Proyecto 1 – INTELIGENCIA DE NEGOCIOS
Etapa 1 – Clasificación de analítica de textos

PRESENTADO POR:

Paula Daza Díaz – 202111276
Sofía Torres Ramírez – 202014872
Juan Camilo Reyes - 201922989

15 de Octubre del 2023

PROFESOR:

Fabián Peña Lozano

UNIVERSIDAD DE LOS ANDES
DPTO. INGENIERÍA DE SISTEMAS Y COMPUTACIÓN
INTELIGENCIA DE NEGOCIOS
BOGOTÁ D.C
2023

Tabla de contenido

Es esencial llevar a cabo un proceso metódico para comprender plenamente el negocio y definir el enfoque analítico para el proyecto asignado. Se llevarán a cabo las siguientes fases para cumplir con esto:

1. **Entendimiento del negocio y enfoque analítico**

Se inicia por comprender la relevancia y el contexto de los Objetivos de Desarrollo Sostenible (ODS) y su importancia en el ámbito global.

- *Definición de ODS*

Los Objetivos de Desarrollo Sostenible (ODS) son un conjunto de 17 objetivos globales adoptados por la Asamblea General de las Naciones Unidas en septiembre de 2015. Estos objetivos se han establecido como un llamado universal a la acción para poner fin a la pobreza, proteger el planeta y garantizar que todas las personas gocen de paz y prosperidad para el año 2030.

- *Definición de ODS Específicos en el Proyecto (ODS 3, 4 y 5)*

Objetivo 3: Garantizar una vida sana y promover el bienestar para todos en todas las edades. Algunas metas de este objetivo son:

3.1 Para 2030, reducir la tasa mundial de mortalidad materna a menos de 70 por cada 100.000 nacidos vivos.

3.2 Fortalecer la prevención y el tratamiento del abuso de sustancias adictivas, incluido el uso indebido de estupefacientes y el consumo nocivo de alcohol.

Objetivo 4: Garantizar una educación inclusiva, equitativa y de calidad y promover oportunidades de aprendizaje durante toda la vida para todos. Algunas metas de este objetivo son:

4.1 De aquí a 2030, asegurar que todas las niñas y todos los niños terminen la enseñanza primaria y secundaria, que ha de ser gratuita, equitativa y de calidad y producir resultados de aprendizaje pertinentes y efectivos.

4.2 De aquí a 2030, asegurar que todas las niñas y todos los niños tengan acceso a servicios de atención y desarrollo en la primera infancia y educación preescolar de calidad, a fin de que estén preparados para la enseñanza primaria.

Objetivo 5: Lograr la igualdad entre los géneros y empoderar a todas las mujeres y las niñas. Algunas metas de este objetivo son:

5.1 Poner fin a todas las formas de discriminación contra todas las mujeres y las niñas en todo el mundo.

5.2 Eliminar todas las formas de violencia contra todas las mujeres y las niñas en los ámbitos público y privado, incluidas la trata y la explotación sexual y otros tipos de explotación.

- *Impacto en Colombia*

El cumplimiento de estos ODS en Colombia puede tener un impacto significativo en la vida de la población colombiana y en el desarrollo sostenible del país:

- **ODS 3 - Salud y Bienestar:** Lograr una vida más saludable reducirá la mortalidad materna y la mortalidad infantil, mejorando la calidad de vida de las madres y los niños colombianos.
- **ODS 4 - Educación de Calidad:** Garantizar una educación inclusiva y de calidad proporciona a los colombianos igualdad de oportunidades para el aprendizaje, lo que puede impulsar el desarrollo educativo y económico del país.
- **ODS 5 - Igualdad de Género:** Al eliminar la discriminación de género y la violencia contra las mujeres y las niñas, Colombia puede avanzar hacia una sociedad más justa y equitativa.

1.1 Definición de Objetivos del Negocio

➢ *Automatización de la Clasificación de Textos:*

- **Objetivo:** Desarrollar un modelo de clasificación de textos basado en técnicas de aprendizaje automático que permita asignar automáticamente un texto a uno de los Objetivos de Desarrollo Sostenible (ODS) específicos.

