Proyecto 1 – INTELIGENCIA DE NEGOCIOS
Etapa 2 – Clasificación de analítica de textos

PRESENTADO POR:

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

28 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

1. Proceso de automatización del proceso de preparación de datos, construcción del modelo, persistencia del modelo y acceso por medio de API.

1.1 Automatización del proceso y construcción del modelo

Para llevar a cabo este proceso, realizamos un análisis de la entrega anterior con el fin de identificar las áreas que requerían modificaciones para automatizar el proceso y determinar qué componentes debían ser eliminados. Basándonos en este análisis, tomamos la decisión de emplear el Modelo Número 3, que consiste en un modelo de Random Forest con hiperparámetros personalizados configurados de la siguiente manera:

- Hiperparámetros Ajustados:
  - Vectorización: Utilizamos TfidfVectorizer con tokenización de palabras y eliminación de palabras vacías.
  - Clasificador: Configuramos 300 estimadores, el criterio de división como 'gini', y una profundidad máxima de 100.
  - Evaluación: La métrica F1 (ponderada) alcanza un impresionante valor de 0.9800.

Una vez seleccionado el modelo, decidimos incluir el proceso de limpieza como parte integral de la automatización. Para lograr esto, creamos un archivo denominado "limpieza.py", que se ejecuta como el primer paso en el pipeline. Este archivo engloba todas las operaciones básicas de limpieza que se utilizaron en la entrega anterior, como la eliminación de caracteres no ASCII, la corrección de problemas de codificación, la tokenización de palabras, la eliminación de "stop words", y la conversión de números a su equivalente en letras, entre otros.

Después de la fase de limpieza, llevamos a cabo la vectorización y el entrenamiento del modelo, siguiendo la misma metodología que se utilizó en la entrega previa. De esta manera, se genera el archivo "pipeline.joblib", el cual puede ser llamado por la API y, posteriormente, ser consumido por la página web. Este enfoque garantiza la eficiencia y precisión en el procesamiento de datos y la generación de resultados.

1.2 Acceso por medio de API

En el archivo "api.py", se construye una API que emplea el framework FastAPI para automatizar un proceso de predicción. Esta API inicia creando una instancia de la aplicación FastAPI y carga un modelo previamente entrenado desde el archivo "pipeline.joblib", el cual es asignado a una variable denominada "pipe". La API define una ruta POST llamada "/predict/" diseñada para gestionar solicitudes de predicción.

La ruta "/predict/" ofrece dos alternativas para proporcionar datos de entrada: la carga de un archivo Excel mediante el parámetro "file" o la provisión directa de

texto a través del parámetro "text_input". Si se elige la opción de archivo, la API verifica si el archivo es un archivo Excel y, en caso afirmativo, lo lee y almacena en un DataFrame de pandas llamado "data". En cambio, si se opta por ingresar texto, se genera un DataFrame utilizando ese contenido textual.

Posteriormente, la API emplea el modelo previamente cargado ("pipe") para realizar predicciones en los datos proporcionados. Si se utiliza un archivo, las predicciones se retornan como una lista en una respuesta JSON. Por otro lado, si se introduce texto de manera directa, las predicciones se almacenan en un archivo Excel y se envían como respuesta en formato de archivo.

Este enfoque ofrece flexibilidad para predecir con datos estructurados o con texto sin procesar, facilitando la generación de resultados en un formato adecuado según la elección del usuario.

2. Desarrollo de la aplicación y justificación

La solución presentada es una aplicación de React que clasifica textos relacionados con los Objetivos de Desarrollo Sostenible (ODS). La aplicación permite a los usuarios ingresar texto directamente o cargar un archivo en formato Excel con textos para ser clasificados. Después de clasificar el texto, muestra la categoría del ODS a la que pertenece y brinda una descripción correspondiente.

2,1 Desarrollo de la aplicación y justificación:

La aplicación se centra en la clasificación automatizada de textos en relación con los ODS. Utiliza un modelo analítico basado en aprendizaje automático para procesar la información ingresada y vincularla con una de las categorías específicas de los ODS. La justificación principal de esta aplicación es brindar una herramienta ágil y accesible que permita a los usuarios identificar rápidamente la categoría a la que pertenece un texto en relación con los ODS. Estas son algunas razones que justifican el uso de esta plataforma para beneficiar al usuario interesado:

- **Facilidad de clasificación y uso:** La aplicación simplifica y agiliza la identificación de la categoría de un texto relacionado con los ODS, lo cual es fundamental para cualquier profesional involucrado en la implementación, seguimiento o evaluación de iniciativas relacionadas con los ODS. Además, Al ser una aplicación web o móvil, brinda accesibilidad y facilidad de uso, lo que amplía su utilidad y alcance a una variedad de usuarios.
- **Apoyo en la toma de decisiones:** Al clasificar textos de manera automatizada, la aplicación puede respaldar la toma de decisiones informadas, permitiendo a los usuarios entender rápidamente la relevancia de un texto específico en el contexto de los ODS.
- **Eficiencia en el análisis**: Para analistas de datos y profesionales involucrados en proyectos relacionados con los ODS, esta aplicación puede acelerar y

mejorar el proceso de análisis de grandes cantidades de textos, facilitando la identificación de áreas de enfoque en sus respectivos campos de trabajo.

