Proyecto 1 – Analítica de Texto (Etapa 1)

Inteligencia de Negocios

Universidad de los Andes

Juan Andrés Vargas Bolaños
ja.vargasb1@uniandes.edu.co

Juan Felipe Ledesma Velázquez
j.ledesma@uniandes.edu.co

13 de septiembre de 2025 — Bogotá

Contenido

Resumen Ejecutivo

En este proyecto clasificamos textos en español de la ciudadanía dentro de tres Objetivos de Desarrollo Sostenible: **ODS1 (Fin de la pobreza)**, **ODS3 (Salud y bienestar)** y **ODS4 (Educación de calidad)**. La idea es sencilla: a partir de lo que escribe la gente, decidir a qué ODS pertenece el mensaje para que el equipo adecuado lo atienda y lo analice más rápido.

Trabajamos con un conjunto de entrenamiento de **2.424 textos** (columnas textos y labels) y un conjunto de prueba de **152 textos** (Textos_espanol) que la entrega pide etiquetar. A labels le hicimos un mapeo directo para facilitar la lectura: **1→ODS1**, **3→ODS3**, **4→ODS4**.

Usamos una representación clásica y efectiva para texto: **TF-IDF** con **n-grams (1–2)** y normalización de acentos. Probamos tres modelos supervisados: **Naive Bayes**, **Regresión Logística** y **Linear SVM**. Evaluamos con **validación estratificada 5-fold** y tomamos como métrica principal el **macro-F1** (así cada clase pesa lo mismo, incluso si están algo desbalanceadas). El resultado fue claro: **Linear SVM** obtuvo **macro-F1 = 0.9724**, superando a **Regresión Logística (0.9494)** y **Naive Bayes (0.8034)**.

Como entregable práctico, generamos el archivo **data/test_etiquetado.xlsx** agregando la columna **prediccion_modelo** con la etiqueta de ODS para cada texto del test. Además, dejamos evidencia del rendimiento (reporte de clasificación y matriz de confusión “out-of-fold”) y un listado de **términos más influyentes por clase**, que ayuda a explicar por qué el modelo toma cada decisión.

Sección 1 — Canvas de ML (20%)

Archivo adjunto en el repositorio docs/***CanvasML v1.2.1.docx***

Sección 2 — Datos (20%)

El dataset de entrenamiento viene en ***Datos_proyecto.xlsx*** con 2.424 filas y dos columnas: textos (el contenido del mensaje) y labels (la clase numérica 1/3/4). El dataset de prueba para entregar está en ***Datos de prueba_proyecto.xlsx*** con 152 filas en la columna Textos_espanol. Para hacer el informe más legible, mapeamos labels a nombres de clase: **1→ODS1**, **3→ODS3**, **4→ODS4**.

Antes de modelar hicimos una preparación mínima pero efectiva para español: pasamos a minúsculas, **normalizamos acentos** y representamos el texto con **TF-IDF** considerando **unigramas y bigramas (1–2)**. Decidimos **no aplicar stopwords genéricas** para español porque, en este dominio, palabras que a veces se catalogan como "comunes" pueden cargar señal semántica útil (por ejemplo, en frases como "acceso a **la** educación"). Con TF-IDF y n-grams la mayoría del ruido se controla bien.

La **distribución por clase** en el train es balanceada de forma razonable, aunque con un leve sesgo hacia ODS4:

- **ODS1:** 505 ejemplos (20.8%)
- **ODS3:** 894 ejemplos (36.9%)
- **ODS4:** 1025 ejemplos (42.3%)

No encontramos valores nulos en las columnas clave y las longitudes de texto son variables, como es esperable en este tipo de recolecciones.

Sección 3 — Modelado y evaluación (20%)

Nuestro pipeline es directo: **TF-IDF (1–2, con normalización de acentos)** seguido de un clasificador lineal. Comparamos tres algoritmos clásicos para texto:

- **Naive Bayes (MultinomialNB)**: rápido y sencillo, suele ser un buen punto de partida.
- **Regresión Logística**: probabilístico, con buen desempeño en muchos escenarios de texto.
- **Linear SVM**: muy sólido para márgenes entre clases y robusto con TF-IDF.

Para evaluar de forma justa usamos **validación cruzada estratificada de 5 particiones**. La métrica principal fue **macro-F1**, porque nos importa un rendimiento equilibrado por clase y no solo la exactitud global.

