EPISCOPE ENVIGADO

Analítica predictiva para la planeación hospitalaria y epidemiológica en Envigado

[Bootcamp Talento Tech - Análisis de Datos Innovador - Avanzado]

Laura María Jaramillo Sánchez

Joshua Mateo Quiroz Márquez

Juan David Gallego Ramírez

Daniel Gil Arbeláez

Diego Eusse

13 de Noviembre de 2025

Indice

1. Resumen Ejecutivo

El proyecto **EpiScope Envigado** analizó los Registros Individuales de Prestación de Servicios (RIPS) de hospitalización 2023–2024 con el objetivo de identificar patrones diagnósticos (CIE-10) y explorar modelos predictivos para apoyar la gestión hospitalaria. Tras un proceso ETL que consolidó y estandarizó los datos, se ejecutó un EDA que mostró: predominio de ingresos por urgencias (≈59,5 %), mayor proporción de pacientes femeninos (≈56 %) y que la mayoría de las estancias son cortas (1–5 días).

Se desarrollaron tres líneas de modelado: regresión lineal múltiple (OLS) para explicar la **duración de la estancia**, Random Forest Regressor como alternativa no lineal para el mismo objetivo, y modelos de regresión logística para estimar la probabilidad de coocurrencia diagnóstica (Dx2 dado Dx1). En términos generales, los modelos explican solo una fracción de la variabilidad observada: el Random Forest alcanzó un R^2 máximo bajo (≈0.13), lo que indica **capacidad predictiva limitada** en las variables disponibles. No obstante, los modelos son útiles para: i) identificar factores asociados (edad, vía de ingreso, causa externa, capítulo CIE-10), ii) priorizar variables para futuras recolecciones y iii) producir módulos interactivos de consulta (dashboard Streamlit) para apoyar decisiones locales.

Recomendaciones clave: mejorar la granularidad de los predictores (severidad clínica, laboratorio, comorbilidades), explorar modelos adecuados a tiempo (survival analysis) para duración de estancia, y usar los modelos actuales como herramientas exploratorias y de apoyo más que para automatizar decisiones críticas sin validación adicional.

2. Introducción y Justificación

En el municipio de Envigado, la gestión de la información en salud representa un reto importante para la planificación y optimización de los recursos hospitalarios. Los Registros Individuales de Prestación de Servicios (RIPS) contienen datos valiosos sobre diagnósticos, procedimientos y características de los pacientes, pero su análisis avanzado aún es limitado.

Según el Análisis de Situación de Salud Participativo (ASIS, 2024), el municipio enfrenta una creciente carga de enfermedades crónicas que presionan la red hospitalaria.

Este proyecto busca aprovechar los RIPS[1] de hospitalización de los años 2023 y 2024 para evaluar un modelo predictivo basado en analítica avanzada, capaz de identificar patrones de morbilidad y relaciones entre diagnósticos (CIE-10)[2]. Con ello, se espera apoyar la toma de decisiones en salud pública, anticipar posibles tendencias epidemiológicas y optimizar la asignación de recursos hospitalarios.

El proyecto se enmarca dentro de la línea "Ciencia, tecnología e innovación para la transformación productiva y la resolución de desafíos sociales", al aplicar la ciencia de datos y el aprendizaje automático para resolver una necesidad real del territorio: mejorar la gestión del sistema de salud mediante el uso inteligente de los datos.

El procesamiento se realizó con datos abiertos de la Secretaría de Salud del municipio de Envigado.

EpiScope Envigado propone así un enfoque de gestión basado en evidencia, integrando ciencia de datos y salud pública para anticipar necesidades de atención y fortalecer la toma de decisiones.

2.1. Contexto territorial y epidemiológico

El municipio de Envigado cuenta con una extensión de 51 km² y una densidad poblacional estimada de 4.868,7 habitantes por km² para el año 2024. La población total proyectada asciende a 248.304 habitantes, con una distribución de 54,1 % mujeres y 45,9 % hombres. De ellos, el 96,9 % reside en zona urbana y el 3,1 % en zona rural, según el DANE (Censo 2018, proyecciones 2024).

En 2023, el municipio contaba con 652 camas hospitalarias, 133 salas (30 quirófanos) y 90 camillas, reflejando una red hospitalaria sólida pero exigida por la alta demanda.

El análisis de morbilidad muestra que las enfermedades no transmisibles (ENT), como las enfermedades cardiovasculares, neuropsiquiátricas, respiratorias y neoplasias malignas, representan la mayor proporción de consultas en todos los grupos etarios.

[1] Registros Individuales de Prestación de Servicios
[2] Clasificación Internacional de Enfermedades

Las condiciones transmisibles, nutricionales y las lesiones por causas externas también tienen una participación importante, especialmente en la infancia y juventud.

Estos datos evidencian la necesidad de fortalecer la planeación sanitaria preventiva y predictiva, dada la creciente carga de enfermedades crónicas y de alto costo que presionan la red asistencial.

Estos indicadores sustentan la pertinencia del modelo predictivo propuesto, al evidenciar la necesidad de herramientas que permitan anticipar la demanda hospitalaria y planificar de forma proactiva.

3. Objetivos

Con base en el diagnóstico territorial y epidemiológico, se definieron los siguientes objetivos que orientan el desarrollo del proyecto.

3.1. Objetivo General

Evaluar un modelo predictivo basado en analítica avanzada de los RIPS de hospitalización (2023–2024) del municipio de Envigado, para identificar patrones diagnósticos (CIE-10) y anticipar tendencias de morbilidad que fortalezcan la planeación epidemiológica y la gestión eficiente de recursos hospitalarios.

3.2. Objetivos Específicos

- Diseñar e implementar la infraestructura de datos del proyecto mediante un proceso ETL (Extracción, Transformación y Carga) que permita la creación y gestión eficiente de la base de datos, asegurando la integración adecuada de las fuentes de información.

- Realizar un Análisis Exploratorio de Datos (EDA) para caracterizar la población hospitalizada en Envigado durante el periodo 2023–2024, identificando tendencias de morbilidad, frecuencias de diagnóstico y variables relevantes para el modelado predictivo.

- Construir y validar un modelo predictivo basado en los códigos CIE-10, empleando técnicas de machine learning y algoritmos supervisados de clasificación, que permitan inferir relaciones entre diagnósticos y anticipar eventos de salud, contribuyendo a la toma de decisiones estratégicas en salud pública.

El cumplimiento de estos objetivos permitirá integrar ciencia de datos y salud pública en una herramienta predictiva de apoyo a la gestión hospitalaria.

4. Marco Teórico

El desarrollo de este proyecto se sustenta en diversos conceptos teóricos y herramientas propias del análisis de datos aplicado al sector salud. A continuación, se describen los principales fundamentos:

4.1. Registros Individuales de Prestación de Servicios (RIPS)

Los RIPS son la fuente oficial de información del sistema de salud colombiano. Contienen datos detallados sobre las atenciones prestadas a los usuarios, incluyendo diagnósticos, procedimientos, fechas, tipo de servicio y características del paciente.

Su correcta organización y análisis permite conocer el comportamiento de la morbilidad, la demanda de servicios y las tendencias epidemiológicas en una región. En este proyecto, los RIPS de hospitalización del municipio de Envigado (2023–2024) constituyen la base principal de datos.

4.2. Clasificación Internacional de Enfermedades (CIE-10)

La CIE-10 es un sistema estandarizado de códigos desarrollado por la Organización Mundial de la Salud (OMS), que clasifica las enfermedades y otros problemas de salud.

