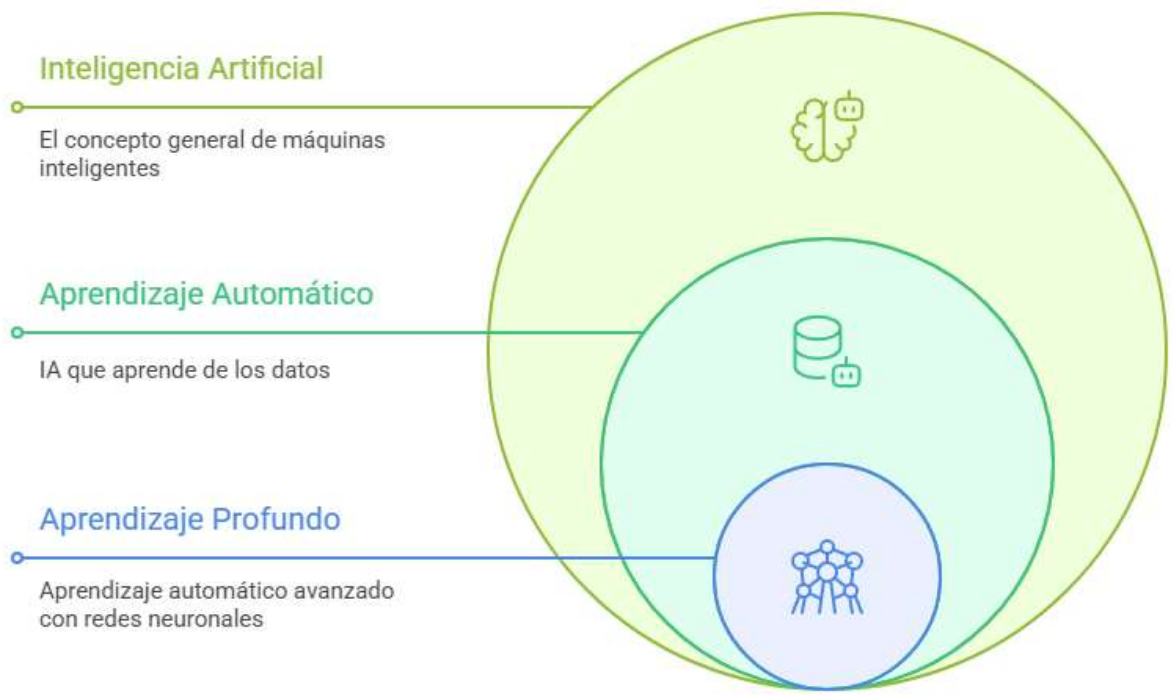
Inteligencia Artificial en el Contexto Geoespacial

Contents

La Inteligencia Artificial (IA) es una disciplina de las ciencias de la computación que estudia el diseño y desarrollo de sistemas capaces de percibir su entorno, razonar sobre la información disponible y ejecutar acciones orientadas a un objetivo, emulando aspectos del comportamiento inteligente humano. En términos formales, un agente de IA busca maximizar una función de rendimiento o perdida a partir de la percepción del entorno y la experiencia previa (Russell & Norvig, 2021).

La IA combina fundamentos de lógica, probabilidad, estadística, optimización, neurociencia y aprendizaje automático, lo que permite construir modelos que pueden aprender representaciones y patrones a partir de grandes volúmenes de datos.

Fig.1.



Nota.

El contexto geoespacial se refiere a todos aquellos datos que contienen una referencia espacial explícita, es decir, que están asociados a una posición o coordenada sobre la superficie terrestre. Esto significa que cada registro o píxel no solo tiene un valor, sino también una ubicación en el espacio.

En el ámbito geoespacial, existen principalmente dos tipos de datos:

Datos ráster: se organizan en forma de celdas o píxeles, como las imágenes satelitales (Sentinel, Landsat, Planet), los modelos digitales del terreno o los mapas de temperatura. Cada píxel representa un valor continuo (por ejemplo, reflectancia, elevación o intensidad).

Fig.2.
Dato Raster

Nota. Ortoimagen

Datos vectoriales: representan objetos definidos geométricamente como puntos, líneas o polígonos. Ejemplos:

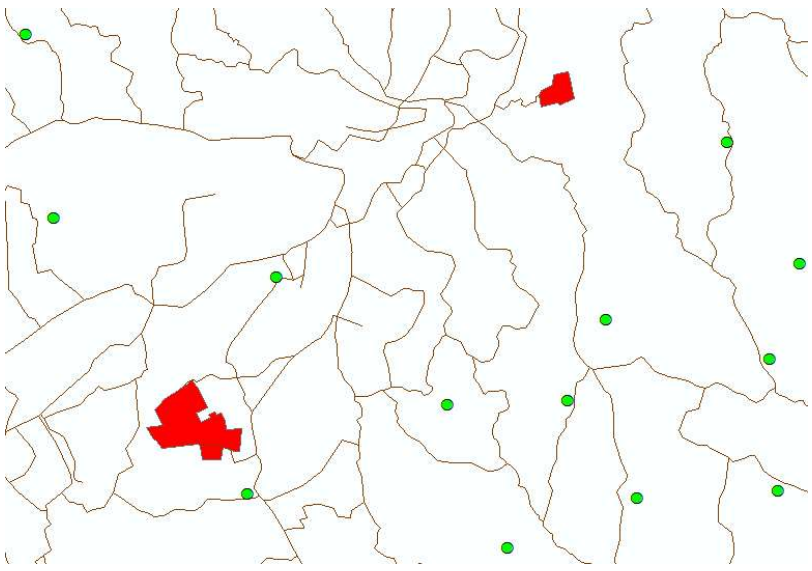
Puntos: ubicación de estaciones meteorológicas

Líneas: redes de carreteras o ríos.

