Informe: **Data Challenge Pro - 202501**

El presente informe resume el desarrollo de una solución predictiva orientada a anticipar la demanda de servicios ofrecidos por las IPS a SURA. El objetivo principal fue diseñar un modelo capaz de generar predicciones confiables a partir de datos históricos y variables contextuales relevantes, contribuyendo así a una planificación más eficiente de recursos y servicios.

# **1. Enfoque General**

El enfoque adoptado comprendió las siguientes etapas:

### Recolección y análisis de datos

Se recopiló información proveniente de la base de datos compartida por parte de SURA, la cual contenía registros históricos detallados relacionados con diversos factores, como oficinas, municipios y tipos de atención. Esta base de datos abarcó un periodo de varios años (especificar el rango temporal exacto), lo que permitió obtener una visión amplia de la evolución de las variables en estudio. Las fuentes de datos incluyeron bases internas, registros históricos de atención y otras fuentes relevantes proporcionadas por la empresa.

### 1.2 Preprocesamiento

En esta etapa, se llevó a cabo un exhaustivo proceso de limpieza de los datos. Se trataron los valores nulos y se codificaron las variables categóricas mediante técnicas como la codificación One-Hot y Label Encoding. También se ajustaron las variables de fecha para facilitar la manipulación temporal, separando la información en componentes como año, mes, día y tipo de atención. Este preprocesamiento fue esencial para asegurar la calidad de los datos y su preparación para el análisis posterior.

### 1.3 Exploración y selección de características

A través de análisis estadísticos y visuales, se identificaron las variables con mayor correlación e influencia sobre la variable objetivo, que en este caso es la cantidad registrada en las distintas oficinas y tipos de atención. Las variables como el municipio, las oficinas y los tipos de atención fueron determinantes en el modelo, mostrando una clara relación con los patrones de comportamiento de los datos. Se utilizó un enfoque exploratorio que incluyó histogramas, distribuciones y análisis de medias móviles para comprender mejor el impacto de estas variables en el modelo.

### 1.4 Entrenamiento y validación de modelos

Se entrenaron diversos algoritmos de aprendizaje automático, con un enfoque en modelos como XGBoost, para predecir la variable objetivo. Los modelos fueron evaluados utilizando métricas como RMSE (Root Mean Squared Error), R² (coeficiente de determinación), y otras métricas relevantes para medir la precisión y el desempeño general del modelo. Este proceso permitió identificar el modelo con mejor rendimiento en términos de predicción y ajuste a los datos reales.

### 1.5 Selección del modelo final

El modelo **XGBoost** fue seleccionado como el modelo final debido a su balance entre precisión y capacidad de generalización. XGBoost demostró ser robusto frente a los datos ruidosos y logró una buena capacidad para predecir incluso en presencia de variables complejas y relaciones no lineales. Su capacidad para manejar grandes volúmenes de datos y su rendimiento superior en las métricas de evaluación lo convirtieron en la opción más confiable para este análisis.

# **2. Desafíos Encontrados**

Durante el desarrollo se presentaron varios desafíos relevantes. Uno de los principales fue la gran cantidad de datos y la selección de variables, lo cual se abordó mediante un análisis exploratorio exhaustivo y la determinación de la relación entre las variables en los modelos. Este proceso permitió identificar las variables más influyentes y relevantes, facilitando así la reducción de la complejidad sin perder información crucial.

Otro reto importante fue el ajuste de un modelo como XGBoost a un problema multivariado. La solución implementada consistió en realizar un ajuste de hiperparámetros utilizando técnicas de optimización como Optuna, lo que permitió encontrar la configuración ideal para manejar múltiples características de entrada de manera eficiente. Además, se utilizaron técnicas como la codificación de variables y el manejo adecuado de los datos para asegurar que el modelo pudiera generalizar correctamente en todas las variables involucradas.

# **3. Evaluación de la Solución**

El modelo seleccionado mostró un rendimiento satisfactorio, con un valor de RMS de 1.37288 y un R² de 0.97239 en el conjunto de validación. Los resultados indican que la solución propuesta es **robusta** y aplicable en escenarios reales, especialmente en contextos donde se requiere anticipar la demanda de servicios o productos, optimizando recursos y mejorando la toma de decisiones a nivel de operación. Además, se identificaron oportunidades para su mejora y expansión futura, incluyendo la incorporación de datos en tiempo real para un modelo más dinámico o el ajuste del modelo según estacionalidades y tendencias emergentes en los datos históricos.