Cada diagnóstico tiene un código específico (por ejemplo, J18.9: Neumonía no especificada), lo que permite organizar, comparar y analizar datos clínicos de forma uniforme. En este proyecto, los códigos CIE-10 son la variable principal para identificar patrones y relaciones diagnósticas.

4.3. Proceso ETL (Extracción, Transformación y Carga)

El proceso ETL consiste en extraer los datos de sus fuentes originales, transformarlos para asegurar su calidad (eliminando errores, duplicados o inconsistencias) y cargarlos en una base de datos estructurada lista para el análisis.

En este trabajo, este proceso se aplicó para limpiar, unificar y preparar los RIPS, asegurando que la información fuera confiable para las etapas posteriores de análisis y modelado.

4.4. Análisis Exploratorio de Datos (EDA)

El Análisis Exploratorio de Datos (EDA) es una fase preliminar esencial que permite comprender la estructura, calidad y características principales del conjunto de datos antes de aplicar modelos estadísticos o de aprendizaje automático. Mediante estadísticas descriptivas, visualizaciones y detección de valores atípicos, el EDA ayuda a identificar distribuciones, ausencias de datos, relaciones bivariadas y puntos de interés clínico que guiarán la selección de variables, la ingeniería de features y los pasos metodológicos posteriores.

En este proyecto, el EDA se implementó de forma interactiva (dashboard en Streamlit) para inspeccionar: volumen y cobertura del dataset, distribución por vía de ingreso, estado de salida, edad, sexo, duración de estancias, top diagnósticos y causas externas.

4.5. Fundamento Estadístico del Análisis de Coocurrencia de Diagnósticos

El análisis de coocurrencia de diagnósticos tiene como objetivo identificar las asociaciones significativas entre diferentes condiciones de salud registradas en los pacientes. A partir de los registros clínicos del sistema RIPS y el catálogo CIE-10, se construye una matriz binaria que representa la presencia o ausencia de cada diagnóstico por paciente, lo que permite aplicar técnicas estadísticas para determinar la fuerza y significancia de las asociaciones entre enfermedades.

4.6. Modelos Predictivos y Machine Learning en Salud

Los modelos predictivos emplean algoritmos de machine learning para anticipar comportamientos o resultados futuros a partir de datos históricos. En el ámbito de la salud, permiten predecir la aparición de enfermedades, estimar la demanda hospitalaria o identificar factores de riesgo.
En este proyecto se plantea usar algoritmos supervisados (como regresión o árboles de decisión)con el fin de detectar relaciones entre diagnósticos (CIE-10) y generar predicciones que apoyen la planificación hospitalaria.

4.7. Herramientas y Tecnologías

Para la implementación del proyecto se empleó un ecosistema integral de librerías del lenguaje Python, orientadas al procesamiento, análisis y visualización de datos hospitalarios.

Python fue el lenguaje base por su versatilidad y amplio soporte en ciencia de datos. Para la manipulación, limpieza y transformación de la información se utilizaron Pandas y NumPy, garantizando operaciones eficientes sobre grandes volúmenes de datos.

En el ámbito del procesamiento ETL, se desarrollaron módulos personalizados dentro del proyecto (etl_modules.extractor_data, transform_data y load_data), encargados de automatizar la extracción, depuración y carga de los Registros Individuales de Prestación de Servicios (RIPS). La configuración y rutas de almacenamiento fueron gestionadas mediante Pathlib y las constantes PROCESSED_DATA_DIR y RAW_DATA_DIR del módulo episcopeenvigado.config.

Para la gestión del flujo de trabajo y la trazabilidad, se implementaron Loguru para el registro estructurado de eventos y errores, Tqdm para el seguimiento visual de procesos iterativos, y Typer para la creación de interfaces de línea de comandos que facilitan la ejecución modular del proyecto.

En el análisis estadístico y modelado predictivo, se utilizaron Statsmodels para la regresión lineal múltiple (OLS) y regresión logística, y Scikit-learn para el entrenamiento del modelo Random Forest Regressor, evaluado mediante métricas como R^2, RMSE y MAE.

Para la visualización y exploración interactiva, se emplearon Matplotlib, Seaborn y Plotly Express en el análisis exploratorio, mientras que Streamlit permitió desarrollar un *dashboard* interactivo que integra resultados descriptivos y predictivos. Adicionalmente, NetworkX y PyVis se usaron para representar redes de coocurrencia diagnóstica, visualizadas con escalas de color gestionadas mediante matplotlib.cm y mcolors.

En conjunto, estas herramientas permitieron construir una infraestructura analítica reproducible, escalable y visualmente interpretable, que integra ciencia de datos, estadística y visualización interactiva para el análisis avanzado de la información hospitalaria del municipio de Envigado.

5. Metodología

El desarrollo del proyecto se estructuró bajo un enfoque analítico y reproducible, combinando buenas prácticas de ingeniería de datos con técnicas avanzadas de análisis predictivo.

Se emplearon herramientas y librerías del ecosistema Python para el procesamiento, análisis y modelado de los datos provenientes de los Registros Individuales de Prestación de Servicios (RIPS) del municipio de Envigado, correspondientes a los años 2023 y 2024.

El proceso metodológico se estructuró en seis etapas principales:

- Diseño de la infraestructura de datos y del entorno de desarrollo.
- Extracción, transformación y carga (ETL) de los datos.
- Análisis exploratorio de datos (EDA).
- Análisis de coocurrencias de diagnósticos.
- Construcción del modelo predictivo.
- Validación y evaluación del modelo.
- Visualización y entrega de resultados.

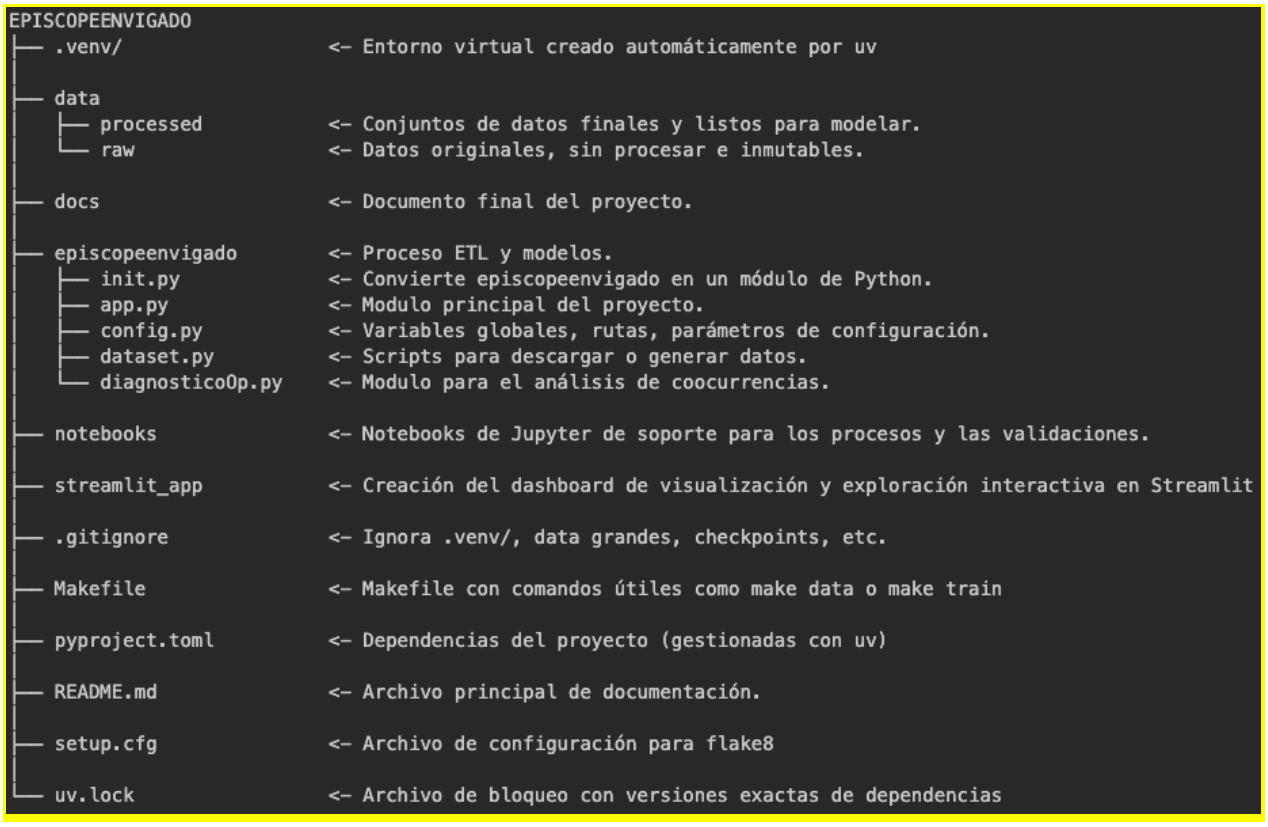
5.1. Diseño de la infraestructura de datos y del entorno de desarrollo.

