# Clasificación y Análisis de Artículos Biomédicos

Informe Técnico Final

Yulián Bedoya
ybedoyab@unal.edu.co

26 de agosto de 2025

## Índice

# 1. Resumen

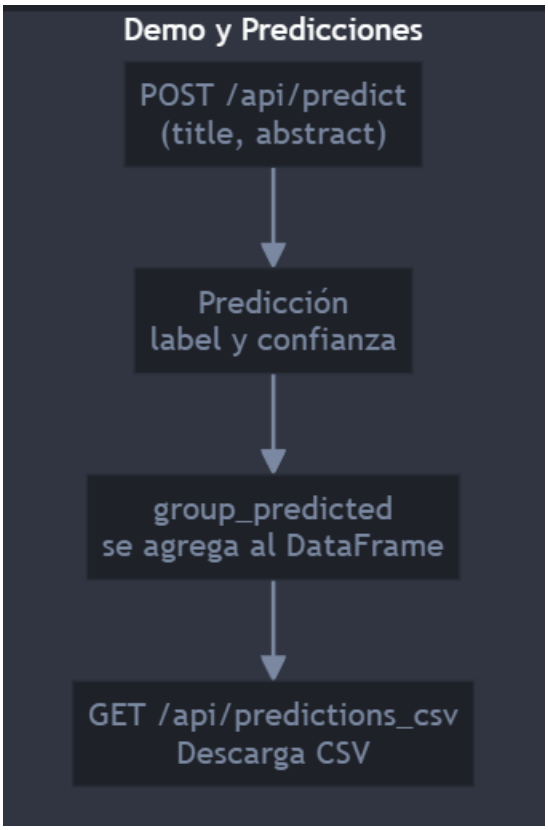
Este informe presenta la solución desarrollada para la clasificación y análisis de artículos médicos. Incluye la arquitectura general del sistema (backend en Python y frontend en Next.js), el flujo de datos, el pipeline de preprocesamiento, el modelo utilizado (Regresión Logística con TF–IDF), y la interfaz de usuario con visualización de métricas, análisis y demo.

# 2. Diseño de la Solución

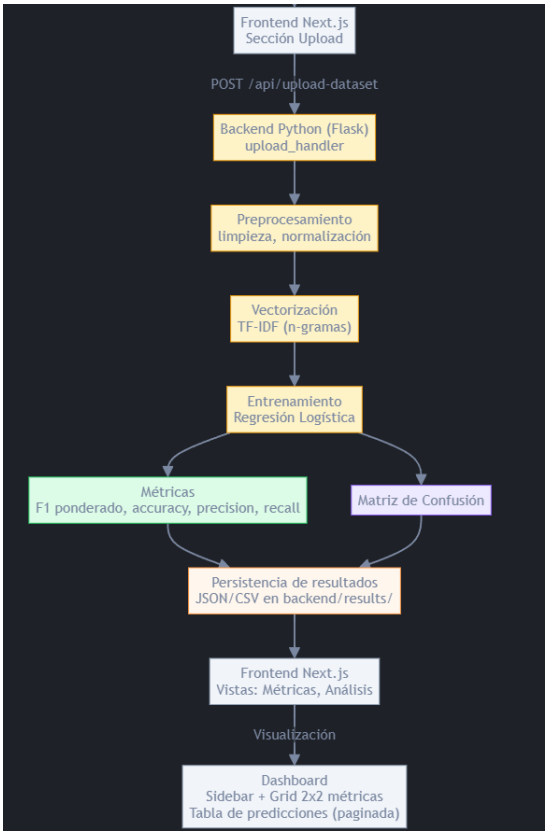
## 2.1. Arquitectura General

Describir brevemente el diagrama de arquitectura (incluir imagen una vez generada):

Diagrama de Arquitectura




(a) Arquitectura general




(b) Flujo de datos

## 2.2. Flujo de Datos

Desde la carga del CSV hasta la generación de métricas y predicciones con columna `group_predicted`.

# 3. Backend (Python)

## 3.1. Tecnologías y Dependencias

Python 3.10+, Flask/FastAPI (según implementación), NumPy, pandas, scikit-learn, etc. (ver `backend/requirements.txt`).

## 3.2. Endpoints Principales

- `/api/uploaddataset`: carga de CSV (`title, abstract, group`).
- `/api/starttraining`: entrena el modelo con TF–IDF + Regresión Logística.
- `/api/modelmetrics`: expone métricas (F1 ponderado, accuracy, precision, recall).
- `/api/confusionmatrix`: devuelve matriz de confusión.
- `/api/predictions`: devuelve predicciones y genera `group_predicted`.
- `/api/predictionscsv`: descarga CSV con `group_predicted`.

