

INTRODUCCIÓN A *MACHINE LEARNING*

Minia Manteiga

Carlos Dafonte

Marco A. Álvarez

Guillermo Torralba

Raúl Santoveña

Lara Pallas

Daniel Garabato

ÍNDICE

- I. CONTEXTO
- II. APRENDIZAJE MÁQUINA
- III. APRENDIZAJE SUPERVISADO
- IV. APRENDIZAJE NO SUPERVISADO
- V. ESTIMACIÓN DE PARÁMETROS EN EL MÓDULO
GSP-SPEC

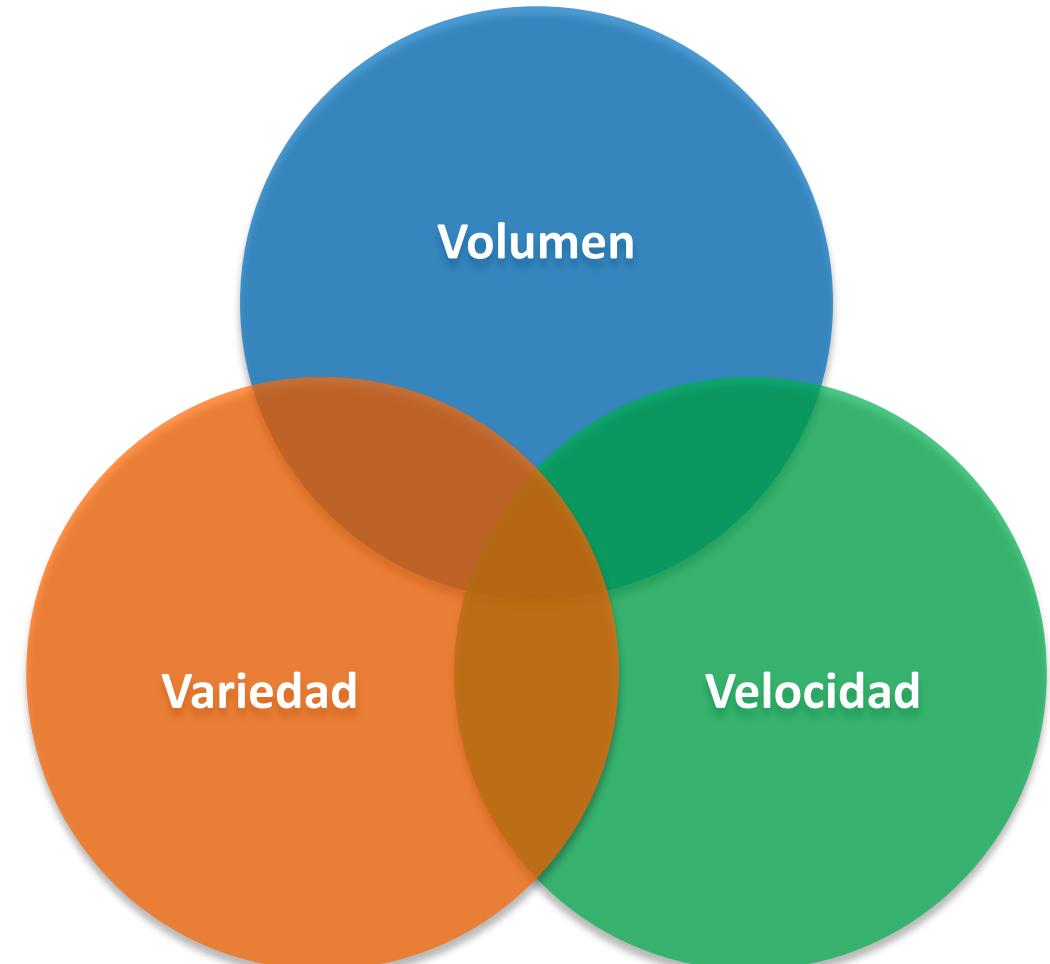
CONTEXTO

- I. *BIG DATA*
- II. MINERÍA DE DATOS (*DATA MINING*)
- III. INTELIGENCIA ARTIFICIAL

BIG DATA

El término **Big Data** hace referencia a la disciplina de las Tecnologías de Información y las Comunicaciones que se ocupa de las tareas asociadas a **volúmenes masivos de información** que no pueden ser tratados mediante técnicas convencionales:

- › Captura
- › Transferencia
- › Almacenamiento
- › Gestión, mantenimiento y consulta
- › **Análisis (extracción de conocimiento)**
- › Visualización

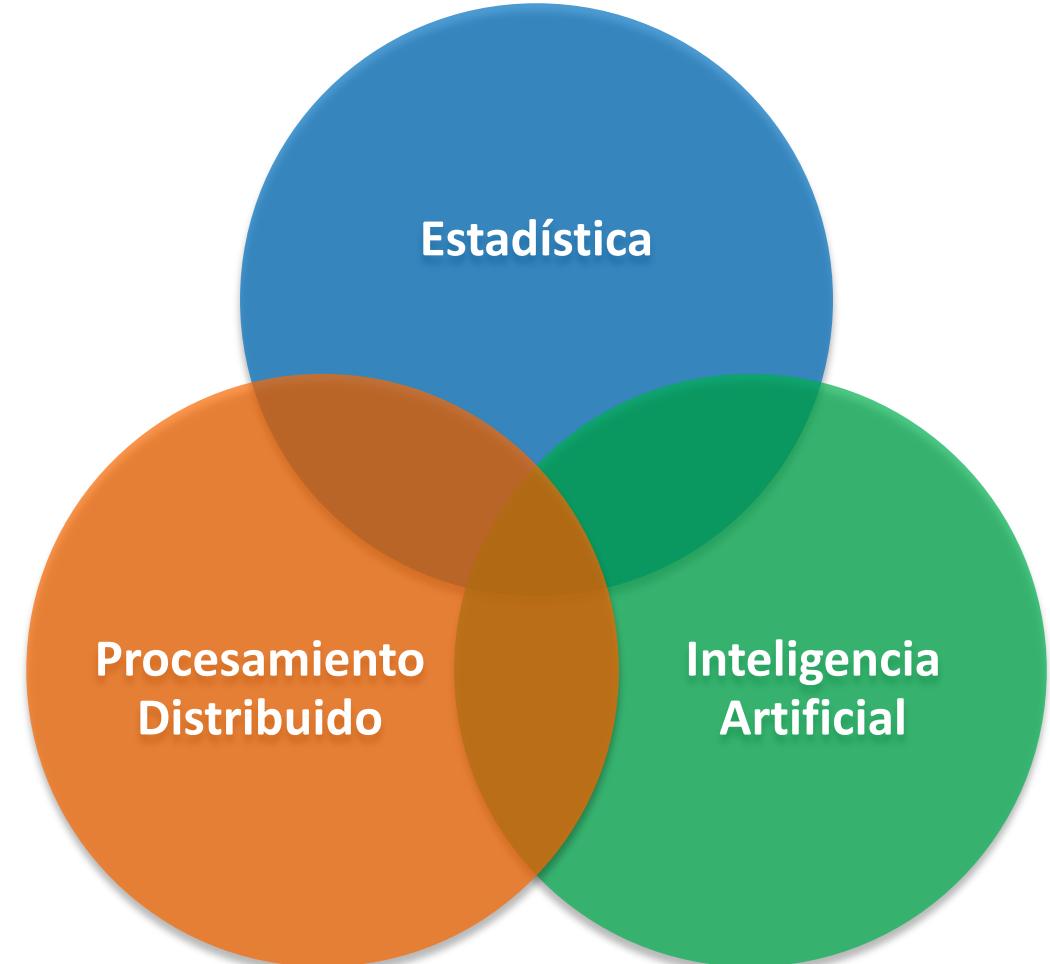


MINERÍA DE DATOS (*DATA MINING*)

Las tareas de análisis de datos en entornos *Big Data* se realiza mediante técnicas de **Minería de Datos (*Data Mining*)**, principalmente para la extracción de conocimiento o patrones ocultos entre los datos.

