



Instituto Tecnológico y de Estudios Superiores de Monterrey

Maestría en Inteligencia Artificial Aplicada

Proyecto Integrador

TC5035.10

Semana 4

Avance 2. Ingeniería de características

“Generalización de modelos nacionales de pronóstico epidemiológico hacia un enfoque modular con desagregación por sexo y entidad federativa en México”

Equipo No 1:

Programa	Matrícula	Estudiante
MNA	A01795838	Javier Augusto Rebull Saucedo
MNA	A01795941	Juan Carlos Pérez Nava
MNA	A01232963	Luis Gerardo Sánchez Salazar

Equipo docente:

Dra. Grettel Barceló Alonso – Profesora Titular y Asesora del Proyecto

Dr. Luis Eduardo Falcón Morales – Director Nacional de MNA

Mtra. Verónica Sandra Guzmán de Valle – Profesora Asistente

Patrocinador:

Dra. Ruth Pérez-Hernández – Líder del Proyecto IMSS

Dra. Lina Díaz Castro – Investigadora en Psiquiatría

Febrero 08, 2026

Índice

Consulta el Análisis Completo	5
1. Setup y Configuración	8
2. Carga de Datos	8
2.1. Boletín Epidemiológico	8
3. Construcción del Dataset INEGI	8
3.1. Obtención de datos demográficos y territoriales (INEGI)	8
3.2. Ajustando el Dataset INEGI	8
3.2.1. Revisión de estructura y tipos de datos	8
3.3. Ajuste y transformación del dataset poblacional	9
3.4. Integración poblacional–territorial	9
4. Generación de variables derivadas y categorizaciones estructurales	10
5. Integración de variables INEGI al Dataset epidemiológico	10
5.1. Merge por entidad	10
5.2. Validación del dataset epidemiológico enriquecido	10
6. Visualización exploratoria de variables demográficas y territoriales	11
6.1. Interpretación rápida de las visualizaciones estructurales	12
6.1.1. Variables numéricas	12
6.1.2. Variables categóricas	13
6.1.3. Conclusión exploratoria	13
7. Series de tiempo por categoría territorial	14
7.1. Región de salud mental (<code>region_salud_mental</code>)	14
7.2. Densidad poblacional por percentiles (<code>densidad_poblacional_percentil</code>)	15
7.3. Tamaño poblacional por rangos fijos (<code>tamano_poblacional_predefinido</code>)	16
7.4. Tamaño poblacional por percentiles (<code>tamano_poblacional_grupo_percentil</code>)	17
7.5. Extensión territorial por percentiles (<code>extension_territorial_percentil</code>)	18
7.6. Ratio hombres/mujeres categorizado (<code>ratio_h_m_cat</code>)	19
8. Visualización de datos	21
8.1. Uso de Tableau Public	21
8.2. Tabla de datos consolidada	21
8.3. Mapa de México con variables demográficas	22
8.4. México en categorías	23
8.5. Casos por año	24
8.6. Casos semanales	25
8.7. Predicciones epidemiológicas	26

9. Síntesis del Avance: Ingeniería de Características y Visualización	28
9.1. Pipeline de Datos y Enriquecimiento	28
9.2. Hallazgos Clave de la Ingeniería de Características	28
9.3. Contribución al Modelado Futuro	29
9.4. Limitaciones de esta Etapa	29
Conclusiones Finales del Equipo	30
Reflexiones del Equipo	32
Referencias	33

Índice de figuras

1.	Histogramas de variables numéricas (superficie, ratio H/M, densidad poblacional).	11
2.	Boxplots de variables numéricas.	11
3.	Distribución de variables categóricas (región, tamaño poblacional, ratio H/M categorizado).	12
4.	Mapa de calor de correlaciones entre variables numéricas.	12
5.	Región de salud mental – Incrementos semanales (Hombres).	14
6.	Región de salud mental – Incrementos semanales (Mujeres).	14
7.	Región de salud mental – Incrementos semanales totales.	15
8.	Densidad poblacional – Incrementos semanales (Hombres).	15
9.	Densidad poblacional – Incrementos semanales (Mujeres).	15
10.	Densidad poblacional – Incrementos semanales totales.	16
11.	Tamaño poblacional (rangos fijos) – Incrementos semanales (Hombres).	16
12.	Tamaño poblacional (rangos fijos) – Incrementos semanales (Mujeres).	16
13.	Tamaño poblacional (rangos fijos) – Incrementos semanales totales.	17
14.	Tamaño poblacional (percentiles) – Incrementos semanales (Hombres).	17
15.	Tamaño poblacional (percentiles) – Incrementos semanales (Mujeres).	17
16.	Tamaño poblacional (percentiles) – Incrementos semanales totales.	18
17.	Extensión territorial – Incrementos semanales (Hombres).	18
18.	Extensión territorial – Incrementos semanales (Mujeres).	18
19.	Extensión territorial – Incrementos semanales totales.	19
20.	Ratio H/M – Incrementos semanales (Hombres).	19
21.	Ratio H/M – Incrementos semanales (Mujeres).	19
22.	Ratio H/M – Incrementos semanales totales.	20
23.	Tabla de datos consolidada en Tableau.	22
24.	Mapa de México con variables demográficas.	23
25.	México en categorías territoriales y demográficas.	24
26.	Casos por año, desagregados por sexo.	25
27.	Incremento de casos semanales por sexo.	26
28.	Predicciones epidemiológicas (simuladas).	27

⊕ Dashboard Interactivo — Versión Preliminar

△ NOTA IMPORTANTE

La siguiente visualización corresponde a una **versión preliminar (beta)** del dashboard, publicada en **Tableau Public** con fines demostrativos y exploratorios. Los datos de **2026** son simulados y las predicciones se encuentran en fase de validación.

► Acceder al Dashboard en vivo ◀

<https://proyectointegrador.org/epidashboard>

Característica	Detalle
Plataforma	Tableau Public
Interactividad	Filtros por año, padecimiento y sexo
Cobertura	32 entidades federativas
Padecimientos	Depresión (F32) · Parkinson (G20) · Alzheimer (G30)
Fuentes de datos	SINAVE (boletines epidemiológicos) + INEGI (demográficos)
Estatus	• Beta — sujeto a iteraciones

Consulta el Análisis Completo

Accede a la versión extendida de este **Análisis de características** y el reporte ejecutivo en:

https://proyectointegrador.org/avance2_equipo01

Página Web del Proyecto

Hemos diseñado una página web institucional donde se pueden consultar los objetivos estratégicos, el stack tecnológico y los datos fundamentales del proyecto:

<https://proyectointegrador.org/>

Repositorio del Proyecto

El código fuente, los pipelines de procesamiento de datos, los notebooks de análisis y la documentación técnica del proyecto **EpiForecast-MX** se encuentran disponibles en el repositorio oficial de GitHub. Este repositorio centraliza el desarrollo del proyecto y garantiza reproducibilidad, versionado y trazabilidad de cada etapa del análisis.

<https://github.com/IntegradorIMSS2026Team01/EpiForecast-MX>

En el repositorio se incluye: el pipeline completo de extracción, limpieza, transformación y enriquecimiento de datos; los notebooks de análisis exploratorio, ingeniería de características y visualización; los scripts de automatización y configuración; y la documentación técnica necesaria para replicar y extender el proyecto.

Contexto

El presente documento corresponde al **Avance 2: Ingeniería de Características (Feature Engineering)** del proyecto **EpiForecast-MX**, desarrollado en colaboración entre el Tecnológico de Monterrey y el Instituto Mexicano del Seguro Social (IMSS). Esta etapa se construye sobre los resultados del Análisis Exploratorio de Datos (EDA) y del pipeline de limpieza previamente definidos, y corresponde a la fase de **Preparación de los Datos** dentro de la metodología **CRISP-ML**.

El proyecto tiene como objetivo la construcción de modelos de pronóstico epidemiológico para tres condiciones neurológicas y de salud mental en México, clasificadas conforme a la CIE-10: **Depresión (F32)**, **Enfermedad de Parkinson (G20)** y **Enfermedad de Alzheimer (G30)**. Para ello, se parte de los registros semanales del Sistema Nacional de Vigilancia Epidemiológica (SINAVE) y se integran variables estructurales de contexto demográfico y territorial provenientes del Instituto Nacional de Estadística y Geografía (INEGI).

En esta fase, el énfasis no está en describir nuevamente los datos, sino en **transformarlos, enriquecerlos y estructurarlos** de manera que resulten adecuados para algoritmos de aprendizaje automático, reduciendo sesgos, mejorando la estabilidad numérica y facilitando la convergencia de los modelos predictivos.

Objetivo de la Ingeniería de Características

La ingeniería de características en este avance cumple los siguientes propósitos:

1. **Generación de variables informativas:** construir nuevas características que capturen patrones estructurales relevantes, tales como densidad poblacional, tamaño poblacional, composición por sexo y contexto socio-urbano, las cuales no están explícitas en los registros epidemiológicos originales.
2. **Reducción de sesgos y escalas:** mitigar distorsiones derivadas de acumulados, diferencias territoriales extremas y escalas heterogéneas entre entidades federativas, preparando el dataset para modelos estadísticos y de aprendizaje automático.
3. **Preparación para el modelado:** facilitar la selección de características, reducir la complejidad del espacio de variables y acelerar la convergencia de los algoritmos que se emplearán en etapas posteriores del proyecto.

Dataset base y enriquecimiento

El dataset de partida corresponde a la versión **limpia y transformada** de los registros epidemiológicos del SINAVE, obtenida al cierre del Avance 1. Sobre este conjunto se integran variables demográficas y territoriales a nivel estatal, obtenidas directamente de fuentes oficiales de INEGI, mediante un proceso reproducible de descarga, validación y combinación por entidad federativa.

Cuadro 1: Tipos de variables en el dataset

Tipo de variable	Descripción general
Epidemiológicas	Casos acumulados por sexo, año y semana epidemiológica
Demográficas	Población total, hombres y mujeres por entidad
Territoriales	Superficie estatal y densidad poblacional
Categóricas derivadas	Tamaño poblacional, percentiles, ratio H/M, región

Estructura del Documento

Sección	Contenido
1. Setup y Configuración	Importaciones, constantes, funciones auxiliares
2. Carga de Datos	Lectura del dataset epidemiológico limpio
3. Construcción del Dataset INEGI	Dataset demográfico y territorial base
4. Ingeniería de Características	Variables derivadas y categorizaciones
5. Integración INEGI + Epidemiológico	Merge y dataset enriquecido
6. Visualización Exploratoria	Ánálisis gráfico de variables
7. Series de Tiempo	Evolución temporal por categoría
8. Visualización en Tableau	Workbook interactivo
9. Síntesis del Avance	Hallazgos clave

Recursos y Reproducibilidad

Como pilar de transparencia y reproducibilidad bajo el marco de la metodología **CRISP-ML(Q)**, la totalidad del código fuente, los activos de datos procesados y el historial de versiones de este análisis se encuentran centralizados en el repositorio oficial. Este espacio no solo sirve como el motor de desarrollo de **EpiForecast-MX**, sino también como una plataforma de colaboración abierta donde se documenta cada fase del ciclo de vida del proyecto.

