Herramientas de Big Data
Proyecto final

Santiago Carrillo (285761)
Alejandro Gómez (274547)
Juan José Rodríguez (289227)

Herramientas de Big Data
Universidad de La Sabana
Carrera: Ingeniería Informática
Profesor: Hugo Franco

Septiembre 2025

6 de septiembre de 2025

Resumen Ejecutivo

El presente proyecto se desarrolló en el marco de la asignatura Coterminal de la Maestría en Analítica Aplicada, con el propósito de aplicar herramientas de Big Data a la optimización de la gestión y visualización de la inversión pública en Colombia. La problemática identificada se centra en la dificultad de acceder de manera ágil a grandes volúmenes de información, los procesos manuales de auditoría y la limitada transparencia de los datos disponibles para la ciudadanía.

Para abordar esta situación, se integraron fuentes de datos provenientes de bases relacionales on-premise (Oracle y SQL Server) y de servicios en la nube, utilizando procesos de ingesta mediante SSIS, modelado con SSAS y visualización con Power BI. La metodología permitió consolidar información de múltiples sistemas, garantizando coherencia en las variables críticas y facilitando la generación de indicadores, gráficos interactivos y filtros de búsqueda.

Los resultados muestran mejoras sustanciales en los tiempos de procesamiento y en la accesibilidad de la información, lo cual habilita a los ciudadanos para consultar proyectos por filtros y mapas interactivos, y a las entidades de control para disponer de información centralizada y confiable. El sistema implementado no solo optimiza la gestión de datos, sino que también fortalece la transparencia y la capacidad de toma de decisiones informadas.

En conclusión, este proyecto constituye un avance significativo en la aplicación práctica de herramientas de Big Data para la gestión pública, aportando soluciones que responden a necesidades reales de eficiencia, control y acceso democrático a la información.

Abstract

This study examines the relationship between digital connectivity and economic growth, measured by Gross National Income (GNI) per capita, through a comparative analysis of multiple countries. Open-access datasets from the World Bank Group were processed using Python-based tools for data cleaning, visualization, and statistical modeling. The methodological approach included data extraction for the period 2020–2024, normalization, correlation analysis, and clustering through the K-Means algorithm.

The results revealed a positive and consistent association between Internet penetration and GNI per capita: countries with higher connectivity tend to show higher income levels. Cluster analysis distinguished three categories of economies: (1) low-income countries with limited digital access, (2) emerging economies with moderate levels of income and connectivity, and (3) developed economies with high income and near-universal access. Temporal analysis highlighted that Internet growth is particularly pronounced in emerging economies, while developed ones show signs of saturation. Furthermore, strong correlations were confirmed between digital infrastructure (secure servers and broadband subscriptions) and economic performance.

In conclusion, the findings demonstrate that digital inclusion is a key driver of economic development. Reducing the digital divide can contribute not only to narrowing income gaps but also to fostering sustainable growth, particularly in developing and emerging economies.

Introducción

En esta sección se desarrollan los principales elementos que orientan el estudio. Primero, se presenta la formulación del problema y las necesidades de información asociadas. Luego, se expone el marco conceptual, donde se definen los términos y variables clave. Posteriormente, se revisan los antecedentes a partir de trabajos previos relacionados con la temática. Finalmente, se establecen los objetivos del proyecto, tanto generales como específicos, que guían el análisis realizado.

Formulación del problema y las necesidades de información asocia das

Comprender cómo los países con mayor desarrollo económico han alcanzado ese nivel constituye una de las grandes preguntas de la humanidad, pues este conocimiento permite orientar a los gobiernos en la definición de estrategias para impulsar su propio crecimiento. En este contexto, el presente estudio se centra en analizar la relación entre el desarrollo económico —medido a través del Ingreso Nacional Bruto (GNI)— y el porcentaje de acceso a internet y a servicios de conectividad. La pregunta que guía esta investigación es: ¿De qué manera el porcentaje de acceso a internet y servicios de conectividad contribuye al crecimiento económico de los países, medido a través del GNI?

