Instituto de Física de Cantabria

CSIC
Consejo Superior de Investigaciones Científicas

UC
Universidad de Cantabria

Inteligencia Artificial y sus posibles aplicaciones en el ámbito de la ciencia y la gestión de la biodiversidad

Fernando Aguilar Gómez

1) Algoritmos de aprendizaje automático - Supervisados

- El aprendizaje supervisado es cuando el conjunto de datos contiene **anotaciones** con **clases de salida**.
- Busca **patrones** dentro de las etiquetas de valor asignadas a datos específicos.
- Algunos algoritmos populares de aprendizaje automático para el aprendizaje supervisado incluyen SVM (support-vector machines) para problemas de clasificación, regresión lineal para problemas de regresión y **random forest** para problemas de regresión y clasificación.
- P.ej. En el caso del análisis de sentimiento, las clases de salida son feliz, triste, enojado, etc.

2) Algoritmos de aprendizaje automático - No Supervisados

- **No hay etiquetas** asociadas a los datos.
- Estos algoritmos de aprendizaje automático **organizan los datos** en un grupo para describir su estructura y hacer que los datos complejos parezcan simples y organizados para el análisis.
- El mejor ejemplo de tal clasificación es el **agrupamiento** o clustering. La agrupación en clústeres agrupa objetos/datos similares, formando así clústeres segregados.
- La agrupación también ayuda a **encontrar sesgos** en el conjunto de datos.
- El aprendizaje no supervisado es relativamente más difícil y, en ocasiones, los grupos obtenidos son difíciles de entender debido a la falta de etiquetas o clases.

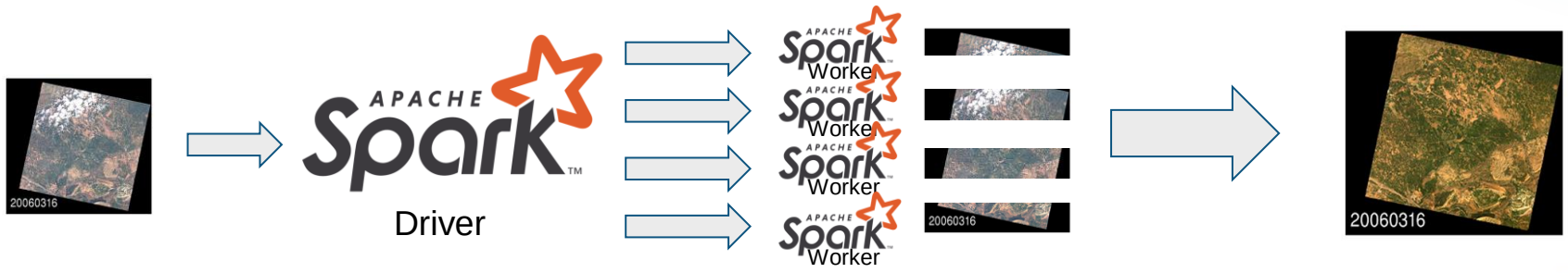
Soluciones de análisis - Big Data

Apache Spark: plataforma de computación distribuida muy popular en Big Data

- Provee una sencilla serie de APIs para procesar paralelamente grandes volúmenes de datos mientras el usuario no necesita tener gran conocimiento sobre cómo se gestionan los recursos

En computación distribuida, las tareas de los usuarios son distribuidas a través de un clúster de máquinas. Tanto los recursos como las tareas deben ser gestionados para tener un control del clúster

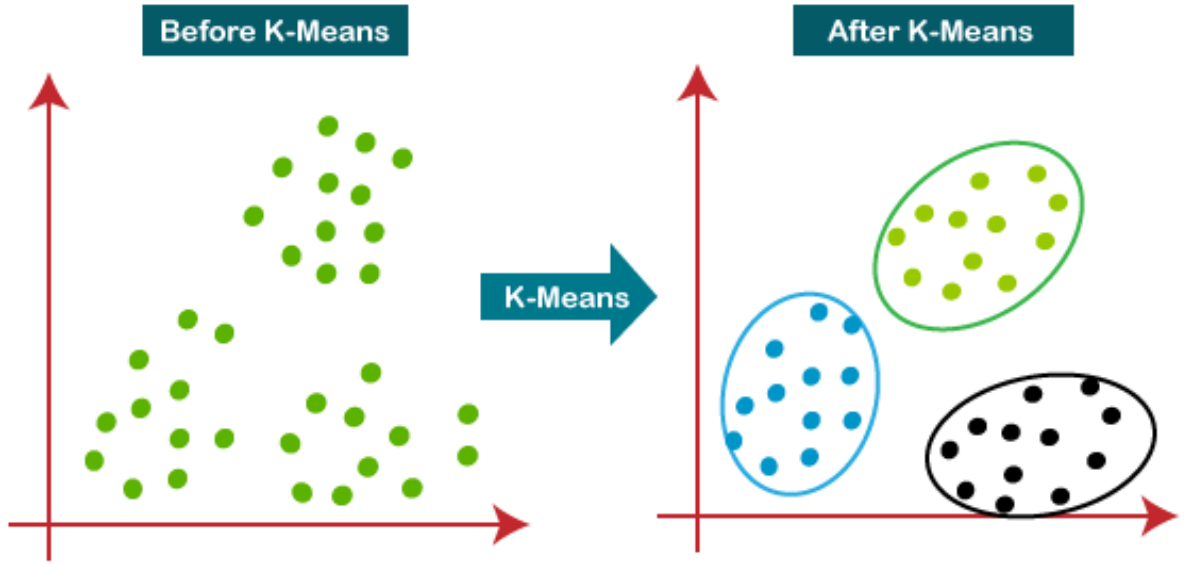
- Gestión de los recursos: Adquirir los servidores disponibles para ejecutar tu tarea
- Gestión de tareas: Sincronizar el código y los datos a través de los servidores del clúster



