université PARIS-SACLAY

IMPUTACIÓN DE DATOS HIDROLÓGICOS USANDO ALGORITMOS DE MACHINE LEARNING

Mayo 2025

Julian Agudelo con elementos de Antoine Cornuéjols y de Gelman & Hill
julian.agudeloacosta@agroparistech.fr

MIA PARIS-SACLAY
EKINOCS

AGROPARISTECH
Institut des Sciences et Industries du Vivant et de l'Environnement

Introducción

Contenidos

¿De qué vamos a estar hablando hoy?

- ¿Qué es la Inteligencia Artificial y que es el Aprendizaje Automático (ML)?
- Valores faltantes en hidrología: tipos, razones y consecuencias.
- **Técnicas de ML para la imputación de datos hidrológicos.**
  - K-Nearest Neighbors (KNN)
  - **MissForest:** Una técnica de imputación basada en bosques aleatorios.
  - Perceptrones multicapa (MLPs).
    - Classical substitution.
    - Network Reduction.
  - Comentario sobre la imputación de datos con técnicas de aprendizaje profundo moderno.
- **Caso práctico !**

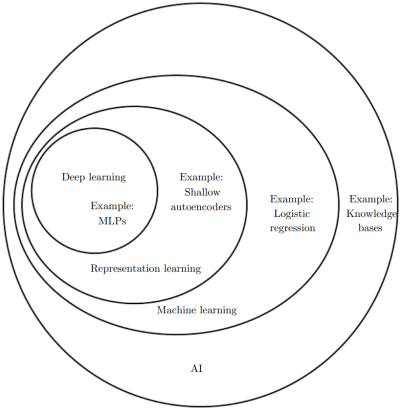
Inteligencia Artificial

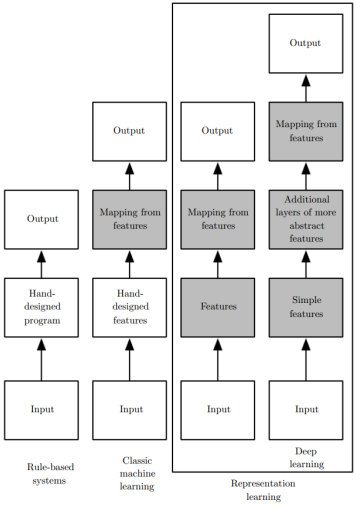
¿Puede pensar una máquina?

El origen de la Inteligencia Artificial como disciplina científica

Máquinas inteligentes

El mapa de la Inteligencia Artificial





Valores faltantes

Datos faltantes en hidrología

Los datos faltantes para una variable suelen comprometer los datos observados de las otras, lo que desencadena una pérdida global de información.

Históricamente, se han usado **métodos estadísticos** como la Regresión Lineal Múltiple (MLR) o **modelos físicos**, como los modelos hidrológicos *per se* para imputar valores faltantes.

En los últimos años, el aprendizaje automático se ha usado en todo tipo de aplicaciones y ha mostrado ser eficaz para tareas de imputación.

¡La elección de la metodología para imputar datos debe basarse en la naturaleza de los mismos!

Tipos de valores faltantes - I

MCAR: Missing Completely at Random

Los valores faltantes completamente al azar no tienen relación con ningúna otra variable observada. **La probabilidad de ausencia es la misma para todas las observaciones.**En este caso **La imputación es aconsejable. Descartar la observación no sesgará los datos,** aunque supondrá una pérdida de tamaño de la muestra.

Tipos de valores faltantes - II

MAR: Missing at Random

Los valores faltantes de manera aleatoria son aquellos valores cuya probabilidad de ausencia depende de una o varias de las otras variables observadas. **Dada esta dependencia, estos valores deben imputarse.**

Tipos de valores faltantes - III

MNAR: Missing Not at Random

La probabilidad de ausencia depende de la variable en cuestión.

Tipos de valores faltantes - Résumen

MCAR: Missing Completely at Random

La probabilidad de ausencia es la misma para todas las observaciones de una variable.

✓ **Imputar o descartar**

MAR: Missing at Random

La probabilidad de ausencia está vinculada a otra u otras variables observadas.

✓ **Imputar**

MNAR: Missing Not at Random

La probabilidad de ausencia depende de la variable en cuestión.

✓ **Imputar y hacer un análisis de sensibilidad**

Tipos de valores faltantes - Formalismo

Siendo $X = (x_{ij}) \in \mathbb{R}^{m \times n}$ la matriz rectangular de datos para n variables $\{X_1 ... X_n\}$ y m observaciones. Consideremos $F = (f_{ij})$ la matriz de indicación de los valores faltantes, que va a definir la repartición de los mismos.

Definamos a los valores observados como $X_{obs} = X\mathbb{1}_{\{F=0\}}$ y a los valores faltantes como $X_{miss} = X\mathbb{1}_{\{F=1\}}$. De modo que el conjunto de datos será $X = \{X_{obs}, X_{miss}\}$

$\mathbb{1}$ **es la función de Kronecker.**

Tipos de valores faltantes - Formalismo II

MCAR

$$p(F|X) = p(F) \text{ para todo } X$$

MAR

$$p(F|X) = p(F|X_{obs}) \text{ para todo } X_{miss}$$

MNAR

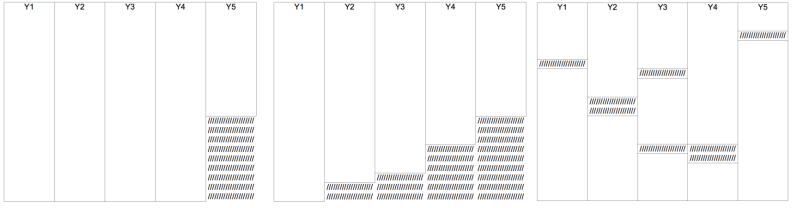
$$p(F|X) = p(F|X_{obs}, X_{miss}) \text{ para todo } X$$

Identificando la distribución de los valores faltantes

a) **Univariados:** para una única variable X_k, si una observación x_{ki} es un valor faltante, no habra mas observaciones de dicha variable.

b) **Monotonos:** si x_{ki} es un valor faltante, esto implica que $\{X_k\}_{k>j}$ seran datos faltantes para dicha observación.

c) **Arbitrarios.**



(a) (b) (c)

Técnicas de ML para la imputación de datos hidrológicos

Caso práctico