UNIVERSIDAD DEL VALLE DE GUATEMALA
Facultad de Ingeniería

Proyecto 2

Análisis Exploratorio

Andy Fuentes, 22944
Gabriel Alberto Paz, 221087
Davis Alejandro Roldan, 22672
José Rodrigo Marchena, 22398

Luis Roberto Furlan
Data Science

Repositorio

https://github.com/Andyfer004/P2-DS

Introducción

Los avances en la ciencia de datos han permitido abordar problemáticas cada vez más complejas de diversas áreas de estudio. Con modelos convolucionales y redes neuronales, el análisis de imágenes y estadísticas se ha convertido en una realidad de nuestro tiempo, con modelos incluso capaces de realizar diagnósticos médicos con una alta precisión.

El conjunto de datos **Mayo Clinic STRIP AI Challenge** provee imágenes digitales de alta calidad de WSI (Whole Slided Image) correspondientes a la presencia de coágulos en el cerebro de diferentes pacientes. La clasificación del origen de estos trombos o trombos es tarea de los médicos y expertos en patología para guiar el tratamiento a futuro, sin embargo, la identificación del origen posee alta incertidumbre.

Por tanto, en este proyecto, se aplicarán los conocimientos adquiridos en este curso para la implementación de modelos de aprendizaje profundo con el fin de realizar un diagnóstico de las imágenes WSI mencionadas, para poder clasificar el origen de un accidente cardiovascular isquémico

Planteamiento de problemática

La problemática radica en la clasificación binaria de dos posibles orígenes de un coagulo cerebral. En la práctica, se obtienen capturas WSI que permiten observar la presencia de fibrina y células rojas en el cerebro para localizar e identificar el coagulo por sus manifestaciones en las siguientes 2 categorías

- **Cardioembólico (CE):** Implica un origen del corazón, que puede requerir el uso de anticoagulantes
- **Ateroesclerosis de arteria grade (LA):** formado en las arterias principales, y puede sugerir el tratamiento de antiagregantes

Según estudios clínicos e imagenológicos, el origen de un trombo puede quedar indeterminado o tener una alta incertidumbre. Por tanto, se plantea el uso de redes neuronales y modelos de aprendizaje profundo para poder realizar esta clasificación.

Objetivos

Para esta entrega, se tiene como objetivo desarrollar un análisis exploratorio de datos (EDA) sobre las imágenes histopatológicas de trombos, con el fin de identificar patrones y características relevantes que preparen la base para un modelo de clasificación CE vs LAA.

- *Objetivos específicos*
  1. Describir la estructura del dataset, incluyendo número de casos, distribución por clase y metadatos asociados.
  2. Implementar técnicas de preprocesamiento para filtrar tiles no informativos (fondo, blur, bajo tejido) y normalizar la tinción.
  3. Explorar métricas básicas de color, textura y cobertura de tejido para cada clase.
  4. Visualizar los datos mediante histogramas, diagramas de caja y bigote y reducciones de dimensionalidad (UMAP/t-SNE) para detectar posibles separaciones entre clases.

Investigación Preliminar

- *Contexto Medico:*

  El **accidente cerebrovascular isquémico (ACV)** ocurre cuando un coágulo interrumpe el flujo sanguíneo al cerebro. Existen varias causas, pero dos de las más frecuentes son: **Cardioembólico (CE)**, que consiste en coágulos que se originan en el corazón, por ejemplo, a raíz de fibrilación auricular o insuficiencia cardíaca. Así como también **Ateroesclerosis de arteria grande (LAA)**, donde trombos formados en arterias cerebrales principales debido a placas ateroscleróticas.

  La distinción entre ambos es esencial, ya que determina el tratamiento: anticoagulación para CE vs. antiagregación y control de factores de riesgo para LAA.

- *Imágenes histopatológicas:*

  En el reto se utilizan **Whole Slide Images (WSI)** de trombos extraídos mediante trombectomía. Estas imágenes se obtienen con tinción **Hematoxilina & Eosina (H&E)**, que resalta estructuras celulares:

  - Hematoxilina: tiñe núcleos de color azul/púrpura.
  - Eosina: tiñe citoplasma y componentes extracelulares de rosado.

