UNIVERSIDAD DEL NORTE

GRUPO sura

# DATA CHALLENGE PRO

*RETO 2025-01*


**Presentado por:**

David Jesús Roa Aníbal

Dariana Sanguino Cuello

Abrahan Elias Basto Martinez

# ÍNDICE

## Contenido del informe

# INTRODUCCIÓN

Fundado en Medellín en el año 1944 bajo el nombre Compañía Suramericana de Seguros Generales, se dedicó a la actividad aseguradora en Colombia. Con aproximadamente 5.3 millones de afiliados, ocupando una gran parte del mercado de aseguradoras colombianas, se encuentran hoy en día bajo el nombre Grupo de Inversiones Suramericana o mayormente conocidos como Grupo SURA.

En el marco del Data Challenge propuesto por la ARL SURA, se plantea una problemática crítica y de alta relevancia para el sistema de salud laboral colombiano: la previsión de la demanda de servicios en los municipios del país. ARL SURA, una de las Administradoras de Riesgos Laborales más importantes de Colombia, tiene como misión proteger y acompañar a los trabajadores en la prevención, atención y rehabilitación de accidentes y enfermedades laborales. En este contexto, anticipar cuándo, dónde y cuántos servicios serán necesarios no es solo una cuestión de eficiencia logística, sino de impacto directo en la vida de las personas.

El reto se enfoca en desarrollar una solución predictiva robusta, capaz de identificar patrones complejos en los datos históricos de atención registrados entre enero de 2019 y diciembre de 2024. La naturaleza urgente, a veces impredecible, de estos eventos exige enfoques que vayan más allá de los métodos tradicionales. Nuestra propuesta combina técnicas de análisis de series temporales, aprendizaje automático y visualización avanzada para construir un modelo de previsión capaz de adaptarse a la variabilidad de los entornos municipales y responder con precisión ante escenarios cambiantes.

El enfoque no se limita a obtener un buen puntaje en precisión: apunta a construir una herramienta útil, interpretable y accionable para la toma de decisiones. Porque en salud laboral, predecir no es solo anticiparse, es cuidar mejor.

# OBJETIVO DEL ESTUDIO

## GENERAL

Desarrollar una solución predictiva que permita estimar la demanda futura de servicios de salud laboral a nivel municipal, usando datos históricos del grupo SURA.

## ESPECÍFICOS

- Limpiar, transformar y explorar la base de datos entregada.
- Probar y comparar modelos predictivos como ARIMA, Prophet, LSTM, XGBoost , Random Forest, entre otros.
- Proporcionar visualizaciones útiles y una propuesta accionable.

# METODOLOGÍA

## FUENTE DE DATOS:

Datos entregados por SURA, periodo 2019–2024.

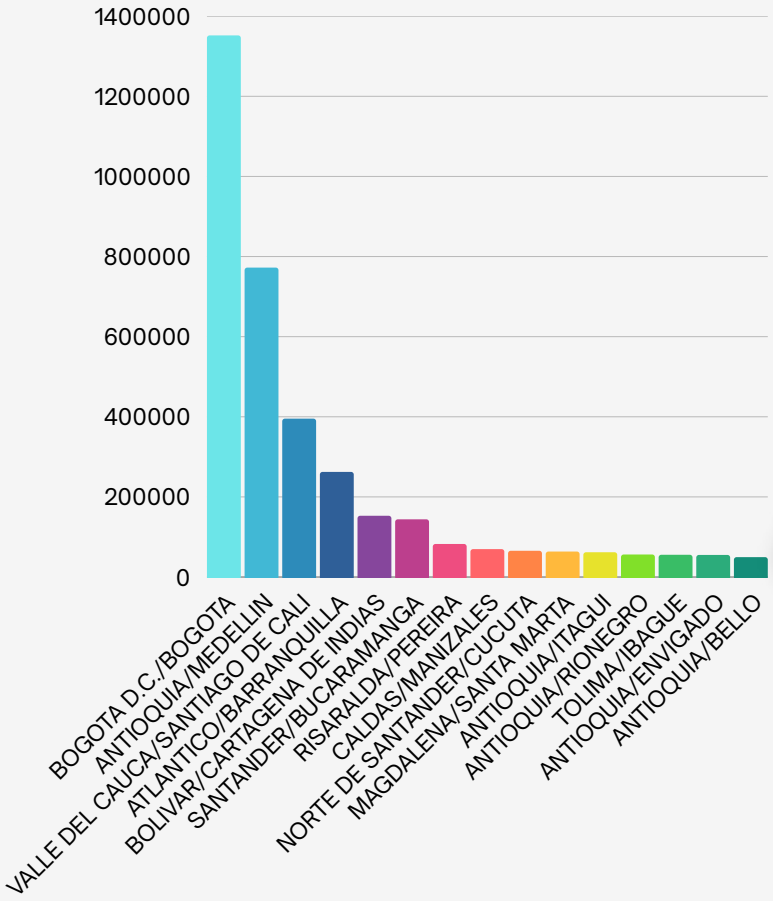
| Nombre_Oficina_Arp | Numero_Uen_Arp | Siniestro_Arp_Id | Nombre_Sinies_Diagnosti_Princi | ... | MUNICIPIO | HOMOLOGACION NIT | ... |
|---|---|---|---|---|---|---|---|
| OFICINA BOGOTA | 419 | bf0bf0c0e3 | G560 | | BOGOTA | 3d1546fd2b | |
| OFICINA BOGOTA | 411 | 96e0b92217 | S626 | | BOGOTA | 3d1546fd2b | |
| OFICINA BOGOTA | 414 | 47999f6929 | S835 | | BOGOTA | 465b9adf14 | |

