Universidad Del Norte

Data Challenge Pro 2025-1

Ana Meza García
Danier Conde
Alexander Rangel

24 de Abril, 2025

1 Limpieza de Datos

1.1 Eliminación de columnas irrelevantes

Se eliminaron múltiples columnas que no aportaban valor al análisis deseado o contenían información redundante. Entre estas columnas estaban identificadores, nombres de médicos, información de localización detallada, códigos internos, fechas no necesarias, y otros campos administrativos.

1.2 Transformación de fecha

La columna **Fecha_Atencion** contenía tanto la fecha como la hora. se separaron ambos componentes utilizando el espacio como delimitador, se creó una nueva columna **Fecha** con formato **YYYY-MM-DD** y la columna **Hora** fue descartada, ya que no era relevante para el análisis (Dado que haremos las predicciones por día)

1.3 Filtrado de registros

Se aplicaron filtros para refinar los datos relevantes al análisis:

- Se excluyeron los registros donde el Concepto_Factura_Des "Consulta No Programada"
- Se eliminaron los registros que no tuvieran fecha
- Se eliminaron los registros correspondientes a los años 2019, 2020 y el primer semestre del 2021, ya que debido al impacto de la pandemia por COVID-19 durante ese periodo, los datos podrían presentar comportamientos atípicos que afectarían la representatividad y precisión del modelo.

1.4 Resultado final

Después del proceso de limpieza, el conjunto de datos quedó reducido y estructurado con las siguientes columnas:

- **Concepto_Factura_Desc:** Descripción del tipo de servicio facturado.
- **Cantidad:** Número de veces que se registró el servicio.
- **Municipio:** Ubicación geográfica del servicio.
- **Fecha:** Fecha en el que se atendió.

2 Selección del Modelo adecuado

En el marco de este proyecto, se exploraron diversos enfoques de modelado para predecir la demanda diaria de servicios de salud en distintos municipios de Colombia. Inicialmente, se probaron modelos avanzados de series temporales como **NHITS y DeepAR**, pertenecientes a la librería NeuralForecast. Aunque estos modelos son altamente efectivos en tareas con patrones complejos y relaciones no lineales, presentaron limitaciones en este caso específico. Particularmente, mostraron dificultades al capturar comportamientos irregulares en la demanda y tendieron a aplanar las predicciones, especialmente en contextos con baja demanda y alta dispersión, como lo evidencian los servicios con registros esporádicos.

