Challenge de Clasificación Biomédica con IA

Hackatón Source Meridian - Techsphere Colombia

Integrantes del equipo Andrés Julián Múnera Julio César Posada

Medellín, Colombia - 2025

Análisis exploratorio y comprensión del problema

El análisis exploratorio tuvo como finalidad comprender la estructura del dataset de artículos médicos, identificar retos para el modelado y definir las primeras hipótesis de trabajo en torno al problema de clasificación multi-etiqueta.

Descripción general del dataset

El conjunto consta de **3,565 artículos médicos**, cada uno descrito por:

- **title** (título),
- **abstract** (resumen científico),
- **group** (categoría(s) médica(s), con estructura multi-label).

No se evidencian valores nulos y los duplicados hallados corresponden a distintas versiones de un mismo estudio, lo que confirma la calidad general de los datos.

Análisis de etiquetas y distribución de clases

El campo *group* presenta hasta cuatro categorías simultáneas: *neurological, cardiovascular, hepatorenal* y *oncological*.

- Predomina la clase **neurological (≈50%)**, seguida por **cardiovascular (≈36%)**, **hepatorenal (≈31%)** y **oncological (≈17%)**.
- El **69% de los artículos tienen una sola etiqueta**, el 28% poseen dos y un número reducido alcanza tres o cuatro.
- La matriz de co-ocurrencia muestra fuertes asociaciones entre *neurological–cardiovascular* y *cardiovascular–hepatorenal*, lo que sugiere traslape semántico.

Este desbalance y la multi-etiqueta implican un reto importante: modelos tradicionales tienden a favorecer las clases mayoritarias, reduciendo la capacidad de generalización hacia categorías menos representadas.

Análisis textual

- **Longitud de títulos y abstracts:** los títulos son cortos (5–15 tokens), mientras que los abstracts son mucho más extensos y ricos en información semántica,

consolidándose como la principal fuente de características.

- **Palabras frecuentes:** términos como *patients, study, methods, disease* confirman la naturaleza científica y clínica del corpus.

- **Top términos discriminativos (TF-IDF):**

    - *Cardiovascular*: *heart, disease, investigation*.

    - *Hepatorenal*: *renal, methods*.

    - *Neurological*: *outcomes, implications*.

    - *Oncological*: *cancer, breast, mutations*.

Estos patrones validan la presencia de vocabulario característico por categoría, aunque con superposición temática en palabras generales (*patients, results*).

Principales retos identificados

1. **Clasificación multi-label**, con necesidad de arquitecturas especializadas.

2. **Desbalance de clases**, que puede afectar el F1-score ponderado si no se aplican técnicas de balanceo.

3. **Solapamiento semántico** entre categorías, que exige modelos capaces de capturar dependencias contextuales.

4. **Heterogeneidad en longitudes de texto**, lo que requiere tokenización y normalización cuidadosa.

Conclusión

El EDA permitió identificar la estructura, calidad y retos centrales del dataset. Se confirma que el problema no se limita a una clasificación estándar, sino a un desafío **multi-label desbalanceado y semánticamente complejo**. Esto justifica el uso de técnicas avanzadas de NLP y validación robusta, y orienta las fases posteriores hacia la implementación de embeddings contextuales y métricas adecuadas como el F1 ponderado.

Preparación, preprocesamiento y justificación

El preprocesamiento del dataset se diseñó con el objetivo de transformar los textos médicos en representaciones aptas para modelos de clasificación multi-etiqueta, manteniendo un balance entre simplicidad computacional y robustez en el manejo semántico.

1. Limpieza y consistencia de datos

- Eliminación de duplicados exactos en títulos y abstracts.
- Revisión manual de duplicados parciales (ejemplo: artículos Cochrane en versiones diferentes) para evitar pérdida de información relevante.
- Normalización de textos: conversión a minúsculas, eliminación de caracteres especiales, números aislados y espacios redundantes.

2. Tokenización y normalización semántica

- **Tokenización basada en expresiones regulares**, eliminando *stopwords* en inglés para reducir ruido.
- **Lematización** de términos biomédicos (con NLTK y SciSpacy) para agrupar variaciones morfológicas (*e.g.*, “mutations” → “mutation”).
- Generación de métricas de longitud de títulos y abstracts para ajustar hiperparámetros como *max sequence length*.

3. Representación vectorial

Se exploraron dos enfoques:

- **TF-IDF (baseline):** permitió obtener rápidamente términos característicos por clase, facilitando la interpretación inicial.
- **Embeddings contextuales (modelo principal):** se adoptó **BioBERT** (un modelo entrenado en literatura biomédica), dado que mejora la captura de contexto clínico frente a TF-IDF o Word2Vec. Esto se justifica porque abstracts médicos contienen terminología especializada que requiere representaciones profundas.