## PROCESAMIENTO:

Limpieza, análisis exploratorio, agrupación por municipio.

## FOCO TERRITORIAL

# TOP 15 MUNICIPIOS CON MÁS AFILIADOS



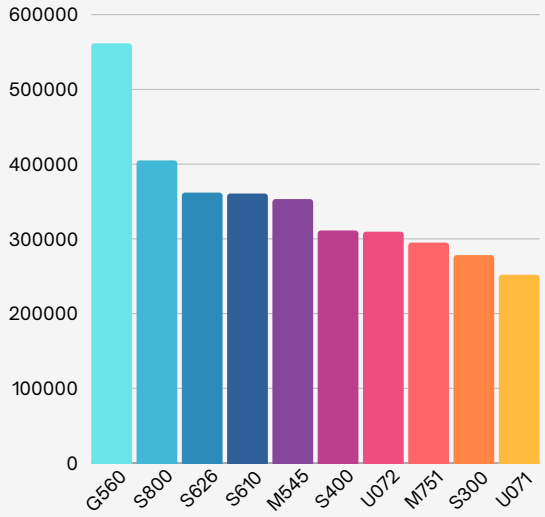


# + 5 millones de Afiliados

Para comprender mejor la distribución territorial de la demanda en salud laboral, se identificaron los 15 municipios con el mayor número de afiliados. Estos municipios representan focos estratégicos para la gestión del riesgo, no solo por su volumen poblacional, sino también por la concentración de actividad económica y laboral. Analizar su comportamiento permite enfocar esfuerzos predictivos y operativos donde más impacto puede generar la intervención oportuna.

