1era Sesión: Preguntas y Respuestas

Por favor escriban 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), Sean McCartney (sean.mcartney@nasa.gov o Armando Marino (armando.marino@stir.ac.uk)

Pregunta 1: ¿Es muy pesado Anaconda?
[Eng.] Is Anaconda very large to install?
Response 1: Thanks for the question, not in my experience. It is around 800 MB and generally it works without crashing.
Respuesta 1: Gracias por la pregunta, no en mi experiencia. Es de alrededor de 800 MB y generalmente funciona bien.

Pregunta 2: ¿Podemos usar además de Anaconda otro motor? Por ejemplo, SageMaker Studio Lab, Máquina planetaria de office, etc?
[Eng.] Can we use another engine besides Anaconda? For example, SageMaker Studio Lab, Planetary Office Machine, etc.?
Response 2: These are all very interesting platforms for doing ML in the Cloud and you are very welcome to use them. Anaconda is not a cloud platform, it is an installer that helps get Python installed on your machine. The processing you are doing today happens on your machine and not on the cloud. If you are interested in cloud computing then please try one of those, including Google Colab, where you can run Jupyter Notebook files.
Respuesta 2: Todas esas plataformas son muy interesantes para hacer ML en la nube y le invitamos a usarlas. Anaconda no es una plataforma en la nube, es un instalador que ayuda a instalar Python en su computadora. El procesamiento que se hizo hoy fue en su computadora y no en la nube. Si está interesado en computaciones en la nube, pruebe uno de los softwares que menciono, incluyendo Google Colab, donde puede ejecutar archivos de Jupyter Notebook.

Pregunta 3: ¿Cómo enlazo Python con SNAP para el preprocesamiento de datos Sentinel-1?
[Eng.] How do I link Python with SNAP for Sentinel-1 data preprocessing?

Response 3: There are different ways to do this. One solution is to use SNAP functions in Python with SNAPpy for instance. However, sometimes this can be fiddly. My suggestion would be to prepare the graph in SNAP and then run it in Python using SNAP GPT Graph Processing Tool (GPT). Here is a good introduction to this https://www.youtube.com/watch?v=CD5Fvoy4PWU. Make sure your graph works properly in SNAP before using the GPT, because it is much harder (I find it very frustrating) to debug graphs from Python.
Respuesta 3: Hay diferentes maneras de hacer esto. Una solución es usar funciones de SNAP en Python con SNAPpy, por ejemplo, pero a veces esto puede ser complicado. Mi sugerencia sería preparar el gráfico en SNAP y luego ejecutarlo en Python usando la herramienta de procesamiento de gráficos SNAP GPT (GPT). Aquí hay una buena introducción a esto https://www.youtube.com/watch?v=CD5Fvoy4PWU
Asegúrese de que su gráfico funcione correctamente en SNAP antes de usar GPT, porque es mucho más difícil (lo encuentro muy frustrante) depurar gráficos desde Python.

Pregunta 4: No entendí qué hace el procesamiento, ¿lo podrás repetir?
[Eng.] I didn't understand what the processing does, can you repeat it?
Response 4: I guess you mean the preparation of data using the SNAP graph. This may take a bit of time to show in this answer, and my suggestion is to look at a SNAP tutorial, where you can see where you need to click in SNAP. In this training we use simple pre-processing. A tutorial on this can be found in the following link: https://www.youtube.com/watch?v=30x6XE6U4CU (until about minute 30) which is the ARSET training: *SAR Polarimetry with Sentinel-1, RCM, & SAOCOM Imagery for Agriculture, Part 2/4*.
Respuesta 4: Supongo que se refiere a la preparación de datos usando el gráfico SNAP. Esto puede tomar un poco de tiempo para explicar en esta respuesta, y mi sugerencia sería ver un tutorial de SNAP, donde se explica que debe hacer en SNAP. En esta capacitación utilizamos un preprocesamiento simple. Se puede encontrar un tutorial sobre esto en el siguiente enlace: https://www.youtube.com/watch?v=30x6XE6U4CU (hasta aproximadamente el minuto 30) que es de la capacitación de ARSET: Polarimetría SAR con Sentinel-1, RCM y SAOCOM Imágenes para la agricultura, Parte 2/4.

