UNIVERSIDAD
DE BURGOS

JUAN JOSÉ RODRÍGUEZ DIEZ

Catedrático de Universidad del Área de Lenguajes y Sistemas Informáticos

APRENDIZAJE AUTOMÁTICO EN CIENCIA DE DATOS

LECCIÓN INAUGURAL DEL CURSO ACADÉMICO 2023-2024

BURGOS

2023

UNIVERSIDAD
DE BURGOS

JUAN JOSÉ RODRÍGUEZ DIEZ

Catedrático de Universidad del Área de Lenguajes y Sistemas Informáticos

APRENDIZAJE AUTOMÁTICO EN CIENCIA DE DATOS

LECCIÓN INAUGURAL DEL CURSO ACADÉMICO 2023-2024

BURGOS

2023

Edita: UNIVERSIDAD DE BURGOS. SECRETARÍA GENERAL

SERVICIO DE PUBLICACIONES E IMAGEN INSTITUCIONAL
Edificio de Administración y Servicios
C/ Don Juan de Austria, 1
09001 Burgos - España

Depósito legal: BU 221-2023
Imprime: Rico Adrados S.L.

Índice

Lección

1. INTRODUCCIÓN

Usamos el término **aprendizaje automático** (*machine learning*) [Zhou, 2021] para referirnos a que las máquinas, ordenadores u otros dispositivos, aprendan o al menos pueda parecer que lo hacen[1] . También se usan los términos equivalentes **aprendizaje máquina**, **aprendizaje computacional** o **aprendizaje algorítmico**.[2]

Hay distintos tipos de aprendizaje, también en el de personas. Por ejemplo, podemos aprender de memoria. Pero almacenar algo en una base de datos de un ordenador no se suele considerar aprendizaje. Las personas también aprendemos algoritmos para realizar algunas operaciones matemáticas como la multiplicación, pero que los ordenadores ejecuten algoritmos programados para realizar esos cálculos de fórmulas tampoco se considera aprendizaje.

Un planteamiento operativo para definir aprendizaje automático[3] es que en una máquina hay que hacer una tarea o resolver un problema. Si la propia má-

[1] Por tanto, no nos estamos refiriendo al aprendizaje humano apoyado en tecnologías. Aunque dentro de las tecnologías para el aprendizaje humano también entra el automático [Romero and Ventura, 2020].

[2] Aunque pueden tener matices distintos.

[3] Para esquivar cuestiones filosóficas, psicológicas, pedagógicas...

quina[4] mejora en esa tarea o problema, se puede considerar que hay aprendizaje automático. A menudo las tareas se pueden resolver mejor o peor. En esos casos el aprendizaje puede ser gradual si vamos mejorando. No solo es una cuestión de obtener una solución correcta, o al menos razonable, también de cuánto tiempo o recursos necesitamos. El conseguir hacer algo de manera más eficiente también es aprendizaje humano.

Al realizar tareas o resolver problemas con ordenadores usamos programas.[5] Estos programas los hacen personas con las herramientas adecuadas.[6] El hacer un programa que realice una tarea o que resuelva un problema también se puede considerar otra tarea o problema. En principio, no tiene que haber aprendizaje automático aunque el programa parezca inteligente.

El aprendizaje automático es una parte de la Inteligencia Artificial (IA). En los seres vivos también el aprendizaje es una parte de la inteligencia. Hay áreas y sistemas en IA en los que no hay necesariamente aprendizaje: la representación del conocimiento, el razonamiento, los sistemas expertos o basados en conocimiento, la diagnosis (en seres vivos o dispositivos), la planificación, la búsqueda...

Por ejemplo, en IA, la búsqueda se entiende como el proceso de encontrar una secuencia de pasos que resuelve un problema. Para resolver sudokus podemos hacer un programa que considere todas las posibilidades, descartando las que no respetan las restricciones. Un programa que juegue muy bien al ajedrez no tiene por qué aprender o haberse construido con aprendizaje automático. Puede ser suficiente con explorar posibles jugadas o movimientos o los más prometedores, con tantas jugadas como el tiempo permita, teniendo en cuenta las reglas del juego y valores de las piezas y sus posiciones.

Aunque el aprendizaje automático no es el único campo de Inteligencia Artificial, sí que es uno de los más relevantes y actuales. Hay distintos tipos de aprendizaje automático. Nos vamos a centrar en el caso, muy habitual, en el que se dispone de unos datos sobre los que se quiere obtener algún conocimiento. Hay otros tipos: en el aprendizaje por refuerzo (*reinforcement learning*) [Sutton and Barto, 2018], el programa tiene que elegir entre distintas acciones que

[4] Al menos parcialmente, también puede haber intervención humana [Wu et al., 2022].

[5] Aunque los ordenadores y los programas también crean problemas...

[6] Que también son programas.

tienen una recompensa o castigo y el objetivo es aprender qué acciones tomar en cada situación. Por ejemplo, en un juego puede aprender jugando contra otros adversarios o contra sí mismo. En estos procesos se recibirán y almacenarán datos, pero estos no estaban disponibles inicialmente.

Si revisamos los antecedentes o inicios de la Informática, está por un lado la capacidad de realizar automáticamente cálculos complejos. Pero otra parte intrínseca es el almacenamiento y tratamiento de datos. De hecho, el término Informática se refiere al tratamiento automático de la información. Y la información es o está en los datos. Un tipo de tratamiento automático es el aprendizaje.

Así que queremos aprender automáticamente a partir de datos. A los datos se asocian distintas profesiones o actividades. Por supuesto, la estadística. Pero hay actividades a las que, por analogía, se plantean las correspondientes para datos: [7] analista de datos, ingeniería de datos, *fontanería* de datos, limpieza de datos, minería de datos, periodismo de datos, ciencia de datos... Entre estos términos hay solapes y el usar un término u otro puede tener su parte de moda (*buzzwords*) o marketing. [8]

Con ingeniería de datos nos referimos al trabajo que consiste en conseguir los datos y tenerlos almacenados de manera adecuada y que permita su uso y actualización eficientes. Esto es particularmente relevante cuando tenemos cantidades enormes de datos (*big data*) [Günther et al., 2017]. La limpieza de datos [Ilyas and Chu, 2019] aparece porque los datos pueden tener errores o inconsistencias y hay métodos para intentar detectarlos, corregirlos o paliar su influencia.

El término **minería de datos** [Gupta and Chandra, 2020] se utiliza en el sentido de que en los datos puede haber algo valioso, que sería conocimiento, pero para conseguirlo hay que extraerlo a partir de los datos. La extracción puede ser costosa, complicada y requiere métodos y herramientas. No está garantizado que encontremos algo de valor, puede que no lo haya o que no estemos usando las herramientas adecuadas.

Un término más actual es el de **ciencia de datos** [Özsu, 2023], que puede englobar los anteriores. Tiene un planteamiento interdisciplinar, se localiza la intersección de tres áreas (figura 1): por un lado matemáticas y estadística,

[7] Para datos personales hay responsables y encargados de su tratamiento.

[8] O incluso usarse despectivamente.

por otro informática, y la tercera es variable ya que es la propia de los datos concretos. Por ejemplo, los datos pueden ser de cualquier rama de la medicina o de cualquier industria.

En cualquier ciencia se usan y se trabaja con datos, no se trata de eso. En ciencia de datos el campo de estudio son los propios datos y los métodos semiautomáticos para trabajar con ellos, sin limitarse al aprendizaje. Aquí la palabra ciencia no se usa solo, ni principalmente, en el ámbito académico. La ciencia de datos aparece entre los perfiles demandados por empresas, la mayoría de las cuales no tienen objetivos o intereses científicos. El conseguir que un cliente o usuario[9] lo siga siendo o gaste más (dinero o tiempo) no parece un problema científico,[10] pero también son tareas de ciencia de datos. Se llegó a decir que la ciencia de datos era la profesión más *atractiva* del siglo XXI [Patil and Davenport, 2012]. [11]

Los datos se asimilan al petróleo. En el sentido de que son un recurso valioso pero también de que necesitan un procesado para que puedan ser usados. También se asimilan los datos con la electricidad, es algo que damos por hecho, pero que resulta imprescindible hoy en día.[12]

2. CLASIFICACIÓN

Una de las tareas más habituales que se puede abordar con aprendizaje automático es la clasificación. Se dispone de ejemplos (también denominados objetos, instancias...) de distintas clases, categorías o tipos. Por ejemplo, podemos tener correos que son o no *spam*, imágenes de caracteres manuscritos que se corresponden con distintos símbolos, fotos de caras de distintas personas o distintos tipos de insectos (como las que se muestran en las figuras 2 y 3), sonidos de distintos animales, vídeos de personas haciendo gestos en lenguaje de signos, síntomas y resultados de pruebas que se corresponden con enfermedades...

Un caso concreto, solo a efectos ilustrativos, es el de determinar si un cristal es o no de una ventana, en función del índice de refracción y valores de de

[9] «Si no eres el cliente, eres el producto» [Papadopoulos et al., 2017].

[10] Aunque puede ser objeto de estudios científicos en economía, psicología...

[11] Pero hace más de una década. Ahora podría ser la **ingeniería de consultas** (*prompt engineering*) a sistemas de IA generativa.

[12] Y que tiene un coste.