## LA HUELLA DEL TRABAJO

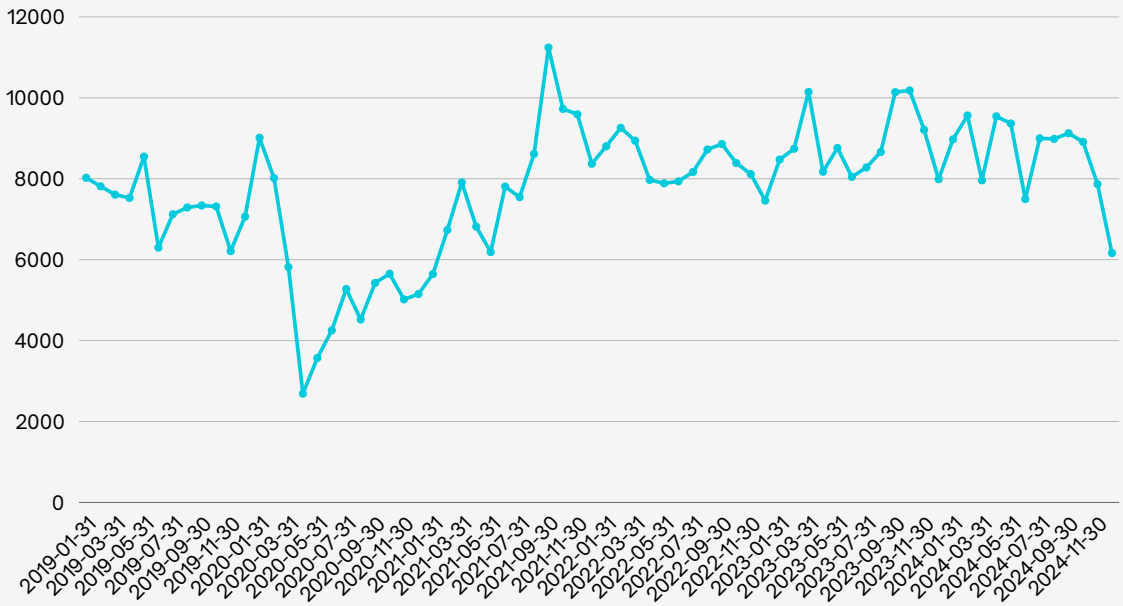
# TOP 10 CAUSAS DE ATENCIÓN REGISTRADAS




1. Síndrome del túnel carpiano (G560)
2. Contusión de rodilla (S800)
3. Fractura de otro dedo de la mano (S626)
4. Herida de dedo(s) de la mano (S610)
5. Lumbago no especificado (M545)
6. Contusión del hombro y del brazo (S400)
7. COVID-19, virus no identificado (U072)
8. Síndrome del manguito rotatorio (M751)
9. Contusión de la región lumbosacra y de la pelvis (S300)
10. COVID-19, virus identificado (U071

El análisis de los diagnósticos más reportados en el conjunto de datos permite identificar los principales motivos de atención en el sistema de riesgos laborales. A través del conteo y clasificación de los siniestros más frecuentes a nivel nacional, se pueden reconocer patrones de salud ocupacional que orientan la priorización de políticas preventivas. Esta información es clave para diseñar estrategias focalizadas que reduzcan la incidencia de enfermedades laborales y accidentes recurrentes, mejorando así la eficiencia del sistema y el bienestar de los trabajadores.

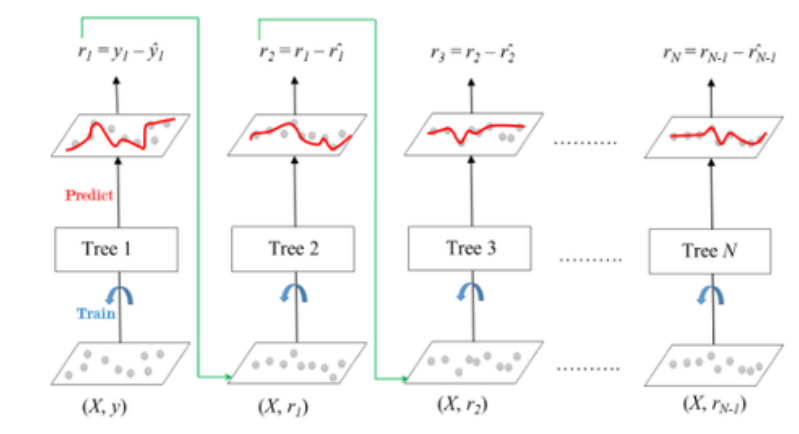
## DEMANDA PARA DIAGNÓSTICO G560 (MENSUAL)



## MODELADO PREDICTIVO:

Para abordar el reto de prever la demanda futura de servicios de salud laboral, se implementó una selección diversa de modelos predictivos, tanto tradicionales como basados en aprendizaje automático y deep learning. La elección de los modelos respondió al objetivo de capturar diferentes patrones temporales y no lineales presentes en los datos municipales.
Los modelos empleados fueron:

# Gradient Boosting




Este modelo fue interesante porque su lógica de combinación de árboles se sale de lo tradicional en series temporales. Nos costó entender cómo incorporar el componente de tiempo, ya que no es nativo para predicción temporal. Al aplicarlo, vimos que tendía a subestimar los picos, como si jugara a lo seguro. Eso nos enseñó que estos modelos son potentes, pero requieren un trabajo cuidadoso en la ingeniería de variables para no perder la dinámica temporal real.

