Maestría en

# CIENCIA DE DATOS Y MÁQUINAS DE APRENDIZAJE CON MENCIÓN EN INTELIGENCIA ARTIFICIAL

**Trabajo previo a la obtención de título de Magister en Ciencia de Datos y Máquinas de Aprendizaje con mención en Inteligencia Artificial**


**AUTORES:**

Adriana Stefany Mendoza López

Gonzalo Paul Rodríguez Galarza

Iván Patricio Torres Ramírez

Víctor German Huilca Revelo

**TUTORES:**

Karla Estefanía Mora Cajas

Bryan Steven Vinueza Bustamante


**Tema:**

Modelo predictivo para la identificación temprana de la demanda territorial de atención social en la provincia de Manabí a partir de registros administrativos de desarrollo humano


Quito - Ecuador

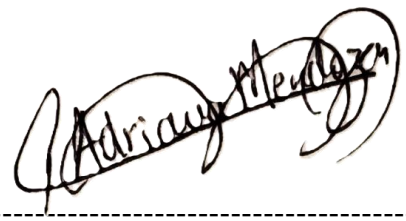
Julio– 2026


REINVENTEMOS EL FUTURO

**Certificación de autoría**

Nosotros, **Adriana Stefany Mendoza López, Gonzalo Paul Rodríguez Galarza, Iván Patricio Torres Ramírez, Víctor German Huilca Revelo**, declaramos bajo juramento que el trabajo aquí descrito es de nuestra autoría; que no ha sido presentado anteriormente para ningún grado o calificación profesional y que se ha consultado la bibliografía detallada.

Cedemos nuestros derechos de propiedad intelectual a la Universidad Internacional del Ecuador (UIDE), para que sea publicado y divulgado en internet, según lo establecido en la Ley de Propiedad Intelectual, su reglamento y demás disposiciones legales.

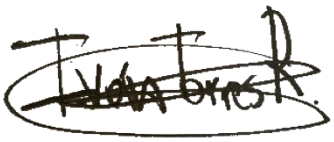


---

**Firma**
**Adriana Stefany Mendoza López**

---

**Firma**
**Gonzalo Paul Rodríguez Galarza**



---

**Firma**
**Iván Patricio Torres Ramírez**

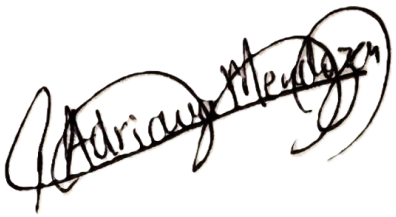
---

**Firma**
**Víctor German Huilca Revelo**

**Autorización de Derechos de Propiedad Intelectual**

Nosotros, **Adriana Stefany Mendoza López, Gonzalo Paul Rodríguez Galarza, Iván Patricio Torres Ramírez, Víctor German Huilca Revelo**, en calidad de autores del trabajo de investigación titulado **Modelo predictivo para la identificación temprana de la demanda territorial de atención social en la provincia de Manabí a partir de registros administrativos de desarrollo humano**, autorizamos a la Universidad Internacional del Ecuador (UIDE) para hacer uso de todos los contenidos que nos pertenecen o de parte de los que contiene esta obra, con fines estrictamente académicos o de investigación. Los derechos que como autores nos corresponden, lo establecido en los artículos 5, 6, 8, 19 y demás pertinentes de la Ley de Propiedad Intelectual y su Reglamento en Ecuador.

D. M. Quito, julio 2026.

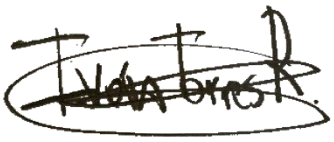


---

**Firma**
**Adriana Stefany Mendoza López**

---

**Firma**
**Gonzalo Paul Rodríguez Galarza**



---

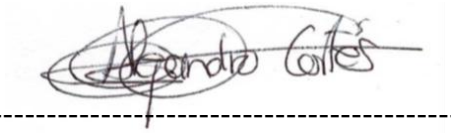
**Firma**
**Iván Patricio Torres Ramírez**

---

**Firma**
**Víctor German Huilca Revelo**

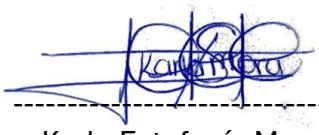
**Aprobación de dirección y coordinación del programa**

Nosotros, **Alejandro Cortés López Director EIG y Karla Estefanía Mora Cajas Coordinadora UIDE,** declaramos que: **Adriana Stefany Mendoza López, Gonzalo Paul Rodríguez Galarza, Iván Patricio Torres Ramírez, Víctor German Huilca Revelo** son los autores exclusivos de la presente investigación y que ésta es original, auténtica y personal de ellos.



---

Alejandro Cortés López
Director de la
Maestría en Ciencia de datos y máquinas de aprendizaje con mención en Inteligencia Artificial



---

Karla Estefanía Mora Cajas
Coordinadora de la
Maestría en Ciencia de datos y máquinas de aprendizaje con mención en Inteligencia Artificial

## DEDICATORIA

Dedico este trabajo a mis padres Klever Mendoza y Vanessa López, por su amor, esfuerzo y apoyo incondicional durante toda mi formación académica, por impulsarme a seguir creciendo y por confiar siempre en mí y en cada uno de mis logros.

A mis abuelos, a mi hermana, a mi novio, y en especial, a mi hermanito Klener, por llenar mis días de alegría y por compartir conmigo incontables aventuras entre bloques de LEGO y Minecraft, recordándome que siempre hay espacio para soñar y crear.

*Adriana Stefany Mendoza López*

Dedico este trabajo a mis padres, Gonzalo Rodríguez y Rocío Galarza, por su amor incondicional, esfuerzo y apoyo constante a lo largo de todo este proceso. A mis hermanas, Erika y Evelyn, cuyo cariño y aliento diario han constituido una fuente permanente de motivación y fortaleza. Finalmente, a todas las personas que, de una u otra manera, me han impulsado a continuar y culminar este proceso con éxito.

*Gonzalo Paul Rodríguez Galarza*

Le dedico este trabajo a mis padres Ernesto Torres y Myriam Ramírez, que siempre han sido una parte fundamental de mi desarrollo como persona y como profesional.

*Iván Patricio Torres Ramírez*

Dedico este trabajo a mis hijos, Mateo, Isaac y Josué, quienes son la razón más importante de cada esfuerzo, sacrificio y sueño alcanzado. Si algo he aprendido en el camino, es que las respuestas más valiosas de la vida no se encuentran únicamente en la lógica o la razón, sino en las personas que llenan de significado nuestro existir. Recuerden que somos como los robles, difíciles de derrumbar.

*Víctor German Huilca Revelo*

**AGRADECIMIENTOS**

Expreso mi sincero agradecimiento a todas las personas que han contribuido a mi crecimiento profesional y personal durante este camino. Agradezco especialmente a las mujeres que he conocido en mi trayectoria laboral, quienes con su ejemplo, liderazgo y determinación han sido fuente de inspiración para continuar alcanzando nuevas metas.

*Adriana Stefany Mendoza López*

Agradezco profundamente a mis padres y hermanas por el apoyo incondicional y el acompañamiento constante brindado durante el desarrollo de este trabajo de titulación. De igual manera, expreso mi gratitud a mis familiares y amigos más cercanos, por su estímulo permanente y por la motivación que me brindaron en cada una de las etapas de este proceso.

*Gonzalo Paul Rodríguez Galarza*

Agradezco a mis familiares, amigos y compañeros de carrera en la maestría. Mis especiales agradecimientos a mis padres Ernesto Torres y Myriam Ramírez por todo el apoyo que me han brindado tanto en mi vida personal como profesional.

*Iván Patricio Torres Ramírez*

Agradezco profundamente a Dios por su infinita gracia, por la salud, la fortaleza y la sabiduría que me concedió para culminar esta importante etapa de mi vida. Le agradezco también por haber sembrado en mi corazón la pasión por el conocimiento, la tecnología y la búsqueda constante de soluciones que contribuyan al bienestar de la sociedad. Cada aprendizaje, desafío superado y logro alcanzado durante este camino ha sido una muestra de su guía, propósito y bendición.

*Víctor German Huilca Revelo*

## RESUMEN

Este proyecto de titulación tiene como objetivo desarrollar modelos predictivos basados en técnicas de aprendizaje automático y minería de datos que permitan identificar y anticipar de forma temprana la demanda territorial de atención social en los cantones y parroquias de la provincia de Manabí, utilizando información proveniente del Sistema de Información Local (SIL). La investigación se desarrolló bajo un enfoque cuantitativo, con un diseño no experimental y un alcance descriptivo y predictivo. Las fuentes principales de información fueron registros administrativos de violencia de género, acciones de salud integral y preventiva, y atención y asistencia humanitaria integral. La metodología adoptada fue CRISP-DM, estándar para el desarrollo de proyectos de minería de datos y ciencia de datos, abarcando las fases de comprensión del negocio, comprensión de los datos, preparación de los datos, modelado, evaluación y despliegue. El proyecto demuestra que los registros administrativos pueden transformarse en información útil para el desarrollo de modelos predictivos de demanda social. Los resultados obtenidos proporcionan una metodología aplicable para fortalecer la planificación y la toma de decisiones basada en evidencia dentro de la gestión pública.

**Palabras clave:** ciencia de datos, minería de datos, aprendizaje automático, demanda territorial, Sistema de Información Local (SIL), registros administrativos, gestión pública.

**ABSTRACT**

This graduation project aims to develop predictive models based on machine learning and data mining techniques that allow for the early identification and anticipation of the territorial demand for social services in the cantons and parishes of the province of Manabí, using information obtained from the Local Information System (LIS). The research was conducted under a quantitative approach, with a non-experimental design and a descriptive and predictive scope. The main sources of information were administrative records related to gender-based violence, comprehensive and preventive health actions, and comprehensive humanitarian assistance. The methodology adopted was CRISP-DM, a standard framework for data mining and data science projects, encompassing the phases of business understanding, data understanding, data preparation, modeling, evaluation, and deployment. The project demonstrates that administrative records can be transformed into valuable information for developing predictive models of social service demand. The results provide an applicable methodology to strengthen planning and evidence-based decision-making within public administration.

**Keywords:** data science, data mining, machine learning, territorial demand, Local Information System (LIS), administrative records, public management.

# Índice

## Índice de tablas

## Índice de figuras

# CAPÍTULO 1: Introducción

## 1.1 Definición de proyecto

La atención social en los Gobiernos Autónomos Descentralizados (GAD) provinciales de Ecuador son competencias obligatorias orientadas a garantizar los derechos de los grupos de atención prioritaria. En la provincia de Manabí, esto se gestiona a través de la Dirección de Desarrollo Humano del Gobierno Autónomo Descentralizado Provincial de Manabí [GADP-Manabí] (2025), donde abarcan áreas vitales como salud preventiva, asistencia humanitaria, atención a víctimas de violencia de género, vivienda rural y el desarrollo de actividades educativas, deportivas y culturales.

A pesar de la importancia de estas acciones, uno de los principales desafíos en cualquier gestión pública actual, consiste en anticipar las necesidades futuras de atención social para fortalecer procesos de planificación y asignación de recursos. Debido a esto, el uso de información histórica mediante técnicas de análisis predictivo representa una oportunidad para complementar los mecanismos tradicionales de toma de decisiones y orientar de manera más eficiente las intervenciones en los territorios.

El uso estratégico de información es fundamental para superar estos desafíos. La Organización para la Cooperación y el Desarrollo Económicos [OCDE] (2020) señala que la toma de decisiones basada en datos empíricos permite a las instituciones adaptarse de una gobernanza reactiva a una proactiva, mitigando sesgos en los registros y centrando la planificación en las necesidades de la ciudadanía. Asimismo, la analítica de datos es fundamental para convertir grandes volúmenes de datos gubernamentales en herramientas de gestión estratégica que mejoren la asignación de recursos, optimicen la administración y anticipen riesgos (Pérez Lara et al., 2025).

En Ecuador, los Sistemas de Información Local (SIL) son plataformas de implementación obligatoria para los GAD, diseñadas para generar estadísticas territoriales, mapear demandas ciudadanas y dar seguimiento a los planes de desarrollo (Lema Changoluisa & Romo, 2022). En la Prefectura de Manabí, su Sistema de Información Local (SIL) reúne y administra los registros administrativos, los cuales son capturados en diferentes periodos de tiempo de los programas y actividades que realiza la institución (GADP-Manabí, 2026). Cuando estos registros son sometidos a procesos de limpieza y estructuración, dejan de ser simples archivos de trámites para convertirse en fuentes de información estadística, de bajo costo y esenciales para evidenciar la gestión pública (Instituto Nacional de Estadística y Censos [INEC], 2022).

Considerando las limitaciones de la planificación reactiva, este proyecto propone el desarrollo de modelos predictivos para la identificación temprana de la demanda territorial de atención social en Manabí, utilizando los registros administrativos provistos por el SIL de la Prefectura de Manabí. Mediante la metodología CRISP-DM y técnicas de aprendizaje automático, se busca transformar los registros administrativos en conocimiento útil, incorporando modelos generativos como CTGAN para el aumento y balanceo de datos cuando sea necesario. Esta solución proporcionará a las autoridades las herramientas necesarias para estimar demandas futuras, garantizando intervenciones sociales oportunas y focalizadas.

### 1.2 Justificación e importancia del trabajo de investigación

Actualmente, la gestión pública se enfrenta al desafío de transitar hacia un modelo de toma de decisiones basado en evidencia, aprovechando los grandes volúmenes de información generados diariamente. Lemus-Delgado y Pérez Navarro (2020) sostienen que el nuevo tipo

de información derivado de los datos masivos ha originado nuevas formas de recopilar y analizar el entorno social, permitiendo ir más allá de la simple lectura de información y convirtiendo los registros administrativos en verdaderas herramientas de gestión estratégica diseñadas para optimizar recursos públicos y anticipar riesgos.

Esta transición tecnológica se justifica mediante la adopción de modelos predictivos que usan técnicas de minería de datos y aprendizaje automático para estimar futuros escenarios a partir del comportamiento pasado, lo cual resulta ser un pilar fundamental para la anticipación de demanda social. Por lo tanto, el proyecto usa un enfoque analítico avanzado que supera la perspectiva puramente descriptiva, utilizando algoritmos para descubrir patrones existentes en la información.

Desde una perspectiva institucional, este trabajo adquiere relevancia al promover el uso estratégico del Sistema de Información Local (SIL) gestionado por el GADP-Manabí. Lema Changoluisa y Romo (2022) argumentan que los registros administrativos institucionales constituyen una fuente primaria vital para el seguimiento de las acciones locales y para evidenciar el resultado de las políticas públicas, brindando a las instituciones públicas herramientas para una gestión transparente y sustentada.

La justificación de esta investigación se enfoca en la mejora significativa de la planificación de los servicios sociales destinados a las poblaciones vulnerables. Tradicionalmente, la atención social suele enfocarse de manera retrospectiva, lo que en ocasiones genera un desequilibrio de la cobertura territorial y una distribución poco eficiente de presupuesto y personal. La aplicación de modelos predictivos sobre registros de atención social permite la

identificación temprana de las necesidades territoriales, mucho antes de que estas se manifiesten de forma crítica.

Tal como lo señala Ortega (2025), la ciencia de datos funciona como un motor de transformación social orientado a mejorar la calidad de vida y a dirigir a las comunidades hacia un futuro más justo y equitativo. En el contexto de Manabí, prever los patrones temporales y territoriales de los requerimientos ciudadanos garantiza una respuesta gubernamental más ágil y justa para quienes más lo necesitan.

Por otro lado, la investigación se justifica desde el ámbito académico al plantear un caso de estudio aplicado al sector público. Aunque existe una vasta literatura en ciencia de datos, su aplicación práctica para la analítica predictiva de registros administrativos locales a nivel gubernamental es un área con gran potencial de exploración.

Finalmente, este trabajo enriquece la literatura sobre la modernización tecnológica del país, demostrando cómo las disciplinas tecnológicas y matemáticas contribuyen directamente a la resolución de problemas públicos y al bienestar social.

### 1.3 Alcance

Este proyecto se limita a la exploración de datos provenientes del Sistema de Información Local (SIL) del Gobierno Autónomo Descentralizado de Provincial de Manabí, una plataforma tecnológica administrada por la Dirección de Planificación para el Desarrollo, la cual consolida la información de servicios de atención social en el territorio entre todas las actividades que realiza la institución. El uso de estos registros administrativos, generados a partir de las operaciones cotidianas de la institución, constituye la base empírica para

transitar de un análisis puramente descriptivo hacia la generación de conocimiento estadístico y predictivo riguroso.

A la fecha de desarrollo de este proyecto, se dispone de 7199 registros correspondientes al período 2024–2025. En términos conceptuales, esta delimitación estratégica ha priorizado tres fuentes y ejes críticos de la gestión social provincial: la atención a víctimas de violencia de género, las acciones de salud integral y preventiva, y los servicios de atención y asistencia humanitaria integral.

Finalmente, este proyecto reconoce sus limitaciones técnicas enfocándose en las variables y la calidad estructural predefinidas por el SIL para los tres registros administrativos seleccionados, limitándose estrictamente a la entrega de herramientas analíticas y recursos técnicos para optimizar la planificación territorial y mitigar la gestión reactiva. Por lo tanto, no se contempla la evaluación de impacto a largo plazo de los servicios prestados, el análisis cualitativo particular de casos, ni la creación de políticas públicas.

## 1.4 Objetivos

### 1.4.1 Objetivo General

Desarrollar modelos predictivos fundamentados en técnicas de aprendizaje automático y minería de datos que permitan identificar y anticipar tempranamente la demanda territorial de atención social en los cantones y parroquias de la provincia de Manabí, a partir de la integración y el aumento sintético de registros administrativos históricos, con el propósito de optimizar la asignación de recursos y fortalecer un modelo de gestión pública.

### 1.4.2 Objetivos Específicos

1. Consolidar un conjunto de datos analítico mediante la integración y depuración de los registros administrativos del Sistema de Información Local (SIL) correspondientes a violencia de género, salud integral y preventiva, y asistencia humanitaria.

2. Implementar la arquitectura de aprendizaje profundo CTGAN (Conditional Tabular Generative Adversarial Network) para generar datos sintéticos que permitan balancear e incrementar el volumen de los registros administrativos.

3. Desarrollar modelos predictivos de clasificación para categorizar los cantones y parroquias de la provincia de Manabí según su nivel de demanda futura de atención social.

4. Evaluar el desempeño de los modelos predictivos mediante métricas estadísticas que permitan validar su capacidad de generalización y su utilidad para apoyar la planificación de la atención social.