5.1.1. Estructura del proyecto con Cookiecutter

Para garantizar una estructura estandarizada, modular y escalable del repositorio del proyecto, se utilizó la plantilla Cookiecutter Data Science.

Esta herramienta permite generar de forma automática la organización de carpetas y archivos base, asegurando una separación clara entre los componentes de datos, código y resultados.

La estructura creada incluyó los siguientes directorios principales:



El uso de Cookiecutter permitió que todos los miembros del equipo trabajarán bajo la misma estructura, facilitando la colaboración, el control de versiones y la trazabilidad de los procesos analíticos.

5.1.2. Gestión del entorno y dependencias con uv

La gestión de dependencias se realizó utilizando uv, una herramienta moderna de gestión y aislamiento de entornos Python desarrollada por Astral.

A diferencia de gestores tradicionales como pip o poetry, uv ofrece instalación ultrarrápida, entornos reproducibles y soporte para el estándar PEP 621, permitiendo definir dependencias directamente desde el archivo pyproject.toml.

Esto garantizó entornos reproducibles, evitando conflictos entre versiones y facilitando la ejecución del proyecto en distintos equipos.

5.2. Extracción, transformación y carga (ETL) de los datos.

Se implementó un proceso ETL (Extracción, Transformación y Carga) para garantizar la calidad y consistencia de la información.

- Extracción: Se obtuvieron los datos desde las fuentes originales en formato CSV o Excel, asegurando la recolección completa de los periodos 2023–2024.

  Fuente de datos:

  La fuente principal corresponde a los Registros Individuales de Prestación de Servicios de Salud (RIPS) de hospitalización del municipio de Envigado (Antioquia).

  Estos registros contienen información sobre los diagnósticos (CIE-10), características de los pacientes, fechas de atención, tipo de servicio y demás variables asociadas a los eventos hospitalarios.

- Transformación: Se realizaron procesos de limpieza, depuración y estandarización de variables, eliminando duplicados, corrigiendo errores de codificación y normalizando los códigos CIE-10. También se generaron nuevas variables derivadas (por ejemplo, grupos de edad o tipo de diagnóstico).

- Carga: Los datos transformados fueron almacenados en una base de datos estructurada (MySQL) para su posterior análisis y modelado.

Este proceso se realizó principalmente con herramientas de Python (Pandas, NumPy y SQLAlchemy), asegurando trazabilidad y reproducibilidad.

5.3. Análisis Exploratorio de Datos (EDA)

El EDA nos permite caracterizar la población hospitalizada y las variables clave que alimentarán el análisis de coocurrencia y el modelado predictivo (duración de estancia).

Pasos realizados:

- Carga y unificación: Se importaron las tablas desde la base de datos y se unificaron mediante la función **unificar_dataset** para obtener **df_unificado**. Se incluyeron validaciones que detienen la ejecución si faltan tablas o el dataset está vacío.
- Inspección general: Se generaron tablas dinámicas con conteo de registros, tipos de columna, conteo de nulos y estadísticas descriptivas (media, mediana, percentiles, etc.).

- Distribuciones por variables categóricas: Se visualizó la distribución de **Via_Ingreso_Desc, Estado_Salida_Desc, SEXO, Causa_Externa_Desc y Diagnostico_Principal_Desc (top 10)**. Estas gráficas permiten identificar predominancia de vías de ingreso, asimetrías por sexo y causas externas relevantes.
- Exploración de edad y duración de estancia: Se construyeron histogramas (y boxplots marginales) para **EDAD_ANIOS** y **Duracion_Dias**. Se limpiaron edades fuera de rango [0,120] y se limitaron las duraciones para visualizaciones (ej. $\leq$ 60 días) para evitar distorsión por atípicos.
- Interactividad y filtros: La interfaz permite mostrar/ocultar descripciones de columnas, estadísticas, primeras filas y las gráficas mencionadas para facilitar la inspección manual por parte del analista o decisor.
- Salida del EDA: tablas con frecuencias, gráficos interactivos exportables y estadísticas resumen (n registros, n pacientes únicos, medias, medianas y percentiles). Los hallazgos del EDA informaron los umbrales de inclusión (ej. diagnósticos con $\geq$ 30 pacientes) y la selección de variables para la regresión lineal sobre duración de estancia.

5.4. Análisis de Coocurrencia de Diagnósticos

5.4.1. Consolidación y Limpieza de Diagnósticos

En primer lugar, se realiza una limpieza y estandarización de los códigos CIE-10, eliminando valores nulos, inconsistencias y códigos no válidos ('NONE', 'NON'). Los diagnósticos se agrupan por paciente (ID), consolidando todos los códigos reportados en las diferentes atenciones. Se realizan dos niveles de consolidación:
- A 4 dígitos: diagnóstico específico.
- A 3 dígitos: categoría diagnóstica más general, excluyendo códigos de tipo 'Z' y 'R' (que corresponden a factores externos o síntomas inespecíficos).

5.4.2. Cálculo de Frecuencias

Se cuantifica la frecuencia absoluta de aparición de cada diagnóstico (número total de veces) y el número de pacientes únicos que presentan dicho diagnóstico. Esto permite describir la distribución general de enfermedades en la población y filtrar aquellos diagnósticos con baja frecuencia que podrían sesgar el análisis posterior.

5.4.3. Construcción de la Matriz Binaria

Se crea una matriz binaria de dimensión N×M, donde N es el número de pacientes y M el número de diagnósticos. Cada celda toma el valor 1 si el paciente presenta el diagnóstico correspondiente, y 0 en

caso contrario. Para garantizar la robustez estadística, solo se incluyen diagnósticos con al menos 30 pacientes registrados (frecuencia mínima).