Setup y Configuración

Importaciones, constantes globales y funciones auxiliares reutilizables para todo el análisis. El código fuente completo se encuentra disponible en el notebook y en el repositorio del proyecto.

Carga de Datos

Boletín Epidemiológico

El dataset se obtiene mediante DVC (`dvc pull`) y se encuentra en la ruta `data/processed/dataset_boletin_epidemiologico.csv`. Contiene registros semanales de casos de Depresión (F32), Parkinson (G20) y Alzheimer (G30) extraídos de los Boletines Epidemiológicos del SINAVER (2012–2025), desglosados por entidad federativa y sexo.

Construcción del Dataset INEGI

Obtención de datos demográficos y territoriales (INEGI)

En esta sección se descargan y preparan los **insumos demográficos y territoriales** que servirán como base para la generación de nuevas características estructurales.

Primero, se consulta la API PxWeb de INEGI para obtener información de **población por entidad federativa**, desagregada por sexo y periodo. Los datos se reciben en formato JSON-STAT y se convierten a un DataFrame tabular en formato largo, manteniendo la estructura original definida por la fuente oficial.

De manera complementaria, se descargan los datos de **extensión territorial por entidad federativa** desde los servicios de indicadores de INEGI. La variable de superficie se limpia y transforma a formato numérico para asegurar su uso posterior en cálculos de densidad y categorizaciones territoriales.

Ajustando el Dataset INEGI

Revisión de estructura y tipos de datos

Antes de avanzar con integraciones y generación de características, se realiza una **revisión general de la estructura de los datasets** descargados desde INEGI. Este paso permite asegurar que los tipos de datos sean consistentes y que la información esté lista para su uso analítico.

Para cada conjunto de datos se revisa: estructura del DataFrame (número de filas y columnas, tipo de dato de cada variable), consistencia general (detección temprana de variables mal tipadas o con conversiones incompletas) y resumen descriptivo (tabla resumen que sintetiza la información principal del dataset).

Ajuste y transformación del dataset poblacional

En esta sección se realizan transformaciones básicas sobre el dataset poblacional con el objetivo de **estandarizar su estructura** y prepararlo para su integración con otras fuentes de información.

Las operaciones realizadas son las siguientes:

1. **Conversión a valores numéricos:** La variable de interés se convierte explícitamente a tipo numérico. Cualquier valor no convertible se asigna como nulo, evitando errores silenciosos en etapas posteriores.
2. **Selección del total poblacional:** Se filtra el dataset para conservar únicamente el total poblacional, eliminando desagregaciones por edad que no se utilizan en esta etapa.
3. **Pivoteo por sexo:** El DataFrame se reestructura para que cada fila represente una entidad federativa, con columnas separadas para la población de hombres, mujeres y total.

Integración poblacional–territorial

En esta sección se realizan tres pasos clave para asegurar la **consistencia interna** de los datos poblacionales y su correcta integración con la información territorial.

1. **Validación poblacional por sexo:** Se verifica que, para cada entidad federativa, la suma de la población de hombres y mujeres sea igual a la población total.
2. **Selección del año 2020 como corte transversal:** El dataset poblacional se filtra para conservar únicamente el año 2020, el cual corresponde al Censo de Población y Vivienda más reciente realizado en México.
3. **Merge con superficie territorial:** Se integra la información de extensión territorial a la tabla poblacional mediante un *left join* por entidad federativa normalizada.

Generación de variables derivadas y categorizaciones estructurales

En esta sección se construyen **nuevas características** a partir de la información poblacional y territorial integrada, con el objetivo de capturar diferencias estructurales relevantes entre entidades federativas y facilitar su uso en análisis epidemiológicos y modelos predictivos.

Las transformaciones realizadas se agrupan en cuatro bloques principales:

1. **Regionalización socio-urbana en salud mental:** Se asigna a cada entidad federativa una categoría de región de salud mental, definida previamente con un enfoque socio-urbano y estructural. Esta variable categórica busca capturar diferencias en urbanización, acceso a servicios y capacidad diagnóstica.
2. **Discretización de variables continuas:** Se generan versiones categóricas de la densidad poblacional, el tamaño poblacional y la extensión territorial, tanto por percentiles como por rangos fijos predefinidos.
3. **Ratio hombres/mujeres:** Se calcula el ratio continuo entre la población masculina y femenina, y se discretiza en categorías interpretables.
4. **Densidad poblacional:** Se calcula la relación entre población y superficie, obteniendo la densidad en habitantes por kilómetro cuadrado.

Integración de variables INEGI al Dataset epidemiológico

Merge por entidad

En esta sección se integra la capa estructural de INEGI al dataset epidemiológico limpio. El objetivo es generar un **dataset Enriquecido** que conserve toda la información del boletín y, adicionalmente, incorpore variables demográficas y territoriales.

Este procedimiento replica **exactamente la lógica** implementada en el pipeline del proyecto (`mapea.py` / `mapea_inegi.py`), con la diferencia de que aquí se trabaja directamente con DataFrames en memoria.

El proceso consta de tres pasos: copias defensivas de los DataFrames originales, normalización de nombres de entidades federativas y *left join* por entidad federativa.

Validación del dataset epidemiológico Enriquecido

Una vez integrada la información epidemiológica del boletín con las variables estructurales de INEGI, se realiza una **revisión formal del dataset resultante**. Esta etapa funciona como un conjunto de *sanity checks*: verificar que la integración se haya realizado correctamente y documentar la estructura del dataset que se utilizará en las etapas posteriores del proyecto.

Visualización exploratoria de variables demográficas y territoriales

En esta sección se presenta una **visualización exploratoria rápida (EDA visual)** del dataset epidemiológico enriquecido con información de INEGI. El objetivo es entender la forma, escala y distribución de las variables estructurales que posteriormente se utilizarán como contexto en el análisis epidemiológico y en el modelado.

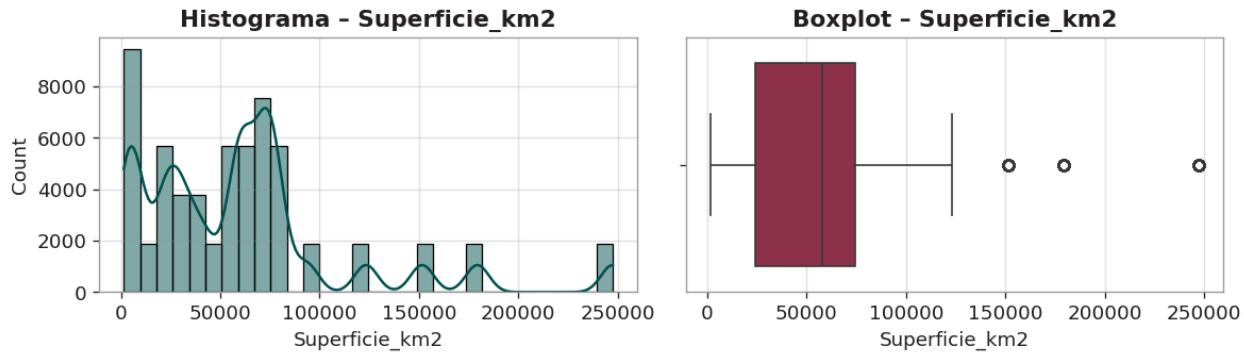


Figura 1: Histogramas de variables numéricas (superficie, ratio H/M, densidad poblacional).

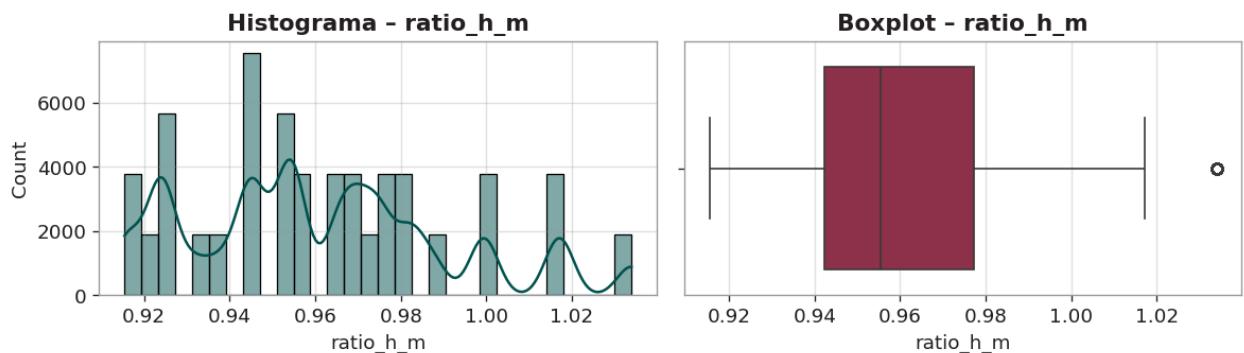


Figura 2: Boxplots de variables numéricas.

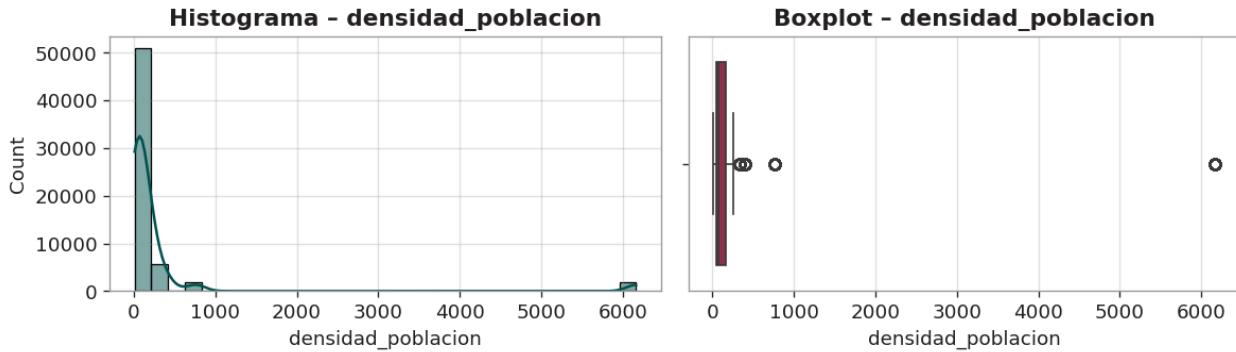


Figura 3: Distribución de variables categóricas (región, tamaño poblacional, ratio H/M categorizado).