Polígonos: límites de parcelas, municipios o coberturas de suelo.

Fig.3.

Dato Vectorial



Nota.

Ambos tipos de datos son complementarios y esenciales para la toma de decisiones territoriales. Por ejemplo, una capa ráster puede mostrar la vegetación actual (NDVI), mientras que una capa vectorial puede delimitar las zonas agrícolas o urbanas para análisis comparativo.

Principales tareas geoespaciales impulsadas por IA

Tarea	Qué hace	Ejemplo práctico
Clasificación	Asigna una etiqueta o clase a cada píxel o zona.	Clasificar una imagen en “urbano”, “agua”, “vegetación” o “suelo desnudo”.
Segmentación	Delimita objetos dentro de una imagen.	Identificar techos, vías o parcelas agrícolas en una ortofoto.
Detección de cambios	Compara imágenes de distintas fechas para encontrar transformaciones.	Detectar expansión urbana entre 2010 y 2025.
Predicción espacial	Estima la probabilidad de que algo ocurra en una zona.	Predecir riesgo de inundaciones o incendios.
Integración multimodal	Combina distintos tipos de datos.	Unir imágenes satelitales con datos de censos o sensores para analizar vulnerabilidad social.

Metodología CRISP-DM

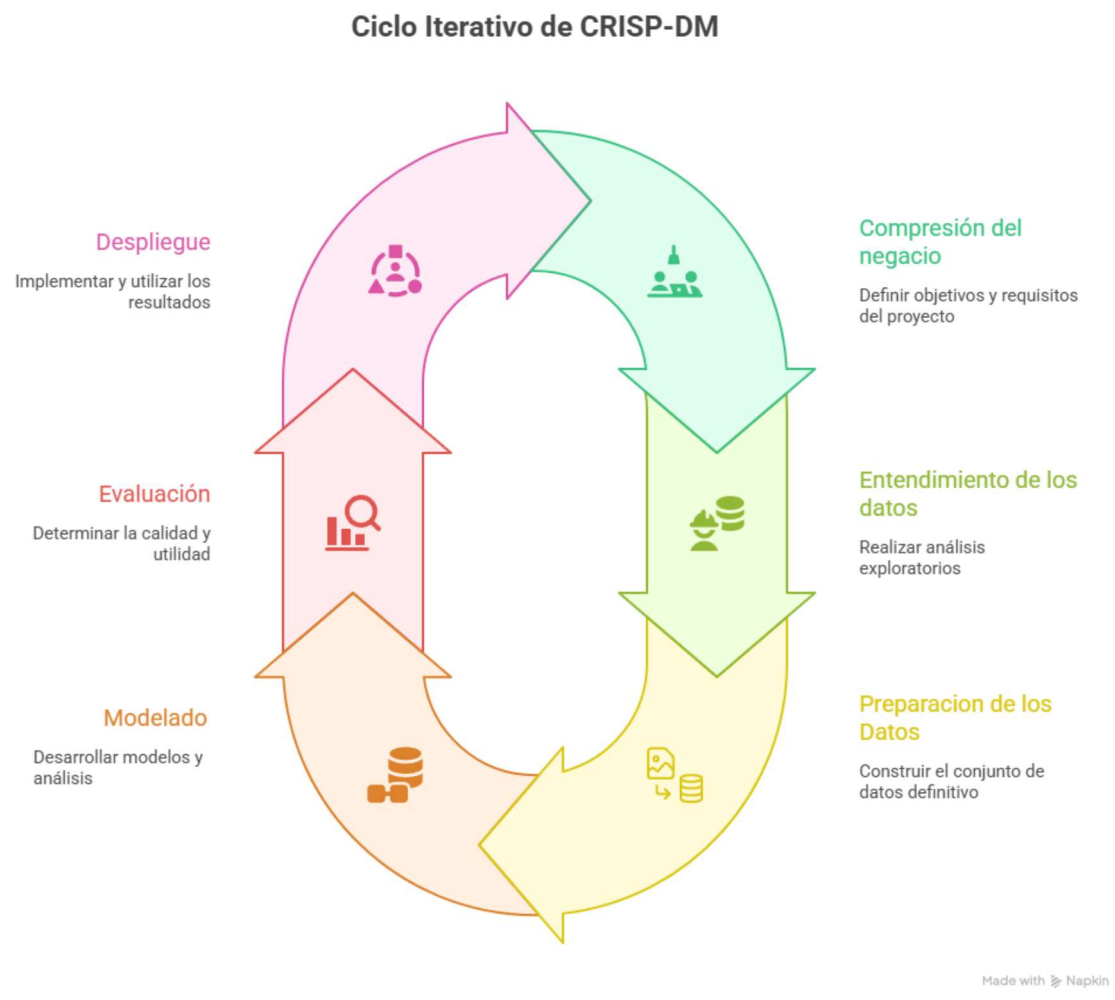
CRISP-DM (Cross-Industry Standard Process for Data Mining) es una metodología estándar desarrollada en 1996 por un consorcio europeo liderado por SPSS, Daimler-Benz, NCR y OHRA. Su propósito fue crear una guía común, práctica y flexible para aplicar técnicas de minería de datos (Data Mining) en diferentes industrias, de forma estructurada, repetible y comprensible.

