Pontificia Universidad Católica del Ecuador

“APLICACIÓN DE TECNICAS DE MACHINE LEARNING PARA PREDECIR LA DESNUTRICION INFANTIL EN EL ECUADOR”.

Trabajo de titulación previo a obtener título de master en sistemas de información mención data science

Cleber Damian Puente Tiscama

Director: Msc. Eduardo Montero Bermúdez

Quito – Ecuador

2024

DEDICATORIA

A mi esposa Jimena y a mis hijos, quienes siempre han estado a mi lado brindándome su amor y apoyo incondicional. A mis padres, Damián y María Carlota, cuya sabiduría y ejemplo me han guiado en cada paso, y a mis hermanos Edison, Lucia y Javier, por su constante aliento.

Cleber Damián Puente Tiscama

AGRADECIMIENTO

A mi esposa Jimena, por su inquebrantable apoyo, amor y comprensión durante todo este proceso. Tu paciencia y aliento han sido fundamentales para alcanzar esta meta. A mis queridos hijos, Emily e Ismael, gracias por su alegría y energía, que me han inspirado a seguir adelante cada día. Este logro es tanto suyo como mío, y estoy inmensamente agradecido por su amor y presencia en mi vida.

A mis padres y hermanos ya que ellos de igual manera son las personas que han estado en las buenas y en las malas siempre apoyándome.

Al la Pontificia Universidad Católica del Ecuador en especial a los que conforman al área de Ingeniería, A mi director Msc. Eduardo Montero Bermúdez, quien gracias a su asesoramiento pude culminar el presente trabajo de titulación.

Cleber Damián Puente Tiscama

ÍNDICE DE CONTENIDOS

PÁGINA

ÍNDICE DE TABLAS

PÁGINA

ÍNDICE DE FIGURAS

PÁGINA

ÍNDICE DE ANEXOS

PÁGINA

RESUMEN

El presente trabajo de titulación establece un modelo optimizado de machine learning que permite predecir la desnutrición crónica infantil con un nivel de accuracy lo suficientemente alto como para tomar decisiones. Se utiliza la base de datos de la Encuesta Nacional de Salud y Nutrición (ENSANUT)

El proyecto realizado tiene como objetivo principal implementar un modelo de aprendizaje supervisado que permita predecir la desnutrición crónica infantil en el Ecuador, este modelo tendrá muy buenas métricas de performance e.g.roc auc superior a 0.8, acurracy por encima del 80%.

Como resultado se obtendrá un modelo computacional de clasificación binaria de machine learning en lenguaje de programación Python con un nivel alto de accuracy que permita caracterizar y clasificar a los niños con desnutrición crónica infantil.

Se recomienda proveer una estrategia de implementación y uso del modelo, así como un análisis del grado de influencia de las variables.

Palabras Claves: Desnutrición crónica infantil, INEC, ENSANUT, clasificación, Python, variables machine learning.

ABSTRACT

This degree work establishes an optimized machine learning model that allows predicting childhood chronic malnutrition with a level of accuracy high enough to make decisions. The database of the National Health and Nutrition Survey (ENSANUT) is used.

The main objective of the project carried out is to implement a supervised learning model that allows predicting childhood chronic malnutrition in Ecuador. This model will have very good performance metrics, for example, roc auc greater than 0.8, precision above 80%.

As a result, a computational machine learning binary classification model will be obtained in the Python programming language with a high level of accuracy that allows the characterization and classification of children with chronic childhood malnutrition.

It is recommended to provide a strategy for implementation and use of the model, as well as an analysis of the degree of influence of the variables.

Keywords: Chronic childhood malnutrition, ENSANUT, classification, Python, machine learning variables.

1. Introducción

La desnutrición crónica infantil es uno de los problemas sociales que impactan a la sociedad a nivel global. En 2022, aproximadamente 149 millones de niños menores de 5 años presentaban retraso en el crecimiento (eran demasiado pequeños para su edad), 45 millones sufrían emaciación (eran demasiado delgados para su estatura). La desnutrición está asociada con alrededor de la mitad de las muertes en este grupo etario, siendo estas defunciones más comunes en países de ingresos bajos y medianos. Las repercusiones de la malnutrición son graves y perdurables, afectando el desarrollo infantil y teniendo consecuencias económicas, sociales y médicas significativas para los individuos afectados, sus familias, comunidades y países (Organización Mundial de la Salud, 2022)

En Ecuador, según los resultados obtenidos de la 1ra. Encuesta Especializada sobre Desnutrición Infantil (ENDI), el 1% de los niños menores de 2 años sufre Desnutrición Crónica Infantil (DCI). La región rural de la sierra tiene el mayor porcentaje, con un 27.7%. La DCI afecta al 24% de los niños en los hogares más pobres, mientras que en los más ricos alcanza el 15.2%. Chimborazo, Bolívar y Santa Elena presentan los niveles más altos, con 35.1%, 30.3% y 29.8%, respectivamente. Entre los niños indígenas, el 33.4% padece DCI, en contraste con el 2% de mestizos, 15.7% de afroecuatorianos y 15% de montubios. Ecuador ocupa el cuarto lugar en la región en índice de DCI, después de Honduras, Haití y Guatemala. (INEC, 2023).

Además de las graves consecuencias para quienes la padecen, la desnutrición ejerce un impacto significativo en el desarrollo económico y social de las naciones. En Ecuador, los costos relacionados con la malnutrición, como los gastos en salud, educación y la disminución de la productividad, representan aproximadamente el 4,3% del Producto Interno Bruto (PIB) (UNICEF, 2021).

Diferentes organizaciones tanto nacionales como internacionales han establecido planes activos para combatir la desnutrición infantil crónica. González (2021), representante de UNICEF en Ecuador, sostiene que "Al no atender el derecho a la nutrición, el país no solo incumple un compromiso con la niñez, sino que está hipotecando su desarrollo a futuro".

1.1. Antecedentes

La desnutrición crónica infantil ha sido un tema de creciente preocupación a nivel global, como lo evidencian diversos informes de UNICEF. La desnutrición crónica afecta a millones de niños menores de cinco años en todo el mundo, con un impacto devastador en su crecimiento y desarrollo. El informe destaca que las deficiencias en el crecimiento comienzan durante el periodo de gestación y se prolongan hasta aproximadamente los 24 meses de edad, con una probabilidad mínima de reversión de los daños una vez que se establecen, también informa que aproximadamente ocho millones de niños menores de cinco años están en riesgo de muerte debido a bajo peso en relación con su altura si no reciben tratamiento adecuado. (UNICEF, 2022). En este contexto, mi investigación busca abordar estas cuestiones mediante la aplicación de algoritmos de machine learning para predecir las causas más relevantes con respecto a la DCI y sobre todo construir un modelo que permita predecir con un nivel de precisión alto que ayudaría, sin lugar a dudas a reducir el número de niños con desnutrición crónica.

Entre los antecedentes importantes para este trabajo se encuentran dos tesis que han examinado la desnutrición crónica infantil en Ecuador desde diversas perspectivas. La primera, titulada "Análisis Exploratorio de Datos e Identificación de Agentes que Influyen en la Desnutrición Crónica de Niños Menores a Cinco Años del Ecuador Mediante la Aplicación de Técnicas de Ciencia de Datos" (Yánez, 2023), utiliza técnicas de ciencia de datos para identificar los factores que contribuyen a la desnutrición crónica, ofreciendo un análisis exhaustivo de las variables que

afectan a esta problemática. La segunda tesis, "Comparación de Modelos Logísticos y Árboles de Decisión para Identificar y Predecir Factores Asociados a la Desnutrición Crónica Infantil Basados en la Encuesta Nacional de Salud y Nutrición – ENSANUT 2018-2019" (Congacha, 2020), se enfoca en la aplicación de modelos estadísticos y de machine learning para predecir y analizar los factores relacionados con la desnutrición crónica infantil. Estas investigaciones proporcionan un marco metodológico y conceptual que sirve de base y guía para el presente estudio, que tiene como objetivo implementar un modelo de aprendizaje supervisado para predecir la desnutrición crónica infantil en Ecuador, con el propósito de ofrecer nuevas perspectivas y soluciones para abordar este grave problema de salud pública.

1.2. Justificación

El interés de este proyecto se encuentra anclado al trabajo de titulación de la maestría en Sistemas de información, mención Data Science de la Pontificia Universidad Católica del Ecuador. El desarrollo de este análisis, beneficia principalmente a la sociedad y a la comunidad universitaria, ya que a través del estudio se pretende implementar la analítica predictiva para identificar factores asociados a la desnutrición crónica. Esto no afecta solo a un país o un continente, más bien, podríamos hablar de una problemática de carácter global.

Se han desarrollado trabajos de investigación aplicando técnicas de machine learning con el objetivo de identificar los factores que contribuyen al DCI tanto a nivel nacional (Yánez, 2023) como internacional (S. Kar, 2021) (Ashis Talukder, 2020). Sin embargo, cuando hablamos de Ecuador hace falta un modelo optimizado de machine learning que permita predecir la DCI con un nivel de accuracy lo suficientemente alto como para tomar decisiones, desarrollar e implementar políticas públicas proactivas en función de este.

Los beneficiarios inmediatos del presente trabajo serán la sociedad en si para erradicar la desnutrición infantil por medio de la tecnología.

Es factible porque se cuente con el apoyo suficiente recursos, materiales y bibliográficos también con el conocimiento necesario para que lo planificado en este proyecto sea lo correcto.

1.3. Pregunta de investigación

¿Es posible desarrollar un modelo de machine learning que permita diferenciar a las personas con desnutrición crónica infantil de las que no?

1.3.1. Objetivo general

Desarrollar un modelo de aprendizaje supervisado que permita predecir la desnutrición crónica infantil en el Ecuador.

1.3.2. Objetivos específicos

Realizar preprocesamiento del banco de datos ENSANUT 2018

Aplicar técnicas de feature engineering, en la creación de variables relacionadas con la desnutrición crónica infantil.

Crear varios modelos de clasificación aplicando diferentes algoritmos de machine learning y seleccionar sus mejores hiperparàmetros

Comparar los modelos en relación varias métricas de performance y seleccionar el modelo ganador.

Aplicar técnicas de visualización en el análisis y detección de insights.

Proponer alguna estrategia de implementación del modelo ganador.

2. Marco teórico

2.1. Desnutrición crónica infantil

2.1.1. Desnutrición

La desnutrición se refiere a un estado patológico derivado de una ingesta insuficiente de nutrientes esenciales necesarios para el mantenimiento de la salud y el bienestar. Este fenómeno puede manifestarse en diversas formas, como malnutrición por déficit calórico, deficiencias específicas de vitaminas o minerales, o desnutrición proteico-calórica. Los efectos de la desnutrición abarcan una amplia gama de problemas de salud, que van desde el retraso en el crecimiento y desarrollo en los niños, hasta una mayor vulnerabilidad a enfermedades infecciosas en todas las edades. La desnutrición no solo afecta a la salud física, sino que también puede influir en el desarrollo cognitivo, el rendimiento escolar y la calidad de vida general (Gómez, 2020).

La desnutrición ocurre principalmente de cuatro maneras: emaciación, retraso del crecimiento, insuficiencia ponderal y carencia de vitaminas y minerales. Los niños desnutridos están especialmente en peligro para la enfermedad y la mortalidad. La emaciación es bajo peso para la altura y generalmente es causada por una mala dieta o una enfermedad recurrente. El retraso del crecimiento, o baja estatura para la edad, ha sido asociado con la desnutrición crónica y malas circunstancias socioeconómicas y tiene implicaciones para el desarrollo físico y cognitivo. Las carencias de vitaminas y minerales también conocida como "hambre oculta", este tipo de desnutrición se debe a la falta de micronutrientes esenciales en la dieta, como el hierro, la vitamina A, y el zinc, lo que afecta el desarrollo y la salud general. (Organización Mundial de la Salud, 2022)

2.1.2. Desnutrición crónica infantil

La desnutrición es el problema más crítico entre millones de niños en todo el mundo, especialmente en los países pobres. Ocurre a partir del consumo inadecuado de alimentos en cantidad y calidad, lo que resulta en un insuficiente crecimiento y desarrollo físico y mental. También aumenta el riesgo de enfermedades y el riesgo de muerte entre los niños (Wisbaum, 2011).

La desnutrición crónica infantil sigue siendo un problema persistente en Ecuador, con serias repercusiones para el desarrollo de los niños y niñas. Según un informe del Fondo de las Naciones Unidas para la Infancia (UNICEF), aproximadamente uno de cada cuatro niños menores de cinco años en el país padece desnutrición crónica, afectando su crecimiento y desarrollo integral. Esta situación pone de manifiesto las desigualdades socioeconómicas y la necesidad urgente de implementar políticas efectivas para abordar las causas subyacentes de la desnutrición, una condición que impacta negativamente en toda la población (UNICEF, 2022)

2.1.3. Causas de la desnutrición crónica

La desnutrición en los niños suele tener múltiples causas, siendo la pobreza, la ignorancia y la falta de alimentos las principales en nuestro entorno. Aunque los niños alimentados con leche materna suelen prosperar durante los primeros seis a siete meses, problemas surgen posteriormente. La falta de conocimiento o recursos para complementar la alimentación puede llevar al estancamiento del crecimiento y eventualmente a la desnutrición.

La desnutrición infantil surge de una combinación de factores interrelacionados:

Ingesta inadecuada de nutrientes: Se debe a una dieta desequilibrada o carente en nutrientes esenciales, influenciada por la pobreza, la falta de acceso a alimentos variados o desconocimiento sobre nutrición adecuada.

Factores socioeconómicos: Las limitaciones económicas impiden el acceso a alimentos de calidad y a servicios de salud adecuados.

Problemas de salud: Enfermedades como diarrea, infecciones respiratorias y parásitos pueden aumentar las necesidades nutricionales y reducir la capacidad del cuerpo para absorber nutrientes.

Acceso deficiente a atención médica: La falta de servicios de salud accesibles y de calidad puede llevar a diagnósticos y tratamientos tardíos, afectando el estado nutricional.

Lactancia materna insuficiente: La falta de apoyo para la lactancia materna exclusiva durante los primeros seis meses puede contribuir a problemas nutricionales en los lactantes.

Saneamiento e higiene deficientes: La ausencia de agua potable y malas condiciones higiénicas elevan el riesgo de infecciones que afectan la nutrición.

Falta de educación y conocimientos: El desconocimiento sobre la nutrición adecuada y el cuidado infantil entre los cuidadores aumenta el riesgo de desnutrición.

Eventos catastróficos y crisis: Desastres naturales, conflictos y crisis humanitarias interrumpen el acceso a alimentos y servicios de salud, agravando la desnutrición.

Factores culturales: Algunas prácticas culturales relacionadas con la alimentación y el cuidado infantil pueden impactar negativamente la ingesta nutritiva.

En la desnutrición, la única curva que sigue su curso normal es la de la edad, lo que causa una notable divergencia con las curvas de peso y talla. A los doce meses, el niño a menudo mantiene el peso que tenía a los seis meses. Al dejar la lactancia y la transición a una alimentación mixta, que puede ser inadecuada tanto en cantidad como en calidad, deteriora la capacidad del organismo para absorber nutrientes. Este deterioro contribuye a un descenso significativo en la curva de peso. Además, el terreno debilitado facilita la aparición de infecciones, que pueden afectar

tanto el sistema digestivo como otras áreas, complicando aún más la situación con diarreas recurrentes que agotan las ya limitadas reservas del organismo.

2.1.4. El cálculo de la desnutrición

Puede variar dependiendo de los criterios utilizados y la población específica que se esté evaluando. Peso para la talla (niños menores de 5 años): Se compara el peso del niño con una referencia de peso estándar para su talla. Si el peso del niño está por debajo de ciertos percentiles establecidos, se considera que el niño está desnutrido.

2.1.4.1. *Talla para la edad (niños menores de 5 años):*

Se compara la altura del niño con una referencia de altura estándar para su edad. Si la altura del niño está por debajo de ciertos percentiles establecidos, puede indicar desnutrición crónica.

2.1.4.2. *Índice de masa corporal (IMC) (adultos y niños mayores de 5 años):*

Se calcula dividiendo el peso (en kilogramos) entre la altura (en metros) al cuadrado. Un IMC por debajo de cierto umbral (por ejemplo, 18.5 kg/m^2) se considera indicativo de desnutrición en adultos.

2.1.4.3. *Circunferencia del brazo:*

Se mide la circunferencia del brazo en la parte superior del brazo (por encima del músculo) con una cinta métrica. Una circunferencia del brazo por debajo de cierto umbral puede indicar desnutrición.

2.1.4.4. *Peso perdido (en un período de tiempo específico):*

Se compara el peso actual con el peso previo de la persona. Una pérdida significativa de peso en un período de tiempo relativamente corto puede indicar desnutrición.

2.1.4.5. *Evaluación clínica y bioquímica:*

Los médicos también pueden realizar evaluaciones clínicas y análisis de laboratorio para determinar el estado nutricional de una persona, incluyendo la evaluación de los niveles de nutrientes en la sangre.

2.1.5. Tipos de desnutrición

2.1.5.1. *Desnutrición de Primer Grado:*

En esta etapa inicial, el niño puede mostrar un cambio sutil en su comportamiento, pasando de estar alegre y bien dormido a volverse irritable y descontento. Aunque es difícil notar una pérdida de peso significativa sin una balanza, se observa un estancamiento en el crecimiento ponderal durante varias semanas. El niño puede presentar ligera constipación, sin diarrea ni vómitos notables. Aunque las infecciones son poco comunes en esta fase, el peso del niño tiende a estancarse o a decrecer lentamente.

2.1.5.2. *Desnutrición de Segundo Grado:*

En esta fase, la pérdida de peso se vuelve más evidente, alcanzando una reducción del 10 al 15% o más. El niño presenta hundimiento de la fontanela y de los ojos, con tejidos corporales flácidos y menos elásticos. Hay una mayor susceptibilidad a infecciones respiratorias y otitis, y el niño muestra una irritabilidad notable. También pueden surgir trastornos diarreicos y signos de deficiencia de proteínas, como edemas. La falta de atención adecuada puede llevar a una intolerancia a los alimentos, complicando aún más la situación nutricional.

