**Expansión Estratégica de Biogenesys con Python**

**Nombre del autor: Laura O.**

**Email:**

**Cohorte: DRAFT-06**

**Fecha de entrega:** 20-09-2024

**Institución:**

****

**Introducción**

Este informe presenta un análisis estratégico para Biogenesys, una empresa farmacéutica enfocada en la expansión de laboratorios en Latinoamérica. El proyecto se centra en identificar las ubicaciones óptimas para futuros laboratorios farmacéuticos y centros de vacunación en países clave como Colombia, Argentina, Chile, México, Perú y Brasil. El análisis se basa en datos sobre la incidencia del COVID-19, tasas de vacunación y la disponibilidad de infraestructura sanitaria.

El objetivo principal es optimizar la respuesta de Biogenesys ante situaciones de pandemia y postpandemia, mejorando el acceso a las vacunas en estas regiones. Para lograrlo, se llevará a cabo un análisis exploratorio de datos, limpieza de los datos relevantes, y se desarrollarán dashboards interactivos para facilitar la toma de decisiones estratégicas. Este informe proporciona insights que permitirán a la directiva de Biogenesys evaluar oportunidades para una expansión exitosa.

**Desarrollo del proyecto**

**Avance 1**

#### Metodología de Recopilación y Selección de Datos

1. **Recopilación de Datos**:
   * El archivo original del dataset tiene un tamaño aproximado de 21 GB y contiene alrededor de 22 millones de registros y 707 columnas. Debido a su magnitud, se creó un dataset específico por parte del equipo de ingeniería de datos, centrado en los países de Latinoamérica.
   * Este nuevo archivo en formato CSV, titulado data\_latinoamerica.csv, incluye 12,216,057 filas y 50 columnas, optimizando así la manipulación y el análisis de los datos pertinentes a la región.
2. **Selección de Datos**:
   * Se llevó a cabo una revisión del archivo readme.txt, que proporciona información detallada sobre las columnas seleccionadas y su significado, permitiendo un mejor entendimiento del contenido del dataset.
   * Los países seleccionados para el análisis son Colombia, Argentina, Chile, México, Perú y Brasil.
3. **Filtrado de Datos**:
   * Se filtraron los registros para incluir únicamente aquellos con fechas posteriores al 1 de enero de 2021, asegurando la relevancia temporal de la información.

#### Transformaciones y Limpieza de Datos

1. **Verificación de Datos**:
   * Se comprobó que el dataset cargado contuviera la cantidad de registros y columnas especificadas, utilizando funciones de verificación en Python.
2. **Manejo de Valores Faltantes**:
   * Se realizó un análisis a nivel de país para identificar y llenar los valores faltantes, utilizando técnicas como la imputación con valores medios, valores anteriores o siguientes según corresponda.
3. **Limpieza Preliminar**:
   * Se eliminaron registros nulos y se corrigieron los tipos de datos donde era necesario, asegurando que todas las variables estuvieran en el formato adecuado para el análisis.
4. **Análisis de Características Básicas**:
   * Se examinaron las características básicas del dataset para comprender la distribución de variables clave, como la incidencia de COVID-19, y se identificaron aquellas variables consideradas clave para el análisis posterior.
5. **Guardado de Datos Filtrados**:
   * Los datos filtrados se guardaron en un archivo titulado DatosFinalesFiltrado.csv, permitiendo su uso en etapas posteriores del proyecto sin necesidad de repetir el proceso de filtrado y limpieza.
6. **Cálculo de Estadísticas Descriptivas**:
   * Se aplicaron bucles for y/o while para el cálculo de estadísticas descriptivas y otras métricas ofrecidas por la librería pandas de forma predeterminada.
   * Se analizó la implicancia de estas métricas en el análisis de datos, y se exploró si todas las estadísticas estaban disponibles para cada columna, así como la razón detrás de cualquier limitación en la presentación de estas estadísticas.
7. **Funciones para Métricas**:
   * Se creó una función personalizada para calcular la mediana, varianza y rango de las variables relevantes. Se analizaron las medidas en función de su representación y su relación con la consistencia y variabilidad de los datos.
8. **Uso de Funciones de Orden Superior** (Extra Credit):
   * Se exploró el uso de funciones de orden superior para facilitar la manipulación eficiente de los datos, aplicando métodos como map, filter y reduce para mejorar la legibilidad y eficacia en el manejo de datos.