Resultados (macro-F1 en CV 5-fold):

- **Linear SVM → 0.9724**
- Regresión Logística → 0.9494
- Naive Bayes → 0.8034

Con esto, **seleccionamos Linear SVM** como modelo final para la entrega. Además de la cifra global, generamos un **reporte de clasificación "out-of-fold"** (precisión, recall y F1 por clase) y una **matriz de confusión** también "out-of-fold". Estas evidencias son útiles para ver dónde se confunde el modelo y si alguna clase queda rezagada.

	precision	recall	f1-score	support
ODS1	0.9755	0.9465	0.9608	505
ODS3	0.9754	0.9765	0.9760	894
ODS4	0.9740	0.9873	0.9806	1025
accuracy			0.9748	2424
macro avg	0.9750	0.9701	0.9725	2424
weighted avg	0.9748	0.9748	0.9748	2424

Ilustración 1. evaluation/cv_classification_report.txt

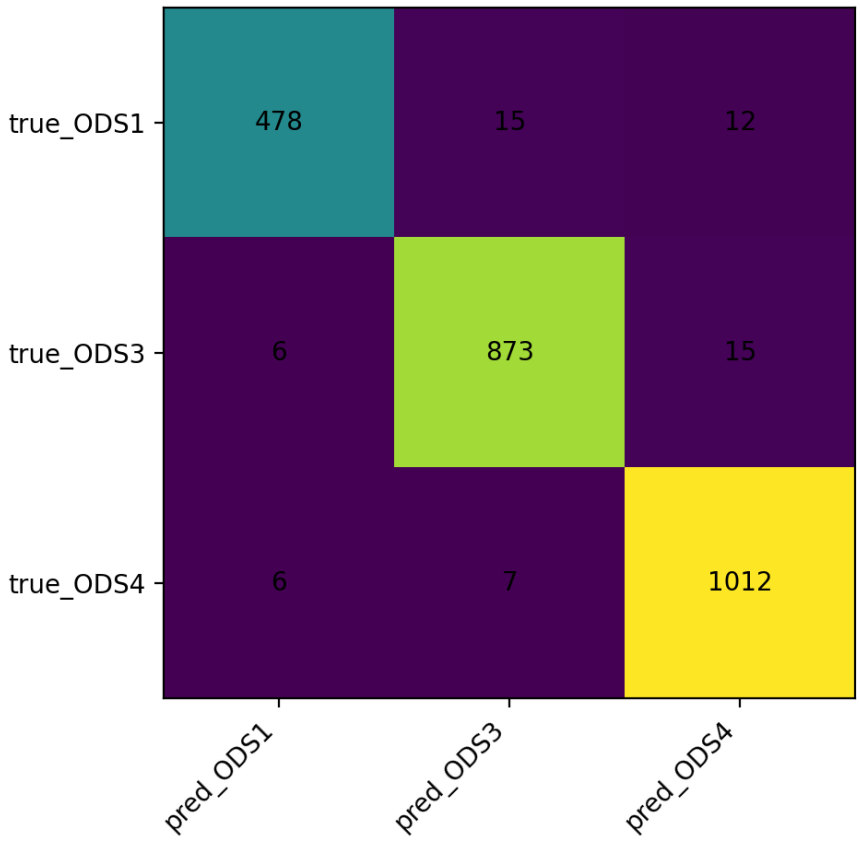

Ilustración 2. evaluation/cv_confusion_matrix.png

Sección 4 — Resultados y uso (20%)

El objetivo práctico de la Etapa 1 es etiquetar el archivo de prueba del curso. Para eso cargamos el **modelo ganador (Linear SVM)** y generamos **data/test_etiquetado.xlsx**, que es exactamente el archivo original, pero con una columna adicional llamada **prediccion_modelo**. Respetamos el orden y la cantidad de filas para que la comparación externa sea directa.

Desde el punto de vista de negocio, esto permite **enrutar rápidamente** cada texto al ODS correcto. Con un macro-F1 tan alto, el modelo reduce el tiempo de clasificación manual y ayuda a construir tableros de seguimiento por ODS (por ejemplo: volumen semanal, picos de ciertos temas, etc.). Los **casos frontera** —textos muy cortos, ambiguos o con lenguaje muy local— se pueden marcar para **revisión experta**. Así combinamos la velocidad del modelo con la calidad del criterio humano.

Para dar **explicabilidad**, extraemos los **términos con mayor peso** por clase (coeficientes de la SVM). Esto muestra qué palabras y bi-gramas empujan a cada ODS. En el informe, conviene listar entre 10 y 15 términos por clase y comentar por qué hacen sentido.