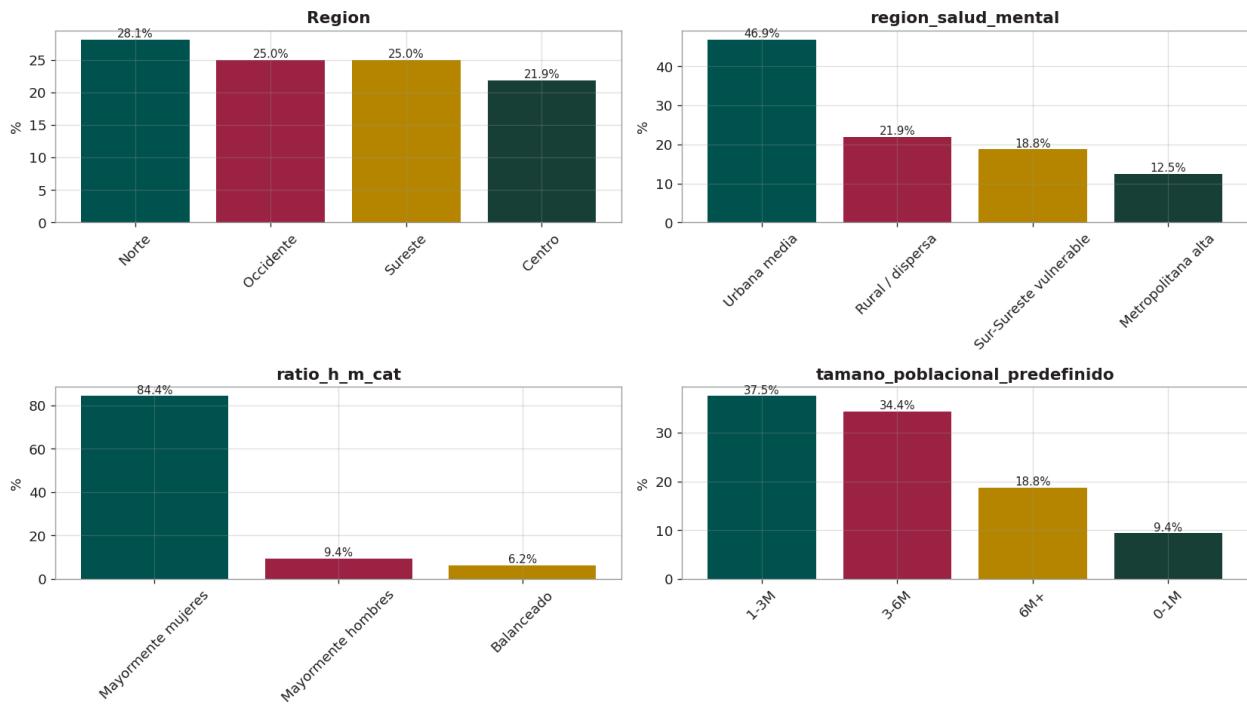


Figura 4: Mapa de calor de correlaciones entre variables numéricas.

Interpretación rápida de las visualizaciones estructurales

Variables numéricas

Superficie territorial (Superficie_km2). La distribución de la superficie territorial muestra una asimetría positiva marcada, con varios outliers naturales asociados a estados de gran extensión territorial (Chihuahua, Sonora, Coahuila). Este comportamiento justifica el uso posterior de percentiles o categorizaciones, ya que el uso directo de la variable puede dominar análisis multivariados.

Densidad poblacional (`densidad_poblacion`). La densidad poblacional presenta una asimetría extrema, con un outlier dominante claramente identificado: la Ciudad de México (entidad con superficie muy pequeña y población muy alta). Este resultado confirma que la densidad poblacional es una variable altamente informativa en epidemiología, pero su uso directo sin transformación puede introducir sesgos fuertes.

Ratio hombres/mujeres (`ratio_h_m`). El ratio hombres/mujeres muestra una distribución compacta, centrada alrededor de 1, lo que indica que las entidades federativas presentan composiciones por sexo relativamente similares. No se observan valores extremos relevantes.

Variables categóricas

Clasificación hombres/mujeres (`ratio_h_m_cat`). La versión categórica del ratio revela una limitación importante de diseño: la categoría “Mayormente mujeres” concentra más del 84 % de las entidades, lo que indica que la categorización actual no discrimina adecuadamente y podría requerir umbrales más estrictos.

Regiones territoriales. Ambas clasificaciones (`Region` y `region_salud_mental`) muestran distribuciones razonablemente balanceadas, sin concentraciones extremas. Son candidatas adecuadas para análisis estratificados y variables exógenas en modelos.

Tamaño poblacional (`tamano_poblacional_predefinido`). La categorización por rangos fijos de población muestra una distribución gradual entre grupos, reflejando la heterogeneidad demográfica del país. Esta variable conserva significado absoluto, lo que la hace más interpretable en términos de política pública.

Conclusión exploratoria

Las visualizaciones confirman que: las asimetrías territoriales y demográficas son estructurales (no errores); algunas variables categóricas (como `ratio_h_m_cat`) requieren revisión metodológica; otras (región, tamaño poblacional) están bien definidas y aportan contexto valioso. Este análisis visual funciona como un *sanity check* previo al modelado.

Series de tiempo por categoría territorial

En esta sección se visualiza la evolución temporal de los **incrementos semanales de casos**, diferenciados por categorías territoriales derivadas de INEGI. El objetivo es analizar la **dinámica epidemiológica a lo largo del tiempo**, incorporando un contexto estructural que permita comparar patrones entre distintos tipos de regiones.

Región de salud mental (`region_salud_mental`)

Tipo: Categórica nominal. **Fuente:** Clasificación definida en el proyecto, apoyada en criterios socio-urbanos.

Esta categoría se asigna mediante un mapeo explícito entidad → región. Las regiones son: Metropolitana alta, Urbana media, Transición y Rural-dispersa.

Segregado por género

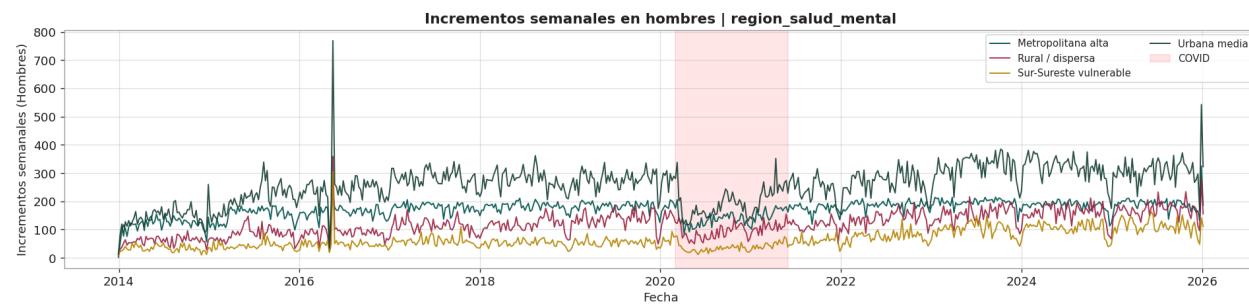


Figura 5: Región de salud mental – Incrementos semanales (Hombres).

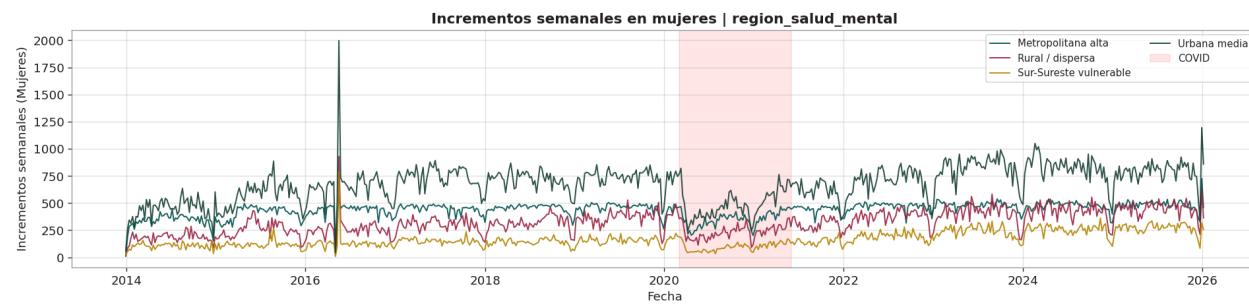


Figura 6: Región de salud mental – Incrementos semanales (Mujeres).

Incrementos totales

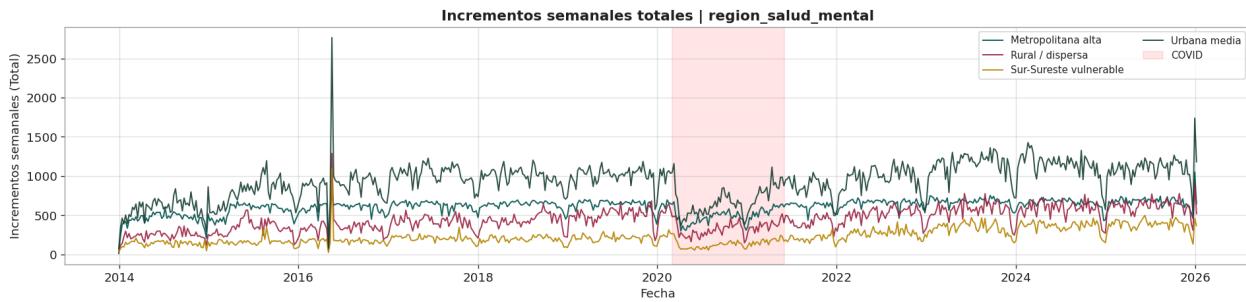


Figura 7: Región de salud mental – Incrementos semanales totales.

Densidad poblacional por percentiles (densidad_poblacional_percentil)

Tipo: Categórica ordinal. **Fuente:** Variable derivada de población y superficie territorial (INEGI, 2020).

