Universidad de los Andes
Docente: Diego Escobar
Curso: ISIS-3301
Juan David Pérez Díaz & Tomas Pardo Ramos

Contenido

Universidad de los Andes
Docente: Diego Escobar
Curso: ISIS-3301
Juan David Pérez Díaz & Tomas Pardo Ramos

1. Entendimiento y preparación de datos

El primer paso para el desarrollo de este proyecto consistio en llevar a cabbo un proceso de perfilamiento del conjunto de datos, con el fin de conocer en detalle sus caracteristicas y evaluar su calidad antes de la construcción de modelos de aplrendizaje. El dataset analizado estivo compuesto por 2424 registros correspondientes a opiniones ciudadanas redactadas en español, cada una asociada a una etiqueta perteneciente a los Objetivos de Desarrollo Sostenible (ODS) 1, 3 y 4. Este análisis inicial permitió identificar que la distribución de las clases presentaba un desbalance moderado: el ODS 4 representaba aproximadamente el 42 % de las instancias, el ODS 3 el 37 % y el ODS 1 el 21 %. Esta proporción constituye un aspecto relevante, dado que la representación desigual de categorías puede introducir sesgos en los algoritmos de clasificación si no se controla adecuadamente.

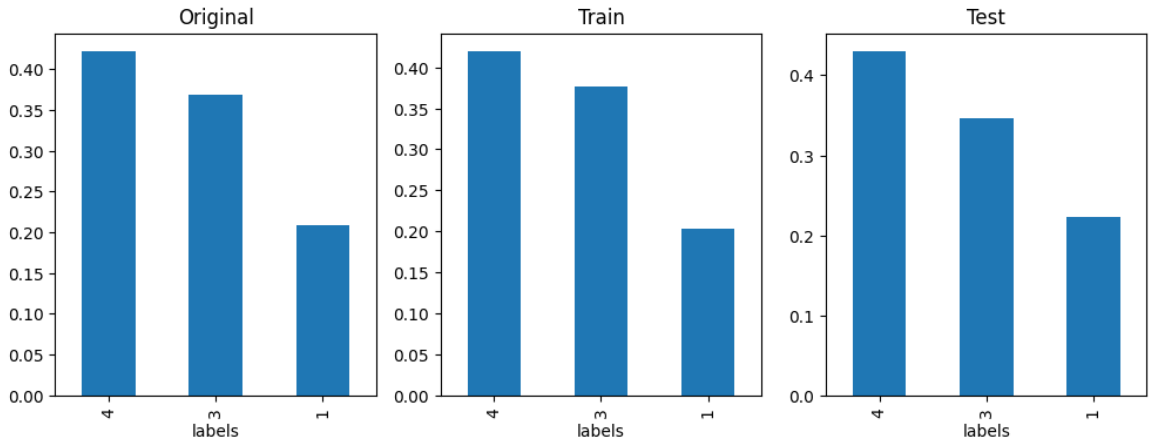

Ilustración 1 Distribución de los ODS

A continuación, se calcularon métricas descriptivas de los textos para caracterizar su estructura. Se incluyó el número total de caracteres por documento, la longitud máxima de palabra en cada texto, la longitud mínima de palabra y la longitud promedio de palabras. El análisis reveló que los textos presentan gran variabilidad, es decir algunos más extensos que otros. El promedio de longitud de palabras por documento oscila entre 5 y 6 caracteres, mientras que la longitud máxima de palabras alcanza hasta 25 caracteres en algunos casos. Este perfil estadístico muestra que las opiniones ciudadanas son heterogéneas en extensión y estilo, lo que justifica la necesidad de normalización previa al modelado.

En cuanto a la calidad de los datos, se comprobó la ausencia de valores nulos en las variables principales, y no se detectaron duplicados significativos. No obstante, se evidenció la presencia de acentos, caracteres no ASCII, signos de puntuación y números expresados tanto en dígitos como en palabras, lo cual introduce ruido y heterogeneidad léxica que debían ser abordados en la etapa de preparación.

Un aspecto clave en la preparación fue la eliminación de *stopwords* del español. A diferencia de aproximaciones tradicionales, en este proyecto las stopwords fueron procesadas de manera coherente con el algoritmo de *stemming*. Primero, cada token se redujo a su raíz morfológica mediante el *SnowballStemmer* en español, y posteriormente se descartaron aquellas raíces que correspondían a stopwords previamente stemmeadas, evitando así que variantes deformadas de palabras funcionales sobrevivieran en el vocabulario final. Este ajuste permitió obtener una representación más limpia y discriminativa de los documentos.

