

## ACTIVIDADES

1. Haga una pequeña investigación del tema para que tenga idea de qué buscar en un análisis exploratorio. En el caso de los problemas médicos, describa la enfermedad a detectar, los síntomas y como se diagnostica (especialmente diagnóstico basado en imágenes). Esto le va a servir para entender cual es el patrón que deben reconocer los algoritmos. En el caso de problemas de Procesamiento del Lenguaje Natural, investigue las técnicas que se usan para detectar patrones en lenguaje escrito. En cuanto al tema de las recomendaciones investigue que se necesita para hacer un sistema de recomendaciones.
2. Analice el problema planteado y los datos.
3. Describa las tareas de limpieza y preprocesamiento que llevó a cabo.
4. Haga un análisis exploratorio de los datos:
  - a. Comience describiendo cuantas variables y observaciones tiene disponible, el tipo de cada una de las variables.
  - b. Haga un resumen de las variables numéricas y tablas de frecuencia para las variables categóricas, escriba lo que vaya encontrando, si aplica.
  - c. Cruce las variables que considere que son las más importantes para hallar los elementos clave que lo pueden llevar a comprender lo que está causando el problema encontrado.
  - d. Haga gráficos exploratorios que le de ideas del estado de los datos.
5. Escriba unas conclusiones con los hallazgos encontrados durante el análisis exploratorio

## EVALUACIÓN

**NOTA:** La evaluación de cada integrante del grupo será de acuerdo con sus contribuciones al trabajo grupal

- **(10 puntos) Situación Problemática:** Describe la situación problemática que da lugar al problema.
- **(10 puntos). Problema científico:** Se enuncia el problema científico que se desprende de la situación planteada. Se comprende bien cuál es el problema.
- **(10 puntos). Objetivos:** Se plantean los objetivos a cumplir para darle solución al problema planteado. Se enuncia al menos un objetivo general y 2 específicos. Los objetivos deben ser medibles y alcanzables durante la investigación.
- **(20 puntos). Descripción de los datos:** Se describen los datos, tanto las variables y observaciones como las operaciones de limpieza que se le hicieron si fueron necesarias.
- **(30 puntos). Análisis Exploratorio:**
  - o Estudia las variables cuantitativas mediante técnicas de estadística descriptiva

DE GUATEMALA

CC3084 – Data Science

- o Hace gráficos exploratorios como histogramas, diagramas de cajas y bigotes, gráficos de dispersión que ayudan a explicar los datos
- o Analiza las correlaciones entre las variables, trata de explicar los outliers (puntos atípicos) y toma decisiones acertadas ante la presencia de valores faltantes.
- o Estudia las variables categóricas
- o Elabora gráficos de barra, tablas de frecuencia y de proporciones
- o Explica muy bien todos los procedimientos y los hallazgos que va haciendo.
- **(20 puntos). Hallazgos y conclusiones:**
  - o Hace un resumen de los hallazgos en el análisis exploratorio
  - o Llega a conclusiones sobre los siguientes pasos a seguir.

## MATERIAL A ENTREGAR

## **Situación Problemática**

Según la OMS, las enfermedades cerebrovasculares -ECV- son fenómenos agudos que se deben principalmente a obstrucciones que imposibilitan la circulación de la sangre hacia el corazón y el cerebro (Organización Mundial de la Salud -OMS-, 2017). Las ECV son la tercera causa de muerte, la segunda causa de discapacidad en adulto y la segunda causa de demencia en el mundo (Bleder, 2019).

Se estima que alrededor de 15 millones de personas sufren un ictus (derrame cerebral, embolia, trombosis o apoplejía) anualmente en el mundo, de las cuales 5.5 millones mueren y 5 millones más quedan con alguna discapacidad permanente. Así, en el mundo la incidencia promedio de las enfermedades cerebrovasculares es de 200 casos por cada 100.000 habitantes anualmente (Blender, 2019).

Según MayoClinic, las enfermedades cerebrovasculares pueden ser isquémicas o hemorrágicas, de acuerdo con sus características fisiopatológicas:

- Accidente cerebrovascular Isquémico: Este es el tipo más común de accidente cerebrovascular. Ocurre cuando los vasos sanguíneos del cerebro se estrechan o se bloquean, lo que causa una importante reducción del flujo sanguíneo (isquemia). Los vasos sanguíneos se bloquean o se estrechan debido a la acumulación de depósitos de grasa o de coágulos sanguíneos u otros desechos que se desplazan por la sangre (con más frecuencia desde el corazón) y se alojan en los vasos sanguíneos del cerebro. Algunas investigaciones iniciales muestran que la infección por COVID-19 puede aumentar el riesgo de sufrir un accidente cerebrovascular isquémico, pero se requieren más estudios.
- Accidente cerebrovascular hemorrágico: El accidente cerebrovascular hemorrágico ocurre cuando un vaso sanguíneo en el cerebro gotea o se rompe. Las hemorragias cerebrales pueden ser el resultado de muchas afecciones que afectan los vasos sanguíneos. Los factores relacionados con el accidente cerebrovascular hemorrágico incluyen los siguientes:
  - Presión arterial alta no controlada
  - Sobretratamiento con anticoagulantes
  - Bultos en puntos débiles de las paredes de los vasos sanguíneos (aneurismas)
  - Traumatismo (como un accidente automovilístico)
  - Depósitos de proteína en las paredes de los vasos sanguíneos que provocan debilidad en la pared del vaso (angiopatía amiloide cerebral)
  - Accidente cerebrovascular isquémico que causa hemorragia

De acuerdo con el Centros para el Control y la Prevención de Enfermedades -CDC-, los síntomas de un accidente cerebrovascular:

Uno de los síntomas que puede manifestarse en los individuos es el desafío de expresarse verbalmente y comprender lo que comunican los demás. Esto puede resultar en un estado de desorientación, dificultad para enunciar palabras o dificultad para captar el significado del lenguaje hablado. Además, el usuario puede experimentar una pérdida de sensación o movimiento en la cara, el brazo o la pierna, lo que se conoce como parálisis o entumecimiento. Puede ocurrir una aparición repentina de entumecimiento, debilidad o parálisis en un solo lado del cuerpo.

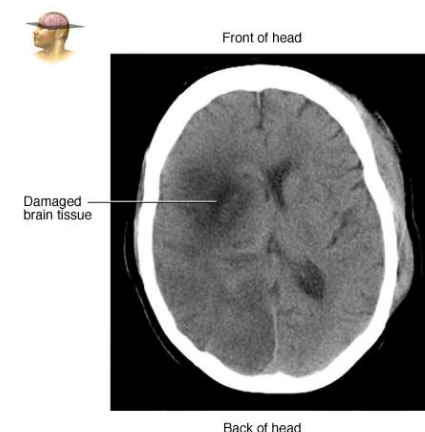
También uno de los posibles síntomas de un problema relacionado con los ojos es la dificultad para ver, ya sea en un ojo o en ambos. Esto podría manifestarse como una pérdida repentina de claridad, lo que resulta en una visión borrosa u oscurecida en uno o ambos ojos, o la percepción de visión doble.

Si experimenta un dolor de cabeza intenso y repentino, también puede experimentar síntomas que lo acompañan, como náuseas, vértigo o incluso desmayos. Estos síntomas pueden indicar que actualmente o que en ese preciso momento está sufriendo un derrame cerebral.

Por último, uno de los principales síntomas de los problemas de movilidad es la dificultad para caminar, que puede manifestarse como tropiezos o pérdida del equilibrio. Además, quienes experimentan estos problemas también pueden sufrir pérdida de coordinación.

Según MayoClinic el diagnóstico de un accidente cardiovascular consta de algunas pruebas que incluyen lo siguiente:

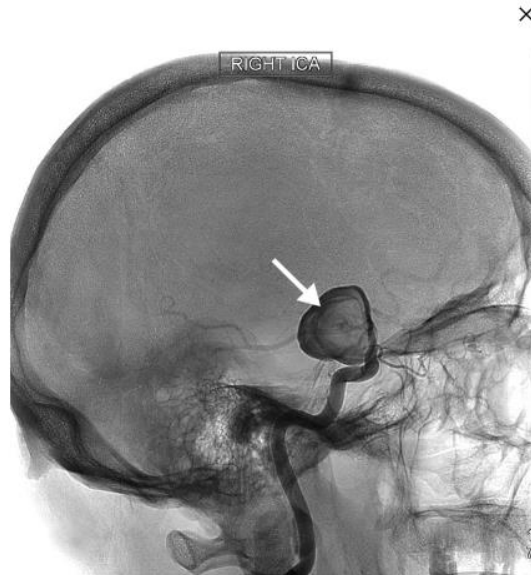
- Examen físico
- Análisis de sangre: En esta prueba se verifica la velocidad a la que coagula la sangre, si existe alguna infección y los niveles de glucosa.
- Tomografía computarizada: Esta prueba puede mostrar sangrado en el cerebro, un accidente cerebrovascular isquémico, un tumor u otras afecciones.