Está metodología nació para ordenar y estandarizar ese trabajo, estableciendo un lenguaje y flujo de referencia que posteriormente se adaptó a proyectos de Inteligencia Artificial (IA)

Está compuesta por seis fases principales, que se ejecutan de forma iterativa (es decir, pueden repetirse o ajustarse en cualquier momento del proyecto).

Fases del modelo CRISP-DM

Fig.4.Fases CRISP-DM



Nota.

1. **Comprensión del negocio**: Se definen los objetivos del proyecto desde una perspectiva práctica. En el contexto geoespacial, implica responder preguntas como:

- ¿Qué fenómeno territorial o ambiental se quiere analizar?
- ¿Qué impacto tiene el modelo en la gestión o planificación del territorio?

2. **Comprensión de los datos**: Se estudian los datos disponibles para conocer su contenido, calidad y estructura. En proyectos geoespaciales esto incluye:

- Identificar fuentes de imágenes (Sentinel-2, Planet, Ortoimagenes, etc.)
- Revisar bandas espectrales, resolución, fechas y condiciones de nubosidad.
- Analizar shapefiles, clases de cobertura o zonas de interés.

3. **Preparación de los datos**: Es la fase donde se construye el conjunto de datos que el modelo utilizará. En imágenes geoespaciales, se incluyen tareas como:

- Recorte por área de estudio (AOI).
- Corrección radiométrica o atmosférica.
- Generación de patches o ventanas de entrenamiento.
- Etiquetado de píxeles o regiones con base en clases conocidas.

4. **Modelado**: Se selecciona y entrena el modelo de inteligencia artificial.
5. **Evaluación**: Se mide el rendimiento del modelo con métricas cuantitativas. En el caso geoespacial, las más comunes son:

- Exactitud global (Overall Accuracy)
- Índice Kappa
- F1-Score o IoU (Intersection over Union)
- Matriz de confusión espacial Aquí se determina si el modelo cumple los objetivos planteados en la fase de negocio.

1. **Despliegue**: Se aplica el modelo a todo el territorio o área de estudio. El resultado se presenta como un mapa clasificado (GeoTIFF) o una capa vectorial lista para análisis en SIG. También puede integrarse en flujos automatizados o plataformas web de monitoreo.

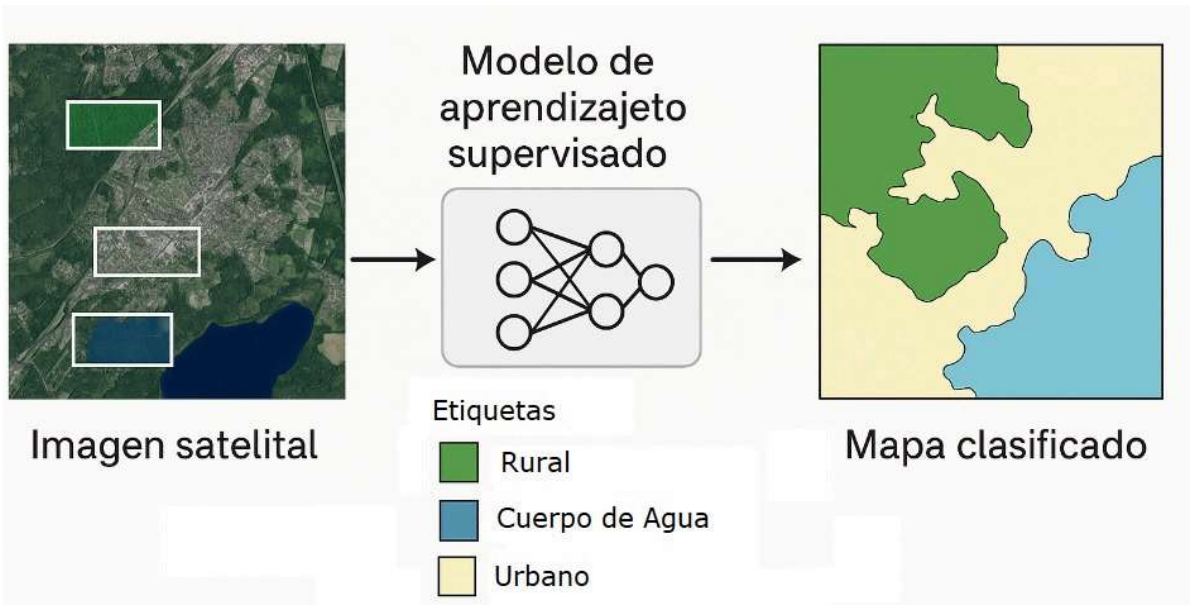
Machine learning supervisado

Aprendizaje supervisado

En el aprendizaje supervisado, los datos de entrenamiento incluyen tanto las características (inputs) como los resultados esperados (outputs). Es decir, el modelo recibe un conjunto de datos de entrenamiento etiquetados con la respuesta correcta. El objetivo es que el modelo pueda aprender a relacionar las características con las respuestas correctas, y posteriormente hacer predicciones precisas sobre nuevos datos que no han sido vistos previamente.

