Introducción a Machine Learning

Aprendizaje de Máquina – Árboles de Decisión

MSc. Marco Sobrevilla

Objetivo

- Aprender conceptos relacionados a Árboles de decisión
  - Entropía
  - *Ganancia de Información*

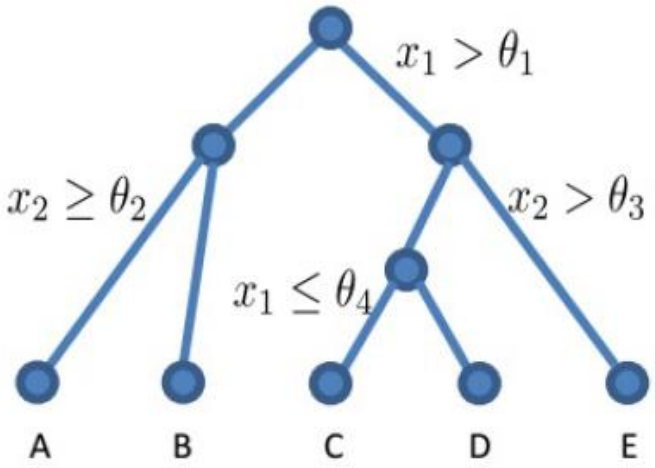
Agenda

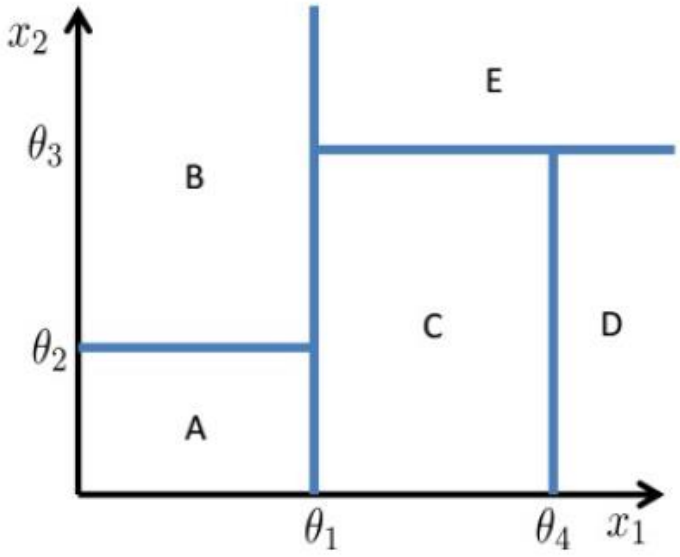
- Árbol de Decisión
- Entropía
- Ganancia de Información
- Ejemplo de algoritmo ID3

Árboles de Decisión

- Método para **inferencia inductiva**
  - Ayuda a predecir la clase de un objeto en estudio con base en un entrenamiento previo
- Un árbol representa una función discreta para aproximar/representar los datos de entrenamiento
- Árboles de Decisión clasifican instancias ordenándolas **desde la raíz hasta algún nodo hoja**
  - Cada **nodo del árbol** representa un **atributo**

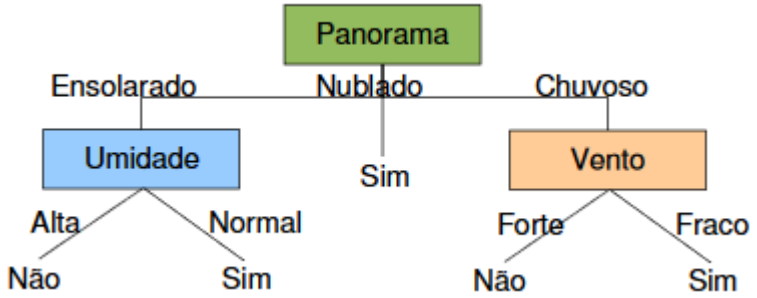
Árboles de Decisión





Árboles de Decisión

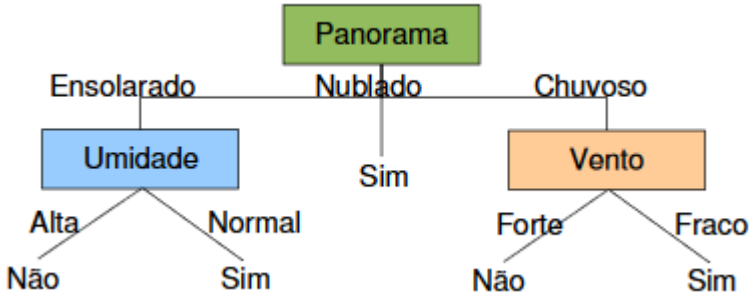
- Consideremos la toma de decisión para el problema «Jugar Tenis»
  - Clasificar si un día es adecuado o no para jugar tenis



  - Por ejemplo, si tenemos la instancia:
    - Panorama: Ensolarado, Umidade: Alta
    - ¿Cuál sería nuestra salida?