➢ *Mejora de la Eficiencia en la Evaluación de Políticas Públicas:*

- **Objetivo:** Facilitar la identificación y evaluación de políticas públicas relacionadas con los ODS mediante el análisis automatizado de opiniones y comentarios de la población local.

➢ *Identificación de Problemas y Soluciones de Manera Más Rápida y Precisa:*

- **Objetivo:** Proporcionar una herramienta que permita identificar problemas y soluciones en relación con los ODS en un contexto territorial de manera más rápida y precisa.

➢ *Criterios de Éxito:*

- **Precisión de la Clasificación de Textos:** Medir la precisión del modelo de clasificación de textos en asignar correctamente los textos a los ODS específicos. Un alto nivel de precisión es fundamental para la utilidad de la herramienta.
- **Tiempo de Respuesta en la Identificación de Problemas y Soluciones:** Medir cuánto tiempo se ahorra en la identificación de problemas y soluciones en comparación con enfoques anteriores.
- **Impacto en la Toma de Decisiones:** Evaluar cómo el proyecto contribuye a la toma de decisiones más informadas y efectivas en la implementación de políticas relacionadas con los ODS.

Oportunidad/Problema del negocio	Enfoque analítico	Organización y rol que se beneficia	Contacto con experto externo del proyecto
Automatización de la Clasificación de Textos relacionados con los ODS	Se propone utilizar técnicas de aprendizaje automático, incluyendo K-Nearest Neighbors (KNN), Árboles de Decisión y Random Forest. Se aplicará preprocesamiento de datos, que incluye tokenización, eliminación de stop words, lematización, y vectorización de texto. El	Fondo de Poblaciones de las Naciones Unidas (UNFPA)	Alberto Rueda Díaz – Ingeniero de Sistemas y Computación

	modelo se entrenará con datos etiquetados y se evaluará utilizando métricas como prEcisión, recall y F1-score.		
Mejora de la Eficiencia en la Evaluación de Políticas Públicas relacionadas con los ODS	Se utilizarán técnicas de procesamiento de lenguaje natural y análisis de sentimientos para automatizar la evaluación de opiniones y comentarios de la población local. Esto permitirá identificar problemas y soluciones más rápido y con mayor precisión.	UNFPA y otras entidades públicas involucradas en la evaluación de políticas públicas relacionadas con ODS	Alberto Rueda Díaz – Ingeniero de Sistemas y Computación
Identificación de Problemas y Soluciones relacionados con los ODS de manera eficiente	Se aplicará un modelo de clasificación de textos para identificar problemas y soluciones en un contexto territorial. Esto acelerará el proceso de identificación de desafíos y oportunidades.	UNFPA y otras entidades interesadas en el desarrollo sostenible a nivel territorial.	Alberto Rueda Díaz – Ingeniero de Sistemas y Computación

2. Entendimiento y preparación de los datos

El preprocesamiento de datos es crucial para el procesamiento de texto. Aquí se describen algunas de las tareas de preprocesamiento que se llevarán a cabo.

➢ ***Perfilamiento y Entendimiento de los Datos:***

Se inicia el proceso de perfilamiento y entendimiento de los datos mediante la lectura del archivo Excel llamado "cat_345.xlsx" utilizando la librería pandas. Los datos se almacenan en un DataFrame llamado "data_t" que contiene 3000 filas y 2 columnas.

➢ ***Limpieza de Datos:***

- Se observa que la columna "sdg" contiene tres valores (3, 4 y 5) con una distribución equitativa de un tercio en cada categoría.
- La columna "Textos_espanol" se convierte a tipo de dato string para asegurar la coherencia en el procesamiento.
- *Correción de palabras mal codificadas:* se aplica una función llamada "fix_malformed_words" que utiliza la librería "ftfy" para corregir problemas de codificación en el texto de la columna "Textos_espanol".
- *Eliminación de Caracteres Especiales y Puntuación:* Se procede a eliminar caracteres especiales y puntuación en el texto de la columna "Textos_espanol". Para esto, se utilizan las siguientes técnicas:

  - Se eliminan caracteres no ASCII para asegurar que los datos estén en formato legible.
  - Se convierten todas las palabras a minúsculas para asegurar uniformidad.
  - Se eliminan los signos de puntuación del texto.
  - Se reemplazan los números con su representación textual.
  - Se eliminarán las palabras comunes (stop words) que no aportan información significativa a la clasificación.
- *Normalización de Texto:* Se lleva a cabo la normalización del texto, que incluye:

- Uso de stemmers y lematizadores para reducir las palabras a su forma base.