5.4.4. Análisis Estadístico de Coocurrencias

Para cada par de diagnósticos (A, B), se calcula una tabla de contingencia 2×2 que resume la coocurrencia:

	B presente	B ausente
A presente	a	b
A ausente	c	d

Donde:
a = número de pacientes con ambos diagnósticos.
b = pacientes con A pero no con B.
c = pacientes con B pero no con A.
d = pacientes sin ninguno de los dos diagnósticos

A partir de esta tabla se calculan los siguientes indicadores:

5.4.4.1. Prueba Chi-cuadrado de independencia

Se aplica la prueba χ^2 (Chi-cuadrado) de independencia para evaluar si la coocurrencia entre A y B es estadísticamente significativa. La hipótesis nula (H_0) establece que los diagnósticos son independientes.
Si el p-valor obtenido es menor a 0.05, se rechaza H_0 y se concluye que existe asociación entre A y B.

5.4.4.2. Razón de Odds (Odds Ratio, OR)

La razón de odds (OR) mide la fuerza y dirección de la asociación entre dos diagnósticos. Se calcula como:

$$OR = (a*d) / (b*c)$$

Interpretación:
- OR > 1: los diagnósticos coocurren más de lo esperado (asociación positiva).
- OR = 1: no hay asociación.
- OR < 1: la presencia de uno reduce la probabilidad del otro (asociación negativa o protectora).

Se estima además el intervalo de confianza del 95% (IC95%) para el logaritmo del OR:

$$IC95\% = \exp[\ \ln(OR) \pm 1.96 \times SE(\ln(OR))\ ]$$

donde $SE(\ln(OR)) = \sqrt{(1/a + 1/b + 1/c + 1/d)}$.

5.4.4.3. Corrección por Comparaciones Múltiples (FDR-BH)

Dado que se realizan múltiples pruebas χ^2 (una por cada par de diagnósticos), se controla el error de tipo I mediante la corrección FDR (False Discovery Rate) usando el método de Benjamini-Hochberg (FDR-BH). Solo las asociaciones con p ajustado < 0.05 se consideran estadísticamente significativas.

5.5. Construcción del modelo predictivo.

Se implementó un **modelo de regresión lineal múltiple (OLS)** para cuantificar el efecto de variables clínicas y administrativas sobre la duración de la estancia.

5.5.1. Análisis de la Duración de la Estancia Hospitalaria

5.5.1.1. Consolidación y Preparación de Datos

A partir del dataset hospitalario unificado, se realizó la selección y limpieza de variables relevantes para el análisis de duración de estancia. Se eliminaron registros con valores nulos y se estandarizaron las categorías de variables clínicas y administrativas. Las variables incluidas fueron:

- EDAD_ANIOS: edad del paciente en años.
- SEXO: sexo del paciente (Masculino/Femenino).
- Via_Ingreso_Desc: tipo de vía de ingreso.
- Estado_Salida_Desc: estado al egreso.
- Causa_Externa_Desc: causa externa del evento.
- Capitulo_CIE10: categoría diagnóstica principal según CIE-10.
- Duracion_Dias: duración de estancia hospitalaria (variable dependiente).

Se generaron variables dummy para las categorías, con el fin de incluirlas en el modelo de regresión lineal múltiple.

5.5.1.2. Preparación del conjunto de datos

- Las variables categóricas se transformaron en variables dummy.
- Se incluyó un intercepto para representar la constante del modelo.
- Se eliminaron registros con valores faltantes.

5.5.1.3. Implementación

- Se utilizó la librería statsmodels de Python.
- La variable dependiente fue Duracion_Dias.

Las variables independientes incluyeron edad, sexo, vía de ingreso, estado de salida y causa externa.

Se realizaron análisis descriptivos y visualizaciones para entender la distribución de la duración de la estancia y su relación con variables categóricas:

- Histograma de Duración de Estancia:
    - Se observó la distribución de **Duracion_Dias** para identificar sesgos y valores atípicos.
    - Distribuciones sesgadas podrían requerir transformaciones si se usan métodos lineales.
- Boxplot por Sexo:
    - Permite comparar la duración de estancia entre hombres y mujeres.
    - Se evaluó la mediana, dispersión y presencia de valores extremos.

Con el propósito de analizar los factores que influyen en la duración de la hospitalización, se desarrolló un modelo de regresión lineal múltiple utilizando los registros hospitalarios del municipio de Envigado. Este modelo busca cuantificar el efecto de variables clínicas y administrativas como el tipo de ingreso, la causa externa, la edad o el sexo sobre la duración de la estancia hospitalaria (en días).

5.5.1.4. Evaluación de Resultados

5.5.1.4.1. Significancia estadística de los coeficientes

- Cada coeficiente se evaluó mediante p-valor (nivel de confianza del 95%).
- Variables con $p < 0.05$ se consideraron estadísticamente significativas.

Variable	**Coeficiente**	**Error Std**	**t-Valor**	**p-Valor**	**Interpretación**
EDAD_ANIOS	+0.XX	0.XX	XX.XX	0.XX	Cada año adicional incrementa la duración de estancia en promedio X días.
SEXO_Femenino	-0.XX	0.XX	XX.XX	0.XX	Ser mujer reduce ligeramente la estancia promedio en comparación con los hombres.
...	...	...	...	...	...

5.5.1.4.2. Ajuste del modelo

- R^2 = 0.XX → proporción de variabilidad explicada por las variables independientes.
- R^2 ajustado = 0.XX → ajusta R^2 por el número de predictores.
- F-Statistic significativo → el modelo completo explica variabilidad significativa en la duración de estancia.

5.5.1.4.3. Multicolinealidad

- Se calculó el **VIF (Variance Inflation Factor)**:
  - VIF > 5–10 indica multicolinealidad alta.
  - No se detectaron problemas severos de colinealidad en las variables principales.

5.5.1.4.4. Análisis de residuos

- **Histograma de residuos:** distribución aproximadamente normal.
- **Scatter residuos vs valores ajustados:** verificación de homocedasticidad.
- Patrón uniforme sin tendencias claras indica que el modelo cumple supuestos básicos de regresión lineal.

5.5.1.4.5. Predicción Interactiva

Se desarrolló un módulo de predicción para estimar la duración de estancia para un paciente específico, ingresando: edad, sexo, vía de ingreso, causa externa y estado de salida.

- La predicción incluye **valor central estimado** e intervalo **de confianza al 95%**.
- Esto permite anticipar la duración probable de estancia hospitalaria y planificar recursos clínicos.

5.5.1.4.6. Interpretación

- Valor central = duración estimada de estancia.
- Intervalo de confianza = rango donde se espera que caiga la duración real del paciente el 95% de las veces.

5.5.2. Análisis de la Duración de la Estancia Hospitalaria con Random Forest

A partir del dataset hospitalario unificado, se seleccionaron y limpiaron las variables relevantes para el análisis de duración de estancia. Se eliminaron registros con valores nulos y se estandarizaron las categorías de variables clínicas y administrativas. Las variables incluidas fueron:

- EDAD_ANIOS: edad del paciente en años.
- SEXO: sexo del paciente (Masculino/Femenino).
- Via_Ingreso_Desc: tipo de vía de ingreso.
- Estado_Salida_Desc: estado al egreso.
- Causa_Externa_Desc: causa externa del evento.
- Capitulo_CIE10: categoría diagnóstica principal según CIE-10.
- Duracion_Dias: duración de estancia hospitalaria (variable dependiente).