Se logró cargar un conjunto de datos significativo, reduciendo su tamaño a un formato manejable que facilitó el análisis posterior. Se filtraron los datos para incluir solo aquellos relevantes para los países de Latinoamérica y se eliminaron registros nulos, asegurando así la calidad de la información. La revisión inicial de las columnas permitió identificar las variables más relevantes para el análisis de la incidencia de COVID-19 y la planificación de la expansión de laboratorios. Además, se calcularon métricas descriptivas que proporcionan una visión general del estado de la pandemia y la efectividad de las respuestas sanitarias en los países seleccionados.

**Avance 2**

#### Metodología de Recopilación y Selección de Datos

1. **Recopilación de Datos**:
   * En esta fase del proyecto, se utilizan los datos previamente limpios y filtrados en el avance 1, que incluyen información relevante sobre la incidencia de COVID-19, tasas de vacunación y disponibilidad de infraestructura sanitaria en países de Latinoamérica.
2. **Selección de Datos**:
   * Se centra el análisis en las variables críticas relacionadas con la salud pública, las cuales fueron identificadas durante el avance anterior. Esto incluye la incidencia de COVID-19, tasas de vacunación y características demográficas relevantes.

#### Transformaciones y Limpieza de Datos

1. **Análisis Estadístico con Pandas y Numpy**:
   * Se realiza un análisis estadístico exploratorio para calcular medidas de tendencia central (media, mediana) y dispersión (desviación estándar, rango) de las variables seleccionadas.
   * Se examinan las correlaciones entre variables utilizando métodos estadísticos apropiados, lo que permite entender mejor la relación entre la incidencia de COVID-19 y otros factores, como tasas de vacunación y temperatura.
2. **Visualización de Datos con Matplotlib y Seaborn**:
   * Se generan diversas visualizaciones para representar los hallazgos de manera clara y comprensible:
     + **Histogramas y Gráficos de Densidad**: Para entender la distribución de la incidencia de COVID-19 y tasas de vacunación.
     + **Gráficos de Barras**: Para comparar diferentes regiones y visualizar las dosis administradas de vacunas por país.
     + **Mapas de Calor**: Para identificar correlaciones entre diferentes variables de interés.
     + **Gráficos de Dispersión**: Para explorar posibles relaciones entre variables, como temperatura media contra casos confirmados y muertes.
     + **Matriz de Correlación**: Para observar las relaciones entre múltiples variables, con un enfoque en aquellos valores que superan 0.5.
     + **Boxplots y Violinplots**: Para visualizar la distribución de variables con cambios significativos.
3. **Identificación de Tendencias y Patrones**:
   * Se analizan los datos relacionados con la incidencia de COVID-19 y las tasas de vacunación, buscando identificar tendencias a largo plazo y patrones estacionales o geográficos.
   * Se investiga la influencia de factores ambientales, como la temperatura, en la propagación de la enfermedad y se examinan las variaciones en las tasas de vacunación en diferentes regiones y períodos del año.
4. **Personalización de Visualizaciones**:
   * Se personalizan las visualizaciones para mejorar su claridad e impacto visual. Esto incluye la modificación de paletas de colores, adición de etiquetas y leyendas, ajuste del tamaño de los gráficos, y utilización de títulos descriptivos que faciliten la comprensión de los datos representados.