1	Textos_espanol	prediccion_modelo
2	El rector, que es el representante local del Ministerio de Educación, tiene la responsabilidad principal de procesar la eva	ODS4
3	Tenga en cuenta que todos los programas antipobreza tienen condiciones, incluso aquellos caracterizados como 'incor	ODS4
4	Debido a que son en gran medida invisibles, estas formas de trabajo infantil son las más difíciles de abordar. Las trabaja	ODS1
5	Los recursos aún son limitados en este sector. Los servicios privados con fines de lucro y comunitarios (religiosos y sec	ODS4
6	Durante el período 1985-2008, la educación primaria, secundaria y terciaria experimentó un aumento sin precedentes (	ODS4
7	En la región de Asia y el Pacífico, casi el 87% de los niños de 1 año o menos están inmunizados contra el sarampión. Sin	ODS3
8	Esta combinación representa una oportunidad para que tanto los alumnos indígenas como los no indígenas se familiari	ODS4
9	Además, muchos llevan a cabo prácticas de seguimiento de la calidad del servicio con el objetivo de informar al públicc	ODS4
10	El alcance de esta visión holística se basa en los avances logrados durante el período de los ODM, durante el cual las ta	ODS4
11	Véase C. Correa, "Protecting Test Data for Pharmaceutical and Agrochemical Products under Free Trade Agreements", (	ODS3
12	En 2012, alrededor del 10 % de la variación en la puntuación de matemáticas en la prueba PISA se explicó por el entorn	ODS4
13	Esto reduce el desempeño promedio del país (ver Figura 3.3). Además, la proporción muy pequeña de estudiantes (0,9 !	ODS4
14	En general, la literatura sugiere que una reducción significativa de las tasas de pobreza es consecuencia del crecimientc	ODS1
15	Con el tiempo, considere alejarse del financiamiento compartido y aumentar los subsidios de cupones lo suficiente par	ODS4
16	La concesión de un título es, por lo tanto, un reconocimiento oficial de los logros de un estudiante por parte de un orga	ODS4
17	En muchos aspectos, la "calidad" está en el corazón del sistema de salud de Gales, este capítulo describe la ya rica arqu	ODS3
18	Con el objetivo de convertirse en una de las instituciones de educación superior de mejor rendimiento en todo el mund	ODS4
19	Recientemente, el Instituto Australiano de Docencia y Liderazgo Escolar (AITSL, por sus siglas en inglés) desarrolló un cc	ODS4
20	La mayoría de los indicadores incluidos en la medida de pobreza propuesta ya se consideran "indicadores básicos" en l	ODS1

