Universidad Internacional de Valencia

# Implementación y comparativa de métodos semi-supervisados


Titulación:
Máster en Inteligencia Artificial

Curso académico
2023-2024

Alumno: Martínez Acha, David
D.N.I: 71310644H

Directora de TFM: Irma Sanabria


Convocatoria:

x

X 2024


De:
Planeta Formación y Universidades

*El ayer es historia, el mañana es un misterio y el hoy es un regalo… por eso se llama presente.*

Eleanor Roosevelt

# Agradecimientos

Me gustaría agradecer...

También quiero destacar...

Por último...

# Índice general

# Índice de figuras

# Índice de tablas

# Índice de algoritmos

# Resumen

# Introducción 1

El aprendizaje automático o *machine learning* como disciplina de la inteligencia artificial resulta ser uno de los campos más cotizados y que despierta más interés en prácticamente cualquier aplicación (investigación, automatización, sistemas de ayuda, detección...). Existe una división muy clara del aprendizaje automático que consta de: aprendizaje supervisado y el no supervisado. Pero existe otra división que no suele mencionarse, y que puede ser muy beneficiosa, este es el aprendizaje semi-supervisado.

De forma resumida, el aprendizaje supervisado trata de aprender de datos de los que se sabe lo que representan para después poder inferir este conocimiento para nuevos datos (por ejemplo, dadas las características de una flor, se intenta predecir de qué clase concreta es), el aprendizaje no supervisado trata de aprender de datos de los que **no** se sabe lo que representan, se utiliza en tareas en las que es necesario realizar agrupaciones o divisiones en base a las similitudes/disimilitudes de los ejemplos (por ejemplo, podría distinguir entre animales que tienen plumaje de los que no sin tener el conocimiento de qué animales son concretamente). En el caso del aprendizaje supervisado, el etiquetado de los datos suele ser un proceso costoso (es posible imaginar, por ejemplo, la cantidad de tiempo y recursos que podría suponer el etiquetado masivo de millones de muestras de posibles cánceres). En la realidad, la mayor parte de los datos no están etiquetados. Ante esta necesidad aparece el aprendizaje semi-supervisado, que se encuentra a caballo entre el supervisado y no supervisado y que permite aprovechar los escasos datos etiquetados para inferir su conocimiento a los no etiquetados.

## 1.1. Aprendizaje automático

El aprendizaje automático (*machine learning*) según Martel (2020) es una rama de la Inteligencia Artificial y se trata de una técnica de análisis de datos que enseña a las computadoras a aprender de la **experiencia** (como los humanos). Para llevar a cabo este proceso, el aprendizaje automático requiere de una amplia cantidad de datos, o los necesarios para el problema específico en cuestión. Estos datos son procesados mediante algoritmos, los cuales se alimentan de ejemplos (también conocidos como instancias o prototipos). A través de estos ejemplos, los algoritmos tienen la capacidad de generalizar comportamientos ocultos.

Estos algoritmos mencionados mejoran su rendimiento iterativamente y de forma automática durante su entrenamiento e incluso también durante su aprovechamiento/explotación. El aprendizaje automático ha adquirido una gran relevancia en una amplia variedad de áreas

como la visión artificial, automoción, detección de anomalías o automatización, entre otras. El aprendizaje automático generalmente se clasifica en tres tipos: aprendizaje supervisado, aprendizaje no supervisado y aprendizaje por refuerzo. Sin embargo, ha surgido una nueva disciplina que se sitúa entre el aprendizaje supervisado y el no supervisado, utilizando tanto datos etiquetados como no etiquetados durante el proceso de entrenamiento van Engelen y Hoos (2020).

En la figura 1.1 se presenta una clasificación del aprendizaje automático.

