PROYECTO 1: ANALÍTICA DE DATOS

Etapa 1. Construcción de modelos de analítica de textos

TABLA DE CONTENIDO

SECCIÓN 1: ENTENDIMIENTO DEL NEGOCIO Y ENFOQUE ANALÍTICO

OPORTUNIDAD O PROBLEMA DE NEGOCIO

El análisis de las opiniones recolectadas a partir de herramientas de participación es un proceso que requiere tiempo y la intervención de expertos para interpretar y clasificar los datos en relación con los Objetivos de Desarrollo Sostenible (ODS). Tal

procedimiento, realizado de manera manual, implica un problema significativo para la UNFPA al ser ineficiente, costoso y limitar la capacidad de la organización para tomar decisiones rápidas basadas en datos sobre temas críticos como la salud, la educación y la igualdad de género.

Se identifica una oportunidad clara para optimizar el proceso anteriormente descrito mediante el uso de un modelo de *machine learning* capaz de clasificar automáticamente las opiniones ciudadanas según su relevancia para los ODS 3, 4 y 5.

OBJETIVOS Y CRITERIOS DE ÉXITO DESDE EL PUNTO DE VISTA DEL NEGOCIO

Bajo el contexto del negocio, se reconocen objetivos tales como: automatizar el análisis de las opiniones ciudadanas, disminuir el tiempo y los recursos necesarios para categorizar opiniones, garantizar que la herramienta de automatización permanezca precisa en la clasificación de nuevos registros y, por último, favorecer integración de un volumen de opiniones ciudadanas mucho mayor al proceso.

Para determinar el éxito de la solución planteada, se espera lograr un nivel de precisión superior al 90% en la clasificación de las opiniones ciudadanas con su respectivo ODS. Así mismo, se propone alcanzar una reducción en el tiempo de análisis de más del 75%, respecto al procedimiento manual de clasificación. Por último, se espera que el modelo implementado mantenga constante su precisión a través del tiempo mediante los ciclos de reentrenamiento.

ROL BENEFICIADO POR LA OPORTUNIDAD DEFINIDA

La UNFPA será directamente beneficiada en la implementación de la solución a la oportunidad identificada. Gracias al modelo, analistas de datos dentro de la organización podrán ser más eficientes, así como también los directivos y especialistas podrán contar con más información actualizada para definir políticas, asignar recursos y diseñar programas acordes con los ODS para lograr un mayor impacto. En general, los tomadores de decisiones dentro de la UNFPA se benefician con la implementación de una herramienta valiosa para mejorar la eficiencia, la calidad del análisis y la capacidad de respuesta en la formulación e implementación de políticas.

IMPACTO QUE PUEDE TENER EN COLOMBIA ESTE PROYECTO

En primer lugar, el análisis automatizado de las opiniones de los ciudadanos permitirá identificar de manera más ágil las preocupaciones relacionadas con la salud pública. Esto podría resultar en una asignación más eficiente de recursos y en la implementación de políticas más efectivas para abordar cuestiones como el acceso a la atención médica, la salud materna y el control de enfermedades, especialmente en comunidades vulnerables. Por otro lado, al clasificar las opiniones de los ciudadanos sobre el sistema educativo, el modelo ayudará a identificar las áreas clave que requieren mejoras y a comprender las percepciones sobre la calidad de la educación. Además, esta herramienta permitirá detectar de manera más efectiva las preocupaciones y

barreras relacionadas con la igualdad de género, proporcionando datos esenciales para diseñar políticas que promuevan el empoderamiento de mujeres y niñas en Colombia.

ENFOQUE ANALÍTICO

Este proyecto se enmarca en el análisis predictivo. Su objetivo es entrenar un modelo capaz de predecir la categoría a la que pertenece una nueva instancia de datos, utilizando un histórico de registros etiquetados recolectados. En particular, el modelo se entrena con datos relacionados con opiniones ciudadanas para predecir el Objetivo de Desarrollo Sostenible (ODS) al que podría referirse cada opinión. Este tipo de aprendizaje se clasifica como supervisado, ya que se basa en un conjunto de datos previamente etiquetados, y la tarea específica es la categorización de las opiniones ciudadanas según su ODS correspondiente.