Para ello, se hace uso de **técnicas interdisciplinarias**, como la estadística, la Inteligencia Artificial, o el *Machine Learning*:

- › Redes Bayesianas
- › Máquinas de Soporte Vectorial (SVM)
- › Árboles de Decisión (DT)
- › Redes de Neuronas Artificiales (RNA)
- › *Deep Learning* (DL)



MINERÍA DE DATOS (*DATA MINING*)

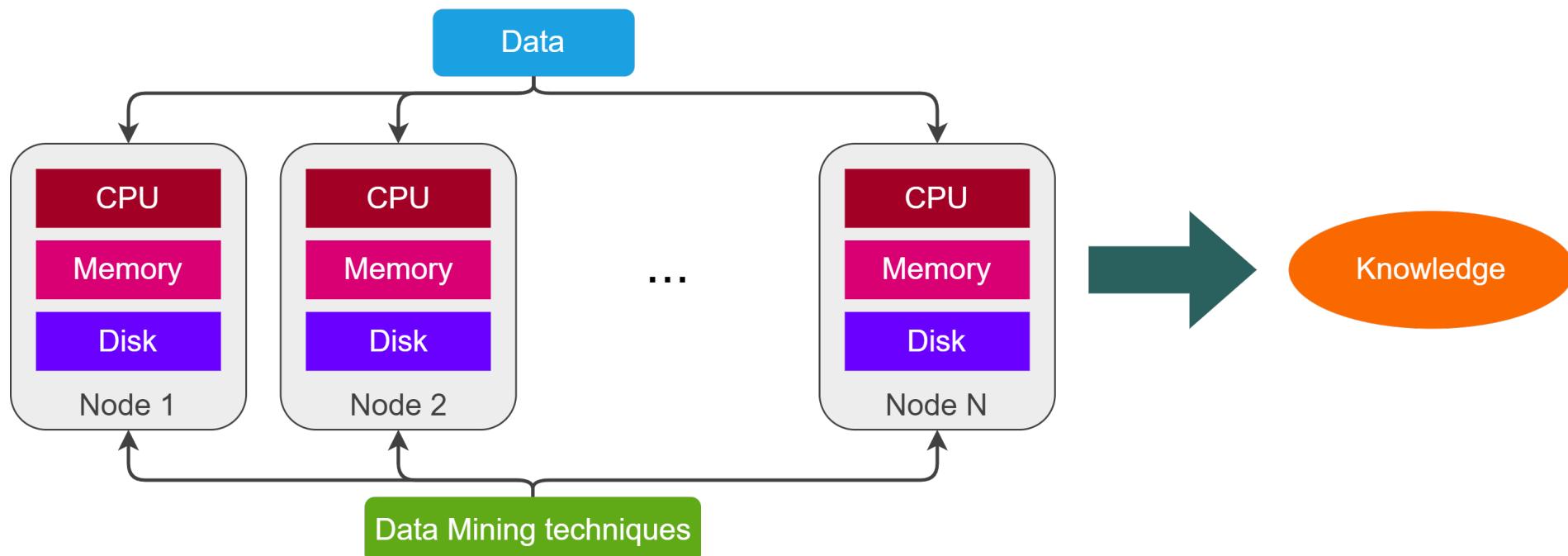
Para lograr procesar tales volúmenes de información, será necesario utilizar estrategias de almacenamiento y procesamiento distribuido, aplicando paradigmas como Map-Reduce o GP-GPU.

Se trata de un modelo de computación basado en el uso de **múltiples ordenadores interconectados** mediante redes de área local o extensa orientados al **almacenamiento y procesamiento de grandes volúmenes de información**.



MINERÍA DE DATOS (*DATA MINING*)

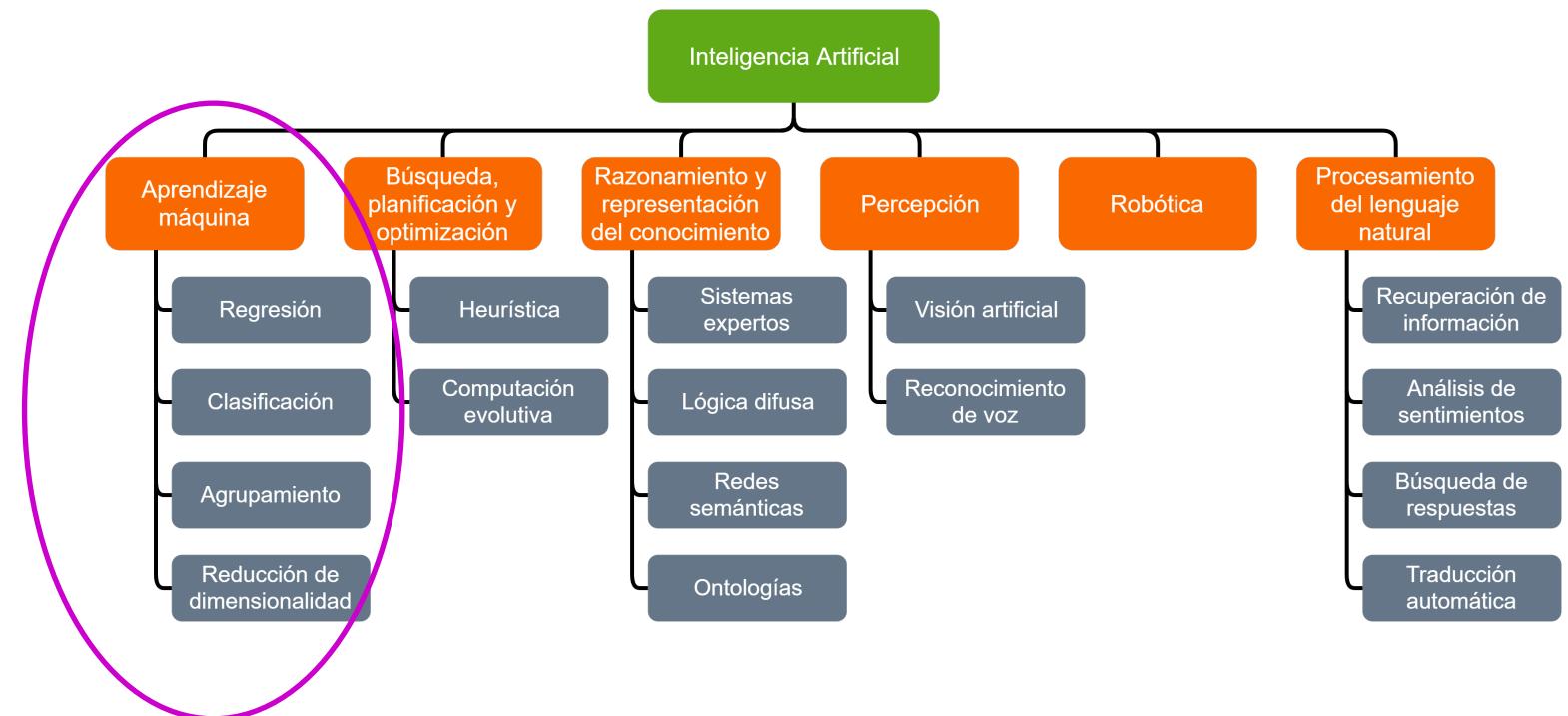
Para lograr procesar tales volúmenes de información, será necesario utilizar estrategias de almacenamiento y procesamiento distribuido, aplicando paradigmas como Map-Reduce o GP-GPU.



INTELIGENCIA ARTIFICIAL

Capacidad de un sistema o una aplicación para “simular” los procesos cognitivos propios de los seres humanos: aprendizaje, resolución de problemas, toma de decisiones, etc.

- › Su objetivo principal consiste en dotar a las máquinas de cierta capacidad para actuar de manera “inteligente”.
- › La IA no busca la solución perfecta para un problema, sino una solución “aceptable”.



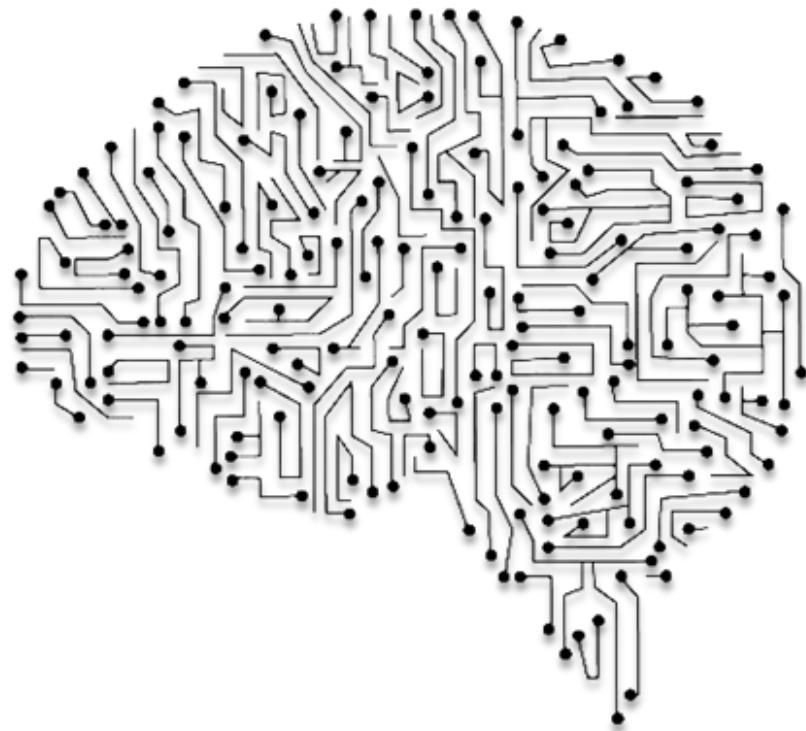
APRENDIZAJE MÁQUINA

- I. CONCEPTOS GENERALES
- II. TRATAMIENTO DE LOS DATOS
- III. FASE DE APRENDIZAJE O ENTRENAMIENTO
- IV. AJUSTE DE HIPERPARÁMETROS
- V. VALIDACIÓN DE LOS MODELOS
- VI. TIPOS DE APRENDIZAJE

CONCEPTOS GENERALES

El término *Machine Learning* se refiere a modelos capaces de realizar una tarea sin recibir instrucciones explícitas para resolverla.