Pregunta 5: El procedimiento que están explicando es muy tedioso y confuso. ¿Sería mejor utilizar Google Earth Engine (GEE) con la ventaja que todo se realiza en la nube?

[Eng.] The procedure you are explaining is very tedious and confusing. Would it be better to use Google Earth Engine (GEE) with the advantage that everything is done in the cloud?

Response 5: I guess the point here is, why one wants to use Python instead of GEE, and Python functions seem more difficult to use than Java functions. The extra effort of using Python is about being in control of the data and procedures (in this practical you know what is going on and you set all parameters… you are more in control of it). For instance in GEE you cannot use SLC data and you can only apply the functions that GEE provides with the data formatting that is supported. Also, you will see the hardest part is to produce the features extracted from the temporal analysis. GEE is very simple for doing things like telling GEE to take a collection and running a classification, but it is harder to say, put these elements in the feature vector when it comes to PolSAR features and temporal data. So, to answer your question, the extra complexity pays back in accuracy, since you can use more methods, more data structure for feature vectors, more variants of ML and parameter settings. This tutorial is focused on teaching how to do things from first principles (the hard way maybe :)), but indeed you can use GEE tools if what you need to do is already included there.

Respuesta 5: Supongo que el punto aquí es por qué uno quisiera usar Python en lugar de GEE ya que las funciones de Python parecen más difíciles de usar que las funciones de Java. El esfuerzo adicional de usar Python es para tener control de los datos y los procedimientos (en esta práctica, usted sabe lo que está pasando y establece todos los parámetros... tiene más control sobre ello). Por ejemplo, en GEE no se puede usar datos SLC y solo puede aplicar las funciones que proporciona GEE con el formato de datos compatible. Además, verá que lo más difícil es generar las características extraídas del análisis temporal. GEE es muy simple para hacer cosas como decirle a GEE que tome una colección de datos y haga una clasificación, pero es más difícil decirle que ponga estos elementos en el vector de características cuando se trata de PolSAR y datos temporales. Entonces, para responder a su pregunta, la complejidad adicional se recupera en precisión, ya que puede usar más métodos, más estructura de datos para vectores de características, más variantes de ML y configuración de parámetros. Este tutorial se enfoca en enseñarles conceptos básicos (quizás de la manera difícil :)), pero de hecho, puede usar las herramientas GEE si lo que necesita hacer ya está incluido allí.

Pregunta 6: ¿ARSET cuenta con entrenamientos con Jupyter Notebook en diferentes niveles?, o ¿una ruta de aprendizaje usando Jupyter Notebook desde básico hasta avanzado?

[Eng.] Does ARSET have Jupyter Notebook training at different levels, or a learning path using Jupyter Notebook from basic to advanced?
Response 6: I am only familiar with the ARSET training I gave: *SAR Polarimetry with Sentinel-1, RCM, & SAOCOM Imagery for Agriculture, Part 2/4*.
ARSET will have a machine learning training within the next month.
Respuesta 6: Solo estoy familiarizado con la capacitación de ARSET que presente: Polarimetría SAR con imágenes Sentinel-1, RCM y SAOCOM para agricultura, Parte 2/4 Próximamente ARSET estará impartiendo una capacitación solamente sobre Machine Learning.

Pregunta 7: ¿Se recomienda tener igual o similar cantidad de píxeles para cada ROI de entrenamiento?
[Eng.] Is it recommended to have the same or a similar number of pixels for each training ROI?
Response 7: Yes, I would use the same number of pixels. These are called Balanced datasets. Unbalanced datasets may have issues unless other methods are used to moderate this. So indeed, try to use a similar number of pixels if you can.
Respuesta 7: Sí, usaría la misma cantidad de píxeles. Estos se denominan conjuntos de datos equilibrados. Los conjuntos de datos desequilibrados pueden tener problemas a menos que se utilicen otros métodos para moderar esto. Entonces, de hecho, intente usar una cantidad similar de píxeles si puede.