Árboles de Decisión

- Puede crearse una expresión para verificar cuándo es posible jugar tenis.



- Si (panorama: ensolarado AND umidade: Normal) **OR** panorama: nublado **OR** (panorama: chuvoso AND vento: Fraco)
- Podemos crear árboles de decisión y **extraer reglas** que nos ayuden a clasificar **instancias nunca vistas**

Árboles de Decisión

- Son adecuados para problemas en que:
  - Instancias son representadas por pares atributos-valor
    - Hay un conjunto fijo de atributos (Ejm: Umidade) y sus valores (Ejm: Alta, Normal)
    - Situación ideal es cuando cada atributo puede asumir pocos valores (discretos)
  - La función a ser aproximada tiene valores discretos
    - En el ejemplo: “Sí o no”
    - Puede ser fácilmente extendible para producir más de 2 valores de salida
    - Se vuelven más complejas y menos utilizables en contextos con variables continuas

Árboles de Decisión

- Aplicaciones Comunes
  - Diagnóstico de Pacientes
  - Análisis de Crédito
  - Problemas en equipamientos mecánicos

- Algoritmos más conocidos
  - **ID3** (Quinlan, 1986) e C4.5 (Quinlan, 1993)

Árboles de Decisión

- Algoritmo ID3
  - Considera un conjunto de datos para el entrenamiento
  - Construye un árbol usando un enfoque **top-down** considerando la pregunta: **“¿Cuál es el atributo más importante?”** -> Raíz del árbol
  - Cada atributo es probado y su capacidad para volverse nodo raíz es evaluada
  - Se crean tantos nodos hijos de la raíz cuanto valores posibles pueda asumir el atributo (caso discreto)
  - Se repite el proceso para cada nodo hijo de la raíz y así sucesivamente

Árboles de Decisión

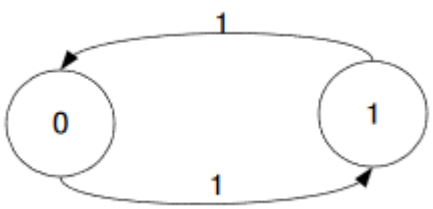
- ¿ID3 cómo evalúa cuál es el atributo más adecuado?
  - Usando la medida **ganancia de información**

Un momento!

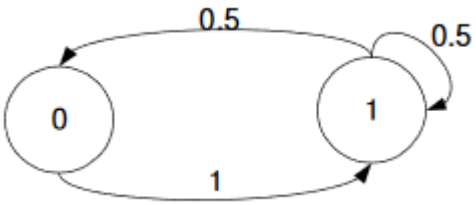
  - Para conocer qué es ganancia de información primero tenemos que saber **¿Qué es entropía?**

Entropía en Teoría de la Informaciór

- Para entender entropía, consideremos el siguiente sistema:



- Ahora considere que el sistema alteró su comportamiento:



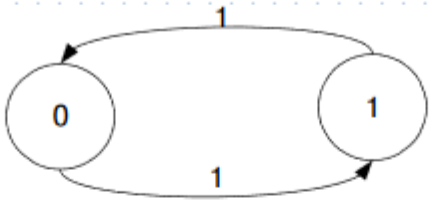
Entropía

- Ahora consideremos la fórmula de Shannon para calcular la entropía:

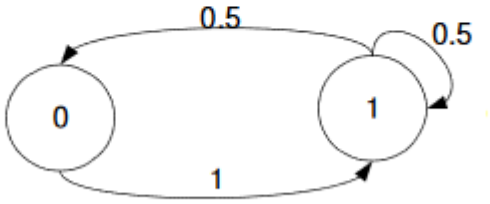
$$E = -\sum_i \sum_j p_{ij} \log_2 p_{ij}$$

- Esa expresión mide la energía total de un sistema:
  - Considerando que el sistema está en el estado «i» y ocurre una transición al estado «j»
  - La función log2 es usada para cuantificar la entropía en términos de **bits**