2.1.5.3. *Desnutrición de Tercer Grado:*

En este estado avanzado, los síntomas de las etapas anteriores se intensifican. El niño presenta una apariencia debilitada, con hundimiento prominente de los ojos y una piel seca y arrugada que revela los huesos subyacentes. Aunque algunos niños en esta etapa no muestran

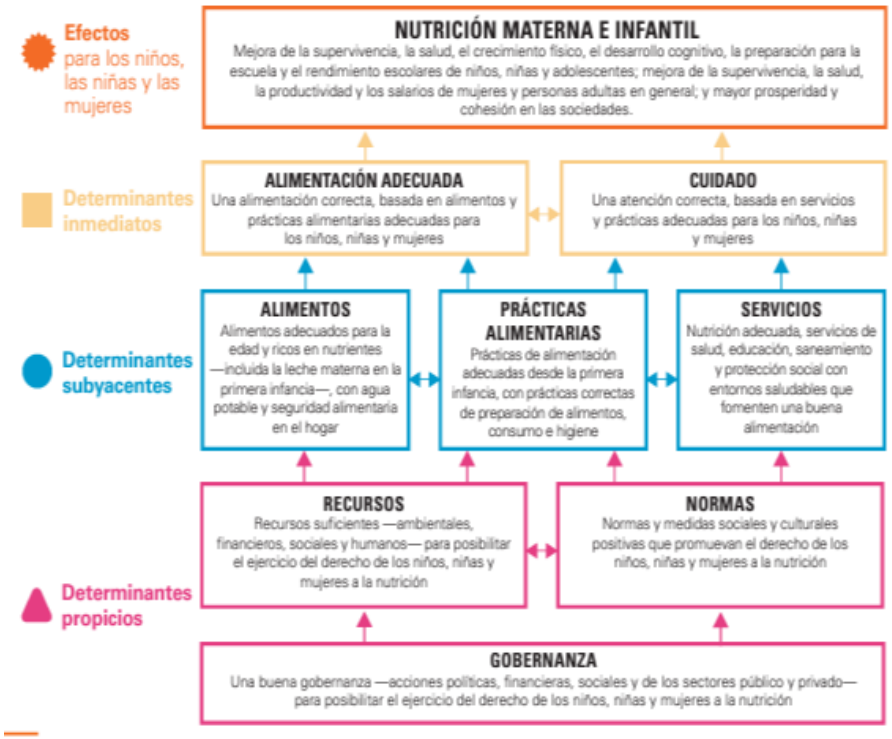
edemas, otros pueden desarrollar hinchazón y cambios en la piel. La intolerancia a los alimentos y los vómitos son frecuentes, y las infecciones se vuelven más severas.

2.1.6. Prevención de la desnutrición crónica infantil

La prevención de la desnutrición crónica en los niños es realmente esencial para asegurar un desarrollo adecuado y un futuro sostenible para los niños. Para hacer frente a este problema de manera efectiva, es muy necesario adquirir alimentos saludables y tener una dieta equilibrada desde una edad temprana.

Por una dieta equilibrada, se entiende la lactancia materna exclusiva durante los primeros seis meses de la edad y luego una dieta complementaria con alimentos sólidos adecuados. Además, la educación nutricional y el desarrollo de hábitos alimentarios racionales también deben ser animados entre los padres y cuidadores. El acceso a agua potable y un entorno higiénico adecuado no son menos importantes y son factores críticos para prevenir enfermedades que pueden afectar indirectamente la nutrición. El acceso a instalaciones de tratamiento de la enfermedad competentes y el monitoreo de cualquier deficiencia nutricional latente no son de ninguna manera menos importantes. Además, las acciones deben abordar las condiciones socioeconómicas, que afectan la seguridad alimentaria con medidas que mejoren la vida de las familias necesitadas. El riesgo de desnutrición crónica y la calidad de vida de los niños pueden asegurarse mucho más con un esfuerzo totalmente coordinado. (ONU, 2022)

Figura 1 *Determinantes de la nutrición materna infantil*



Fuente: (UNICEF, 2022)
Nota. El gráfico representa un marco para prevenir la desnutrición crónica infantil.

2.2. El Instituto Nacional de Estadística y Censos (INEC)

El Instituto Nacional de Estadística y Censos de Ecuador se encarga de publicar uno de los informes críticos para los estudios sobre Salud y Nutrición Infantil; la encuesta ENSANUT 2018 -2019, cuyo informe final expone datos e informes amplios del estado de salud y nutrición en Ecuador, incluidos los indicadores correspondientes a la desnutrición en la infancia. También público el Boletín de Salud Infantil 2021 es una guía de actualización en cuanto a la salud infantil, especialmente con respecto a la nutrición y la desnutrición. Las Estadísticas de Salud Pública y Nutrición 2022, por otro lado, son hechos relacionados con la salud pública, en este caso, la desnutrición en varias cohortes de edades. Así, los dos documentos son de extrema importancia en un análisis detallado del estado nutricional y los problemas dentro del marco de Ecuador.

2.2.1. Encuesta Nacional de Salud, Salud reproductiva y Nutrición

Las Encuestas Nacionales de Salud, Salud Reproductiva y Nutrición (ENSANUT) son herramientas cruciales para la recopilación de una base de datos. Las encuestas proporcionan información detallada sobre diversos aspectos, como el estado de salud general, las prácticas de salud reproductiva, y los niveles de nutrición en la población. ENSANUT recopila datos a través de encuestas representativas a nivel nacional, que incluyen entrevistas y mediciones físicas, permitiendo a los investigadores y responsables de políticas diseñar e implementar intervenciones más efectivas para mejorar la salud pública. Además, estas encuestas son esenciales para monitorear los progresos en la lucha contra problemas de salud y nutrición, y para ajustar las estrategias y políticas en función de los resultados obtenidos.

2.2.1.1. *Objetivo de la encuesta*

La encuesta ENSANUT tiene como propósito principal ofrecer una visión completa de las condiciones de salud, nutrición y salud reproductiva en la población. Su objetivo es identificar deficiencias y problemas en estos ámbitos para orientar la creación de políticas y programas de salud pública. Proporcionando datos precisos y actualizados, ENSANUT permite evaluar el impacto de intervenciones anteriores y detectar nuevas necesidades. Esta información facilita la toma de decisiones fundamentadas, promoviendo la adopción de estrategias más efectivas y adaptadas a las condiciones locales, lo que contribuye a mejorar la calidad de vida y la salud en general.

2.2.1.2. *Metodología*

El marco conceptual y metodológico de ENSANUT se basa en un enfoque integral para la evaluación de la salud y la nutrición. Conceptualmente, se estructura en torno a los determinantes de la salud y nutrición, que incluyen factores socioeconómicos, ambientales, y de acceso a

servicios de salud. Metodológicamente, la encuesta utiliza un diseño muestral representativo que permite generalizar los resultados a nivel nacional. Se emplean técnicas de recolección de datos cuantitativos, como entrevistas estructuradas y mediciones físicas, además de métodos cualitativos para captar información contextual. La integración de estos enfoques permite obtener una visión completa y precisa de los factores que influyen en la salud y nutrición de la población. (INEC, 2018)

2.2.1.3. *Delimitación del estudio*

La delimitación del estudio en el contexto de ENSANUT se refiere a los límites específicos que se establecen para enfocar el análisis y asegurar la precisión y relevancia de los datos. Esto incluye la definición del ámbito geográfico, que puede ser nacional, regional o local, y el rango de población estudiada, que puede abarcar diferentes grupos etarios, socioeconómicos y demográficos. También implica la selección de variables de interés y el período de tiempo durante el cual se recopilan los datos.

Estas delimitaciones son esenciales para garantizar que los resultados sean aplicables a las áreas de interés y para ajustar las intervenciones y políticas en consecuencia.

2.3. Ciencia de datos

La teoría de ciencias de datos se centra en el análisis, la interpretación y la extracción de conocimientos a partir de grandes volúmenes de datos. Esta disciplina integra técnicas de estadística, informática y matemáticas para modelar y entender fenómenos complejos. Los conceptos clave incluyen la recolección de datos, la limpieza y preparación de datos, el análisis exploratorio, y el uso de algoritmos de machine learning y modelos predictivos. La ciencia de datos permite a las organizaciones y a los investigadores hacer predicciones, descubrir patrones y tomar decisiones informadas basadas en datos empíricos.

2.3.1. Conversión de datos sin procesar en conocimientos prácticos

2.3.1.1. *Tipos de análisis*

Análisis Descriptivo: Responde "¿Qué pasó?" usando datos históricos y actuales.

Análisis de Diagnóstico: Responde "¿Por qué sucedió esto?" para entender éxitos o fracasos.

Análisis Predictivo: Predice eventos futuros basados en modelos complejos.

Análisis Prescriptivo: Optimiza procesos recomendando acciones basadas en predicciones.

Análisis prescriptivo: facilita la toma de decisiones informadas y maximiza el valor que se puede obtener a partir del conocimiento derivado de los datos.

2.3.1.2. *Desafíos comunes en el análisis*

Uno de los principales es la dificultad para encontrar empleados que posean las habilidades técnicas necesarias, como el manejo de herramientas analíticas y la capacidad para interpretar grandes volúmenes de datos. Además, incluso los analistas con experiencia a menudo enfrentan el reto de comunicar sus hallazgos de manera clara y comprensible para los gerentes y tomadores de decisiones, quienes pueden no tener el mismo nivel de conocimiento técnico. Esta brecha de comunicación puede limitar la efectividad de las recomendaciones basadas en datos y dificultar la implementación de estrategias informadas.

2.3.1.3. *Data Wrangling*

Extracción de datos: Implica identificar los conjuntos de datos más relevantes para el problema que se desea resolver y extraer una cantidad suficiente de datos que permitan abordar el problema con precisión. (Alteryx, 2024)

Preparación de datos: Consiste en limpiar los datos crudos, eliminando inconsistencias, valores atípicos y errores, para asegurarse de que los datos sean adecuados para el análisis.

Gobernanza de datos: Se deben aplicar estándares de gobernanza de datos para garantizar que los datos almacenados tengan la granularidad y calidad adecuadas, facilitando su uso en análisis posteriores.

Arquitectura de datos: Es fundamental contar con una arquitectura de TI eficiente, ya que los datos aislados en repositorios fijos y separados pueden dificultar su acceso y análisis.

Cuando se prepare para analizar datos, siga este proceso de 6 pasos para la preparación de datos:

Importar: Lea conjuntos de datos relevantes en su aplicación.

Limpiar: Elimine registros perdidos, duplicados y fuera de rango, y también estandarice la carcasa.

Transformar: En este paso, trata los valores faltantes, maneja los valores atípicos y escala sus variables.

Procesamiento: Implica el análisis de datos, la recodificación de variables, la concatenación y otros métodos para reformatear su conjunto de datos para prepararlo para el análisis.

Acceso: En este paso, simplemente crea un registro que describe su conjunto de datos.

Haga una copia de seguridad: Es almacenar una copia de seguridad de este conjunto de datos procesados para que tenga una versión limpia y nueva pase lo que pase. (Staff, 2023)

2.3.2. Tomar medidas sobre los conocimientos empresariales

Es absolutamente necesario que una organización esté preparada y equipada para cambiar, evolucionar y progresar cuando estén disponibles nuevos conocimientos empresariales se debe asegurar de contar con las siguientes personas y sistemas implementados y listos para funcionar:

Datos correctos, momento correcto, lugar correcto

Científicos de datos y analistas de negocios centrados en los negocios

Gestión educada y entusiasta

Cultura organizacional informada y entusiasta

Procedimientos escritos con cadenas de responsabilidad claramente designadas

Avance en tecnología

2.3.3. Diferenciar entre inteligencia empresarial y ciencia de datos.

En la Tabla 1 compara a los científicos de datos y los analistas de negocio. Los científicos de datos se centran en el análisis avanzado de datos estructurados y no estructurados, utilizando herramientas como Hadoop y Python, para decisiones estratégicas a largo plazo. Los analistas de negocio optimizan procesos empresariales a corto plazo, trabajando principalmente con datos estructurados y generando informes tabulares y paneles. Ambos roles difieren en enfoque, tecnologías y productos entregables, pero contribuyen al análisis de datos en la empresa

Tabla 1. *Diferencia entre ciencia de datos y analista de datos*

Aspecto	Ciencia de Datos	Analista de Negocio
Objetivo	Convertir datos sin procesar en conocimientos valiosos para decisiones empresariales a largo plazo.	Optimizar procesos y mejorar la eficiencia y efectividad de los negocios a corto y mediano plazo.
Enfoque	Utiliza análisis y modelado avanzados para descubrir nuevos paradigmas y perspectivas.	Analiza datos empresariales y crea especificaciones para mejorar procesos y sistemas.
Fuentes de Datos	Datos estructurados y no estructurados (datos generados por máquinas, redes sociales).	Principalmente datos estructurados de bases de datos relacionales.
Resultados	Visualizaciones avanzadas de datos y paneles de análisis, pero rara vez informes tabulares.	Informes tabulares, paneles interactivos y especificaciones escritas.
Tecnología	Hadoop, MapReduce, procesamiento paralelo masivo, Python, R, plataformas en la nube.	Bases de datos relacionales, almacenes de datos, tecnologías OLAP y ETL.

Habilidades Requeridas	Estadística, matemáticas, programación, conocimiento del negocio y entorno.	Gestión de requisitos, análisis de procesos de negocio, planificación de mejoras.
Rol en la Organización	Generar conocimientos valiosos a partir de grandes cantidades de datos comerciales complejos.	Identificar procesos que necesitan mejoras y detallar los cambios necesarios para optimizar resultados.
Productos de Trabajo	Paneles de análisis, visualizaciones de datos, descubrimientos basados en análisis avanzados.	Especificaciones escritas, informes tabulares, paneles de control de apoyo a la toma de decisiones.
Comunicación de Resultados	A través de visualizaciones de datos y descripciones detalladas de hallazgos complejos.	Informes y recomendaciones fáciles de entender para gerentes de negocios.

Fuente: (Wiley & Sons, 2017).

Nota. Esta tabla muestra las principales diferencias entre un analista de datos y un científico de datos.

2.3.4. Análisis exploratorio de datos (eda)

El análisis exploratorio de datos (EDA) es una fase crucial en el proceso de análisis de datos que se centra en la comprensión inicial de los datos mediante la generación de estadísticas descriptivas y visualizaciones. En el contexto de la variable dependiente, EDA implica examinar la distribución de variable dependiente dentro de una muestra de datos, identificar patrones, y detectar anomalías. Esto puede incluir la creación de histogramas, diagramas de caja y gráficos de dispersión para visualizar cómo varía la variable dependiente en relación con las variables independientes.

El objetivo es obtener una comprensión preliminar de las características de la variable dependiente en el conjunto de datos, lo que puede informar el desarrollo de modelos analíticos más sofisticados y guiar la interpretación de los resultados.

2.3.4.1. *Limpieza de datos*

La limpieza de datos es fundamental para garantizar la calidad y fiabilidad del análisis. Este proceso incluye la identificación y corrección de errores como valores faltantes, duplicados, inconsistencias y datos irrelevantes. Los valores faltantes se pueden manejar mediante la

eliminación de registros incompletos o la imputación de datos mediante métodos estadísticos como la media, la mediana o modelos predictivos. Los datos duplicados se eliminan para evitar distorsiones en el análisis, mientras que las inconsistencias y errores tipográficos se corrigen para asegurar la uniformidad. Además, se estandarizan formatos y se verifica la coherencia de los datos. Un tratamiento adecuado de estos aspectos garantiza que los datos sean consistentes, precisos y aptos para el análisis, lo que contribuye a la obtención de resultados más fiables y a una toma de decisiones informada

2.3.4.2. *Eliminación de valores a típicos (outliers)*

Los valores atípicos, o outliers, son observaciones que se desvían significativamente del resto de los datos y pueden influir negativamente en los resultados del análisis. La eliminación de estos valores se basa en la identificación de datos que son inusuales o extremos en comparación con el resto del conjunto. Métodos comunes para detectar outliers incluyen el uso de gráficos de dispersión, el análisis de z-scores, y el cálculo de percentiles para establecer umbrales. Es importante considerar la causa de los outliers antes de eliminarlos, ya que algunos pueden ser indicativos de fenómenos importantes o errores de medición. La decisión de eliminar outliers debe ser justificada y basada en una comprensión clara de cómo estos datos afectan el análisis y las conclusiones

2.3.5. La ingeniería de características (Feature Engineering)

Es un proceso fundamental en el desarrollo de modelos de aprendizaje automático y análisis de datos. Consiste en transformar los datos crudos en características (features) que mejor representen el problema que se quiere resolver, permitiendo que los algoritmos de aprendizaje automático trabajen de manera más eficiente y efectiva.

Aquí algunos aspectos clave de la ingeniería de características:

2.3.5.1. *Selección de características*

Identificar y elegir las características más relevantes y útiles para el modelo, descartando aquellas que no aportan valor o que pueden introducir ruido.

La selección de variables es una etapa crucial en el proceso de modelado de datos que busca identificar las características más relevantes para construir modelos predictivos eficaces. En Python, existen varias bibliotecas que facilitan esta tarea. Entre ellas, Scikit-learn es una de las más populares, ofreciendo herramientas como SelectKBest, RFE (Recursive Feature Elimination), y FeatureImportances para la selección y evaluación de características. Statsmodels también proporciona métodos para la selección de variables basados en modelos estadísticos. Boruta es otra biblioteca que se especializa en la selección de características utilizando un enfoque basado en el algoritmo de Random Forest. Cada una de estas herramientas ofrece distintos enfoques y técnicas, permitiendo a los analistas elegir la más adecuada según la naturaleza de sus datos y el problema en cuestión (Ojeda, 2016).

2.3.5.2. *Transformación de características*

Modificar o combinar características existentes para crear nuevas que sean más representativas del problema, como aplicar logaritmos, normalización o estandarización de datos.

2.3.5.3. *Creación de nuevas características*

Generar nuevas características a partir de las existentes, como la extracción de estadísticas (media, mediana, etc.) o la combinación de múltiples columnas para formar una nueva.

2.3.5.4. *Codificación de variables categóricas*

Transformar variables categóricas en un formato numérico que el modelo pueda procesar, como One-Hot Encoding o Label Encoding.

2.3.5.5. *Reducción de dimensionalidad*

Simplificar el conjunto de características manteniendo la información esencial, por ejemplo, utilizando técnicas como Análisis de Componentes Principales (PCA).

2.3.6. Aprendizaje automático (machine learning)

El aprendizaje automático, o machine learning, es una rama de la inteligencia artificial que se centra en el desarrollo de algoritmos y modelos capaces de aprender de los datos para realizar predicciones o tomar decisiones, sin necesidad de una programación específica para cada tarea. Este proceso implica el ajuste de modelos basados en datos históricos, mejorando así su precisión en la predicción de resultados futuros. Las tres principales categorías de aprendizaje automático son:

2.3.6.1. *Supervisado*

Los modelos se entrenan con datos etiquetados, algunos algoritmos de clasificación utilizados son los siguientes:

Regresión Logística

Árboles de Decisión

Máquinas de Vectores de Soporte (SVM)

K-Vecinos Más Cercanos (K-NN)

Redes Neuronales

Naive Bayes

Bosques Aleatorios (Random Forest)

Gradient Boosting Machines (GBM)

Redes Neuronales Convolucionales (CNN)

Redes Neuronales Recurrentes (RNN)

2.3.6.2. *No supervisado*

Identifica patrones en datos sin etiquetas, para esto se puede utilizar los siguientes algoritmos:

K-means

Jerárquico

DBSCAN

Expectación-Maximización (EM)

PCA

t-SNE

Mapas Auto-Organizativos (SOM)

Clustering Basado en Modelos

Factorización de Matrices

Algoritmos de Asociación

2.3.6.3. *Por refuerzo*

Los modelos aprenden a tomar decisiones optimizando una función de recompensa a través de la interacción con un entorno.

Q-Learning

Deep Q-Network (DQN)

SARSA (State-Action-Reward-State-Action)

Policy Gradient Methods

2.3.7. Región logística

La Regresión Logística es un algoritmo de clasificación utilizado para predecir la probabilidad de una variable dependiente categórica. Es útil para problemas de clasificación

binaria. Se destaca por su capacidad para manejar relaciones lineales y proporcionar probabilidades ajustadas, lo que facilita la interpretación de los resultados. Es menos propensa al sobreajuste comparado con modelos más complejos, y es ampliamente aplicable en diversas áreas como medicina y economía. Aunque es eficaz en muchos casos, puede necesitar técnicas adicionales para manejar relaciones no lineales o multicolinealidad entre variables.