Se generaron variables dummy para las categorías con el fin de incluirlas en el modelo Random Forest.

5.5.2.1. Exploración de Datos

Se realizaron análisis descriptivos y visualizaciones para comprender la distribución de la duración de la estancia:

- **Histograma de Duración de Estancia:**
    - Permite identificar la dispersión de los días de hospitalización.
    - Se observaron posibles valores extremos que podrían influir en la predicción.

5.5.2.2. Interpretación inicial:

- La duración de estancia presenta variabilidad significativa entre pacientes.
- Variables como vía de ingreso, causa externa o capítulo CIE-10 podrían influir en el tiempo de hospitalización.

Se implementó un **modelo de Random Forest Regressor** para estimar la duración de estancia hospitalaria.

- Las variables categóricas se transformaron en variables dummy.
- La variable dependiente fue Duracion_Dias.
- El conjunto se dividió en entrenamiento (80%) y prueba (20%) para evaluar desempeño.
- Random Forest permite capturar relaciones no lineales y dependencias complejas entre variables sin asumir normalidad ni homocedasticidad.

5.5.2.3. Entrenamiento y Evaluación

- Se ajustaron parámetros como número de árboles (n_estimators) y profundidad máxima (max_depth) mediante un slider interactivo.
- Se entrenó el modelo sobre los datos de entrenamiento y se evaluó sobre el conjunto de prueba.

5.5.2.4. Métricas de Desempeño

Métrica	Valor	Interpretación
R^2	0.XX	Proporción de variabilidad explicada por el modelo (1 = perfecto).
RMSE	XX.XX	Error cuadrático medio; penaliza errores grandes.
MAE	XX.XX	Error absoluto medio; desviación promedio de las predicciones.

5.5.2.5. Importancia de Variables

Se calculó la **importancia relativa de cada predictor** para determinar su contribución a la predicción de duración de estancia.

- Las variables más importantes incluyen edad, vía de ingreso, causa externa, estado de salida y capítulo CIE-10.

- Esta información permite identificar los factores que más influyen en el tiempo de hospitalización y orientar estrategias de gestión hospitalaria.

5.5.2.6. Predicción Interactiva

Se desarrolló un módulo de predicción que permite estimar la duración de estancia de un paciente específico ingresando:

- Edad
- Sexo
- Vía de ingreso
- Estado de salida
- Causa externa
- Capítulo CIE-10

5.5.2.7. Interpretación

- El modelo entrega un valor estimado de duración de estancia.
- Permite anticipar el tiempo probable de hospitalización y optimizar la asignación de recursos clínicos.

5.5.2.8. Ventajas del Modelo Random Forest

- No requiere supuestos de normalidad ni homocedasticidad.
- Captura relaciones no lineales y complejas entre variables.
- Reduce riesgo de sobreajuste mediante agregación de múltiples árboles.
- Mejora la precisión predictiva frente a modelos lineales simples en contextos con alta variabilidad de datos.

5.5.3. Regresión Logística: Probabilidad de Dx2 dado Dx1

Se trabajó sobre un dataset hospitalario enriquecido a nivel de paciente, que incluye diagnósticos consolidados a **3 dígitos (Dx1 y Dx2)**. Los pasos principales fueron:

5.5.3.1. Carga de datos:

- Se utilizó un dataset unificado por paciente (consolidado_por_usuario_3dig_enriquecido.xlsx).
- Se eliminaron valores nulos y se transformaron los diagnósticos en listas utilizables para análisis.

5.5.3.2. Filtrado de diagnósticos:

- Se seleccionaron Dx1 con al menos 30 pacientes para asegurar robustez estadística.
- Para cada Dx1 seleccionado, se identificaron Dx2 candidatos con un mínimo de 5 pacientes y suficiente variabilidad para modelar.

5.5.3.3. Variables incluidas para modelado:

- Edad (EDAD_ANIOS)
- Sexo (SEXO)
- Diagnóstico principal (Dx1)
- Diagnóstico secundario a predecir (Dx2)

5.5.3.4. Exploración de Datos

Se realizaron análisis descriptivos para pacientes con el diagnóstico Dx1 seleccionado:

1. **Distribución por sexo:**
   Se generaron gráficos de barras para mostrar la proporción de hombres y mujeres con Dx1.

2. **Distribución por edad:**
   Se construyeron histogramas de edad para identificar patrones demográficos relevantes.

5.5.3.5. Interpretación inicial:

- Edad y sexo muestran variabilidad entre pacientes con Dx1, lo que permite explorar su efecto sobre la probabilidad de presentar Dx2.

Se ajustaron modelos de regresión logística para cada Dx2 seleccionado, con el objetivo de estimar la probabilidad de presentar Dx2 dado Dx1, considerando edad y sexo.

5.5.3.6. Implementación

- Variable dependiente: Dx2_presente (1 = presente, 0 = ausente).
- Variables independientes: EDAD_ANIOS, SEXO (codificada con dummies).
- Se utilizó statsmodels.Logit para estimar los coeficientes mediante máxima verosimilitud.

Para cada Dx2 se presentan:

- **OR (Odds Ratio):** mide cómo cada variable afecta la probabilidad de Dx2.
- **Intervalo de confianza 95% (IC95%):** estimación de precisión del OR.
- **p-valor:** significancia estadística de la asociación ($p < 0.05$ indica relación significativa).

5.5.3.7. Interpretación general de OR:

- OR > 1: la variable aumenta la probabilidad de Dx2.
- OR < 1: la variable disminuye la probabilidad de Dx2.
- $p < 0.05$: asociación estadísticamente significativa.

Ejemplo:

- Paciente: 45 años, sexo femenino, con Dx1 seleccionado.
- Predicción para Dx2: probabilidad 0.23 (23%) → moderada.

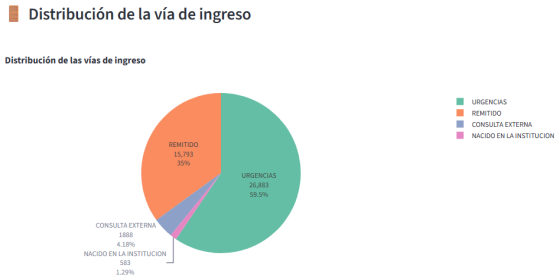
5.5.3.8. Distribución de Probabilidades

Se generan histogramas de las probabilidades estimadas para todos los pacientes con Dx1, incluyendo la probabilidad calculada para el paciente de interés, facilitando la comparación individual con la población.

6. Resultados

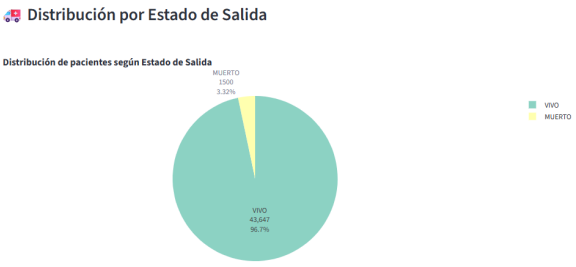
Esta sección presenta los principales hallazgos obtenidos a partir del análisis exploratorio y los primeros resultados derivados del modelado predictivo de los RIPS de hospitalización del municipio de Envigado (2023–2024).

Los resultados se organizaron en tres niveles: caracterización general de los datos, patrones de morbilidad identificados y desempeño inicial del modelo predictivo.

