,2	6	0,10	0,06	0,06	0,06	0,16	0,18	0,17	0,09	0,12	0,11
g1_m	0,8	0,2	6	0,10	0,07	0,08	0,07	0,16	0,18	0,17	0,08	0,10	0,09
g1_m	0,9	0,2	6	0,10	0,07	0,08	0,07	0,16	0,18	0,17	0,08	0,10	0,09
g1_m	0,7	0,2	6	0,11	0,07	0,07	0,07	0,16	0,18	0,17	0,09	0,11	0,10
g1_m	0,8	0,2	6	0,11	0,07	0,06	0,07	0,16	0,18	0,17	0,08	0,12	0,10
g1_m	0,9	0,2	6	0,11	0,07	0,07	0,07	0,16	0,18	0,17	0,09	0,10	0,10
g1_m	0,7	0,2	6	0,12	0,08	0,03	0,07	0,16	0,18	0,17	0,08	0,14	0,10
g1_m	0,8	0,2	6	0,12	0,07	0,06	0,07	0,16	0,18	0,17	0,08	0,12	0,09
g1_m	0,9	0,2	6	0,12	0,07	0,08	0,07	0,16	0,18	0,17	0,08	0,10	0,09
g2_h	0,7	0,2	6	0,08	0,25	0,21	0,24	0,27	0,30	0,27	0,02	0,09	0,03
g2_h	0,8	0,2	6	0,08	0,26	0,23	0,25	0,27	0,30	0,27	0,01	0,07	0,02
g2_h	0,9	0,2	6	0,08	0,25	0,24	0,25	0,27	0,30	0,27	0,02	0,06	0,02
g2_h	0,7	0,2	6	0,09	0,19	0,20	0,19	0,27	0,30	0,27	0,08	0,09	0,08
g2_h	0,8	0,2	6	0,09	0,21	0,19	0,20	0,27	0,30	0,27	0,06	0,10	0,07
g2_h	0,9	0,2	6	0,09	0,19	0,19	0,19	0,27	0,30	0,27	0,07	0,10	0,08
g2_h	0,7	0,2	6	0,10	0,17	0,15	0,17	0,27	0,30	0,27	0,10	0,15	0,10
g2_h	0,8	0,2	6	0,10	0,17	0,17	0,17	0,27	0,30	0,27	0,09	0,13	0,10
g2_h	0,9	0,2	6	0,10	0,18	0,16	0,17	0,27	0,30	0,27	0,09	0,14	0,10
g2_h	0,7	0,2	6	0,11	0,15	0,18	0,16	0,27	0,30	0,27	0,12	0,12	0,11
g2_h	0,8	0,2	6	0,11	0,14	0,15	0,14	0,27	0,30	0,27	0,13	0,15	0,13
g2_h	0,9	0,2	6	0,11	0,15	0,15	0,15	0,27	0,30	0,27	0,12	0,15	0,12
g2_h	0,7	0,2	6	0,12	0,15	0,14	0,15	0,27	0,30	0,27	0,12	0,16	0,13
g2_h	0,8	0,2	6	0,12	0,15	0,16	0,15	0,27	0,30	0,27	0,12	0,14	0,12
g2_h	0,9	0,2	6	0,12	0,14	0,14	0,14	0,27	0,30	0,27	0,13	0,16	0,13

