Prof David Miguel Avila Cruz
Mario Solano

22/04/2025

Tarea Semana 1-2: Pipeline de MLOps + Imagen de Docker

Descripción de un Pipeline de MLOps

Dados los avances tecnológicos, en el campo de la medicina la cantidad de información que existe de los pacientes es muy abundante. Sin embargo, para algunas enfermedades no tan comunes, llamadas huérfanas, los datos que existen escasean. Se requiere construir un modelo que sea capaz de predecir, dados los datos de síntomas de un paciente, si es posible o no que este sufra de alguna enfermedad. Esto se requiere tanto para enfermedades comunes (muchos datos) como para enfermedades huérfanas (pocos datos).

Diseño del pipeline

El objetivo es generar un modelo que pueda predecir la probabilidad de múltiples enfermedades, comunes y huérfanas, dado un conjunto de síntomas. Debe de ser capaz de generar una respuesta inclusive en condiciones donde no se tenga toda la información.

Existen restricciones con respecto a la cantidad y calidad de datos tales como:

- Desbalance de datos entre las enfermedades comunes y huérfanas. Esto afectará directamente el entrenamiento y la evaluación del modelo.
- La calidad y heterogeneidad de los datos variarán ya que vamos a considerar diferentes fuentes de datos con formatos y niveles de detalle faltantes así como la existencia de datos faltantes o inconsistencias.
- Al trabajar con información sensible de cada cliente se debe de considerar la privacidad de los mismos utilizando algún tipo de anonimización.

Los datos que esperamos obtener para la detección de enfermedades pueden ser:

- Datos estructurados o semiestructurados que representen los síntomas presentes, ausentes y la severidad de los mismos
- Datos estructurados con respecto a las caracteristicas de los mismos como edad, sexo, etc.
- Notas médicas con las anotaciones hechas por el doctor encargado.
- Las probabilidades asociadas a cada enfermedad a considerar o la enfermedad diagnosticada.

Desarrollo del pipeline

Las fuentes usadas podrían ser:

- Historias Clínicas
- Registros especializados de enfermedades raras o huérfanas
- Bases de datos de investigación médica
- Literatura científica para las relaciones entre enfermedad y síntomas

En estas fuentes de datos se debe de realizar un proceso de ingesta y procesamiento. Se debería de crear un repositorio centralizado con las fuentes de datos y luego aplicar procesos de limpieza y homogeneización. Además, en esta etapa deberíamos de incluir procesos robustos de anonimización para proteger la privacidad de los pacientes.

Adicionalmente, en este proceso se deberían de incluir características adicionales tales como conversión de síntomas crudos a variables categóricas haciendo encoding o agrupaciones de variables numéricas como la temperatura.

Los modelos de machine learning a usar podría ser un modelo de clasificación multi etiqueta siendo cuidadosos con el imbalance en las enfermedades huérfanas. Deberíamos de usar técnicas como sobremuestreo y pesos para las clases raras.

Por último, el protocolo de evaluación debería ser hold out con una división estratificada en las enfermedades huérfanas y asegurarnos que estén presente proporcionalmente en los conjuntos de entrenamiento, validación y testeo. Consiguientemente el modelo deberá ser evaluado sobre estos conjuntos de datos usando métricas como la precisión, el recall y F1-Score prestando especial atención al rendimiento en las enfermedades huérfanas. Estás métricas de evaluación se deberían de someter a una validación cruzada durante el entrenamiento para asegurar robustez.

Producción del pipeline

La solución se va a desplegar como un sitio web, usando la librería flask, donde el médico ingrese los parámetros de ojos clientes y que este nos proporcione de forma inmediata un listado de estados en los que puede estar el paciente. Esto se implementa mediante un despliegue contenerizado usando docker y los aprendido en el módulo.

Idealmente, se debería monitorear continuamente el desempeño del modelo para detectar con precisión las enfermedades por los cambios en las mismas y en cómo se presentan. Al realizar este monitoreo será posible establecer disparadores para realizar un reentrenamiento al detectar cambios significativos en los datos de entrada.

Diagrama del proceso