Pregunta 8: Tengo curiosidad por saber si el enfoque sintetizado supera al enfoque de pila completa debido a la alta dimensionalidad de la pila completa y, en caso afirmativo, si convertir la fecha de la imagen en una variable (es decir, utilizar una matriz con sólo seis columnas en lugar de 95, pero con millones de filas) mejoraría la precisión. También tengo curiosidad por saber hasta qué punto la rotación de cultivos supone un problema para el enfoque sintetizado, y si hay alguna adaptación que mitigue la asimetría de la media y la desviación estándar causada por los cambios en las series temporales?
[Eng] I'm curious whether the synthesized approach outperforms the full stack approach due to the full stack's high dimensionality, and if so, whether making a variable out of the image date (i.e., using a matrix with only six columns instead of 95, but with millions of rows) would improve accuracy. I'm also curious to know how much of a problem crop rotation poses for the synthesized approach, and if there are any adaptations that mitigate skewness of the mean and standard deviation caused by changes in the time series?

Response 8: Excellent points. Indeed, adding many features can produce issues in ML. I didn't have time to show how the performance would have improved if we applied a dimension reduction method before running the RF (e.g., a PCA). Dimension reduction normally helps, but my experience in the case of time series is that the same crops can be seeded at different times and therefore the time trends for those parcels do not match anymore (one of the trends is delayed). Crop rotation is indeed an issue if you have more crops inside the same time frame you use for the model. If you apply this to a location where they have more rotations, then the time series should be cut to times when this happens. A much more complex problem indeed. :) I am not sure what you mean with skewness of mean and std, maybe ask more please.
Respuesta 8: Excelentes puntos. De hecho, utilizar muchas características puede producir problemas en ML. No tuve tiempo de mostrar cómo habría mejorado el rendimiento si hubiéramos aplicado un método de reducción de dimensiones antes de ejecutar el RF (por ejemplo, un PCA). La reducción de dimensiones normalmente ayuda, pero en mi experiencia en el caso de las series temporales es que los mismos cultivos pueden sembrarse en diferentes momentos y, por lo tanto, las tendencias temporales de esas parcelas ya no coinciden (una de las tendencias se retrasa). La rotación de cultivos es de hecho un problema si tiene más cultivos dentro del mismo período de tiempo que usa para el modelo. Si aplica esto a una ubicación donde tienen más rotaciones, entonces la serie de tiempo debe reducirse al periodo en que esto sucede. Un problema mucho más complejo de hecho. :) No estoy seguro de lo que quiere decir con sesgo de media y estándar, tal vez pregunte más, por favor.

Pregunta 9: Hola, ¿podrían explicar por favor porque al realizar la validación cruzada la precisión disminuye?
[Eng.] Hello, could you please explain why when performing cross validation the accuracy decreases?
Response 9: Generally this happens, and it is because you consider different areas for testing and training, the algorithm is tested over areas that the algorithm has not seen before. So you need to split testing and training to reduce overfitting, which otherwise may happen. Another very good point about splitting in more parts is that you reduce the randomness of selecting a test area that is particularly favorable or unfavorable for calculating the accuracy.
Respuesta 9: Suele ocurrir, y se debe a que se consideran áreas diferentes para las pruebas y el entrenamiento, el algoritmo se prueba en áreas que el algoritmo no ha visto antes. Así que hay que dividir los datos en datos de prueba y datos de entrenamiento para reducir el sobreajuste que, de lo contrario, podría producirse. Otro

punto muy bueno de dividir en major cantidad de partes es que se reduce la aleatoriedad de seleccionar un área de prueba que sea particularmente favorable o desfavorable para calcular la precisión.

