Proyecto 1 – Etapa 1

Sergio Ramírez Vélez

Sergio Gonzalez Mateus

Luis Castelblanco Quintero

Universidad de Los Andes

Programa Ingeniería de Sistemas

7 septiembre de 2024

a. Análisis de los resultados obtenidos

El objetivo principal del proyecto es automatizar la clasificación de opiniones ciudadanas en tres categorías basadas en los **Objetivos de Desarrollo Sostenible (ODS)**:

- **ODS 3**: Salud y Bienestar.
- **ODS 4**: Educación de Calidad.
- **ODS 5**: Igualdad de Género.

Al clasificar automáticamente estas opiniones, se espera que la organización pueda entender mejor los problemas prioritarios de la población y tomar decisiones más informadas y rápidas.

Entendimiento de los datos

Los datos consisten en 4049 opiniones clasificadas en los tres ODS mencionados:

- **ODS 5 (Igualdad de Género)**: 1451 opiniones (35.8%).
- **ODS 4 (Educación de Calidad)**: 1354 opiniones (33.4%).
- **ODS 3 (Salud y Bienestar)**: 1244 opiniones (30.7%).

Los textos tienen una longitud media de 699 caracteres con una desviación estándar de 228 caracteres. Esto sugiere que la mayoría de las opiniones están dentro de un rango razonablemente compacto en términos de longitud, lo que puede influir en el rendimiento de los modelos de clasificación.

Preparación de datos

El preprocesamiento de los textos incluye las siguientes etapas:

- **Tokenización y eliminación de palabras vacías**: Esto es crucial para eliminar ruido en los datos y enfocarse en las palabras clave que aportan valor.
- **Vectorización utilizando TF-IDF:** Tras limpiar y normalizar los textos, se aplicó la técnica de vectorización TF-IDF (Term Frequency - Inverse Document Frequency) para convertir los textos en representaciones numéricas. TF-IDF asigna un valor a cada palabra basado en su frecuencia en un documento y su rareza en todo el conjunto de datos, lo que permite que las palabras más relevantes (y menos comunes) tengan mayor peso en el modelo de clasificación. Esta representación numérica es esencial para que los algoritmos de aprendizaje automático puedan procesar y clasificar las opiniones de manera eficiente.

Este preprocesamiento permite al modelo trabajar con representaciones más limpias y homogéneas, lo que es fundamental para mejorar la precisión de la clasificación.

Modelos utilizados y su comparación

Random Forest

- **Precisión general**: 97%
- **F1-Score promedio**: 0.97
- **Precisión por clase**:
    - ODS 3: Precisión 0.98, Recall 0.98, F1-Score 0.98.
    - ODS 4: Precisión 0.96, Recall 0.97, F1-Score 0.97.
    - ODS 5: Precisión 0.98, Recall 0.96, F1-Score 0.97.

El modelo de **Random Forest** es el más robusto de los tres, alcanzando una precisión general del 97%. Este alto rendimiento es resultado de la capacidad del modelo para manejar grandes cantidades de características (palabras) y combinaciones de decisiones.

Impacto en el negocio: La alta precisión y el excelente equilibrio entre precisión y recall aseguran que casi todas las opiniones ciudadanas se clasifiquen correctamente. Esto implica que la organización puede confiar en el modelo para tomar decisiones informadas sin requerir demasiada intervención humana, lo que optimiza los recursos y reduce tiempos de procesamiento.

K-Nearest Neighbors (KNN)

- **Precisión general**: 94%
- **F1-Score promedio**: 0.94
- **Precisión por clase**:
    - ODS 3: Precisión 0.93, Recall 0.93, F1-Score 0.93.
    - ODS 4: Precisión 0.94, Recall 0.95, F1-Score 0.95.
    - ODS 5: Precisión 0.94, Recall 0.93, F1-Score 0.94.

El modelo **KNN** con 3 vecinos tiene un buen desempeño, aunque ligeramente inferior al de Random Forest. Este modelo es fácil de implementar y ajusta las predicciones basándose en los textos más cercanos en términos de características (usando la representación TF-IDF).

Impacto en el negocio: Si bien este modelo es más simple, su precisión es ligeramente inferior. Esto puede ser útil en entornos donde se requiere una

implementación rápida, pero no es tan efectivo para garantizar la precisión en la clasificación de grandes volúmenes de datos.

Árbol de Decisión (CART)

- **Precisión general**: 92%
- **F1-Score promedio**: 0.92
- **Precisión por clase**:
    - ODS 3: Precisión 0.89, Recall 0.91, F1-Score 0.90.
    - ODS 4: Precisión 0.94, Recall 0.91, F1-Score 0.92.
    - ODS 5: Precisión 0.94, Recall 0.95, F1-Score 0.95.

