

## Taller 2 Árboles de Decisión

- 1 Ajustar un modelo para el conjunto de datos Saber Pro. El conjunto de datos contiene información de una población de estudiantes que presentaron los exámenes saber Pro con la cual se intenta predecir si su nota esta por encima de la media o no.
  - 1.1.0 Prepare el conjunto de datos elimine los valores incompletos para este conjunto de datos y transforme las variables categóricas a dummies.
  - 1.1.1 Separe las poblaciones en entrenamiento y validación en forma aleatoria en 70/30 (genere una semilla aleatoria e indique el valor de la semilla en el software que utilizó). Indique que cantidad de casos quedaron para cada ambiente.
  - 1.1.2 Ajuste el mejor modelo posible explotando los parámetros del árbol .
  - 1.1.3 Visualizar el árbol generado y explicar dos reglas de este que mas le llamen la atención.
  - 1.1.4 Obtener la matriz de confusión tanto para el ambiente de entrenamiento como de validación.
  - 1.1.5 Calcular la accuracy en ambiente de entrenamiento y validación.
  - 1.1.6 Calcular el AUC y el gráfico ROC en entrenamiento y validación.
  - 1.1.7 Utilizar el conjunto de entrenamiento calcule el promedio de AUC y accuracy con el método de validación cruzada con 10 folds o particiones.
  - 1.1.8 Selecciones el 25% de los individuos en el ambiente de validación de acuerdo a la siguiente lógica. Entregue los resultados indicados:
    - Al azar e indique la cantidad de individuos que están por encima de la media.
    - Utilizando el modelo desarrollado en el punto 1.1.2 e indique la cantidad de individuos que están por encima de la media.
  - 1.1.9 Lea el conjunto de datos Saber\_pro\_predecir.csv utilice el modelo generado para predecir los registros de este archivo, genere archivo.csv con los resultados de las predicciones de su modelo y enviarlo por vía mail.