1era Sesión: Preguntas y Respuestas

Por favor escriba sus preguntas en el cuadro para preguntas. Si tiene preguntas adicionales, por favor comuníquese con cualquiera de los siguientes instructores:

Erika Podest (erika.podest@jpl.nasa.gov)

Jordan A. Caraballo-Vega (jordan.a.caraballo-vega@nasa.gov)

Pregunta 1: Según su experiencia, ¿qué da mejores resultados en la clasificación de imágenes de percepción remota, el uso de algoritmos de Aprendizaje automático, o el uso de redes neuronales profundas? ¿Qué tanto aumenta la dificultad de la aplicación de una sobre la otra técnica?
[Translation] In your experience, which has better results when classifying remote sensing images, the use of Machine Learning algorithms or the use of Deep Neural Networks? How much does the difficulty of applying one technique increase over the other?
Respuesta 1: Las redes neuronales profundas pertenecen al grupo de algoritmos de aprendizaje automático, solo que son en parte de los algoritmos que necesitan más datos para poder ser eficientes. Estos algoritmos han sido utilizados extensivamente y han demostrado que son superiores en comparación con otros algoritmos más simples como los Random Forest. El contraste de este se va a basar en cuántos datos de entrenamiento tienes disponible. Si no tienes suficiente datos, puede que la red neuronal no pueda aprender lo suficiente de los datos de entrenamiento, y un algoritmo como el Random Forest provea una mejor clasificación porque muchas veces no necesitan tantos datos de entrenamiento para ser eficientes. Ahora bien, si tienes datos de entrenamiento suficientes, la red neuronal podrá encontrar patrones que otros algoritmos no son capaces de encontrar.

Pregunta 2: ¿El uso de GPU permite utilizar esa capacidad "ociosa" de procesamiento de una computadora, ya que al trabajar con datos no estamos procesando imágenes?

[Translation] Does the use of GPU allow us to use that "idle" processing capacity of a computer, since when working with data we are not processing images?
Respuesta 2: Sí, originalmente el GPU fue diseñado para acelerar imágenes, pero también se puede utilizar como un hardware de tu sistema que pueda utilizar su capacidad de procesamiento para otros usos. Muchos de los servidores de alto rendimiento configuran los GPUs para procesamiento de datos sin necesidad de transmitir imágenes. El GPU es capaz de hacer cómputos de datos que no sean necesariamente provenientes de imágenes.

Pregunta 3: ¿Puede darnos un par de ejemplos más de un problema continuo y uno discreto?
[Translation] Can you give us a couple of more examples of a continuous and discrete problem?
Respuesta 3: Otros problemas discretos podrían ser: 1) clasificar el uso del suelo entre agua, árbol, o cultivo; 2) clasificar el tipo de combustible del suelo entre una lista de estos; 3) clasificar animales dentro de una imagen satelital. Ejemplos de problemas continuos: 1) pronóstico de pulgadas de lluvia, 2) pronóstico de la trayectoria de un huracán, 3) medida de profundidad de un lago.

Pregunta 4: Los datos de entrenamiento supervisados, por ejemplo de cultivos, ¿se pueden usar en otras regiones (Sudamérica), o es preferible generarlos para cada estudio? ¿Enseñarán cómo generar datos de entrenamiento?
[Translation] Can supervised training data, for example from crops, be used in other regions (South America), or is it preferable to generate them for each study area? Will you teach how to generate training data?
Respuesta 4: Siempre es bueno tener datos del mismo lugar de estudio. Sin embargo, existe un auge increíble en estos modelos para desarrollar modelos que no necesitan datos del lugar de estudio para poder hacer predicciones de alta calidad. La idea es utilizar modelos fundacionales que permitan predecir datos que el modelo nunca ha visto, pero que tiene una noción de lo que son. No entraremos en cómo desarrollar datos de entrenamiento, pero mencionamos un poco el concepto como parte del ejercicio 1.

Pregunta 5: ¿Qué significa que el algoritmo tenga un “accuracy” alto, eso quiere decir que el algoritmo dará una precisión alta en la predicción?

[Translation] What does it mean that the algorithm has a high accuracy? Does it mean that the algorithm will have high prediction accuracy?
Respuesta 5: "Accuracy" es solo una métrica para evaluar los modelos de aprendizaje automático. Es simplemente la división entre los datos acertados sobre el total de los datos. Hay otras medidas como precisión, coeficiente de determinación para casos continuos etc.

