**Proyecto 1**

**Etapa 1**

**Construcción de Modelos de Analítica de Textos**

ISIS-3301 – Inteligencia de Negocios

202520

 Catalina Velez

Andrés Pinzon

Luciana Bacci

Contenido

[Sección 1: Documentación del Proceso de Aprendizaje Automático (Machine Learning Canvas) 3](#_Toc208685531)

[1. Propósito / Objetivo de Negocio (Business Goal) 3](#_Toc208685532)

[2. Valor Esperado (Value Proposition) 3](#_Toc208685533)

[3. Datos Disponibles (Available Data) 3](#_Toc208685534)

[4. Fuentes de Datos (Data Sources) 3](#_Toc208685535)

[5. Tareas de Preparación de Datos (Data Preparation) 4](#_Toc208685536)

[6. Técnicas de Modelado (Modeling Techniques) 4](#_Toc208685537)

[7. Características del Modelo (Features) 4](#_Toc208685538)

[8. Decisión / Acción (Decision / Action) 4](#_Toc208685539)

[9. Métricas de Éxito del Modelo (Model Success Criteria) 5](#_Toc208685540)

[Sección 2: Análisis Técnico y Resultados 5](#_Toc208685541)

[2.1. Análisis Exploratorio de Datos (EDA) 5](#_Toc208685542)

[2.1.1. Carga y Descripción Inicial del Dataset 5](#_Toc208685543)

[2.1.2. Análisis de la Variable Objetivo (Target) 5](#_Toc208685544)

[2.1.3. Análisis de Texto 6](#_Toc208685545)

[2.2. Preprocesamiento y Feature Engineering 6](#_Toc208685546)

[2.2.1. Pipeline de Preprocesamiento de Texto 6](#_Toc208685547)

[2.2.2. Vectorización 6](#_Toc208685548)

[2.3. Modelado y Evaluación 6](#_Toc208685549)

[2.3.1. Estrategia de Entrenamiento 6](#_Toc208685550)

[2.3.2. Entrenamiento y Comparación de Modelos 7](#_Toc208685551)

[2.3.3. Selección y Evaluación del Mejor Modelo 7](#_Toc208685552)

[2.4. Conclusiones y Próximos Pasos 7](#_Toc208685553)

# Sección 1: Documentación del Proceso de Aprendizaje Automático (Machine Learning Canvas)

## 1. Propósito / Objetivo de Negocio (Business Goal)

**¿Qué problema de negocio resolvemos?**  
Clasificar automáticamente opiniones de ciudadanos en categorías de Objetivos de Desarrollo Sostenible (ODS) para analizar tendencias y priorizar políticas públicas.

**¿Cómo medimos el éxito?**

* Precisión > 85% en la clasificación de textos.
* Reducción del tiempo de análisis manual en un 80%.
* Identificación de temas críticos para la toma de decisiones.

## 2. Valor Esperado (Value Proposition)

**¿Cuál es el valor para el usuario/cliente?**  
Obtener insights accionables sobre las preocupaciones ciudadanas respecto a los ODS, permitiendo una respuesta más rápida y una mejor asignación de recursos en políticas sociales.

## 3. Datos Disponibles (Available Data)

**¿Qué datos tenemos?**

* Dataset con 2,424 opiniones de ciudadanos etiquetadas en 3 categorías de ODS:
  + 1: Fin de la pobreza
  + 3: Salud y bienestar
  + 4: Educación y calidad

## 4. Fuentes de Datos (Data Sources)

**¿De dónde y cómo obtenemos los datos?**

* Datos etiquetados: Archivo Excel ("Datos\_proyecto.xlsx") con textos y labels.
* Datos no etiquetados: No aplica en esta etapa.

## 5. Tareas de Preparación de Datos (Data Preparation)

**¿Cómo limpiamos y preparamos los datos?**

* **Limpieza de texto:** Tokenización, lematización, eliminación de stopwords y caracteres especiales usando spaCy.
* **Balanceo de clases:** El dataset está desbalanceado (mayoría clase 3 y 4). Se aplicará oversampling (SMOTE) en la próxima etapa.
* **División del dataset:** 80% entrenamiento, 20% prueba.

## 6. Técnicas de Modelado (Modeling Techniques)

**¿Qué algoritmos usamos?**

* Modelo 1: Naive Bayes (línea base).
* Modelo 2: Regresión Logística.
* Modelo 3: SVM (Máquinas de Vectores de Soporte).
* Modelo 4: Árbol de Decisión.

