Laboratorio Selección y Calibración de modelos

Ejercicio 1: Esta pregunta utiliza los datos sobre muerte del corazón en Sudáfrica (SAheart.csv). La variable que queremos predecir es chd que es un indicador de muerte coronaria basado en algunas variables predictivas (factores de riesgo) como son el fumado, la obesidad, las bebidas alcohólicas, entre otras.

a) El objetivo de este ejercicio es calibrar el método de Potenciación. Aquí interesa predecir el "Si" en la variable chd, para esto genere 10 Validaciones Cruzadas con 6 grupos calibrando el modelo de acuerdo con los tres tipos de algoritmos que permite, discrete, real y gentle, para medir la calidad de método sume la cantidad de Si detectados en los diferentes grupos. Luego grafique las 10 iteraciones para los tres algoritmos en el mismo gráfico. ¿Se puede determinar con claridad cuál algoritmo es el mejor? Para generar los modelos predictivos use las siguientes instrucciones:

```
modelo<ada(chd~.,data=taprendizaje,iter=20,nu=1,type="discr
ete")

modelo<- ada(chd~.,data=taprendizaje,iter=20,nu=1,
type="real")

modelo<-ada(chd~.,data=taprendizaje,iter=20,nu=1,
type="gentle")
```

b) Repita el ejercicio anterior, pero esta vez en lugar de sumar los "Si" detectados, promedie los errores globales cometidos en los diferentes grupos (folds). Luego grafique las 10 iteraciones para los tres algoritmos en el mismo gráfico. ¿Se puede determinar con claridad cuál algoritmo es el mejor?

c) ¿Cuál algoritmo usaría con base en la información obtenida en los dos ejercicios anteriores?
d) El objetivo de esta etapa es comparar todos los métodos predictivos vistos en el curso con esta tabla de datos. Aquí interesa predecir el "Si" en la variable chd, para esto genere 10 Validaciones Cruzadas con 6 grupos para los métodos SVM, KNN, Bayes, Árboles, Bosques, Potenciación y Redes Neuronales. Luego grafique las 10 iteraciones para todos los métodos en el mismo gráfico. ¿Se puede determinar con claridad cuál métodos es el mejor?

Ejercicio 2: **(Reto Predictivo)** En este ejercicio usted tiene una tabla de datos Seguros.csv con información sobre fraudes en seguros, esta tabla tiene 16 variables y 6413 casos, se trata de predecir la variable Fraude que indica si hubo o no fraude. Este ejercicio es un verdadero reto predictivo ya que se trata de un problema muy desbalanceado, se tienen 6146 no fraudes y a penas 267 fraudes, esto hace que sea muy difícil el aprendizaje para cualquier modelo.

Para este ejercicio usted recibe además el archivo SegurosNuevosVE150.csv en el cual la variable Fraude viene con un NA para todos sus registros. El reto consiste en predecir para este archivo los valores de la variable Fraude, para esto haga lo siguiente:

a) Usando Validación Cruzada determine cuál de los modelos estudiados en el curso funciona mejor para estos datos, antes debe calibrar los modelos, por ejemplo, para Redes Neuronales debe determinar el Números de Capas Ocultas, para Árboles debe determinar La Profundidad Máxima, etc.
b) Incluya una explicación del procedimiento seguido, el método escogido y la justificación de su elección. Debe entregar el archivo con las predicciones.