Profesor: Dr. Oldemar Rodríguez Rojas
Minería de Datos 1
Fecha de Entrega: Jueves 4 de mayo - 8am
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 7

- **Pregunta 1:** [30 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).

  Realice lo siguiente:

  1. Cargue la tabla de datos `tumores.csv` en `R` y genere en **R** usando la función `createDataPartition(...)` del paquete `caret` la tabla de testing con una 25 % de los datos y con el resto de los datos genere una tabla de aprendizaje.
  2. Usando Máquinas de Soporte Vectorial, con todos los núcleos (kernel) (en `traineR`) genere un modelos predictivos para la tabla de aprendizaje.
  3. Construya una tabla para los índices anteriores que permita comparar el resultado de Máquinas de Soporte Vectorial con respecto a los métodos generados en las tareas anteriores ¿Cuál método es mejor?

- **Pregunta 2:** [30 puntos] Esta pregunta utiliza los datos sobre la conocida historia y tragedia del Titanic, usando los datos `titanicV2020.csv` de los pasajeros se trata de predecir la supervivencia o no de un pasajero.

  La tabla contiene 12 variables y 1309 observaciones, las variables son:

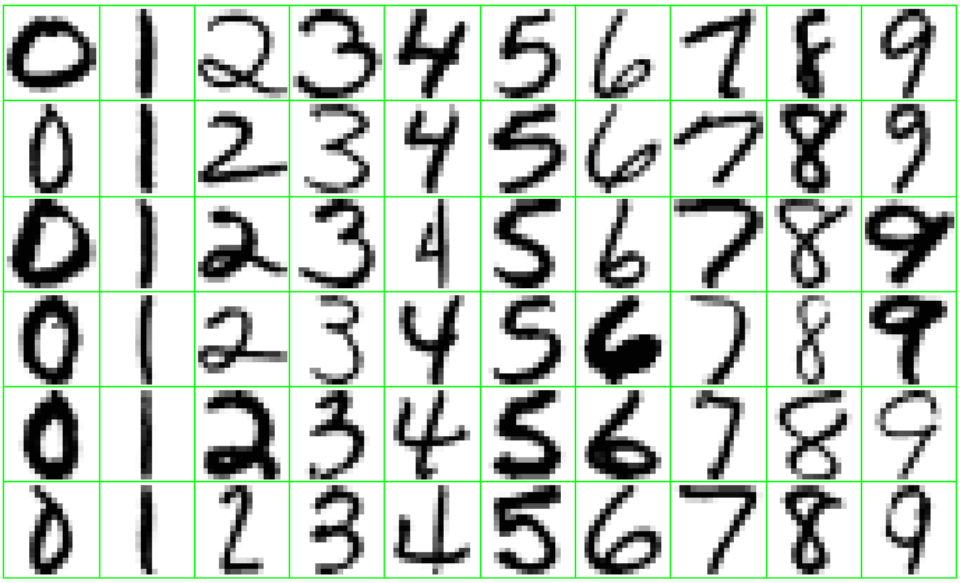
  - `PassegerId`: El código de identificación del pasajero (valor único).
  - `Survived`: Variable a predecir, 1 (el pasajero sobrevivió) 0 (el pasajero no sobrevivió).
  - `Pclass`: En que clase viajaba el pasajero (1 = primera, 2 = segunda , 3 = tercera).
  - `Name`: Nombre del pasajero (valor único).
  - `Sex`: Sexo del pasajero.

- `Age`: Edad del pasajero.
- `SibSp`: Cantidad de hermanos o cónyuges a bordo del Titanic.
- `Parch`: Cantidad de padres o hijos a bordo del Titanic.
- `Ticket`: Número de tiquete (valor único).
- `Fare`: Tarifa del pasajero.
- `Cabin`: Número de cabina (valor único).
- `Embarked`: Puerto donde embarco el pasajero (C = Cherbourg, Q = Queenstown, S = Southampton).

Realice lo siguiente:

1. Cargue la tabla de datos `titanicV2020.csv`, asegúrese re-codificar las variables cualitativas y de ignorar variables que no se deben usar.
2. Usando el comando `sample` de **R** genere al azar una tabla aprendizaje con un 80 % de los datos y con el resto de los datos genere una tabla de aprendizaje.
3. Genere un Modelo Predictivo usando SVM, con el paquete `traineR`, luego para este modelo calcule la matriz de confusión, la precisión, la precisión positiva, la precisión negativa, los falsos positivos, los falsos negativos, la acertividad positiva y la acertividad negativa. Utilice el Kernel que dé mejores resultados.
4. Construya una tabla para los índices anteriores que permita comparar el resultado de los métodos SVM con respecto a los métodos de las tareas anteriores ¿Cuál método es mejor?