La densidad poblacional continua se discretiza en cuartiles. Las categorías son: Baja, Media-baja, Media-alta y Alta.

Segregado por género

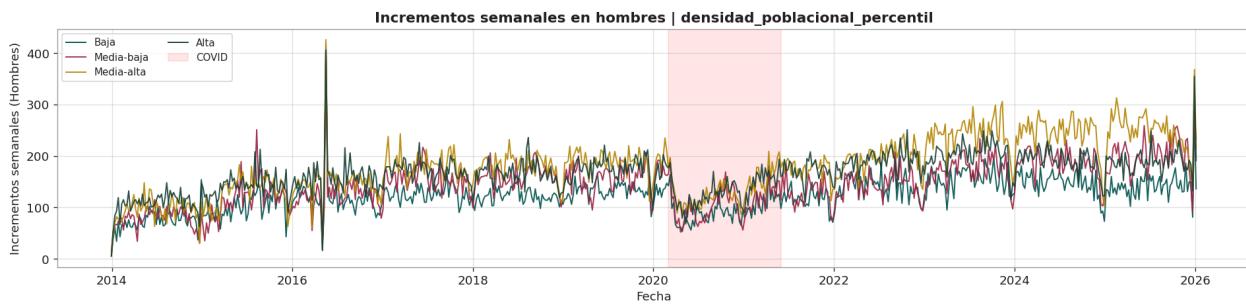


Figura 8: Densidad poblacional – Incrementos semanales (Hombres).

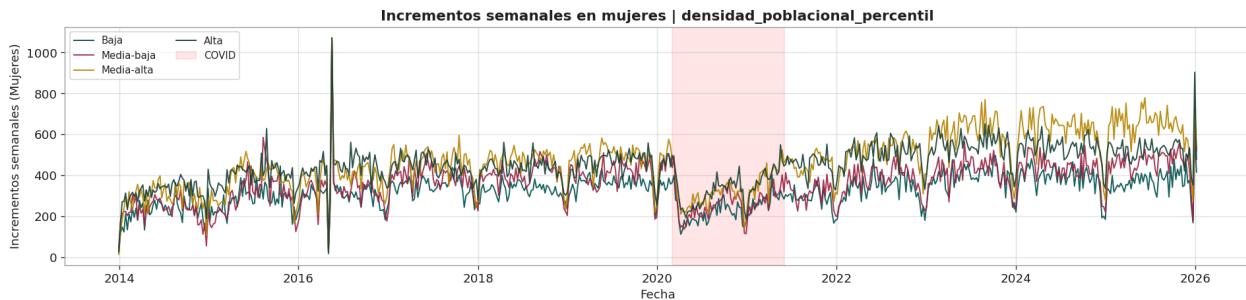


Figura 9: Densidad poblacional – Incrementos semanales (Mujeres).

Incrementos totales

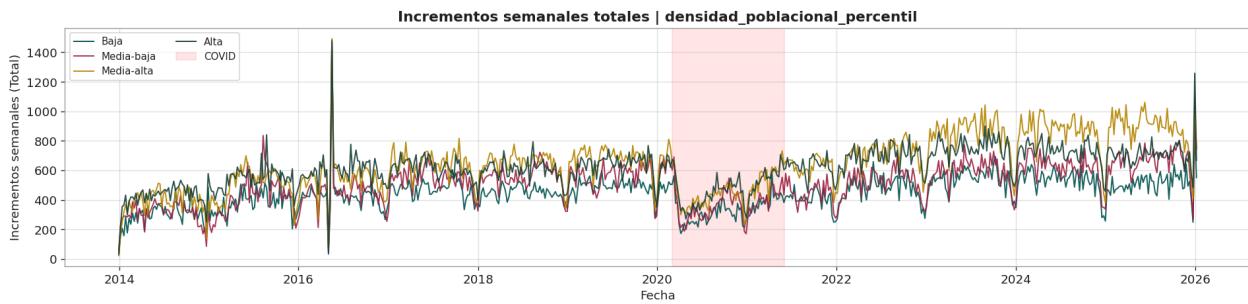


Figura 10: Densidad poblacional – Incrementos semanales totales.

Tamaño poblacional por rangos fijos (tamano_poblacional_predefinido)

Tipo: Categórica ordinal. **Fuente:** INEGI, población total (Censo 2020).

La población total se discretiza en rangos absolutos predefinidos: 0 a 1M, 1M a 3M, 3M a 6M y más de 6M habitantes.

Segregado por género

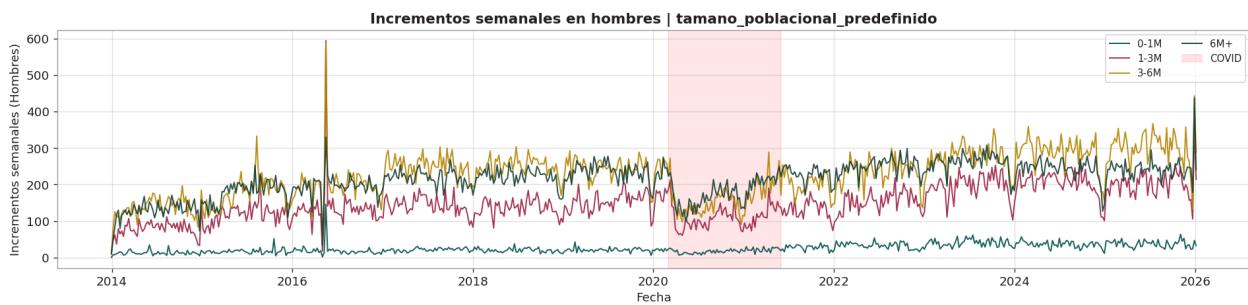


Figura 11: Tamaño poblacional (rangos fijos) – Incrementos semanales (Hombres).

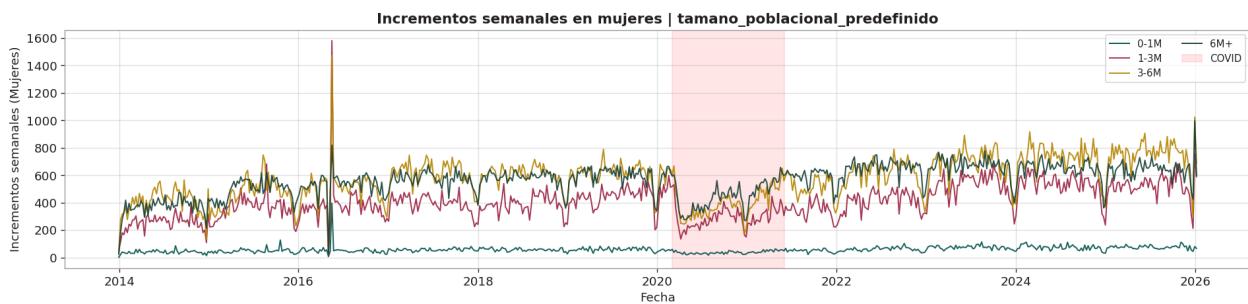


Figura 12: Tamaño poblacional (rangos fijos) – Incrementos semanales (Mujeres).

Incrementos totales

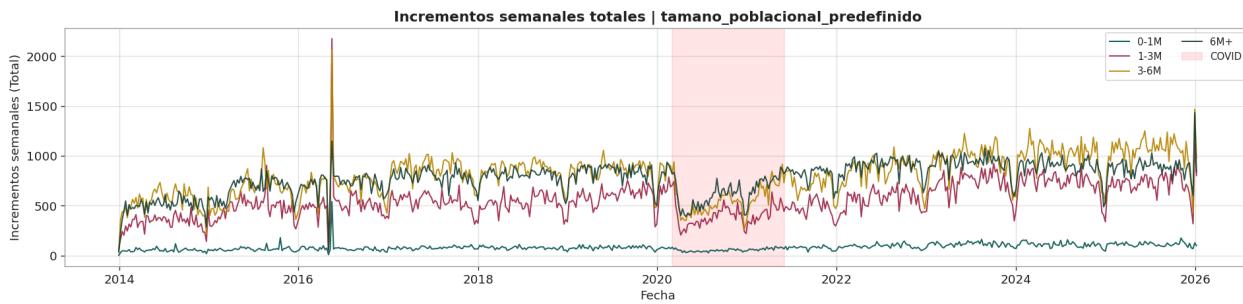


Figura 13: Tamaño poblacional (rangos fijos) – Incrementos semanales totales.

Tamaño poblacional por percentiles (tamano_poblacional_grupo_percentil)

Tipo: Categórica ordinal. **Fuente:** INEGI, población total (Censo 2020).

La población total se agrupa mediante discretización por cuantiles, generando cuatro grupos de tamaño similar: Población baja, Media-baja, Media-alta y Alta.

Segregado por género

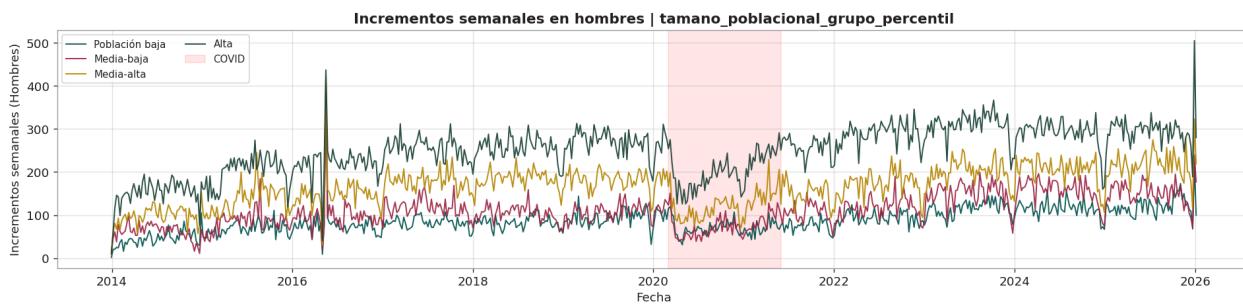


Figura 14: Tamaño poblacional (percentiles) – Incrementos semanales (Hombres).

