Inteligencia de Negocios

Integrantes:

- Santiago Páez Márquez 202014644 – Modelo Naive Bayes
- Luis Ángel Angarita Coba 201910393 – Modelo Regresión Logística
- David Mateo Barbosa Monsalve 202110756 – Modelo Random Forest

Proyecto 1 – Etapa 1

Entendimiento de negocio y enfoque analítico

Oportunidad/problema Negocio	El problema principal es la necesidad de analizar grandes volúmenes de opiniones ciudadanas sobre temas relacionados con los ODS 3 (Salud), 4 (Educación) y 5 (Igualdad de género), sin depender de expertos humanos para hacerlo manualmente. Actualmente, este proceso consume mucho tiempo y recursos, y el Fondo de Poblaciones de las Naciones Unidas (UNFPA) busca una solución automatizada que pueda clasificar de forma precisa y rápida estas opiniones para identificar problemas clave y evaluar soluciones más eficientemente.
Objetivos y criterios de éxito desde el punto de vista del negocio.	El principal objetivo es automatizar la clasificación de opiniones ciudadanas en relación con los ODS 3, 4 y 5, utilizando técnicas de procesamiento de lenguaje natural (NLP) y aprendizaje automático. El éxito se medirá en función de la precisión del modelo, idealmente alcanzando al menos un 90% de precisión en la clasificación. Un criterio de éxito adicional sería la capacidad del modelo de manejar grandes volúmenes de datos sin perder eficiencia, permitiendo su reentrenamiento periódico para mantenerse actualizado.
Organización y rol dentro de ella que se beneficia con la oportunidad definida	La organización principal beneficiada es el Fondo de Poblaciones de las Naciones Unidas (UNFPA), que utiliza los resultados del análisis para identificar problemas relacionados con la salud, la educación y la igualdad de género. Otros beneficiarios incluyen entidades públicas, analistas de políticas y tomadores de decisiones gubernamentales, quienes utilizarán los resultados para diseñar e implementar políticas que mejoren la calidad de vida en las comunidades colombianas.
Impacto que puede tener en Colombia este proyecto.	Este proyecto puede tener un impacto significativo en Colombia al mejorar la capacidad de las organizaciones públicas y el UNFPA para entender los problemas que enfrentan los ciudadanos en relación con los ODS 3, 4 y 5. Al automatizar el análisis de las opiniones, se reducirá el tiempo de respuesta para identificar las necesidades prioritarias, lo que permitirá una toma de decisiones más rápida y precisa. Esto puede conducir a políticas más efectivas y centradas en las áreas más críticas para el desarrollo sostenible en el país, especialmente en salud, educación y equidad de género.

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.	El enfoque es predictivo, basado en una tarea de clasificación supervisada para asignar las opiniones ciudadanas a los ODS 3, 4 o 5. Las técnicas utilizadas incluyen Bag of Words y TF-IDF para vectorizar el texto, seguidas de la aplicación de algoritmos de aprendizaje automático como Naive Bayes y Regresión Logística. El proceso de optimización se realiza mediante GridSearchCV para encontrar los mejores hiperparámetros, asegurando que los modelos ofrezcan la mayor precisión posible en la clasificación.