Pregunta 6: Los métodos clásicos de clasificación (máxima verosimilitud por ejemplo), ¿se pueden considerar como aprendizaje automático?
[Translation] Can classical classification methods (maximum likelihood for example) be considered machine learning?
Respuesta 6: El algoritmo de máxima verosimilitud es probabilístico y por sí solo no es considerado un algoritmo de aprendizaje automático. Muchos algoritmos de aprendizaje automático utilizan máxima verosimilitud para combinarla con sus otras ecuaciones lineales y no lineales para hacer predicciones, y muchas veces para agrupar esas decisiones.

Pregunta 7: ¿Discreto es clasificación y continuo es regresión?
[Translation] Discrete is classification and continuous is regression?
Respuesta 7: Sí, sin embargo discreto también puede ser detección de objetos.

Pregunta 8: ¿Qué librería es la más adecuada para aplicar un algoritmo de aprendizaje automático de validación geoestadística supervisada como modelo de precisión que incluya muestreo en campo o in situ, como cambios en la cobertura de tierra por deforestación?
[Translation] Which library is the most appropriate to apply a supervised geostatistical validation machine learning algorithm as a precision model that includes sampling in the field or in situ, such as changes in land cover due to deforestation?
Respuesta 8: Si estas son observaciones individuales (puntos en un mapa), vas a necesitar un algoritmo que pueda procesar los datos de forma de columna y fila. Si tienes menos de 1000 observaciones, te recomiendo un algoritmo como el Random Forest o el XGBoost. De tener más de 1000 observaciones, podrías tratar con una red neuronal, pero aun así tal vez tengas pocos datos para que la red neuronal generalice lo suficiente.

Pregunta 9: ¿Dónde están ubicados los ráster que se descargan en el ítem 4.1 Data Download?
[Translation] Where are the rasters that are downloaded in item 4.1 Data Download located?

Respuesta 9: Estos se encuentran en HuggingFace https://huggingface.co/datasets/nasa-cisto-data-science-group/modis-lake-powell-toy-dataset.

Pregunta 10: En Kmeans clustering, ¿cómo determinas el k? ¿Utilizas un número por defecto, o haces una revisión preliminar para saber un número aproximado de clases?
[Translation] In K Means clustering, how do you determine the K? Do you use a default number or do you do a preliminary review to know an approximate number of classes?
Respuesta 10: Todo dependerá del caso que estés tratando de resolver. Por ejemplo, si deseas identificar si los píxeles son agua o no agua, mi primer intento va a ser tener una K de 2 porque solo quiero identificar 2 clases. Si el clustering no fue lo suficiente para agrupar los datos entre agua y no agua, puedo continuar incrementando la K dependiendo de los grupos que quiera encontrar. Cuando queremos crear datos de entrenamiento de datos satelitales, podemos definir la K en 100, y comenzar a agrupar las clases en clases más pequeñas que nos puedan dar información suficiente para diferenciar las clases que nos interesan identificar.

Pregunta 11: ¿Cuál será el balance entre la precisión del modelo AI con el número de datos y el tiempo de entrenamiento?
[Translation] What is the balance between the precision of the AI model with the number of data and the training time?
Respuesta 11: A medida que incrementas la cantidad de datos, al modelo le tomará más tiempo entrenar. A medida que el modelo incrementa, le tomará más tiempo entrenar y más tiempo hacer predicciones. Por ende, esto dependerá de la aplicación que tengas a la mano. Si te interesa ser extremadamente rápido, tal vez un modelo más pequeño como Random Forest sea de utilidad. Si te interesa un poco más de precisión, tal vez un modelo más complejo podría ser la mejor opción.

Pregunta 12: Consulta: en el Random Forest, ¿en qué etapa introducimos los datos recolectados en campo (ej: firmas espectrales de tipo de cobertura del suelo colectado con un espectrorradiómetro)?
[Translation] Consultation: in the Random Forest, at what stage do we introduce the data collected in the field (e.g.: spectral signatures of land cover type collected with a spectroradiometer)?

Respuesta 12: Los datos recolectados en campo se introducen al principio. Van a necesitar las firmas para hacer el entrenamiento y para la predicción. Desde el principio estamos utilizando esa información para entender cómo el modelo va a pensar y para hacer las predicciones que necesitamos.

Pregunta 13: ¿Cómo se combinan los datos espaciales y tabulares para entrenar modelos? ¿Debo pasar los datos espaciales a tabular o viceversa?
[Translation] How do you combine spatial and tabular data to train models? Should I convert the spatial data to tabular or vice versa?
Respuesta 13: Muchas veces lo que hacemos es que si tenemos un modelo que solo acepta datos tabulares, entonces convertimos los datos espaciales a tabulares. Si el modelo acepta datos en forma espacial, entonces introducimos los datos espaciales directamente. Al final, cuando el modelo predice en forma tabular, entonces se puede hacer un reshape para poner los datos en forma espacial como producto final.