Se llevaron a cabo análisis exploratorios profundos de los datos, revelando patrones, tendencias y anomalías en la incidencia de COVID-19 y las tasas de vacunación. Las visualizaciones generadas ayudaron a comunicar hallazgos clave de manera clara y efectiva, facilitando la identificación de áreas de intervención prioritaria. Se identificaron tendencias a largo plazo y patrones estacionales que son fundamentales para la planificación estratégica de la expansión de laboratorios farmacéuticos. Además, el análisis de correlaciones proporcionó información valiosa sobre la relación entre la cobertura de vacunación y la reducción de casos, lo que puede guiar decisiones futuras en la asignación de recursos.

**Avance 3**

#### Metodología de Recopilación y Selección de Datos

1. **Recopilación de Datos**:
   * Se utilizaron los datos previamente filtrados y limpios del avance anterior, asegurando que la información mantenía su relevancia y calidad.
2. **Selección de Datos**:
   * El enfoque se centró en variables clave como la incidencia de COVID-19, tasas de vacunación, condiciones demográficas y otros factores que son cruciales para la expansión de laboratorios farmacéuticos en Latinoamérica.

#### Análisis Exploratorio de Datos (EDA)

1. **Análisis Detallado**:
   * Se llevó a cabo un análisis exploratorio detallado utilizando técnicas avanzadas de Pandas y Numpy, enfocándose en la evolución de los datos a lo largo del tiempo. Se identificaron tendencias, estacionalidad y patrones temporales que son relevantes para entender la propagación del COVID-19.
2. **Generación de Gráficas**:
   * Se generaron diversas gráficas que ilustran la evolución de casos, tasas de crecimiento y la relación entre variables clave. Estas visualizaciones facilitaron la comprensión de la situación actual y las necesidades en las áreas estudiadas.
3. **Investigación de Correlaciones**:
   * Se examinaron posibles correlaciones entre las diferentes características del conjunto de datos. Esto permitió identificar relaciones significativas que pueden influir en la planificación de la expansión de laboratorios.
4. **Preparación para el Modelado**:
   * Se realizó una limpieza y estructuración de los datos, asegurando que estén listos para análisis más complejos, incluyendo modelado predictivo y análisis geoespacial. Este paso es crucial para ofrecer recomendaciones basadas en datos sólidos que guiarán las decisiones de inversión de la empresa.
5. **Exploración Geoespacial (Opcional)**:
   * Como parte del desafío extra, se propuso la creación de mapas utilizando la información de latitud y longitud, lo que enriquecería el análisis al proporcionar una perspectiva geográfica de los datos analizados.

Se realizó un análisis exploratorio detallado utilizando técnicas avanzadas de Pandas y Numpy, enfocándose en la evolución de los datos a lo largo del tiempo. Se identificaron tendencias, estacionalidad y patrones temporales relevantes para entender la propagación del COVID-19. Se generaron diversas gráficas que ilustran la evolución de casos, tasas de crecimiento y la relación entre variables clave, facilitando la comprensión de la situación actual y las necesidades en las áreas estudiadas. Asimismo, se investigaron posibles correlaciones entre las diferentes características del conjunto de datos, lo que permitió identificar relaciones significativas que pueden influir en la planificación de la expansión de laboratorios. Finalmente, se realizó una limpieza y estructuración de los datos, asegurando que estén listos para análisis más complejos, incluyendo modelado predictivo y análisis geoespacial.

### Avance 4

* **Recopilación de Datos:**
  + El archivo original del dataset tiene un tamaño aproximado de 21 GB y contiene alrededor de 22 millones de registros y 707 columnas.
  + Debido a su magnitud, se creó un dataset específico por parte del equipo de ingeniería de datos, centrado en los países de Latinoamérica.
  + Este nuevo archivo en formato CSV, titulado "DatosFinalesFiltrado.csv", incluye 12,216,057 filas y 50 columnas, optimizando así la manipulación y el análisis de los datos pertinentes a la región.
* **Selección de Datos:**
  + Se filtraron los datos para incluir solo registros relevantes desde enero de 2021.
  + Se priorizaron variables relacionadas con la incidencia de COVID-19 y tasas de vacunación en países de Latinoamérica.

