Pablo Arranz ropero y Juan Alberto camino Sáez



Si me lo como, ¿me muero?

**Índice**

* [Descripción del proyecto](#Descripción del proyecto)…………………….………………………………………………………….……3
* Apéndices……………………………..

# **Descripción del proyecto**

El propósito de este proyecto es aplicar diferentes algoritmos de aprendizaje automático a un conjunto de datos de setas encontrado en la siguiente URL:

<https://www.kaggle.com/uciml/mushroom-classification>

El conjunto de datos presenta un total de 8124 instancias con 23 atributos cada una (incluyendo el resultado como atributo).

Cada uno de los datos representa lo explicado en el Excel adjunto.

# **Consideraciones previas al desarrollo de los algoritmos**

Como podemos observar en el conjunto de datos original tenemos un conjunto de datos que maneja caracteres para definir el valor de cada atributo. Para una mayor facilidad de cómputo transformaremos dichos atributos a dígitos numéricos según el Excel adjunto. Nos ayudaremos de ciertas expresiones regulares similares a:

sed -i -E 's/^([0-9]+\,[0-9]+\,[0-9]+\,[0-9]+\,)t/\11/g' mushrooms.csv

Para adaptar la expresión regular, si queremos sustituir el atributo n repetiremos “[0-9]+\,” n-1 veces y después del cierre de paréntesis escribiremos el carácter que queremos sustituir seguido de “/\1” y el dígito por el que lo sustituiremos.

En el caso de la expresión regular del ejemplo se sustituirá, en la columna del 5º atributo, todos los caracteres t por el dígito 1.

A continuación, crearemos un archivo llamado *mushroomdata* en el que se guardarán los datos de manera que ejecutando la instrucción load se cargarán automáticamente en la variable *data*.

También hemos creado una función que recibe un nombre de fichero y carga los datos y los separa en 3 subconjuntos: Subconjunto de entrenamiento, subconjunto de cross-validation y subconjunto de test.

El código de dicho método es el siguiente:

<https://es.mathworks.com/matlabcentral/answers/101996-how-can-i-sort-a-vector-in-a-random-manner-in-matlab>