2.3.8. Random forest

Es un algoritmo de aprendizaje supervisado que utiliza una colección de árboles de decisión para realizar predicciones. Cada árbol se construye a partir de una muestra aleatoria de datos y características, y el resultado final se obtiene mediante la agregación de las predicciones de todos los árboles en el bosque. Este enfoque mejora la precisión y robustez del modelo al reducir el riesgo de sobreajuste y aumentar la capacidad de generalización.

Random Forest es especialmente útil para manejar datos con alta dimensionalidad y es resistente al ruido en los datos. Su principal desventaja es la falta de interpretabilidad en comparación con modelos más simples, ya que el proceso de agregación de múltiples árboles puede ser complejo de desglosar.

2.3.9. El algoritmo de Máquinas de Vectores de Soporte (SVM)

Es un poderoso modelo de clasificación utilizado para encontrar el hiperplano óptimo que separa distintas clases en un espacio de características. Su principal fortaleza radica en su capacidad para manejar problemas de clasificación lineal y no lineal mediante el uso de funciones kernel, que transforman los datos a espacios de mayor dimensión. SVM es eficaz en escenarios con alta dimensionalidad y es robusto frente al sobreajuste, especialmente en casos de datos con un número limitado de muestras. Sin embargo, puede ser computacionalmente costoso y menos interpretable en comparación con otros modelos.

2.3.10. Hiperparámetros

Los hiperparámetros son parámetros que se establecen antes del proceso de entrenamiento de un modelo de aprendizaje automático y no se actualizan durante el entrenamiento. Estos parámetros influyen en el funcionamiento del algoritmo y en cómo se ajusta a los datos, afectando directamente el rendimiento y la eficiencia del modelo. A diferencia de los parámetros del modelo, que se aprenden directamente de los datos durante el entrenamiento (como los pesos en una red neuronal), los hiperparámetros deben definirse previamente. (Muñoz & Romero, 2021)

2.3.11. Métricas para evaluar los algoritmos de clasificación

Es fundamental evaluar el rendimiento del modelo y realizar los ajustes necesarios. Algunas de las métricas más utilizadas para este propósito incluyen la matriz de confusión, el accurancy, la Curva ROC (Receiver Operating Characteristic) y el AUC (Área Bajo la Curva ROC) , las curvas de precisión-recall son especialmente valiosas en escenarios con clases desbalanceadas, y los gráficos de importancia de características permiten identificar qué variables tienen un mayor impacto en las predicciones del modelo (Sancho, 2021).

2.3.11.1. *La precisión, o accuracy*

Es una métrica comúnmente utilizada para evaluar el rendimiento de un modelo de clasificación. Se define como la proporción de predicciones correctas (tanto verdaderos positivos como verdaderos negativos) entre el total de predicciones realizadas. En otras palabras, mide qué tan bien el modelo clasifica correctamente las instancias en comparación con el número total de instancias.

2.3.11.2. *Matriz de confusión*

La matriz de confusión es una herramienta fundamental para evaluar el rendimiento de un modelo de clasificación. Esta matriz proporciona una tabla que resume las predicciones del modelo

comparadas con los valores reales, mostrando la cantidad de verdaderos positivos, verdaderos negativos, falsos positivos y falsos negativos. La matriz de confusión permite calcular métricas clave como la precisión, la sensibilidad (recall), la especificidad y la puntuación F1. Analizar esta matriz ayuda a identificar el tipo de errores que el modelo está cometiendo y a entender mejor cómo se comporta en diferentes clases.

2.3.12. Curva ROC

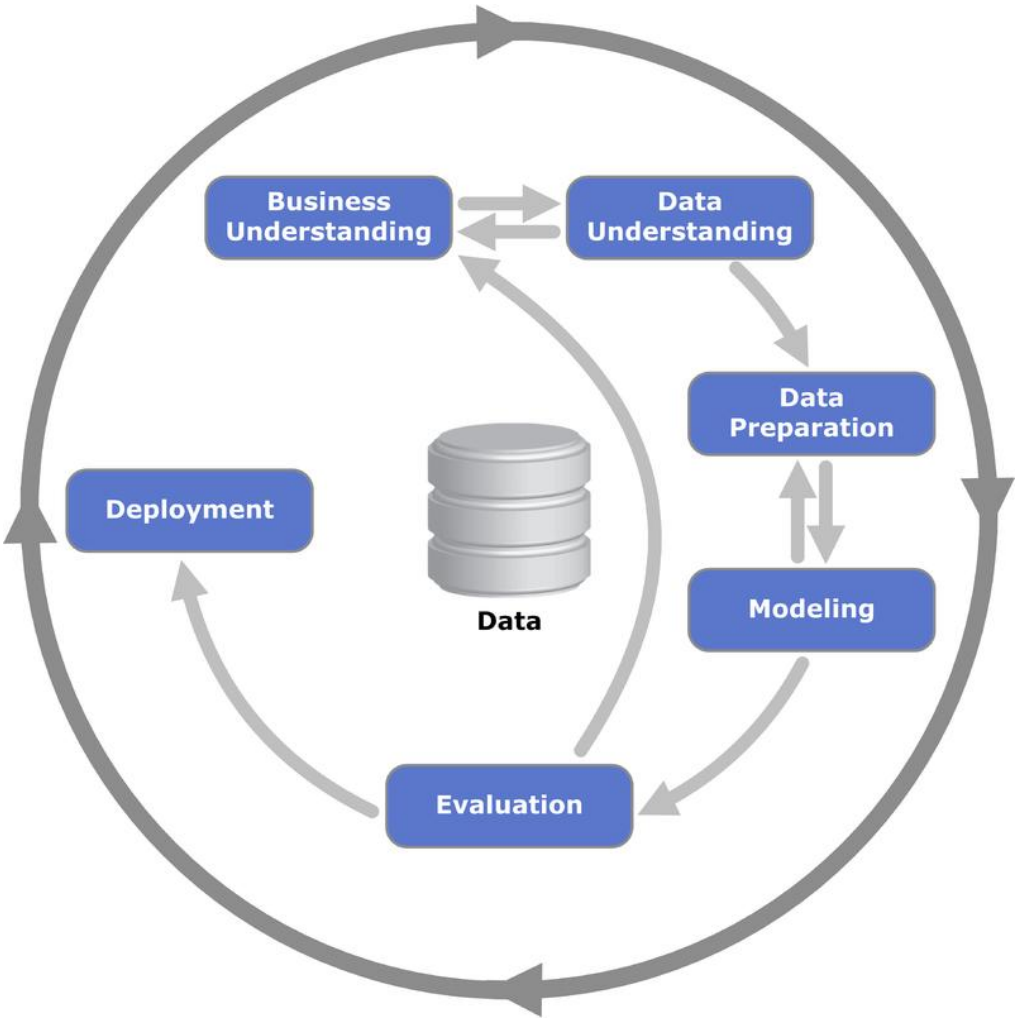
La Curva ROC es una representación gráfica que ilustra el rendimiento de un modelo de clasificación en función de distintos umbrales de decisión. Esta curva traza la tasa de verdaderos positivos (sensibilidad) contra la tasa de falsos positivos (1 - especificidad) para cada umbral. Proporciona una perspectiva integral sobre la habilidad del modelo para diferenciar entre clases. El Área Bajo la Curva ROC (AUC-ROC) mide esta capacidad de discriminación, donde valores más cercanos a 1 indican un mejor rendimiento. Esta herramienta es particularmente útil para comparar diferentes modelos y para determinar el umbral de decisión más adecuado.

2.4. Metodología CRISP DM

La metodología CRISP-DM tal como se muestra en la figura 2, es un estándar ampliamente aceptado en proyectos de minería de datos, que se estructura en seis fases: comprensión del negocio, comprensión de los datos, preparación de los datos, modelado, evaluación y despliegue. El proceso comienza con la identificación de los objetivos del negocio y luego se enfoca en la exploración y preparación de los datos. Posteriormente, se desarrollan modelos predictivos que son evaluados para asegurar que cumplan con los objetivos del proyecto. Finalmente, estos modelos se implementan en un entorno de producción, asegurando su integración y valor para el negocio. Esta metodología es iterativa, permitiendo ajustes y refinamientos en cualquiera de sus fases según sea necesario.

Este enfoque guía a los investigadores desde la definición del objetivo hasta la generación de resultados prácticos y aplicables. Es útil para analizar grandes volúmenes de datos, evaluar modelos predictivos y resolver problemas complejos en diversos campos como salud, economía y ciencias sociales. Además, facilita la documentación y la replicabilidad de los estudios, promoviendo la transparencia y la confiabilidad en los resultados obtenidos. Su flexibilidad lo hace adaptable a múltiples contextos, apoyando la toma de decisiones basadas en datos.

Figura 2 *Metodología CRISP DM*



Fuente: (Vallalta, 2024)

3. Metodología

3.1. Enfoque metodológico

El presente trabajo se sustenta en el enfoque mixto no experimental de tipo longitudinal, porque los resultados de la investigación de campo provienen de la encuesta ENSANUT 2018, los cuales serán sometidos a análisis numéricos con técnicas estadísticas. Este enfoque es cualitativo porque los datos numéricos serán interpretados críticamente a través del Marco Teórico para evaluar su viabilidad. Además, es aplicado, dado que el análisis de datos se centrará en identificar los factores que afectan la desnutrición crónica infantil (DCI) en menores de cinco años. Los datos utilizados se obtuvieron del repositorio del Instituto Nacional de Estadísticas y Censos (INEC), descargados en formato CSV desde su página oficial sobre Salud, Salud Reproductiva y Nutrición. Se utiliza la metodología **CRISP-DM** (Cross-Industry Standard Process for Data Mining), que ofrece un marco estructurado para abordar proyectos de minería de datos. Esta metodología permite analizar, transformar e interpretar datos para identificar patrones relacionados con la desnutrición crónica infantil (DCI).

El enfoque es aplicado, ya que busca generar un modelo predictivo que identifique factores asociados a la DCI en menores de cinco años, con el objetivo de contribuir al diseño de estrategias que combatan este problema.

3.2. Descripción de la metodología CRISP-DM

La metodología CRISP-DM se compone de seis fases iterativas que aseguran un análisis profundo y estructurado de los datos. A continuación, se describe cómo se aplicarán estas fases en la investigación:

3.2.1. Entendimiento del negocio

De la página del Instituto Nacional de Estadísticas y Censos (INEC) se encuentra información estadística que va a servir en el análisis de la DCI como son los tabulados que

contienen los resultados de la encuesta en forma de tablas y cuadros estadísticos y la Base de datos – periodo vigente donde tiene acceso a la base de datos y documentación adicional que permite la interpretación y comprensión de los datos.

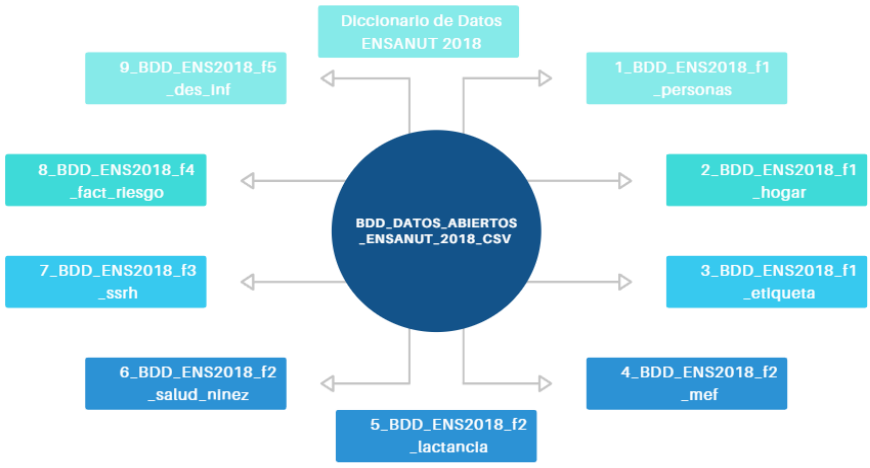
3.2.2. Entendimiento de los datos

Se analizará la base de datos ENSANUT 2018 para identificar las variables clave.

3.2.2.1. *Revisión de archivos CSV proporcionados por el INEC.*

La Figura 1 presenta los archivos CSV que conforman la base de datos ENSANUT 2018, a partir de los cuales se realiza un análisis para seleccionar aquellos que son relevantes en la predicción de la desnutrición crónica infantil en Ecuador.

Figura 3 *Base de datos ENSANUT*



Fuente: (INEC, 2018)

Nota. Esta figura muestra el conjunto de archivos csv que contiene la base de datos ENSANUT

1_BDD_ENS2018_f1_personas.csv: Contiene información detallada sobre características demográficas, estado de salud, acceso a servicios de salud, entre otros.

2_BDD_ENS2018_f1_hogar.csv: Contiene información sobre las condiciones del hogar, el acceso a servicios básicos, y la composición familiar.

3_BDD_ENS2018_f1_etiqueta.csv: Contiene datos sobre información si las personas compran productos basándose en el empaque viendo el semáforo nutricional.

4_BDD_ENS2018_f2_mef.csv: Contiene datos relacionados con las mujeres en edad fértil (MEF). Podría incluir información sobre salud reproductiva, planificación familiar, y otros indicadores relacionados.

5_BDD_ENS2018_f2_lactancia.csv: Contiene datos sobre prácticas de lactancia, incluyendo duración de la lactancia materna, alimentación complementaria, y aspectos relacionados con la salud infantil.

6_BDD_ENS2018_f2_salud_ninez.csv: Contiene datos específicos sobre la salud infantil, cubriendo aspectos como el acceso a servicios de salud para niños, estado nutricional, vacunación, y otros indicadores de salud.

7_BDD_ENS2018_f3_ssrh.csv: Contiene datos sobre temas de salud sexual y reproductiva, incluyendo acceso a métodos anticonceptivos, información sobre infecciones de transmisión sexual, y otros indicadores relacionados.

8_BDD_ENS2018_f4_fact_riesgo.csv: Contiene datos sobre factores de riesgo para la salud, como consumo de tabaco, alcohol, actividad física, y alimentación, que pueden influir en la aparición de enfermedades crónicas.

9_BDD_ENS2018_f5_des_inf.csv: Contiene información sobre el desarrollo infantil, evaluando aspectos como el crecimiento, desarrollo cognitivo, y otros indicadores clave en la primera infancia.

Diccionario de Datos ENSANUT 2018.ods: Es un documento que proporciona detalles sobre las variables y códigos utilizados en los conjuntos de datos de la ENSANUT 2018. Es esencial para la correcta interpretación de los datos.

3.2.2.2. *Preselección de variables:*

Los datos seleccionados contiene información de ENSANUT 2018, se selecciona 5 formularios relacionadas con la salud infantil, características del hogar, persona, mujer en edad fértil y prácticas de lactancia.de los cuales se realiza una preselección de variables de interés de acuerdo a las causas clasificadas por la Organización del Fondo de las Naciones Unidas para la Infancia (UNICEF) , entre las cuales se incluyen variables numéricas y categóricas

3.2.2.3. *Identificación de la variable objetivo:*

El objetivo es predecir la variable "dcronica", que es una variable binaria que toma dos valores 1 si tiene desnutrición crónica y 0 si no tiene desnutrición crónica.

3.2.3. Preparación de los datos

Esta fase implica:

- Limpieza de datos para manejar valores nulos, duplicados y outliers.
- Transformación de variables categóricas en formatos utilizables por los algoritmos de machine learning.
- Selección final de variables de interés utilizando una herramienta avanzada de Python (Featurewiz), diseñada para la selección automática de características (variables) en conjuntos de datos, optimizando el rendimiento del modelo y reduciendo la dimensionalidad

3.2.4. Modelado:

3.2.4.1. *Selección de técnicas de modelado*

Desarrollo de modelos de machine learning, aplicando algoritmos tales como:

Regresión logística

Random Forest,

SVM (Support Vector Machine)

3.2.4.2. *Generar diseño de prueba*

Con función de scikit-learn se divide el conjunto de datos X (características) e y (etiquetas o variable objetivo) en dos subconjuntos: uno para entrenamiento y otro para prueba. Se utilizará el 70% de datos para entrenamiento y el 30% para pruebas. Al realizar esto creamos una base sólida para tomar decisiones sobre si el modelo está listo para el despliegue.

3.2.5. Evaluación

Las Métricas que se utilizan para evaluar el modelo son la curva ROC, acurancy y matriz de confusión.

3.2.5.1. *Curva ROC y AUC para evaluar los modelos*

En la fase de evaluación la curva ROC y el AUC se utilizan para medir la capacidad del modelo para discriminar entre las clases positivas y negativas. Esta métrica es particularmente útil cuando se tienen datos desequilibrados, ya que ofrece una visión más completa del rendimiento del modelo en diferentes umbrales.

3.2.5.2. *Precisión (Accuracy) para evaluar el modelo*

En la fase de evaluación, la precisión es una métrica fundamental para entender qué tan bien el modelo clasifica correctamente las instancias en general. Es especialmente útil cuando las clases están equilibradas. Sin embargo, en casos de desbalance de clases, la precisión puede ser engañosa, por lo que se deben considerar otras métricas en paralelo.

3.2.5.3. *Matriz de Confusión para evaluar el modelo*

En la fase de evaluación del modelo, la matriz de confusión proporciona una visión detallada de los tipos de errores que el modelo está cometiendo. Permite calcular otras métricas derivadas, como precisión, recall y F1-score, y ayuda a identificar posibles áreas de mejora en el

modelo. Es particularmente útil para entender los patrones de error y cómo se distribuyen entre las clases.

El rendimiento de los modelos se medirá mediante métricas como la precisión, el área bajo la curva ROC (AUC), y la matriz de confusión.

3.2.6. Implementación

Finalmente, los resultados serán presentados de manera comprensible, destacando las variables más influyentes en la DCI. Este análisis podrá ser utilizado para sugerir estrategias que reduzcan la incidencia de la desnutrición crónica infantil en Ecuador.

3.3. Fuentes de datos

Se utilizarán los conjuntos de datos ENSANUT 2018 proporcionados por el INEC, organizados en diferentes formularios relacionados con la salud, características del hogar, lactancia y desarrollo infantil.

3.4. Herramientas y software

Para el análisis y modelado se emplearán herramientas como:

- **Python**: Para la limpieza y análisis de datos utilizando librerías como pandas, numpy, y sklearn.
- **Visual Studio Code**: Para el desarrollo del código.
- **Power Bi**: Para la exploración preliminar de los datos.

3.5. Alcance y limitaciones

El alcance del estudio se limita a menores de cinco años en Ecuador, basándose exclusivamente en datos recolectados en ENSANUT 2018. Las limitaciones incluyen posibles sesgos, actualización en los datos y restricciones en la generalización de los resultados a contextos diferentes.

4. Resultados y discusión

4.1. Comprensión del negocio:

- Objetivo: Implementar un modelo de aprendizaje supervisado que permita predecir la desnutrición crónica infantil en el Ecuador.
- Valor: Proponer alguna estrategia de implementación del modelo ganador.

