Desarrollo de un Pipeline de IA para Cáncer Cerebral:

Un Recorrido por el Proceso, Desafíos y Aprendizajes

Juan Pablo Restrepo Urrea

13 de junio de 2025

Índice

Preámbulo: Una Guía para esta Documentación

A través de este documento, quiero ofrecer una perspectiva complementaria al detalle técnico y analítico ya presente en las versiones HTML de los notebooks del proyecto (`01_EDA_and_Preprocessing.html`, `02_MRI_Tumor_Classification_Model.html`, y `03_Treatment_Recommendation_Model.html`). Si bien dichos notebooks contienen un registro exhaustivo de cada paso, análisis y resultado, mi intención aquí es compartir una narrativa más personal sobre el flujo de trabajo, los razonamientos detrás de las decisiones tomadas, los desafíos encontrados y los aprendizajes clave obtenidos durante el desarrollo de este pipeline de inteligencia artificial para la predicción de tratamiento a través de imagenes e historias clinicas.

Este ejercicio ha sido una oportunidad valiosa para aplicar y fortalecr mis conocimientos, y espero que este relato ofrezca una visión más profunda de mi enfoque y motivación. Consideren este escrito como un complemento que busca humanizar el proceso técnico, ofreciendo un vistazo a las reflexiones y los momentos de predicamento y descubrimiento que son inherentes a cualquier proyecto de ciencia de datos.

1. Inmersión Inicial: Comprendiendo el Dominio y los Datos

Mi primer paso, fundamental en cualquier proyecto de ciencia de datos, fue sumergirme en el contexto del problema. Antes de escribir una sola línea de código, dediqué tiempo a investigar sobre los tipos de tumores cerebrales en cuestión (Glioma, Meningioma y la categoría "Otros Tumores"), buscando entender sus características distintivas, especialmente aquellas observables en imágenes de resonancia magnética (MRI). Quería comprender la naturaleza de las patologías y los tratamientos comunes asociados a cada una, para así abordar el análisis de datos con una base de conocimiento más sólida.

Esta fase de referenciación me permitió no solo familiarizarme con la terminología médica, sino también comenzar a formular hipótesis sobre qué aspectos de los datos podrían ser más relevantes para los modelos predictivos.

2. Exploración y Preprocesamiento de Datos: De la Observación a la Acción

Con una comprensión inicial del dominio, procedí a la inspección técnica de los datos, tanto tabulares como de imágenes.

2.1. Datos de Imágenes: Observaciones y Primeras Decisiones

La exploración de las imágenes MRI reveló varios puntos interesantes. Noté que, si bien la documentación del dataset no declaraba un preprocesamiento exhaustivo más allá del reescalado a 512×512 píxeles, una inspección visual sugería ciertas transformaciones adicionales:

- **Preprocesamiento Aparente:** Observé rotaciones in-plane y un posible padding o recorte para centrar las imágenes. Esto me llevó a pensar en los potenciales artefactos que podrían introducirse y cómo podrían influir en la inferencia del modelo, haciendo que se centre en características no determinantes del tumor.

- **Imágenes Duplicadas o Aumentadas:** Identifiqué la presencia de imágenes que parecían ser copias o versiones aumentadas dentro del dataset (ej. con "copy" en el nombre). Decidí filtrar estas imágenes para el análisis y modelado principal, con el fin de evitar posibles sesgos en el entrenamiento.
- **Variabilidad en Información Diagnóstica:** No todos los cortes (slices) de RM mostraban la lesión tumoral con la misma claridad; algunos incluso parecían carecer de información patológica relevante. Esto planteó la pregunta sobre la utilidad de cada imagen individual para el entrenamiento.

Curiosamente, investigando más a fondo, encontré otro conjunto de datos de la misma fuente que parecía más enriquecido y con transformaciones más explícitas. Esto me dio ideas y también un conjunto de datos externo (PMRAM: Bangladeshi Brain Cancer - MRI Dataset / DOI: 10.17632/m7w55sw88b.1) con el que, más adelante, pude realizar algunas pruebas informales de mi clasificador de imágenes.

2.2. Datos Tabulares: El Desafío de la Discriminación y el Poder de las Notas Clínicas

El Análisis Exploratorio de Datos (EDA) de los datos tabulares fue revelador.

- **Limpieza Inicial:** Identifiqué un pequeño porcentaje de filas con `Case ID` nulos (aproximadamente 0.73 %). Dado que un `Case ID` válido es crucial para la integridad del dato y su vinculación con las imágenes, estas filas fueron eliminadas. No se encontraron otros `Case ID` duplicados válidos.
- **Distribuciones Univariadas:**
  - La variable `Condition` (tipo de tumor) estaba notablemente balanceada entre las tres clases, lo cual es un excelente punto de partida para un modelo de clasificación.
  - La variable `Sex` mostró una distribución bastante equilibrada.
  - La `Age` (edad) mostró una distribución similar entre los diferentes tipos de tumores y, como veríamos más adelante, también entre los diferentes tratamientos.
- **Las Notas Clínicas:** Sabía que la columna `Clinical Note` contenía información rica y potencialmente discriminatoria. Mi primera inclinación fue considerar modelos de Procesamiento de Lenguaje Natural (NLP). Sin embargo, al inspeccionar detenidamente el texto, noté una estructura sintáctica tremendamente consistente: descripción de síntomas, seguida de una referencia a la duración/persistencia y, finalmente, una indicación de la intensidad global.