Para responder a este interrogante es necesario establecer un marco comparativo entre el nivel de conectividad de los países y su desempeño económico, identificando hasta qué punto el internet puede actuar como un motor de desarrollo en el contexto actual, marcado por la digitalización. Asimismo, se requiere tener en cuenta que los países analizados enfrentan realidades diversas y que existen factores externos —como la estabilidad política, la calidad institucional, la distribución de la riqueza o el nivel educativo— que también influyen en la relación entre conectividad y crecimiento económico. Estos elementos deben ser considerados para reducir sesgos en el análisis y contextualizar adecuadamente los resultados.

Marco conceptual

El análisis se fundamenta en la relación entre el acceso a internet y el desarrollo económico, por lo cual es necesario precisar los conceptos clave.

Acceso a internet: se mide como el porcentaje de la población que utiliza internet, ya sea mediante banda ancha fija, redes móviles u otras tecnologías. Este indicador refleja inclusión digital y el grado de participación de una sociedad en la economía global (International Telecommunication Union [ITU], s.f.-a).

Servicios de conectividad: comprenden la infraestructura que permite acceso y comunicación digital eficiente y segura. Entre ellos se destacan:

Broadband (banda ancha): conexión de alta velocidad que facilita la transmisión de grandes volúmenes de datos y habilita procesos productivos, educativos y de innovación tecnológica. Su medición proviene de la base de datos World Telecommunication/ICT Indicators Database (ITU, s.f.-b).

Internet Secure Servers (servidores seguros de internet): servidores que

utilizan protocolos de seguridad como SSL/TLS para proteger transacciones digitales. La densidad de estos servidores se obtiene a partir de la Secure Server Survey realizada por Netcraft (Netcraft, s.f.).

Ingreso Nacional Bruto (GNI): indicador económico que corresponde al total de ingresos generados por los residentes de un país, sumando el PIB y los ingresos netos desde el exterior. Es ampliamente utilizado para clasificar países en categorías de ingreso (Banco Mundial, s.f.).

Brecha digital: desigualdad en el acceso y uso de las TIC, que limita la participación en la economía digital y restringe oportunidades de desarrollo (ITU, 2022).

Antecedentes

Diversos estudios han explorado la relación entre el acceso a internet, la infraestructura digital y el crecimiento económico. La literatura coincide en señalar que la conectividad digital se constituye como un factor clave para la productividad y el desarrollo inclusivo.

Por ejemplo, Koutroumpis (2019) encontró que la penetración de banda ancha fija tiene un efecto positivo y estadísticamente significativo en el crecimiento del PIB, destacando su papel como infraestructura esencial en economías modernas. En la misma línea, Qiang et al. (2009) demostraron que un aumento del 10 % en la penetración de banda ancha se asocia con un crecimiento adicional del 1,21 % en economías desarrolladas y del 1,38 % en países en desarrollo.

En cuanto a la seguridad digital, indicadores como el número de servidores seguros se han empleado para aproximar la confianza en transacciones electrónicas, elemento indispensable para el comercio electrónico y la economía digital. Un estudio del Banco Mundial (2016) subraya que la confianza digital es determinante para la adopción de servicios financieros en línea y el dinamismo del sector servicios.

Por otra parte, la brecha digital ha sido ampliamente discutida como un factor que amplía desigualdades. La Unión Internacional de Telecomunicaciones (ITU, 2022) resalta que, pese al incremento global en acceso a internet, persisten brechas entre países de bajos y altos ingresos, lo que limita el potencial de desarrollo económico en los primeros.

Metodológicamente, los estudios previos han utilizado modelos econométricos, análisis de correlación y técnicas de clustering para agrupar países con

trayectorias digitales y económicas similares. Estas aproximaciones permiten tanto cuantificar relaciones como clasificar contextos comparables, lo cual es consistente con el enfoque adoptado en este proyecto.

En conjunto, la evidencia previa respalda la hipótesis de que la conectividad digital —medida en acceso, banda ancha y seguridad— se asocia positivamente con el desarrollo económico, aunque su magnitud varía según el nivel de ingreso y el contexto institucional de los países.

Objetivos Del Proyecto

Se estableció un objetivo general y cuatro objetivos específicos en relación a este.

Objetivo general

Analizar la relación entre el acceso y los servicios de conectividad digital con el crecimiento económico medido en el Ingreso Nacional Bruto (GNI) per cápita, mediante una comparación entre países.

