**Guía de ejercicios Capítulo VI**

**Árboles de decisión**

1. **Objetivos de la Unidad**

| **#** | **Descripción** | **Preguntas** |
| --- | --- | --- |
| 1. | Comprender la generación de un árbol de decisión. |  |
| 2. | Cuantificar la ganancia de información para un atributo en un conjunto de datos. |  |
| 3. | Comprender los algoritmos de generación de los árboles de decisión. |  |
| 4. | Establecer los mecanismos de poda de los árboles de decisión. |  |
| 5. | Comprender los mecanismos de equivalencia de reglas y la generalización de reglas simples. |  |

1. **Preguntas**

P1. Razón de ganancia

Para la siguiente base de datos muestre los cálculos de la razón de ganancia propuestos por Quinlan, para seleccionar el atributo que presenta mejor discriminación de la clase.

| **A** | **B** | **Clase** |
| --- | --- | --- |
| 0 | 0 | Sí |
| 0 | 0 | No |
| 0 | 1 | No |
| 0 | 1 | Sí |
| 1 | 0 | Sí |
| 1 | 1 | Sí |
| 1 | 0 | Sí |
| 1 | 1 | Sí |
| 2 | 0 | No |
| 2 | 0 | No |
| 2 | 1 | Sí |
| 2 | 1 | Sí |

P2. Clientes bancarios

La BD de la tabla muestra dos variables seleccionadas por un banco, para clasificar clientes. Salario y si tiene hijos. Usando el método de máxima entropía propuesto por Quinlan, determine cuál de las dos variables discrimina mejor si es un buen cliente.

| **Salario** | **Hijos** | **Buen cliente** |
| --- | --- | --- |
| Alto | Si | Si |
| Alto | Si | No |
| Alto | No | No |
| Alto | No | Si |
| Medio | Si | Si |
| Medio | No | Si |
| Medio | Si | Si |
| Medio | No | Si |
| Bajo | Si | No |
| Bajo | Si | No |
| Bajo | No | Si |
| Bajo | No | Si |

P3. Posiciones promotor genético

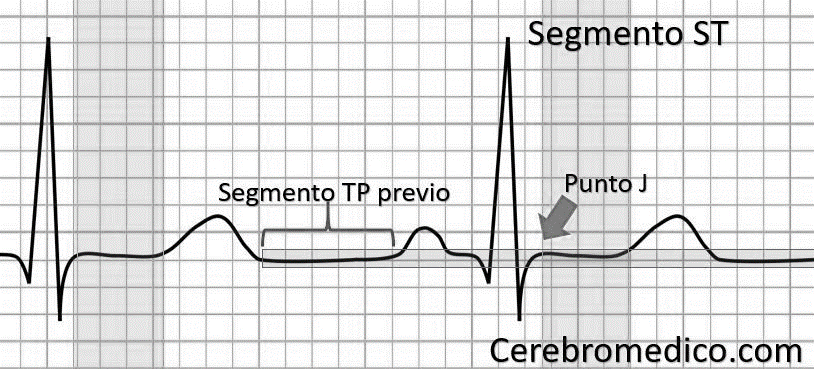
En genética es muy importante determinar si un gen está expresado en un organismo particular. Una forma de averiguar si el gen está expresado o no, es saber si existe un *promotor* antes del gen. La existencia de un determinado *promotor* se puede medir por la existencia de aminoácidos en el inicio y final del *promotor*. Para el caso de un organismo específico, se sabe que en el inicio del *promotor* existen dos valores posibles: Glicina (G) y Prolina (P); y en el fin del *promotor* existen tres alternativas: Histidina (H), Metionina (M) y Lisina (L).