Entropía



$$E = -(1\log_2(1) + 1\log_2(1)) = 0$$



$$E = -(1\log_2(1) + 0.5\log_2(0.5) + 0.5\log_2(0.5)) = 1$$

Al modificar su comportamiento, el sistema agregó mayor nivel de incerteza o mayor energía

Uso de Entropía en el ID3

- Considere una colección S de instancias com ejemplos **positivos** y **negativos**
  - Con 2 clases distintas

- En este caso, se asume la probabilidad de pertenecer a una clase (positiva o negativa) de S
  - Entonces, la **entropia**, en ese contexto, es dada por:

$$E(S) = -p_{\oplus} \log_2 p_{\oplus} - p_{\ominus} \log_2 p_{\ominus}$$

Uso de Entropía en el ID3

- Considere el conjunto S con 14 ejemplos de algún concepto boleano
  - 9 positivos
  - 5 negativos

- Entonces, la entropía de esse conjunto es dada por:

$$E(S) = -\frac{9}{14}\log_2\frac{9}{14} - \frac{5}{14}\log_2\frac{5}{14} = 0.94$$

Uso de Entropía en el ID3

- En otros casos:
  - Para [7+, 7-]

$$E(S) = -\frac{7}{14}\log_2\frac{7}{14} - \frac{7}{14}\log_2\frac{7}{14} = 0.99\ldots \approx 1$$

  - Para [0+, 14-] ou [14+, 0-]

$$E(S) = -\frac{14}{14}\log_2\frac{14}{14} = 0$$

- La entropía mide el **nivel de certeza** que tenemos sobre un evento

Uso de Entropía en el ID3

- Podemos generalizar para más de 2 posibles clases:

$$E(S) = \sum_{i=1}^{c} -p_i \log_2 p_i$$

- ¿Por qué el uso de la función *Log*?
  - En Teoría de la Información se mide la información proveniente de una fuente en bits

Ganancia de Información en el ID3

- Ahora sí… ¿Qué es ganancia de información?
  - Mide la efectividad de un atributo en clasificar un conjunto de entrenamiento
  - Ganancia de información de un atributo A:
    - **Mide la reducción de la entropía, causada por la partición de ejemplos de acuerdo con este atributo**

$$\mathbf{GI}(S, A) = E(S) - \sum_{v \in \mathbf{Valores}(A)} \frac{S_v}{S} E(S_v)$$

    - El segundo término mide la entropía partiendo el conjunto de entrenamiento de acuerdo con el atributo A

  - **GI mide la reducción de la entropía (la incerteza) al seleccionar el atributo A**

Ganancia de Información en el ID3

- Por ejemplo, considere S un conjunto de entrenamiento conteniendo el atributo viento («*vento*») que puede asumir los valores de «*fraco*» y «*forte*»
  - S contiene 14 ejemplos (9 positivos y 5 negativos)
  - Ahora considere que:
    - 6 de los ejemplos positivos y 2 de los ejemplos negativos son definidos por vento = fraco (8 en total)
    - 3 ejemplos definidos por vento = forte tanto para la clase positiva cuanto para la clase negativa (6 en total)
  - La ganancia de información al seleccionar el atributo «*vento*» en la raíz del árbol es dada por:

$$S = [9+, 5-]$$
$$S_{fraco} \leftarrow [6+, 2-]$$
$$S_{forte} \leftarrow [3+, 3-]$$
$$\mathbf{GI}(S, A) = E(S) - \sum_{v \in \mathbf{Valores}(A)} \frac{S_v}{S} E(S_v)$$

Ganancia de Información en el ID3

$$S = [9+, 5-]$$
$$S_{fraco} \leftarrow [6+, 2-]$$
$$S_{forte} \leftarrow [3+, 3-]$$

$$\mathbf{GI}(S, A) = E(S) - \sum_{v \in \mathbf{Valores}(A)} \frac{S_v}{S} E(S_v)$$

$$\mathbf{GI}(S, A) = 0.94 - \frac{8}{14} E(S_{fraco}) - \frac{6}{14} E(S_{forte})$$

$$S = [9+, 5-]$$
$$S_{fraco} \leftarrow [6+, 2-]$$
$$S_{forte} \leftarrow [3+, 3-]$$

$$E(S_{fraco}) = -\frac{6}{8}\log_2\frac{6}{8} - \frac{2}{8}\log_2\frac{2}{8} = 0.811$$

$$E(S_{forte}) = -\frac{3}{6}\log_2\frac{3}{6} - \frac{3}{6}\log_2\frac{3}{6} = 1.00$$

$$\mathbf{GI}(S, A) = 0.94 - \frac{8}{14} 0.811 - \frac{6}{14} 1.00 = 0.048$$

- Ganancia de Información es usada por ID3 en cada paso de la generación del árbol de decisión
  - **Reducimos muy poco el nivel de incerteza.**
  - **«Vento» no es buen atributo**

