Documentación Etapa 2 Proyecto 1.

Cesar Avellaneda 202214746

Santiago Tinjacá 202215991

Jorge Bustamante 202210581

Sección 1: Automatización de la preparación de datos, construcción y persistencia del modelo y acceso por medio de la API.

Para la automatización de la preparación de los datos debemos resumir, en funciones que puedan ser llamadas cuando se tenga una nueva entrada, el proceso de tokenizar los párrafos, limpiar y estandarizar caracteres. Para esto usamos las funciones de preparación de datos que programamos para la Etapa 1 y las juntamos en una sola función “prep” que las llama secuencialmente. A su vez hicimos uso de FunctionTransformer para agregar la función “prep” al pipeline, con el que hacemos la limpieza, la vectorización y posteriormente el modelo SVC. Después exportamos el modelo en formato “.joblib” para su posterior uso.

El modelo generado es utilizado en una aplicación que implementa la lógica con capacidad de recibir texto, usar los funcionar de transformación y el modelo para retornor una respuesta. Como resultado, la aplicación backend hace uso de la librería de fastapi, consta de un DataModel, PredictionModel, el preprocesamiento, el entrenamiento de nuevos modelos y el main, el cual tiene los endpoints que permiten recibir las peticiones post y get que necesitamos.

Para el front usamos el FrameWork de React.

Sección 2: Desarrollo de la aplicación y justificación.

La aplicación consta de dos partes, viene con el modelo entrenado en la Etapa 1 de base al inicializarla, haciendo uso de este modelo el cliente puede enviar texto para que la aplicación lo envié al api, se haga la predicción y retorne al usuario el ODS al que se puede asociar su texto con la probabilidad que el modelo asegura que si pertenece. Y la otra parte de la aplicación recibe un CSV con el mismo formato del original, para reentrenar el modelo haciendo uso del Pipeline que definimos al principio.

De esta manera, la aplicación sirve para que el negocio pueda usar el modelo para automatizar la clasificación de textos en alguno de los ODS y saber cuál deberían priorizar porque le compete más a la ciudadanía. Además, en el momento que se tenga un registro más exacto o amplio de datos de entrenamiento (marcados), estos se le pueden enviar a la aplicación para que el modelo sea entrenado nuevamente y pueda ser más eficaz en esta tarea.

Sección 3: Resultados.

Video con la aplicación en uso en el padlet.

Validaciones de usabilidad.

Hablando con otros estudiantes y familiares llegamos a las siguientes conclusiones:

La aplicación es muy fácil de usar en la medida de que solo tiene dos funcionalidades/partes, una para enviarle texto y que te devuelva la predicción y probabilidad dada por el modelo y la otra que es un poco más complicada es la de reentrenar el modelo y su complejidad viene porque el XLSX que se envía tiene que estar en un formato especifico, por lo que una de las soluciones era separar estas dos funcionalidades en dos páginas distintas, pero también concluimos que al estar relativamente automatizado, no serán muchas personas las que usen la aplicación, seguramente sea una persona que tenga conocimientos técnicos suficientes como para que esta distinción no sea necesaria del todo.

Sección 4: Trabajo en equipo.

Líder de proyecto: Santiago Tinjacá.

Roles y Tareas:

- Gestión del Proyecto: Coordinación general, definición de fechas y asignación de tareas.

- Documentación y Entrega: Preparación y subida de la entrega final.

- Resolución de Conflictos: Toma de decisiones finales.

Tiempos Dedicados:

- Gestión y Coordinación: 10 horas

- Documentación: 8 horas

- Resolución de Conflictos: 5 horas

Retos y Soluciones:

- Desacuerdos: Implementar reuniones frecuentes y comunicación clara.

Distribución de Puntos:

- 33.3 puntos

Ingeniero de datos: Cesar Avellaneda.

Roles y Tareas:

- Automatización de la preparación de datos: Encargado de hacer todas las funciones requeridas para el preprocesamiento.

- Automatización del entrenamiento del modelo: Encargado de hacer la función general de entrenamiento del modelo.

- Pipeline del modelo: Encargado de juntar todo el preprocesamiento y entrenamiento en el pipeline para exportar el modelo ya entrenado.

- Documentar: Encargado de hacer parte del documento.

Tiempos Dedicados:

- Programación del preprocesamiento: 5 horas

- Programación del entrenamiento: 6 horas

- Construcción del Pipeline: 3 horas

-Documentación: 5 horas.

Distribución de Puntos:

- 33.3 puntos

Ingeniero de software responsable del diseño de la aplicación: Jorge Bustamante.

Roles y Tareas:

-Implementación del Back: Encargado de implementar el back.

-Implementación del Front: Encargado de implementar la aplicación front para los usuarios

Tiempos Dedicados:

Diseño del back: 4 horas

Diseño del front: 4 horas

Implementación: 5 horas

Distribución de Puntos:

- 33.3 puntos