- › Requieren de una fase de aprendizaje en base a ejemplos, mediante la búsqueda de patrones subyacentes en los datos.
- › Deben adquirir una cierta capacidad de generalización: generar una respuesta aceptable ante nuevas entradas (datos desconocidos).



TRATAMIENTO DE LOS DATOS

Antes de comenzar a procesar y analizar cualquier conjunto de datos mediante técnicas de *Machine Learning*, es necesario estudiar sus propiedades y características. En general, siempre será necesario llevar a cabo un preprocesado de los datos en bruto y que implicará tareas como:

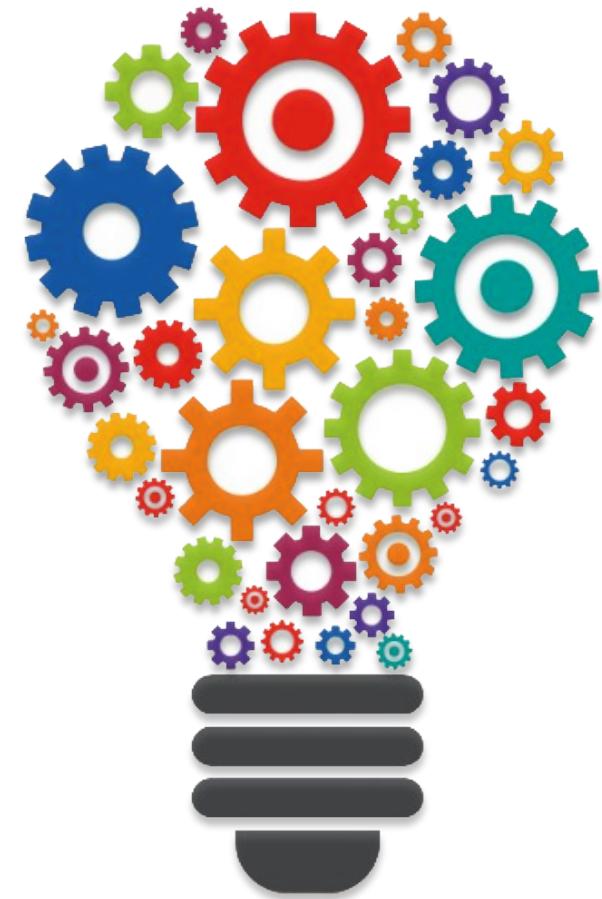
- › Eliminación de ruido y/o valores anómalos
- › Imputación de valores no disponibles (*missing values*)
 - › Media, mediana, *clustering*, ...
- › Extracción y selección de características
 - › Análisis de componentes principales (PCA), Eliminación recursiva de características (RFE), ...
- › Normalización/Estandarización/Escalado
 - › Típicamente se suelen presentar los datos en el rango [0,1] ó [-1, 1]

FASE DE APRENDIZAJE O ENTRENAMIENTO

Se trata de un **proceso iterativo** en el que se utiliza una muestra de los datos para ajustar el modelo progresivamente.

- › Condición de parada:
 - › Máximo de iteraciones permitidas
 - › El error de aprendizaje alcanza un cierto umbral
 - › La red no consigue mejorar el ajuste/aprendizaje
- › El proceso parte de una inicialización aleatoria del modelo

La muestra o conjunto de entrenamiento debe ser **significativo** y **representativo**



AJUSTE DE HIPERPARÁMETROS

En general, las técnicas de Machine Learning cuentan con una serie de **parámetros de configuración** o **hiperparámetros**, que se deberán establecer adecuadamente para cada problema a resolver.

- › Éstos variarán entre las diferentes técnicas:
 - › Máquinas de soporte vectorial: penalización (C), función de kernel, gamma, ...
 - › Redes de Neuronas Artificiales: número de capas, número de neuronas por capa, función de activación, ...
- › Procedimiento para ajustar su valor:
 - › Manualmente
 - › Reglas o guías de aproximación
 - › Regla de la Pirámide Geométrica (RNAs)
 - › **Métodos de optimización**
 - › Grid Search
 - › Random Search
 - › Algoritmos evolutivos

VALIDACIÓN DE LOS MODELOS

En general, siempre será necesario llevar a cabo un **proceso de validación** sobre los modelos ya entrenados, para poder evaluar si estos son capaces de proporcionar una respuesta adecuada, especialmente ante entradas desconocidas.

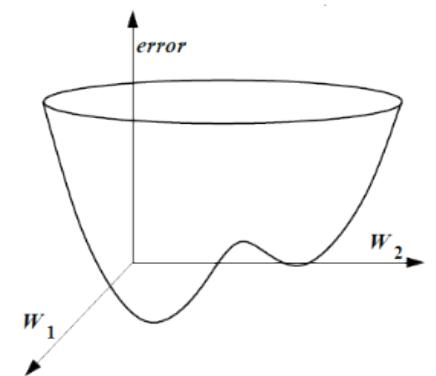
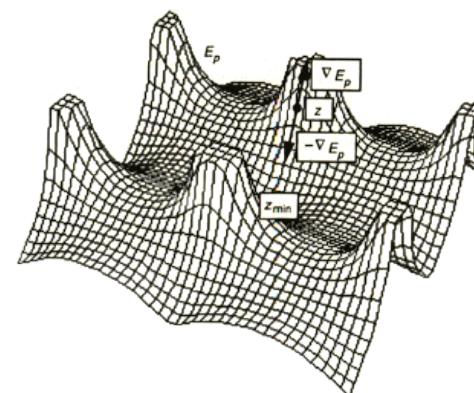
De esta manera, podremos detectar determinados **problemas derivados de un mal proceso de entrenamiento:**

› Sobreentrenamiento

- › El modelo se especializa y pierde la capacidad de generalización y, por tanto, no será capaz de proporcionar una respuesta adecuada ante datos desconocidos.

› Mínimos locales / Saturación

- › El modelo se estanca porque no es capaz de reducir el error de aprendizaje “todo lo posible”.



TIPOS DE APRENDIZAJE

› Supervisado

- › Se dispone de una muestra etiquetada de los datos, donde para cada entrada conocemos previamente su valor esperado de salida
- › Los modelos tratarán de aproximar la función que permite establecer una relación entre las muestras y su salida ideal

› No supervisado

- › No disponemos de una muestra de datos etiquetada
- › Los modelos buscarán patrones y relaciones internas entre los datos



TIPOS DE APRENDIZAJE

› Semi-supervisado

- › “Combinación” de los anteriores, donde solo conocemos la salida deseada para una parte de la muestra.

› Por refuerzo

- › Se basan en una metodología de ensayo y error, premiando o reforzando las respuestas acertadas y penalizando las incorrectas.



APRENDIZAJE SUPERVISADO

- I. VALIDACIÓN DE MODELOS SUPERVISADOS
- II. MÁQUINAS DE SOPORTE VECTORIAL (SVM)
- III. REDES DE NEURONAS ARTIFICIALES (RNA)

