Informe Final del Proyecto de Clasificación Biomédica con IA

TRIPLE M

Duván santiago Mendivelso, Andrés Camilo Moreno, Juan Diego Muñoz

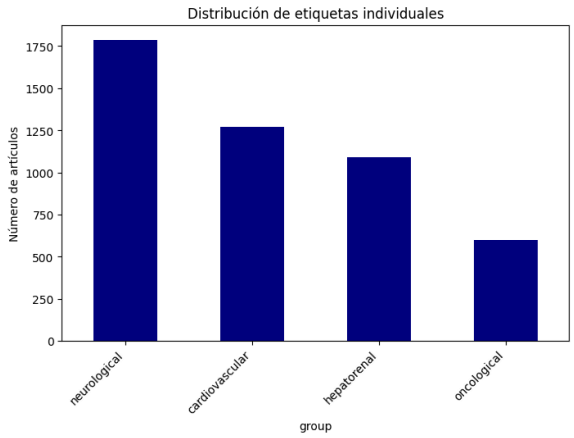
Introducción

El presente proyecto tuvo como objetivo desarrollar un sistema automático de clasificación de artículos biomédicos a partir de sus títulos y resúmenes (abstracts). El reto consistió en asignar cada texto a un dominio médico (neurological, cardiovascular, hepatorenal u oncological) utilizando técnicas de aprendizaje automático (Machine Learning).

Análisis Exploratorio

El dataset estaba compuesto por 3.565 artículos y contenía tres columnas: *title, abstract y group.*

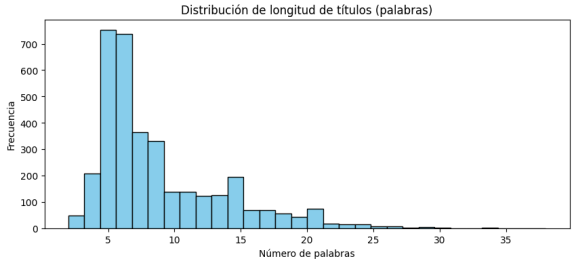
No se encontraron valores nulos ni duplicados.

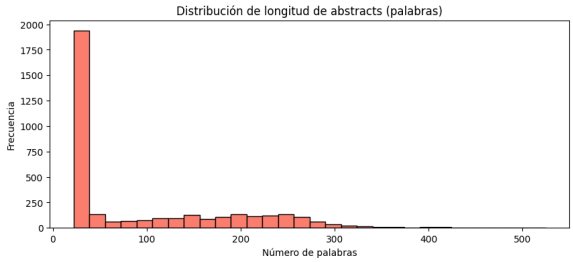


Al realizar el conteo por etiquetas se evidencia que el dataset presenta un desbalance teniendo la etiqueta oncological casi la mita de datos que la etiqueta con más datos que es neurological, por ende más adelante necesitamos usar una estrategia de balanceo

Análisis estadístico

	title_len	abstract_len
count	3565.000000	3565.000000
mean	8.729032	100.056662
std	4.878152	93.066759
min	2.000000	22.000000
25%	5.000000	31.000000
50%	7.000000	37.000000
75%	11.000000	172.000000
max	38.000000	525.000000

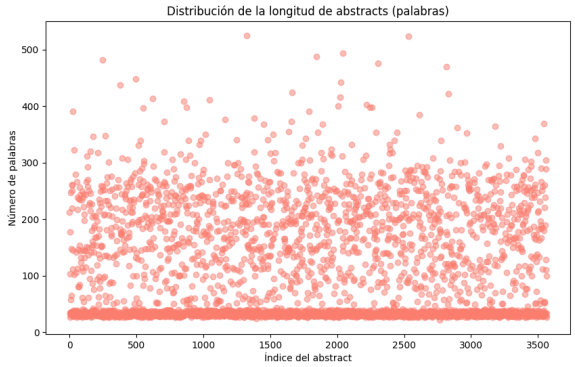




Al analizar las estadísticas descriptivas de la longitud de los títulos y resúmenes, se observa que los títulos presentan en promedio 9 palabras, con un rango entre 2 y 38. En cuanto a los resúmenes, la media se sitúa en aproximadamente 100 palabras, con una variabilidad considerable (desviación estándar de 93) y un rango que va desde 22 hasta 525 palabras.

Aunque se identificaron resúmenes con más de 400 palabras, no se consideran valores atípicos en sentido estricto, ya que la extensión de un abstract depende en

gran medida del estilo del autor y de las normas editoriales de cada revista.



Resúmenes con longitudes significativamente mayores, llegando en casos excepcionales a superar las 500 palabras. Estos valores, aunque poco frecuentes, no necesariamente constituyen errores en los datos, ya que la extensión del abstract puede variar en función de las normas de cada publicación o del estilo de los autores.