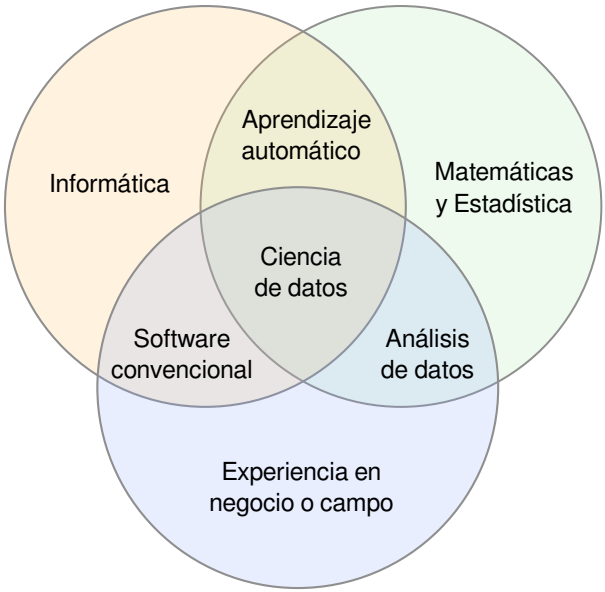

Figura 1: Diagrama de Venn para la ciencia de datos [Finzer, 2013].

algunos elementos químicos.[13] La tabla 1 muestra algunos datos.

Para algunos de estos ejemplos se desconoce su clase y estamos interesados en saber cuál es esta. Los ejemplos tendrán algunas propiedades, atributos o características. Posiblemente haya algún tipo de relación entre las propiedades de los ejemplos y su tipo. Estamos interesados en descubrir estas relaciones.

Para un conjunto de ejemplos, se sabe de qué clase son y se dice que los ejemplos tienen etiquetas y el conjunto está etiquetado. El objetivo es construir un **clasificador** o **modelo**, que dado un ejemplo de clase desconocida asigne, prediga[14, 15] o *adivine* cuál es su clase. Se dice que el ejemplo a clasificar es la entrada del clasificador y que su salida es la predicción.

[13] Es una versión simplificada del conjunto de datos disponible en [German, 1987].

[14] Usamos el término predecir no solo para lo que va a ocurrir en el futuro, también para lo que ocurre en el presente, pero es desconocido. Por ejemplo, un paciente puede tener ya una enfermedad, pero no sabemos cuál y podemos usar un clasificador para predecirla. Aunque también podemos usar clasificadores para predecir la probabilidad de que alguien desarrolle una enfermedad.

[15] «Es difícil hacer predicciones, especialmente sobre el futuro» [Schrodi et al., 2014].

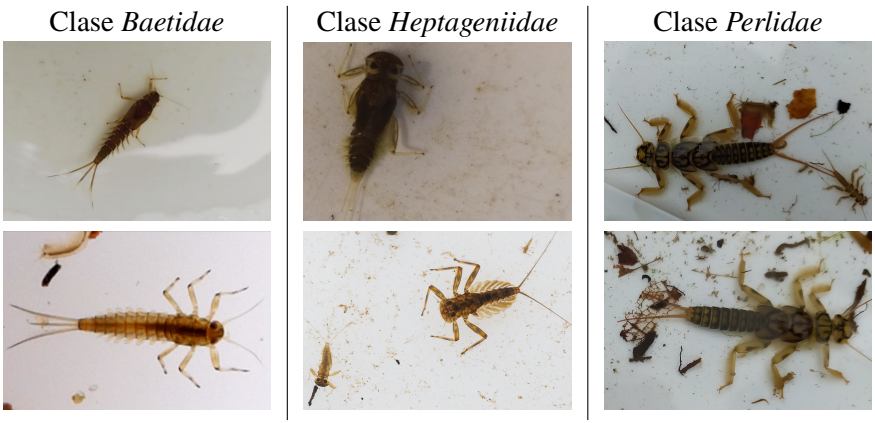

Figura 2: Ejemplos de un conjunto de entrenamiento para clasificación. Para conseguir un clasificador adecuado se necesitarían bastantes más ejemplos de cada clase. Se muestran tres clases, cada una con dos ejemplos. Imágenes cedidas por el proyecto Aquacolab (FCT-21-16785).

Ejemplos de entrenamiento

Refracción	Sodio	Magnesio	Aluminio	Silicio	Potasio	Calcio	Bario	Hierro	Ventana
1.518	12.79	3.50	1.12	73.03	0.64	8.77	0.00	0.00	sí
1.515	14.14	0.00	2.68	73.39	0.08	9.07	0.61	0.05	no
1.516	12.16	3.52	1.35	72.89	0.57	8.53	0.00	0.00	sí
1.518	13.21	3.48	1.41	72.64	0.59	8.43	0.00	0.00	sí
1.513	14.40	1.74	1.54	74.55	0.00	7.59	0.00	0.00	no

Ejemplos para clasificar

Refracción	Sodio	Magnesio	Aluminio	Silicio	Potasio	Calcio	Bario	Hierro	Ventana
1.517	13.14	3.45	1.76	72.48	0.6	8.38	0.00	0.17	?
1.517	13.48	3.48	1.71	72.52	0.62	7.99	0.00	0.00	?
1.517	14.75	0.00	2.00	73.02	0.00	8.53	1.59	0.08	?

Tabla 1: Algunos ejemplos de conjuntos de datos con los que construir un clasificador y con los que aplicarlo. Para obtener un clasificador adecuado hacen falta bastantes más ejemplos.

Figura 3: Ejemplo de imágenes sobre las que usar el clasificador obtenido a partir de los ejemplos de la Figura 2.

Ese proceso de construir el clasificador se considera un aprendizaje a partir de ejemplos.[16] Se dice que se ha **entrenado**[17] al clasificador y que los ejemplos de los que se aprende son los de entrenamiento.

Los métodos que se usan para construir clasificadores son los **algoritmos de aprendizaje**. Un algoritmo es una secuencia de instrucciones que resuelven un problema. El problema que se resuelve con los métodos de aprendizaje automático es construir un clasificador a partir de unos ejemplos de entrenamiento. Otros algoritmos resuelven otros problemas, como ordenar un conjunto de elementos o encontrar la mejor ruta entre dos localidades.

Normalmente los clasificadores no van a acertar la clase de todos los ejemplos, pero si tienen una precisión elevada,[18] pueden ser de utilidad.[19]

3. TIPOS DE CLASIFICADORES

Un clasificador predice una clase dados los valores de las propiedades de los ejemplos. Dependiendo de cómo se consideran y combinan los atributos

[16] También es un tipo de aprendizaje (y enseñanza) humano.

[17] En vez de **enseñado**, que sería más apropiado con la idea de aprendizaje.

[18] Lo que se considere elevado depende del problema, entre otros aspectos del número de clases. Una precisión de solo el 1 % para acertar el número de la lotería estaría bastante bien.

[19] «Todos los modelos son incorrectos, pero algunos son útiles» [Box, 1976].

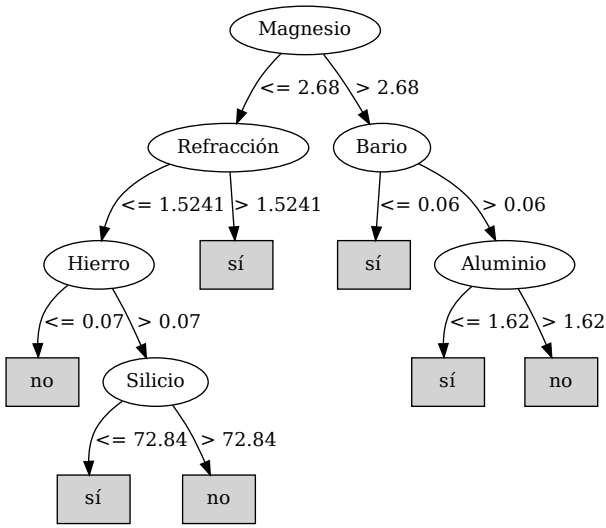

Figura 4: Árbol de decisión para el conjunto de datos de la tabla 1.

tenemos distintos tipos de clasificadores.

Un tipo son los árboles de decisión [Maimon and Rokach, 2014, Costa and Pedreira, 2023] que son una forma de ordenar de forma jerárquica un conjunto de decisiones de forma que se plantean tomar unas u otras decisiones en función de qué respuestas se ha dado a decisiones previas, así hasta llegar a una decisión final que consiste, típicamente, en asignar una clase al ejemplo analizado.

Para el conjunto de datos de los cristales, en un árbol (como el que se muestra en la figura 4) se podría empezar consultando si el valor de Magnesio es menor o mayor que un número. Si es menor se podría consultar la refracción, mientras que si es mayor se podría consultar el Bario. Así sucesivamente hasta llegar a las predicciones de si el cristal es o no de ventana.

Los modelos así obtenidos se basan exclusivamente en los datos con los que se han construido, sin ningún tipo de conocimiento, sentido común o *inteligencia* adicional. Así que no tienen porque ser coherentes con la realidad o el conocimiento existente. Una de las diferencias de este tipo de sistemas con el aprendizaje humano es que cada conjunto de datos se empieza desde cero, sin

Si Magnesio≤2.41, Hierro≤0.09 y Aluminio≤1.19 entonces Ventana=no
Si Calcio≤6.65 entonces Ventana=no
En otro caso: Ventana=sí

Figura 5: Reglas de clasificación para el conjunto de datos de la tabla 1.

tener en cuenta lo aprendido previamente.[20]

Otro tipo de clasificadores, similares a los árboles, son las reglas de decisión [Angelino et al., 2017]. Una regla está formada por varias condiciones sobre los valores de los atributos y si se cumplen esas condiciones, se predice una de las clases. Varias reglas pueden formar un clasificador, como las que se muestran en la figura 5. Se pueden convertir los árboles a reglas y viceversa, pero el resultado de la conversión puede ser bastante más grande que el clasificador de partida.