Las transformaciones realizadas incluyeron la limpieza exhaustiva del dataset, que permitió reducir significativamente su tamaño y mejorar la calidad de los datos. Se corrigieron tipos de datos y se imputaron valores faltantes, lo que facilitó la integración de los datos en Power BI. Este enfoque asegura que los insights generados sean precisos y relevantes para la toma de decisiones estratégicas en la expansión de laboratorios y centros de vacunación. La implementación de dashboards interactivos en Power BI proporciona una visualización clara y accesible, permitiendo a los directivos explorar y comprender la información de manera intuitiva.

**EDA e insights**

**Análisis Exploratorio de Datos (EDA)** El EDA se centró en analizar el comportamiento de los casos de COVID-19 en América Latina a lo largo del tiempo. Se agruparon los datos semanal y anualmente para observar tendencias en la evolución de los casos confirmados. Los siguientes insights fueron destacados:

1. **Tendencias Semanales y Anuales**: Se observó que, a pesar de la disminución de casos en general, todavía hay picos en ciertas semanas, indicando que la pandemia no ha terminado y que las olas pueden ser influenciadas por factores estacionales o políticas públicas.
2. **Impacto de la Vacunación**: La comparación entre dosis de vacunas administradas y nuevos casos confirmó que la vacunación tiene un efecto positivo en la reducción de casos activos y fallecimientos. Sin embargo, países como Colombia y Brasil mostraron un desempeño deficiente en la vacunación, lo que correlacionó con mayores tasas de mortalidad.
3. **Relación con Factores Socioeconómicos**: Se analizó la densidad de población y la urbanización, mostrando que los países con mayor urbanización y densidad, como Chile, experimentaron una propagación más rápida del virus.
4. **Distribución por Edad**: Se identificó que ciertos grupos de edad presentan una mayor vulnerabilidad, lo que sugiere que las políticas de vacunación deben enfocarse en estos segmentos para mitigar la mortalidad.

**Análisis del dashboard**

**Navegación del Dashboard en Power BI** El dashboard desarrollado en Power BI permite visualizar y explorar interactivamente los datos de COVID-19 en América Latina. Las características clave incluyen:

* **Visualizaciones Dinámicas**: Gráficos interactivos que muestran la evolución de casos nuevos, muertes y vacunaciones a lo largo del tiempo. Al seleccionar diferentes países, los usuarios pueden comparar rápidamente las métricas entre ellos.
* **Filtros y Segmentaciones**: Los usuarios pueden filtrar por país, fecha y otros indicadores, permitiendo un análisis más detallado según las necesidades específicas.
* **Insights Rápidos**: Cada gráfico incluye un resumen de insights clave, destacando la relación entre vacunación y reducción de casos, así como las diferencias entre países.

**Posibles Conclusiones del Dashboard** El dashboard permite concluir que las políticas de vacunación y el comportamiento social tienen un impacto directo en la evolución de la pandemia. Los datos resaltan la necesidad de continuar con campañas de vacunación y el seguimiento de las tendencias en grupos demográficos vulnerables.

**Conclusiones y Recomendaciones**

**Síntesis de Resultados** El análisis integral de datos y visualizaciones ha revelado que, aunque la situación de COVID-19 en América Latina ha mejorado, se deben mantener vigilancias constantes ante posibles rebrotes. Las políticas de salud pública deben seguir ajustándose en función de los datos disponibles para garantizar una respuesta efectiva.

**Conclusiones Estratégicas**

1. **Mejora en Políticas de Vacunación**: Se recomienda aumentar la inversión en campañas de vacunación, especialmente en países con baja cobertura, como Colombia y Brasil. La creación de centros de vacunación y la promoción de la vacunación en comunidades vulnerables son cruciales.
2. **Monitoreo de Tendencias**: Establecer un sistema de monitoreo continuo que permita la identificación temprana de picos de casos y el ajuste rápido de las políticas sanitarias.
3. **Enfoque en Grupos Vulnerables**: Priorizar la vacunación y atención médica en grupos de edad que han mostrado mayores tasas de mortalidad, así como en áreas urbanas densamente pobladas.
4. **Educación y Concientización**: Implementar campañas educativas para informar sobre la importancia de la vacunación y las medidas de prevención, especialmente en países con altas tasas de escepticismo hacia la vacunación.