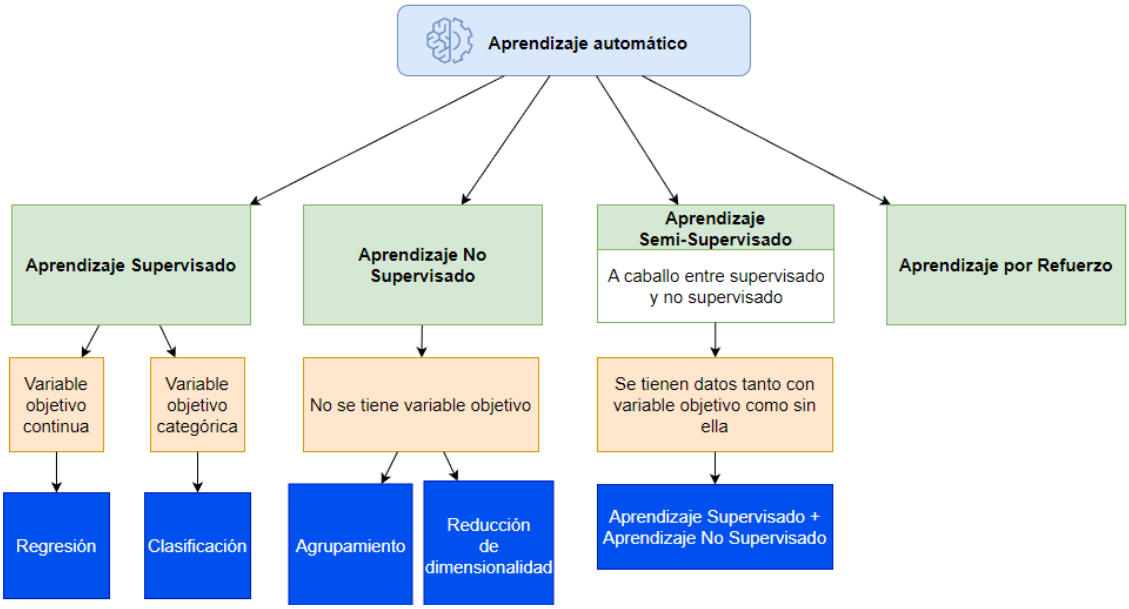


**Figura 1.1**: ***Clasificación de aprendizaje automático, basado en Solutions (2018).***

### 1.1.1. Aprendizaje supervisado

Los algoritmos de aprendizaje supervisado utilizan datos etiquetados durante su proceso de entrenamiento Alexander S. Gillis (2021). Un ejemplo popular de datos etiquetados podría ser un conjunto flores de iris y las posibles etiquetas podrían ser: setosa, versicolor y virginica. Estos datos estarán formados por un conjunto de características (en el caso de las flores de iris podrían ser la longitud y ancho del sépalo y del pétalo). Estas características podrían ser categóricas, continuas o binarias Dridi (2021a).

Para generar un modelo correcto, estos datos son divididos en varios subconjuntos: conjunto de entrenamiento (*training data set*), conjunto de validación (*validation data set*) y conjunto de test (*test data set*). El conjunto de entrenamiento corresponde con la porción de los datos que el algoritmo utilizará para aprender un modelo que generalice los patrones ocultos subyacentes. El conjunto de validación permite comprobar, durante el proceso de entrenamiento, que el modelo que se está generando no memoriza los datos (fenómeno conocido como sobreajuste), también sirve para finalizar el entrenamiento (e.g. el error en validación aumenta durante varias iteraciones). Una vez que el algoritmo ha generado un modelo, se utiliza el conjunto de test para comprobar el rendimiento real (una estimación) Wikipedia contributors (2024). Ningún

dato de este último conjunto ha sido "visto" por el modelo previamente.

En la figura 1.2 se encuentra un diagrama con el funcionamiento general.

