## **NOMBRE: Nicolas Nahuel Belinaux**

**EMAIL: nico.belinaux@gmail.com**

**COHORTE: DAFT- 08**

**FECHA DE ENTREGA: 18/11/2024**

## 

## 

## **EXPANSIÓN ESTRATÉGICA DE BIOGENESYS:**

## **Introducción**

La empresa farmacéutica **BIOGENESYS** está interesada en identificar las ubicaciones estratégicas para expandir sus laboratorios en Latinoamérica. Este proyecto tiene como objetivo principal utilizar el análisis de datos para fundamentar la expansión, centrándose en factores como la incidencia de COVID-19, las tasas de vacunación y la disponibilidad de infraestructuras sanitarias en países clave. La iniciativa busca optimizar la respuesta a la pandemia y postpandemia, mejorando el acceso a vacunas en las regiones más vulnerables.

En este contexto, mi rol como **Data Analyst** es desarrollar un análisis integral que permita tomar decisiones estratégicas basadas en datos. Para ello, el estudio se enfoca en los países de Argentina, Brasil, Chile, Colombia, México y Perú. BIOGENESYS plantea invertir en ubicaciones prioritarias mediante la identificación de patrones, tendencias y áreas de oportunidad, aprovechando datos confiables y visualizaciones interactivas.

Los objetivos específicos del proyecto son claros y están alineados con las necesidades de la empresa:

1. **Explorar y analizar los datos** sobre incidencia de COVID-19 y factores asociados, extrayendo tendencias, patrones y oportunidades mediante estadísticas descriptivas, visualizaciones y mediciones clave.
2. **Asegurar la calidad de los datos**, implementando técnicas avanzadas de limpieza y transformación para garantizar la fiabilidad de los análisis.
3. **Optimizar los procesos ETL (Extracción, Transformación y Carga)** para estructurar y procesar la información de forma eficiente, maximizando la eficacia del análisis y la toma de decisiones.
4. **Diseñar dashboards interactivos en Power BI**, permitiendo una exploración dinámica y una comprensión visual de los datos para orientar estrategias empresariales.

Este análisis no solo busca proporcionar insights valiosos sobre las ubicaciones ideales para los laboratorios, sino también identificar cómo las variables demográficas y de salud influyen en la incidencia del COVID-19 y la efectividad de las campañas de vacunación.

## **Desarrollo del proyecto**

### **Primer Avance: Exploración y Limpieza de Datos**

El proyecto comenzó con la exploración de un extenso dataset proporcionado por Ingeniería de Datos. El archivo original contenía **12,216,057 filas y 50 columnas**, ocupando un tamaño de 21 GB. Para facilitar su manipulación, se redujo su tamaño a un CSV más manejable, junto con un documento auxiliar que detallaba el significado de las columnas seleccionadas.

El análisis inicial se llevó a cabo en **Python**, utilizando librerías como **Pandas**, **Numpy** y herramientas de visualización como **Matplotlib** y **Seaborn**. El primer paso fue comprobar la estructura del archivo utilizando np.shape, validando que los datos cargados coincidían con el tamaño especificado.

Posteriormente, se aplicó un **filtro geográfico** para limitar el dataset a los países seleccionados: Argentina, Brasil, Chile, Colombia, México y Perú. Este filtro excluyó regiones y subregiones (identificadas por la columna location\_key) para centrarnos únicamente en datos a nivel nacional. También se restringieron los datos temporalmente, manteniendo solo los registros posteriores al **1 de enero de 2021**, periodo más relevante para analizar la evolución de la pandemia y las campañas de vacunación.

Durante la limpieza, se evaluó la cantidad de valores faltantes en cada columna. Se implementaron las siguientes estrategias:

* **Columnas con pocos valores nulos**: Los vacíos se rellenaron con la media de los valores disponibles.
* **Columnas con una alta proporción de nulos**: Los valores se reemplazaron por ceros, priorizando la conservación de la información existente.