4.2. Entendimiento de los datos:

Este proyecto se enfoca en niños y niñas menores de 5 años de nacionalidad ecuatoriana, para quienes se dispone de información verídica proveniente de la Encuesta Nacional de Salud y Nutrición (ENSANUT) 2018. Se seleccionaron cinco formularios de los cuales se realizó una preselección de variables de interés, siguiendo las causas clasificadas por el Fondo de las Naciones Unidas para la Infancia (UNICEF). Estas variables incluyen tanto datos numéricos como categóricos. El objetivo es predecir la variable "dcronica", que es binaria y toma el valor 1 si el niño presenta desnutrición crónica, y 0 si no la presenta.

4.2.1. Importar las siguientes datasets

Antes de importar los datasets a Visual Studio Code, se renombraron los archivos CSV.

4.2.1.1. *Aensanut_personas.csv*

Esta base de datos cuenta con 168,747 registros y 264 variables. Sin embargo, no se utilizaron todas las variables; en la Tabla 2 se presentan las variables seleccionadas, con un dataset de 168,747 registros y 16 variables.

Tabla 2 *Variables seleccionadas dataset personas*

Variables	Descripción
'_id_viv'	Código único de la vivienda
'_id_hogar'	Código único del hogar
'_id_per'	Código único de la persona
'_f1_s2_9'	Identificación cultural y de costumbres según la persona
'_f1_s2_14'	Presencia del padre en el hogar
'_f1_s2_15'	Presencia de la madre en el hogar

'_f1_s4_41'	Realización de chequeo de salud preventiva en los últimos 30 días
'_f1_s7_4_1'	Registro de peso - primera medición
'_f1_s7_4_2'	Registro de peso - segunda medición
'_f1_s7_4_3'	Registro de peso - tercera medición
'_f1_s7_5_1'	Registro de longitud - primera medición
'_f1_s7_5_2'	Registro de longitud - segunda medición
'_f1_s7_5_3'	Registro de longitud - tercera medición
'_f1_s7_6_1'	Registro de talla - primera medición
'_f1_s7_6_2'	Registro de talla - segunda medición
'_f1_s7_6_3'	Registro de talla - tercera medición

Fuente: (INEC, 2018).

4.2.1.2. *Bensanut _hogar*

La base de datos de hogares tiene un tamaño de 43,311 registros y 120 variables. No se utilizaron todas las variables; en la Tabla 3 se muestran las variables seleccionadas, con un dataset de 43,311 registros y 10 variables.

Tabla 3. *Variables seleccionadas dataset hogar*

Variables	**Descripción**
'id_viv'	Código único de la vivienda
'id_hogar'	Código único del hogar
'f1_s1_3'	Principal material utilizado en la construcción del techo de la vivienda.
'f1_s1_4'	Principal material utilizado en la construcción del piso de la vivienda.
'f1_s1_5'	Principal material utilizado en la construcción de las paredes de la vivienda.
'f1_s1_13'	Tipo de instalación sanitaria disponible en el hogar.
'f1_s1_25'	Fuente principal de agua potable utilizada en el hogar.
'f1_s1_27'	Tiempo necesario para llegar a la fuente de agua potable.
'f1_s1_28'	Disponibilidad de agua potable suficiente en las últimas dos semanas.
'f1_s6_1_6'	Incidencia de escasez de alimentos en el hogar.

Fuente: (INEC, 2018).

4.2.1.3. *Densanut_mef.csv*

La base de datos MEF cuenta con 48,700 registros y 536 variables. No se utilizaron todas las variables; en la Tabla 4 se presentan las variables seleccionadas, junto con su descripción, resultando en un dataset de 48,700 registros y 13 variables.

Tabla 4. *Variables seleccionadas dataset mef*

Variables	Descripción
'_id_viv'	Código único de la vivienda
'_id_hogar'	Código único del hogar
'_id_per'	Código único de la persona
'_f2_s1_101'	Edad actual de la persona en años.
'_f2_s2_208_3'	Número total de hijos que viven actualmente con el encuestado.
'_f2_s2_211_3'	Total de hijos que han fallecido.
'_f2_s2_212_2'	Número de hijos que fallecieron antes del nacimiento (mortinatos).
'_f2_s2a_224'	Cantidad de sesiones educativas sobre alimentación complementaria recibidas para el niño/niña.
'_f2_s5_513_2'	Episodios de desmayos experimentados.
'_f2_s5_513_4'	Episodios de convulsiones experimentados.
'_f2_s5_513_8'	Casos de preeclampsia/eclampsia debido a presión arterial alta.
'_f2_s5_513_9'	Casos de infección en las vías urinarias.
'_f2_s5_516_7'	Casos de infección generalizada (sepsis).

Fuente: (INEC, 2018).

4.2.1.4. *Eensanut_lactancia.csv*

Esta *dataset* contiene 11,293 registros y 88 variables. No se utilizaron todas las variables; en la Tabla *5* se presentan las variables seleccionadas, con un dataset de 11,293 registros y 22 variables.

Tabla 5. *Variables seleccionadas dataset lactancia*

Variables	Descripción
'_id_viv'	Código único de la vivienda
'_id_hogar'	Código único del hogar
'_id_per'	Código único de la persona
'_id_hijo'	Código único del hijo

'_f2_s3a_302'	Indica si el último hijo/a recibió leche materna al nacer.
'_f2_s3a_303a'	Indica si el niño recibió leche materna de otra fuente, además de la madre.
'_f2_s3a_304'	Tiempo desde el nacimiento hasta el inicio de la lactancia materna.
_'f2_s3c_307_2'	Duración en meses de la lactancia materna exclusiva.
'_f2_s3d_311f_2'	Frecuencia de consumo de yogurt durante el día o noche anterior.
'_f2_s3d_311b_2'	Frecuencia de consumo de leche de fórmula durante el día o noche anterior.
'_f2_s3d_311c_2'	Frecuencia de consumo de leche en polvo durante el día o noche anterior.
'_f2_s3d_311d_2'	Frecuencia de consumo de jugos naturales durante el día o noche anterior.
'_f2_s3d_311e_2'	Frecuencia de consumo de sopa durante el día o noche anterior.
'_f2_s3d_311g_2'	Frecuencia de consumo de colada durante el día o noche anterior.
_'f2_s3d_312'	Indica si se consumió algún alimento sólido o semisólido durante el día o noche anterior.
'_f2_s3d_313_1'	Tipo de alimento sólido consumido durante el día o noche anterior.
'_f2_s3d_315'	Frecuencia de consumo de alimentos sólidos, semisólidos o suaves durante el día o noche anterior.
'_f2_s3_322'	Indica si se recibieron al menos dos tomas de leche artificial o animal durante el día o noche anterior.
'_f2_s3c_307_1'	Duración en años de la lactancia materna exclusiva.
'_f2_s3c_307_3'	Duración en días de la lactancia materna exclusiva.
'_f2_s3d_311a_1'	Indica si se consumió agua pura durante el día o noche anterior.
'_f2_s3d_311a_2'	Frecuencia de consumo de agua pura durante el día o noche anterior.

Fuente: (INEC, 2018).

4.2.1.5. *Fensanut_salud_ninez.csv*

Esta base de datos cuenta con 20,510 registros y 350 variables. No se utilizaron todas las variables; en la Tabla 5 se muestran las variables seleccionadas, con un dataset de 20,510 registros y 44 variables. Se aplicó un filtro a estos datos, seleccionando únicamente los casos de niños menores de 5 años y nacidos vivos, lo que resultó en un total de 20,356 registros y 44 variables.

Tabla 6. *Variables seleccionadas dataset salud ninez*

Variables	**Descripción**
'_area'	Zona
'_id_viv'	Código único de la vivienda
_'id_hogar'	Código único del hogar
'_id_per'	Código único de la persona
'_id_hijo'	Código único del hijo
'_f2_s4a_402_'	Estado de vida del niño
'_f2_s4b_406_'	Presencia de controles prenatales durante el embarazo
'_f2_s4b_418_'	Consumo de micronutrientes durante el embarazo
'_f2_s4b_419a_'	Frecuencia de vacunación contra el tétano
'_f2_s4b_419a_'	Asesoría sobre lactancia materna durante el control prenatal
'_f2_s4b_419d_'	Asesoría sobre higiene en la preparación de alimentos
'_f2_s4b_419b_'	Asesoría sobre el uso de micronutrientes (hierro, ácido fólico)
'_f2_s4e_441a_'	Tiempo transcurrido hasta el primer control post parto (días)
'_f2_s4e_441b_'	Tiempo transcurrido hasta el primer control post parto (semanas)
'_f2_s4e_441c_'	Tiempo transcurrido hasta el primer control post parto (meses)
'_f2_s4d_433b_	Número de registros de curva de crecimiento en el carnet
'_f2_s4f_449dias_'	Tiempo hasta el primer control médico después del nacimiento (días)
'_f2_s4f_449semanas_'	Tiempo hasta el primer control médico después del nacimiento (semanas)
'_f2_s4f_449meses_'	Tiempo hasta el primer control médico después del nacimiento (meses)
'_f2_s4f_451num_1'	Número de controles en el carnet del último nacido vivo

'_f2_s4f_454meses_'	Edad en meses en que se comenzó a dar leche materna
'_f2_s4f_454dias_'	Edad en días en que se comenzó a dar leche materna
_'f2_s4g_458_'	Duración de la diarrea en días
'_f2_s4g_461_'	Presencia de sangre en la diarrea
'_f2_s4h_474_'	Días de enfermedad en el último período
'_f2_s4i_482_'	Administración de desparasitantes en los últimos 6 meses
'_f2_s4i_483_'	Recepción de hierro en polvo para prevenir anemia en el último año
'_f2_s4j_4871a1_'	Registro de vacunación BCG
'_f2_s4j_4872a1_'	Registro de vacunación contra Hepatitis B
'_f2_s4j_4875a1_'	Registro de vacunación Pentavalente 1
'_f2_s4j_4876a1_'	Registro de vacunación Pentavalente 2
'_f2_s4j_4877a1_'	Registro de vacunación Pentavalente 3
'_f2_s4j_4878a1_'	Registro de vacunación OPV 1
'_f2_s4j_4879a1_'	Registro de vacunación OPV 2
_'f2_s4j_48710a1_'	Registro de vacunación OPV 3
'_f2_s4j_48711a1_'	Registro de vacunación Neumococo 1
'_f2_s4j_48712a1_'	Registro de vacunación Neumococo 2
'_f2_s4j_48713a1_'	Registro de vacunación Neumococo 3
'_f2_s4j_48714a1_'	Registro de vacunación SRP 1
'_f2_s4j_48715a1_'	Registro de vacunación SRP 2
'_edadmeses'	Edad calculada en meses
'_nivins_mef'	Nivel de educación de la madre
'_dcronica'	Indicador de desnutrición crónica infantil
'_f2_s4b_420_'	Semanas de embarazo en el primer control prenatal
'_f2_s4b_421_'	Número de controles prenatales realizados antes del parto

Fuente: (INEC, 2018).

4.2.2. Unión de la base de datos

Las bases de datos se unieron utilizando la función pd.merge() de la biblioteca pandas.La fusión se basa en las columnas comunes 'id_viv', 'id_hogar', 'id_per' y ‘id hijo’ que deben estar presentes en los Dataset. Se usa la opción 'inner' que asegura que solo se incluyan las filas donde los valores de estas columnas coincidan en ambos Dataset, también se usa la opción 'left' que asegura que todas las filas del Dataset se mantengan en el resultado, y solo se agregan datos de donde haya coincidencias en las columnas especificadas. Las filas que no tienen coincidencia en

el Dataset recibirán valores NaN. El resultado de esta operación se almacena en un nuevo dataset llamado bddtotal, que contiene únicamente las filas con coincidencias entre los conjuntos de datos, resultando en un total de 20,356 registros y 93 variables en estudio.

4.3. Preparación de los datos:

Limpieza de datos aplicando técnicas tales como:

4.3.1. Verificar datos faltantes

Para visualizar los datos faltantes, se utilizó la función missing_data = bddtotal.isnull().que crea un nuevo Dataset missing_data, que contiene valores booleanos (True o False). Cada celda en missing_data indica si el valor correspondiente en bddtotal es nulo (NaN). Si el valor en bddtotal es nulo, la celda en missing_data será True; si no es nulo, será False. Esto permite identificar y analizar fácilmente los valores faltantes en el Dataset bddtotal y con la función isnull().sum() que suma los valores True nos da el resultado de los datos faltantes de cada columna.

4.3.2. Imputación de valores faltantes.

Eliminación de valores faltantes

Se procede a eliminar las variables cuyo porcentaje de datos nulos supera el 50% de las observaciones. Luego, se eliminan los registros con valores faltantes en la variable dcronica. Como resultado, el dataset queda con 19,273 registros y 62 variables.

Imputación con la media:

Para las variables numéricas restantes, con un porcentaje de valores nulos inferior al 50%, se reemplazan los valores faltantes con la media de los valores existentes en cada columna.

Imputación por la Moda

Para las variables categóricas restantes, con un porcentaje de valores nulos inferior al 50%, se reemplazan los valores faltantes y las respuestas "no sabe/no responde" con el valor más

frecuente en la columna. Esta técnica se utiliza bajo la suposición de que la categoría más común es representativa de los datos ausentes.

4.3.3. Feature engineering

4.3.3.1. *Transformación de Datos:*

Se procederá a unificar las variables _f2_s4f_449meses_, _f2_s4f_449semanas_ y _f2_s4f_449dias_ en una sola variable llamada control_total_dias, que representará el tiempo hasta el primer control médico después del nacimiento en días.

Asimismo, se combinarán las variables _f2_s4f_454meses_ y _f2_s4f_454dias_ en una nueva variable denominada lactancia_total_dias, que indicará la edad en días en la que se comenzó a dar leche materna.

Transformar la variable ‘edadmeses’ a ‘edad_dias'

4.3.3.2. *Binarización*

Se convierte variables categóricas en valores numéricos binarios, como 0 y 1. Esta transformación es especialmente útil ya que se tiene datos como "si" y "no". Al reemplazar estos valores con números, la binarización facilita el uso de los datos en modelos de machine learning, que requieren entradas numéricas. Se transforman *los valores "si" por 1 y "no" por 0.*

4.3.3.3. *Reducción de variables*

Se realiza este proceso para disminuir el número de características en el conjunto de datos, manteniendo la información relevante para los modelos de machine learning, para esto utilizamos la herramienta Featurewiz que utiliza métodos avanzados, como la importancia de características basada en modelos (model-based feature importance), para seleccionar automáticamente las características más relevantes del dataset como resultado nos da 20 características importantes para el estudio.

Tabla 7. *Características seleccionadas en featurewiz*

Variable	Descripción
'_f1_s1_28'	Disponibilidad de agua potable suficiente en las últimas dos semanas.
'_f1_s7_4_1'	Registro de peso - primera medición
'_f1_s7_4_2'	Registro de peso - segunda medición
'_f1_s7_6_1'	Registro de talla - primera medición
'_f1_s7_6_2'	Registro de talla - segunda medición
'_f2_s1_101'	Edad actual de la persona en años.
'_f2_s4b_420_'	Semanas de embarazo en el primer control prenatal
'_f2_s4b_421_'	Número de controles prenatales realizados antes del parto
'_f2_s2_208_3'	Número total de hijos que viven actualmente con el encuestado.
'lactancia_total_dias'	Edad en días en que se comenzó a dar leche materna
'control_total_dias'	Tiempo hasta el primer control médico después del nacimiento (días)
'edad_dias'	Edad días
'_f1_s2_9'	Identificación cultural y de costumbres según la persona
'_f1_s1_25'	Fuente principal de agua potable utilizada en el hogar.
_'nivins_mef'	Nivel de educación de la madre
'_f1_s1_4'	Principal material utilizado en la construcción del piso de la vivienda.
'f1_s1_13'	Tipo de instalación sanitaria disponible en el hogar.
'f1_s1_3'	Principal material utilizado en la construcción del techo de la vivienda.
'f1_s1_5'	Principal material utilizado en la construcción de las paredes de la vivienda.
'area'	Zona

Fuente: (INEC, 2018).

4.3.3.4. *Codificación de Variables Categóricas:*

Convertir variables categóricas en un formato numérico que los algoritmos de machine learning puedan procesar. Se realiza la técnica label encoding en las siguientes variables:

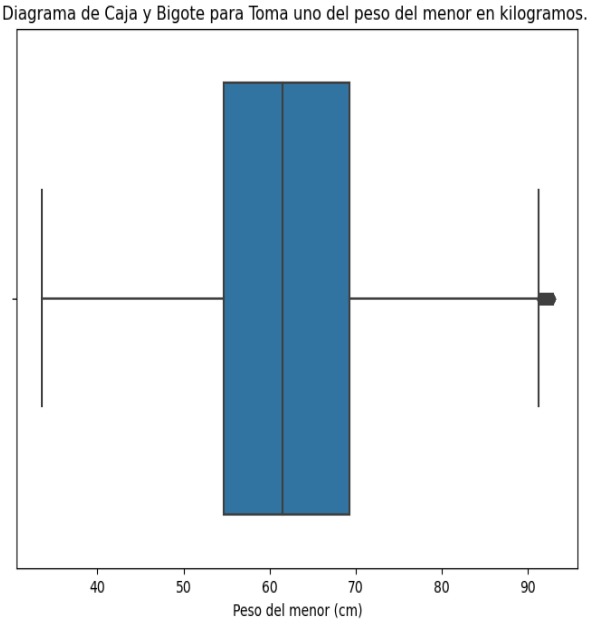
- '_f1_s2_9': Identificación cultural y de costumbres según la persona