- *Relevancia del análisis computacional*

El análisis manual de WSI es complejo y requiere mucho tiempo de especialistas. Aquí entra la ciencia de datos, que nos permite implementar una forma estructurada de analizar y extraer información de estas imágenes

Por ejemplo, se considera el uso de fragmentación de las imágenes en tiles para su estudio, y el uso de técnicas de preprocesamiento y normalización para eliminar el ruido

También, esto facilita la extracción de características cuantitativas y la generación de representaciones que faciliten la clasificación con modelos de Vision por Computadora

Descripción de los Datos

El conjunto de datos de entrenamiento consta 754 imágenes de WSI asociadas a un diagnóstico de un accidente cerebrovascular. La implementación se basa en 2 subconjuntos separados: las imágenes en sí, y la meta data o información asociada a ellas.

Metadata / Informacion adicional:

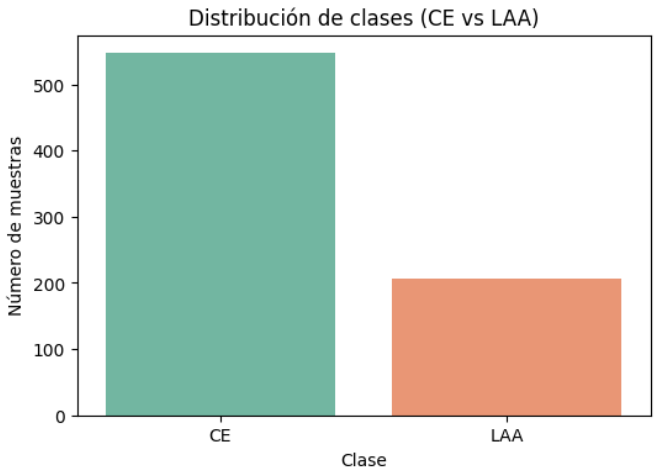
- **Image_id:** cadena de caracteres que identifica a una imagen, siendo una concatenacion del id el paciente y del número de imagen
- **Center id:** número identificador del centro médico donde se realizó la toma el WSI
- **Patient Id:** cadena hexadecimal de 6 caracteres que identifica a los pacientes
- **Image Num:** cantidad de imágenes asociadas a un paciente, indexadas desde cero. Es decir, 0 significa una imagen, 1 significa 2 imágenes, etc.
- **Label:** clasificación del diagnóstico de origen del coagulo, como se mencionó anteriormente puede ser CE o LAA.

Como se puede observar, mucha de la información es identificadora y no realmente descriptiva, pues el análisis se realizará principalmente en las imágenes asociadas a un paciente.

En las numéricas, podemos ver que el identificador del centro médico es un numero entero que varía entre 1 y 11, pero igualmente es parte de una asignación. Asimismo, el número de imágenes se mantiene en una media menor a 1, aunque existe un paciente con 4 imágenes asociadas a su diagnóstico.

