

**Integrantes:**

Jaime Torres – 202014866

María Luisa Rodríguez - 202121549

Juan Guillermo Gutiérrez - 202122659

**Link al Repositorio de GitHub:**

<https://github.com/BI-DATASCIENCE/Proyecto-1.git>

**Entendimiento del Negocio y Enfoque Analítico****1. Oportunidad/problema Negocio:**

El problema principal es la necesidad de procesar y analizar eficientemente grandes volúmenes de opiniones ciudadanas relacionadas con los ODS 3 (Salud y Bienestar), ODS 4 (Educación de Calidad) y ODS 5 (Igualdad de Género). Actualmente, este proceso consume muchos humanos y es muy lento. La oportunidad sería automatizar este análisis mediante un modelo analítico que clasifique las opiniones ciudadanas, optimizando el uso de recursos y mejore la toma de decisiones basadas en datos.

**2. Objetivos y criterios de éxito desde el punto de vista del negocio.:**

El objetivo principal sería desarrollar un modelo que permita clasificar automáticamente las opiniones ciudadanas y asociarlas a los ODS 3, 4 y 5.

Por otro lado, los criterios de éxito serían los siguientes:

- Reducción significativa del tiempo y recursos que toma analizar grandes volúmenes de datos de texto.
- Precisión y consistencia en la clasificación de opiniones de los ODS.
- Capacidad del sistema para identificar palabras clave que vinculen directamente las opiniones a un ODS específico.
- Integración del modelo en una plataforma web o móvil que permita alimentar el sistema con nuevos datos.

**3. Organización y rol dentro de ella que se beneficia con la oportunidad definida:**

Los beneficiarios principales serían los siguientes:

- Fondo de Poblaciones de las Naciones Unidas (UNFPA):
  - o Al automatizar el análisis de opiniones ciudadanas, pueden asignar recursos de manera más eficiente y obtener información clave sobre problemas sociales.
- Entidades públicas y tomadores de decisiones en Colombia:
  - o Al obtener datos estructurados y análisis rápidos, pueden desarrollar políticas más enfocadas y basadas en las preocupaciones y necesidades de la población, relacionadas con los ODS 3, 4 y 5.
- Ciudadanos y comunidades:
  - o El análisis eficiente de sus opiniones puede generar una respuesta más ágil y pertinente a sus necesidades por parte de las autoridades.

#### 4. Impacto que puede tener en Colombia este proyecto:

El impacto principal estaría relacionado con la mejora de las políticas públicas y la gestión de recursos para abordar los problemas sociales más urgentes vinculados a la salud, la educación y la igualdad de género. El proyecto permitirá identificar con precisión las necesidades de la población, empoderando a los ciudadanos al demostrar que sus opiniones se analizan y se consideran para tomar decisiones.

5. Enfoque analítico: Descripción de la categoría de análisis (descriptivo, predictivo, etc.), tipo y tarea de aprendizaje e incluya las técnicas y algoritmos que propone utilizar.

Descripción de la categoría de análisis: El análisis es predictivo y clasificatorio, ya que el modelo debe predecir a cuál de los ODS (3, 4, 5) se asocia cada opinión ciudadana. Además, tiene un componente descriptivo cuando se identifican las palabras clave que llevan a esta clasificación.

Tipo y tarea de aprendizaje: Esto sería un ejemplo de aprendizaje supervisado, ya que existe un conjunto de opiniones ya clasificadas por expertos en función de los ODS, el modelo debe aprender a partir de estos datos etiquetados. El modelo debe categorizar cada opinión en una de las tres categorías (ODS 3, 4 o 5), esto al identificar las palabras y frases que están más asociadas a cada ODS para poder explicar las predicciones.

Técnicas y algoritmos propuestos:

- K-Nearest Neighbor (KNN): Es un algoritmo de clasificación que asigna una clase a un punto nuevo basado en las clases de sus vecinos más cercanos en el espacio de características.

- Random Forest: Es un modelo de conjunto que crea múltiples árboles de decisión y los combina para mejorar la precisión y reducir el riesgo de sobreajuste.
- Multi-layer Perceptron (MLP): Es un tipo de red neuronal artificial compuesta por capas de neuronas que pueden aprender relaciones no lineales complejas entre las entradas y las salidas.

#### Mapa de actores relacionado con el producto de datos creado:

Rol en la Empresa	Tipo de Actor	Beneficio	Riesgo
Dirección de análisis y políticas públicas del UNFPA	Usuario/Cliente	El modelo facilita la identificación eficiente de problemas sociales emergentes en las áreas de salud, educación e igualdad de género, optimizando la asignación de recursos y esfuerzos para abordar esos problemas.	Si el modelo no tiene buen desempeño, puede generar políticas públicas basadas en análisis erróneos, lo que llevaría a decisiones incorrectas que no reflejan las verdaderas necesidades de la población.
Equipo de TI del UNFPA	Proveedor	Aseguran que el modelo se integre de manera adecuada en las plataformas tecnológicas (web o móvil).	Si la integración tecnológica no cumple con los estándares de seguridad, podría haber vulnerabilidades en el manejo de datos sensibles o comprometer la eficiencia del sistema.
Entidades públicas (ministerios de salud, educación, etc.)	Usuario/Cliente	Obtienen información rápida y precisa sobre las necesidades de la población, lo que les permite crear políticas más informadas centradas en los ODS 3, 4 y 5.	Si el modelo no refleja adecuadamente las opiniones ciudadanas, podrían crear políticas públicas que no atiendan de manera efectiva las verdaderas necesidades de los ciudadanos.
Financiadores externos	Financiador	Pueden respaldar un proyecto innovador que mejora la eficiencia en la recopilación y análisis de datos sociales.	Si el proyecto no tiene éxito, podría afectar futuras colaboraciones o inversiones.

Ciudadanos	Beneficiado	Se benefician de una respuesta más ágil y precisa por parte del gobierno y de otras instituciones respecto a los problemas que expresan.	Si el modelo malinterpreta las opiniones, los problemas prioritarios de la población podrían no ser atendidos adecuadamente.
------------	-------------	--	--

### **Trabajo en Equipo:**

Líder de negocio: Juan Guillermo Gutiérrez (33.33 puntos)

Es responsable de velar por resolver el problema o la oportunidad identificada y estar alineado con la estrategia del negocio para el cual se plantea el proyecto. Debe garantizar que el producto se puede comunicar de forma apropiada.

Líder de datos: Jaime Torres (33.33 puntos)

Se encarga de gestionar los datos que se van a usar en el proyecto y de las asignaciones de tareas sobre datos. Debe dejarlos disponibles para todo el grupo y garantizar la entrega en el repositorio de git.

Líder de analítica: María Luisa Rodríguez (33.33 puntos)

Se encarga de gestionar las tareas de analítica del grupo. Se encarga de verificar que los entregables cumplen con los estándares de análisis y que se tiene el “mejor modelo” según las restricciones existentes.

### **Retos enfrentados en el proyecto:**

En la etapa de vectorización, hubo algunos problemas para la representación de datos. Por ejemplo, el hecho de que varios de estos modelos no estaban pensados para trabajar con strings, por lo que nos tomó trabajo representar a estos últimos de manera numérica.

### **Duración del proyecto:**

Por ahí unas 18 horas (no las medimos con cronómetro).