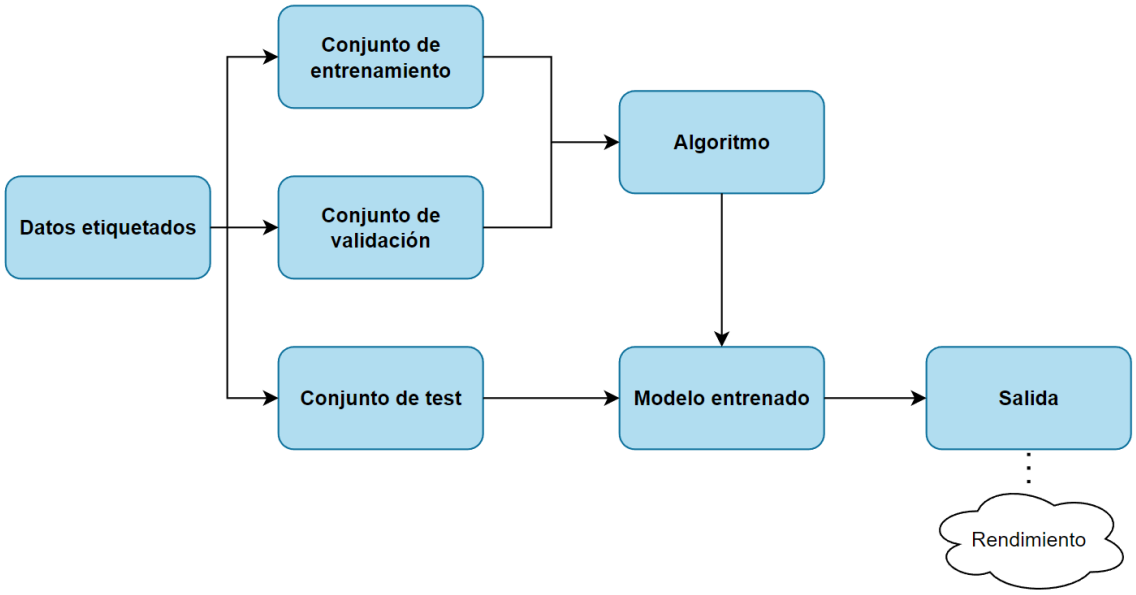


**Figura 1.2**: ***Funcionamiento general del aprendizaje supervisado, basado en Dridi (2021a).***

Partiendo del concepto de etiqueta de un dato, el problema será de **clasificación** si los valores que puede tomar la etiqueta representan un conjunto finito. Por otro lado, si estos valores son continuos, el problema será de **regresión**.

- **Clasificación**: Un modelo entrenado en un problema de clasificación se denomina clasificador. Ante un nuevo dato, el clasificador predecirá su etiqueta correspondiente. Por lo general, a cada valor de etiqueta se le suele llamar clase. Dependiendo de la cantidad de valores, se referirá a un problema binario o multiclase.
- **Regresión**: En este caso, ante un nuevo dato, el modelo predecirá un valor continuo. La idea subyacente es evaluar una función (ajustada/aprendida durante el entrenamiento) dado un dato como variables de entrada.

### 1.1.2. Aprendizaje no supervisado

A diferencia del aprendizaje supervisado, el no supervisado no trabaja con datos etiquetados y clases. Según Dridi (2021b) esto quiere decir que nosotros no "supervisamos" el algoritmo. No se le añade ese conocimiento extra. Estos algoritmos intentarán descubrir patrones que se encuentren en la propia estructura de los datos (de sus características). La idea del aprendizaje no supervisado es estudiar las similitudes/disimilitudes que hay entre los datos y, por ejemplo, obtener una separación o agrupación de los mismos (e.g. separación de especies en imágenes de animales sin conocer el animal concreto).

Entre las principales aplicaciones del aprendizaje no supervisado se encuentran las siguientes:

1. **Agrupamiento (Clustering)**: Divide los datos en grupos. Los ejemplos de un grupo tendrán cierta similitud entre ellos, mientras que todos los ejemplos de ese grupo serán disimilares a los de otro grupo (y por eso se genera esa división). Algunos algoritmos necesitan conocer de antemano el número de grupos en los que dividir lo datos, otros son capaces de descubrir cuántos grupos existen Dridi (2021b).

2. **Reducción de la dimensionalidad**: Los conjuntos de datos generalmente tiene un número bastante grande de características. Esto hace que los algoritmos de aprendizaje sean más lentos. La reducción de dimensionalidad hace referencia a la reducción de número de características tratando de no perder información al hacerlo. Según javaTpoint se denomina como:

   > «*Una forma de convertir conjuntos de datos de alta dimensionalidad en conjunto de datos de menor dimensionalidad, pero garantizando que proporciona información similar.*»

   Algunos ejemplos concretos de reducción de dimensionalidad son:

   - Análisis de Componentes Principales (PCA).
   - Cuantificación vectorial.
   - Autoencoders.

