Guía para de Proyecto 2.
Análisis Exploratorio

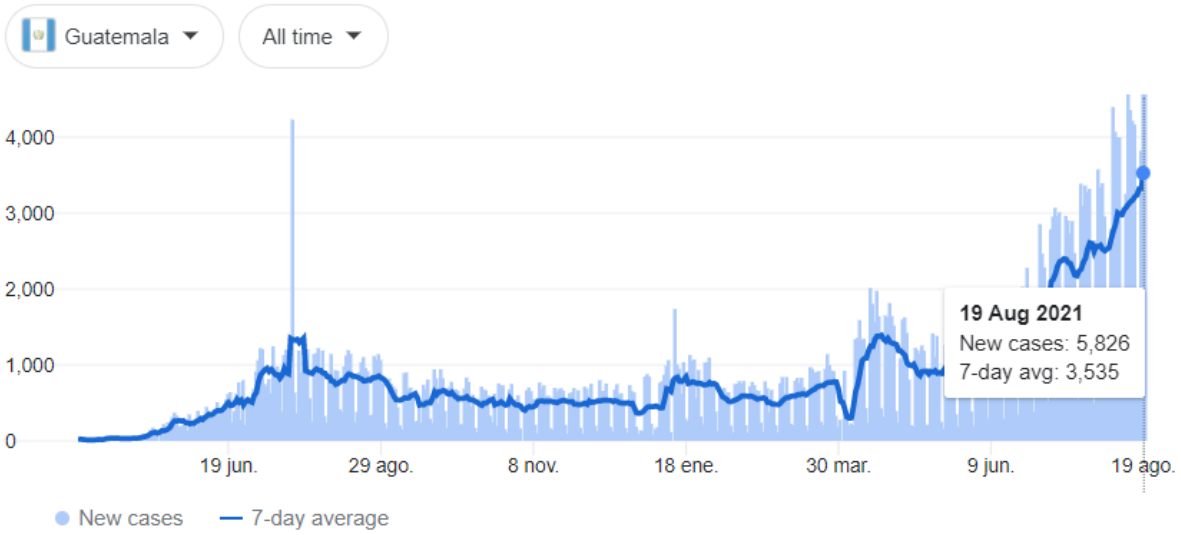
INTRODUCCIÓN

Dado que existen tantos problemas complejos que se pueden atacar con la Ciencia de Datos. Es por esto por lo que este año se plantean alternativas de proyecto. La actividad se realizará en grupos de 4 personas que deberán seleccionar el tema que trabajarán de los que aquí se plantean:

- Detección de covid en radiografías de tórax
- Clasificación radiogenómica de tumores cerebrales RSNA-MICCAI
- Procesamiento del lenguaje natural con tweets de desastres

Detección de covid en radiografías de tórax:

El virus SARS-COV 2 causante de la enfermedad COVID 19 ataca principalmente los pulmones causando neumonía viral. En una gran parte de los casos, los pacientes terminan en una unidad de cuidados intensivos con ventilación mecánica invasiva. En estos momentos en nuestro país se vive un incremento de casos asociados a la entrada de la variante delta, 60% más infecciosa que la alfa.

