










<p><b>TAREA DE APRENDIZAJE</b></p> <p></p> <p><b>Tipo:</b> Clasificación supervisada multiclase.</p> <p><b>Qué se predice:</b> la etiqueta ODS asociada a un texto en español que expresa una necesidad/observación ciudadana.</p> <p><b>Clases/Resultados posibles:</b> ODS1 (Fin de la pobreza), ODS3 (Salud y bienestar), ODS4 (Educación de calidad).</p> <p><b>Cuándo se observa el resultado:</b> inmediatamente al recibir el texto (inferencia en milisegundos); el modelo se entrena offline y se usa en batch para la entrega de la Etapa 1.</p> <p><b>Datos base:</b> 2.424 registros de</p>	<p> <b>DECISIONES</b></p> <p><b>Objetivo operativo:</b> clasificar y <b>enrutar</b> cada texto al equipo responsable del ODS correspondiente (pobreza, salud, educación) para priorización y análisis.</p> <p><b>Reglas de acción sugeridas:</b></p> <ol style="list-style-type: none"> <li>1. <b>Asignar ODS</b> = predicción del modelo (SVM).</li> <li>2. <b>Revisión manual</b> cuando el texto sea muy corto (&lt;10 caracteres) o contenga solo números/URLs; (si se dispone del margen de decisión del SVM, revisar los percentiles más bajos).</li> <li>3. <b>Tablero de priorización:</b> conteos por ODS, top términos por ODS, y lista de ejemplos “dudosos” para curaduría.</li> </ol> <p><b>Beneficio:</b> reduce tiempos de lectura manual y mejora la asignación de casos a programas/secretarías.</p>	<p> <b>PROPUESTA DE VALOR</b></p> <p><b>Beneficiario:</b> área de planeación/social de una entidad local (alcaldía/universidad/ONG) que analiza voz ciudadana.</p> <p><b>Problema resuelto:</b> clasificación consistente y rápida de opiniones a ODS 1/3/4 para <b>priorizar</b> políticas y dirigir recursos.</p> <p><b>Valor:</b> +productividad analítica, métricas comparables en el tiempo, detección de tendencias por ODS.</p> <p><b>Riesgos y mitigación:</b></p> <ul style="list-style-type: none"> <li>• <b>Riesgo de sesgo/errores:</b> términos locales o jerga pueden producir confusiones; mitigar con revisión de “casos frontera” y ejemplos representativos.</li> </ul>	<p> <b>RECOLECCIÓN DE DATOS – NO SE DEBE DILIGENCIAR</b></p> <p>¿Cómo se obtiene el conjunto inicial de entidades y resultados (por ejemplo, extractos de bases de datos, extracciones de API, etiquetado manual)?</p> <p>¿Qué estrategias se aplican para actualizar los datos continuamente, controlando los costos y manteniendo la vigencia?</p>	<p> <b>FUENTES DE DATOS</b></p> <p><b>Entrenamiento:</b> data/Datos_proyecto.xlsx (Sheet1) con columnas <b>textos</b> (cadena) y <b>labels</b> (1/3/4).</p> <p><b>Prueba (entrega):</b> data/Datos de prueba_proyecto.xlsx (Sheet1) con columna <b>Textos_espanol</b>; se devuelve test_etiquetado.xlsx añadiendo la columna <b>prediccion_modelo</b>.</p> <p><b>Derivadas (features):</b> matriz TF-IDF con n-grams (1–2), normalización de acentos; sin stopwords embebidas para no perder señales específicas del dominio.</p> <p><b>Uso:</b> los datos son suficientes para cumplir el objetivo de la Etapa 1 (construcción y evaluación del modelo y etiquetado del test).</p>
--	--	--	---	--