K Means Clustering Algorithm

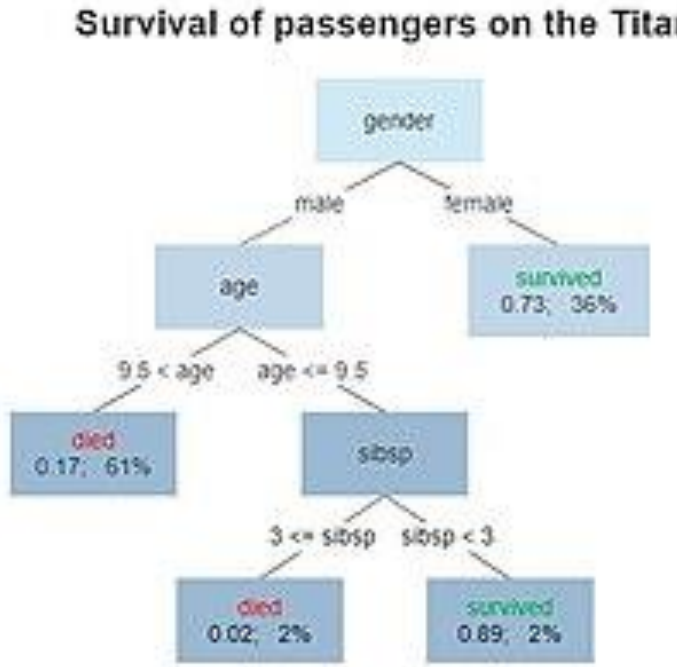
K-means is a popularly used unsupervised machine learning algorithm for cluster analysis. K-Means is a non-deterministic and iterative method.

The algorithm operates on a given data set through a pre-defined number of clusters, k (iterative, we could find the best K if needed).

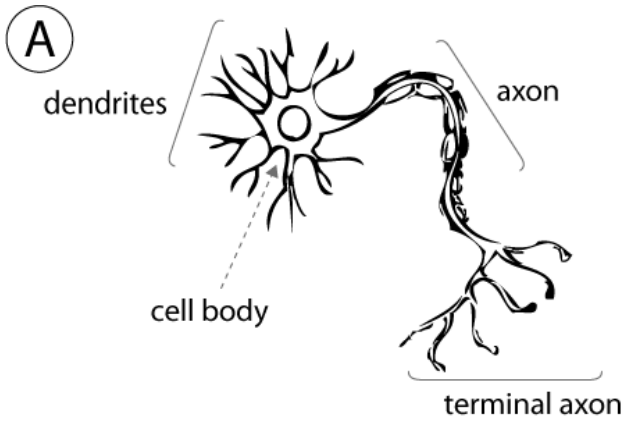


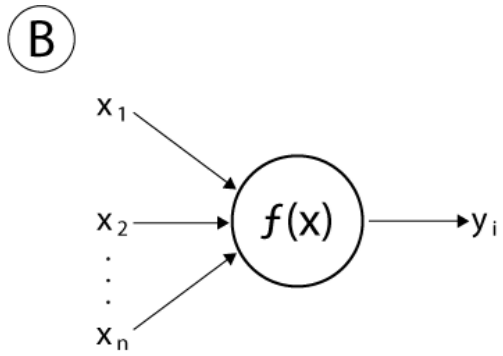
Decision Tree Machine Learning Algorithm

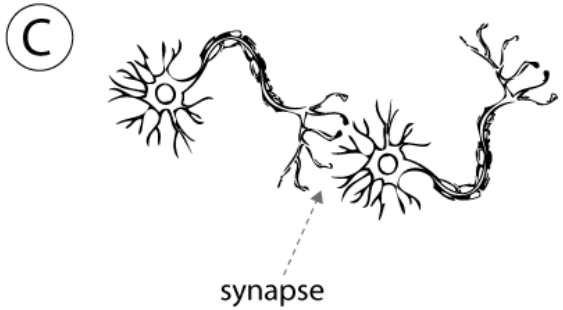
A decision tree is a graphical representation that makes use of branching methodology to exemplify all possible outcomes of a decision, based on certain conditions.
The internal node represents a test on the attribute, each branch of the tree represents the outcome of the test and the leaf node represents a particular class label i.e. the decision made after computing all of the attributes.
The classification rules are represented through the path from root to the leaf node.

