**Tarea 1 de TÉCNICAS DE CLASIFICACIÓN**

**MASTER EN DATA SCIENCE. CUNEF**

**Notas**: Se deben incluir las respuestas como un fichero pdf o Word, incluyendo como anexo el script del programa R correspondiente a la resolución de cada una de las partes de la tarea. También se permitirá intercalar el código R y el análisis e interpretación de resultados en cada apartado. En el documento de respuesta escrita se debe incluir la información necesaria con los resultados fundamentales, gráficos, tablas, etc. así como el **análisis e interpretación de los mismos**. En cuanto a este fichero, debe remitir como archivo único a la tarea que se ha habilitado al efecto y cuya hora de entrega finaliza el 10 de noviembre a las 23:59.

**Enunciado**: Una empresa del sector financiero pretende utilizar técnicas de clasificación en la selección preliminar de solicitudes de crédito. De registros anteriores la empresa dispone de información sobre 3 tipos de adjudicatarios de créditos: a) alto riesgo y b) riesgo medio y c) bajo riesgo. También se dispone de información adicional que se recoge en la tabla con el siguiente encabezado.

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **TIPO** | **I=Ingreso anual (miles de euros)** | **Edad** | **Sexo** | **EC=Estado civil** | **H=Hijos (número)** | **P=Patrimonio (miles de euros)** | **R=Ratio endeudamiento sobre patrimonio (%)** | **A=Grado de aversión al riesgo** |

Se pide:

1. Analizar descriptivamente cada una de las variables en cada uno de los dos grupos. Calcular medidas más relevantes, comprobar normalidad y comentar los principales resultados.
2. Aplicar un análisis discriminante a este caso justificando la selección de variables discriminantes. Interpretar la función discriminante y evaluar la importancia que tiene cada una de las variables discriminantes. Evaluar la capacidad predictiva del modelo.
3. Aplicar un árbol de clasificación a este caso justificando la selección de variables predictoras, utilizando una muestra del 80% para la estimación y del 20 % para la muestra de validación (seed=1234). Interpretar el árbol o árboles y evaluar la importancia que tiene cada una de las variables predictoras. Evaluar la capacidad predictiva del modelo.