

**Carrera: Data Analytics**

**Módulo 4**

**Nombre del autor: Cristian Andres Herrera Rendón Email:** [cristianherrera1130@gmail](mailto:angelicasantamaria@live.com)**[.com](mailto:angelicasantamaria@live.com)**

**Cohorte: DAPT-03**

# INFORME

****

# Introducción

En el contexto actual de una pandemia global, **Biogenesys** se encuentra comprometida en ofrecer soluciones basadas en el análisis de datos para apoyar la toma de decisiones estratégicas en el ámbito de la salud pública. Este informe presenta un análisis detallado de los factores que influyen en la propagación del COVID-19 en varios países de América Latina, seleccionando específicamente a Argentina, Brasil, Chile, Colombia, México y Perú. A través de técnicas avanzadas de análisis y visualización de datos, se busca identificar patrones y correlaciones entre el comportamiento de la pandemia y variables ambientales y demográficas, como la temperatura, la humedad relativa y las cifras de vacunación.

El proyecto utiliza una amplia gama de herramientas para llevar a cabo una limpieza rigurosa y estandarización de datos, así como análisis estadísticos que permiten evaluar el impacto de los factores locales en el desarrollo de la pandemia. Esta labor no solo busca entender el pasado, sino también sentar las bases para modelos predictivos futuros que puedan anticipar brotes y apoyar a los gobiernos en la implementación de políticas de salud efectivas.

La metodología empleada por Biogenesys en este análisis respalda su compromiso con la innovación y el valor agregado de sus soluciones, buscando transformar datos complejos en información clara y útil. La aplicación de estas técnicas en el contexto de América Latina permite a Biogenesys posicionarse como un referente en la industria de la salud, promoviendo un enfoque regionalizado que optimiza la gestión de recursos y el desarrollo de estrategias específicas para cada país.

# Objetivos Generales

# Este informe tiene como propósito realizar un análisis detallado de datos relacionados con los países seleccionados en América Latina: Argentina, Chile, Colombia, México, Perú y Brasil. A través de un análisis exploratorio de datos (EDA), se identifican patrones y correlaciones entre variables clave que permiten comprender mejor la evolución de casos confirmados y otras métricas importantes en la región.

### 1. Carga y Exploración de Datos

**Descripción de la Carga de Datos**  
El primer paso en este análisis consistió en cargar el conjunto de datos desde un archivo CSV y verificar su estructura. Se usaron las librerías pandas y numpy para facilitar la manipulación y análisis de datos. La carga de datos iniciales incluyó la verificación de las primeras filas y la estructura (número de filas y columnas) del conjunto de datos para obtener una visión general de su tamaño y formato.

data = pd.read\_csv("data\_latinoamerica.csv", sep=',')

data.head()

np.shape(data)

**Selección de Países**  
Para centrarse en los países seleccionados, se realizó un filtro de los datos, limitando la información a Argentina, Chile, Colombia, México, Perú y Brasil. Esto redujo el conjunto de datos y permitió un análisis más específico de estos mercados.

paises\_seleccionados = ['Colombia', 'Argentina', 'Chile', 'Mexico', 'Peru', 'Brazil']

data\_latinoamerica = data[data['country\_name'].isin(paises\_seleccionados)]

### 2. Limpieza de datos

**Identificación de Valores Nulos**  
Uno de los pasos cruciales fue identificar y tratar los valores nulos en el conjunto de datos. Los datos faltantes pueden afectar negativamente el análisis, y por ello, se cuantificaron los valores nulos en cada columna. Se decidió establecer un umbral de 4.000.000 para identificar las columnas con más valores nulos.

data\_latinoamerica.isnull().sum()[data\_latinoamerica.isnull().sum()>4000000]

**Tratamiento de Datos Faltantes**  
Para columnas claves relacionadas con la pandemia (como new\_confirmed, new\_deceased, cumulative\_confirmed, etc.), los valores faltantes fueron completados utilizando el promedio específico de cada país. Este enfoque proporciona estimaciones razonables para datos faltantes, evitando sesgos al utilizar valores medios nacionales.

for pais in paises\_seleccionados:

data\_latinoamerica\_paises\_Fecha['new\_confirmed'] = data\_latinoamerica\_paises\_Fecha['new\_confirmed'].fillna(data\_latinoamerica\_paises\_Fecha[data\_latinoamerica\_paises\_Fecha['country\_name']==pais]['new\_confirmed'].mean())

Para ciertas columnas como new\_recovered, se rellenaron los valores faltantes con 0, ya que algunos países no registran esta métrica en cada fecha específica.

