2da 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: ¿Los datos de estos DAACs difieren de los disponibles en catálogos como GEE o Planetary Computer?
[Translation] Do the data from these DAACs differ from those available in catalogs such as GEE or Planetary Computer?
Respuesta 1: Los DAAC tienen los productos de misión “oficiales”. GEE y Planetary han descargado copias de estos datos, pero también tienen productos derivados de usuarios que pueden diferir de los productos "oficiales". Para asegurarse de tener los archivos más actualizados y apropiados, siempre es mejor utilizar los datos de origen seleccionados de los DAAC.

Pregunta 2: ¿Qué recomendarías para hacer un modelo de machine learning utilizando datos satelitales para un río urbano estrecho de entre 10-20 metros de ancho?
[Translation] What would you recommend in order to create a machine learning model using satellite data for a narrow urban river between 10-20 meters wide?
Respuesta 2: Para ríos que tienen menos de 20 m de ancho, necesitará usar datos de un satélite de resolución más alta como Sentinel o posiblemente datos comerciales de alta resolución. Los datos aéreos también podrían ser útiles en este caso.

Pregunta 3: ¿Tienen ejercicios de interpretación de cuerpos de agua con fenómenos de mezcla de vegetación en agua como contenido de macrófitas, algas, cianobacterias y material flotante como residuos orgánicos e inorgánicos?

[Translation] Do you have exercises for interpreting bodies of water with mixed vegetation, such as macrophyte content, cyanobacteria algae, and floating material such as organic and inorganic residues?
Respuesta 3: Los ejercicios de esta capacitación se centraron en un caso de uso "típico" de aguas abiertas, por lo que no tenemos un ejercicio aquí para identificar el agua con vegetación emergente o proliferación de algas.

Pregunta 4: Creo que Pandas y GeoPandas son muy buenos para tratar los datos y es el más utilizado, pero me gustaría saber qué piensas de la librería Polars (escrita en Rust) debido a la gran velocidad de cómo trata los datos.
[Translation] I think that Pandas and GeoPandas are very good at handling data and are the most widely used, but I would like to know, what do you think of the Polars library (written in Rust)? due to the great speed of how it treats the data.
Respuesta 4: Las estadísticas de rendimiento de Polars han sido sorprendentes hasta el momento, sin embargo, es muy temprano para saber si la comunidad continuará brindando apoyo para continuar su desarrollo. Por el momento, utilizamos cudf para acelerar la manipulación de dataframes usando GPUs y estamos monitoreando poco a poco la madurez de Polars.

Pregunta 5: ¿Existen metodologías para determinar el número mínimo de datos que se deben garantizar para ejecutar procesos de clasificación de imágenes de satélite?
[Translation] Are there methodologies to determine the minimum number of data that must be guaranteed to run a satellite image classification?
Respuesta 5: Es difícil determinar un número "mínimo" de entradas porque realmente depende del tamaño y la complejidad del área de estudio, su tolerancia al error y el número total de clases en los archivos de salida. Puede hacer un modelo con muy pocos puntos de entrenamiento, pero es poco probable que el modelo funcione bien. Generalmente trato de mantener la cantidad de datos de entrenamiento pequeña al principio, entreno el modelo y lo aplico a los datos para hacer un mapa. Evalúo el mapa y luego itero agregando más datos de entrenamiento o ajustando el modelo hasta que los resultados sean satisfactorios.

Pregunta 6: ¿Qué tipo de tratamiento se puede realizar con valores atípicos?
[Translation] What type of treatment can be performed with outliers?

Respuesta 6: Dependiendo del algoritmo de aprendizaje automático, los valores atípicos se pueden añadir a sus datos de entrenamiento para que el modelo tenga datos de entrenamiento representativos de las observaciones. En otros casos, se pueden aplicar técnicas de preprocesamiento que disminuyan estos valores atípicos, de esta manera disminuimos la varianza de estas observaciones al momento de hacer la predicción.