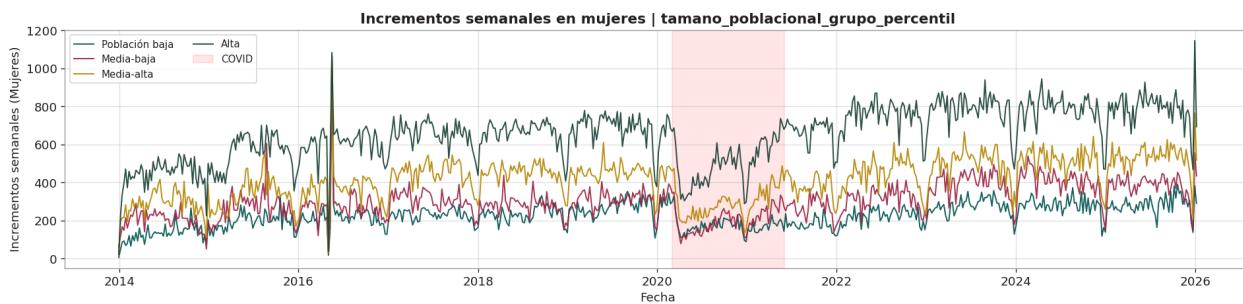


Figura 15: Tamaño poblacional (percentiles) – Incrementos semanales (Mujeres).

Incrementos totales

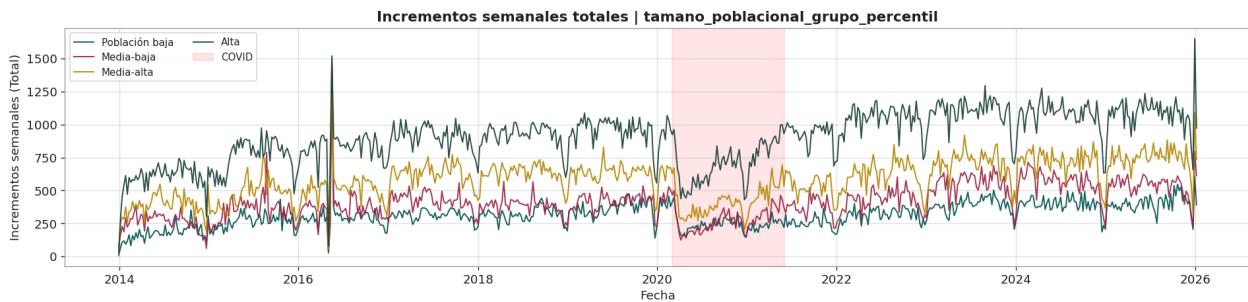


Figura 16: Tamaño poblacional (percentiles) – Incrementos semanales totales.

Extensión territorial por percentiles (extension_territorial_percentil)

Tipo: Categórica ordinal. **Fuente:** INEGI, superficie territorial (km^2).

La superficie territorial se discretiza en cuartiles: Territorio pequeño, Medio-pequeño, Medio-grande y Grande.

Segregado por género

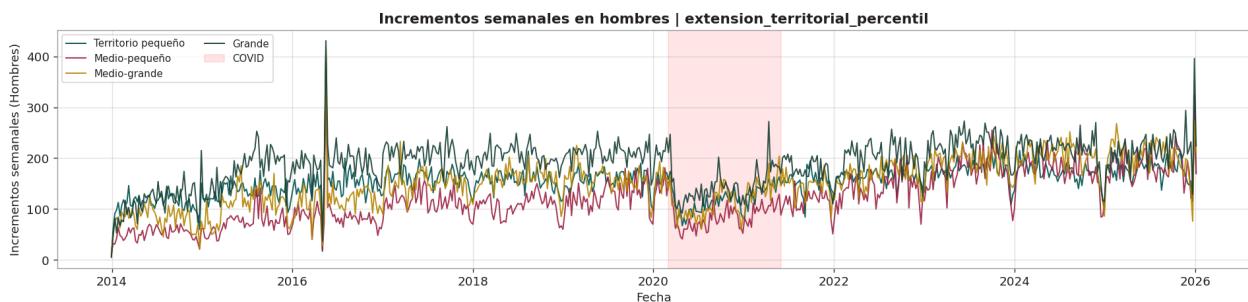


Figura 17: Extensión territorial – Incrementos semanales (Hombres).

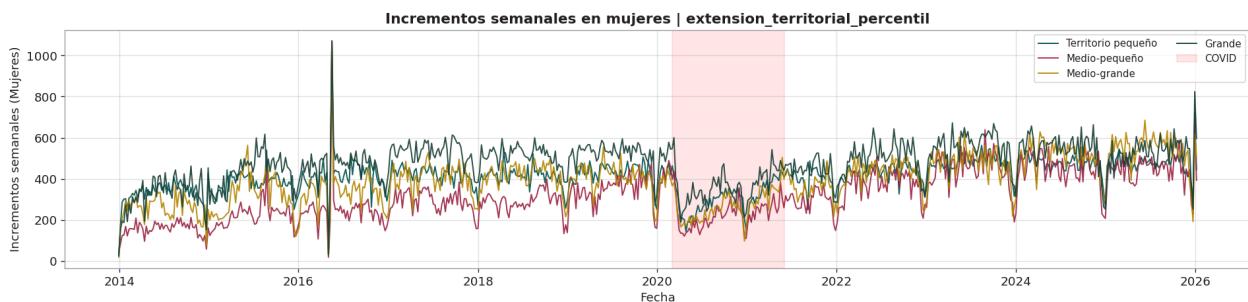


Figura 18: Extensión territorial – Incrementos semanales (Mujeres).

Incrementos totales

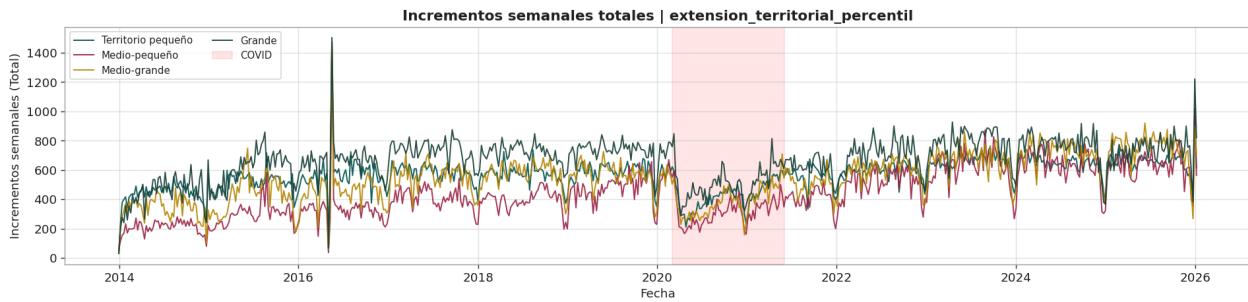


Figura 19: Extensión territorial – Incrementos semanales totales.

Ratio hombres/mujeres categorizado (ratio_h_m_cat)

Tipo: Categórica ordinal. **Fuente:** INEGI, población por sexo (Censo 2020).

Se calcula el ratio continuo entre población masculina y femenina, y se discretiza: menor a 0.99 (Mayormente mujeres), entre 0.99 y 1.01 (Balanceado) y mayor a 1.01 (Mayormente hombres).

Segregado por género

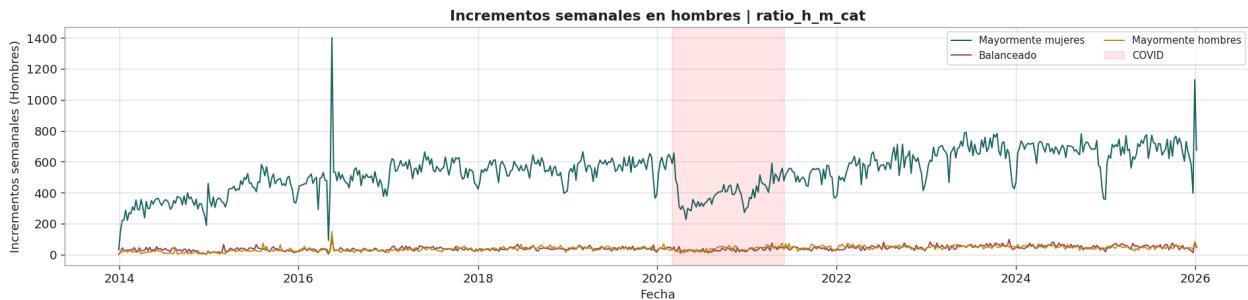


Figura 20: Ratio H/M – Incrementos semanales (Hombres).



Figura 21: Ratio H/M – Incrementos semanales (Mujeres).

Incrementos totales

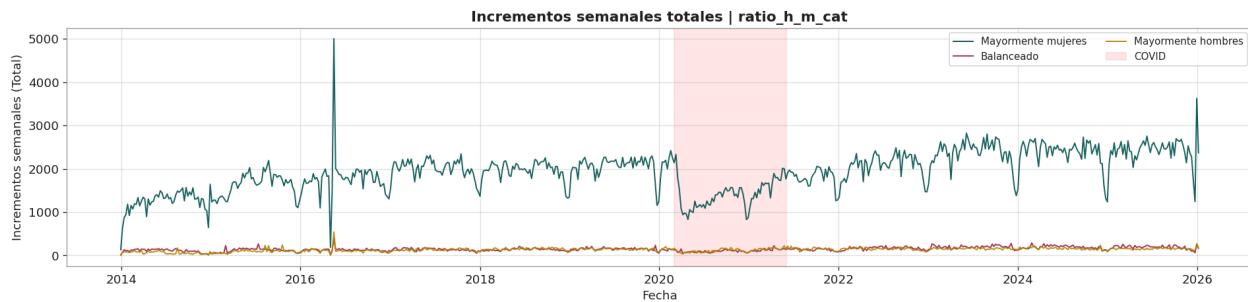


Figura 22: Ratio H/M – Incrementos semanales totales.

Visualización de datos

El pipeline implementado en el proyecto **EpiForecast-MX** se mantiene consistente a lo largo de todas las etapas del análisis. Como salida de este proceso, además de los datasets en formato CSV, se genera un archivo en formato Excel (**.xlsx**) que sirve como fuente de datos para la capa de visualización en **Tableau**.

Uso de Tableau Public

Para esta etapa se creó un *workbook* utilizando **Tableau Public**. Esta plataforma permite desarrollar visualizaciones interactivas y publicarlas directamente en la web. No obstante, Tableau Public presenta una limitante: no permite conectarse directamente a bases de datos ni a servicios de almacenamiento como S3. Se buscará la obtención de licencias completas de Tableau para una integración más robusta.

Tabla de datos consolidada

Esta vista presenta el dataset consolidado utilizado en el análisis. La tabla permite filtrar dinámicamente por año y padecimiento. Nota: la información correspondiente a 2026 es simulada y se incluye únicamente con fines informativos y exploratorios.

