MiDaS Millenium Nucleus Center for the Discovery of Structures in Complex Data

FACULTAD DE MATEMÁTICAS
PONTIFICIA UNIVERSIDAD
CATÓLICA DE CHILE

Agencia Nacional de Investigación y Desarrollo
Ministerio de Ciencia, Tecnología, Conocimiento e Innovación
Gobierno de Chile

milenio INICIATIVA CIENTÍFICA

# Métodos de clasificación

María José García

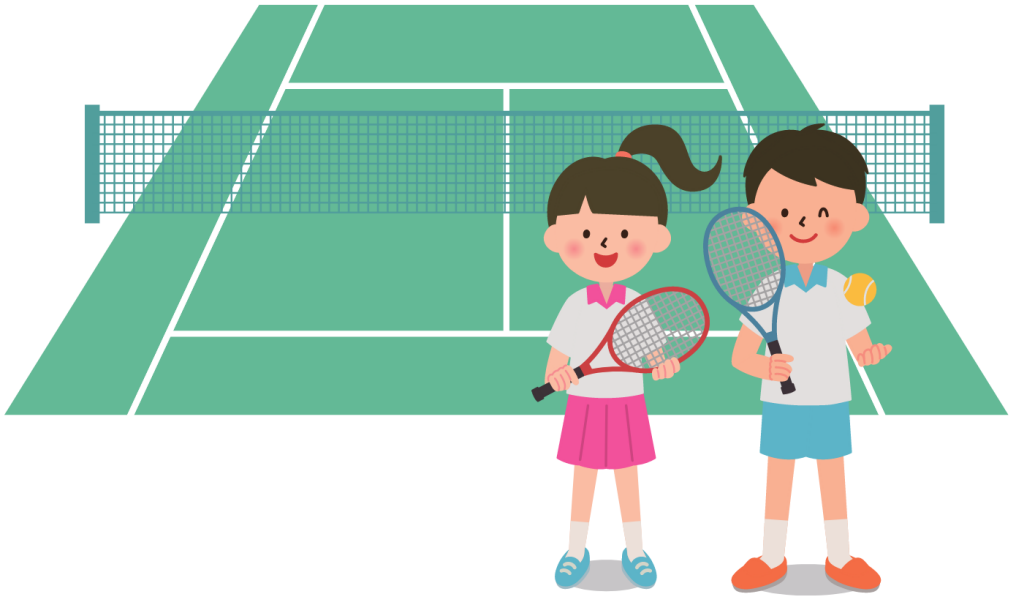
Olimpiada de Big Data

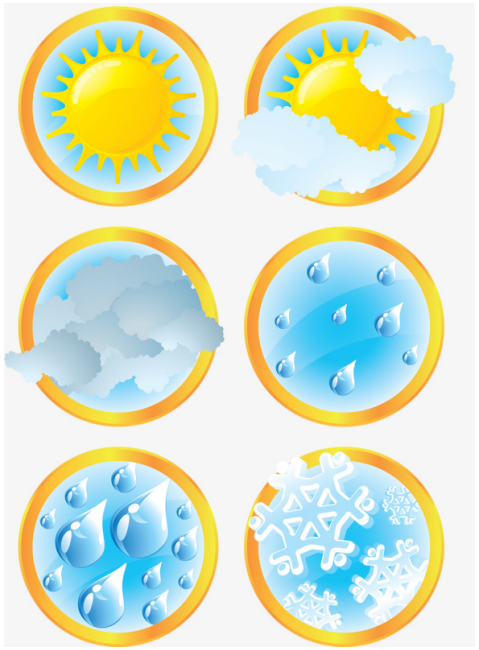
# Motivación


MiDaS

# Ejemplo

Suponga que usted juega tenis y quiere saber qué tan probable es que mañana juegue, ya que tiene que planificar una actividad familiar.





