Santiago Alonso Molina
François Morales Segura
Juan Leonardo Rangel Barrera

Proyecto Inteligencia de Negocios – Etapa 1 2024-2

1. Caso:

El Fondo de Poblaciones de las Naciones Unidas (UNFPA1) junto con entidades públicas y haciendo uso de diferentes herramientas de participación ciudadana, busca identificar problemas y evaluar soluciones actuales, relacionando la información dada por los ciudadanos con los diferentes Objetivos de Desarrollo Sostenible (ODS). Los ODS fueron adoptados por las Naciones Unidas en 2015 como un llamamiento universal para poner fin a la pobreza, proteger el planeta y garantizar que para el 2030 todas las personas disfruten de paz y prosperidad2 . En este contexto, uno de los procesos que requiere mayor esfuerzo es el análisis de la información textual recopilada, ya que consume muchos recursos, que incluyen la participación de un experto. Es así como el UNFPA quiere desarrollar un proyecto con ustedes, donde el objetivo principal es relacionar de forma automática opiniones de los ciudadanos con los ODS 3, 4 y 5. A nivel de la solución a plantear deben aplicar la metodología de desarrollo de aplicaciones analíticas para crear un modelo analítico que sea utilizado y reentrenado por medio de una aplicación web o móvil a partir de un conjunto de opiniones que contienen texto en lenguaje natural.

