A continuación, se presenta un resumen de la conversación y una tabla con los tiempos y fases del proyecto, junto con las ideas importantes que se discutieron:

**Resumen de la Conversación**

La reunión se centró en la planificación del proyecto, abarcando desde la selección de *datasets* hasta el despliegue final y la gestión del equipo.

**1. Selección y Calidad de *Datasets***:

* Se discutió la complejidad de entender y extraer datos de los *datasets*.
* Se identificó que el *dataset* **Aptos 2019 es repetido** y se podría eliminar, sugiriendo **IDRID** como una buena opción, aunque con pocos datos, útil para complementar.
* El *dataset* **Mesidor** presenta una diferencia en el rango de valoración de RD (0-3 vs. 0-4 en otros), lo que podría complicar la unificación de datos.
* Se mencionó que el *dataset* **OCT de DL** (subido por Sofía) tiene metadatos con muchos nulos de edad/sexo, pero campos útiles como la detección temprana/tardía de la enfermedad.
* Se acordó que los dos primeros *datasets* (posiblemente los mencionados como "el grueso") e **IDRID de Cation** son los mejores para trabajar.
* La **falta de metadatos** es un desafío importante, siendo IDRID el único con algunos metadatos (ej., categorización, condiciones, pacientes, sexo, años), aunque limitados a 147 casos de una enfermedad específica y binaria (0 y 1).
* Se señaló la necesidad de **diferenciar imágenes OCT de Fundus** y no mezclarlas para evitar complicaciones.
* Se reconoce que la parte más complicada de los *datasets* es encontrar **datos clínicos tabulares**.

**2. Propuesta de *Pipeline* y Metodología de Trabajo**:

* Sofía presentó un *pipeline* detallado dividido en clases de imágenes y metadatos clínicos.
  + **Imágenes**: Analizar calidad, añadir ruido a imágenes buenas para evitar *overfitting*, mejorar imágenes malas (o eliminarlas/incluirlas para variabilidad), aplicar normalización, redimensionamiento, *data augmentation* y estandarización de etiquetas.
  + **Datos clínicos**: Análisis descriptivo, ajuste de tipos, manejo de nulos y *outliers*, normalizaciones, manejo de variables categóricas, estandarización de etiquetas y visualización/correlación.
* **Modelado**: Sugirió un **CNN para imágenes y un *Machine Learning* para metadatos**, concatenándolos para un clasificador.
* **Despliegue**: En un *dashboard* (interfaz web para interacción con el cliente) y una API.
* **Trabajo en la nube vs. local**: Se propuso **empezar en local con muestras pequeñas** para pruebas y escalar a la nube una vez que el modelo sea óptimo, para **evitar costos excesivos** y pérdida de tiempo en la nube con modelos no funcionales. Se discutió el uso de **MLFlow, Airflow** (para orquestación y automatización de procesos como ETL) y **Spark/Dask** (para procesamiento paralelo de grandes datos).
* **Integración de IA (RAG)**: La IA (específicamente RAG para sugerencias o tratamientos) se vería como un **plus final**, integrándose en el *dashboard* como un botón de "sugerencias". Se buscarían guías clínicas y *papers* para nutrir el modelo RAG.
* **Organización del equipo**: Se debatió entre un enfoque semanal donde todos tocan todos los módulos para maximizar el aprendizaje y una división por equipos/células para mayor agilidad. La decisión fue adoptar un **modelo mixto/híbrido**, con trabajo en parejas o tríos y reuniones semanales para puesta en común y revisión.

**3. Objetivos y Alcance**:

* Los *targets* prioritarios serían **retinopatía diabética (RD) de 0 a 4** y **edema macular diabético de 0 a 2**, basándose en los *datasets* disponibles.
* El usuario final de la aplicación serían **profesionales de la medicina** (oftalmólogos). Se valoró la posibilidad de generar informes para pacientes.
* Se estima que el proyecto tendrá un coste en horas directas de **910 horas** en equipo, que se traduciría en un coste económico de **60.000 EUR**, con un factor de aumento del 30-50% por horas indirectas.