- Se utiliza el stemmer Lancaster para reducir las palabras a su raíz.
- Se lematizan los verbos en función de su tipo de palabra.
- Se mapeo de etiquetas POS a las categorías de WordNet para el lematizador.

- *Tokenización*: Se realiza la tokenización del texto de la columna "Textos_espanol" después de corregir las contracciones y eliminar ruido. Las palabras se dividen en tokens, y se aplican las técnicas de limpieza mencionadas anteriormente.
- *Vectorización de Texto:* Los textos se convertirán en vectores numéricos utilizando técnicas como TF-IDF (Term Frequency-Inverse Document Frequency) para representar la importancia de las palabras en los textos.

3. Modelado y evaluación:

Para la tarea designada se crearon cinco modelos basados en árboles. Los modelos basados en árboles son una elección adecuada debido a su capacidad para manejar datos estructurados y no estructurados, como el texto, y su flexibilidad para adaptarse a diferentes configuraciones. Aquí se describen los modelos, sus hiperparámetros y se justifica su selección:

3.1 Random Forest con Parámetros por Defecto:

- Modelo: Se utilizó un Random Forest con los parámetros por defecto.
- Justificación: Este es un punto de partida sólido para la construcción de modelos. El Random Forest es conocido por su capacidad para manejar características en texto y, al utilizar los valores predeterminados, se puede evaluar la capacidad del modelo sin ajustes manuales.
- Evaluación: F1 (weighted): 0.9733
  Este modelo con parámetros por defecto muestra un rendimiento sólido, con una alta precisión, recall y puntuación F1. Esto sugiere que el modelo es capaz de predecir con precisión la pertenencia de los textos a los ODS, y tiene un buen equilibrio entre la precisión y el recall.
- Miembro encargado: Paula Daza.

3.2 Random Forest con Hiperparámetros Ajustados (1):

- Modelo: Se entrenó otro Random Forest, pero se ajustaron los hiperparámetros de la vectorización y el clasificador.
- Hiperparámetros Ajustados:
  - Vectorización: TfidfVectorizer con tokenización de palabras y eliminación de palabras vacías.
  - Clasificador: 100 estimadores, criterio 'gini' y profundidad máxima de 50.
- Justificación: Se mejoró la vectorización y se optimizaron algunos hiperparámetros clave, como el número de estimadores y la profundidad máxima, para evaluar si estos cambios mejoran el rendimiento del modelo.
- Evaluación: F1 (weighted): 0.9694

Aunque ligeramente inferior en comparación con el modelo por defecto, el modelo con hiperparámetros ajustados todavía muestra un rendimiento sólido. La ligera disminución en la precisión y el recall podría deberse a la especificidad de los hiperparámetros utilizados.

- Miembro encargado: Juan Camilo Reyes.

3.3 Random Forest con Hiperparámetros Ajustados (2):

- Modelo: Otro Random Forest con hiperparámetros específicos.
- Hiperparámetros Ajustados:
  - Vectorización: TfidfVectorizer con tokenización de palabras y eliminación de palabras vacías.
  - Clasificador: 300 estimadores, criterio 'gini' y profundidad máxima de 100.
- Justificación: Se aumentó el número de estimadores y la profundidad máxima para ver si el modelo puede beneficiarse de un mayor poder predictivo.
- Evaluación: F1 (weighted): 0.9800
- Este modelo con hiperparámetros ajustados muestra un rendimiento excelente, con una precisión, recall y puntuación F1 de alrededor del 98%. Esto sugiere que el aumento en el número de estimadores y la profundidad máxima contribuyó significativamente a mejorar la capacidad predictiva del modelo.
- Miembro encargado: Sofia Torres.