2. Entendimiento del negocio:

Universidad de los Andes

Santiago Alonso Molina
François Morales Segura
Juan Leonardo Rangel Barrera

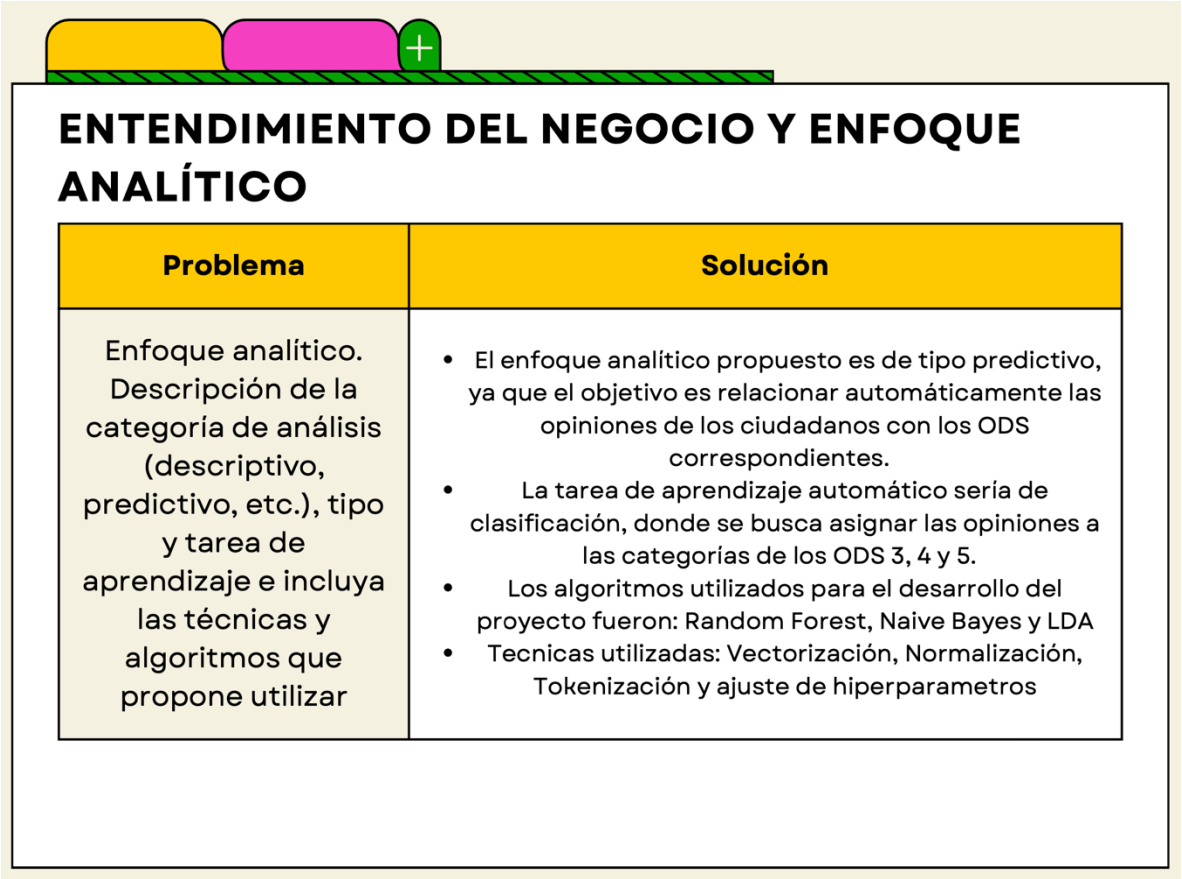
ENTENDIMIENTO DEL NEGOCIO Y ENFOQUE ANALÍTICO

Problema	Solución
Oportunidad/ problema Negocio	El UNFPA busca identificar problemas y evaluar soluciones relacionando la información dada por los ciudadanos (opiniones) con los Objetivos de Desarrollo Sostenible (ODS). Uno de los principales retos es el análisis de la información textual recopilada.
Objetivos y criterios de éxito desde el punto de vista del negocio.	El objetivo principal del proyecto es relacionar de forma automática las opiniones de los ciudadanos con los ODS 3, 4 y 5. Los criterios de éxito incluyen la capacidad del modelo analítico para identificar efectivamente las relaciones entre las opiniones y los ODS correspondientes.

ENTENDIMIENTO DEL NEGOCIO Y ENFOQUE ANALÍTICO

Problema	Solución
Organización y rol dentro de ella que se beneficia con la oportunidad definida	• La organización que se beneficia es el UNFPA (Fondo de Poblaciones de las Naciones Unidas). • El rol o función dentro del UNFPA que se beneficiaría sería el de analizar y procesar la información textual recopilada de los ciudadanos.
Impacto que puede tener en Colombia este proyecto.	El proyecto podría tener un impacto significativo a nivel nacional en Colombia, ya que permitiría al UNFPA y otras entidades públicas identificar más efectivamente los problemas y soluciones relacionados con los ODS 3, 4 y 5.

Santiago Alonso Molina
François Morales Segura
Juan Leonardo Rangel Barrera



ENTENDIMIENTO DEL NEGOCIO Y ENFOQUE ANALÍTICO

Problema	Solución
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 analítico propuesto es de tipo predictivo, ya que el objetivo es relacionar automáticamente las opiniones de los ciudadanos con los ODS correspondientes. • La tarea de aprendizaje automático sería de clasificación, donde se busca asignar las opiniones a las categorías de los ODS 3, 4 y 5. • Los algoritmos utilizados para el desarrollo del proyecto fueron: Random Forest, Naive Bayes y LDA • Tecnicas utilizadas: Vectorización, Normalización, Tokenización y ajuste de hiperparametros

3. Mapa de actores

Rol dentro de la empresa	Tipo de actor	Beneficio	Riesgo
UNFPA	Cliente	Recibe un mecanismo automatizado para relacionar opiniones ciudadanas con los ODS, facilitando la toma de decisiones sobre intervenciones necesarias.	Si el modelo es inexacto, las decisiones basadas en él podrían estar mal informadas, impactando los objetivos de desarrollo sostenible.
Entidades públicas participantes	Colaboradores	Pueden mejorar la calidad de los servicios ofrecidos al tener una mejor comprensión de los problemas ciudadanos.	Si el modelo no es eficiente, se pueden perder recursos y esfuerzos en intervenciones que no

			solucionan los problemas identificados.
Científicos de datos del equipo	Desarrolladores	Desarrollan un modelo de análisis de texto que proporciona valor, automatizando un proceso que antes dependía de la intervención manual de expertos.	Si los datos no se preparan correctamente, el modelo podría fallar, afectando su desempeño y credibilidad.
Ingenieros de datos	Soporte Técnico	Automatizan la actualización y reentrenamiento del modelo, garantizando que esté actualizado con los nuevos conjuntos de datos.	Si la aplicación no se mantiene correctamente, el modelo puede volverse obsoleto y generar resultados inexactos con el tiempo.
Habitantes locales	Beneficiarios	Sus opiniones son tomadas en cuenta para mejorar las políticas públicas alineadas con los ODS 3, 4 y 5.	Si el análisis es defectuoso, sus preocupaciones pueden no ser abordadas de manera efectiva.
Entidades encargadas de los Objetivos de Desarrollo Sostenible (ODS)	Supervisores	Utilizan los resultados del análisis para monitorear los avances y adecuar estrategias hacia el cumplimiento de los ODS.	La incorrecta interpretación de los resultados puede llevar a una falta de progreso hacia los ODS 3, 4 y 5.