Finalmente, los tokens resultantes de la normalización fueron concatenados en cadenas de texto procesado, listas para ser utilizadas en los métodos de vectorización. De esta forma, se logró un conjunto de datos depurado, homogéneo y adaptado a las exigencias de las técnicas de representación (BOWy TF-IDF) y a los algoritmos seleccionados (Logistic Regression y KNN), garantizando la validez metodológica del modelado posterior.

2. Modelado y evaluación

Para la construcción de los modelos de clasificación orientados a identificar los ODS, se emplearon 3 algoritmos fundamentales de aprendizaje supervisado: Regresión Logística Multinomial Y K-Nearest Neighbors. Dichos algoritmos se implementaron dentro de pipelines en los que integraron los pasos de vectorización de texto (mediante BOW y TF-IDF) con el clasificador respectivo, garantizando así un flujo reproducible y libre de data leakage.

El primer modelo corresponde a la Regresión Logística Multinomial, utilizada en conjunto con la representación TF-IDF de los textos. Este algoritmo se caracteriza por su robustez en problemas de clasificación multiclase, su capacidad para manejar datos de alta dimensionalidad y su interpretabilidad a través de los coeficientes asociados a cada término. El modelo logró un accuracy en el conjunto de prueba de 97,2 %. Al analizar las métricas por clase, se observaron valores de precisión, recall y f1-score superiores al 0,90 en las tres categorías, con un macro-F1 promedio cercano al 0,93. Además, se identificaron términos clave que contribuyeron a la clasificación de cada ODS lo cual evidencia la capacidad del modelo para capturar patrones semánticos relevantes.

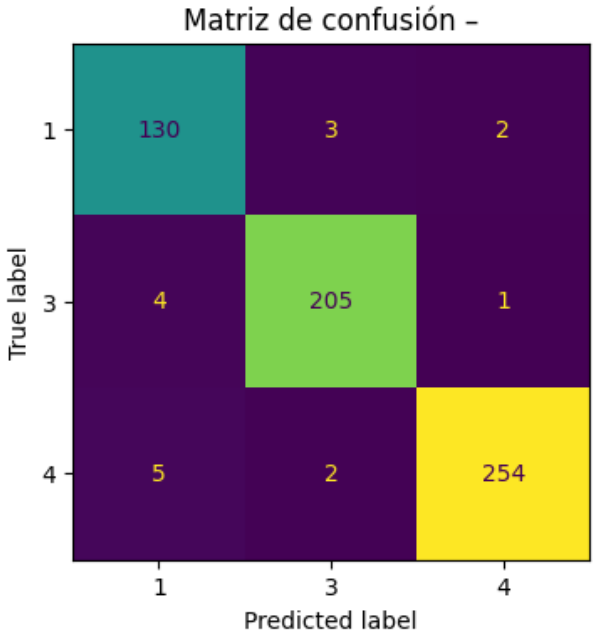

Ilustración 2 Matriz de confusión LR

El segundo algoritmo implementado fue K-Nearest Neighbors, el cual se evaluó en dos configuraciones: (i) con representación BoW binaria y métrica de similitud coseno, y (ii) con representación TF-IDF y distancia euclídea. En la primera variante (BoW + KNN, k=5), el modelo alcanzó un accuracy de 92,4 %, con un macro-F1 de 0,916. En la segunda variante (TF-IDF + KNN, k=5), los resultados mejoraron hasta un accuracy de 94,9 %, macro-F1 de 0,944 y un ROC-AUC macro de 0,9908, mostrando un desempeño competitivo y adecuado para comparar con la regresión logística. Sin embargo, se identificó que KNN presenta mayores costos computacionales para predicciones en tiempo real, debido a que requiere almacenar todo el conjunto de entrenamiento y calcular distancias para cada nuevo ejemplo.