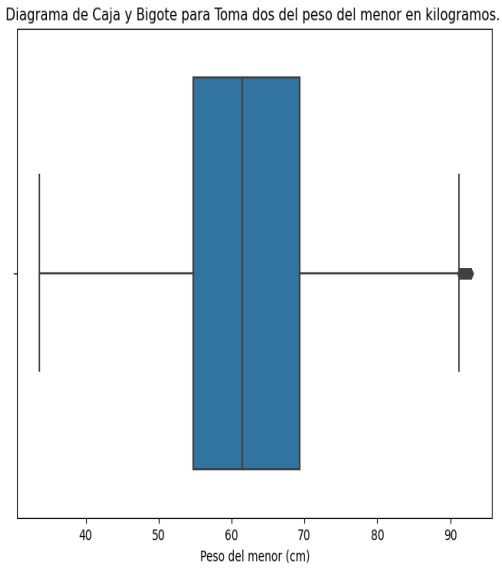
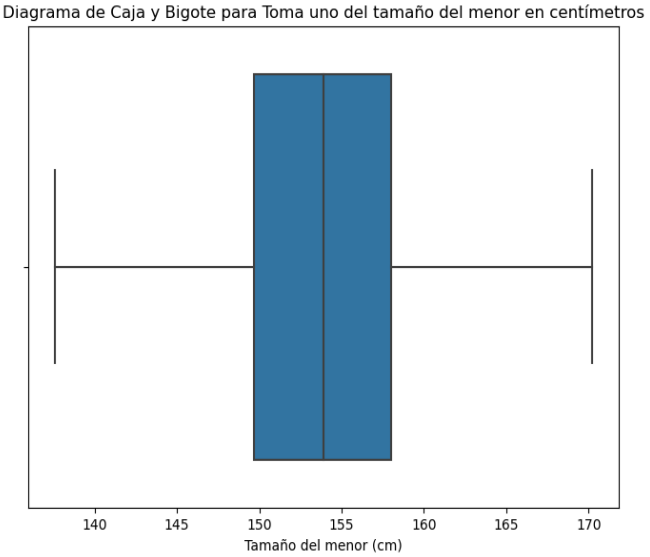
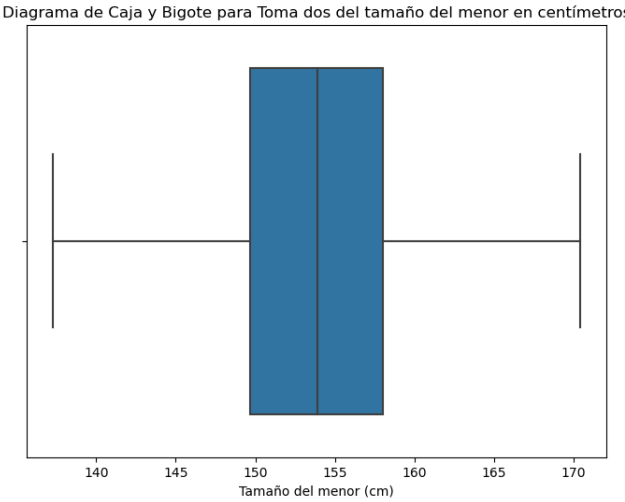
- '_f1_s1_25': Fuente principal de agua potable utilizada en el hogar.
- _'nivins_mef': Nivel de educación de la madre
- '_f1_s1_4': Principal material utilizado en la construcción del piso de la vivienda.
- 'f1_s1_13': Tipo de instalación sanitaria disponible en el hogar.
- 'f1_s1_3': Principal material utilizado en la construcción del techo de la vivienda.
- 'f1_s1_5': Principal material utilizado en la construcción de las paredes de la vivienda.
- 'area': Zona

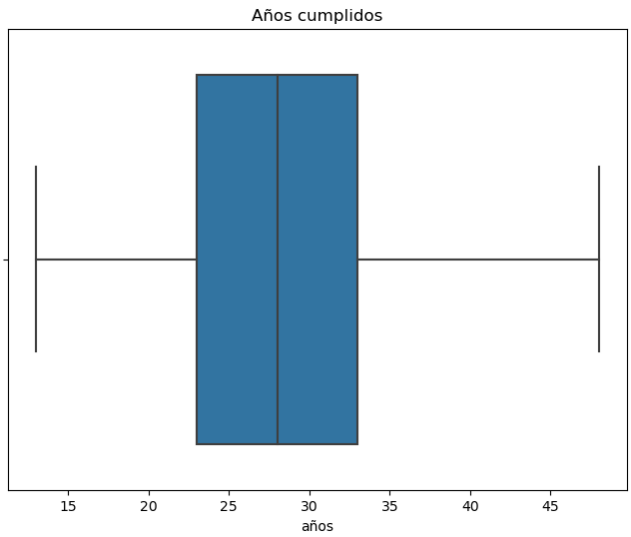
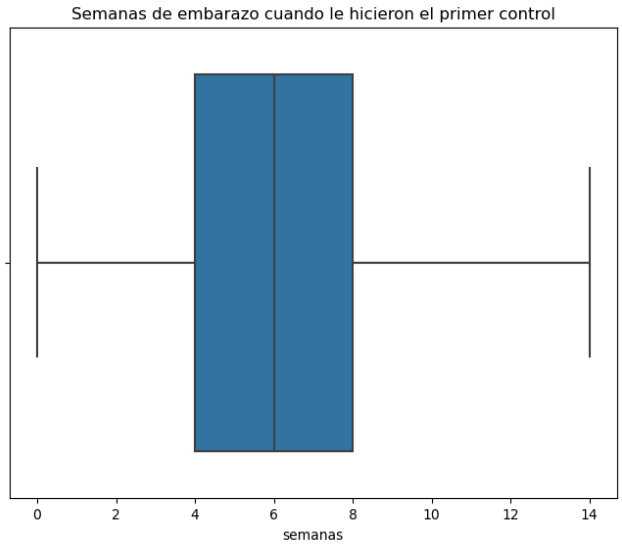
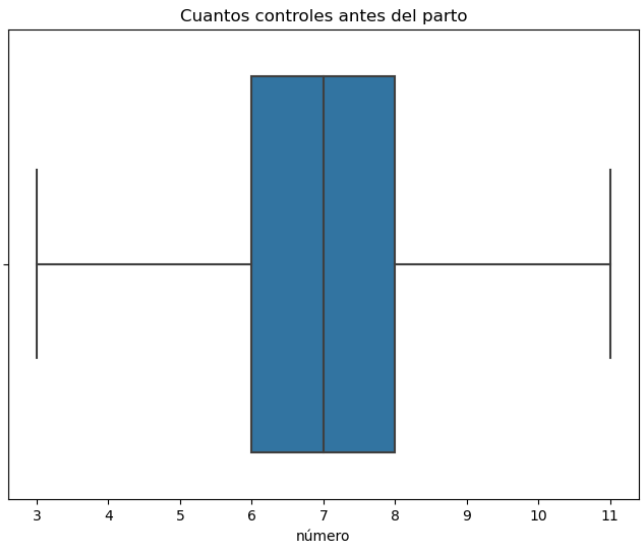
4.3.4. Análisis de variables numéricas

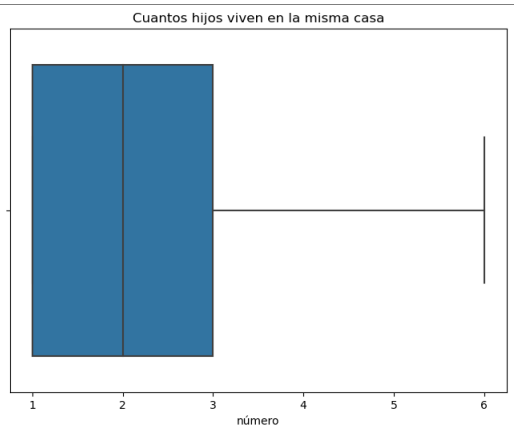
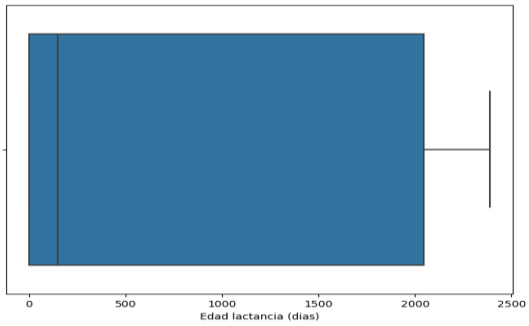
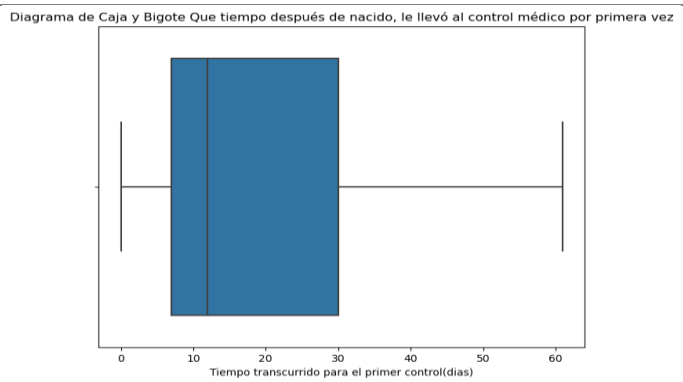
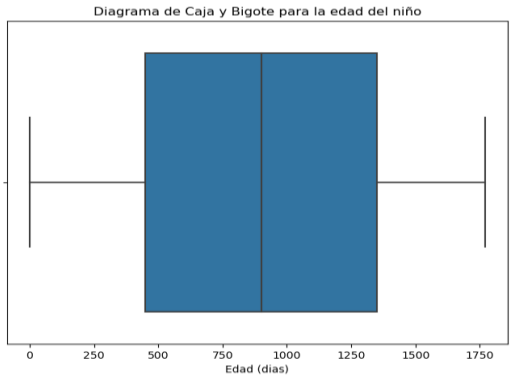
Los outliers son valores atípicos que se alejan significativamente del resto de los datos y pueden afectar negativamente el rendimiento de los modelos de machine learning. Para eliminación de estos se utiliza el rango intercuartil (IQR) que sirve para identificar y eliminar outliers. Se eliminan los valores que están más allá de 1.5 veces el IQR por encima del tercer cuartil o por debajo del primer cuartil. En la tabla 10 se muestra las 11 variables donde se realizó la eliminación de outliers.

Tabla 8. *Variables numéricas analizadas con el diagrama de caja y bigote*

Descripción	**Gráfico**
Registro de peso - primera medición	

Registro de peso - segunda medición	
Registro de talla - primera medición	
Registro de talla - segunda medición	

Edad actual de la persona en años.	
Semanas de embarazo en el primer control prenatal	
Número de controles prenatales realizados antes del parto	

Número total de hijos que viven actualmente con el encuestado.	
Edad en días en que se dio leche materna	
Tiempo hasta el primer control médico después del nacimiento (días)	
Edad días	

Fuente: Tabla realizada por Cleber Puente

Nota. Esta tabla muestra los diagramas de caja y bigote para las variables numéricas.

4.3.5. Análisis de variables categóricas

Para realizar un análisis exhaustivo de las variables categóricas incluidas en el estudio, se desarrolló un dashboard interactivo en Power BI. Este dashboard permitió visualizar y explorar de manera dinámica la distribución y relación de variables como:

Educación: Nivel educativo alcanzado por la madre del niño, categorizado en diferentes niveles

Raza: Identificación cultural y de costumbres.

Área: Zona urbana o rural.

Construcción: Categorías de materiales empleados en paredes, techo y piso.

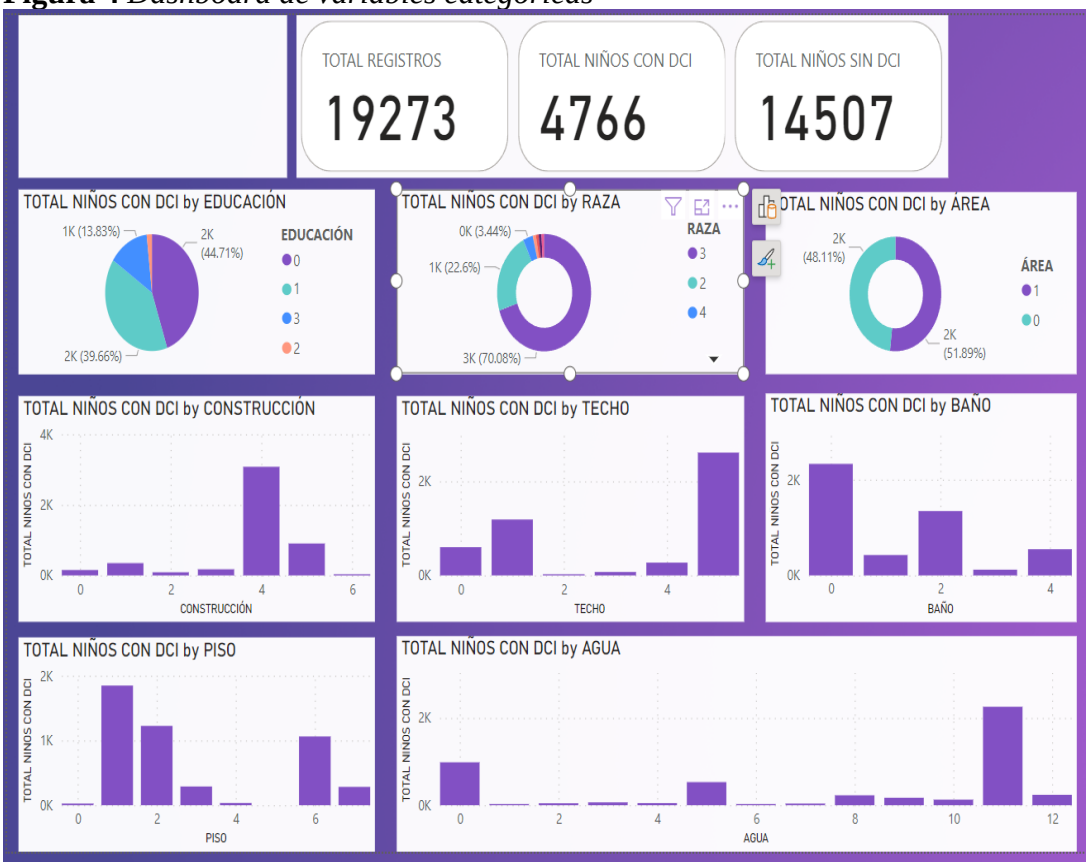
Agua: Disponibilidad y fuente principal de agua.

Baño: Tipo de instalación sanitaria disponible en el hogar.

También se puede visualizar el número total de registros en el conjunto de datos, diferenciando entre los niños que presentan desnutrición crónica infantil (DCI) y aquellos que no la presentan. Se incluyeron filtros interactivos para examinar cómo las variables categóricas se distribuyen entre los grupos de niños con y sin DCI. Después de la preparación de los datos, se cuenta con un total de 19,273 registros de niños ecuatorianos menores de 5 años vivos. De estos, 4,766 presentan desnutrición crónica infantil, lo que representa un 24.8% del total.

En la figura 4 se puede identificar patrones en los datos categóricos, como nivel de educación de la madre, la raza o etnia de la madre, diferencias de la DCI entre zonas urbanas y rurales, desigualdades en el acceso a servicios básicos, características del material utilizado para la construcción de las viviendas

Figura 4 *Dashboard de variables categóricas*



Fuente: Figura realizada por Cleber Puente

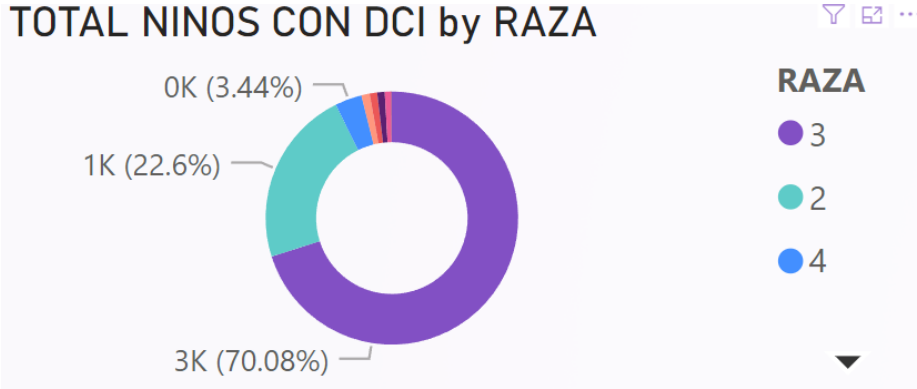
La Tabla 9 muestra la codificación de la variable raza en relación con las características, y la Figura 5 ilustra que la desnutrición crónica infantil se presenta más frecuentemente en las categorías de mestizo con un 70% e indígena con un 22%.

Tabla 9. *Codificación para la variable raza*

Raza	Codificación
Afroecuatoriano/a	0
Blanco/a	1
Indígena	2
Mestizo/a	3
Montuvio/a	4
Mulato/a	5
Negro/a	6
Otra	7

Fuente: Tabla realizada por Cleber Puente

Figura 5 *Total niños con DCI en la variable raza*



Fuente: Figura realizada en Power BI por Cleber Puente

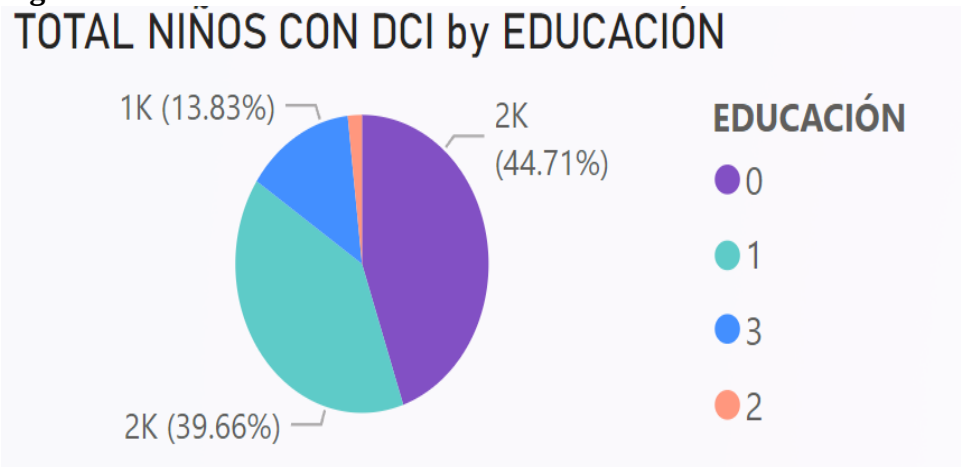
La tabla 10 muestra la codificación de la variable educación respecto a las características y en la figura 6 se observa que donde más se presenta la DCI es en las características de educación básica con un 45% y Bachillerato con un 40%.

Tabla 10. *Codificación del nivel de educación*

Nivel de educación	Codificación
Educación básica	0
Bachillerato	1
Ninguno	2
Superior	3

Fuente: Tabla realizada por Cleber Puente

Figura 6 *Total niños con DCI en la variable educación*



Fuente: Figura realizada en Power BI por Cleber Puente

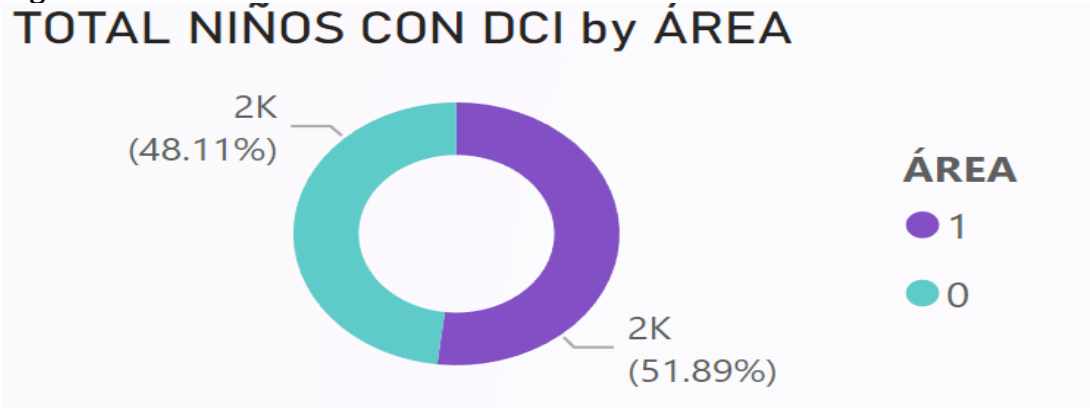
La tabla 11 muestra la codificación de la variable área respecto a las características y en la figura 7 se observa que donde más se presenta la DCI es en el área urbana con un 51%.