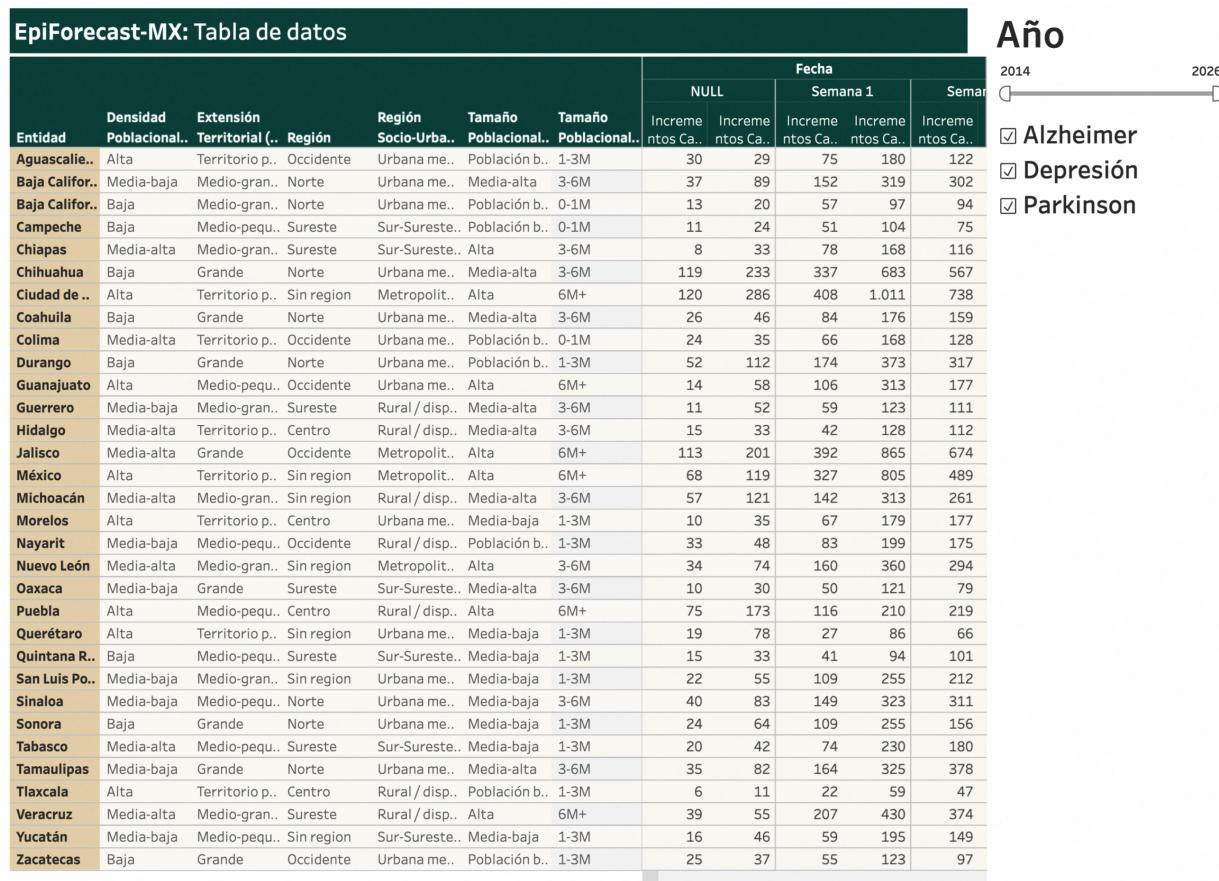


Figura 23: Tabla de datos consolidada en Tableau.

Mapa de México con variables demográficas

Representación espacial de México a nivel de entidad federativa, incorporando variables demográficas clave provenientes de INEGI. Mediante escalas de color se visualizan indicadores como la densidad poblacional y la superficie territorial, permitiendo contrastar entidades densamente pobladas frente a territorios extensos y dispersos. Adicionalmente, el mapa integra los casos epidemiológicos desagregados por sexo.

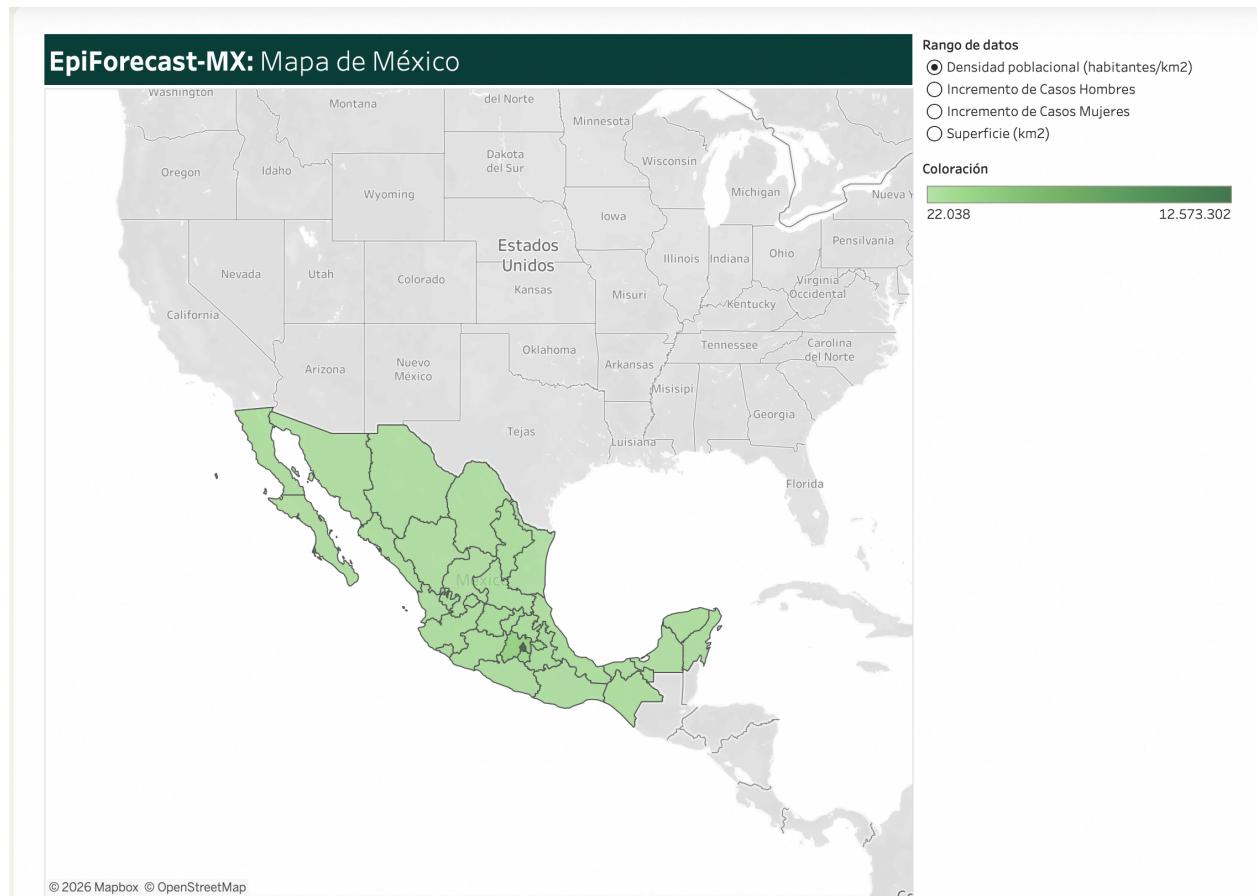


Figura 24: Mapa de México con variables demográficas.

Méjico en categorías

Mapa categórico del país, construido a partir de las clasificaciones territoriales y demográficas definidas durante la fase de ingeniería de características: región de salud mental, ratio H/M, extensión territorial, densidad poblacional y tamaño poblacional por rangos fijos y por percentiles.



Figura 25: México en categorías territoriales y demográficas.

Casos por año

Incremento anual de casos, desagregado por sexo. Permite filtrar por padecimiento y comparar tendencias entre distintos años, facilitando el análisis de cambios de mediano y largo plazo.

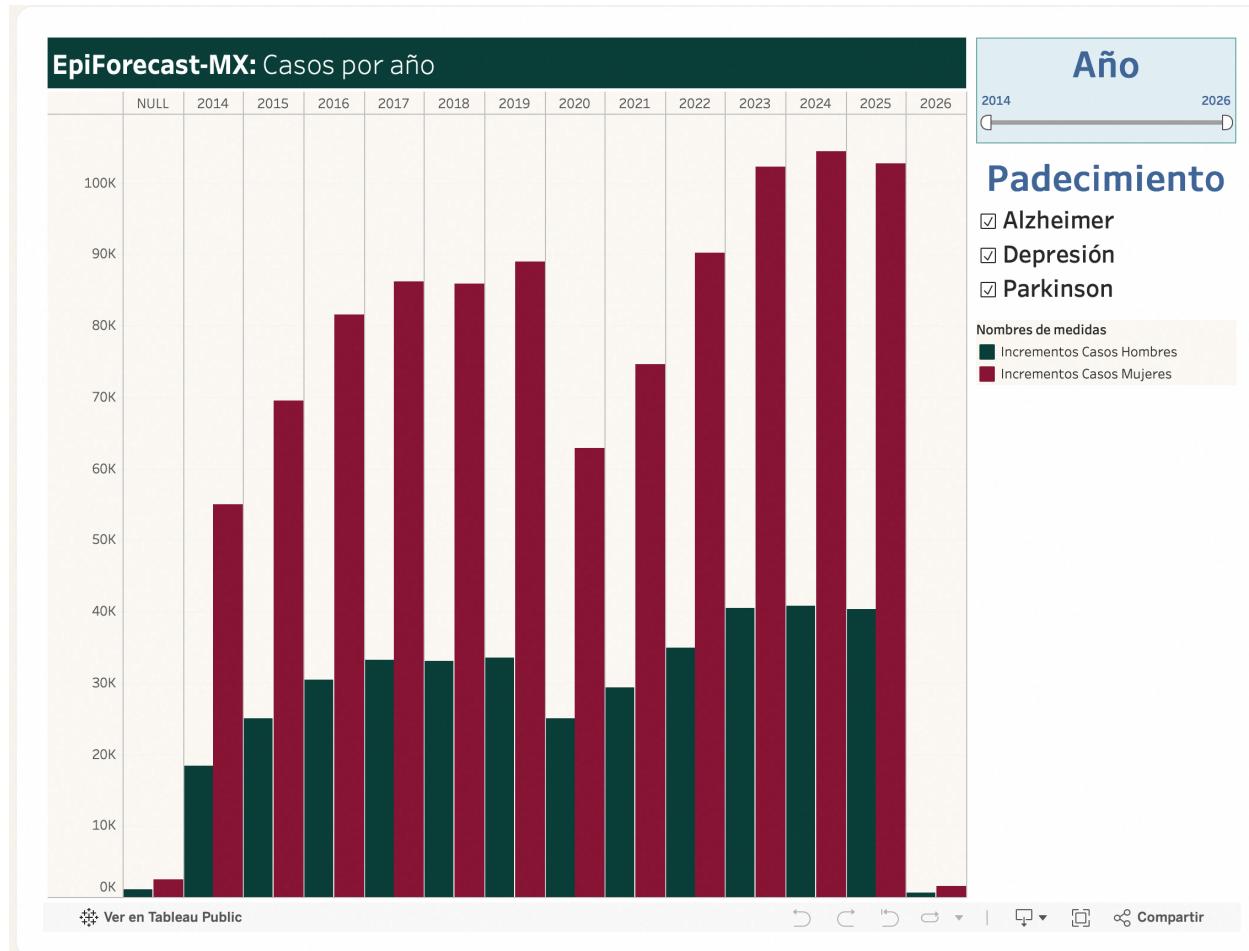


Figura 26: Casos por año, desagregados por sexo.

