La reunión se centró en organizar y planificar las siguientes fases del proyecto, que incluyen la creación del *dataset* final y el entrenamiento de modelos.

Ideas y decisiones clave:

- **Creación del *Dataset* y Pruebas de Modelos:**
  - La siguiente parte del proyecto implica **trabajar todos juntos** para crear el *dataset* final y empezar a entrenar con diferentes modelos.
  - Se considera la **instalación de ML Flow** para subir reportes, aunque hay dudas sobre la versión y si causará problemas en local.
  - El proceso general será **crear el *dataset* y luego probar el modelo**.
- **Preprocesamiento de Imágenes (Fase Crítica):**
  - **Recorte a cuadrado y redimensionado:** Es indispensable que todas las imágenes sean cuadradas y tengan una resolución específica, inicialmente **224x224** para el modelo ResNet18.
    - Se propone usar la función `center crop` para mantener la proporción y preservar la región central sin relleno artificial, y `resize`. Se probará esta combinación.
    - Se acordó que el preprocesado de imágenes es la **tarea prioritaria y unificada** que se debe completar rápidamente (en uno o dos días).
  - **Orientación y Rotación de Imágenes:**
    - Se discutió la idea de girar las imágenes para que queden todas en la misma dirección, pero no es lo más fácil.
    - Se mencionó el uso de `random horizontal flip` (para orientación izquierda/derecha) y `random rotation` (para que no aprenda patrones de rotación y generalice mejor).
    - **Conclusión:** Por ahora, **no se orientarán todas las imágenes a un mismo perfil**; se probará sin esta orientación al principio. La orientación y alineación se considera una tarea "extra" para probar más adelante si hay tiempo.
  - **Manejo de Imágenes de Baja Calidad:**
    - Se identificó que un 20% de las imágenes de glaucoma tienen baja calidad (oscuras o bajo contraste).
    - Están etiquetadas en el *dataset* (`quality DF`).
    - **Conclusión:** Las imágenes con brillo y contraste bajo o alto **se mantendrán** inicialmente en el *dataset*. Se podría aplicar transformaciones solo a las de baja calidad, pero se empezará sin transformaciones especiales y se probará.
  - **Limitaciones Computacionales:** El procesamiento de imágenes grandes requiere mucho cómputo, por lo que se **priorizarán imágenes más pequeñas** para evitar el ruido y problemas de generalización del modelo.
    - No es recomendable redimensionar imágenes pequeñas a tamaños más grandes, ya que introduce relleno falso.
    - Se utilizará Google Colab para el cómputo, aunque se reconoce que puede haber limitaciones con gráficas menos potentes.
  - **Modelo Propuesto:** El modelo inicial a utilizar es **ResNet18** con imágenes de 224x224 píxeles.
- **Manejo de Datos Tabulares / Metadatos:**

    - **Casos de un año de edad:** Eran 16 casos de pacientes de un año, todos femeninos.
        - Se consideró que podrían generar confusión o *overfitting* en el modelo debido a su bajo número y su edad muy específica.
        - **Conclusión:** Se decidió **eliminar estos 16 casos** inicialmente, con la opción de probar con ellos si el tiempo lo permite.
    - **Imágenes con "Lendas" (polvo/artefactos):** Se identificaron 480 entradas con comentarios de "lendas".
        - Estas imágenes tienen un diagnóstico asociado además del comentario de "lendas".
        - **Conclusión:** Se decidió **mantener estas imágenes**, ya que dan información y el artefacto no necesariamente impide un buen diagnóstico.
    - **Selección de la Columna Target (Variable a Predecir):**
        - Se debatió sobre usar la columna `ndje` (letras de las enfermedades) o los *embeddings* numéricos (`detail`) que se habían generado, los cuales representaban combinaciones de enfermedades y otras informaciones.
        - Los *embeddings* generaron más de 100 niveles, lo que preocupó sobre la dispersión de datos y la pérdida de potencia del algoritmo.
        - **Conclusión:** Para empezar y por simplicidad, se acordó usar la columna `label` (o `ndje`), que contiene las letras de las enfermedades y es un **target binario/multiclase más sencillo** (0s y 1s). Se puede dejar el *embedding* como una opción para probar si hay tiempo o en pruebas paralelas.
    - **Balanceo de Clases:** Se mencionó la importancia de considerar el balanceo de clases (algunas enfermedades tienen muchas menos imágenes) si se usan todas las enfermedades, pero si se empieza con una sola clase (como la retinopatía diabética), no es tan urgente.
- **Gestión del Proyecto y Flujo de Trabajo (Git):**
    - **Unificación de ramas:** Existe un poco de incertidumbre sobre cómo unificar las ramas de Git, pero se confía en que se puede revertir si hay problemas. Miguel Ángel se encargará de esto con ayuda.
    - **Nueva fase de trabajo:** Se creará una nueva rama para el preprocesado y los modelos.
    - **División de tareas:** Se dividirán las tareas: un grupo se enfocará en el **preprocesado de imágenes** (que debería ser rápido), y otro en **construir un modelo** para probarlo cuanto antes. Una vez terminado el preprocesado, el grupo de imágenes se unirá a la fase de modelos.
    - Se avisará cuando el preprocesado de imágenes esté terminado para que se puedan unificar las ramas sin conflictos.
- **Idioma de la Documentación y el Código:**
    - Se debatió si el proyecto, informes y código deben estar en español o inglés.
    - **Conclusión:** Se decidió que la documentación y los informes estarán en **español** para mayor comodidad. Sin embargo, los **nombres de funciones y variables en el código se mantendrán en inglés**.
- **Próximos Pasos:**

- Se propone una próxima reunión para el lunes por la tarde, alrededor de las 18:00 o 18:30, para revisar el avance del preprocesado.
- Los ausentes (David y Sofía) serán informados de las decisiones.

En resumen, la reunión sentó las bases para una **fase de trabajo intensa y unificada en el preprocesamiento de imágenes**, con decisiones claras sobre cómo manejar los datos (imágenes y metadatos) y una hoja de ruta para avanzar rápidamente hacia las pruebas de modelos. Se prioriza la simplicidad y la agilidad para obtener un primer modelo funcional lo antes posible.