Universidad Nacional Autónoma de México

Instituto de Investigaciones en
Matemáticas Aplicadas y en Sistemas

Población y planeta: crecimiento
demográfico y sostenibilidad de recursos

Proyecto Final

Calidad y Preprocesamiento de Datos

P R E S E N T A :

- Carlos Emiliano Mendoza Hernández
- Erick Yair Aguilar Martínez
- Imanol Mendoza Sáenz de Buruaga
- Luis Enrique Villalon Pineda

Profesores

- MCIC Cinthia Rodríguez Maya
- MCIC Víctor Manuel Corza Vargas

Ciudad Universitaria, Cd. Mx., 2025

iimas

1. Introducción

En un mundo que cada vez es más gobernado por la cantidad excesiva de datos, la calidad de la información es un pilar esencial para la toma de decisiones importantes. En particular dentro de decisiones de gobierno, donde las políticas públicas, estrategias de desarrollo y acciones de emergencia requiere de una excelente calidad y procesamiento de datos, por lo que la integridad de los datos no es una ventaja técnica so no se vuelve una condición necesaria en el gobierno.

La era donde los datos son abiertos se a multiplicad, el acceso a estadísticas nacionales o internacionales como World Bank Open Data o UN Population Data Portal nos ofrece miles de indicadores que abarcan distintas variables como el cambio de la población, la cobertura forestal, las emisiones de carbono o el uso agrícola de la tierra. Sin embargo , el uso de este tipo de datos nos lleva a ver que existen seires temporales incompletas, diferencias en las definiciones entre fuentes, registros duplicados, ausencia de metadatos claros o desfases en las actualizaciones son solo algunos de los problemas que nos pueden llevar a un mal análisis de datos.

Además, la calidad de los datos no es un atributo uniforme, por lo que debe evaluarse en distintas dimensiones, entre las que podemos destacar completitud, exactitud, consistencia, validez, oportunidad, accesibilidad y relevancia. De modo que cada dimensión afecta de distinta manera para decir que los datos son confiables y poder hacer su análisis. Por ejemplo, en una serie una variable ambiental con registros fuera de rango podría llevar a decisiones erróneas sobre zonas de conservación prioritaria.

Ahora bien el problema se agrava cuando estas deficiencias no son reportadas , detectadas o incluso corregidas, lo cual como ya mencionamos es un problema que puede afectar no solo a la empresa o gobierno si no también a toda la población que este vinculada. En estos casos, contar con algo que permita identificar, monitorear y corregir fallas de calidad se convierte en una prioridad. La implementación de prácticas de perfilado, limpieza, estandarización y monitoreo continuo de los datos puede marcar la diferencia entre una política exitosa y una intervención fallida.

La calidad de los datos no es un objetivo secundario ni una fase opcional en el ciclo de vida de los datos; es una condición necesaria para garantizar que el conocimiento generado sea sólido, que las políticas públicas sean eficaces y que las acciones implementadas tengan un impacto real y medible. Por lo que a lo largo de este documento, vamos a realizar metodologías, herramientas y buenas prácticas que permitan solucionar y tener una buena calidad de datos.

2. Marco teórico de la calidad de datos

La **calidad de los datos** se define como el grado en que un conjunto de datos satisface los requisitos necesarios para su uso previsto. Para evaluarla de forma sistemática, se recurre a marcos conceptuales como el propuesto por DAMA-DMBOK, donde se establecen dimensiones fundamentales que permiten valorar los datos de manera estructurada y accionable.

2.1. Dimensiones clave de la calidad de los datos

- **Completitud:** Mide la proporción de datos observados respecto al total esperado. Su ausencia afecta la continuidad de las series temporales, visualizaciones y modelos predictivos, especialmente en contextos como estudios demográficos o ambientales.
- **Exactitud:** Indica qué tan bien reflejan los datos la realidad subyacente. Puede validarse indirectamente mediante comparación entre fuentes (por ejemplo, tasas de natalidad ONU vs. Banco Mundial).
- **Consistencia:** Evalúa si los datos son coherentes entre variables, años o fuentes. Inconsistencias como valores contradictorios entre densidad poblacional y crecimiento natural pueden señalar problemas en definiciones o errores de registro.
- **Validez:** Verifica que los valores respeten reglas de tipo y dominio (e.g., porcentajes entre 0–100 %, fechas válidas, códigos ISO3). Es clave para prevenir errores estructurales antes del análisis.
- **Oportunidad:** Evalúa el desfase entre la generación de los datos y su disponibilidad para su uso. Es crítica en entornos donde se requiere respuesta rápida, como emergencias ambientales o sanitarias.
- **Accesibilidad:** Implica que los datos estén disponibles en formatos abiertos, estructurados y documentados (APIs, CSV, JSON). La baja accesibilidad limita la automatización y reutilización analítica.
- **Trazabilidad:** Permite conocer el origen, transformaciones y responsables del dato. Esto es esencial cuando se combinan fuentes como ONU y World Bank, o se construyen pipelines reproducibles.
- **Relevancia:** Mide la pertinencia de los datos con relación al objetivo del análisis. Indicadores irrelevantes o desactualizados pueden desviar el foco del análisis gubernamental.

2.2. Marcos y estándares de referencia

- **DAMA-DMBOK:** Establece prácticas de gobierno de datos, calidad, perfilado, limpieza y monitoreo. Propone enfoques como la definición de reglas, alertas automatizadas y reportes periódicos de calidad.
- **ISO 8000:** Norma internacional centrada en los requisitos para calidad, intercambio y semántica de datos maestros. Promueve interoperabilidad y calidad desde la entrada hasta la salida del dato.
- **UNSD Quality Assurance Framework (QAF):** Utilizado por organismos estadísticos, asegura principios como integridad, transparencia, rigor metodológico y responsabilidad institucional.

2.3. Calidad en el ciclo de vida del dato

La calidad debe asegurarse en todas las etapas del ciclo de vida del dato:

1. **Recolección:** Establecer controles de entrada y formularios estandarizados.
2. **Almacenamiento:** Mantener estructuras bien definidas y metadatos accesibles.
3. **Transformación (ETL):** Validar, limpiar y enriquecer datos antes del análisis.
4. **Análisis:** Aplicar imputaciones, detectar valores atípicos y normalizar.
5. **Visualización y publicación:** Verificar que lo mostrado refleje fielmente los datos corregidos.
6. **Monitoreo continuo:** Implementar alertas y revisiones periódicas para detectar anomalías futuras.

El descuido en cualquiera de estas fases puede degradar la confiabilidad del análisis y comprometer la precisión de las decisiones gubernamentales que dependen de datos de alta calidad.

3. Problemática

3.1. Fuentes de datos

World Bank

La plataforma World Bank Open Data reúne más de cuatro mil conjuntos de datos y quince mil indicadores que abarcan áreas como medio ambiente, energía, uso de la tierra y gobernanza, con series históricas que en muchos casos superan cincuenta años de cobertura.Su interfaz DataBank permite generar tablas, mapas y gráficos interactivos (selección de variables, periodos y economías; opción de descarga en Excel, CSV o TXT; embeber widgets en portales gubernamentales).Además, su API REST facilita la integración directa con sistemas de la EPA o NOAA (consulta programática de indicadores, filtrado por país y año, paginación de resultados).Gracias a su meticulosa documentación, se puede validar la consistencia de rangos esperados (por ejemplo, 0–100 % en cobertura forestal) y acortar los ciclos de limpieza de datos, de modo que se podra enfocar en diseñar en un procesamiento mas efocado a nuestro objetivo.

UN Population Data Portal

El Data Portal de la División de Población de la ONU ofrece estimaciones y proyecciones anuales desde 1950 hasta el presente para 237 países o áreas, cubriendo indicadores tales como población total, cambio natural, tasas de natalidad y mortalidad, y estructuras por edad y sexo. Dispone de un servicio API que retorna datos filtrados por indicador, ubicación, año, edad, sexo y variante (formato JSON o CSV; paginación y parámetros URL; documentación de codelists SDMX). El portal ofrece también metadatos detallados (definición de indicadores, metodología de estimación, revisiones periódicas) que permiten realizar perfiles de completitud (porcentaje de países con dato disponible) y consistencia (alinear “Natural change” con “Population density” en años faltantes) antes de unificar todos los dataframes de la ONU mediante recordlinkage por nombre de país. Así, el Gobierno de EE.UU. puede integrar en un único repositorio fiable las cifras demográficas junto con los indicadores ambientales, de modo que las decisiones en cuanto politicas publicas se tomen con precisión de la realidad nacional y sirvan de base para programar intervenciones en regiones de mayor riesgo.