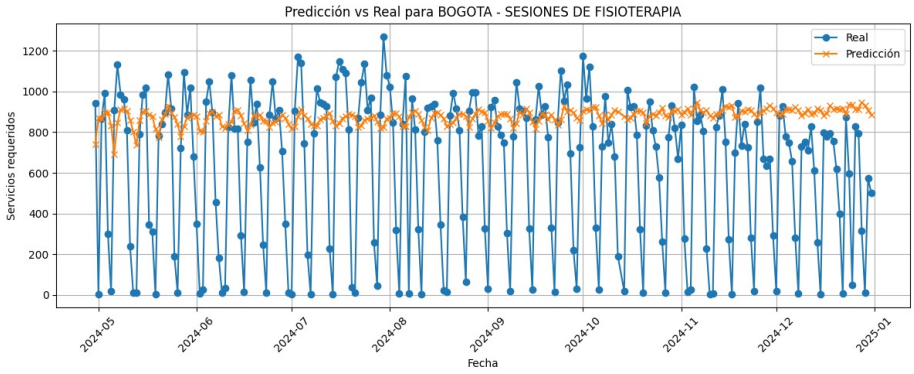

Figure 1: Predicción de servicios usando NHITS

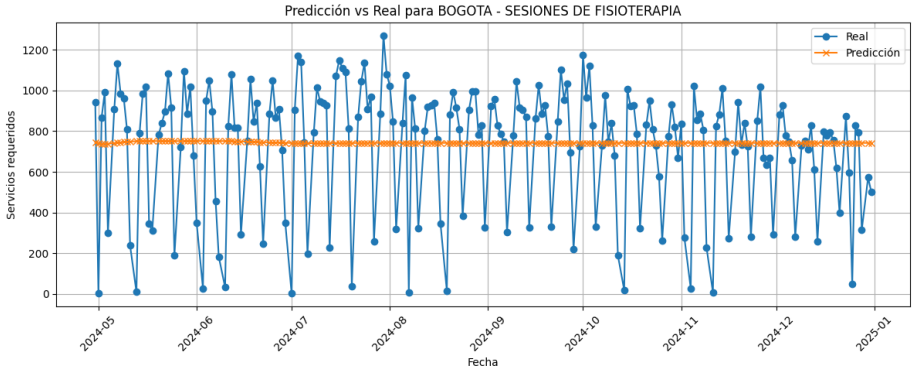

Figure 2: Predicción de servicios usando DeepAR

Por el contrario, el modelo **Prophet**, desarrollado por **Facebook**, demostró una mayor capacidad para adaptarse a la estacionalidad semanal y a los cambios abruptos en los datos históricos. Prophet ofrece ventajas clave en entornos donde los datos presentan patrones repetitivos con componentes de tendencia y estacionalidad bien definidos, y permite incorporar de forma sencilla variables como días festivos o eventos especiales. Su flexibilidad y facilidad de interpretación facilitaron la evaluación y ajuste de parámetros, lo que condujo a mejores métricas de error y mayor coherencia visual en las predicciones, consolidándolo como la mejor alternativa para la solución propuesta.

Finalmente, con base en el análisis de la matriz de correlación entre las diferentes series temporales, se observa que los niveles de correlación entre la mayoría de las series son bajos, con valores cercanos a cero en su mayoría. Esto sugiere que las series son prácticamente independientes entre sí, lo cual implica que no existe una relación lineal significativa que pueda ser aprovechada por modelos globales como DeepAR o NHITS, los cuales dependen de la existencia de patrones compartidos entre las series para mejorar el desempeño general del modelo. Por esta razón, se optó por utilizar Prophet, un modelo de series temporales que trabaja de manera individual por serie y que permite capturar dinámicas locales específicas de cada municipio y tipo de servicio, resultando ser una opción más adecuada para la naturaleza de nuestros datos.

Figure 3: Grafico de correlación de series

3 Transformación de variables

Para implementar correctamente los modelos de regresión, es fundamental que los datos de entrenamiento tengan una estructura clara y estandarizada. Por ello, se realizaron algunos procesos de creación y renombramiento de variables. A continuación, se describen los cambios más relevantes:

- **Nueva variable 'unique_id':** Creamos una nueva variable en la que se combinan los diferentes municipios de colombia con los servicios solicitados en el periodo considerado.
- **Renombramiento de la variable fecha a 'ds':** Para garantizar la compatibilidad con el modelo Prophet, fue necesario renombrar la variable que representa la fecha a **'ds'**, ya que este modelo requiere específicamente dicha nomenclatura para identificar el componente temporal de la serie.
- **Renombramiento de la variable objetivo a 'y':** Los modelos implementados requieren que la variable dependiente o de salida sea identificada con el nombre 'y' para su correcto funcionamiento.

unique_id	ds	y
BOGOTA - SESIONES DE FISIOTERAPIA	2021-07-01	505.0
BOGOTA - SESIONES DE FISIOTERAPIA	2021-07-02	459.0
BOGOTA - SESIONES DE FISIOTERAPIA	2021-07-03	158.0
BOGOTA - SESIONES DE FISIOTERAPIA	2021-07-04	4.0
BOGOTA - SESIONES DE FISIOTERAPIA	2021-07-05	30.0

Figure 4: Dataset final con los cambios realizados

4 Entrenamiento y validación del modelo

4.1 Tratamiento de datos faltantes

En el conjunto de datos original, las fechas en las que no se prestaron servicios no estaban registradas, lo cual implicaba la ausencia total de datos para esos días. Esta ausencia podría llevar a interpretaciones erróneas durante el análisis temporal, como asumir que no hubo días sin actividad.

Para resolver esto, se procedió a completar el dataset agregando explícitamente todas las fechas dentro del rango temporal de análisis. Para aquellas fechas en las que no se brindaron servicios, se asignaron valores de cero en las variables correspondientes.

Esto nos permite mantener la continuidad temporal, lo que es crucial para análisis de series de tiempo, cálculos de promedios móviles, estacionalidad y

tendencias y evitar sesgos, ya que la omisión de días sin servicio podría dar una falsa impresión de actividad continua.

	ds	unique_id	y	weekday	month	year
0	2021-07-01	ABEJORRAL - CALIFICACIÓN DE ORIGEN AT	0.0	3	7	2021
1	2021-07-02	ABEJORRAL - CALIFICACIÓN DE ORIGEN AT	0.0	4	7	2021
2	2021-07-03	ABEJORRAL - CALIFICACIÓN DE ORIGEN AT	0.0	5	7	2021
3	2021-07-04	ABEJORRAL - CALIFICACIÓN DE ORIGEN AT	0.0	6	7	2021
4	2021-07-05	ABEJORRAL - CALIFICACIÓN DE ORIGEN AT	0.0	0	7	2021
...	...	...	...	...	...	...
15331835	2024-12-27	ZIPAQUIRA - URGENCIAS ORTOPEDISTA	0.0	4	12	2024
15331836	2024-12-28	ZIPAQUIRA - URGENCIAS ORTOPEDISTA	0.0	5	12	2024
15331837	2024-12-29	ZIPAQUIRA - URGENCIAS ORTOPEDISTA	0.0	6	12	2024
15331838	2024-12-30	ZIPAQUIRA - URGENCIAS ORTOPEDISTA	0.0	0	12	2024
15331839	2024-12-31	ZIPAQUIRA - URGENCIAS ORTOPEDISTA	0.0	1	12	2024