El modelo **CART** tiene la menor precisión de los tres, aunque sigue siendo útil. Los árboles de decisión son fáciles de interpretar y pueden ser beneficiosos cuando se requiere mayor transparencia en las predicciones.

Impacto en el negocio: Este modelo es más interpretativo, lo que lo hace útil para auditorías o explicaciones a usuarios no técnicos. Sin embargo, debido a su precisión relativamente baja, puede no ser la mejor opción para tareas donde se prioriza la exactitud.

Análisis de matriz de confusión de Random Forest

La matriz de confusión muestra que los errores de clasificación son bajos, pero están presentes:

- **ODS 5** presenta la mayor confusión, con un 3.08% de opiniones clasificadas erróneamente como ODS 4 en el modelo de Random Forest.
- **ODS 4** tiene un 1.87% de confusión hacia ODS 5.
- **ODS 3** es la clase con menos errores de clasificación.

Aunque los errores son bajos, es importante tener en cuenta que la mayor confusión ocurre entre ODS 4 y ODS 5, lo que podría indicar que los textos relacionados con estos dos objetivos tienen palabras clave o temas en común. Esto podría requerir ajustes adicionales en el preprocesamiento o el uso de modelos más complejos para reducir la confusión entre estas clases.

Contribución a los objetivos del negocio

Los modelos entrenados permiten a la organización clasificar automáticamente las opiniones ciudadanas con una precisión superior al 90% en todos los casos. Esto tiene varias implicaciones directas para el negocio:

1. **Optimización de recursos**: La clasificación automática reduce significativamente el tiempo y los recursos humanos necesarios para procesar grandes volúmenes de texto.
2. **Toma de decisiones más rápida**: Al identificar automáticamente los temas críticos relacionados con los ODS, la organización puede priorizar las áreas de intervención (salud, educación, igualdad de género) y tomar decisiones más informadas.
3. **Adaptabilidad**: El modelo de Random Forest, en particular, se puede reentrenar fácilmente con nuevos datos, lo que asegura que las decisiones sigan siendo precisas a lo largo del tiempo.

Modelo seleccionado

Decidimos utilizar **Random Forest** como modelo principal debido a su capacidad para manejar grandes volúmenes de datos textuales y su robustez frente al sobreajuste, lo que lo convierte en una opción confiable para clasificar las opiniones ciudadanas relacionadas con los ODS. Este modelo ofrece una alta precisión (97%), manteniendo un equilibrio excelente entre la precisión y el recall en todas las categorías de los ODS. Además, Random Forest puede manejar la complejidad inherente a los datos textuales, ya que combina múltiples árboles de decisión, lo que mejora la generalización del modelo y reduce el riesgo de clasificaciones erróneas. Su rendimiento superior en comparación con otros modelos probados, como KNN y CART, justifica su elección para garantizar predicciones más precisas y consistentes, lo que es fundamental para tomar decisiones basadas en datos en el contexto de la organización.

El uso de modelos de aprendizaje automático, en particular Random Forest, permite a la organización clasificar opiniones con una alta precisión, lo que facilita el logro de los objetivos de negocio relacionados con los ODS 3, 4 y 5. Los resultados obtenidos garantizan que las intervenciones se basen en datos precisos, lo que mejora la eficiencia y efectividad de las acciones tomadas.

b. Analisis de palabras

Este análisis integra las palabras y bigramas identificados en las nubes correspondientes a los ODS 3, 4 y 5, con el objetivo de proporcionar una visión más profunda de las preocupaciones ciudadanas y sugerir estrategias que la organización puede implementar. A través de la combinación de palabras clave y bigramas, obtenemos un entendimiento claro de los temas prioritarios para la población y cómo estos están relacionados con la consecución de los Objetivos de Desarrollo Sostenible.

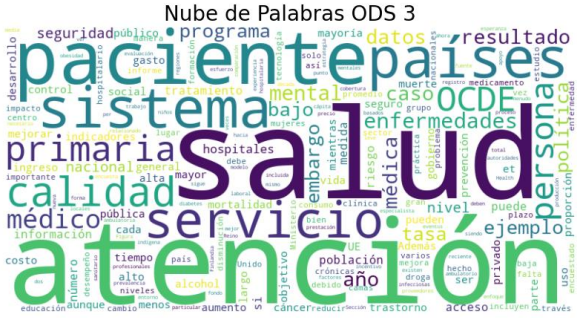
ODS 3: Salud y Bienestar

Palabras clave: Atención, salud, paciente, países, sistema, servicio, médico, mental, enfermedades, calidad.

