Indicaciones para el trabajo final

En el archivo "**Todas las bases.xlsx**" se encuentran las bases de datos. A cada alumno le fue asignado un par de bases, una para clasificación supervisada, cuya variable target se destaca en rojo, y otra para clasificación no supervisada.

El archivo "**Listado de asignación de bases.xlsx**" contiene la lista que especifica con qué bases deberá trabajar cada alumno, y las indicaciones correspondientes para obtener una muestra de cada conjunto.

Cada alumno **no analizará la base tal cual está en el Excel**. Dependiendo del caso se deberá considerar sólo algunas variables y/o tomar una muestra del total de registros fijando una semilla. Para eso, una vez seleccionadas las variables, el alumno construirá el conjunto de datos tomando, según se indique, su número de DNI como semilla.

A continuación se muestra un ejemplo en R de cómo construir el conjunto de datos final a partir de una selección aleatoria sin reemplazo de registros de la base completa.

```
dni=31234567

n=round(0.9* nrow(datos))      #si se quisiera el 90% de los
datos

set.seed(dni);cuales= sample(1:nrow(datos), size=n,
replace=FALSE)

misdatos=datos[cuales,]
```

Situaciones a considerar según la base

nafta diesel:	90% de los datos según la semilla seteada
accidentes:	90% de los datos según la semilla seteada
Bancarrotasosana:	90% de los datos según la semilla seteada
Insectos:	elegir 5 variables cualesquiera
Próstata:	80% de los datos según la semilla seteada
bajo peso:	80% de los datos según la semilla seteada
Esófago:	100 registros según la semilla seteada
Renal:	80% de los datos según la semilla seteada
Seguros:	75% de los datos según la semilla seteada
Médicos:	90% de los datos según la semilla seteada
Spotify:	70% de los datos según la semilla seteada
Venta_autos:	80% de los datos según la semilla seteada
Telecomunicaciones:	Elegir 5 variables cualesquiera
Paises:	90% de los datos según la semilla seteada y excluir una variable

Consigna

a) Ejercicio de Clasificación Supervisada

a.1- Elija y justifique las variables que considere pertinentes para la clasificación.

a.2- Pruebe al menos dos métodos de clasificación distintos y compare la bondad de clasificación de cada uno de los métodos mediante un algoritmo no ingenuo (matriz de confusión mínimamente).

a.3- Analice el cumplimiento de los supuestos si fuera necesario y en caso de cumplirse utilice un algoritmo robusto.

a.4- Concluya en términos del problema con qué algoritmo se quedaría y por qué.

b) Ejercicio de Clasificación no Supervisada

b.1- Seleccione las variables pertinentes y la distancia que va a utilizar.

b.2- Utilice al menos dos algoritmos para realizar una clusterización.

b.3- Decida criteriosamente la cantidad de clusters.

b.4- Elija una clusterización y explique las características de los agrupamientos logrados.

Nota 1: en todo el trabajo use libremente los contenidos de la asignatura, si cree necesario incorporar algún contenido no explicado en la cursada, cite la referencia correspondiente.

Nota 2: el script puede ser en R o Python como anexo. El trabajo debe ser presentado en PDF tipo informe (no puede ser .RMD o similar). La longitud máxima es de 5 carillas. No colocar gráficos que no sean estrictamente necesarios.

Entrega

La fecha límite de entrega es el **1° de septiembre de 2020**. El alumno deberá enviar un mail al correo debiechan@gmail.com con el asunto "**TP-AID-31234567**", donde los ocho números corresponden a su DNI. En él se adjuntará el informe en formato PDF, y el código como anexo.