# Métodos de Aprendizaje Supervisado y No Supervisado[[1]](#footnote-1) en Clasificación de Tweets

Los métodos de ***aprendizaje supervisado*** pretenden asignar una etiqueta *t* a una instancia de entrada representada por un vector de características .

Por su parte, el ***aprendizaje no supervisado*** persigue algo distinto: una buena representación interna de los datos de entrada.

En el primer caso, utilizaremos conjuntos de entrenamiento para obtener los valores del parámetro quemejor asigne las etiquetas para el modelo especificado por .

Este aprendizaje supone ajustar los parámetros del modelo de modo que se *minimice la discrepancia* entre la etiqueta real de la instancia ***t***y la predicha por el modelo para cada caso del conjunto de entrenamiento.

El *aprendizaje no supervisado* ha sido prácticamente ignorado por la comunidad de aprendizaje automático con la excepción del análisis de conglomerados (*cluster*). De hecho, algunas definiciones de aprendizaje máquina excluyen este tipo de aprendizaje. Una razón para ello es que resulta difícil decir cuál es el objetivo del aprendizaje supervisado.

* Un objetivo importante, es obtener una representación interna de la entrada quesea útil *para un posterior aprendizaje supervisado*. Es decir, el problema de aprendizaje se descompone en dos partes: una sin supervisión y otra supervisada.
* Permitir una representación con *baja dimensionalidad*, más compacta *de la entrada*. P.e. en una imagen con millones de píxeles, solo algunos cientos de estos son importantes. PCA se utiliza para esta finalidad.
* Permite representar las *características de entrada* de alta dimensionalidad de un *modo económico*:
  1. Características binarias.
  2. Valores reales, pero con muchos valores nulos.
* Encontrar conglomerados razonables en la entrada. Esto equivale a encontrar características dispersas en que solo una de las características es distinta de cero.

Lo anterior muestra que, en definitiva, lo que debemos realizar es una ***clasificación supervisada***, único medio de asignar a cada instancia - *tweet* - una etiqueta - *odio*/*no\_odio* - aunque los métodos no supervisados - cluster, gensim,.. - en nuestro caso podrían utilizarse como *etapa* *previa* para reducir el tamaño del conjunto de entrenamiento y/o la dimensionalidad del espacio de atributos , lo que ya hemos hecho en el proyecto:

* Lo primero, mediante el *filtrado de* *tweets* por el vocabulario de odio y lo segundo
* *Seleccionando atributos* por la χ2

pero ***no es de ningún modo posible*** *utilizar aprendizaje no supervisado para clasificar* los tweets.

1. Prescindimos del *aprendizaje reforzado* en que la salida es una acción o un conjunto de acciones que persiguen maximizar una recompensa. [↑](#footnote-ref-1)