© MAYO FOUNDATION FOR MEDICAL EDUCATION AND RESEARCH. ALL RIGHTS RESERVED.

Tomografía computarizada de tejido cerebral dañado por accidente cerebrovascular

- Imágenes por resonancia magnética: Esta prueba se detecta si el tejido cerebral está dañado por un accidente cerebrovascular isquémico y por hemorragias cerebrales.
- Angiografía cerebral: En esta prueba el médico inyecta un tinte en los vasos sanguíneos para hacerlos visibles mediante imágenes por rayos X.



Angiografía cerebral

Una angiografía cerebral que muestra un aneurisma de la arteria carótida asociado con un accidente cerebrovascular.

- Ecocardiograma: Esta prueba puede encontrar el origen o una fuente de coágulos en el corazón que pueden haberse desplazado desde el corazón hasta el cerebro y haber provocado un accidente cerebrovascular.

## **Problema científico**

Luego del primer ictus, puede ocurrir uno segundo empeorando las posibilidades de supervivencia del paciente, sin embargo, esto podría mitigarse si los médicos pudieran determinar la etiología del accidente cerebrovascular, lo que influye en el manejo terapéutico después de un accidente cerebrovascular (Kaggle, 2022).

Durante la última década, la trombectomía mecánica se ha convertido en el tratamiento estándar para el accidente cerebrovascular isquémico agudo por oclusión de grandes vasos. Como resultado, los coágulos recuperados se volvieron susceptibles de análisis. Se han intentado aplicar métodos basados en el aprendizaje profundo para predecir la etiología del accidente cerebrovascular isquémico y el origen del coágulo. Lamentablemente, los formatos de datos únicos, los tamaños de archivos de imágenes y la cantidad de diapositivas de patología disponibles han dificultado solucionar el problema (Kaggle, 2022).

Para disminuir las posibilidades de sufrir accidentes cerebrovasculares posteriores, se debe desarrollar un modelo de inteligencia artificial altamente preciso y efectivo que pueda diferenciar de manera confiable entre las dos principales etiologías de los accidentes cerebrovasculares isquémicos agudos (AIS): la causa cardíaca y la aterosclerosis de arteria grande, utilizando imágenes completas de patología digital. Esto podría tener un gran impacto para salvar las vidas de los sobrevivientes de accidentes cerebrovasculares y ayudar a comprender mejor la segunda causa de muerte en el mundo.

## **Objetivos**

### **Objetivo general**

Desarrollar un modelo de inteligencia artificial efectivo y preciso que pueda distinguir entre las dos principales etiologías de los accidentes cerebrovasculares isquémicos agudos (AIS) - causa cardíaca y aterosclerosis de arteria grande - utilizando imágenes completas de patología digital, con el propósito de mejorar la identificación de los orígenes de los coágulos sanguíneos en accidentes cerebrovasculares mortales y facilitar la prescripción de un tratamiento terapéutico post-AIS más efectivo, reduciendo así la probabilidad de un segundo AIS

### **Objetivos específicos**

- Crear un conjunto de algoritmos de procesamiento de imágenes que permitan la extracción y caracterización de las características patológicas relevantes en las imágenes completas de patología digital relacionadas con los coágulos sanguíneos en AIS de origen cardíaco y aterosclerosis de arteria grande.
- Entrenar y evaluar un modelo de inteligencia artificial utilizando el conjunto de datos de imágenes completas de patología digital, con el objetivo de alcanzar una precisión mínima del 95% en la clasificación de la etiología de los coágulos sanguíneos en AIS, diferenciando con alta sensibilidad y especificidad entre las causas cardíacas y la aterosclerosis de arteria grande.

## Descripción de los datos

Se cuenta con 3 carpetas con imágenes distribuidas de la siguiente manera

- Train: imágenes a utilizar para el proceso de entrenamiento en formato TIFF.
- Test: 280 imágenes comprimidas para utilizar en el proceso de testeo.
- Other: suplement de imágenes con una etiología distinta a CE o LAA o desconocida.

