Profesor: Dr. Oldemar Rodríguez Rojas
Minería de Datos 1
Fecha de Entrega: Jueves 27 de abril - 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 6

- **Ejercicio 1:** [20 puntos] Dada la siguiente Tabla de Testing de un Scoring de Crédito:

MontoCredito	IngresoNeto	CoefCreditoAvaluo	MontoCuota	GradoAcademico	BuenPagador	PrediccionKNN
30690	2	11	Medio	Bachiller	No	Si
27110	2	11	Medio	Licenciatura	Si	Si
12804	2	11	Medio	Bachiller	No	No
13512	2	11	Bajo	Licenciatura	Si	Si
14077	1	9	Bajo	Licenciatura	Si	Si
118143	2	12	Medio	Bachiller	Si	Si
26577	2	5	Alto	Licenciatura	Si	Si
28088	1	1	MuyBajo	Bachiller	Si	No
51366	1	12	Alto	Licenciatura	No	No
287668	1	12	Bajo	Bachiller	No	No
29842	2	5	Alto	Licenciatura	Si	Si
45385	1	12	Medio	Bachiller	Si	Si
38131	1	4	Bajo	Licenciatura	Si	Si
39958	1	12	Alto	Licenciatura	No	No
33277	2	11	Alto	Licenciatura	Si	Si
53501	2	1	Alto	Licenciatura	Si	Si
19366	2	11	Medio	Licenciatura	Si	Si
12867	1	10	Medio	Licenciatura	Si	Si
40125	2	12	Medio	Bachiller	Si	No
12722	1	12	Bajo	Bachiller	Si	Si
12771	2	12	Alto	Licenciatura	Si	Si
18407	1	11	Medio	Licenciatura	Si	Si
32537	1	10	Bajo	Bachiller	No	Si
48598	1	12	MuyBajo	Bachiller	No	Si
12562	2	11	Medio	Licenciatura	Si	Si

1. Usando la columna `BuenPagador` en donde aparece el verdadero valor de la variable a predecir y la columna `PrediccionKNN` en donde aparece la predicción del Método `KNN` para esta tabla de Testing, calcule la Matriz de Confusión.
2. Con la Matriz de Confusión anterior calcule "a mano" la Precisión Global, el Error Global, la Precisión Positiva (PP), la Precisión Negativa (PN), la Proporción de Falsos Positivos (PFP), la Proporción de Falsos Negativos (PFN), la Asertividad Positiva (AP) y la Asertividad Negativa (AN).

- **Ejercicio 2:** [20 puntos] Resuelva lo siguiente:

1. Programe en lenguaje **R** una función que reciba como entrada la matriz de confusión (para el caso 2×2) que calcule y retorne en una lista: la Precisión Global, el Error Global, la Precisión Positiva (PP), la Precisión Negativa (PN), los Falsos Positivos (FP), los Falsos Negativos (FN), la Asertividad Positiva (AP) y la Asertividad Negativa (NP).

	No	Sí
No	892254	212
Sí	8993	300

2. Supongamos que tenemos un modelo predictivo para detectar Fraude en Tarjetas de Crédito, la variable a predecir es `Fraude` con dos posibles valores `Sí` (para el caso en que sí fue fraude) y `No` (para el caso en que no fue fraude). Supongamos la matriz de confusión es:
   - Calcule la Precisión Global, el Error Global, la Precisión Positiva (PP), la Precisión Negativa (PN), los Falsos Positivos (FP), los Falsos Negativos (FN), la Asertividad Positiva (AP) y la Asertividad Negativa (NP).
   - ¿Es bueno o malo el modelo predictivo? Justifique su respuesta.

- **Ejercicio 3:** [20 puntos] 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. Use el método de K vecinos más cercanos en el paquete **traineR** para generar un modelo predictivo para la tabla `tumores.csv` usando el 75 % de los datos para la tabla aprendizaje y un 25 % para la tabla testing. No olvide recodificar, desde **R**, la variable a predecir como categórica.
  2. Genere un Modelo Predictivo usando K vecinos más cercanos para cada uno de los siguientes núcleos: `rectangular`, `triangular`, `epanechnikov`, `biweight`, `triweight`, `cos`, `inv`, `gaussian` y `optimal` ¿Cuál produce los mejores resultados en el sentido de que predice mejor los tumores, es decir, `Tumor = 1`.