Preparación y Preprocesamiento

Con el fin de evaluar la utilidad de las palabras en el proceso de clasificación, se realizó un conteo de frecuencia tanto en títulos como en resúmenes. Este análisis permitió identificar la presencia recurrente de artículos, preposiciones y conectores comunes en inglés (por ejemplo, and, in, of, the), los cuales no aportan información semántica relevante para discriminar entre categorías específicas. La detección de estos términos resulta clave, ya que su alta frecuencia podría introducir ruido en el modelo y dificultar la correcta asociación de un texto con su clase correspondiente.

Se aplicó TF-IDF por separado a title y abstract.

Se eliminaron stopwords para reducir ruido.

No se usaron técnicas de oversampling/undersampling debido al riesgo de generar incoherencias semánticas. En su lugar, se incorporó class_weight="balanced" en los modelos.

Selección y Diseño de la Solución

5 razones por las que el Machine Learning supera a la IA y a los agentes inteligentes en la clasificación de artículos médicos

En este problema de clasificación de artículos médicos a partir del title y el abstract es más adecuado emplear modelos de Machine Learning (ML) en lugar de enfoques basados en IA genérica o agentes de IA.

1. Naturaleza del problema

El reto consiste en un problema de clasificación supervisada de texto, donde se cuenta con datos de entrada (títulos y resúmenes) y etiquetas de salida (dominio médico).

Este tipo de problema se ajusta perfectamente al marco de Machine Learning, que busca aprender patrones estadísticos a partir de datos para realizar predicciones precisas.

2. Eficiencia y simplicidad

Los modelos de Machine Learning (como regresión logística, SVM o redes neuronales simples) son menos costosos computacionalmente que agentes de IA complejos.

Además, requieren menos recursos de entrenamiento, menos infraestructura y permiten un despliegue más sencillo y rápido en entornos reales.

3. Evita sobreingeniería

Usar un agente de IA (capaz de razonar, interactuar dinámicamente o tomar decisiones autónomas) no aporta valor adicional en este caso, ya que el objetivo no

es interactuar con el usuario ni resolver múltiples tareas, sino clasificar textos.

Introducir agentes aquí sería innecesariamente complejo y podría incluso añadir ruido en lugar de mejorar el desempeño.

4. Transparencia y explicabilidad

Los algoritmos de Machine Learning tradicionales ofrecen métricas claras (precisión, recall, F1) y permiten analizar la importancia de palabras o características.

Esto es clave en el ámbito médico, donde la interpretabilidad es un requisito importante.

5. Control de los datos

Con Machine Learning, los datos se preprocesan de manera controlada (TF-IDF, stopwords, normalización), mientras que un agente de IA podría generar, sintetizar o transformar información de formas no verificables, lo cual no es deseable en un dominio sensible como la medicina.

Se compararon tres modelos:

- Linear SVC
- Logistic Regression
- XGBoost

Los tres modelos obtuvieron resultados similares (F1 macro ≈ 0.85), pero se seleccionó Support Vector Machine (SVM) por:

Manejo robusto de alta dimensionalidad.

Posibilidad de ajustar pesos para clases desbalanceadas.

Mayor estabilidad frente a ruido semántico.

5. Resultados y Métricas

En el conjunto de prueba, el modelo final obtuvo:

Accuracy: 0.87 | F1 macro: 0.86

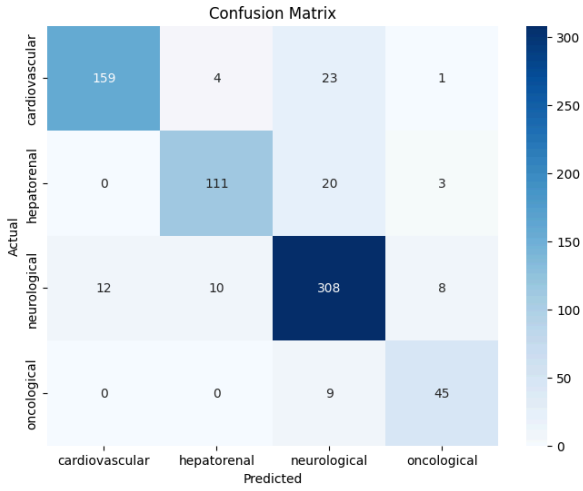
Por clase:

cardiovascular → F1 = 0.89

hepatorenal → F1 = 0.86

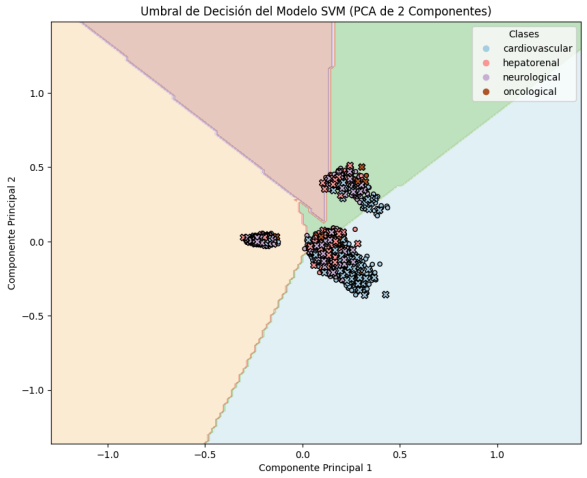
neurological → F1 = 0.88

oncological → F1 = 0.81



Se observa un mejor desempeño en categorías con más datos (neurological, cardiovascular) y un menor rendimiento en oncological, donde la cantidad de ejemplos es limitada.