Pregunta 10: ¿Qué se debe incluir en los scripts para la clasificación de cultivos en zonas de altas pendientes?
[Eng.] What should be included in the scripts for the classification of crops in areas of high slopes?
Response 10: This is an excellent question, thanks for asking. SAR data are impacted by terrain and one needs to correct this before applying a classifier. Here we do terrain correction by extracting the gamma naught. That is generally not enough, because it corrects the data radiometrically, but not polarimetrically (the phase is unchanged by the correction). If you had quad-pol, then you should correct for orientation angle as well. If not, another trick that often helps is to include the local incidence angle as a feature in your model. The machine will learn that a specific target behaves a certain way in certain slope conditions and use this info to correct for terrain-induced changes. You get the local incidence angle from SNAP. It's not an easy task to deal with mountainous terrain, but lots of progress has been made recently and we are getting there. :)
Respuesta 10: Esta es una pregunta excelente, gracias por preguntar. Los datos SAR se ven afectados por el terreno y es necesario corregirlo antes de aplicar una clasificación. En este caso, la corrección del terreno se realiza extrayendo gamma naught. Esto no suele ser suficiente, porque corrige los datos radiométricamente pero no polarimétricamente (la fase no cambia con la corrección). Si tuvieras quad-pol, entonces deberías corregir el ángulo de orientación también. Si no, otro truco que a menudo ayuda es incluir el ángulo de incidencia local como una característica en el modelo. El algoritmos aprenderá que un objetivo específico se comporta de cierta manera en ciertas condiciones de pendiente y utilizará esa información para corregir los errores inducidos por cambios por el terreno. El ángulo de incidencia local se obtiene de SNAP. No es fácil trabajar en áreas con terreno montañoso, pero recientemente ha habido mucho progreso en este tipo de análisis.

Pregunta 11: Se podría usar K-means como un mecanismo para definir clases para el entrenamiento del ML model?
[Eng.] Could K-means be used as a mechanism to define classes for ML model training?

Response 11: It can be used as an initial step that will be refined later, iteratively, with another ML method (like RF). However, it is best if you know what targets you have on the scene and you can select them manually. But yes, it helps to have some insight on the data.
Respuesta 11: Se puede usar como un paso inicial que se refinará más adelante, de manera iterativa, con otro método de ML (como RF). Sin embargo, es mejor si sabe qué objetivos (o clases) están en la imagen y seleccionarlos manualmente. Pero sí, ayuda tener una idea de los datos.

Pregunta 12: ¿El código o algoritmo automatizado puede incluir el índice Normalizado de Área Quemada (NBR)?
[Eng.] Can the automated code or algorithm include the Normalized Burned-area Rate (NBR)?
Response: Yes, that is a good point, you can include this or other vegetation indexes if this makes sense for your specific problem. PolSAR opens up the chance for using many observables and indeed you can select the features that best suit your application.
Respuesta 12: Sí, ese es un buen punto, puede incluir este u otros índices de vegetación si tiene sentido para su aplicación específica. PolSAR abre la posibilidad de usar muchas características y, de hecho, puede seleccionar las características que mejor se adapten a su aplicación.

Pregunta 13: Para reproducir los ejercicios con otros datos, ¿requiere utilizar otro paquete adicional a Python y SNAP?
[Eng.] To reproduce the exercises with other data, do I need to use another package in addition to Python and SNAP?
Response 13: No, you should be able to use only the libraries I showed here (the ones that came with Anaconda). If you want to improve the code and make it more automatic, you may need to install other libraries like GDAL, but this is not strictly needed. It depends on your application.
Respuesta 13: No. Debería poder usar solo las bibliotecas que mostré en la sesión de hoy (las que vinieron con Anaconda). Si desea mejorar el código y hacerlo más automático, es posible que deba instalar otras bibliotecas como GDAL, pero esto no es estrictamente necesario. Depende de su aplicación.