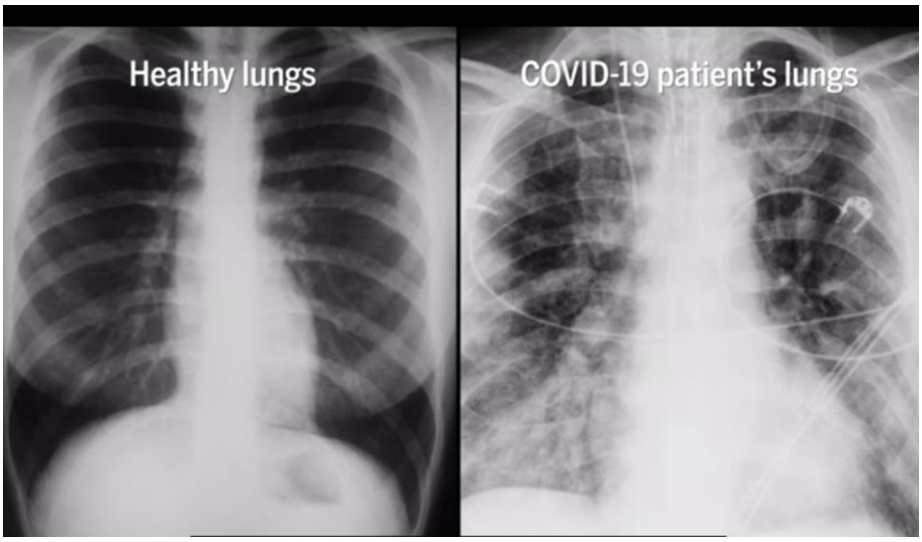


Las imágenes de neumonías producidas por virus son variadas y poco específicas, pudiéndose ver en otros procesos infecciosos o inflamatorios pulmonares. La radiografía simple de tórax muestra un patrón intersticial en más del 85% de los casos, y se presenta de forma bilateral en más del 75% Por otra parte, se ha relacionado la mayor afectación en la radiografía simple con la necesidad de ingreso en una unidad de cuidados intensivos.

Existen hallazgos en imágenes que obligan a realizar un diagnostico diferencial entre covid 19 y otras afecciones pulmonares como el Síndrome Respiratorio de Oriente Medio (MERS). En comparación con otros virus de la misma familia, parece existir una mayor tendencia a la afectación bilateral que

en el SARS (síndrome respiratorio agudo grave humano) y una ausencia de un patrón de distribución específico similar al MERS (González-Castro et al., 2020).

En la siguiente imagen se puede ver la diferencia en una radiografía de tórax de un paciente con covid de uno sano:



Fuente: Describen cómo el coronavirus SARS-CoV-2 afecta los órganos del cuerpo

No obstante, es posible que no se encuentren hallazgos específicos en las pruebas de imagen, por lo que no es una prueba válida para excluir pacientes con enfermedad, aunque las imágenes normales en la Tomografía Computarizada se han descrito en pocos casos incluso en estadios iniciales de la enfermedad (González-Castro et al., 2020).

Actualmente, COVID-19 se puede diagnosticar mediante la reacción en cadena de la polimerasa para detectar material genético del virus o una radiografía de tórax. Sin embargo, pueden pasar algunas horas y, a veces, días antes de que se obtengan los resultados de las pruebas moleculares. Por el contrario, las radiografías de tórax se pueden obtener en minutos. Si bien existen pautas para ayudar a los radiólogos a diferenciar el COVID-19 de otros tipos de infección, sus evaluaciones varían. Además, los no radiólogos podrían recibir apoyo con una mejor localización de la enfermedad, como con un cuadro delimitador visual.

La Sociedad de Informática por Imágenes en Medicina (SIIM) promueve la informática de imágenes médicas a través de la educación, la investigación y la innovación. En sociedad con Fundación para el Fomento de la Salud y la Investigación Biomédica de la Comunitat Valenciana (FISABIO), el Banco de Datos de Imágenes Médicas de la Comunitat Valenciana (BIMCV) y la Sociedad Radiológica de Norteamérica (RSNA) han provisto este conjunto de datos. El objetivo es identificar y localizar anomalías causadas por covid-19 en radiografías de tórax.

Descripción del Conjunto de datos:

El conjunto de datos está separado en conjunto de entrenamiento (train_image_level.csv, train_study_level.csv) y conjunto de prueba. Para este proyecto se trabajará únicamente con los archivos de entrenamiento.

El conjunto de entrenamiento comprende 6,334 exploraciones de tórax en formato DICOM, que se anonimizaron para proteger la privacidad del paciente. Todas las imágenes fueron etiquetadas por un panel de radiólogos experimentados para la presencia de opacidades, así como la apariencia general.