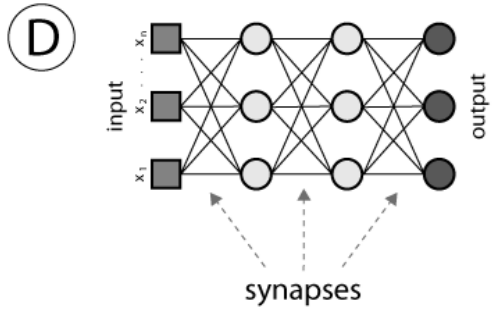


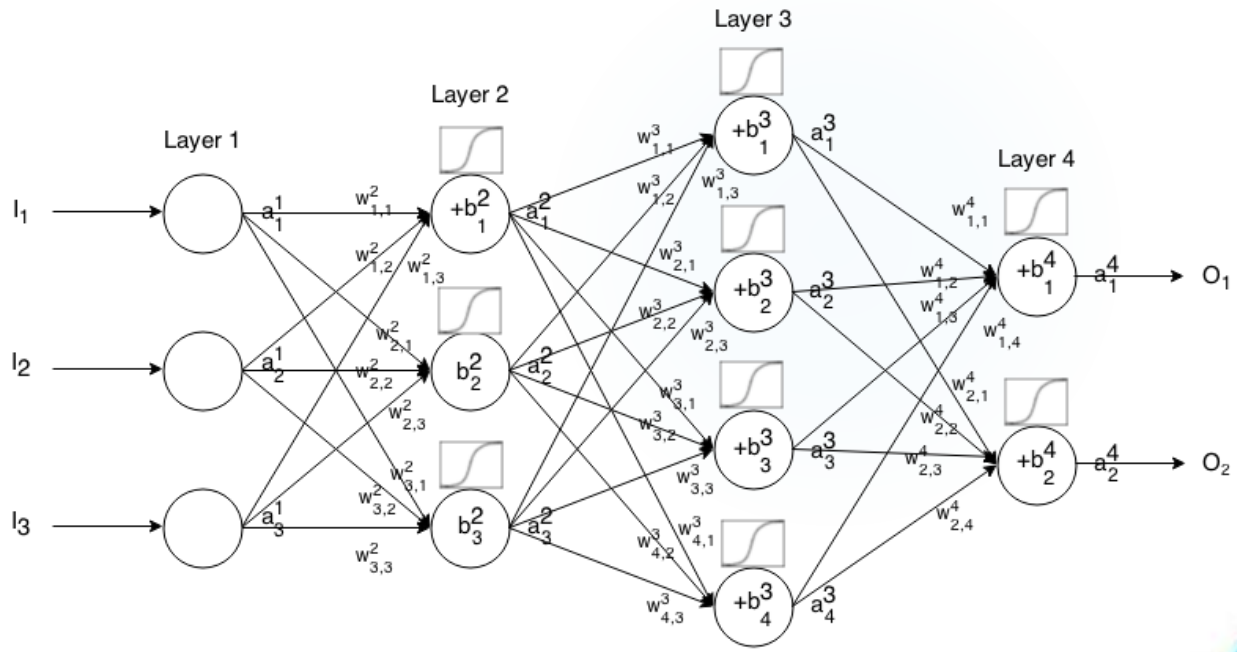
Artificial Neural Networks Machine Learning Algor







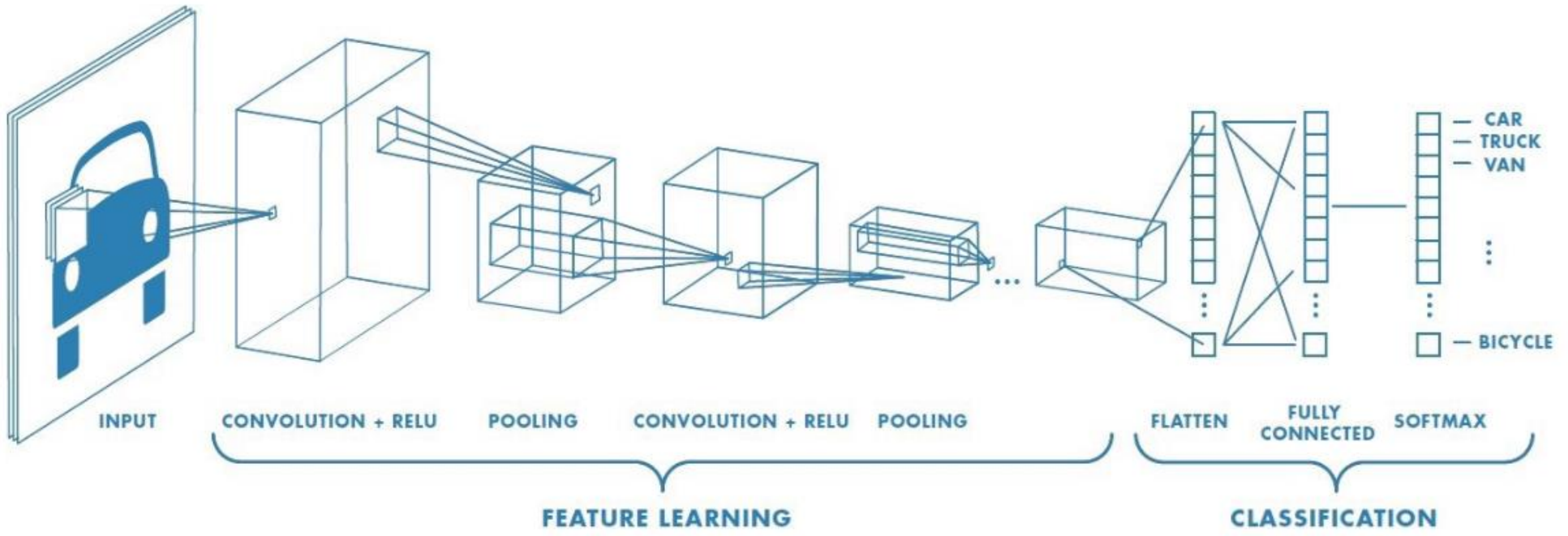




ANN PySpark: https://towardsdatascience.com/artificial-neural-network-using-pyspark-324cf47e8d0a