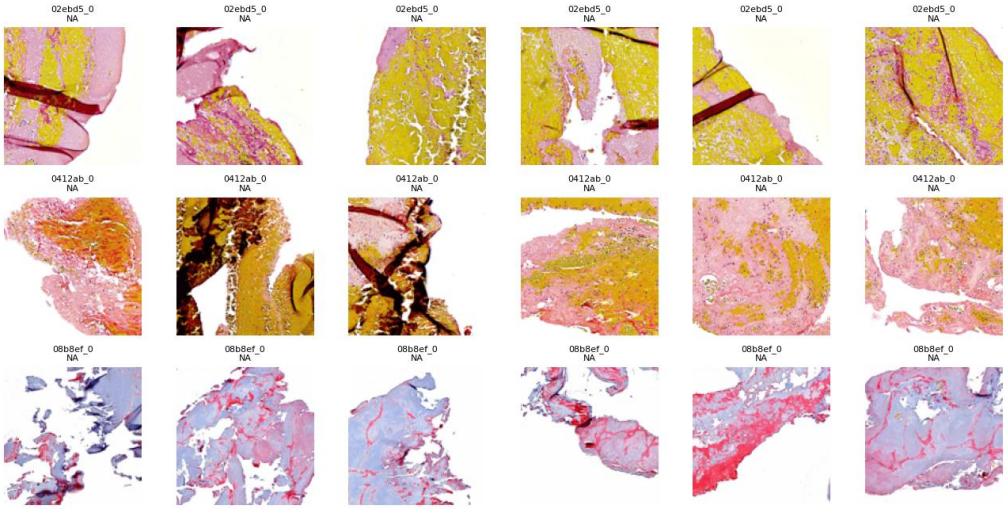
	center_id	image_num
count	754.000000	754.000000
mean	7.115385	0.226790
std	3.504306	0.599046
min	1.000000	0.000000
25%	4.000000	0.000000
50%	7.000000	0.000000
75%	11.000000	0.000000
max	11.000000	4.000000

Con respecto a la clasificación de los diagnósticos, podemos observar el siguiente grafico de barras, donde se puede observar que cerca de 500 observaciones (representativas del 70% de nuestro conjunto de datos) es perteneciente al diagnóstico CE, mientras 200 son de LAA.



Imágenes:

Las imágenes del conjunto de datos están identificadas por un nombre el cual es la concatenación del identificador de paciente con el número de imagen. Estas están en formato png con canales rojo, verde y azul.



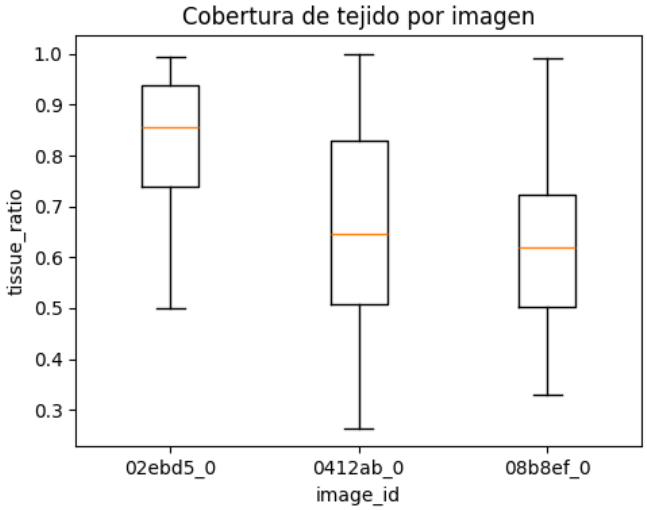
Análisis exploratorio

En el manejo de imágenes, especialmente aquellos correspondientes a análisis de patologías, se consideran diferentes aspectos relevantes para el análisis de las muestras. A

continuación, se muestra un análisis exploratorio de las fotografías del conjunto de datos con un enfoque en las tendencias y comportamientos de estas mismas

Cobertura de tejido

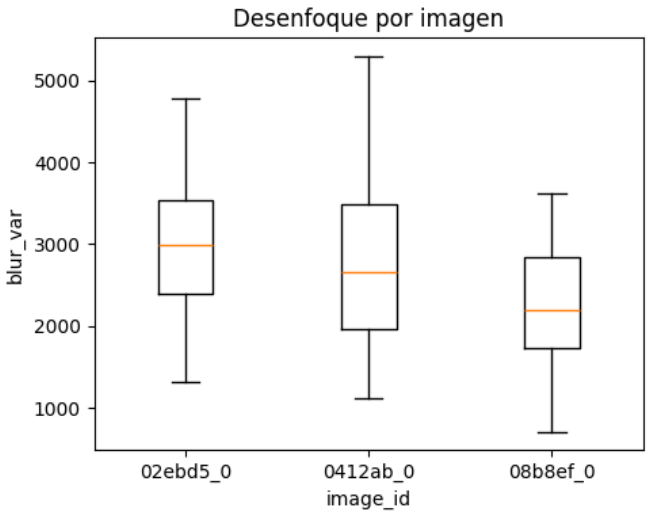
la cobertura de tejido indica que porcentaje de una imagen está compuesto por información relevante sobre el tejido. Dado que se manejan formatos PNG, la diferenciación con el fondo se puede realizar de manera sencilla