Los árboles y las reglas son fáciles de entender e interpretar, aunque esto puede depender de cuantos elementos lo formen. Una característica compartida por árboles y reglas es que posiblemente no usen todos los atributos. Eso puede ser deseable si hay atributos que son irrelevantes o redundantes para determinar la clase. El no usar todos los atributos también ayuda a la comprensibilidad de los clasificadores.

Dado que a menudo los problemas son complejos, los árboles o reglas no tienen una precisión suficiente, incluso siendo excesivamente grandes. Por eso, otros tipos de modelos pueden ser más adecuados.

En vez de seleccionar solo las propiedades más importantes, se puede considerar que cada una de ellas aporta algo de información sobre la clase y combinar de alguna manera esos aportes. En los clasificadores **bayesianos** la información de los atributos se interpreta, combina y trata como probabilidades [Xuan et al., 2019]. La tabla 2 y la figura 6 muestran ejemplos.

Otra manera de clasificar ejemplos es comparando con los ejemplos para los que conocemos su clase, seleccionando el más parecido y asignado la clase que

[20] Aunque hay métodos hay métodos que intentan incorporar de alguna manera conocimiento existente [Willard et al., 2022] o aprovechar lo aprendido en un conjunto de datos sobre otros conjuntos de datos relacionados [Zhuang et al., 2020].

		Ventana	
		sí	no
Refracción	≤1.52	0.485	0.566
	>1.52	0.515	0.434
Sodio	≤13.3	0.594	0.226
	>13.3	0.406	0.774
Magnesio	≤3.49	0.370	0.981
	>3.49	0.630	0.019
Aluminio	≤1.37	0.624	0.132
	>1.37	0.376	0.868
Silicio	≤72.8	0.552	0.358
	>72.8	0.448	0.642
Potasio	≤0.56	0.424	0.736
	>0.56	0.576	0.264
Calcio	≤8.61	0.564	0.321
	>8.61	0.436	0.679
Bario	≤0.03	0.933	0.453
	>0.03	0.067	0.547
Hierro	≤0.01	0.618	0.830
	>0.01	0.382	0.170

Tabla 2: Tabla de probabilidades para un clasificador bayesiano, sobre los datos de la tabla 1. Por ejemplo, si el cristal no es de ventana la probabilidad de que el sodio sea mayor que 13.3 es 0.774.

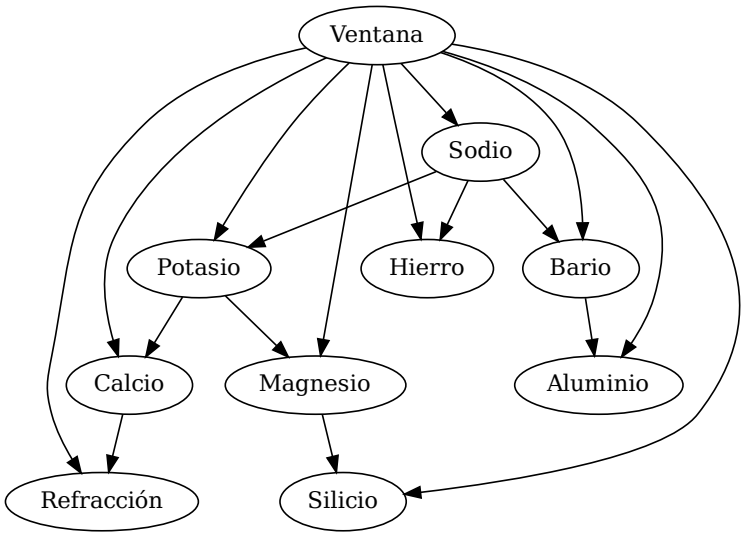

Figura 6: Ejemplo de la estructura de una red bayesiana, para los datos de la tabla 1. Cada elemento de la red tiene asociada una distribución de probabilidad.

tenga ese ejemplo. Este estrategia se denomina **método del vecino más cercano** y en general las aproximaciones que se basan en comparar con los ejemplos de entrenamiento se denominan **basados en instancias** [Cunningham and Delany, 2021]. Para poder determinar en un ordenador cuánto se parecen los ejemplos necesitamos alguna fórmula o función de similitud (distancia),[21] como puede ser la euclídea. Dependiendo del tipo de datos, puede haber distancias bastante más apropiadas.

Los valores de los atributos se pueden combinar usando cualquier fórmula. Por ejemplo, podemos sumar los valores de los atributos multiplicados por números y en función de que el resultado sea mayor o menor que otro número asignar una u otra clase.[22] O podemos asociar a cada clase una de estas formulas y predecir la clase para la que su fórmula proporcione el mayor valor. La figura 7 muestra una de estas funciones.

[21] También se pueden aprender estas funciones de distancia, en lo que se conoce como **aprendizaje de métricas** [Bellet et al., 2022].

[22] Como ocurre con los métodos de **máquinas de vectores soporte** (SVM) [Cervantes et al., 2020].

$$
\begin{array}{lll}
+\,0.0050 \times \text{Refracción} & +\,0.5870 \times \text{Sodio} & -\,0.9026 \times \text{Magnesio} \\
+\,0.8935 \times \text{Aluminio} & +\,0.4777 \times \text{Silicio} & +\,0.0238 \times \text{Potasio} \\
-\,0.3337 \times \text{Calcio} & +\,0.4295 \times \text{Bario} & -\,0.5925 \times \text{Hierro} \\
-\,39.2935 & &
\end{array}
$$

Figura 7: Ejemplo de función para clasificar los datos de la tabla 1. La decisión se toma teniendo en cuenta el signo del resultado.

Además de multiplicar los atributos por números y sumarlos, podemos usar cualquier otra operación o función matemática para su combinación.[23]

Estas funciones matemáticas aparecen en las **redes de neuronas artificiales** [Abiodun et al., 2018], inspiradas en las naturales. Las *neuronas* que las componen reciben valores numéricos como entradas, los multiplican por otros números denominados pesos y aplican alguna función matemática que produce otro valor numérico como resultado. Las neuronas se organizan en capas, con una de entrada con las propiedades de los ejemplos, otra de salida con el valor de la confianza de cada clase y producen una salida como puede ser la confianza en algunas de las clases. Puede haber varias capas intermedias, en las que las salidas de las neuronas de una capa son las entradas de las neuronas de las siguientes capas. Dado que cada neurona se corresponde con una función matemática, el resultado final son funciones definidas en términos de funciones intermedias o auxiliares. Dado el número de capas y el número de neuronas de cada una, el entrenamiento consiste en determinar, ajustar, esos pesos numéricos. Las figuras 8 y 9 muestran la estructura de dos de estas redes.

Los métodos de **aprendizaje profundo** (*deep learning*) [LeCun et al., 2015] son la base de muchos de los recientes y espectaculares avances de la IA. Son redes de neuronas en las que hay muchas capas intermedias. Se denomina profundo porque hay muchas capas, nada que ver con lo que se puede entender por profundo en el aprendizaje humano. Al haber muchas capas y muchísimas neuronas, en este tipo de modelos puede haber una cantidad enorme de pesos numéricos, por lo que pueden requerir mucha memoria y tiempo para su uso, pero especialmente de entrenamiento para encontrar valores adecuados de esos pesos. Esto no sería posible hoy sin los avances en los ordenadores, pero tampoco

[23] En el caso de las máquinas de vectores soporte se puede hacer usando funciones *kernel* (núcleo) [Tharwat, 2019].

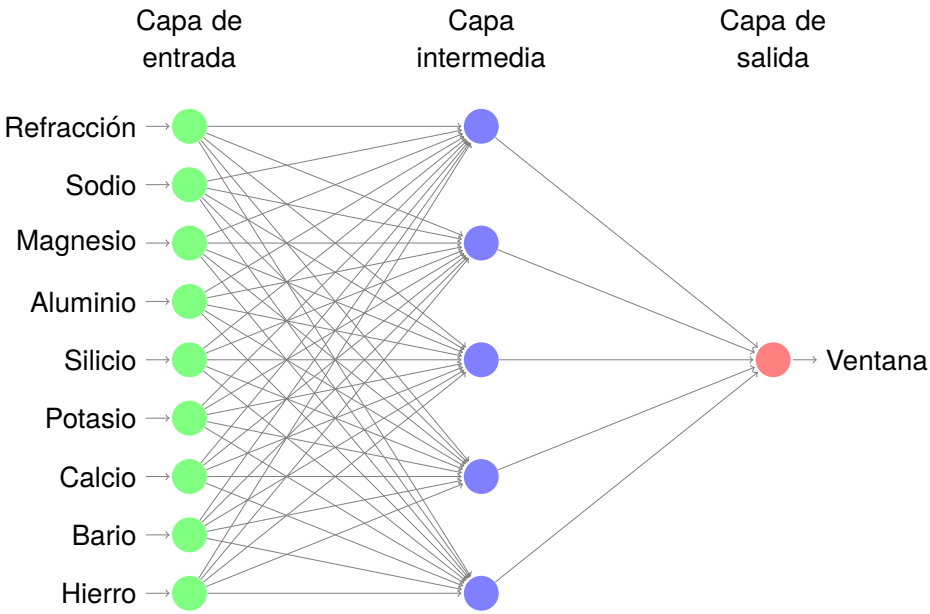

Figura 8: Arquitectura de red de neuronas para los datos de la tabla 1. Las conexiones entre neuronas tienen valores numéricos (pesos) y en cada neurona se calcula una función sobre sus entradas multiplicada por los correspondientes pesos.

sin la aparición de mejores métodos y algoritmos.