Objetivos Específicos

1. Medir los niveles de acceso a internet en distintos países y su evolución en los últimos años.
2. Examinar cómo la conectividad digital contribuye al crecimiento económico en términos de GNI per cápita.
3. Evaluar la importancia de la infraestructura y la seguridad digital en el desarrollo económico.
4. Comparar patrones de conectividad y desempeño económico entre países con diferentes niveles de ingreso, identificando brechas y oportunidades.

Datos Empleados

Los datos utilizados en este estudio fueron extraídos de archivos en formato CSV provistos por el *World Bank Group*, entidad internacional creada en 1944 con el objetivo de reducir la pobreza y promover el desarrollo sostenible mediante la provisión de financiamiento, asistencia técnica y generación de conocimiento. Actualmente, recopila y publica indicadores económicos, sociales y tecnológicos de alcance global, ampliamente utilizados en la investigación académica y la formulación de políticas públicas (World Bank Group, 2023).

Variables consideradas para el análisis:

- **Nombre de países**: identificación de las unidades de observación.
- **GNI (Ingreso Nacional Bruto)**: valor total de los ingresos obtenidos por los residentes de un país, sumando el PIB y los ingresos netos recibidos del exterior.
- **GNI per capita**: ingreso nacional bruto dividido entre la población, utilizado como medida de bienestar económico promedio.
- **GNI growth**: tasa de crecimiento anual del GNI.
- **Percent Individuals Using Internet**: porcentaje de la población con acceso a internet.
- **Secure Internet Servers (per 1 million people)**: número de servidores seguros por cada millón de habitantes, indicador de la infraestructura de seguridad digital.
- **Fixed Broadband Subscriptions (per 1 million people)**: número de suscripciones a banda ancha fija por cada millón de habitantes, indicador de conectividad de alta velocidad.

En el marco del estudio, las variables dependientes corresponden al GNI y GNI per capita, como medidas del desarrollo económico. Las variables independientes corresponden al porcentaje de individuos que usan internet, la densidad de servidores seguros y las suscripciones de banda ancha fija, como indicadores de conectividad digital.

Materiales y Métodos

Materiales

El presente estudio se fundamentó en bases de datos abiertas del *Banco Mundial*, específicamente aquellas relacionadas con el Ingreso Nacional Bruto (GNI) en sus modalidades total, per cápita y tasas de crecimiento, así como con indicadores de conectividad digital, incluyendo porcentaje de usuarios de Internet, número de servidores seguros y penetración de banda ancha (World Bank Group, 2023).

Los archivos en formato CSV fueron procesados en un entorno de desarrollo Python (versión 3.9), utilizando las siguientes librerías: *Pandas* para la gestión y limpieza de datos, *Matplotlib* y *Seaborn* para la visualización, *Scikit-learn* para el análisis estadístico y los algoritmos de *clustering*, y *SQLAlchemy* para la integración con bases de datos.

La infraestructura de almacenamiento se implementó en un sistema gestor de bases de datos PostgreSQL 16, desplegado mediante contenedores Docker a través de un archivo `docker-compose.yml`. La orquestación del flujo de trabajo de extracción, transformación y carga (ETL) se realizó con la herramienta *Prefect*.

Métodos

El procedimiento metodológico contempló tres fases: extracción, transformación y análisis de los datos.

1. **Extracción**: Se recolectaron las series históricas correspondientes al período 2020–2024 desde el portal de datos del *Banco Mundial*.
2. **Transformación**:
    - Se efectuó una depuración de los registros, priorizando los países miembros del G20 y el caso particular de Colombia.
    - Los valores ausentes fueron tratados mediante interpolación lineal y sustitución por promedios, garantizando la coherencia temporal de las series.
    - Posteriormente, se realizó la normalización de variables y la consolidación de las bases en tablas integradas.

3. **Análisis**:

   - Se efectuó un análisis exploratorio de datos, con la construcción de matrices de correlación y representaciones gráficas para identificar relaciones entre indicadores económicos y digitales.
   - Para el análisis de patrones, se aplicó un modelo de clustering K-Means. El número óptimo de grupos fue determinado a través del método del codo y validado con el índice de silueta.
   - Finalmente, los resultados fueron representados gráficamente y almacenados en la base de datos PostgreSQL para su posterior consulta y análisis comparativo.