Teniendo en cuenta lo anterior, las técnicas a utilizar serán, en primer lugar, las relacionadas con el procesamiento de lenguaje natural (NLP). Específicamente, se emplearán técnicas de tokenización, lematización y stemming, así como la eliminación de palabras vacías (stop words) y la vectorización del texto. Dadas las características del problema, se propone la utilización de los siguientes algoritmos de machine learning: Naive Bayes, Regresión Logística y MLPClassifier (Red neuronal)."

SECCIÓN 2: ENTENDIMIENTO Y PREPARACIÓN DE LOS DATOS

PERFILAMIENTO DE DATOS

Durante la etapa de perfilamiento de datos, se documenta que se ha obtenido por parte de la UNFPA, una colección extensa de datos donde se clasifica una lista de comentarios respecto al objetivo de desarrollo sostenible ODS (3, 4 o 5) al que corresponden. Los datos han sido recolectados en un archivo de *Excel* en el cual cada registro corresponde a un comentario de un ODS. En total, se obtuvieron 4049 filas, cada una detallada por 2 columnas, donde en la primera (*Textos_espanol*) se muestra el comentario, y en la segunda (*sdg*) el ODS al que corresponden (3, 4 o 5). De estas columnas, la primera es de tipo texto y la segunda es de tipo categórica, respectivamente. Ninguna columna identifica de manera única cada opinión ciudadana, sin embargo, la coincidencia en la columna tipo texto de dos opiniones que se registren como distintas es improbable. Además, se puede observar que la distribución de los valores de la columna *sdg* es la siguiente: para el ODS 5 un 36% de los registros (1451), para el ODS 4 un 33% de los registros (1354) y para el ODS 3 un 30% de las opiniones (1244).

ANÁLISIS DE CALIDAD DE LOS DATOS

Los datos presentan buena calidad, con alta completitud y sin duplicados, lo cual es fundamental para un análisis confiable. La validez de las opiniones se respalda por su longitud, pero se necesita un manejo cuidadoso de los caracteres especiales para mantener la integridad del texto. En términos de consistencia, aunque existe una correlación entre las opiniones y los ODS asignados, hay margen para mejorar mediante técnicas de procesamiento de texto más avanzadas y un enfoque más

detallado en la identificación de patrones de consistencia semántica. Mantener la calidad de los datos en estos aspectos asegurará una mejor fidelidad y precisión en los resultados de modelos predictivos y análisis futuros.

1. COMPLETITUD:

En la base de datos no hay valores nulos, lo cual garantiza que cada fila y columna contiene un valor válido. Esto es esencial para asegurar la robustez de los análisis posteriores, ya que la ausencia de valores faltantes reduce la necesidad de técnicas de imputación y evita sesgos en los resultados. La completitud de datos también facilita la integridad y consistencia en el preprocesamiento, análisis y modelado de datos.

2. UNICIDAD:

No se han encontrado duplicados totales, lo que indica que cada fila en la base de datos representa una observación única. En particular, en la columna de opinión, no se encuentran valores repetidos. Esto sugiere que cada opinión es única y no hay redundancia directa en el contenido.

En la columna de objetivo (ODS), existen repeticiones, pero esto es esperado ya que se trata de una columna categórica que clasifica opiniones en los (ODS) relevantes.

3. VALIDEZ:

La validez de las opiniones se ha asegurado mediante el análisis del número de caracteres por cada opinión, que varía entre 294 y 1513 caracteres. Esto sugiere que cada opinión es lo suficientemente extensa para ofrecer información significativa. Sin embargo, se identificaron caracteres especiales (como la ñ y las tildes), que pueden causar problemas de codificación. Para manejar estas inconsistencias se implementarán técnicas de limpieza y normalización de texto, como lematización y eliminación de caracteres especiales, para asegurar que todos los caracteres sean interpretados correctamente por cualquier modelo de procesamiento de lenguaje natural. Además, utilizar librerías que soporten adecuadamente caracteres especiales y textos en español.