VALIDACIÓN DE MODELOS SUPERVISADOS

› División del conjunto de entrenamiento

- › Entrenamiento → ~60%
- › Validación → ~15%
- › Test → ~25%

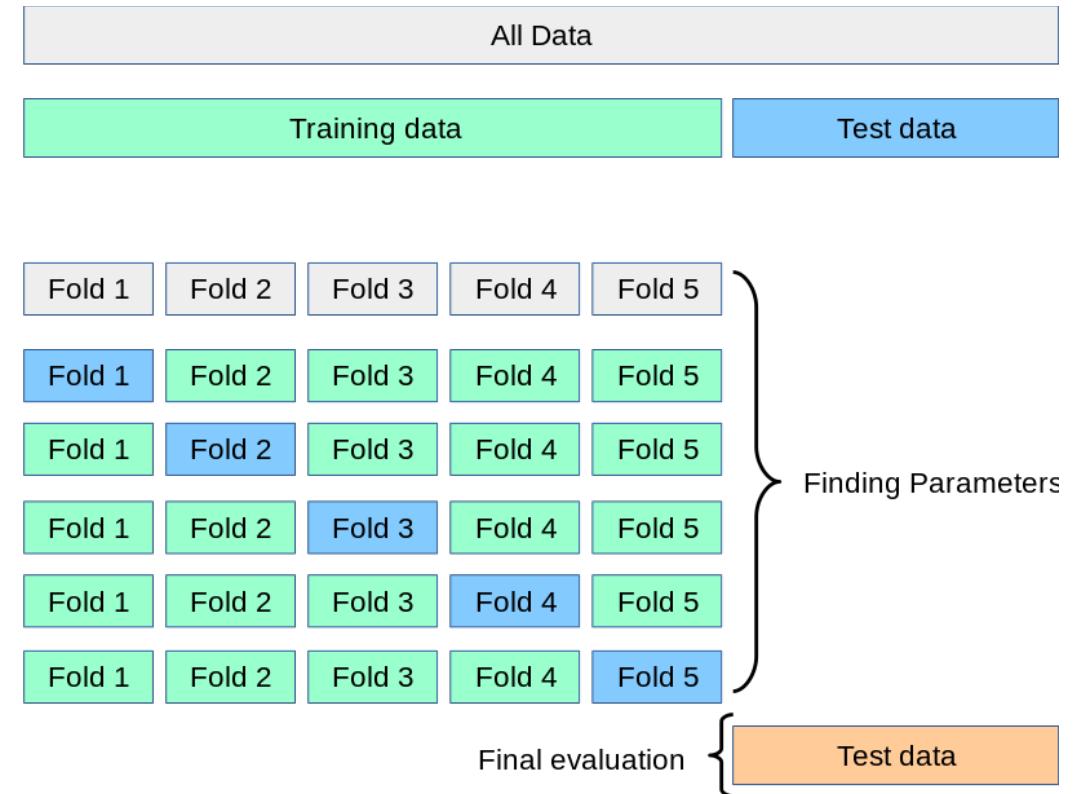
› Métricas de rendimiento

- › Problemas de clasificación
 - › Accuracy, precision, recall, $F_1 - score$, ...
- › Problemas de regresión
 - › Error cuadrático medio (MSE, RMSE), error absoluto medio (MAE), R^2

VALIDACIÓN DE MODELOS SUPERVISADOS

› Validación cruzada (Cross-validation)

- › Generalmente, se sigue un procedimiento *K-Fold*, donde se divide el conjunto de entrenamiento en *K* partes.
- › El proceso de entrenamiento se repite tantas veces como *K-Folds* tengamos, reservando en cada pasada una de las partes para validación.
- › Tras cada pasada del entrenamiento, se evalúa el modelo contra el conjunto de validación.
- › Tras concluir el proceso, se selecciona el mejor modelo y se evalúa contra el conjunto de *test*, reservado previamente y que no ha formado parte del conjunto de entrenamiento en ninguna pasada.
- › En general, todo el proceso de entrenamiento se suele repetir *N* veces, para seleccionar el mejor modelo



Fuente: <https://scikit-learn.org>

MÁQUINAS DE SOPORTE VECTORIAL (SVM)

- › Método desarrollado por Vapnik en los años 90
- › Aplicables tanto a problemas de regresión como de clasificación
- › Si el problema no es linealmente resoluble, mediante una **función kernel** se pueden proyectar los datos a una dimensión superior en la cual el problema se convierte en linealmente resoluble.

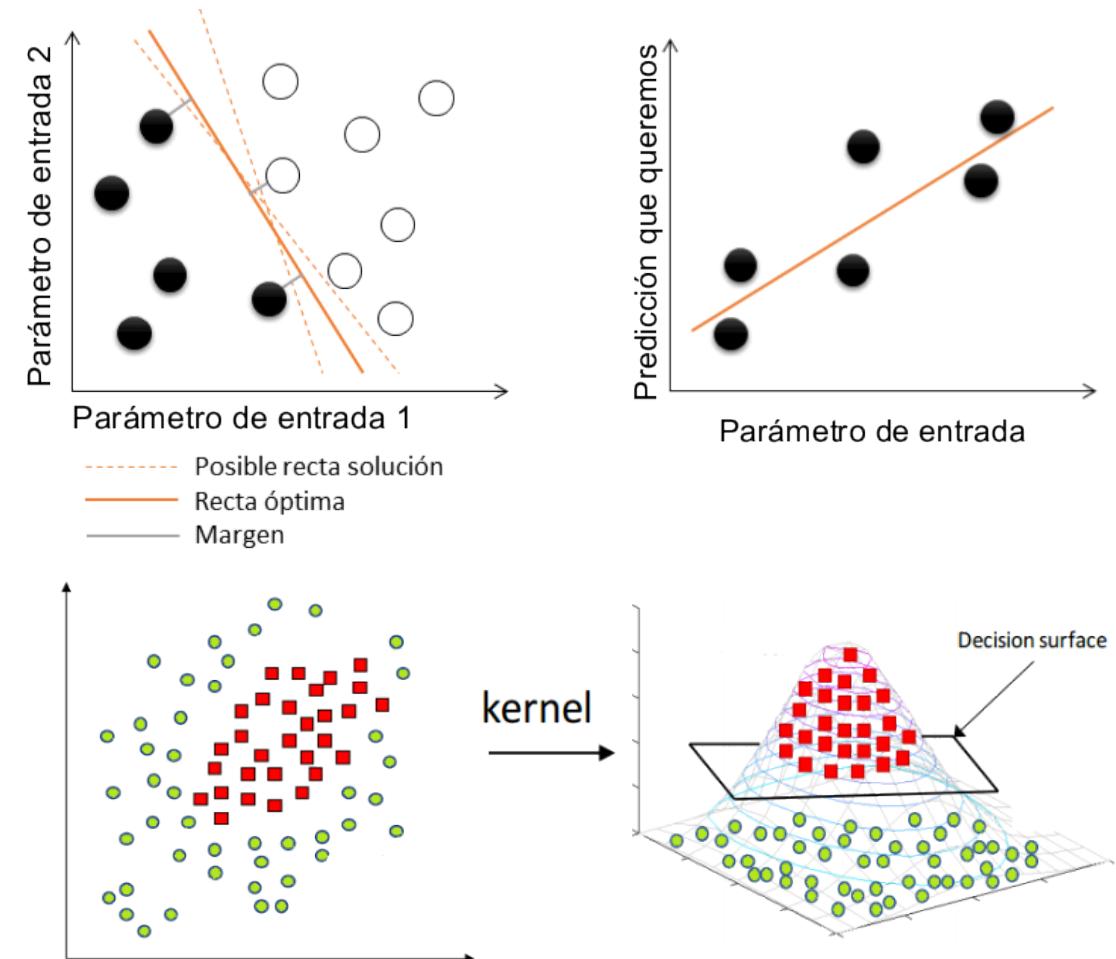
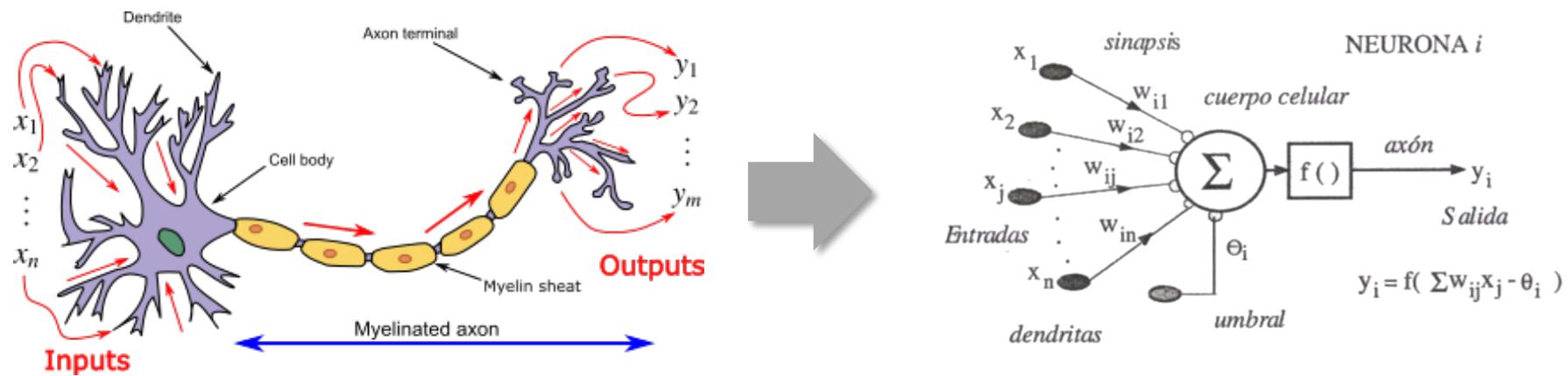


Imagen: S. Van Vaerenbergh, I. Santamaría

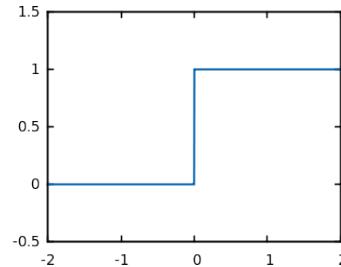
REDES DE NEURONAS ARTIFICIALES (RNA)

- › Planteadas por McCulloch y Pitts en 1943
- › Están inspiradas en las redes neuronales biológicas
- › Existen multitud de modelos y de planteamientos diferentes

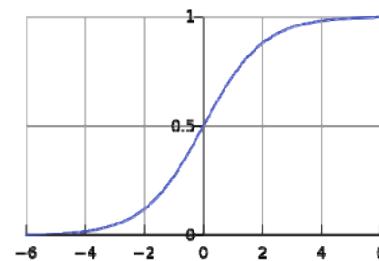


REDES DE NEURONAS ARTIFICIALES (RNA)