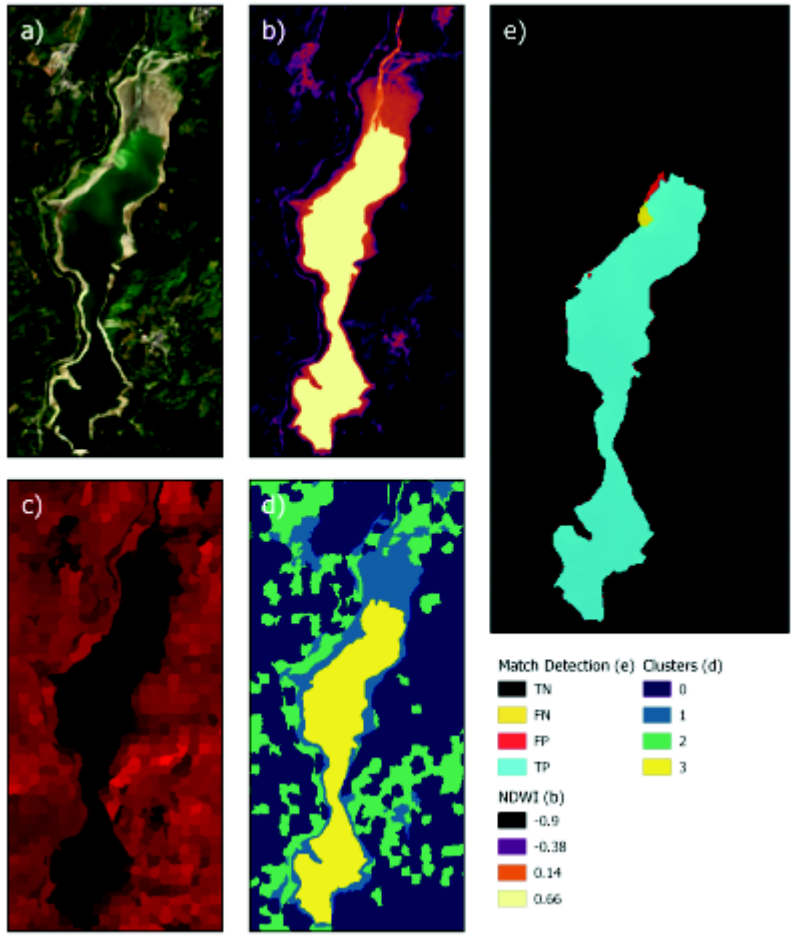
Convolutional Neural Networks (CNN)

- Las CNNs son un tipo de Red Neuronal Artificial que pretende funcionar como las neuronas visuales.
- Matrices bidimensionales como entrada - Visión artificial
- Clasificación y segmentación de imágenes