- **Ejercicio 4:** [20 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:

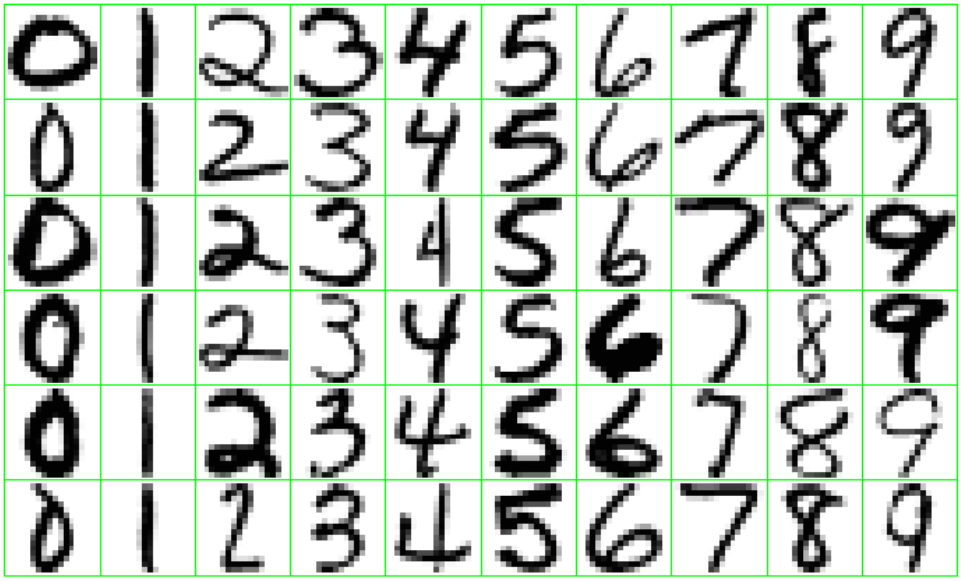
  - `PassegerId`: El código de identificación del pasajero (valor único).
  - `Survived`: Variable a predecir, 1 (el pasajero sobrevivió) 0 (el pasajero no sobrevivió).
  - `Pclass`: En que clase viajaba el pasajero (1 = primera, 2 = segunda , 3 = tercera).
  - `Name`: Nombre del pasajero (valor único).
  - `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. Realice un análisis exploratorio (estadísticas básicas) que incluya: el resumen numérico (media, desviación estándar, etc.), los valores atípicos, la correlación entre las variables, el poder predictivo de las variables predictoras. Interprete los resultados.
3. ¿Es este problema equilibrado o desequilibrado? Justifique su respuesta.
4. Use el método de K vecinos más cercanos en el paquete **traineR**, con los parámetros que logren el mejor resultado, para generar un modelo predictivo con la tabla `titanicV2020.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 dos categorías. ¿Son buenos los resultados? Explique.
5. Repita el item *4)*, pero esta vez, seleccione las 5 variables que, según su criterio, tienen mejor poder predictivo. ¿Mejoran los resultados?
6. Usando la función programada en el ejercicio 1, los datos `titanicV2020.csv` y los modelos generados arriba construya un `DataFrame` de manera que en cada una de las filas aparezca un modelo predictivo y en las columnas aparezcan los índices *Precisión Global, Error Global Precisión Positiva (PP), Precisión Negativa (PN), Falsos Positivos (FP), los Falsos Negativos (FN), la Asertividad Positiva (AP) y la Asertividad Negativa (AN).* ¿Cuál de los modelos es mejor para estos datos?

- **Ejercicio 5:** [20 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. ¿Es este problema equilibrado o desequilibrado? Justifique su respuesta.
3. Use el método de K vecinos más cercanos en el paquete **traineR** (con los parámetros por defecto) 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.

oldemar rodríguez

CONSULTOR en M1NƐR14 DƐ D4T0S