Pregunta 7: En el análisis de correlación entre bandas, ¿cuál sería el valor límite del coeficiente de correlación a considerar para descartar bandas correlacionadas y evitar la colinealidad?
[Translation] In correlation analysis between bands, what would be the limit value of the correlation coefficient to consider in order to rule out correlated bands and avoid collinearity?
Respuesta 7: No existen pautas específicas para el umbral del coeficiente de correlación para eliminar bandas. Realmente está impulsado por la tolerancia del investigador principal del proyecto al error en el modelo. Existen modelos capaces de encontrar relaciones no lineales que tal vez no necesiten la eliminación de bandas que tengan alta colinealidad.

Pregunta 8: En los casos en que los datos que vamos a utilizar presenten datos muy correlacionados, ¿cómo podemos escoger el modelo que mejor nos brinde resultados?
[Translation] In cases where the data that we will be using is highly correlated, how can we choose the model that provides the best results?
Respuesta 8: Muchos de los métodos muestran la importancia de las características que se pueden usar para eliminar características que no contribuyeron en gran medida al modelo. Esto a veces puede eliminar las variables correlacionadas. De cualquier manera, la selección del método depende más de la pregunta científica y los datos que de por cuánto están correlacionados los datos de entrada.

Pregunta 9: En las cuestiones agrícolas-forestal, ¿qué tipo de algoritmo y limpieza de datos se puede hacer?
[Translation] In agricultural-forestry issues, what kind of algorithm and data cleaning can be done?
Respuesta 9: El tema de agricultura y bosques podría ser abarcado con muchos o con la mayoría de los métodos de ML. Cuál es el mejor depende casi siempre de la pregunta

científica en particular. Del mismo modo, los métodos de limpieza de datos dependen explícitamente de los datos particulares que se estén utilizando. Una buena práctica es buscar literatura en referencia a su problema y ver qué algoritmos se han utilizado para responder algunas de sus preguntas.

Pregunta 10: A manera de ejemplo cómo puedo diferenciar o dividir los datos de prueba, de validación y de entrenamiento?
[Translation] As an example, how can I differentiate or divide the test, validation and training data?
Respuesta 10: Hay varias opiniones sobre la división de datos de entrenamiento y validación. Yo normalmente los divido en 80/20, pero muchas personas los dividen en 50/50.

Pregunta 11: En un algoritmo de Random Forest en donde tenía 5700 datos, usé 4000 de training y 1700 de validación. Ya que RF no usa la validación para el entrenamiento, ¿es necesario usar datos de prueba?
[Translation] In a Random Forest algorithm where I had 5,700 data points, I used 4,000 for training and 1,700 for validation. Since RF does not use validation for training, is it necessary to use test data?
Respuesta 11: Muchas personas adoptan el enfoque de usar solo la división de entrenamiento y validación y luego simplemente validan el mapa de salida. Yo apoyo este enfoque.

Pregunta 12: Si yo espero que una de las clases tenga menos representación que otras, ¿igualmente tienen que estar balanceadas las muestras? Por ejemplo, si hago una clasificación de cultivos y sé que soja y maíz ocupan el 80% de la superficie, ¿debería igualmente tener la misma cantidad de datos de otros cultivos (como girasol o sorgo)?
[Translation] If I expect one of the classes to have less representation than the others, do the samples still have to be balanced? For example, if I generate a crop classification and I know that soybean and corn occupy 80% of the surface, should I also have the same amount of data for other crops (such as sunflower or sorghum)?
Respuesta 12: Si sus clases están desequilibradas, corre el riesgo de que las clases con menos cobertura no se entrenen bien y, por lo tanto, no se representen adecuadamente en el modelo y la salida del mapa posterior. Dependiendo del algoritmo que elijas, el balance de las clases será de gran importancia para evitar sesgos en el entrenamiento e inferencia del modelo.

Existen técnicas en algoritmos como las redes neuronales donde puede usar funciones de penalización para disminuir el sesgo en el proceso de aprendizaje.

