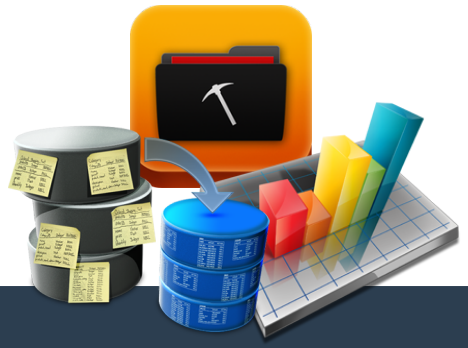
Universidad Nacional Autónoma de México
Facultad de Ciencias

Profesores:
Dra. Amparo López Gaona
M. I. Gerardo Avilés Rosas

Laboratorista:
José Luis Vázquez Lázaro



1. Objetivo

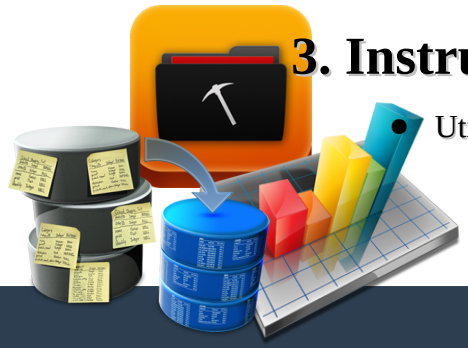
Aplicar la construcción de agrupamiento a un problema de identificación de categorías y futura clasificación, así como su uso como técnica de exploración de datos.

2. Marco teórico

El ***Clustering*** (**Agrupamiento**) es un procedimiento de agrupación de una serie de tuplas de acuerdo con un criterio. Esos criterios son por lo general **distancia** o **similitud**. La cercanía se define en términos de una determinada **función de distancia**, como la euclidiana, aunque existen otras más robustas o que permiten extenderla a variables discretas. La medida más utilizada para medir la similitud entre los casos es la matriz de correlación entre los nxn casos. Sin embargo, también existen muchos algoritmos que se basan en la maximizaciónde una propiedad estadística llamada verosimilitud.

Generalmente, las tuplas de un mismo grupo (o clústers) comparten propiedades comunes. El conocimiento de los grupos puede permitir una descripción sintética de un conjunto de datos multidimensional complejo. De ahí su uso en **Minería de Datos**. Esta descripción sintética se consigue sustituyendo la descripción de todos los elementos de un grupo por la de un representante característico del mismo.

En algunos contextos, como el de la **Minería de Datos**, se lo considera una técnica de **aprendizaje no supervisado** puesto que busca encontrar relaciones entre variables descriptivas pero no la que guardan con respecto a una variable objetivo (como una etiqueta de clase).



3. Instrucciones

Utilizar **R Studio** y el dataset (adjunto a la práctica) **Adult.cvs**.

Elaborado por José Luis Vázquez Lázaro / 23 de abril de 2018.

4. Actividades

1. Cree un agrupamiento de los datos utilizando el algoritmo ***k-medoids***, con ***k*** = 5, utilizando la función ***pam*** de la biblioteca ***cluster***.
2. Detecte los grupos mas representativos del censo.
3. Indique cuáles son los valores atípicos encontrados bajo esta configuración. Justifique.

5. Entregables

Deberás enviar un archivo `.zip`, con nombre `<número de cuenta>_practica13`, que contenga lo siguiente:

- El proyecto `RStudio` con el script de la primer actividad solicitada.
- Un archivo `.pdf` con el desarrollo y resultado de las actividades 2 y 3.
- Un archivo `README.txt` que contenga tu nombre completo, tu número de cuenta y tu correo.

a la dirección de correo `luis_lazaro@ciencias.unam.mx` con el asunto `[A&MD2018-2]Practica13`.