En la naturaleza, el aprendizaje y la inteligencia aparecen en los individuos, pero también en las especies. Hay comportamientos inteligentes que son innatos y fruto de la evolución. Hay métodos automáticos inspirados en la evolución [Al-Sahaf et al., 2019]. El comportamiento inteligente también aparece en colectivos de individuos, como las colonias de hormigas y los enjambres de abejas, donde la inteligencia emerge de la combinación de los comportamientos de los integrantes del colectivo.

Hay métodos que se inspiran en el funcionamiento de otros sistemas biológicos, como el sistema inmunitario [Bayar et al., 2015]. También incluso en procesos físicos, como los que ocurren en procesos térmicos o en la difusión de gases [Salcedo-Sanz, 2016].

Si se siguen considerando distintos tipos de métodos es porque no hay un

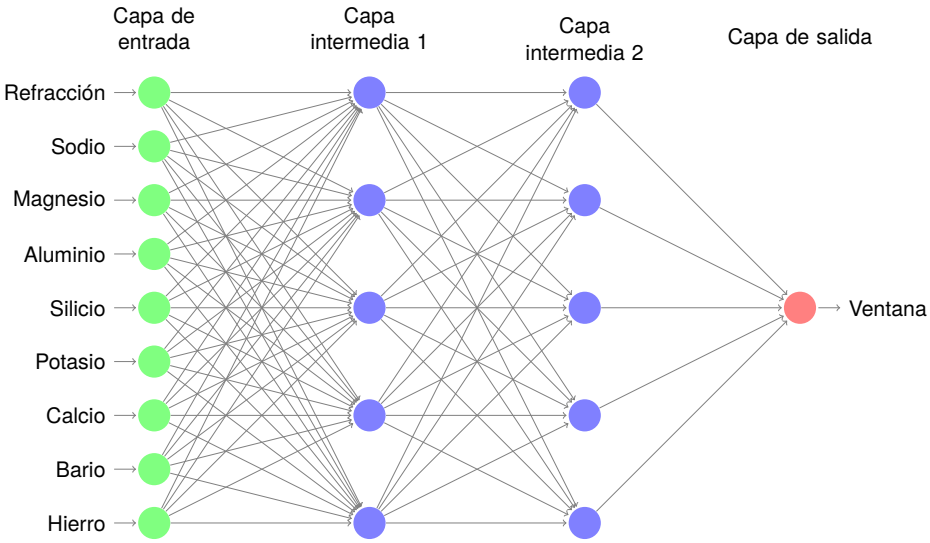

Figura 9: Arquitectura de red de neuronas con dos capas intermedias.

mejor método universal para todos los conjuntos de datos. Incluso para un mismo conjunto de datos distintos métodos pueden tener sus ventajas e inconvenientes.

Por supuesto, también se pueden combinar distintos métodos, en lo que se conoce como combinaciones, grupos o conjuntos (*ensembles*) de clasificadores[24] [Kuncheva, 2014]. También podemos tener combinaciones de clasificadores donde todos ellos son del mismo tipo, aunque diferentes. Por ejemplo, podemos combinar varios árboles en un bosque [González et al., 2020], como el que se muestra en la figura 10. Es bastante natural consultar distintas opiniones de distintos expertos al tener que tomar una decisión. Un tribunal puede ser más justo que una sola persona. Al resolver un examen conjuntamente se pueden obtener mejores resultados que individualmente, a cada uno se le puede dar mejor un tipo de preguntas o temas.

[24]Estos métodos son una de las especialidades de nuestro grupo de investigación, tanto en el desarrollo de algoritmos como en su aplicación.

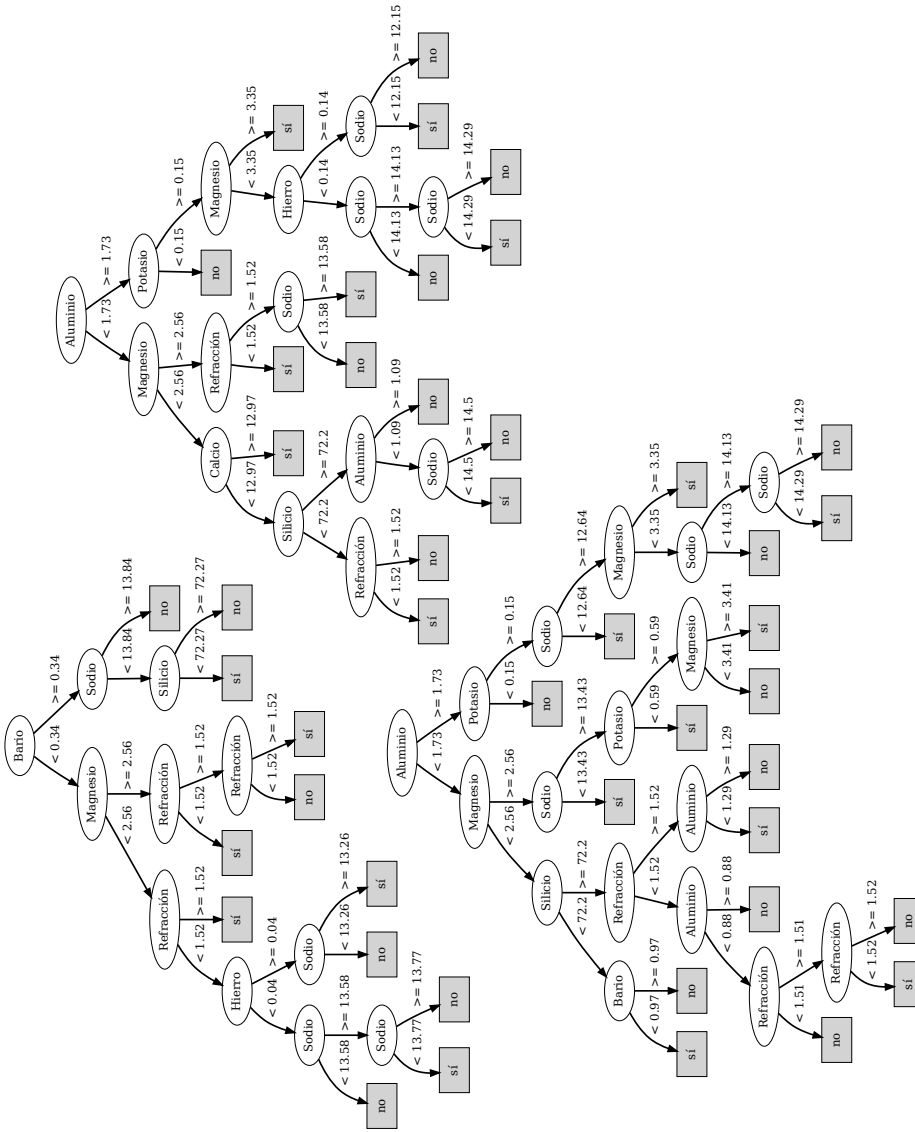

Figura 10: Varios árboles de un bosque. Cada árbol hace una predicción, la predicción del bosque es el resultado de la votación.

Regresión

A menudo el valor que queremos predecir no es una clase, sino un número. Por ejemplo, una temperatura, valor monetario o demanda de un producto. En esos casos se habla de regresión. Muchos métodos de clasificación tienen sus correspondientes versiones para regresión [Fernández-Delgado et al., 2019]. Es más, un problema de clasificación se puede replantear como un problema de predicción numérica donde el número es la probabilidad de la clase, con lo que también los métodos de regresión se podrían utilizar para resolver problemas de clasificación.

Además de producir como resultado un valor numérico, un mismo modelo puede dar como resultado muchos valores. Por ejemplo, una imagen está formada por píxeles, cada uno de ellos con sus valores numéricos.

4. TIPOS DE DATOS

Para poder aprender de ejemplos, necesitamos que tengan algunas propiedades, características o atributos. El caso más simple es que todos los ejemplos tengan los mismos atributos y que se diferencien en los valores de estos atributos. En ese caso podemos organizar los datos en una tabla o en una hoja de cálculo, colocando los ejemplos en las filas y los atributos en columnas. A menudo se asume esta estructura, pero cuando se quiere indicar explícitamente, se puede decir que son datos tabulares [Shwartz-Ziv and Armon, 2022].

Los valores de una propiedad pueden ser numéricos o textuales. En el caso de los textuales, para un atributo concreto a menudo los posibles valores están restringidos a unos cuantos. Por ejemplo, para el estado civil solo hay unos pocos valores posibles. También hay atributos que solo pueden tomar los valores cierto y falso. En estos casos se dice que los datos categóricos o nominales. Muchos métodos asumen que los datos son numéricos y en ese caso los categóricos tendrán que representarse de alguna manera [Hancock and Khoshgoftaar, 2020].

Es habitual que para algunos ejemplos se desconozcan los valores de algunas propiedades. O que en algunos ejemplos alguna propiedad no tenga sentido. Si nuestros ejemplos son publicaciones, un artículo estará en alguna revista, mientras que los libros no lo están. Estas situaciones se pueden representar en los datos con algún valor especial, en una hoja de cálculo podríamos dejarlo en blanco.

Pero habrá que tenerlo en cuenta de algún modo a la hora de aprender, y también cuando queremos predecir la clase de ejemplos con valores desconocidos [Lin and Tsai, 2020].