4. Codificación de etiquetas (multi-label)

- Se aplicó **binarización multi-etiqueta** (*MultiLabelBinarizer*), expandiendo el campo *group* en vectores binarios (ejemplo: `[1, 0, 1, 0]` para un artículo *neurological* + *hepatorenal*).

- Esta representación es indispensable para algoritmos compatibles con clasificación multi-etiqueta.

5. Manejo del desbalance

- Se calcularon pesos de clase proporcionales al inverso de su frecuencia.

- Estos **class weights** fueron integrados en la función de pérdida, con el fin de reducir el sesgo hacia categorías mayoritarias como *neurological*.

6. División del dataset

Se implementó un **split estratificado** en proporción 70/15/15 para entrenamiento, validación y prueba, asegurando que la distribución de etiquetas múltiples se mantuviera balanceada en cada subconjunto.

Conclusión

El pipeline de preprocesamiento combina técnicas clásicas (TF-IDF) con representaciones avanzadas (BioBERT), asegurando tanto interpretabilidad como desempeño. La inclusión de estrategias de balanceo y división estratificada del dataset responde directamente a los retos identificados en el EDA, consolidando una base sólida para el entrenamiento de modelos multi-label robustos.

Selección y diseño de la solución

El problema planteado consiste en una **clasificación multi-etiqueta de artículos médicos** a partir de títulos y abstracts. Esto implica que cada muestra puede pertenecer simultáneamente a varias categorías (*neurological, cardiovascular, hepatorenal, oncological*), lo cual demanda un diseño específico tanto en la representación del texto como en el algoritmo de clasificación.

1. Enfoque inicial (baseline)

Como línea de base, se empleó un pipeline con:

- **TF-IDF** como vectorización,

- **One-vs-Rest Logistic Regression** como clasificador.

Este enfoque sirvió para establecer un punto de comparación cuantitativo, evaluando métricas de accuracy y F1 ponderado. Sin embargo, se observó que las limitaciones de TF-IDF (pérdida de contexto semántico y dificultad para capturar relaciones entre etiquetas) reducían la capacidad de generalización del modelo.

2. Enfoque avanzado (modelo principal)

Con el fin de superar las limitaciones del baseline, se adoptó un **modelo basado en transformers biomédicos**:

- **BioBERT** (Bidirectional Encoder Representations from Transformers entrenado en PubMed y PMC), seleccionado por su capacidad de capturar contexto en lenguaje biomédico.

- Se utilizó la salida del *[CLS] token* como embedding de frase, sobre el cual se añadió una capa densa con activación *sigmoid* para predecir la probabilidad de pertenencia a cada clase.

3. Estrategia multi-label

Se implementó una arquitectura de salida con **activaciones sigmoid independientes** y función de pérdida **binary cross-entropy con class weights**, lo que permite al modelo asignar múltiples etiquetas simultáneamente.

Se compararon distintas variantes:

- **Binary Relevance** (independencia entre etiquetas),

- **Classifier Chains** (dependencia secuencial entre etiquetas),

- **Red neuronal multi-salida** (implementada sobre BioBERT).

El mejor desempeño se obtuvo con el modelo BioBERT fine-tuneado con *binary cross-entropy* y pesos de clase.

4. Justificación técnica

La elección de BioBERT frente a TF-IDF se fundamenta en:

1. **Mayor capacidad semántica**: capta dependencias a nivel de contexto médico que no se reflejan en el conteo de palabras.

2. **Adaptación al dominio**: al estar entrenado en corpus biomédico, BioBERT supera modelos generales como BERT-base en tareas de clasificación médica.

3. **Escalabilidad**: la arquitectura permite transfer learning y ajuste fino con un dataset relativamente pequeño (≈3.5k abstracts).

5. Integración futura

El diseño planteado permite integrar:

- **Explicabilidad** mediante *LIME/SHAP*, mostrando qué términos influyen en cada predicción.

- **Demo interactiva en tiempo real**, donde el usuario ingresa un abstract y el sistema devuelve categorías con sus probabilidades.

Conclusión

El diseño de la solución se estructuró en dos niveles: (i) un baseline interpretable (TF-IDF + regresión logística), y (ii) un modelo avanzado basado en **BioBERT fine-tuneado para clasificación multi-etiqueta**. Esta combinación permite medir el aporte de representaciones profundas y justifica la elección de una arquitectura moderna y escalable, capaz de abordar los retos de desbalance, solapamiento semántico y complejidad propia del texto biomédico.