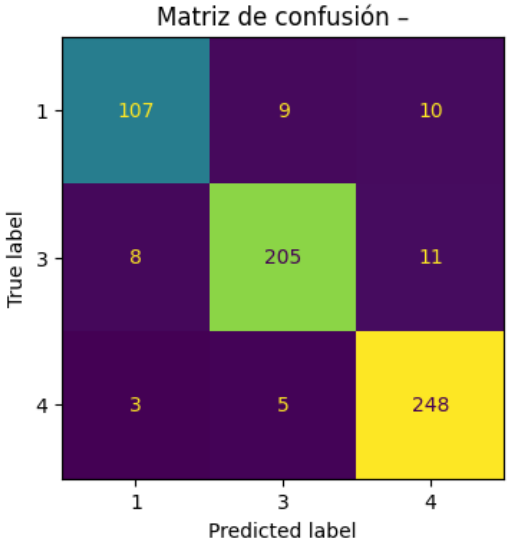

Ilustración 3 modelo KNN y BOW

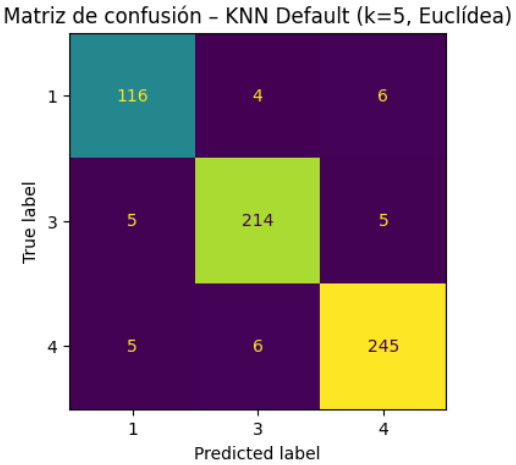

Ilustración 4 Modelo KNN y TF_IDF

El tercer algoritmo implementado fue el **Árbol de Decisión**, un modelo supervisado basado en la construcción de reglas jerárquicas del tipo “if-then”. En este enfoque, el espacio de características se divide de manera recursiva en nodos internos que representan condiciones sobre los atributos del texto (en este caso, las representaciones TF-IDF de las palabras), hasta llegar a hojas que asignan la etiqueta de salida correspondiente a cada ODS. Para este proyecto, se configuró un clasificador con criterio **Gini** y profundidad máxima de 3 niveles, buscando un balance entre interpretabilidad y capacidad predictiva.

En cuanto a desempeño, el Árbol de Decisión, alcanzó un accuracy del 72% en el conjunto de prueba, con un macro-F1 de 73%. El análisis detallado mostró que el modelo tiende a favorecer ciertas clases (ODS 3) en términos de recall, mientras que para otras (ODS 4) logra mayor precisión, lo cual evidencia un sesgo en la forma como segmenta el espacio de decisión. Pese a estas limitaciones, el árbol aporta alta interpretabilidad, ya que permite identificar directamente las palabras más relevantes para cada clasificación a través de las medidas de importancia de características y la visualización de la estructura del árbol. Esto lo convierte en una herramienta útil para explicar el razonamiento del modelo, aunque en términos de rendimiento bruto resulta menos competitivo que la regresión logística y KNN.

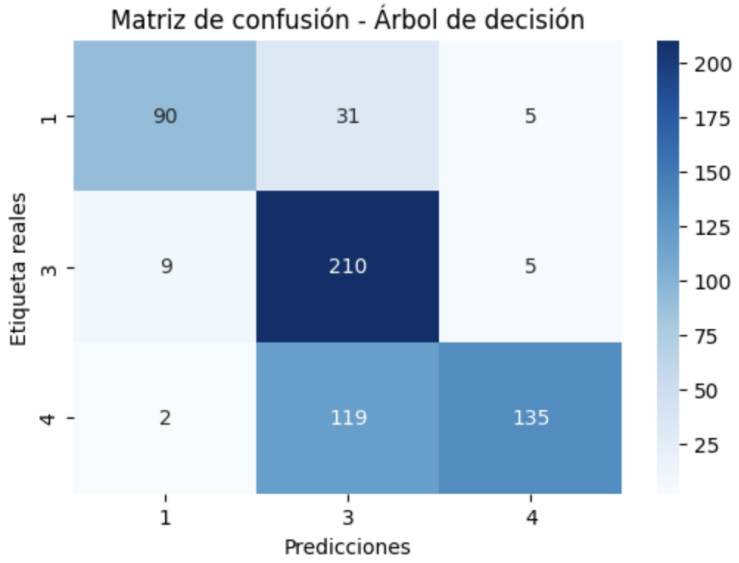

Ilustración 5. Mapa de confusión modelo árbol de decisión

3. Resultados

3.1. Descripción de los resultados obtenidos

El análisis de estas métricas permite comprender que ambos modelos responden a la necesidad central del negocio: agilizar el análisis de información ciudadana y asegurar una clasificación confiable de las opiniones en torno a los ODS prioritarios. La alta precisión y el balance en la clasificación reducen el riesgo de interpretaciones erróneas y facilitan la generación de reportes estratégicos para la toma de decisiones. Concretamente, estas métricas apoyan al UNFPA en la identificación rápida de problemáticas sociales en los territorios, permiten priorizar políticas en pobreza, salud y