Archivos:

- train_study_level.csv: Contiene los metadatos, cada observación corresponde a un estudio, incluyendo las etiquetas correctas
    - variables:
        - id: identificador único del estudio
        - Negative for Pneumonia: -1 si es negativo para neumonía, 0 en caso contrario
        - Typical Appearance: -1 si el estudio tiene una apariencia típica, 0 en caso contrario
        - Indeterminate Appearance: -1 si el estudio tiene una apariencia indeterminada, 0 en caso contrario
        - Atypical Appearance: -1 si el estudio tiene una apariencia atípica, 0 en caso contrario
- train_image_level.csv: Contiene los metadatos de nivel de imagen del conjunto de entrenamiento, con una fila para cada imagen, incluidas las etiquetas correctas y los cuadros delimitadores en un formato de diccionario. Algunas imágenes tienen varios cuadros delimitadores.
    - Variables:
        - id - identificador de imagen único
        - boxes - cuadros delimitadores en formato de diccionario de fácil lectura
        - label - la etiqueta de predicción correcta para los cuadros delimitadores proporcionados

El objetivo final del proyecto es que para cada imagen se pueda predecir un cuadro delimitador y una clase para todos los hallazgos. Si se predice que no se encontraron hallazgos de la enfermedad se debería crear una categoría de la siguiente forma “ninguno 1 0 0 1 1” "ninguno" es el ID de clase para ningún hallazgo, y esto proporciona un cuadro delimitador de un píxel con una confianza de 1.0). Para cada estudio del conjunto de prueba debe clasificar la imagen en:

- Negativo para neumonía
- Aspecto típico
- Aspecto indeterminado
- Aspecto atípico

Se debe crear una cadena de predicción de forma similar a la categoría “ninguno”. Por ejemplo atypical 1 0 0 1 1.

El formato en que se encuentran las imágenes (DICOM) contiene datos adicionales que pueden ser útiles para visualizar y clasificar. ¡Explórelas!

Clasificación radiogenómica de tumores cerebrales RSNA-MICCAI:

El glioblastoma (GB) es el tumor maligno primario del sistema nervioso central (SNC) más común en adultos (supone más del 50%) y se asocia invariablemente a un mal pronóstico. Solo el 33% de los pacientes sobrevive al año y el 5% de los pacientes llegan a vivir más de 5 años tras el diagnóstico. La Organización Mundial de la Salud (OMS) clasifica los gliomas fundamentalmente por criterios histopatológicos en: astrocitomas, oligodendrogliomas, oligoastrocitomas y ependimomas. Además, establece una gradación relacionada con el pronóstico de la enfermedad que identifica a los de alto grado como astrocitomas de grado III (astrocitoma anaplásico) y grado IV (GB).

Los GB pueden comprometer cualquier estructura neuroanatómica, pero en adultos es más común en los hemisferios cerebrales, mientras que en los niños lo es en la fosa posterior. Su crecimiento infiltrativo es extremadamente rápido. Histológicamente está compuesto de células de gran variabilidad morfológica, algunas bizarras, pleomórficas y multinucleadas; con actividad mitótica elevada; proliferación microvascular; severa y característica hiperplasia endotelial; microtrombos intravasculares, y necrosis extensas de carácter isquémico o en forma de pseudoempalizadas. La denominación multiforme se debe a la gran heterogeneidad que lo caracteriza con variados patrones y rasgos citológicos. En general, muestran focos bien diferenciados que alternan con otros pobremente diferenciados. Tradicionalmente se han clasificado en dos subtipos morfológicamente idénticos: primarios (GB1) y secundarios (GB2) (Castañeda et al., 2015).

Un biomarcador forma parte de una subcategoría de signos médicos que pueden ser medidos y reproducidos con precisión, y tienen la capacidad potencial de predecir un desenlace. Los procesos biológicos están conformados por tejidos, células o fluidos. El uso potencial de un biomarcador consiste en la capacidad de identificar la predisposición a cierta enfermedad tomando en cuenta la variabilidad y validez. La sistematización de los procesos puede reducir los costos operativos. Actualmente se han descrito cuatro biomarcadores para los gliomas de alto grado: deleción 1p/19q, metilación del sitio promotor de O6-methylguanine-DNA methyltransferase (MGMT), mutación isocitrato deshidrogenasa 1/isocitrato deshidrogenasa 2 (IDH1/IDH2) y micro-ARN (Manrique-Guzmán, 2004).