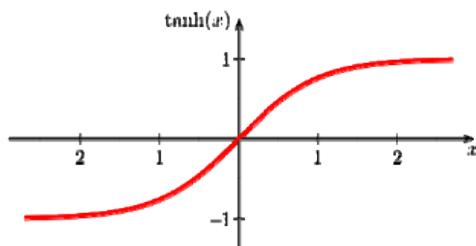
› Funciones de activación típicas:



Heaviside step function: $f(x) = \begin{cases} 0, & n < 0 \\ 1, & n \geq 0 \end{cases}$



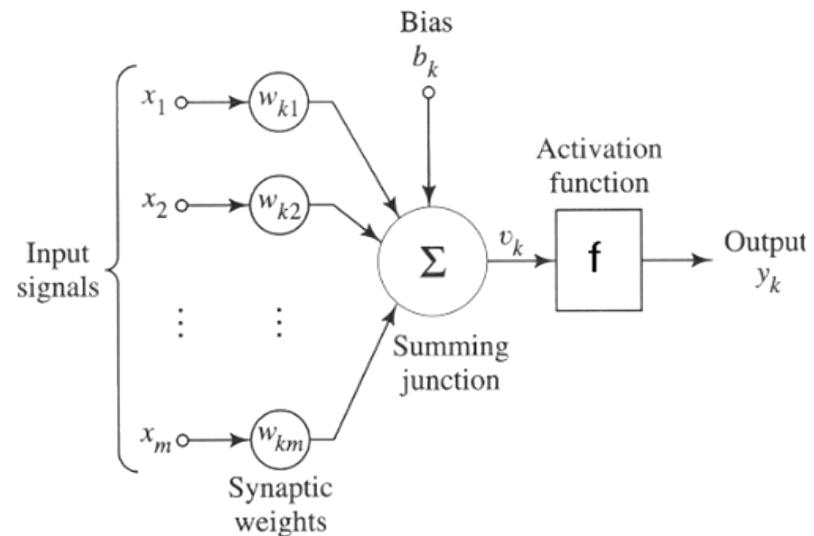
Logistic function: $f(x) = \frac{1}{1+e^{-x}}$



Hyperbolic tangent function: $f(x) = \frac{e^x - e^{-x}}{e^x + e^{-x}}$

PERCEPTRÓN

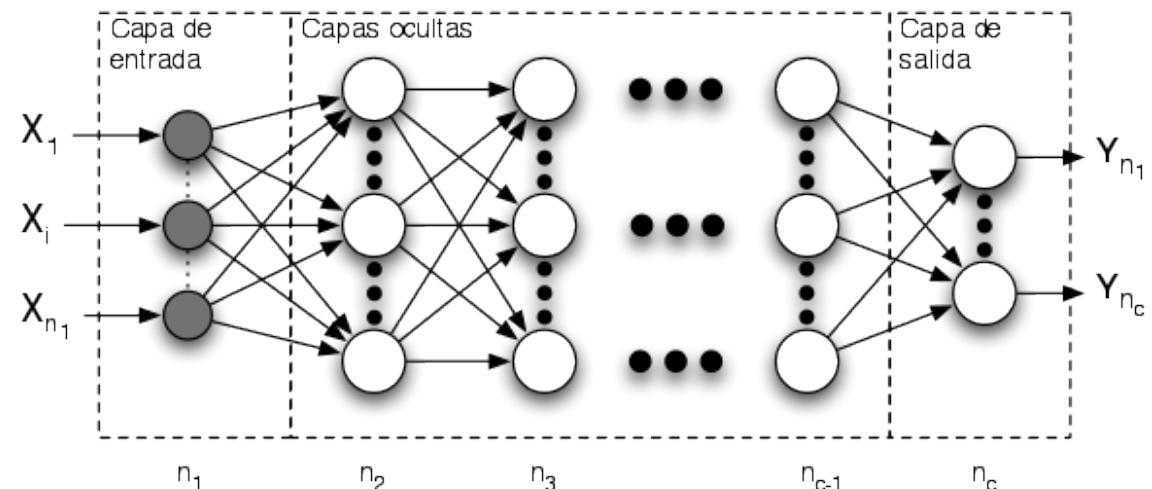
- › Planteado por Frank Rosenblatt en 1958
- › Primer modelo de Red de Neuronas Artificiales
- › Únicamente resuelve problemas de clasificación lineal (aprendizaje supervisado)
- › Características
 - › Monocapa
 - › Entrada continua
 - › Salida binaria
 - › Función de activación: escalón



$$y_k = f \left(\sum_{j=1}^m (w_{kj} x_j) + b_k \right)$$

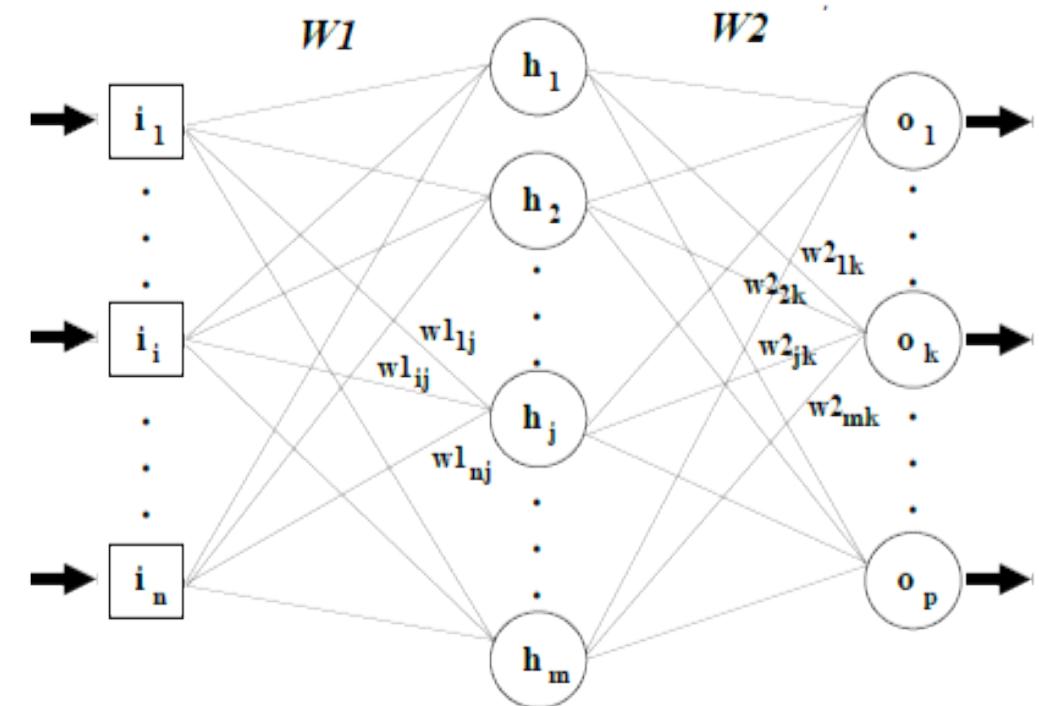
PERCEPTRÓN MULTICAPA (MLP)

- › Planteadas por Paul Werbos en 1974
- › Conexiones dirigidas hacia adelante, también denominadas ***feedforward networks***.
- › Aplicable a problemas de clasificación y regresión
- › Múltiples capas y múltiples neuronas



PERCEPTRÓN MULTICAPA (MLP)

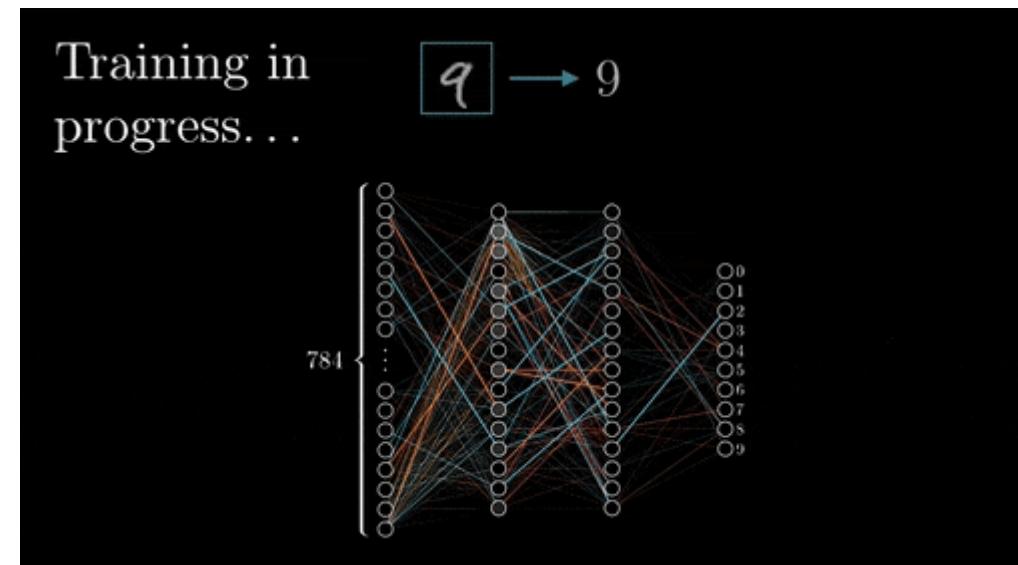
- › Planteadas por Paul Werbos en 1974
- › Conexiones dirigidas hacia adelante, también denominadas ***feedforward networks***.
- › Aplicable a problemas de clasificación y regresión
- › Múltiples capas y múltiples neuronas
 - › Típicamente se suele utilizar una arquitectura de 3 capas con N elementos en la capa oculta