<p>entrenamiento (textos, labels), mapeando labels {1→ODS1, 3→ODS3, 4→ODS4}; 152 textos de prueba para etiquetar.</p> <p><b>Métrica principal:</b> F1 macro por desbalance de clases.</p> <p><b>Resultados</b> (CV 5-fold): Linear SVM (0.9724), Logistic Regression (0.9494), Naive Bayes (0.8034). Modelo seleccionado: Linear SVM (TF-IDF bigramas + acentos normalizados).</p>		<ul style="list-style-type: none"><li><b>Riesgo de uso indebido:</b> el modelo <b>no</b> reemplaza criterio experto; incluir nota visible “asistente de apoyo, requiere verificación”.</li></ul>		
<div><div>SIMULACIÓN DE IMPACTO</div></div> <p><b>Costos/beneficios (intuición):</b></p> <p>Acierto → texto bien enrutado al ODS → ahorro de tiempo y mejor focalización (beneficio alto).</p>	<div><div>APRENDIZAJE (USO DEL MODELO)</div></div> <p><b>Modalidad de uso (Etapa 1):</b> batch sobre el Excel de prueba → se genera test_etiquetado.xlsx.</p> <p><b>Frecuencia:</b> bajo demanda (cada corte de recolección).</p>		<div><div>CONSTRUCCIÓN DE MODELOS</div></div> <p><b>Modelos evaluados:</b> Naive Bayes, Regresión Logística y <b>Linear SVM</b>.</p> <p><b>Ganador:</b> SVM (F1 macro <b>0.9724</b>) con TF-IDF (1–2).</p> <p><b>Actualización:</b> cada nuevo lote etiquetado</p>	<div><div>INGENIERÍA DE CARACTERÍSTICAS</div></div> <p><b>Variables de entrada:</b> texto libre en español.</p> <p><b>Transformaciones:</b></p> <ul style="list-style-type: none"><li>Normalización: minúsculas y <b>remoción de acentos</b>.</li><li><b>TF-IDF con n-grams (1–2)</b> para capturar palabras y bi-gramas semánticos (p. ej., “salud pública”, “educación primaria”).</li><li>Opcional (no usado en baseline): lematización ligera y lista de stopwords en español si se valida que no degrade F1.</li></ul>

<p>Error → puede desviar la atención a un programa equivocado (costo moderado; mitigable con revisión de casos frontera).</p> <p><b>Criterios de éxito (para pasar a app en Etapa 2):</b> Macro-F1 <math>\geq</math> 0.90 sostenido y F1 por clase <math>\geq</math> 0.90.</p> <p>Estabilidad del rendimiento en muestras de semanas distintas.</p> <p><b>Equidad:</b> revisar diferencias por longitud del texto y por presencia de términos regionales; si se detecta sesgo sistemático, ampliar datos o ajustar representación.</p>	<p>Latencia de predicción: milisegundos por texto en CPU estándar (Python + scikit-learn).</p> <p><b>Etapas 2 (referencia del enunciado):</b> se prevé una API con /predict y, para rol experto, /retrain.</p>		<p>(o mensual) si varía la distribución de textos; en Etapa 2 se comparará contra el baseline y se actualizará solo si mejora macro-F1.</p> <p><b>Tiempo de construcción:</b> preparación + entrenamiento &lt; 2 min en CPU estándar para 2.4k textos; inferencia instantánea.</p> <p><b>Trazabilidad:</b> se guardan métricas CV, matriz de confusión y top-términos por clase para auditoría. (El Canvas hace parte de la Sección 1 del documento.</p>	<p><b>Importancia de términos:</b> para SVM, se reportan coeficientes por clase (top-términos) como explicación cualitativa de cada ODS.</p>
	<div data-bbox="416 1070 1046 1362"><p><b>MONITOREO NO SE DEBE DILIGENCIAR</b></p></div> <div data-bbox="1046 1070 1281 1362"><p>¿Qué métricas y KPI se utilizan para hacer un seguimiento del impacto de la solución de ML una vez desplegada, tanto para los usuarios finales como para la</p></div>			

empresa? ¿Con qué frecuencia  
deben revisarse?

## Adaptación de OWNML MACHINE LEARNING CANVAS



Version 1.2. Created by Louis Dorard, Ph.D. Licensed under a [Creative Commons Attribution-ShareAlike 4.0 International License](https://creativecommons.org/licenses/by-sa/4.0/).  
Please keep this mention and the link to [ownml.co](https://ownml.co) when sharing.

**OWNML.CO**