4. CONSISTENCIA:

Se evaluó la consistencia entre las opiniones y los ODS asignados mediante listas de palabras clave para cada ODS:

- Para el ODS 3 (Salud y Bienestar), se encontró una consistencia del 69,9% con una palabra clave y del 30,9% con dos palabras clave.
- Para el ODS 4 (Educación de Calidad), se obtuvo un 61,8% con una palabra clave y un 20,8% con dos palabras clave.
- Para el ODS 5 (Igualdad de Género), la consistencia fue del 90,4% con una palabra clave y del 27,0% con dos palabras clave.

Estos resultados sugieren que el uso de dos palabras clave reduce la consistencia detectada. Esto puede deberse a la variabilidad en la expresión de las ideas y la posibilidad de que las opiniones utilicen sinónimos o diferentes formas gramaticales.

PREPARACIÓN DE LOS DATOS

Los datos proporcionados por el cliente presentan una gran variabilidad en su formato, lo que hace necesario prepararlos adecuadamente antes de utilizarlos como entrada en los algoritmos. Idealmente, las transformaciones aplicadas a los datos deberían ser reutilizables para todos los algoritmos. Por lo tanto, decidimos procesar los datos mediante un pipeline de preparación que pueda ser compartido por todos los algoritmos. Este pipeline consta de cuatro pasos principales: limpieza, tokenización, normalización y vectorización.

En la primera etapa, el proceso de limpieza se encarga de estandarizar el texto, corrigiendo caracteres especiales y normalizando la ortografía. El texto se convierte a minúsculas, se eliminan números y caracteres no alfabéticos, y se eliminan stopwords en español. Esto asegura que el texto esté en un formato consistente y libre de ruido para las siguientes etapas.

En seguida, la tokenización divide el texto en palabras individuales o "tokens". Esta etapa incluye la eliminación de filas con valores nulos, lo que garantiza que solo se procesen textos válidos. La tokenización permite que las palabras individuales sean manejadas por los algoritmos, facilitando su análisis y procesamiento posterior.

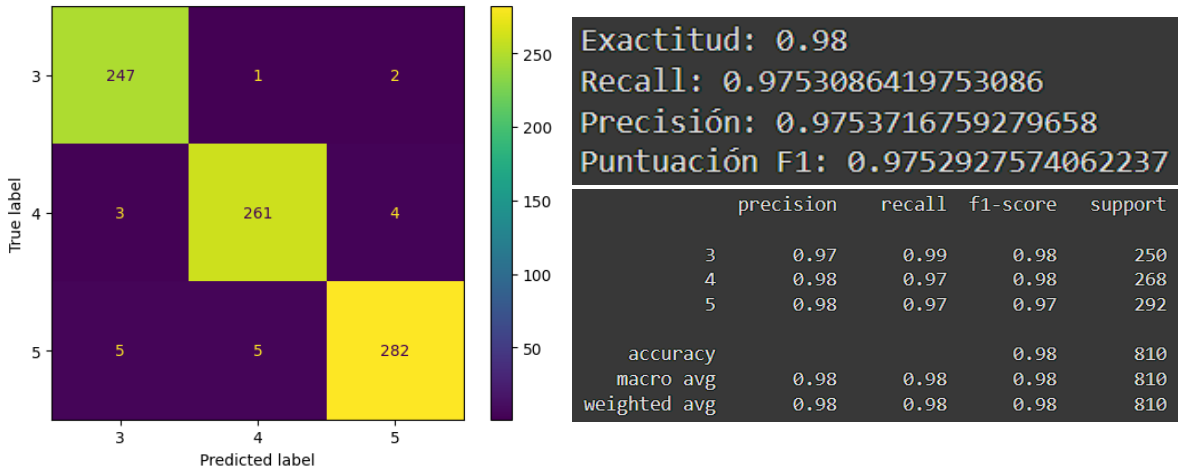
En la etapa de normalización, se aplican técnicas de lematización, que convierten las palabras a sus formas básicas o "lemas". Esto reduce la variabilidad lingüística, ayudando a los algoritmos a reconocer palabras diferentes que tienen el mismo significado. La normalización se realiza utilizando un modelo de spaCy para español, asegurando una lematización precisa y contextual.

