





<p>TAREA DE APRENDIZAJE</p> <p>El modelo se basa en aprendizaje no supervisado, específicamente en clustering de textos. En lugar de predecir una categoría previamente definida, el objetivo es agrupar automáticamente las opiniones ciudadanas según sus similitudes en el lenguaje utilizado.</p> <p>Los posibles resultados de la tarea de aprendizaje son clusters que representan temas o patrones de conversación comunes entre los ciudadanos, los cuales pueden asociarse posteriormente a los ODS (1, 3 y 4) mediante interpretación de los grupos formados.</p> <p>Los resultados se observan inmediatamente después de procesar el conjunto de datos, ya que el modelo entrega las agrupaciones en el mismo momento en que se ejecuta el algoritmo. Esto significa que el análisis puede actualizarse de forma rápida (en horas o menos) cada vez que se incorporen nuevas opiniones, permitiendo tener una visión dinámica de los temas emergentes en la población.</p>	<p> DECISIONES</p> <p>Los resultados del modelo se convierten en herramientas para la toma de decisiones en el ámbito de políticas públicas y de gestión social. La clasificación automática de las opiniones ciudadanas permite identificar patrones de necesidad o percepción de la población y agruparlos de acuerdo con los ODS. Esto hace posible que las entidades públicas y organizaciones sociales prioricen iniciativas y recursos según la categoría más crítica. Por ejemplo, si se detecta un volumen mayor de opiniones clasificadas en ODS 3 (Salud y bienestar), se pueden orientar proyectos y presupuestos hacia esa área. De esta manera, el modelo facilita la conversión de datos dispersos en información clara y accionable</p>	<p> PROPUESTA DE VALOR</p> <p>El beneficiario principal es el Fondo de Población de las Naciones Unidas (UNFPA), en el marco de la Agenda 2030 para el Desarrollo Sostenible, junto con las entidades encargadas de dar seguimiento a los ODS 1 (Fin de la pobreza), 3 (Salud y bienestar) y 4 (Educación de calidad).</p> <p>El modelo de clustering aporta valor porque automatiza el análisis de grandes volúmenes de opiniones ciudadanas, una tarea que de otra forma demandaría muchos recursos y expertos. Esto permite vincular de forma más rápida y objetiva la voz ciudadana con los ODS, facilitando decisiones de política pública más informadas.</p> <p>El principal riesgo es que los clusters generados no sean totalmente interpretables o se confundan entre categorías cercanas, lo que puede afectar la precisión al relacionarlos con los ODS. Sin embargo, su uso ofrece un beneficio significativo en eficiencia y en la capacidad de tomar decisiones basadas en evidencia real de la ciudadanía.</p>	<p> RECOLECCIÓN DE DATOS – NO SE DEBE DILIGENCIAR</p> <p>¿Cómo se obtiene el conjunto inicial de entidades y resultados (por ejemplo, extractos de bases de datos, extracciones de API, etiquetado manual)? ¿Qué estrategias se aplican para actualizar los datos continuamente, controlando los costos y manteniendo la vigencia?</p>	<p> FUENTES DE DATOS</p> <p>Las fuentes de datos corresponden a bases de datos de opiniones ciudadanas recolectadas y puestas a disposición en el proyecto. Los datos provienen de un corpus textual en español con miles de ejemplos etiquetados en los diferentes ODS de interés. Estas fuentes son externas, ya que no se generan dentro de la organización, pero resultan confiables y suficientes para entrenar y validar los modelos de clasificación. Los textos provienen de ejercicios de participación ciudadana y sirven de insumo para evaluar el estado de percepción de la población sobre pobreza, salud y educación.</p>
--	--	--	--	--



SIMULACIÓN DE IMPACTO

En términos de costo/beneficio, una decisión correcta implica ahorro de tiempo y recursos en el análisis de opiniones y una mejor orientación de las políticas hacia los ODS 1, 3 y 4, generando mayor impacto social. Una decisión incorrecta, en cambio, puede llevar a dirigir recursos hacia un ODS equivocado, reduciendo la efectividad de las acciones y aumentando los costos de corrección.

Los criterios de éxito del modelo para su despliegue incluyen: lograr métricas de desempeño altas y estables (ej. f1-score superior al 0.90), baja confusión entre clusters cercanos y capacidad de generalización en nuevos datos.

Respecto a restricciones de equidad, es fundamental garantizar que el modelo no privilegie un ODS sobre otro por la cantidad de datos disponibles. Los clusters deben representar de manera



APRENDIZAJE (USO DEL MODELO)

El modelo se usaría en modo batch (por lotes), ya que se procesan bloques de opiniones ciudadanas recolectadas en periodos determinados y no en tiempo real. La frecuencia de uso depende de la disponibilidad de nuevas encuestas u opiniones, pero puede plantearse como un proceso semanal o mensual. Este esquema asegura que el modelo se actualice de manera periódica y siga siendo útil para apoyar decisiones basadas en la voz ciudadana vinculada a los ODS.



CONSTRUCCIÓN DE MODELOS


Para este proyecto se requieren al menos tres modelos de referencia, con el fin de comparar su rendimiento y seleccionar el más adecuado para el despliegue. Estos modelos deben actualizarse cada vez que se recolecten nuevas opiniones ciudadanas o cuando se observe una caída significativa en el desempeño, asegurando así que el sistema se mantenga vigente y robusto.

El tiempo estimado para generar el modelo, incluyendo la ingeniería de características (limpieza, lematización, TF-IDF) y el análisis de desempeño (matriz de confusión, métricas y curvas de aprendizaje), es de horas a pocos días, dependiendo del tamaño del lote de datos y la complejidad de la actualización. Esto permite que el modelo pueda reentrenarse periódicamente sin afectar los tiempos de entrega de información al negocio.



INGENIERÍA DE CARACTERÍSTICAS

Las variables utilizadas en el modelo corresponden a las palabras presentes en las opiniones ciudadanas, que se transforman en características numéricas mediante técnicas de procesamiento de lenguaje natural. A partir de los datos originales se aplicaron procesos de limpieza de texto (eliminación de tildes, caracteres especiales, URLs, menciones y stopwords), normalización a minúsculas, así como lematización y stemming para reducir las palabras a su forma base. Finalmente, se utilizó la técnica de TF-IDF para vectorizar los textos y asignar un peso relativo a cada término, lo que permite capturar la importancia de las palabras más representativas en el corpus y facilitar que los modelos identifiquen patrones relevantes.

equilibrada las opiniones ciudadanas para evitar sesgos que puedan distorsionar las decisiones de política pública.			
	<p>MONITOREO NO SE DEBE DILIGENCIAR</p>  <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 empresa? ¿Con qué frecuencia deben revisarse?</p>		

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 when sharing.

OWNML.CO