# CAPÍTULO 2: Revisión de Literatura

## 2.1 Estado del arte

### 2.1.1. Uso de registros administrativos para análisis social

El rol de los registros administrativos ha evolucionado dentro de un entorno político y social; dejando de ser herramientas operativas para transformarse en un pilar fundamental para el diseño, monitoreo y evaluación de las políticas públicas locales. Lema Changoluisa y Romo (2022) mencionan que en el ámbito de los Gobiernos Autónomos Descentralizados (GAD), los Sistemas de Información Local (SIL) resultan vitales para organizar y aprovechar los datos generados cotidianamente por las administraciones, permitiendo transparentar la gestión y generar evidencia oportuna alineada con las realidades y necesidades específicas de cada territorio.

Sin embargo, la utilidad de estos datos no es automática, el INEC (2022) por su lado, establece que la transformación de un registro administrativo en un registro estadístico requiere procesos de recopilación, perfilamiento, estandarización e integración, con el fin de garantizar la calidad y consistencia de la información para su análisis. Por su lado el Fondo de las Naciones Unidas para la Infancia [UNICEF], la Comisión Económica para América Latina y el Caribe [CEPAL], et al. (2024), plantea que el uso sistemático de los registros administrativos constituye un recurso clave para fortalecer los procesos de formulación, seguimiento y evaluación de políticas públicas orientadas a la atención de poblaciones en situación de vulnerabilidad. Bajo este contexto, ambos planteamientos muestran que el valor de los registros administrativos no se basa únicamente en su disponibilidad, sino en los procesos metodológicos aplicados para garantizar su calidad y posterior aprovechamiento analítico.

### 2.1.2. Aplicaciones de ciencia de datos en el sector público

La disponibilidad de registros administrativos ha permitido la posibilidad de que las instituciones públicas puedan incorporar progresivamente técnicas de ciencia de datos para fortalecer la toma de decisiones basada en evidencia. En este contexto, Ortega (2025) sostiene que la ciencia de datos podría transformar la información generada por las entidades gubernamentales en conocimiento útil para optimizar la gestión pública, mejorar la eficiencia institucional y orientar la formulación de políticas públicas.

Por su lado, Pérez y Covarrubias (2025) señalan que el verdadero aprovechamiento de la información pública no se limita a la publicación de datos abiertos, sino que requiere la aplicación de técnicas analíticas que permitan describir, comprender y anticipar el comportamiento de los fenómenos de interés.

Por lo que si bien, los registros administrativos además de ser únicamente un mecanismo de seguimiento de acciones institucionales y una fuente generadora de conocimiento, son en gran parte un insumo estratégico que, mediante la aplicación de técnicas de ciencia de datos permite generar información útil para apoyar la planificación y la gestión pública basada en evidencia.

### 2.1.3. Modelos predictivos para demanda de servicios

Dentro de las técnicas analíticas y técnicas de minería de datos, los algoritmos predictivos emergen como herramientas indispensables para la anticipación de las necesidades ciudadanas. Sepúlveda y Benavides (2023) argumentan que la predicción matemática y probabilística de la volumetría de atenciones es indispensable para generar una gestión

eficiente de la disponibilidad, capacidad y productividad de los servicios de las instituciones donde sea aplicado.

Es por eso que la implementación de modelos de regresión y clasificación permite estimar la proporción de servicios que se requerirán a futuro, optimizando la asignación de presupuestos y la focalización de las estrategias operativas. Siendo esta capacidad predictiva el principal enfoque transversal del presente proyecto, pues proporciona una forma de catalogar los cantones y parroquias según sus niveles futuros de prioridad, permitiendo a cualquier institución pública que maneje estos programas sociales, garantizar una respuesta estatal oportuna, dimensionada e inteligente ante las demandas territoriales de atención social.

### 2.1.4. Generación de datos sintéticos en investigación

A pesar del vasto potencial de la ciencia de datos en la gestión pública, la construcción de modelos predictivos precisos frecuentemente tropieza con limitaciones metodológicas originadas por la escasez, la irregularidad temporal o el fuerte desbalance de los registros históricos disponibles. Para superar estas deficiencias, se recurre a la generación de datos sintéticos mediante el uso de inteligencia artificial. Los datos sintéticos son observaciones creadas algorítmicamente que imitan las propiedades, varianzas y distribuciones matemáticas de registros con los que se alimenta, pero sin contener información genuina que comprometa la identidad o confidencialidad de los individuos (Amazon Web Services [AWS], s.f.).

Autores como Sainz-Aja (2025) y Arroni del Riego (2024) demuestran que el uso de arquitecturas generativas profundas, tales como las Redes Generativas Antagónicas (GAN,

por sus siglas en inglés) y los modelos de difusión, permite incrementar artificialmente el tamaño de las bases de datos tabulares mientras se preservan las correlaciones multivariantes estructurales. Esta aumentación artificial de datos resulta vital en entornos de registros limitados, ya que resuelve la insuficiencia de observaciones para el entrenamiento algorítmico, reduciendo el sesgo y aumentando drásticamente el rendimiento, la estabilidad y la capacidad de generalización de los modelos de predicción (Sánchez, 2022).

Por consiguiente, la síntesis de información viabiliza la creación de algoritmos predictivos robustos para el ámbito del desarrollo humano, garantizando resultados fiables aun cuando el repositorio administrativo original carezca de la profundidad histórica ideal.

## 2.2. Marco Teórico

### 2.2.1. Sistemas de Información y Gestión Pública Basada en Datos

Los sistemas de información constituyen uno de los principales soportes para la gestión de las organizaciones, ya que permiten recopilar, almacenar, procesar y transformar datos en información útil para apoyar la toma de decisiones, permitiendo responder a las necesidades de la organización (Trasobares, 2003). En este sentido, la importancia de un sistema de información reside en su capacidad para transformar los datos en conocimiento útil que contribuya al logro de los objetivos institucionales.

Dentro del sector público, los sistemas de información adquieren mayor relevancia, debido que usan grandes volúmenes de información generada por la prestación de servicios y la ejecución de programas gubernamentales. Considerando esto, la transformación digital en

el sector público constituye un fenómeno de alcance global que ha impulsado la evolución de los sistemas de información hacia modelos de gestión pública basados en datos.

El uso de información institucional mediante herramientas tecnológicas y analíticas ha dejado de ser una posible alternativa a convertirse en un componente esencial de la modernización administrativa (Flores-Cano et al., 2026), generando valor público cuando la información producida es utilizada para comprender problemas sociales y apoyar la toma de decisiones de manera colaborativa y eficiente (Quintanilla y Gil-García, 2016). Sumado al contexto de los gobiernos locales y provinciales, los sistemas de información territorial permiten integrar y analizar información proveniente de distintos ámbitos de gestión, facilitando la identificación de necesidades ciudadanas y fortaleciendo la planificación de intervenciones públicas.

### 2.2.2. Registros Administrativos como Fuente de Información para la Toma de Decisiones

Los registros administrativos están conformados por el conjunto de actividades, datos y recursos que recopila una entidad pública de manera continua en el ejercicio legal de sus funciones o en su interacción diaria con la ciudadanía (Lema y Romo, 2022), conformando un enorme potencial estadístico y analítico. Su utilización presenta ventajas significativas frente a los métodos de recolección tradicionales (como censos o encuestas), con menores costos de producción, disponibilidad de información oportuna y continua, e indicadores más desagregados (Caminos-Maldonado y Vásquez, 2022).

Para que los registros administrativos puedan utilizarse de manera confiable en la toma de decisiones y en la formulación de políticas públicas, es necesario someterlos a un proceso

técnico que permita su conversión en registros estadísticos. Este proceso implica tareas de recopilación estandarizada, validación de variables de identificación, limpieza, seudonimización y estandarización de metadatos (INEC, 2022). Como resultado, la disponibilidad de información estructurada, continua y prácticamente en tiempo real fortalece la capacidad de la administración pública local para monitorear y evaluar la efectividad de las políticas sociales con base en evidencia empírica.

### 2.2.3. Aprendizaje Automático y Analítica Predictiva

Arthur Samuel (1959) definió el aprendizaje automático (Machine Learning) como la capacidad de un sistema para aprender a partir de la experiencia y mejorar su desempeño de forma autónoma, sin requerir que cada comportamiento sea programado de manera explícita. Este planteamiento sentó las bases de una de las áreas más importantes de la inteligencia artificial, experimentado un notable desarrollo con el paso de los años.

En la actualidad, el crecimiento acelerado de los registros administrativos y de otras fuentes de información ha generado la necesidad de incorporar herramientas analíticas más avanzadas, capaces de procesar grandes volúmenes de datos y complementar las técnicas estadísticas tradicionales. En este contexto, el aprendizaje automático se ha consolidado como una tecnología que permite identificar patrones, relaciones complejas y comportamientos no lineales de manera automatizada, facilitando la generación de conocimiento útil para apoyar la toma de decisiones basada en datos (Sepúlveda y Benavides, 2023).

Cuando estas técnicas se integran con la analítica predictiva, es posible desarrollar modelos capaces de estimar la ocurrencia de eventos futuros a partir del comportamiento

histórico de las variables analizadas. De esta manera, la analítica predictiva deja de ser únicamente una herramienta tecnológica para convertirse en un elemento estratégico en los ámbitos de la gestión pública y organizacional, permitiendo respaldar la toma de decisiones mediante proyecciones fundamentadas en evidencia, reduciendo la dependencia de criterios basados en la intuición (Casanova-Villalba et al., 2023).

### 2.2.4. Calidad y Preparación de Datos

La calidad de los datos es un aspecto determinante en cualquier proceso de análisis, ya que la validez de los resultados depende directamente de la confiabilidad de la información utilizada. De acuerdo con Herrera (2016), diversos factores, como la presencia de ruido, valores faltantes, inconsistencias, datos redundantes o un número excesivo de variables, pueden afectar la calidad de los datos y limitar la obtención de conocimiento útil. Por esta razón, contar con grandes volúmenes de información no asegura mejores resultados si previamente no se realizan procesos adecuados de limpieza, organización y preparación.

En el ámbito del aprendizaje automático, la calidad de los datos influye de manera directa en el desempeño de los modelos predictivos. Por ello, antes de entrenar un algoritmo, es necesario aplicar técnicas de preprocesamiento que permitan mejorar la consistencia de la información. Entre las más utilizadas se encuentran la normalización de los datos, la imputación de valores faltantes y la transformación de variables categóricas. Estas actividades contribuyen a generar un conjunto de datos más uniforme y confiable, reduciendo posibles sesgos y favoreciendo un mejor proceso de aprendizaje del modelo (Cabrera-Sánchez et al., 2024).

Como etapa inicial del análisis, resulta fundamental desarrollar un Análisis Exploratorio de Datos (EDA, por sus siglas en inglés), cuyo propósito es comprender las características y el comportamiento del conjunto de datos antes de construir los modelos predictivos. Este análisis incluye el cálculo de estadísticas descriptivas, el estudio de la distribución y dispersión de las variables, así como el análisis de correlaciones, lo que permite identificar patrones, relaciones entre variables y posibles anomalías que podrían influir en los resultados del modelado.

### 2.2.5. Generación de Datos Sintéticos para Datos Tabulares (CTGAN)

En el contexto de la gestión pública y social, el desarrollo de modelos analíticos suele verse limitado por la escasez de datos disponibles para determinados fenómenos, así como por las normativas que regulan la privacidad y protección de la información de la ciudadanía. Ante esta situación, la inteligencia artificial generativa ha favorecido el uso de datos sintéticos, los cuales son generados artificialmente con el propósito de reproducir las características, relaciones y patrones estadísticos presentes en los datos originales, sin comprometer la identidad ni la información personal de las personas (Del Riego, 2024; Velilla, 2025). Además, estos datos son utilizados para la creación de conjuntos de prueba, la validación de modelos y el entrenamiento de algoritmos de aprendizaje automático. Su principal ventaja radica en que permiten compartir y utilizar información con fines analíticos sin revelar datos confidenciales, convirtiéndose en una alternativa segura para proteger la privacidad y, al mismo tiempo, conservar el valor analítico de la información (Pathare et al., 2023).

Sin embargo, la generación de datos sintéticos en conjuntos de datos tabulares representa un reto importante. A diferencia de imágenes o textos, este tipo de datos suele integrar

variables continuas y categóricas, distribuciones desbalanceadas y relaciones complejas entre sus atributos. Estas particularidades dificultan que los métodos estadísticos convencionales y diversos modelos de aprendizaje profundo reproduzcan de forma precisa el comportamiento de los datos originales (Xu et al., 2019).

En este contexto, CTGAN (Conditional Tabular Generative Adversarial Network) se ha consolidado como una de las técnicas más efectivas para generar datos sintéticos en información tabular. Su capacidad para preservar la estructura y la distribución de los datos permite ampliar y equilibrar los conjuntos de información utilizados durante el entrenamiento de modelos predictivos. En la presente investigación, esta técnica contribuyó a incrementar el número de registros disponibles, manteniendo la coherencia estadística del conjunto de datos y favoreciendo el desarrollo de modelos con mayor capacidad de aprendizaje y generalización.

### 2.2.6. Balanceo de Datos

En el análisis de registros provenientes del sector público es común encontrar conjuntos de datos desbalanceados, donde algunos eventos de interés, como los casos de mayor vulnerabilidad social, representan una proporción mucho menor en comparación con las atenciones más frecuentes. Esta situación puede afectar el desempeño de los modelos de clasificación, ya que estos tienden a aprender principalmente de la clase mayoritaria, obteniendo una alta precisión general (Accuracy), pero mostrando dificultades para identificar correctamente los casos menos frecuentes, que suelen ser los de mayor relevancia para la investigación (Morales et al., 2022).

Para reducir este sesgo y mejorar la capacidad de los modelos para reconocer las clases minoritarias, especialmente en términos de sensibilidad (Recall), es necesario aplicar técnicas de balanceo durante la etapa de preprocesamiento de los datos. Entre las estrategias más utilizadas se encuentran el submuestreo (undersampling), que disminuye la cantidad de registros de la clase mayoritaria, y el sobremuestreo (oversampling), que incrementa la representación de la clase minoritaria mediante la creación de nuevos registros sintéticos. Una de las técnicas más empleadas para este propósito es SMOTE (Synthetic Minority Over-sampling Technique), la cual genera nuevas instancias a partir de las características de los datos existentes, contribuyendo a mejorar el aprendizaje y la capacidad predictiva de los modelos (Morales et al., 2022). El equilibrio adecuado de las clases en las bases de datos de disponibles asegurará que los algoritmos sean capaces de detectar de forma temprana las zonas o grupos demográficos específicos con alta demanda de atención, independientemente de que históricamente hayan representado un volumen estadístico menor.

### 2.2.7. Metodología CRISP-DM

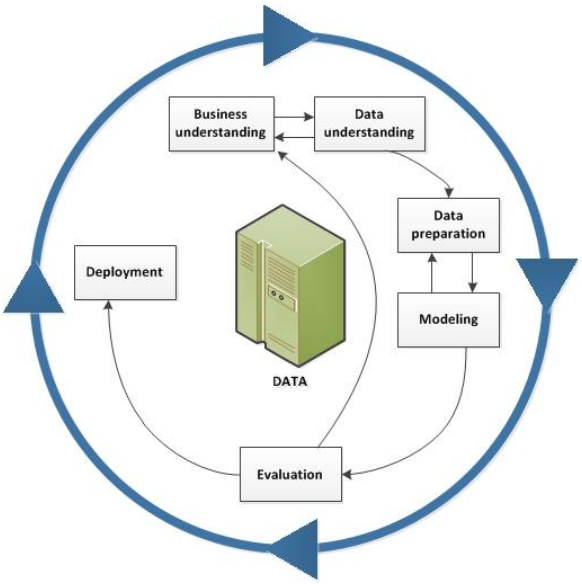
Para garantizar que el ciclo completo de la analítica de datos se desarrolle de forma ordenada y cumpla con los objetivos institucionales, es necesario adoptar un estándar procedimental comprobado. La metodología CRISP-DM (Cross Industry Standard Process for Data Mining) se ha conoslidado como el marco de referencia más utilizado en proyectos de ciencia de datos, debido a que proporciona una guía estructurada para planificar, desarrollar y evaluar soluciones basadas en datos (Shearer, 2000; Espinosa, 2020).

De acuerdo con Chapman et al. (1999), CRISP-DM se encuentra estructurada mediante un modelo jerárquico de cuatro niveles de abstracción: fases, tareas genéricas, tareas especializadas e instancias del proceso. Esta organización permite definir un marco metodológico suficientemente general para adaptarse a distintos tipos de proyectos de minería de datos, al mismo tiempo que facilita su personalización según el contexto, las características de los datos y los objetivos específicos de cada investigación.

Según IBM (s. f.), la metodología CRISP-DM define el ciclo de vida de un proyecto de minería de datos a través de seis fases principales: comprensión del negocio, comprensión de los datos, preparación de los datos, modelado, evaluación y despliegue. Aunque estas etapas siguen un orden lógico, la metodología se caracteriza por su enfoque iterativo y flexible, permitiendo regresar a fases previas cuando los resultados obtenidos evidencian la necesidad de realizar ajustes, replantear decisiones o mejorar tanto los datos como los modelos desarrollados.

**Figura 1**

*Ciclo de vida de minería de datos*




Fuente: IBM

En el contexto de este proyecto, una de las principales fortalezas de la metodología CRISP-DM es que, desde su etapa inicial, promueve la definición clara de los objetivos del proyecto antes de iniciar el tratamiento de los datos. Este enfoque garantiza que las técnicas y herramientas analíticas utilizadas respondan a las necesidades de la investigación y aporten información útil para apoyar la gestión y la planificación en el ámbito local.

Asimismo, la adopción de esta metodología permitió organizar de manera sistemática cada una de las etapas del estudio, desde la recopilación y depuración de los registros administrativos hasta la aplicación de técnicas como CTGAN y los métodos de balanceo de datos, culminando con el desarrollo del modelo predictivo para la atención social. De esta forma, CRISP-DM no solo facilitó el proceso técnico de análisis, sino que también contribuyó a que los resultados obtenidos pudieran convertirse en una herramienta práctica para respaldar la toma de decisiones y fortalecer la planificación territorial basada en evidencia.

