Evaluación Proyecto:

1. Entregables del Proyecto (Revisión)

1. Resumen del Problema (Bio-Actives)

Data source

La identificación de compuestos bioactivos en fuentes botánicas food-grade a partir de datos crudos de alto rendimiento (como LC-MS) es un cuello de botella significativo. La información sobre la bioactividad de los metabolitos existe en bases de datos públicas (PubChem, etc.) y en la literatura científica, pero está desconectada de los flujos de trabajo de análisis de datos. Actualmente, vincular una *feature* cromatográfica (ej. un pico con un m/z y tiempo de retención) a una *anotación putativa* (un compuesto químico) y luego a su *actividad biológica* es un proceso manual, lento y que requiere alta especialización.

2. Objetivo del Proyecto

Desarrollar un pipeline RAG que automatice la anotación funcional de *features* metabolómicas. El sistema permitirá a un investigador consultar sobre una *feature* específica y recibir un informe resumido que incluya su identidad putativa y sus bioactividades potenciales, citando las fuentes de datos públicos e internos.

Caso de Uso Práctico (Flujo de trabajo)

Paso 1: Consulta del Usuario (Query) El analista ingresa: `"Tengo una feature con m/z 449.107, RT 8.2 min, detectada en mi muestra de 'Té Verde'. ¿Qué es y qué bioactividad tiene?"`

Paso 2: Recuperación (Retrieval) El sistema RAG descompone la consulta y busca en su base de datos vectorial:

- **Búsqueda de m/z:** Encuentra `m/z 449.107` en la base de datos de compuestos.
    - *Resultado 1 (Público):* "Fórmula C21H20O12. Compuesto: Myricetina 3-galactósido."
    - *Resultado 2 (Interno):* "Feature `mz449.1_rt8.1` anotada como Myricetina-derivado en 'Muestra Arándano 004'."
- **Búsqueda de Bioactividad (para Myricetina):**
    - *Resultado 3 (PubChem BioAssay):* "Ensayo ID 5678: Myricetina inhibe la agregación plaquetaria..."

- *Resultado 4 (PubMed Abstract):* "Myricetin... posee fuertes propiedades antidiabéticas..."

Paso 3: Generación Aumentada (Generation) El LLM recibe todos estos resultados como contexto y sintetiza la respuesta:

> "La *feature* (m/z 449.107, RT 8.2 min) en su muestra de 'Té Verde' se anota putativamente como **Myricetina 3-galactósido** (basado en coincidencia de masa C21H20O12).
>
> **Actividades Biológicas Potenciales reportadas para Myricetina:**
>
> - Inhibición de la agregación plaquetaria (Fuente: PubChem BioAssay 5678).
> - Propiedades antidiabéticas (Fuente: PubMed ID [X]).
>
> **Contexto Interno:** Una feature similar (m/z 449.1, RT 8.1 min) fue previamente detectada y anotada como un derivado de Myricetina en nuestros análisis de 'Muestra Arándano 004'."

1. Resumen del Problema (Preservantes)

Data source

La industria de alimentos está en una transición hacia el "clean label", buscando reemplazar conservantes sintéticos (ej. benzoatos, sorbatos) por alternativas naturales (ej. aceites esenciales, extractos vegetales, nisina). Sin embargo, la eficacia de *cualquier* conservante depende críticamente de parámetros fisicoquímicos como el pH, la actividad de agua (a_w), la concentración y el microorganismo objetivo.

Esta información (la "matriz comparativa") existe, pero está dispersa en miles de artículos de investigación y fichas técnicas. Un formulador no puede consultar eficientemente esta "base de conocimiento" para responder a una pregunta como: *"¿Qué alternativa natural a 1000 ppm de Sorbato de Potasio puedo usar en un producto con pH 4.5 y aw 0.95 para inhibir mohos?"*

2. Objetivo del Proyecto

Desarrollar un pipeline RAG que actúe como un "experto en formulación" capaz de recomendar alternativas de conservantes naturales. El sistema recibirá como entrada un producto (con sus parámetros de pH, aW) y/o un conservante sintético a reemplazar, y generará una recomendación de una alternativa natural, incluyendo su concentración de uso sugerida y la fuente bibliográfica que respalda su eficacia en esas condiciones.

