Profesor: Dr. Oldemar Rodríguez Rojas
Minería de Datos 1
Fecha de Entrega: Jueves 15 de junio - 8am
Notas Finales: Antes del 29 de junio aparecerán en el Aula Virtual
Instrucciones:

- Las tareas serán revisadas en clase, no pueden ser enviadas por correo.
- Las tareas son estrictamente individuales.
- Tareas idénticas se les asignará cero puntos.
- Todas las tareas tienen el mismo valor en la nota final del curso.

Tarea Número 13

- **Ejercicio 1:** [40 puntos] Esta pregunta también utilizan nuevamente los datos `tumores.csv`. El objetivo de este ejercicio es comparar todos los métodos predictivos vistos en el curso con esta tabla de datos. Aquí interesa predecir en la variable `tipo`, para los métodos `SVM, KNN, Árboles, Bosques, Potenciación, eXtreme Gradient Boosting, LDA, Bayes, Redes Neuronales, KNN` y `Potenciación` se desea determinar cuál método produce mejores resultados usando la curva ROC. Para esto realice lo siguiente:

  1. Grafique la curva ROC para todos modelos predictivos generados en la tarea anterior. Use el parámetro `type = ''prob''` en las funciones `predict` del paquete `traineR` para retornar la probabilidad en lugar de la clase. ¿Cuál método parece ser mejor?
  2. Calcule en área bajo curva ROC para todos los modelos predictivos generados en la tarea anterior. ¿Cuál método es el mejor según este criterio?

- **Ejercicio 2:** [30 puntos] Dada la siguiente tabla:

Individuo	Clase	Score
7	P	0.61
8	N	0.06
1	N	0.80
2	P	0.11
3	N	0.66
6	N	0.46
4	P	0.40
10	N	0.19
5	N	0.00
9	P	0.91

  1. Usando la definición de curva ROC calcule y grafique "a mano" la curva ROC, use un umbral $T = 0$ y un paso de 0,05. Es decir, debe hacerlo variando el umbral y calculando las matrices de confusión.
  2. Verifique el resultado anterior usando el código visto en clase.

3. Usando el algoritmo eficiente para la curva ROC calcule y grafique "a mano" la curva ROC, use un umbral $T = 0{,}1$ y un paso de 0,1.
4. Verifique el resultado anterior usando el código visto en clase para el algoritmo eficiente.

- **Ejercicio 3:** [30 puntos] Esta pregunta utiliza los datos `SAheart.csv` sobre muerte del corazón en Sudáfrica. 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. Las variables son:

  - `sbp`: systolic blood pressure (numérica).
  - `tobacco`: cumulative tobacco (kg) (numérica).
  - `ldl`: low densiity lipoprotein cholesterol (numérica).
  - `Adiposity`: Adiposity level (numérica).
  - `famhist`: family history of heart disease (Present, Absent) (categórica).
  - `typea`: type-A behavior (numérica).
  - `Obesity`: Obesity of the person (numérica).
  - `alcohol`: current alcohol consumption (numérica).
  - `age`: age at onset (numérica).
  - `chd`: coronary heart disease (categórica).

  Realice lo siguiente:

  1. Usando todos los métodos vistos en el curso, para la tabla `SAheart.csv` con el 80 % de los datos para la tabla aprendizaje y un 20 % para la tabla testing determine la mejor **Probabilidad de Corte** para cada uno de estos métodos, de forma tal que se prediga de la mejor manera posible la categoría `Si` de la variable `chd`, pero sin desmejorar de manera significativa la precisión global.
  2. El objetivo de este ejercicio es comparar todos los métodos predictivos vistos en el curso con esta tabla de datos. Aquí interesa predecir en la variable `chd`, Usando los paquetes `snow` y `traineR` **programe en paralelo** 3 Validaciones Cruzadas con 10 grupos para los métodos `SVM, KNN, Árboles, Bosques, Potenciación, eXtreme Gradient Boosting, LDA, Bayes` y `Redes Neuronales`, `KNN` y `Potenciación`, **para esto use en cada método la probabilidad de corte obtenida en el ejercicio anterior** (puede usar los parámetros por defecto en cada método). Luego grafique las 5 iteraciones para todos los métodos en el mismo gráfico. ¿Se puede determinar con claridad cuál métodos es el mejor?
  3. Además de calibrar cada uno de los métodos ¿Cómo se podría mejorar la predicción en el ejercicio anterior?