# CAPÍTULO 3: Desarrollo del trabajo

## 3.1. Metodología de Investigación

La investigación se desarrolla bajo un enfoque cuantitativo, debido a que emplea registros administrativos estructurados y técnicas de análisis estadístico y aprendizaje automático para la construcción de modelos predictivos. Corresponde a una investigación aplicada, ya que busca resolver un problema específico de planificación territorial mediante el desarrollo de un modelo de predicción de la demanda social. El diseño es no experimental, longitudinal y retrospectivo, porque analiza datos históricos sin manipular las variables de estudio y considera varios períodos de tiempo. Su alcance es descriptivo, correlacional y predictivo, al describir la demanda territorial, analizar la relación entre variables y generar predicciones para apoyar la toma de decisiones. La unidad de análisis está constituida por los cantones de la provincia de Manabí observados mensualmente durante los años seleccionados, representados mediante variables agregadas obtenidas de los registros administrativos del Sistema de Información Local (SIL).

### 3.1.1. Enfoque de investigación

Este proyecto se desarrolló bajo un enfoque metodológico cuantitativo, dado que su ejecución se fundamentó en la recolección, estructuración y procesamiento de datos numéricos y categóricos medibles. A través de este enfoque, se aplicaron técnicas de análisis estadístico y algoritmos de ciencia de datos con el propósito de identificar patrones y tendencias dentro de la información. Asimismo, el estudio se enmarca en la clasificación de investigación aplicada; esta categorización responde a que el proyecto buscó resolver un problema práctico e institucional mediante el aprovechamiento de datos reales para apoyar la toma de decisiones en la gestión pública provincial. Es importante señalar que el trabajo

se diseñó como una prueba de concepto orientada al diseño y desarrollo de una herramienta analítica replicable; por lo que el propósito principal no fue la obtención o implementación de resultados predictivos reales en el entorno de producción, sino la validación metodológica del proceso de construcción del modelo a partir de los datos administrativos disponibles.

### 3.1.2. Diseño metodológico

El diseño usado para la presente investigación fue de carácter no experimental, debido a que en ninguna etapa del estudio existió una manipulación o control de las variables. El proyecto se enfocó en un análisis de registros administrativos provenientes del Sistema de Información Local (SIL) del GADPM, los cuales fueron recopilados, integrados, depurados y transformados para su posterior utilización en procesos de modelado predictivo. En base a esto, las observaciones analizadas corresponden a hechos ya ocurridos y registrados previamente, por lo que el estudio se desarrolló en un contexto de observación y análisis de información existente.

Referente al alcance, la investigación se comprendió en dos niveles complementarios. En una primera etapa se desarrolló un alcance descriptivo, orientado a la comprensión de las características de los datos mediante procesos de exploración, perfilamiento, evaluación de calidad e identificación de patrones presentes en las diferentes fuentes de información. Esta fase permitió conocer la estructura de los registros, detectar inconsistencias y establecer las condiciones necesarias para su preparación e integración.

Posteriormente, el estudio adquirió un alcance predictivo, al incorporar técnicas de aprendizaje automático y minería de datos con el propósito de desarrollar modelos capaces

de identificar patrones asociados a la demanda territorial de atención social. Para ello se ejecutaron procesos de preparación de datos, balanceo de observaciones y generación sintética mediante CTGAN, con la finalidad de disponer de un conjunto de datos adecuado para el entrenamiento y evaluación de modelos de clasificación orientados a anticipar escenarios de demanda futura en los cantones y parroquias de la provincia de Manabí.

### 3.1.3. Aplicación de CRISP-DM

Para estructurar, sistematizar y guiar el desarrollo de la prueba de concepto, se implementó la metodología CRISP-DM (Cross Industry Standard Process for Data Mining). Este marco de trabajo es ampliamente reconocido y utilizado a nivel global en proyectos de minería y ciencia de datos, pues proporciona una aproximación estructurada y estandarizada para la resolución de problemas complejos mediante la explotación de información. El uso de esta metodología permitió organizar el ciclo de vida de la investigación en fases lógicas e iterativas, enfocando los mayores esfuerzos en las etapas del proceso analítico.

En la primera etapa, correspondiente a la Comprensión del Negocio, se definieron los objetivos del proyecto y se analizó a profundidad el contexto del problema gubernamental. Durante esta fase se determinó el potencial del Sistema de Información Local (SIL) de la provincia de Manabí no solo como un repositorio de registros administrativos, sino como la fuente principal de datos para crear herramientas de análisis avanzado.

Posteriormente, en la fase de Comprensión de los Datos, se procedió con la recolección de tres fuentes de datos independientes extraídas directamente del SIL. En este punto se ejecutó un primer Análisis Exploratorio de Datos (EDA) con el fin de examinar la naturaleza preliminar de los registros. Esta exploración inicial permitió la identificación de las variables

relevantes, la revisión de los tipos de datos, la detección de valores atípicos y la estructuración del diccionario de datos que serviría de guía para las siguientes intervenciones.

La mayor carga metodológica y operativa de la investigación recayó sobre la fase de Preparación de los Datos, lo cual es una característica común en los proyectos de ciencia de datos basados en casos reales. Durante esta etapa se llevó a cabo la integración y unificación de las tres fuentes originales en una única estructura tabular consolidada. A partir de dicha integración, se desarrollaron procesos de limpieza, homologación de formatos y transformación de variables para garantizar la coherencia del conjunto de datos. Una característica distintiva de esta fase fue la evaluación continua de la calidad de los datos, la cual se gestionó mediante la aplicación iterativa del EDA a lo largo del proyecto. La exploración de datos se repitió cíclicamente para validar cada corrección, comprender las distribuciones tras el tratamiento de valores faltantes y perfeccionar de forma progresiva la calidad de la información unificada.

Durante la etapa de preparación de los datos se identificó una limitación importante relacionada con la cantidad de registros históricos disponibles en el SIL. Esta situación representaba un desafío para el entrenamiento de los modelos predictivos, ya que un volumen reducido de datos puede afectar su capacidad de aprendizaje y, en consecuencia, su desempeño.

Con el fin de superar esta limitación sin alterar la calidad ni la confiabilidad de la información, se aplicaron técnicas de balanceo y generación de datos sintéticos. Para ello, se utilizó la arquitectura CTGAN (Conditional Tabular Generative Adversarial Network), la

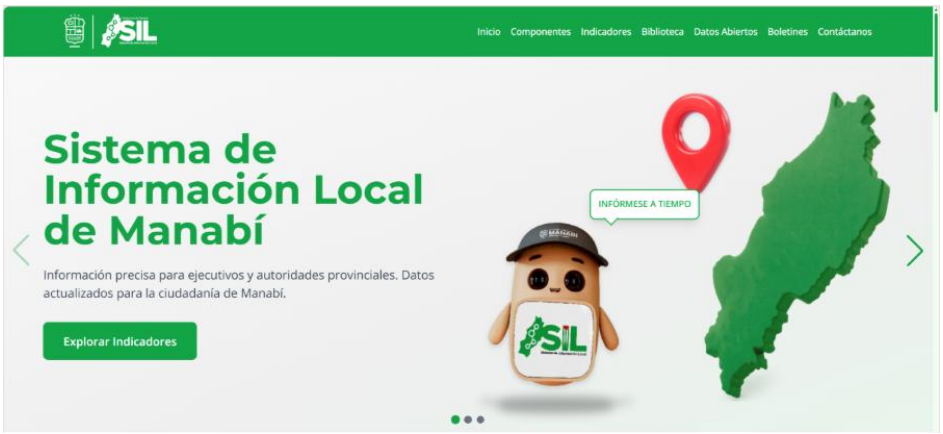
cual permitió generar nuevos registros artificiales que conservaron las características, distribuciones y relaciones estadísticas presentes en los datos originales, incrementando así el tamaño del conjunto de datos disponible para el entrenamiento.

Finalmente, la ejecución secuencial y cíclica de las tareas de limpieza, integración, exploración iterativa y aumento de datos permitió concluir la fase de preparación de manera exitosa, dando como resultado la construcción de un dataset analítico depurado, balanceado y listo para ser utilizado en la creación del modelo predictivo propuesto.

### 3.2. Fase I: Comprensión del Negocio

En el contexto de la gestión pública del Gobierno Autónomo Descentralizado Provincial de Manabí (GADPM), el Sistema de Información Local (SIL) ejerce un papel fundamental como la plataforma tecnológica y el medio institucional para gestionar, organizar y acceder a la información. En este sentido, la fuente oficial y primaria de información del SIL son los registros administrativos, los cuales se generan de manera continua a partir de la gestión operativa diaria y la prestación de servicios a la ciudadanía.

**Figura 2**
*Página web del SIL*




Fuente. Página oficial del Sistema de Información Local de Manabí (SIL)

Esta plataforma tecnológica permite centralizar y consolidar de manera estructurada los diversos registros administrativos generados por las diferentes áreas y programas institucionales. Específicamente, dentro del eje de desarrollo humano gestionado en el SIL, reposan registros administrativos asociados a líneas de acción críticas dirigidas por la Dirección de Desarrollo Humano, entre las cuales destacan las mostradas en la Tabla 1.

**Tabla 1**

*Registros administrativos del Sistema de Información Local de Manabí (SIL)*

| Id RA | Registro Administrativo | Descripción |
|---|---|---|
| RA_020 | VIOLENCIA DE GENERO | Registra las acciones enfocadas en la prevención de la violencia intrafamiliar dirigidas a grupos de atención prioritaria. |
| RA_039 | ACCIONES DE SALUD INTEGRAL Y PREVENTIVA | Registra los servicios de salud integral y preventiva que se brindan a través de brigadas médicas y centros de rehabilitación física |
| RA_040 | ATENCIÓN Y ASISTENCIA HUMANITARIA INTEGRAL | Registra la entrega de ayudas técnicas y humanitarias orientadas específicamente a personas en situación de vulnerabilidad o grupos prioritarios. |

Fuente. Sistema de Información Local de Manabí (SIL)

Estos registros no solo documentan el accionar operativo y burocrático de la entidad, sino que representan una valiosa información histórica sobre las intervenciones realizadas, constituyendo una fuente potencial para análisis de patrones y comportamientos continuo de demanda social en la provincia. A partir de la disponibilidad de esta información, se identifica una clara oportunidad para aprovechar el potencial de los registros administrativos y transformar la información histórica en insumos analíticos avanzados;

fortaleciendo significativamente la toma de decisiones basada en evidencia y posibilitando la transición hacia una gobernanza proactiva.

El propósito de esta iniciativa analítica es identificar los patrones relacionados a la demanda territorial de atención social, estimando su comportamiento en los diferentes cantones y parroquias de la provincia de Manabí. Si bien el proyecto reconoce limitaciones prácticas referente a la disponibilidad y al volumen temporal de la información histórica recopilada en los repositorios de la institución, el enfoque estratégico de este proyecto prioriza aprovechar al máximo los datos administrativos ya existentes en el SIL.

### 3.3. Fase II: Comprensión de los Datos

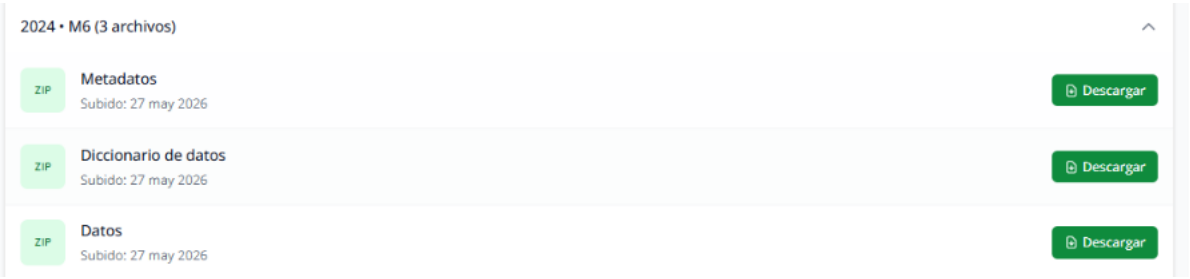
#### 3.3.1. Recolección y verificación de datos

La fase de comprensión de los datos tuvo como objetivo identificar, recopilar y verificar las fuentes de información necesarias para el desarrollo del modelo predictivo de demanda territorial de atención social en la provincia de Manabí. Para ello, se utilizó información publicada en el Sistema de Información Local de Manabí (SIL), correspondiente a distintos registros administrativos generados por la Dirección de Desarrollo Humano del Gobierno Autónomo Descentralizado Provincial de Manabí.

Cada registro administrativo disponible en el SIL se encuentra organizado por períodos de reporte y, para cada período publicado, se dispone de un conjunto de recursos conformado por los datos en formato CSV, su respectivo diccionario de datos y un archivo de metadatos. Esta estructura permitió no solo acceder a la información operativa registrada, sino también comprender la definición de las variables, los criterios de registro y las características generales de cada conjunto de datos.

**Figura 3**

*Archivos disponibles en el Sistema de Información Local de Manabí (SIL)*




Fuente. Página oficial del Sistema de Información Local de Manabí (SIL)

Dado que la información histórica de un mismo registro administrativo se encontraba distribuida en múltiples períodos independientes, fue necesario realizar una etapa inicial de recopilación y consolidación. Para cada registro se descargaron los archivos correspondientes a todos los períodos disponibles, verificando previamente la consistencia de sus estructuras mediante la revisión de los diccionarios de datos y metadatos asociados. Posteriormente, los archivos CSV pertenecientes a un mismo registro administrativo fueron integrados en una única base de datos acumulada, aprovechando que mantenían una estructura homogénea de variables entre períodos.

Como resultado de este proceso de recopilación y consolidación, se obtuvieron tres conjuntos de datos principales, cada uno asociado a una línea de intervención social específica:

- **Violencia de Género:** conformado a partir de la integración de los diferentes archivos históricos reportados para este registro administrativo.
- **Acciones de Salud Integral y Preventiva**: construido mediante la unificación de los archivos correspondientes a los distintos períodos disponibles.

- **Atención y Asistencia Humanitaria Integral:** generado a partir de la consolidación de los reportes históricos publicados para este registro administrativo.

Estos tres conjuntos de datos constituyeron las fuentes principales de información utilizadas en la investigación y representaron el punto de partida para las actividades posteriores de exploración, evaluación de calidad, integración y preparación de datos dentro de la metodología CRISP-DM.

Una vez consolidadas las fuentes de información, se realizó una verificación inicial orientada a garantizar la correcta lectura y carga de los conjuntos de datos en el entorno de análisis. Esta actividad permitió confirmar la disponibilidad de los registros recopilados y asegurar la integridad básica de los archivos obtenidos durante la etapa de recolección.

Como resultado de esta revisión, se determinaron las dimensiones correspondientes a cada uno de los registros administrativos consolidados, las cuales se presentan a continuación:

**Figura 4**
*Registros administrativos consolidados*

| | Fuente de Datos | Registros | Variables |
|---|---|---|---|
| 0 | RA020 - Violencia de Género | 267 | 20 |
| 1 | RA039 - Salud Integral | 5628 | 22 |
| 2 | RA040 - Asistencia Humanitaria | 1305 | 19 |

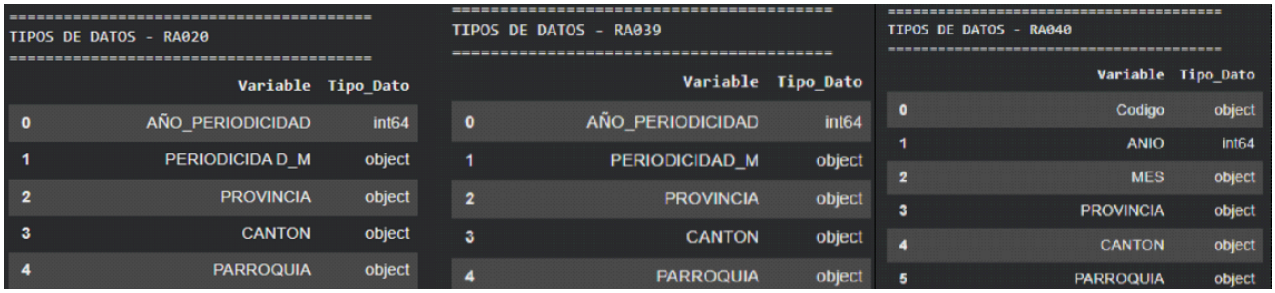
### 3.3.2. Exploración preliminar

Con el objetivo de obtener una primera comprensión de las fuentes de información, se realizó una revisión de la estructura de cada conjunto de datos consolidado. Para ello, se identificaron las variables disponibles y sus respectivos tipos de datos, permitiendo

conocer la naturaleza de la información contenida en cada registro administrativo y verificar la consistencia inicial de su estructura.

**Figura 5**

*Estructura de los conjuntos de datos consolidados*



TIPOS DE DATOS - RA020

| | Variable | Tipo_Dato |
|---|---|---|
| 0 | AÑO_PERIODICIDAD | int64 |
| 1 | PERIODICIDA D_M | object |
| 2 | PROVINCIA | object |
| 3 | CANTON | object |
| 4 | PARROQUIA | object |

TIPOS DE DATOS - RA039

| | Variable | Tipo_Dato |
|---|---|---|
| 0 | AÑO_PERIODICIDAD | int64 |
| 1 | PERIODICIDAD_M | object |
| 2 | PROVINCIA | object |
| 3 | CANTON | object |
| 4 | PARROQUIA | object |

TIPOS DE DATOS - RA040

| | Variable | Tipo_Dato |
|---|---|---|
| 0 | Codigo | object |
| 1 | ANIO | int64 |
| 2 | MES | object |
| 3 | PROVINCIA | object |
| 4 | CANTON | object |
| 5 | PARROQUIA | object |

Analizando todas las variables presentes en las fuentes, se identificaron variables relacionadas con la localización geográfica de las intervenciones, características sociodemográficas de los beneficiarios y tipos de atención social brindada. Entre las variables más relevantes se encontraron:

- Año y mes de la intervención.
- Cantón y parroquia.
- Género.
- Edad.
- Etnia.
- Tipo de discapacidad.
- Categoría y subcategoría de atención social.

### 3.3.3. Verificación de duplicados

Con el objetivo de garantizar la calidad de la información, se efectuó una búsqueda de registros duplicados dentro de cada conjunto de datos. Como resultado se identificaron registros repetidos en la base correspondiente a las acciones de salud integral.

**Figura 6**
*Registros duplicados*

```
print(f"Registros duplicados para RA_020: {dataRA020.duplicated().sum()}")
print(f"Registros duplicados para RA_039: {dataRA039.duplicated().sum()}")
print(f"Registros duplicados para RA_040: {dataRA040.duplicated().sum()}")
```

```
Registros duplicados para RA_020: 0
Registros duplicados para RA_039: 1
Registros duplicados para RA_040: 0
```

Los registros duplicados fueron eliminados antes de la fase de integración, evitando así la sobreestimación de la demanda social y garantizando la consistencia de los análisis posteriores.