Pregunta 14: ¿Hay bibliotecas de algoritmos organizadas por aplicación? ¿Cuáles son las herramientas de búsquedas más recomendadas para dar con el algoritmo indicado?
[Translation] Are there algorithms organized by application? What are the most recommended search tools to find the right algorithm?
Respuesta 14: De los algoritmos que identificamos en esta sesión, sklearn y cuML (la versión de GPU) es la librería que te va a ayudar con aprendizaje automático sencillo. Si deseas aplicar aprendizaje profundo, entonces librerías como TensorFlow o PyTorch son las adecuadas. Esas son las dos distinciones principales. Las mejores búsquedas se pueden llevar a cabo con la librería de https://arxiv.org/ que genera resultados de publicaciones con los diferentes algoritmos y te dará una idea de los algoritmos más utilizados para distintas aplicaciones. Si deseas buscar los diferentes softwares entonces lo mejor es explorar las páginas web de las diferentes librerías, las cuales contienen tutoriales que te ayudarán a entender si el algoritmo es aplicable a tu caso.

Pregunta 15: ¿Se podría realizar trabajos o investigaciones con respecto a proyecciones climáticas con variables como temperatura y precipitación utilizando estas tecnologías?
[Translation] Could work or research be carried out regarding climate projections with variables such as temperature and precipitation using these technologies?

Respuesta 15: Sí. Ahora mismo estoy trabajando en modelación de incendios utilizando variables de temperatura y de precipitación, así que muchas de estas técnicas se utilizan para eso. Incluso, hay modelos que son "data driven" (impulsados por datos) que estamos utilizando para hacer predicciones del clima en los próximos 100 años.

Pregunta 16: La técnica para seleccionar el algoritmo parte al manejar la técnica de seleccionar un píxel vecino, ¿así es?
Respuesta 16: Es muy importante saber si tu problema necesita tener información acerca de sus vecinos. En tal caso, para poder, por ejemplo, hacer una clasificación o regresión, entonces necesitas considerar cuál algoritmo toma en cuenta los píxeles vecinos. Si piensas que tu problema no necesita información sobre sus vecinos entonces puedes seleccionar cualquier algoritmo que sea de observaciones simples.

Pregunta 17: Como una forma académica de enseñar Aprendizaje Profundo usando Python desde el nivel básico y nivel avanzado, ¿puede ser afectado por la tendencia de ChatGPT?
Respuesta 17: En mi opinión, en estos momentos estamos combinando lo que enseñamos con la inteligencia artificial -Chat GPT. No deben de utilizarse individualmente sino combinados. Yo creo que ChatGPT puede ser una fuente inmensa de información pero que se debe de combinar con el conocimiento básico del ser humano. ChatGPT puede ayudar a acelerar nuestro proceso de enseñanza de aprendizaje automático, pero debemos de evaluar los resultados de ChatGPT ya que muchas veces los modelos se equivocan. Así que tenemos que utilizarlo de una forma que nos ayude pero también validarlo.

Pregunta 18: ¿Cuáles son las limitantes más importantes del aprendizaje automático?
Respuesta 18: La más importante son los datos de entrenamiento. Si no tenemos datos de entrenamiento de calidad o suficientes, entonces es muy difícil conseguir buenos resultados. Muchas veces el 40-50% del tiempo es dedicado a obtener buenos datos de entrenamiento. A medida que tenemos más datos de entrenamiento, reducimos las limitaciones del aprendizaje automático.

Pregunta 19: ¿Cuáles son las principales ventajas y desventajas de trabajar algoritmos de machine learning para imágenes Landsat y Sentinel en marcos de aprendizaje en la nube, respecto a programas instalados?

Respuesta 19: Muchas de las ventajas son que muchas de las aplicaciones que utilizamos en la nube tienen muchas de las librerías instaladas. Como ventajas, ya tenemos una fuente de información para comenzar con estas aplicaciones rápidamente. Con respecto a las desventajas, muchos de los datos que tenemos contienen datos de entrenamiento, pero no tienen las características que necesitamos. Las ventajas de utilizar aprendizaje automático para estos problemas son que la mayoría de los algoritmos tradicionales hacen medidas utilizando umbrales o métodos clásicos. Machine learning ha demostrado que puede inferir características que no podemos hacer manualmente.