Finalmente, la vectorización transforma el texto tokenizado y normalizado en una representación numérica que los algoritmos pueden procesar. Se utiliza el método TF-IDF (Term Frequency-Inverse Document Frequency), que pondera la relevancia de las palabras en el contexto de todo el corpus, mejorando la capacidad del modelo para distinguir entre términos importantes y comunes.

SECCIÓN 3: MODELADO Y EVALUACIÓN

MLP CLASIFIER (DAVID SAMUEL ROJAS SANCHEZ)

Este algoritmo utiliza una red neuronal con una capa oculta para clasificar las opiniones de los ciudadanos en relación con los ODS. La red neuronal es capaz de aprender patrones complejos en los datos, lo que le permite realizar predicciones precisas. Sin embargo, su entrenamiento puede ser más lento y requiere más recursos en comparación con otros modelos.

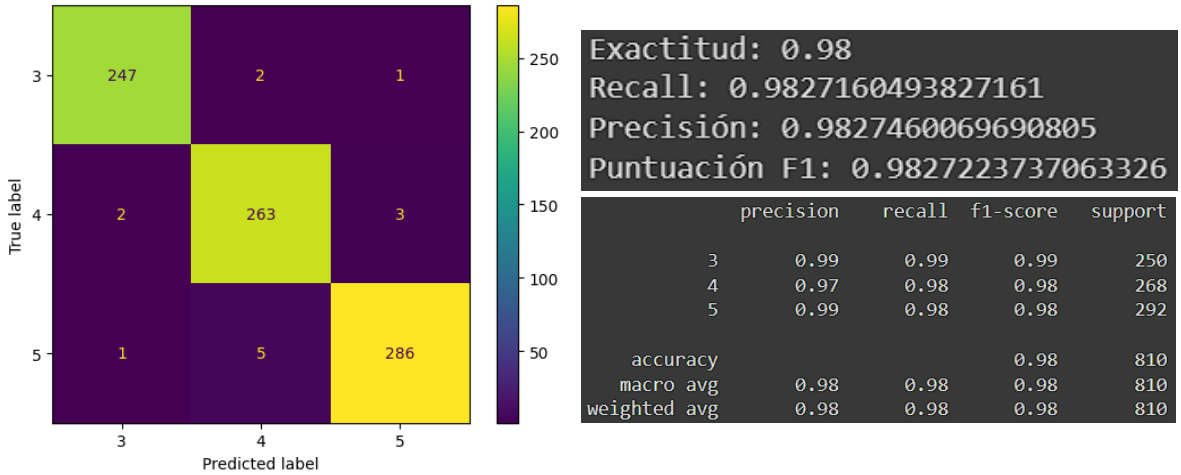


	precision	recall	f1-score	support
3	0.97	0.99	0.98	250
4	0.98	0.97	0.98	268
5	0.98	0.97	0.97	292
accuracy			0.98	810
macro avg	0.98	0.98	0.98	810
weighted avg	0.98	0.98	0.98	810

Fig. 1 *Métricas de calidad correspondientes al algoritmo de la red neuronal sobre los datos de prueba. Se evidencia a la izquierda la matriz de confusión y a la derecha el reporte completo de clasificación.*

REGRESIÓN LOGÍSTICA (GABRIELA GARCÍA SUAREZ)

La regresión logística es un modelo lineal que se utiliza para predecir la probabilidad de que una entrada pertenezca a una clase específica. En este proyecto, se aplica para clasificar las opiniones en las categorías de los ODS. Es un modelo rápido, eficiente y fácil de interpretar, ideal para situaciones donde las relaciones entre las características y las clases son lineales.



	precision	recall	f1-score	support
3	0.99	0.99	0.99	250
4	0.97	0.98	0.98	268
5	0.99	0.98	0.98	292
accuracy			0.98	810
macro avg	0.98	0.98	0.98	810
weighted avg	0.98	0.98	0.98	810

Fig. 2 *Métricas de calidad correspondientes al algoritmo de regresión sobre los datos de prueba. Se evidencia a la izquierda la matriz de confusión y a la derecha el reporte completo de clasificación.*

NAIVE BAYES (ELKIN CUELLO)

Este es un algoritmo probabilístico que asume la independencia condicional entre las características. En el contexto del proyecto, se utiliza para clasificar el texto basándose en la probabilidad de que un conjunto de palabras pertenezca a una categoría relacionada con los ODS. Es muy eficiente, especialmente en clasificación de texto, pero

su rendimiento puede verse afectado si las características no son realmente independientes.