Posibles aplicaciones

Algoritmos no supervisados para capas

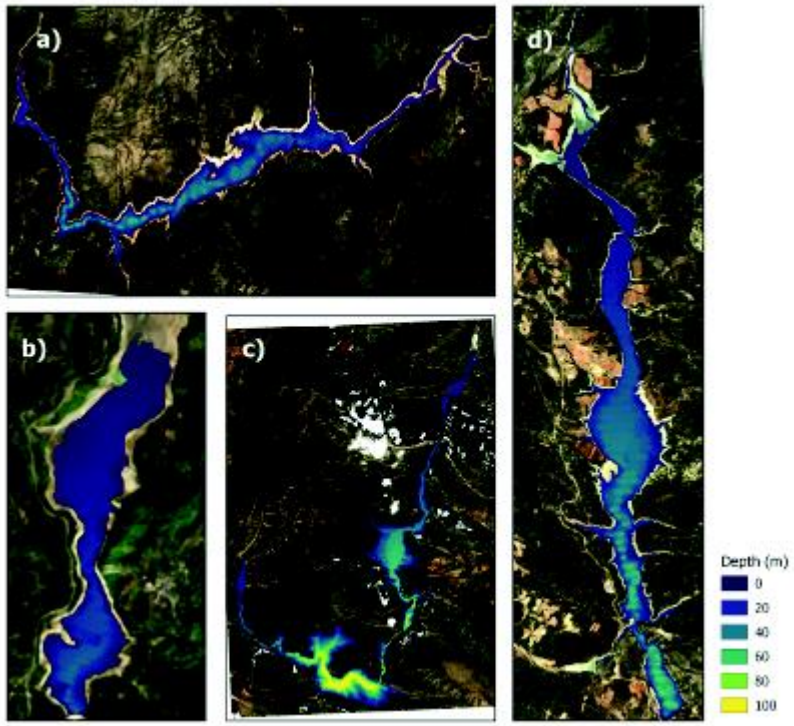


Identificación de objetos

Superpixeles - Características similares

Clustering para agrupación

Estimación de valores



Inferir a partir del entrenamiento con datos dados

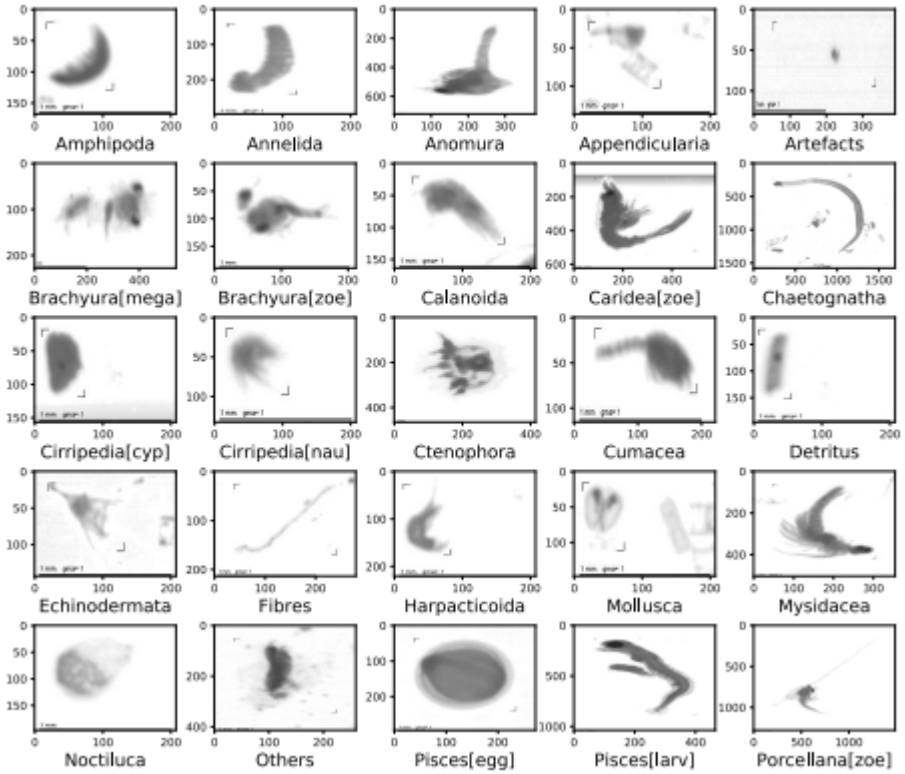
Cálculo de Batimetría

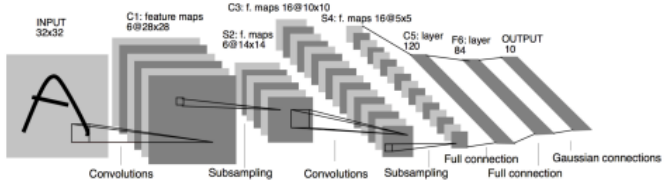
Concentración de algas

Temperatura

Necesidad de monitorización in-situ para obtener los datos de entrenamiento

Clasificación - Zooplancton



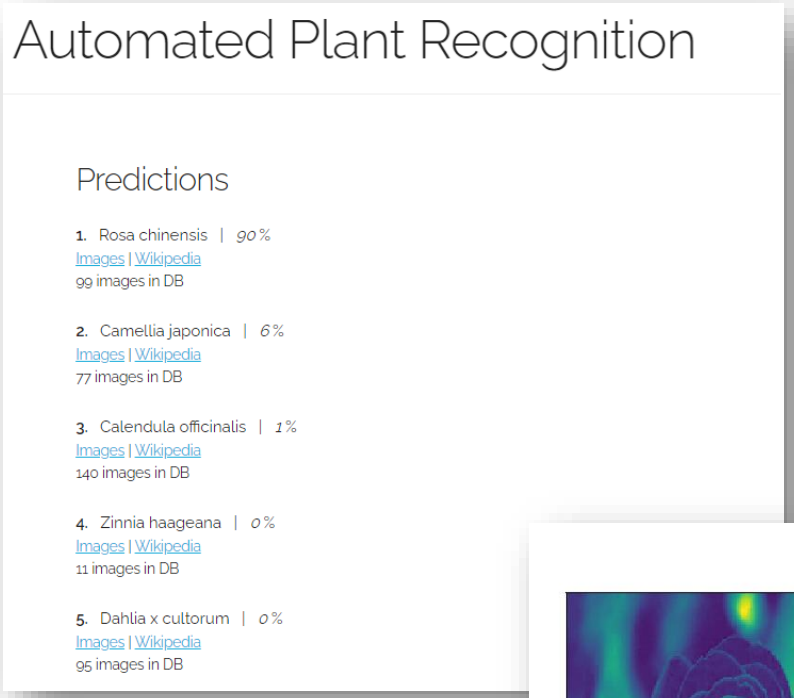


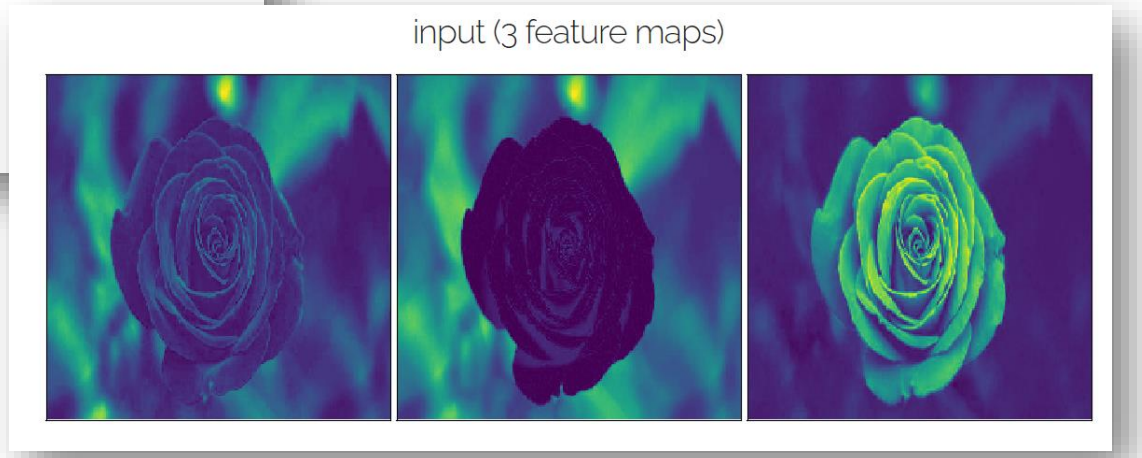
CNN

Aprende de las características

Necesidad de banco de imágenes para el entrenamiento

Clasificación - Plantas





CNN

Aprende de las características

Necesidad de banco de imágenes para el entrenamiento

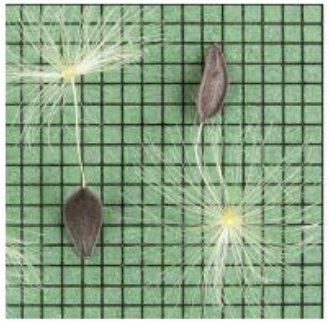
Clasificación - Semillas

This Docker container contains a trained Convolutional Neural network optimized for seeds identification using images. The architecture used is an Xception [1] network using Keras on top of Tensorflow.

The PREDICT method expects an RGB image as input (or the url of an RGB image) and will return a JSON with the top 5 predictions.

As training dataset we have used a collection of images from the Spanish Royal Botanical Garden which consists of around 28K images from 743 species and 493 genera.