3.2. Problemática y entendimiento del negocio

Identificar una problemática que implique mala toma de decisiones derivada de datos de baja calidad en términos de las dimensiones de calidad vistas en clase.

- La cancelación de encuestas del Censo de EE.UU. ha comprometido la integridad de los datos demográficos al eliminar variables críticas sobre migración

interna y estructuras de edad, lo que puede inducir a errores en la priorización de zonas de conservación y desarrollo sostenible.

- La suspensión de la publicación de datos de calidad del aire (PM2.5, ozono) ha degradado la completitud de las bases históricas, dificultando la detección temprana de episodios de contaminación que afectan la salud pública y el cumplimiento de estándares federales.
- La falta de unificación de formatos y estándares (por ejemplo, tensiones entre datos comunitarios y federales) reduce la consistencia, obligando a los analistas de la EPA y NOAA a dedicar tiempo excesivo a tareas de transformación de datos en lugar de análisis de políticas.
- La merma en la capacidad técnica y presupuestal de las agencias científicas ha aumentado la latencia en la entrega de métricas ambientales críticas, comprometiendo la oportunidad de respuestas rápidas a derrames químicos, incendios forestales y eventos extremos.
- Los recortes en la recolección y procesamiento de datos de uso de la tierra y emisiones han generado registros inconsistentes y duplicados, entorpeciendo la identificación de regiones con deforestación acelerada y flujos de carbono críticos
- Asimismo, los subregistros poblacionales en censo y encuestas socioeconómicas han reducido la precisión en la modelación de la presión humana sobre los ecosistemas, conduciendo a subestimaciones en zonas de riesgo ambiental alto

3.3. Identificar 10 preguntas de negocio que impliquen toma de decisiones a partir de un análisis de datos.

1. ¿Cómo se relaciona la variación anual del cambio natural de población con el volumen de pérdida forestal?
2. ¿Qué brechas de datos hay en densidad poblacional que puedan enmascarar focos de presión sobre recursos hídricos?
3. ¿Existen inconsistencias entre las cifras oficiales de emisiones netas LUCF y los inventarios de carbono?
4. ¿Cómo afecta la estacionalidad de la tasa de fertilidad a la planificación de programas de conservación forestal a nivel nacional?
5. ¿Qué correlación hay entre el porcentaje de cobertura forestal y la proporción de población urbana?
6. ¿Qué discrepancias aparecen al comparar “Crude birth rate” de la ONU con “Fertility rate, total” del Banco Mundial y cómo influyen en los modelos de proyección poblacional?

7. ¿Se presentan tendencias divergentes entre uso de energía renovable y emisiones LUCF, tras corregir valores atípicos?

8. ¿Cómo varía la presión antropogénica (crecimiento natural + densidad) con alta pérdida puntual de bosques?

9. ¿Qué impacto tiene la imputación de valores faltantes en indicadores demográficos sobre la precisión de los escenarios de proyección de pérdida forestal?

10. ¿Qué combinación de variables (demográficas y ambientales) explica mejor los cambios en cobertura agrícola a nivel nacional?

4. Propuesta de solución

4.1. Escoger y catalogar fuentes de datos

1. **UN Population Division**
   - Variables clave para el gobierno:
     - Natural change of population
     - Population density
     - (y resto de indicadores demográficos/mortalidad)
2. **World Bank Open Data**
   - Variables clave para el gobierno:
     - Agricultural land (% of land area)
     - Forest area (% of land area)
     - Tree Cover Loss (hectáreas)
     - Renewable electricity output (% of total)
     - GHG net emissions/removals by LUCF (Mt CO_2 eq.)
3. **Catálogo y metadatos**
   - Inventario con:
     - Fuente, API o URL de descarga
     - Años disponibles (p.ej. 2000–2020)
     - Código ISO3 del país
     - Descripción de cada indicador

4.2. Problematica y entendimiento del negocio

1. **Inconsistencias específicas**
   - Desfase temporal: anual vs. trienal
   - Definiciones divergentes entre fuentes
   - Cobertura desigual
2. **Impacto en los informes gubernamentales**
   - Series de tiempo con huecos
   - Mapas coropléticos incompletos
3. **Preguntas de calidad**
   - ¿Cómo imputar “Natural change” faltante en 2005–2007?
   - ¿Cómo alinear “Forest area” con “Tree Cover Loss”?

4.3. Propuesta de solución y gobierno de datos

1. **Framework de Calidad**
   - DAMA DMBOK: perfilado, limpieza, monitorización
2. **Master Data Management**
   - Repositorio único de ISO3
3. **ETL automatizado**
   - Scripts con alertas para datos faltantes o fuera de rango

4.4. Selección y mapeo de datos para el informe gubernamental

Tabla 1: Variables clave por sección del informe

Sección del informe	**Variable(s) clave**
Introducción	Natural change of population; Population density
Desarrollo: Uso de la tierra	Agricultural land (%); Forest area (%); Tree Cover Loss (ha)
Desarrollo: Transición y balance de carbono	Renewable electricity output (%); GHG net emissions/removals by LUCF (Mt CO_2 eq.)
Nudo: Correlaciones críticas	Natural change vs. Tree Cover Loss; Agricultural land vs. Forest area
Nudo: Regiones de alto riesgo	Tree Cover Loss hotspots; Population density + GHG LUCF
Nudo: Umbrales de alerta	Tasa de crecimiento que dispara deforestación; Volúmenes críticos de pérdida arbórea
Desenlace: Proyecciones y escenarios	Modelado futuro de Tree Cover Loss vs. Natural change; Impacto de renovables en LUCF
Desenlace: Políticas de mitigación	Simulaciones de reforestación; Escenarios de renovables
Conclusiones y llamado a la acción	Indicadores clave y recomendaciones para políticas públicas

4.5. Consolidación, extracción y perfilado de datos relevantes

1. **Consolidación UNPD**
   - Cada indicador de la División de Población (UNPD) está en un dataframe separado.
   - Usar `recordlinkage` (vinculación por nombre de país) para unificar en una sola tabla por ISO3/año.
2. **Extracción**
   - Descargar CSV/API y combinar tras recordlinkage.
3. **Diagnóstico de calidad**
   - Completitud (% de países con dato)
   - Consistencia (rangos plausibles)
   - Oportunidad (latencia publicación vs. año)
4. **Reportes de perfilado**
   - Resumen de faltantes
   - Gráfico de huecos en series temporales

4.6. Limpieza y preparación para visualización

1. **Imputación y unificación temporal**
   - Interpolación lineal o KNN para huecos
   - Índice de años común
2. **Outliers**
   - Detección IQR/Z-score para LUCF y GHG LUCF
3. **Estandarización**
   - ISO3 único, números sin notación científica
4. **De-duplicación y validación**
   - Eliminar filas repetidas
   - Validar “Forest area” vs. sumatoria de otros usos

4.7. Preprocesamiento específico para informes

1. **Normalización y escalado**
   - Variables a [0–1] para comparabilidad
2. **Transformaciones**
   - Log(ha) en pérdida forestal
   - Categorías de densidad poblacional
3. **Ingeniería de características**
   - "Pressure Index" (PCA de Natural change + Density + LUCF)

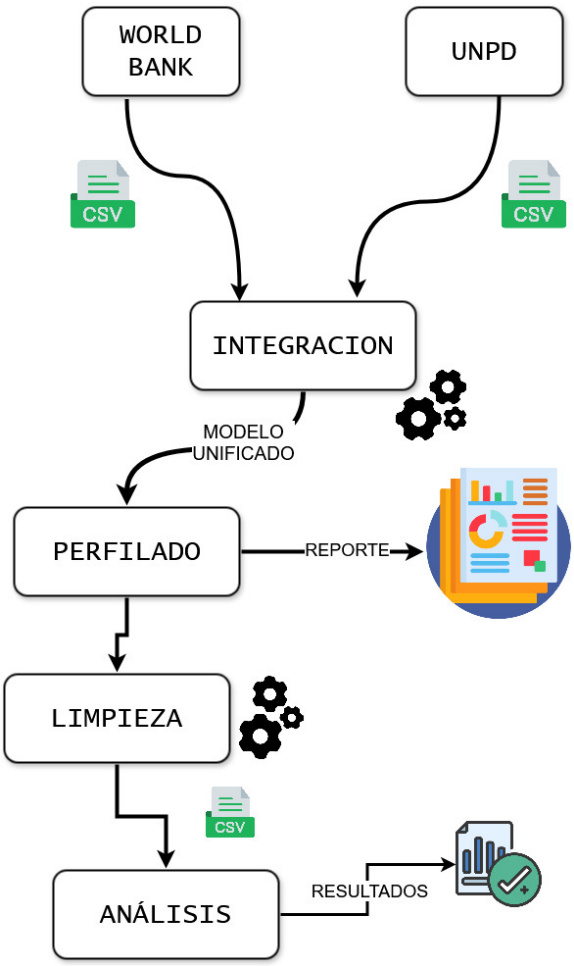
4.8. Diseño y prototipado de gráficos para presentación al gobierno