Caso de Uso Práctico (Flujo de Reemplazo)

Paso 1: Consulta del Usuario (Query)

"Necesito reemplazar 800 ppm de Benzoato de Sodio en una salsa con pH 4.2 y aw 0.97. El objetivo es inhibir levaduras (como Zygosaccharomyces bailii). ¿Qué alternativas naturales tengo?"

Paso 2: Recuperación Híbrida (Retrieval)

1. **Filtro por Metadatos (Condiciones):**
    - type == "Natural"
    - target_microorganism CONTAINS "Zygosaccharomyces" OR "Levadura"
    - condition_ph BETWEEN 4.0 AND 4.5
    - condition_aw BETWEEN 0.95 AND 0.98
2. **Búsqueda Vectorial (Contexto):**
    - Busca embeddings similares a "alternativa a benzoato" o "eficacia contra levaduras osmotolerantes".

Paso 3: Generación Aumentada (Generation)

El LLM recibe los "hechos" recuperados y genera la recomendación:

> **Respuesta de Preserv-RAG:**
>
> "Para reemplazar el Benzoato de Sodio (800 ppm) en su producto (pH 4.2, aW 0.97) contra levaduras como *Zygosaccharomyces bailii*, se sugieren las siguientes alternativas naturales:
>
> 1. **Extracto de Clavo (Eugenol):**
>   - **Concentración Sugerida:** 400-600 ppm.
>   - **Respaldo:** Un estudio (Fuente: [X]) demostró una inhibición completa de *Z. bailii* a 500 ppm en un sistema modelo con pH 4.0 y aW 0.97.
> 2. **Nisina (Péptido Bacteriano):**
>   - **Concentración Sugerida:** 250-500 ppm (en combinación).
>   - **Respaldo:** Aunque es primariamente antibacteriana, la Nisina ha mostrado eficacia contra levaduras a pH bajo (Fuente: [Y]). Se recomienda un estudio de desafío, posiblemente en combinación con extracto de romero.
>
> **Consideración:** El Benzoato de Sodio es altamente efectivo a pH < 4.5. Las alternativas naturales pueden requerir concentraciones más altas o una estrategia de "obstáculos múltiples" (hurdle technology) para lograr la misma vida útil."

El proyecto final, que se entrega el **Miércoles 20 de diciembre**, se divide en dos hitos principales y la presentación final:

- **Hito 1(35% de la nota final): Creación del sistema base (Baseline)** (Entrega: Viernes 22 de noviembre): Un **RAG Naive** que debe incluir:

    - Un *pipeline* de procesamiento de datos (ingesta, organización y limpieza).
    - *Chunking* de documentos.
    - Configuración de una base de datos de vectores con embeddings.
    - Un *retriever* simple para recuperar *chunks*.
    - Un **benchmark** con métricas base para establecer el **baseline** del sistema. La funcionalidad esperada es: dada una *query*, se devuelven *Documents* relevantes y métricas de rendimiento.
    - Todo en una interfaz de Streamlit donde:
        - Se puedan hacer consultas y se retornen los documentos relevante
        - Se vean las metricas de rendimiento del baseline
- **Hito 2(35% de la nota final): Mejorando el baseline** (Entrega: Miércoles 20 de diciembre):
    - Integración de técnicas avanzadas de **retrieval** (preprocesamiento de *queries*, *hybrid search*, *self-query*, *reranking*, *repacking*).
    - Implementación de un *pipeline* de **generación aumentada** para mejorar la calidad de las respuestas.
    - **Reporte de métricas** que justifique y valide las mejoras en el rendimiento respecto al *baseline*.
    - Todo en una interfaz de Streamlit donde:
        - Se puedan hacer consultas y se retornen las respuestas aumentadas por el LLM
        - Se vean las metricas de rendimiento del baseline versus las versiones mejoradas

- **Presentación Final de la Solución(30% de la nota final)** (Miércoles 20 de diciembre):
    - Una **interfaz de usuario en *Streamlit*** que corra localmente y demuestre el *pipeline* completo.
    - Cada componente (ingesta, recuperación, generación) debe estar **modularizado y visualmente separado** en la interfaz.
    - Entrega del código en un **repositorio público de GitHub**, bien documentado y reproducible.