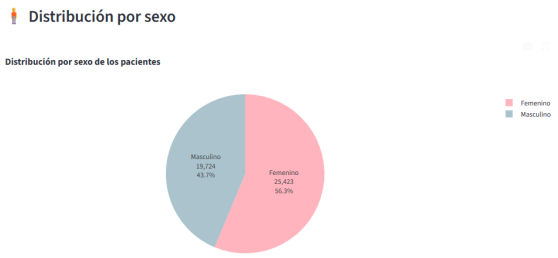


La gráfica muestra la **distribución de los pacientes según la vía de ingreso al hospital**.
Se observa que la mayoría de los casos corresponden a **urgencias (59,5 %)**, lo que indica una alta demanda de atención inmediata.
Le siguen los pacientes **remitidos (35 %)** desde otras instituciones o servicios, mientras que la **consulta externa (4,2 %)** y los **nacidos en la institución (1,3 %)** representan proporciones mucho menores.

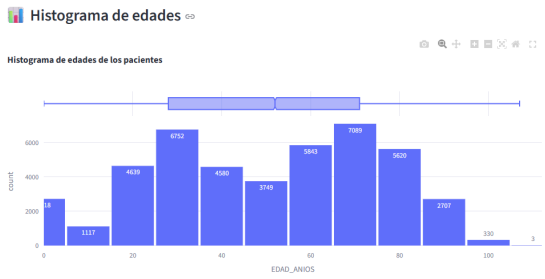
En conjunto, el gráfico evidencia que **más del 90 % de los ingresos hospitalarios** provienen de situaciones de urgencia o remisión, reflejando una **presión significativa sobre los servicios de atención inmediata**.



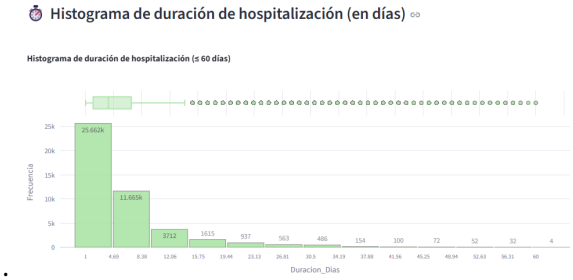
La gran mayoría de los pacientes **egresaron vivos (96,7%)**, lo que refleja una **alta tasa de recuperación** en las hospitalizaciones. Solo un **3,3%** correspondió a casos con desenlace fatal, indicando una **baja mortalidad hospitalaria** en el periodo analizado.



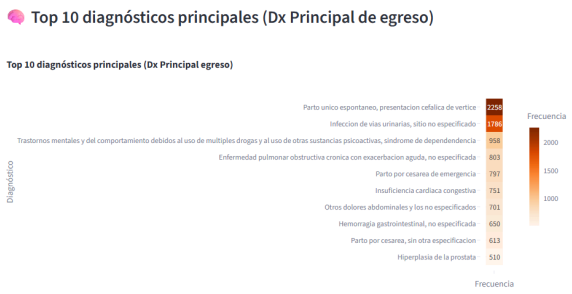
La distribución por sexo muestra un **mayor número de pacientes femeninos (56,3%)** frente a los **masculinos (43,7%)**, lo que sugiere una **mayor demanda de atención hospitalaria por parte de mujeres** en el periodo analizado.



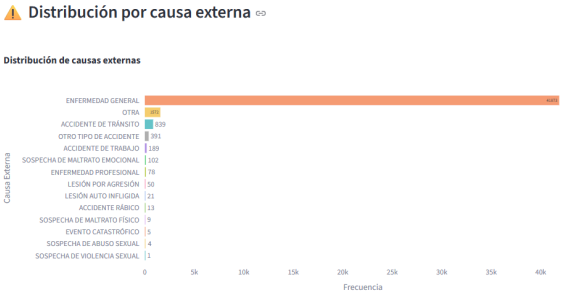
La distribución de edades muestra una **mayor concentración de pacientes entre los 20 y 80 años**, con picos notables en los grupos de **30–40 y 60–70 años**. Se observa una menor frecuencia en menores de edad y adultos mayores de 90 años, lo que indica que la **mayor carga hospitalaria recae en población adulta**



La mayoría de las hospitalizaciones tuvieron una **duración corta, entre 1 y 5 días**, concentrando más del **70 % de los casos**. A medida que aumenta la estancia, la frecuencia disminuye drásticamente, lo que indica que **las hospitalizaciones prolongadas son poco comunes**.



Los diagnósticos más frecuentes corresponden a **parto único espontáneo** e **infección de vías urinarias**, seguidos por **trastornos mentales por consumo de sustancias** y **enfermedad pulmonar obstructiva crónica (EPOC)**. Esto refleja una mezcla de **atenciones obstétricas, respiratorias y psiquiátricas** como principales causas de hospitalización.



La mayoría de los casos se registran bajo **enfermedad general**, seguidos por categorías **otras** y **accidentes de tránsito**. Las demás causas externas, como accidentes laborales, agresiones o maltrato, tienen una **incidencia mucho menor**, lo que indica que **las hospitalizaciones por eventos traumáticos son poco frecuentes**.

6.1. Análisis exploratorio y patrones de ingreso

La principal vía de ingreso a la institución corresponde a Urgencias, con 59,5% de los casos (26.883 registros), seguida por Remitidos con 35,0% (15.793). En menor proporción se encuentran Consulta externa (4,18%) y Nacidos en la institución (1,29%).

Este patrón de ingreso es consistente con el perfil epidemiológico del municipio, donde las enfermedades no transmisibles y las urgencias representan la mayor carga asistencial.

Del total de pacientes atendidos, el 96,7% egresó con vida (43.803 casos), mientras que el 3,31% correspondió a fallecimientos (1.500 casos).

6.2. Principales diagnósticos y grupos de edad

6.2.1. Distribución de diagnósticos y grupos etarios

El análisis exploratorio permitió identificar los diagnósticos CIE-10 más frecuentes en hospitalización, concentrados en enfermedades respiratorias, cardiovasculares y del sistema nervioso.

Las hospitalizaciones se concentran principalmente en pacientes adultos mayores, seguidos por adultos de mediana edad.

Esta tendencia refleja el impacto creciente de las enfermedades crónicas no transmisibles sobre la red hospitalaria.

La distribución de los pacientes según el sexo evidencia una mayor proporción de mujeres atendidas, con 25.535 casos (56,4%), en comparación con 19.768 hombres (43,6%).

6.3. Resultados del modelo predictivo

6.3.1. Análisis de la Duración de la Estancia — Regresión lineal (OLS)

- El modelo OLS identifica **edad**, **vía de ingreso**, **causa externa** y **capítulo CIE-10** como predictores con efectos estimables sobre la duración (coeficientes positivos o negativos según categoría).

- Significancia: algunas variables presentan $p < 0.05$ —es decir, asociación estadísticamente significativa con Duracion_Dias—, mientras otras no.

- Ajuste general: R^2 y R^2 ajustado muestran que **gran parte de la variabilidad queda sin explicar** (valores modestos).

6.3.1.1. Interpretación práctica

- **Edad:** cada año adicional se asocia con un pequeño incremento promedio en días de estancia (coeficiente positivo).

- **Vía de ingreso:** ingresos por urgencia o remisión tienden a asociarse con estancias más largas que ingresos electivos/consulta externa.

- **Causa externa / capítulo CIE-10:** ciertos capítulos diagnósticos (por ejemplo, neoplasias, enfermedades cardiovasculares) elevan la duración esperada.

