Inferencia Estadística y Reconocimiento de Patrones - **Segundo Cuatrimestre 2025**

Trabajo Práctico 2 a ser resuelto con Python y R

Condiciones formales

- La consigna debe ser resuelta usando Python y R, según corresponda.
- Fecha de entrega: hasta el miércoles 26/10 23 hs.
- Modalidad de trabajo: en pareja.
- Modalidad de entrega: cada pareja deberá subir **3** archivos a la tarea correspondiente en el aula virtual de la materia.

Archivo 1 en pdf: informe sobre la resolución del ejercicio. El informe detallará lo realizado en el ejercicio, demostrando la comprensión de los temas abordados.

Archivo 2 notebook implementada en lenguaje Python, de extensión ipynb. Contendrá el código implementado para resolver parte del problema planteado y su nombre será `pareja0.ipynb` según el número de pareja que sea.

Archivo 3 implementado en lenguaje R. Contendrá el código implementado para resolver parte del problema planteado cuyo nombre será `pareja0.R` según el número de pareja que sea.

Es importante que tengan en cuenta que los archivos 2 y 3 no resuelven lo mismo. En el archivo 2 deben resolver lo que se pueda con Python, y en el archivo 3, lo que se pueda con R. **Los archivos 2 y 3 son complementarios, no equivalentes.** Aunando los archivos 2 y 3 se debe encontrar una completa resolución del problema planteado.

---

1. Para el conjunto de datos trabajado en el TP1 clasificar utilizando:
   - árboles de clasificación podados (pareja 5).
   - bosques aleatorios (pareja 2).
   - boosting (pareja 4).
   - SVM multiclase (pareja 3).

   La pareja que no figure en los ítems anteriores deberá calcular un intervalo bootstrap de la MAD (*median absolute deviation*) de la variable `pH`, justificando la elección y el procedimiento. El intervalo bootstrap construirlo a partir de una submuestra (representativa) de tamaño 70.

2. En el archivo `tabla_nutricional.csv` se encuentra una tabla del contenido nutricional para un total de 2700 calorías diarias.

   a) Leer los datos de la tabla. Completar las celdas faltantes con ceros. Poner todos los datos en la misma unidad.

   b) Realizar un Análisis en Componentes Principales (ACP), tomando el valor nutricional de cada alimento por gramo. ¿Cuántas componentes principales eligiría para reducir la dimensión del conjunto de datos original?

   c) Graficar cada alimento como un punto en ejes coordenados de un nuevo espacio creado por los autovectores generadores del ACP.

   d) En el gráfico realizado en el ítem anterior, analizar si se visualizan agrupamientos de los alimentos.

3. Los exoplanetas son planetas fuera del Sistema Solar. El primero de este tipo fue descubierto en 1995 por Mayor y Queloz (1995). El planeta, similar en masa a Júpiter, se encontró orbitando una estrella relativamente ordinaria, 51 Pegasus. En el período intermedio se han descubierto más de cien exoplanetas, casi todos detectados indirectamente, utilizando la influencia gravitacional que ejercen sobre sus estrellas centrales asociadas. Las propiedades de los exoplanetas encontradas hasta ahora parecen desafiar la teoría del desarrollo planetario construida para los planetas del sistema solar.

   Los exoplanetas no se parecen en nada a los nueve planetas locales que conocemos tan bien. Un primer paso en el proceso de comprensión de los exoplanetas podría ser tratar de agruparlos con respecto a sus propiedades conocidas y este será el objetivo en este ejercicio. En el paquete `HSAUR2` de **R** mediante el comando `data("planets")` se carga el conjunto de datos que contiene la masa (en Júpiter masa, mass), el período (en días terrestres, period) y la excentricidad (eccen) de los exoplanetas descubiertos hasta octubre de 2002.

   a) Cuando las variables están en escalas muy diferentes, se necesitará usar alguna forma de estandarización. Realizar un gráfico tridimensional usando las observaciones escaladas.

   b) Aplicar, a los datos escalados, el comando `kmeans` usando 4 centroides (de manera aleatoria) y obtener el valor mínimo de la función objetivo. Inspeccionar visualmente los agrupamientos generados.

   c) Repetir el ítem anterior varias veces y comparar.

   d) Ejecutar el método K-medias eligiendo los centroides a partir de algún criterio.

   e) Aplicar el método K-medoides utilizando el comando `pam` del paquete `cluster`.

   f) Utilizar el método del codo para identificar la cantidad de clusters en cada uno de los métodos. ¿Cuántos grupos le sugiere?

   g) Utilizar el método Silhouette para identificar la cantidad de clusters en cada uno de los métodos. ¿Cuántos grupos le sugiere?

   h) Concluir.