Un tipo de datos habitual en aprendizaje automático son las imágenes [Voulodimos et al., 2018]. Las imágenes se pueden representar de manera que los valores sean los colores de los puntos (píxeles) y los atributos los distintos puntos que conforman las imágenes. Podemos tener imágenes de distintos tamaños, en cuyo caso se podrían redimensionar para que todas tuvieran las mismas dimensiones. Pero utilizar métodos de aprendizaje que trabajan sobre datos tabulares cualesquiera con esta representación de las imágenes no es la mejor opción. Por eso, lo que se hace es extraer o construir otras propiedades de las imágenes que puedan ser más útiles que el color de un píxel concreto. Si podemos detectar líneas, elipses u otras formas, podríamos usar su número, tamaño, color o posición para clasificar. También hay métodos diseñados específicamente para aprender sobre imágenes.

Los textos en lenguaje natural también pueden ser datos. Y pueden ser de distintas clases. Pueden estar en distintos idiomas, estar a favor o en contra, ser de distintos temas, ser publicidad no deseada... Si bien los textos se almacenan como un carácter detrás de otro, está representación no es adecuada para aprender. Se pueden extraer propiedades del texto y usarlas para aprender con métodos que trabajan con datos tabulares. La presencia o frecuencia de cada palabra[25] puede ser una propiedad. Otra opción es usar métodos específicos para trabajar con textos [Kadhim, 2019].

Los textos en lenguaje natural son secuencias de caracteres. Pero hay otras secuencias de caracteres que no son de lenguaje natural. Por ejemplo, en bioinformática [Min et al., 2017] tenemos las secuencias de ADN o las de aminoácidos, como las que forman las proteínas. Estas secuencias pueden ser de distintas clases y se puede intentar aprender estas clases.

En ocasiones los ejemplos pueden estar formados por un número variable de elementos con distintas conexiones entre estos elementos. Tanto los elementos como las conexiones pueden tener sus propiedades. Este tipo de datos se denominan grafos [Zhang et al., 2019]. Por ejemplo, las moléculas pueden representarse como grafos, donde tenemos átomos de distintos tipos con enlaces.

[25] Idealmente también grupos o secuencias de palabras, al menos los más frecuentes.

Las moléculas pueden ser de distintas clases y podemos intentar aprender las relaciones entre su estructura y su clase.

5. APRENDIZAJE NO SUPERVISADO

Hemos considerado el caso en el que los ejemplos a partir de los que se aprende están etiquetados con valores o clases conocidas o predefinidas. Pero también hay casos donde los ejemplos no tienen asociada ninguna etiqueta. También hay tareas y métodos de aprendizaje automático para este tipo de datos.

Clustering

Podemos tener textos manuscritos en un alfabeto desconocido. ¿Cuántos caracteres o símbolos forman ese alfabeto? Al escribir un mismo carácter habrá variaciones y distintos caracteres pueden ser muy parecidos. También es un problema de clasificación, pero no sabemos cuáles son las clases.

Cuando las clases son conocidas se dice que el aprendizaje es supervisado y cuando no lo son es no supervisado [Berry et al., 2019]. Se usa el término supervisado en el sentido de que hay instrucciones o un maestro que nos indica de qué clase es un ejemplo. En el caso del supervisado podemos tener una idea de qué bien lo estamos haciendo, en cuanto tenemos ejemplos para los que sabemos la clase real y por tanto podemos determinar si las predicciones son correctas.

Por establecer una analogía con nuestro aprendizaje, es como tener ejercicios con y sin respuesta. O tener o no a alguien que nos diga si nuestras respuestas son correctas.

La clasificación no supervisada también se denomina agrupamiento (*clustering*) [Ezugwu et al., 2022]. Dado un conjunto de ejemplos, queremos dividirlos en grupos, pero sin saber cuáles son, ni siquiera cuantos. El objetivo es que los ejemplos de un mismo grupo se parezcan o tengan más afinidad entre sí que los ejemplos de distintos grupos. Para ello, se requiere algún criterio o función que nos indique cuánto se parecen dos ejemplos.

La figura 11 muestra varios conjuntos de datos bidimensionales sobre los que identificar grupos. En este caso se puede hacer visualmente porque cada ejemplo solo tiene dos valores, sus coordenadas. Para el conjunto de los cristales

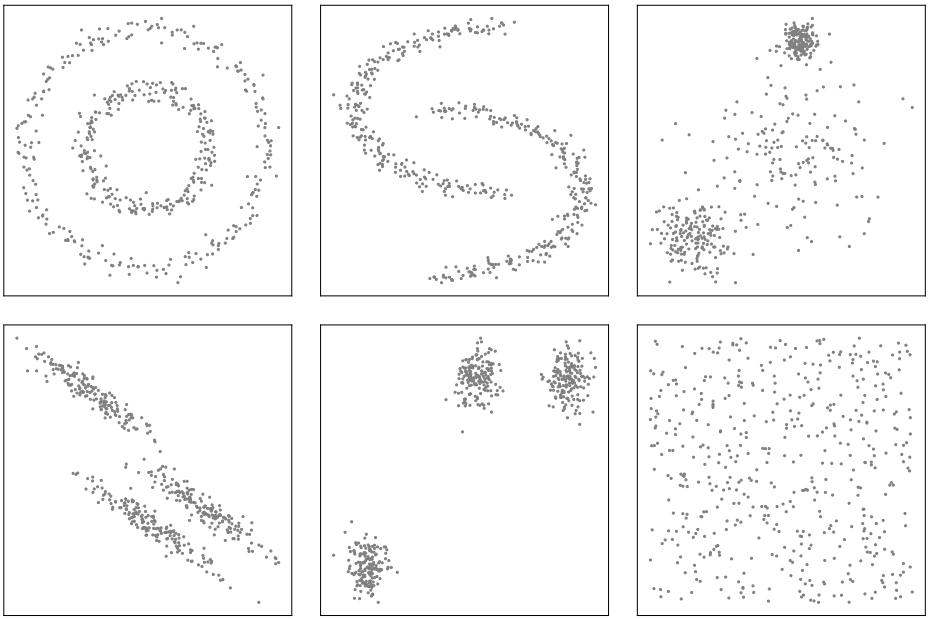

Figura 11: Ejemplos de conjuntos datos sobre los que aplicar *clustering*. En cada recuadro hay un conjunto de puntos y el objetivo es dividirlos en varios grupos.

no se puede hacer visualmente salvo que usemos algún método para seleccionar o reducir el número de propiedades[26] [Reddy et al., 2020].

Detección de anomalías

Otra situación en la que tenemos ejemplos sin etiquetar es la detección de anomalías [Pang et al., 2021]. Dado un conjunto de datos queremos identificar los ejemplos atípicos, anómalos o *raros*.

Podría parecer que es un problema de clasificación supervisada, puesto que hay dos clases, los normales y los atípicos. Pero no lo es porque los ejemplos no están etiquetados como normales y atípicos.

Aun en el caso en el que se dispusiera de ejemplos etiquetados como normales y atípicos, posiblemente la opción de plantearlo como un problema de

[26] Al reducir se pierde información, el objetivo es perder la mínima.

clasificación supervisada no sería la más adecuada, porque los ejemplos atípicos tampoco tienen por qué parecerse entre sí.

También se puede plantear como un problema de clustering. Si tenemos grupos con muy pocos elementos, incluso con uno solo, esos elementos se pueden considerar anomalías.

Por otro lado, la rareza no es un concepto absoluto, se puede ser muy raro o algo raro. Se podría ver como un problema de ordenar los ejemplos de más a menos raro.

Elementos frecuentes y asociaciones

Dado un conjunto de ejemplos sin etiquetar, además de buscar relaciones entre ejemplos, como se hace en clustering, podemos buscar relaciones entre las propiedades de los ejemplos. Ciertas combinaciones de valores pueden ser más frecuentes y esa información puede ser útil [Luna et al., 2019]. En una regla de asociación lo que se indica es que ciertos valores de atributos hacen que sea más probable que otros atributos tomen otros valores.

Un caso típico es el de la cesta de la compra. ¿Qué productos aparecen simultáneamente con más frecuencia en las compras? Hay combinaciones que parecen evidentes, cuando tenemos productos relacionados, como la lechuga y el tomate. Dada cualquier combinación de productos que se nos ocurra, si tenemos datos podemos calcular su frecuencias. Pero de lo que se trata es de encontrar combinaciones inesperadas, que nos permitan descubrir conocimiento nuevo y posiblemente útil. El ejemplo típico es que el comprar pañales implica comprar cerveza, lo que podría estar justificado porque esos compradores pueden tener menos opciones de tomar la cerveza fuera de casa.

Una cuestión es que las combinaciones frecuentes que lo sean solo porque los productos individuales sean frecuentes no son interesantes. Lo que interesa es que la presencia de un producto o productos incremente (o reduzca) la probabilidad de que aparezca otro. Si en una panadería es frecuente comprar pan y comprar leche, la combinación de pan y leche será frecuente. Pero eso solo interesaría si el comprar pan cambia sustancialmente la probabilidad de comprar leche o viceversa.

Dos productos muy poco frecuentes individualmente, necesariamente no

podrán tener una alta frecuencia conjuntamente. Pero la combinación puede ser interesante si su frecuencia es notablemente mayor que la que se tendría si la compra de un producto no afectara de ninguna manera a la compra del otro.