Esta decisión permitió mantener información crucial, evitando eliminar columnas completas o registros con datos parciales. Se tomó en cuenta que los valores faltantes podrían completarse más adelante con información adicional.

Finalmente, los datos filtrados y transformados se guardaron en un nuevo archivo para su reutilización sin necesidad de repetir los pasos de limpieza.

### **Segundo Avance: Análisis Exploratorio de Datos (EDA)**

Una vez limpia la información, se profundizó en el análisis exploratorio para extraer insights clave. Para esto, se utilizaron técnicas estadísticas avanzadas, además de herramientas de visualización para identificar patrones, tendencias y anomalías.

#### **Correlaciones entre variables**

Se creó una matriz de correlación para detectar relaciones significativas entre variables. Utilizando un **heatmap**, se analizaron correlaciones mayores a |0.5|, identificando aquellas con asociaciones fuertes.

#### **Visualización de tendencias y patrones**

* **Barplots**: Permitieron comparar las variables seleccionadas entre los países. Por ejemplo, se observó que **Brasil**, debido a su tamaño poblacional, presentaba cifras notablemente superiores, mientras que **Chile** mostraba la menor incidencia de enfermedades como la diabetes y el tabaquismo.
* **Histogramas**: Revelaron la distribución de las tasas de vacunación y la incidencia de casos confirmados.
* **Gráficos de dispersión**: Relacionaron variables como la temperatura promedio con el número de casos confirmados, mostrando diferencias en la influencia de factores climáticos entre países.

Se realizó un análisis longitudinal, observando el comportamiento de las variables en el tiempo. Por ejemplo, se analizó la evolución mensual de las dosis de vacunas administradas, detectando un aumento general en los seis países, aunque con variaciones según factores como la disponibilidad de vacunas y la logística de distribución.

#### **Insights clave del EDA**

* **Chile** fue identificado como el país con mejor manejo de la pandemia, reflejado en una menor tasa de casos confirmados y decesos, además de un alto porcentaje de población vacunada.
* **Brasil** y **México** destacaron como los países con mayor número de casos confirmados y decesos, aunque con grandes diferencias en sus estrategias de vacunación.
* **Enfermedades preexistentes**: El tabaquismo y la diabetes fueron las condiciones más comunes asociadas a la mortalidad en los países analizados.

### **Avances finales: Presentación y Visualización en Power BI**

Se integraron los hallazgos analíticos en **Power BI** mediante la creación de un dashboard interactivo con 4 secciones:

1. **Data**: Visualización de los datos poblacionales y de vacunación por grupos etarios.
2. **Performance**: Métricas sobre evolución de casos confirmados, recuperados y decesos acumulados por país, además de mapas interactivos con porcentajes de vacunación.
3. **Índices**: Comparaciones de tasas de fatalidad y evolución mensual de casos.
4. **Expansión**: Índices compuestos que combinan la tasa de fatalidad con la proporción de vacunas administradas, destacando las áreas prioritarias para la expansión de laboratorios.

## **Conclusiones y Recomendaciones**

El análisis sugiere que la empresa BIOGENESYS debería priorizar su expansión en **Brasil**, **México** y **Colombia**, ya que presentan:

* **Altas tasas de fatalidad**.
* **Baja proporción de vacunas administradas** en comparación con la población.
* **Elevado número de casos confirmados y decesos acumulados**.

En contraste, **Chile** y **Perú** muestran mejores indicadores de vacunación, lo que los posiciona como países menos prioritarios en esta etapa de la estrategia.

## **Reflexión Final**

Este proyecto representó un desafío significativo, especialmente en la integración de técnicas avanzadas de análisis con herramientas de visualización. Aunque Python es indispensable para el procesamiento y limpieza de datos, Power BI resultó más intuitivo para la creación de dashboards interactivos, reflejando los insights de manera clara y accesible.

El aprendizaje obtenido en este proyecto demuestra el valor del análisis de datos en la toma de decisiones estratégicas, sentando las bases para futuros estudios que incluyan factores adicionales como la logística de distribución y la infraestructura sanitaria en los países seleccionados.