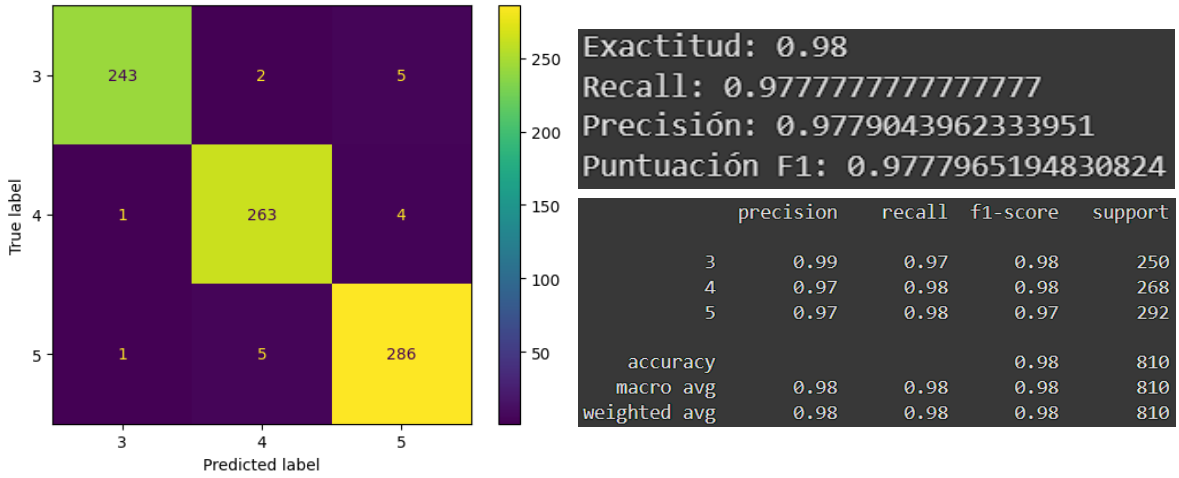


Fig. 3 *Métricas de calidad correspondientes al algoritmo de Naive Bayes sobre los datos de prueba. Se evidencia a la izquierda la matriz de confusión y a la derecha el reporte completo de clasificación.*

SECCIÓN 4: RESULTADOS

DESCRIPCIÓN DE RESULTADOS

En el contexto del proyecto, las métricas de rendimiento como **precisión**, **recall**, **exactitud** y **f1-score** son fundamentales para evaluar la efectividad de los modelos en la clasificación de opiniones ciudadanas en relación con los Objetivos de Desarrollo Sostenible (ODS).

Se debe tener en cuenta que, la **precisión** mide la proporción de predicciones correctas sobre todas las predicciones realizadas para una clase específica, lo que indica cuántas de las etiquetas predichas como positivas son realmente positivas. En este proyecto, una alta precisión significa que el modelo está evitando falsos positivos, es decir, no clasifica incorrectamente las opiniones en una categoría de ODS a la que no pertenecen. Así mismo, el **recall** es la proporción de verdaderos positivos sobre el total de instancias relevantes (positivos reales). Esto refleja la capacidad del modelo para capturar todas las instancias de una clase. Un alto recall en este contexto asegura que el modelo está identificando la mayoría de las opiniones relevantes para cada ODS, minimizando los falsos negativos. Por su parte, la exactitud o **accuracy** es la proporción de predicciones correctas (tanto verdaderos positivos como verdaderos negativos) sobre el total de predicciones. En este caso, una alta exactitud significa que el modelo en general clasifica correctamente la mayoría de las opiniones en las categorías correctas de ODS. Para balancear las medidas, se utiliza el **F1-score,** el cual es la media armónica entre precisión y recall, y proporciona una medida equilibrada de la precisión y el recall del modelo. Es particularmente útil en situaciones donde existe un desequilibrio entre las clases. Un alto f1-score sugiere que el modelo logra un buen balance entre capturar la mayoría de las instancias relevantes (recall) y ser preciso en sus predicciones (precisión).