g2_m	0,7	0,2	6	0,08	0,17	0,13	0,16	0,19	0,17	0,18	0,02	0,05	0,02
g2_m	0,8	0,2	6	0,08	0,18	0,13	0,17	0,19	0,17	0,18	0,01	0,04	0,01
g2_m	0,9	0,2	6	0,08	0,19	0,16	0,18	0,19	0,17	0,18	0,00	0,02	0,00
g2_m	0,7	0,2	6	0,09	0,15	0,10	0,14	0,19	0,17	0,18	0,04	0,07	0,04
g2_m	0,8	0,2	6	0,09	0,14	0,11	0,14	0,19	0,17	0,18	0,05	0,07	0,05
g2_m	0,9	0,2	6	0,09	0,14	0,11	0,14	0,19	0,17	0,18	0,05	0,06	0,05
g2_m	0,7	0,2	6	0,10	0,12	0,09	0,12	0,19	0,17	0,18	0,07	0,08	0,07
g2_m	0,8	0,2	6	0,10	0,14	0,09	0,13	0,19	0,17	0,18	0,05	0,08	0,06
g2_m	0,9	0,2	6	0,10	0,14	0,10	0,13	0,19	0,17	0,18	0,05	0,07	0,05
g2_m	0,7	0,2	6	0,11	0,13	0,08	0,12	0,19	0,17	0,18	0,06	0,10	0,07
g2_m	0,8	0,2	6	0,11	0,13	0,08	0,12	0,19	0,17	0,18	0,06	0,10	0,06
g2_m	0,9	0,2	6	0,11	0,13	0,07	0,12	0,19	0,17	0,18	0,06	0,10	0,06
g2_m	0,7	0,2	6	0,12	0,11	0,07	0,10	0,19	0,17	0,18	0,08	0,11	0,08
g2_m	0,8	0,2	6	0,12	0,12	0,08	0,11	0,19	0,17	0,18	0,07	0,09	0,07
g2_m	0,9	0,2	6	0,12	0,12	0,09	0,12	0,19	0,17	0,18	0,07	0,08	0,07
g3_h	0,7	0,2	6	0,08	0,49	0,50	0,49	0,35	0,41	0,35	0,14	0,09	0,14
g3_h	0,8	0,2	6	0,08	0,49	0,51	0,49	0,35	0,41	0,35	0,14	0,10	0,14
g3_h	0,9	0,2	6	0,08	0,50	0,49	0,49	0,35	0,41	0,35	0,15	0,08	0,14
g3_h	0,7	0,2	6	0,09	0,39	0,40	0,39	0,35	0,41	0,35	0,04	0,01	0,04
g3_h	0,8	0,2	6	0,09	0,40	0,40	0,40	0,35	0,41	0,35	0,06	0,01	0,05
g3_h	0,9	0,2	6	0,09	0,40	0,43	0,41	0,35	0,41	0,35	0,06	0,01	0,05
g3_h	0,7	0,2	6	0,10	0,34	0,34	0,34	0,35	0,41	0,35	0,01	0,07	0,02
g3_h	0,8	0,2	6	0,10	0,33	0,34	0,33	0,35	0,41	0,35	0,01	0,07	0,02
g3_h	0,9	0,2	6	0,10	0,33	0,37	0,34	0,35	0,41	0,35	0,02	0,05	0,02
g3_h	0,7	0,2	6	0,11	0,31	0,32	0,31	0,35	0,41	0,35	0,04	0,09	0,04
g3_h	0,8	0,2	6	0,11	0,29	0,31	0,29	0,35	0,41	0,35	0,05	0,10	0,06
g3_h	0,9	0,2	6	0,11	0,30	0,31	0,31	0,35	0,41	0,35	0,04	0,10	0,05
g3_h	0,7	0,2	6	0,12	0,29	0,31	0,30	0,35	0,41	0,35	0,05	0,10	0,06
g3_h	0,8	0,2	6	0,12	0,29	0,29	0,29	0,35	0,41	0,35	0,05	0,12	0,06
g3_h	0,9	0,2	6	0,12	0,28	0,30	0,29	0,35	0,41	0,35	0,06	0,11	0,07
g3_m	0,7	0,2	6	0,08	0,29	0,32	0,30	0,28	0,25	0,27	0,01	0,07	0,02
g3_m	0,8	0,2	6	0,08	0,28	0,30	0,28	0,28	0,25	0,27	0,00	0,05	0,01
g3_m	0,9	0,2	6	0,08	0,29	0,30	0,29	0,28	0,25	0,27	0,01	0,05	0,02
g3_m	0,7	0,2	6	0,09	0,23	0,22	0,23	0,28	0,25	0,27	0,05	0,03	0,04
g3_m	0,8	0,2	6	0,09	0,23	0,23	0,23	0,28	0,25	0,27	0,06	0,02	0,04
g3_m	0,9	0,2	6	0,09	0,23	0,22	0,23	0,28	0,25	0,27	0,05	0,03	0,04
g3_m	0,7	0,2	6	0,10	0,19	0,18	0,19	0,28	0,25	0,27	0,09	0,08	0,08
g3_m	0,8	0,2	6	0,10	0,19	0,18	0,19	0,28	0,25	0,27	0,09	0,07	0,08
g3_m	0,9	0,2	6	0,10	0,20	0,18	0,20	0,28	0,25	0,27	0,08	0,07	0,07
g3_m	0,7	0,2	6	0,11	0,18	0,16	0,18	0,28	0,25	0,27	0,10	0,09	0,09
g3_m	0,8	0,2	6	0,11	0,18	0,15	0,17	0,28	0,25	0,27	0,10	0,11	0,10
g3_m	0,9	0,2	6	0,11	0,19	0,16	0,18	0,28	0,25	0,27	0,09	0,10	0,09
g3_m	0,7	0,2	6	0,12	0,17	0,15	0,17	0,28	0,25	0,27	0,11	0,11	0,10
g3_m	0,8	0,2	6	0,12	0,17	0,13	0,16	0,28	0,25	0,27	0,11	0,12	0,11
g3_m	0,9	0,2	6	0,12	0,18	0,14	0,18	0,28	0,25	0,27	0,10	0,12	0,10

@InecEcuador
@ecuadorencifras
@ecuadorencifras
INECEcuador

IMPUTACIÓN TALLA
ENSANUT 2018



www.ecuadorencifras.gob.ec