Pregunta 13: Cuando tienes datos nulos, ¿eliminas ese punto o realizas un tipo de asignación, por ejemplo, valor medio de la población?
[Translation] When you have null data, do you eliminate that point or do you do a type of assignment, for example, mean value of the population?
Respuesta 13: Esto también depende de las preferencias del investigador principal del proyecto. Personalmente, no estoy a favor de interpolar valores nulos, por lo que lo haría a través de esos puntos (se introducen datos que ya no tienen la misma calidad). Otras personas tienen más tolerancia al error, por lo que están más dispuestas a permitir la interpolación para remover algunos de estos valores nulos.

Pregunta 14: ¿Los outliers afectan las correlaciones o distribuciones?
[Translation] Do outliers affect correlations or distributions?
Respuesta 14: A veces, los valores atípicos pueden representar una clase real en los datos, por lo que puede ser importante mantenerlos para hacer que el modelo sea más preciso. Recuerde que sus datos de entrenamiento son solo una "representación" del conjunto de datos completo, por lo que solo debe asegurarse de tener todos los casos importantes representados en el conjunto de entrenamiento, incluso si algunos parecen valores atípicos. Por lo tanto, puede usar los valores atípicos para regresar e investigar sus datos para determinar qué está causando el valor atípico y si es una clase que le interesa investigar más a fondo.

Pregunta 15: ¿Las características correlacionadas pueden afectar el modelo de clasificación de ML? Otra duda, ¿cómo se filtran o identifican las características más relevantes al crear un modelo de ML?
[Translation] Can correlated features affect the ML classification model? Another question, how do you filter or identify the most relevant characteristics when creating an ML model?
Respuesta 15: Hay modelos que son sensibles a características altamente correlacionadas (ej., regresión logística, árboles de decisión) y podrían afectar el rendimiento del modelo. Otros modelos, como las redes neuronales, son más resistentes a características altamente correlacionadas, ya que encuentran patrones no lineales dentro de los datos. Las características de entrada correlacionadas pueden engañar al modelo para que piense que es

más preciso de lo que realmente es, especialmente los modelos basados en árboles. Puede usar técnicas de importancia de permutación para visualizar la importancia de algunas características en sus datos que influyen en el rendimiento de la predicción y así poder filtrar algunas de las entradas de su modelo.

Pregunta 16: ¿Los problemas derivados de la diferencia en la cantidad de registros entre clases puede solucionarse incluyendo en el modelo el parámetro class_weight?
[Translation] Can problems derived from the difference in the number of records between classes be solved by including the class_weight parameter in the model?
Respuesta 16: Esta es una técnica que podría ayudar, sin embargo, los parámetros de class_weight introducidos en los parámetros del modelo implicarán una gran cantidad de permutaciones de prueba y error para encontrar el modelo de mayor rendimiento. Incluso con el conjunto de parámetros class_weight, todavía hay posibilidades de que su modelo tenga un rendimiento inferior dada la poca representación de algunas de estas clases.

Pregunta 17: ¿Es evidente que cuando se tratan con predictores o datos continuos pierden parte de su información al categorizar el tipo de cobertura en el momento de la división de los nodos?
Respuesta 17: Esto dependerá de la generalización de su problema. En algunas preguntas científicas, no nos importa mucho la precisión flotante de los valores, pero sí nos importa categorizar algunos de estos en un rango general. Por lo tanto, si solo le interesan ciertas categorías dentro de su problema científico, categorizarlas en rangos podría no afectar su flujo de trabajo. Si desea rangos de valores específicos, entonces sí, perdería información al categorizar predictores continuos.