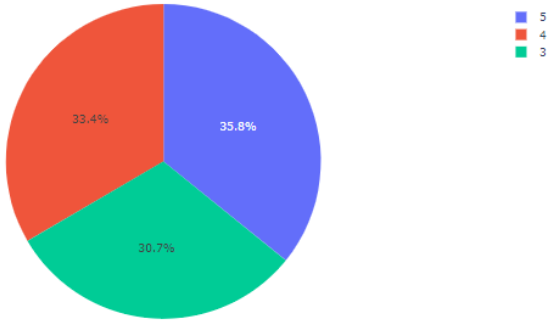
Entendimiento y preparación de datos

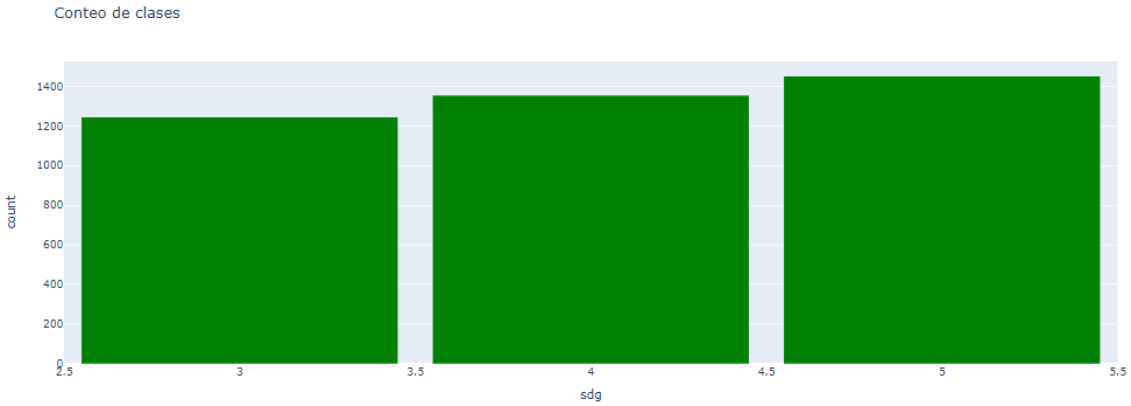
Perfilamiento de los datos

Se analizará el comportamiento de los datos sin modificarlos para luego procesarlos por el algoritmo. En primer lugar, los datos tienen una dimensión de (4049, 2). Lo anterior nos dice que se trabaja con 4049 textos las cuales tiene una clasificación de Objetivos de Desarrollo Sostenible (ODS) entre 3 y 5. Los datos no muestran ningún tipo de error de sintaxis ni textos repetidos por lo que no se verá tan afectado la implementación de los modelos.

Respecto a cada ODS se debe destacar las siguientes cifras y visualizaciones:

```
Conteo de clases:
 sdg
5    1451
4    1354
3    1244
```

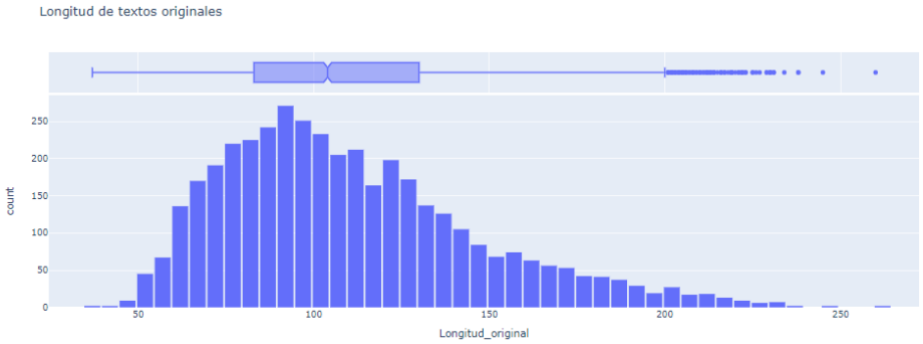




De las anteriores gráficas y datos podemos afirmar que la proporción de los ODS es bastante pareja siendo ODS 5 la de mayor cantidad seguido de ODS 4 y ODS 3 respectivamente.

Ahora se puede analizar la longitud de los textos que nos da las siguientes estadísticas:

	Longitud_original
count	4049.000000
mean	110.294147
std	36.291055
min	37.000000
25%	83.000000
50%	104.000000
75%	130.000000
max	260.000000



Se puede ver que la longitud de los textos sigue aparentemente una distribución normal por la forma y que existen muchos outliers o longitudes que se alejan bastante del promedio. Para saber a qué ODS afecta más podemos verlo por clase:

sdg	mean	max	min
3	109.823151	238	38
4	106.351551	260	48
5	114.376981	245	37

Se evidencia que el ODS 5 tiene el promedio más grande que las demás y es entendible ya que para designar una clasificación alta pueda requerir textos más largos y descriptivos. Por otro lado, el ODS 3 tiene un promedio menor lo cual nos dice que no se requiere tanto texto para poder diferenciarlo de las demás.

