Johan Alexis Bautista Quinayas
Danny Camilo Muñoz Sanabria
Juan Camilo López Cortés

Universidad de los Andes

ISIS 3301 – Inteligencia de Negocios

Proyecto 1 Etapa 2

Link Backend API: https://github.com/BI-Grupo32/Proy1-BI.git

Link Frontend: https://github.com/Pandlop/ProyectoBI-front.git

Sección 2. (40%) Justificación. Descripción del rol de la organización.

El Fondo de Poblaciones de las Naciones Unidas (UNFPA) tiene como uno de sus principales objetivos la identificación de problemáticas y la evaluación de soluciones relacionadas con los Objetivos de Desarrollo Sostenible (ODS), específicamente el ODS 3 (Salud y bienestar), ODS 4 (Educación inclusiva y de calidad) y ODS 5 (Igualdad de género). Para alcanzar estos objetivos, el UNFPA depende de la participación ciudadana y del análisis eficiente de las opiniones y percepciones de las personas sobre estos temas.

Es por esto, que el rol propuesto para el usuario de la aplicación es el de un analista de la organización. Es decir, del Fondo de Poblaciones de las Naciones Unidas (UNFPA). Este usuario es responsable de cargar y analizar conjuntos de opiniones recopiladas para alinearlas con los Objetivos de Desarrollo Sostenible (ODS) 3, 4 y 5. El objetivo principal de este rol es cargar y procesar grandes cantidades de opiniones ciudadanas a través de la aplicación. Al analizar las opiniones y generar predicciones automáticas sobre cuál de los ODS se alinea con cada opinión, así como la distribución de estas opiniones, el analista puede extraer información clave que permitirá a la organización ajustar sus estrategias y políticas para abordar las problemáticas identificadas.

Conexión con el Proceso de Negocio:

El proceso de negocio que apoya la aplicación está centrado en el análisis de datos textuales relacionados con los ODS. Este analista utilizará la aplicación para cargar archivos que contienen opiniones de los ciudadanos, y a través del pipeline de predicción, podrá identificar automáticamente qué ODS se alinean con dichas opiniones.

El uso de un archivo Excel para cargar los datos en lugar de realizar predicciones individuales en tiempo real está alineado con la naturaleza del trabajo del analista, quien normalmente trabajará con grandes volúmenes de datos obtenidos a lo largo de fuentes masivas de información. El analista puede generar predicciones para múltiples opiniones simultáneamente y obtener un archivo con los resultados y probabilidades asociadas a los mismos, así como una visualización rápida de la distribución de estas clasificaciones lo que permite ahorrar tiempo y recursos en el análisis.

Importancia de la Aplicación:

La existencia de esta aplicación es útil para el rol del analista, ya que automatiza un proceso que anteriormente requería tiempo y esfuerzo humano. La capacidad de hacer predicciones sobre grandes volúmenes de datos textuales permite que se enfoquen los recursos en interpretar los resultados y generar recomendaciones estratégicas para la organización en lugar de realizar análisis manuales de las opiniones.

En este sentido, la aplicación no solo automatiza el proceso de análisis de opiniones, sino que también acelera el cumplimiento de los objetivos del negocio. El UNFPA necesita actuar con rapidez y precisión para implementar soluciones que respondan a las necesidades sociales en temas críticos (salud, educación e igualdad de género). Al reducir significativamente el tiempo y los recursos necesarios para analizar estas opiniones, la aplicación permite que el UNFPA responda de manera más efectiva y oportuna a los desafíos asociados con los ODS, desde el rol de Analista de datos para el cual fue desarrollada la aplicación.

Además. este proceso no solo mejora la eficiencia, sino que también aumenta la precisión de las recomendaciones, asegurando que las estrategias de la organización estén alineadas con las necesidades y percepciones reales de la ciudadanía.

Sección 1. (20%) Proceso de automatización del proceso de preparación de datos, construcción del modelo, persistencia del modelo y acceso por medio de API:

Tipos de Reentrenamiento del Modelo

Para el reentrenamiento del modelo, se evaluaron tres enfoques. A continuación, se presenta una descripción de cada uno de ellos, sus ventajas y desventajas, y finalmente se detalla la opción implementada.