Resultados

El análisis gráfico y estadístico arrojó varios hallazgos relevantes:

- La relación entre penetración de Internet y GNI per cápita de la gráfica 6 resultó ser positiva y consistente. Los países con mayor conectividad tienden a mostrar mayores ingresos.

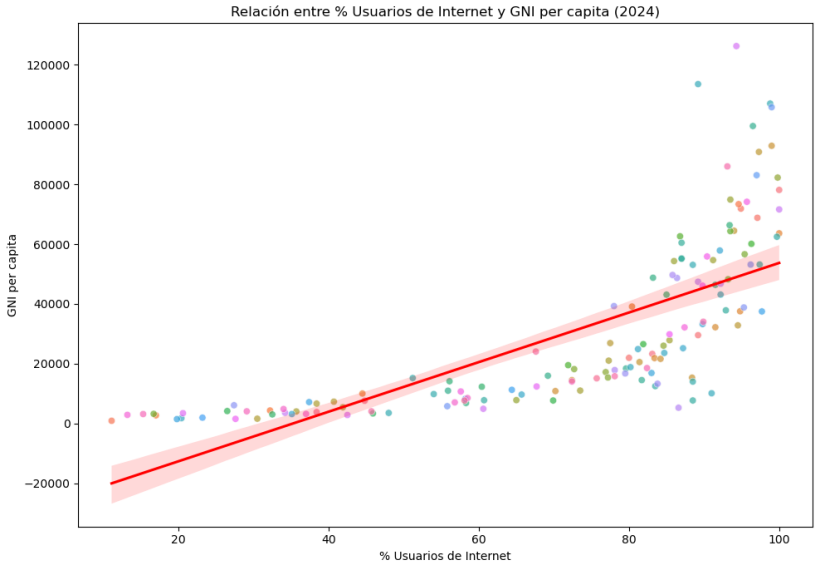

Figura 1: Grádifo scatter plot entre el porcentaje de usuarios de internet y el GNI Per Capita de los países alrededor del mundo.

- El análisis de clustering mostrado en la 6 identificó tres grupos de países:

    1. Economías con bajos ingresos y bajo acceso a Internet.
    2. Economías emergentes con niveles medios en ambas variables.
    3. Economías desarrolladas con altos ingresos y acceso casi universal a Internet.

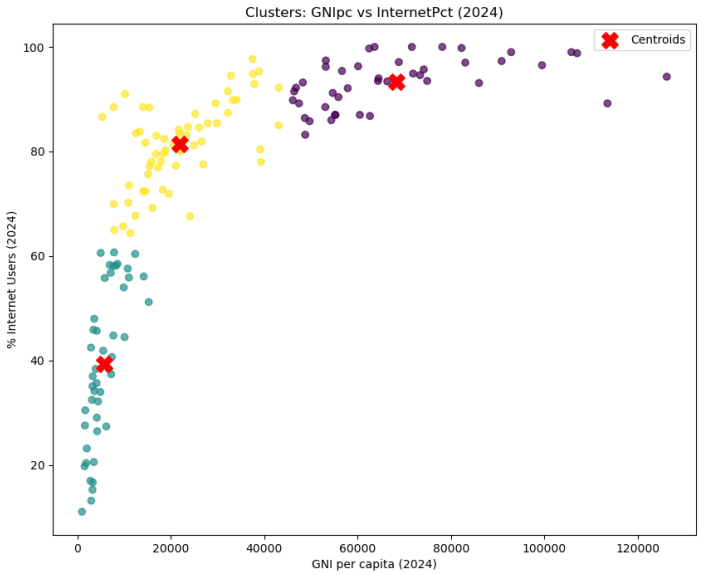

Figura 2: Clúser clasificando los países por GNI y baja conectividad.

- En la evolución del GNI per cápita frente al porcentaje de individuos que utilizan Internet, representada en 6, se evidencia una clara relación entre el nivel de desarrollo económico y la conectividad digital.