Al analizar las métricas de los tres algoritmos utilizados en el proyecto (MLPClassifier, Regresión Logística y Naive Bayes), podemos observar que todos presentan un desempeño notable en términos de exactitud, recall, precisión y puntuación F1. Todos los resultados poseen una cantidad pertinente de aciertos distribuidos de manera equilibrada para las tres categorías de ODS. Sin embargo, pequeñas diferencias en las métricas y el contexto de los objetivos del negocio son cruciales para interpretar los resultados y seleccionar el mejor modelo.

MLPClassifier es un modelo robusto que, a pesar de su complejidad y necesidad de más recursos computacionales, tiene la capacidad de capturar patrones complejos en los datos. El desempeño general del MLPClassifier, con una puntuación F1 de 0.975, lo hace altamente confiable para la clasificación de las opiniones de los ciudadanos en relación con los ODS. Sin embargo, en el contexto de este proyecto, donde la facilidad de implementación y la eficiencia son clave, el ligero descenso en las métricas frente a la Regresión Logística podría no justificar su uso, especialmente si el entrenamiento más largo no aporta beneficios significativos adicionales.

La **Regresión Logística**, por otro lado, muestra un rendimiento ligeramente superior con una puntuación F1 de 0.9827. Este modelo lineal es más fácil de interpretar y rápido de entrenar, lo que lo hace ideal para situaciones donde la simplicidad y la velocidad son prioritarias. En el contexto del negocio, este modelo es valioso porque permite a los analistas y tomadores de decisiones entender de manera clara las relaciones entre las variables y las categorías de ODS, asegurando una rápida implementación y mantenimiento continuo del modelo en la aplicación web o móvil. La alta precisión y recall en este caso indican que el modelo logra identificar correctamente la mayoría de las clases sin generar demasiados falsos positivos o negativos, lo cual es crucial para mantener la confianza en las predicciones automáticas que se utilizarán para la toma de decisiones.

Finalmente, **Naive Bayes** es un modelo muy eficiente y fácil de implementar, que también obtiene una puntuación F1 cercana (0.977). Sin embargo, su suposición de independencia condicional entre las características puede no capturar de manera óptima la complejidad de los datos de texto, lo que podría ser una desventaja en escenarios donde las relaciones entre palabras son más complejas. A pesar de su eficiencia, el hecho de que Naive Bayes tenga un rendimiento ligeramente inferior en comparación con los otros modelos sugiere que, aunque es útil, podría no ser la mejor opción cuando se busca maximizar la precisión en la clasificación en relación con los ODS.

En conclusión, aunque los tres modelos presentan buenos resultados, la **Regresión Logística** emerge como la mejor opción en este contexto específico. Su balance entre simplicidad, velocidad de entrenamiento, y rendimiento ligeramente superior en métricas clave, hace que sea la mejor opción para ser implementada en una aplicación web o móvil, facilitando el cumplimiento de los objetivos de UNFPA en términos de eficiencia y efectividad en la clasificación de opiniones ciudadanas en relación con los ODS.

ANÁLISIS DE PALABRAS, UTILIDAD DE LA INFORMACIÓN Y PLANTEAMIENTO DE ESTRATEGIAS

Las palabras clave identificadas a través del modelo de regresión logística proporcionan un análisis profundo sobre los temas recurrentes en las opiniones ciudadanas, relacionadas con los Objetivos de Desarrollo Sostenible (ODS). Al observar las palabras más importantes en cada clase (ODS 3, 4 y 5), la organización puede extraer información valiosa para dirigir sus estrategias de manera más efectiva hacia los problemas que son más prioritarios para los ciudadanos.

