

Guía para elaborar el reporte de las prácticas de Árboles de decisión

Nombre: _____ Grupo: _____

Índice

1. Introducción.
Agregar una breve descripción del estudio que se realizó así como los resultados encontrados.
2. Describir la intención de la recopilación de los datos.
Agregar la descripción que sugiere la intención que se tuvo para recopilar los datos que contiene el conjunto de datos (*dataset*) que se analiza.
3. Diccionario de datos.
Agregar el diccionario de datos del conjunto de datos a analizar. Considerando: nombre de la variable, tipo de dato (nominal, ordinal, numérico discreto o continuo), dominio de valores.
4. Objetivo del análisis del conjunto de datos.
Establecer el objetivo para el análisis de datos mediante un árbol. Identificando variables dependientes e independientes.
5. Árbol.
La imagen del árbol generada. Especificando el tipo de muestra que tomó para el entrenamiento y la fase de prueba. Indicar la herramienta en la que se desarrolló.
6. Medidas.
Incluir las siguientes medidas y la explicación del significado de cada una de ellas:
 - Matriz de confusión - Sensibilidad - Precisión
 - Tasa de error - Especificidad - Explicación de los positivos
 - Exactitud verdaderos y positivos falsos
7. Análisis de los resultados.
Describir los resultados encontrados proporcionando la explicación de los mismos. Así como las conclusiones que obtenga.

Anexo de las medidas:

Medidas e Imágenes tomadas de Rokach, L. & Maimon, O. (2015).

La **sensibilidad** (*Sensitivity*) (también conocida como recuerdo (*recall*)) evalúa qué tan bien el clasificador puede reconocer muestras positivas y se define como:

$$Sensitivity = \frac{true_positive}{positive}, \quad (4.2)$$

La medida de **especificidad** (*Specificity*) mide que tan bien el clasificador puede reconocer las muestras negativas. Se define como:

$$Specificity = \frac{true_negative}{negative}, \quad (4.3)$$

La **precisión** (*Precision*) mide cuántos ejemplos clasificados como clase "positiva" son realmente "positivos". Esta medida es útil para evaluar clasificadores nítidos que se utilizan para clasificar un conjunto de datos completo. Formalmente:

$$Precision = \frac{true_positive}{true_positive + false_positive}. \quad (4.4)$$

La **matriz de confusión** contiene el número de elementos que se han clasificado correcta o incorrectamente para cada clase. Un beneficio de una matriz de confusión es que es fácil ver si el sistema confunde dos clases (es decir, etiquetar comúnmente una como otra).

Table 4.1 A confusion matrix.

	Predicted negative	Predicted positive
Negative Examples	A	B
Positive Examples	C	D

- Accuracy is: $(a + d)/(a + b + c + d)$
- Misclassification rate is: $(b + c)/(a + b + c + d)$
- Precision is: $d/(b + d)$
- True positive rate (Recall) is: $d/(c + d)$
- False positive rate is: $b/(a + b)$
- True negative rate (Specificity) is: $a/(a + b)$
- False negative rate is: $c/(c + d)$