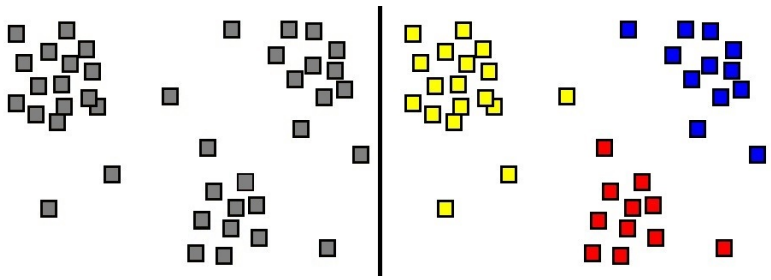

**Figura 1.3**: Clusters. *A la izquierda los datos sin agrupar y a la derecha los datos coloreados según la pertenencia a los distintos grupos. By hellisp - Own work, Public Domain, https://commons.wikimedia.org/w/index.php?curid=36929773.*

### 1.1.3. Aprendizaje semi-supervisado

Según van Engelen y Hoos (2020), el aprendizaje semi-supervisado es la rama del aprendizaje automático que utiliza tanto datos etiquetados como no etiquetados durante el entrenamiento. Es por esto que se dice que está a medio camino entre el aprendizaje supervisado y el no supervisado. Como se ha comentado, el problema al que todos los algoritmos se enfrentan en la realidad es a la escasez de datos etiquetados, pues es un proceso costoso. Gracias a la naturaleza del semi-supervisado, hace que sea una buena aproximación para esos casos. Por lo general, suele aplicarse en problemas de clasificación.

#### 1.1.3.1. Suposiciones

¿Y por qué utilizar aprendizaje semi-supervisado? Lo cierto es que algunos de los algoritmos existentes de aprendizaje supervisado funcionan bastante bien incluso con pocos datos etiquetados. Sin embargo, los datos no etiquetados podrían aprovecharse para mejorar el rendimiento.

El objetivo, por tanto, del aprendizaje semi-supervisado será obtener clasificadores que obtengan mejores resultados que los del aprendizaje supervisado. En van Engelen y Hoos (2020) se especifican unas condiciones que han de cumplirse.

La primera premisa que se debe cumplir es que la distribución $p(x)$ de entrada contenga información sobre la distribución posterior $p(y|x)$ van Engelen y Hoos (2020).

**Smoothness assumption**

Probablemente, si dos ejemplos se encuentran próximos en el espacio, comparten la misma etiqueta.

**Low-density assumption**

La frontera de decisión en un problema de clasificación se encontrará en una zona del espacio en el que existan pocos ejemplos.

**Manifold assumption**

Los ejemplos suele encontrarse en una estructuras de dimensionalidad baja (algunas características no son útiles), denominadas *manifolds*. Los ejemplos que se encuentren en una misma *manifold* comparten la misma etiqueta Lukas Huber (2022); van Engelen y Hoos (2020).

**Cluster assumption**

Los ejemplos que se encuentren en un mismo grupo compartirán la misma etiqueta.

El concepto clave de todas estas suposiciones es el de la “similitud” (ejemplos próximos en el espacio, ejemplos en misma manifold, mismo grupo...). Es por esto que la *Cluster assumption* es una generalización del resto (o el resto son versiones de esta).

Por esto, para que el **el aprendizaje semi-supervisado mejore al supervisado** es necesario que se cumpla dicha suposición generalizada. Si no fuese así (i.e. datos no agrupables), el aprendizaje semi-supervisado no mejorará al supervisado van Engelen y Hoos (2020).

En la figura 1.4 se presenta la taxonomía general del aprendizaje semi-supervisado.