### 3. Análisis Estadístico Descriptivo

Se calcularon estadísticas descriptivas como la media, mediana, varianza y rango para todas las columnas numéricas. Esto permite una comprensión general de los valores centrales, variabilidad y dispersión en cada variable.

def describir\_datos(data):

for i in data.columns:

if data[i].dtype=='float64' or data[i].dtype=='int64':

print(i)

print('Mediana:',data[i].median())

print('varianza:',data[i].var())

print('Rango', data[i].max()-data[i].min())

print("Moda",data[i].mode())

### 4. Análisis de correlación

Se realizó un análisis de evaluación entre las variables numéricas del conjunto de datos, visualizado en un mapa de calor (heatmap) para destacar las relaciones con correlaciones superiores a 0.5 o inferiores a -0.5. Las relaciones más fuertes ayudan a identificar patrones significativos, como la posible asociación entre el número de casos nuevos y las variables climáticas (temperatura y humedad relativa).

correlation\_matrix = data\_latinoamerica\_paises[numeric\_columns].corr()

sns.heatmap(correlation\_matrix, mask=mask, annot=False, cmap='coolwarm', fmt=".2f")

### 5. Visualización de datos

**Gráficos de Barras para Comparación entre Países**  
Se utilizaron gráficos de barras para mostrar los datos agregados de cada país en variables clave como new\_confirmed, new\_deceased, y rainfall\_mm. Esta visualización permite comparar de manera efectiva la carga de cada país en términos de casos nuevos y factores ambientales, proporcionando una visión clara de la magnitud de los datos en cada región.

**Histogramas de Distribución**  
Se generaron histogramas para observar la distribución de variables como new\_confirmed, relative\_humidity, y average\_temperature\_celsius. Estos gráficos muestran la variabilidad y la concentración de cada métrica, ayudando a identificar sesgos en los datos y patrones de frecuencias.

**Gráfico de Dispersión (Scatter Plot)**  
El gráfico de dispersión entre average\_temperature\_celsiusy new\_confirmedpermite analizar la relación entre el clima y la propagación de la enfermedad. Este tipo de gráfico puede ser útil para detectar tendencias o correlaciones entre las condiciones ambientales y el número de casos confirmados.

sns.scatterplot(x='average\_temperature\_celsius', y='new\_confirmed', data=data\_latinoamerica\_paises, hue='country\_name')

### 6. Exportación de datos

El conjunto de datos limpio y filtrado fue exportado a un nuevo archivo CSV para garantizar la persistencia de los datos procesados ​​y permitir su uso en análisis futuros.

data\_latinoamerica\_paises\_Fecha.to\_csv("DatosFinalesFiltrado.csv", index=False)

**Conclusiones:**

**Altos Niveles de Valores Faltantes**

Se observará una gran cantidad de valores nulos en algunas columnas, especialmente en métricas ambientales. El relleno basado en promedios por país fue una decisión adecuada para reducir el impacto de estos valores sin introducir sesgos significativos.

**Correlación entre Variables**

El análisis de evaluación reveló asociaciones significativas entre varias métricas relacionadas con los casos de COVID-19 y factores climáticos, lo cual sugiere que condiciones como la temperatura y la humedad relativa podrían estar relacionadas con los niveles de propagación en cada país.

**Diferencias Regionales en Casos Nuevos y Condiciones Climáticas**

Las visualizaciones mostraron diferencias en el número de casos nuevos y las condiciones ambientales entre los países. Argentina y Brasil, por ejemplo, mostraron diferentes patrones en sus valores de new\_confirmed, lo cual podría reflejar variaciones en políticas de salud pública y condiciones demográficas.

**Importancia de Análisis Continuo**

Dado el comportamiento dinámico de los casos de COVID-19 y los cambios en las condiciones climáticas, es recomendable mantener un monitoreo continuo y agregar datos recientes a estos análisis para lograr una interpretación más precisa y relevante.

**Recomendaciones:**

**Implementación de Modelos Predictivos**

Con las relaciones establecidas, se podrían crear modelos predictivos que incluyan factores climáticos para anticipar el comportamiento de los casos nuevos de COVID-19 en la región.

**Monitoreo de Indicadores Ambientales y de Salud Pública**

Se recomienda utilizar el análisis de evaluación para identificar rápidamente cambios en las condiciones que puedan afectar los casos confirmados, lo cual puede ser útil para la toma de decisiones en salud pública.

**Automatización del Proceso de Limpieza de Datos**

Dado el alto volumen de valores nulos en este tipo de datos, una automatización en el proceso de limpieza permitiría un análisis más eficiente y una mayor precisión en los resultados.