El proceso de validación se centró en evaluar el rendimiento del modelo en un escenario de **clasificación multi-etiqueta**. Para ello, se analizaron métricas por clase y se construyó una matriz de confusión binaria para cada etiqueta.

1. Estrategia de validación

- Se realizó una división en conjuntos de entrenamiento y prueba preservando la distribución de etiquetas (estratificación).

- La métrica principal fue el **F1-score ponderado**, complementado con **precision** y **recall** por clase para identificar fortalezas y debilidades del modelo.

2. Resultados por categoría

Cardiovascular

- TP = 210, FN = 46, FP = 5, TN = 452

- Precision = **97.7%**, Recall = **82.0%**
  El modelo presenta un recall aceptable, pero con 46 falsos negativos que evidencian cierta dificultad en capturar todos los casos de esta categoría.

Neurological

- TP = 317, FN = 40, FP = 30, TN = 326

- Precision = **91.3%**, Recall = **88.8%**
  Excelente recall, aunque con la mayor cantidad de falsos positivos (30), lo que indica que el modelo tiende a **sobredetectar esta clase**, posiblemente por su frecuencia alta y vocabulario compartido con otras categorías.

Oncological

- TP = 89, FN = 47, FP = 1, TN = 576

- Precision = **98.9%**, Recall = **65.4%**
  El modelo es altamente preciso, pero con un recall bajo: más de un tercio de los artículos oncológicos no fueron detectados. Esto sugiere que el modelo es **conservador** en esta clase y pierde ejemplos con lenguaje menos típico.

Hepatorenal

- TP = 166, FN = 57, FP = 2, TN = 488

- Precision = **98.8%**, Recall = **74.4%**
  La precisión es muy alta, pero el recall moderado refleja una cantidad considerable de falsos negativos.

3. Interpretación global

- **Precisión general elevada:** el modelo clasifica correctamente cuando asigna una categoría (≈ 92% – 99% en todas las clases).

- **Recall variable:** las clases *oncological* y *hepatorenal* presentan más falsos negativos, lo que indica que el modelo **prefiere no predecir antes que equivocarse**, penalizando la cobertura.

- **Tendencia a sobredetectar "neurological"**: consistente con el análisis exploratorio, donde esta categoría es la más frecuente y presenta alta co-ocurrencia con otras.

4. Conclusión

El modelo logra un desempeño sólido, con un balance favorable en **precision**, aunque requiere mejoras en **recall** para evitar la pérdida de artículos relevantes, especialmente en categorías con menor frecuencia. Estos hallazgos orientan a dos posibles mejoras:

1. Ajustar los **umbrales de decisión** por clase en lugar de usar 0.5 fijo.

2. Implementar **técnicas de data augmentation o balanceo** (oversampling, focal loss) para mejorar la detección de etiquetas menos representadas.

Presentación y reporte

El presente informe se estructuró con el objetivo de garantizar claridad, coherencia narrativa y trazabilidad en el proceso de construcción de la solución de clasificación médica multilabel. La organización de los contenidos responde a un flujo lógico: **análisis exploratorio de los datos, preparación y preprocesamiento, diseño de la solución, validación con métricas y conclusiones finales**, lo que facilita la comprensión tanto a audiencias técnicas como no técnicas.

Evidencias y recursos utilizados

Una característica diferencial de este trabajo fue la inclusión de técnicas de **IA generativa para la revisión crítica y la documentación**. Con el fin de asegurar un análisis profundo y reproducible, se utilizaron prompts cuidadosamente diseñados que guiaron la evaluación del EDA y orientaron las decisiones posteriores. El prompt inicial empleado fue el siguiente:

Prompt utilizado

Contexto

Este documento PDF contiene el análisis exploratorio de datos (EDA) para un reto de clasificación médica multilabel. El dataset está compuesto por títulos y abstracts de artículos biomédicos, clasificados en categorías como *neurological, cardiovascular, oncological* y *hepatorenal*.