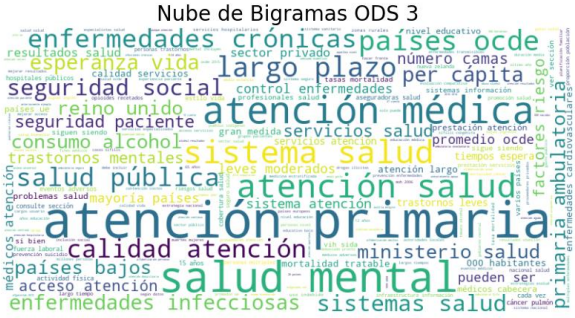
Bigramas clave: Atención primaria, salud mental, enfermedades crónicas, sistema salud, calidad atención.

Atención primaria y salud son términos fundamentales en las opiniones sobre el ODS La población está preocupada por la capacidad del sistema de salud para proporcionar atención básica de calidad y gestionar adecuadamente las enfermedades crónicas. El enfoque en salud mental sugiere que el bienestar psicológico es un tema de creciente relevancia, mientras que calidad atención indica que existe un interés en mejorar los estándares de los servicios médicos. Los bigramas refuerzan estas preocupaciones, especialmente con la mención de la atención primaria y la salud mental, lo que subraya la necesidad de una intervención en estos ámbitos para mejorar la infraestructura y la eficiencia del sistema de salud.

Nube de Palabras ODS 3



Nube de Bigramas ODS 3



Estrategias Sugeridas:

1. Reforzar la atención primaria: Dada la repeticion de la frase atención primaria, es necesario invertir en la mejora de los servicios básicos de salud, asegurando que la población tenga acceso a atención médica oportuna y de calidad.
2. Mejorar la atención en salud mental: El aumento de la mención a salud mental refleja una preocupación creciente que la organización debería abordar mediante la creación de programas de apoyo psicológico y campañas de sensibilización. 3.Enfocarse en las enfermedades crónicas: Dado que la calidad de atención y la gestión de enfermedades crónicas son temas recurrentes, la organización podría diseñar programas específicos para el seguimiento de enfermedades a largo plazo y la capacitación de profesionales médicos en este campo.

La salud es una prioridad para la población, y mejorar el acceso a atención primaria, así como la atención a salud mental y enfermedades crónicas, asegurará que el

sistema de salud responda mejor a las necesidades de la ciudadanía. Estas intervenciones facilitarán el logro del ODS 3.

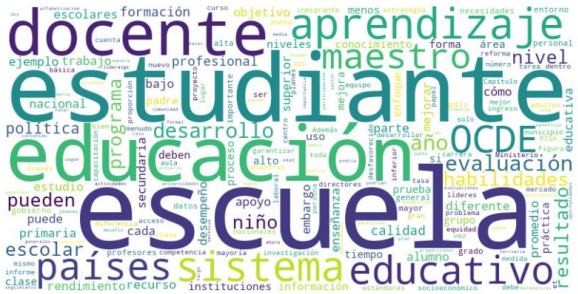
ODS 4: Educación de Calidad

Palabras clave: Escuela, estudiante, educación, docente, aprendizaje, evaluación, nivel, programa.

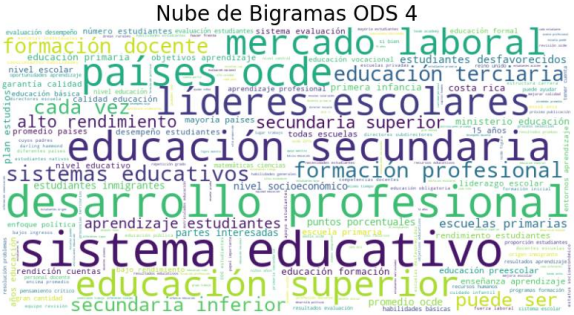
Bigramas clave: Educación secundaria, educación superior, sistema educativo, desarrollo profesional, líderes escolares.

Las palabras clave como escuela, estudiante y educación sugieren que la calidad de la enseñanza y el acceso a la educación son preocupaciones centrales para los ciudadanos. La aparición de docente y aprendizaje indica que existe un enfoque en mejorar la enseñanza y el rendimiento de los estudiantes. Los bigramas como educación secundaria, educación superior y desarrollo profesional refuerzan la importancia de la formación en niveles más avanzados y la necesidad de preparar a los estudiantes para el mercado laboral. El sistema educativo en su conjunto también es objeto de preocupación, lo que sugiere que se perciben problemas estructurales en la organización y gestión de la educación.

