Universidad del Norte

Proceso de Desarrollo Data Challenge Pro Sura

Entregable 4

Luis Alfonso Herrera, Francesca Emilsen Martínez, Henry David Saenz

April 25, 2025

Objetivo

Identificar y documentar las principales dificultades, limitaciones y aprendizajes obtenidos durante el desarrollo del modelo predictivo, así como reflexionar sobre el proceso de modelado, la calidad del conjunto de datos y la relevancia de las variables empleadas, con el fin de fortalecer la comprensión crítica del trabajo realizado.

El desarrollo del modelo predictivo para estimar la demanda de servicios médicos por tipo y municipio durante los próximos 12 meses fue una experiencia de alto valor formativo, tanto en términos técnicos como metodológicos. A lo largo del proceso, se enfrentaron múltiples desafíos que exigieron capacidad de análisis, toma de decisiones basada en evidencia y un enfoque reflexivo sobre las limitaciones inherentes al trabajo con datos reales. Desde el inicio, el proyecto partió de una pregunta fundamental: ¿es posible predecir con precisión cuántos servicios médicos serán requeridos en el futuro a partir de la información histórica disponible? Para responder esta interrogante, se propuso la implementación de un modelo predictivo en Python que integrara técnicas de análisis de datos, ingeniería de características y aprendizaje automático.

Conforme avanzaba el proyecto, se hizo evidente que el reto no constaba solo en programar un modelo, sino en comprender profundamente la estructura, distribución, naturaleza y el comportamiento de los datos.

Uno de los aspectos más complejos fue la identificación y selección de variables explicativas. Si bien el conjunto de datos proporcionado incluía un número significativo de variables, muchas de ellas no aportaban significancia estadística para las predicciones y pronostico de la demanda. Esto generó una primera gran reflexión: la calidad del modelo depende en gran medida de la calidad y pertinencia de las variables utilizadas. A raíz de ello, se consideró la posible incorporación de variables externas no disponibles en el dataset original, tales como:

- Indicadores socioeconómicos del municipio
- Nivel de acceso a servicios de salud.
- Condiciones climáticas (temperatura, humedad, lluvias)
- Comportamiento estacional de enfermedades y datos demográficos más detallados.

Otro punto crítico del proceso fue la comprensión de la naturaleza de la variable objetivo. Inicialmente, se interpretó como un conteo simple de servicios solicitados, pero el análisis más profundo reveló que en muchos casos un mismo individuo podía requerir múltiples tipos de atención en un mismo periodo y varios servicios en una misma solicitud. Este descubrimiento llevó a replantear los métodos estadísticos y de modelado utilizados, seleccionando solo las obserbvaciones donde se demanda un único servicio.

Para el modelamiento se inició con un benchmark de modelos de regresión, sin embargo, se analizaron tambien modelos específicos para conteo como regresión Poisson y binomial negativa, en lugar de asumir una distribución normal estándar.

En cuanto a la implementación técnica, se realizaron procesos cuidadosos de limpieza, transformación y validación de datos. Esto incluyó la eliminación de

registros inconsistentes y la codificación de variables categóricas. A pesar de estos esfuerzos, se identificaron limitaciones que no podían ser resueltas únicamente desde el ámbito técnico, sino que requerían una mejora estructural en la recolección y disponibilidad de datos.

La validación del modelo se realizó mediante técnicas de validación cruzada para reducir el riesgo de sobreajuste, el método implementado fue una validación cruzada de tipo ventana. Si bien el modelo alcanzó métricas aceptables de rendimiento, fue evidente que su capacidad para generalizar en el conjunto de test no alcanzó un rendimiento tan relevante.

Durante el proceso también se aprendió que, más allá de la precisión del modelo, la interpretabilidad y la utilidad práctica de los resultados son elementos clave en proyectos aplicados. Por ello, se priorizó la generación de visualizaciones claras y comprensibles que permitieran a los usuarios no técnicos entender los patrones detectados y las proyecciones realizadas.

En términos generales, este proyecto no solo dejó un modelo funcional, sino que también proporcionó valiosos aprendizajes sobre los principios fundamentales de la ciencia de datos: la importancia del contexto, el papel crítico de los datos en la modelación y la necesidad de adaptar las herramientas tecnológicas a la naturaleza del problema. Aunque las predicciones generadas pueden ser útiles como referencia, se reconoce que para una implementación real y efectiva, sería necesario continuar mejorando la calidad del conjunto de datos, así como explorar modelos más complejos e integradores que combinen múltiples fuentes de información.