**Informe de Análisis de Nacimientos en el Municipio de Charalá**

**Santiago David Garzón Rodríguez**

**Ingeniería Mecatrónica**

**Docente: Mag. Andrés Fabian Leal Archila**

**Estadística Inferencial**

**Universidad Autónoma de Bucaramanga**

**2025-2**

# 1. Introducción

Este informe presenta un análisis exploratorio y un proceso de limpieza de datos aplicados al conjunto de datos sobre nacimientos en el municipio de Charalá, el cual se encuentra disponible en el repositorio de datos abiertos del Gobierno de Colombia. Esta base de datos contiene información relevante sobre variables demográficas y clínicas asociadas a los nacimientos, como el sexo del recién nacido, peso al nacer, tipo de parto, institución prestadora de salud, entre otras.

Dado el objetivo de realizar un análisis enfocado, se seleccionaron 6 variables representativas de las más de 20 disponibles en el conjunto original. Esta selección se basó en criterios de relevancia estadística, completitud de los datos y pertinencia para el análisis exploratorio.

El propósito principal de este trabajo fue comprender la estructura y calidad del conjunto de datos, identificar patrones y tendencias generales. Para ello, se realizaron diversas tareas fundamentales: detección y tratamiento de valores faltantes, identificación y eliminación de valores atípicos mediante métodos estadísticos, y generación de visualizaciones gráficas que faciliten la interpretación de los datos categóricos y numéricos.

Además, se llevaron a cabo pruebas de normalidad sobre variables cuantitativas para evaluar su distribución estadística y determinar la aplicabilidad de métodos paramétricos en análisis posteriores. Todo este proceso permitió no solo mejorar la calidad del conjunto de datos, sino también extraer conclusiones preliminares útiles.

# 2. Metodología

El análisis se llevó a cabo en Python usando la ayuda de la inteligencia artificial Gemini, presente en el software de Google Colab y utilizando bibliotecas como pandas, matplotlib, seaborn, scipy y statsmodels. Los pasos principales fueron los siguientes:

* - Carga del conjunto de datos en formato CSV.
* - Revisión inicial del contenido, tipos de datos y presencia de valores nulos.
* - Limpieza de datos: eliminación de filas con valores faltantes.
* - Selección de datos de la base a usar (Departamento Nacimiento, Municipio Nacimiento, Sexo, Peso, Talla, Tiempo Gestación")
* - Visualización de distribuciones categóricas y numéricas mediante gráficos de barras, tortas, histogramas y diagramas de caja.
* - Eliminación de valores atípicos usando el método del rango intercuartílico (IQR).
* - Pruebas de normalidad (Shapiro-Wilk, Kolmogorov-Smirnov, Anderson-Darling, Jarque-Bera) y gráficos QQ.

# 3. Resultados

# 

# 

# 

# 

# 

# 

# 

# 

# 

# 

# 

# 

# 

# 4. Conclusiones

Este análisis permitió limpiar y explorar adecuadamente el conjunto de datos de nacimientos en Charalá. Se identificaron problemas comunes en datos reales como valores nulos y outliers, los cuales fueron tratados de forma sistemática. El análisis de normalidad aportó información útil para la elección de futuras pruebas estadísticas. Así mismo, se logra concluir lo siguiente:

* El departamento de nacimiento es exclusivamente Santander, como era de esperarse.
* Aunque el municipio de análisis es Charalá, la mayoría de los nacimientos ocurrieron en San Gil, lo que sugiere una posible centralización de servicios de salud.
* El sexo de los recién nacidos está equilibrado entre hombres y mujeres.
* La limpieza y eliminación de datos atípicos mejoró significativamente la distribución de variables numéricas.
* A través de pruebas como Shapiro-Wilk, Kolmogorov-Smirnov, Anderson-Darling y Jarque-Bera, complementadas por gráficos QQ Plot, se concluyó que la única variable que sigue una distribución normal es el peso, mientras que la talla y el tiempo de gestación no.