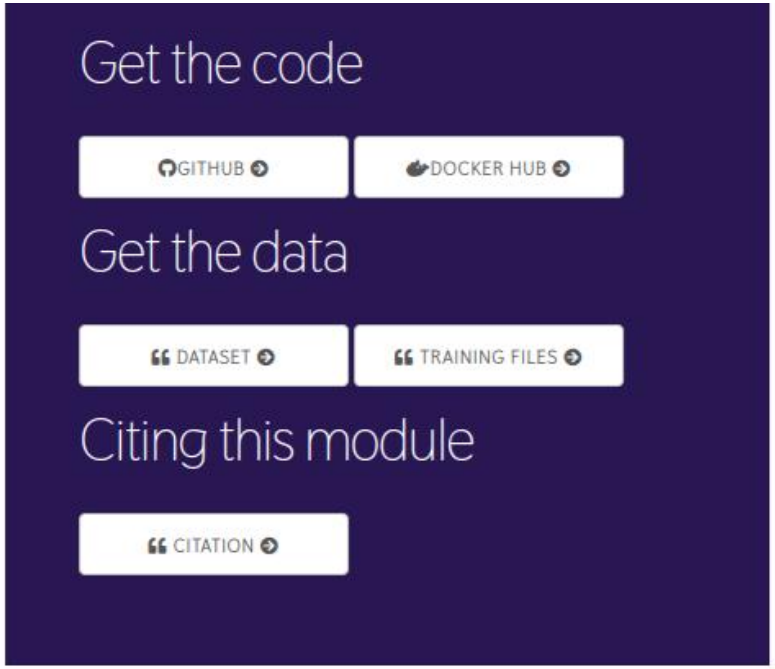




This service is based in the Image Classification with Tensorflow model.

References

[1]: Chollet, Francois. Xception: Deep learning with depthwise separable convolutions arXiv preprint (2017): 1610-02357.

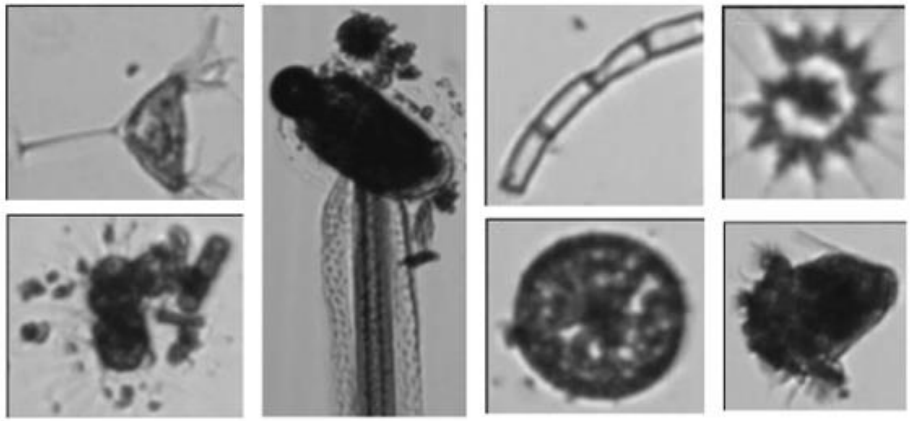


https://marketplace.deep-hybrid-datacloud.eu/modules/deep-oc-seeds-classification-tf.html

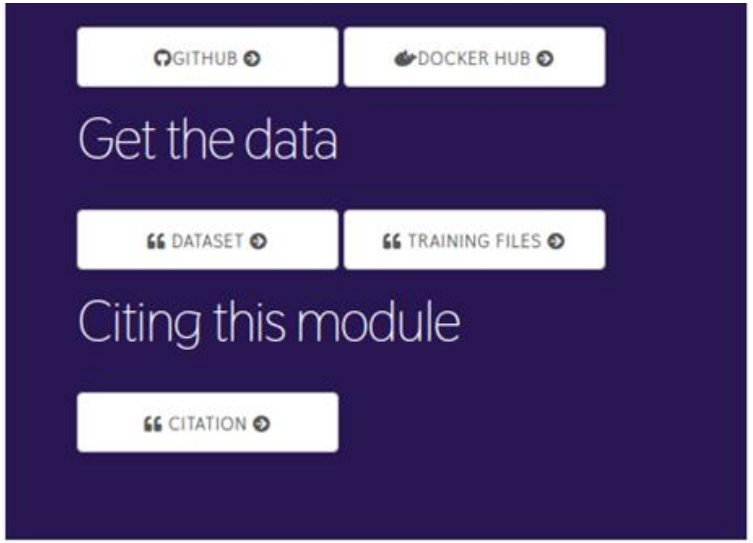
Clasificación - Phytoplankton

The PREDICT method expects an RGB image as input (or the url of an RGB image) and will return a JSON with the top 5 predictions.

As training dataset we have used a collection of images from the Vlaams Instituut voor de Zee which consists of around 650K images from 60 classes of phytoplankton.



This service is based in the Image Classification with Tensorflow model.