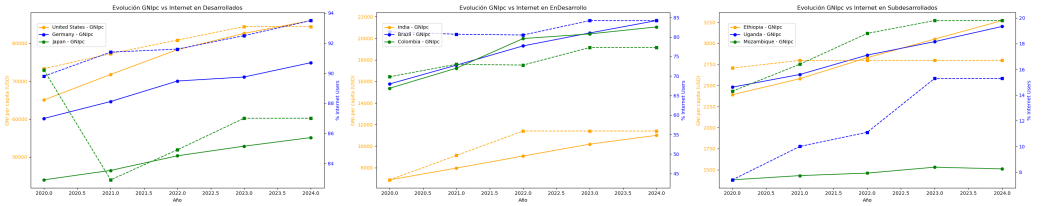

Figura 3: Gráficos de comparación porcentaje de usuarios de internet y GNI per cápita clasificado por el nivel de desarrollo. Las líneas punteadas representan el porcentaje de usuarios de internet y la sólida el GNI per cápita

En los países desarrollados, los niveles de acceso a Internet superan el 85 %, mientras que en los países en desarrollo se observa una tendencia ascendente que los acerca progresivamente a dicho umbral. Por el contrario, en los países subdesarrollados los porcentajes se mantienen por debajo del 20 %, reflejando limitaciones estructurales en infraestructura digital y acceso a tecnologías de información.

Un caso particular lo constituye India, que pese a ser considerada una economía emergente, presenta un nivel de penetración de Internet cercano al 55 %. Este valor relativamente bajo se explica por la magnitud de su población y las desigualdades socioeconómicas que persisten en el país.

- En los diagramas de barras del 6 que comparan los niveles de suscripciones de banda ancha fija y la disponibilidad de servidores seguros de Internet, se evidencia una marcada diferencia entre los grupos de países.

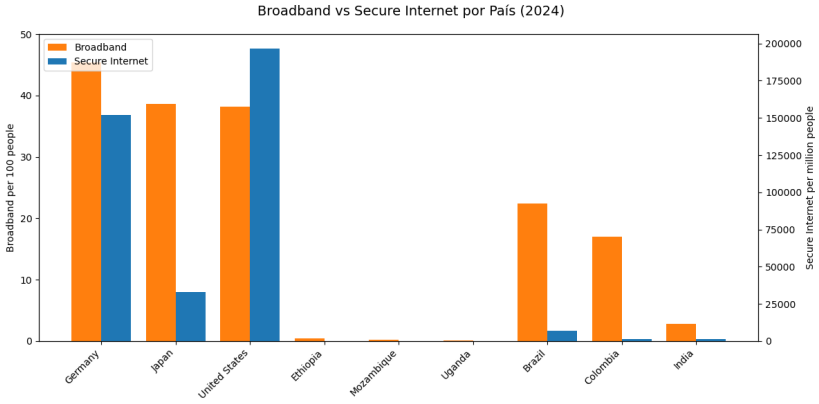

Figura 4: Diagrama de barras analisando el nivel de ancha banda(broadband) y de servicios seguros de internet por países.

En las economías desarrolladas, ambos indicadores presentan valores significativamente más altos, reflejando una infraestructura digital sólida y segura. En contraste, los países en vía de desarrollo muestran un crecimiento progresivo en el acceso a banda ancha; sin embargo, el nivel de servidores seguros sigue siendo reducido, lo que limita el fortalecimiento de la confianza digital y la seguridad en transacciones en línea.

Finalmente, en los países subdesarrollados, tanto la penetración de banda ancha como la densidad de servidores seguros son prácticamente nulas en comparación con los demás grupos, lo cual refuerza la existencia de una amplia brecha digital y tecnológica.

- La matriz de correlación de la 6 confirmó asociaciones positivas entre las variables digitales y económicas. Destaca la fuerte relación entre el GNI per cápita y el porcentaje de usuarios de Internet, así como con infraestructura de banda ancha y servidores seguros.