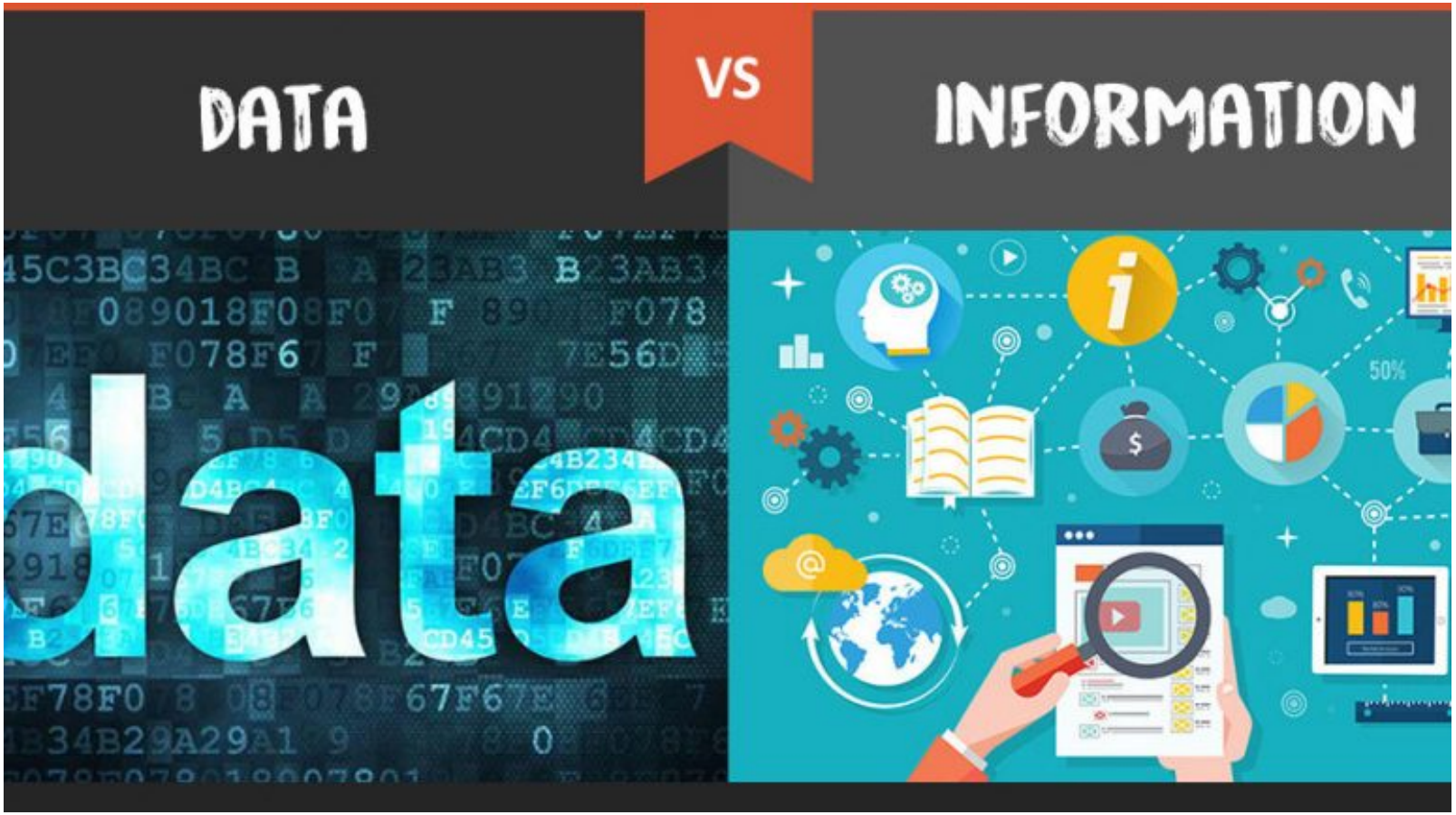
MiDaS

# Datos

Para tomar una decisión cuenta con información de diferentes aspectos meteorológicos correspondiente a 14 jugadores de tenis y si jugaron o no con esas condiciones climáticas.

