Universidad Católica Boliviana “San Pablo”

Sede Santa Cruz

Exploración de Indicadores Socioeconómicos Mundiales

Asignatura: Adquisición, Análisis y Procesamiento de datos

Docente: Msc. Ing. Marin Salazar Carmen Rosa

Estudiante: Nicole Lozada León

Fecha: 8 de abril de 2025

1. Introducción

En un mundo cada vez más interconectado, la comprensión de los indicadores socioeconómicos a nivel global se vuelve esencial para la formulación de políticas efectivas y el desarrollo sostenible. Este proyecto explora diversas métricas que reflejan el estado económico y social de diferentes países, utilizando un enfoque basado en datos. A través de la recolección y análisis de información relevante, se busca identificar patrones y tendencias que informen decisiones estratégicas en el ámbito socioeconómico.

Para llevar a cabo este análisis, se emplean variadas técnicas de recolección de datos, incluyendo web scraping, acceso a APIs y el consumo de archivos planos. Estas metodologías permiten compilar un conjunto de datos robusto que abarca indicadores clave como la densidad poblacional (Densidad P/Km²), la esperanza de vida (EV), el acceso a Internet, la inflación y el Producto Interno Bruto (PIB) de 2023, además de clasificar la información por continente.

Una vez recopilados los datos, se realiza un proceso exhaustivo de limpieza y transformación. Este paso asegura la calidad y la integridad de la información, al involucrar la unión de múltiples archivos CSV, la gestión de valores faltantes, y la estandarización de nombres de países y formatos de datos. A través de esta fase, se establece una base sólida para el análisis posterior.

Finalmente, se lleva a cabo un Análisis Exploratorio de Datos (EDA) que permite descubrir hallazgos significativos y visualizaciones que facilitan la comprensión de las diferencias socioeconómicas entre países y regiones. Con este enfoque, se contribuye al debate sobre cómo los indicadores económicos y sociales pueden influir en el desarrollo global y en la formulación de políticas públicas efectivas.

2. Recolección de Datos

La recolección de datos constituye un elemento fundamental en el desarrollo de este proyecto, ya que permite construir un conjunto de información sólido y confiable. El proceso comenzó con una búsqueda exhaustiva de archivos planos que contuvieran indicadores socioeconómicos relevantes. Tras una revisión cuidadosa, se encontró un archivo CSV en Kaggle titulado “Google Country Information Dataset 2023”, creado por el Data Scientist Nidula Elgiriyewithana. Este conjunto de datos está disponible bajo la licencia “Attribution 4.0 International”, lo que permite su uso y distribución con el debido reconocimiento.

De este conjunto, se decidió utilizar la columna de densidad poblacional (Densidad P/Km²) debido a que no presentaba valores nulos, lo que garantiza la integridad de los datos. Este indicador se consideró especialmente valioso, ya que obtener información

actualizada sobre densidad poblacional a través de otros métodos, como web scraping y APIs, resultaría complicado y poco confiable.

Posteriormente, se realizó una búsqueda en la página web de IndexMundi, donde se accedió a la columna de esperanza de vida, la cual se basa en fuentes confiables de Estados Unidos. Esta elección fue estratégica, ya que esta fuente también carecía de valores nulos, lo que la hacía adecuada para el análisis en comparación con la información disponible en otros sitios, que a menudo presentaba inconsistencias.

El siguiente paso consistió en buscar datos sobre el Producto Interno Bruto (PIB). Para ello, se consultó una página web del Fondo Monetario Internacional (FMI), que es reconocida por su credibilidad y por la calidad de la información que proporciona. La utilización de esta fuente asegura que los datos sobre PIB sean actualizados y precisos, lo que es crucial para el análisis socioeconómico.

Finalmente, se recurrió a la API de Databank para acceder a la información sobre el acceso a Internet y la inflación. Durante este proceso, se presentó un desafío relacionado con el idioma: los nombres de los países estaban inicialmente en inglés, lo que dificultaba su integración con otras fuentes de datos en español. Para resolver este inconveniente, se utilizó la librería de Deep Translator, que facilitó la traducción de los nombres de los países, asegurando así la coherencia en el conjunto de datos final.

3. Limpieza y Transformación

La limpieza y transformación de datos es un paso crítico que garantiza la calidad y la coherencia del conjunto de información utilizado en este proyecto. Como se mencionó en la sección de recolección, se priorizó la utilización de columnas que no presentaran valores nulos, dado que los datos provienen de diferentes fuentes y abarcan años que varían entre 2020 y 2023. Esta variabilidad en las fuentes y los años de los datos introduce desafíos que deben ser abordados cuidadosamente.

Uno de los principales problemas que se encontró durante el proceso de limpieza fue la estandarización de los nombres de los países. Se observó que algunos nombres estaban escritos con tildes, como “Bangladés”, mientras que otros utilizaban una representación sin tilde, como “Bangladesh”. Esta inconsistencia dificultaba la integración de los datos, ya que los nombres de los países debían coincidir exactamente para poder realizar un análisis efectivo.