2.2 Usuario/Rol de la organización que utilizará la aplicación:
El usuario/rol de la organización que se beneficiaría de esta aplicación puede variar y abarcar desde investigadores, profesionales en desarrollo sostenible, responsables de políticas públicas, analistas de datos, hasta cualquier persona interesada en clasificar textos en relación con los ODS.

2.3 Conexión entre la aplicación y el proceso de negocio:
La conexión con el proceso de negocio se establece en el apoyo a la toma de decisiones informadas y en la identificación rápida de la categoría de un texto específico relacionado con los ODS. Si se implementa en una organización, esta herramienta podría respaldar el análisis de datos, la generación de informes, la evaluación de proyectos o políticas en términos de su alineación con los ODS.

2.4 Tabla de actores actualizada

Rol dentro de la empresa	**Tipo de actor**	**Beneficio**	**Riesgo**
UNFPA	Financiador/Usuario	Mejora en la toma de decisiones fundamentada en la clasificación precisa de problemáticas y soluciones relacionadas con los ODS.	Riesgo de toma de decisiones erróneas si el modelo no funciona correctamente, lo que podría conducir a políticas ineficaces o decisiones no fundamentadas en datos sólidos. Se necesita una revisión constante para asegurar la calidad de los resultados.
Gobierno Colombiano	Proveedor/Usuario	Uso más eficiente de los recursos para implementar políticas que impacten positivamente en la calidad de vida de la población, basado en la identificación precisa de las áreas problemáticas relacionadas con los ODS.	Riesgo de implementar políticas incorrectas si el modelo no identifica adecuadamente las áreas problemáticas relacionadas con los ODS, lo que podría derivar en una asignación inadecuada de recursos y una posible pérdida de confianza por parte de la población
Población	Proveedor/Beneficiado	Mejora sustancial en la calidad de vida al lograr el cumplimiento efectivo de los ODS, lo que se traduce en políticas más relevantes y precisas que abordan directamente las necesidades de la población.	Riesgo de falta de apoyo para mejorar la calidad de vida si el modelo no identifica adecuadamente las áreas problemáticas o soluciones relacionadas con los ODS. Existe la posibilidad de no abordar las necesidades reales de la población si la información proporcionada no es precisa.

2.5 Decisiones de diseño
La justificación de las decisiones de diseño en la aplicación de clasificación de textos relacionados con los Objetivos de Desarrollo Sostenible (ODS) se basa en

varios factores clave, principalmente centrados en la influencia del usuario objetivo, los objetivos de la persona de estadística (Andrea Molano) y las necesidades del proyecto. Estos son algunos puntos que respaldan las decisiones de diseño:

1**. Interfaz de usuario intuitiva:**
Se diseñó una interfaz sencilla y amigable para el usuario final. Esta decisión se tomó considerando que el usuario puede ser diverso en conocimientos técnicos y se necesitaba una interfaz que fuera fácil de usar para cualquier usuario interesado en clasificar textos en relación con los ODS.

2. Funcionalidad de entrada flexible:
- Influencia del grupo de estadística: Se permitió a los usuarios ingresar texto directamente o cargar un archivo en formato Excel para clasificar. Esta decisión consideró la conveniencia del usuario final que puede tener datos en diferentes formatos y optar por la forma más conveniente para ellos.

3. Retroalimentación visual y comprensión de resultados:
La presentación de los resultados a través de una tarjeta con imagen, título y descripción proporciona una comprensión rápida y visual de la clasificación del texto en relación con los ODS. Esto se hizo teniendo en cuenta que los usuarios podrían necesitar información clara y rápida sobre la categorización del texto.

4. Acceso a información adicional:
Se incorporó un enlace a los Objetivos de Desarrollo Sostenible (ODS) para proporcionar información adicional sobre los mismos. Esto se hizo reconociendo que los usuarios pueden necesitar contexto y detalles más amplios sobre los ODS para comprender mejor el significado de la clasificación.