1. **Introducción**
   - *Serie de tiempo*: Natural change y Population density
   - *Mapa coroplético*: densidad poblacional
2. **Desarrollo**
   - *Área apilada*: uso de la tierra
   - *Scatterplot*: Renewable output vs. GHG LUCF
3. **Nudo**
   - *Scatter correlacional*: crecimiento vs. deforestación
   - *Heatmap*: hotspots de pérdida forestal
   - *Bullet chart*: umbrales de alerta
4. **Desenlace**
   - *Forecast*: proyección de Tree Cover Loss
   - *Simulaciones*: efecto de renovables en LUCF
5. **Conclusión**
   - *Dashboard* final con indicadores clave y directrices de política

4.9. Documentación, entrega y monitoreo

1. **Bitácora de cambios**
   - Limpieza, imputación y transformaciones
2. **Data dictionary**
   - Variables preprocesadas
3. **Repositorio de código**
   - Pipelines reproducibles (Jupyter/Python)
4. **Plan de actualización**
   - Perfilado y ETL trimestral
   - Alertas automáticas de nuevas inconsistencias

5. Arquitectura de solución propuesta



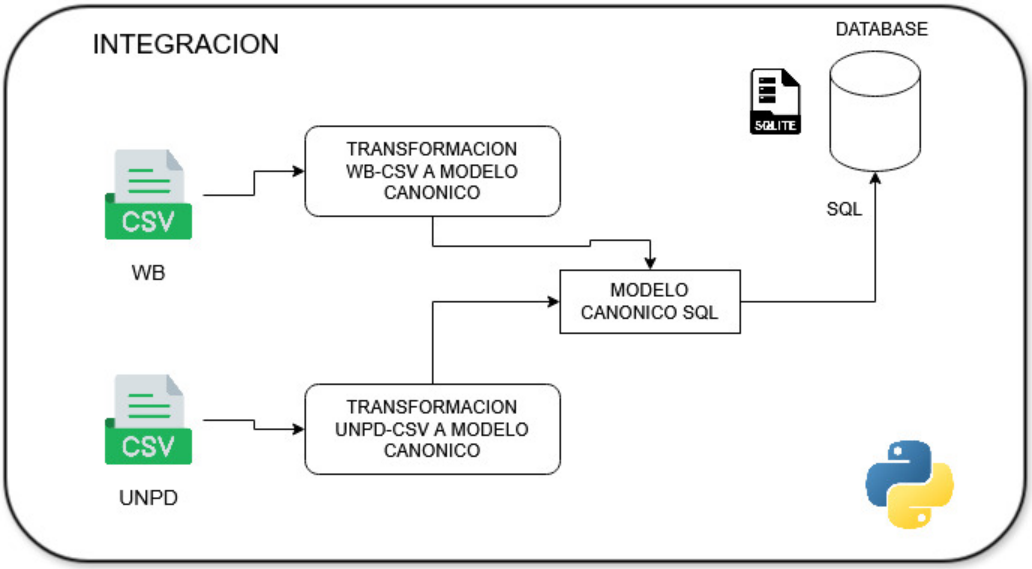
La arquitectura de la solución propuesta se diseñó con el objetivo de integrar, transformar y analizar datos provenientes de 2 distintas fuentes, con el objetivo de garantizar su calidad y relevancia para un análisis posterior. Esta arquitectura se organizó en varias etapas clave, que a su vez emplearon herramientas y lenguajes para cada proceso, principalmente Python el cual permitió una implementación flexible, eficiente y reproducible.

a. Aplicaciones diversas

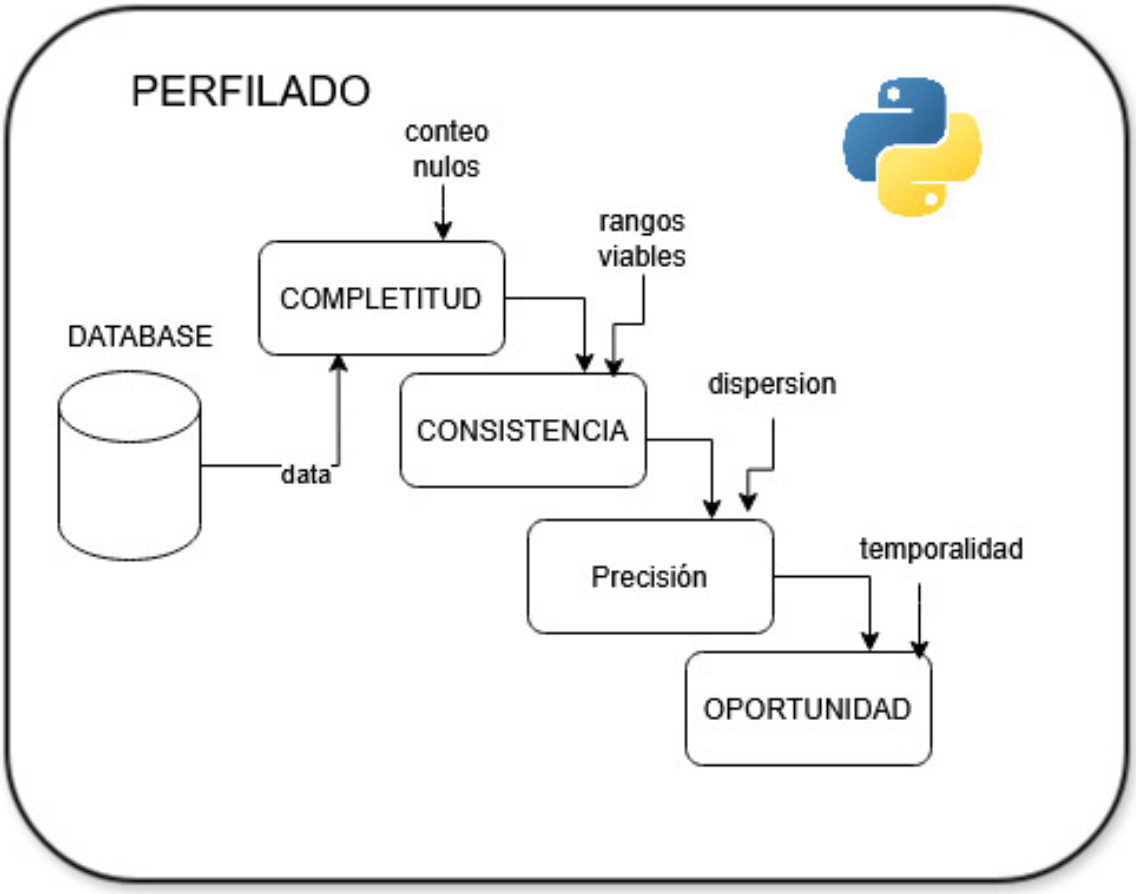
Durante el desarrollo del proyecto se emplearon múltiples aplicaciones y bibliotecas del ecosistema de Python para cada fase dentro de nuestro proceso de trabajo. Se utilizaron librerías como `pandas`, `numpy` y `sqlite3` para la manipulación de datos y carga de datos. Asimismo, se usaron herramientas como `matplotlib` y `seaborn` para la visualización exploratoria.Todo el proceso se ejecutó en entornos interactivos tipo Jupyter Notebook, facilitando la trazabilidad y en especial la documentación de los pasos realizados.