- Predecir un valor continuo para una variable. Se utilizan los valores de otras variables conocidas. Por ejemplo: Predecir el valor de un predio.
- Predecir una clase para un elemento. Por ejemplo: Realizar clasificación de fotografías de fachadas por si tipología constructiva.

Fig.1.
Representación del aprendizaje supervisado en SIG



Nota. Imagen generada con inteligencia artificial por ChatGPT (OpenAI), 2025.

Consideraciones procesamiento espacial para Machine Learning

Consistencia espacial

Componente	Explicación
CRS unificado	El sistema de referencia define cómo se representan los puntos del planeta en una proyección 2D. Si los datasets tienen CRS distinto, las coordenadas representan lugares físicamente diferentes, creando errores de solapamiento entre datos. La consistencia del CRS protege la integridad física del análisis espacial y evita asignaciones incorrectas píxel-etiqueta.
Resolución homogénea	El tamaño físico del píxel determina el nivel de detalle observacional. Remuestrear altera la señal espectral original y puede mezclar objetos diferentes dentro de un píxel. La coherencia de resolución asegura comparabilidad espectral y evita pérdida de información crítica para diferenciar clases.
Alineación de píxeles	Supone coincidir exactamente el origen y transform del grid espacial. Una mínima desalineación cambia la asociación espectro-clase, generando ruido sistemático en el dataset supervisado. Es especialmente crítico en imágenes de alta resolución.
Recorte al Área de Interés	El área de estudio filtra regiones irrelevantes, bordes y NoData. Evita incluir patrones no representativos del fenómeno y previene fugas espaciales que inflan artificialmente la precisión durante validación.

Calidad del dato vectorial

Componente	Explicación
Geometrías válidas	Las etiquetas representan la verdad de terreno. Errores topológicos (self-intersections, huecos) inducen asignaciones incorrectas de píxeles a categorías. La topología correcta garantiza un entrenamiento confiable y científicamente válido.
Consistencia semántica de etiquetas	Cada clase debe tener significado físico estable. Ambigüedad semántica deriva en clases espectralmente superpuestas, aumentando la confusión del modelo y reduciendo la separabilidad de las categorías.
Balance de clases	Las clases dominantes tienen mayor impacto en la función de pérdida. Si no se corrige, el modelo aprende a ignorar clases minoritarias, generando un sesgo estadístico severo.
Eliminación de ruido geométrico	Los objetos mínimos irrelevantes desde la resolución del sensor generan ruido no explicable por el modelo. Eliminar geometría irrelevante mejora la generalización y reduce el overfitting espacial.

Preprocesamiento ráster

Componente	Explicación
Corrección radiométrica	Convierte radiancia en reflectancia, eliminando efectos atmosféricos y geométricos. Esto garantiza que el modelo aprenda propiedades físicas del material y no ruido atmosférico.
Normalización de valores espectrales	Ajusta bandas a escalas numéricas equivalentes para que ninguna domine la función de pérdida. Mejora estabilidad y rendimiento del algoritmo.
Manejo de NoData	Los valores sin información introducen artefactos si se incluyen en el entrenamiento. Deben ser detectados y excluidos para mantener integridad del dataset.
Índices espectrales y transformaciones	Extraen información física compacta del espectro (NDVI, NDBI) o reducen dimensiones redundantes (PCA), mejorando la relación señal-ruido para ML.

Ingeniería espacial de características

Componente	Explicación
Texturas y patrones espaciales	La distribución espacial de valores captura organización del terreno. Métricas como GLCM permiten diferenciar clases similares espectralmente mediante estructura espacial.
Variables topográficas	La forma del relieve condiciona procesos ecológicos y urbanos. Integrar pendiente, orientación y curvatura aporta conocimiento físico del territorio.
Distancias y proximidad	La ubicación relativa a elementos clave (vías, ríos, límites urbanos) añade semántica geográfica que mejora la capacidad de clasificación del modelo.
Estadísticos de vecindario	La información agregada del entorno reduce ambigüedad espectral y mejora robustez en áreas heterogéneas.

Estructuración del dataset final

Componente	Explicación técnica profunda y razón de importancia
Representación de X e y	Los predictores y las etiquetas deben estar organizados en estructuras compatibles con el modelo (matrices o tensores). La correcta alineación dimensional es crucial para reproducibilidad científica y operacional.
Metadatos GIS	Información como CRS, resolución y transform debe preservarse para aplicar el modelo sobre nuevos datos sin pérdida de referencia espacial.
División espacial de datos	La forma en que se separan los conjuntos de entrenamiento y prueba afecta directamente la estimación de desempeño y la capacidad de generalización espacial del modelo.

Evaluación de los modelos de machine learning

El objetivo del aprendizaje automático es lograr que los modelos generalicen, es importante poder medir de manera confiable el poder de generalización del modelo.