### 3.3.4. Construcción del diccionario de datos

Como parte de la fase de comprensión de los datos, se realizó la construcción de un diccionario de datos consolidado que permitiera documentar y estandarizar la información proveniente de los diferentes registros administrativos analizados. Esta actividad tuvo como propósito generar una referencia única para la comprensión de las variables disponibles y facilitar los procesos posteriores de integración y preparación de datos.

Para su elaboración, se revisaron los diccionarios de datos originales publicados junto a cada uno de los registros administrativos del Sistema de Información Local (SIL), identificando las variables comunes, las diferencias de nomenclatura y las particularidades propias de cada fuente. A partir de este análisis, se definió una estructura unificada que permitió homologar conceptos, estandarizar nombres de variables y documentar sus respectivas descripciones, tipos de datos y posibles valores.

***Tabla 2***
*Diccionario consolidado de los datos*

| Nombre del campo | Descripción del campo |
|---|---|
| **Id RA** | Identificación del Registro Administrativo |
| **CATEGORÍA DE ATENCIÓN SOCIAL** | Nombre del Registro Administrativo que representa el tipo de categoría de atención social |
| **AÑO** | Describe el año del periodo en que se ejecuta la intervención |
| **MES** | Describe el mes del periodo en que se ejecuta la intervención |
| **CANTON** | Describe la localidad a nivel cantonal del proyecto |
| **PARROQUIA** | Describe la localidad a nivel parroquial del proyecto |
| **GENERO** | Femenino (género femenino), Masculino (género masculino), N/A (no aplica) |
| **RANGO DE EDAD** | Describe el rango de edad al que pertenece el beneficiario: "A. Menor de Edad (0 a 17 años)", "B. Joven Adulto (18 a 29 años)", "C. Adulto Joven (30 a 44 años)", "D. Adulto Medio (45 a 64 años)", "E. Adulto Mayor (65 años o más)", "Z. Desconocido" (edad no registrada, vacía o con un valor diferente) |
| **ETNIA** | Indígena (se autoidentifica indígena), Afroecuatoriano/afrodescendiente (se autoidentifica afroecuatoriano o afrodescendiente), Montubio/a (se autoidentifica montubio o montubia), Mestizo/a (se autoidentifica mestizo o mestiza), Blanco/a (se autoidentifica blanco o blanca), Otros (tiene otro tipo de autoidentificación), No específicado |
| **SUBCATEGORÍA DE ATENCIÓN O AYUDA SOCIAL** | a. TIPO_ATENCION RA_020 (Campo valido solo para registros del RA_020): Ambulatorio (paciente visita el establecimiento), Acogida (hospitalidad que ofrece un lugar), Otra Atención (el tipo de atención no se refiere a Acogida ni a Ambulatoria)<br>b. SERVICIO_ATENCION RA_039 (Campo valido solo para registros del RA_039): Describe el tipo de servicio que se brinda al beneficiario: Odontología - Nutrición - Medicina General - Optometría - Terapia de Lenguaje - Terapia Ocupacional - Rehabilitación Física - Rehabilitación Psicológica - Otro Servicio de Atención<br>c. TIPO_AYUDA RA_040 (Campo valido solo para registros del RA_040): Tipo de ayuda que se entrega al beneficiario: Ayuda Humanitaria - Ayuda Técnica - Ambas – Otro Tipo de Ayuda |

El diccionario consolidado se convirtió en un insumo fundamental para el presente proyecto, ya que permitió establecer criterios uniformes para la integración de las fuentes de datos de violencia de género, salud integral y preventiva, y atención y asistencia humanitaria integral. Asimismo, sirvió como guía para las actividades posteriores de limpieza, transformación y construcción del conjunto de datos analítico final.

### 3.4. Fase III: Preparación de los Datos

#### 3.4.1. Integración y unificación de datos

Se procedió a integrar las tres fuentes de datos en un único conjunto de datos consolidado. Para ello se unieron las estructuras de las bases de datos RA_020, RA_039 y RA_040, generando un conjunto de datos compuesto por las variables territoriales, temporales y sociodemográficas. También se incorporó una variable identificadora de la fuente de origen con el propósito de preservar la trazabilidad de los registros una vez realizada la unificación.

**Figura 7**
*Registros iniciales de dataset integrado*

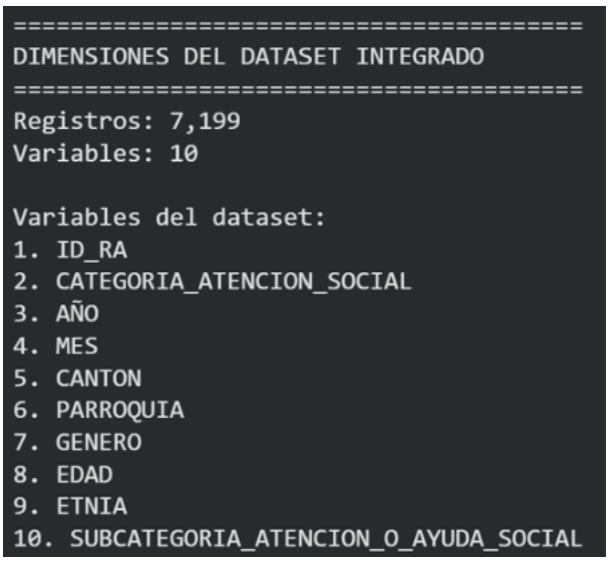
(7199, 10)

| | ID_RA | CATEGORIA_ATENCION_SOCIAL | AÑO | MES | CANTON | PARROQUIA | GENERO | EDAD | ETNIA | SUBCATEGORIA_ATENCION_O_AYUDA_SOCIAL |
|---|---|---|---|---|---|---|---|---|---|---|
| 0 | RA_020 | Violencia de Genero | 2024 | Enero | Manta | Manta | Femenino | NaN | Mestizo/a | Acogida |
| 1 | RA_020 | Violencia de Genero | 2024 | Enero | Manta | Manta | Femenino | NaN | Mestizo/a | Acogida |
| 2 | RA_020 | Violencia de Genero | 2024 | Enero | Bolívar | Calceta | Femenino | 10.0 | Mestizo/a | Acogida |
| 3 | RA_020 | Violencia de Genero | 2024 | Enero | Pedernales | Pedernales | Femenino | 9.0 | Mestizo/a | Acogida |
| 4 | RA_020 | Violencia de Genero | 2024 | Enero | Jipijapa | Jipijapa | Femenino | 4.0 | Mestizo/a | Acogida |

#### 3.4.2. Revisión de estructura y dimensiones

Una vez integradas las tres fuentes en un conjunto de datos, se realizó una nueva revisión de sus dimensiones y estructura. Esto permitió verificar que todas las variables requeridas para el análisis territorial y temporal estuvieran presentes y correctamente ingresadas.

**Figura 8**
*Dimensiones del dataset integrado*




Adicionalmente, se validó la coherencia entre las variables provenientes de las diferentes fuentes y se comprobó la correcta incorporación de los registros en el conjunto de datos integrado. Los registros encontrados por fuente fueron 5627 registros para RA_039, 1305 registros para RA_040 y 267 registros para RA_020.

### 3.4.3. Análisis de calidad de datos

La evaluación de calidad permitió identificar problemas comunes en datos administrativos, tales como valores faltantes, registros incompletos, inconsistencias categóricas y errores de codificación de caracteres. Se realizó una revisión de los valores faltantes presentes en el dataset integrado. Como parte de este proceso se identificaron tanto valores nulos reales como valores que representaban ausencias de información mediante cadenas de texto tales como “NaN”, “None”, “NULL” o espacios vacíos.

**Figura 9**
*Valores faltantes del dataset*

| | Variable | Valores_Faltantes | Porcentaje_Faltantes (%) |
|---|---|---|---|
| 8 | ETNIA | 309 | 4.29 |
| 9 | SUBCATEGORIA_ATENCION_O_AYUDA_SOCIAL | 94 | 1.31 |
| 7 | EDAD | 64 | 0.89 |
| 0 | ID_RA | 0 | 0.00 |
| 1 | CATEGORIA_ATENCION_SOCIAL | 0 | 0.00 |
| 2 | AÑO | 0 | 0.00 |
| 5 | PARROQUIA | 0 | 0.00 |
| 4 | CANTON | 0 | 0.00 |
| 3 | MES | 0 | 0.00 |
| 6 | GENERO | 0 | 0.00 |

En variables como etnia y subcategoría de atención social se aplicaron procedimientos de imputación basados en reglas de negocio y categorías predominantes, garantizando la conservación de la mayor cantidad posible de información. También efectuó un análisis de las variables categóricas con el fin de identificar la cantidad de categorías existentes, su frecuencia de aparición y posibles inconsistencias en su representación.

Durante esta etapa se detectaron diferencias de escritura, variaciones en el uso de mayúsculas y problemas derivados de la codificación de caracteres especiales, especialmente en nombres de cantones y parroquias. Esto demostró la necesidad de aplicar procedimientos de limpieza orientados a mejorar la consistencia y confiabilidad de la información antes de su utilización en procesos estadísticos.

### 3.4.4. Limpieza y transformación

La fase de limpieza incluyó diversas actividades orientadas a mejorar la calidad y consistencia de los datos:

- Tratamiento de valores faltantes.
- Eliminación de registros duplicados.

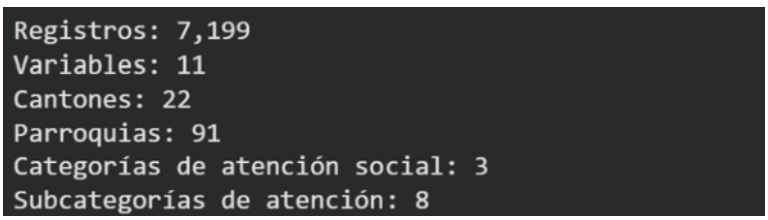
- Corrección de problemas de codificación de caracteres.
- Estandarización de variables categóricas.
- Homologación de categorías de atención social.
- Normalización de formatos de texto.
- Eliminación de acentos y caracteres especiales.

Adicionalmente, se construyó la variable RANGO_EDAD a partir de la edad registrada, agrupando los registros en categorías etarias homogéneas que facilitan el análisis demográfico. Estas transformaciones permitieron disponer de un conjunto de datos consistente, estructurado y adecuado para el desarrollo de análisis exploratorios y modelos predictivos.

### 3.4.5. Análisis Exploratorio de Datos consolidado

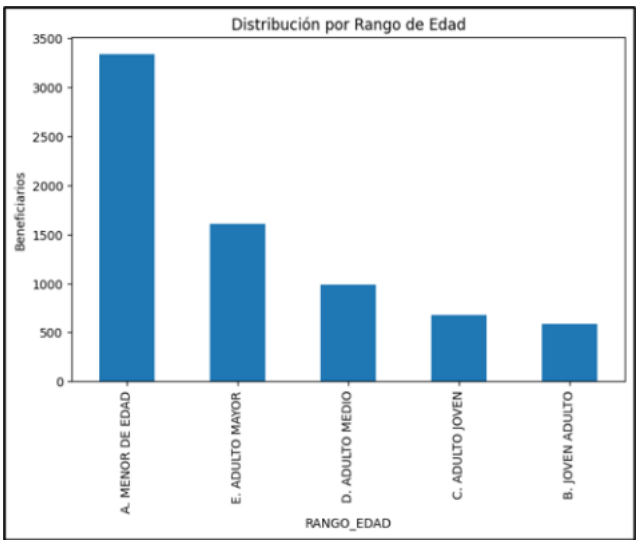
Una vez finalizada la preparación de los datos, se realizó un análisis exploratorio consolidado con el objetivo de identificar patrones temporales, territoriales y sociodemográficos asociados a la demanda de atención social. Se calcularon métricas generales del dataset, incluyendo número total de registros, cantidad de cantones y parroquias representadas, categorías de atención social existentes y diversidad de grupos poblacionales atendidos.

**Figura 10**
*Resumen general del dataset*



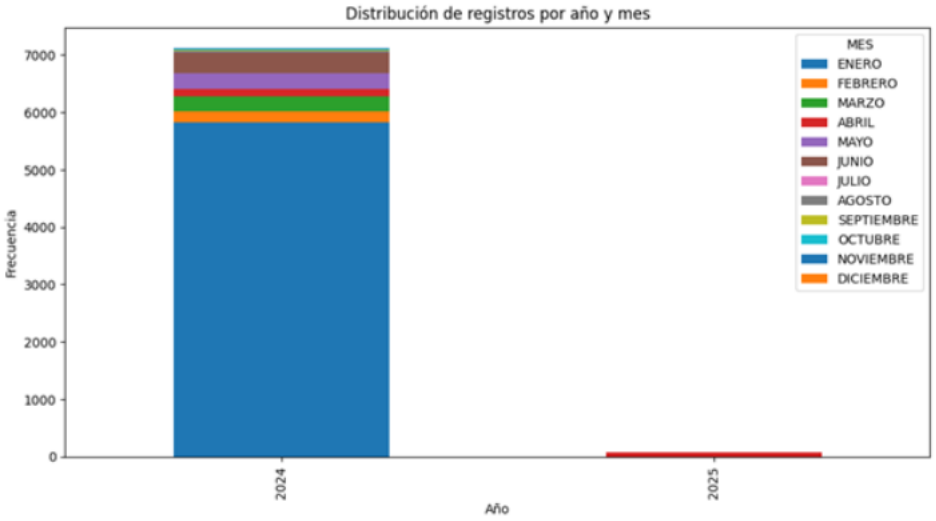
También se analizaron las distribuciones de variables clave como género, etnia, rango de edad, categoría y subcategoría de atención social. Este análisis permitió caracterizar la composición de la población beneficiaria y detectar grupos con mayor presencia dentro de los programas sociales.  Por ejemplo, entre las distribuciones encontradas, la distribución por rango de edad muestra que la mayor cantidad de beneficiarios son menores de edad, como se ve a continuación:

**Figura 11**
*Distribución de registros por rango de edad*




Se evaluaron las evoluciones de las intervenciones sociales por año y mes, identificando variaciones temporales en la demanda de atención. Este análisis es uno de los pasos previos para la construcción de modelos predictivos orientados a anticipar necesidades futuras de intervención.  Por ejemplo, la gráfica de distribución de registros por año y mes muestra que la mayor frecuencia de registros fue en enero del 2024.

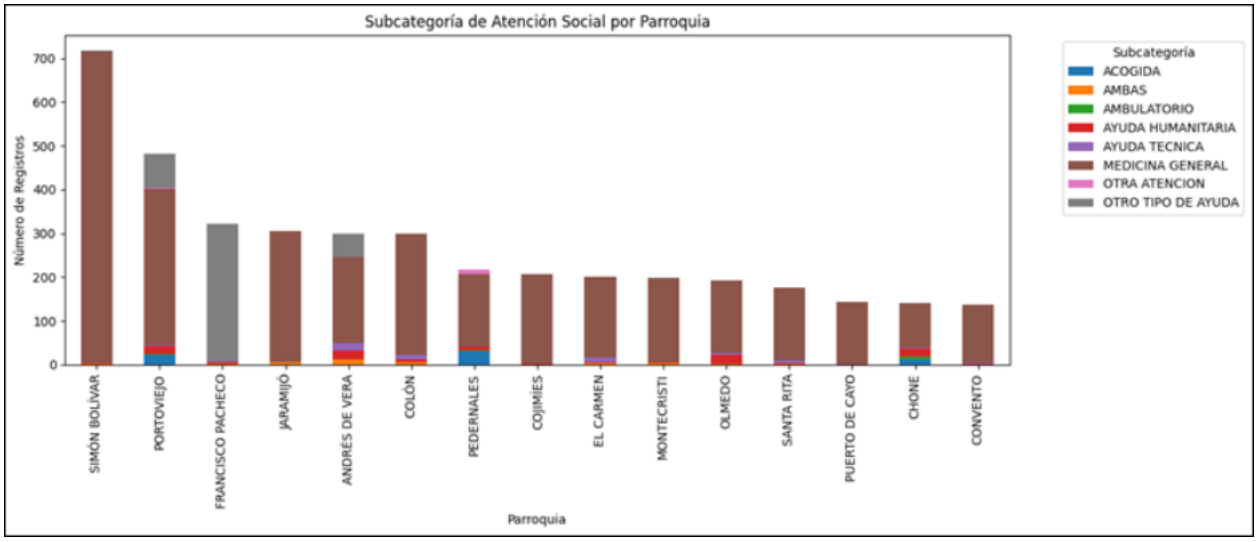
**Figura 12**
*Distribución de registros por año y mes*




Se examinó la distribución de las atenciones sociales a nivel de cantón y parroquia. Asimismo, se analizaron las relaciones entre territorio, categoría y subcategoría de atención, permitiendo identificar zonas con mayores niveles de demanda y posibles concentraciones territoriales de vulnerabilidad. Por ejemplo, a continuación, la gráfica de subcategoría de atención social por parroquia muestra que el mayor número de registros son de la subcategoría de medicina general en la parroquia Simón Bolívar.

**Figura 13**

*Número de registros por cada subcategoria de Atención Social por Parroquia*



En base a dicho análisis, se construyeron tablas agregadas de demanda social utilizando distintas unidades territoriales (cantón, parroquia y combinaciones de ambas), las cuales servirán como base para el desarrollo del modelo predictivo. A continuación, se presenta una de las tablas generadas que corroboran todo este proceso.

**Figura 14**
*Demanda de atención social por cantón, parroquia, categoría y subcategoría de atención social*

| | AÑO | MES | CANTON | PARROQUIA | CATEGORIA_ATENCION_SOCIAL | SUBCATEGORIA_ATENCION_O_AYUDA_SOCIAL | DEMANDA |
|---|---|---|---|---|---|---|---|
| 169 | 2024 | ENERO | PORTOVIEJO | SIMÓN BOLÍV | colab.research.google.com – To exit full screen, press Esc | MEDICINA GENERAL | 710 |
| 161 | 2024 | ENERO | PORTOVIEJO | PORTOVIE | | MEDICINA GENERAL | 357 |
| 105 | 2024 | ENERO | JARAMIJÓ | JARAMIJÓ | ACCIONES DE SALUD INTEGRAL Y PREVENTIVA | MEDICINA GENERAL | 300 |
| 155 | 2024 | ENERO | PORTOVIEJO | COLÓN | ACCIONES DE SALUD INTEGRAL Y PREVENTIVA | MEDICINA GENERAL | 277 |
| 135 | 2024 | ENERO | PEDERNALES | COJIMÍES | ACCIONES DE SALUD INTEGRAL Y PREVENTIVA | MEDICINA GENERAL | 204 |

*Nota*. Se presentan las cinco combinaciones con mayor cantidad de registros.