educación, y garantizan que la voz ciudadana se integre de manera sistemática en el diseño y evaluación de políticas públicas.

3.2. Análisis de palabras

Es importante incluir el análisis de las palabras identificadas para relacionar las opiniones con los ODS y posibles estrategias que la organización debe plantear utilizando los resultados obtenidos en los modelos analíticos y una justificación de por qué esa información es útil para ellos.

Además de las métricas de calidad obtenidas por los modelos, resulta fundamental interpretar las palabras clave que los algoritmos identifican como determinantes para la clasificación de las opiniones ciudadanas. Estas palabras no solo confirman la robustez del proceso analítico, sino que también ofrecen a la organización información directa y accionable sobre cómo la ciudadanía expresa sus preocupaciones en torno a los Objetivos de Desarrollo Sostenible (ODS).

El análisis léxico se apoyó tanto en los coeficientes de la Regresión Logística como en las divisiones jerárquicas del árbol de decisión, los cuales coincidieron en resaltar términos característicos de cada ODS.

Para el ODS 1 (Fin de la pobreza), se destacan raíces como transferent, niñ, hog, proteccion, social, privacion, hogar, ingres, pobr y pobrez. Estas palabras reflejan que las opiniones relacionadas con la pobreza giran en torno a la protección social de familias y niños, la privación de recursos básicos y la falta de ingresos en el hogar. El modelo capta que la ciudadanía concibe la pobreza no solo como ausencia de dinero, sino como un fenómeno multidimensional vinculado a la seguridad social y la protección de grupos vulnerables. Para el UNFPA, esta información orienta estrategias de fortalecimiento de programas de protección social, subsidios focalizados y políticas de apoyo directo a hogares en condiciones de privación.

Universidad de los Andes
Docente: Diego Escobar
Curso: ISIS-3301
Juan David Pérez Díaz & Tomas Pardo Ramos

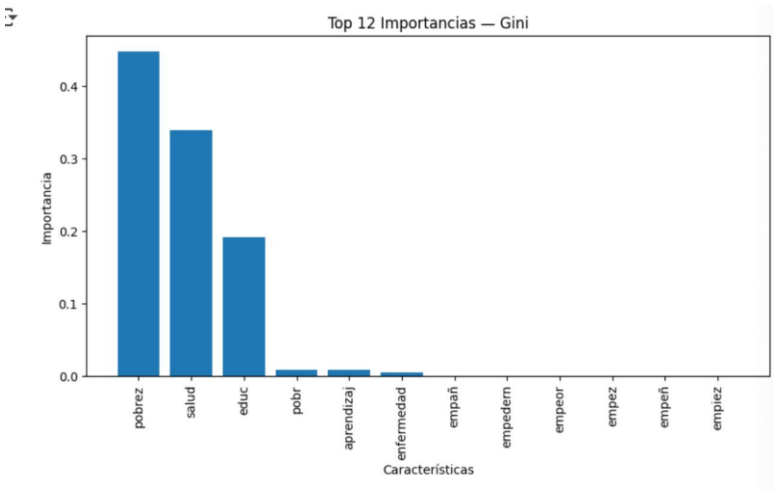

Ilustración 6. Importancia de palabras clave en la clasificación de ODS según el Árbol de Decisión (criterio Gini).

En el caso del ODS 3 (Salud y bienestar), las palabras más representativas incluyen alcohol, hospital, mortal, mental, enfermedad, sanitari, pacient, medic, atencion y salud. El modelo evidencia que los ciudadanos asocian la salud tanto con problemas estructurales del sistema sanitario como con condiciones específicas que afectan la vida cotidiana (salud mental, consumo de alcohol, acceso a medicamentos y mortalidad). Estos hallazgos sugieren que las estrategias de la organización deben centrarse en fortalecer la atención hospitalaria, promover campañas de prevención en salud mental y adicciones, así como ampliar la cobertura en servicios médicos esenciales.