**Conclusiones y Aprendizajes del Proyecto**

**Conclusiones principales**

**Impacto de las Condiciones Ambientales en la Propagación del COVID-19**

La compensación entre variables ambientales (como la temperatura y la humedad relativa) y el número de casos nuevos sugiere que las condiciones climáticas pueden influir en la propagación de la enfermedad. Países con climas variados como Argentina y Brasil presentan diferencias en los patrones de contagio, lo cual puede deberse en parte a estas condiciones ambientales. Esto subraya la importancia de considerar factores climáticos en los modelos epidemiológicos de la región.

**Desigualdad en los Datos y Registros entre Países**

Se detectaron diferencias significativas en la disponibilidad y calidad de los datos entre los países analizados. Algunos países muestran un registro completo de métricas relacionadas con la pandemia, mientras que otros presentan grandes cantidades de datos faltantes, particularmente en métricas ambientales y casos de recuperación. Esto puede estar vinculado a diferencias en la infraestructura de salud y en las políticas de registro, lo cual plantea desafíos para lograr una comparación precisa y homogénea.

**Valor de la Limpieza y Estandarización de Datos**

La limpieza de los datos fue esencial para manejar la gran cantidad de valores nulos y posibles inconsistencias. La estrategia de rellenar valores faltantes con promedios específicos por país evitó sesgos significativos, mientras que la eliminación de datos muy incompletos mantuvo la calidad general del análisis. Este proceso permite asegurar que las métricas derivadas sean representativas y que las visualizaciones y modelos de evaluación se basen en datos confiables.

**Patrones Regionales en el Comportamiento de la Pandemia**

Las visualizaciones demostraron patrones específicos en cada país, con Brasil y Argentina mostrando un volumen de casos particularmente alto en comparación con otros países. Estas diferencias pueden reflejar variaciones en políticas públicas, densidad poblacional y la efectividad de las medidas de control de salud implementadas. Este análisis regional puede facilitar el desarrollo de estrategias personalizadas de respuesta a la pandemia.

**Importancia de Monitoreo Continuo y Actualización de los Datos**

Dado el comportamiento dinámico del COVID-19 y las condiciones cambiantes en cada país, el análisis debe actualizarse periódicamente para reflejar las nuevas tendencias y ajustes en las políticas de cada país. Este proyecto demostró que un monitoreo continuo puede mejorar la precisión de los modelos predictivos y las estrategias de respuesta a la pandemia.

**Aprendizajes Clave**

Importancia de una Estrategia de Limpieza de Datos Rigurosa

La calidad del análisis depende en gran medida de la integridad y consistencia de los datos. Este proyecto demuestra que una estrategia de limpieza y estandarización efectiva no solo mejora la precisión del análisis, sino que también facilita la identificación de patrones válidos. Este aprendizaje destaca la importancia de implementar procesos automatizados de limpieza y verificación de datos en futuros proyectos.

**Capacidad para Extraer Información Relevante de Datos Incompletos**

Aprendimos que, aun con una cantidad significativa de datos faltantes, es posible extraer información relevante al utilizar técnicas de imputación, como el relleno de promedios específicos por país. Esta práctica ayuda a minimizar la pérdida de información sin introducir sesgos significativos, un aprendizaje valioso para futuros análisis donde la integridad de los datos no esté completamente asegurada.

**Uso de Visualizaciones para Interpretación Rápida y Eficaz**

Las visualizaciones facilitaron la interpretación de grandes volúmenes de datos y permitieron la identificación de patrones clave. Gráficos como mapas de calor y gráficos de dispersión fueron esenciales para observar correlaciones y comportamientos entre variables. Este aprendizaje reafirma la importancia de las visualizaciones como herramienta en el análisis exploratorio de datos para la toma de decisiones.

**Necesidad de un Enfoque Regionalizado en el Análisis de Pandemias**

Las diferencias observadas entre los países en América Latina resaltan la importancia de analizar la pandemia desde una perspectiva regionalizada. Las estrategias de respuesta deben ser específicas para cada país, tomando en cuenta las condiciones locales y el contexto de cada región. Este enfoque es aplicable no solo en estudios de pandemia, sino en cualquier análisis que incluya múltiples variables contextuales.

**Preparación de Datos para Modelos Predictivos**

La estructuración de los datos a través de su limpieza, normalización y análisis de correlación permite una base sólida para la construcción de modelos predictivos. El proyecto enseñó que un análisis exploratorio adecuado puede revelar variables claves que deben considerarse en modelos futuros, especialmente en análisis de tendencias y predicción de casos.

En resumen, este proyecto proporcionó una profunda comprensión de los factores que afectan la propagación de COVID-19 en América Latina y proporcionó herramientas para mejorar los análisis futuros. La integración de estos aprendizajes ayudará a desarrollar estrategias más efectivas para gestionar y anticipar situaciones en el ámbito de la salud pública.