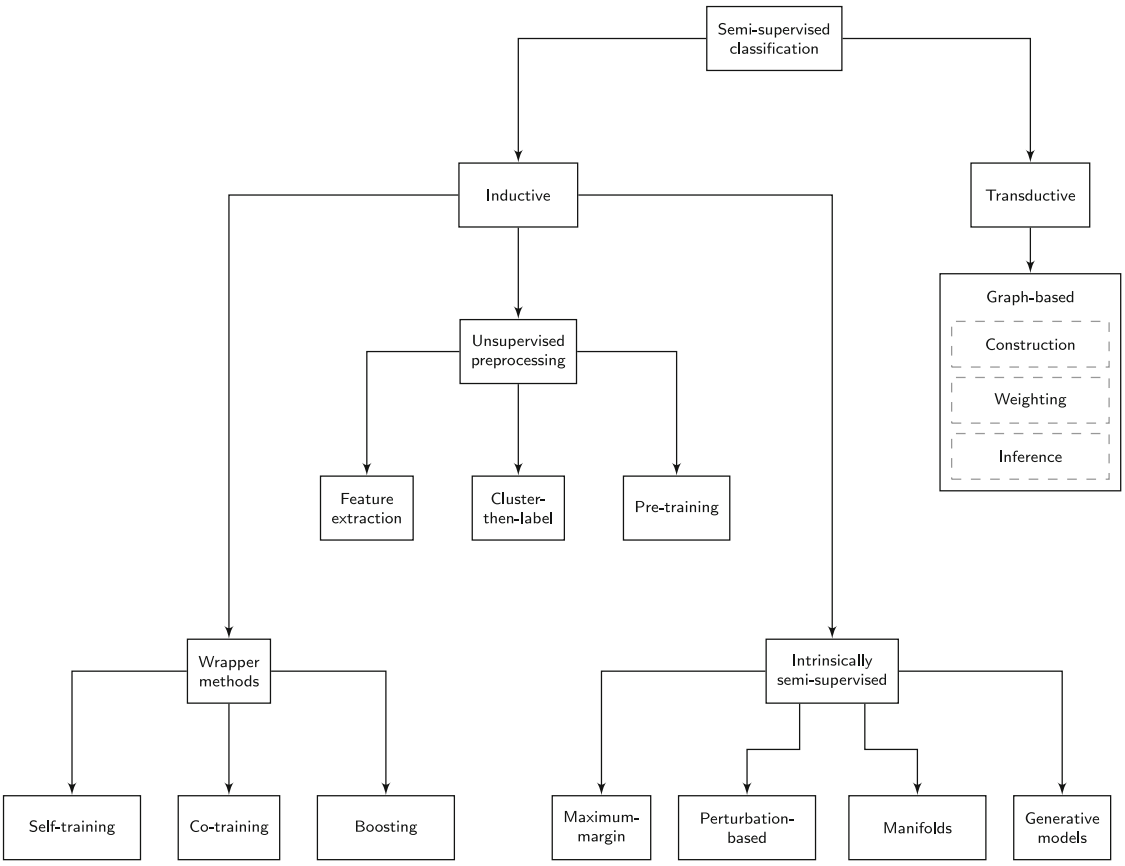


**Figura 1.4**: ***Taxonomía de métodos semi-supervisados van Engelen y Hoos (2020).***

Sin pérdida de generalidad, este trabajo estará centrado en métodos semi-supervisados basados en grafos y árboles (intrínsecamente semi-supervisados) con la comparación con otros métodos enmarcados en esta taxonomía.

## 1.2. Árboles de decisión (CART)

Antes de entrar en las explicaciones teóricas es conveniente indicar que existen multitud de algoritmos que permiten la creación de árboles (ID3 Quinlan (1986), C4.5 Quinlan (2014), C5.0 Quinlan (2004) y CART Breiman (2017), entre otros).

El algoritmo en el que se centrará este desarrollo será CART (Classification and regression trees) para árboles de clasificación.

## 1.3. Grafos

### 1.3.1. Inferencia

# Objetivos

2

Los algoritmos semi-supervisados suponen un área de mucha utilidad dentro del *machine learning*, sin embargo, así como para otras ramas (como el aprendizaje supervisado y no supervisado) es posible encontrar numerosas bibliotecas y algoritmos bien desarrollados y probados, para el semi-supervisado todavía hay una gran cantidad de investigación que no se ha materializado (o que si lo ha hecho, no se ha publicado).