Ejemplo - ID3

- **Consideremos que queremos saber cuándo se jugará tenis**

Dia	Panorama	Temperatura	Umidade	Vento	Jogar Tênis
1	Ensolarado	Quente	Alta	Fraco	Não
2	Ensolarado	Quente	Alta	Forte	Não
3	Nublado	Quente	Alta	Fraco	Sim
4	Chuvoso	Intermediária	Alta	Fraco	Sim
5	Chuvoso	Fria	Normal	Fraco	Sim
6	Chuvoso	Fria	Normal	Forte	Não
7	Nublado	Fria	Normal	Forte	Sim
8	Ensolarado	Intermediária	Alta	Fraco	Não
9	Ensolarado	Fria	Normal	Fraco	Sim
10	Chuvoso	Intermediária	Normal	Fraco	Sim
11	Ensolarado	Intermediária	Normal	Forte	Sim
12	Nublado	Intermediária	Alta	Forte	Sim
13	Nublado	Quente	Normal	Fraco	Sim
14	Chuvoso	Intermediária	Alta	Forte	Não

Ejemplo - ID3

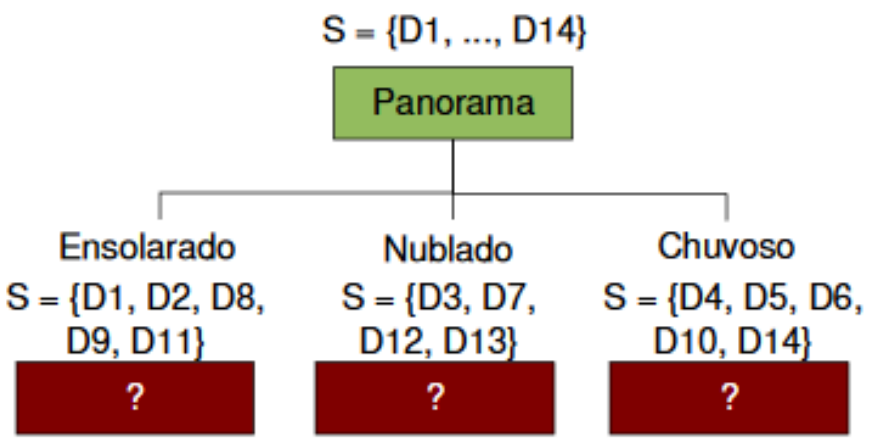
- Primer paso:
  - Calculamos la ganancia de información de cada atributo

$$\mathbf{GI}(S, \mathbf{Panorama}) = 0.246$$
$$\mathbf{GI}(S, \mathbf{Umidade}) = 0.151$$
$$\mathbf{GI}(S, \mathbf{Vento}) = 0.048$$
$$\mathbf{GI}(S, \mathbf{Temperatura}) = 0.029$$

  - Atributo con mayor ganancia de información es seleccionado como raíz del árbol
    - **El que más reduce el nivel de incerteza**
    - **Panorama es escogido**
    - **Creamos los nodos hijos a partir de la raíz con los posibles valores asumidos por Panorama**

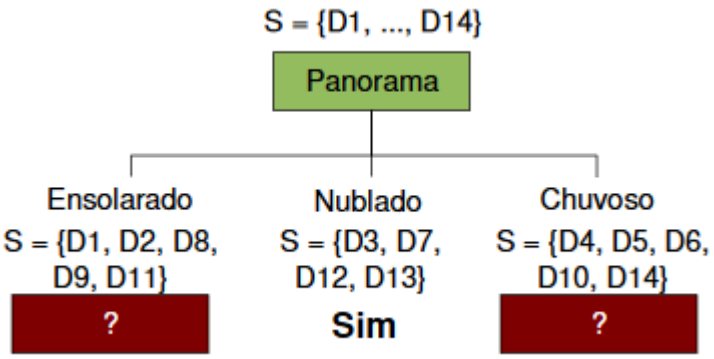
Ejemplo - ID3

- Ahora tenemos la raíz
  - Proceder de la misma forma por las ramas de la raíz
  - En cada rama consideramos solo los ejemplos contenidos en ella
    - Desde que haya divergencia entre las clases de salida