PERCEPTRÓN MULTICAPA (MLP)

Utiliza el algoritmo de aprendizaje basado en la retropropagación del error

- › Minimización del error cuadrático medio (MSE) en la salida
- › Se utiliza el método del gradiente descendiente
- › El error se propaga desde la capa de salida hacia la capa de entrada, ajustando los pesos W en cada iteración
- › Se introduce un hiperparámetro α que regula la **velocidad del aprendizaje**



APRENDIZAJE NO SUPERVISADO

- I. PARTICULARIDADES DEL APRENDIZAJE NO SUPERVISADO
- II. MODELOS MÁS HABITUALES

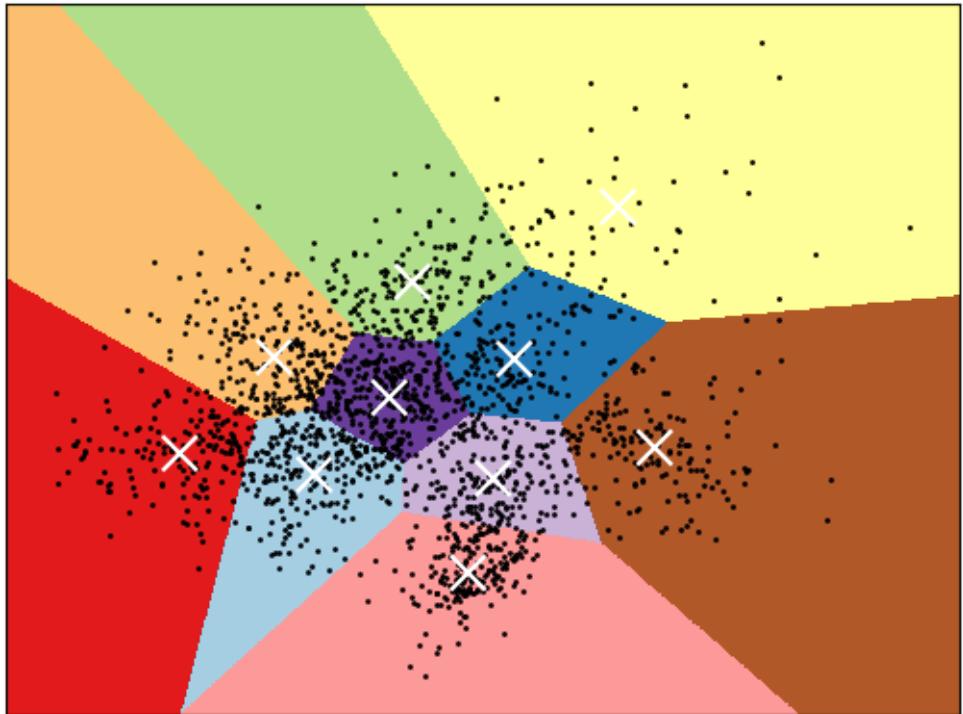
PARTICULARIDADES DEL APRENDIZAJE NO SUPERVISADO

- › Únicamente disponemos de información limitada sobre los datos
- › Evaluación de modelos
 - › Existen diferentes métricas basadas en información interna del modelo, aunque son muy costosas de calcular
 - › Se puede utilizar una aproximación híbrida, recuperando información acerca de una pequeña muestra para poder llevar a cabo una evaluación mediante métricas supervisadas
- › Configuración de hiperparámetros
 - › Ensayo y error
 - › Reglas de aproximación
 - › Métodos de optimización

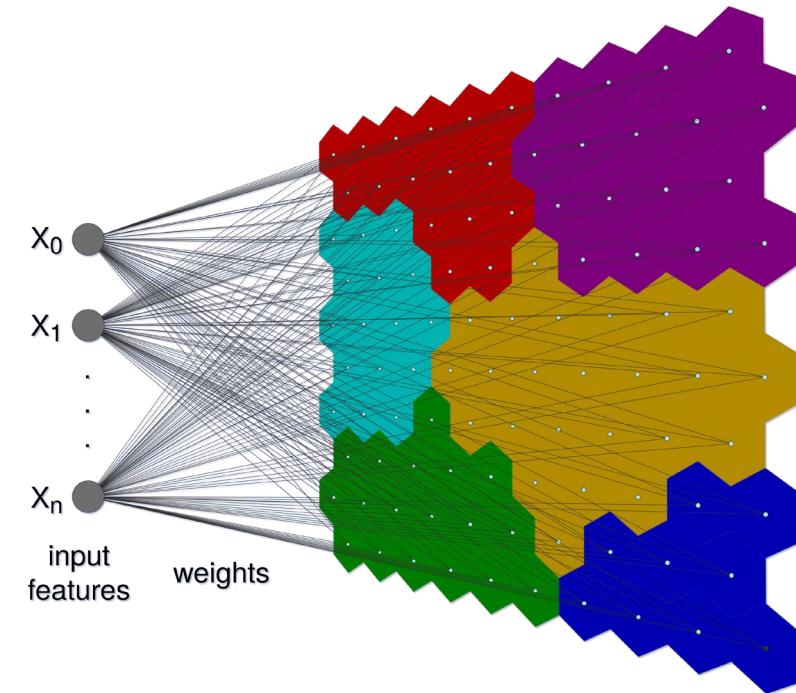


[Esta foto](#) de Autor desconocido está bajo licencia [CC BY](#)

MODELOS MÁS HABITUALES



K-Means



Mapas Auto-Organizativos (SOM)

ESTIMACIÓN DE PARÁMETROS EN EL MÓDULO GSP-SPEC

- I. DEFINICIÓN DEL PROBLEMA
- II. DATOS DE ENTRADA
- III. TÉCNICAS EMPLEADAS
- IV. PREPROCESADO DE LOS DATOS
- V. ANÁLISIS DE LOS RESULTADOS

DEFINICIÓN DEL PROBLEMA

El paquete GSP-Spec de CU8 pretende la estimación de parámetros astrofísicos para las estrellas más brillantes ($G_{RVS} \leq 14$) mediante el uso de espectroscopía RVS:

- › Temperatura efectiva (T_{eff})
- › Gravedad superficial logarítmica ($\log g$)
- › Metalicidad ($[Fe/H]$)
- › Abundancia de elementos α ($[\alpha/Fe]$)