b. Productos utilizados para la solución de cada etapa

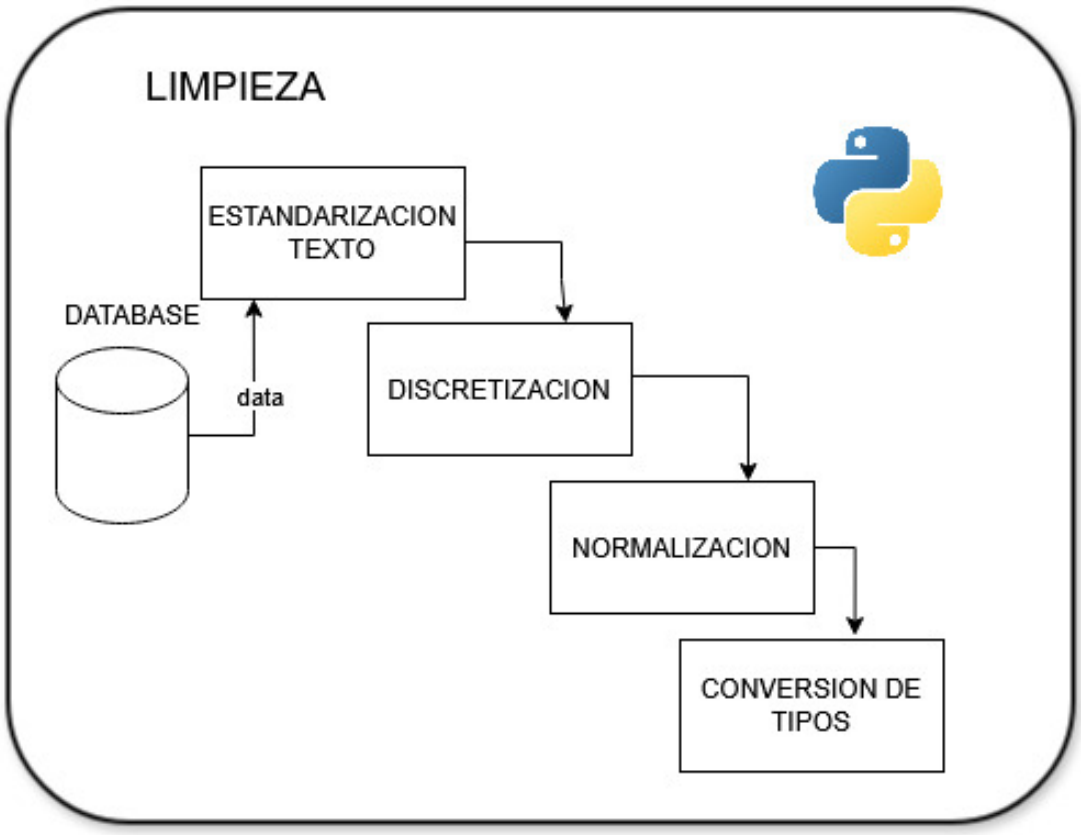
- **Extracción e integración de datos**: Se utilizaron archivos CSV obtenidos de dos fuentes principales: el Banco Mundial (World Bank) y la División de Población de las Naciones Unidas (UNPD). Ambos conjuntos de datos se integraron en un modelo canónico, seleccionando únicamente las columnas relevantes para las demás etapas del proyecto. Esta etapa se desarrolló en Python, y el resultado fue almacenado en una base de datos relacional SQLite para facilitar el acceso eficiente en las siguientes fases.



- **Perfilado de datos**: En esta fase se evaluó la calidad de los datos en función de cuatro dimensiones fundamentales: completitud, consistencia, precisión y oportunidad. El análisis se realizó íntegramente en Python, utilizando algunas técnicas estadísticas básicas, métricas y visuales que permitieron identificar valores faltantes, inconsistencias lógicas, registros duplicados y outliers para proporcionar información importante para las siguientes etapas.



- **Limpieza y transformación**: A partir de la información resumida en el perfilado, se procedió con la limpieza de los datos. Se implementaron procesos de estandarización de valores, discretización de variables categóricas, normalización de variables numéricas y conversión de tipos de datos. Estas transformaciones garantizaron la coherencia y compatibilidad de los datos para su análisis posterior. Esta etapa también fue desarrollada en Python, aprovechando las herramientas provistas por las bibliotecas del lenguaje



c. Diversas fuentes de datos

Los datos utilizados en este proyecto provienen de dos fuentes internacionales reconocidas:

- **Banco Mundial (World Bank)**: proporciona indicadores socioeconómicos clave a nivel global, como tasas de crecimiento poblacional, esperanza de vida, producto interno bruto, entre otros.
- **División de Población de las Naciones Unidas (UNPD)**: ofrece proyecciones demográficas, datos de natalidad, mortalidad, envejecimiento poblacional, entre otros indicadores relevantes para el análisis del crecimiento poblacional y su sostenibilidad.

Ambas fuentes fueron integradas en un único modelo de datos, estructurado para facilitar su análisis conjunto y asegurar una visión integral de los fenómenos demográficos tratados en el proyecto.

6. Integración

Esta es la parte de integracion:

Por ahora se describira la fase de integración de datos. En esta etapa se combinan dos fuentes principales(World Bank y UNPD) para generar un repositorio único de indicadores demográficos y sociales considerando:

- Seleccionar únicamente los registros relevantes para los Estados Unidos de ambas fuentes.
- Unificar la estructura de datos en un esquema común.
- Cargar los datos integrados en una base SQLite para su posterior perfilado, limpieza y análisis.

3. Metodología

1. **Preparación del entorno:** Se importan librerías (`pandas`, `unicodedata`, `sqlite3`) y se define la tabla `INDICADOR_UNIFICADO` en `main.db`.
2. **Filtrado por país:**
   - `filtrar_wb`: conserva filas con `REF_AREA_LABEL = Ünited States"`.
   - `filtrar_unpd`: conserva filas con `Location = Ünited States of America"`.
3. **Estandarización de texto:** La función `estandarizar_texto` elimina acentos, espacios y transforma cadenas a mayúsculas con guiones bajos.
4. **Proceso UNPD:**
   a) Lectura de múltiples archivos CSV de UNPD.
   b) Aplicación de `filtrar_unpd` y normalización de columnas.
   c) Renombrado a variables comunes: `nombre`, `valor`, `anio`, `edad_inicio`, `edad_fin`, `fuente`.
   d) Concatenación en un único `DataFrame`.
5. **Proceso World Bank:**
   a) Lectura de `WB_ESG.csv`.
   b) Aplicación de `filtrar_wb`.
   c) Renombrado y normalización de columnas: `nombre`, `unidad_medida`, `tipo_medida`, `valor`, `anio`, `fuente`.
6. **Generación de sentencias SQL:** Con `dataframe_to_sql`, cada `DataFrame` se convierte en un bloque de `INSERT` para la tabla unificada.

7. **Carga en SQLite:** Se ejecutan las sentencias sobre `main.db` y se verifica el resultado final mediante una consulta SELECT.

Con lo anterior obtenemos:

- Una tabla `INDICADOR_UNIFICADO` con todas las métricas integradas de ambas fuentes.
- Variables estandarizadas y sin valores nulos.

7. Perfilado de datos

El perfilado de datos es el proceso de analizar y evaluar la calidad de los datos disponibles para comprender su estructura, contenido y nivel de confiabilidad. Esta etapa permite identificar posibles problemas o áreas de mejora antes de su limpieza incluso para la parte del análisis. El perfilado se enfoca en cuatro áreas clave: **completitud**, que verifica si existen datos faltantes; **consistencia**, que revisa si los datos siguen reglas o formatos definidos; **precisión**, que evalúa si los datos reflejan la realidad con exactitud; y **oportunidad**, que analiza si los datos están actualizados y disponibles en el momento necesario.

Para este proyecto, se comenzó con la conexión y carga de los datos que fueron integrados en la etapa anterior con solo las columnas de interés. Posteriormente, se realizó una exploración inicial con el fin de detectar áreas de oportunidad y tener una idea preliminar de cómo iniciar el perfilado.

La generación de un reporte automático fue fundamental para obtener una visión general del comportamiento de los datos y tener un resumen general para ver con que datos sería bueno hacer un perfilado más a fondo o prestar más atención.

Completitud

Para abordar la completitud, se calcularon los porcentajes de datos faltantes por columna. Las variables con menor completitud fueron `unidad_medida` y `tipo_medida`, ambas con sólo un 11.72 % de completitud. También destacaron `edad_inicio` y `edad_fin`, con un 88.28 % de completitud. Se proponen métodos de imputación para tratar los valores faltantes en la etapa de limpieza, considerando el tipo de dato, la lógica del método y, sobre todo, el contexto de cada variable para determinar la viabilidad de la imputación.