Análisis de la calidad de los datos

Unicidad: Se verificó la presencia de filas duplicadas en el conjunto de datos, basándose en las columnas Textos_espanol y sdg, y no se encontraron duplicados.

Completitud: No se encontraron valores nulos en las columnas Textos_espanol y sdg, ni textos o etiquetas vacíos.

Consistencia: Se comprobó que los valores en sdg estuvieran dentro del rango permitido (3, 4 o 5), y que los valores en Textos_espanol fueran cadenas de texto válidas. No se encontraron valores fuera de rango o que no fueran cadenas de texto.

Validez: Se verificó que todos los valores en Textos_espanol fueran cadenas de texto y que todos los valores en sdg fueran enteros. El análisis reveló que todos los valores son válidos, sin registros no textuales en Textos_espanol ni valores no enteros en sdg.

Luego, sigue la preparación de datos de cada algoritmo dependiendo de los parámetros necesarios.

Preparación de los datos (Naives Bayes y Logistic Regression)

Limpieza de textos: Se realizó una limpieza de los textos para eliminar palabras de relleno y caracteres irrelevantes. Esto incluyó la eliminación de enlaces, menciones, hashtags, puntuación, y caracteres especiales. Además, se eliminaron las stopwords (palabras comunes que no aportan valor analítico) y se aplicó lemmatización, es decir, la reducción de las palabras a su forma base. Posteriormente, se comparó la longitud de los textos antes y después de la limpieza, observando una reducción en la longitud de los textos, aunque algunos aún se encuentran alejados del promedio.

Inicialización de Bag of Words (BoW): Se utilizó la técnica Bag of Words (BoW) para convertir los textos en vectores numéricos basados en la frecuencia de palabras. Esto crea un vocabulario único con las palabras que aparecen en los textos y representa cada documento como un vector que indica cuántas veces aparece cada palabra. BoW no tiene en cuenta el orden de las palabras ni el contexto.

TF-IDF: Adicionalmente, se empleó la técnica TF-IDF para ponderar la importancia de cada palabra dentro de un documento en función de su frecuencia y rareza en el corpus. TF-IDF convierte los textos en vectores numéricos más adecuados para el análisis, reflejando la relevancia de las palabras en el documento específico en comparación con todo el conjunto.

Preparación de Datos para Random Forest

Limpieza de textos: Se aplicó un proceso de limpieza que incluyó la eliminación de acentos, signos de puntuación y números adjuntos a palabras, asegurando que el texto estuviera en un formato limpio y uniforme para el análisis. El texto también se transformó a minúsculas para evitar diferencias entre mayúsculas y minúsculas.

Tokenización: Después de la limpieza, los textos fueron tokenizados, es decir, se dividieron en palabras individuales. Se eliminaron las stopwords en español, que son palabras comunes que no aportan valor en el análisis.

Selección de las palabras más comunes: Se seleccionaron las 100 más comunes en el conjunto de datos usando la frecuencia de aparición de los tokens, para reducir el vocabulario y mejorar la eficiencia del modelo.

Generación de variables: Para cada palabra seleccionada, se añadió una columna al conjunto de datos que indicaba si la palabra estaba presente en cada texto. Esto permitió que el modelo Random Forest pudiera utilizar las palabras más frecuentes como características para el entrenamiento.

