Profesor: Dr. Oldemar Rodríguez Rojas
Minería de Datos 1
Fecha de Entrega: Jueves 1 de junio - 8am
Instrucciones:

Instrucciones:

- Las tareas deben ser subida la Aula Virtual antes de las 12 media noche. Luego de esta hora pierde 20 puntos y cada día de retraso adicional perderá 20 puntos más.
- 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.
- Cada día de entrega tardía tendrá un rebajo de 20 puntos.

Tarea Número 11

- **Ejercicio 1:** [25 puntos] En esta pregunta utiliza los datos (`tumores.csv`). Se trata de un conjunto de datos de características del tumor cerebral que incluye cinco variables de primer orden y ocho de textura y cuatro parámetros de evaluación de la calidad con el nivel objetivo. La variables son: Media, Varianza, Desviación estándar, Asimetría, Kurtosis, Contraste, Energía, ASM (segundo momento angular), Entropía, Homogeneidad, Disimilitud, Correlación, Grosor, PSNR (Pico de la relación señal-ruido), SSIM (Índice de Similitud Estructurada), MSE (Mean Square Error), DC (Coeficiente de Dados) y la variable a predecir `tipo` (1 = Tumor, 0 = No-Tumor).

  1. Cargue la tabla de datos `tumores.csv` en **R** y ejecute un `str(...)`, `summary(...)` y un `dim(...)`, verifique la correcta lectura de los datos.
  2. El objetivo de este ejercicio es analizar la variación del error (usando el enfoque training-testing) para la predicción de variable `tipo` (que indica 1 = Tumor, 0 = No-Tumor), para esto repita 5 veces el cálculo de error global de predicción usando el método de los k vecinos más cercanos (use `kmax=50`) y con un 75 % de los datos para tabla aprendizaje y un 25 % para la tabla testing. Grafique los resultados.
  3. El objetivo de este ejercicio es medir el error para la predicción de variable `tipo`, utilizando validación cruzada con K grupos ($K-$fold cross-validation). Para esto usando el método de los k vecinos más cercanos (use `kmax=50`) realice una validación cruzada 5 veces con 10 grupos (folds) y grafique el error obtenido en cada iteración, agregue en este gráfico los 5 errores generados en el ejercicio anterior.
  4. ¿Qué se puede concluir?

- **Ejercicio 2:** [25 puntos] Para esta pregunta también usaremos los datos `tumores.csv`.

  1. El objetivo de este ejercicio es calibrar el método de `ADA` para esta Tabla de Datos. Aquí interesa predecir en la variable `tipo`. Para esto genere 5 Validaciones Cruzadas con 10 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 1's detectados en los diferentes grupos. Luego grafique las 5 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<-train.ada(tipo~.,data=taprendizaje,iter=80,nu=1,type="discrete")
modelo<-train.ada(tipo~.,data=taprendizaje,iter=80,nu=1,type="real")
modelo<-train.ada(tipo~.,data=taprendizaje,iter=80,nu=1,type="gentle")
```

2. Repita el ejercicio anterior, pero esta vez en lugar de sumar la cantidad de 1's, promedie los errores globales cometidos en los diferentes grupos (folds). Luego grafique las 5 iteraciones para los tres algoritmos en el mismo gráfico. ¿Se puede determinar con claridad cuál algoritmo es el mejor?
3. Para estar realmente seguros de cuál de los tipos algoritmos es mejor, modifique (una en uno solo código) los códigos de los dos ejercicios anteriores de manera que, en lugar de sumar la cantidad de 1's detectados y de promediar los errores globales, guarde en cada iteración (en listas) las matrices de confusión de cada uno de los 3 algoritmos. Luego al final de los ciclos ejecutados para cada algoritmo calcule una Matriz de Confusión Promedio de todas las matrices de confusión guardadas en la lista del respectivo algoritmo; así con estas Matrices de Confusión Promedio, mediante gráficos de barras, determine el método que en promedio detecta mayor porcentaje de 1's, determine el método que en promedio detecta mayor porcentaje de 0's y determine el método que en promedio tiene menor error global.
4. ¿Cuál algoritmo usaría con base en la información obtenida en el ejercicio anterior?

- **Ejercicio 3:** [25 puntos] Para esta pregunta usaremos nuevamente los datos `tumores.csv`.

1. El objetivo de este ejercicio es calibrar el método de `kknn` para esta Tabla de Datos. Aquí interesa predecir en la variable `tipo`. Para esto genere 5 Validaciones Cruzadas con 10 grupos calibrando el modelo de acuerdo con todos los tipos de algoritmos que permite `train.kknn` en el parámetro `kernel`, estos algoritmos son: `rectangular`, `triangular`, `epanechnikov`, `biweight`, `triweight`, `cos`, `inv`, `gaussian` y `optimal`. Para medir la calidad de método sume la cantidad de 1's detectados en los diferentes grupos. Luego grafique las 5 iteraciones para todos algoritmos en el mismo gráfico. ¿Se puede determinar con claridad cuál algoritmo es el mejor?
2. Repita el ejercicio anterior, pero esta vez en lugar de sumar la cantidad de 1's, promedie los errores globales cometidos en los diferentes grupos (folds). Luego grafique las 5 iteraciones para todos los algoritmos en el mismo gráfico. ¿Se puede determinar con claridad cuál algoritmo es el mejor?
3. Para estar realmente seguros de cuál de los tipos algoritmos es mejor, modifique (una en uno solo código) los códigos de los dos ejercicios anteriores de manera que, en lugar de sumar la cantidad de 1's detectados y de promediar los errores globales, guarde en cada iteración (en listas) las matrices de confusión de cada uno de los 9 algoritmos. Luego al final de los ciclos ejecutados para cada algoritmo calcule una Matriz de Confusión Promedio de todas las matrices de confusión guardadas en la lista del respectivo algoritmo; así con estas Matrices de Confusión Promedio, mediante gráficos de barras, determine el método que en promedio detecta mayor porcentaje de 1's, determine el método que en promedio

detecta mayor porcentaje de 0's y determine el método que en promedio tiene menor error global.

4. ¿Cuál algoritmo usaría con base en la información obtenida en el ejercicio anterior?

- **Ejercicio 4:** [25 puntos] Esta pregunta también utilizan nuevamente los datos `tumores.csv`.

1. 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 esto genere 5 Validaciones Cruzadas con 10 grupos para los métodos `SVM, KNN, Árboles, Bosques, Potenciación, eXtreme Gradient Boosting y Redes Neuronales`, para `KNN` y `Potenciación` use los parámetros obtenidos en las calibraciones realizadas en los ejercicios anteriores. Luego grafique las 5 iteraciones para todos los métodos en el mismo gráfico. ¿Se puede determinar con claridad cuál métodos es el mejor?
2. Para estar realmente seguros de cuál de los métodos es mejor, modifique (una en uno solo código) los códigos de los dos ejercicios anteriores de manera que, en lugar de sumar la cantidad de 1's detectados y de promediar los errores globales, guarde en cada iteración (en listas) las matrices de confusión de cada uno de los 11 métodos. Luego al final de los ciclos ejecutados para cada algoritmo calcule una Matriz de Confusión Promedio de todas las matrices de confusión guardadas en la lista del respectivo algoritmo; así con estas Matrices de Confusión Promedio, mediante gráficos de barras, determine el método que en promedio detecta mayor porcentaje de 1's, determine el método que en promedio detecta mayor porcentaje de 0's y determine el método que en promedio tiene menor error global.
3. ¿Cuál algoritmo usaría con base en la información obtenida en el ejercicio anterior?