3.4 Random Forest con Hiperparámetros Ajustados (3):

- Modelo: Otra instancia de Random Forest con diferentes configuraciones.
- Hiperparámetros Ajustados:
  - Vectorización: TfidfVectorizer con tokenización de palabras y sin conversión a minúsculas.
  - Clasificador: 350 estimadores, criterio 'gini' y profundidad máxima de 100.
- Justificación: Se eliminó la conversión a minúsculas en la vectorización y se ajustaron los hiperparámetros para evaluar si la sensibilidad a las mayúsculas y una mayor complejidad del modelo influyen en el rendimiento.
- Evaluación: F1 (weighted): 0.9787
  Este modelo también muestra un rendimiento muy bueno, con una alta precisión, recall y puntuación F1. La eliminación de la conversión a minúsculas en la vectorización no parece haber afectado negativamente el rendimiento del modelo.
- Miembro encargado: Sofia Torres.

3.5 Decision Tree Classifier con Parámetros por Defecto:

- Modelo: Se utilizó un Decision Tree Classifier con los parámetros por defecto.

- Justificación: Este modelo se incluye como comparación con los Random Forest, ya que es un clasificador basado en árboles más simple y puede ayudar a determinar si un modelo más complejo es necesario en este contexto.
- Evaluación: F1 (weighted): 0.9520
  A pesar de que este modelo utiliza un árbol de decisión más simple en comparación con los Random Forest, todavía muestra un rendimiento aceptable. La precisión es ligeramente más baja que en los modelos anteriores, lo que sugiere que la complejidad adicional de los Random Forest puede ser beneficiosa en este caso.
- Miembro encargado: Paula Daza.

4. Resultados - video

Basándonos en los resultados del numeral anterior, que se refiere al modelado y evaluación, es posible concluir que, según el resultado de la métrica F1 Score, el mejor algoritmo es el tercero (Random Forest con Hiperparámetros Ajustados (2)), ya que obtuvo un puntaje de 0.98 en esta métrica. Por lo tanto, se utiliza este algoritmo para predecir los valores del dataframe. Con este resultado, es posible observar que, dado un conjunto de valores esperados de SDGs de 3, 4 y 5, el modelo predijo y clasificó correctamente las diferentes frases en estas tres categorías.
La distribución de los datos predecidos fue la siguiente:

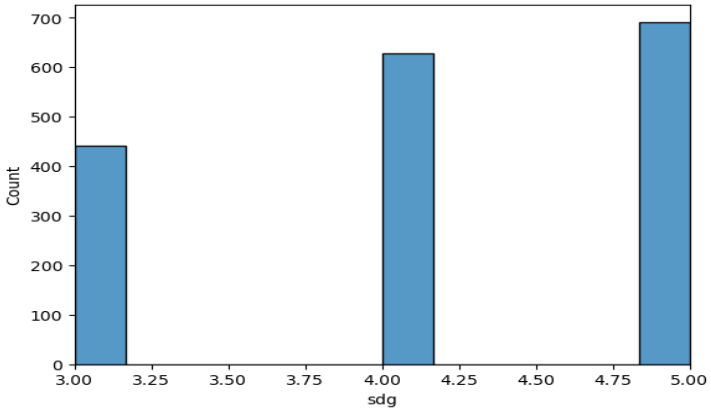

Ilustración 1: Datos predecidos utilizando el modelo 3

Dados estos resultados se le realizan las siguientes recomendaciones a la organización:

a. La cantidad de soluciones y problemáticas relacionados con el SDG 3 son menores a los relacionados con los SDG 4 y 5 por lo tanto es recomendable evaluar la razón por la cual este objetivo no ha sido trabajado.
b. Es necesario establecer un proceso de seguimiento continúo utilizando el modelo para poder evaluar los resultados a lo largo del tiempo.
c. Al momento de formular soluciones a las problemáticas es recomendable utilizar el modelo propuesto ya que este facilitara la verificación de la correlación entre la problemática y solución con el SDG a cumplir.