1. **Reentrenamiento Incremental (Implementado)**
   El modelo se reentrena utilizando tanto los datos originales como los nuevos proporcionados por el usuario. Los nuevos datos se combinan con los existentes, y luego el modelo se entrena desde cero utilizando todo el conjunto de datos actualizado.

   Ventajas:

   - ✓ Mantiene el conocimiento previo, ya que, al reutilizar los datos antiguos junto con los nuevos, el modelo sigue aprendiendo de toda la información acumulada. Es decir, se garantiza que no se pierda el conocimiento adquirido previamente.
   - ✓ El modelo se puede adaptar a nuevos patrones, debido a que permite que el modelo se ajuste a los cambios presentes en los nuevos datos, claro sin olvidar lo aprendido anteriormente.

   Desventajas:

   - ✓ El crecimiento de los datos es importante, puesto que a medida que se agregan más datos, el tamaño total del conjunto de datos crece, lo que puede volverse ineficiente en términos de tiempo de procesamiento y recursos de almacenamiento.
   - ✓ No contempla que los datos antiguos pueden volverse obsoletos. Entonces, si los datos viejos ya no reflejan la realidad actual, se puede afectar negativamente las predicciones.

2. **Reentrenamiento desde cero**
   El modelo se reentrena completamente desde cero utilizando solo los datos nuevos proporcionados, ignorando todos los datos antiguos y patrones previamente aprendidos.

   Ventajas:

- ✓ El modelo tiene una adaptación inmediata a los nuevos datos, útil cuando los datos antiguos ya no son relevantes.
- ✓ Tiene una menor carga de recursos, debido a que solo utiliza los nuevos datos. Es decir, se reduce el tiempo de procesamiento y requiere menos recursos de almacenamiento.

Desventajas:

- ✓ El modelo pierde el conocimiento acumulado, al entrenar desde cero, se pierden todos los patrones previamente aprendidos, así como toda la información previamente extraída.
- ✓ Surge un posible problema con pocos datos, ya que, si los datos nuevos no son suficientes, el modelo puede no tener suficiente información para generalizar correctamente.

3. **Reentrenamiento con Amortiguación de Datos**
   Este enfoque es un punto intermedio de los dos anteriores, donde se utiliza un subconjunto representativo de los datos antiguos junto con los nuevos para entrenar el modelo. E este sentido, en lugar de incluir todos los datos antiguos, se seleccionan aquellos que aún son relevantes para combinarlos con los nuevos datos.

   Ventajas:

   - ✓ Existe un balance entre lo nuevo y lo antiguo, ya que, se permite que el modelo mantenga parte de su conocimiento anterior mientras se adapta a los nuevos datos.

   Desventajas:

   - ✓ La selección de datos es extremadamente compleja, puesto que se requiere una estrategia adecuada para seleccionar qué datos antiguos mantener. Si se seleccionan mal, el rendimiento del modelo puede empeorar.

Elección del enfoque de reentrenamiento:

El enfoque implementado es el Reentrenamiento Incremental, que combina los datos nuevos con los históricos, lo que permite al modelo aprender continuamente sin perder el conocimiento previo. Se eligió porque es óptimo para el UNFPA, puesto que garantiza que el análisis de las opiniones siga alineado sin ambigüedad con los ODS 3, 4 y 5. En este sentido, mantener los datos anteriores permite reconocer patrones persistentes, mientras se adapta a nuevos cambios en la población, lo que es importante para ajustar políticas en temas clave como salud, educación e igualdad de género. Teniendo en cuenta que no se corre el riesgo de sesgar el modelo con pocos datos (en el caso de reentrenamiento desde cero) y tampoco de fallar en la ambigua elección de datos históricos relevantes (en el caso de reentrenamiento con Amortiguación de Datos).

Sección 4. (10%) Trabajo en equipo.

En esta etapa del proyecto, los miembros del equipo asumieron roles específicos que permitieron el desarrollo completo tanto del backend (API) como del frontend de la aplicación de clasificación de textos. Las reuniones del equipo ayudaron a gestionar la distribución equitativa de tareas y el progreso

del proyecto, manteniendo los objetivos del Fondo de Poblaciones de las Naciones Unidas (UNFPA) como el enfoque central.