## 3.3. Pipeline de Entrenamiento

1. Limpieza y normalización de texto (minúsculas, signos, espacios).
2. Vectorización TF–IDF con n-gramas.
3. Entrenamiento de Regresión Logística (clase uno contra resto).
4. Validación (hold-out o CV). Almacenar métricas y matriz de confusión.

# 4. Frontend (Next.js)

## 4.1. Vistas y Navegación

- Resumen y carga de dataset.
- Entrenamiento con barra de progreso.
- Métricas (grid 2x2, tarjeta destacada, indicador de rendimiento).
- Análisis (estadísticas de dataset, distribuciones, texto).
- Demo (clasificación manual).
- Predicciones (tabla paginada, filtros, descarga de CSV con `group_predicted`).

# 5. Evaluación del Desempeño

## 5.1. Métrica Principal

F1 ponderado (*weighted F1*). También se reportan accuracy, precision y recall.

Matriz de Confusión

| | Pred: Cardiovascular | Pred: Hepatorenal | Pred: Neurological | Pred: Oncological | Total |
|---|---|---|---|---|---|
| Real: Cardiovascular | 152 | 6 | 28 | 1 | 187 |
| Real: Hepatorenal | 0 | 100 | 33 | 1 | 134 |
| Real: Neurological | 11 | 3 | 320 | 4 | 338 |
| Real: Oncological | 0 | 1 | 25 | 28 | 54 |
| Total | 163 | 110 | 406 | 34 | 713 |

Figura 2: Enter Caption

### 5.2. Matriz de Confusión

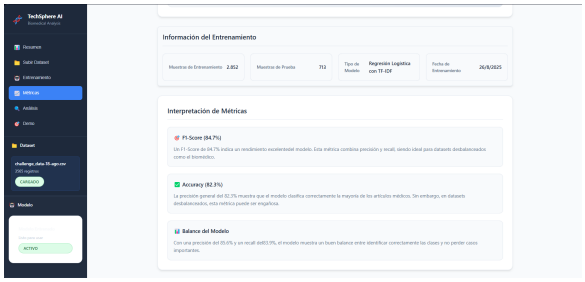
### 5.3. Procedimiento de Evaluación

1. Cargar CSV con `title, abstract, group`.
2. Entrenar el modelo.
3. Realizar predicción para obtener `group_predicted`.
4. Calcular métricas y matriz de confusión.
5. Descargar CSV con predicciones.

## 6. Resultados y Evidencias

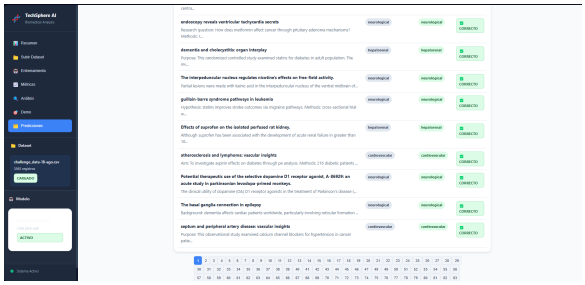
### 6.1. Métricas del Modelo


(a) Resumen de métricas




(b) Indicador de rendimiento y detalles

### 6.2. Predicciones del Modelo



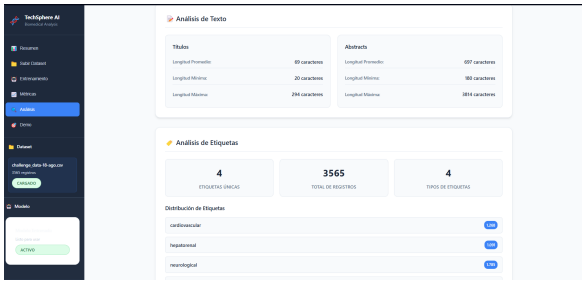
(a) Vista de predicciones en el frontend

```
1  title,abstract,group,group_predicted
2  "Adrenoleukodystrophy: survey of 303 cases: bi
3  endoscopy reveals ventricular tachycardia secr
4  dementia and cholecystitis: organ interplay,"P
5  The interpeduncular nucleus regulates nicotine
6  guillain-barre syndrome pathways in leukemia,"
7  Effects of suprofen on the isolated perfused r
8  atherosclerosis and lymphoma: vascular insight
9  "Potential therapeutic use of the selective do
```

(b) CSV descargado con columna `group_predicted`

### 6.3. Análisis del Dataset


(a) Resumen del dataset




(b) Distribuciones y estadísticas de texto

## 7. Reflexiones y Trabajo Futuro

Limitaciones, posibles mejoras (mejoras del preprocesamiento, modelos más robustos, interpretabilidad, despliegue, MLOps).

## Repositorio y Reproducibilidad

- URL del repositorio: github.com/tu-repo-publico
- Instrucciones de ejecución en README (backend y frontend).
- Versionado y evidencias de pruebas.