Nube de Palabras ODS 4



Nube de Bigramas ODS 4



Estrategias Sugeridas:

1. Capacitar a los docentes y líderes escolares: Con el énfasis en docentes y líderes escolares, la organización debe implementar programas de formación continua para mejorar las competencias pedagógicas y de gestión de los líderes educativos.
2. Fortalecer la educación secundaria y superior: Los términos educación secundaria y educación superior destacan la necesidad de mejorar la calidad de la enseñanza en estos niveles. Se sugiere diseñar programas que preparen mejor a los estudiantes para los retos académicos y laborales.
3. Desarrollar sistemas de evaluación más efectivos: La palabra evaluación sugiere que los ciudadanos no están satisfechos con los sistemas actuales. La

organización podría rediseñar los métodos de evaluación para medir de manera más precisa el progreso y las habilidades de los estudiantes.

Mejorar la calidad de la educación en todos los niveles y ofrecer oportunidades de desarrollo profesional tanto para estudiantes como para docentes ayudará a cerrar las brechas educativas y contribuirá al cumplimiento del ODS 4. Al abordar estas preocupaciones, la organización estará mejorando la preparación académica y profesional de los estudiantes.

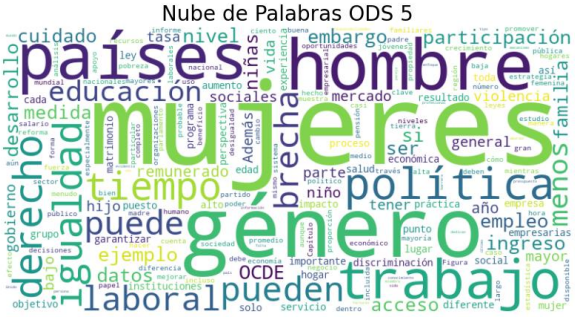
ODS 5: Igualdad de Género

Palabras clave: Mujeres, hombres, género, igualdad, trabajo, política, derechos, brecha, participación.

Bigramas clave: Hombres mujeres, brecha género, igualdad género, mujeres empresarias, mercado laboral.

Los términos como mujeres, hombres y género reflejan preocupaciones sobre la igualdad de género y la necesidad de cerrar la brecha de género, especialmente en el ámbito laboral. El bigrama hombres mujeres sugiere que las comparaciones entre ambos géneros en términos de acceso a oportunidades y derechos son un tema de debate central. Mujeres empresarias destaca que las mujeres enfrentan barreras para participar plenamente en el sector empresarial. Además, igualdad género y mercado laboral refuerzan la necesidad de crear políticas inclusivas que promuevan la equidad en el empleo y los salarios.

Nube de Palabras ODS 5



Nube de Bigramas ODS 5

Estrategias Sugeridas:

1. Reducir la brecha de género en el mercado laboral: Dado que brecha género e igualdad género son temas repetitivos, la organización debe abogar por políticas que promuevan la igualdad salarial y el acceso equitativo a oportunidades de liderazgo para las mujeres.
2. Apoyar a las mujeres en el emprendimiento: Con la mención frecuente de mujeres empresarias, la organización podría desarrollar programas que faciliten

el acceso a financiamiento y formación para mujeres que deseen iniciar o hacer crecer sus negocios.

3. Promover la participación política de las mujeres: La palabra política sugiere que los ciudadanos perciben una falta de representación femenina en la toma de decisiones. La organización debería impulsar programas que promuevan la participación política de las mujeres.

El análisis de los bigramas refuerza la necesidad de eliminar las barreras de género en el trabajo y el emprendimiento. Abordar la brecha de género y promover la igualdad en el mercado laboral contribuirá directamente al cumplimiento del ODS 5, mejorando la equidad de género en todos los niveles.

Conclusiones generales:

El análisis de palabras y bigramas para los ODS 3, 4 y 5 revela áreas clave de intervención que la organización puede abordar para mejorar el bienestar de la población y contribuir al logro de los Objetivos de Desarrollo Sostenible.

ODS 3: Las áreas prioritarias incluyen mejorar la atención primaria, fortalecer los programas de salud mental y gestionar mejor las enfermedades crónicas.

ODS 4: Es crucial mejorar la calidad de la educación en todos los niveles, capacitar a docentes y líderes escolares, y desarrollar sistemas de evaluación más efectivos.

ODS 5: Se deben reducir las desigualdades de género en el mercado laboral y apoyar a las mujeres empresarias, además de promover su participación política.

c. Datos de prueba con categoría asignada

En el siguiente link se encuentran los datos generados por el modelo Random Forest, modelo que seleccionado por su desempeño.

Link: Resultados de archivo Test