Distribución del trabajo y roles:

1. **Johan Alexis Bautista Quinayas:**

   Roles:

   - ✓ Líder del proyecto
   - ✓ Ingeniero de datos y desarrollador backend

   Tareas realizadas:

   - ✓ Desarrollo de la API en FastAPI, incluyendo el diseño de todos los endpoints para predicción y reentrenamiento.
   - ✓ Preprocesamiento de los datos (tokenización, lematización, vectorización con TF-IDF).
   - ✓ Exportación y persistencia del modelo de clasificación entrenado (SVC) mediante joblib.
   - ✓ Creación del pipeline de clasificación y automatización del proceso de reentrenamiento.
   - ✓ Pruebas y validación del pipeline de predicción y reentrenamiento.

   Retos:

   - ✓ Asegurar que la API cumpliera con los requerimientos del UNFPA, en términos de predicción y reentrenamiento.
   - ✓ Manejar la optimización de los tiempos de ejecución y la comunicación eficiente entre la API y el frontend.

   Formas para resolver retos:

   - ✓ Se optimizó el código de la API y se ajustó el pipeline de preprocesamiento para mejorar la eficiencia del modelo.
   - ✓ Se realizaron varias iteraciones de prueba para asegurar la correcta integración entre el frontend y la API.

   Puntos asignados: 33.3 puntos

   Horas de trabajo: 16 horas

2. **Danny Camilo Muñoz Sanabria:**

   Roles:

   - ✓ Ingeniero de software encargado del diseño y desarrollo del frontend

   Tareas realizadas:

   - ✓ Diseño de la interfaz del frontend que interactúa con la API para permitir la carga de archivos y visualización de resultados.
   - ✓ Implementación de la conexión entre el frontend y la API, asegurando una comunicación fluida para la predicción de textos y la descarga de resultados.

- ✓ Pruebas del frontend para validar la funcionalidad con diferentes casos de uso y archivos Excel.

Retos:

- ✓ Garantizar que la interfaz fuera intuitiva y que la comunicación entre la API y el frontend fuera eficiente.
- ✓ Manejo de errores y mensajes de retroalimentación desde el frontend hacia el usuario.

Formas para resolver retos:

- ✓ Se realizaron varias pruebas de conexión y manejo de errores para asegurar que los usuarios pudieran cargar archivos Excel sin problemas.
- ✓ Se estableció una estructura clara de comunicación entre la API y el frontend, devolviendo mensajes adecuados al usuario.

Puntos asignados: 33.3 puntos

Horas de trabajo: 10 horas

3. **Juan Camilo López Cortes:**

Roles:

- ✓ Ingeniero de software encargado del desarrollo y pruebas del frontend

Tareas realizadas:

- ✓ Colaboración en el diseño del frontend, asegurando que la interfaz cumpliera con los requisitos del proyecto.
- ✓ Desarrollo de la funcionalidad de conexión del frontend con la API para las funcionalidades de predicción y reentrenamiento.
- ✓ Pruebas del sistema completo (frontend + backend) para garantizar la usabilidad y funcionalidad esperadas.

Retos:

- ✓ Integrar correctamente el frontend con la API y manejar grandes volúmenes de datos desde los archivos Excel.
- ✓ Asegurar que la interfaz refleje correctamente los resultados y métricas obtenidas desde la API.

Formas para resolver retos:

- ✓ Se realizaron pruebas exhaustivas con archivos de diferentes tamaños para asegurar la estabilidad de la conexión.
- ✓ Se implementó una interfaz amigable que permitiera al usuario descargar los resultados de las predicciones de forma sencilla.

Puntos asignados: 33.3 puntos

Horas de trabajo: 10 horas

Johan Alexis Bautista Quinayas
Danny Camilo Muñoz Sanabria
Juan Camilo López Cortés

Reuniones del grupo:

1. **Reunión de lanzamiento y planeación:** Se discutieron los objetivos generales del proyecto y se asignaron los roles de cada integrante. Se planificaron las tareas iniciales, centrándose en el diseño del backend y del frontend.
2. **Reuniones de seguimiento:** Se realizó una reunión semanal para monitorear el avance, ajustar las tareas y resolver problemas.

Distribución de Puntos:

Cada miembro del equipo fue evaluado equitativamente y contribuyó de manera balanceada al éxito del proyecto. Se repartieron 100 puntos entre los tres integrantes, asignando 33.3 puntos a cada uno basado en sus contribuciones.