Además, de 3 archivos en csv con la siguiente información

- Train: contiene 5 variables con 754 observaciones:
  - Image\_id: identificador único que se forma como {patient\_id}\_{image\_num} y corresponde a la imagen {image\_id}.tif.
  - Center\_id: identificador del centro médico donde se obtuvo la slide.
  - Patient\_id: identificador del paciente de quien se obtuvo la slide.
  - Image\_num: enumera las imágenes de los coágulos obtenidos por el mismo paciente.
  - Label: etiología del coágulo, es decir, si es CE o LAA.
- Test: contiene 4 variables, idénticas a las de Train, a excepción del label con 4 observaciones.
- Other: contiene 5 variables, idénticas a la de Train a excepción de Center\_id que es reemplazada por other\_specificied que, en caso se desconozca la etiología, se etiqueta como “otro”, cuenta con 396 observaciones

## Análisis exploratorio

La base de datos tiene las siguientes cinco variables, cuyos tipos son:

- image\_id: object
- center\_id: int64
- patient\_id: object
- image\_num: int64
- label: object

### Tablas de frecuencia

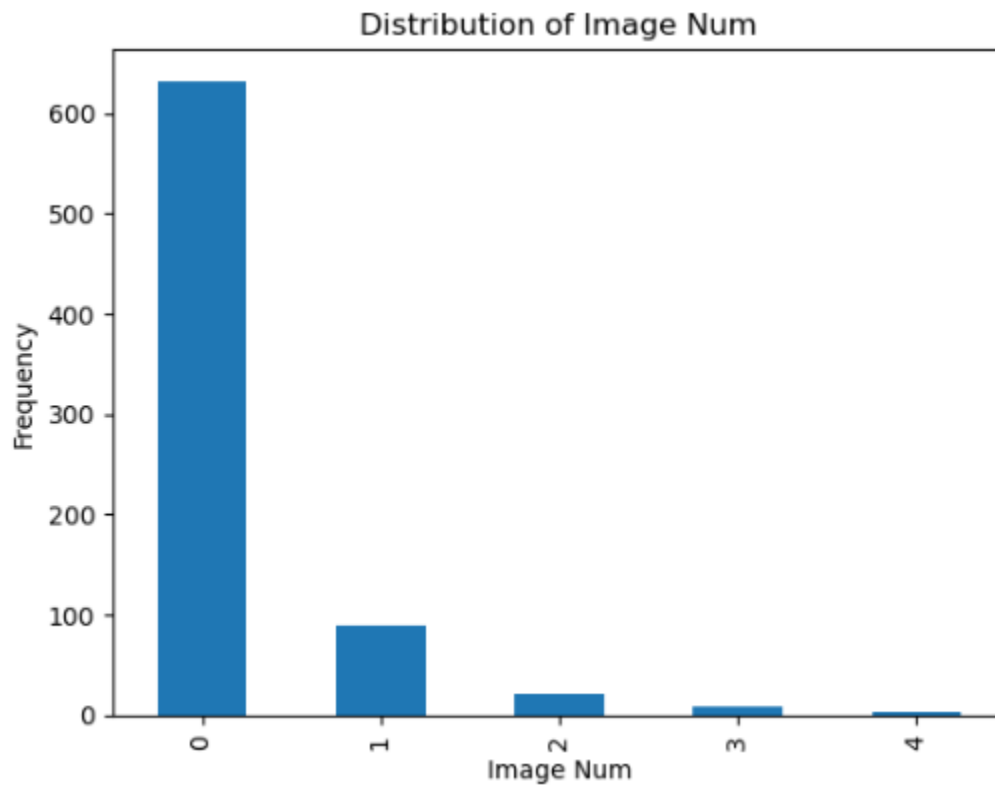
```
11    257
4     114
7      99
1      54
3      49
10     44
5      38
6      38
2      29
8      16
9      16
Name: center_id, dtype: int64
CE     547
LAA    207
Name: label, dtype: int64
```