# ARIMA

$$y_t^* = \Delta^d y_t$$

$$y_t^* = \mu + \underbrace{\sum_{i=1}^{p} \phi_i y_{t-i}^*}_{\text{AR}} + \underbrace{\sum_{i=1}^{q} \theta_i \epsilon_{t-i}}_{\text{MA}} + \epsilon_t$$

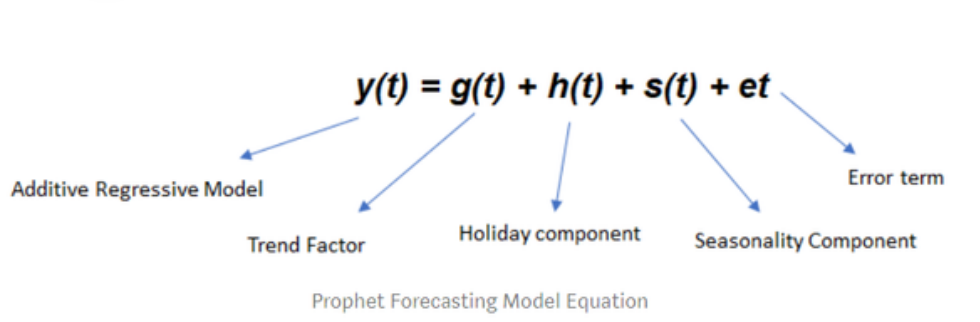
Este modelo fue uno de los primeros que implementamos debido a su enfoque clásico en predicción de series temporales. Si bien fue útil para comprender los fundamentos del modelado univariado, su capacidad para adaptarse a las fluctuaciones abruptas o tendencias no lineales fue limitada. En la práctica, su predicción terminó siendo una línea casi plana, lo que evidenció sus restricciones frente a una serie con cambios estructurales como los provocados por la pandemia. El mayor aprendizaje fue entender que, aunque ARIMA es sólido teóricamente, no siempre ofrece buenos resultados en escenarios reales complejos.

# SARIMAX

$$y_t = c + \sum_{n=1}^{p} \alpha_n y_{t-n} + \sum_{n=1}^{q} \theta_n \epsilon_{t-n} + \sum_{n=1}^{P} \phi_n y_{t-sn} + \sum_{n=1}^{Q} \eta_n \epsilon_{t-sn} + \epsilon_t$$

El modelo SARIMAX nos permitió explorar una de las técnicas más robustas dentro de los enfoques estadísticos tradicionales. A diferencia de ARIMA, este modelo tiene la ventaja de incorporar regresores externos, lo cual lo convierte en una herramienta muy útil cuando se dispone de variables adicionales que puedan explicar el comportamiento de la serie. Sin embargo, implementar SARIMAX también nos enfrentó a un reto: entender cómo ajustar correctamente la estructura estacional y los términos exógenos para evitar sobreajustes o errores sistemáticos. El modelo logró capturar parte de la tendencia general, pero presentó limitaciones al adaptarse, lo que resalta su sensibilidad a la especificación inicial y a la calidad de los regresores externos.