5. Facilidad de acceso a través de múltiples dispositivos:
- Influencia del grupo de estadística: el desarrollo de una aplicación web permite el acceso a la herramienta a través de diversos dispositivos (computadoras, tablets, teléfonos móviles). Esto se hizo teniendo en cuenta que los usuarios pueden acceder a la herramienta desde distintos entornos y dispositivos según su conveniencia.

Finalmente, las decisiones de diseño se tomaron considerando en primer lugar la experiencia del usuario final que pudimos discutir con la persona de estadística (Andrea Molano). Se priorizó la accesibilidad, la facilidad de uso y la presentación clara de los resultados, reconociendo que los usuarios pueden provenir de diferentes ámbitos y tener diferentes niveles de familiaridad con la tecnología y los conceptos relacionados con los ODS. El diseño de la aplicación se orientó hacia una experiencia de usuario fluida y comprensible para satisfacer las necesidades y expectativas del usuario objetivo.

3. Resultados - video

Se ha creado una aplicación web utilizando React como framework frontend. Mediante el uso de una API desarrollada en Python con el framework FastAPI, es capaz de hacer que la interacción de un usuario con el modelo sea mucho más amena. El modelo se ha desplegado en una instancia de AWS que se mantiene activa durante los periodos de tiempo solicitados por el usuario de estadística.

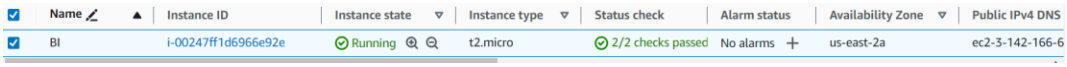

Ilustración 1: instancia AWS

Ilustración 2: App web desplegada

Inicialmente, el usuario ingresaba un texto en el recuadro y este se clasificaba en el modelo de manera individual. Luego, el usuario de estadística indicó en una reunión preliminar que sería una mejora sustancial poder procesar una gran cantidad de textos al mismo tiempo para realizar los análisis estadísticos pertinentes. Por lo tanto, se agregó una opción para enviar directamente un archivo Excel a la API, la cual devuelve el archivo con los textos clasificados en conjunto.

Ilustración 3: Texto clasificado por el modelo

Ilustración 4: Texto clasificado por el modelo

Finalmente, durante el uso del modelo, el usuario de estadística enfatizó que, durante la recolección de datos, es probable que haya recopilado datos erróneos, ya que no sabe cuál es la longitud mínima que debe tener un texto para que pueda ser clasificado correctamente. Por lo tanto, esta es una consideración que se debe tener en cuenta en el futuro.

Ilustración 5: Esta clasificación es ambigua, ya que la palabra "Desigualdad" puede ser clasificada en varios ODS

4. Trabajo en equipo
(pequeño resumen de que nos fue muy bien trabajando con Andrea bla bla bla)

4.4. Roles

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.

4.5. Reuniones y fases

Fase I: Entendimiento del requerimiento y planeación

- LP: Coordinar la reunión de lanzamiento y panear la forma de trabajo del grupo (5 puntos).

- LN: Contactar al grupo de estadística y programar la reunión para saber sus requerimientos y qué espera de la aplicación (5 puntos).
- LD: Realizar análisis para entender cómo se llevaría a cabo la automatización del proceso (5 puntos).
- LA: Identificar el enfoque analítico para alcanzar los objetivos del negocio (10 puntos).

Fase II: Desarrollo de la aplicación y evaluación

- LD: Automatizar el proceso de limpieza de datos para incluirlo en el modelo (5 puntos).
- LA: Evaluar los resultados obtenidos por el modelo (10 puntos).
- LN: Validar los resultados con el grupo de estadística para generar correcciones en caso de ser necesario (5 puntos).
- LN: Creación de la API para poder ser consumida por la aplicación (5 puntos).
- LP: Coordinar reuniones de seguimiento y asegurar el avance del proyecto (5 puntos).
- LA: Desarrollo de la aplicación según las preferencias de diseño del usuario (grupo de estadística)(10 puntos)
- LD: Despliegue de la aplicación en la nube (2 puntos).

Fase III: Resultados y Sustentación

- LP: Coordinar la reunión de finalización (4 puntos).
- LN: Habilitar al grupo de estadística para probar la aplicación final (10 puntos).
- LD: Asegurar que los entregables estén completos y listos para la presentación (1 punto).
- LN: Generar un video de 5 minutos explicando el proyecto y resultados (8 puntos).
- LA: Contribuir a la modificación del mapa de actores relacionados con el producto de datos (10 puntos).

- *Distribución de Puntos entre los Integrantes:*

Líder de Proyecto (LP): 14 puntos
Líder de Negocio (LN): 33 puntos
Líder de Datos (LD): 13 puntos
Líder de Analítica (LA): 40 puntos