Carrera: Data Analytics
Módulo 4
Nombre del autor: Carlos Ivan Prieto Carrillo
Email: cipc333.1999@gmail.com
Cohorte: DAFT-12
Fecha de entrega: Abril 07, 2025

Análisis Estratégico para la Expansión de Laboratorios BIOGENESYS en Latinoamérica

Introducción

BIOGENESYS es una empresa farmacéutica enfocada en el desarrollo y distribución de vacunas, busca expandir su presencia en Latinoamérica. Como parte de su estrategia postpandemia, se ha propuesto realizar un análisis de datos sobre COVID-19, vacunación y condiciones sanitarias en seis países clave: Colombia, Argentina, Chile, México, Perú y Brasil.

El objetivo de este proyecto es proporcionar información basada en datos para ubicar estratégicamente nuevos laboratorios, optimizando así el acceso a las vacunas y fortaleciendo la respuesta sanitaria ante futuras emergencias epidemiológicas.

Primer Avance

Exploración del Dataset

Se realizó un análisis inicial del dataset "data_latinoamerica.csv" para comprender mejor las variables disponibles y su estructura. Se identificaron las siguientes variables clave para el estudio:

- **Casos confirmados**: `new_confirmed`, `cumulative_confirmed`
- **Muertes**: `new_deceased`, `cumulative_deceased`
- **Vacunación**: `cumulative_vaccine_doses_administered`
- **Población total**: `population` (para calcular tasas por millón de habitantes)

Filtrado de Datos

Dado que el análisis está enfocado en la expansión en Latinoamérica, se filtraron los datos para incluir exclusivamente los siguientes países:

- Colombia
- Argentina
- Chile
- México
- Perú
- Brasil

Además, se estableció un filtro temporal para considerar solo los registros con fecha posterior a `2021-01-01`.

Limpieza de Datos

Para garantizar la calidad del dataset, se realizaron las siguientes acciones de limpieza:

- Se definió una lista de columnas clave, excluyendo aquellas con más del 80% de valores nulos, y se eliminaron filas con más del 60% de valores ausentes.
- Imputación por grupo país
    - Clima: Se rellenaron valores usando el último valor conocido (`ffill`).
    - Demografía: Se usó la mediana por país.
    - Casos COVID: Se usó `ffill` y luego se reemplazaron nulos restantes con 0.
- Se aseguró que la columna de fechas estuviera en el formato adecuado con `pd.to_datetime`.
- Se confirmó que no quedaran valores faltantes en el dataset con `df.isnull().sum()`.
- Se guardaron los datos en un nuevo archivo `data_latinoamerica_limpio.csv`.

Estadísticas Descriptivas

Se aplicaron bucles para calcular manualmente estadísticas como media, mediana, desviación estándar, varianza y rango para variables clave. Algunos hallazgos relevantes incluyen:

- La media de nuevos casos confirmados es 29.46, pero la mediana es 0, lo que indica que la distribución es sesgada y que algunas regiones tienen valores extremadamente altos mientras que otras reportan pocos casos.
- En el caso de fallecimientos acumulados, la media (336.65) es considerablemente superior a la mediana (10), lo que sugiere que la mortalidad se concentra en ciertas áreas.
- La población presenta una media de 164,591, mientras que la mediana es 13,432, reflejando grandes diferencias demográficas entre países.

- El índice de desarrollo humano muestra una mediana de 0.709 y una varianza de 0.007, indicando diferencias moderadas en calidad de vida y acceso a servicios esenciales.
- La temperatura promedio tiene una mediana de 23.69°C, con un rango de 83.86°C, evidenciando una notable diversidad climática entre los países estudiados.

¿Qué implican estas métricas y cómo pueden ayudar en el análisis de datos?
Estas métricas permiten comprender la distribución de los datos, identificar patrones, detectar valores atípicos y tomar decisiones fundamentadas. Por ejemplo:

  - La media indica el valor promedio, aunque puede verse influenciada por outliers.
  - La mediana proporciona el valor central sin ser afectada por extremos.
  - La desviación estándar señala qué tan dispersos están los datos respecto a la media.
  - Mínimos y máximos ayudan a detectar errores o datos atípicos.

¿Qué representa la mediana?

La mediana es el valor central de un conjunto de datos, dividiendo la distribución en dos partes iguales. En el análisis realizado, se observan diferencias significativas entre la media y la mediana en variables como casos confirmados y fallecidos, lo que indica la presencia de valores extremos. Esto sugiere que la mayoría de las regiones tienen cifras bajas, pero hay algunas con valores mucho más altos.

¿Se muestran todas las estadísticas en todas las columnas?

No, ya que algunas métricas no aplican a variables categóricas o con valores constantes.

¿Qué nos indica esto sobre la consistencia o la variabilidad de los datos en relación con la mediana?

La alta varianza y los rangos amplios reflejan una **gran dispersión en los datos**, lo que sugiere que las condiciones sanitarias y la incidencia de COVID-19 son **altamente variables entre regiones**. Como la mediana es mucho menor que la media en varios indicadores (como fallecimientos acumulados), se confirma que la distribución está influenciada por valores atípicos, lo que puede afectar la toma de decisiones. Esto indica que la planificación de la expansión farmacéutica debe considerar estas diferencias regionales para una **respuesta más eficaz y equitativa**.

Segundo Avance

Con los datos limpios y filtrados del avance anterior, el segundo avance se enfocó en el análisis exploratorio utilizando visualizaciones para obtener patrones relevantes, comparar variables y facilitar la interpretación de la información.