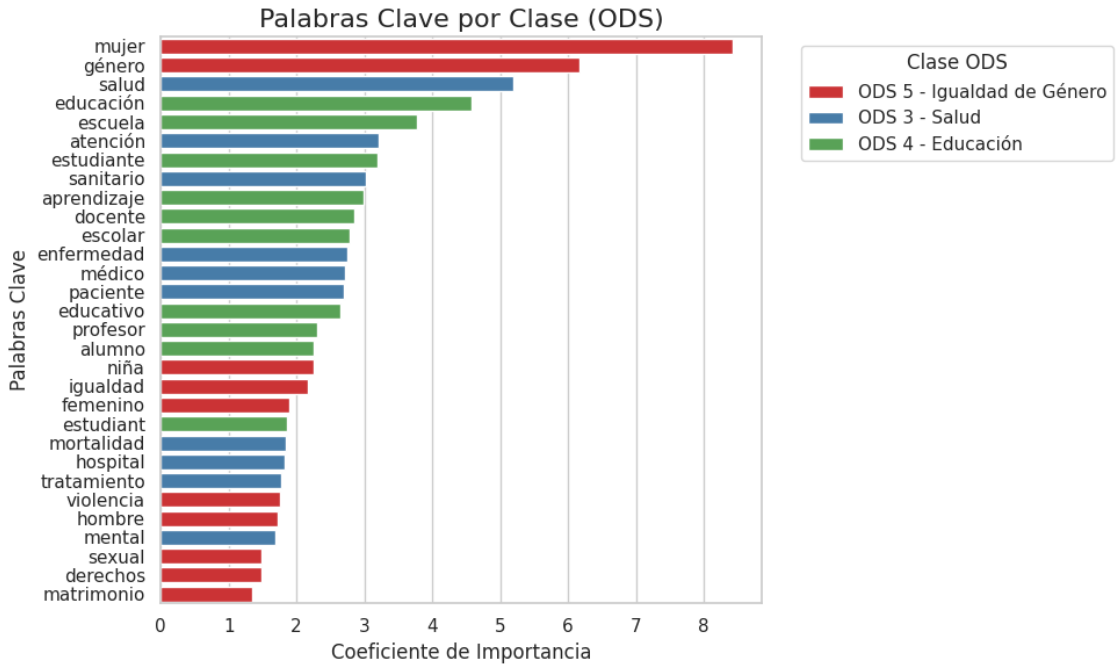


Fig. 4 *Gráfico de barras apiladas para la comparación de las palabras clave de las diferentes clases (ODS 3, 4 y 5). Se evidencia la relevancia de cada palabra una según su coeficiente y clase ODS.*

En la clase 3, relacionada con salud (ODS 3), términos como "salud", "atención", "sanitario", y "médico" resaltan que gran parte de las preocupaciones ciudadanas están enfocadas en los servicios de atención médica, la infraestructura sanitaria y las enfermedades. Palabras como "mortalidad", "hospital" y "tratamiento" indican una inquietud sobre la calidad y accesibilidad del sistema de salud, así como la respuesta ante enfermedades graves. Estrategias futuras podrían enfocarse en mejorar la atención hospitalaria, reducir las tasas de mortalidad y aumentar el acceso a tratamientos médicos, especialmente en áreas vulnerables.

En la clase 4, relacionada con educación (ODS 4), las palabras clave como "educación", "escuela", "estudiante" y "aprendizaje" demuestran que las preocupaciones se centran en el acceso y calidad de la educación. Términos como "docente" y "escolar" sugieren que la relación entre maestros y alumnos, junto con los recursos educativos, son temas críticos para la ciudadanía. Para abordar estas inquietudes, la organización debería enfocarse en mejorar los recursos educativos, capacitar a los docentes y asegurar

que los estudiantes reciban una educación de calidad, lo que podría traducirse en una mejora en los resultados educativos a largo plazo.