### Tabla de contingencia

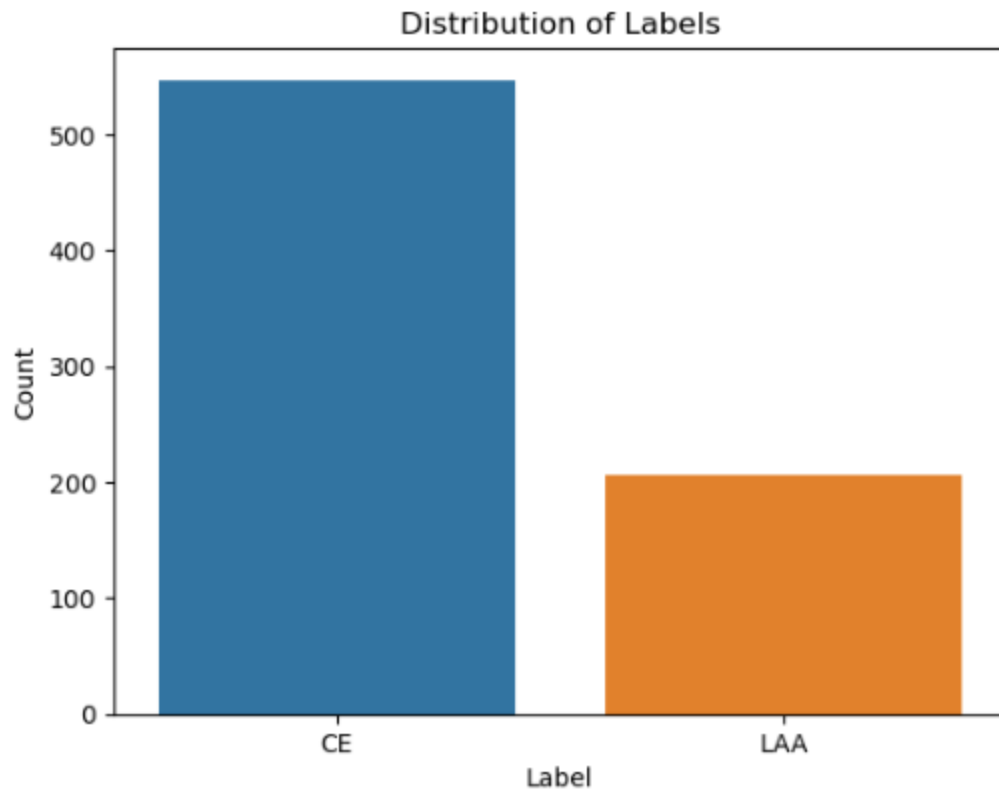
label	CE	LAA
center_id		
1	44	10
2	26	3
3	22	27
4	88	26
5	29	9
6	24	14
7	70	29
8	14	2
9	14	2
10	37	7
11	179	78



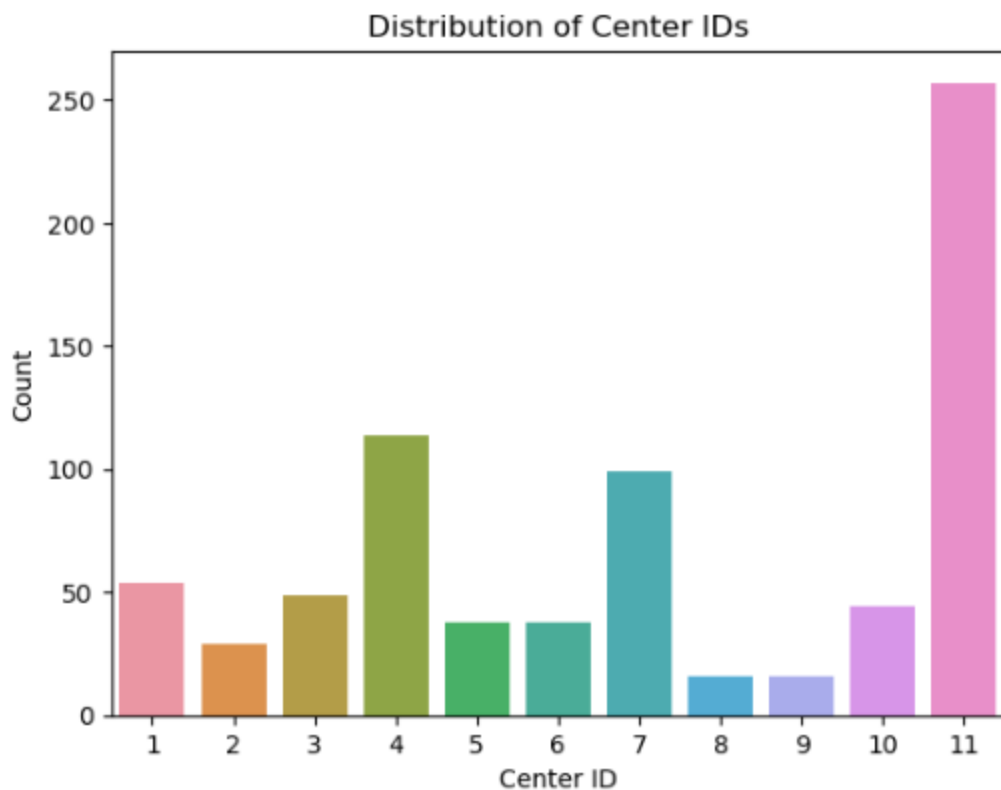
Primero estudiamos como se distribuyen las variables:



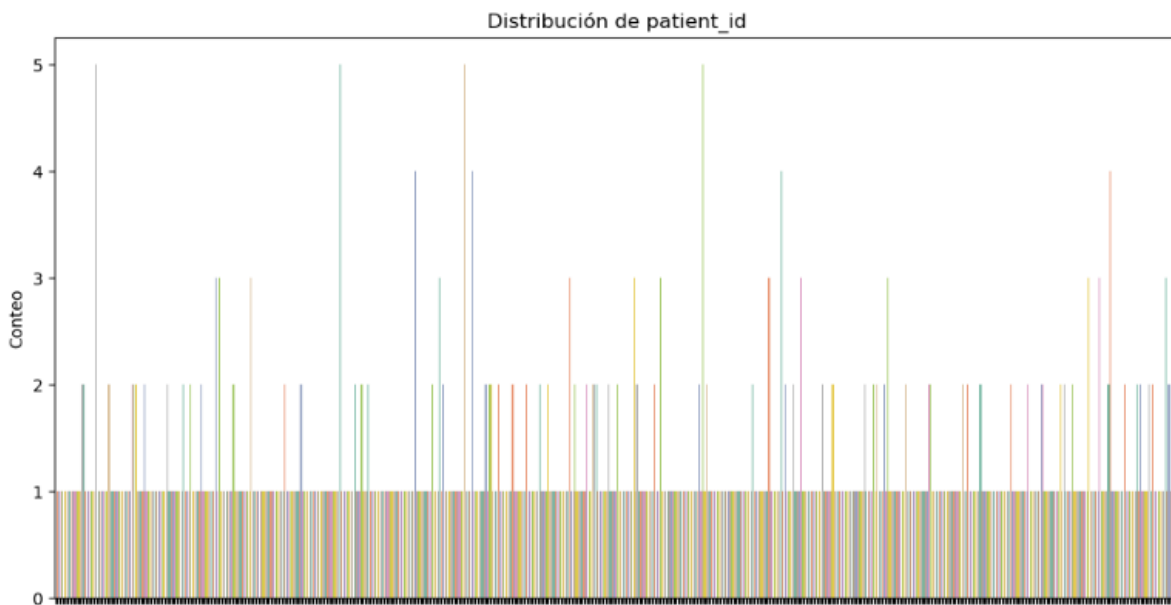
Se puede observar que la mayoría de los datos de la variable `image_num` son 0.

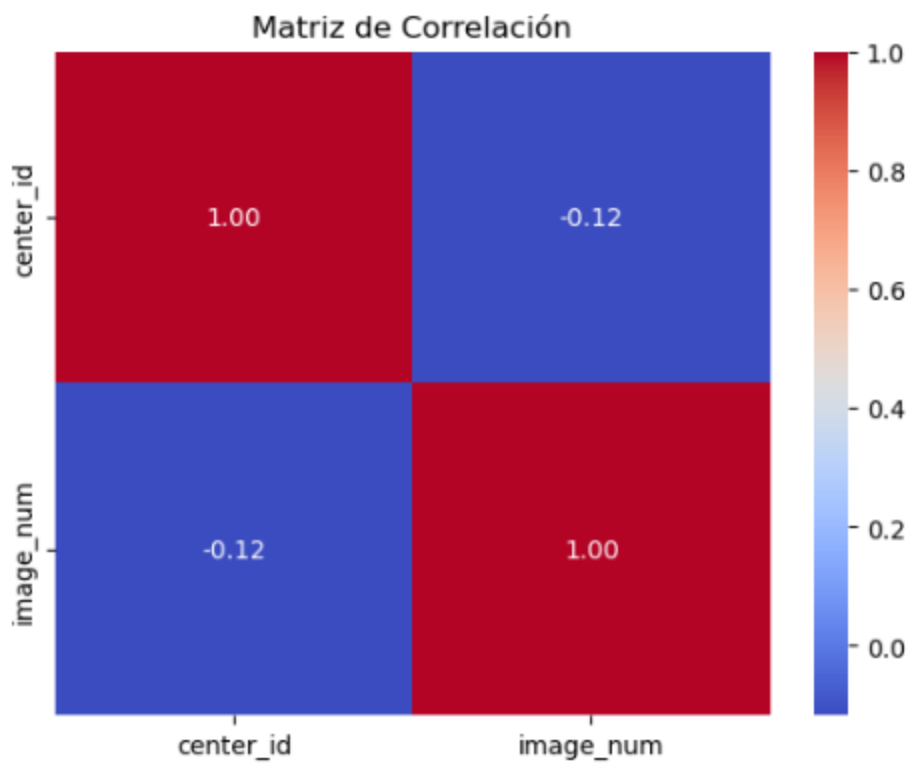


Vemos también que la mayoría de los datos tienen valor de label “CE”.