4. Entendimiento y preparación de los datos.

a. Entendimiento:

Los datos presentan dos variables: una de texto en español y una variable categórica denominada sdg. La variable de texto contiene 3,239 valores únicos, lo que indica una alta diversidad y singularidad en el contenido, lo que puede representar un desafío para el análisis, ya que la identificación de patrones repetidos podría ser difícil. En cuanto a la variable categórica sdg, esta tiene tres categorías con distribuciones desiguales: el valor 5 es el más frecuente con un 36.3% de las observaciones, seguido por el valor 4 con un 33.6% y el valor 3 con un 30.1%. Aunque la variable sdg no está perfectamente balanceada, las

diferencias en las proporciones no son extremadamente marcadas, lo que sugiere que el entrenamiento de modelos predictivos no debería requerir técnicas de balanceo intensivo. La presencia de una distribución relativamente equitativa en la variable categórica y la alta diversidad en la variable de texto son factores importantes a considerar en el desarrollo de modelos analíticos y predictivos.

b. Preparación de los datos

En el proceso de limpieza de los datos, se realizaron varias operaciones para asegurar la calidad de los textos antes de proceder con su análisis. Primero, se intentó la eliminación de los caracteres no ASCII para garantizar que el texto estuviera en un formato estándar. Luego, se transformó todo el texto a minúsculas para evitar problemas de diferenciación entre mayúsculas y minúsculas. A continuación, se eliminaron los signos de puntuación y se reemplazaron los números por su forma textual para asegurar la consistencia de los datos. Finalmente, se eliminaron las palabras vacías o "stopwords" en español, con el objetivo de reducir el ruido y mantener únicamente las palabras relevantes para el análisis.

5. Evaluación de los modelos:

a. Modelo de Random-Forest (Leonardo Rangel) :

Random Forest es un algoritmo de aprendizaje automático utilizado principalmente para tareas de clasificación y regresión. Es una extensión del método de árboles de decisión, pero lo mejora al construir múltiples árboles y combinar sus resultados para obtener predicciones más robustas.
Los Hiperparametros se sitúan entre 40 y 45 para la profundidad y 100 a 103 para los caminos. Al probarse varias veces y el algoritmo determinar por si mismo los mejores hiperparametros, determinamos usar ese rango para hacer que el modelo no tarde mucho en ejecutarse.
El modelo de Random Forest muestra un excelente rendimiento, con una exactitud del 100% en el conjunto de entrenamiento y del 97% en el conjunto de prueba, lo que sugiere un posible sobreajuste leve. Las métricas de clasificación son muy altas, con precisiones entre 97% y 99%, un recall entre 97% y 98%, y un

F1-Score que oscila entre 97% y 98%, indicando un buen equilibrio entre precisión y sensibilidad. Las variables más importantes para la predicción son "Mujeres" (0.086480), "Género" (0.042242), "Salud" (0.029701), "Estudiantes" (0.025426) y "Educación" (0.023537). El modelo tiene una optima generalización, pero habría que tener cuidado por si hay un sobreajuste, al ver que hay 100% en exactitud con los datos de entrenamiento.

b. **Modelo de Naive Bayes (François Morales):**

Naive Bayes es un algortimo de aprendizaje supervisado popular, el cual aplica el teorema de Bayes con la suposición ingenua de las probabilidades de las caracteristicas para un conjunto de datos.

La construcción del modelo se llevó a cabo en dos fases principales. En la primera iteración, se desarrolló un modelo base utilizando Naive Bayes sin realizar ningún ajuste en los hiperparámetros. Este enfoque inicial permitió obtener un conjunto de resultados que sirvieron como referencia para evaluar el rendimiento del modelo en su estado más simple. Posteriormente, en la segunda iteración, se implementó un proceso de ajuste de hiperparámetros con el fin de optimizar el modelo y mejorar su precisión. El objetivo era identificar posibles mejoras en el rendimiento del modelo ajustado en comparación con el modelo base inicial, evaluando si los cambios en los hiperparámetros lograban un aumento en la capacidad predictiva del modelo.

Entrenamiento: 98.80%

Prueba: 95.56%

Estos resultados indican un muy buen desempeño con una ligera disminución en la exactitud al pasar de entrenamiento a prueba, lo que sugiere una buena generalización sin un alto riesgo de sobreajuste.

Reporte de Clasificación:

Precisión: Entre 92% y 100% en las diferentes clases.

Recall: Entre 93% y 97%.

F1-Score: Valores que oscilan entre 95% y 97%.

Este reporte muestra un balance adecuado entre precisión y recall, con un buen desempeño en todas las clases. Este nuevo modelo presenta un buen

rendimiento, con una exactitud del 98.80% en el entrenamiento y del 95.56% en la prueba, lo que sugiere que generaliza bien sin indicios de sobreajuste significativo. Las métricas de precisión, recall y F1-Score muestran un rendimiento equilibrado y consistente entre las clases, lo que indica que el modelo tiene un buen desempeño en todas las categorías analizadas.

c. Modelo Linear Discriminant (Santiago Molina):

El Linear Discriminant Analysis (LDA) es un algoritmo de clasificación supervisada que se utiliza principalmente para separar clases linealmente. LDA crea un único espacio proyectado, donde busca maximizar la separación entre las clases.
El modelo LDA presenta una exactitud del 89% en el conjunto de entrenamiento y del 74% en el conjunto de prueba, lo que indica dificultades en la generalización. Las métricas de precisión, recall y F1-Score oscilan entre el 71% y el 78%, mostrando un rendimiento aceptable, aunque la clase 3 tiene un recall bajo del 65%. La matriz de confusión revela que el modelo confunde frecuentemente la clase 3 con las clases 4 y 5. En general, el modelo podría beneficiarse de ajustes adicionales para mejorar su capacidad predictiva, especialmente en la clase 3.

