

UNIVERSIDAD PERUANA DE CIENCIAS APLICADAS  
CIENCIAS DE LA COMPUTACIÓN

MACHINE LEARNING  
Laboratorio Repaso  
(Primer Semestre del 2019)

Objetivos de aprendizaje:

- Entrenar algoritmo de Machine Learning.
  - Describir modelos algorítmicos.
  - Identificar paradigmas de aprendizaje de máquina.
- 

## 1. Pregunta 1

Se desea encontrar el modelo más óptimo para predecir el conjunto de datos denominado **Fertility**. La descripción completa de este conjunto de datos la puede encontrar en la URL <https://archive.ics.uci.edu/ml/datasets/Fertility>. Deberá usar como datos de entrenamiento el archivo `fertility_Diagnosis.csv` usando la técnica de la validación cruzada. El archivo se encuentra en el aula virtual.

## 2. Pregunta 2

En el archivo `Energy_efficiency.csv` se almacena un conjunto de datos en donde se registran ciertos parámetros de construcción, la carga de calefacción y la carga de refrigeración. Los parámetros de construcción están representados por los atributos **X1** (*Relative Compactness*), **X2** (*Surface Area*), **X3** (*Wall Area*), **X4** (*Roof Area*), **X5** (*Overall Height*), **X6** (*Orientation*), **X7** (*Glazing Area*) y **X8** (*Glazing Area Distribution*). La **carga de calefacción** está representado por el atributo **y1** (*Heating Load*) y la **carga de refrigeración** por el atributo **y2** (*Cooling Load*).

La eficiencia energética se puede definir en función de los valores de la carga de calefacción y la carga de refrigeración. Se desea saber si es que es posible definir la eficiencia energética en términos de los parámetros de construcción. Se desea encontrar el modelo de regresión lineal más óptimo para predecir el conjunto de datos. La descripción completa de este conjunto de datos la puede encontrar en la URL <https://archive.ics.uci.edu/ml/datasets/Energy+efficiency>.

## 3. Pregunta 3

El conjunto de datos **Pima Indians Diabetes Database**, que se encuentra en el archivo `diabetes.csv`, posee los siguientes atributos, todos ellos numéricos:

- `preg` (Number of times pregnant)
- `plas` (Plasma glucose concentration a 2 hours in an oral glucose tolerance test)
- `pres` (Diastolic blood pressure (mm Hg))
- `skin` (Triceps skin fold thickness (mm))
- `insu` (2-Hour serum insulin ( $\mu$ U/ml))
- `mass` (Body mass index (weight in kg/(height in m)<sup>2</sup>))
- `pedi` (Diabetes pedigree function)

- age (Age (years))

Al atributo meta es `class` el cual puede tener los siguientes valores:

- tested\_positive
- tested\_negative

Se le pide que entrene un modelo de clasificación para este conjunto de datos combinando los modelos algorítmicos que usted crea más conveniente.

## 4. Pregunta 4

Un entrenador físico está analizando la dieta que ingiere uno de los deportista que entrena. Al entrenador le gustaría poder clasificar los alimentos en grupos homogéneos, los más similares posibles, para de esta manera, analizar la composición de los alimentos. Se le pide a Ud. que le ayude al realizar esta tarea usando el conjunto de datos `food.arff` o `food.csv` que se encuentra en el aula virtual.

Monterrico, 25 de junio de 2019.