Pregunta 20: ¿Cuán viable es mezclar diversas fuentes de información para alimentar los modelos de Aprendizaje automático, como el uso de fotografías tomadas por drones sumado a fuentes de imágenes satelitales?
[Translation] How viable is it to mix different sources of information to feed the Machine Learning models, such as the use of photographs taken by drones added to satellite image sources?
Respuesta 20: Hay líneas de investigación en aprendizaje automático que analizan esta aplicación. La respuesta corta es que los modelos de aprendizaje automático son sensibles a la distribución de los datos y, para realizar predicciones precisas, el modelo necesita ver características similares a las que se entrenó. Los avances recientes en los transformadores ("transformers") han mostrado un rendimiento prometedor en datos de diferentes fuentes (consulte el modelo "Segment Anything" de Microsoft).

Pregunta 21: Un algoritmo como XGboost puede mejorar la precisión de mi algoritmo. He estado trabajando en algoritmos de detección de manglar y la clasificación se ha realizado por RF, pero deseo mejorar mi precisión. ¿Cómo puedes definir la cantidad de data suficiente y así elegir tu algoritmo?, ¿Alguna relación área-número de datos o número de datos y resolución de imagen?
[Translation] An algorithm like XGboost can improve the accuracy of my algorithm. I have been working on a mangrove detection algorithm and the classification has been done by RF, but I want to improve my accuracy. How can you define the amount of sufficient data and thus choose your algorithm? Is there a relation between area/quantity of data or quantity of data and image resolution?

Respuesta 21: Todo lo mencionado tiene que ver. En cuanto a la resolución, a medida que tengas una resolución más alta (e.g. 2 metros) tendrás muchos más datos ya que cada pixel que estás observando va a ser mucho más diferente comparado con otros pixeles. A medida que disminuya la resolución, puede que no necesites tanta información porque la diferencia entre píxeles no será tan grande. Con respecto a cambiar de Random Forest a GBoost, podrías intentarlo. No es difícil, solo tendrías que cambiar dos líneas de código y puedes utilizar la misma cantidad de datos. En el análisis que hemos hecho, ninguno de los dos necesita más datos que el otro. Los dos son buenos trabajando con cantidades similares de datos.

Pregunta 22: ¿Se pueden usar Transformers para hacer regresión, por ejemplo para estimar la temperatura de la superficie terrestre usando datos de MODIS?
[Translation] Can Transformers be used to do regression, for example to estimate land surface temperature using MODIS data?
Respuesta 22: Sí. En este entrenamiento no estaremos hablando sobre "transformers". Sin embargo, la idea de los transformers es que utiliza el concepto de "self-supervised" (autosupervisado) lo que significa que el "transformer" aprende por sí mismo para muchas de estas aplicaciones. Uno añade en una segunda parte qué es lo que está haciendo la clasificación o regresión. Así que por ejemplo, con MODIS podrías utilizar un visión "transformer" (VIT) y podrías ponerle en la parte del frente una red convolucional que podría ser la parte de la regresión con sus datos.

Pregunta 23: ¿Los datos procesados son almacenados en la carpeta de Google Colab?
[Translation] Is the processed data stored in the Google Colab folder?
Respuesta 23: Los datos que van a estar creando por medio de este proceso sí se están almacenando allí.

Pregunta 24: ¿Es necesario realizar algoritmos de índices espectrales para desarrollar un modelo de predicción de machine learning?
[Translation] Is it necessary to apply spectral index algorithms to develop a machine learning prediction model?
Respuesta 24: No necesariamente. Los índices espectrales ayudan a los algoritmos de tipo decisión porque agregan información adicional para informar mejor la decisión (e.g. árboles de decisión, random forest, XGBoost). Es posible que los algoritmos como las redes

neuronales no necesiten índices espectrales, ya que encuentran relaciones dentro de sus datos. Aquí hay un ejemplo de una red convolucional que usa RGB+NIR para clasificar nubes a 2m, https://doi.org/10.1016/j.rse.2022.113332.

Pregunta 25: ¿Conocen casos en los cuales a partir de información inicial clasificada como "incendio" o "no incendio", se pueda hacer predicciones en forma continua, por ejemplo, poder predecir la probabilidad de incendio en un rango del 0-1?
[Translation] Are you aware of cases in which, based on initial information classified as "fire" or "non-fire", predictions can be made continuously, for example, being able to predict the probability of fire in a range of 0-1.
Respuesta 25: Algoritmos como Random Forest y XGBoost proporcionan la función predict_proba que le permite generar una probabilidad de cualquier clasificador. Si construye un clasificador binario con fuego o sin fuego, le daría la probabilidad de esa observación en particular en el rango de 0-1. Hay otras formas más sofisticadas de hacer esto mediante el uso de redes neuronales y activaciones que le permiten generar probabilidades.