- **Pregunta 3:** [30 puntos] En este ejercicio vamos a predecir números escritos a mano (Hand Written Digit Recognition), la tabla de de datos está en el archivo `ZipData_2020.csv`. En la figura siguiente se ilustran los datos:



Los datos de este ejemplo vienen de los códigos postales escritos a mano en sobres del correo postal de EE.UU. Las imágenes son de 16×16 en escala de grises, cada píxel va de intensidad de -1 a 1 (de blanco a negro). Las imágenes se han normalizado para tener aproximadamente el mismo tamaño y orientación. La tarea consiste en predecir, a partir de la matriz de 16×16 de intensidades de cada píxel, la identidad de cada imagen $(0, 1, \ldots, 9)$ de forma rápida y precisa. Si es lo suficientemente precisa, el algoritmo resultante se utiliza como parte de un procedimiento de selección automática para sobres. Este es un problema de clasificación para el cual la tasa de error debe mantenerse muy baja para evitar la mala dirección de correo. La columna 1 tiene la variable a predecir `Número` codificada como sigue: 0=‘cero’; 1=‘uno’; 2=‘dos’; 3=‘tres’; 4=‘cuatro’; 5=‘cinco’;6=‘seis’; 7=‘siete’; 8=‘ocho’ y 9=‘nueve’, las demás columnas son las variables predictivas, además cada fila de la tabla representa un bloque 16×16 por lo que la matriz tiene 256 variables predictivas.

Para esto realice lo siguiente (podría tomar varios minutos los cálculos):

1. Cargue la tabla de datos `ZipData_2020.csv` en **R**.
2. Use el método de Máquinas de Soporte Vectorial con el núcleo y los parámetros que usted considere más conveniente para generar un modelo predictivo para la tabla `ZipData_2020.csv` usando el 80 % de los datos para la tabla aprendizaje y un 20 % para la tabla testing, luego calcule para los datos de testing la matriz de confusión, la precisión global y la precisión para cada una de las categorías. ¿Son buenos los resultados? Explique.
3. Compare los resultados con los obtenidos en las tareas anteriores.

- **Pregunta 4:** [10 puntos] Suponga que se tiene la siguiente tabla de datos:

X	Y	Z	Clase
1	0	1	Rojo
1	0	2	Rojo
1	1	2	Rojo
3	1	4	Rojo
1	1	3	Rojo
3	2	3	Azul
1	2	1	Azul
3	2	1	Azul
1	1	0	Azul

Puede observar los puntos con el siguiente código:

```
library(plotly)

datos <- data.frame(x = c(1, 1, 1, 3, 1, 3, 1, 3, 1),
                    y = c(0, 0, 1, 1, 1, 2, 2, 2, 1),
                    z = c(1, 2, 2, 4, 3, 3, 1, 1, 0),
                    clase = c("Rojo", "Rojo", "Rojo",
                              "Rojo", "Rojo", "Azul",
                              "Azul",  "Azul", "Azul"))
```

```
plot_ly(data = datos) %>%
  add_trace(x = ~x, y = ~y, z = ~z, color = ~clase,
            colors = c("#0C4B8E", "#BF382A"),
            mode = "markers", type = "scatter3d")
```

Realice lo siguiente:

1. Dibuje con colores los puntos de ambas clases en $\mathbb{R}^3$.
2. Dibuje el hiperplano óptimo de separación e indique la ecuación de dicho hiperplano de la forma $ax + by + cz + d = 0$. Nota: Se debe observar con detenimiento los puntos de ambas clases para encontrar los vectores de soporte de cada margen y trazar con estos puntos los hiperplanos de los márgenes luego trazar el hiperplano de soporte justo en el centro.
3. Escriba la regla de clasificación para el clasificador con margen máximo. Debe ser algo como lo siguiente: $w = (w_1, w_2, w_3)$ se clasifica como *Rojo* si $ax + by + cz + d > 0$ y otro caso se clasifica como *Azul*.
4. Indique el margen para el hiperplano óptimo y los vectores de soporte.
5. Explique por qué un ligero movimiento de la octava observación no afectaría el hiperplano de margen máximo.
6. Dibuje un hiperplano que no es el hiperplano óptimo de separación y proporcione la ecuación para este hiperplano.
7. Dibuje un hiperplano de separación pero que no es el hiperplano óptimo de separación, y escriba la ecuación correspondiente.
8. Dibuje una observación adicional de manera que las dos clases ya no sean separables por un hiperplano.