6. Trabajo en Equipo - División de roles y tareas

Leonardo Rangel:

- **Roles:**
  - **Lider de proyecto:** A cargo de la gestión del proyecto. Define las fechas de reuniones, pre-entregables del grupo y verifica las asignaciones de tareas para que la carga sea equitativa. Es el encargado de subir la entrega del grupo.
  - **Lider de negocio:** Es responsable de velar por resolver el problema y estar alineado con la estrategia del negocio para el cual se plantea el proyecto. Encargado de garantizar que el producto se puede comunicar de forma apropiada.
- **Tareas:**
  - Realización del enfoque analitico, entendimineto, perfilamiento y calidad de los datos, preparación y limpieza a la para con los demás miembros del grupo

    - Implementación del modelo utilizando Random Forest
    - Analisis de resultados y mapa de actores a la par con los demás miembros del grupo
    - Presentación al negocio
- **Algoritmo:** Random Forest
- **Retos:** Para este algoritmo se me complico la búsqueda de los hiperparametros porque inicie con un rango amplio y random forest tiene un tiempo de ejecución bastante alto.
- **Puntos:** 33,33 puntos
- **Horas trabajadas en el proyecto:** 15

Santiago Molina:

- **Roles:**
    - **Líder de datos:** Se encarga de gestionar los datos que se van a usar en el proyecto y de las asignaciones de tareas sobre datos. Dejandolos disponibles para todo el grupo y garantizando la entrega en el repositorio de git.
- **Tareas:**
    - Realización del enfoque analitico, entendimineto, perfilamiento y calidad de los datos, preparación y limpieza a la para con los demás miembros del grupo
    - Implementación del modelo utilizando
    - Analisis de resultados y mapa de actores a la par con los demás miembros del grupo
    - Presentación al negocio
- **Algoritmo:**
- **Retos:** Se me dificultó mucho el hecho de vectorizar los datos para poder aplicar el algoritmo.
- **Puntos:** 33,33 puntos
- **Horas trabajadas en el proyecto:** 15

François Morales:

- **Roles:**

    - **Líder de analítica:** 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.

- **Tareas:**
    - Realización del enfoque analitico, entendimineto, perfilamiento y calidad de los datos, preparación y limpieza a la para con los demás miembros del grupo
    - Implementación del modelo utilizando Niave Bayes
    - Analisis de resultados y mapa de actores a la par con los demás miembros del grupo
    - Presentación al negocio
- **Algoritmo:** Naive Bayes
- **Retos:** Uno de los mayores retos durante la realización del modelo fue encontrar los hiperparámetros adecuados que permitieran mejorar el desempeño del modelo. Este proceso fue algo complicado ya que no existe una fórmula única para determinar los mejores valores, sino que depende del tipo de datos y de las características del modelo.

  Para superar este reto, realicé una investigación acerca de las mejores prácticas para ajustar hiperparámetros en modelos Naive Bayes. Consulté diversas fuentes e investigué enfoques como la búsqueda en cuadrícula (grid search) y la búsqueda aleatoria (random search). A partir de esta investigación, implementé técnicas de ajuste que me permitieron explorar diferentes combinaciones de hiperparámetros y, finalmente, encontrar una configuración que mejoró el rendimiento del modelo.
- **Puntos:** 33,33 puntos
- **Horas trabajadas en el proyecto:** 15

Reuniones para el proyecto:

- **Reunión de lanzamiento y planeación:** Definición de roles y forma de trabajo del grupo. Se generó una lluvia de ideas sobre la forma de resolver el proyecto. Se desarrolló una reunión de lanzamiento al inicio de la semana.

- **Reunión de ideación:** Reunión para definir la organización/empresa/institución y el rol dentro de ella y la solución analítica que van a desarrollar. Se desarrolló una reunión de ideación al inicio de la semana.
- **Reuniones de seguimiento:** Reunión para definir el avance según lo planeado. Se realizaron 6 reuniones de seguimiento, una diaria durante la semana de realización del proyecto.
- **Reunión de finalización:** Para consolidar el trabajo final, verificar el trabajo del grupo y analizar los puntos a mejorar para la siguiente etapa del proyecto. Se realizó una reunión de finalización el ultimo día de entrega del proyecto.