Tarea Número 3

Ejercicio 1: Esta pregunta utiliza los datos sobre muerte del corazón en Sudáfrica (SAheart.csv). La variable a predecir es chd, un indicador de muerte coronaria. La predicción se construirá basándose en algunas variables predictivas (factores de riesgo) como son el ser fumador, la obesidad, el consumo de bebidas alcohólicas, entre otras.

1. Use el método de Árboles de Decisión para generar un modelo predictivo para la tabla SAheart.csv usando el 75% de los datos para la tabla de aprendizaje y un 25% para la tabla de prueba.

2. Calcule para los datos de prueba la matriz de confusión, la precisión global, la precisión positiva, la precisión negativa, los falsos positivos, los falsos negativos.

3. Grafique el Árbol de Decisión generado y las Reglas de Decisión.

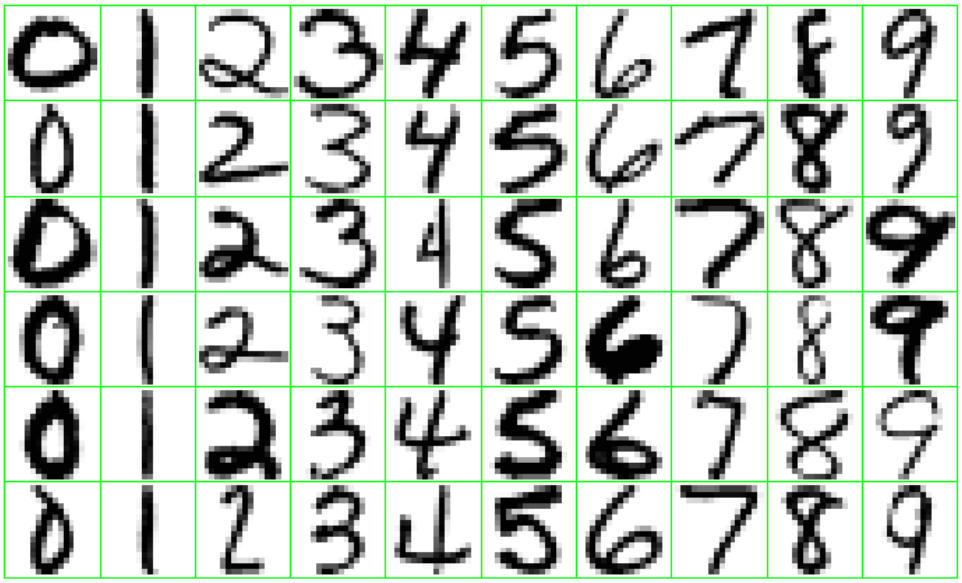
Ejercicio 2: Esta pregunta utiliza los datos "CompraBicicletas.csv", que contienen información de 11 variables (predictoras) y de la variable "PurchasedBike" que da cuenta de si un cliente compró o no una bicicleta.

1. Use el método de Árboles de Decisión para generar un modelo predictivo para la tabla PurchasedBike.csv usando 70% de los datos para tabla de aprendizaje y un 30% para la tabla de prueba.

2. Calcule para los datos de prueba y para toda la tabla la matriz de confusión y la precisión global. Interprete la calidad de los resultados. ¿Qué observa?

3. Grafique el Árbol de Decisión generado y enliste las reglas de decisión.

Ejercicio 3: El objetivo de este ejercicio es utilizar el método de árboles de decisión para predecir números escritos a mano (Hand Written Digit Recognition). Las tablas de aprendizaje y prueba ya han sido previamente preparadas y se encuentran en los archivos "ZipDataTrainCod.csv" y "ZipDataTestCod.csv".



Los datos corresponden a códigos postales escritos a mano en sobres del correo postal de EE.UU. Más precisamente, cada dígito está representado por una imagen de 16 × 16 en escala de grises, cada pıxel con 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 pıxel, la identidad de cada imagen (0, 1, . . . , 9) de forma rápida y precisa. De lograrse el cometido con una buena precisión, el algoritmo resultante se utilizará 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 lo más baja posible de modo de evitar una mala distribución 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, cada fila de la tabla representa un bloque 16 × 16 por lo que la matriz

tiene 256 variables predictivas.

Para este ejercicio realice lo siguiente:

1. Use el método de Árboles de Decisión para generar un modelo predictivo para la tabla ZipDataTrainCod.csv usando 100 % de los datos para tabla aprendizaje. ¿Qué puede comentar con respecto al tiempo de ejecución del algoritmo?

2. Para la tabla de prueba ZipDataTestCod.csv calcule la matriz de confusión y la precisión positiva para cada número ¿Qué tan buenos son los resultados? ¿Puede explicar por qué algunos números inducen una menor precisión?

3. Grafique el Árbol de Decisión generado y las reglas de decisión.

Ejercicio 4: Supongamos que tenemos un Árbol de Decisión con tres clases A, B, C. Se debe decidir cómo dividir el nodo padre:

$$N = \begin{pmatrix} A & 100 \\ B & 50 \\ C & 60 \end{pmatrix}$$

para esto hay dos posibles divisiones. La primera posible división divide el nodo N en los 2 siguientes nodos:

$$N_{1,1} = \begin{pmatrix} A & 62 \\ B & 8 \\ C & 0 \end{pmatrix}, \quad N_{1,2} = \begin{pmatrix} A & 38 \\ B & 42 \\ C & 60 \end{pmatrix}$$

La segunda opción de división para el nodo N es la siguiente en 3 nodos:

$$N_{2,1} = \begin{pmatrix} A & 65 \\ B & 20 \\ C & 0 \end{pmatrix}, \quad N_{2,2} = \begin{pmatrix} A & 21 \\ B & 19 \\ C & 20 \end{pmatrix}, \quad N_{2,3} = \begin{pmatrix} A & 14 \\ B & 11 \\ C & 40 \end{pmatrix}$$

1. Calcule la información ganada usando el índice de Gini para las dos posibles divisiones. ¿Cuál división es la mejor?

2. Repita 1, usando el criterio de la Entropía.

3. Repita 1, usando el criterio del Error Clasificación.