Ilustración 3. data/test_etiquetado.xlsx

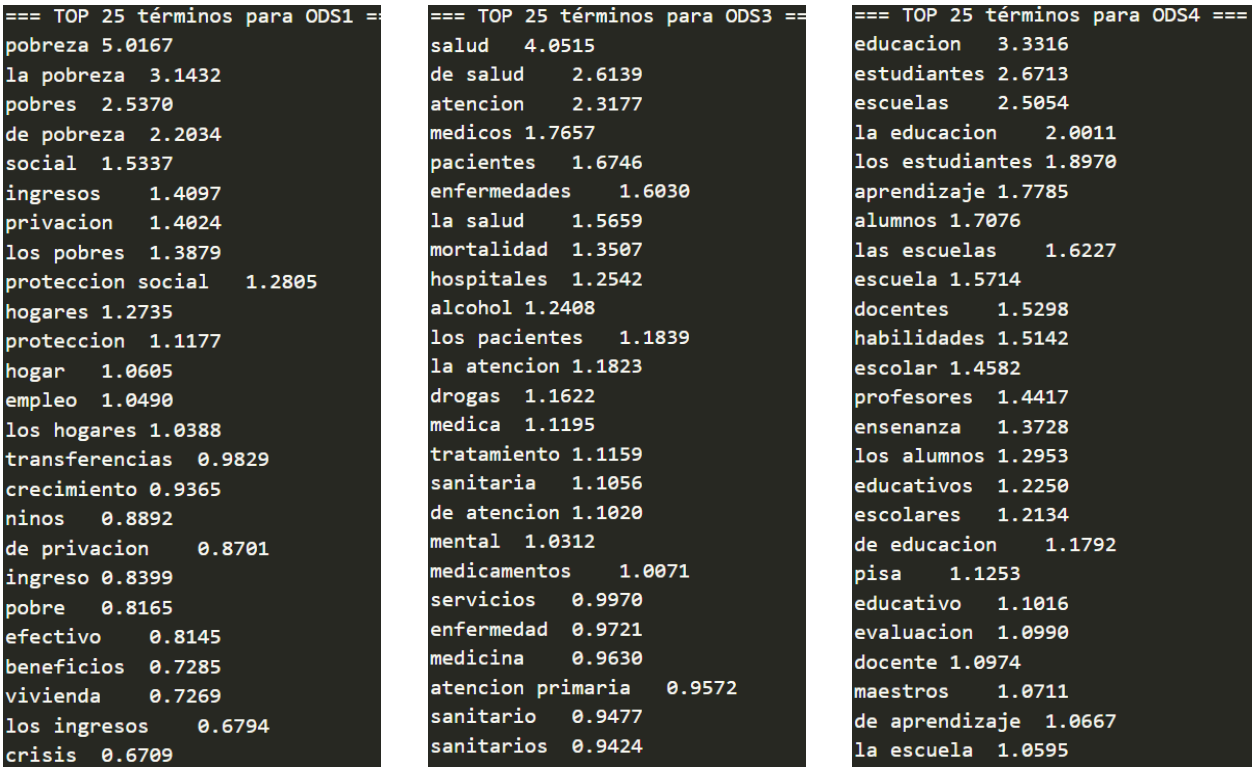

Ilustración 4. evaluation/top_words.txt

ODS3 (Salud y bienestar): Aparecen términos sobre atención médica, salud pública, prevención, etc., lo que concuerda con la temática.

ODS4 (Educación): Surgen expresiones sobre acceso, formación, escuela, currículo, etc.

ODS1 (Pobreza): Destacan vocablos de ayudas, subsidios, ingresos, vulnerabilidad, empleo, entre otros.

Sección 5 — Trabajo en equipo (8%)

Integrantes

- **Juan Felipe Ledesma Velásquez (JFLV) — Data & NB/LR**
- **Juan Andrés Vargas Bolaños (JAVB) — Modelado (SVM) & Evaluación**

5.0 Asignación de algoritmos (obligatoria — 3 modelos implementados)

Algoritmo	Responsable Principal	Evidencias / artefactos	Resultado (macro-F1 CV)
Naive Bayes (MultinomialNB)	JFLV	Implementación en src/pipelines.py ("nb"), corridas en src/train_from_excel.py (logs con "[CV] nb"), registro en evaluation/model_selection.json	0.8034
Regresión Logística	JAVB con revisión de JFLV	Implementación en src/pipelines.py ("lr"), corridas en src/train_from_excel.py (logs "[CV] lr"), registro en evaluation/model_selection.json	0.9494
Linear SVM	JAVB	Implementación en src/pipelines.py ("svm"), corridas en src/train_from_excel.py (logs "[CV] svm"), selección final; reportes en evaluation/	0.9724 (ganador)

5.1 Roles y responsabilidades

JFLV (Data & NB/LR). Preparó los datos (***Datos_proyecto.xlsx / Datos de prueba_proyecto.xlsx***) y el *pipeline* de texto (minúsculas, normalización de acentos y TF-IDF (1–2)) en ***src/pipelines.py***. Implementó y ejecutó Naive Bayes y Regresión Logística, interpretó sus resultados y redactó la Sección 2 (Datos) del informe. Ordenó repo/wiki y publicó el ***data/test_etiquetado.xlsx***.

JAVB (Modelado SVM & Evaluación). Implementó y ejecutó Linear SVM, seleccionó el modelo final por macro-F1, generó la matriz de confusión out-of-fold (***src/make_cv_reports.py*** + ***src/plot_confusion_matrix.py***) y la explicabilidad por ODS (***src/explain_top_words.py***). Ejecutó ***predict_excel.py*** y verificó la columna prediccion_modelo.

5.2 Distribución de tareas (100-split) y horas

Integrante	Aporte	Horas
JFLV	**50%** — datos, **NB + LR**, wiki/repo	
JAVB	**50%** — **SVM**, evaluación, explicabilidad, etiquetado	

5.3 Uso de IA generativa (declaración)

Usamos IA (ChatGPT) como apoyo para: plantillas de código (p. ej., train_from_excel.py, make_cv_reports.py, predict_excel.py), redacción inicial del Canvas y del informe, y buenas

prácticas de .gitignore/organización del repo. Todas las decisiones técnicas (métrica, selección de modelo, interpretación de resultados y de términos TOP por ODS) se validaron manualmente y se ejecutaron localmente con nuestros datos.

5.4 Retos y lecciones

- **Compatibilidad de librerías.** TfidfVectorizer(stop_words="spanish") no es válido en scikit-learn; lo resolvimos con stop_words=None manteniendo **TF-IDF (1–2)**.
- **Reproducibilidad.** Añadimos openpyxl/matplotlib a requirements.txt; dejamos rutas y scripts claros para repetir experimentos.
- **Higiene del repo.** No versionamos datos crudos ni el modelo binario; subimos **código esencial**, **evidencias** y **test_etiquetado.xlsx**.
- **Explicabilidad.** Los coeficientes de SVM muestran léxico coherente con cada ODS (educación: *estudiantes, escuelas, docentes*; salud: vocabulario médico; pobreza: términos socioeconómicos).

5.5 Cierre del equipo

Con los tres algoritmos implementados (NB y LR por JFLV, SVM por JAVB), alcanzamos macro-F1 = 0.9724 con SVM, generamos el Excel etiquetado y dejamos evidencia lista para el informe, la wiki y el video.