Finalmente, en la clase 5, relacionada con igualdad de género (ODS 5), palabras clave como "mujer", "género", "violencia" y "derechos" subrayan la importancia de abordar la desigualdad y la violencia de género. El uso de términos como "igualdad", "femenino" y "niña" sugiere que las conversaciones están centradas en la equidad y en las oportunidades para mujeres y niñas, así como en la protección contra la violencia. Estrategias enfocadas en la promoción de políticas de igualdad de género, campañas de concientización y la implementación de leyes más estrictas contra la violencia de género serían acciones claves para la organización.

La identificación de estas palabras clave es crucial porque permite a la organización desarrollar políticas específicas que aborden los problemas más relevantes para los ciudadanos. Utilizando los resultados del análisis, la organización puede priorizar sus esfuerzos de intervención y realizar un seguimiento continuo del impacto de sus políticas a través de la opinión pública, ajustando sus estrategias de manera ágil y efectiva.

SECCIÓN 5: MAPA DE ACTORES

Descripción de una organización que puede beneficiarse del resultado del modelo analítico planteado.

Rol dentro de la empresa	Tipo de actor	Beneficio	Riesgo
Analistas de datos	Usuario	Reducción significativa del tiempo de análisis de opiniones y mayor eficiencia en el procesamiento de datos.	Si el modelo no es preciso, podría clasificar incorrectamente opiniones, lo que llevaría a una mala interpretación de los datos.
Directores y tomadores de decisiones	Beneficiado	Acceso a análisis más rápidos y basados en datos para diseñar políticas y tomar decisiones estratégicas.	Si el modelo no se mantiene actualizado con los cambios en la opinión pública, podría influir negativamente en la formulación de políticas.
Departamento de Tecnología e Innovación	Proveedor	Promueve la innovación y mejora de herramientas tecnológicas dentro de la organización para aumentar la eficiencia.	Inversiones en tecnología que podrían no proporcionar el retorno esperado si el modelo no logra ser eficaz o precisa en sus resultados.

Departamento de Tecnología e Innovación Financiadores (donantes y patrocinadores)	Financiador	Garantía de que los fondos se utilizan en herramientas que mejoran la toma de decisiones y la eficiencia organizacional.	Riesgo de pérdida de confianza si los fondos no resultan en mejoras tangibles o si la herramienta no tiene el impacto esperado.
Comunidades y población civil	Beneficiado	Mejora en la atención a sus necesidades y preocupaciones gracias a decisiones informadas por análisis precisos de opiniones.	Pueden ser malinterpretadas o subrepresentadas si el modelo tiene sesgos o errores en la clasificación de opiniones.

SECCIÓN 6: TRABAJO EN EQUIPO

GABRIELA GARCÍA ***(33 pts.)***

Roles: Líder de datos, líder de negocio.

Tareas: Análisis de calidad de los datos, estructuración de pipelines, video de presentación, implementación del algoritmo

Algoritmo: Regresión logística.

Horas de trabajo: 10 horas de trabajo.

Retos: Ajuste de hiperparámetros, vectorización de los datos mediante TF-IDF.

ELKIN CUELLO ***(33 pts.)***

Roles: Líder de proyecto, líder de analítica.

Tareas: Estructuración de pipelines, conclusiones de análisis, evaluación cuantitativa, implementación del algoritmo.

Algoritmo: Naive Bayes.

Horas de trabajo: 9 horas de trabajo.

Retos: Métodos de evaluación cuantitativa de los modelos.

DAVID ROJAS ***(33 pts.)***

Roles: Líder de negocio, líder de analítica.

Tareas: Perfilamiento de datos, pipeline de procesamiento de datos, implementación del algoritmo.

Algoritmo: MLP Classifier.

Horas de trabajo: 9 horas de trabajo.

Retos: Familiarización con las librerías para el procesamiento natural como spaCy y proceso de persistencia del modelo con pipelines.

Con el fin de enfrentar los retos identificados, se destinó tiempo y esfuerzo a comprender la tecnología con la cual trabajamos durante el proyecto. Mediante la investigación, la lectura y el apoyo de herramientas de inteligencia artificial, todos los integrantes pudimos aprender lo necesario para mitigar nuestras debilidades.