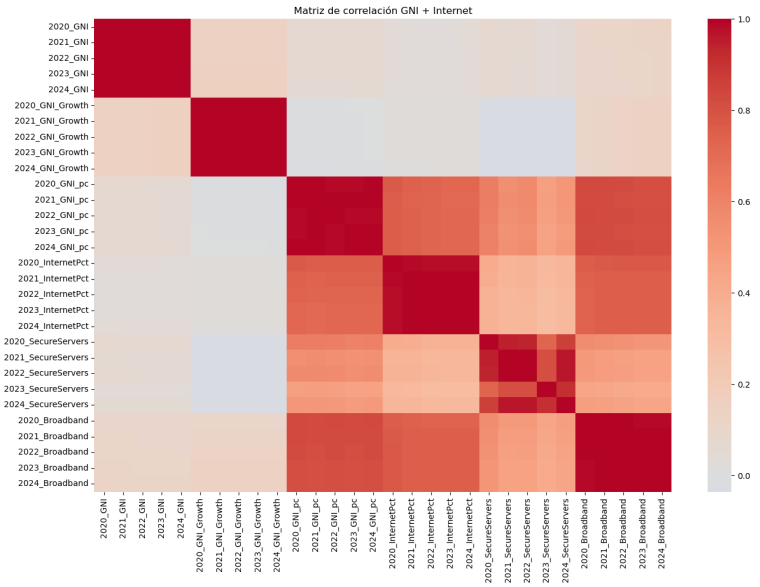

Figura 5: Matriz correlación, variables de estudio.

- La distribución de crecimiento (%) de la 6 reflejó que Internet presenta mayor variabilidad y dispersión, mientras que el GNI per cápita crece de manera más estable y homogénea entre países.

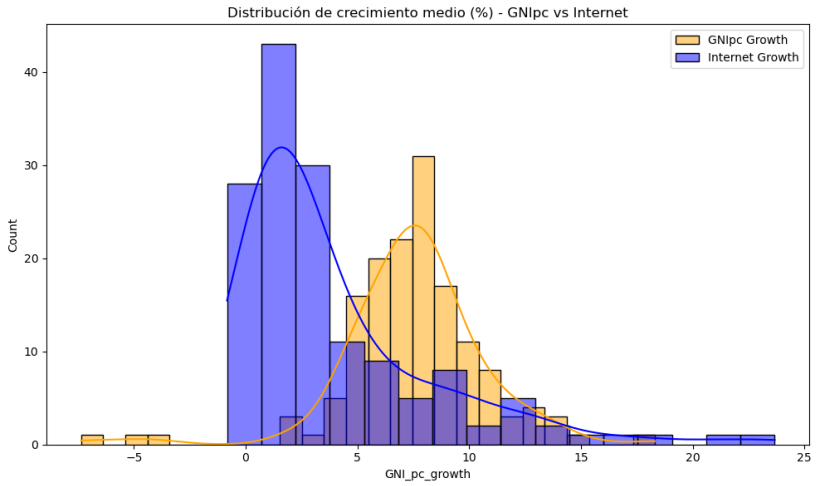

Figura 6: Distribución de crecimiento, de porcentage de individuos con internet y GNI per capita.

Discusión y Conclusiones

Los resultados evidencian la interdependencia entre digitalización y desarrollo económico. A mayor acceso a Internet y mayor disponibilidad de infraestructura tecnológica, los países presentan un nivel más alto de ingresos. Esta relación sugiere que la inclusión digital puede ser un catalizador del crecimiento económico, especialmente en economías emergentes.

Los resultados además constataron que los países desarrollados no solo presentan un mayor porcentaje de usuarios con acceso a Internet, sino que también cuentan con servicios complementarios que fortalecen la competitividad de sus empresas y ciudadanos en la actual era digital. Entre estos destacan la amplia disponibilidad de banda ancha y la presencia de servidores seguros de Internet, elementos clave para la innovación, la productividad y la seguridad en las transacciones en línea.

Por su parte, los países en vía de desarrollo se encuentran en un proceso de transición, con avances significativos en conectividad, aunque aún con limitaciones en la consolidación de servicios digitales avanzados. En contraste, los países subdesarrollados permanecen rezagados, sin acceso suficiente a esta infraestructura, lo que amplía la brecha digital y restringe sus oportunidades de crecimiento económico y social.

El *clustering* permitió distinguir tres trayectorias de desarrollo digital-económico, lo cual es útil para clasificar a los países y orientar políticas públicas. Los países del primer grupo (bajos ingresos y baja conectividad) enfrentan el reto de invertir en infraestructura digital, mientras que los del tercer grupo deben enfocarse en mantener la innovación y seguridad tecnológica.