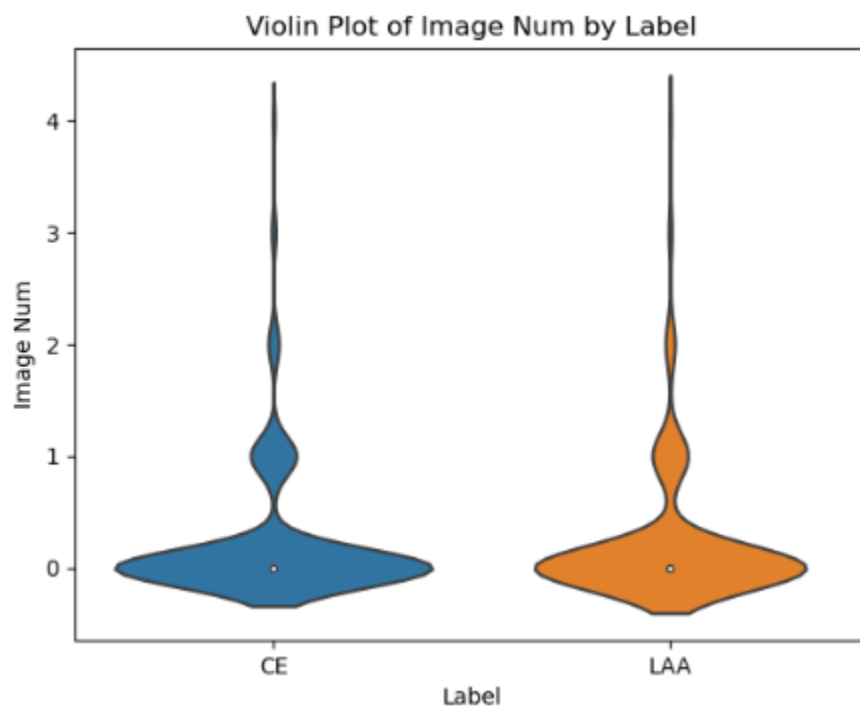


El centro que mas muestras contiene es el 11, mientras que los que menos tienen son el 8 y 9.