Al implementar estas recomendaciones, se espera mejorar la gestión de la pandemia y preparar mejor a los países de América Latina para futuras crisis sanitarias.

**Reflexión personal**

Durante este proyecto, he aprendido a abordar problemas complejos con un enfoque estructurado en el análisis de datos. He mejorado mis habilidades en la manipulación, limpieza y visualización de grandes conjuntos de datos, lo cual fue clave para identificar patrones y tendencias en las regiones objetivo de expansión. Además, desarrollar dashboards interactivos me permitió presentar los resultados de manera clara y accesible para la toma de decisiones estratégicas. Este proyecto me ha reforzado la importancia de contar con datos de calidad y bien estructurados para poder ofrecer insights valiosos y confiables.

Si tuviera que volver a empezar este proyecto, probablemente planificaría con mayor profundidad la fase de limpieza de datos y la creación del pipeline ETL. Este paso resultó ser más crítico de lo que inicialmente anticipé, y dedicar más tiempo a la preparación de los datos hubiera facilitado el análisis posterior. Además, consideraría la implementación de técnicas de machine learning para prever posibles escenarios futuros en base a los datos históricos, lo que habría añadido un valor predictivo al análisis realizado. Sin embargo, en general, estoy satisfecho con los resultados obtenidos y las habilidades que he desarrollado a lo largo del proyecto.

**EXTRA CREDIT**

**Mejoras del Proyecto**

1. **Visualización Geoespacial**:
   * Se implementaron mapas utilizando las coordenadas de latitud y longitud disponibles en los datos, permitiendo una exploración visual de la distribución geográfica de los casos de COVID-19 en América Latina. Esto proporciona una perspectiva clara sobre cómo se distribuyen los contagios, recuperaciones y vacunaciones en diferentes regiones.
2. **Uso de Herramientas Avanzadas**:
   * Se utilizó Folium para crear mapas interactivos y dinámicos que representan los casos confirmados, muertes, vacunas administradas y recuperaciones. Esto permite a los usuarios explorar los datos de manera más intuitiva, facilitando la identificación de áreas críticas y patrones geográficos.
3. **Comparación de Mapas de Calor**:
   * Se generaron múltiples mapas de calor superpuestos, cada uno representando una métrica diferente (confirmados, muertes, vacunas y recuperados). Esta comparación visual ayuda a entender la relación entre estos indicadores y permite identificar correlaciones en tiempo real.
4. **Análisis de Casos Confirmados por País**:
   * Utilizando Geopandas y Matplotlib, se elaboró un mapa temático que ilustra los casos confirmados acumulados por país, ofreciendo un contexto adicional que complementa la información de los mapas de calor. Esto ayuda a los tomadores de decisiones a visualizar el impacto de la pandemia en cada nación.
5. **Mejora de la Interactividad**:
   * Para enriquecer la experiencia del usuario, se pueden agregar elementos interactivos en los mapas, como pop-ups con información detallada sobre cada país al hacer clic, lo que proporciona datos adicionales sin sobrecargar la visualización.
6. **Futuras Integraciones**:
   * Se sugiere explorar otras herramientas de visualización geoespacial, como Plotly o ArcGIS, para incorporar funciones avanzadas como análisis de clústeres o integración de datos históricos, lo que permitirá un análisis más profundo de la evolución de la pandemia.
7. **Actualización Dinámica de Datos**:
   * Implementar un flujo de trabajo automatizado para actualizar los datos en tiempo real y reflejar los cambios en las visualizaciones, asegurando que la información sea relevante y precisa en todo momento.
8. **Documentación y Accesibilidad**:
   * Proporcionar documentación clara y accesible sobre cómo utilizar y navegar los mapas, asegurando que tanto usuarios técnicos como no técnicos puedan aprovechar al máximo las visualizaciones.