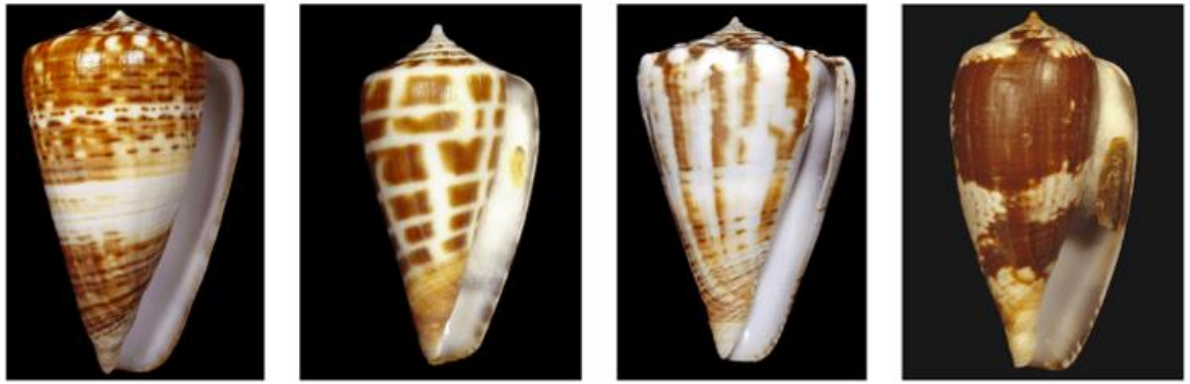


https://marketplace.deep-hybrid-datacloud.eu/modules/deep-oc-phytoplankton-classification-tf.html

Clasificación - Conus

The PREDICT method expects an RGB image as input (or the url of an RGB image) and will return a JSON with the top 5 predictions.

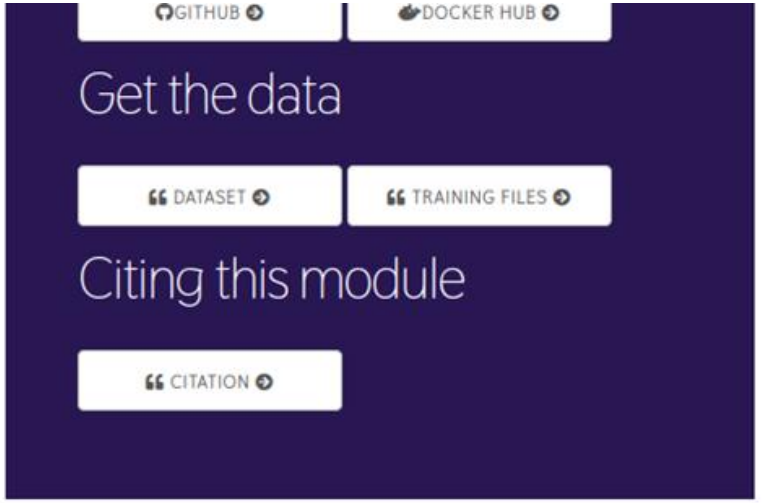
The training dataset has been provided by the Natural Science Museum of Madrid and it consists on a dataset containing images of 68 species of conus covering three different regions: the Panamic region; the South African region; and the Western Atlantic and Mediterranean region.



This service is based in the Image Classification with Tensorflow model.

References

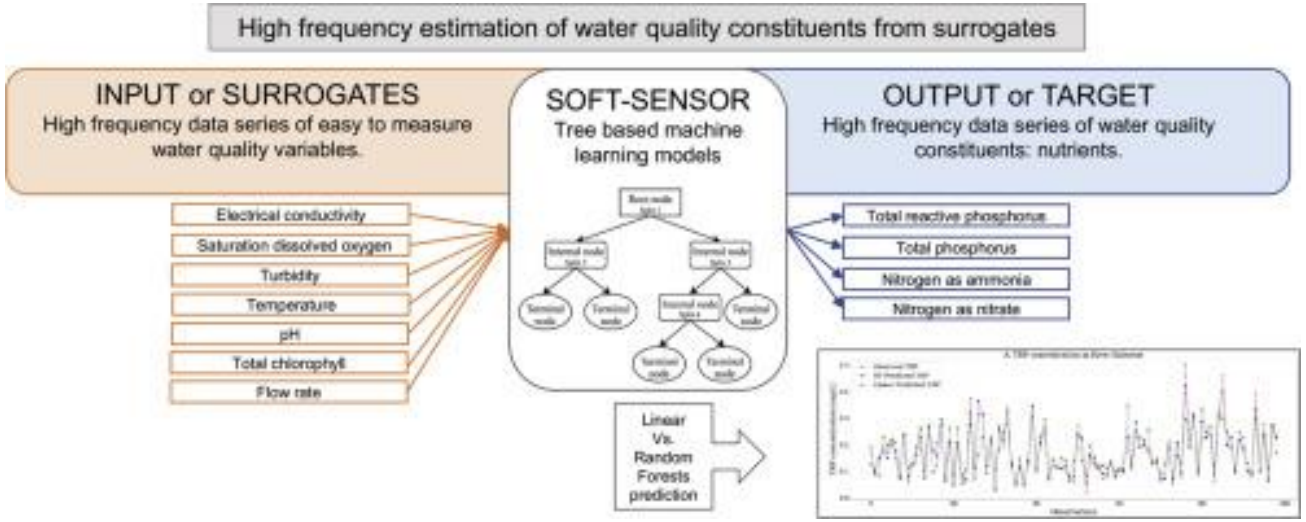
[1]: Puillandre, N.; Duda, T.F.; Meyer, C.; Olivera, B.M.; Bouchet, P. (2014). One, four or 100 genera? A new classification of the cone snails. Journal of Molluscan Studies. 81 (1): 1-23.