Casos semanales

Dinámica temporal de corto plazo, mostrando el incremento de casos semanales por sexo. La posibilidad de filtrar por año y padecimiento permite detectar variaciones intraanuales, picos, estacionalidad y cambios abruptos.

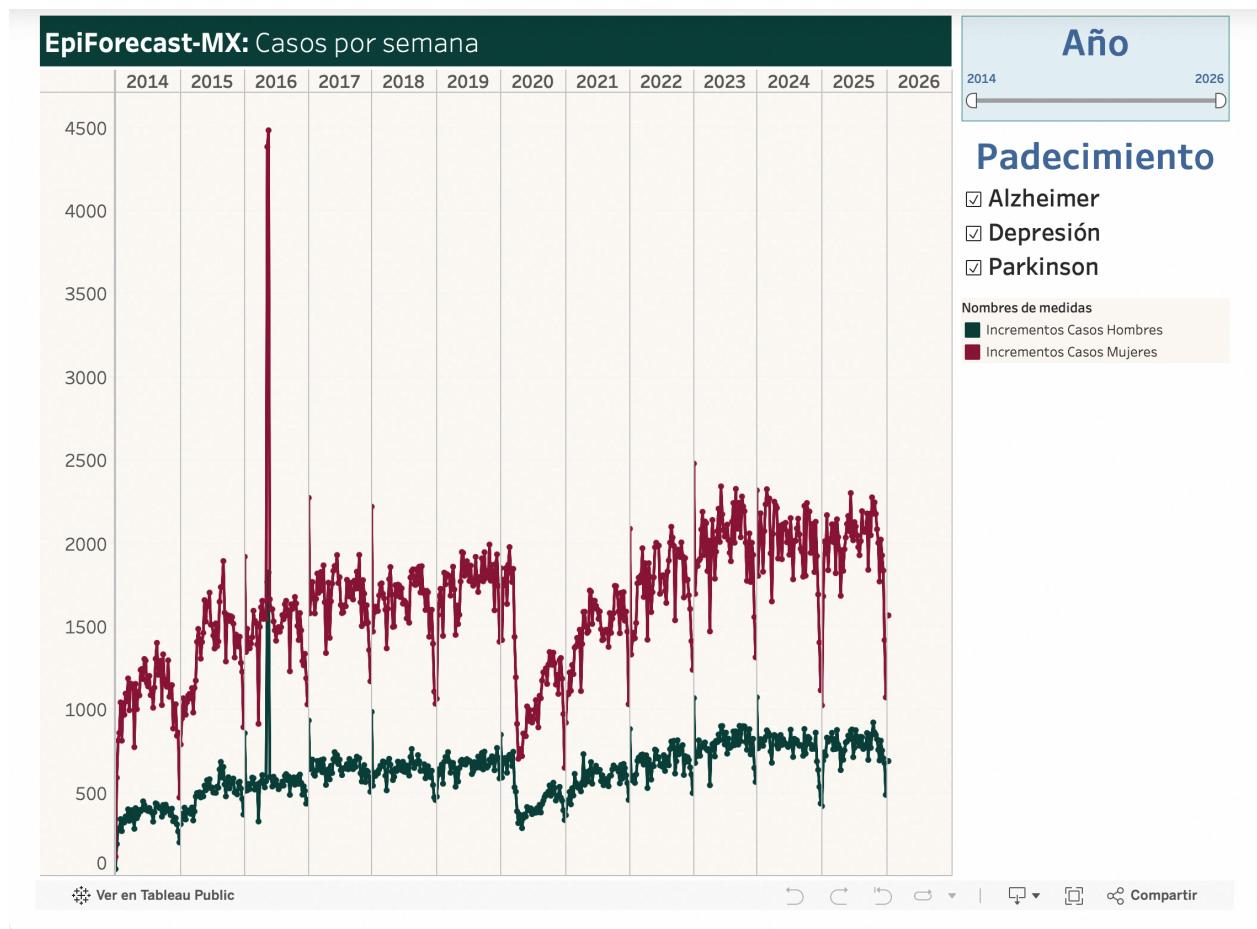


Figura 27: Incremento de casos semanales por sexo.

Predicciones epidemiológicas

Nota: las predicciones mostradas son simuladas en esta etapa.

Esta visualización combina los incrementos históricos de casos anuales por sexo con la predicción generada por el modelo seleccionado para cada padecimiento. Su propósito es exploratorio y demostrativo, y servirá como base para integrar predicciones validadas en entregas posteriores.

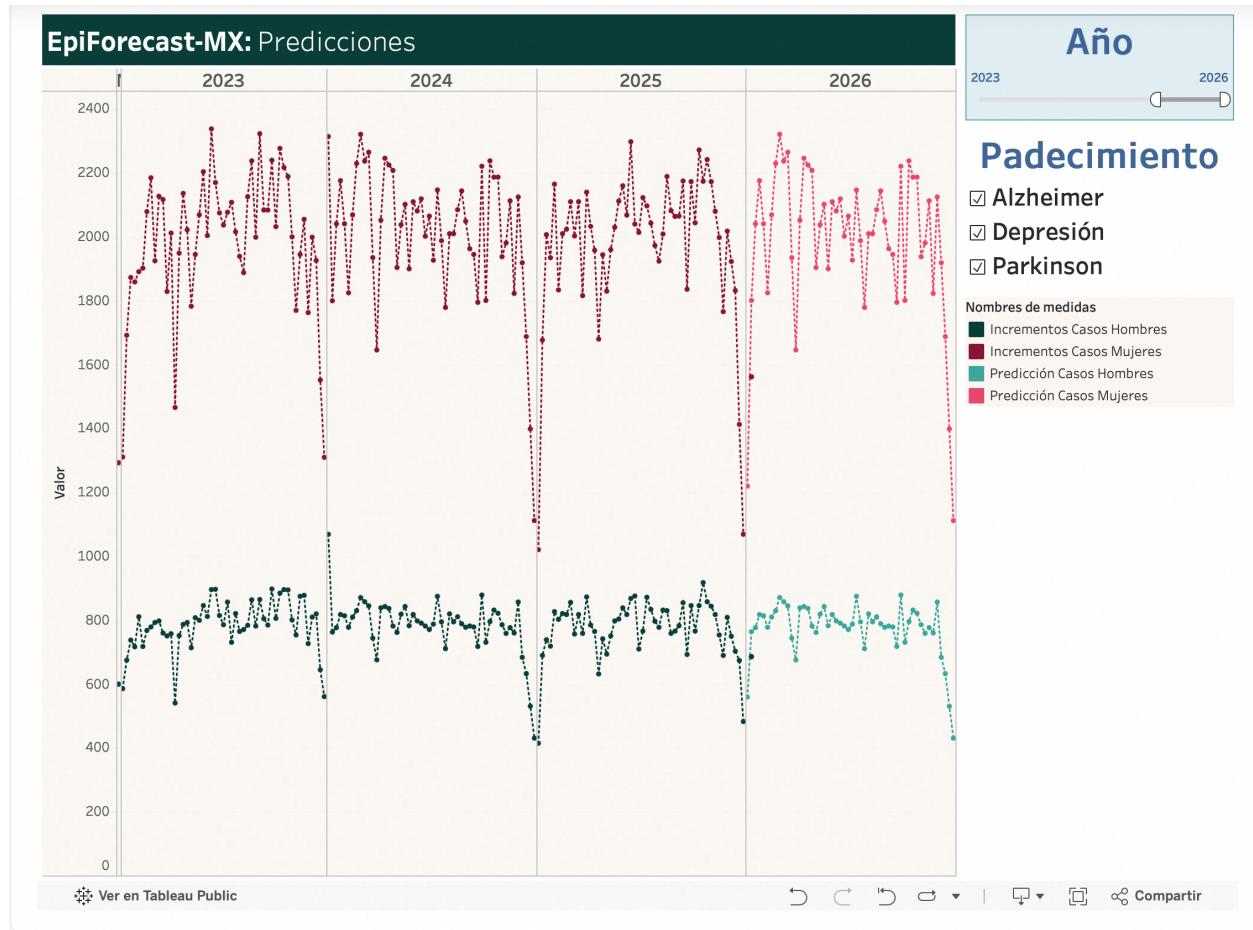


Figura 28: Predicciones epidemiológicas (simuladas).

Síntesis del Avance: Ingeniería de Características y Visualización

El trabajo desarrollado en esta etapa corresponde al **Avance 2: Ingeniería de Características**, y se enfoca en enriquecer estructuralmente el dataset epidemiológico mediante la integración de información demográfica y territorial proveniente de INEGI, así como en formalizar una primera capa de visualización interactiva del proyecto **EpiForecast-MX**.

