

# Seminario de Aplicaciones Actuariales

## Seminario de Estadística I

### Aplicaciones de Ciencia de Datos con Python

Profesor: Dr. Arrigo Coen Coria

Ayudante: Act. Miriam Colín

### Tarea 3: Árboles de Decisión

Instrucciones:

- La entrega será el **viernes 3 de septiembre**. Puede ser de manera individual o en equipos de a lo más 3 alumnos.
- Las preguntas 1-3 se entregarán en un pdf con el nombre:  
T1\_ApPatNom1\_ ApPatNom2\_ ApPatNom3\_1\_3
- Cada una de las preguntas 4-5 se entregarán en un jupyter notebook con los nombres:  
T1\_ApPatNom1\_ ApPatNom2\_ ApPatNom3\_4,  
T1\_ApPatNom1\_ ApPatNom2\_ ApPatNom3\_5, respectivamente.
- Responde las siguientes preguntas y realiza lo que se pide.

1. Explica los siguientes conceptos:

- Árbol de decisión
- Bosque aleatorio
- Ada Boost para bosques aleatorios

Para esto da ejemplos de su aplicación y escribe sus ventajas y desventajas al implementarlo.

2. Explica cómo se calcula la impureza de Gini para los nodos y subnodos (incluye la explicación de qué se hace con las columnas nominales y ordinales).
3. Calcula la impureza de Gini y forma un árbol con los datos del archivo "Heart\_Disease.csv" para clasificar si el paciente tiene o no una enfermedad cardíaca.

4. Con la base de datos del archivo *Placement\_Data\_Full\_Class.csv*  
<https://www.kaggle.com/benroshan/factors-affecting-campus-placement>:
  - a) Realizar un análisis completo de los datos (describir/interpretar variables, gráficas, resultados, ...)
  - b) Predecir el status de la persona (variable *status*) utilizando el modelo de árboles de decisiones y de bosques aleatorios para clasificación. Compara los resultados de cada método y di por que elegirías uno.
  - c) Escribir conclusiones
5. Con la base de datos del archivo *data.csv*  
<https://www.kaggle.com/bricevergnou/spotify-recommendation>:
  - a) Realizar un análisis completo de los datos (describir/interpretar variables, gráficas, resultados, ...)
  - b) Predecir si la canción es gustada o no (variable *liked*) utilizando el modelo de Random Forest y Ada Boost para Random Forest. Compara los resultados de cada método y di por que elegirías uno.
  - c) Escribir conclusiones