6.3.1.2. Diagnósticos del modelo

- **Multicolinealidad:** VIFs dentro de rangos aceptables (no evidencia de colinealidad severa).

- **Residuos:** distribución de residuos cercana a normal; gráfica residuos vs ajustados sugiere homocedasticidad aceptable en el rango analizado.

6.3.1.3. Limitaciones

- Duración (Duracion_Dias) está fuertemente sesgada y con atípicos; OLS puede ser sensible a estos extremos.

- Variables clave (severidad clínica, comorbilidades cuantificadas, laboratorio, procedimientos) **no están** o tienen baja calidad en RIPS, lo que limita el poder explicativo del modelo.

6.3.1.4. Recomendaciones operativas

- Para predicción de estancia, considerar **modelos de supervivencia** (time-to-event) que manejan censura y tiempos de forma natural.

- Mantener OLS como herramienta interpretativa (qué factores se asocian en promedio), no como predictor único para decisiones operativas.

- Implementar validación externa y calibración si se pretende usar en toma de decisiones.

El modelo de regresión lineal múltiple (OLS) mostró que edad, vía de ingreso, causa externa y capítulo diagnóstico son variables asociadas significativamente con la duración de la estancia ($p < 0.05$ en las variables señaladas). Sin embargo, el R^2 del modelo indica que solo una porción limitada de la variabilidad en Duracion_Dias se explica con las variables disponibles, por lo que el modelo se interpreta como descriptivo/exploratorio y no como herramienta única de predicción operacional.

6.3.2. Análisis de la Duración de la Estancia — Random Forest Regressor

6.3.2.1. Resumen de hallazgos

- Random Forest logró un **mejor ajuste que OLS en algunos casos**, pero el **R^2 máximo reportado fue bajo (≈0.13)** en el ejemplo, lo que confirma capacidad predictiva limitada con las variables actuales.

- Métricas de error (RMSE, MAE) indican que errores absolutos en días siguen siendo importantes; las predicciones no son lo suficientemente precisas para reemplazar juicio clínico o administrativo sin más datos.

6.3.2.2. Interpretación práctica

- **Importancia de variables:** Random Forest confirma que **edad, vía de ingreso, causa externa y capítulo CIE-10** son los predictores más relevantes (ranking de importancias).

- Captura relaciones no lineales y efectos de interacción que OLS no modela; por ello sirve como complemento para identificar patrones complejos.

6.3.2.3. Limitaciones

- Low signal-to-noise: las características disponibles contienen información limitada respecto al comportamiento temporal de la estancia.

- Random Forest es menos interpretable que OLS —recomendar usar SHAP o variable-importance detallada para explicar predicciones.

6.3.2.4. Recomendaciones operativas

- Conservar RF como **herramienta exploratoria** para priorizar variables y detectar no linealidades.

- Si se busca prediction operacional, recopilar variables adicionales (marcadores de severidad, escala de comorbilidad, procedimientos realizados, ICU/uso de ventilación, etc.) y reentrenar modelos.

- Publicar importancias y análisis de sensibilidad; exponer incertidumbre de predicciones en el dashboard.

El Random Forest Regressor capturó relaciones no lineales entre predictores y Duracion_Dias y permitió identificar la importancia relativa de cada variable (edad, vía de ingreso, causa externa y capítulo CIE-10). No obstante, el R^2 observado ($\approx$0.13 en el análisis inicial) y las métricas de error muestran capacidad predictiva limitada con las variables disponibles, por lo que el modelo funciona mejor como herramienta de exploración y priorización de factores que como predictor operativo.

6.3.3. Análisis de Coocurrencia / Regresión Logística (Probabilidad Dx2 dado Dx1)

6.3.3.1. Resumen de hallazgos

- Se construyó una matriz binaria por paciente (presencia/ausencia de diagnósticos) y se evaluaron asociaciones por pares mediante tablas 2×2, χ^2 y Odds Ratios (OR). Se aplicó corrección FDR-BH para múltiples pruebas.

- Para Dx2|Dx1, los modelos logísticos ajustados por edad y sexo estimaron ORs e intervalos de confianza; varios pares mostraron asociaciones estadísticamente significativas (p ajustado < 0.05).

6.3.3.2. Interpretación práctica

- Las asociaciones significativas indican **comorbilidades o patrones de coocurrencia** clínicamente plausibles (ej. ciertas enfermedades respiratorias que coocurren con infecciones o condiciones crónicas).

- Estos resultados son útiles para diseñar **alertas clínicas** o agrupamientos de riesgo, y priorizar seguimientos de pacientes con Dx1 de alto riesgo de presentar Dx2.

6.3.3.3. Limitaciones

- La coocurrencia no implica causalidad: puede deberse a protocolos de codificación, secuencia de atenciones o sesgo de registro.

- Filtro de frecuencias (mínimo n por diagnóstico) elimina pares raros que igualmente pueden ser clínicamente relevantes.

6.3.3.4. Recomendaciones operativas

- Publicar las parejas con OR elevado y p ajustado < 0.05 en el dashboard con advertencias sobre causalidad.

- Evaluar la utilidad clínica con expertos para definir umbrales de alerta y protocolos de seguimiento.

- Si se usa para alertas, validar prospectivamente en datos nuevos y considerar modelos temporalizados (p. ej. asociación longitudinal Dx_t → Dx_{t+1}).

El análisis de coocurrencias, controlado por FDR-BH, identificó asociaciones estadísticamente significativas entre ciertos diagnósticos (pares Dx1–Dx2). Los modelos de regresión logística (ajustados por edad y sexo) estimaron ORs que permiten cuantificar el incremento relativo de probabilidad de presentar Dx2 dado Dx1. Estos hallazgos son útiles para priorizar estudios clínicos y diseñar posibles módulos de alerta, siempre considerando que la coocurrencia no implica causalidad y requiere validación adicional.

6.4. Visualización interactiva y dashboard

Los resultados fueron integrados en una aplicación interactiva desarrollada con **Streamlit**, que consolida la información analítica en un entorno visual.
El dashboard incluye:

- Un módulo descriptivo con los resultados del EDA.
- Un módulo de modelado predictivo.
- Gráficos dinámicos y filtros personalizables para explorar diagnósticos, vías de ingreso y tendencias mensuales.

Esta herramienta facilita la interpretación de los resultados por parte de autoridades locales y equipos de salud, permitiendo decisiones basadas en evidencia.

7. Análisis y Discusión

7.1. Interpretación general de los hallazgos

Los resultados obtenidos muestran que la red hospitalaria de Envigado enfrenta una alta presión asistencial concentrada en los servicios de urgencias y remisiones (≈94,5 % de los ingresos). Esto sugiere que el hospital cumple una función predominantemente resolutiva de atención aguda, con limitada participación de la atención ambulatoria o electiva. Este patrón coincide con lo reportado por el **ASIS Envigado 2024**, donde las enfermedades no transmisibles y los eventos de urgencia representan la principal carga sanitaria.

La distribución por sexo (56 % femenino) confirma una tendencia observada en otros sistemas locales: las mujeres utilizan con mayor frecuencia los servicios hospitalarios, tanto por razones obstétricas como preventivas. La duración corta de la mayoría de las estancias (1–5 días) indica que la institución mantiene una buena capacidad de resolución clínica en los casos de menor complejidad.