DATOS DE ENTRADA

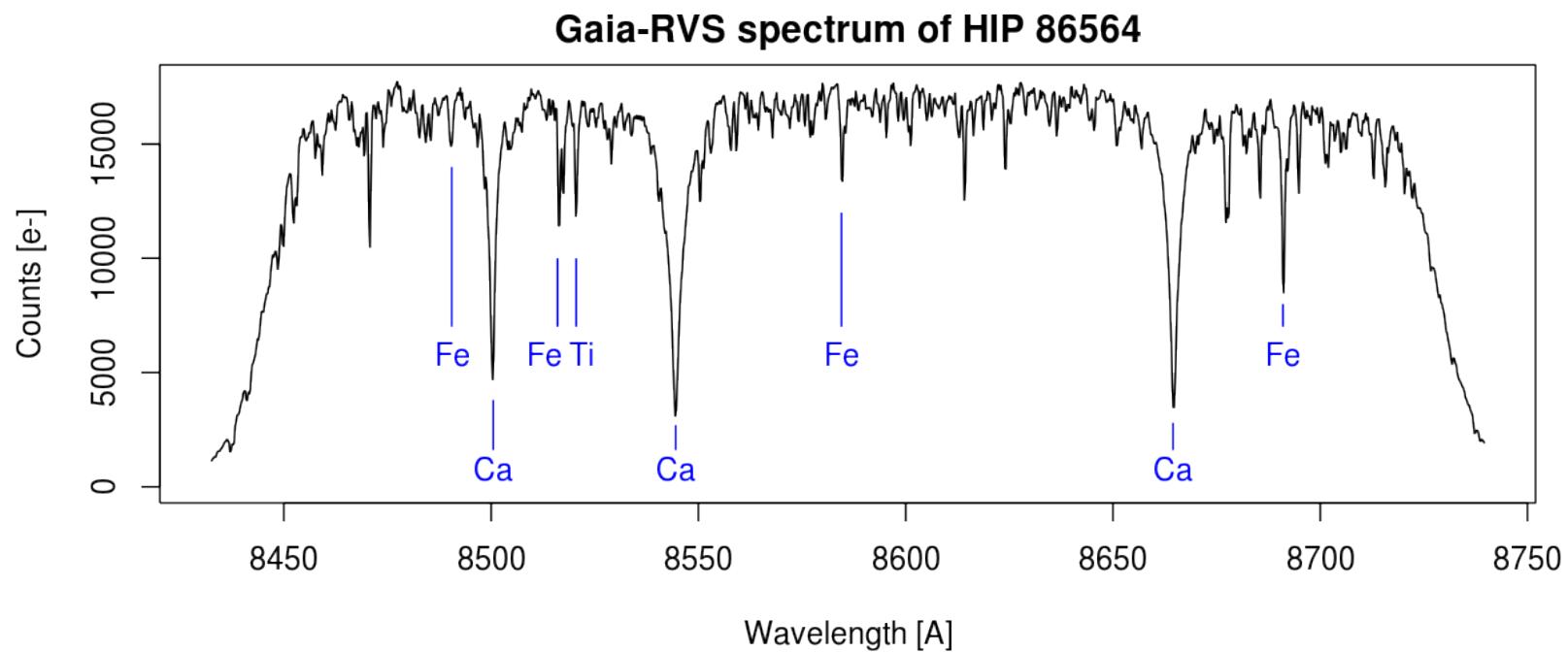
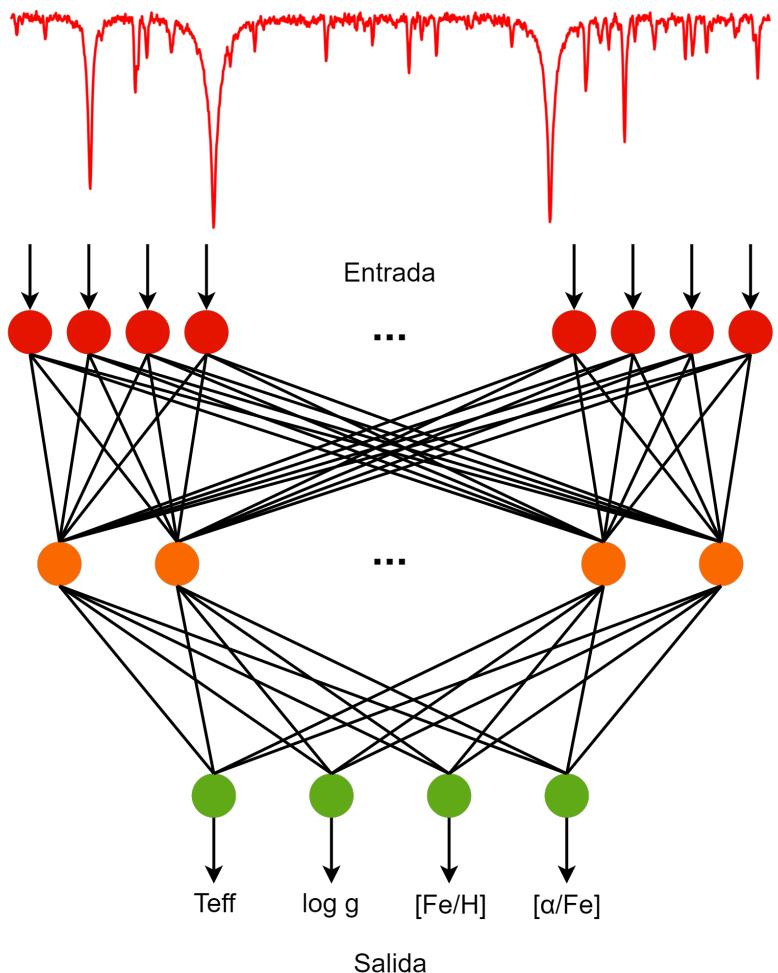


Imagen: ESA/Gaia/DPAC/Airbus DS

TÉCNICAS EMPLEADAS

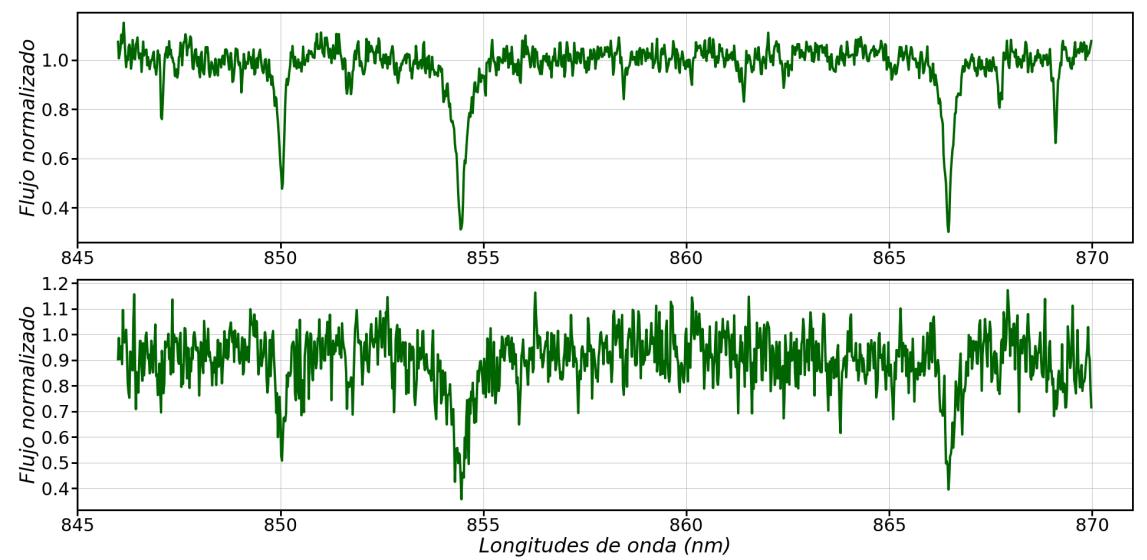
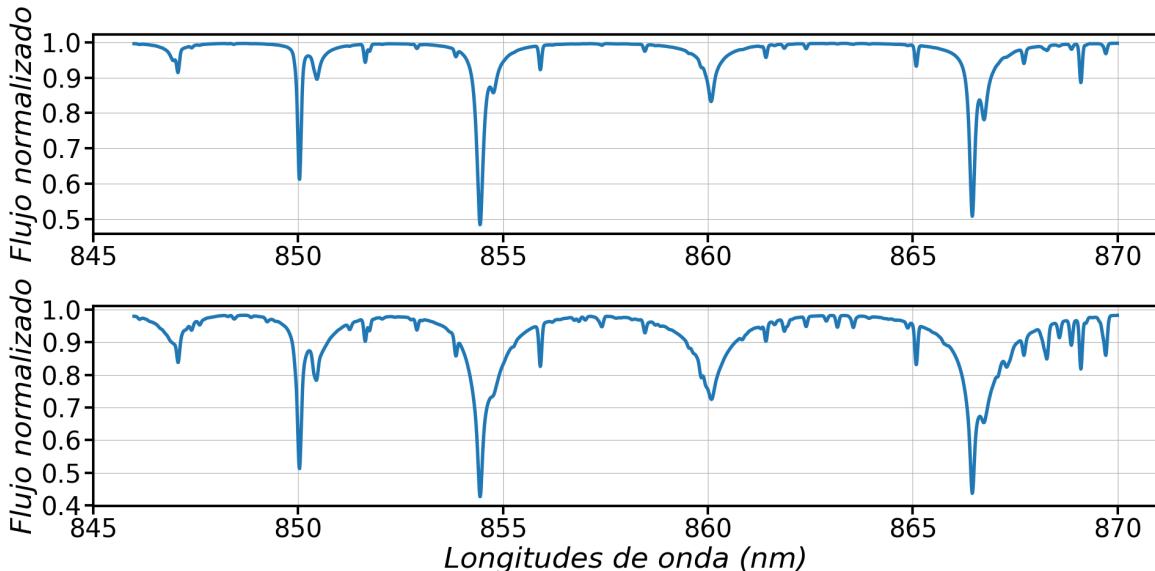
- › MATISSE
 - › Algoritmo basado en álgebra lineal
- › GAUGUIN
 - › Minimiza la distancia entre el espectro a parametrizar y la malla de referencia
- › ANN
 - › Técnica de IA basada en el reconocimiento de patrones para realizar la estimación no lineal de los parámetros