Tabla 11. *Codificación de la variable área*

Área	Codificación
rural	0
urbano	1

Fuente: Tabla realizada por Cleber Puente

Figura 7 *Total niños con DCI en la variable área*



Fuente: Figura realizada en Python por Cleber Puente

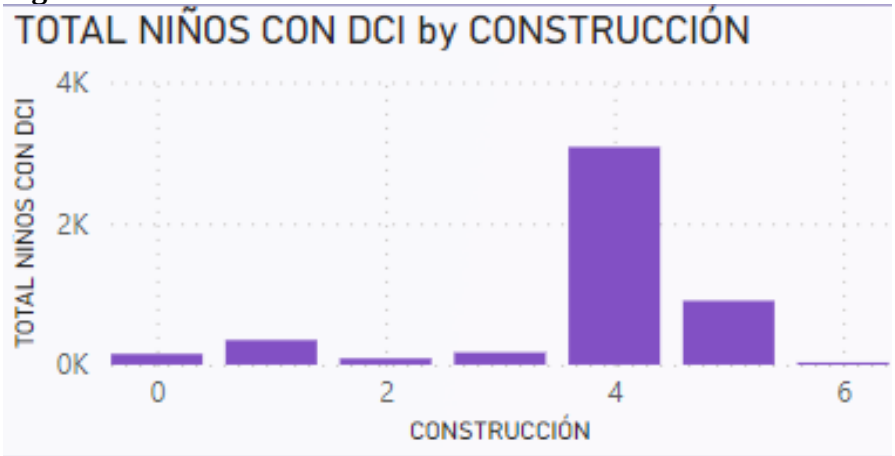
La tabla 12 muestra la codificación de la variable construcción respecto a las características y en la figura 8 se observa que donde más se presenta la DCI son el 65% de casas que fueron construidas con material de hormigón/bloque/ladrillo

Tabla 12. *Codificación de la variable construcción*

Área	Codificación
adobe/tapia	0
asbesto/cemento	1
bahareque	2
caña o estera	3
hormigón/bloque/ladrillo	4
madera	5
otra	6

Fuente: Tabla realizada por Cleber Puente

Figura 8 *Total niños con DCI en la variable construcción*



Fuente: Figura realizada en Python por Cleber Puente

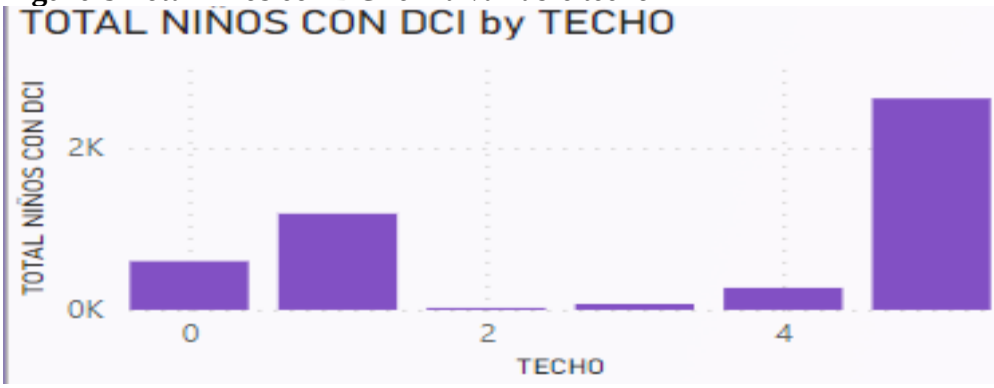
La tabla 13 muestra la codificación de la variable área respecto a las características y en la figura 9 se observa que donde más se presenta la DCI es en las casas donde el techo fue construido con material de zinc.

Tabla 13. *Codificación de la variable techo*

Área	Codificación
asbesto(eternit)	0
hormigón/losa/cemento	1
otro	2
palma/paja/hoja	3
teja	4
zinc	5

Fuente: Tabla realizada por Cleber Puente

Figura 9 *Total niños con DCI en la variable techo*



Fuente: Figura realizada en Python por Cleber Puente

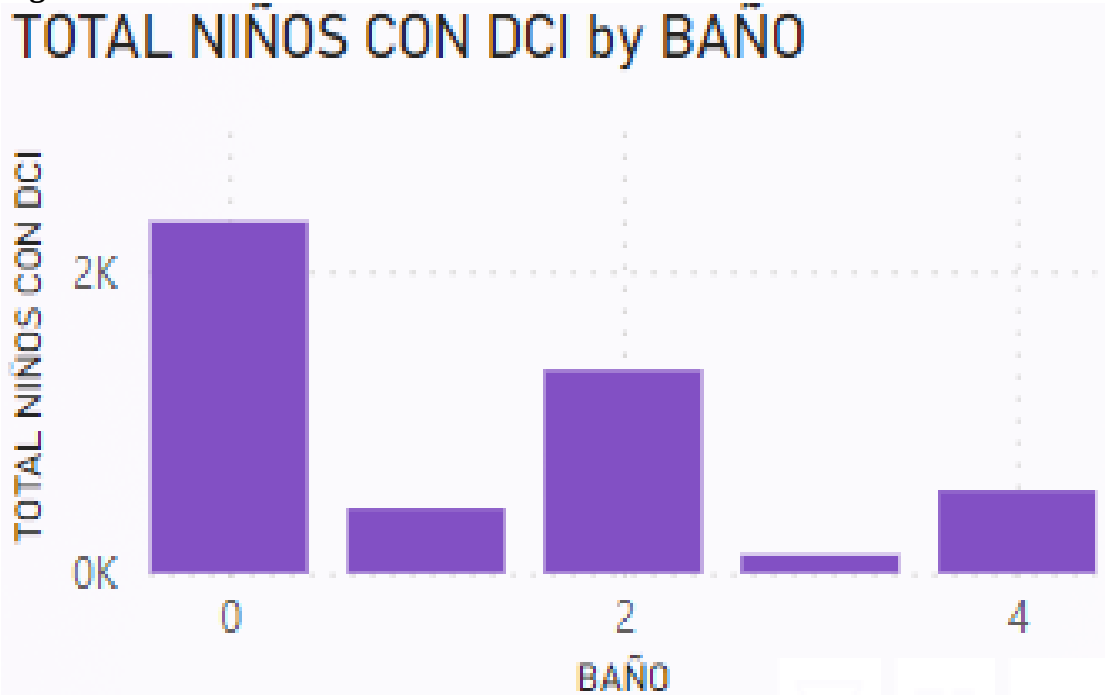
La tabla 14 muestra la codificación de la variable baño respecto a las características y en la figura 10 se observa que donde más se presenta la DCI es en las casas que tienen un baño con excusado y alcantarillado.

Tabla 14. *Codificación de la variable baño*

Área	Codificación
excusado y alcantarillado	0
excusado y pozo ciego	1
excusado y pozo séptico	2
letrina	3
no tiene	4

Fuente: Tabla realizada por Cleber Puente

Figura 10 *Total niños con DCI en la variable baño*



Fuente: Figura realizada en Python por Cleber Puente

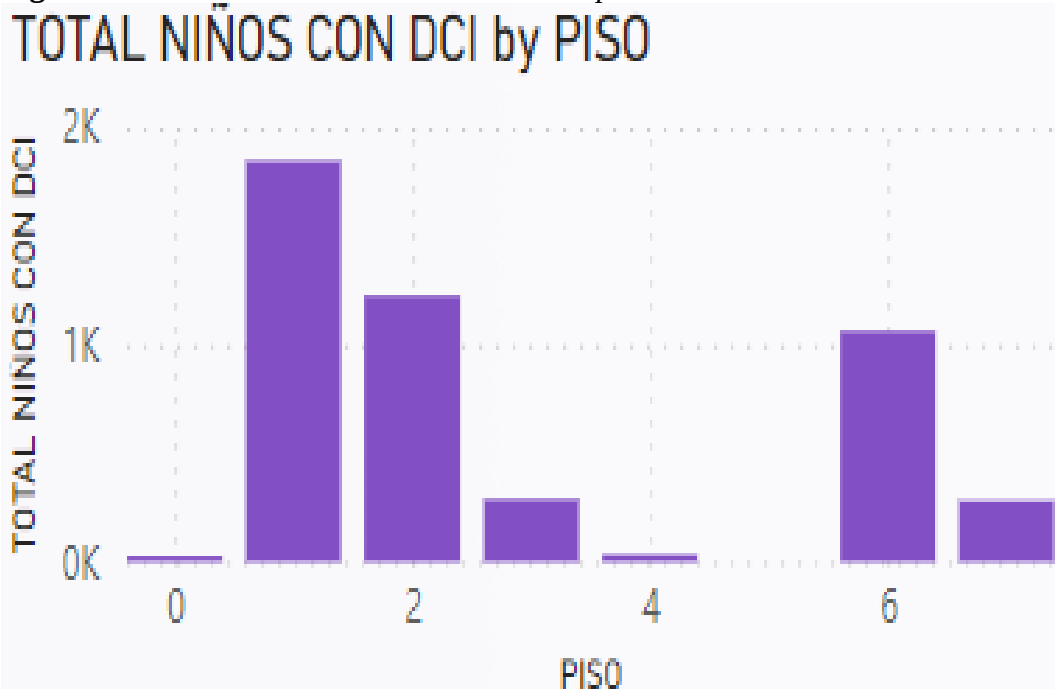
La tabla 15 muestra la codificación de la variable piso respecto a las características y en la figura 11 se observa que donde más se presenta la DCI es donde la construcción del piso de la casa es de cemento o ladrillo.

Tabla 15. *Codificación de la variable piso*

Área	Codificación
caña	0
cemento/ladrillo	1
cerámica/baldosa/vinyl	2
duela/ parquet/	3
mármol	4
otro	5
tabla/tablón no tratado	6
tierra	7

Fuente: Tabla realizada por Cleber Puente

Figura 11 *Total niños con DCI en la variable piso*



Fuente: Figura realizada en Python por Cleber Puente

La tabla 16 muestra la codificación de la variable agua respecto a las características y en la figura 12 se observa que donde más se presenta la DCI es que las personas se abastecen de agua de la red pública.

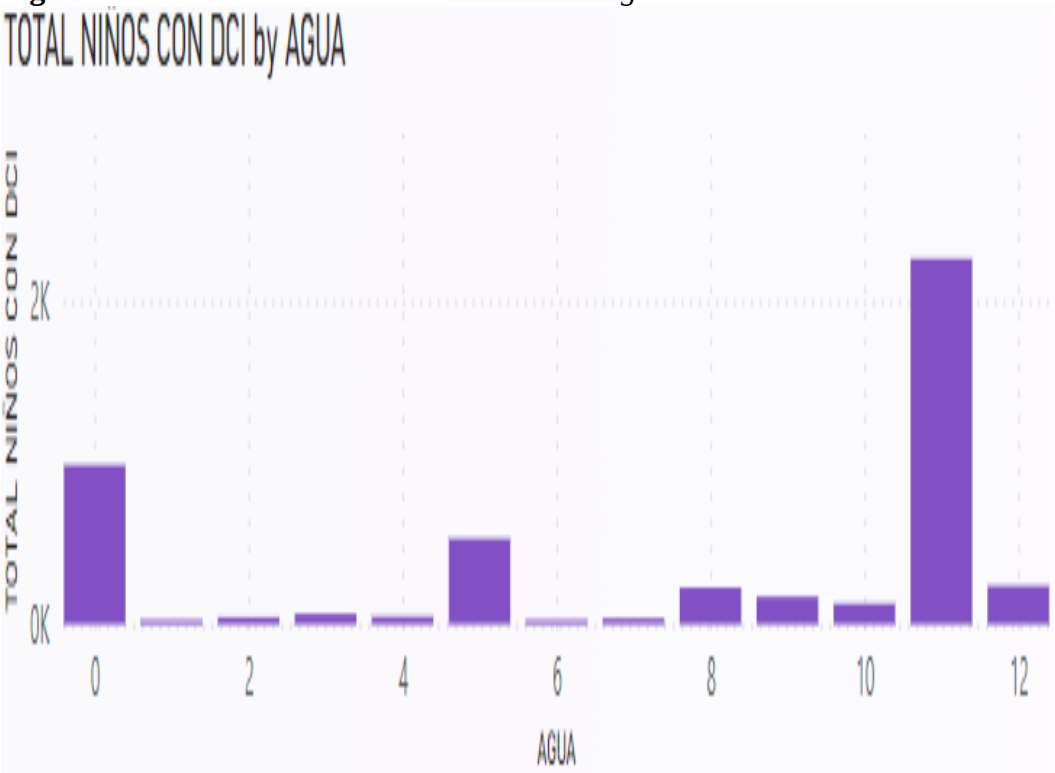
Tabla 16. *Codificación de la variable agua*

Área	Codificación
agua embotellada	0
agua en funda	1

carro repartidor/ tanquero	2
manantial/vertiente no protegida	3
manantial/vertiente protegida	4
otra fuente por tubería	5
Otro	6
pila o llave pública	7
Pozo entubado/pozo protegido	8
pozo no protegido	9
recogen agua de la lluvia	10
red pública	11
acequia	12

Fuente: Tabla realizada por Cleber Puente

Figura 12 *Total niños con DCI en la variable agua*



Fuente: Figura realizada en Python por Cleber Puente

4.4. Regresión logística

4.4.1. Descripción del modelo

El modelo de regresión logística se aplicó con el propósito de predecir e identificar las variables que influyen en la desnutrición crónica infantil.

4.4.2. Entrenamiento del modelo

Para estimar el modelo, se utilizaron los datos de entrenamiento y de prueba.

4.4.3. Evaluación training vs testing

Se obtienen los resultados de la evaluación del modelo de regresión logística para ambos conjuntos: entrenamiento y prueba.

4.4.3.1. *Conjunto de Entrenamiento:*

Accuracy: 75.52%

Matriz de Confusión:

Verdaderos negativos: 10,017

Falsos positivos: 137

Falsos negativos: 3,166

Verdaderos positivos: 171

Área bajo la curva ROC (AUC): 0.6555

4.4.3.2. *Conjunto de Prueba:*

Accuracy: 75.34%

Matriz de Confusión:

Verdaderos negativos: 4,290

Falsos positivos: 63

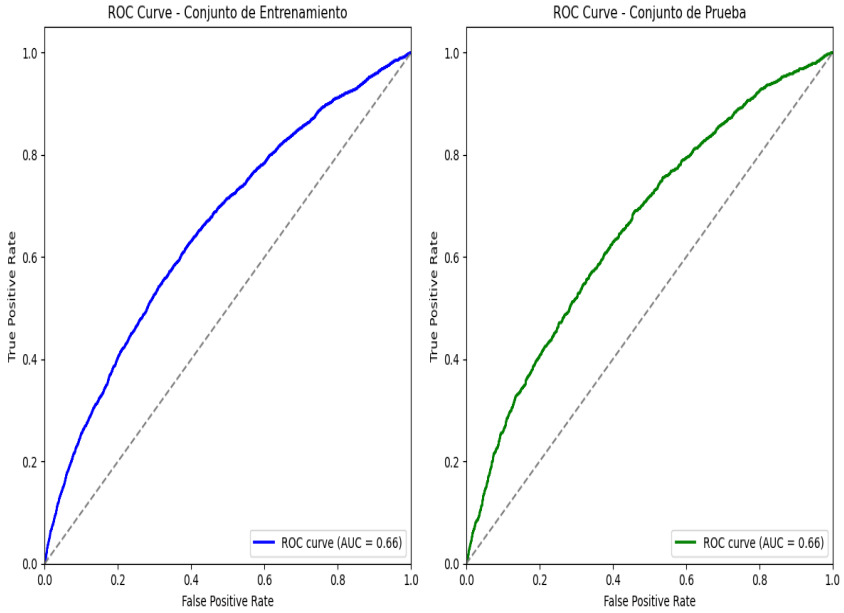
Falsos negativos: 1,363

Verdaderos positivos: 66

Área bajo la curva ROC (AUC): 0.6593

En la Figura 13 se presentan las curvas ROC para ambos conjuntos. La curva ROC del conjunto de entrenamiento muestra un área bajo la curva (AUC) de 0.6555, mientras que la del conjunto de prueba tiene un AUC de 0.6593. Estos valores permiten evaluar la capacidad del modelo para diferenciar entre las clases en ambos conjuntos.

Figura 13 *Conjunto de entrenamiento vs conjunto de prueba (regresión logística)*



Fuente: Figura realizada en Python por Cleber Puente

4.4.3.3. *Resultados obtenidos:*

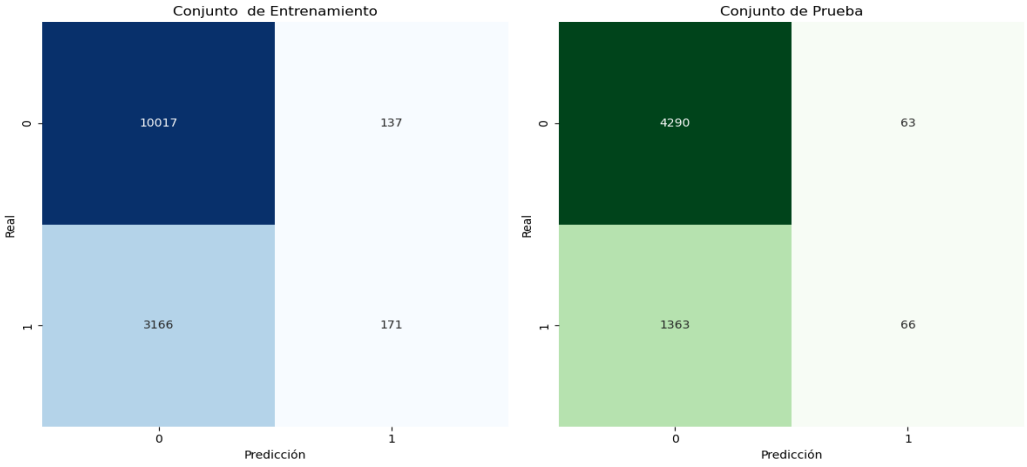
Conjunto de Entrenamiento:

Accuracy: 0.7552

Interpretación: El modelo tiene una precisión del 75.52% en el conjunto de entrenamiento, lo que significa que el 75.52% de las predicciones fueron correctas. Este nivel de precisión sugiere que el modelo está capturando patrones significativos en los datos, aunque aún hay un margen considerable de error.

Matriz de Confusión:

Figura 14 *Matriz de confusión regresión logística*



Fuente: Figura realizada en Python por Cleber Puente

Interpretación: En la figura 14 conjunto de entrenamiento se puede observar la alta cantidad de falsos negativos (3,166) en comparación con los verdaderos positivos (171) sugiere que el modelo tiene dificultades para identificar correctamente la clase de interés (probablemente una clase minoritaria).

AUC: 0.6555

Interpretación: El Área Bajo la Curva (AUC) del 65.55% indica una capacidad moderada del modelo para discriminar entre las clases. Un AUC cercano a 0.5 indicaría un modelo que no es mejor que adivinar al azar, mientras que un AUC de 1 indica una discriminación perfecta. En este caso, el AUC sugiere que el modelo podría beneficiarse de ajustes adicionales o de una mejor selección de características.

Conjunto de Prueba:

Accuracy: 0.7534

Interpretación: La precisión del 75.34% en el conjunto de prueba es muy similar a la del conjunto de entrenamiento, lo que sugiere que el modelo generaliza razonablemente bien y no está sobreajustado (overfitting) a los datos de entrenamiento.