5. **Mapa de actores relacionado con un producto de datos creado con el modelo analítico construido.**

Rol dentro de la empresa	Tipo de actor	Beneficio	Riesgo
UNFPA	Financiador /Usuario	Logra clasificar apropiadamente las problemáticas y soluciones relacionadas con los ODS para mejorar sus mecanismos de toma de decisiones	Si no funciona correctamente el modelo la UNFPA realizara incorrectamente seguimientos y evaluaciones de políticas públicas, posiblemente estimando incorrectamente el impacto que puedan tener en la población.
Gobierno Colombiano	Proveedor /Usuario	El gobierno logra invertir menos dinero en políticas más efectivas que mejoran la calidad de vida de su población.	En el caso en que el modelo no funcione puede que las problemáticas que se intenten solucionar no estén relacionadas con el ODS que se quiere cumplir y por lo tanto causen un perjuicio a la población por el dinero mal invertido.
Población	proveedor /Beneficiado	Recibe una mejora significativa en su calidad de vida gracias a los cumplimientos de los ODS que afectan su día a día.	La falta de apoyo para mejorar la calidad de vida por la dificultad para identificar las problemáticas o soluciones relacionadas con los ODS.

6. **Trabajo en equipo:**

Plan de Trabajo

- *Roles y Tareas del Grupo:*

  1. Líder de Proyecto (LP) – Asignado a Paula Daza: Será responsable de la gestión general del proyecto, definición de fechas de reuniones, pre-entregables, y verificar la asignación equitativa de tareas. El LP subirá la entrega del grupo y tomará decisiones finales en caso de desacuerdo.
  2. Líder de Negocio (LN) - Asignado a Juan Camilo Reyes: Se encargará de alinear el proyecto con los objetivos de negocio y resolver el problema identificado. También coordinará las reuniones con expertos de estadística para revisar los resultados. Será el enlace con UNFPA.
  3. Líder de Datos (LD) – Asignado a Sofía Torres: Gestionará los datos del proyecto y asignará tareas relacionadas con la preparación de los datos. Asegurará que los datos estén disponibles para todo el grupo.
  4. Líder de Analítica (LA) – Sofía Torres y Paula Daza: Gestionará las tareas analíticas del grupo, verificará que los entregables cumplan con los estándares y seleccionará el "mejor modelo" según las restricciones.

- *Fases y Tareas:*

  Fase I: Entendimiento del Negocio y Preparación

- LP: Coordinar la reunión de lanzamiento y definir roles y forma de trabajo del grupo (5 puntos).
- LN: Describir los ODS involucrados en el proyecto asignado y su impacto en Colombia (5 puntos).
- LN: Contactar al grupo de expertos de estadística y programar la reunión de inicio (5 puntos).
- LD: Realizar el perfilamiento y análisis de calidad de los datos (5 puntos).
- LD: Realizar la limpieza y transformación de los datos (5 puntos).
- LA: Identificar el enfoque analítico para alcanzar los objetivos del negocio (5 puntos).

Fase II: Modelado y Evaluación

- LA: Aplicar al menos tres algoritmos de aprendizaje automático (10 puntos).
- LA: Evaluar cuantitativamente los resultados de los modelos (10 puntos).
- LP: Coordinar reuniones de seguimiento y asegurar el avance del proyecto (5 puntos).
- LD: Proporcionar los datos preparados para los modelos (5 puntos).
- LN: Evaluar y analizar los resultados obtenidos y proponer posibles estrategias (5 puntos).

Fase III: Resultados y Sustentación

- LP: Coordinar la reunión de finalización y análisis de puntos a mejorar (5 puntos).
- Todos los miembros: Crear un archivo csv con los resultados y asignar ODS al conjunto de datos (10 puntos).
- LD: Asegurar que los entregables estén completos y listos para la presentación (5 puntos).
- Todos los miembros: Generar un video de 5 minutos explicando el proyecto y resultados (5 puntos).
- Todos los miembros: Contribuir a la creación del mapa de actores relacionados con el producto de datos (5 puntos).

➢ *Distribución de Puntos entre los Integrantes:*

Líder de Proyecto (LP): 15 puntos
Líder de Negocio (LN): 20 puntos
Líder de Datos (LD): 20 puntos
Líder de Analítica (LA): 25 puntos

➢ *Puntos para mejorar en la siguiente entrega:*

1. Mejor coordinación en la asignación de tareas y cumplimiento de plazos.
2. Mayor comunicación y colaboración entre los líderes de cada área.
3. Evaluar la posibilidad de realizar reuniones adicionales para garantizar un seguimiento más cercano del proyecto.
4. Documentar de manera más detallada las tareas y resultados obtenidos en cada fase.