Pregunta 18: ¿Podrías volver a explicar el proceso de reshape en la parte de predicción?
[Translation] Could you explain the process of reshaping in the part about predictions again?
Respuesta 18: El input de entrada al crear el ráster es una matriz (array) que es de m x n. Nuestro algoritmo requiere que tengamos una estructura de datos en forma de columna y fila así que lo que hacemos es que utilizamos cada banda espectral y las colocamos de manera que sean las columnas. Los valores de los píxeles en el array se convierten en filas. De esta manera el algoritmo puede tomar esa columna y fila y hacer las predicciones y después lo ponemos en forma de array para representar el ráster.

Pregunta 19: ¿Las variables ndvi y ndwi* son variables derivadas a partir del valor de las bandas de MODIS?
Respuesta 19: Sí. Si ven la función de los ejercicios que lee el array, pueden ver que allí se calcula utilizando las bandas espectrales 1, 5. 7 etc. dependiendo de si es NDVI o NDWI.

Pregunta 20: Cuando se habla de predicción, ¿se refiere a predecir en años futuros el comportamiento de un objeto o cuerpo de agua o tierra?
[Translation] When talking about prediction, do you mean predicting the behavior of an object or body of water or land in future years?
Respuesta 20: No, aunque predicciones futuras podrían ser incluidas. Cuando hablamos de predicciones - el algoritmo toma los datos de entrada para proveer probabilidades u observaciones de lo que está viendo. Los datos pueden ser del pasado. Si tuviésemos datos del futuro que el algoritmo no ha visto todavía entonces se llama forecasting (pronóstico). Si son datos que se utilizan como entrada al algoritmo entonces estamos haciendo la clasificación de esa imagen.

Pregunta 21: ¿Se puede hacer esto mismo en Google Earth Engine ?
[Translation] Can this same thing be done in Google Earth Engine?
Respuesta 21: Sí. GEE también contiene la habilidad de utilizar algoritmos de aprendizaje automático, así que hay dos opciones. Primero pueden ubicar los datos en GEE, descargarlos y hacer la clasificación local en Google Colab. La segunda opción es utilizar la consola de GEE y hacer la clasificación directamente desde allí.

Pregunta 22: ¿Cómo se podría cambiar el modelo para clasificar varias clases?
[Translation] How could you change the model in order to classify various classes?
Respuesta 22: Tiene que cambiar las variables de entrada. Por ejemplo, cambiar la columna de agua (que es de tipo binaria) a una que tenga sus propias clases. También puede cambiar la función de decisión del algoritmo. Hoy utilizamos Gini pero puede utilizar Entropy que son los básicos. Puede hacer lo mismo con el XGBoost.

Pregunta 23: ¿Cómo puede interpretarse el Out Of Bag score en términos de validación del modelo?

[Translation] How can we interpret the Out of Bag score in terms of model validation?
Respuesta 23: El out of bag score es la precisión que genera el modelo de los datos de entrenamiento. Nos indica la certeza del modelo en relación a los datos de entrenamiento utilizados, lo cual se puede incrementar añadiendo datos de prueba. De tener suficientes datos de validación, el Out of Bag score tal vez no sea necesario. Ahora bien, si sus datos de entrenamiento tal vez no son lo suficientemente grandes, puede que el utilizar el Out of Bag score sea al menos una guía para saber cuán preciso es su modelo.

Pregunta 24: Se puede hacer las predicciones cuando hemos entrenado, pero ¿como se puede poner en producción dentro de un aplicativo web? ¿Puede dar alguna bibliografía acerca de este tema?
[Translation] You can make predictions when we've trained, but how can this be put into production within a web application? Can you provide a bibliography on this subject?
Respuesta 24: La idea es que una vez tengas el modelo entrenado, guardes ese objeto. Entonces puedes utilizar aplicaciones como FLASK de Python, TensoFlow Lite, etc. que le permiten a través de un API o interfaz proveer la data a tu aplicación web o móvil y hacer las predicciones y retornar cualquier visualización o archivo que quieras proveer. Existen otras aplicaciones que le permiten hacer flujos similares.