Es posible observar en las pruebas realizadas en 3 imágenes aleatoriamente elegidas, que la cobertura del tejido tiende a ser cercana al 60% dadas las medianas de los diagramas. Sin embargo, existen varias entradas con una muy baja cobertura de tan solo 30%.

Desenfoque

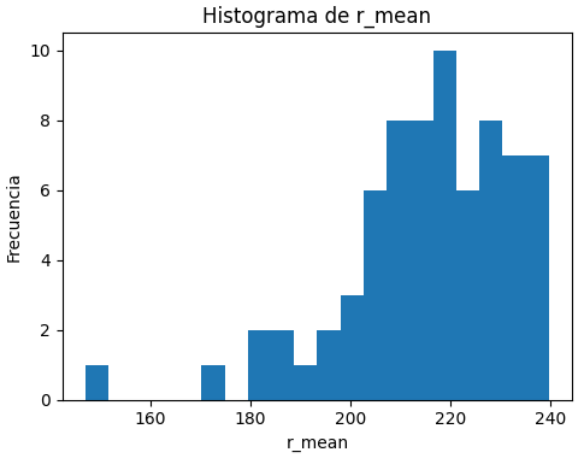
En visión por computadora, es también posible identificar el índice de desenfoque de una imagen, debido a que una imagen con bordes poco definidos puede conllevar a incertidumbre en las predicciones y una peor diferenciación de razgos. Utilizando el índice de varianza de desenfoque a partir del laplaciano de los canales de color, es posible observar la siguiente distribución.

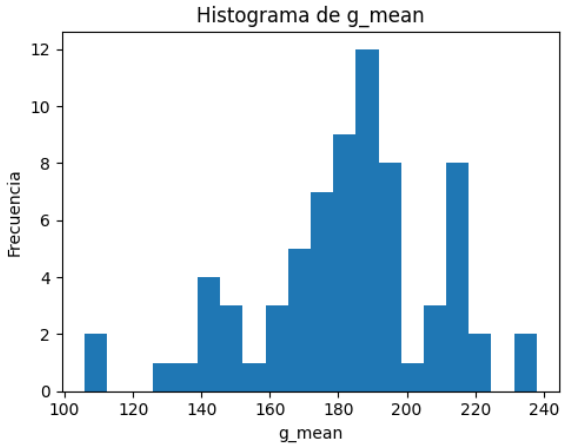


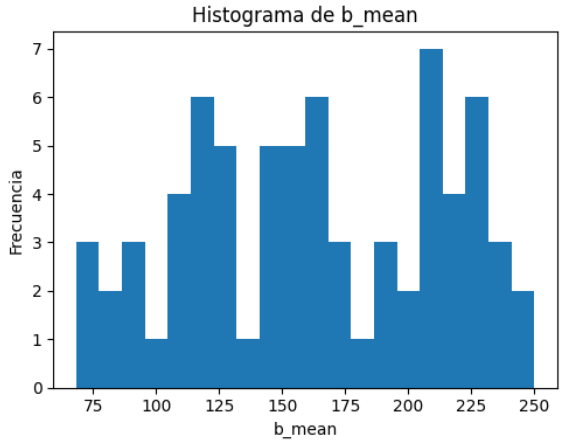
El desenfoque de estas imágenes tiende a estar en su mayoria por encima del 1000. Para contextualizar esta métrica, valores superiores al 100 indican una alta definición y nitidez de bordes. Por tanto, se considera que estas imágenes tienen un alto índice de enfoque

Presencia en canales RGB

Considerando la distribución de frecuencia para los canales de rojo, verde y azul para las imágenes, podemos observar ciertos comportamientos interesantes. En la distribución de rojo, existe un sesgo hacia la derecha, lo cual muestra una acumulación de valores altos en este canal. Asimismo, el histograma de verde parece tener una moda más cercana al medio, lo que nos indica una distribución unimodal este canal. Por último, el azul parece ser más uniforme, con una presencia alta en todas las imágenes y muchas modas.