Pregunta 26: Buenas tardes. Si quisiera clasificar una zona de estudio de acuerdo a las pendientes del terreno, para clasificar zonas susceptibles a alguna remoción en masa, ¿en que herramienta debería interiorizarme?
[Translation] Good afternoon, if I wanted to classify a study area according to the slope of the terrain, to classify areas susceptible to mass removal, what tool should I utilize?
Respuesta 26: La solución no estará muy centrada en el algoritmo, sino más bien en qué datos tiene disponibles para el problema. Según su formato y los datos de entrenamiento disponibles, podría elegir cualquier algoritmo según los criterios que discutimos durante el entrenamiento.

Pregunta 27: ¿Qué nivel de Cálculo Diferencial se debe tener para tratar de mejorar nuestros modelos, dada la importancia de la misma? Lo digo porque no se ha tratado las matemáticas.

[Translation] What level of Differential Calculus should be done to improve our models, given its importance. I say this because mathematics has not been discussed.
Respuesta 27: La mayoría de las operaciones matemáticas que se tratan en el aprendizaje automático están relacionadas con el álgebra lineal y el cálculo diferencial básico para algoritmos más complejos. Debido al tiempo, esta capacitación está enfocada en la aplicación de estos algoritmos a las Ciencias de la Tierra, y no en las matemáticas detrás de ellos. Proporcionamos referencias a lo largo de estas sesiones para que pueda informarse más sobre el tema. Las matemáticas detrás de estos algoritmos serán de gran utilidad cuando se intenta descomponer el comportamiento del modelo y cuando se intenta comprender detalles específicos de cada modelo. Solo tocamos estadísticas básicas de ciencia de datos durante esta capacitación, ya que es lo que necesitará para trabajar con los ejemplos proporcionados.

Pregunta 28: Si hubiese que entrenar modelos en "tiempo real" que requieran alta capacidad de procesamiento/memoria, ¿qué soluciones de software/hardware recomendarían? ¿Es posible realizar algo así en la nube o sería mejor contar con un servidor dedicado?
[Translation] If I were to train models in "real time" that require high processing capacity/memory. What software/hardware solutions would you recommend? Is it possible to do something like this in the cloud or would it be better to have a dedicated server?
Respuesta 28: La recomendación general es tener GPUs disponibles para poder entrenar y predecir los datos de manera eficiente. En el cloud existen servicios como SageMaker que podrían ayudar. También podría utilizar funciones lambda que le permitan someter operaciones de entrenamiento e inferencia en la nube de AWS. Tener servidores dedicados siempre será un beneficio porque no influenciará en la disponibilidad de servicios de la nube.

Pregunta 29: ¿Es mejor agrupar datos de entrenamiento en celdas?
[Translation] Is it better to group training data in cells?
Respuesta 29: No necesariamente, todo dependerá del problema. En casos donde se utilice datos de teledetección en forma de ráster, estos datos se pueden almacenar fácilmente en forma de celda. En el caso donde los datos sean de plataformas aéreas (e.g. LiDAR) o datos que solo tengan observaciones en forma de puntos, es mejor almacenar los datos de entrenamiento en forma de columna y fila.

Pregunta 30: Mediante modelos de Machine Learning, ¿puedo mejorar la resolución de imágenes, por ejemplo de Landsat combinando con imágenes de mayor resolución como de Planet?

[Translation] Using Machine Learning models, can I improve the resolution of images, for example from Landsat, combining them with higher resolution images such as Planet?

Respuesta 30: Sí. Existen algoritmos de aprendizaje automático en el área de Super Resolución que combinan datos de baja resolución con datos de alta resolución para mejorar la resolución de una fuente de datos específica. Adjunto algunas referencias que le podrían ayudar: Kyzivat, E. D., & Smith, L. C. (2022). Contemporary and historical detection of small lakes using cross-sensor super resolution Landsat imagery, Kyzivat, E. D., & Smith, L. C. (2022, December). Landsat at 3 m resolution?!? Applying a super resolution model to surface water detection. In *AGU Fall Meeting Abstracts* (Vol. 2022, pp. IN42A-07), Teo, T. A., & Fu, Y. J. (2021). Spatiotemporal fusion of formosat-2 and landsat-8 satellite images: A comparison of “super resolution-then-blend” and “blend-then-super resolution” approaches. *Remote Sensing*, *13*(4), 606.