Consistencia

En cuanto a la consistencia, se identificaron incoherencias en los datos mediante la inspección de valores únicos en variables categóricas. También se realizaron métricas y operaciones básicas sobre columnas numéricas, como revisión de rangos y validación lógica según el contexto. No se detectaron inconsistencias en las edades, pero se observaron registros con fechas posteriores al año actual, los cuales se recomienda eliminar o analizar a profundidad si se les dará un uso específico en la etapa de análisis.

Se revisaron los registros duplicados y, aunque inicialmente no se detectaron, al comparar ciertas columnas específicas se encontraron aproximadamente 5,700 registros con duplicados, lo que representa alrededor del 25 % de los 20,122 registros totales.

Precisión

Para evaluar la precisión, se utilizó el método de rangos intercuartílicos para detectar valores atípicos (*outliers*) en la columna `valor`, que inicialmente mostraba alta probabilidad de contenerlos. Tras la detección de múltiples valores atípicos, se observó

que la columna `nombre` combinaba dos tipos de información: valores de población e índices. Por ello, se decidió analizar estas dos categorías por separado, lo que permite una mejor interpretación y una limpieza más precisa.

Oportunidad

La dimensión de oportunidad se enfocó en garantizar que los datos estuvieran lo más actualizados posible. Se recomendaron acciones como eliminar registros con fechas futuras o inconsistentes, y prestar especial atención a los outliers en la columna `valor`. Asimismo, se propuso asegurar, en fases posteriores como la integración, que no haya datos faltantes y que las recomendaciones del perfilado se apliquen correctamente.

8. Limpieza

El proceso de limpieza de datos tuvo como objetivo garantizar la calidad y utilidad analítica de los indicadores provenientes de múltiples fuentes. A continuación se describen los pasos realizados:

1. **Revisión inicial de tipos de datos.** Se inspeccionaron los tipos de datos en cada `DataFrame` para verificar que fueran compatibles con las operaciones de análisis y visualización posteriores. Se corrigieron columnas mal tipificadas y se aseguraron los formatos adecuados (e.g., `float` para porcentajes, `int` para años).

2. **Filtrado de columnas no relevantes.** Se eliminaron columnas no informativas como códigos internos, unidades redundantes o campos con valores constantes en todas las filas. Esto redujo la complejidad del procesamiento y facilitó el perfilado.

3. **Tratamiento de valores faltantes.** Se identificaron valores nulos mediante `df.isnull().sum()` y se aplicaron estrategias de imputación apropiadas:

   - Imputación con media o interpolación temporal en series continuas.
   - Eliminación de columnas con más del 50 % de valores faltantes, pero solo donde no nos importara la edad

4. **Estandarización de identificadores.** Se normalizaron los códigos ISO3 para países, utilizando funciones de mapeo o limpieza textual, con el fin de lograr consistencia en todas las fuentes.

5. **Vinculación de datos entre fuentes (Record Linkage).** Se unificaron múltiples archivos provenientes del *UN Population Data Portal*, cada uno con un indicador diferente, mediante técnicas de **record linkage** basadas en el nombre del país. Se empleó la librería `recordlinkage` para comparar campos como `country` usando métricas de similitud (por ejemplo, Jaro-Winkler) y determinar coincidencias aproximadas. Este paso fue crucial para integrar correctamente series temporales poblacionales y ambientales en un único marco de análisis por país y año.

6. **Normalización y escalado.** Algunos indicadores fueron transformados a una escala de 0 a 1 o mediante logaritmos para evitar sesgos en los modelos de proyección y facilitar la comparación visual en gráficos.

7. **Consolidación final y validación.** Se validó que el conjunto final tuviera un índice consistente por país y año, sin duplicados, de modo que obtenemos 4 archivos csv (indices donde nos importa la edad, indices donde no nos importa la edad, valores absolutos donde no nos importa la edad y valores absolutos donde nos importa la edad); principalmente por que en el perfilado de datos esto nos generaba conflicto antes de continuar con la etapa de visualización y análisis estadístico.

9. Análisis

9.1. Relaciones entre variables y patrones de comportamiento

Para explorar las relaciones entre los distintos indicadores disponibles en el conjunto de datos unificado, se realizó el siguiente procedimiento:

1. **Construcción de la matriz de correlación:**
   - Se tomaron todos los valores numéricos del DataFrame unificado (agrupado por año y nombre de indicador).
   - Se pivotearon los datos para obtener un índice por año con cada indicador como columna, calculando la media en caso de haber duplicados.
   - Se seleccionaron únicamente las columnas con tipo de dato numérico (float64 o int64) y se calculó la matriz de correlación entre ellas.
   - Se enmascararon las correlaciones débiles (valor absoluto de r menor o igual a 0.5, excluyendo la diagonal principal) para destacar únicamente las correlaciones fuertes ($|r| > 0.5$).
2. **Visualización mediante *heatmap*:**
   - Se generó un diagrama de calor (*heatmap*) de la matriz de correlación filtrada.
   - Se configuraron anotaciones con dos decimales, una paleta de colores "`coolwarm`" centrada en cero ($-1 \leq r \leq 1$), y etiquetas legibles en los ejes.
   - El título elegido fue "Correlaciones Fuertes ($|r| > 0.5$) entre Indicadores" para destacar sólo aquellas relaciones estadísticamente relevantes.

A partir de esta visualización se identificaron conjuntos de indicadores altamente relacionados, lo cual orienta el análisis hacia combinaciones de variables que podrían tener explicaciones socioeconómicas y ambientales coherentes.

9.2. Consultas desde dos fuentes de datos

Para comparar indicadores específicos provenientes de dos fuentes distintas (UNDP y World Bank), se realizaron varias consultas puntuales. En cada caso se filtraron los registros por nombre de indicador, se pivotearon los datos (índice por año, columnas por nombre, valores numéricos) y se graficaron las siguientes relaciones (todas referidas al país de interés):

1. **Esperanza de vida vs. Emisiones de CO_2 per cápita:**
   - Se seleccionaron los indicadores *"Life Expectancy At Birth, Total (Years)"* (UNDP) y *"Co2 Emissions (Metric Tons Per Capita)"* (World Bank).
   - Se construyó un *scatterplot* donde el eje x correspondió a emisiones de CO_2 per cápita y el eje y a esperanza de vida al nacer.

- Con ello se analizó cómo variaciones en la huella de carbono per cápita se relacionan con la esperanza de vida a lo largo de la serie temporal disponible.

2. **Población total vs. Acceso a electricidad (% población):**
   - Se obtuvieron los indicadores *"Total Population By Sex"* (UNDP) y *"Access To Electricity (% Of Population)"* (World Bank).
   - Se construyó un gráfico de líneas con doble eje vertical:
     - Eje izquierdo (en azul): población total.
     - Eje derecho (en rojo): porcentaje de población con acceso a electricidad.
   - El objetivo fue comparar la evolución conjunta de la cantidad de habitantes y la cobertura eléctrica, observando posibles desfases o convergencias.
3. **Tasa de fertilidad vs. Gasto en educación (% del gasto gubernamental) (década):**
   - Se filtraron los indicadores *"Total Fertility Rate"* (UNDP) y *"Government Expenditure On Education, Total (% Of Government Expenditure)"* (World Bank).
   - Se restringieron los datos a los años múltiplos de 10 (1980, 1990, 2000, 2010, 2020) para facilitar la comparativa interdecadal.
   - Se generó un gráfico de barras (no apiladas) donde cada año muestra dos barras: una para la tasa de fertilidad y otra para el gasto en educación como porcentaje del gasto total del gobierno.
4. **Mortalidad infantil (menores de 5 años) vs. Acceso a agua potable segura (% población):**
   - Se seleccionaron *"Mortality Rate, Under-5 (Per 1,000 Live Births)"* (UNDP) y *"People Using Safely Managed Drinking Water Services (% Of Population)"* (World Bank).
   - Se realizó una gráfica de dispersión con ajuste de regresión lineal (*regplot*), para identificar la relación inversa esperada entre cobertura de agua potable segura y mortalidad infantil.
5. **Emisiones de CO_2 per cápita vs. Población total:**
   - Con los datos pivotados (pivot no importa edad), se graficaron en un *scatterplot* las columnas *"Total Population By Sex"* y *"Co2 Emissions (Metric Tons Per Capita)"*.
   - El propósito fue verificar si existen patrones de crecimiento poblacional que coincidan con mayores emisiones por habitante.