Pipeline de Datos y Enriquecimiento

El pipeline general del proyecto se mantiene intacto y versionado en el repositorio. Sobre este pipeline se añadió una rama específica de ingeniería de características, cuyas etapas principales son:

1. **Consumo del dataset epidemiológico limpio:** Se parte del archivo `data_prepare_<tipo>.csv`, resultado directo del pipeline de limpieza y transformación del SINAVE.
2. **Extracción de datos INEGI:** Descarga programática de información poblacional y territorial desde la API PxWeb de INEGI, utilizando como referencia el Censo de Población y Vivienda 2020.
3. **Construcción del dataset demográfico base:** Población total, población por sexo, superficie territorial por entidad y validación interna (Hombres + Mujeres = Total).
4. **Ingeniería de características territoriales y demográficas:** Generación de variables continuas y categóricas (densidad poblacional, tamaño poblacional, extensión territorial, ratio H/M, región socio-territorial).
5. **Normalización y merge por entidad federativa:** Replicando la lógica del módulo `MapeaInegi`, se normalizan nombres de entidades y se realiza un *left join* entre el boletín epidemiológico y el dataset INEGI.
6. **Salida dual del pipeline:** CSV versionado para análisis y modelado, y archivo Excel como insumo para Tableau.

Hallazgos Clave de la Ingeniería de Características

Estructura territorial y demográfica. La densidad poblacional emerge como una de las variables más discriminantes entre entidades. Las clasificaciones por percentiles generan grupos balanceados, útiles para análisis comparativos. La mayoría de las entidades presentan un ratio hombres/mujeres cercano a 1.

Integración con epidemiología. El merge con INEGI no introduce pérdida de registros, confirmando consistencia en la variable entidad. Las nuevas categorías permiten analizar patrones epidemiológicos desde una perspectiva estructural, no solo temporal. La segmentación territorial revela heterogeneidades que no son visibles en agregados nacionales simples.

Contribución al Modelado Futuro

1. **Variables estructurales listas para usar:** Las categorías territoriales y demográficas pueden emplearse como variables exógenas, estratos de análisis o criterios de segmentación.
2. **Reducción de sesgos por escala:** Normalizar por población y densidad evita conclusiones dominadas por entidades grandes.
3. **Mejor interpretabilidad:** Las variables categóricas derivadas permiten explicar resultados a audiencias no técnicas.
4. **Compatibilidad con modelos temporales:** El dataset enriquecido conserva la estructura temporal necesaria para Prophet, ARIMA o modelos híbridos.

Limitaciones de esta Etapa

La regionalización de salud mental es una asignación estructural inicial, no completamente *data-driven*. La información demográfica es estática (Censo 2020) y no captura cambios poblacionales anuales. Las visualizaciones en Tableau Public no están conectadas a una fuente dinámica. Las predicciones mostradas son simuladas y no deben interpretarse como resultados finales.

Conclusiones Finales del Equipo

El desarrollo del Avance 2 consolida la transición del proyecto **EpiForecast-MX** desde la limpieza y exploración inicial de datos hacia una fase de preparación estructurada y orientada al modelado predictivo. A continuación se presentan las conclusiones colectivas del equipo respecto a esta etapa.

Enriquecimiento con impacto analítico real. La integración de variables demográficas y territoriales de INEGI no fue un ejercicio meramente acumulativo: cada variable añadida responde a una hipótesis sobre los determinantes estructurales de la incidencia epidemiológica en México. La densidad poblacional, la regionalización socio-urbana y el tamaño poblacional por rangos fijos demostraron ser las características con mayor potencial discriminante para estratificar patrones entre entidades federativas. Este enriquecimiento transforma el dataset de una colección de series temporales aisladas en un recurso multidimensional con contexto territorial.

Pipeline reproducible como activo del proyecto. El diseño modular del pipeline, desde la extracción automatizada de boletines hasta la generación del dataset enriquecido con salida dual (CSV y Excel), garantiza que cualquier miembro del equipo o evaluador pueda replicar íntegramente los resultados. La combinación de DVC para versionado de datos, GitHub Actions para automatización y scripts parametrizados refuerza el compromiso del equipo con las mejores prácticas de MLOps y con el marco metodológico CRISP-ML(Q).

Visualización como instrumento de validación y comunicación. Las visualizaciones desarrolladas, tanto en Python como en Tableau, cumplieron una doble función: sirvieron como herramienta de validación interna para detectar inconsistencias o confirmar patrones esperados, y como medio de comunicación hacia los stakeholders del IMSS. La segmentación por categorías territoriales en las series de tiempo reveló heterogeneidades regionales que permanecen ocultas en los agregados nacionales, reforzando la pertinencia de un enfoque desagregado para el pronóstico.

Base sólida para la fase de modelado. El dataset resultante de esta etapa está preparado para alimentar directamente los modelos de pronóstico planificados. Las variables categóricas derivadas podrán utilizarse como regresores exógenos en Prophet o como criterios de segmentación para entrenar modelos diferenciados por región o condición. La normalización por población y densidad mitiga sesgos de escala que, de no tratarse, distorsionarían las predicciones en favor de las entidades más pobladas.

Trabajo colaborativo y aprendizaje continuo. Esta entrega reflejó una dinámica de trabajo donde cada integrante aportó desde su experiencia profesional: la perspectiva de infraestructura de datos del IMSS, la visión de ingeniería de sistemas y la experiencia en operaciones de datos se integraron de forma orgánica. El equipo identificó también áreas de mejora, como la necesidad de una regionalización más robusta basada en datos y la

incorporación de fuentes dinámicas para actualizar las variables demográficas más allá del Censo 2020.

Próximos pasos. Con la ingeniería de características consolidada, el equipo se enfocará en la selección final de variables, la implementación de modelos de pronóstico basados en Prophet con regresores exógenos, y la evaluación de desempeño mediante intervalos de predicción. El objetivo inmediato es cerrar la brecha entre un dataset bien preparado y un sistema de pronóstico funcional que aporte valor estratégico al IMSS para la planificación de recursos en salud mental y neurología a nivel nacional y subnacional.

Reflexiones del Equipo



Juan Carlos Pérez — “*Desde el enfoque de datos, este avance marcó un punto de consolidación del proyecto. El pipeline permitió pasar de un dataset corregido a uno estructuralmente enriquecido, manteniendo consistencia, reproducibilidad y control sobre cada transformación. La ingeniería de características obligó a evaluar cuidadosamente qué variables realmente aportaban información y cuáles solo agregaban complejidad innecesaria. El uso de visualizaciones y series de tiempo fue clave para validar decisiones del pipeline, ya que permitió confirmar visualmente patrones, tendencias y posibles errores. Este avance reforzó la idea de que un pipeline sólido no solo prepara datos para modelar, sino que también funciona como una herramienta de validación continua del análisis.*”



Javier Rebull — “*Pienso que el principal aprendizaje de esta entrega fue pensar el proyecto como un sistema vivo y no como un análisis estático. El diseño del pipeline y de los scripts considera explícitamente la detección de nuevos boletines epidemiológicos y la actualización progresiva del dataset, reduciendo la dependencia de procesos manuales. Este enfoque facilita la escalabilidad del proyecto y garantiza que los resultados puedan mantenerse actualizados conforme se publique nueva información oficial. El trabajo realizado deja sentadas las bases para que el proyecto evolutive hacia un flujo completamente automatizado, alineado con buenas prácticas de ingeniería de datos y operación continua.*”



Luis Sánchez — “*Desde el punto de vista de la extracción y uso de datos de INEGI, este avance permitió integrar información demográfica y territorial de manera coherente con los datos epidemiológicos. La construcción de categorías, tanto continuas como discretizadas, ayudó a traducir variables complejas en indicadores interpretables y comparables. Asimismo, el trabajo en visualización, particularmente mediante Tableau, evidenció la importancia de contar con una capa clara de comunicación que permita explorar los datos sin perder rigor analítico. Este avance mostró que la correcta clasificación de variables y su representación visual son fundamentales para revelar patrones regionales y demográficos que no son evidentes en tablas o estadísticas agregadas y que nos permitirán llevar la predicción al siguiente nivel en etapas posteriores de modelado.*”

Referencias

1. Alegría, M., et al. (2018). *Social determinants of mental health*. Social Psychiatry and Psychiatric Epidemiology.
2. Brownlee, J. (2020). *How to Choose a Feature Selection Method for Machine Learning*. Machine Learning Mastery.
3. CONAPO (2020). *Índices de marginación por entidad federativa y municipio*. Consejo Nacional de Población.
4. CONEVAL (2020). *Índices de marginación por entidad federativa y municipio*.
5. Costa, R. (2022). *The CRISP-ML Methodology: A Step-by-Step Approach to Real-World Machine Learning Projects*.
6. Elliott, P., & Wartenberg, D. (2004). *Spatial epidemiology*. Environmental Health Perspectives, 112(9).
7. Galli, S. (2022). *Python Feature Engineering Cookbook* (2.^a ed.). Packt Publishing.
8. Hernández-Sampieri, R., & Mendoza, C. (2023). *Metodología de la investigación: Las rutas de la investigación cuantitativa, cualitativa y mixta* (3.^a ed.). McGraw-Hill.
9. Huang, C. Y., & Dai, H. L. (2021). *Learning from class-imbalanced data*. Data Science in Finance and Economics, 1(1).
10. INEGI (s. f.). *Áreas geográficas e indicadores territoriales*. Instituto Nacional de Estadística y Geografía.
11. INEGI (s. f.). *Localidades urbanas y rurales en México*. INEGI.
12. INEGI (s. f.). *Marco geoestadístico y datos poblacionales*. INEGI.
13. INEGI (2024). *PxWeb API: Población por entidad federativa y sexo*. INEGI.
14. Kumar Mukhiya, S., & Ahmed, U. (2020). *Hands-On Exploratory Data Analysis with Python*. Packt Publishing.
15. Organización Panamericana de la Salud (2021). *La salud mental en la región de las Américas*. OPS.
16. Organización Mundial de la Salud (2014). *Social determinants of mental health*. WHO.
17. Organización Mundial de la Salud (2022). *World mental health report: Transforming mental health for all*. WHO.
18. Organización Mundial de la Salud (2025). *ICD-11 2025 update*. WHO.
19. Pérez-Hernández, R. *Enfermedades neurológicas y trastornos mentales en México 2014–2024*.
20. Secretaría de Salud (s. f.). *Boletín Epidemiológico: SINAVE*. Gobierno de México.
21. Secretaría de Salud (s. f.). *Directorio de unidades de atención en salud mental*. Gobierno de México.

22. SINAVE (s. f.). *Sistema Nacional de Vigilancia Epidemiológica*. Secretaría de Salud.
23. Visengeriyeva, L., et al. (2023). *CRISP-ML(Q): The ML Lifecycle Process*. INNOQ / MLOps.