### 3.4.6. Aumento y balanceo de datos

El análisis exploratorio de datos evidenció que, si bien el conjunto de datos integrado contenía 7199 registros individuales de atención social, la agregación territorial requerida para el modelado predictivo generó un conjunto de datos significativamente más reducido, compuesto por 455 observaciones. Adicionalmente, se identificó una marcada concentración temporal de registros en determinados períodos, consecuencia de la disponibilidad de información publicada en el Sistema de Información Local (SIL) al momento del desarrollo del proyecto.

Considerando que el objetivo es desarrollar y validar una herramienta predictiva para la identificación temprana de la demanda territorial de atención social, se determinó que el volumen y distribución de los datos disponibles podrían limitar la capacidad de

generalización de los modelos de aprendizaje automático. Por esta razón, se implementó una estrategia de aumento y balanceo de datos en dos etapas complementarias.

### 3.4.7. Generación de datos sintéticos representativos

En esta etapa se generó un conjunto de datos sintéticos representativo mediante técnicas de simulación controlada. Esta generación se fundamentó en las distribuciones observadas en los datos reales, preservando las relaciones territoriales entre cantones y parroquias definidas en el diccionario de datos institucional, así como las proporciones observadas para categorías de atención social, subcategorías de atención, género, grupos etarios y autoidentificación étnica.

Como insumo para la generación sintética se utilizaron el conjunto de datos consolidado obtenido durante las fases previas y el diccionario institucional de cantones y parroquias. Luego se construyó un catálogo de combinaciones territoriales válidas entre cantones y parroquias a partir de la información institucional disponible, garantizando que los registros sintéticos respetaran la organización territorial de la provincia.

Después las variables categóricas fueron generadas respetando las distribuciones observadas en los registros originales, preservando la representatividad de categorías de atención social, género y autoidentificación étnica.  Y finalmente, las edades sintéticas fueron obtenidas mediante muestreo de la distribución observada en los registros reales, permitiendo conservar las características demográficas presentes en la población atendida.

**Figura 15**
*Proceso para la generación de datos sintéticos representativos*


El propósito de este proceso fue ampliar la diversidad de combinaciones territoriales y sociales presentes en el conjunto de entrenamiento, manteniendo coherencia con las características observadas en los registros originales.

### 3.4.8. Balanceo de datos mediante CTGAN

En esta segunda etapa se aplicó la técnica de generación sintética mediante CTGAN (Conditional Tabular Generative Adversarial Network), una red generativa adversarial diseñada para datos tabulares mixtos. El entrenamiento del modelo CTGAN se inició utilizando un conjunto combinado conformado por los registros originales y los registros sintéticos representativos generados en la etapa anterior, permitiendo ampliar la diversidad de patrones presentes en los datos.

Para controlar la distribución temporal de los registros generados, se creó una variable compuesta año-mes que posteriormente fue utilizada como condición durante el proceso de muestreo sintético. Luego se identificaron las variables discretas presentes en el conjunto de datos para que el modelo preservara adecuadamente las distribuciones categóricas durante la generación sintética.

Gracia a esto el modelo CTGAN fue entrenado durante 300 épocas utilizando las variables categóricas previamente definidas, con el propósito de aprender las distribuciones y

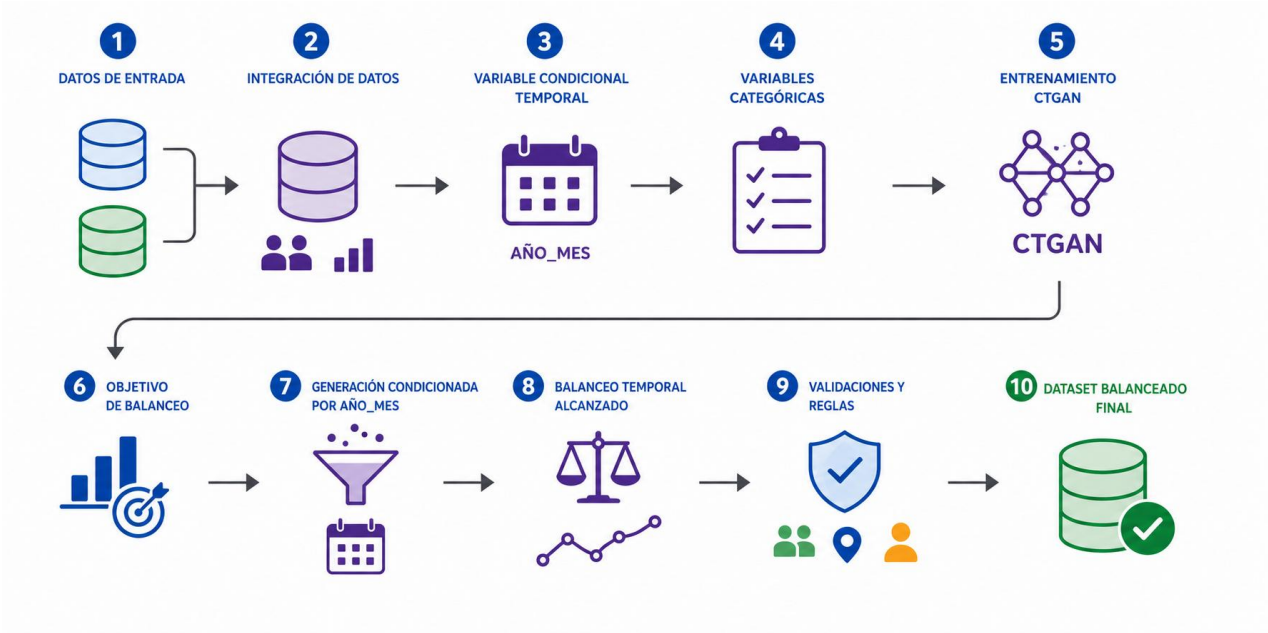
relaciones existentes entre las variables del conjunto de entrenamiento. Se estableció como objetivo alcanzar una cantidad mínima uniforme de registros para cada combinación temporal año-mes, permitiendo reducir los desequilibrios existentes entre períodos.

Por consiguiente, la generación sintética se realizó mediante muestreo condicionado, forzando al modelo a producir observaciones asociadas a períodos específicos de tiempo. Esta estrategia permitió incrementar únicamente los períodos con menor representación dentro del conjunto de datos.

Finalmente, se aplicaron reglas de validación para garantizar la coherencia de los registros generados, verificando restricciones de edad, relaciones entre categorías de atención y consistencia de las variables territoriales. Por lo que esta estrategia permitió obtener un conjunto de datos ampliado, balanceado y estructuralmente consistente, adecuado para las etapas posteriores de entrenamiento, evaluación y validación de modelos predictivos.

**Figura 16**

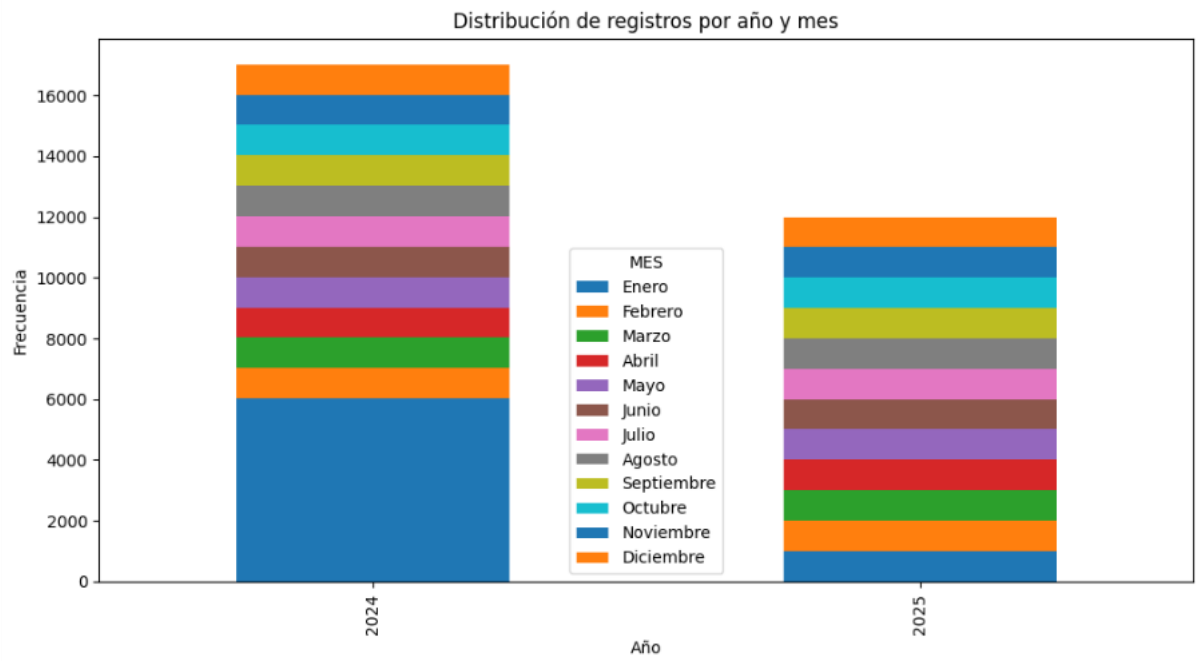
*Proceso para balanceo de datos mediante CTGAN*



### 3.4.9. Validación exploratoria del conjunto de datos ampliado

Una vez finalizado el proceso de generación y balanceo de datos sintéticos, se realizó un análisis exploratorio del conjunto de datos ampliado con el propósito de verificar la coherencia y consistencia de la información obtenida. Esta etapa permitió evaluar si las características generales del conjunto de datos se mantenían alineadas con las distribuciones observadas en los registros originales. Por ejemplo, a continuación, se representa la distribución obtenida por año y mes, demostrando que el número de observaciones disponibles mejoró la representatividad temporal de los registros.

**Figura 17**
*Distribución de registros por año y mes*




### 3.4.10. Consideraciones metodológicas sobre el uso de datos sintéticos

La incorporación de datos sintéticos permitió incrementar el número de observaciones disponibles y reducir el desequilibrio temporal del conjunto de datos, su utilización implica

determinadas consideraciones metodológicas que deben ser reconocidas al interpretar los resultados del proyecto.

Si bien los datos sintéticos permitieron incrementar el volumen de observaciones y preservar las distribuciones estadísticas de las variables mediante CTGAN, no incorporan nuevos patrones de comportamiento ni reemplazan información real. En consecuencia, los modelos predictivos desarrollados deben interpretarse como una validación del funcionamiento de la herramienta propuesta y no como una representación del comportamiento futuro de la demanda social, cuya capacidad de generalización dependerá de la disponibilidad de un mayor volumen y variabilidad de registros administrativos reales.

Por tanto, la capacidad de generalización de los modelos desarrollados se encuentra limitada por la disponibilidad temporal de la información utilizada en el desarrollo de este proyecto, constituyendo una de las principales limitaciones del estudio y una oportunidad para futuros trabajos que incorporen nuevos períodos de información, permitiendo fortalecer la validez externa de los resultados obtenidos.

**3.4.11. Selección y preparación de variables para modelado.**

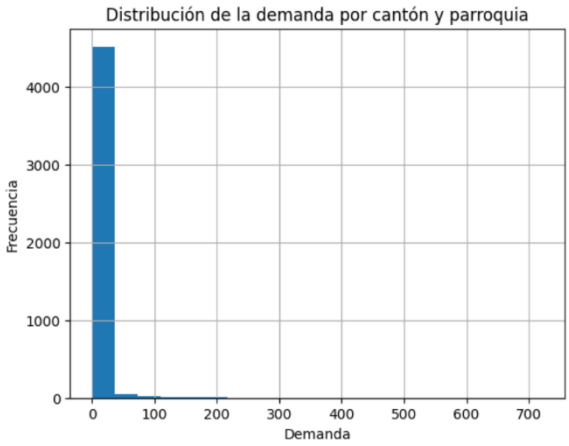
La construcción de la variable de demanda requirió transformar los registros individuales de atención social en unidades de análisis territoriales. Para ello, se realizaron procesos de agregación que permitieron contabilizar las atenciones registradas en diferentes niveles de detalle, generando conjuntos de datos adecuados para las etapas posteriores de modelado predictivo.

En una primera instancia, los registros de demanda fueron agrupados por año, mes, cantón, parroquia y categoría de atención social. A partir de esta agrupación se calculó el número

de atenciones registradas para cada combinación territorial y temporal, obteniéndose la variable DEMANDA como una medida cuantitativa de la atención social observada en cada territorio.

Posteriormente, se analizó la distribución de la variable DEMANDA mediante un histograma, permitiendo identificar la frecuencia de los distintos niveles de atención registrados y obtener una visión general de su comportamiento dentro del conjunto de datos.

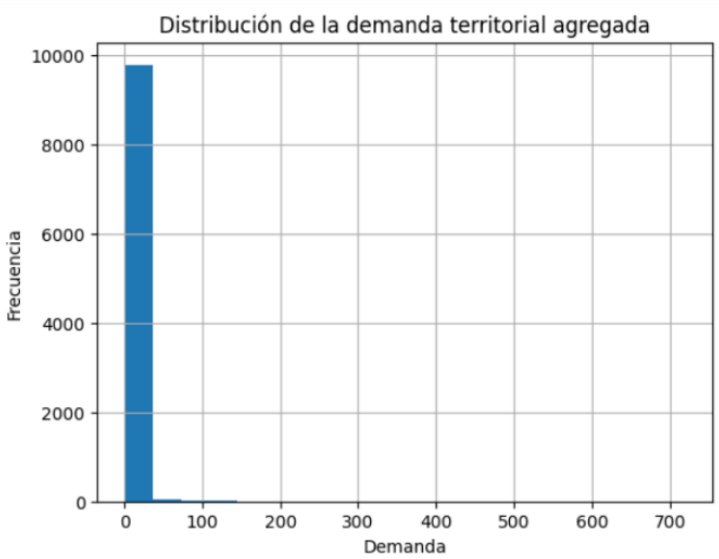
**Figura 18**
*Distribución de demanda por cantón y parroquia*




Para incorporar un mayor nivel de detalle al análisis, se realizó una segunda agregación considerando además la subcategoría de atención o ayuda social. De esta manera, la demanda fue calculada para cada combinación de año, mes, cantón, parroquia, categoría y subcategoría de atención social. Esta nueva estructura permitió representar de forma más específica los diferentes tipos de servicios y ayudas brindadas en cada territorio, generando una base analítica más granular para la construcción de modelos predictivos. Finalmente, se examinó la distribución de la variable DEMANDA resultante con el propósito de verificar la consistencia y comportamiento de los datos agregados.

**Figura 19**

*Distribución de la demanda territorial agregada*




Para conservar el mayor nivel posible de detalle territorial y temático, se seleccionó el dataset de demanda agregada como conjunto de datos base para el modelado predictivo. Este conjunto incorpora información temporal (año y mes), territorial (cantón y parroquia) y características de la atención social (categoría y subcategoría), permitiendo representar de manera más completa la demanda observada en los registros administrativos.

**Figura 20**

*Datasets de los registros actualizados*

| | Dataset | Registros |
|---|---|---|
| 0 | Demanda por Cantón-Parroquia | 4606 |
| 1 | Demanda Agregada Completa | 9842 |

El dataset seleccionado quedó conformado por 9842 observaciones y 7 variables, incluyendo la variable objetivo DEMANDA, correspondiente al número de atenciones registradas para cada combinación territorial y de atención social.

### 3.5. Fase IV: Modelado

#### 3.5.1. Preparación del conjunto analítico

La etapa del modelo fue desarrollada mediante el análisis de los datos consolidados obtenidos a partir de la fase de preparación. La variable objetivo fue nombrada DEMANDA y se refirió al número de atenciones inscritas por cualquier configuración de territorial, cronológica y operativa. Entre las variables de agrupación se incluyeron el distrito, la aldea, la especialidad de asistencia social y el subgrupo de atención o apoyo social.

La serie temporal fue creada utilizando la secuencia numérica de los meses convertida en datos numéricos y se definió la variable 'PERIODO' mediante la concatenación del año con el mes. Luego, los registros se organizaron en orden territorial, operativo y temporal con el fin de mantener la continuidad que es fundamental para la elaboración de las características predictivas.

#### 3.5.2. Variable objetivo y unidad de análisis

La variable objetivo fue la demanda territorial mensual de atención social, expresada en el número de atenciones registradas en cualquier periodo. El entorno de estudio se basó en cada conjunción formada por cantón, parroquia, categoría de atención social, subcategoría de apoyo y mes observado.

Por tanto, cada línea de la colección de datos denotó el volumen de atenciones contabilizado en una región y en un tipo de servicio particular durante una periocidad mensual específica. Esto permitió evaluar simultáneamente la magnitud temporal, territorial e inmediata de la demanda social.

#### 3.5.3. Ingeniería de características temporales

Se desarrolló un proceso de caracterización temporal para incorporar el comportamiento del historial de la demanda. Por cada combinación territorial y operativa se generaron tres variables de retraso, correspondientes a la demanda observada uno, dos y tres períodos atrás.

Además, se calcularon las medias móviles y las desviaciones estándar del período inmediatamente anterior y del período dos períodos antes para representar la demanda reciente y la variabilidad. La tendencia fue estimada mediante la diferencia entre la demanda del período que precede y la registrada en el periodo dos. El objetivo era capturar la tendencia en el comportamiento histórico de la demanda.

La estacionalidad mensual se mostró utilizando ciclos trigonométricos de seno y coseno. Este formato preservó la relación circunferencial entre los meses del año y desacreditó la percepción de que diciembre y enero eran meses numéricamente próximos.

**Tabla 3**
*Variables utilizadas en el proceso de modelado*

| Variable | Tipo | Descripción |
|---|---|---|
| DEMANDA | Objetivo | Número de atenciones registradas en cada combinación territorial, operativa y mensual |
| Lag_1 | Numérica | Demanda registrada en el periodo inmediatamente anterior |
| Lag_2 | Numérica | Demanda registrada dos periodos antes |
| Lag_3 | Numérica | Demanda registrada tres periodos antes |
| Rolling_Mean_3 | Numérica | Promedio de la demanda de los tres periodos anteriores |
| Rolling_STD_3 | Numérica | Desviación estándar de la demanda de los tres periodos anteriores |
| Trend | Numérica | Diferencia entre Lag_1 y Lag_2 |
| Month_Sin | Numérica | Componente seno de la estacionalidad |

| | | |
|---|---|---|
| | | mensual |
| Month_Cos | Numérica | Componente coseno de la estacionalidad mensual |
| Cantón | Categórica | Unidad territorial cantonal |
| Parroquia | Categórica | Unidad territorial parroquial |
| Categoría de atención social | Categórica | Área general del servicio prestado |
| Subcategoría de atención o ayuda | Categórica | Tipo específico de atención brindada |
| Año | Numérica | Año correspondiente al periodo analizado |
| Mes | Numérica | Mes correspondiente al periodo analizado |

### 3.5.4. División temporal de los datos

El conjunto de datos fue dividido según un criterio cronológico. El 80 por ciento inicial de los datos se empleó en el entrenamiento, mientras que el 20 por ciento restante lo constituyó el conjunto de pruebas.