En particular el gen MGMT codifica una proteína reparadora del ADN removiendo el grupo alquilante de la posición O6 de la guanina a causa del tratamiento de quimioterapia alquilante, como la temozolamida. El proceso para identificar el estado de metilación se lleva a cabo a través de PCR específica de metilación mediante la conversión bisulfito (conversión no metilada a uracilo). La MGMT es una enzima reparadora que remueve el agente alquilante de la posición O6 de la guanina, que ocasiona un mal apareamiento durante la replicación celular, induciendo su apoptosis, que conlleva una mayor sobrevida (Manrique-Guzmán, 2004).

Se ha observado que la metilación del sitio promotor de MGMT tiene capacidad predictiva en la respuesta al tratamiento y la sobrevida de los pacientes con gliomas de alto grado. La diferencia en la sobrevida entre los pacientes con alta expresión de MGMT y los de baja expresión es de 8 frente a 29 meses (p = 0.0002). La mutación de IDH1 es un marcador confiable para discernir entre glioblastomas primarios, glioblastomas secundarios y gliomas anaplásicos (Manrique-Guzmán, 2004).

Actualmente, el análisis genético del cáncer requiere cirugía para extraer una muestra de tejido. Luego puede tomar varias semanas determinar la caracterización genética del tumor. Dependiendo de los resultados y el tipo de terapia inicial elegida, una cirugía posterior puede ser necesaria. Si se pudiera desarrollar un método preciso para predecir la genética del cáncer a través de imágenes, (es decir, radiogenómica) potencialmente minimizaría el número de cirugías y refinaría el tipo de terapia requerida.

La Sociedad Radiológica de América del Norte (RSNA) en sociedad con la Medical Image Computing and Computer Assisted Intervention Society (la Sociedad MICCAI) están trabajando día a día para mejorar el diagnóstico y la planificación del tratamiento de los pacientes con glioblastoma. Es por esto por lo que han proporcionado el siguiente conjunto de datos. El objetivo es predecir el subtipo genético del glioblastoma utilizando resonancias magnéticas (MRI) para entrenar y probar los modelos capaces de detectar la presencia de metilación del promotor MGMT. Se debe predecir la probabilidad de presencia de MGMT.

Descripción del Conjunto de datos:

Tanto el conjunto de entrenamiento como el de prueba tienen una carpeta dedicada identificada por un número de cinco dígitos. Dentro de cada una de estas carpetas de "casos", hay cuatro subcarpetas, cada una de ellas correspondiente a cada una de las imágenes de resonancia magnética multiparamétricas estructurales (mpMRI), en formato DICOM. Las exploraciones exactas de mpMRI incluidas son:

- Recuperación de inversión atenuada de fluidos (FLAIR)
- Pre-contraste ponderado por T1 (T1w)
- Post-contraste ponderado por T1 (T1Gd)
- T2 ponderado (T2)

Archivos:

- train/: Contiene las imágenes del conjunto de entrenamiento, cada carpeta contiene 4 subcarpetas con los cuatro tipos de exploraciones detalladas anteriormente.
- train_labels.csv - archivo que contiene el objetivo para cada sujeto en los datos de entrenamiento (por ejemplo, la presencia de metilación del promotor MGMT)
    - MGMT_value: Presencia del marcador o no.

Nota: Este conjunto de datos pertenece a una competencia activa de kaggle, si desea participar y optar por premios, puede inscribirse como grupo y trabajar directamente en su plataforma.

Procesamiento del lenguaje natural con tweets de desastres

En la actualidad, las redes sociales forman parte de la vida diaria de la mayoría de las personas. Siendo una forma de comunicación masiva, que puede resultar muy eficaz para ciertas situaciones críticas. Durante un desastre, ya sea natural o provocado por el hombre, la comunicación es un componente de suma importancia. Dadas las amenazas a la vida humana y la propiedad, las personas necesitan información sobre lo que sucedió y lo que todavía está ocurriendo dentro de un área afectada por un desastre. Debido a que las redes sociales en general, y Twitter en particular, ofrecen una rápida recuperación de información de multitud de fuentes y a su vez son canales de comunicación que permiten llegar a muchas personas ,se han desarrollado estudios sobre su uso sistemático como parte de una respuesta a emergencias (Martínez Quezada, Ortiz Sierra, Martínez Cano, & Lamos Diaz, 2020).