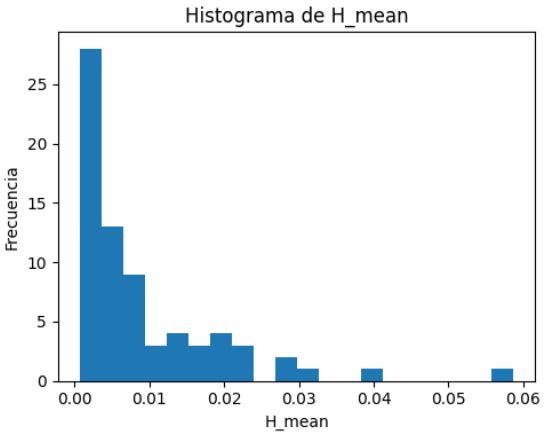


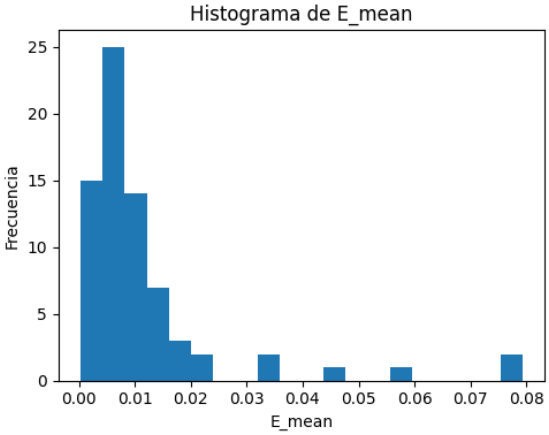




Presencia en canales Hematoxilina y Eosina

En el contexto de imagenes de patologia, se puede hacer una correlacion directa entre los canales azul y rojo con la pesencia de los tintes de hematoxilina y eosina las cuales identician la presencia de nucleo y cuerpo celular, respectivamente.

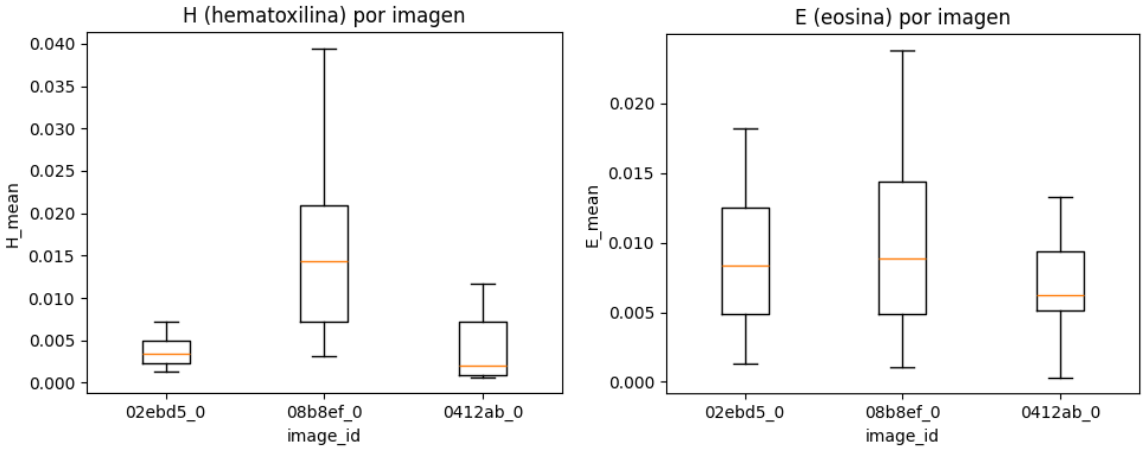




En las distribuciones obtenidas, es posible observar que muchas áreas tienen un bajo índice de estas dos métricas, estas podrían atribuirse al espacio vacío de las mismas imágenes, pues indican una falta completa de presencia celular. Asimismo, la mayor anchura de eosina representa la mayor cobertura de espacio del cuerpo celular comparado con su núcleo