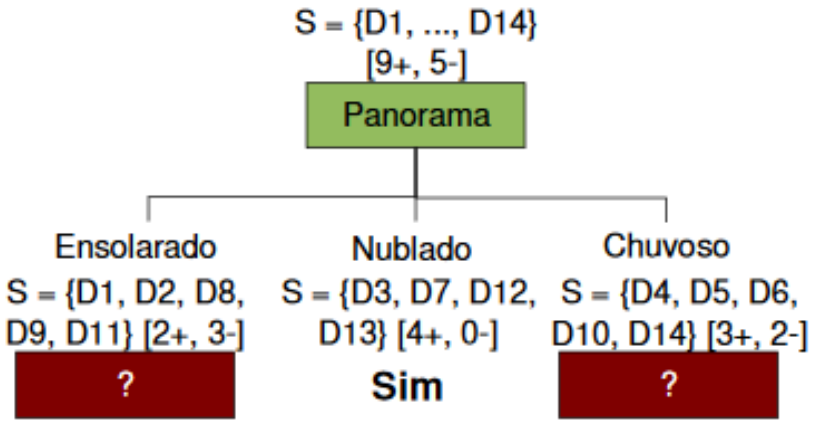
Ejemplo - ID3

- Una de las ramas no tiene divergencia entre las clases de salida, o sea, **Entropía igual a cero**



- Atributos existentes incorporados encima de determinado nodo **no entran la evaluación** de la ganancia de información del siguiente

Ejemplo - ID3



- Calculando la ganancia de información para la rama «ensolarado» tenemos:
  - Calculamos la entropía para *E(S = Ensolarado)*
    - *Nivel de incerteza del panorama «ensolarado»*

$$E(S = Ensolarado) = -\frac{2}{5}\log\frac{2}{5} - \frac{3}{5}\log\frac{3}{5} = 0.97$$

Ejemplo - ID3

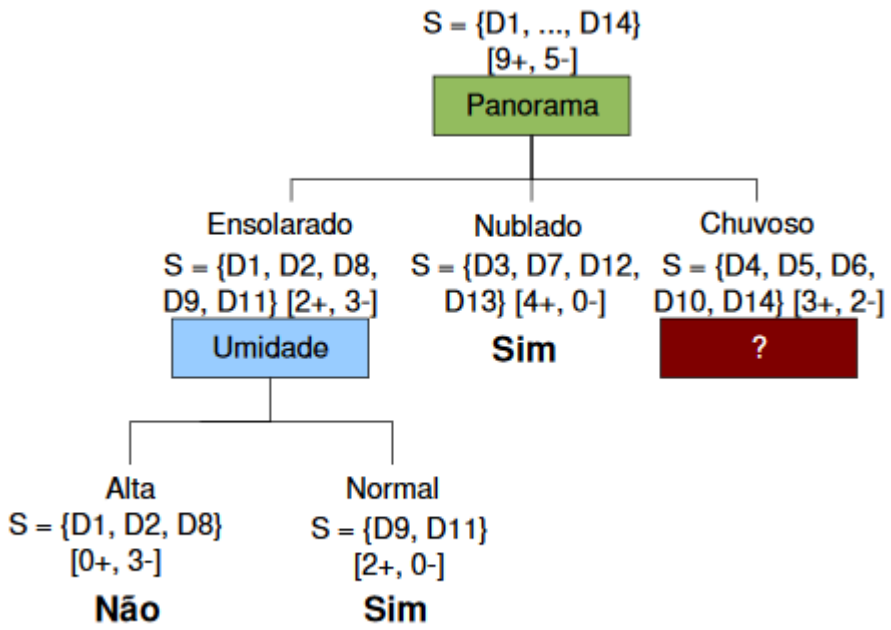
- Calculando la ganancia de información para la rama «ensolarado», tenemos:

$$\mathbf{GI}(S, \mathbf{Umidade}) = 0.97 - \frac{3}{5}0.0 - \frac{2}{5}0.0 = 0.97$$
$$\mathbf{GI}(S, \mathbf{Temperatura}) = 0.97 - \frac{2}{5}0.0 - \frac{2}{5}1.0 = 0.57$$
$$\mathbf{GI}(S, \mathbf{Vento}) = 0.97 - \frac{2}{5}1.0 - \frac{3}{5}0.918 = 0.019$$

- Seleccionamos «Umidade» porque es la que tiene mayor ganancia (reduce más la entropía)

