Profesor: Dr. Oldemar Rodríguez Rojas
Análisis de Datos 1
Fecha de Entrega: Viernes 21 de junio - 8 a.m.
Instrucciones:

- Las tareas deben ser subida la Aula Virtual antes de las 8:00 a.m. Luego de esta hora pierde 10 puntos y cada día de retraso adicional perderá 10 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 10 puntos.

Tarea Número 12

- **Pregunta 1:** [10 puntos] Supongamos que se tiene una nueva fila o registro $12 = (1, 3, 2, 4, ?)$ en la base de datos de la filmina 18 en la presentación del método Bayes, prediga (a mano) si el individuo corresponde a un buen pagador o a un mal pagador.
- **Pregunta 2:** [10 puntos] Supongamos que se tiene una nueva fila o registro $t = (\text{Pedro}, M, 4, ?)$ en la base de datos de la filmina 23 en la presentación del método Bayes, prediga (a mano) si Pedro corresponde a la clase pequeño, mediano o alto.
- **Pregunta 3:** [20 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 Naïve Bayes, LDA y QDA genere un modelos predictivos para la tabla de aprendizaje, puede ser que LDA y QDA generen errores.
  3. Para la tabla de testing calcule la matriz de confusión, la precisión global, el error global y la precisión en cada de las clases. Construya una tabla para los índices anteriores que permita comparar los resultados de Naïve Bayes, LDA y QDA con respecto a los métodos generados en las tareas anteriores ¿Cuál método es mejor?
- **Pregunta 4:** [20 puntos] Para la presentación del Análisis Discriminante complete la demostraciones que quedaron pendientes.

- **Pregunta 5:** [20 puntos] Esta pregunta utiliza los datos sobre la conocida historia y tragedia del Titanic, usando los datos `titanicV2024.csv` de los pasajeros se trata de predecir la supervivencia o no de un pasajero.

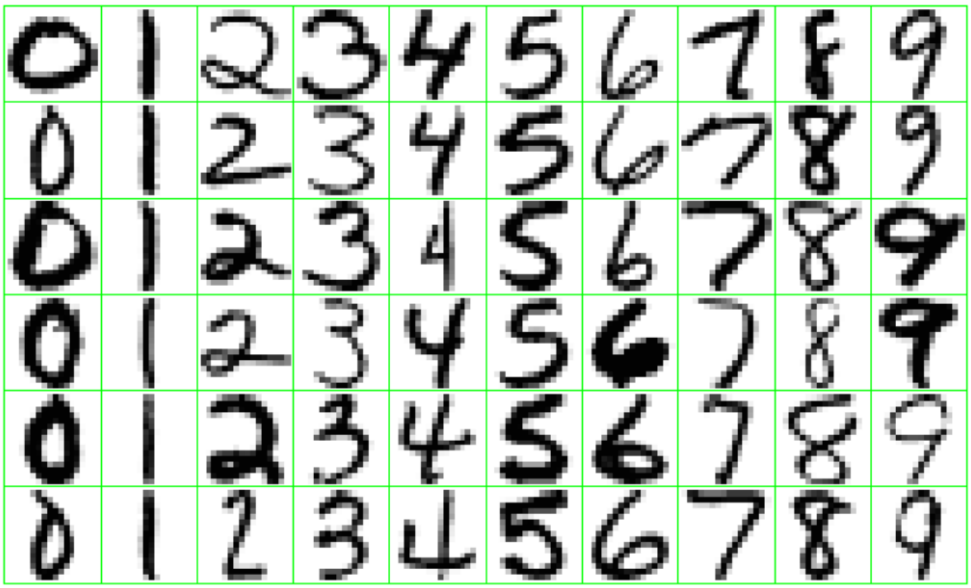
  La tabla contiene 12 variables y 1309 observaciones, las variables son:

  - `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 `titanicV2024.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 Naïve Bayes, 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.
  4. Genere Modelos Predictivos usando LDA y QDA, con el paquete `MASS`, 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, puede ser que QDA falle.
  5. Construya una tabla para los índices anteriores que permita comparar el resultado de los métodos Bayes, LDA y QDA con respecto a los métodos de las tareas anteriores ¿Cuál método es mejor?

- **Pregunta 6:** [20 puntos] En este ejercicio vamos a predecir números escritos a mano (Hand Written Digit Recognition). Unifique las tablas de datos de la tarea anterior `ZipDataTrainCod.csv` y `ZipDataTestCod.csv` en un solo archivo, esto para tener una sola tabla de de datos, nombre este archivo como `ZipData_2024.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_2024.csv` en **R**.
2. Usando KNN, Naïve Bayes, LDA y QDA genere un modelos predictivos y los parámetros que usted considere más conveniente para generar un modelo predictivo para la tabla `ZipData_2024.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.