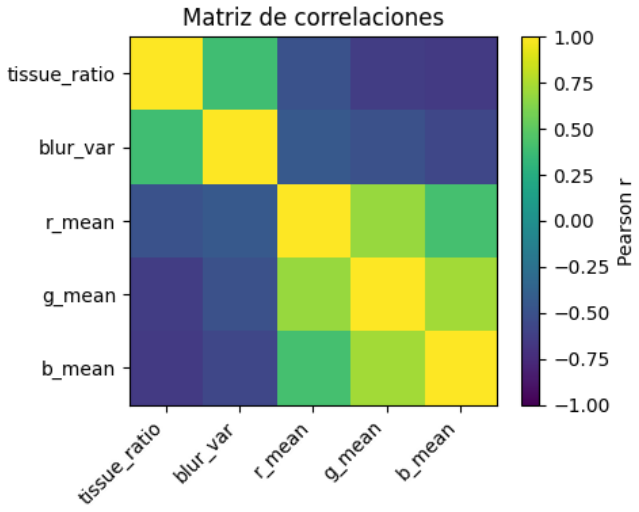
En este caso, el enfoque debería de ser dirigido a los puntos e instancias atípicas de estos dos tintes, pues una distribución poco equivalente de estos puede ser utilizadas para detectar formas anormales de las células, organización de los tejidos o patrones de enfermedades.

En los diagramas de caja y bigotes evaluados en 3 imágenes aleatorias mostrados a continuación, se puede ver que hay una cierta concordancia entre ambas métricas. Usualmente con una mayor presencia de hematoxilina se puede esperar una correspondencia de eosina. También se puede observar que la distribución de los cuartiles tiende a ser poco equitativa en algunas imágenes, lo que nos puede ayudar a identificar problemas previos a evaluar una imagen.



Análisis correlacional

Por último, se considera un analisis correlacional de las diversas métricas evaluadas hasta el momento. Se puede observar una correlación presente entre los canales de color, especialmente entre verde y azul, lo cual se puede correlacionar con la presencia de los tintes mencionados anteriormente, asi como con la intensidad de un pixel dado. También, la cobertura de tejido se ve relacionada con la nitidez de una imagen, lo cual nos indica que el contenido de tejido de una imagen tiende a tener alto enfoque una vez se presenta.



Conclusiones

Analisis Exploratorio

- Se desarrollo un pipeline de extracción de tiles informativos sobre la presencia del tejido y el enfoque o nitidez de las imágenes, en conjunto con información sobre la tinción en canales RGB y presencia de H&E (Hematoxilina y Eosina)
- Se observaron diferencias consistentes de cobertura y nitidez en las imagenes evaluadas
- La presencia de H&E concuerda con la distribución esperada, con cola a la derecha, así como también la dominancia de eosinofílicos en varias regiones

Implicaciones para modelado

- Además del uso de las imágenes en crudo, se considera la posiblidad y beneficio que conlleva el uso de características H&E, enfoque y cobertura para el muestreo de tiles en la futura etapa de modelado

Limitaciones y recomendaciones

- El conjunto de datos posee en ciertas instancias la falta de etiquetas para algunas de las imagenes
- El tamaño reducido del conjunto de imágenes con tan solo 750 entradas, algunas para pacientes recurrentes, puede llegar a ser una limitación a la hora de realizar modelos de aprendizaje profundo. Se recomienda el uso de aumentaciones de data para extraploar la mayor cantidad de información del conjunto de datos.
- Las imágenes de muy alta resolución también pueden llegar a limitar el aprendizaje de un modelo, por lo que se reconoce el beneficio de aplicar previews o downsapling