Esta metodología fue aplicada porque el modelo era capaz de predecir valores futuros basándose en datos históricos. Si se hubiera utilizado una partición aleatoria, habría mezclado registros posteriores con registros anteriores y podríamos producir una medición muy favorable. La división temporal replicó la situación real de aplicación más de cerca, en la cual el modelo se ha entrenado usando los datos disponibles en un período específico, y se valora en periodos posteriores.

### 3.5.5. Tratamiento de variables y valores faltantes

Se transformaron las variables territoriales y operativas en categorías para poder procesarlas con LightGBM y CatBoost. Se eliminaron los valores ausentes en las variables de retraso solo si los registros no contaban con una historia mínima que permitiera calcular

las características temporales.

Esto ocurre, en gran medida, en los primeros datos de cada combinación territorial y operativa. Por tanto, para registrar los efectos de dicha eliminación de valores ausentes, se comparamonan las dimensiones del conjunto de entrenamiento y el conjunto de pruebas antes y después de esa eliminación.

**Tabla 4**
*Registros disponibles antes y después del tratamiento de valores faltantes*

| Conjunto | Registros iniciales | Registros posteriores | Registros eliminados | Porcentaje eliminado |
|---|---|---|---|---|
| Entrenamiento | 4703 | 4703 | 0 | 0,0 % |
| Prueba | 109 | 109 | 0 | 0,0 % |

No se eliminaron registros durante el tratamiento de valores faltantes, debido a que las observaciones de los conjuntos de entrenamiento y prueba conservaron información completa en las variables requeridas para el modelado.

### 3.5.6. Modelo base de referencia

El modelo DummyRegressor se utilizó como referencia básica, donde se predijo automáticamente la media de la demanda observada durante el conjunto de entrenamiento. Este modelo no identifica ni detecta ningún patrón temporal, territorial ni operativo, sino que solo prevé la media histórica.

Esta comparación nos ayudó a ver si los métodos de aprendizaje automatizado ofrecen una mejora significativa sobre una predicción fundamental basada en el promedio del pasado.

### 3.5.7. Modelos predictivos evaluados

Se compararon dos algoritmos de potenciación por gradiente: LightGBM y CatBoost. Se eligieron estos porque son adecuados para capturar relaciones no lineales y analizar conjuntos de datos tabulados que incluyen valores numéricos y categorizados.

LightGBM forma un árbol de decisión en etapas consecutivas; cada nuevo árbol se adapta para disminuir la cantidad de error que generan sus antecesores. La configuración inicial de LightGBM estableció una tasa de aprendizaje de 0,05, 500 estimadores, 31 hojas por árbol y porcentajes de submuestras de fila y variable del 80%.

CatBoost fue adoptado como modelo de sustitución por sus habilidades para trabajar con variables categóricas con una forma interna de manipulación única. La configuración inicial de CatBoost abarcó 500 iteraciones, una tasa de aprendizaje de 0,05, una profundidad máxima de ocho.

**Tabla 5**
*Configuración inicial de los modelos predictivos*

| Modelo | Parámetros principales | Función dentro del análisis |
|---|---|---|
| DummyRegressor | Estrategia basada en la media | Establecer una línea base de comparación |
| LightGBM | 500 estimadores, tasa de aprendizaje 0,05, 31 hojas, submuestreo 0,80 | Capturar relaciones no lineales y patrones temporales |
| CatBoost | 500 iteraciones, profundidad 8, tasa de aprendizaje 0,05 | Modelar variables categóricas y relaciones complejas |

### 3.5.8. Optimización de hiperparámetros

Tras la evaluación inicial, se utilizaron el algoritmo de OPTUNA para optimizar la configuración de los hiperparámetros seleccionados. Este proceso implicó realizar 50

ensayos, usando RMSE como criterio a minimizar.

Los parámetros que se evaluaron incluyeron la tasa de aprendizaje, número de capas, la máxima profundidad, el número mínimo de observaciones por nodo, los valores de muestreo por fila y variables, los parámetros de regularización y el número de árboles.

La optimización fue ejecutada utilizando el enfoque de validación temporal en el conjunto de entrenamiento. Por otro lado, el conjunto de prueba se mantuvo para la medición final del rendimiento. No se permitió que la elección de los hiperparámetros influyera en los datos utilizados para medir el rendimiento final.

### 3.5.9. Métricas de evaluación

El desempeño predicción fue evaluado mediante el Error Absoluto Medio (MAE), el Raíz del Error Cuadrático Medio (RMSE) y el coeficiente de determinación ($R^2$). El MAE refleja la magnitud media de los errores en las mismas unidades de la variable objetivo. El RMSE asocia mayor influencia a las divergencias superiores; por otro lado, el $R^2$ indica la proporción de la variabilidad de la demanda comprendida por el modelo. Se seleccionó conjuntamente los valores bajos del MAE y del RMSE junto con un valor positivo del $R^2$ y una medida de $R^2$ superior al resultado del modelo básico.

### 3.5.10. Selección del modelo

Los resultados del modelo inicial, LightGBM y CatBoost fueron agrupados en una tabla para la comparación. Se valoró el menor RMSE, acompañado de MAE y $R^2$. El método con mayor rendimiento tras la validación pasó a la optimización de parámetros. La configuración final se probó posteriormente con el conjunto de pruebas temporal.

### 3.5.11. Análisis preliminar de errores

Los valores observados fueron comparados con los estimados por la simulación. La gráfica de valores reales contra predichos es una excelente herramienta para evaluar la precisión de las estimaciones frente a la línea de la coincidencia.

El análisis de residuos nos permite identificar posibles patrones sistemáticos, como sobreestimación, subestimación o la variación de error según áreas, periodos o categorías. Se destacaron también las observaciones que tuvieron errores mayores. Estas últimas ofrecen información valiosa acerca de zonas, períodos o tipos donde el modelo mostró menor precisión.

### 3.5.12. Interpretabilidad mediante SHAP

El proceso de interpretación del modelo empleó SHAP (un método basado en valores de Shapley cuyo objetivo es medir la contribución de cada característica en las predicciones).

Fue generado un análisis global sobre el nivel de importancia y gráficos de relación para las características con mayor influencia. Con este procedimiento, se evaluó si la estimación estaba justificada fundamentalmente por la demanda histórica, la estacionalidad, la geografía del territorio o la naturaleza del servicio.

### 3.5.13. Reproducibilidad

Para facilitar la reproducibilidad se definió una semilla aleatoria de 42 y se registraron las bibliotecas, parámetros y etapas de procesamiento empleadas. El modelo seleccionado se almacenó en formato pkl mediante Joblib, mientras que sus hiperparámetros y métricas se conservaron en archivos JSON. El código fuente fue desarrollado en Google Colab y se incorporó como anexo digital de la investigación en un archivo digital.

## 3.6. Fase V: Evaluación

### 3.6.1. Evaluación del desempeño predictivo

La fase de evaluación tuvo por fin comprobar si los modelos desarrollados superaban el rendimiento de la línea base y respondían al objetivo de anticipar la demanda territorial de atención social.

La evaluación completa se hizo en base al conjunto de datos de prueba temporal, el cual no intervino en el entrenamiento ni en la optimización de hiperparámetros. Las medidas empleadas fueron MAE, RMSE y $R^2$ para analizar el error promedio, la gravedad de las distorsiones y la precisión en la explicación de cada modelo.

### 3.6.2. Comparación de modelos

Los resultados del modelo base, LightGBM y CatBoost fueron incluidos en una tabla de comparación. Esta evaluación nos permitió determinar si el algoritmo escogido supuso una mejora significativa en comparación con la predicción mediante la mediana histórica.

**Tabla 6**

*Comparación del rendimiento entre los modelos predictivos*

| Modelo | MAE | RMSE | $R^2$ | Posición |
|---|---|---|---|---|
| LightGBM optimizado | 2,0636 | 3,4162 | 0,8285 | 1 |
| LightGBM | 2,2622 | 4,5204 | 0,6997 | 2 |
| CatBoost | 2,2433 | 5,1273 | 0,6137 | 3 |
| Modelo base | 4,0125 | 8,6624 | -0,1026 | 4 |

Nota. El modelo LightGBM optimizado mostró el mejor rendimiento al lograr los menores valores de MAE y RMSE, así como el mayor coeficiente de determinación. El modelo base se situó en segundo lugar y presentó el peor rendimiento; además, su valor de $R^2$ fue negativo, sugiriendo que sus pronósticos no eran tan precisos como uno obtenido mediante la mediana de las demandas observadas.

### 3.6.3. Capacidad de generalización

La capacidad de generalización se midió comparando el rendimiento en entrenamiento, validación y prueba. Si existía una diferencia significativa en uno o más de estos conjuntos, esto podría sugerir sobreajuste.

Además, se evaluó la estabilidad de los errores entre periodos, cantones y parroquias para determinar si el rendimiento total podía haber ocultado una pobreza de datos en áreas geográficas con menos observaciones.

### 3.6.4. Análisis de errores

La evaluación de errores se realizó mediante el contraste entre los valores observados y los valores generados por el modelo, analizando la distribución de los residuos y las observaciones con mayores desviaciones absolutas. Este análisis permitió determinar la existencia de sesgos sistemáticos en las predicciones, identificando los contextos territoriales, temporales o categóricos donde el modelo presentó mayores niveles de incertidumbre o menor capacidad predictiva.

### 3.6.5. Interpretación del modelo final

Para interpretar el modelo final se aplicó el enfoque basado en valores SHAP, permitiendo identificar cómo cada variable influyó en las predicciones generadas. Los resultados permitieron analizar el aporte de la información histórica, los indicadores derivados mediante medias móviles, los patrones estacionales y las características territoriales y operativas. Así, la evaluación integró tanto las métricas de desempeño predictivo como la capacidad del modelo para aprender relaciones coherentes dentro del contexto analizado.

### 3.6.6. Cumplimiento de los criterios de éxito

Los resultados obtenidos fueron comparados con los criterios técnicos e institucionales establecidos durante la fase de comprensión del problema. Desde el ámbito técnico, se comprobó que el modelo presentara un desempeño superior al de la línea base, mantuviera niveles de error aceptables en el conjunto de prueba y conservara estabilidad en los diferentes territorios analizados. Desde la perspectiva institucional, se verificó la capacidad del modelo para generar estimaciones desagregadas por cantón, parroquia, periodo y tipo de atención social. El cumplimiento de estos criterios permitió establecer la factibilidad del modelo como una herramienta de apoyo para la planificación territorial, considerando que sus resultados complementan, pero no reemplazan, el análisis y la toma de decisiones de los responsables institucionales.

## 3.7. Fase VI: Despliegue

### 3.7.1. Desarrollo de la aplicación web

La fase de despliegue se concretó mediante el desarrollo de una aplicación web interactiva denominada Sistema Inteligente para la Predicción de la Demanda Social de Manabí (SIPP Manabí). La herramienta fue construida con Streamlit y tuvo como propósito facilitar el uso del modelo mediante una interfaz gráfica. La aplicación integró el conjunto de datos agregado, el modelo LightGBM entrenado y distintos componentes de visualización para presentar la demanda histórica y el pronóstico del siguiente periodo mensual.

### 3.7.2. Arquitectura funcional

La aplicación se estructuró en cuatro componentes: carga y transformación de datos, recuperación del modelo, generación de características predictoras y presentación de resultados. En la primera etapa se carga el archivo demanda_aggregated.csv, se transforma el mes a formato numérico y se construye una variable de fecha. Posteriormente, se

recupera el modelo almacenado en best_lightgbm.pkl.

A continuación, el sistema genera las características requeridas a partir de los tres periodos más recientes de la combinación seleccionada. El flujo concluye con la ejecución del pronóstico y la presentación de indicadores, gráficos temporales, representación territorial y valores SHAP.

### 3.7.3. Parámetros de entrada

La interfaz permite seleccionar el cantón, la parroquia, la categoría de atención social y la subcategoría de atención o ayuda. Las parroquias disponibles se actualizan según el cantón elegido, mientras que las subcategorías se filtran en función de la categoría seleccionada. Este comportamiento reduce la generación de combinaciones territoriales u operativas inexistentes.

**Tabla 7**
*Parámetros de entrada de la aplicación web*

| Parámetro | Descripción |
|---|---|
| Cantón | Unidad territorial cantonal |
| Parroquia | Unidad territorial perteneciente al cantón seleccionado |
| Categoría de atención social | Área general del servicio brindado |
| Subcategoría | Tipo específico de atención o ayuda |
| Historial requerido | Tres periodos recientes utilizados para generar las características temporales |

### 3.7.4. Generación de variables para el pronóstico

Para producir una estimación, la aplicación recupera los tres registros mensuales más recientes correspondientes a la combinación seleccionada. Con esta información se generan Lag_1, Lag_2, Lag_3, la media móvil, la desviación estándar, la tendencia y los

componentes cíclicos del mes. La fecha de predicción corresponde al mes inmediatamente posterior al último periodo disponible. Cuando existen menos de tres observaciones históricas, el sistema informa que los datos son insuficientes y detiene el proceso, evitando generar una estimación sin el soporte temporal mínimo.

### 3.7.5. Generación y presentación del pronóstico

El modelo LightGBM recibe las características calculadas y genera una estimación numérica de la demanda para el siguiente mes. Debido a que la variable objetivo representa un conteo, el resultado se redondea al número entero más próximo y se restringe a valores iguales o superiores a cero. La predicción se compara con la demanda del periodo anterior para identificar incrementos, disminuciones o estabilidad.

### 3.7.6. Tablero de indicadores

El tablero presenta cuatro indicadores principales: demanda pronosticada, demanda observada en el mes anterior, RMSE del modelo y $R^2$. Estos indicadores permiten interpretar el resultado en relación con el comportamiento histórico y el desempeño obtenido durante la evaluación. El $R^2$ debe presentarse como proporción de variabilidad explicada y no como porcentaje de confianza.

### 3.7.7. Visualización temporal y territorial

La aplicación incluye un gráfico que representa la evolución mensual de la demanda y añade el pronóstico del siguiente periodo. También incorpora un mapa con la ubicación aproximada del cantón analizado. Debido a que las coordenadas corresponden a los cantones, la representación geográfica debe interpretarse como una referencia cantonal y no como una localización exacta de la parroquia.

### 3.7.8. Interpretabilidad local mediante SHAP

La aplicación incorpora un análisis local mediante valores SHAP. Este componente identifica las variables que incrementaron o redujeron la predicción generada para la combinación seleccionada. La interfaz muestra las ocho características con mayor impacto absoluto y diferencia visualmente sus contribuciones positivas y negativas. Esto permite explicar qué factores temporales, territoriales u operativos influyeron en el pronóstico.

### 3.7.9. Requisitos tecnológicos

La aplicación fue desarrollada en Python y empleó las bibliotecas necesarias para la gestión de datos, ejecución del modelo, visualización e interpretabilidad. Las dependencias fueron documentadas en el archivo requirements.txt, lo que facilita la reproducción del entorno de ejecución.

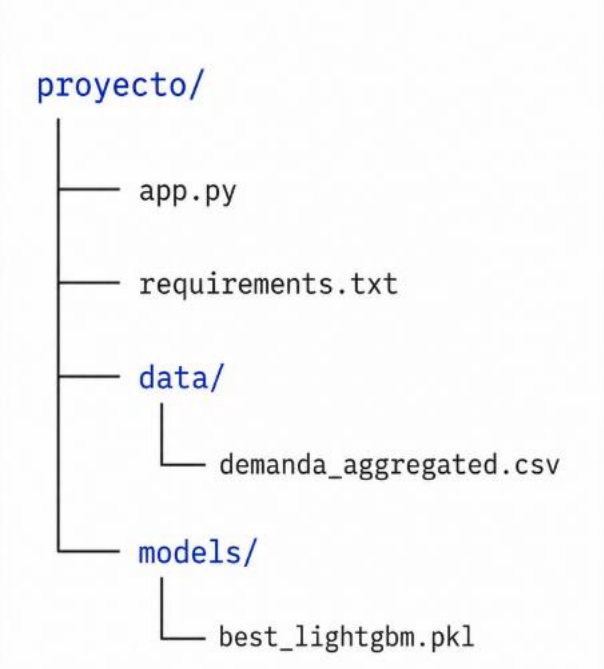
**Tabla 8**
*Bibliotecas utilizadas en la aplicación web*

| Biblioteca | Función |
|---|---|
| Streamlit | Construcción de la interfaz web |
| pandas | Carga, filtrado y transformación de datos |
| NumPy | Cálculos numéricos y variables temporales |
| LightGBM | Ejecución del modelo predictivo |
| Joblib | Carga del modelo serializado |
| Plotly | Elaboración de visualizaciones interactivas |
| SHAP | Interpretación local de las predicciones |
| scikit-learn | Métricas y compatibilidad del modelo |
| Matplotlib | Soporte gráfico complementario |

### 3.7.10. Estructura de archivos del sistema

El sistema se organizó en directorios independientes para separar el código, los datos y el modelo entrenado. Esta estructura facilitó la portabilidad, el mantenimiento y la reproducción del despliegue.

**Figura 21**
*Estructura de directorios y archivos de la aplicación SIPP Manabí*




*Nota*. La figura muestra la organización del proyecto, compuesta por el archivo principal de la aplicación, las dependencias, el conjunto de datos agregado y el modelo LightGBM serializado.

El archivo app.py contiene la lógica de la interfaz y del proceso de inferencia. El directorio data almacena el conjunto agregado, mientras que models contiene el modelo serializado.

### 3.7.11. Validación funcional

La validación funcional consistió en comprobar la carga del conjunto de datos, la recuperación del modelo, la actualización dinámica de filtros, la generación de

características temporales y la ejecución del pronóstico. También se evaluaron los controles implementados para detener el flujo cuando el modelo no está disponible o cuando el historial contiene menos de tres periodos.