La cesta de la compra es un caso donde se plantea el problema de encontrar asociaciones, pero no el único. Si una persona ha sufrido o sufre una enfermedad puede ser más frecuente que también padezca otra. También puede haber relaciones entre síntomas o resultados de pruebas diagnósticas. Los valores de varios sensores pueden estar relacionados entre sí. Las calificaciones en unas asignaturas pueden estar relacionadas con otras. En las redes sociales puede haber personas que compartan muchos seguidores.

6. OTROS TIPOS DE SUPERVISIÓN

En el aprendizaje supervisado convencional tenemos ejemplos con los valores de sus propiedades y las etiquetas (clase o número) que queremos obtener, lo que sería la respuesta correcta. En el no supervisado directamente no hay etiquetas que queramos obtener.

También hay otras situaciones. En ocasiones tenemos una tarea de aprendizaje supervisado, pero además de datos etiquetados también los tenemos sin etiquetar antes de construir el clasificador. El etiquetar todos los datos puede ser costoso, en tiempo o recursos. También podemos tener ejemplos en los que no haya manera fiable de determinar la etiqueta correcta. Precisamente para estos casos es para los que queremos clasificadores. Podemos ignorar los datos no etiquetados y construir un clasificador supervisado. Un problema de los métodos supervisados es que se necesitan bastantes ejemplos etiquetados para obtener clasificadores adecuados. Los datos sin etiquetar pueden aun así tener información valiosa que podría mejorar los clasificadores obtenidos. En el **aprendizaje semisupervisado** [Van Engelen and Hoos, 2020] se usan datos con y sin etiquetas para obtener clasificadores.

En el aprendizaje humano, el semisupervisado se corresponde con tener ejercicios y las soluciones de solo algunos. Podemos ignorar los no resueltos, pero posiblemente podamos aprender más considerando también los no resueltos.

Cuando etiquetar los ejemplos tiene un coste pero podemos permitirnos etiquetar algunos, está la cuestión de cuáles etiquetar. Esta decisión se puede

integrar en el proceso de aprendizaje, lo que se denomina **aprendizaje activo** [Ren et al., 2021]. En el aprendizaje humano se corresponde con el caso en el que podemos pedir que nos resuelvan algún ejercicio, pero no todos, así que tenemos que seleccionar aquellos ejercicios para los que conocer su solución nos será más útil.

En el aprendizaje semisupervisado o activo hay algo de supervisión, pero no es completa. Por eso se pueden considerar como **aprendizaje débilmente supervisado** [Zhou, 2018]. También tenemos supervisión débil cuando tenemos algo de información sobre las etiquetas, pero no es completa. Para un ejemplo podemos tener probabilidades de que sea de alguna clase, no saber de qué clase es pero sí de que no es de algunas clases... Podemos tener varios ejemplos y desconocer la etiqueta de cada uno, pero tener alguna información respecto a las etiquetas de todos: cuántos puede haber de distintas clases, que algunos ejemplos tengan que ser de la misma o distintas clases que otros...

En el aprendizaje automático, en principio cada problema (conjunto de datos) es independiente, y cada uno se resuelve independientemente, sin tener en cuenta los aprendizajes previos. Cuando se trabaja con tipos de datos muy comunes, como textos o imágenes, ya hay muchos modelos disponibles para estos datos. El aprovechar de alguna manera estos modelos para otros problemas o conjuntos de datos, de manera que el aprendizaje no se realice desde cero, es lo que se conoce como **aprendizaje por transferencia** [Zhuang et al., 2020]. Por ejemplo, para identificar individuos de una especie animal se puede intentar adaptar modelos para otras especies. Esto puede permitir obtener clasificadores adecuados con menos ejemplos etiquetados.

Otro tipo de supervisión es la del **aprendizaje autosupervisado** [Liu et al., 2021], que es otra de las bases principales de los avances actuales en el campo. La idea es que en el propio proceso de aprendizaje se decida cuál es la respuesta que se desea, e incluso cuáles son los ejemplos. El problema de partida no es supervisado, se usan los datos disponibles para crear ejemplos y sus etiquetas. Dado un ejemplo de entrada, podemos intentar predecir una parte del ejemplo a partir de otra. Dado un texto, podemos intentar predecir una palabra a partir de las palabras que la rodean. Dada una imagen de partida, podemos intentar predecir un trozo a partir del resto. Podemos aplicar alguna transformación aleatoria a la imagen, como una rotación, e intentar predecir la imagen de partida o alguna propiedad de la transformación, como el ángulo de la rotación.

Con el aprendizaje autosupervisado se pueden aprovechar cantidades ingentes de datos sin etiquetar, por lo que se puede considerar como un tipo de no supervisado. El autosupervisado puede ser un paso previo o estar integrado en otras tareas de aprendizaje, como la clasificación.

7. MODELOS SOBRE DATOS SENSIBLES

Los datos sobre los que aprender pueden ser sensibles, con lo que el uso del aprendizaje automático sobre estos datos tiene sus implicaciones éticas y legales. Pero también dificultades y retos técnicos.

Idealmente los modelos (clasificadores) deberían ser **comprensibles** o **interpretables** [Rudin, 2019]. Hay métodos que producen modelos que lo son, como los árboles de decisión y las reglas, aunque dejan de serlo a medida que su tamaño aumenta y se hace menos manejable. También se pueden entender fórmulas matemáticas simples. Pero normalmente los modelos comprensibles no son los más precisos.

Hay modelos muy complicados, formados por una cantidad enorme de números combinados con distintas funciones, lo que denominamos **modelos de caja negra** y no son comprensibles. Se puede intentar obtener modelos simples a partir de complejos, pero perdiendo precisión [Ming et al., 2018].

El concepto de caja negra se usa en el sentido de que no se ve lo que hay dentro, que no es transparente. Se usa el término **transparencia algorítmica** [Rader et al., 2018] para indicar que los motivos de una decisión son visibles, que los algoritmos están disponibles. Pero lo que ocurre con el aprendizaje automático es que los algoritmos lo que hacen es construir modelos a partir de los datos y el mismo algoritmo con distintos datos producirá modelos distintos. Aunque tuviéramos el algoritmo,[27] el modelo resultante e incluso los datos con los que se ha construido, seguimos sin saber los motivos de las decisiones si el modelo no es comprensible.

Una propiedad deseable que pueden tener algunos modelos, aunque no sean comprensibles, es que sean **explicables**, que puedan dar algún tipo de información de en qué se basan para realizar sus predicciones [Linardatos et al.,

[27] Los programas que implementan los algoritmos pueden ser difícilmente comprensibles por ser enormes y muy complicados .

2020, Bodria et al., 2023]. Esta información puede ser global, respecto al funcionamiento de todo el modelo, como la importancia de las distintas propiedades. Pero también puede ser local, dadas una entrada y una predicción, se puede obtener qué propiedades han sido las que más y menos han contribuido a la decisión. En el caso de textos e imágenes se pueden marcar las regiones o palabras que han sido determinantes a la hora de hacer la predicción. Al clasificar imágenes puede que la predicción sea correcta pero que el modelo no se esté fijando en las partes relevantes de la imagen. Ocasionalmente se puede llegar a una respuesta correcta con un razonamiento equivocado.

Otra manera de obtener algún tipo de información de las predicciones individuales es indicando los menores cambios posibles en las propiedades de la entrada para que la predicción cambie [Verma et al., 2020].

Los modelos también deberían respetar la **privacidad** [Mendes and Vilela, 2017]. Aunque la privacidad y la transparencia parecen objetivos contradictorios. Los datos privados tendrán que estar a buen recaudo, pero puede ser lícito aprender sobre esos datos y obtener modelos que sean útiles, por ejemplo en aplicaciones médicas. Si los modelos no son comprensibles y además no dan explicaciones, puede parecer que los datos privados están a salvo. Pero haciendo predicciones con el modelo y usando datos públicos, se podría acabar obteniendo información privada. En esos casos se dice que el modelo tiene fugas.

Así que en aprendizaje automático puede haber **adversarios** [Wang et al., 2019]. Por tanto, una propiedad deseable de los modelos es la **robustez** frente a estos adversarios. También son adversarios quienes buscan que el modelo se equivoque. Por ejemplo, quienes envían *spam* preferirían que sus mensajes se etiquetaran como que no son *spam*. En el caso de las imágenes, pequeños cambios inapreciables para el ojo humano pueden hacer cambiar la clase predicha. Los adversarios pueden atacar a un clasificador ya entrenado, pero también cabe la posibilidad de que los ataques se produzcan sobre los datos con los que se construye el clasificador. Si un clasificador se entrena con datos públicos y los adversarios pueden modificar o añadir datos, podrían modificar el comportamiento del clasificador.

Relacionado con la privacidad está el **derecho al olvido**. Eliminar datos es fácil, si los tenemos localizados. Se puede volver a entrenar los modelos sin esos datos eliminados. Pero dado el coste de construir ciertos modelos, no es factible

construirlos desde cero. Puede haber modelos en uso en los que los datos con los que se construyeron ya no estén disponibles. Por eso, puede ser necesario que los modelos desaprendan u olviden [Xu et al., 2023]. No es trivial, cuando los modelos son cajas negras formados por una cantidad ingente de números donde no sabemos qué números se corresponden con qué datos, hay datos que se corresponden con muchos números y cada número se corresponde con muchos datos. También en las personas el olvido intencionado tiene sus dificultades.