  Esta observación me llevó a optar por una estrategia de extracción basada en **expresiones regulares y heurísticas**. Aunque un modelo NLP podría ser más robusto ante variaciones no contempladas, el enfoque adoptado permitió una extracción rápida y efectiva de:
  - *Síntomas:* Se identificaron y normalizaron (ej. plurales a singular). Luego, síntomas semánticamente similares como [‘speech difficultie‘, ‘speech difficulty‘, ‘speech problem‘] fueron agrupados bajo una etiqueta canónica como ‘speech_issue‘.

- *Intensidad Global:* Se asignó un valor numérico basado en calificativos como "mild", "moderate", "high-grade".
- *Duración:* Se estimó en días a partir de frases como "a few weeks" o "several months". Consideré crear una variable categórica con intervalos, pero me decidí por días continuos pensando en una mayor granularidad informativa.

Este proceso transformó el texto no estructurado en un conjunto de características cuantitativas, listas para el análisis.

2.3. Análisis Bivariado y Hallazgos Clave (Datos Tabulares Procesados)

Con las nuevas características de síntomas, profundicé en el análisis bivariado, especialmente en relación con la predicción del `treatment`.

- **`Condition` vs. Variables Demográficas/Clínicas:**
  - `Condition` vs. `Age` y `Sex`: Los perfiles de edad y sexo eran muy homogéneos entre los tipos de tumores, sugiriendo poco poder discriminatorio individual.
  - `Condition` vs. `duration_days`: ¡Aquí sí emergieron diferencias notables! Los meningiomas tendían a presentar la menor duración de síntomas, los gliomas una duración intermedia, y la categoría "Otros tumores" la mayor duración. Esto apuntaba a `duration_days` como un predictor potencialmente útil.
- **`Treatment` vs. Variables Demográficas/Clínicas:**
  - `Treatment` vs. `Age` y `Sex`: Nuevamente, las distribuciones eran bastante parejas.
  - `Treatment` vs. `duration_days`: Aunque se observaron algunas diferencias en las medianas (ej. quimioterapia asociada a duraciones más largas), la variabilidad era alta y las distribuciones se solapaban considerablemente.
- **Relación `Condition` - `Treatment`:** Este fue un hallazgo crucial. Ciertos tratamientos eran exclusivos o predominantemente usados para tipos de tumores específicos (ej. ausencia de 'close monitoring' para gliomas, ausencia de quimioterapia para meningiomas). Esto reforzó la idea de que `condition` sería una característica importante para predecir `treatment`.
- **Síntomas vs. `Condition` y `Treatment`:**
  - *Síntomas por `Condition`:* Se identificaron perfiles sintomáticos distintivos. Ciertos síntomas eran notablemente más prevalentes o incluso exclusivos de algunos tipos de tumores, y su intensidad mediana también variaba.
  - *Síntomas por `Treatment`:* Si bien se observaron algunas diferencias, la mayoría de los síntomas tenían alguna presencia en casi todos los tratamientos. Esto, sumado a los hallazgos anteriores, comenzó a generar la presunción de que los datos tabulares podrían ser poco discriminatorios para predecir `treatment` directamente, más allá de la influencia de `condition`.

Desde esta etapa del EDA, comencé a tener la presunción de que los datos tabulares disponibles, incluso después de la ingeniería de características, podrían ser **poco discriminatorios para predecir el tratamiento con alta precisión**.

3. Desarrollo de Modelos: Un Contraste de Rendimientos y Desafíos

Desde el inicio, mi estrategia contemplaba un enfoque de ensamblaje, combinando un modelo de visión para las imágenes MRI con un modelo tabular para los datos clínicos.

3.1. Modelo de Clasificación de Tumores (Imágenes MRI)

Para la clasificación de imágenes, opté por probar un modelo **EfficientNet** (específicamente, `efficientnet_b3a`). Casi de inmediato, observé un rendimiento muy bueno, tanto que incluso me hizo considerar la posibilidad de sobreentrenamiento. Trabajé en la regularización (ajustando `WeightDecay`, la tasa de aprendizaje, etc.) y en la calibración de las curvas de predicción. Aunque una menor cantidad de épocas parecía inicialmente más segura, la versión final, entrenada por unas quince épocas, mostró una pérdida que se estabilizaba bien tanto en el conjunto de entrenamiento como en el de validación.