Matriz de Confusión:

Interpretación: En la figura 14 conjunto de prueba se puede observar una situación similar al conjunto de entrenamiento, con una gran cantidad de falsos negativos en comparación con los verdaderos positivos. Esto refuerza la idea de que el modelo tiene dificultades para identificar correctamente la clase de interés, lo cual podría deberse a un desbalanceo en las clases o a la necesidad de un mejor ajuste de los hiperparámetros.

AUC: 0.6593

Interpretación: El AUC del 65.93% en el conjunto de prueba es similar al del conjunto de entrenamiento, lo que refuerza la idea de una capacidad moderada del modelo para discriminar entre clases.

4.5. Random forest

4.5.1. Descripción del modelo

El modelo de Random Forest se aplicó con el propósito de predecir e identificar las variables que influyen en la desnutrición crónica infantil.

4.5.2. Entrenamiento del modelo

Para estimar el modelo, se utilizaron los datos de entrenamiento y de prueba.

4.5.3. Evaluación training vs testing

Se obtiene los resultados de la evaluación del modelo de Random Forest tanto en el conjunto de entrenamiento como en el de prueba

4.5.3.1. *Conjunto de Entrenamiento:*

Accuracy: 96.52%

Matriz de Confusión:

Verdaderos negativos: 10,043

Falsos positivos: 111

Falsos negativos: 359

Verdaderos positivos: 2978

Área bajo la curva ROC (AUC): 0.9891

4.5.3.2. *Conjunto de Prueba:*

Accuracy: 71.79%

Matriz de Confusión:

Verdaderos negativos: 3815

Falsos positivos: 538

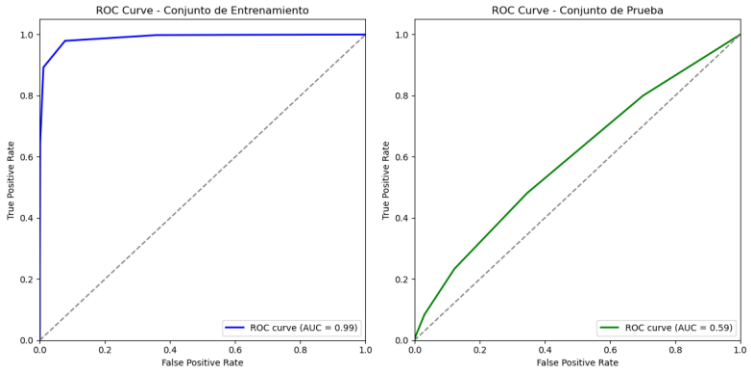
Falsos negativos: 1093

Verdaderos positivos: 336

Área bajo la curva ROC (AUC): 0.5925

En la figura 15 se observa las curvas ROC para ambos conjuntos. La curva ROC para el conjunto de entrenamiento muestra un área bajo la curva (AUC) de 0.9891 y la curva ROC para el conjunto de prueba muestra un AUC de 0.5925. Esto nos permite distinguir la capacidad del modelo para distinguir entre las variables en ambos conjuntos.

Figura 15 *Conjunto de entrenamiento vs conjunto de prueba (random forest)*



Fuente: Figura realizada en Python por Cleber Puente

4.5.3.3. *Resultados obtenidos:*

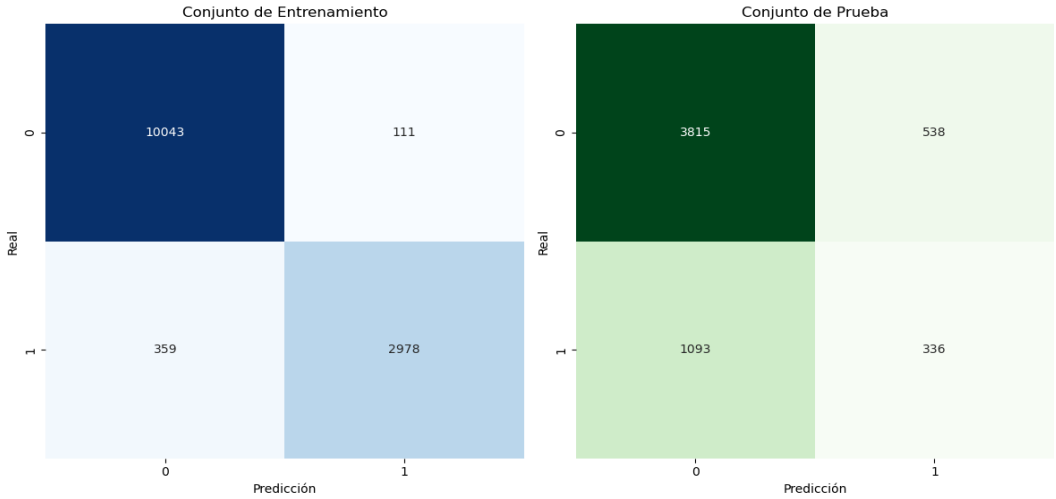
Conjunto de Entrenamiento

Accuracy: 0.9652 (96.52%)

Interpretación: Esto significa que el modelo logró clasificar correctamente el 96.52% de las observaciones en el conjunto de entrenamiento. Este es un valor bastante alto, lo que indica que el modelo está funcionando bien en los datos con los que fue entrenado.

Matriz de Confusión:

Figura 16 *Matriz de confusión Random Forest*



Fuente: Figura realizada en Python por Cleber Puente

Interpretación: En la figura 16 conjunto de entrenamiento se puede observar el bajo número de falsos positivos y falsos negativos en comparación con los verdaderos positivos y negativos sugiere que el modelo tiene un buen rendimiento en el conjunto de entrenamiento.

AUC: 0.9891

Interpretación: El AUC (Área Bajo la Curva ROC) de 0.9891 indica que el modelo tiene una excelente capacidad para discriminar entre las clases positiva y negativa. Un AUC cercano a 1 significa que el modelo es casi perfecto en la distinción entre las clases en el conjunto de entrenamiento.

Conjunto de Prueba

Accuracy: 0.7179 (71.79%)

Interpretación: La precisión en el conjunto de prueba es del 71.79%, significativamente más baja que en el conjunto de entrenamiento. Esto podría sugerir que el modelo no generaliza bien a datos nuevos, lo que es una indicación de posible overfitting (sobreajuste).

Matriz de Confusión:

Interpretación: En la figura 10 conjunto de prueba se observa un aumento considerable en el número de falsos positivos y falsos negativos en comparación con el conjunto de entrenamiento, lo que sugiere que el modelo tiene dificultades para clasificar correctamente las observaciones en el conjunto de prueba.

AUC: 0.5925

Interpretación: El AUC en el conjunto de prueba es de 0.5925, lo que está cerca de 0.5. Esto indica que el modelo tiene casi la misma capacidad para discriminar entre las clases que un modelo aleatorio, lo que refuerza la idea de que el modelo no está generalizando bien y podría estar sobreajustado a los datos de entrenamiento.

4.6. Support vector machine (svm)

4.6.1. Descripción del modelo

El modelo de SVM se aplicó con el propósito de predecir e identificar las variables que influyen en la desnutrición crónica infantil.

4.6.2. Entrenamiento del modelo

Para estimar el modelo, se utilizaron los datos de entrenamiento y de prueba.

4.6.3. Evaluación training vs testing

Se obtiene los resultados de la evaluación del modelo de regresión logística tanto en el conjunto de entrenamiento como en el de prueba

4.6.3.1. *Conjunto de Entrenamiento:*

Accuracy: 75.26%

Matriz de Confusión:

Verdaderos negativos: 10,154

Falsos positivos: 0

Falsos negativos: 3,337

Verdaderos positivos: 0

Área bajo la curva ROC (AUC): 0.5228

4.6.3.2. *Conjunto de Prueba:*

Accuracy: 75.29%

Matriz de Confusión:

Verdaderos negativos: 4353

Falsos positivos: 0

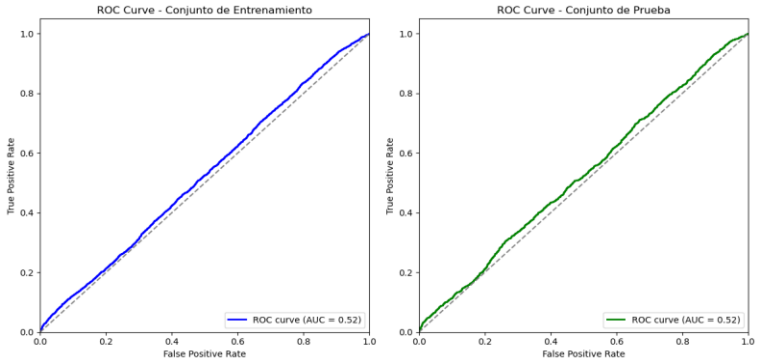
Falsos negativos: 1,429

Verdaderos positivos: 0

Área bajo la curva ROC (AUC): 0.5248

En la figura 17 se observa las curvas ROC para ambos conjuntos. La curva ROC para el conjunto de entrenamiento muestra un área bajo la curva (AUC) de 0.5228 y la curva ROC para el conjunto de prueba muestra un AUC de 0.5248 . Esto nos permite distinguir la capacidad del modelo para distinguir entre las clases en ambos conjuntos.

Figura 17 *Conjunto de entrenamiento vs conjunto de prueba (SVM)*



Fuente: Figura realizada en Python por Cleber Puente

4.6.3.3. *Resultados obtenidos:*

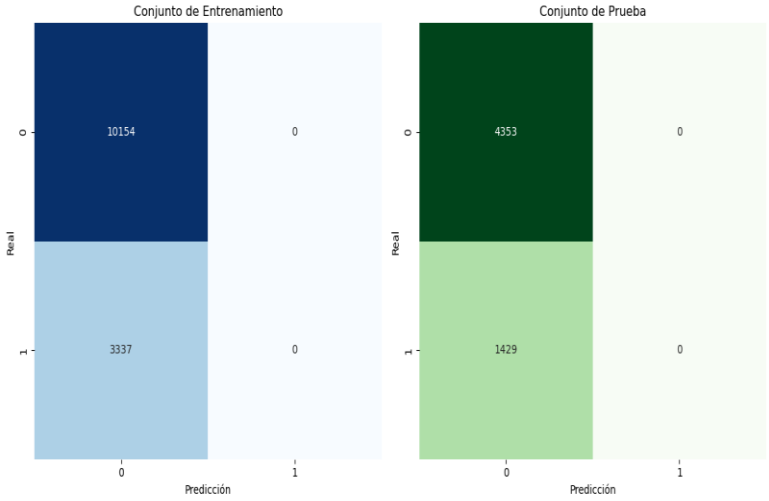
Conjunto de Entrenamiento

Accuracy: 0.7526 (75.26%)

Interpretación: El modelo logró una precisión del 75.26% en el conjunto de entrenamiento, lo que significa que clasificó correctamente el 75.26% de las observaciones. Sin embargo, la precisión por sí sola no nos dice mucho sobre el rendimiento del modelo, especialmente cuando las clases están desbalanceadas.

Matriz de Confusión:

Figura 18 *Matriz de confusión SVM*



Fuente: Figura realizada en Python por Cleber Puente

Interpretación: En la figura 18 conjunto de entrenamiento se puede observar que no hay verdaderos positivos ni falsos positivos (ambas son cero). Esta matriz sugiere que el modelo nunca predijo la clase positiva (es decir, todos los casos se predijeron como clase negativa). Esto puede indicar un modelo altamente sesgado o que las clases están muy desbalanceadas.

AUC: 0.5228

Interpretación: Un AUC de 0.5228 está apenas por encima de 0.5, lo que indica que el modelo tiene poca capacidad para discriminar entre las clases positiva y negativa. Es casi tan efectivo como un modelo que predice al azar.

Conjunto de Prueba

Accuracy: 0.7529 (75.29%)

Interpretación: La precisión en el conjunto de prueba es similar a la del conjunto de entrenamiento, lo que sugiere que el modelo es consistente, aunque no necesariamente bueno.

Matriz de Confusión:

Interpretación: En la figura 12 conjunto de prueba se puede observar nuevamente, no hay verdaderos positivos ni falsos positivos.

Al igual que en el conjunto de entrenamiento, el modelo nunca predice la clase positiva, lo que indica un problema significativo con el modelo.

AUC: 0.5248

Interpretación: El AUC de 0.5248 es nuevamente apenas superior a 0.5, lo que significa que el modelo tiene muy poca capacidad para diferenciar entre las clases. Esto sugiere que el modelo no está funcionando bien para predecir la clase positiva.

4.7. Comparación de los modelos en base a los resultados obtenidos

En la tabla 12 se realiza una comparación clara entre los modelos en términos de sus métricas clave y desempeño en los conjuntos de entrenamiento y prueba.

Tabla 17. *Tabla comparativa entre modelos*

Aspecto	**Regresión Logística**	**Random Forest**	**Support Vector Machine (SVM)**
Descripción del Modelo	Predicción e identificación de variables que influyen en la desnutrición crónica infantil	Predicción e identificación de variables que influyen en la desnutrición crónica infantil	Predicción e identificación de variables que influyen en la desnutrición crónica infantil
Accuracy (Entrenamiento)	75.52%	96.52%	75.26%
Accuracy (Prueba)	75.34%	71.79%	75.29%
Matriz de Confusión (Entrenamiento)	TN: 10,017, FP: 137, FN: 3,166, TP: 171	TN: 10,043, FP: 111, FN: 359, TP: 2,978	TN: 10,154, FP: 0, FN: 3,337, TP: 0
Matriz de Confusión (Prueba)	TN: 4,290, FP: 63, FN: 1,363, TP: 66	TN: 3,815, FP: 538, FN: 1,093, TP: 336	TN: 4,353, FP: 0, FN: 1,429, TP: 0
AUC (Entrenamiento)	0.6555	0.9891	0.5228
AUC (Prueba)	0.6593	0.5925	0.5248
Interpretación Accuracy (Entrenamiento)	Precisión moderada, pero con margen de mejora	Excelente precisión en datos de entrenamiento	Precisión moderada, pero sugiere posibles problemas de sesgo
Interpretación Accuracy (Prueba)	Precisión similar entre entrenamiento y prueba (buena generalización)	Precisión mucho menor que en entrenamiento (indicación de sobreajuste)	Precisión similar al conjunto de entrenamiento, pero problemas de sesgo
Interpretación AUC (Entrenamiento)	Capacidad moderada de discriminar entre clases	Excelente discriminación entre clases	Capacidad casi nula para discriminar entre clases

Interpretación AUC (Prueba)	Capacidad moderada de discriminar entre clases	Capacidad muy baja de discriminar entre clases (posible sobreajuste)	Capacidad muy baja para discriminar entre clases
Problemas Identificados	Dificultades para identificar la clase de interés (muchos falsos negativos)	Sobreajuste en entrenamiento, con rendimiento pobre en prueba	No predice la clase positiva, lo que indica un modelo sesgado o desbalanceado

Fuente: Tabla realizada por Cleber Puente

4.8. Seleccionar el modelo ganador

Con base en estos resultados, el modelo seleccionado es el de Regresión Logística. Aunque tiene un AUC moderado, su consistencia entre entrenamiento y prueba lo hace un mejor candidato, lo que indica que tiene un buen balance entre generalización y precisión. Además, es menos propenso al sobreajuste, como se observó en el modelo de Random Forest, y no presenta problemas serios de sesgo o desbalance como en el caso de SVM.

4.9. Desarrollo de una propuesta de implementación

La provincia de Chimborazo ha sido identificada históricamente como una de las más afectadas por la desnutrición crónica infantil (DCI) en Ecuador. Gracias por los datos actualizados de la 1ra. Encuesta Especializada sobre Desnutrición Infantil (ENDI) diseñada para conocer el estado nutricional de los niños en el Ecuador (INEC, 2023). Con esta nueva información, la provincia de Chimborazo tiene un índice de desnutrición crónica infantil (DCI) del 35.1%, lo que sigue confirmando su posición como la provincia más afectada en Ecuador.

Estos datos fortalecen la urgencia de implementar un modelo predictivo en Chimborazo para abordar la DCI y también sirven de referencia para comparaciones regionales y para identificar posibles factores comunes entre las provincias con altas y bajas tasas de DCI.

4.9.6. Implementación en el Hospital Andino de Chimborazo

En el marco de esta propuesta, se llevó a cabo una aplicación del modelo en colaboración con el Hospital Andino de Chimborazo, el cual proporcionó un entorno adecuado para realizar pruebas y ajustes iniciales según los requerimientos específicos de la institución. Los detalles clave de esta implementación son:

4.9.1.1. *Infraestructura necesaria*

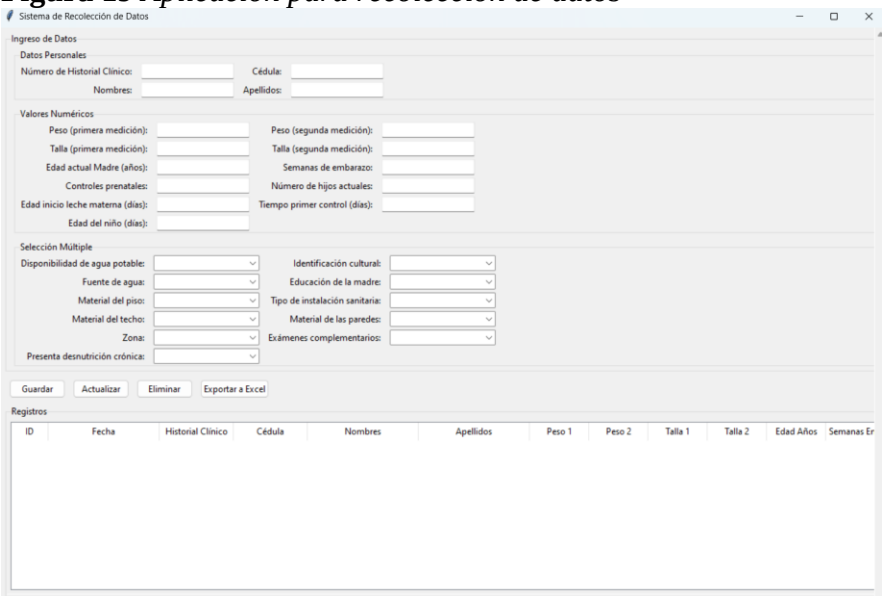
Dado que la aplicación se utilizará únicamente para la recolección de datos y no requiere conexión a internet ni almacenamiento en la nube, los requisitos de infraestructura son modestos:

Equipo local: Se utilizará un computador de escritorio o laptop dentro del Hospital Andino de Chimborazo, donde se instalará la aplicación. Este equipo debe contar con las siguientes características mínimas:

Un procesador Intel Core i3 o superior, memoria RAM de4 GB o más, almacenamiento de 500 GB de disco duro y sistema operativo Windows 10 o superior.

Software:

Figura 19 *Aplicación para recolección de datos*



Fuente: Aplicación realizada en Python por Cleber Puente

En la figura 19 se muestra la aplicación realizada en Visual Studio Code , se utilizo esta herramienta para el diseño y desarrollo de la aplicación, permitiendo una integración eficiente de las bibliotecas necesarias y la creación de una interfaz gráfica intuitiva para el usuario.

