Sebastian Murillas
Mlops

Descripcion de los componentes del pipeline MLOPS

Etapa de Diseño

La etapa de diseño tiene como objetivo establecer las bases del modelo predictivo, identificando las restricciones inherentes al dominio clínico, el tipo de datos disponibles y las acciones necesarias para garantizar su calidad y utilidad para el entrenamiento posterior. Se identifican varias restricciones críticas: en primer lugar, el desbalance de clases, ya que se cuenta con una mayor proporción de casos asociados a enfermedades comunes que a enfermedades huérfanas; en segundo lugar, la sensibilidad de los datos personales y clínicos, que exige su tratamiento bajo normativas de privacidad como HIPAA o GDPR; y finalmente, la posibilidad de contar con una cantidad significativa de datos nulos o mal etiquetados, especialmente en registros históricos o no estructurados.

En cuanto al tipo de datos, se reconocen principalmente dos categorías: los datos estructurados (como síntomas codificados, edad, género o antecedentes), que representan la mayoría de las variables de entrada, y los datos no estructurados (como comentarios médicos o notas clínicas), que pueden integrarse en fases posteriores mediante técnicas de procesamiento de lenguaje natural (NLP).

Durante esta fase, se establecen tres acciones clave: la validación de datos nulos con el apoyo de expertos clínicos, el tratamiento y transformación de los datos mediante técnicas como codificación (One-Hot Encoding), imputación de valores faltantes y generación de embeddings, y el tratamiento del desbalance de clases utilizando estrategias como data augmentation y transferencia de aprendizaje (transfer learning). Todas estas acciones alimentan el proceso de definición de la estructura de datos, donde se decide el formato final de entrada para el modelo y se garantiza la integridad y calidad del dataset inicial.

Etapa de Desarrollo – Descripción

La etapa de desarrollo tiene como objetivo transformar los datos disponibles en un modelo predictivo funcional y evaluado, capaz de identificar enfermedades tanto comunes como huérfanas. El proceso comienza con la identificación de las fuentes de datos, que incluyen registros internos institucionales y bases de datos externas especializadas en enfermedades no comunes. Estos datos son gestionados mediante herramientas de almacenamiento en la nube y sistemas de versionado que aseguran trazabilidad y reproducibilidad.

Una vez asegurada la calidad y disponibilidad de los datos, se realiza un análisis exploratorio (EDA) y un proceso de ingeniería de características para comprender la estructura de los datos y preparar los insumos necesarios para el entrenamiento. Posteriormente, se procede con el uso de modelos supervisados, destacando algoritmos como **XGBoost** y **Random Forest**, que ofrecen un buen rendimiento en problemas con datos tabulares y desbalance de clases. Sobre estos modelos se aplica una fase de ajuste fino (**fine-tuning**) para optimizar hiperparámetros y mejorar el desempeño.

Sebastian Murillas
Mlops

El modelo que arroje las mejores métricas es seleccionado y sometido a una revisión técnica de resultados, donde se validan su estabilidad y capacidad de generalización. Si los resultados son satisfactorios, se considera el modelo como listo para pasar a producción. En caso contrario, el proceso entra en una nueva iteración, ajustando los pasos anteriores con base en los hallazgos obtenidos.

Esta fase garantiza que el modelo entregado esté sustentado en un ciclo riguroso de análisis, entrenamiento, evaluación y validación técnica, cumpliendo criterios clínicos y técnicos antes de su despliegue final.

Etapa de Producción – Descripción

La etapa de producción tiene como objetivo poner en funcionamiento el modelo en un entorno clínico real, permitiendo su uso por parte del personal médico y asegurando su mantenimiento a lo largo del tiempo. Este proceso contempla tres componentes fundamentales: el despliegue del modelo, la predicción diaria y el monitoreo continuo, incluyendo el eventual reentrenamiento automático.

El modelo se expone a través de una **API REST**, ejecutada en un **entorno contenerizado y orquestado** (por ejemplo, usando Docker y Kubernetes), lo cual permite escalar el servicio y asegurar su disponibilidad. A partir de esta API, se realizan predicciones diarias sobre los nuevos datos ingresados, cuyas salidas son almacenadas para auditoría, análisis y evaluación futura.

De manera paralela, se implementa un sistema de **monitoreo doble**: por un lado, el **monitoreo de infraestructura** permite rastrear el estado del servicio (latencia, errores, disponibilidad); por otro lado, el **monitoreo de las métricas del modelo** y de las características de los datos (por ejemplo, cambio en la distribución de síntomas o clases) permite detectar **data drift** o degradación del rendimiento. Esto incluye el análisis de métricas agregadas como precisión, recall o F1-score calculadas sobre los casos con diagnóstico confirmado.

Cuando se detectan cambios relevantes en el comportamiento del modelo o se acumulan suficientes nuevos registros clínicos, se activa el **proceso de reentrenamiento automático**. Este proceso parte de la **ingesta de nuevos datos**, realiza un entrenamiento actualizado del modelo, y compara su desempeño frente al modelo actual. Si los resultados son favorables, se sustituye el modelo en producción garantizando la continuidad operativa.

Con esta estructura, se garantiza que el modelo permanezca actualizado, confiable y trazable, permitiendo su integración efectiva en los flujos de trabajo médicos reales.