Usando el método de máxima entropía propuesto por Quinlan, determine cuál de las dos posiciones (inicio o fin) del *promotor* discrimina mejor su existencia.

| **Inicio** | **Fin** | **Promotor** |
| --- | --- | --- |
| G | H | Sí |
| G | H | No |
| P | H | Sí |
| G | M | Sí |
| P | L | Sí |
| P | M | Sí |
| P | H | No |
| G | M | Sí |
| G | L | No |
| G | L | No |
| P | M | Sí |
| P | L | Sí |

P4. Características del segmento ST

En un electrocardiograma (ECG), el segmento ST representa el período isoeléctrico cuando los ventrículos se encuentran entre la despolarización y la repolarización.



Al evaluar los totales de una BD de ECG, se requiere determinar entre las características: *Depresión de segmento* ST e *Inclinación del segmento* ST, cual clasifica de mejor forma una *cardiopatía isquémica*.

Se cuentan con 1.000 casos para la *Depresión del segmento* ST, y 800 para el caso de *Inclinación* *del segmento* ST, como se muestra en las siguientes tablas.

|  | **Con isquemia** | **Sin isquemia** | **Total** |
| --- | --- | --- | --- |
| **Sin depresión ST** | 200 | 400 | 600 |
| **Con depresión ST** | 200 | 200 | 400 |
| **Total** | 400 | 600 | 1.000 |

|  | **Con Isquemia** | **Sin isquemia** | **Total** |
| --- | --- | --- | --- |
| **Sin inclinación ST** | 48 | 552 | 600 |
| **Con inclinación ST** | 200 | 0 | 200 |
| **Total** | 248 | 552 | 800 |

P5. Riesgo de accidente de tránsito

La SIAT de Carabineros de Chile posee una base de datos reducida (incluida en la siguiente tabla), con las variables edad del conductor y el estado del camino (Bueno / Malo) para predecir el riesgo de accidentes de tránsito. Además, entre los conductores se reconocen tres comportamientos según su rango de edad: desde adolescentes a 26 años, de 27 a 37 años, y mayores de 38 años.

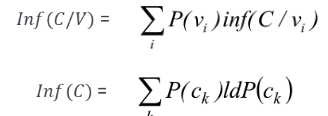
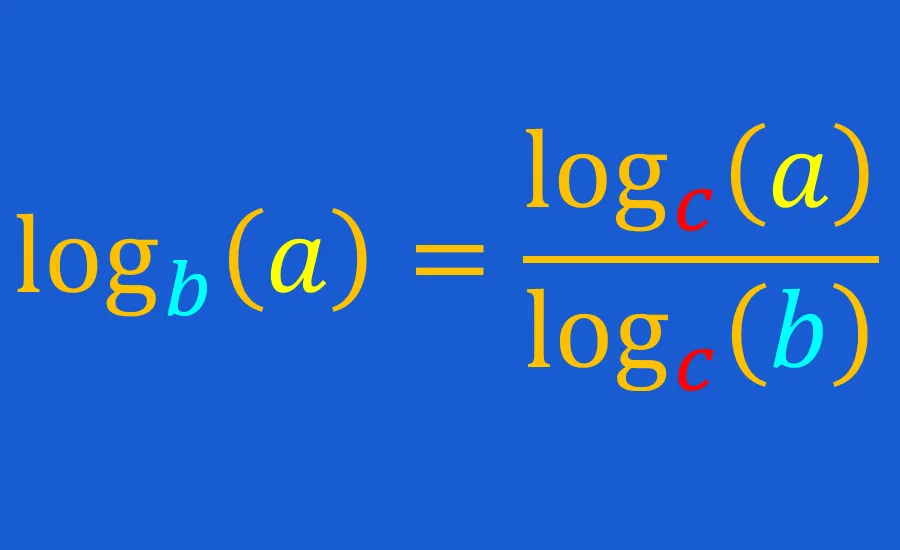
Utilizando el método de árboles de decisión, determine cual atributo es más eficiente para clasificar el riesgo de accidente de tránsito.

| **Edad** | **Estado camino** | **Accidente** |
| --- | --- | --- |
| 40 | Malo | Si |
| 20 | Malo | No |
| 28 | Bueno | No |
| 47 | Bueno | Si |
| 37 | Malo | Si |
| 32 | Bueno | No |
| 41 | Malo | No |
| 38 | Bueno | No |
| 15 | Malo | No |
| 27 | Malo | No |
| 38 | Bueno | Si |
| 42 | Bueno | Si |

1. **Respuestas**

P2.