Funciones Clave Utilizadas en Python

- **Matplotlib (`plt`)**: Configuración del tamaño de los gráficos (`plt.figure(figsize=...)`), etiquetas, títulos, grillas (`plt.grid(True)`) y visualización (`plt.show()`).
- **Seaborn (`sns`)**:
    - `sns.scatterplot()`: Para gráficos de dispersión.
    - `sns.boxplot()`: Para distribución de temperaturas.
    - `sns.violinplot()`: Para distribución del IDH.
    - `sns.barplot()`: Para gráficos de barras.
    - `sns.lineplot()`: Para evolución temporal.
- **Pandas (`pd`)**:
    - Conversión de fechas con `pd.to_datetime`.
    - Agrupación con `groupby()` y `sum()` o `mean()`.
    - Transformaciones como `.reset_index()` y `.T` (transposición).

Hallazgos Clave

El análisis reveló que los casos y muertes por COVID-19 en América Latina se concentraron en regiones con temperaturas templadas, sin una correlación directa clara, pero sí con una alta densidad en climas entre 10 °C y 30 °C. El índice de desarrollo humano (IDH) mostró valores predominantemente entre 0.65 y 0.85, lo que sugiere condiciones medias de desarrollo que podrían afectar la capacidad de respuesta sanitaria. La población está distribuida de manera equilibrada entre hombres y mujeres, y la mayoría de los países tienen una base poblacional joven-adulta (20-49 años), lo cual influye en la planificación de campañas de vacunación y asignación de recursos.

En cuanto a la evolución temporal, enero de 2022 marcó el pico de nuevos casos, con una disminución progresiva hacia 2023. Brasil, México y Argentina registraron los valores más altos en casos y muertes mensuales, lo que sugiere puntos críticos para la expansión sanitaria. También se evidenció una mayor tasa de mortalidad en hombres en todos los países, destacando Perú. Estos hallazgos permiten identificar regiones prioritarias para la instalación de laboratorios y orientar decisiones estratégicas de BIOGENESYS en función de factores climáticos, demográficos y epidemiológicos

Tercer Avance

Se realizó un análisis exploratorio de los datos disponibles. A través de la visualización y el procesamiento de variables clave como casos, muertes y vacunación, fue posible identificar patrones temporales, relaciones entre variables y diferencias significativas entre países

Funciones Clave Utilizadas en el Proyecto

- **Creación de variables nuevas** como `vaccination_coverage` y `mortality_rate` usando operaciones sobre columnas (`dosis_acumuladas / población`, `muertes / casos confirmados * 100`).

- **Transformaciones temporales** con `pd.to_datetime()` y `to_period("M")` para análisis por mes o año.

- **Agrupaciones y agregaciones** con `groupby()` y funciones como `.sum()`, `.mean()`, `.diff()` para calcular acumulados y variaciones.

- **Selección de países representativos** por cantidad de registros o relevancia en la región.

- **Visualización de datos** con `seaborn.scatterplot()`, `seaborn.lineplot()`, `plt.subplots()`, y `seaborn.heatmap()` para crear gráficos claros y comparativos.

- **Personalización de gráficos** con etiquetas, títulos, leyendas y escalas logarítmicas para facilitar la interpretación.

Hallazgos Importantes

El análisis exploratorio reveló patrones clave en la evolución y manejo de la pandemia en Latinoamérica. La cobertura de vacunación muestra una relación inversa con los casos nuevos, respaldando su efectividad, aunque algunos picos persistieron pese a altas tasas de vacunación. Chile y Brasil destacaron por una rápida implementación de campañas, mientras que Perú mostró un proceso más lento. El año 2021 fue el más crítico en casos, con Brasil liderando en impacto. La evolución mensual de casos y muertes evidenció la

gravedad del brote en Brasil, especialmente a mediados de 2021. México presentó la tasa de mortalidad más alta de forma sostenida, en contraste con Chile y Brasil, que tuvieron tasas más bajas y estables. Finalmente, el mapa de calor reveló correlaciones relevantes, como la fuerte relación entre casos y muertes, y una ligera correlación negativa entre desarrollo humano y mortalidad, sugiriendo que el contexto socioeconómico influyó en la capacidad de respuesta sanitaria.

Cuarto Avance

Durante este avance se revisaron y ajustaron posibles errores en las columnas generadas a lo largo del análisis exploratorio, asegurando la integridad y consistencia del dataset. Se verificaron tanto los nombres como la cantidad de columnas mediante el uso de `print(data_filtrada.columns)` y `print(data_filtrada.shape)`.
Una vez validada la estructura final del DataFrame, se procedió a guardar el archivo como `Data_Filtrada_Final.csv`, el cual está listo para ser utilizado en Power BI como base para el análisis visual y la generación de tableros interactivos.

Conclusiones

El análisis desarrollado en este proyecto permitió convertir una base de datos compleja en un conjunto de información estructurada, confiable y útil para la toma de decisiones estratégicas. Desde la limpieza y transformación de los datos hasta la creación de variables clave y visualizaciones exploratorias, se identificaron patrones relevantes en torno al impacto del COVID-19, la cobertura de vacunación, las condiciones sanitarias y las diferencias sociodemográficas entre países latinoamericanos.

Estos hallazgos ofrecen a BIOGENESYS una visión clara sobre qué regiones presentan mayores desafíos y cuáles podrían beneficiarse de la instalación de nuevos laboratorios, considerando variables críticas como la tasa de mortalidad, el desarrollo humano, y la respuesta vacunal.

Como producto final, se generó un archivo depurado (`Data_Filtrada_Final.csv`) que sirvió de base para la construcción de un dashboard interactivo en Power BI, también alojado en la misma carpeta del análisis. Esta visualización resume los principales indicadores y permite explorar los datos de forma dinámica, facilitando la interpretación y aplicación de los resultados por parte del equipo directivo. Con ello, se concluye el proceso analítico, dejando a disposición de BIOGENESYS una herramienta basada en evidencia para guiar su expansión en América Latina.