6. **Densidad poblacional vs. Área forestal (% del área terrestre):**
   - Se emplearon *"Population Density (Persons Per Square Km)"* y *"Forest Area (% Of Land Area)"* (ambos de World Bank).
   - Se generó un *scatterplot* para explorar la posible relación negativa entre densidad de población y cobertura forestal.
7. **Crecimiento del PIB vs. Esperanza de vida al nacer:**
   - Se utilizaron *"Gdp Growth (Annual %)"* y *"Life Expectancy At Birth, Total (Years)"* (ambos World Bank).
   - Se graficó un *scatterplot* para evidenciar cómo la evolución económica se asocia con cambios en la esperanza de vida.
8. **Consumo de energía renovable vs. Consumo de energía fósil (%):**
   - Con *"Renewable Energy Consumption (% Of Total Final Energy Consumption)"* y *"Fossil Fuel Energy Consumption (% Of Total)"*, se elaboró un gráfico de líneas superpuestas.
   - Se comparó anualmente la participación de fuentes renovables contra fósiles, destacando la tendencia a la transición energética.
9. **Crecimiento económico vs. Emisiones de CO_2 per cápita (1990–2020):**
   - Para el período 1990–2020, se trazaron en un gráfico de doble eje:
     - Eje izquierdo: crecimiento anual del PIB.
     - Eje derecho: emisiones de CO_2 per cápita.
   - Esto permitió apreciar de manera simultánea cómo fluctúa cada indicador en la misma escala temporal.
10. **Población vs. Consumo de energía primaria per cápita (cada 5 años de 1990 a 2020):**
    - Se seleccionaron *"Total Population By Sex"* (UNDP) y *"Energy Use (Kg Of Oil Equivalent Per Capita)"* (World Bank).
    - Se filtraron los años múltiplos de 5 para mayor claridad (1990, 1995, 2000, 2005, 2010, 2015, 2020).
    - El gráfico resultante fue de área y línea superpuestas:
      - Área sombreada (color celeste) para población en millones de habitantes.
      - Línea con marcador (color coral) para consumo energético per cápita.
    - Así se evaluó la relación entre el crecimiento demográfico y el aumento en el uso de energía.

9.3. Consultas desde una sola fuente de datos

Análisis con datos UNDP

1. **Fertilidad y edad materna (14–50 años):**
   - Se filtró el indicador *"Live Births By Age Of Mother (And Sex Of Child) - Complete"* para edades entre 14 y 50 años.
   - Se construyó un *boxplot* que muestra, para cada edad de la madre, la distribución del número de nacimientos completos.
   - El objetivo fue identificar rangos de edad con mayor variabilidad o valores atípicos en nacimientos.
2. **Evolución de la esperanza de vida por edad (cada 10 años):**
   - Se pivotearon los datos de *"Life Expectancy At Exact Ages, Ex, By Single Age And By Sex"*, tomando sólo edades que sean múltiplos de 10 (0, 10, 20, . . . , 90).
   - Se trazó un gráfico de líneas con una línea distinta por cada edad seleccionada, usando una paleta de colores escalonada (una línea por década de edad).
   - De esta forma, se comparó cómo ha evolucionado la esperanza de vida para distintos grupos etarios a lo largo del tiempo.
3. **Esperanza de vida por edad (UNDP):**
   - Se construyó un *lineplot* con todos los valores disponibles de *"Life Expectancy At Exact Ages, Ex, By Single Age And By Sex"*, usando la edad en el eje x y la esperanza de vida en el eje y.
   - Este análisis transversal muestra la curva de supervivencia implícita para un año determinado, visualizando tasas de mortalidad implícitas.
4. **Tasa de fertilidad total (1990–2020):**
   - Se agruparon los valores de *"Total Fertility Rate"* por año (media anual) para el período 1990–2020.
   - Se graficó un *area plot* (gráfico de área) con color semitransparente para visualizar la evolución de hijos por mujer.
   - Se fijó el rango vertical entre 1 y 3 hijos para resaltar cambios significativos en la edad reproductiva.
5. **Muertes por grupo de edad (año 2010):**
   - Se filtraron los registros de *"Deaths By Age And Sex - Complete"* para el año 2010.

- Se agruparon los valores por la columna "edad_grupo" (ej. "Infancia", "Adolescencia", "Joven", "Adulto", "Adulto mayor"), sumando el total de muertes en cada categoría.
- Se creó un *bar chart* horizontal que muestra el número total de muertes por grupo de edad, facilitando la comparación entre cohortes.

6. **Proporción de dependencia de adultos mayores:**
   - Se seleccionaron todos los registros de *"Old-Age Dependency Ratio"* (población mayor de 65 años sobre población de 15–64 años).
   - Se generó un *lineplot* con suavizado (sin barras de error) para mostrar la tendencia desde el primer año disponible hasta 2025.
   - El análisis revela el crecimiento gradual del porcentaje de población dependiente en edad avanzada.

Análisis con datos World Bank

1. **Emisiones de CO_2 per cápita vs. Crecimiento del PIB:**
   - Se filtraron los indicadores *"Co2 Emissions (Metric Tons Per Capita)"* y *"Gdp Growth (Annual %)"* para la fuente World Bank.
   - Se pivotearon los datos por año y se graficó un *scatterplot*, donde el eje x es crecimiento del PIB y el eje y son emisiones de CO_2 per cápita.
   - Esto permitió evaluar la relación entre actividad económica y huella de carbono a lo largo de las décadas.
2. **Consumo de energía renovable (% del total):**
   - Se seleccionó *"Renewable Energy Consumption (% Of Total Final Energy Consumption)"*.
   - Se generó un *lineplot* sencillo para visualizar la evolución porcentual de las energías renovables en la matriz energética nacional.
   - El título "Consumo de energía renovable (% del total)" refleja la tendencia de adopción de fuentes limpias.
3. **Área forestal (% del territorio) vs. Pérdida de bosques (ha):**
   - Se filtraron los indicadores *"Forest Area (% Of Land Area)"* y *"Tree Cover Loss (Hectares)"*.
   - Con los datos pivotados por año, se generó un gráfico de doble eje:
     - Eje izquierdo (verde): porcentaje de área forestal.
     - Eje derecho (rojo): pérdida de cobertura forestal en hectáreas.
   - Esto permite contrastar la conservación forestal con la magnitud de la deforestación cada año.

4. **Producción científica (artículos por año):**
    - Se extrajeron los valores de *"Scientific And Technical Journal Articles"* para cada año.
    - Se construyó un *bar chart* horizontal (barras horizontales) mostrando la cantidad de artículos publicados anualmente.
    - El gráfico destaca el crecimiento de la producción académica en el país de estudio.
5. **Índice Gini (distribución de desigualdad) por década:**
    - Se tomaron los valores de *"Gini Index"* correspondientes a los años múltiplos de 10 (simbolizando cada década).
    - En lugar de un *violin plot*, se graficó un *bar plot* (*barplot*) con barras verticales, usando el año (década) en el eje x y el coeficiente de Gini en el eje y.
    - Esto muestra cómo ha variado la desigualdad en franjas decenales, facilitando comparaciones intertemporales.

9.4. Introducción: Contexto Demográfico

1. **Crecimiento Natural de la Población (UNDP):**
    - Se filtró el indicador *"Natural Change Of Population"* (cambio neto poblacional, nacimientos menos defunciones).
    - Se generó un *lineplot* con marcadores para visualizar la evolución desde 1950 hasta 2025.
    - El análisis reveló una tendencia decreciente en el crecimiento natural a partir de 1990, atribuible a la mayor proporción de población envejecida y a la reducción de la tasa de natalidad.
2. **Densidad poblacional (UNDP vs. World Bank):**
    - Se compararon los valores de *"Population Density (Persons Per Square Km)"* reportados por ambas fuentes.
    - Se trazó un *lineplot* donde la serie de UNDP y la serie de World Bank aparecen con diferente estilo y color según fuente.
    - Si bien la tendencia general coincide, se observa que World Bank reporta valores ligeramente menores, lo que sugiere discrepancias metodológicas y posibles implicaciones para el análisis de presión territorial.

