PRACTICA 1: AUDITORIA DE DATOS

ENTREGA: Viernes 6 de Octubre de 2017

Una de las partes más costosa en aprendizaje automático es preparar los datos para ser procesados posteriormente. Lo primero que tenemos que hacer es entender el problema al que nos vamos a enfrentar, invirtiendo tiempo en estudiar/visualizar la base de datos que nos facilita el cliente/colaborador.

En esta práctica se pide realizar una auditoría de los datos de la base suministrada. Esto nos permitirá conocer los datos con los que se va a trabajar.

El cliente (en este caso tus profesores) nos marca unos hitos para realizar el estudio de la base de datos (ten en cuenta que el cliente puede pedir puntos que son irrealizables en su base de datos):

- Descripción de las variables y valores estadísticos (minimo, máximo, media, desviación, mediana, etc.). Estudia qué valores estadísticos son los convenientes según el tipo de variable y procede en consecuencia.
- Describe y realiza modificaciones en la base datos si lo consideras necesario. Por ejemplo, qué harías con valores nominales, si los hubiera.
- Estudia si es necesario normalizar los datos y cómo lo harías. Procede a modificar la base de datos (normalizar) si lo consideras necesario.
- Detección de valores extremos (outliers) y descripción de qué harías en cada caso.
- Detección de valores perdidos (missing values) y descripción de cómo actuarías para solventar el problema.
- Buscar correlaciones entre:
  - las variables predictoras, lo que permitirá ver si hay variables redundantes.
  - variables predictoras y la clase (target).
- Detecta, si hubiera, falsos predictores.
- Estudia si fuera conveniente segmentar alguna de las variables.
- Estudia si fuera conveniente crear nuevas variables sintéticas basada en las variables originales.

El cliente solicita un documento (audit), de un máximo de 12 páginas, que recoga las conclusiones de los puntos anteriores así como otras deducciones inferidas del estudio de la base de datos y que aportan conocimiento al problema.