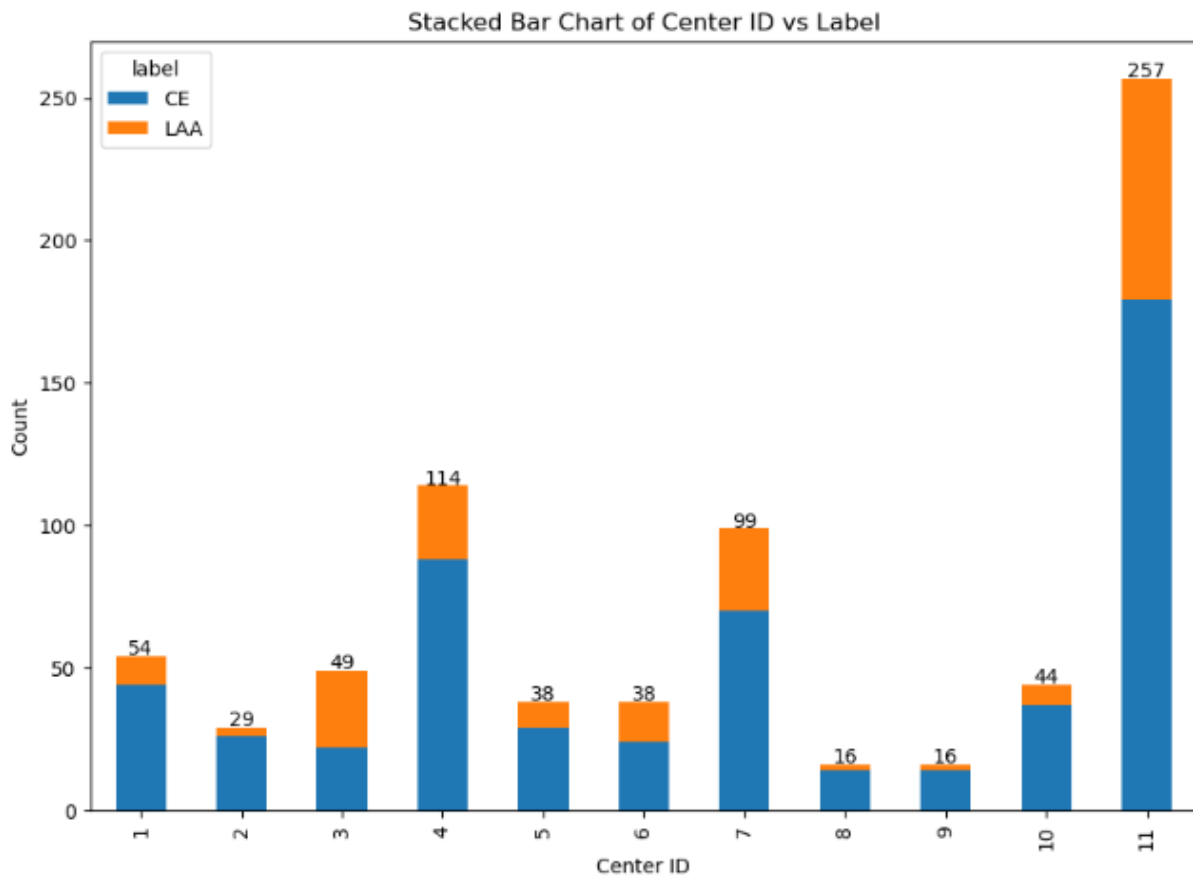




Al cruzar las variables vemos que no hay una correlación entre la variable `image_num` y la variable `center_id`, que son nuestras variables numéricas.



Entre image\_num y label vemos que en ambos valores de label se puede ver que los valores de image\_num tienen comportamientos muy similares, concentrando en ambos casos, la mayoría en el 0.



Entre las variables center\_id y label, vemos que en cada centro predomina la etiqueta “CEE”, y que en cada centro se encuentran ambas labels.



## Hallazgos y conclusiones

### Hallazgos:

Observando el histograma de *image\_num*, se puede determinar que la mayoría de los pacientes solo cuentan con una imagen del coágulo que presentan. Asimismo, existen varios pacientes que tienen el número máximo de imágenes asociadas a ellos dentro del *dataset* las cuales son 5. Dicho esto, al observar la concentración de la cantidad de imágenes por paciente y tipo de coágulo, estas siguen una distribución similar dando a entender que no existe una diferencia entre la cantidad de imágenes tomadas de un paciente dada la etiología.

Además, de todas las imágenes que pueden estar asociadas a un paciente, siempre se presenta el mismo tipo de coágulo, ya sea CEE o LAA, es decir que no hay pacientes que tengan dos tipos de coágulos o que presenten uno y luego el otro. No obstante, existe una gran diferencia en cuanto a la cantidad de personas que presentan CEE a comparación de LAA. La etiología CEE representa un 73% de todos los datos vistos mientras que la LAA tan solo presenta un 27%. Esta diferencia puede verse similar a lo largo de los distintos centros siendo CEE el que tiene la mayor prevalencia siempre.

La cantidad de pacientes u observaciones vistas dentro de los diferentes centros de salud varia en grandes cantidades. El centro que más muestras contiene es el número 11, superando al segundo lugar por casi el doble. Asimismo, los centros de salud que menos muestran presentan en el set de datos son el 8 y 9. Sin embargo, al verificar si existía una correlación entre la cantidad de imágenes por paciente y el número de centro, se encontró que no hay ninguna relación. Es decir, que no hay evidencia que en ciertos centros tomen más imágenes por paciente o que se hayan observado más imágenes de coágulos por cada uno de los pacientes.

### Conclusiones:

- Los coágulos con etiología cardioembólica (CEE) son más frecuentes que los coágulos con etiología aterosteclórica (LAA).
- Los pacientes que han presentado coágulos en más de una ocasión suelen presentar siempre el mismo las veces que se repiten. No se encontró ningún caso en que se presente uno y luego otro.
- El centro que más muestras contiene es el 11, mientras que los que menos tienen son el 8 y 9.

## Bibliografía

Organización Mundial de la Salud -OMS-. (2017). *Enfermedades Cardiovasculares*. <https://www.who.int/es/news-room/fact-sheets/detail/cardiovascular-disea...>(cvds)

Blender, J. (2019). *Las enfermedades cerebrovasculares como problema de salud* . Revista Cubana de Neurología y Neurocirugía, (2):e335.

MayoClinic. (2023). *Accidente cerebrovascular*. <https://www.mayoclinic.org/es/diseases-conditions/stroke/symptoms-causes/syc-20350113#:~:text=Un%20accidente%20cerebrovascular%20isquémico%20se,acumulación%20de%20placas%2C%20denominada%20ateroesclerosis>.

Centros para el Control y la Prevención de Enfermedade. (2023). *Signos y síntomas del accidente cerebrovascular*. [https://www.cdc.gov/stroke/signs\\_symptoms\\_esp.htm](https://www.cdc.gov/stroke/signs_symptoms_esp.htm)