Figure 5: Dataset final una vez tratado

4.2 Partición del dataset

Para entrenar el modelo, se realizó una partición del conjunto de datos de la siguiente manera:

- El **dataset de entrenamiento** contiene los registros con fechas anteriores al 30 de abril de 2024. En total, se utilizaron tres años de datos históricos para entrenar el modelo.
- El **dataset de prueba** incluye los registros posteriores al 30 de abril de 2024. Este conjunto fue empleado para evaluar la capacidad predictiva del modelo en datos no vistos.

4.3 Tratamiento de los días festivos

Para mejorar la predicción del modelo, es importante tratar fechas anómalas, en este caso, días festivos. Cómo se puede apreciar, en estos días, la tendencia del número de servicios es a la baja, por lo que su tratamiento es fundamental para el buen funcionamiento de la solución implementada (ver gráfica).

Para una mayor visualización de este fenómeno, veamos el valor predicho por el modelo vs el valor real de la cantidad de servicios para esta fecha (25/04/2024)

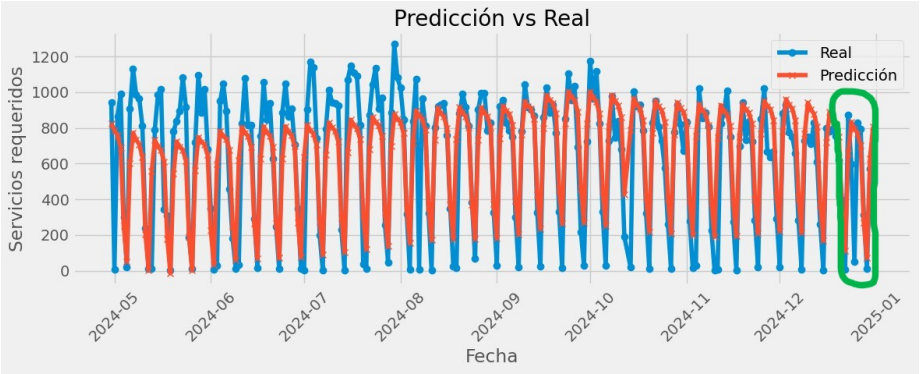

Figure 6: Predicción del 25 de diciembre (encerrada en verde)

```
navidad_prediccion = prediccion[prediccion['ds'] == '2024-12-25']
navidad_real=dataset_Bog_Fisio[dataset_Bog_Fisio['ds'] == '2024-12-25']

navidad= pd.merge(navidad_real, navidad_prediccion, on=['ds'], how='inner')
navidad= navidad[["y","entero_cercano"]]
navidad
```

	y	entero_cercano
0	49.0	796.0

Figure 7: Predicción de servicios para el 25 de diciembre sin incluir 'holidays' al modelo

Luego del tratamiento, la predicción para los servicios requeridos en los días festivos mejoró significativamente, cómo lo evidencia la figura 8.

```
navidad_prediccion = prediccion[prediccion['ds'] == '2024-12-25']
navidad_real=dataset_prueba[dataset_prueba['ds'] == '2024-12-25']

navidad= pd.merge(navidad_real, navidad_prediccion, on=['ds'], how='inner')
navidad= navidad[["y","entero_cercano"]]

navidad
```

	y	entero_cercano
0	49.0	63.0

Figure 8: Predicción de servicios para el 25 de diciembre incluyendo el parámetro 'holidays' al modelo

5 Evaluación del modelo y resultados

Para evaluar la efectividad del modelo, se tomaron los últimos 7 meses del 2024, esto con el fin de comprobar si los valores predichos eran parecidos a los reales.