Para validar de forma más robusta, implementé una pequeña sección de "Playground" en el notebook, que permitía cargar el modelo entrenado y clasificar hasta cinco imágenes descargadas de internet o seleccionadas localmente. Las pruebas con estas imágenes no vistas, junto con el buen rendimiento en el conjunto de test (holdout) y extraidas de internet, me dieron confianza en la capacidad de generalización del clasificador de imágenes.

El modelo EfficientNet demostró una alta capacidad para clasificar los tipos de tumores a partir de las MRI, alcanzando un rendimiento robusto tras un cuidadoso ajuste y validación.

3.2. Modelo de Recomendación de Tratamiento (Datos Tabulares y Ensamblaje)

Aquí es donde el proyecto presentó su mayor desafío. Sabiendo que la variable `condition` (predicha por el modelo de imágenes) sería una entrada clave para este segundo modelo, comencé mis pruebas con un **XGBoost**, utilizando todas las características tabulares procesadas. Los resultados iniciales fueron, francamente, modestos.

Esto me llevó a una fase intensiva de experimentación:

- **Exploración de Modelos Tabulares:** Probé Random Forest, LightGBM, e incluso redes neuronales simples.
- **Ingeniería de Características Avanzada:** Profundicé en diferentes transformaciones y creaciones de características, aunque sin un impacto drástico.
- **Integración de Embeddings de NLP:** Incluso exploré el uso de ClinicalBERT para generar embeddings a partir de las notas clínicas originales, reemplazando mi extracción basada en reglas, pero esto tampoco produjo una mejora significativa sobre el XGBoost inicial.
- **Balanceo de Clases:** Apliqué técnicas como SMOTE a la variable objetivo `treatment`.

- **Tuning de Hiperparámetros:** Realicé una búsqueda aleatoria exhaustiva en un espacio amplio de hiperparámetros para el XGBoost.

A pesar de estos esfuerzos (muchos de los cuales, por su exhaustividad, no incluí en el notebook final para mantener su claridad), el rendimiento del modelo de recomendación de tratamiento no mejoró sustancialmente. Llegué a un punto que interpreté como el "techo" de la información contenida en los datos tabulares disponibles para esta tarea específica.

Decidí, para los fines de esta prueba técnica, presentar el modelo XGBoost con el mejor rendimiento obtenido, reconociendo sus limitaciones. En un entorno real, este resultado impulsaría una búsqueda activa de datos adicionales o características más ricas para mejorar la predicción del tratamiento.

4. Retos y Aprendizajes Clave del Proceso

Este reto técnico ha sido una experiencia de aprendizaje sumamente valiosa. Más allá de los resultados de los modelos, me gustaría destacar algunos retos y reflexiones personales:

- **Estructuración y Calidad del Código:** Uno de los aspectos en los que puse especial atención fue en mejorar la estructura de mi código, moviendo lógica recurrente a funciones y clases en módulos `.py`. Si bien es un área en la que siempre hay espacio para crecer, este proyecto me impulsó a adoptar mejores prácticas de modularización y legibilidad. Fue un reto estimulante organizar el flujo de trabajo de una manera más profesional y mantenible.
- **Enfrentar el "Techo de los Datos":** La fase de optimización del modelo de recomendación de tratamiento fue particularmente intensiva. Experimentar con numerosas técnicas y no ver mejoras sustanciales puede ser frustrante, pero también es una lección importante en ciencia de datos: a veces, los datos disponibles simplemente tienen limitaciones inherentes. Reconocer esto y saber cuándo detener la optimización en favor de estrategias alternativas (como la búsqueda de más datos) es una habilidad importante.
- **La Importancia del Conocimiento del Dominio:** La investigación inicial sobre los tipos de cáncer y tratamientos fue fundamental. Aunque no soy un experto médico, esta comprensión me permitió interpretar mejor los datos, formular hipótesis más informadas y entender las implicaciones de los resultados.

5. Mirando Hacia Adelante y Motivación

Este proyecto ha reforzado mi interés en la aplicación de la inteligencia artificial en el sector salud y ha destacado áreas en las que estoy activamente buscando mejorar, como la profundización en buenas prácticas de desarrollo de software y la industrialización de modelos mediante servicios web.

A pesar de los desafíos y los "puntos flacos" identificados, estoy convencido de que poseo una base sólida y una gran capacidad de aprendizaje que pueden ser de valor para su equipo. Mi experiencia previa con datos y mi interés genuino por resolver problemas complejos , junto con una

gran pasión por trabajar en temas de salud, me hacen valorar mas esta posibilidad. Me entusiasman los retos que implican aprendizaje continuo y estoy seguro de que, de tener la oportunidad de unirme a su equipo, podría adaptarme rápidamente y contribuir significativamente.

Agradezco la oportunidad de participar en este proceso de selección y reitero mi profundo interés y entusiasmo por continuar y, con suerte, unirme a su equipo para enfrentar juntos desafíos mayores y aprender con ustedes.