Solicitud de Revisión Detallada

Por favor, realiza una evaluación crítica y constructiva del EDA en las siguientes dimensiones:

1. Estructura y Organización

- **¿El análisis sigue un flujo lógico y progresivo?**

- **¿Las secciones están claramente definidas, bien justificadas y completas?**
- **¿Falta algún componente esencial para un EDA orientado a Machine Learning?**

2. Rigor Técnico y Código

- **Identifica posibles errores en cálculos estadísticos o interpretaciones.**
- **Evalúa la validez de las visualizaciones y métricas empleadas.**
- **Sugiere mejoras de eficiencia, legibilidad y mejores prácticas en el código.**
- **¿Las librerías y métodos usados son los más adecuados para este contexto biomédico multilabel?**

3. Validez de las Conclusiones

- **CRÍTICO: Verifica que cada conclusión esté respaldada directamente por los datos y visualizaciones.**
- **Detecta conclusiones exageradas, incorrectas o sin soporte.**
- **¿Los números, porcentajes y métricas coinciden con los resultados presentados?**
- **¿Las recomendaciones propuestas se derivan de manera lógica y consistente del análisis?**

4. Completitud del Análisis

- **¿Se incluyeron todos los análisis relevantes para un problema multilabel?**
- **¿Se exploraron relaciones importantes entre variables (ej. co-ocurrencia de etiquetas, longitud de textos)?**
- **¿El análisis del desbalance de clases fue suficiente y bien presentado?**

5. Calidad de Visualizaciones

- **¿Los gráficos comunican de manera efectiva los hallazgos?**

- **¿Son apropiados para la audiencia (equipo técnico, jueces de hackatón)?**
- **Sugiere visualizaciones adicionales que podrían aportar valor (ej. diagramas de co-ocurrencia, distribuciones comparativas).**

6. Preparación para Modelado

- **¿El EDA entrega *insights accionables* para la construcción del modelo?**
- **¿Las recomendaciones de preprocesamiento están justificadas con evidencia?**
- **¿Se identificaron correctamente los principales desafíos del dataset (multi-label, desbalance, longitud, duplicados)?**

Formato de Respuesta Esperado

Estructura tu evaluación en los siguientes apartados:

1. **Fortalezas principales del análisis.**
2. **Errores críticos encontrados (con correcciones propuestas).**
3. **Mejoras sugeridas organizadas por dimensión (estructura, código, visualizaciones, etc.).**
4. **Análisis faltantes que deberían incluirse.**
5. **Recomendaciones finales para la preparación del modelado y siguientes fases del proyecto.**

Expectativas

- **Sé específico con ejemplos concretos tomados del documento.**
- **Prioriza observaciones que impacten en decisiones de modelado.**

- **Ofrece soluciones prácticas, no solo críticas.**

- **Considera en todo momento el contexto de un reto de clasificación médica multilabel.**

Este prompt permitió generar retroalimentación iterativa que fue incorporada directamente en la mejora del análisis y en la construcción del pipeline final.

Adicionalmente, se contó con la asesoría de un profesor experto en la Universidad EAFIT, cuyas recomendaciones se integraron al diseño metodológico. Las discusiones técnicas fueron plasmadas inicialmente en pizarras físicas y luego digitalizadas en diagramas de flujo que representan el proceso de transformación: **texto → embeddings → reducción dimensional (PCA/UMAP) → clasificación supervisada (modelos ML y ensambles) → evaluación con métricas**.

Visualizaciones y recursos gráficos

Para garantizar la comunicabilidad de los resultados, se incluyeron visualizaciones y recursos gráficos como:

- Distribución de etiquetas y análisis multilabel.

- Heatmap de co-ocurrencia entre categorías.

- Nubes de palabras y términos más relevantes por clase según TF-IDF.

- Matriz de confusión por categoría.

- Dashboard interactivo con métricas globales (F1-score ponderado, Accuracy), matriz de confusión y demostrador de clasificación en tiempo real.

Estas evidencias se complementaron con diagramas vectorizados que resumen el pipeline técnico, aportando claridad y profesionalismo en la presentación.

Estilo y diseño visual

El informe empleó un estilo uniforme, con tipografías limpias, esquemas cromáticos consistentes y gráficas estandarizadas para asegurar legibilidad. Las imágenes de los tableros originales se incorporan como referencia histórica del proceso creativo, pero fueron acompañadas por diagramas digitalizados que refuerzan la claridad visual.

Transparencia y trazabilidad

Se documentaron los prompts, los resultados intermedios y las decisiones técnicas de manera explícita. Esto asegura que el proceso pueda ser auditado, reproducido y mejorado por terceros, fortaleciendo la transparencia de la solución propuesta.

Conclusión

La presentación del proyecto integra claridad narrativa, trazabilidad del proceso y evidencias visuales robustas (ver en el repositorio de GitHub). La combinación de prompts, diagramas, métricas y un prototipo interactivo permite no solo mostrar los resultados, sino también transmitir el razonamiento técnico y estratégico que guió el desarrollo de la solución.