4.9.1.2. *Funcionalidad de la aplicación*

La aplicación estará diseñada exclusivamente para la recolección de datos relacionados con las variables analizadas en este estudio. Sus principales características son:

Ingreso de datos manual: Permitir al personal autorizado del hospital registrar información sobre las variables seleccionadas, como acceso a agua potable, registros de peso y talla, lactancia materna, condiciones de vivienda, entre otras.

Validación de datos: Incluir validaciones en tiempo real para evitar errores en la entrada de datos, como valores fuera de rango o datos incompletos.

Exportación de datos: La aplicación generará archivos en formatos compatibles, que serán enviados al equipo de análisis por correo electrónico para su procesamiento.

4.9.1.3. *Implementación local*

La aplicación será utilizada exclusivamente dentro del Hospital Andino de Chimborazo, lo que asegura un control completo sobre los datos recolectados. En el futuro, este modelo podría extenderse a otros puntos de atención local como:

Centros de Desarrollo Infantil (CDI).

Dispensarios médicos estatales.

Hospitales infantiles públicos de la provincia.

Sin embargo, cualquier expansión deberá garantizar el cumplimiento estricto de las normativas legales relacionadas con la protección de datos.

4.9.1.4. *Certificación y colaboración*

El Hospital Andino de Chimborazo se ha comprometido a recolectar los datos necesarios y enviarlos para el análisis por correo electrónico. Como respaldo de esta colaboración, se adjunta en el Anexo 2 un certificado otorgado por el hospital, que valida su compromiso con este proyecto.

Con esta solución, se busca garantizar la viabilidad de la implementación en un entorno local, asegurando el cumplimiento legal y facilitando la recolección y análisis de datos clave para la lucha contra la desnutrición crónica infantil.

4.9.2. Intervenciones específicas basadas en resultados

Los resultados del modelo deben servir como base para diseñar intervenciones específicas en salud y nutrición. Por ejemplo, en las zonas donde se prediga una alta prevalencia de desnutrición crónica, se pueden implementar programas de suplementación nutricional o mejorar la infraestructura.

4.9.2.1. *Uso de las Características Seleccionadas en el Modelo y la Intervención*

- ***Disponibilidad de agua potable (_f1_s1_28)***

  En el modelo: El acceso al agua es una de las principales variables predictivas de DCI, ya que la calidad del agua influye directamente en la salud de los niños. Las familias sin acceso a agua potable tienen un mayor riesgo de sufrir enfermedades gastrointestinales, que afectan la nutrición.

  En la intervención: Se debe priorizar la mejora de infraestructura de agua y saneamiento. Esto puede incluir la instalación de sistemas de filtrado de agua o la mejora del acceso a fuentes seguras en comunidades rurales de Chimborazo.

- ***Registros de peso y talla (_f1_s7_4_1, _f1_s7_4_2, _f1_s7_6_1, _f1_s7_6_2)***

  En el modelo: Los registros de peso y talla son indicadores directos del estado nutricional de los niños. Un crecimiento por debajo del estándar puede ser un indicador claro de DCI. Estos datos serán fundamentales para calibrar el modelo, ya que permiten una evaluación más precisa de la salud infantil.

  En la intervención: Los niños identificados con desnutrición en estas mediciones deben recibir atención inmediata con suplementación nutricional, monitoreo frecuente del crecimiento e intervenciones en programas alimentarios para garantizar una ingesta adecuada.

- **Edad (_f2_s1_101)**

  En el modelo: La edad influye en los requerimientos nutricionales. El modelo debe tomar en cuenta que los niños más pequeños (particularmente menores de 2 años) son más vulnerables a la DCI, ya que es un periodo crítico de crecimiento.

  En la intervención: Se deben dirigir los recursos y esfuerzos hacia los niños en estas etapas más vulnerables, promoviendo la lactancia materna exclusiva hasta los 6 meses y la correcta introducción de alimentos complementarios.

- **Controles prenatales (_f2_s4b_420_, _f2_s4b_421_)**

  En el modelo: La cantidad y momento de los controles prenatales son factores clave que afectan la salud del recién nacido y su desarrollo inicial. Un número bajo de controles prenatales está correlacionado con un mayor riesgo de desnutrición infantil.

  En la intervención: Se debe garantizar el acceso a controles prenatales de calidad. Las madres que no han asistido a los controles prenatales necesarios deben ser priorizadas para recibir asistencia médica y seguimiento postnatal para sus hijos.

- **Número de hijos que viven con el encuestado (_f2_s2_208_3)**

En el modelo: Un mayor número de hijos puede reducir los recursos disponibles por niño, aumentando el riesgo de desnutrición. El modelo debe considerar este factor para identificar familias en riesgo.

En la intervención: Las políticas de intervención deben estar dirigidas a familias numerosas, proporcionando asistencia alimentaria y programas educativos que fomenten el control de la natalidad y la planificación familiar.

- **Lactancia materna (lactancia_total_dias)**

  En el modelo: La duración de la lactancia es un predictor clave de desnutrición. Los niños que no reciben lactancia materna exclusiva tienen un mayor riesgo de sufrir DCI.

  En la intervención: Promover programas que fomenten la lactancia materna exclusiva durante los primeros 6 meses de vida y continuar con la lactancia complementaria hasta los 2 años o más. Se deben ofrecer asesorías a madres sobre los beneficios de la lactancia y cómo superar posibles dificultades.

- **Primer control médico (control_total_dias)**

  En el modelo: El retraso en el primer control médico postnatal puede ser un indicador de negligencia o falta de acceso a servicios de salud, aumentando el riesgo de DCI.

  En la intervención: Mejorar la infraestructura de salud y garantizar que los recién nacidos reciban su primer control médico dentro de los primeros días de vida. Se deben lanzar campañas de concientización sobre la importancia de los primeros controles postnatales.

- **Identificación cultural (_f1_s2_9)**

  En el modelo: Las prácticas culturales pueden influir en las decisiones de alimentación y cuidado infantil. Por ejemplo, ciertas prácticas tradicionales podrían estar asociadas con mayores tasas de desnutrición.

En la intervención: Se deben adaptar las intervenciones a las particularidades culturales de Chimborazo, colaborando con líderes comunitarios y respetando las costumbres locales. Se pueden lanzar programas de sensibilización culturalmente adecuados que promuevan la nutrición infantil sin confrontar creencias locales.

- **Condiciones de vivienda (_f1_s1_25, _f1_s1_4, _f1_s1_13, _f1_s1_3, _f1_s1_5)**

  En el modelo: Las condiciones materiales de la vivienda, como la fuente de agua potable, el tipo de material de las paredes, techo y piso, reflejan el nivel socioeconómico de las familias, que está directamente relacionado con el riesgo de DCI.

  En la intervención: Mejorar las condiciones de vivienda es esencial para reducir la DCI. Las políticas públicas deben incluir subsidios para la mejora de infraestructuras básicas en las viviendas más vulnerables.

- **Nivel de educación de la madre (_nivins_mef)**

  En el modelo: El nivel de educación de la madre es un fuerte predictor de la salud infantil. Las madres con mayor nivel educativo suelen tener mejores conocimientos sobre nutrición y cuidado infantil.

  En la intervención: Implementar programas de educación para madres con bajo nivel educativo, centrándose en temas de nutrición, lactancia materna y cuidado infantil. Estos programas deben ser accesibles en las áreas rurales y en los idiomas locales.

- **Zona geográfica (area)**

  En el modelo: El área (urbana o rural) influye en el acceso a servicios de salud y recursos. En áreas rurales, los niños tienen un mayor riesgo de desnutrición debido al acceso limitado a infraestructura y servicios.

  En la intervención: Las zonas rurales deben ser priorizadas en la implementación del modelo. Se deben mejorar los servicios de salud, transporte y acceso a recursos esenciales

como agua potable y alimentos. Las intervenciones específicas, como brigadas de salud móviles, pueden ser cruciales para alcanzar a las poblaciones más aisladas.

4.9.3. Mejoras propuestas para el modelo

Para mejorar el rendimiento del modelo de Regresión Logística, implementado para predecir la desnutrición crónica infantil en la provincia de Chimborazo, considerando que los datos serán obtenidos del Hospital Andino, se podrían realizar los siguientes ajustes:

Selección de Características: Implementar una técnica más robusta de selección de variables para garantizar que las características más relevantes que influyen en la desnutrición crónica infantil sean priorizadas.

Manejo del Desbalance de Clases: Debido a que la clase positiva (casos de desnutrición crónica infantil) es minoritaria, se debe aplicar técnicas de balanceo, como el uso de un algoritmo de muestreo (oversampling o undersampling) para mejorar la predicción de la clase positiva.

4.9.4. Monitoreo y ajustes continuos del modelo

Una vez implementado el modelo, y con la recolección constante de datos desde el Hospital Andino, se deben realizar actividades de monitoreo y ajuste continuo para garantizar su eficacia:

Evaluación mediante métricas clave: Monitorear el desempeño del modelo a través de métricas como la precisión (*accuracy*), el área bajo la curva ROC (*AUC*), y la matriz de confusión. Estas métricas proporcionarán información valiosa sobre su capacidad para clasificar correctamente los casos.

Actualización periódica del modelo: A medida que se recolecten más datos en el Hospital Andino, será fundamental recalibrar el modelo, ajustando hiperparámetros y actualizando las características incluidas para reflejar cualquier cambio en los patrones de los datos.

5. Conclusiones y recomendaciones

5.1. Conclusiones

- Al analizar los factores asociados a la desnutrición crónica infantil, como se observa en las Figuras 4, se identificaron los siguientes factores significativos: para la variable raza, las características predominantes son mestizo e indígena; para el nivel de educación, la categoría destacada es educación básica; y para el área, el factor significativo es el entorno urbano. La construcción de las casas en su mayoría son las paredes de hormigón, boque/ ladrillo , el techo de zinc, el baño con excusado y alcantarillado, el piso es de cemento o bloque y los hogares se abastecen de agua de la red pública.
- El modelo de regresión logística muestra una precisión razonable tanto en el conjunto de entrenamiento como en el conjunto de prueba, lo que indica que no hay sobreajuste significativo. Sin embargo, la capacidad discriminativa del modelo (AUC en torno al 65%) sugiere que podría no estar capturando adecuadamente las diferencias entre las clases, especialmente en lo que respecta a identificar correctamente la clase de interés (baja tasa de verdaderos positivos y alta tasa de falsos negativos).
- El modelo Random Forest parece estar sobreajustado al conjunto de entrenamiento, lo que se manifiesta en una alta precisión y un excelente AUC en los datos de entrenamiento, pero un rendimiento mucho peor en los datos de prueba. Este tipo de comportamiento sugiere que el modelo ha capturado patrones específicos del conjunto de entrenamiento que no se generalizan bien a nuevos datos.
- El modelo muestra un rendimiento deficiente, especialmente en la predicción de la clase positiva. La precisión del 75.26% y 75.29% en los conjuntos de entrenamiento y prueba, respectivamente, parece alta, pero es engañosa debido al hecho de que el modelo no está prediciendo correctamente ninguna observación de la clase positiva. El bajo valor de AUC en ambos conjuntos sugiere que el modelo no es útil para la discriminación entre clases.
- El uso de las características seleccionadas en el modelo de regresión logística permitirá identificar con precisión a las familias y comunidades con mayor riesgo de DCI en Chimborazo.

5.2. Recomendaciones

- Para el modelo de regresión logística se tiene la necesidad de ajustar el modelo, explorar técnicas de balanceo de clases, o considerar diferentes enfoques de modelado.
- Para el modelo Random Forest es posible que se necesiten ajustes adicionales, como regularización, reducción de complejidad del modelo, o más datos de entrenamiento, para mejorar el rendimiento en el conjunto de prueba y lograr un modelo más robusto.
- Para el modelo SVM el conjunto de datos esta altamente desbalanceado, lo que lleva al modelo a sesgarse fuertemente hacia la clase mayoritaria. Para mejorar, se podría considerar el uso de técnicas de balanceo de clases, como el sobre muestreo de la clase minoritaria, el submuestreo de la clase mayoritaria, o el uso de modelos más avanzados o ajustados específicamente para manejar clases desbalanceadas
- La implementación debe enfocarse en mejorar el acceso a recursos esenciales (agua, salud, vivienda), promover la educación y cuidado materno-infantil, y diseñar intervenciones culturalmente adecuadas para mitigar la desnutrición crónica infantil

Bibliografía

Alteryx. (2024). *Organizacion de datos.* https://www.alteryx.com/es/glossary/data-wrangling.

Congacha, G. (2020). *Comparación de Modelos Logísticos y Árboles de Decisión para Identificar y Predecir Factores Asociados a la Desnutrición Crónica Infantil Basados en la Encuesta Nacional de Salud y Nutrición – ENSANUT 2018-2019.* Riobamba.

INEC, I. N. (2018). *Documento metodológico de la encuesta nacional de salud y nutrición (ENSANUT).* Quito.

Muñoz, D., & Romero, J. (2021). *Optimización de los hiperparametros de una máquina de regresion de soporte vectorial utilizando enjambre de partículas para el pronóstico de casos de CIVID-19.* Bogotá.

Muñoz, D., & Romero, J. (2021). *Optimización de los hiperparámetros de una máquina de regresión de soporte vectorial utilizando enjambre de partículas para el pronóstico de casos de COVID-19.* Bogotá.

ONU, O. d. (2022). *El desafío de la nutrición: Soluciones desde los sistemas alimentarios.*

Organización Mundial de la Salud, O. (2022). *Malnutrición.*

Staff, C. (2023). *¿Que es data wrangling y por qué es importante?* https://www.coursera.org/mx/articles/data-wrangling.

UNICEF. (2022). *Informe anual de UNICEF 2022.*

Vallalta, J. (2024). *CRISP DM: Una metodologá para minería de datos en salud* . México.

Wiley, J., & Sons. (2017). *Data Science for Dummies.* Pierson.

Wisbaum, W. (2011). *LA DESNUTRICIÓN INFANTIL. Causas, consecuencias y estrategias para su prevención y tratamiento.* Madrid: punto&coma.

Yanez, C. (2023). *ANÁLISIS EXPLORATORIO DE DATOS E IDENTIFICACIÓN DE AGENTES QUE INFLUYEN EN LA DESNUTRICIÓN CRÓNICA DE NIÑOS MENORES A CINCO AÑOS DEL ECUADOR MEDIANTE LA APLICACIÓN DE TÉCNICAS DE CIENCIA DE DATOS.* Quito: PUCE.

6. Anexo

Anexo 1 Código y ejecución de featurewiz para selección de variables

Código

```
pip install featurewiz

!pip install "dask[dataframe]" --upgrade

from featurewiz import featurewiz

X=df.drop('dcronica',axis=1)

y=df.dcronica

print(X)

print(y)

target = 'dcronica'

feats = featurewiz(df, target, corr_limit=0.85, verbose=1)

len(feats)
```

Ejecución:

```
#######################################################################################
###########      F A S T  F E A T U R E  E N G G  A N D  S E L E C T I O N !  ########
# Be judicious with featurewiz. Don't use it to create too many un-interpretable features! #
#######################################################################################
featurewiz has selected 0.85 as the correlation limit. Change this limit to fit your needs...
```

Skipping feature engineering since no feature_engg input...

Skipping category encoding since no category encoders specified in input...

Single_Label Binary_Classification problem

Loaded train data. Shape = (19273, 55)

Single_Label Binary_Classification problem

No test data filename given...

Classifying features using a random sample of 10000 rows from dataset...

Single_Label Binary_Classification problem

loading a random sample of 10000 rows into pandas for EDA

##
###############

######################## C L A S S I F Y I N G V A R I A B L E S
####################

1 variable(s) to be removed since ID or low-information variables

variables removed = ['f2_s4a_402_']

train data shape before dropping 1 columns = (19273, 55)

train data shape after dropping columns = (19273, 54)

No GPU active on this device

Tuning XGBoost using CPU hyper-parameters. This will take time...

Removing 1 columns from further processing since ID or low information variables

After removing redundant variables from further processing, features left = 53

...

Selected 20 important features:

['f1_s7_6_1', 'edad_dias', 'f1_s7_4_1', 'f2_s1_101', 'f1_s7_6_2', 'control_total_dias', 'f2_s4b_420_', 'f2_s4b_421_', 'f1_s7_4_2', 'f2_s2_208_3', 'lactancia_total_dias', 'f1_s2_9', 'f1_s1_25', 'nivins_mef', 'f1_s1_4', 'f1_s1_13', 'f1_s1_3', 'f1_s1_5', 'area', 'f1_s1_28']

Anexo 2 Manual de usuario aplicación

Registrar un Nuevo Paciente

En la figura 19 se muestra la pantalla principal de la aplicación en la cual se debe realizar las siguientes actividades:

- Complete los campos de Datos Personales con la información del paciente.
- Llene los valores numéricos correspondientes.
- Seleccione las opciones relevantes en los menús desplegables de Selección Múltiple.
- Presione el botón Guardar para almacenar la información.

Modificar un Registro Existente

- Seleccione un registro en la tabla.
- Edite los valores en los campos correspondientes.
- Presione el botón Actualizar.

Eliminar un Registro

- Seleccione el registro que desea eliminar.
- Presione el botón Eliminar.

Exportar Registros a Excel

- Haga clic en el botón Exportar a Excel.
- Guarde el archivo generado en la ubicación deseada.

Anexo 3 Certificado otorgado por el Hospital Andino

RIOBAMBA // Pastaza S/N y Manabí - Ciudadela 24 de Mayo
info@hospitalandino.com
2600 153 / 0989 711 807

HOSPITAL ANDINO
HOSPITAL GENERAL

Hospital General Andino de Riobamba
Riobamba, Ecuador

CERTIFICACIÓN

En calidad de Gerente General del **HOSPITAL GENERAL ANDINO DE RIOBAMBA**, la **ING. VERONICA ELIZABETH MANCERO REVELO** con cedula de identidad **060251925-8** autoriza el uso del programa de **recolección de datos** en el área de **Estadística**, en el marco del proyecto titulado **"Aplicación de técnicas de machine learning para predecir la desnutrición infantil en el Ecuador"**, realizado por **CLEBER DAMIAN PUENTE TISCAMA**, con cédula de identidad **1721997920.**

El Hospital Andino de Riobamba, comprometido con la lucha contra la desnutrición infantil, apoya cualquier proyecto orientado a prevenir, analizar y mitigar esta problemática que afecta a nuestra comunidad. Este programa de recolección de datos es una contribución más en este esfuerzo.

El objetivo de este programa es recopilar información relevante para la prevención y análisis de la desnutrición infantil, tomando en cuenta factores sociales, económicos y de salud.

Estos datos serán utilizados para el desarrollo y entrenamiento de un modelo de Machine Learning, cuya finalidad será predecir y analizar el riesgo de desnutrición crónica infantil en la población atendida.

El proceso de recolección de datos se realizó conforme a los principios éticos y bajo estrictas normativas de confidencialidad y protección de datos, en apego a las regulaciones vigentes.

Dado en Riobamba, a los 18 días del mes de noviembre de 2024.

HOSPITAL GENERAL ANDINO
RUC: 0691780465001

Ing. VERONICA ELIZABETH MANCERO REVELO
GERENTE GENERAL
HOSPITAL GENERAL ANDINO DE RIOBAMBA