Para ello, se tomó el caso particular del número de servicios de **Sesiones de Fisioterapia** solicitados en **Bogotá** y se obtuvo la siguiente predicción:

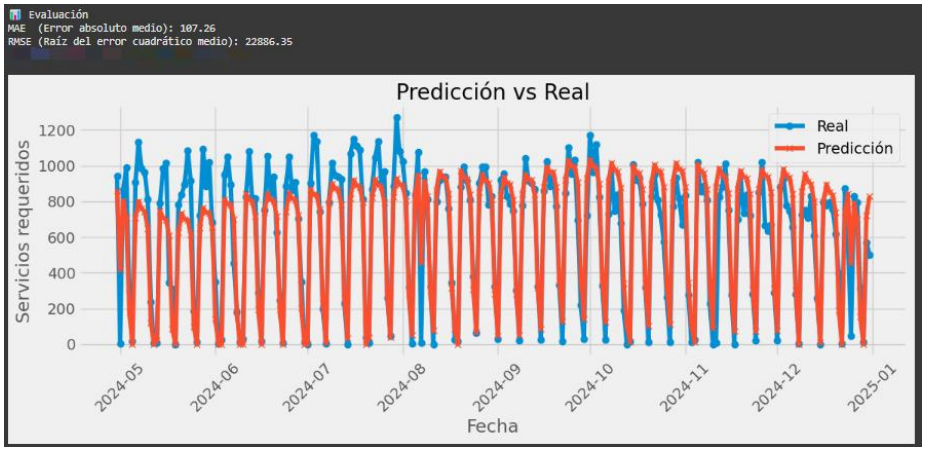

Figure 9: Cantidad de sesiones de fisioterapia solicitadas en Bogotá 2021-2024

5.1 Análisis de Componentes del Modelo

Con la gráfica 10 podemos observar cómo se descomponen los distintos componentes que influyen en el comportamiento de la serie de tiempo analizada. En primer lugar, la tendencia muestra una caída progresiva desde finales de 2024 hasta finales de 2025, lo que indica una disminución constante en los valores a lo largo del año. Además, se evidencia que los días festivos tienen un impacto negativo importante, generando caídas marcadas en la serie durante esas fechas.

Por otro lado, el patrón semanal refleja que los días con mejor comportamiento son de lunes a jueves, destacándose especialmente los martes, mientras que los domingos muestran una baja considerable en el valor, lo que sugiere menor actividad o rendimiento ese día. Finalmente, el componente anual revela que durante los meses de mitad de año (especialmente entre julio y noviembre) hay una recuperación, en contraste con caídas que se presentan al inicio y hacia mediados del año, como en mayo.

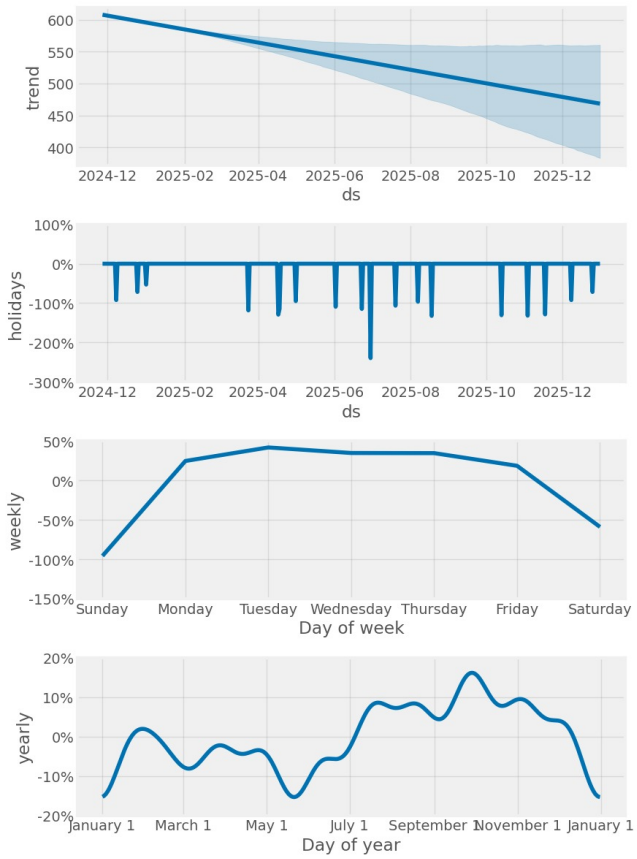

Figure 10: Gráficas para el análisis del modelo

6 Conclusiones

Luego de evaluar diferentes enfoques, Prophet resultó ser el modelo más adecuado para este análisis, principalmente porque permite trabajar de forma independiente entre municipios y tipos de servicios, lo cual se ajusta muy bien a la estructura de los datos disponibles. Además, el hecho de poder incorporar los días festivos fue un factor clave, ya que estos tienen un impacto significativo en la demanda, como se evidenció en la descomposición del modelo.

A través de este análisis, fue posible identificar una tendencia general a la baja en el número de sesiones de fisioterapia en Bogotá, así como patrones semanales bien definidos, donde los días de mayor actividad son entre lunes y jueves, mientras que domingos y festivos presentan caídas notables. También se observó una recuperación en la demanda durante la segunda mitad del año, lo que puede ser útil para ajustar estrategias operativas y de planificación.