Pregunta 14: ¿Se han hecho estudios con otros algoritmos de Machine Learning además de Random Forest? Ejemplo: Support Vector Machine (SVM)

[Eng.] Have studies been done with other Machine Learning algorithms besides Random Forest? Example: Support Vector Machine (SVM)
Response 14: Many studies. SVM is a big name but also Neural Networks, Naive Bayes, Nearest Neighbour, Gaussian Processes, and more… have a look at the sklearn library and you can try many more https://scikit-learn.org/stable/modules/classes.html.
Respuesta 14: Muchos estudios. SVM es uno de los mas populares pero también está Neural Networks, Naive Bayes, Nearest Neighbour, Gaussian Processes y más... Explore la biblioteca de sklearn para aprender sobre los diferentes algoritmos que contiene y poder aplicarlos: https://scikit-learn.org/stable/modules/classes.html.

Pregunta 15: ¿Pueden darnos un hub con repositorios de ejemplos?
[Eng.] Can you give us a hub with example repositories?
Response 15: You can find several GitHub links online, there are a lot.
Respuesta 15: Puede encontrar varios enlaces de GitHub en línea, hay muchos.

Pregunta 16: Entre Deep Learning y RF, ¿qué algoritmo es de mayor uso para la clasificación y monitoreo de cultivos?
[Eng.] Between Deep Learning and RF, which algorithm is most used for crop classification and monitoring?
Response 16: To my knowledge, RF is possibly the most used algorithm for crop monitoring with SAR, but there is also a lot done using DL, especially recently. As a rule of thumb, DL will help extract information about the texture of the crop when this is expected to be useful to discriminate between different crops. It also can help identify time trends. One disadvantage of DL is that it generally needs more training than RF, and it is therefore harder to get it trained. But I am sure in the future we will see more and more advances in DL.
Respuesta 16: Que yo sepa, RF es posiblemente el algoritmo más utilizado para el monitoreo de cultivos con SAR, pero también se usa mucho Deep Learnning (aprendizaje profundo o DL), especialmente recientemente. En general, DL ayudará a extraer información sobre la textura del cultivo, cuando se espera que sea útil para discriminar entre diferentes cultivos. También puede ayudar a identificar tendencias temporales. Una desventaja de DL es que generalmente necesita más entrenamiento que RF y, por lo tanto, es más difícil entrenarlo. Pero estoy seguro de que en el futuro veremos más y más avances en DL.

Pregunta 17: ¿Se podría dar salida a los resultados de los scripts mediante un webmap interactivo?

[Eng.] Could the results of the scripts be output via an interactive webmap?
Response 17: Absolutely yes. You can build a processing chain where the output is then pulled by somebody else with an API and displayed on a webpage. You will need to deal with image sizes so don't make it too heavy and slow… but absolutely possible.
Respuesta 17: Absolutamente que sí. Puede crear una cadena de procesamiento en la que otra persona extraiga el resultado con una API y lo visualize en una página web. Tendrá que lidiar con el tamaño de las imágenes, así que no lo haga demasiado pesado y lento... pero absolutamente que sí es posible.

Pregunta 18: Hablando de la visualización de los datos predecidos por el cubo 500x500. ¿Cuál es el máximo de reshape de la imagen obtenida que se puede hacer en la predicción de los datos, siendo un limitante el uso de python para esto?
[Eng.] Speaking of the visualization of the predicted data by the 500x500 cube. What is the maximum reshape that can be done with the predicted data, being that the use of Python is limited for this?
Response 18: I am not sure I understand, can you please repeat? If you are asking about the size of cubes, then this depends on your RAM. You should try to not have arrays too big that will exceed your available RAM or the processing will become tremendously slow. You can check the RAM usage while you process, e.g. in Spyder or just opening Task Manager in Windows.
Respuesta 18: No estoy seguro de haber entendido esta pregunta, ¿puedes repetirlo? Si estás preguntando por el tamaño de los cubos, entonces esto depende de tu memoria RAM. Debes de tratar de no tener matrices demasiado grandes que excedan tu RAM disponible o el procesamiento se volverá extremadamente lento. Puede verificar el uso de RAM durante el procesamiento, por ejemplo en Spyder o simplemente abriendo el Administrador de Tareas en Windows.