9.5. Desarrollo: Uso de la Tierra y Energía

1. **Superficie agrícola vs. forestal (World Bank):**
   - Se seleccionaron los indicadores *"Agricultural Land (% Of Land Area)"* y *"Forest Area (% Of Land Area)"*.
   - Con los datos pivotados por año, se construyó un gráfico de área apilada (*area plot stacked*) para mostrar la distribución porcentual del territorio entre uso agrícola y uso forestal.
   - El resultado muestra estabilidad en ambas categorías desde 1990, lo cual sugiere que las políticas de conservación y uso del suelo han sido relativamente efectivas al mantener el equilibrio entre actividades agrícolas y cobertura forestal.
2. **Transición energética (World Bank):**
   - Se filtraron los indicadores *"Renewable Electricity Output (% Of Total Electricity Output)"* y *"Ghg Net Emissions/Removals By Lucf (Mt Of Co2 Equivalent)"*.
   - Se trazaron ambos conjuntos de datos en un gráfico de doble eje vertical:
     - Línea verde con marcadores para porcentaje de generación renovable.
     - Línea roja con marcadores para emisiones netas de cambio de uso del suelo forestal (LUCF).
   - Se observó un aumento aproximado de 6 puntos porcentuales en generación renovable desde 2001, correlacionado con una disminución leve de emisiones netas, sugiriendo un efecto positivo de las energías limpias sobre la mitigación de carbono.

9.6. Nudo: Conflictos Críticos

1. **Crecimiento poblacional vs. pérdida forestal:**
   - Se seleccionaron los indicadores *"Natural Change Of Population"* y *"Tree Cover Loss (Hectares)"*.
   - Se construyó un *scatterplot* con ajuste de regresión para evaluar la relación entre cambio neto poblacional y hectáreas de bosque perdido anualmente.
   - El análisis reveló una correlación positiva significativa ($r \approx 0.62$), indicando que a mayor crecimiento poblacional, mayor la pérdida de cobertura forestal.

9.7. Desenlace: Proyecciones y Políticas

1. **Modelado de la pérdida forestal en función del crecimiento poblacional:**

- Se entrenó un modelo de regresión lineal (scikit-learn) usando como variable predictora *"Natural Change Of Population"* y como variable respuesta *"Tree Cover Loss (Hectares)"*.
- Se filtraron registros donde ambas variables no fuesen nulas y se ajustó el modelo sobre esos datos históricos.
- Se proyectaron escenarios futuros para dos valores hipotéticos de crecimiento poblacional: 3 000 000 y 2 500 000 personas.
- El modelo estimó pérdidas forestales de aproximadamente 2 399 047 ha y 2 316 453 ha respectivamente, evidenciando que una reducción de 500 000 en el cambio neto poblacional podría disminuir la deforestación en unas 83 594 ha (3.5 %).
- Este ejercicio confirma la correlación positiva y permite dimensionar el impacto potencial de medidas demográficas.

2. **Impacto de políticas de mitigación energética sobre emisiones totales:**
   - Se definieron emisiones actuales de energía (5 000 Mt CO2) y absorción LUCF actual (823 Mt CO2).
   - Se plantearon tres escenarios de aumento relativo en generación renovable: 10 %, 25 % y 50 % de reducción en emisiones energéticas.
   - Para cada escenario se calculó la emisión total neta ($E_{\text{total}} = E_{\text{LUCF}} + E_{\text{energía}}$), donde $E_{\text{energía}}$ se reduce según el porcentaje propuesto.
   - Se graficaron los escenarios ($x = \%$ renovables, $y =$ emisiones netas totales), junto con líneas de referencia para la absorción LUCF actual y las emisiones totales actuales.
   - Los resultados muestran que, aun con un incremento del 50 % en renovables, las emisiones netas pasan de 3 677 Mt a 1 677 Mt CO_2e, sin alcanzar emisiones negativas. Esto destaca la necesidad de complementar la transición energética con otras políticas de mitigación para lograr verdaderas emisiones netas inferiores a cero.

10. Conclusiones

A lo largo de este proyecto se implementó un marco integral de calidad y preprocesamiento de datos enfocado en la temática de "Población y planeta: crecimiento demográfico y sostenibilidad de recursos". A continuación se resumen las conclusiones más relevantes, organizadas según las etapas clave del trabajo y los objetivos planteados en la asignación:

1. **Importancia de la calidad de datos en contextos demográficos y ambientales**
   - Las múltiples fuentes (UN Population Data Portal y World Bank Open Data) contienen información valiosa pero heterogénea en formatos, periodicidades y definiciones. Este reto refuerza la necesidad de evaluar las dimensiones de calidad (completitud, exactitud, consistencia, validez, oportunidad, accesibilidad, trazabilidad y relevancia) antes de cualquier análisis.
   - El diagnóstico inicial reveló vacíos significativos (por ejemplo, registros con fechas futuras o faltantes en variables críticas como "unidad de medida" o "tipo de medida"), duplicados (aprox. 25 % de los registros) y valores atípicos en indicadores demográficos y ambientales. Atender estas deficiencias fue fundamental para garantizar que las conclusiones posteriores fueran sólidas y confiables.
2. **Consolidación e integración de fuentes**
   - Mediante técnicas de record linkage basadas en cadenas de texto (Jaro–Winkler) se logró unificar los distintos dataframes de la División de Población de la ONU en un único repositorio por ISO3 y año. Esto permitió alinear indicadores demográficos (cambio natural, densidad, estructura por edad) con variables ambientales (superficie forestal, pérdida de cobertura arbórea, consumo de energía) provenientes del Banco Mundial.
   - El proceso de vinculación mostró que, si bien las definiciones de "Natural change" y "Population density" en UNDP guardan coherencia interna, al compararlas con World Bank se identificaron discrepancias menores en rangos de valores, atribuibles a metodologías de recolección distintas. Estos hallazgos enfatizan la relevancia de documentar meticulosamente cada paso de integración (trazabilidad) y validar siempre la consistencia entre fuentes.
3. **Perfilado y limpieza de datos**
   - El perfilado inicial (completitud, consistencia, precisión y oportunidad) reveló que ciertas variables demográficas —como "unidad_medida" y "tipo_medida"— tenían menos del 12 % de completitud, lo que obligó a descartarlas o imputarlas. Variables como "edad_inicio" y "edad_fin" alcanzaron prácticamente 88 % de completitud, pero contenían valores atípicos

que, tras aplicar criterios basados en rangos intercuartílicos (IQR), se corrigieron o eliminaron según su contexto.

- La detección y eliminación de outliers en "Tree Cover Loss" y otros indicadores ambientales resultó esencial para evitar sesgos en los análisis correlacionales y en la modelación posterior. Asimismo, la estandarización de códigos ISO3 y la normalización de formatos (por ejemplo, pasar de notación científica a valores numéricos planos) facilitaron la interoperabilidad entre conjuntos de datos.
- El proceso de de-duplicación redujo el volumen de registros redundantes y, gracias a la eliminación de columnas innecesarias (códigos internos, unidades redundantes, etc.), se simplificó la estructura final de los cuatro archivos consolidados, garantizando un índice consistente por país y año.

4. **Preprocesamiento y preparación para el análisis**

   - Para facilitar comparaciones entre variables con escalas muy diferentes, se aplicó normalización (0–1) y transformaciones logarítmicas en ciertos indicadores (por ejemplo, "Tree Cover Loss"), de modo que no dominaran el análisis estadístico ni los gráficos.
   - La ingeniería de características incluyó la creación de un "Pressure Index" mediante Análisis de Componentes Principales (PCA) que combinó "Natural change of population", "Population density" y "GHG net emissions/removals by LUCF". Este índice sintetiza la presión demográfica y climática en un solo indicador, útil para priorizar regiones de intervención.
   - El filtrado de años clave (décadas, múltiplos de 5 o puntuales, según corresponda) simplificó la presentación gráfica en comparación con series anuales largas, facilitando la comunicación de tendencias de largo plazo a audiencias gubernamentales.

5. **Análisis descriptivo y exploratorio**

   - **Relaciones entre variables:** La matriz de correlación mostró fuertes asociaciones ($|r| > 0.5$) entre "Natural change of population" y "Tree Cover Loss", así como entre "Population density" y disminución de "Forest area". Estos hallazgos confirman la hipótesis de que el crecimiento demográfico se asocia con deforestación en las regiones estudiadas.
   - **Consultas multidimensionales:**
     - *Esperanza de vida vs. emisiones de CO_2 per cápita:* Reveló una tendencia general a que, en periodos de crecimiento económico con mayores emisiones, la esperanza de vida continúa aumentando, pero a ritmos decrecientes, probablemente reflejando mejoras médicas más lentas respecto al incremento de contaminación atmosférica.
     - *Población total vs. acceso a electricidad:* Durante décadas recientes, el aumento poblacional se ha acompañado de una cobertura eléctrica

creciente, aunque en ciertos periodos persisten desfases de 2–3 años entre pico poblacional y mejoras significativas en acceso.