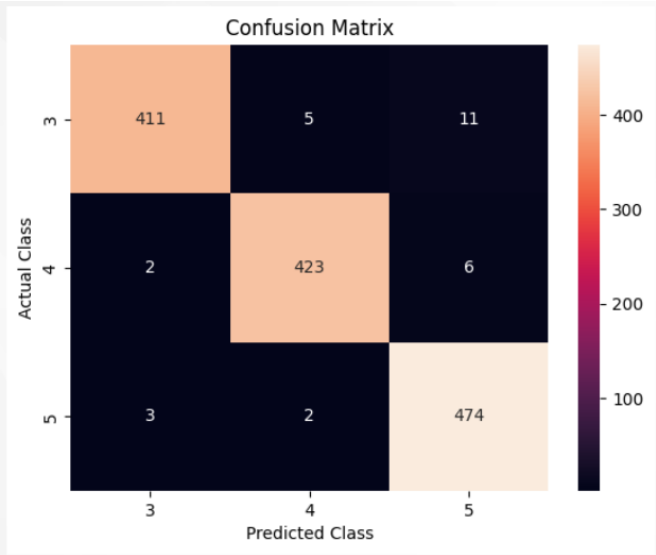
Modelado y evaluación

Algoritmo 1 – Naive Bayes – Santiago Paez

Después de la preparacion de los datos, se aplicó el modelo que se enfoca en calcular los hiperparametros de los datos para ejecutar el modelo y clasificar las opiniones de los

habitantes por ODS. Se obtuvo una precisión del 98% lo que nos dice que es un desempeño bastante alto. Por otro lado, se mostró en la matriz de confusión que casi todos los datos fueron asignados de manera correcta en cada clase.

En cuanto a las medidas de desempeño, se puede apreciar tanto la precisión como el recall y el f1 en 98% en todo el modelo lo cual nos dice que el modelo es bastante eficaz sobre todo en el ODS 4 que tiene valores mayores.

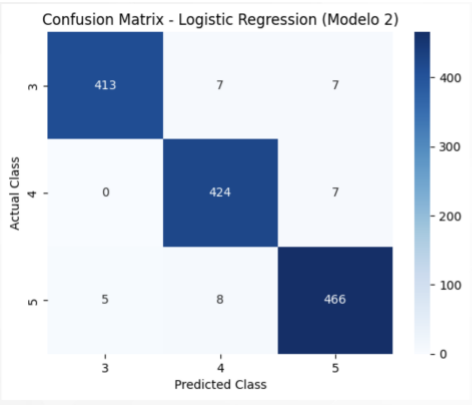


	Precision	Recall	F1
3	0.99	0.96	0.98
4	0.98	0.98	0.98
5	0.97	0.99	0.98
Macro avg	0.98	0.98	0.98

Algoritmo 2 – Logistic Regression – Angel Angarita

La regresión logística es un método estadístico utilizado para predecir variables categóricas, en este caso, la clasificación de opiniones ciudadanas en tres clases: ODS 3, ODS 4 y ODS 5. Utiliza la función logística para modelar la relación entre las características del texto y la probabilidad de pertenecer a una de estas categorías. El modelo fue entrenado y probado, obteniendo una precisión global del 97%, lo que indica un buen desempeño. La matriz de confusión mostró que la mayoría de las opiniones fueron correctamente clasificadas en su categoría correspondiente, con leves confusiones entre las clases ODS 3 y 5, así como entre ODS 4 y 5.

En cuanto a las métricas de evaluación, se alcanzaron altos valores de precisión, recall y f1-score, todos cercanos al 97% en las tres clases. Estos resultados sugieren que el modelo es eficaz para distinguir entre las diferentes opiniones, siendo especialmente preciso en la predicción de la clase ODS 3.



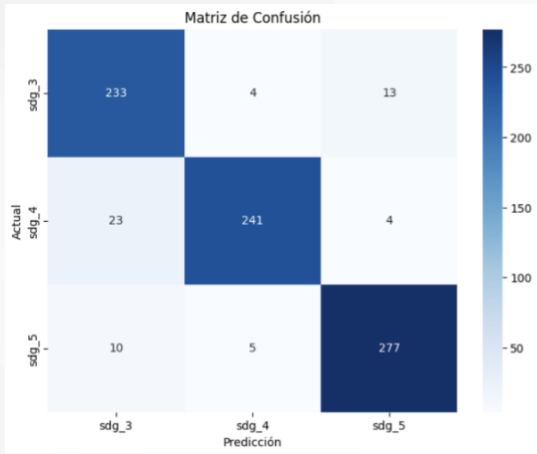
	Precision	Recall	F1
3	0.99	0.97	0.98
4	0.97	0.98	0.97
5	0.97	0.97	0.97
Macro avg	0.97	0.97	0.97

Algoritmo 3 – Random Forest – David Mateo Barbosa Monsalve

El algoritmo Random Forest es una técnica de clasificación que utiliza múltiples árboles de decisión para tomar decisiones basadas en la votación mayoritaria de los árboles. En este caso, fue utilizado para clasificar las opiniones en tres categorías: ODS 3, ODS 4 y ODS 5. Al combinar varios árboles, el modelo se vuelve más robusto ante el sobreajuste y logra una clasificación más precisa.