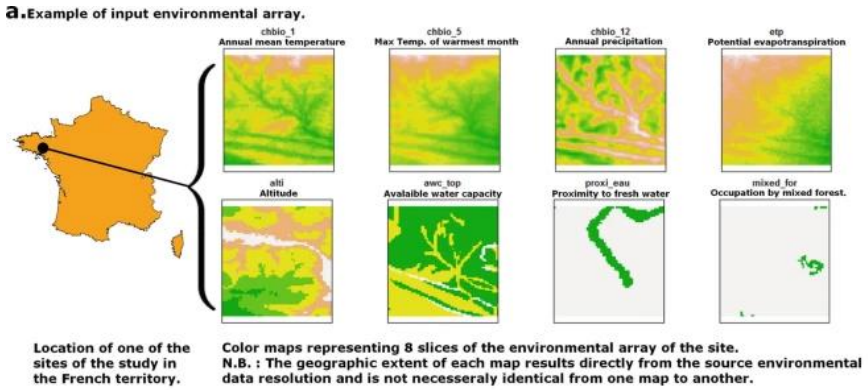
https://marketplace.deep-hybrid-datacloud.eu/modules/deep-oc-conus-classification-tf.html

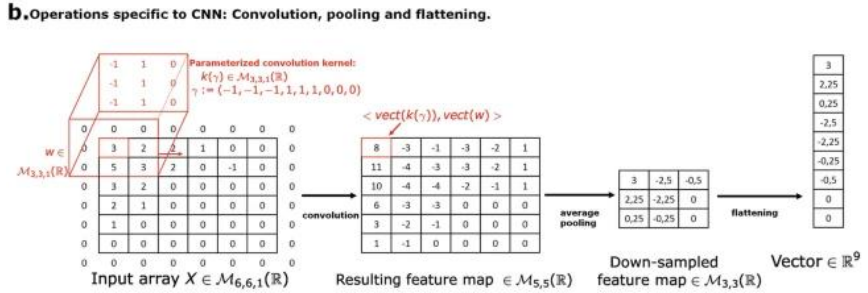
Inferencia de nutrientes

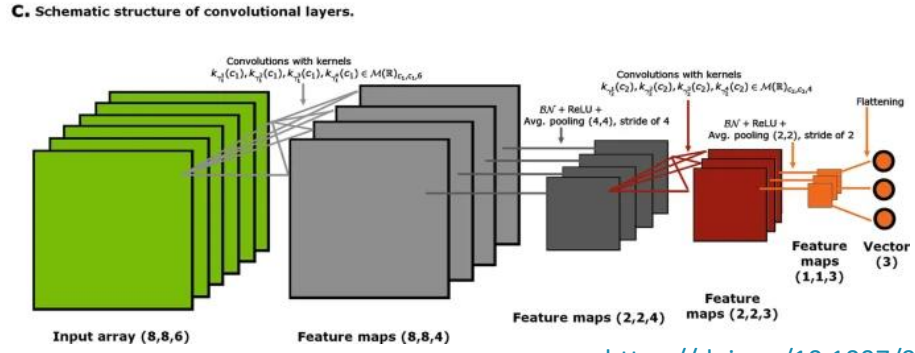
A partir de variables más sencillas de medir



Distribución de especies



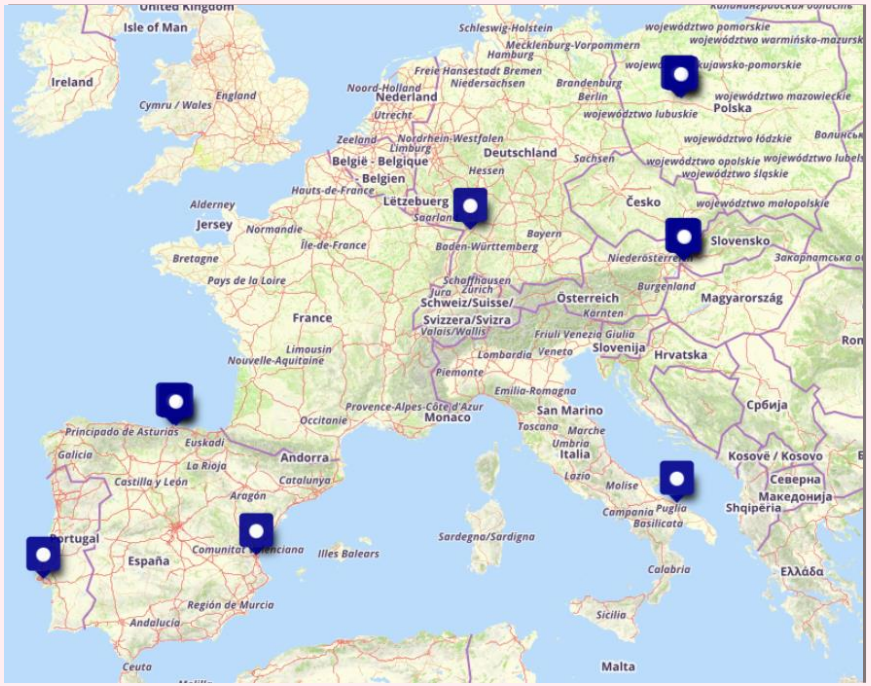




En base a datos de presencia (ausencia)

Capas medioambientales, características

https://doi.org/10.1007/978-3-319-76445-0_10



AI4EOSC

Artificial Intelligence for the #EOSC

- Evolution of the DEEP Hybrid DataCloud platform
- HORIZON-INFRA-2021-EOSC-01-04 call
- Runs September 1st 2022 – August 2025 (36 months)
- 7 academic partners
  + 2 SME
  + 1 non-profit organization

Advanced features for distributed, federated, composite learning, metadata provenance, MLOps, event-driven data processing, and provision of AI/ML/DL services

19

AI4EOSC Expected results

Cloud based AI platform, integrated into the EOSC, with distributed training capabilities

Reusable AI/ML applications offered through AI4EOSC exchange, with easy deployment paths

Best practices and recommendations for AI practitioners and data scientists

MLOps technological framework providing drift detection capabilities

Model provenance metadata framework, covering the whole AI/M

Instituto de Física de Cantabria

CSIC
Consejo Superior de Investigaciones Científicas

UC
Universidad de Cantabria

Inteligencia Artificial y sus posibles aplicaciones en el ámbito de la ciencia y la gestión de la biodiversidad

Fernando Aguilar Gómez