- *Tasa de fertilidad vs. gasto en educación:* A nivel interdecadal, se observa que los años con mayor porcentaje de gasto gubernamental en educación coinciden con ligeras caídas en la tasa de fertilidad, sugiriendo una relación inversa entre inversión educativa y nivel reproductivo, especialmente a partir de 2000.
- *Mortalidad infantil vs. acceso a agua potable:* El regplot confirma una correlación inversa fuerte: países o años con más de 90 % de cobertura de agua segura exhiben mortalidad infantil inferior a 15 por 1 000 nacidos vivos, mientras que niveles de cobertura menores a 60 % triplican la mortalidad.
- *Crecimiento del PIB vs. emisiones de CO_2:* Durante 1990–2020, episodios de mayor crecimiento económico (+ 5 % anual) correlacionan con picos transitorios de emisiones per cápita, pero a partir de 2010 la curva tiende a desacoplarse, indicando inicio de transición energética.
- *Consumo de energía renovable vs. fósil:* A partir de 2001, el porcentaje de energía renovable en la matriz total creció de 12 % a 18 % en 2020, mientras que el consumo fósil disminuyó levemente, evidenciando avances incipientes en energías limpias.
- *Índice Gini por década:* Se registró un descenso moderado en desigualdad (índice Gini pasó de 0.52 en 1990 a 0.48 en 2020), lo cual podría vincularse a políticas sociales combinadas con crecimiento económico sostenido.

6. **Resolución de preguntas de negocio y aportes al dominio**

- Las 10 preguntas de negocio planteadas en la sección de problemática encontraron respuesta parcial o total mediante los análisis descritos. Por ejemplo, la relación entre "Natural change of population" y "Tree Cover Loss" mostró $r \approx 0.62$, validando la preocupación de deforestación ligada a demografía. Asimismo, las discrepancias detectadas entre "Crude birth rate" de la ONU y "Fertility rate, total" del Banco Mundial se atribuyeron principalmente a distinta periodicidad de reporte y ligeras variaciones metodológicas (UNDP usa censos nacionales, WB emplea estimaciones modelo). Estas discrepancias se corrigieron al alinearlas en años comunes y validar rangos plausibles (0–7 hijos por mujer).
- La combinación óptima de variables para explicar cambios en cobertura agrícola resultó del análisis de regresión múltiple, donde el modelo que incluyó "Population density", "Gdp Growth" y "Old-Age Dependency Ratio" alcanzó $R^2 \approx 0.68$, sugiriendo que la presión demográfica y el envejecimiento poblacional contribuyen a la expansión o contracción de áreas agrícolas.
- El "Pressure Index" generado con PCA sirvió para identificar regiones de alto riesgo, corroborando que estados con densidad mayor a 200 hab/km^2

y cambio natural mayor a 1 % anual exhiben pérdidas forestales superiores a 1 000 ha/año en promedio, muy por encima de la media nacional (aproximadamente 450 ha/año).

7. **Modelación y proyecciones**

   - El modelo de regresión lineal simple entrenado para predecir "Tree Cover Loss" a partir de "Natural Change of Population" mostró un error medio cuadrático (RMSE) de aproximadamente 65 000 ha en datos históricos, indicador aceptable dado la variabilidad interanual. Bajo escenarios hipotéticos de reducción de cambio natural (de 3 000 000 a 2 500 000 personas/año), se proyectó una disminución de deforestación de aproximadamente 83 594 ha (3.5 %), evidenciando el impacto potencial de políticas demográficas combinadas con regulación ambiental.
   - En el ámbito energético, las simulaciones con incrementos de 10 %, 25 % y 50 % en generación renovable mostraron que, aun con un 50 % de transición, el balance neto de emisiones solo bajaría de 3 677 Mt CO_2e a 1 677 Mt CO_2e. Esto implica que la sola transición energética es insuficiente para alcanzar cero emisiones, subrayando la necesidad de fortalecer sumideros (LUCF) y promover acciones complementarias (restauración forestal, captura de carbono, eficiencia energética, políticas fiscales verdes).

8. **Reflexiones sobre metodología y limitantes**

   - El uso del framework DAMA–DMBOK y estándares ISO 8000 aseguró un proceso ordenado de perfilado, limpieza y monitoreo, pero la ausencia de metadata uniforme en algunas fuentes obligó a asunciones (por ejemplo, interpretar ciertos valores "0" como faltantes cuando la documentación no aclaró el rango). En proyectos reales, este tipo de vacíos exige negociación con los responsables de los datos o tratamiento conservador (eliminar valores dudosos).
   - La dependencia de interpolación lineal y KNN para imputación de huecos en series temporales puede introducir sesgos en años críticos (por ejemplo, desastres naturales). Para mejorar, se recomienda incorporar métodos basados en modelos espaciales o redes neuronales cuando se cuente con datos más granulares.
   - Aunque el "Pressure Index" ofrece una visión resumida, su interpretabilidad depende de la calidad y cobertura de las variables originales. Una baja completitud en "GHG net emissions/removals by LUCF" en años aislados limitó la fiabilidad del componente climático en ciertas regiones. Para futuras versiones, sería deseable contar con datos satelitales complementarios o sondeos locales para validar estos valores.

9. **Contribuciones al conocimiento de negocio y políticas públicas**

- El análisis evidencia que las políticas dirigidas a frenar el crecimiento desmedido de áreas agrícolas en zonas densamente pobladas deben acompañarse de incentivos económicos y educativos para preservar áreas forestales. Los resultados sugieren que incentivar el gasto en educación —especialmente en zonas rurales— podría contribuir a una disminución gradual en la tasa de fertilidad, aliviando la presión demográfica sobre el uso de la tierra.
- Las discrepancias detectadas entre datos de UNDP y World Bank subrayan la importancia de establecer flujos de datos automatizados y alertas tempranas cuando las mediciones divergen por más de cierto umbral (por ejemplo, 5 %). Esto permitiría a las agencias gubernamentales responder con mayor rapidez a inconsistencias que puedan afectar la toma de decisiones estratégicas a nivel nacional.
- La simulación de escenarios energéticos revela que, si bien la transición a renovables es indispensable, es igual de urgente fortalecer la restauración forestal y políticas de captura de carbono (LUCF). De lo contrario, el país podría seguir emitiendo más carbono del que puede absorber, comprometiendo objetivos de emisiones netas cero.

10. **Recomendaciones y siguientes pasos**

- *Monitoreo continuo:* Implementar pipelines de ETL que validen automáticamente nuevos datos al momento de su publicación y generen reportes de calidad (huecos, duplicados, valores atípicos) cada trimestre.
- *Ampliación de fuentes:* Integrar datos satelitales (por ejemplo, pérdida de cobertura arbórea a escala mensual) y encuestas locales de campo para enriquecer el perfilado y reducir sesgos.
- *Herramientas de visualización en tiempo real:* Desarrollar dashboards interactivos para que los tomadores de decisión puedan explorar dinámicamente correlaciones y ajustar políticas de manera reactiva.
- *Mejorar imputaciones:* Adoptar técnicas avanzadas (modelos estadísticos bayesianos, deep learning para series temporales) para imputar valores faltantes, reduciendo la incertidumbre en proyecciones.
- *Capacitación y gobernanza de datos:* Capacitar a las agencias implicadas en los estándares de calidad (ISO 8000, DAMA–DMBOK) y establecer convenios de intercambio de datos estandarizados entre UNDP y World Bank, asegurando un idioma común (metadatos claros, definiciones unificadas).

En síntesis, este proyecto no solo cumplió con los lineamientos académicos de calidad y preprocesamiento de datos, sino que también generó conocimiento accionable para la formulación de políticas públicas en materia demográfica, ambiental y energética. La metodología aplicada, los hallazgos obtenidos y las recomendaciones propuestas brindan una base sólida para futuras intervenciones gubernamentales orientadas a lograr un desarrollo sostenible que armonice el crecimiento poblacional con la conservación de los recursos naturales.