université PARIS-SACLAY

IMPUTACIÓN DE DATOS HIDROLÓGICOS USANDO ALGORITMOS DE MACHINE LEARNING

Junio 2025

Julian Agudelo con elementos de Antoine Cornuéjols y de Gelman & Hill
julian.agudeloacosta1@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 Imputer (KNN-I)
  - **MissForest:** Una técnica de imputación basada en bosques aleatorios.
  - Perceptrones multicapa (MLPs).
    - Classical substitution (En una segunda sesión dedicada al DL).
    - Network Reduction (En una segunda sesión dedicada al DL).
  - 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

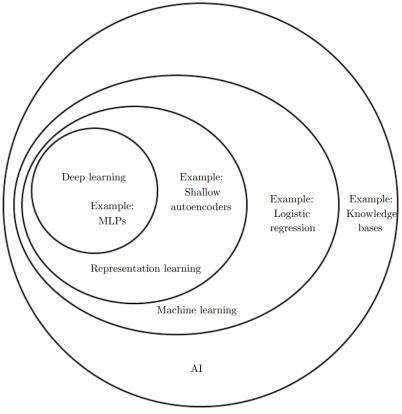
El objetivo de la inteligencia artificial es explorar formalmente la posibilidad de crear máquinas capaces de simular la inteligencia humana. **El objetivo de crear dichas maquinas es entender los mecanismos inherentes que gobiernan el cerebro de humanos y otros animales.**

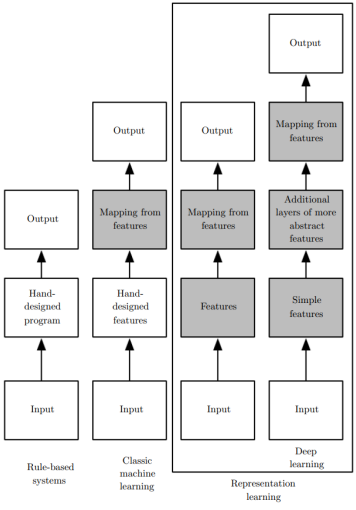
Máquinas inteligentes

≪Se dice que un programa aprende de la experiencia E con respecto a una tarea T y una medida de rendimiento P, si su rendimiento resolviendo T, medido por P, mejora con la experiencia E.≫

Machine Learning, 2003
Tom Mitchell

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

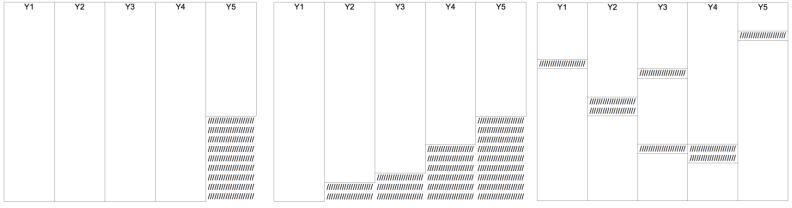
$$p(F|X) = p(F|X_{obs}, X_{miss}) \text{ para todo } X$$

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

K-Nearest Neighbors Imputer - I

KNN es uno de los algoritmos de aprendizaje supervisado más extendidos.

Nacido como un algoritmo de clasificación ($\sim$1967), cuanta con margenes teóricas bien conocidas:

Un algoritmo 1-NN tiene un error de clasificación que es, como mucho, el doble del riesgo óptimo.

Nearest Neighbor Pattern Classification, 1967
Cover and Hart

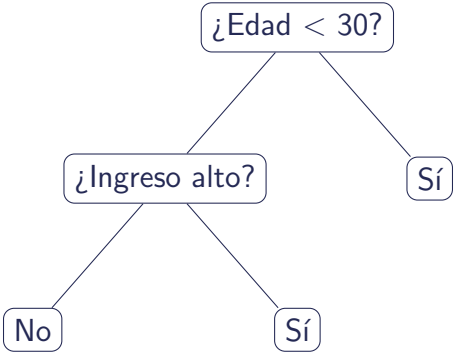
K-Nearest Neighbors Imputer - III

$X = (x_{ij}) \in \mathbb{R}^{m \times n}$, un conjunto de datos
$q = x_{ij} | f_{ij} = 1$, un valor faltante
k, un número arbitrario de vecinos

1: **function** $\mathrm{KNN}(X, q, k)$
2: **for all** (x_i) in D **do**
3: Compute distance d between q and x_i using
4: Store (x_i, d)
5: **end for**
6: Sort all (d) pairs by d
7: Take the k nearest individuals
8: Compute the absent values of q using aggregations of the values from k
9: **end function**

Árboles de Decisión

Los árboles de decisión son modelos o algoritmos no paraméticos que se utilizan principalmente para la resolución de problemas de clasificación, en los que hay que predecir las distintas categorías de la variable objetivo. **Se construye dividiendo el espacio de características recursivamente.**



Árboles de Decisión: Entropia

Los arboles de décision pueden utilizarse para tareas de regresión o clasificación. **A continuación, estudiaremos cómo se forma un árbol de decisión para clasificación utilizando el criterio de ganancia de información.**

La entropía mide el desorden de un conjunto S de observaciones:

$$H(S) = -\sum_{c \in C} p(c) \log_2 p(c)$$

donde:

- C: conjunto de clases posibles.
- $p(c)$: proporción de ejemplos en S que pertenecen a la clase c.

Árboles de Decisión: Ganancia de información

La ganancia de información al dividir un conjunto S usando un atributo A que genera particiones $S_1, \ldots, S_k$:

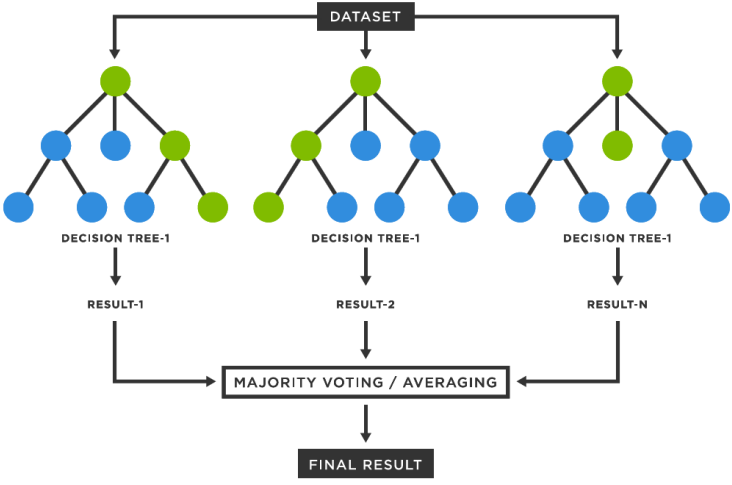
$$IG(S, A) = H(S) - \sum_{i=1}^{k} \frac{|S_i|}{|S|} H(S_i)$$

donde:

- $|S_i|$: Número de observaciones i del atributo A.
- $H(S_i)$: Entropía del subconjunto S_i.

En cada paso de la construcción del árbol, el algoritmo selecciona como nodo el atributo con la mayor ganancia de información.

Random Forest



MissForest

MissForest es una técnica basada en **MICE** (Multivariate Imputation by Chained Equations).

- Imputa valores faltantes construyendo un modelo por columna con NaNs.
- Cada columna con valores faltantes se trata como una **variable dependiente** y se predice usando el resto.
- Se repite el proceso iterativamente para refinar las imputaciones.

Caso práctico