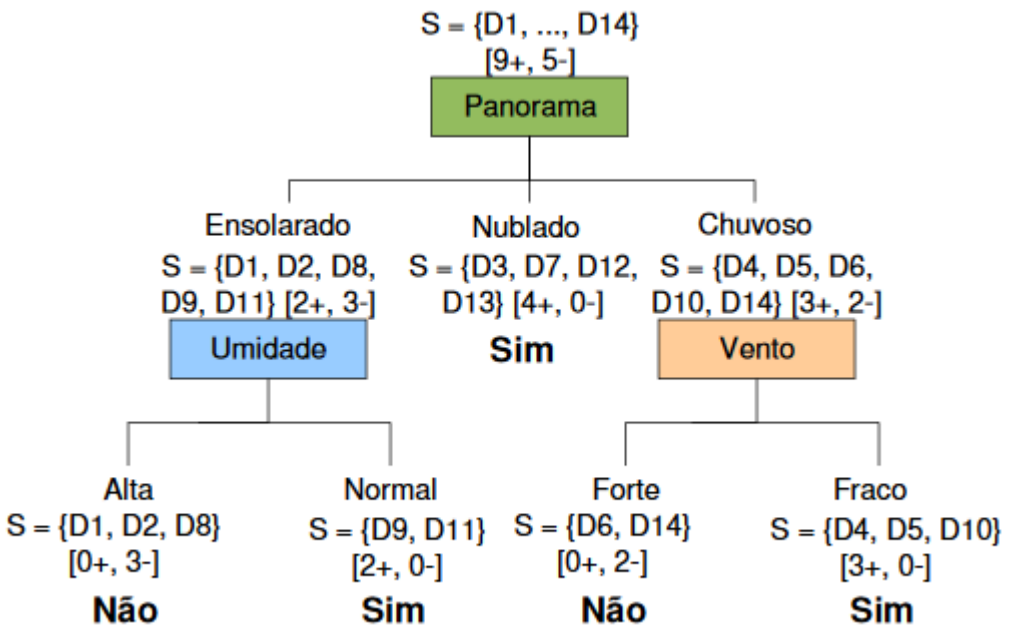
Ejemplo - ID3

- Continuando…



Ejemplo - ID3

- Continuando…



Ejemplo - ID3

- El algoritmo continua hasta que una de 2 condiciones sea satisfecha:
  - Todos los atributos fueron incluídos en el camino de la raíz hasta las hojas
  - Ejemplos de entrenamiento asociados con una rama presentan el mismo valor de salida

Observaciones- ID3

- ID3 crece lo suficiente para abarcar todos los ejemplos del entrenamiento
  - ¿Esto es bueno?
  - ¿Qué puede ocurrir?
    - *Overfitting* a los datos de entrenamiento
    - Problemas de generalización
- Ruído en los datos
  - Construcción de árbol más compleja
  - Se adecua a los ejemplos pero no generalizará

Observaciones- ID3

- Otro problema:
  - Underfitting -> Pocos datos de entrenamiento

- ¿Cómo solucionarlo?
  - Podando ☺
  - C4.5 supera estas limitaciones

Ejercicio

- Diseñar el árbol ID3 para el problema de clasificación, considerando que la clase de salida decide si administramos o no tratamiento

Paciente	Presión Aterial	Urea en sangre	Gota	Hipotiroidismo	Administrar Tratamiento
1	Alta	Alta	Sí	No	No
2	Alta	Alta	Sí	Sí	No
3	Normal	Alta	Sí	No	Sí
4	Baja	Normal	Sí	No	Sí
5	Baja	Baja	No	No	Sí
6	Baja	Baja	No	Sí	No
7	Normal	Baja	No	Sí	Sí
8	Alta	Normal	Sí	No	No
9	Alta	Baja	No	No	Sí
10	Baja	Normal	No	No	Sí
11	Alta	Normal	No	Sí	Sí
12	Normal	Normal	Sí	Sí	Sí
13	Normal	Alta	No	No	Sí
14	Baja	Normal	Sí	Sí	No

Fin ☺

Bibliografía

- Samuel, A. L. (2000). Some studies in machine learning using the game of checkers. IBM Journal of research and development, 44 (1.2), 206-226.
- Anderson, J. R. (1986). Machine learning: An artificial intelligence approach (Vol. 2). R. S. Michalski, J. G. Carbonell, & T. M. Mitchell (Eds.). Morgan Kaufmann