1. Desafíos con los Datasets Se identificaron varias dificultades al trabajar con datasets:

- **Acceso y Metadatos:** Es muy difícil acceder a buenos datasets y conseguir la información de metadatos. Muchos son exclusivos para profesionales de la salud o requieren pago y cursos previos.
- **Calidad y Homogeneidad:** Las imágenes de diferentes datasets pueden variar mucho en calidad, resolución, tamaño y tipo de corte (2D vs 3D/OCT), lo que complica su combinación. Algunos datasets tienen imágenes con defectos o pre-procesadas de formas desconocidas.
- **Redundancia:** Muchos datasets son copias o segmentos de otros más grandes, como el de Kaggle.
- **Información de Metadatos Limitada:** La mayoría de los datasets solo tienen información binaria (si tiene o no la enfermedad), y los metadatos más comunes son **edad y sexo**, con muchos valores nulos. Algunos incluyen localización o elementos de la fotografía, pero son escasos.

2. Selección del Dataset Principal Después de discutir varias opciones, el equipo acordó centrarse en el dataset **ODIR-5K** como base para el proyecto.

- **Razones de la elección de ODIR-5K:**
  - Tiene **buenas imágenes** y **buena calidad**.
  - Contiene **14.000 imágenes** (aunque la tabla de metadatos pre-procesados tiene 6.393 entradas) y se confirmó que el original tiene hasta 8.000 imágenes y 3.501 metadatos, con ojo derecho e izquierdo para cada caso.
  - Incluye **diferentes enfermedades oculares** (hasta ocho categorías, como glaucoma y diabetes) y la posibilidad de diferenciar entre niveles de enfermedad (ej. temprano, intermedio).
  - Ofrece **metadatos de edad y sexo**, y lo más importante, **observaciones diagnósticas escritas por clínicos** (aunque cortas), lo que permite trabajar con Procesamiento de Lenguaje Natural (NLP).
- **Datasets considerados/descartados:**
  - **Messidor:** Descartado por la mala calidad o rareza de las imágenes (tonos azules, tamaño pequeño, imágenes tratadas), y porque el link derivaba a un challenge.
  - **Kaggle y IDRiD:** Eran opciones iniciales por su similitud de formato de imagen y buena calidad, pero IDRiD carecía de los niveles detallados de enfermedad (solo binario) que Kaggle sí ofrecía.
  - **RFMiD:** Similar a Kaggle, buena calidad, pero con pocos casos de retinopatía diabética (DR) y sin edema macular diabético (DME).

3. Objetivo y Enfoque del Proyecto El grupo busca ir **más allá de un simple modelo de convolución** con imágenes. La **"magia está en intentar trabajar con los metadatos"** para darle un **"salto"** y un **"plus de inteligencia artificial"** al proyecto, haciéndolo más complejo y atractivo que una práctica básica de *deep learning*. Se busca implementar **RAC (Retrieval Augmented Generation)** y **NLP** para trabajar con los textos clínicos y potencialmente ofrecer **tratamientos o recomendaciones** a partir de PDFs o información en línea sobre las patologías.

4. Plan de Trabajo y Organización de Equipos El proyecto se dividirá en fases:

- **Fase 1: Exploración y Selección de Dataset (Finalizada)**.
- **Fase 2: Limpieza y Procesado (Fecha límite: 5 de agosto)**.

Para la Fase 2, se decidió dividir el equipo en dos subgrupos:

- **Equipo 1 (Leticia, David, Sara):** Se encargará de las **imágenes**, lo que incluye la limpieza, pre-procesado, estandarización de tamaño, simetría (macula y focus en la misma posición), y posiblemente *data augmentation*. Se considera que esta parte tendrá **más carga de trabajo**.
- **Equipo 2 (Javier, Miguel Ángel, Sofía, Nause):** Se encargará del **EDA (Análisis Exploratorio de Datos)**, los **metadatos** (limpieza, manejo de nulos), y la **preparación para el RAC/NLP**, que implica la búsqueda de PDFs y textos clínicos de las enfermedades del dataset.

Metodología de Trabajo:

- **Modularidad:** Se optará por un **enfoque modular** utilizando **scripts** para facilitar el mantenimiento y la resolución de errores.
- **EDA en Notebooks:** La parte de EDA se comenzará en notebooks para visualización y explicaciones detalladas, y luego las funciones se trasladarán a scripts.
- **Pre-procesado en Scripts:** Las funciones de transformación de imágenes se desarrollarán directamente en scripts.
- **Control de Versiones:** Se trabajará en **ramas separadas** en Git para evitar conflictos al *mergear*.
- **Colaboración en Notebooks:** No se recomienda que dos personas trabajen simultáneamente en el mismo notebook debido a la complejidad de *mergear* los archivos Json. Si se trabaja juntos, será en videollamada, con una persona escribiendo y otra aportando.
- **Comunicación:** Se sugiere avisar al grupo cuando se empiece a trabajar y subir los cambios (commits) para coordinar y evitar pisarse el trabajo. Los comentarios en el código se usarán para anotaciones y sugerencias.
- **Manejo del Dataset:** Debido al gran tamaño (10GB), subir el dataset completo a Git es un problema. Se considerará una cuenta Git Pro/Team si es viable ($4) o que cada miembro lo descargue y lo use localmente, enlazándolo desde fuera del repositorio.

Próximas Reuniones:

- **Equipo 2 (EDA/Metadatos):** Se reunirá **mañana (miércoles)** de 9:30 a 11:30 AM para comenzar a trabajar.
- **Reunión General:** Se agendó una reunión corta de seguimiento para el **próximo viernes**, aproximadamente a las 9:30 AM, para revisar el progreso y resolver dudas.

El grupo mostró una actitud positiva y predispuesta a adaptarse y experimentar con la metodología de trabajo.