Para resolver este problema, se llevó a cabo una comparación manual de la columna “país” en cuatro dataframes diferentes. Este proceso permitió identificar los nombres que estaban fuera de lugar y requerían corrección. Dado que el número de países con discrepancias no superaba los diez por dataframe, se optó por reescribir manualmente los nombres incorrectos para asegurar la precisión.

Una vez corregidos los nombres, se generó un diccionario en el que la llave correspondía al nombre del país y el contenido al continente al que pertenece. Este diccionario resultó ser una herramienta valiosa para facilitar la integración de los diferentes dataframes. A través de la función `merge`, se unieron los dataframes, incorporando la información del continente a cada entrada correspondiente.

Finalmente, se añadió la columna de continente al dataframe principal, denominado "data".

4. EDA

En esta fase del proyecto, se llevó a cabo un análisis exhaustivo de los indicadores socioeconómicos para extraer información significativa y patrones relevantes.

Primero, se obtuvo la media, mediana y desviación estándar de las variables cuantitativas del dataframe. Este resumen estadístico proporciona una visión general del comportamiento de cada indicador, permitiendo identificar tendencias y variaciones dentro de los datos.

A continuación, se generaron histogramas y diagramas de caja (boxplots) para cada variable cuantitativa. Estos gráficos no solo facilitan la visualización de la distribución de los datos, sino que también permiten identificar outliers de manera efectiva. Durante este análisis, se observó un hallazgo interesante en el histograma de la esperanza de vida: existían países con edades por debajo de los 60 años. Esto suscitó una investigación más profunda, y se determinó que el país con la menor esperanza de vida era Chad, un país africano. Este dato contrasta notablemente con Japón, que tiene una esperanza de vida cercana a los 85 años, lo que resalta las disparidades en salud y bienestar entre diferentes regiones del mundo.

Adicionalmente, se analizó la matriz de correlación entre las variables. Este análisis reveló una fuerte relación positiva entre el acceso a Internet y la esperanza de vida. Esta correlación sugiere que a medida que aumenta el acceso a Internet, también tiende a aumentar la esperanza de vida en los países, lo que podría reflejar la influencia de la tecnología en la mejora de la calidad de vida y el acceso a servicios de salud.

En conjunto, estos hallazgos no solo aportan una comprensión más profunda de los indicadores socioeconómicos analizados, sino que también plantean preguntas importantes sobre las políticas de salud y desarrollo que podrían implementarse para abordar las disparidades observadas.

5. Conclusiones

El análisis exploratorio de datos realizado sobre los indicadores socioeconómicos mundiales ha revelado disparidades significativas en la esperanza de vida entre

diferentes países. Chad, con una esperanza de vida por debajo de los 60 años, destaca como un caso crítico que refleja los desafíos en salud que enfrenta. En contraposición, Japón, con casi 85 años de esperanza de vida, simboliza un contexto donde el acceso a servicios de salud es más efectivo. Estas diferencias subrayan la importancia de entender los factores que contribuyen a tales desigualdades y la necesidad de políticas que aborden estas realidades.

Además, se ha observado una fuerte correlación positiva entre el acceso a Internet y la esperanza de vida. Este hallazgo sugiere que el desarrollo tecnológico y la conectividad pueden influir de manera directa en la salud y el bienestar de la población. Promover el acceso a Internet es, por lo tanto, un componente crucial para mejorar la calidad de vida en diversas regiones, ya que puede facilitar el acceso a información vital y servicios de salud.

El proceso de identificación de outliers a través de histogramas y diagramas de caja ha permitido detectar valores atípicos que merecen un análisis más profundo. Comprender las causas detrás de estos outliers no solo enriquece el análisis, sino que también proporciona oportunidades para desarrollar intervenciones específicas que aborden los problemas socioeconómicos subyacentes.

6. Recomendaciones

Para este estudio, se utilizó la información de 125 países, lo cual proporciona una base sólida para el análisis, pero también limita la capacidad de generalización de los hallazgos. Por lo tanto, una recomendación clave es expandir el conjunto de datos incluyendo información de más países. Esta ampliación no solo enriquecería el análisis, sino que también permitiría identificar tendencias y patrones más representativos a nivel global.

Además de aumentar la cantidad de países, es fundamental considerar la inclusión de datos históricos para cada indicador. Esto permitiría realizar análisis de tendencias a lo largo del tiempo, facilitando la identificación de cambios significativos y el impacto de políticas implementadas en diferentes regiones.

También se sugiere diversificar las fuentes de datos, explorando bases de datos de organizaciones internacionales, como la Organización Mundial de la Salud (OMS), el Banco Mundial y otros organismos que monitorean indicadores socioeconómicos. La triangulación de datos de diversas fuentes puede mejorar la fiabilidad y la precisión de la información recopilada.

Por último, es recomendable implementar un sistema de actualización periódica de los datos, para asegurar que la información utilizada en futuros análisis esté siempre al día. Esto es especialmente importante en un contexto global en constante cambio, donde

los indicadores socioeconómicos pueden variar rápidamente debido a factores como crisis económicas, pandemias o cambios en políticas públicas.

7. Anexos

Anexo 1 “Gráfico de correlación”