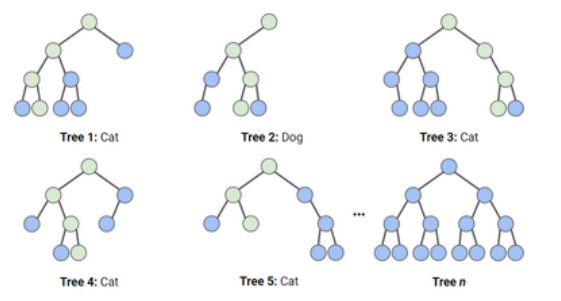
# Prophet



Prophet Forecasting Model Equation

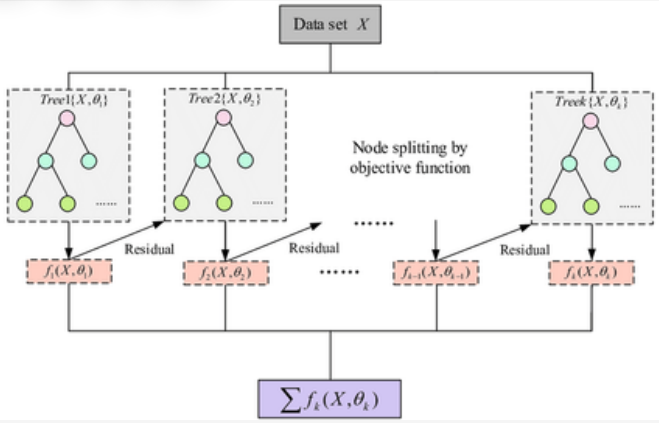
Prophet destacó por su flexibilidad y facilidad de implementación, la curva de aprendizaje también fue más sueve. Esta opción Nos permitió modelar componentes estacionales, tendencias y eventos atípicos con relativa claridad. Lo más retador fue interpretar cómo ajustaba la incertidumbre y cómo reaccionaba ante rupturas en la serie. Al compararlo con los datos reales de 2024, Prophet ofreció una proyección bastante cercana, capturando la estacionalidad y parte de la dinámica general. Esta experiencia nos mostró el valor de los modelos híbridos que combinan componentes estructurados con adaptabilidad.

# Random Forest




Random Forest presentó retos similares: si bien es robusto y fácil de entrenar, su naturaleza no secuencial lo limita para capturar dependencias temporales complejas. Al evaluar sus predicciones, notamos una marcada tendencia a la subestimación, sobre todo en los puntos altos de la serie. Esta experiencia reforzó la importancia de considerar estructuras temporales explícitas cuando se trabaja con series cronológicas.

# XGBoots




El modelo XGBoost mostró un desempeño aceptable en la predicción mensual, aunque con ciertas limitaciones visibles al contrastar con los datos reales de 2024. Si bien logró mantener una tendencia general coherente, se notó una tendencia a suavizar los extremos, especialmente en los picos más altos de demanda, donde subestimó los valores reales. Esto sugiere que el modelo fue capaz de aprender la estructura básica de la serie, pero tuvo dificultades para capturar los cambios abruptos o estacionales más marcados. A pesar de esto, la cercanía de varias predicciones al comportamiento observado lo posiciona como uno de los modelos con mejor capacidad de ajuste entre los que no son específicamente diseñados para series temporales. Esta implementación nos permitió ver de primera mano la importancia de ajustar bien los hiperparámetros y seleccionar variables que reflejen adecuadamente los patrones temporales.

# CONCLUSIONES Y RECOMENDACIONES

A lo largo de este análisis, logramos no solo implementar y comparar diversos enfoques de modelado predictivo, sino también comprender profundamente las dinámicas que subyacen en la demanda de servicios de salud laboral en los municipios analizados. Cada modelo aplicado nos enseñó algo distinto: desde los límites de los métodos clásicos como ARIMA, hasta la versatilidad y precisión que ofrece Prophet al incorporar estacionalidades, tendencias y eventos externos. Este proceso no fue solo técnico, sino formativo, y nos permitió identificar patrones relevantes que, en otros contextos, podrían pasar desapercibidos.

Una de las principales conclusiones es que no existe un modelo universalmente óptimo, sino que la calidad de la predicción depende del entendimiento del problema, de la estructura de los datos y de una calibración cuidadosa. En nuestro caso, Prophet se destacó como el modelo con mayor capacidad de adaptación a la complejidad de la serie, mostrando un comportamiento cercano al observado en 2024, sin requerir transformaciones agresivas ni supuestos rígidos.

A partir de esto, proponemos a SURA:

- **Implementar Prophet como motor base para la predicción mensual por municipio**, combinándolo con alertas automáticas en caso de desviaciones atípicas entre predicción y realidad.
- Diseñar un **tablero interactivo** donde se visualicen las predicciones futuras junto con intervalos de confianza, comparables en tiempo real con la evolución real. Esto permitiría tomar decisiones de planeación anticipada sobre asignación de recursos, refuerzo de atención en ciertas regiones o detección temprana de cambios estructurales.
- Incorporar **variables externas y contextuales** (clima, movilidad, días festivos, actividades económicas por municipio) en futuros modelos para capturar mejor los determinantes de la demanda.

Más allá de las métricas, nuestro objetivo fue generar valor accionable. Este ejercicio no solo fortaleció nuestras capacidades analíticas, sino que también nos permitió acercarnos a una problemática real con impacto tangible en la salud laboral del país. Creemos que este tipo de soluciones integradas, predictivas y visuales pueden marcar una diferencia sustancial en la manera como se planean y entregan los servicios de salud.

UNIVERSIDAD DEL NORTE

# DATA CHALLENGE PRO

*RETO 2025-01*

GRUPO sura