**4. Gestión del Proyecto**:

* Se acordó utilizar una **herramienta gráfica como Trello** para organizar tareas y visualizar el progreso.
* Se enfatizó la importancia de la **documentación continua** del proyecto, incluyendo errores y soluciones, posiblemente por módulo o en un documento compartido. Se sugirió el uso de **capturas de pantalla (ej., GreenShot)** para enriquecer la documentación visual.
* Se definieron **fechas límite** clave para cada fase del proyecto (ver tabla a continuación).

**Tabla de Tiempos y Fases del Proyecto**

| **Fase del Proyecto** | **Fecha Límite / Periodo** | **Responsabilidad (Modalidad)** | **Objetivos Clave** |
| --- | --- | --- | --- |
| **Fase 1: Exploración y Selección de *Datasets*** | **Martes, 29 de Julio** | Todo el equipo (conjunta) | Definir *datasets* (imágenes y tabulares) y *targets* definitivos (Retinopatía Diabética 0-4, Edema Macular Diabético 0-2). Buscar documentos para RAG. |
| **Fase 2: Limpieza y Preprocesado (ETL)** | **Martes, 5 de Agosto** | Equipos divididos (Metadatos/Imágenes) | Limpieza general de datos (borrosas, nulos). Aplicar normalización, redimensionamiento, *data augmentation*, filtros. Preparar datos limpios para el modelo. Trabajar localmente con muestras pequeñas. |
| **Fase 3: *Feature Engineering* y Modelado** | **Martes, 12 de Agosto** (Feature Eng.)**Viernes, 22 de Agosto** (Modelos) | Equipos divididos (Modelo/Arquitectura) | Testear que los datos son correctos. Desarrollar y probar varios modelos (CNN, ML, *transfer learning*, *fine-tuning*). Evaluar resultados y decidir el modelo final. |
| **Fase 4: Despliegue, IA (RAG) y Entrega Final** | **Viernes, 5 de Septiembre** | Equipos divididos (API/Interfaz/RAG) | Implementar el modelo en la nube. Desarrollar una API y un *dashboard* interactivo. Integrar funcionalidades de IA (RAG). Finalizar la documentación, presentación y vídeo del proyecto. |
| **Presentación Final** | Lunes, 11 de Septiembre | Todo el equipo | Presentar el proyecto y sus resultados. |

*Nota: Las fechas son estimaciones y pueden ajustarse según el progreso del equipo y las dificultades encontradas.*

**Ideas Importantes Discutidas**

* **Prioridad a la Calidad de los Datos**: Se enfatizó repetidamente que "si los datos son basura, la salida es basura". Una buena elección y un preprocesado exhaustivo son fundamentales para el éxito del proyecto.
* **Modelo Multimodal**: La visión es combinar el poder de los datos de imagen con los datos clínicos tabulares (metadatos) para un diagnóstico más completo.
* **Optimización de Recursos**: Dada la preocupación por el tiempo y los costos, se prefiere probar los modelos y el preprocesado intensivo localmente con muestras pequeñas antes de escalar a la nube.
* **Aprendizaje y Crecimiento Profesional**: Varios miembros del equipo expresaron el deseo de tocar todas las fases del proyecto (a través del modelo mixto) para maximizar su aprendizaje en diferentes áreas (ciencia de datos, ingeniería de datos, ML Ops).
* **Incorporación de Herramientas de Vanguardia**: Se propuso utilizar herramientas como Airflow para la orquestación de datos y Spark/Dask para el procesamiento paralelo de grandes volúmenes, lo que aportaría un "plus" y demostraría la aplicación de conocimientos adquiridos en el *bootcamp*.
* **Foco en el Usuario Profesional**: La interfaz y los resultados deben ser **fácilmente interpretables por médicos**, incluso considerando la generación de informes en PDF para el historial clínico o sugerencias de tratamiento.
* **Gestión Ágil y Comunicación Constante**: La adopción de metodologías como Scrum/Sprints y el uso de herramientas como Trello, junto con reuniones semanales, buscan facilitar la coordinación y el seguimiento del progreso en un proyecto con tiempos ajustados.

La planificación de un proyecto es como la construcción de un puente: no solo necesitas saber dónde quieres llegar, sino también cómo distribuir el peso, asegurar los cimientos (los *datasets*), y coordinar a los constructores (el equipo) para asegurar que cada sección (fase) se conecte sin problemas, incluso cuando el terreno (los datos y el tiempo) es incierto.