El método **hold- out** para entrenar un modelo de machine learning es el proceso de dividir los datos en dos partes y usar una división para entrenar el modelo y la otra probar los modelos. Normalmente se utiliza 80-20

Conjuntos de entrenamiento y prueba

- **Conjunto de entrenamiento:** Un subconjunto para entrenar un modelo. Debe ser lo suficientemente grande como para generar resultados significativos desde el punto de vista estadístico
- **Conjunto de prueba:** Un subconjunto para probar el modelo entrenado. Debe ser representativo del conjunto de datos en su totalidad. En otras palabras, no elijas un conjunto de prueba con características diferentes a las del conjunto de entrenamiento.

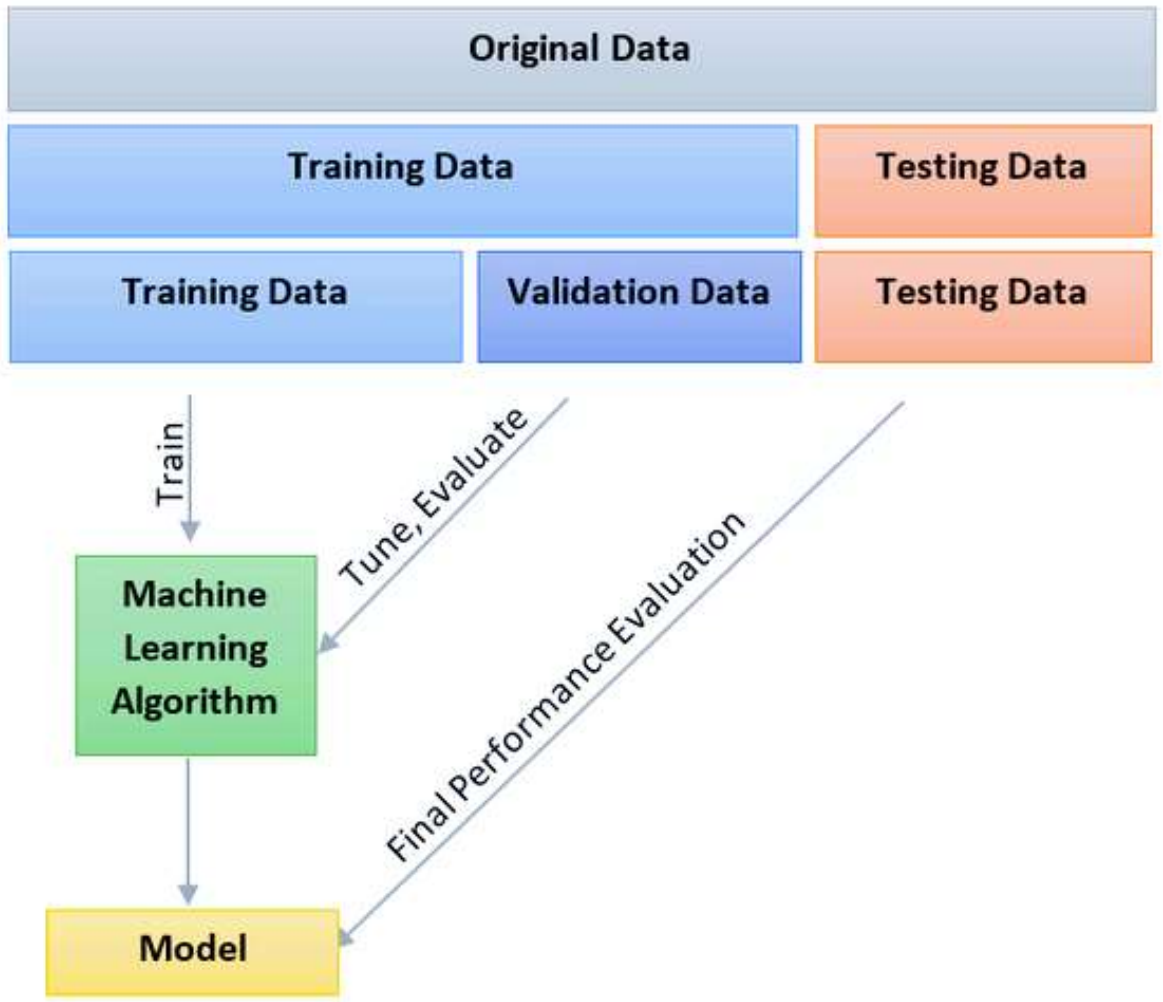
Nota

Al dividir el conjunto de datos es importante tener en cuenta estos aspectos.

Representatividad de los datos: es deseable que los tres conjuntos sean representativos de los datos.

Redundancia en sus datos: Los 3 o 2 conjuntos de datos deben se disjuntos.

Fig.2.
Conjunto de entrenamiento



Nota. Imagen generada con inteligencia artificial por ChatGPT (OpenAI), 2025.

Caso de Estudio: Clasificación de la Cobertura del Suelo en 6 Municipios de Cundinamarca

Se busca clasificar la cobertura del suelo a Nivel 1 en seis municipios de Cundinamarca para apoyar la gestión territorial y ambiental. Se utilizan imágenes Landsat 8 (bandas B2-B7) y capas vectoriales de referencia, aplicando un modelo de clasificación supervisada.

Las clases a identificar son:

Nivel 1	Nombre
1	Territorios Artificializados
2	Territorios Agrícolas
3	Bosques y Áreas Seminaturales
4	Áreas Húmedas
5	Superficies de Agua

Tabla 1. Clasificación de Uso del Suelo - Nivel 1

Análisis del Problema

Objetivo: Implementar modelos de Machine Learning supervisado para clasificar coberturas.