El modelo fue entrenado y probado, obteniendo una precisión global del 94%, lo que demuestra un buen rendimiento en la clasificación de opiniones. La matriz de confusión muestra que las opiniones fueron en su mayoría correctamente clasificadas, aunque se presentan ligeras confusiones entre las clases ODS 3 y ODS 5, así como entre ODS 4 y ODS 5.

En cuanto a las métricas de evaluación, se alcanzaron altos valores de precisión, recall y f1-score, con valores entre el 92% y el 96% en las tres clases. Estos resultados destacan la eficacia del modelo para diferenciar las opiniones entre los diferentes ODS, siendo especialmente fuerte en la predicción de ODS 5, con un f1-score del 95%.



	Precision	Recall	F1
3	0.92	0.92	0.92
4	0.96	0.90	0.93
5	0.94	0.95	0.95
Macro avg	0.94	0.92	0.93

Resultados

Para el análisis de resultados y definición de estrategias, se decidió que el modelo sobre el cual se hará es Naive Bayes por tener las métricas más altas. Sin embargo, podemos apreciar resultados de otros modelos para poder complementar los resultados.

Análisis de palabras

```
Palabra:educacin, Log Probabilidad: -4.96122043898301
Palabra:estudiantes, Log Probabilidad: -5.059736132910661
Palabra:escuelas, Log Probabilidad: -5.187100461267654
Palabra:aprendizaje, Log Probabilidad: -5.4256519732814406
Palabra:docentes, Log Probabilidad: -5.537313924633853
Palabra:alumnos, Log Probabilidad: -5.5868106955718275
Palabra:evaluacin, Log Probabilidad: -5.627317605197717
Palabra:escuela, Log Probabilidad: -5.7158727246012555
Palabra:profesores, Log Probabilidad: -5.819295913113921
Palabra:ocde, Log Probabilidad: -5.820406362700293
```

```
Palabra: salud, Coeficiente: 4.77362632457497
Palabra: mujeres, Coeficiente: -3.2479169201563423
Palabra: atencin, Coeficiente: 3.1435811111801915
Palabra: educacin, Coeficiente: -2.9588951972262327
Palabra: gnero, Coeficiente: -2.880441547535671
Palabra: pacientes, Coeficiente: 2.22097995801211
Palabra: enfermedades, Coeficiente: 1.94023394988994
Palabra: estudiantes, Coeficiente: -1.851485965157678
Palabra: mdicos, Coeficiente: 1.819435213133996
Palabra: mortalidad, Coeficiente: 1.7363663276969703
```

Se puede ver que las palabras más repetidas están relacionadas con el contenido de cada ODS. En primer lugar, tenemos palabras como “salud”,” enfermedades”,” pacientes”, “médicos”, “mortalidad” lo denota que los problemas asociados con el ODS 3 se basan en las enfermedades y la posible mortalidad de los habitantes. Por otro lado, tenemos palabras como “educación”, “estudiantes”, “docentes", "alumnos”, “evaluacion” que evidencia que un objetivo basado en el ODS 4 es enfocarse en los actores de la educación como los docentes o estudiantes ya que de estos depende que el sistema educativo sea optimo. Por último, tenemos “genero” y “mujeres” de la ODS 5 que tal vez no es algo que tengan en enfocado como objetivo debido a que son palabras menos frecuentes a los demás ODS.

Estrategias

A partir del análisis de palabras, se puede definir estrategias dando un énfasis en aquellos problemas que más pueden afectar la población. Primero, se puede dar más relevancia a las condiciones que viven hombres y mujeres debido a la poca frecuencia de palabras en comparación con otros objetivos. También se debe hacer un énfasis de mejora en la educación especializándose en los estudiantes y profesores brindándoles herramientas para mejorar el sistema educativo. Finalmente es crucial tomar acción sobre el sistema de salud en cuanto a la calidad de tratamientos para evitar la propagación de enfermedades y posibles mortalidades.

