**Documentación de los datasets explorados**

**1. Dataset: *Student Habits vs Academic Performance***

* **Fuente y tipo de datos**
  + Fuente: [Kaggle – Student Habits vs Academic Performance](https://www.kaggle.com/datasets/jayaantanaath/student-habits-vs-academic-performance/data)
  + Tipo de datos: Tabular (CSV)
  + Origen: Terciario
* **Características básicas**
  + Número de registros: 1,000
  + Número de atributos: 16
  + Tamaño del archivo: 0.072 MB
  + Variables principales: student\_id, age, gender, study\_hours\_per\_day, social\_media\_hours, netflix\_hours, attendance\_percentage, sleep\_hours, diet\_quality, exercise\_frequency, mental\_health\_rating, exam\_score.
* **Nivel de documentación disponible**
  + Incluye descripción de cada columna (ej. género, horas de estudio, calidad de dieta, etc.).
  + Tiene una explicación general del propósito del dataset: simular cómo los hábitos de vida afectan el rendimiento académico.
  + Nivel de documentación: **Alto**
* **Posibles aplicaciones**
  + **Orientación educativa**: diseñar programas de mejora de hábitos para estudiantes.
  + **Concientización**: elaborar informes o visualizaciones que muestren el impacto del sueño y uso de redes sociales en las notas.
  + **Validación de hipótesis**: contrastar supuestos sobre ejercicio, dieta y rendimiento académico.

**2. Dataset: *SMS Spam Collection***

* **Fuente y tipo de datos**
  + Fuente: [UCI Machine Learning Repository – SMS Spam Collection](https://archive.ics.uci.edu/dataset/228/sms+spam+collection)
  + Tipo de datos: Texto
  + Origen: Terciario
* **Características básicas**
  + Número de registros: 5,574
  + Número de atributos: 2 (label, sms)
  + Tamaño del archivo: 0.466 MB
* **Nivel de documentación disponible**
  + Incluye una descripción clara: mensajes de texto clasificados como “ham” (legítimos) o “spam”.
  + Contiene README con información del corpus.
  + Estructura simple: solo dos columnas.
  + Nivel de documentación: **Medio**
* **Posibles aplicaciones**
  + **Detección de spam**: crear un clasificador de mensajes similar a los filtros de correo electrónico.
  + **Análisis del lenguaje**: estudiar diferencias lingüísticas entre mensajes legítimos y spam.
  + **Seguridad digital**: detectar intentos de phishing en mensajes.
  + **Educación**: sensibilizar sobre cómo identificar spam en la vida diaria.

**3. Dataset: *Global Forest Change 2000–2014***

* **Fuente y tipo de datos**
  + Fuente: [Global Forest Change (Hansen et al., 2013)](https://earthenginepartners.appspot.com/science-2013-global-forest/download_v1.2.html)
  + Tipo de datos: Geoespacial (GeoJSON)
  + Origen: Terciario
* **Características básicas**
  + Resolución espacial: 30 metros por píxel.
  + Cobertura: global, organizada en mosaicos de 10x10 grados.
  + Variables principales:
    - treecover2000: cobertura arbórea en el año 2000 (0–100%).
    - loss: pérdida de cobertura forestal (2000–2014).
    - gain: ganancia de cobertura forestal (2000–2012).
    - lossyear: año específico de pérdida (2001–2014).
    - datamask: áreas sin datos, tierra o cuerpos de agua.
  + Tamaño: depende de la selección de mosaicos, pero el conjunto completo es de varios GB.
* **Nivel de documentación disponible**
  + Incluye artículo científico de referencia (Hansen et al., *Science*, 2013).
  + Explica las metodologías de procesamiento con Landsat.
  + Datos acompañados de licencia y metadatos detallados.
  + Nivel de documentación: **Alto**
* **Posibles aplicaciones**
  + **Monitoreo ambiental**: analizar la deforestación y recuperación de bosques a nivel global.
  + **Gestión de recursos naturales**: apoyar la planificación de conservación forestal.
  + **Investigación climática**: estudiar el impacto de la cobertura arbórea en la captura de carbono.
  + **Visualización geoespacial**: crear mapas interactivos de cambio de uso de suelo.

**Justificación de la base de datos seleccionada**

Tras la exploración de tres datasets de distinta naturaleza (csv, text y GeoJSON), el equipo decidió seleccionar como base final de trabajo el conjunto de datos **Student Habits vs Academic Performance**. Esta elección se fundamenta en los siguientes criterios:

1. **Completitud y calidad de los datos**: el dataset cuenta con 1,000 registros y 16 atributos bien documentados, sin presentar un tamaño excesivo que complique su manipulación. Además, cada variable está claramente descrita, lo cual facilita su comprensión y análisis.
2. **Relevancia temática**: aborda una problemática cercana y significativa en el ámbito académico: la relación entre los hábitos de vida de los estudiantes y su rendimiento académico. Esto permite formular hipótesis interesantes y de gran aplicabilidad, como el impacto de las horas de estudio, el sueño o el uso de redes sociales en las calificaciones.
3. **Manejabilidad técnica**: al estar en formato CSV y con un tamaño reducido (0.072 MB), resulta fácilmente manipulable con herramientas de análisis de datos (Python, R, Excel), lo cual permite concentrar esfuerzos en la exploración y el análisis en lugar de en la preparación o limpieza extensiva.
4. **Documentación disponible**: incluye una explicación general del dataset y un diccionario de variables detallado, lo que proporciona un **alto nivel de documentación** en comparación con los otros conjuntos explorados.
5. **Potencial de aplicación**: su contenido permite abordar diferentes técnicas de análisis exploratorio, preprocesamiento y reducción de dimensionalidad, además de ser un recurso ideal para generar visualizaciones claras y extraer insights con valor práctico.