Conceptos:

- Árboles de Decisión
- Random Forest
- Conjunto de árboles de decisión que mejora la precisión y evita el sobreajuste.

Recursos

Recursos

Capas Raster (Landsat 8 SR):

- LC08_L2SP_008056_20200322_20200822_02_T1_SR_B2_MUESTRA.TIF
- LC08_L2SP_008056_20200322_20200822_02_T1_SR_B3_MUESTRA.TIF
- LC08_L2SP_008056_20200322_20200822_02_T1_SR_B4_MUESTRA.TIF
- LC08_L2SP_008056_20200322_20200822_02_T1_SR_B5_MUESTRA.TIF
- LC08_L2SP_008056_20200322_20200822_02_T1_SR_B6_MUESTRA.TIF
- LC08_L2SP_008056_20200322_20200822_02_T1_SR_B7_MUESTRA.TIF

Capa vectorial:

- e_cobertura_tierra_2020_muestra.shp

Cuaderno

- D2M1_2_supervisado_ejercicio.ipynb

Nota

El nombre del archivo LC08_L2SP_008056_20200322_20200822_02_T1_SR_B2_MUESTRA.TIF sigue la convención de nomenclatura oficial de Landsat 8.

Fragmento	Significado
LC08	Satélite Landsat 8 OLI/TIRS
L2SP	Producto Nivel 2 Surface Reflectance (corregida atmosféricamente)
008056	Código de trayectoria/escena (Path 008, Row 056)
20200322	Fecha de adquisición de la imagen: 22 de marzo de 2020
20200822	Fecha de procesamiento: 22 de agosto de 2020
02	Versión del producto de procesado
T1	Tier 1: alta calidad geométrica
SR	Surface Reflectance (reflectancia superficial)
B2	Banda 2 (Azul)
.TIF	Formato raster GeoTIFF

Resolución espacial: 30 m

- **Bandas utilizadas para clasificación:**
  - **B2** Azul
  - **B3** Verde
  - **B4** Rojo
  - **B5** NIR
  - **B6** SWIR 1
  - **B7** SWIR 2

En el campo del Machine Learning supervisado, los algoritmos basados en árboles de decisión son herramientas habitualmente usadas para la clasificación y predicción. Entre estos, se tienen : Árboles de Decisión, Random Forest y XGBoost.

Árbol de decisión

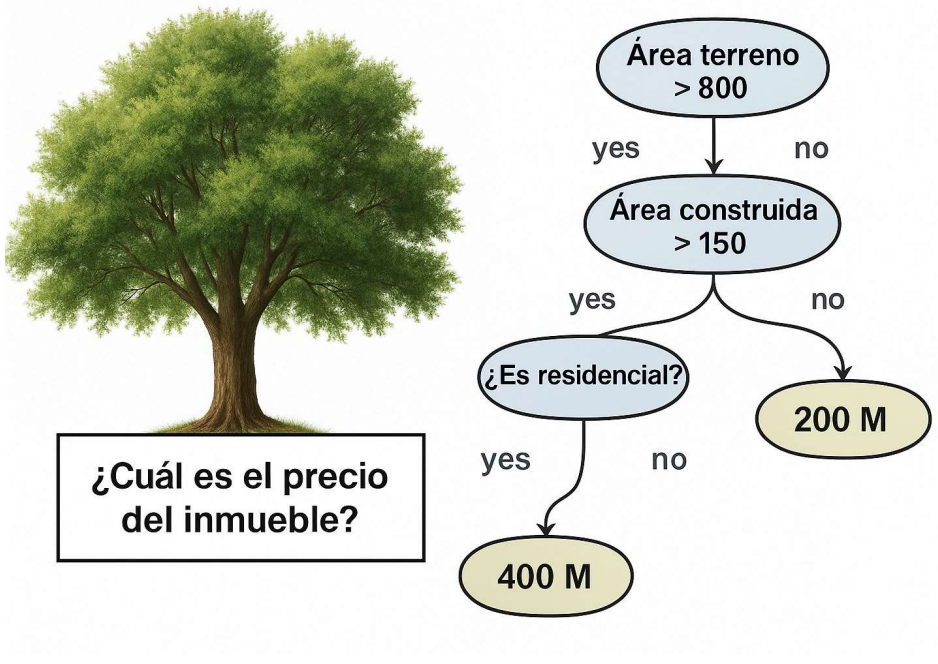
Es una estructura jerárquica de nodos de decisión y hojas que se utiliza para modelar relaciones no lineales entre variables y predecir el valor de una variable objetivo en función de un conjunto de variables predictoras.

El objetivo es clasificar de una forma simple mediante análisis estadístico y teoría de la información. Cada nodo de decisión se basa en una regla que divide el conjunto de datos en dos o más subconjuntos, y cada hoja representa una predicción para la variable objetivo.

Los árboles de decisión son útiles porque son fáciles de interpretar, permiten manejar datos faltantes y son resistentes a valores atípicos. Además, se pueden utilizar para clasificación o regresión, dependiendo de si la variable objetivo es categórica o numérica, respectivamente.

Interpretabilidad Gráfica.