PREPROCESADO DE LOS DATOS

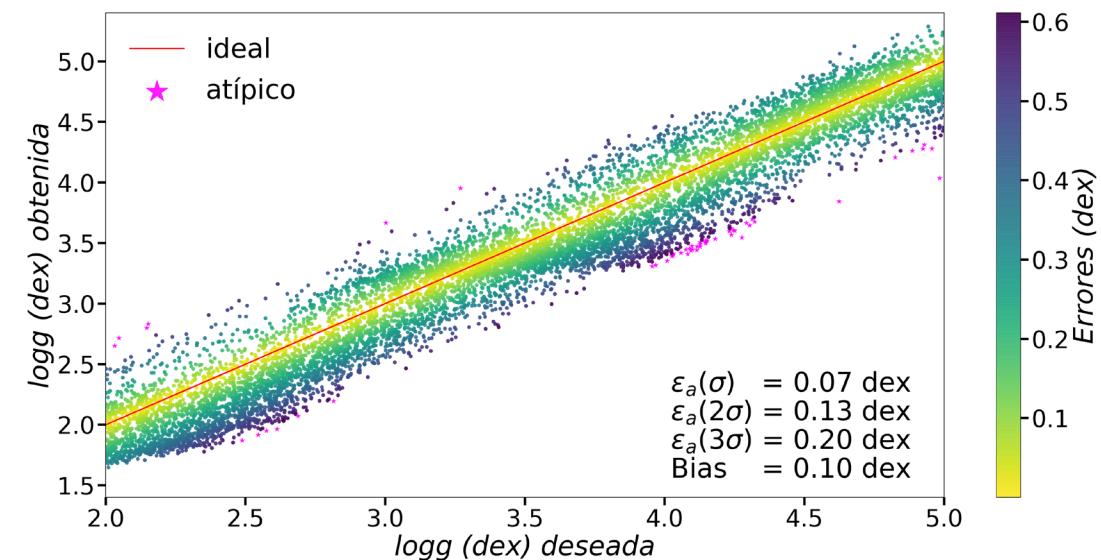
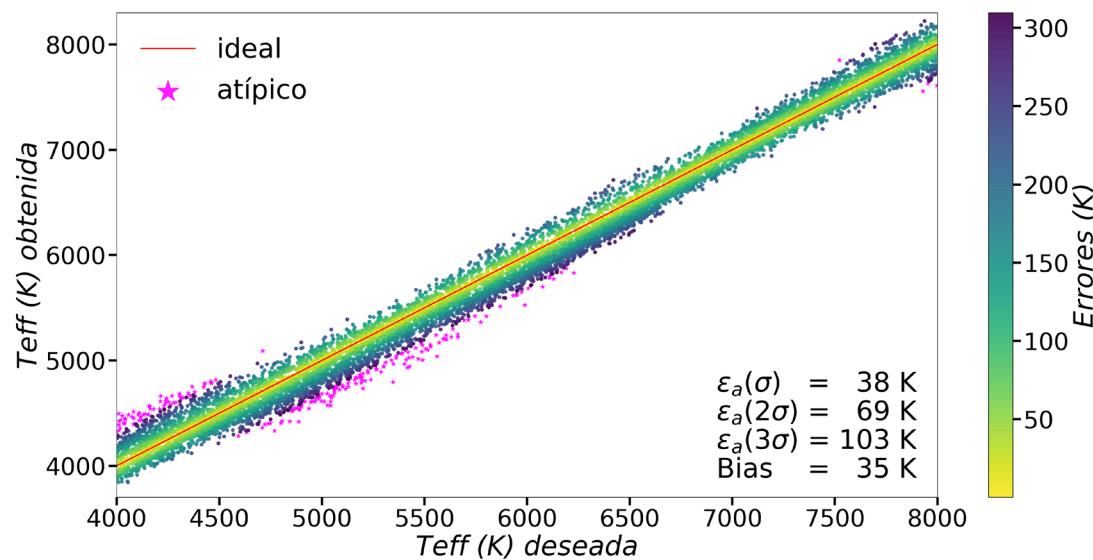
Para entrenamiento se utiliza un conjunto sintético de espectros simulados mediante modelos de atmósferas (Kurucz, MARCS, etc.), a los que se añade ruido, mientras que para validar se utiliza un conjunto de espectros observacionales.

Los espectros se normalizan para evitar sesgos geométricos y que todas las dimensiones del espectro estén en un rango acotado y bien definido para que tengan la misma importancia durante el entrenamiento.



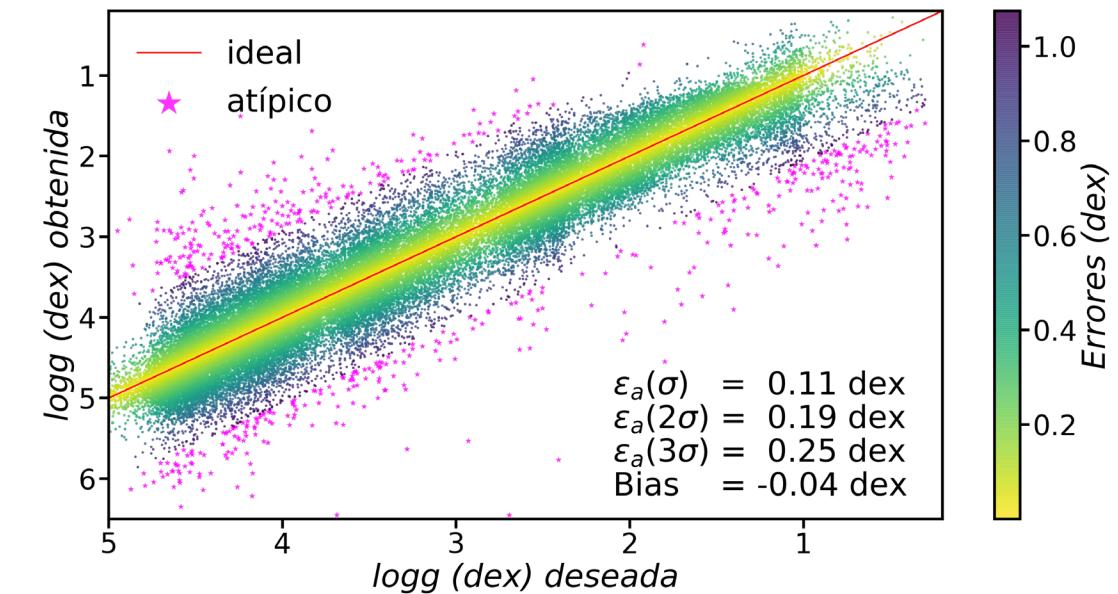
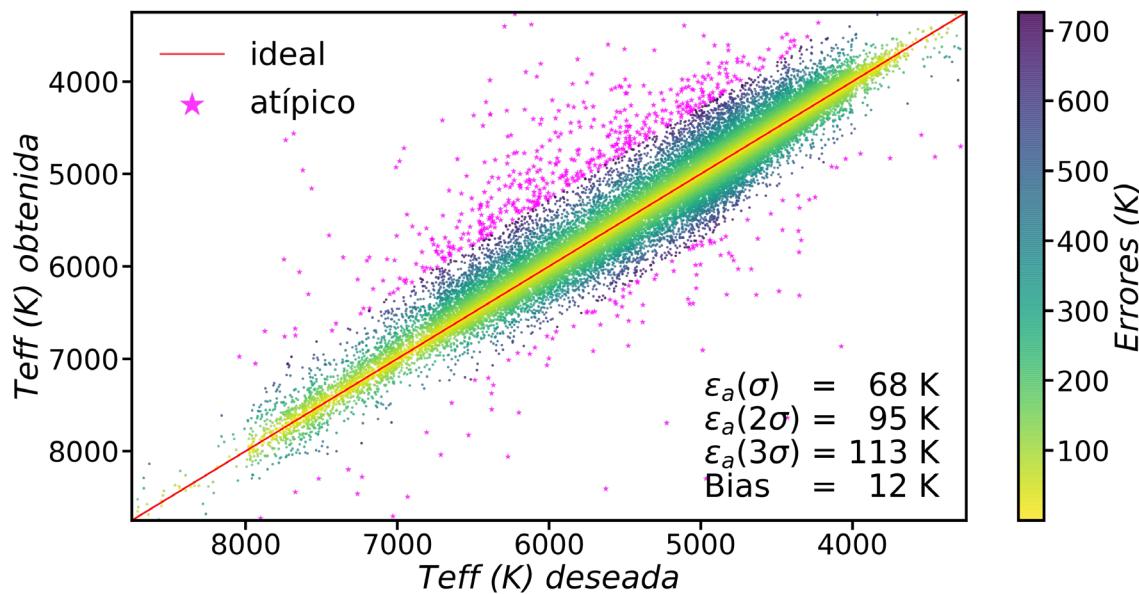
ANÁLISIS DE RESULTADOS

Errores internos: se obtienen evaluando la red con una muestra con las mismas características del conjunto de entrenamiento.



ANÁLISIS DE RESULTADOS

Errores externos: se obtienen evaluando los resultados de la red con un nuevo conjunto de datos.



ANÁLISIS DE COHERENCIA

Se realizan diagramas de diagnóstico como el diagrama H-R para comprobar si las estimaciones de parámetros son consistentes y representan adecuadamente las características esperadas en las diferentes poblaciones estelares de nuestra galaxia.