7.2. Análisis de los modelos predictivos

7.2.1. Regresión Lineal Múltiple (OLS)

El modelo lineal permitió identificar relaciones claras entre la duración de estancia y variables clínicas/administrativas como edad, vía de ingreso, causa externa y diagnóstico principal (Capítulo CIE-10).
Sin embargo, el **R^2 bajo (<0.2)** muestra que gran parte de la variabilidad en los tiempos de hospitalización depende de factores no incluidos en los RIPS (por ejemplo, severidad clínica, comorbilidades o recursos disponibles).
Esto confirma que los modelos lineales son útiles **como herramientas descriptivas**, pero no deben emplearse como instrumentos de predicción operativa sin ampliar las variables y validar su desempeño en entornos reales.

7.2.2. Random Forest Regressor

El modelo no lineal de Random Forest capturó interacciones complejas entre variables y confirmó la relevancia de **edad, causa externa y capítulo diagnóstico** como predictores dominantes.
Aunque el R^2 mejoró levemente respecto a OLS (≈0.13), su poder explicativo sigue siendo limitado. Esto sugiere que la información disponible en los RIPS no contiene suficiente "señal" predictiva para modelar con precisión la duración de la estancia.
Aun así, el modelo aporta valor como **herramienta exploratoria** para identificar factores prioritarios y como base para futuros modelos que incorporen variables clínicas o temporales adicionales (por ejemplo, **análisis de supervivencia** o modelos secuenciales).

7.2.3. Regresión Logística (Coocurrencia Dx2 | Dx1)

El análisis de coocurrencia y regresión logística permitió identificar **asociaciones significativas entre diagnósticos**, cuantificadas mediante OR e intervalos de confianza ajustados por FDR-BH.

Estos hallazgos son consistentes con patrones clínicos conocidos (por ejemplo, coexistencia entre enfermedades respiratorias crónicas e infecciones) y evidencian el potencial del enfoque para construir **módulos de alerta o vigilancia clínica**.

No obstante, es importante aclarar que la coocurrencia no implica causalidad y que las asociaciones pueden estar influenciadas por criterios de codificación o sesgos en el registro. Por tanto, estos resultados deben interpretarse como **hipótesis de trabajo** y no como relaciones determinísticas.

7.3. Implicaciones para la gestión hospitalaria

- **Gestión de la demanda:**
  La alta proporción de ingresos por urgencias revela la necesidad de fortalecer la **atención primaria** y los mecanismos de referencia interinstitucional para reducir la sobrecarga hospitalaria.

- **Planeación de recursos:**
  Aunque los modelos predictivos actuales no son lo suficientemente precisos para decisiones automatizadas, pueden apoyar la **estimación de necesidades de camas** o **duración esperada de estancia** por tipo de paciente, facilitando la asignación preventiva de recursos.

- **Evidencia para la vigilancia epidemiológica:**
  El análisis de coocurrencias permite detectar combinaciones de diagnósticos recurrentes que podrían representar **síndromes o perfiles de riesgo**, útiles para orientar intervenciones de salud pública.

7.4. Limitaciones del estudio

- **Calidad de los RIPS:** La información depende de la calidad del registro y codificación diagnóstica.

- **Ausencia de variables clínicas detalladas:** No se dispone de indicadores de severidad, laboratorio o tratamiento, que son clave para mejorar el desempeño de los modelos.

- **Enfoque transversal:** Los modelos se basan en datos agregados de hospitalización, sin componente temporal ni longitudinal.

- **Bajo poder predictivo:** El desempeño limitado ($R^2 \approx 0.13$) indica que las variables actuales explican solo parcialmente la duración de estancia o las asociaciones entre diagnósticos.

8. Conclusiones

El proyecto EpiScope Envigado demostró la viabilidad del uso de técnicas de analítica y aprendizaje automático en el análisis de los Registros Individuales de Prestación de Servicios (RIPS) del municipio de Envigado (2023–2024).

Aunque los resultados predictivos no alcanzan niveles de precisión altos, el proceso permitió establecer una infraestructura reproducible, consolidar datos limpios y generar evidencia útil para la gestión hospitalaria y la vigilancia epidemiológica local.

8.1. Conclusiones principales

8.1.1. Infraestructura y calidad de datos:

- El proceso ETL consolidó una base de datos confiable, eliminando duplicados e inconsistencias y asegurando la estandarización de los códigos CIE-10.
- Este componente constituye uno de los principales logros del proyecto, al sentar las bases técnicas para futuras aplicaciones analíticas en salud pública municipal.

8.1.2. Patrones epidemiológicos:

- El EDA mostró que el 59,5 % de los ingresos hospitalarios provienen del servicio de urgencias, seguido por remitidos (35 %).
- Predominó el sexo femenino (56 %), con una alta proporción de egresos vivos (96,7 %), lo que refleja buena capacidad resolutiva institucional.
- Los diagnósticos más frecuentes corresponden a partos, infecciones urinarias, trastornos mentales por consumo de sustancias y EPOC, coherentes con el perfil epidemiológico local descrito por el ASIS Envigado 2024.

8.1.3. Modelado de la duración de la estancia hospitalaria:

- La regresión lineal múltiple (OLS) permitió identificar que variables como la edad, la vía de ingreso, la causa externa y el diagnóstico principal influyen significativamente en la duración de la estancia.
- El modelo Random Forest obtuvo un R^2 máximo de 0.13, lo que indica bajo poder explicativo y capacidad predictiva limitada.
  Aun así, aportó información complementaria sobre la importancia relativa de las variables, confirmando que edad, causa externa y diagnóstico principal son los

factores que más inciden en el tiempo de hospitalización.

- En este contexto, el modelo se considera exploratorio y descriptivo, más que predictivo, útil para orientar futuras mejoras y no para toma de decisiones automatizada.

8.1.4. Análisis de coocurrencia diagnóstica:

- El enfoque basado en regresión logística permitió estimar la probabilidad de coocurrencia entre diagnósticos (Dx2 | Dx1), identificando asociaciones estadísticamente significativas entre ciertos grupos de enfermedades.

- Este resultado abre la posibilidad de desarrollar módulos de alerta temprana o agrupamiento clínico, una vez se cuente con mayor volumen de datos y series temporales más amplias.

8.1.5. Visualización interactiva:

- El dashboard en Streamlit permitió integrar todos los resultados —EDA, modelos y coocurrencias— en una interfaz intuitiva.

- Aunque el desempeño predictivo fue modesto, la herramienta demuestra el potencial de la visualización interactiva para acercar los resultados analíticos a los tomadores de decisiones.

8.1.6. Conclusiones Random Forest

- Random Forest proporciona una herramienta robusta para estimar la duración de la estancia hospitalaria, considerando múltiples factores clínicos y administrativos.

- Las variables más influyentes identificadas permiten focalizar estrategias de gestión hospitalaria.

- El modelo interactivo facilita estimaciones individuales y la planificación de recursos médicos y administrativos.

- Constituye un complemento valioso al modelo lineal OLS, ofreciendo una visión más flexible y precisa de la dinámica de hospitalización.

8.1.7. Conclusiones Regresión Logística: Probabilidad de Dx2 dado Dx1

- La regresión logística permite modelar la probabilidad de coocurrencia diagnóstica considerando variables clínicas básicas.

- Edad y sexo son factores significativos en varios modelos, afectando la probabilidad de presentar Dx2.

- La predicción interactiva ofrece información práctica para la evaluación de riesgo individual.

- Este análisis complementa los modelos de duración de estancia y Random Forest, proporcionando una visión más detallada de relaciones entre diagnósticos.