Fig.3.
Representación del aprendizaje supervisado en SIG



Nota. Imagen generada con inteligencia artificial por ChatGPT (OpenAI), 2025.

En el nodo raíz se elige aquella variable que aporta más información a la predicción. Los ejemplos se dividen en grupos con distintos valores para esta clase.

- El algoritmo continúa dividiendo los nodos con la elección de la mejor variable hasta que se alcance algunos de los siguientes criterios de parada:

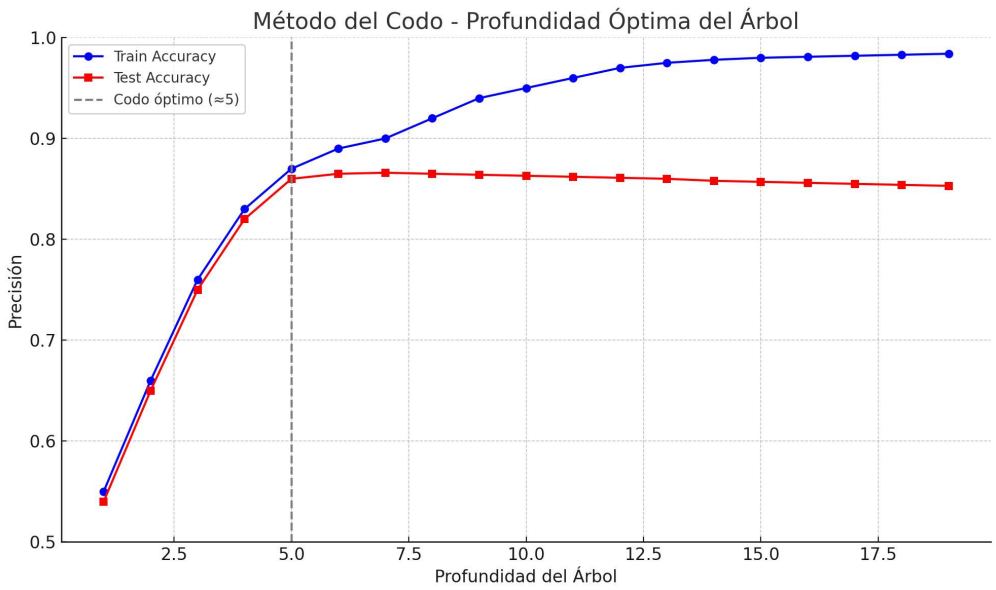
1. Todos (casi todos) los ejemplos del nodo son de la misma clase.
2. No existen variables para distinguir entre los ejemplos.
3. El árbol ha alcanzado un tamaño predefinido.

Sin embargo, usualmente se desea detener este proceso antes de que alguno de los anteriores puntos suceda a esto se le llama poda del árbol.

Poda

Podar un nodo de un árbol de decisión consiste en eliminar un subárbol anidado en ese nodo transformándolo en una hoja y asignándole la clasificación más común de los ejemplos de entrenamiento considerados en ese nodo. Existe un método para hallar el número de ramas optima en la poda llamado el método del codo.

Fig.4.
Gráfico del método del codo



Nota. Fuente propia.

📌 Ejemplo interactivo

Random Forest

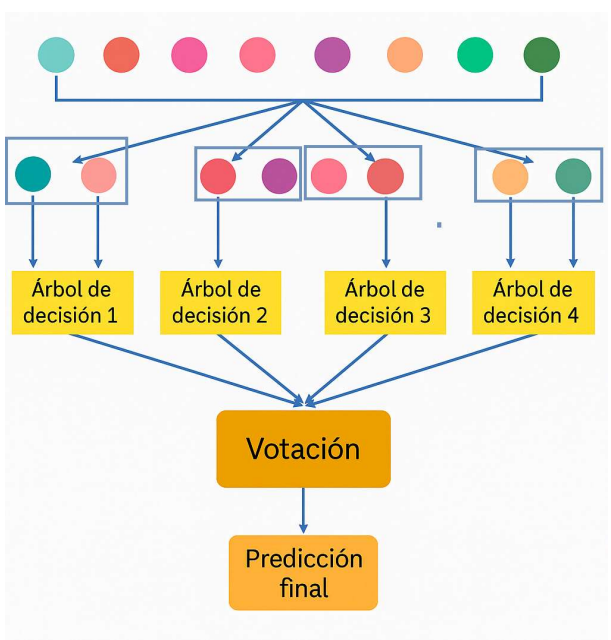
Random Forest es un algoritmo basado en árboles de decisión, diseñado para mejorar la precisión y reducir el sobreajuste. Funciona mediante la creación de múltiples árboles de decisión y la combinación de sus resultados para obtener predicciones más confiables. En primer lugar,comienza con la construcción de un conjunto de árboles de decisión. Pero en lugar de entrenar un solo árbol con todos los datos, Random Forest introduce aleatoriedad:

- **Selección aleatoria de datos (Bootstraping)**: Random Forest genera múltiples subconjuntos de datos, tomados de manera aleatoria y con reemplazo. Esto significa que algunas muestras pueden repetirse en diferentes árboles, mientras que otras pueden no ser seleccionadas en absoluto.
- **Selección aleatoria de características (Feature Bagging):** Cuando un árbol de decisión debe hacer una división en los datos, en lugar de considerar todas las variables disponibles, se elige aleatoriamente un subconjunto más pequeño. Esto evita que todos los árboles sean demasiado similares y hace que el modelo sea más robusto.