**Tabla 9**
*Pruebas funcionales de la aplicación web*

| Prueba | Resultado esperado |
|---|---|
| Carga del conjunto de datos | Lectura correcta del archivo sin errores |
| Carga del modelo | Modelo disponible para generar predicciones |
| Selección de cantón | Actualización automática de las parroquias |
| Selección de categoría | Actualización automática de las subcategorías |
| Historial inferior a tres periodos | Mensaje de datos insuficientes |
| Historial suficiente | Generación del pronóstico mensual |
| Interpretabilidad SHAP | Visualización de la influencia de las variables |
| Visualización temporal | Presentación del historial y del pronóstico |
| Representación territorial | Ubicación aproximada del cantón seleccionado |

### 3.7.12. Alcance del despliegue

La implementación de la aplicación permitió comprobar la integración funcional de los componentes desarrollados durante las etapas de modelado, evaluación y despliegue del sistema. El dashboard SIPP Manabí fue ejecutado en un entorno local mediante Streamlit, permitiendo la carga adecuada del conjunto de datos agregado, el modelo LightGBM optimizado y los recursos necesarios para la generación de predicciones.

La interfaz desarrollada incorpora un panel lateral que permite al usuario seleccionar variables de consulta como cantón, parroquia, categoría de asistencia y subcategoría de atención. A partir de estos parámetros, el sistema recupera la información histórica asociada, construye las características temporales requeridas y ejecuta la predicción correspondiente al siguiente periodo mensual.

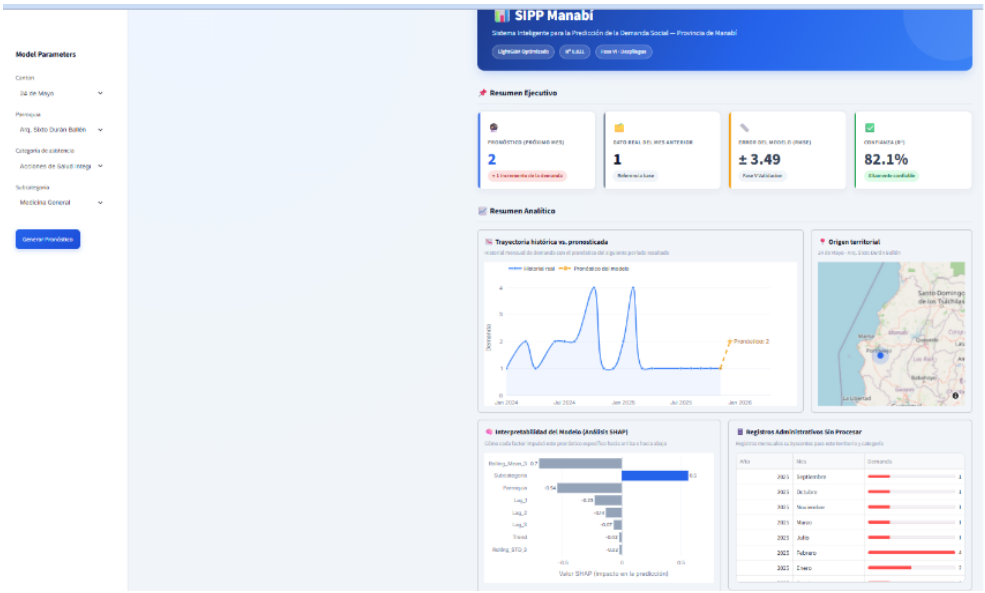
En el panel principal se muestran indicadores asociados a la demanda estimada, el valor registrado en el periodo anterior, así como las métricas de evaluación del modelo, incluyendo el RMSE y el $R^2$. La comparación entre la demanda pronosticada y el último valor observado facilita la identificación de posibles variaciones en el comportamiento esperado, como tendencias de crecimiento, reducción o estabilidad.

El dashboard integra elementos de visualización como la evolución histórica y proyectada de la demanda, la representación geográfica del cantón seleccionado, la interpretación local mediante valores SHAP y una tabla con los registros mensuales utilizados como referencia. La combinación de estos componentes facilita la comprensión de los resultados al permitir analizar de manera conjunta la dinámica temporal, la distribución territorial, los factores que influyeron en la predicción y la información histórica disponible.

La ejecución del sistema evidenció la capacidad de la aplicación para transformar los resultados predictivos del modelo en información visual accesible para usuarios institucionales. Sin embargo, el dashboard corresponde a una prueba de concepto, por lo que sus resultados deben considerarse como insumos de apoyo para la planificación territorial y estar sujetos a validación técnica y contextual antes de su aplicación operativa.

**Figura 22**

*Ejecución y evaluación funcional del dashboard predictivo SIPP Manabí*




*Nota*. La figura muestra la interfaz funcional del dashboard predictivo SIPP Manabí, donde se integran componentes de consulta territorial, indicadores de desempeño, comportamiento histórico y pronosticado de la demanda, visualización geográfica, explicación del modelo mediante valores SHAP y registros históricos de referencia.

# CAPITULO 4: Análisis de esultados

## 4.1. Resultados del modelo predictivo

En este capítulo se describen los resultados obtenidos durante la evaluación de los modelos predictivos desarrollados para estimar la demanda territorial de atención social en la provincia de Manabí. El análisis comparativo consideró un modelo base basado en la media histórica, el modelo LightGBM con su configuración inicial, CatBoost y una versión optimizada de LightGBM mediante ajuste de hiperparámetros. Para medir el desempeño de los modelos se emplearon las métricas Error Absoluto Medio (MAE), Raíz del Error Cuadrático Medio (RMSE) y coeficiente de determinación ($R^2$), calculadas sobre el conjunto de prueba temporal definido para la evaluación final.

### 4.1.1. Desempeño de los modelos

Los resultados mostraron diferencias relevantes entre los modelos evaluados. La versión optimizada de LightGBM presentó el mejor desempeño, alcanzando un MAE de 2,0636, un RMSE de 3,4162 y un coeficiente de determinación ($R^2$) de 0,8285. Estos resultados indican que el modelo presentó un error promedio cercano a dos atenciones y logró explicar aproximadamente el 82,85 % de la variabilidad de la demanda observada en el conjunto de prueba.

Por su parte, el modelo LightGBM con configuración inicial obtuvo un MAE de 2,2622, un RMSE de 4,5204 y un $R^2$ de 0,6997. Si bien este modelo mostró un desempeño adecuado, presentó mayores niveles de error en comparación con la versión optimizada y una menor capacidad para explicar la variabilidad de la demanda.

El modelo CatBoost alcanzó un MAE de 2,2433, un RMSE de 5,1273 y un $R^2$ de 0,6137. Aunque su error absoluto medio fue ligeramente inferior al registrado por LightGBM inicial, el incremento en el RMSE evidenció una mayor sensibilidad ante errores de mayor magnitud dentro de las predicciones realizadas.

Finalmente, el modelo base obtuvo el menor desempeño entre las alternativas analizadas, con un MAE de 4,0125, un RMSE de 8,6624 y un $R^2$ de −0,1026. Este comportamiento evidencia que una estrategia de predicción constante basada únicamente en la media histórica no logró representar adecuadamente la dinámica temporal, territorial y operativa asociada a la demanda de atención social.

**Tabla 10**
*Evaluación comparativa del desempeño de los modelos predictivos*

| Modelo | MAE | RMSE | $R^2$ | Posición |
|---|---|---|---|---|
| LightGBM optimizado | 2,0636 | 3,4162 | 0,8285 | 1 |
| LightGBM | 2,2622 | 4,5204 | 0,6997 | 2 |
| CatBoost | 2,2433 | 5,1273 | 0,6137 | 3 |
| Modelo base | 4,0125 | 8,6624 | −0,1026 | 4 |

*Nota*. La posición se determinó principalmente a partir del menor RMSE, complementado con el MAE y el $R^2$.

### 4.1.2. Comparación de resultados

La comparación de resultados evidenció que los tres algoritmos de aprendizaje automático superaron el desempeño del modelo base en las métricas analizadas. Este comportamiento demuestra que la incorporación de variables temporales, territoriales y operativas permitió capturar patrones adicionales de la demanda, proporcionando una capacidad predictiva superior frente a una estimación basada únicamente en el promedio histórico.

El modelo LightGBM optimizado presentó la mayor mejora respecto al modelo base, reduciendo el MAE aproximadamente en un 48,57 % y el RMSE cerca de un 60,56 %. Estos resultados reflejan una disminución significativa tanto en el error promedio de predicción como en la magnitud de las desviaciones más elevadas, evidenciando una mayor precisión del modelo.

En comparación con la configuración inicial de LightGBM, la versión optimizada logró reducir el MAE en aproximadamente 8,78 % y el RMSE en 24,43 %. Además, el coeficiente de determinación ($R^2$) se incrementó de 0,6997 a 0,8285, representando una mejora de 0,1288 puntos porcentuales en la capacidad del modelo para explicar la variabilidad observada en la demanda.

Aunque CatBoost obtuvo un MAE similar al de LightGBM inicial, presentó un RMSE superior, lo que indica que ambos modelos generaron errores promedio comparables, pero CatBoost presentó algunas predicciones con desviaciones más significativas. En consecuencia, LightGBM alcanzó un mejor desempeño global dentro de los modelos evaluados.

Por otro lado, el valor negativo del $R^2$ obtenido por el modelo base evidenció que su capacidad predictiva fue inferior incluso frente a una estrategia de estimación constante. Este resultado confirma que el promedio histórico no fue suficiente para representar la dinámica compleja de la demanda territorial de atención social, donde intervienen factores temporales, espaciales y operativos.

#### .4.1.3. Selección del mejor modelo

El modelo LightGBM optimizado fue seleccionado como modelo final debido a que presentó el menor valor de MAE y RMSE, además del mayor coeficiente de determinación ($R^2$) entre

las alternativas evaluadas. Esta combinación de métricas evidenció una mayor precisión en las predicciones, una menor presencia de errores de mayor magnitud y una mejor capacidad para representar la variabilidad de la demanda analizada.

La selección del modelo no se fundamentó únicamente en el valor del coeficiente de determinación, sino también en la reducción de los errores obtenidos frente al modelo base y las configuraciones iniciales evaluadas. El RMSE de 3,4162 demostró que las desviaciones más significativas fueron inferiores respecto a las registradas por los demás algoritmos, mientras que el MAE de 2,0636 reflejó un error promedio reducido en relación con las unidades de la variable objetivo.

Posteriormente, el modelo seleccionado fue integrado en la aplicación web SIPP Manabí, desde la cual se generan estimaciones mensuales desagregadas por cantón, parroquia, categoría y subcategoría de atención social. Su implementación se plantea como una herramienta de apoyo técnico para fortalecer la planificación institucional, sin sustituir el análisis profesional ni convertirse en un mecanismo automático para la asignación de recursos.

### 4.2. Interpretación de los resultados para la gestión social

Los resultados obtenidos demuestran que el modelo LightGBM optimizado representa una herramienta técnica con potencial para apoyar el fortalecimiento de la gestión social del Gobierno Autónomo Descentralizado Provincial de Manabí. La capacidad del modelo para explicar el 82,85 % de la variabilidad observada en la demanda evidencia que la integración de variables temporales, territoriales y operativas permitió identificar patrones relevantes asociados al comportamiento de las atenciones sociales.

Este desempeño proporciona una base para anticipar posibles variaciones en la demanda futura y generar estimaciones desagregadas por cantón, parroquia, categoría y subcategoría de atención social. En este sentido, el modelo puede contribuir a mejorar los procesos de planificación y asignación estratégica de capacidades institucionales, siempre considerando la validación técnica y el análisis contextual por parte de los responsables de la gestión pública.

#### .4.2.1. Aporte al Sistema de Información Local

El principal aporte al Sistema de Información Local (SIL) radica en ampliar la utilidad de los registros administrativos, trascendiendo su función tradicional como fuente de almacenamiento operativo y descripción de eventos pasados. La incorporación del modelo predictivo permite transformar los datos históricos disponibles en el SIL en estimaciones orientadas a comprender el comportamiento futuro de la demanda social. De esta manera, la información adquiere una dimensión analítica que facilita la identificación anticipada de necesidades y fortalece los procesos de planificación institucional.

La integración del modelo también favorece un mejor aprovechamiento de los datos institucionales al estructurar la información considerando variables territoriales, temporales y relacionadas con el tipo de atención. Esto permite reconocer tendencias, variaciones crecientes o decrecientes y comportamientos atípicos dentro de los servicios registrados. La aplicación SIPP Manabí amplía estas capacidades mediante la presentación de resultados a través de indicadores, gráficos, representaciones geográficas y explicaciones basadas en valores SHAP, facilitando la interpretación de los resultados por parte de los usuarios institucionales.

El sistema también puede contribuir al fortalecimiento de la calidad de los registros. La ejecución periódica del modelo permite detectar combinaciones territoriales con información insuficiente, periodos incompletos o variaciones poco consistentes. Estas alertas pueden orientar la revisión de los procesos de registro, actualización y estandarización de la información almacenada en el SIL.

#### 4.2.2. Utilidad para la planificación territorial

El modelo puede incorporarse como una herramienta de apoyo para la planificación territorial mediante la estimación mensual de la demanda desagregada por cantón, parroquia, categoría y subcategoría de atención social. Esta capacidad permite superar las limitaciones de los análisis basados únicamente en promedios provinciales, los cuales pueden ocultar diferencias significativas en la dinámica y necesidades específicas de cada territorio.

Cuando el sistema proyecta un aumento en la demanda, los responsables institucionales pueden anticipar requerimientos relacionados con personal, insumos, ayudas técnicas, brigadas de atención o capacidad operativa. Por otro lado, en territorios donde se prevé estabilidad o reducción de la demanda, las estimaciones pueden contribuir a revisar la distribución de recursos, analizar variaciones en la cobertura de servicios o identificar posibles inconsistencias en el registro de atenciones.

La herramienta también fortalece los procesos de priorización territorial al facilitar la identificación de zonas con mayores niveles de demanda esperada y permitir la comparación de su comportamiento respecto a periodos anteriores. Esta información puede apoyar la planificación de intervenciones, la asignación estratégica de recursos y la

articulación entre las unidades responsables de áreas como violencia de género, salud integral y asistencia humanitaria.

No obstante, la planificación sustentada en los resultados del modelo debe complementarse con información demográfica, presupuestaria, logística y contextual. Las predicciones generadas representan señales anticipadas sobre posibles escenarios de demanda; sin embargo, las decisiones institucionales deben considerar las particularidades de cada territorio y la disponibilidad efectiva de recursos

### 4.2.3. Mejora de la toma de decisiones

El modelo fortalece los procesos de toma de decisiones al generar estimaciones cuantitativas fundamentadas en información histórica, reduciendo la dependencia exclusiva de apreciaciones subjetivas o análisis retrospectivos. La comparación entre los valores observados y los pronosticados permite identificar posibles variaciones en la demanda antes de que estas tengan un impacto significativo en la gestión operativa institucional.

La incorporación de métricas de evaluación como MAE, RMSE y $R^2$ proporciona información sobre el nivel de precisión alcanzado y las limitaciones asociadas al modelo. Complementariamente, el análisis mediante valores SHAP permite identificar la contribución de cada variable en las predicciones generadas, aportando mayor transparencia y facilitando la interpretación de los resultados antes de considerarlos como apoyo para la toma de decisiones.

La aplicación web desarrollada facilita el acceso a los pronósticos mediante una interfaz que permite a los responsables de planificación consultar estimaciones según el territorio y

el tipo de atención requerido. De esta manera, se reduce la complejidad asociada al modelo predictivo y se transforma la salida técnica del algoritmo en información comprensible para usuarios sin conocimientos especializados en aprendizaje automático.

La contribución del modelo a la toma de decisiones se refleja en la capacidad de anticipar escenarios, comparar dinámicas territoriales, sustentar técnicamente la planificación de recursos y evaluar la correspondencia entre las acciones institucionales y los patrones identificados. No obstante, las predicciones deben interpretarse como insumos complementarios y no como mecanismos automáticos de decisión, debido a que el criterio de los equipos técnicos, las condiciones territoriales y la disponibilidad presupuestaria continúan siendo factores esenciales en la planificación social.

En conjunto, el modelo incorpora al SIL una capacidad predictiva orientada a la planificación territorial mediante estimaciones desagregadas, fortaleciendo la toma de decisiones a través de información cuantitativa, interpretable y anticipatoria. Su aplicación contribuye a evolucionar desde una gestión predominantemente reactiva hacia un enfoque de planificación basado en evidencia, con mayor capacidad para responder oportunamente a las necesidades sociales de la provincia

### 4.3. Prueba de concepto: escenarios de predicción territorial

La prueba de concepto tuvo como finalidad comprobar la viabilidad funcional del modelo predictivo dentro de un entorno orientado a la consulta y análisis territorial. Para ello, el modelo LightGBM optimizado fue integrado en la aplicación web SIPP Manabí, permitiendo generar estimaciones de demanda correspondientes al siguiente periodo mensual

mediante combinaciones específicas de cantón, parroquia, categoría y subcategoría de atención social.

La aplicación emplea los tres periodos históricos más recientes asociados a la selección realizada por el usuario. A partir de estos datos se generan variables derivadas relacionadas con rezagos temporales, promedios móviles, desviación estándar, tendencia y componentes estacionales. Posteriormente, esta información es procesada por el modelo predictivo para obtener una estimación cuantitativa de la demanda esperada para el mes siguiente. Finalmente, el resultado generado se contrasta con la demanda registrada en el periodo anterior, permitiendo identificar posibles variaciones en su comportamiento.

### 4.3.1. Configuración de los escenarios territoriales

Los escenarios de predicción fueron definidos a partir de cuatro variables de entrada: cantón, parroquia, categoría de atención social y subcategoría de atención o ayuda. Esta parametrización permitió que cada estimación generada correspondiera a una combinación territorial y operativa específica, evitando la utilización de valores promedio generales que podrían limitar la identificación de diferencias asociadas a las características particulares de cada territorio y tipo de servicio.

**Tabla 11**

*Variables de entrada utilizadas en la configuración de escenarios de predicción territorial*

| Variable | Función dentro del escenario |
|---|---|
| Cantón | Delimita el territorio provincial de análisis |
| Parroquia | Especifica la unidad territorial local |
| Categoría de atención social | Identifica el área general de intervención |
| Subcategoría de atención o ayuda | Define el servicio o ayuda específica |

| | |
|---|---|
| Historial reciente | Proporciona los tres periodos previos requeridos para la predicción |
| Periodo de pronóstico | Corresponde al mes posterior al último registro disponible |

La aplicación implementó un proceso de filtrado automático que vincula las parroquias disponibles con el cantón seleccionado y las subcategorías de atención con la categoría previamente elegida. Esta funcionalidad permite reducir posibles inconsistencias en la configuración de los escenarios y garantiza que las predicciones sean generadas únicamente para combinaciones territoriales y operativas existentes dentro de los registros administrativos.