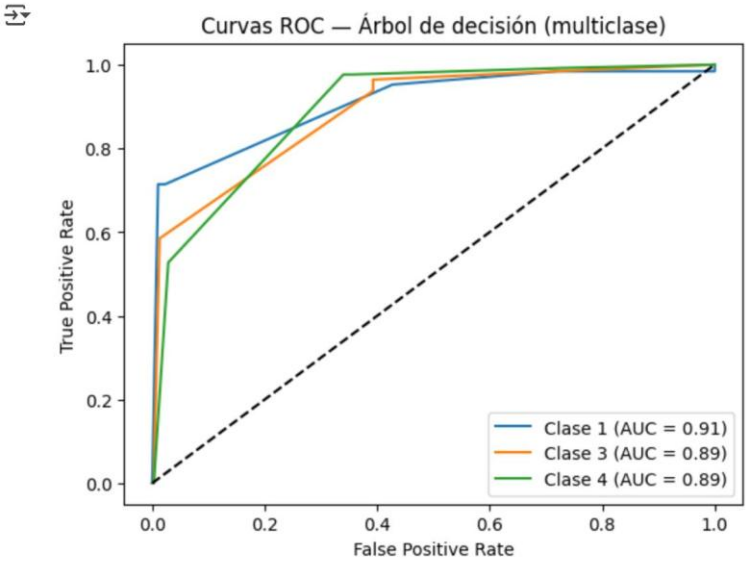

Ilustración 7. Curvas ROC por clase (ODS 1, 3 y 4); se resalta el alto AUC obtenido en el ODS 3 (Salud y bienestar).

Por su parte, el ODS 4 (Educación de calidad) se vincula con términos como evalu, enseñ, habil, profesor, aprendizaj, alumn, docent, estudi, escuel y educ. Estas palabras reflejan que la ciudadanía percibe la calidad educativa no únicamente en el acceso a escuelas, sino en el proceso mismo de enseñanza-aprendizaje, la preparación de los docentes y la formación integral de los estudiantes. A partir de esta evidencia, el UNFPA y los aliados institucionales pueden impulsar programas de capacitación docente, sistemas de evaluación más inclusivos y políticas que promuevan el aprendizaje significativo y la permanencia escolar.

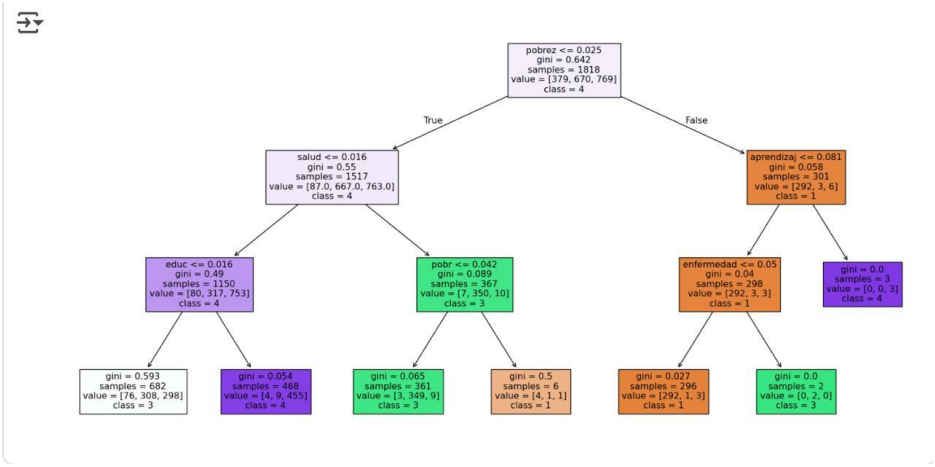

Ilustración 8. Representación del Árbol de Decisión (criterio Gini, profundidad 3) mostrando las divisiones por palabras clave en la clasificación de ODS.

La utilidad de este análisis radica en que traduce la salida estadística de los modelos en conocimiento práctico para la toma de decisiones. Las palabras clave actúan como indicadores temáticos que permiten entender cómo la población verbaliza sus necesidades en torno a pobreza, salud y educación. De esta manera, la organización puede diseñar políticas y programas más alineados con la voz ciudadana, mejorar la focalización de recursos y fortalecer el impacto de las intervenciones en la consecución de los objetivos de la Agenda 2030.