Finalmente, una vez que todos los árboles han sido entrenados, se realizan las predicciones.

- Para problemas de clasificación, cada árbol vota por una categoría y la que recibe más votos se convierte en la predicción final. Es como si un grupo de jueces emitiera su veredicto y se tomara la decisión por mayoría.
- Para problemas de regresión, cada árbol genera un valor numérico y se calcula el promedio de todas las predicciones, obteniendo así un resultado más estable y preciso.

Fig.5.
Random Forest



Nota. Imagen generada con inteligencia artificial por ChatGPT (OpenAI), 2025.

La razón de utilizar un gran número de árboles es para que cada característica tenga la oportunidad de aparecer en varios modelos.

- Reduce la varianza cuando se toma el promedio de los árboles.
- En random forests, a la hora de hacer un split se realiza una selección aleatoria de m predictores del total de p. (defecto m == sqrt(p))

XGBoost: Un Algoritmo de Gradient Boosting Optimizado

Este algoritmo sigue un enfoque secuencial: cada nuevo árbol se construye para corregir los errores del anterior. El proceso se basa en la técnica de Gradient Boosting, donde los árboles se ajustan minimizando el error residual en cada iteración.

XGBoost comienza con la construcción de un árbol de decisión inicial, que realiza una primera predicción sobre los datos. Sin embargo, este primer intento no es perfecto, por lo que se comparan sus predicciones con los valores reales, calculando así los errores o residuos. Cada nuevo árbol que se agrega al modelo tiene la tarea de corregir los errores del árbol anterior. De esta forma, el modelo aprende progresivamente, enfocándose en los casos más difíciles de predecir en cada iteración.

Finalmente, para la clasificación, XGBoost predice probabilidades y usa funciones como sigmoide o softmax para tomar una decisión final. Para regresión, el modelo ajusta valores numéricos sumando pequeñas mejoras hasta llegar a una buena estimación.

Fig.6.
Estructura del modelo XGBoost

Nota. Tomada de *Forecasting Multi-Step Ahead Monthly Reference Evapotranspiration Using Hybrid Extreme Gradient Boosting with Grey Wolf Optimization Algorithm*, s. f., https://www.researchgate.net/figure/The-structure-of-XGB-model_fig2_346246036.

Conclusiones

La selección de algortimo consiste en elegir el modelo más adecuado según el tipo de problema y los datos disponibles. Para tomar esta decisión, se deben considerar factores como si el problema es de clasificación o regresión, si hay un desbalance en las clases, la cantidad de datos y la necesidad de interpretabilidad.

Característica	Árbol de Decisión	Random Forest	XGBoost
Estrategia	Un solo árbol de reglas	Bagging (Árboles independientes)	Boosting (Árboles secuenciales)
Construcción	Se entrena de forma única	Se entrenan en paralelo	Se entrenan de forma secuencial
Corrección de errores	No ajusta errores previos	No ajusta errores previos	Cada árbol corrige errores del anterior
Velocidad	Rápido en conjuntos pequeños	Más lento en grandes volúmenes de datos	Optimizado para ser más rápido
Manejo de sobreajuste	Tiende a sobreajustarse	Usa aleatoriedad para reducir sobreajuste	Incluye regularización para evitar sobreajuste
Uso recomendado	Problemas simples con interpretabilidad	Datos con muchas variables y patrones diversos	Problemas complejos con datos estructurados y ruido reducido

Referencias

- r2d3.us. (s.f.). *Una introducción visual al machine learning*. http://www.r2d3.us/una-introduccion-visual-al-machine-learning-1/
- IBM. (s.f.). *Supervised learning*. https://www.ibm.com/think/topics/supervised-learning
- Amazon Web Services. (s.f.). *Diferencias entre aprendizaje supervisado y no supervisado*. https://aws.amazon.com/es/compare/the-difference-between-machine-learning-supervised-and-unsupervised/
- Song, Y. Y., & Lu, Y. (2015). Decision tree methods: Applications for classification and prediction. *Shanghai Archives of Psychiatry*, *27*(2), 130–135. https://doi.org/10.11919/j.issn.1002-0829.215044
- scikit-learn.org. (s.f.). *Árboles de decisión*. https://scikit-learn.org/1.5/modules/tree.html
- scikit-learn.org. (s.f.). *Métodos ensemble*. https://scikit-learn.org/1.5/modules/ensemble.html
- Paperspace. (s.f.). *XGBoost: A comprehensive guide to model overview, analysis and code demo*. https://blog.paperspace.com/xgboost-a-comprehensive-guide-to-model-overview-analysis-and-code-demo-using/

- ChatGPT (OpenAI). (2025). Representación del aprendizaje supervisado en SIG [Imagen generada por IA].
- The structure of XGB model [Figura]. En Forecasting Multi-Step Ahead Monthly Reference Evapotranspiration Using Hybrid Extreme Gradient Boosting with Grey Wolf Optimization Algorithm. ResearchGate. Recuperado de https://www.researchgate.net/figure/The-structure-of-XGB-model_fig2_346246036

-Russell, S. J., & Norvig, P. (2021). *Artificial intelligence: A modern approach* (4th ed.). Pearson.