Además de la robustez frente a los adversarios, también es deseable que lo sean frente al ruido [Song et al., 2022]. Es habitual que algunos de los datos con los que se construyen los clasificadores estén incorrectamente etiquetados, aun sin adversarios. O que alguna de las propiedades de los datos no sea correcta. En estos casos se dice que los datos tienen ruido. Este ruido puede ser introducido por adversarios, pero también pueden ser simples errores. Si los datos están repletos de errores, no vamos a poder aprender nada útil de ellos, pero si la proporción de errores es limitada, se debería poder obtener clasificadores útiles.

El uso de modelos de aprendizaje automático puede afectar a las personas. De hecho, incluso a quién vive, dado que estos modelos se pueden utilizar para seleccionar embriones. Se han usado modelos para conceder hipotecas, asignar salarios e incluso para establecer las condenas por delitos. Si los datos con los que se construyen los modelos reflejan alguna discriminación o injusticia, los modelos propagarán estas situaciones. La transparencia ayudaría, pero idealmente los modelos deberían garantizar la **equidad** [Mehrabi et al., 2021]. Esto es, que las predicciones no penalicen la pertenencia a ciertos colectivos.

Si los datos con los que se entrenan los clasificadores no tienen información explícita de los colectivos, o se puede excluir esa información de los datos con los que se construyen los modelos, podría parecer que no hay problema. Pero lo hay, porque la pertenencia al colectivo se podría predecir con una determinada precisión a partir de los datos explícitos disponibles. Por ejemplo, técnicamente podría ser factible un sistema que, dada solo una foto de la cara, determinara alguna decisión sobre esa persona, y sobre los datos históricos podría tener una precisión alta. Por eso, hay métodos que comprueban la equidad de los modelos y que permiten adaptarlos o construirlos para que sean más equitativos.

8. CONCLUSIONES

Para concluir, hemos intentado presentado algunas nociones sobre aprendizaje automático y ciencia de datos.

Estos temas están relacionados y son una parte fundamental de la Inteligencia Artificial, que tiene tantos avances recientes, posibilidades y riesgos. Sobre IA se pueden plantear lecciones mucho más entusiastas[28] y excitantes, pero para eso ya están las noticias sobre el tema que continuamente aparecen.

Los métodos que construyen los modelos que se han presentado aquí son una muestra de algoritmos propuestos en la academia que han terminado utilizándose en muchas aplicaciones prácticas de gran impacto (detectores de *spam*, sistemas de recomendación, mantenimiento predictivo, publicidad personalizada, detección precoz de enfermedades, ...).

Actualmente los métodos de aprendizaje profundo tienen un gran éxito, generando noticias continuamente, cuando hace unos pocos años solo eran conocidos en ámbitos académicos muy restringidos. La IA es un término que incluye muchos campos, pero el aprendizaje automático es uno de los más relevantes.

Los avances actuales en la IA se deben al incremento de las capacidades de cómputo, la mejora de los ordenadores, la disponibilidad de grandes volúmenes de datos, pero también en el desarrollo de nuevos algoritmos de aprendizaje.

La investigación en aprendizaje automático hoy en día está liderada por las grandes empresas tecnológicas. Podemos usar sus servicios en la *nube* para construir modelos con sus métodos en nuestros datos y estos modelos pueden tener una alta precisión. Pero no todos los datos, por muy *anonimizados* que estén, es conveniente que estén almacenados en sus servidores. Y aun en los casos en los que se puedan usar los servicios de estas empresas, no está de más tener algunas nociones de cómo funcionan estos modelos.

Aun con todos los avances y recursos, el construir estos sistemas sobre nuestros propios datos está muy lejos de ser completamente automático. No funcionan sobre los datos originales, *crudos*. Estos necesitan ser procesados para que se pueda aprender sobre ellos. Por tanto, son necesarios conocimientos y trabajo en ingeniería y ciencia de datos.

[28] O reivindicativas. Aunque también catastrofistas.

Incluso con los datos ya procesados, construir modelos adecuados con nuestros datos y recursos limitados es un reto. Podemos tener muy pocos datos. O datos que ocupen mucho, pero que realmente tengan muy poca información. Por ejemplo, en clasificación podemos tener muy pocas etiquetas. Conseguir que uno de estos sistemas aprenda, incluso con los datos adecuados, tiene su *ciencia*.

El uso de estos sistemas tiene sus implicaciones éticas y legales, pero dar soporte a estos aspectos también tiene sus dificultades y retos técnicos.

AGRADECIMIENTOS

Muchas gracias.

A las y los asistentes y lectores, por su paciencia.

A la Dirección de la Escuela Politécnica Superior por su propuesta para impartir esta lección inaugural y al Rectorado por aceptar esta propuesta.

A César García Osorio, Álvar Arnaiz y Carlos Pardo.

A las y los colegas del grupo de investigación, Departamento, Escuela y Universidad, tanto de los actuales como de los previos. Personal docente, investigador, de administración, de servicios, y estudiantado.

A las y los colaboradores en trabajos de investigación, especialmente durante las estancias de investigación en el extranjero y en concreto a la catedrática Ludmila Kuncheva.

A los organismos (Ministerios, Junta, Empresas, Fundaciones, Universidad...) que han financiado nuestra investigación.

A mi familia.

REFERENCIAS

[Abiodun et al., 2018] Abiodun, O. I., Jantan, A., Omolara, A. E., Dada, K. V., Mohamed, N. A., and Arshad, H. (2018). State-of-the-art in artificial neural network applications: A survey. *Heliyon*, 4(11). (Citado en la página 16).

[Al-Sahaf et al., 2019] Al-Sahaf, H., Bi, Y., Chen, Q., Lensen, A., Mei, Y., Sun, Y., Tran, B., Xue, B., and Zhang, M. (2019). A survey on evolutionary machine learning. *Journal of the Royal Society of New Zealand*, 49(2):205–228. (Citado en la página 17).

[Angelino et al., 2017] Angelino, E., Larus-Stone, N., Alabi, D., Seltzer, M., and Rudin, C. (2017). Learning certifiably optimal rule lists. In *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 35–44. (Citado en la página 13).

[Bayar et al., 2015] Bayar, N., Darmoul, S., Hajri-Gabouj, S., and Pierreval, H. (2015). Fault detection, diagnosis and recovery using artificial immune systems: A review. *Engineering Applications of Artificial Intelligence*, 46:43–57. (Citado en la página 17).

[Bellet et al., 2022] Bellet, A., Habrard, A., and Sebban, M. (2022). *Metric learning*. Springer Nature. (Citado en la página 15).

[Berry et al., 2019] Berry, M. W., Mohamed, A., and Yap, B. W. (2019). *Supervised and unsupervised learning for data science*. Springer. (Citado en la página 22).

[Bodria et al., 2023] Bodria, F., Giannotti, F., Guidotti, R., Naretto, F., Pedreschi, D., and Rinzivillo, S. (2023). Benchmarking and survey of explanation methods for black box models. *Data Mining and Knowledge Discovery*, pages 1–60. (Citado en la página 28).

[Box, 1976] Box, G. E. (1976). Science and statistics. *Journal of the American Statistical Association*, 71(356):791–799. (Citado en la página 11).

[Cervantes et al., 2020] Cervantes, J., Garcia-Lamont, F., Rodríguez-Mazahua, L., and Lopez, A. (2020). A comprehensive survey on support vector machine classification: Applications, challenges and trends. *Neurocomputing*, 408:189–215. (Citado en la página 15).

[Costa and Pedreira, 2023] Costa, V. G. and Pedreira, C. E. (2023). Recent advances in decision trees: An updated survey. *Artificial Intelligence Review*, 56(5):4765–4800. (Citado en la página 12).

[Cunningham and Delany, 2021] Cunningham, P. and Delany, S. J. (2021). k-nearest neighbour classifiers-a tutorial. *ACM computing surveys (CSUR)*, 54(6):1–25. (Citado en la página 15).

[Ezugwu et al., 2022] Ezugwu, A. E., Ikotun, A. M., Oyelade, O. O., Abualigah, L., Agushaka, J. O., Eke, C. I., and Akinyelu, A. A. (2022). A comprehensive survey of clustering algorithms: State-of-the-art machine learning applications, taxonomy, challenges, and future research prospects. *Engineering Applications of Artificial Intelligence*, 110:104743. (Citado en la página 22).

[Fernández-Delgado et al., 2019] Fernández-Delgado, M., Sirsat, M. S., Cernadas, E., Alawadi, S., Barro, S., and Febrero-Bande, M. (2019). An extensive experimental survey of regression methods. *Neural Networks*, 111:11–34. (Citado en la página 20).

[Finzer, 2013] Finzer, W. (2013). The data science education dilemma. *Technology Innovations in Statistics Education*, 7(2). (Citado en la página 9).

[German, 1987] German, B. (1987). Glass Identification. UCI Machine Learning Repository. DOI: https://doi.org/10.24432/C5WW2P. (Citado en la página 9).

[González et al., 2020] González, S., García, S., Del Ser, J., Rokach, L., and Herrera, F. (2020). A practical tutorial on bagging and boosting based ensembles for machine learning: Algorithms, software tools, performance study, practical perspectives and opportunities. *Information Fusion*, 64:205–237. (Citado en la página 18).

[Günther et al., 2017] Günther, W. A., Mehrizi, M. H. R., Huysman, M., and Feldberg, F. (2017). Debating big data: A literature review on realizing value from big data. *The Journal of Strategic Information Systems*, 26(3):191–209. (Citado en la página 7).

[Gupta and Chandra, 2020] Gupta, M. K. and Chandra, P. (2020). A comprehensive survey of data mining. *International Journal of Information Technology*, 12(4):1243–1257. (Citado en la página 7).

