

UNIVERSIDAD DEL NORTE
Data Challenge
Semestre: 2025-01

Entregable #2- Documentación detallada

Presentado por:

- Luis Alfonso Herrera
- Francesca Emilsen Martínez
- Henry Sáenz Valverde

Fecha de entrega: 23/04/2025

Objetivo General

Desarrollar un modelo de aprendizaje supervisado que permita predecir la **cantidad de servicios de salud demandados** por municipio, categoría del servicio, tipo de atención, y periodo temporal (año y mes), mediante la aplicación de técnicas de análisis exploratorio de datos, selección de variables relevantes y evaluación comparativa (benchmarking) de modelos de regresión, con el fin de identificar patrones de comportamiento en la demanda y aportar información útil para la planificación estratégica del sistema de salud.

2. Proceso de interpretación de variables y construcción del modelo predictivo

El desarrollo del modelo predictivo se basó en una secuencia estructurada de pasos que combinó análisis exploratorio de datos, depuración y selección de variables, entendimiento del contexto del dominio y la implementación de múltiples enfoques de modelado. El objetivo central fue construir un modelo capaz de predecir con un alto nivel de precisión la **cantidad de servicios médicos demandados** por municipio, mes, categoría de atención y otros factores contextuales disponibles en el dataset.

2.1. Análisis exploratorio inicial del conjunto de datos


El proceso comenzó con un análisis exploratorio completo de todas las variables disponibles en el conjunto original. Este análisis inicial permitió:

- Comprender la naturaleza de cada variable (si era categórica, numérica o temporal).
- Observar su distribución, cardinalidad y valores atípicos.
- Identificar variables que tenían demasiados valores nulos, redundancias o escasa variabilidad, las cuales se eliminaron por no aportar significativamente a la construcción del modelo.

Durante esta etapa, se creó una matriz de frecuencia cruzada para visualizar el comportamiento de la variable dependiente Cantidad en combinación con variables explicativas clave como el municipio, la categoría del servicio, el tipo de atención y el año o mes del servicio. Esto permitió revelar ciertos **patrones cíclicos o estacionales** en la demanda, así como diferencias territoriales que justificaban el uso de modelos capaces de capturar interacciones complejas.

2.2. Selección y justificación de variables utilizadas

Luego de la limpieza y el análisis exploratorio, se seleccionaron un conjunto de variables que se consideraron claves para el modelo. Las variables finales, utilizadas en la construcción del modelo predictivo, fueron las siguientes:



```
vars_modelo = [  
    'Siniestro_Diagnosti_Princi_Id',  
    'Nombre_Tipo_Atencion_Arp',  
    'Nombre_Municipio_IPS',  
    'TIPIFICACION',  
    'categoria_servicios',  
    'MUNICIPIO2',  
    'AÑO_PER',  
    'MES',  
    'Cantidad'  
]
```

A continuación, se describe el papel de cada una en la lógica del modelo:

- **Siniestro_Diagnosti_Princi_Id:** Esta variable representa el diagnóstico médico principal asociado al siniestro reportado. Su inclusión permitió al modelo capturar patrones de demanda que pueden estar correlacionados con ciertas patologías o condiciones de salud recurrentes. Aunque su cardinalidad es alta, se mantuvo por el valor informativo que puede aportar a la predicción.
- **Nombre_Tipo_Atencion_Arp:** Identifica el tipo de atención prestada, como urgencias, consultas generales, procedimientos especializados, etc. Este campo es crucial ya que los distintos tipos de atención tienen patrones de demanda muy distintos, tanto en frecuencia como en estacionalidad.
- **Nombre_Municipio_IPS:** Indica el municipio donde se encuentra ubicada la institución prestadora del servicio. Esta variable es importante para entender la **oferta de servicios** y su posible relación con la demanda. Puede además correlacionarse con la densidad poblacional y la infraestructura de salud en ciertas zonas.
- **TIPIFICACION:** Esta variable agrupa los servicios o eventos según una lógica interna de clasificación. Ayuda a organizar la información en bloques más interpretables y útiles para el análisis de tendencias.
- **categoria_servicios:** Complementa a la variable anterior con una categorización más funcional o administrativa de los servicios prestados. Ambas variables juntas permiten capturar distintos niveles de abstracción en los tipos de servicios médicos.
- **MUNICIPIO2:** Representa el municipio desde el cual proviene el paciente o donde reside. Esta variable es clave para modelar la **demandas geográfica**, que no siempre coincide con el municipio de prestación del servicio. Permite modelar fenómenos como la migración por atención médica.

- **AÑO_PER** y **MES**: Variables temporales que fueron utilizadas para establecer una estructura cronológica en la demanda, lo cual es fundamental para cualquier modelo de pronóstico. Estas variables también permitieron identificar estacionalidades, tendencias crecientes o decrecientes y cambios de comportamiento en ciertos períodos.
- **Cantidad**: Es la variable objetivo (target) del modelo. Representa la cantidad de servicios reportados por cada combinación única de características. Al tratarse de una variable de conteo, se definió que la naturaleza del problema era de **regresión continua discreta**, lo cual condicionó las técnicas de modelado a aplicar.

2.3. Preparación de los datos para el modelado

Antes de entrenar los modelos, se realizaron varias transformaciones importantes:

- Las variables categóricas fueron codificadas mediante técnicas como **Label Encoding** o **One-Hot Encoding**, dependiendo de la cardinalidad de la variable y el algoritmo a utilizar.
- Se normalizaron algunas variables numéricas si el modelo lo requería (aunque no fue el caso para los árboles).
- Se separó el dataset en conjuntos de entrenamiento y validación (train-test split), considerando particiones temporales para evitar fugas de información. También se realizó un análisis de **correlaciones cruzadas** y **análisis de importancia de variables** en modelos preliminares, para validar que las variables seleccionadas tuvieran un aporte significativo al rendimiento del modelo.

2.4. Modelos de regresión implementados y evaluación

La tarea de modelado se abordó como un problema de **regresión supervisada**, dado que se deseaba predecir una variable numérica continua. Para asegurar la mejor precisión posible, se llevó a cabo un **benchmarking** entre múltiples modelos:

- **Regresión lineal**: Para establecer una línea base simple, fácil de interpretar.
- **Regresión Ridge y Lasso**: Para controlar sobreajuste y analizar penalizaciones sobre las variables menos relevantes
- **Random Forest**: Por su robustez ante datos categóricos y su capacidad para manejar interacciones no lineales.
- **XGBoost**: Por su eficiencia, velocidad y precisión en problemas estructurados.
- **LightGBM**: Por su rendimiento con datasets grandes y su enfoque en velocidad de entrenamiento.

Cada modelo fue evaluado usando métricas de error como:

- **Error Cuadrático Medio (MSE)**
- **Raíz del Error Cuadrático Medio (RMSE)**
- **MAPE (Error Porcentual Absoluto Medio)**

Estas métricas permitieron comparar el rendimiento relativo de los modelos y elegir aquel que mejor se adaptaba al comportamiento de los datos, sin caer en sobreajuste.

2.5. Conclusión del proceso de modelado

El proceso completo de análisis de variables, limpieza de datos, codificación, selección de algoritmos y evaluación de rendimiento permitió obtener un modelo predictivo sólido y bien fundamentado. Este modelo es capaz de generar estimaciones precisas sobre la demanda futura de servicios de salud por municipio, tipo de atención y categoría de servicio, brindando un **insumo valioso para la planificación estratégica y la asignación eficiente de recursos en el sistema de salud.**