Ganancia = Inf(C)-Inf(C/V)

**

*P(C = si) = 8/12 = 0,67*

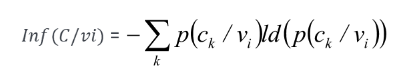
*P(C = no) = 4/12 = 0,33*

*inf(C) = -[P(C = si) \* log\_2P(C= si) + P(C= no)\* log\_2(P(C=no))]*

*inf(C) = 0,67\*log\_2(0,67) +0,33\*log\_2(0,33)*

*= -(0,67 \* ln (0,67)/ ln (2) + 0,33 \* ln(0,33)/ln(2))*

*= 0.914926*

**

*inf(C/ V = Salario):*

*P(salario = alto)\* (-P(C = si / salario = Alto) \* log2(P(C = si / salario = Alto )) - P(C = no / salario = Alto ) \* log2(P(C = no / salario = Alto )) +*

*P(salario = medio) \* (- P(C = si / salario = medio) \* log2(P(C = si / salario = medio )) - P(C = no / salario = medio ) \* log2(P(C = no / salario = medio )) +*

*P(salario = bajo) \* (- P(C = si / salario = bajo) \* log2(P(C = si / salario = bajo )) - P(C = no / salario = bajo ) \* log2(P(C = no / salario = bajo ))*

*inf(C/ V = Hijos):*

*P(hijos = Si)\* (-P(C = si / hijos = Si) \* log2(P(C = si / hijos = Si )) - P(C = no / hijos = Si ) \* log2(P(C = no / hijos = Si )) +*

*P(hijos = No) \* (- P(C = si / hijos = No) \* log2(P(C = si / hijos = No )) - P(C = no / hijos = No ) \* log2(P(C = no / hijos = No ))*

*Ganancia(Salario) = inf(c) - inf(C/ V = Salario)*

*Ganancia(Hijos) = inf(c) - inf(C/ V = Hijos)*

*P(C = si / salario = Alto ) = 2/4 P(C = no / salario = Alto ) = 2/4*

*P(C = si / salario = Medio ) = 4/4 P(C = no / salario = Medio ) = 0/4*

*P(C = si / salario = Bajo ) = 2/4 P(C = no / salario = Bajo ) = 2/4*

*P(C = si / hijos = Si ) = 3/6 P(C = no / hijos = Si ) = 3/6*

*P(C = si / hijos = No ) = 5/6 P(C = no / hijos = No ) = 1/6*

*P(C = si / salario = Alto ) \* log\_2(P(C = si / salario = Alto ) = 2/4 \* log\_2(2/4)*

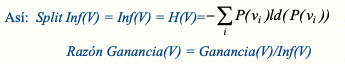
*P(C = si / salario = medio ) \* log\_2(P(C = si / salario = medio ) = 4/4 \* log\_2(4/4)*

*P(C = si / salario = bajo ) \* log\_2(P(C = si / salario = bajo ) = 2/4 \* log\_2(2/4)*

*P(C = no / salario = Alto ) \* log\_2(P(C = no / salario = Alto ) = 2/4 \* log\_2(2/4)*

*P(C = no / salario = medio ) \* log\_2(P(C = no / salario = medio ) = 4/4 \* log\_2(4/4)*

*P(C = no / salario = bajo ) \* log\_2(P(C = no / salario = bajo ) = 2/4 \* log\_2(2/4)*

**

*Inf(salario) = -(P(Salario = Alto)\*log2(P(Salario = Alto)) + P(Salario = Medio)\*log2(P(Salario = Medio)) + P(Salario = Bajo)\*log2(P(Salario = Bajo))*