Umbral de decisión



Otra manera de visualizar el rendimiento y cómo se está comportando el modelo es mediante la visualización del umbral de decisión. El gráfico representa los umbrales de decisión del modelo SVM proyectados en un espacio bidimensional tras aplicar PCA (2 componentes principales). Cada punto corresponde a un artículo médico, coloreado según su clase real, mientras que las regiones de fondo muestran cómo el modelo divide el espacio para clasificar nuevas instancias. Se observa que el modelo logra separar de manera clara grupos como oncological y cardiovascular, los cuales aparecen bien delimitados en sus regiones. Sin embargo, existe cierta superposición entre las clases neurological y hepatorenal.

Integración con V0

Prompt utilizado: “Ayudame a crear esta página, con colores oscuros: Quiero una aplicación web con la siguiente funcionalidad:

1. Una interfaz con dos secciones:

- *Subida de CSV: permitir cargar un archivo con estructura `title;abstract;group`.*

- Mostrar el contenido del archivo en una tabla.

- Enviar cada fila al modelo HuggingFace: https://huggingface.co/kgemera/Biomedic_Text_Classifier/blob/main/classification_model.pkl

- Añadir una columna "Predicted Group" con la clase asignada por el modelo.

*- **Predicción manual**: un formulario con dos campos de texto ("Title", "Abstract") y un botón "Clasificar".*

- Al hacer clic, enviar estos textos al modelo HuggingFace.

- Mostrar la clase predicha en un recuadro.

2. Los resultados deben visualizarse en una tabla clara con las columnas:

- Title

- Abstract

- Actual Group (si existe en el CSV)

- Predicted Group

3. Estilo: interfaz moderna, minimalista, con un área central para los resultados.

4. Backend:

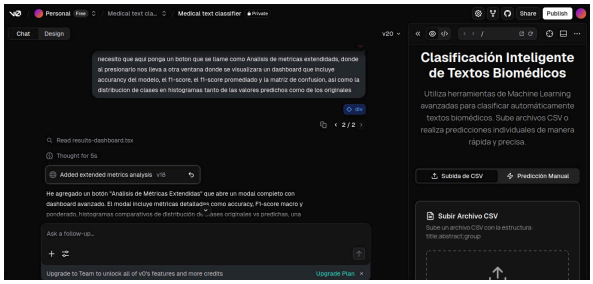
- Usar Python con FastAPI o Next.js API Routes.

- Cargar el modelo desde HuggingFace (`classification_model.pkl`).

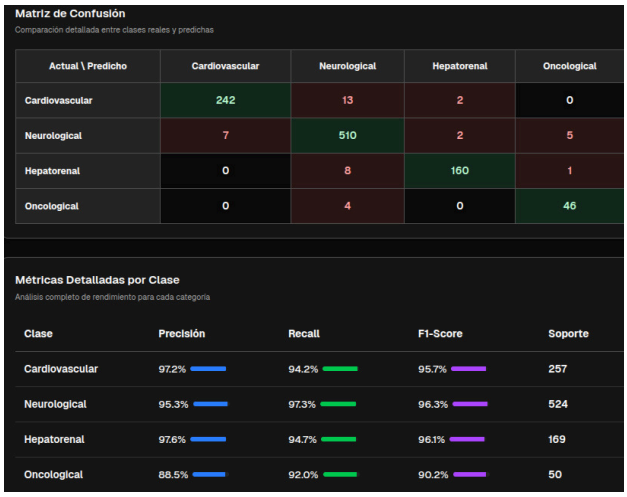
- Endpoint `/predict` que reciba JSON con {title, abstract} y devuelva {predicted_group}.

- Endpoint `/batch_predict` que reciba lista de textos (del CSV) y devuelva lista de predicciones.

5. Integrar la UI con estos endpoints para mostrar predicciones al usuario.”



Imágenes de la interfaz



Matriz de Confusión

Comparación detallada entre clases reales y predichas

Actual \ Predicho	Cardiovascular	Neurological	Hepatorenal	Oncological
Cardiovascular	242	13	2	0
Neurological	7	510	2	5
Hepatorenal	0	8	160	1
Oncological	0	4	0	46

Métricas Detalladas por Clase

Análisis completo de rendimiento para cada categoría

Clase	Precisión	Recall	F1-Score	Soporte
Cardiovascular	97.2%	94.2%	95.7%	257
Neurological	95.3%	97.3%	96.3%	524
Hepatorenal	97.6%	94.7%	96.1%	169
Oncological	88.5%	92.0%	90.2%	50