### 4.3.2. Escenario de incremento de la demanda

El escenario de incremento se configura cuando la demanda proyectada para el periodo mensual siguiente supera el valor registrado en el mes anterior. Este comportamiento puede interpretarse como una alerta temprana sobre un posible aumento en la presión operativa de los servicios de atención social dentro del territorio analizado.

Desde la perspectiva de la gestión institucional, este escenario proporciona información útil para evaluar anticipadamente la disponibilidad de personal, insumos, ayudas técnicas, brigadas de atención y recursos logísticos necesarios. No obstante, la estimación generada por el modelo no establece de manera automática una redistribución presupuestaria, sino que constituye un elemento de apoyo para advertir posibles requerimientos de fortalecimiento de la capacidad operativa en un territorio o servicio específico.

### 4.3.3. Escenario de disminución de la demanda

El escenario de disminución se establece cuando la demanda pronosticada para el siguiente periodo mensual presenta un valor inferior al registrado en el mes anterior. Este comportamiento puede asociarse con una posible reducción temporal en el volumen de atenciones; sin embargo, su interpretación debe considerar las condiciones particulares del territorio y la confiabilidad de los registros administrativos disponibles.

Una disminución proyectada en la demanda no implica necesariamente una reducción efectiva de las necesidades sociales de la población. Este comportamiento también puede estar relacionado con factores como cambios en la cobertura de los servicios, interrupciones operativas, limitaciones en los procesos de registro o variaciones propias de la estacionalidad. Por esta razón, los resultados del modelo deben contrastarse con información institucional complementaria antes de realizar ajustes en la planificación o distribución de recursos.

#### 4.3.4. Escenario de estabilidad

El escenario de estabilidad se configura cuando la demanda pronosticada presenta valores iguales o cercanos a los registrados durante el último periodo analizado. Este comportamiento indica que, bajo condiciones similares a las observadas recientemente, el nivel de atención podría mantenerse dentro de un rango comparable en el periodo futuro.

Desde la perspectiva de la planificación territorial, este escenario permite mantener una capacidad operativa cercana a la utilizada en periodos anteriores. No obstante, una estabilidad en el valor agregado de la demanda no implica necesariamente un comportamiento uniforme entre los diferentes niveles de análisis, debido a que pueden existir variaciones específicas entre categorías, subcategorías o parroquias. Por ello, resulta

necesario complementar la interpretación con una revisión desagregada del comportamiento de los servicios.

### 4.3.5. Escenario de información insuficiente

La prueba de concepto incorpora un mecanismo de validación para aquellos casos en los que una combinación territorial y operativa cuenta con menos de tres periodos históricos disponibles. En estas situaciones, el sistema restringe la generación de predicciones y presenta una alerta indicando que la información disponible es insuficiente para realizar la estimación.

Esta limitación evita la generación de pronósticos basados en historiales reducidos que no permiten calcular adecuadamente las variables temporales requeridas por el modelo. Además, facilita la identificación de territorios o servicios donde resulta necesario fortalecer la continuidad, completitud y calidad de los registros administrativos utilizados como fuente de información.

### 4.3.6. Presentación e interpretación del escenario

Una vez generado el pronóstico, la aplicación muestra la demanda estimada, el valor registrado en el periodo anterior y la diferencia entre ambos resultados. Además, el dashboard integra una visualización de la trayectoria histórica, una representación territorial del cantón seleccionado y un análisis local de interpretabilidad basado en valores SHAP.

La representación temporal permite analizar si la predicción mantiene el comportamiento observado en los periodos recientes o si evidencia cambios relevantes en la tendencia de la demanda. Por su parte, el componente SHAP facilita la comprensión de los factores que contribuyeron al aumento o disminución de la estimación, proporcionando mayor

transparencia en el proceso predictivo. De esta manera, el escenario generado no se limita a presentar un valor numérico, sino que incorpora elementos que permiten interpretar su comportamiento y apoyar el análisis institucional.

**Tabla 12**
*Escenarios de predicción territorial considerados en la prueba de concepto*

| Escenario | Condición | Interpretación operativa |
|---|---|---|
| Incremento | Pronóstico superior al último valor observado | Posible aumento de la presión sobre los servicios |
| Disminución | Pronóstico inferior al último valor observado | Posible reducción de atenciones o cambio en cobertura y registro |
| Estabilidad | Pronóstico igual o cercano al último valor observado | Continuidad probable del nivel reciente de demanda |
| Información insuficiente | Menos de tres periodos históricos | Imposibilidad de generar un pronóstico con soporte temporal mínimo |

### 4.3.7. Alcance de la prueba de concepto

La prueba de concepto permitió comprobar que el modelo predictivo puede integrarse en una aplicación web funcional, recibir parámetros territoriales y operativos, generar estimaciones y presentar resultados con mecanismos de interpretación. Su principal contribución radica en validar la viabilidad técnica del flujo completo, desde la configuración del escenario de consulta hasta la generación y visualización del pronóstico.

Los escenarios obtenidos deben interpretarse como estimaciones aproximadas sujetas a un margen de incertidumbre propio de los modelos predictivos. La herramienta constituye un apoyo para el análisis institucional, pero no sustituye el criterio de los equipos técnicos ni considera de manera directa factores externos como disponibilidad presupuestaria,

características demográficas o situaciones coyunturales que pueden influir en la demanda social.

# CAPÍTULO 5: Conclusiones y recomendaciones

## 5.1. Conclusiones

La integración, depuración y estructuración de los registros administrativos provenientes del Sistema de Información Local (SIL) permitió consolidar un conjunto de datos analítico orientado al estudio de la demanda territorial de atención social en la provincia de Manabí. Este proceso evidenció que la información institucional, después de aplicar procedimientos de limpieza, homologación y transformación, puede convertirse en un insumo relevante para el desarrollo de modelos predictivos destinados a fortalecer la gestión pública.

La aplicación de técnicas de balanceo y generación de datos sintéticos mediante CTGAN contribuyó a ampliar la disponibilidad de observaciones y mejorar la representación del conjunto de datos empleado durante la etapa de modelado. No obstante, los registros sintéticos deben entenderse como un recurso metodológico complementario para mitigar limitaciones asociadas al volumen y distribución de los datos, sin reemplazar la incorporación progresiva de registros administrativos reales correspondientes a nuevos periodos de observación.

El desarrollo y evaluación de los modelos predictivos confirmaron la viabilidad técnica del uso de algoritmos de aprendizaje automático para identificar patrones asociados a la demanda territorial de atención social. Entre las alternativas analizadas, el modelo LightGBM optimizado obtuvo el mejor desempeño, con un MAE de 2,0636, un RMSE de 3,4162 y un $R^2$ de 0,8285, evidenciando una mayor capacidad para explicar la variabilidad de la demanda y menores niveles de error frente al modelo base, LightGBM inicial y CatBoost.

La comparación entre los modelos permitió establecer que la incorporación de variables temporales, territoriales y operativas proporcionó información relevante para representar la dinámica de la demanda social. La reducción del MAE y RMSE alcanzada por el modelo optimizado respecto al modelo base demuestra que los algoritmos de aprendizaje automático presentan una mayor capacidad predictiva frente a enfoques sustentados únicamente en valores históricos promedio.

Desde una perspectiva metodológica, la investigación aporta un procedimiento estructurado para aprovechar registros administrativos en escenarios caracterizados por restricciones relacionadas con disponibilidad, calidad y volumen de información. La integración de la metodología CRISP-DM, técnicas de preparación de datos, generación sintética y evaluación comparativa de modelos constituye una referencia metodológica aplicable a futuras investigaciones orientadas al desarrollo de soluciones predictivas en servicios públicos.

El modelo desarrollado representa un prototipo funcional de apoyo analítico para la exploración de escenarios de demanda territorial de atención social. En consecuencia, sus resultados deben interpretarse como estimaciones orientativas y no como una validación definitiva del comportamiento futuro de la demanda ni como un mecanismo sustituto de los procesos institucionales de planificación. Su principal contribución radica en demostrar la viabilidad técnica de construir herramientas predictivas a partir del aprovechamiento de registros administrativos disponibles.

### 5.2. Limitaciones

La principal limitación de la investigación estuvo asociada con la disponibilidad temporal de los registros administrativos empleados. El periodo histórico disponible presentó restricciones para capturar variaciones de largo plazo en la demanda social, lo que podría limitar la capacidad del modelo para representar cambios estructurales o comportamientos emergentes en escenarios futuros.

La generación de datos sintéticos mediante CTGAN permitió ampliar el conjunto analítico utilizado durante el modelado; sin embargo, estos datos representan patrones derivados de la información original y no incorporan nuevas dinámicas sociales que no hayan sido registradas previamente. Por ello, los resultados obtenidos deben interpretarse considerando que el entrenamiento del modelo combinó información real y sintética, lo que implica reconocer las particularidades y limitaciones asociadas a esta estrategia.

El alcance del proyecto se circunscribió al desarrollo y evaluación de un prototipo predictivo bajo condiciones controladas de investigación. En consecuencia, no se realizó una validación operativa en un entorno de producción con usuarios institucionales ni una evaluación longitudinal del comportamiento del modelo frente a nuevos periodos de registros de atención social generados por la institución.

### 5.3. Recomendaciones

Recomendaciones institucionales Se recomienda fortalecer los procesos institucionales relacionados con el registro, actualización y estandarización de la información generada por las entidades responsables de la atención social. Esto permitirá disponer de bases históricas más completas, consistentes y representativas, favoreciendo el desarrollo de futuros estudios predictivos con mayor capacidad de análisis y generalización.

Asimismo, se recomienda consolidar la continuidad del Sistema de Información Local como una fuente estratégica para la generación de conocimiento territorial, mediante la implementación de mecanismos orientados a garantizar la calidad, consistencia, trazabilidad y disponibilidad periódica de los registros administrativos. Estas acciones contribuirán a mejorar el aprovechamiento de los datos institucionales para la planificación y toma de decisiones basada en evidencia.

**Recomendaciones metodológicas**

Se recomienda que futuras implementaciones del modelo incorporen nuevos periodos de información real antes de su utilización como herramienta de apoyo operativo. Esta ampliación permitirá evaluar su estabilidad, capacidad de adaptación y desempeño frente a diferentes condiciones temporales y territoriales.

Asimismo, se recomienda establecer procesos periódicos de evaluación del desempeño del modelo mediante métricas estadísticas actualizadas, con el objetivo de verificar si mantiene niveles adecuados de precisión y confiabilidad al incorporar nuevos registros administrativos generados por la institución.

Finalmente, se recomienda complementar las estimaciones predictivas con análisis cualitativos desarrollados por especialistas del ámbito social, considerando que la demanda de atención está influenciada por factores sociales, económicos e institucionales que pueden no encontrarse completamente reflejados en los registros administrativos disponibles.

**Líneas de investigación futura**

Se recomienda explorar nuevos enfoques de modelado que permitan comparar el desempeño de diferentes algoritmos de aprendizaje automático y analizar su comportamiento utilizando bases de datos institucionales con mayor amplitud y diversidad de información. Esto contribuirá a identificar alternativas metodológicas que puedan mejorar la capacidad predictiva y la adaptación del modelo a distintos escenarios territoriales.

Se recomienda incorporar progresivamente variables externas relacionadas con aspectos socioeconómicos, demográficos y territoriales que permitan enriquecer la representación de los factores asociados a la demanda social. La integración de estas fuentes complementarias podría favorecer una comprensión más amplia de las dinámicas que influyen en la necesidad de atención social.

Se recomienda desarrollar investigaciones posteriores orientadas a validar el modelo mediante registros administrativos reales acumulados durante periodos prolongados. Esta validación permitirá analizar su capacidad de generalización, evaluar su comportamiento bajo diferentes condiciones operativas y determinar la viabilidad de una incorporación progresiva como herramienta complementaria para la planificación institucional.

## Bibliografía

Amazon Web Services [AWS]. (s.f.). ¿Qué son los datos sintéticos? - Explicación de los datos sintéticos. Recuperado de Amazon Web Services.

Bahamonde Morales, D. I., Tapia Pisarro, W. S., & Morillo Alcívar, P. A. (2022). Análisis comparativo del rendimiento de algoritmos de clasificación binaria en un conjunto de datos desbalanceados. Universidad Politécnica Salesiana.

Cabrera-Sánchez, J. F., Cruz-Corona, C., Yáñez Escolano, A., & Silva-Ramírez, E. L. (2024, May). A benchmark for missing data imputation techniques: Development perspectives and performance comparative. En International Conference on Optimization and Learning (pp. 140–153). Springer Nature Switzerland.

Caminos-Maldonado, F. E., & Quiroz, P. T. V. (2022). La gestión pública eficaz: fundamentos, factores y casos. Pro Sciences: Revista de Producción, Ciencias e Investigación, 6(45), 282–296.

Casanova-Villalba, C. I., Herrera-Sánchez, M. J., & Proaño-González, E. A. (2023). Impacto de la analítica predictiva en la toma de decisiones gerenciales. Revista Científica Ciencia y Método, 1(3), 16–30.

Chapman, P., Clinton, J., Kerber, R., Khabaza, T., Reinartz, T., Shearer, C., & Wirth, R. (1999). CRISP-DM 1.0

Del Riego, S. A. (2024). Generador de datos sintéticos tabulares basado en modelos de difusión. Universidad Politécnica de Madrid.

Espinosa-Zúñiga, J. J. (2020). Aplicación de metodología CRISP-DM para segmentación geográfica de una base de datos pública. Ingeniería, Investigación y Tecnología, 21(1).

Flores-Cano, Y. d. P. Y., Neira Ortega, I. M., Cairo Méndez, S. Y., & Garay Mercado, M. C. (2026). Gobierno digital y confianza ciudadana: una revisión sistemática orientada a la construcción de una agenda metodológica basada en evidencia. Aula Virtual, 7(14), e692. https://doi.org/10.5281/zenodo.18937506

Fondo de las Naciones Unidas para la Infancia, Comisión Económica para América Latina y el Caribe, Instituto Nacional de Estadísticas de Chile, & Instituto Nacional de Estadística de Uruguay. (2024). Diagnóstico y recomendaciones para integrar registros administrativos relacionados con la niñez: Uso de registros administrativos para mejorar las estadísticas sobre niñas, niños y adolescentes. Naciones Unidas.

Gobierno Autónomo Descentralizado Provincial de Manabí. (2025). Desarrollo Humano - Prefectura de Manabí. Recuperado de https://www.manabi.gob.ec/

Gobierno Autónomo Descentralizado Provincial de Manabí, Subdirección de Planificación Territorial. (2026). Plan de Gestión de la Información del Sistema de Información Local 2026. Portoviejo, Ecuador.

Herrera, F. (2016). Big Data: Preprocesamiento y calidad de datos. Novática, 237(1), 17–20.

IBM. (s. f.). CRISP-DM overview. IBM Documentation. https://www.ibm.com/docs/es/spss-modeler/saas?topic=dm-crisp-help-overview

Instituto Nacional de Estadística y Censos [INEC]. (2022). Metodología para transformar registros administrativos en registros estadísticos.

https://www.ecuadorencifras.gob.ec/documentos/web-inec/Bibliotecas/Libros/Metod_para_transformar_registros_admin_en_registros_estad.pdf

Lema Changoluisa, K., & Romo, E. (2022). Registros administrativos para los sistemas de información y política pública local. Estudios de la Gestión: Revista Internacional de Administración, (12), 13–30. https://doi.org/10.32719/25506641.2022.12.1

Lemus-Delgado, D., & Pérez Navarro, R. (2020). Ciencia de datos y estudios globales: aportaciones y desafíos metodológicos. Colombia Internacional, 102, 41–62. https://doi.org/10.7440/colombiaint102.2020.03

Organización para la Cooperación y el Desarrollo Económicos. (2019–2020). The Path to Becoming a Data-Driven Public Sector. OECD Digital Government Studies. OECD Publishing. https://doi.org/10.1787/059814a7-en

Ortega Ovalle, M. T. (2025). Importancia de la Ciencia de los Datos en la Sociedad. Ciencia Latina Revista Científica Multidisciplinar, 9(6), 2804–2823. https://doi.org/10.37811/cl_rcm.v9i6.21409

Pathare, A., Mangrulkar, R., Suvarna, K., Parekh, A., Thakur, G., & Gawade, A. (2023). Comparison of tabular synthetic data generation techniques using propensity and cluster log metric. International Journal of Information Management Data Insights, 3(2), 100177.

Pérez Lara, J. E., Covarrubias Moreno, O. M., & Tolentino Sanjuan, A. V. (2025). Analítica de datos y gobierno abierto: hacia una gestión pública basada en evidencia. RIESED - Revista Internacional de Estudios sobre Sistemas Educativos, 3(16), 762–780. https://doi.org/10.5281/zenodo.15697620

Quintanilla, G., & Gil-García, J. R. (2016). Gobierno abierto y datos vinculados: conceptos, experiencias y lecciones con base en el caso mexicano. Revista del CLAD Reforma y Democracia, 65, 69–102. https://doi.org/10.69733/clad.ryd.n65.a923

Sainz-Aja Guerra, J. A. (2025). Generación y validación de datos sintéticos para la mejora de modelos predictivos en entornos de datos limitados [Trabajo de fin de máster, Universidad de Cantabria].

Samuel, A. L. (1959). Some studies in machine learning using the game of checkers. IBM Journal of Research and Development, 3(3), 210–229.

Sánchez Pérez, A. (2022). Generación de datos sintéticos para el refuerzo de modelos de aprendizaje automático en entornos reales para la extracción de información [Trabajo de fin de grado, Universidad Politécnica de Madrid]. Archivo Digital UPM. https://oa.upm.es/70940/

Sepúlveda Calle, C. A., & Benavides Posso, M. T. (2023). Análisis predictivo de demanda de servicios bajo series temporales [Trabajo de grado de especialización, Universidad de Antioquia]. Repositorio Institucional Universidad de Antioquia.

Shearer, C. (2000). The CRISP-DM model: The new blueprint for data mining. Journal of Data Warehousing, 5(4), 13–22.

Trasobares, A. H. (2003). Los sistemas de información: evolución y desarrollo. Proyecto Social: Revista de Relaciones Laborales, (10), 149–165.

Velilla Recio, D. (2025). Métodos de generación de datos sintéticos tabulares basados en inteligencia artificial para aumentación de datos [Trabajo de Fin de Grado, Universitat Politècnica de Catalunya].

Xu, L., Skoularidou, M., Cuesta-Infante, A., & Veeramachaneni, K. (2019). Modeling tabular data using conditional GAN. Advances in Neural Information Processing Systems, 32.

## Anexos

Repositorio de Proyecto: ProyectoUIDE_Grupo7.rar