[Hancock and Khoshgoftaar, 2020] Hancock, J. T. and Khoshgoftaar, T. M. (2020). Survey on categorical data for neural networks. *Journal of Big Data*, 7(1):1–41. (Citado en la página 20).

[Ilyas and Chu, 2019] Ilyas, I. F. and Chu, X. (2019). *Data cleaning*. Morgan & Claypool. (Citado en la página 7).

[Kadhim, 2019] Kadhim, A. I. (2019). Survey on supervised machine learning techniques for automatic text classification. *Artificial Intelligence Review*, 52(1):273–292. (Citado en la página 21).

[Kuncheva, 2014] Kuncheva, L. I. (2014). *Combining pattern classifiers: methods and algorithms*. John Wiley & Sons. (Citado en la página 18).

[LeCun et al., 2015] LeCun, Y., Bengio, Y., and Hinton, G. (2015). Deep learning. *nature*, 521(7553):436–444. (Citado en la página 16).

[Lin and Tsai, 2020] Lin, W.-C. and Tsai, C.-F. (2020). Missing value imputation: a review and analysis of the literature (2006–2017). *Artificial Intelligence Review*, 53:1487–1509. (Citado en la página 21).

[Linardatos et al., 2020] Linardatos, P., Papastefanopoulos, V., and Kotsiantis, S. (2020). Explainable ai: A review of machine learning interpretability methods. *Entropy*, 23(1):18. (Citado en la página 28).

[Liu et al., 2021] Liu, X., Zhang, F., Hou, Z., Mian, L., Wang, Z., Zhang, J., and Tang, J. (2021). Self-supervised learning: Generative or contrastive. *IEEE transactions on knowledge and data engineering*, 35(1):857–876. (Citado en la página 26).

[Luna et al., 2019] Luna, J. M., Fournier-Viger, P., and Ventura, S. (2019). Frequent itemset mining: A 25 years review. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, 9(6):e1329. (Citado en la página 24).

[Maimon and Rokach, 2014] Maimon, O. Z. and Rokach, L. (2014). *Data mining with decision trees: theory and applications*, volume 81. World scientific. (Citado en la página 12).

[Mehrabi et al., 2021] Mehrabi, N., Morstatter, F., Saxena, N., Lerman, K., and Galstyan, A. (2021). A survey on bias and fairness in machine learning. *ACM computing surveys (CSUR)*, 54(6):1–35. (Citado en la página 29).

[Mendes and Vilela, 2017] Mendes, R. and Vilela, J. P. (2017). Privacy-preserving data mining: methods, metrics, and applications. *IEEE Access*, 5:10562–10582. (Citado en la página 28).

[Min et al., 2017] Min, S., Lee, B., and Yoon, S. (2017). Deep learning in bioinformatics. *Briefings in bioinformatics*, 18(5):851–869. (Citado en la página 21).

[Ming et al., 2018] Ming, Y., Qu, H., and Bertini, E. (2018). Rulematrix: Visualizing and understanding classifiers with rules. *IEEE transactions on visualization and computer graphics*, 25(1):342–352. (Citado en la página 27).

[Özsu, 2023] Özsu, M. T. (2023). Data science—a systematic treatment. *Commun. ACM*, 66(7):106–116. (Citado en la página 7).

[Pang et al., 2021] Pang, G., Shen, C., Cao, L., and Hengel, A. V. D. (2021). Deep learning for anomaly detection: A review. *ACM computing surveys (CSUR)*, 54(2):1–38. (Citado en la página 23).

[Papadopoulos et al., 2017] Papadopoulos, P., Kourtellis, N., Rodriguez, P. R., and Laoutaris, N. (2017). If you are not paying for it, you are the product: How much do advertisers pay to reach you? In *Proceedings of the 2017 Internet Measurement Conference*, pages 142–156. (Citado en la página 8).

[Patil and Davenport, 2012] Patil, T. and Davenport, T. (2012). Data scientist: The sexiest job of the 21st century. *Harvard business review*, 90(10):70–76. (Citado en la página 8).

[Rader et al., 2018] Rader, E., Cotter, K., and Cho, J. (2018). Explanations as mechanisms for supporting algorithmic transparency. In *Proceedings of the 2018 CHI conference on human factors in computing systems*, pages 1–13. (Citado en la página 27).

[Reddy et al., 2020] Reddy, G. T., Reddy, M. P. K., Lakshmanna, K., Kaluri, R., Rajput, D. S., Srivastava, G., and Baker, T. (2020). Analysis of dimensionality reduction techniques on big data. *Ieee Access*, 8:54776–54788. (Citado en la página 23).

[Ren et al., 2021] Ren, P., Xiao, Y., Chang, X., Huang, P.-Y., Li, Z., Gupta, B. B., Chen, X., and Wang, X. (2021). A survey of deep active learning. *ACM computing surveys (CSUR)*, 54(9):1–40. (Citado en la página 26).

[Romero and Ventura, 2020] Romero, C. and Ventura, S. (2020). Educational data mining and learning analytics: An updated survey. *Wiley interdisciplinary reviews: Data mining and knowledge discovery*, 10(3):e1355. (Citado en la página 5).

[Rudin, 2019] Rudin, C. (2019). Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nature machine intelligence*, 1(5):206–215. (Citado en la página 27).

[Salcedo-Sanz, 2016] Salcedo-Sanz, S. (2016). Modern meta-heuristics based on nonlinear physics processes: A review of models and design procedures. *Physics Reports*, 655:1–70. (Citado en la página 17).

[Schrodi et al., 2014] Schrodi, S. J., Mukherjee, S., Shan, Y., Tromp, G., Sninsky, J. J., Callear, A. P., Carter, T. C., Ye, Z., Haines, J. L., Brilliant, M. H., et al. (2014). Genetic-based prediction of disease traits: prediction is very difficult, especially about the future. *Frontiers in genetics*, 5:162. (Citado en la página 9).

[Shwartz-Ziv and Armon, 2022] Shwartz-Ziv, R. and Armon, A. (2022). Tabular data: Deep learning is not all you need. *Information Fusion*, 81:84–90. (Citado en la página 20).

[Song et al., 2022] Song, H., Kim, M., Park, D., Shin, Y., and Lee, J.-G. (2022). Learning from noisy labels with deep neural networks: A survey. *IEEE Transactions on Neural Networks and Learning Systems*. (Citado en la página 29).

[Sutton and Barto, 2018] Sutton, R. S. and Barto, A. G. (2018). *Reinforcement learning: An introduction*. MIT press. (Citado en la página 6).

[Tharwat, 2019] Tharwat, A. (2019). Parameter investigation of support vector machine classifier with kernel functions. *Knowledge and Information Systems*, 61:1269–1302. (Citado en la página 16).

[Van Engelen and Hoos, 2020] Van Engelen, J. E. and Hoos, H. H. (2020). A survey on semi-supervised learning. *Machine learning*, 109(2):373–440. (Citado en la página 25).

[Verma et al., 2020] Verma, S., Boonsanong, V., Hoang, M., Hines, K. E., Dickerson, J. P., and Shah, C. (2020). Counterfactual explanations and

algorithmic recourses for machine learning: A review. *arXiv preprint arXiv:2010.10596*. (Citado en la página 28).

[Voulodimos et al., 2018] Voulodimos, A., Doulamis, N., Doulamis, A., Protopapadakis, E., et al. (2018). Deep learning for computer vision: A brief review. *Computational intelligence and neuroscience*, 2018. (Citado en la página 21).

[Wang et al., 2019] Wang, X., Li, J., Kuang, X., Tan, Y.-a., and Li, J. (2019). The security of machine learning in an adversarial setting: A survey. *Journal of Parallel and Distributed Computing*, 130:12–23. (Citado en la página 28).

[Willard et al., 2022] Willard, J., Jia, X., Xu, S., Steinbach, M., and Kumar, V. (2022). Integrating scientific knowledge with machine learning for engineering and environmental systems. *ACM Computing Surveys*, 55(4):1–37. (Citado en la página 13).

[Wu et al., 2022] Wu, X., Xiao, L., Sun, Y., Zhang, J., Ma, T., and He, L. (2022). A survey of human-in-the-loop for machine learning. *Future Generation Computer Systems*, 135:364–381. (Citado en la página 6).

[Xu et al., 2023] Xu, H., Zhu, T., Zhang, L., Zhou, W., and Yu, P. S. (2023). Machine unlearning: A survey. *ACM Computing Surveys*. (Citado en la página 29).

[Xuan et al., 2019] Xuan, J., Lu, J., and Zhang, G. (2019). A survey on bayesian nonparametric learning. *ACM Computing Surveys (CSUR)*, 52(1):1–36. (Citado en la página 13).

[Zhang et al., 2019] Zhang, S., Tong, H., Xu, J., and Maciejewski, R. (2019). Graph convolutional networks: a comprehensive review. *Computational Social Networks*, 6(1):1–23. (Citado en la página 21).

[Zhou, 2018] Zhou, Z.-H. (2018). A brief introduction to weakly supervised learning. *National science review*, 5(1):44–53. (Citado en la página 26).

[Zhou, 2021] Zhou, Z.-H. (2021). *Machine learning*. Springer Nature. (Citado en la página 5).

[Zhuang et al., 2020] Zhuang, F., Qi, Z., Duan, K., Xi, D., Zhu, Y., Zhu, H., Xiong, H., and He, Q. (2020). A comprehensive survey on transfer learning. *Proceedings of the IEEE*, 109(1):43–76. (Citado en las páginas 13, 26).


IN ITINERE VERITAS
BVRGENSIS VNIVERSITAS