| Jugador | Temperatura | Pronóstico | Humedad | Viento | Tenis |
|---|---|---|---|---|---|
| 1 | caluroso | nublado | alta | no | si |
| 2 | caluroso | nublado | normal | no | si |
| 3 | caluroso | soleado | alta | no | no |
| 4 | caluroso | soleado | alta | si | no |
| 5 | frio | lluvia | normal | no | si |
| 6 | frio | lluvia | normal | si | no |
| 7 | frio | nublado | normal | si | si |
| 8 | frio | soleado | normal | no | si |
| 9 | templado | lluvia | normal | no | si |
| 10 | templado | lluvia | alta | si | no |
| 11 | templado | lluvia | alta | no | si |
| 12 | templado | nublado | alta | si | si |
| 13 | templado | soleado | alta | no | no |
| 14 | templado | soleado | normal | si | si |

MiDaS




MiDaS

## Objetivo

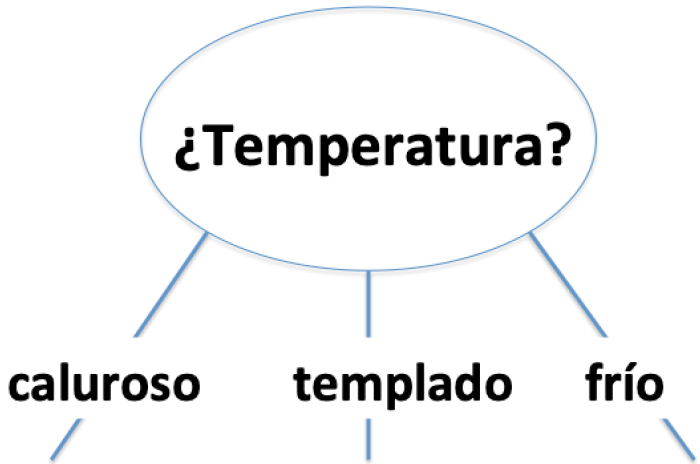
Encontrar un modelo que permita decidir mañana jugará tenis o no.

## Insumos

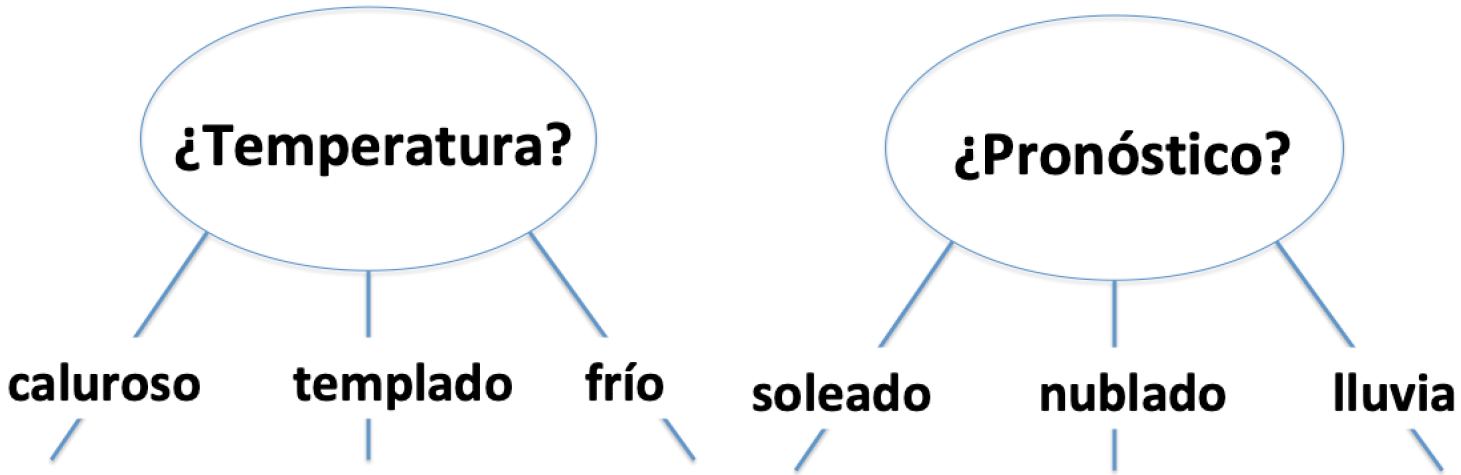
¿Qué tenemos para llevar a cabo nuestra tarea?