Pregunta 25: ¿Cómo puedo hacer para clasificar variables categóricas de sistemas con más de dos variables, por ejemplo: vegetación, agua y suelo utilizando Random Forest?
[Translation] What do I need to do to classify categorical variables with more than two variables, for example: vegetation, water, and soil using Random Forest?
Respuesta 25: Cambie el número de clases definidas en el código. Vea la respuesta número 22.

Pregunta 26: ¿Es posible realizar con estos cuadernos la clasificación de cultivos utilizando Machine Learning con bandas de Sentinel 2?
[Translation] Is it possible to run a crop classification with these notebooks using Machine Learning with Sentinel-2 bands?
Respuesta 26: Sí. Lo único que debe de hacer es reemplazar los datos de entrada con los de Sentinel-2 de su proyecto.

Pregunta 27: En la sesión anterior preguntaba sobre Transformers, respecto al acondicionamiento de los datos, no se si puede explicar sobre "input embeddings" en los Transformers. Gracias.
[Translation] Good afternoon. Last session, I asked about Transformers, with respect to data conditioning. I don't know if you could explain about "input embeddings" in the Transformers. Thank you.
Respuesta 27: Esto no se cubrió en este entrenamiento porque nos hemos enfocado en algoritmos fundamentales que le provean la base para entender algoritmos más complejos. Los transformers utilizan el input de entrada para crear "tokens" donde se almacena la data y se extraen patrones y features (características) de esos tokens del texto o las imágenes de entrada. El input embedding es una combinación de esos tokens para entender o extraer features de los datos de entrada para poder hacer las predicciones de esos mismos features. En el tiempo, es una base de datos de donde se aprende cómo está la data y se puede utilizar con cualquier algoritmo.

Pregunta 28: ¿Lo ideal es que la información satelital tenga las bandas del espectro rojo e infrarrojo cercano?
[Translation] Is it ideal for satellite information to have bands in the red or near-infrared spectrum?
Respuesta 28: Depende del problema científico. El algoritmo de aprendizaje automático simplemente utiliza las bandas de entrada especificadas y utiliza las bandas que contienen más información para el problema definido. Las bandas del espectro rojo e infrarrojo muchas veces proveen información adicional como por ejemplo, cuando trabajamos con datos de vegetación.

Pregunta 29: Me gustaría que comentaran ejemplos de uso de ML en determinación de la estructura tridimensional de la vegetación. ¿Podrían sugerir librerías disponibles con data SAR y Lidar, y algoritmos existentes, que son indicados en la bibliografía para evaluar el volumen de vegetación y ajustarlo para la determinación de la biomasa sobre el suelo?
[Translation] I would like you to comment on examples of the use of ML in determining the three-dimensional structure of the vegetation. Could you suggest available libraries with SAR and Lidar data, and existing algorithms, which are indicated in the bibliography to

evaluate the volume of vegetation and adjust it for the determination of aboveground biomass?

Respuesta 29: Hay varios algoritmos utilizados para SAR, incluyendo transformers y redes convolucionales neuronales. Adjunto algunas referencias: Li, Y., Ma, L., Zhong, Z., Liu, F., Chapman, M. A., Cao, D., y Li, J. (2020). Deep learning for lidar point clouds in autonomous driving: A review. *IEEE Transactions on Neural Networks and Learning Systems*, *32*(8), 3412-3432, Guan, H., Yu, Y., Ji, Z., Li, J., y Zhang, Q. (2015). Deep learning-based tree classification using mobile LiDAR data. *Remote Sensing Letters*, *6*(11), 864-873.

Pregunta 31: ¿Se puede hacer un Data hub automatizado con un modelo de machine learning elaborado?

[Translation] Can you do an automated Data hub with an elaborated machine learning model?

Respuesta 31: Sí, en definitiva. Depende del tiempo necesario para hacer las predicciones (si lo necesitas hacer inmediatamente, etc.) y del problema que tengas. Pero los algoritmos de aprendizaje automático lo podrás utilizar para cualquiera de las tareas que los entrenes a hacer.