PRACTICA 1: AUDITORIA DE DATOS (Parte I)

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 (mínimo, 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 10 páginas, que recoja las conclusiones de los puntos anteriores así como otras deducciones inferidas del estudio de la base de datos y que aportan conocimiento al problema.

PRACTICE 1: DATA AUDIT (Part I)

One of the most expensive parts of machine learning is preparing the data for further processing. The first issue we have to do is to understand the problem we are going to face, investing time to study/visualize the database provided by the client/collaborator.

In this practice we are asked to carry out an audit of the data of the supplied database. This will allow us to know the data we are going to work with.

The client (in this case your teachers) sets us some milestones to carry out the study of the database (bear in mind that the client can ask for points that are unachievable in their database):

(1) Description of the variables and statistical values (minimum, maximum, mean, deviation, median, etc.). It studies which statistical values are suitable according to the type of variable and proceeds accordingly.
(2) Describe and modify the database if necessary. For example, what would you do with nominal values, if any.
(3) Study if it is necessary to normalize the data and how you would do it. Proceed to modify the database (normalize) if you consider it necessary.
(4) Detection of extreme values (outliers) and description of what you would do in each case.
(5) Detection of missing values and description of how you would act to solve the problem.
(6) Search for correlations between:
    * the predictor variables, which will allow to see if there are redundant variables.
    * Predictor variables and the class (target).
(7) Detection of false predictors, if any.
(8) Study if it is convenient to segment some of the variables.
(9) Study if it is convenient to create new synthetic variables based on the original variables.

The client requests a document (audit), of a maximum of 10 pages, that gathers the conclusions of the previous points as well as other inferred deductions of the study of the database and that contribute knowledge to the problem.