En conclusión:

- Existe una correlación positiva y sólida entre la penetración digital y el GNI per cápita, lo que confirma que el acceso a Internet es un factor clave en el desarrollo económico.
- La brecha digital refleja también una brecha económica: los países con menor acceso a Internet tienden a tener ingresos más bajos y limitadas oportunidades de crecimiento.
- La inversión en infraestructura digital no solo contribuye a reducir desigualdades tecnológicas, sino que también impulsa el crecimiento económico sostenible y fortalece la inclusión social.

- Los países desarrollados destacan no solo por sus altos niveles de acceso a Internet, sino también por la consolidación de servicios complementarios —como la amplia disponibilidad de banda ancha y la densidad de servidores seguros— que incrementan la competitividad de empresas y ciudadanos en la economía digital.

- Los países en vía de desarrollo muestran avances importantes, aunque aún enfrentan limitaciones en servicios de seguridad y en infraestructura tecnológica robusta, lo cual ralentiza su transición hacia una economía plenamente digital.

- Los países subdesarrollados permanecen rezagados, con niveles muy bajos tanto en acceso a Internet como en servicios digitales complementarios, lo que profundiza la brecha tecnológica y económica frente a las demás regiones del mundo.

Referencias

International Telecommunication Union. (s.f.-a). *Individuals using the Internet (% of population)*. World Telecommunication/ICT Indicators Database. Recuperado de `https://datahub.itu.int/`

International Telecommunication Union. (s.f.-b). *Fixed broadband subscriptions*. World Telecommunication/ICT Indicators Database. Recuperado de `https://datahub.itu.int/`

Netcraft. (s.f.). *Secure Server Survey*. Netcraft. Recuperado de `https://www.netcraft.com/`

Banco Mundial. (s.f.). *GNI (current US$)*. World Bank Data. Recuperado de `https://data.worldbank.org/indicator/NY.GNP.MKTP.CD`

International Telecommunication Union (ITU). (2022). *Measuring digital development: Facts and Figures 2022*. ITU. `https://www.itu.int/en/ITU-D/Statistics/Pages/facts/default.aspx`

Koutroumpis, P. (2019). The economic impact of broadband: Evidence from OECD countries. *Technological Forecasting and Social Change, 148*, 119719. `https://doi.org/10.1016/j.techfore.2019.119719`

Qiang, C. Z., Rossotto, C. M., & Kimura, K. (2009). Economic impacts of broadband. In *Information and Communications for Development 2009: Extending Reach and Increasing Impact* (pp. 35–50). World Bank. `https://openknowledge.worldbank.org/handle/10986/2636`

Banco Mundial. (2016). *World Development Report 2016: Digital Dividends*. World Bank. https://www.worldbank.org/en/publication/wdr2016

Anexo A. Archivos del Proyecto

El proyecto incluye dos archivos principales disponibles en el repositorio de GitHub:

- `docker-compose.yaml`: archivo con la configuración del contenedor Docker utilizado para ejecutar la solución.
- `CodigoProyectoFinal.py`: archivo en Python que contiene el flujo ETL (Extract, Transform, Load), análisis exploratorio y clustering.

A continuación, se muestra un ejemplo representativo de ambos archivos.

Ejemplo de configuración Docker (`docker-compose.yaml`)

```
version: "3.9"
services:
  postgres:
    image: postgres:14
    container_name: postgres_bd
    restart: always
    environment:
      POSTGRES_USER: psqluser
      POSTGRES_PASSWORD: psqlpass
      POSTGRES_DB: bigdatatools1
    ports:
      - "5433:5432"
    volumes:
      - ./data:/var/lib/postgresql/data
```

Ejemplo de código Python (`CodigoProyectoFinal.py`)

```
@flow(name="ETL GNI + Internet + Clusters")
def etl_gni_internet():
    gni, gni_growth, gni_pc, internet_pct, internet_secure, broadband = extract_da
    gni_df, internet_df = transform_data(
```

```
        gni, gni_growth, gni_pc, internet_pct, internet_secure, broadband
    )
    cluster_df = analyze_data(gni_df, internet_df)
    load_data(gni_df, internet_df, cluster_df)

if __name__ == "__main__":
    etl_gni_internet()
```

El código completo y la configuración detallada se encuentran en el repositorio de GitHub asociado al proyecto.