## 2.1. Objetivo general

El objetivo general del presente trabajo es realizar una revisión bibliográfica (guiada) sobre métodos de aprendizaje semi-supervisado, centrándose en los ámbitos de grafos y árboles para realizar posteriormente una implementación y validación con otros algoritmos bien afianzados en este ámbito como Self-Training o Co-Forest.

## 2.2. Objetivos específicos

Se proponen una serie de objetivos específicos que surgen a raíz de una revisión bibliográfica para evaluar los algoritmos más prometedores tanto para grafos como árboles:

1. Implementación completa y desde cero de un algoritmo de construcción de árboles que permita trabajar con datos etiquetados y no etiquetados (semi-supervisado). Incluye también de la adición de algoritmos complementarios como *post-pruning*.

2. Debido a la naturaleza de los algoritmos basados en grafos que necesitan de dos pasos separados (construcción de grafo y propagación de etiquetas), se requiere la implementación desde cero de varios algoritmos tanto de construcción de grafo como de *label propagation*.

3. Seleccionar el/los conjuntos de datos adecuados para la experimentación de estos algoritmos, tanto los que cumplen las suposiciones del aprendizaje semi-supervisado como los que no, para una validación exhaustiva que refleje la utilización de estos algoritmos en muy diversos ámbitos. Al menos, 20 *datasets*.

4. Codificación de experimentos adecuados para algoritmos. Previsiblemente incluirá: preprocesado de datos, codificación de validaciones cruzadas, experimentación de parámetros específicos (influencia) o graficar resultados, entre otros...

5. Comparación, gracias a experimentos, con algoritmos afianzados del estado del arte para la extracción de resultados y conclusiones.

# Metodología

3

# Resultados y Discusión

# 4

# Conclusiones

5

1. Conclusión 1.
2. Conclusión 2.
3. Conclusión 3.

# Limitaciones y Perspectivas de Futuro

# 6

# Apéndize A

A

# Apéndize B

B

# Bibliografía


Alexander S. Gillis, D. P. (2021). Supervised learning. https://www.techtarget.com/searchenterpriseai/definition/supervised-learning. [Internet; descargado 18-abril-2024].

Breiman, L. (2017). *Classification and regression trees*. Routledge.

Dridi, S. (2021a). Supervised learning - a systematic literature review. *ResearchGate*.

Dridi, S. (2021b). Unsupervised learning - a systematic literature review. *ResearchGate*.

javaTpoint. Unsupervised machine learning. https://www.javatpoint.com/unsupervised-machine-learning. [Online; accessed 18-April-2024].

Lukas Huber (2022). A friendly intro to semi-supervised learning. https://towardsdatascience.com/a-friendly-intro-to-semi-supervised-learning-3783c0146744. [Online; accessed 18-April-2024].

Martel, J. (2020). Machine learning: qué es y cómo funciona. https://itelligent.es/es/machine-learning-que-es-como-funciona/. [Internet; descargado 18-abril-2024].

Quinlan, J. R. (1986). Induction of decision trees. *Machine learning*, 1:81–106.

Quinlan, J. R. (2004). C5. 0. *http://www. rulequest. com/see5-info. html*.

Quinlan, J. R. (2014). *C4. 5: programs for machine learning*. Elsevier.

Solutions, N. T. (2018). Machine learning algorithms: Beginners guide part 1. https://www.neovasolutions.com/2018/06/06/machine-learning-algorithms-beginners-guide-part-1/. [Internet; descargado 18-abril-2024].

van Engelen, J. E. y Hoos, H. H. (2020). A survey on semi-supervised learning. *Machine Learning*, 109(2):373–440.

Wikipedia contributors (2024). Training, validation, and test data sets — Wikipedia, the free encyclopedia. https://en.wikipedia.org/w/index.php?title=Training,_validation,_and_test_data_sets&oldid=1218746717. [Online; accessed 18-April-2024].