En particular Twitter se ha convertido en un importante canal de comunicación en tiempos de emergencia. La ubicuidad de los teléfonos inteligentes permite a las personas anunciar una emergencia que están observando en tiempo real. Debido a esto, más agencias están interesadas en monitorear programáticamente Twitter (es decir, organizaciones de socorro en casos de desastre y agencias de noticias). Pero, no siempre está claro si las palabras de una persona realmente están anunciando un desastre.

El objetivo de este proyecto es crear modelos de aprendizaje automático que permita predecir qué Tweets son sobre desastres reales y cuáles no. Para esto se tiene un conjunto de datos de 10, 000 tweets clasificados a mano.

Descripción del Conjunto de datos:

Los conjuntos de datos tienen la siguiente información:

- El texto de tweet
- Una palabra clave de este tweet
- La ubicación desde donde se envió el tweet.

Archivos:

Para este proyecto se trabajará únicamente con los archivos de entrenamiento.

- train.csv: Conjunto de datos de entrenamiento.
    - Variables:
        - id: identificador único de cada tweet
        - text: el texto del tweet
        - ubicación: la ubicación desde donde se envió el tweet
        - keyword: una palabra clave del tweet
        - target: 1 si es un tweet sobre un desastre real, 0 en caso contrario.

INSTRUCCIONES

Archivos:

- Deben contar con un repositorio en el cual versionen los archivos y demostrar trabajo de TODOS los miembros del grupo a lo largo del tiempo que van a dedicar al proyecto.
- Deben contar con un archivo en línea que permita comprobar la participación de todos en la generación del informe.

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 del problema de Procesamiento del Lenguaje Natural, investigue las técnicas que se usan para detectar patrones en lenguaje escrito.
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:**
    - Estudia las variables cuantitativas mediante técnicas de estadística descriptiva

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

MATERIAL A ENTREGAR

- Archivo .pdf con el informe de análisis exploratorio.
- Script de R (.r o .rmd) o de Python que utilizó para responder las preguntas con el código utilizado.
- Link de Google Drive donde trabajó el grupo
- Link del repositorio usado para versionar el código.
- Presentación de Power Point a usar para presentar resultados.

FECHAS DE ENTREGA

- **AVANCES:** Análisis Exploratorio de cada una de las variables de los conjuntos de datos. miércoles **6 de septiembre de 2021**.
- **PRESENTACIÓN Y DOCUMENTO FINAL COMPLETO: 10 de septiembre de 2021**
- **NOTA:** Solo se calificará el Documento Final si está entregado el avance con todo lo que se pide.

BIBLIOGRAFIA

Castañeda, C. A., Casavilca, S., Orrego, E., García-Corrochano, P., Deza, P., Heinike, H., … Ojeda, L. (2015). Glioblastoma: Análisis molecular y sus implicancias clínicas. Revista Peruana de Medicina Experimental y Salud Publica, 32, 316–325. Recuperado de http://www.scielo.org.pe/scielo.php?script=sci_arttext&pid=S1726-46342015000200017&nrm=iso

González-Castro, A., Escudero-Acha, P., Peñasco, Y., Leizaola, O., Martínez de Pinillos Sánchez, V., & García de Lorenzo, A. (2020). Cuidados intensivos durante la epidemia de coronavirus 2019. Medicina Intensiva, 44(6), 351–362. https://doi.org/10.1016/J.MEDIN.2020.03.001

Manrique-Guzmán, S. (2004). Biomarcadores en gliomas de alto grado: revisión sistemática. GACETA MÉDICA DE MÉXICO, 152, 87–93. Recuperado de www.anmm.org.mx

Martínez Quezada, D. O., Ortiz Sierra, R., Martínez Cano, J. G., & Lamos Diaz, H. (2020). Identificación de actores en un desastre a través de Twitter: Caso de estudio SINABUNG 2018. Stakeholders Identification in a Disaster Through Twitter: Study Case SINABUNG 2018., 30(1), 48–64. Recuperado de http://10.0.71.183/rcin.3938