Pregunta 19: No logré observar los parámetros de ajuste (tuning) de los algoritmos utilizados (p. ej. en random forest, mtry, ntree, etc.) ¿Son posibles estos ajustes de manera manual y automática? Por otra parte, ¿cuál es su concepto con el uso de ensambles de modelos para estas clasificaciones?
[Eng.] I did not manage to observe the tuning parameters of the algorithms used (e.g., in random forest, mtry, ntree, etc.) Is it possible to do these adjustments manually and automatically? On the other hand, what is your opinion about the use of model ensembles for these classifications?

Response 19: Good point, here we used default values that work nicely most of the time. If you want to bring this to another level then you may want to tune the parameters. You can have a look at https://scikit-learn.org/stable/modules/generated/sklearn.ensemble.RandomForestClassifier.html for more info on the parameters. Also look at methodologies that can set the parameters optimally based on your data (often referred as Hyperparameters tuning).
Respuesta 19: Buen punto. Aquí usamos valores por defecto que funcionan bien la mayor parte del tiempo. Si desea llevar esto a otro nivel, entonces es posible que desee ajustar los parámetros. Puede consultar https://scikit-learn.org/stable/modules/generated/sklearn.ensemble.RandomForestClassifier.html para obtener más información sobre los parámetros. También busque metodologías que puedan ajustar los parámetros de forma óptima en función de sus datos (a menudo denominado ajuste de hiperparámetros).

Pregunta 20: Still, regarding the areas with high slopes, is it useful to apply terrain flattening in SNAP? Or is this a tool for something else?
Response 20: Yes, you should apply it if you have topography in your scene. It makes the processing a bit slower, but more robust. My suggestion is to use gamma naught if you can afford the extra time processing. For more info see reply to question 10.
Respuesta 20: Sí, debería aplicarlo si hay topografía en su escena. Hace que el procesamiento sea un poco más lento, pero más robusto. Mi sugerencia es usar gamma naught pero el procesamiento tomará más tiempo. Para más información por favor consulte la respuesta número 10.

Pregunta 21: ¿Los scripts que nos están dando en python son suficiente para procesar una imagen que descargue desde cero?
***[Eng.]*Are the scripts you are giving us in Python enough to process an image that you download from scratch?**
Response 21: No. The scripts start with preprocessed images that have been done in SNAP. Dr. Marino included a link on how to do this preprocessing in his presentation slides.
Respuesta 21: No. Los scripts comienzan con imágenes preprocesadas que se han hecho en SNAP. El Dr. Marino incluyó un enlace sobre cómo hacer este preprocesamiento en su presentación.

Pregunta 22: Mediante el uso de archivos de forma que representan diferentes cultivos y diferentes tamaños, con un área tan grande como un país entero, ¿será fácil adaptar su código Python?
***[Eng.]* By using shapefiles that have different crops and different sizes for an area as big as an entire country, will it be easy to adapt your Python code?**
Response 22: I guess this question refers to using shapefiles and the processing time. Indeed I agree that the optimal way to train the data is to use some shapefile of known locations and import those in Python. You can do it using GDAL and RasterIO for instance. A bit too complex to show in this practical, but a good way to proceed indeed. In terms of processing time, working at country scale with multitemporal data is generally very onerous and you will probably need to do this on the Cloud, unless you have access to some APC facilities.
Respuesta 22: Supongo que esta pregunta se refiere al uso de shapefiles y al tiempo de procesamiento. De hecho, estoy de acuerdo en que la mejor manera de entrenar los datos es utilizar algún shapefile de ubicaciones conocidas e importarlas en Python. Puedes hacerlo usando GDAL y RasterIO por ejemplo. Hubiera sido un poco demasiado complejo haberlo demostrado en esta práctica, pero es una buena manera de proceder de hecho. En términos de tiempo de procesamiento, trabajar a escala de país con datos multitemporales es generalmente muy oneroso y probablemente necesitarás hacerlo en la Nube, a menos que tengas acceso al servicio de APC.