Mapa de actores relacionado con el producto de datos creado

Rol dentro de la organización	Tipo de actor	Beneficio	Riesgo
Dirección de Políticas Públicas	Usuario-cliente	Facilita la identificación de problemas clave en salud, educación y género, permitiendo crear políticas más eficientes y enfocadas en áreas prioritarias.	Si el modelo tiene un mal desempeño, podría priorizar problemas irrelevantes y descuidar cuestiones urgentes.
Fondo de Poblaciones de las Naciones Unidas (UNFPA)	Financiador	Reduce costos operativos mediante la automatización del análisis, mejorando la eficiencia en la toma de decisiones basadas en datos.	Si el modelo no alcanza la precisión esperada, los recursos invertidos podrían no generar el impacto esperado.
Ministerio de Tecnologías de la Información y las comunicaciones	Proveedor	Asegura el cumplimiento de estándares de calidad, seguridad y privacidad en el manejo de datos ciudadanos durante el desarrollo del modelo.	El manejo incorrecto o la filtración de datos podría generar problemas legales y pérdida de confianza pública.
Ciudadanos	Beneficiado	Sus problemas relacionados con salud, educación y género pueden ser abordados más rápidamente y con mayor precisión gracias a las políticas públicas generadas a partir del modelo.	Si el modelo no funciona adecuadamente, podría ignorar problemas críticos en las comunidades más vulnerables.

Trabajo en equipo

Roles y tareas de cada integrante

Líder de proyecto: Luis Ángel

- **Tareas**: Luis estuvo a cargo de la gestión general del proyecto, definiendo fechas clave para las reuniones y asegurando que cada tarea fuera asignada de manera equitativa entre los miembros del equipo. También verificó que cada entregable fuera presentado en los tiempos acordados y gestionó la carga final del proyecto en la plataforma correspondiente.
- **Tiempo dedicado**: 27 horas.
- **Retos enfrentados**: Uno de los principales desafíos fue coordinar los tiempos de trabajo de los demás integrantes, asegurando que todos cumplieran con los plazos sin sobrecargarse. También tuvo que tomar decisiones importantes cuando no había consenso en algunas soluciones analíticas, garantizando el avance del proyecto.

Líder de negocio y Líder de analítica: Santiago

- **Tareas**: Santiago se encargó de garantizar que el enfoque analítico del proyecto estuviera alineado con la estrategia del negocio y que resolviera el problema planteado. Además, lideró la parte de análisis de datos, asegurándose de que los modelos fueran optimizados y cumplieran con los requisitos del proyecto.
- **Tiempo dedicado**: 35 horas.
- **Retos enfrentados**: Los mayores retos estuvieron relacionados con la selección y optimización del mejor modelo para los datos, enfrentándose a restricciones de tiempo y recursos computacionales.

Líder de datos: David Mateo

- **Tareas**: David gestionó todo lo relacionado con la recopilación, limpieza y preparación de los datos utilizados en el proyecto. Aseguró que los datos estuvieran correctamente organizados y disponibles para todo el equipo a través de un repositorio compartido.
- **Tiempo dedicado**: 20 horas.
- **Retos enfrentados**: El principal reto fue la limpieza de los datos, ya que algunos estaban mal codificados o incompletos, lo que dificultó su uso inmediato en los modelos analíticos.

Distribución de los puntos

Según la contribución de cada integrante y los retos enfrentados, el equipo ha decidido repartir los 100 puntos de la siguiente manera:

- **Luis Ángel**: 34 puntos.
- **Santiago**: 37 puntos.
- **David Mateo**: 29 puntos.

Reflexión y puntos para mejorar

En general, el equipo ha funcionado bien, con una buena distribución de las tareas. Sin embargo, para la próxima entrega, se sugiere mejorar la comunicación en las fases iniciales

del proyecto, para evitar retrasos en la definición de los algoritmos y la limpieza de datos. Además, se propondrán reuniones más regulares para el seguimiento de los avances.