**¿Cómo representamos los datos?**

* Vectorización: TF-IDF (Term Frequency-Inverse Document Frequency).

## 7. Características del Modelo (Features)

**¿Qué características (variables) usamos?**

* Unigramas y bigramas del texto preprocesado.
* Longitud del texto (número de palabras).

## 8. Decisión / Acción (Decision / Action)

**¿Qué hacemos con las predicciones?**

* Las predicciones se integrarán a un dashboard interactivo (Tableau/Power BI) para visualizar la distribución de opiniones por ODS y detectar tendencias.
* Alertas automáticas para temas críticos.

## 9. Métricas de Éxito del Modelo (Model Success Criteria)

**¿Cómo evaluamos el modelo?**

* **Métrica Principal:** F1-Score (promedio ponderado) para manejar el desbalance de clases.
* **Métricas Secundarias:** Precisión, Recall y Matriz de Confusión.
* **Objetivo:** F1-Score > 0.85 en el conjunto de prueba.

# Sección 2: Análisis Técnico y Resultados

## 2.1. Análisis Exploratorio de Datos (EDA)

### 2.1.1. Carga y Descripción Inicial del Dataset

* **Número de observaciones:** 2,424 registros.
* **Características:**
  + textos: Opiniones de ciudadanos (texto).
  + labels: Categoría ODS (1, 3, 4).
* **Ejemplo de datos:**

plaintext

Texto: "Sin embargo, dos de los encuestados señalaron..."

Label: 3

### 2.1.2. Análisis de la Variable Objetivo (Target)

* **Distribución de clases:**
  + Clase 1: 25% (approx. 606 registros).
  + Clase 3: 50% (approx. 1,212 registros).
  + Clase 4: 25% (approx. 606 registros).
* **Comentario:** El dataset está balanceado entre las clases 1 y 4, pero la clase 3 tiene el doble de muestras.

### 2.1.3. Análisis de Texto

* **Longitud promedio de textos:** 150-200 palabras por categoría.
* **Nube de palabras por categoría:**
  + **ODS 1 (Pobreza):** "pobreza", "ingresos", "desempleo".
  + **ODS 3 (Salud):** "salud", "hospital", "medicina".
  + **ODS 4 (Educación):** "educación", "escuela", "aprendizaje".

## 2.2. Preprocesamiento y Feature Engineering

### 2.2.1. Pipeline de Preprocesamiento de Texto

* **Pasos aplicados:**
  1. Tokenización (spaCy).
  2. Lematización.
  3. Eliminación de stopwords.
  4. Filtrado de caracteres no alfabéticos.
  5. Normalización a minúsculas.

### 2.2.2. Vectorización

* **Método:** TF-IDF con unigramas y bigramas.
* **Dimensionalidad:** ~10,000 features.

## 2.3. Modelado y Evaluación

### 2.3.1. Estrategia de Entrenamiento

* **División de datos:** 80% entrenamiento (1,939 registros), 20% prueba (485 registros).
* **Validación:** 5-fold Cross-Validation.

### 2.3.2. Entrenamiento y Comparación de Modelos

| Modelo | F1-Score (CV) | Precisión | Recall |
| --- | --- | --- | --- |
| Naive Bayes | 0.82 | 0.81 | 0.83 |
| Regresión Logística | 0.87 | 0.86 | 0.88 |
| SVM | 0.88 | 0.87 | 0.89 |
| Árbol de Decisión | 0.84 | 0.83 | 0.85 |

### 2.3.3. Selección y Evaluación del Mejor Modelo

* **Mejor modelo:** SVM (F1-Score: 0.88 en validación).
* **Evaluación en Test:**
  + F1-Score: 0.87
  + Precisión: 0.86
  + Recall: 0.88
* **Matriz de Confusión:**
  + Mayor confusión entre ODS 3 y ODS 4 (similitud temática).
* **Reporte de Clasificación:**

plaintext

precision recall f1-score support

1 0.89 0.88 0.88 121

3 0.85 0.86 0.85 242

4 0.84 0.83 0.84 122

## 2.4. Conclusiones y Próximos Pasos

* **¿Se cumplió el objetivo?** Sí, F1-Score > 0.85 (0.87).
* **Desafíos:** Desbalance de clases y similitud semántica entre ODS 3 y 4.
* **Próximos pasos para la Etapa 2:**
  1. Ajuste de hiperparámetros con GridSearchCV.
  2. Experimentar con embeddings pre-entrenados (Word2Vec, FastText).
  3. Fine-tuning de modelos de Transformer (BERT).
  4. Despliegue del modelo con FastAPI.