Propósito General de la Reunión: La reunión fue para tomar decisiones de forma colectiva y aunarlas, especialmente en el preprocesado de las imágenes, ya que hay implicaciones importantes para la arquitectura del modelo. Se busca establecer un preprocesado inicial bajo una hipótesis y luego iterar.

Puntos Clave y Conclusiones sobre los Metadatos (Datos Tabulares):

- **Estado del Dataset:** El dataset de metadatos estaba bastante trabajado y no presentaba demasiadas complicaciones iniciales. No hay valores nulos ni duplicados significativos.
- **Problema con la Edad (Casos de 1 Año):**
    - **Observación:** Se detectaron 16 casos en el dataset de pacientes de 1 año de edad, todos ellos niñas. Esto es inusual, ya que el siguiente valor de edad en el dataset salta a 14-15 años.
    - **Hipótesis:** Se considera la posibilidad de que sean **errores en la introducción de datos**. Aunque ChatGPT sugiere que algunos casos de enfermedades en niños de 1 año podrían darse, son muy raros y en el límite, lo que refuerza la idea de un posible error en la adición de los datos. Por ejemplo, se mencionó que la degeneración macular no debería existir en niños.
    - **Conclusión (Pendiente): No se tomó una decisión final** sobre si eliminar estos 16 casos de 1 año o mantenerlos. Se propuso probar el modelo con y sin ellos para ver cómo interfiere en los resultados finales. Se sugiere consultar a un experto (optometrista) para validar la lógica subyacente.
- **Problema con Anotaciones en Imágenes (Lens Dust / Low Image Quality):**
    - **Observación:** Se encontraron casos de pacientes "normales" (sin enfermedades) que tenían comentarios en las keywords como "lens dust" (mota de polvo en las lentes) o "low image quality".
    - **Hipótesis:** Estas anotaciones indican **defectos o artefactos en las imágenes**, como lentes sucias o baja calidad. Hay 480 entradas con "Lens Dust".
    - **Conclusión:** Se decidió **mantener estas imágenes por ahora y tratarlas como ruido**, ya que podrían arreglarse con técnicas de preprocesado de imágenes. La idea es "tirar para adelante con todo" y luego considerar la eliminación si es necesario.
- **Diagnósticos y Etiquetas de Enfermedades:**
    - Se discutió la complejidad de que un ojo pueda tener múltiples enfermedades (separadas por una "coma china").
    - **Propuesta de Tratamiento:** Las enfermedades se transformarán de letras a **números y se añadirán a una lista** por cada caso (embedding). Esto se considera un paso más avanzado que el enfoque categórico original del dataset y podría dar mejores resultados.
- **Almacenamiento de Datos:** Se decidió guardar el dataset preprocesado en formato **Parquet** por su peso y rendimiento, aunque se puede cambiar.
- **Documentación:** Se acordó la necesidad de **documentar todo el proyecto** de forma exhaustiva, incluyendo el análisis de metadatos, el análisis de imágenes, las decisiones tomadas y las herramientas utilizadas.

Puntos Clave y Conclusiones sobre el Análisis y Preprocesado de Imágenes:

- **Calidad de Imagen (Brillo y Contraste):**
    - **Observación:** Muchas imágenes son muy oscuras o tienen bajo contraste, y estas están **ligadas a enfermedades específicas, especialmente el glaucoma**. También hay imágenes con mucho brillo (en la cola derecha de la distribución).
    - **Conclusión: No se eliminarán las imágenes oscuras o con bajo contraste**, ya que están relacionadas con el glaucoma y posibles derrames, lo cual afecta la luminosidad. Se explorará aplicar un **preprocesado especial** (como ecualización) a estas imágenes con baja media de brillo. Se decidió también **visualizar y analizar las imágenes con mucho brillo** para determinar si deben ser tratadas o eliminadas.
- **Orientación y Recorte de Imágenes:**
    - **Discusión:** Se planteó la idea de rotar las imágenes para que todas miren en la misma dirección y unificar la presentación. También se mencionó la posibilidad de hacer recortes para centrarse en áreas de interés (mácula, nervio óptico).
    - **Conclusión (Pendiente):** La rotación se considera importante, pero el recorte de partes específicas de la imagen se ve como una **transformación más severa** que podría aplicarse en una segunda o tercera ronda de iteración, ya que podría eliminar información relevante para distintas dolencias.
- **Resolución y Redimensionamiento de Imágenes:**
    - **Observación:** Las imágenes tienen tamaños muy variados, siendo la mayoría de gran tamaño (1728x2592). Los modelos de redes neuronales requieren tamaños de entrada específicos.
    - **Discusión:** Se reconoce que es mejor **reducir el tamaño** de las imágenes grandes, pero no aumentar el tamaño de las pequeñas, ya que se perdería detalle o se distorsionaría la imagen (salvo con técnicas avanzadas de IA). El padding (rellenar alrededor) se mencionó como una técnica para mantener la coherencia sin distorsionar.
    - **Conclusión:** Se decidió **usar inicialmente 224x224 para el modelo base (ResNet)**, que es un tamaño común de entrada. Se harán pruebas con otros tamaños para otras arquitecturas.
- **Arquitectura del Modelo:**
    - **Conclusión:** Se partirá de un **modelo preentrenado como ResNet18 como base inicial**. Se realizarán pruebas con diferentes arquitecturas y formatos para ver el rendimiento.

Aspectos Generales del Proyecto:

- **Lenguaje:** Se decidió que la **documentación se realizará en español**. Las variables pueden estar en inglés y los comentarios en español, para una mayor visibilidad del proyecto en GitHub, aunque esto implica un trabajo extra.
- **RAG (Retrieval Augmented Generation):** Se ha avanzado en la construcción de un sistema RAG para recopilar información sobre enfermedades. Se decidió enfocarse en **PDFs en español** para evitar problemas con múltiples idiomas y formatos, y se utilizará una herramienta para limpiar el texto de los PDFs.

Próximos Pasos/Acciones Pendientes (a concretar para el jueves):

- **Metadatos:** Decidir el tratamiento final para los 16 casos de edad de 1 año.
- **Imágenes:**
    - Visualizar y analizar las imágenes con **mucho brillo** (cola derecha de la distribución de brillo) para decidir si son válidas o necesitan tratamiento.
    - Confirmar el tratamiento inicial para las imágenes oscuras y con bajo contraste (mantenerlas y/o aplicar preprocesado especial como ecualización).
    - Aunar criterios para la decisión sobre la redimensión de las imágenes y la arquitectura inicial del modelo (ResNet18 con 224x224 como base).
- **General:** Dedicar un tiempo para que todo el equipo revise y dé el "okay" a los planes de preprocesado.