**¿Sugerencias...?**

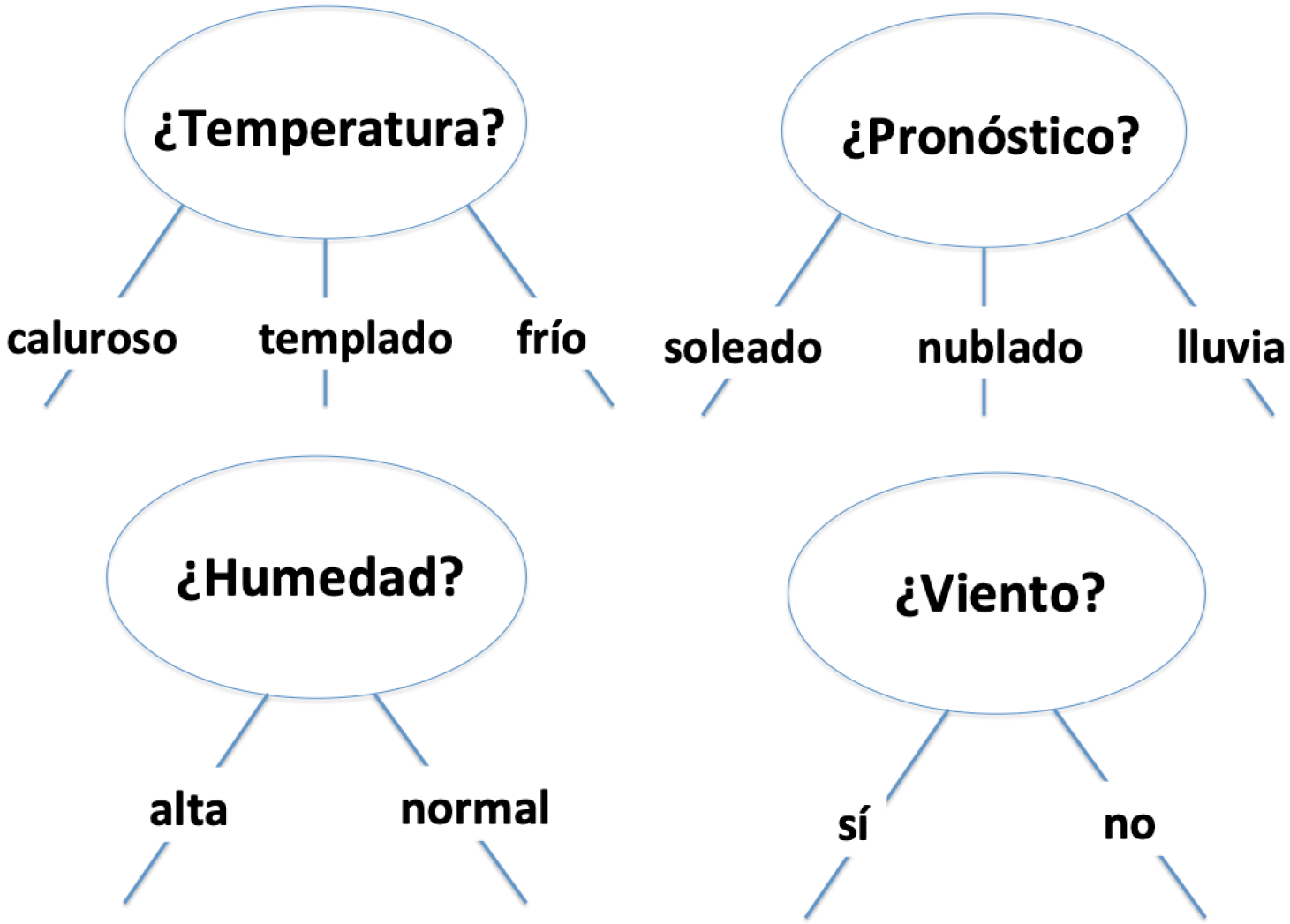
MiDaS



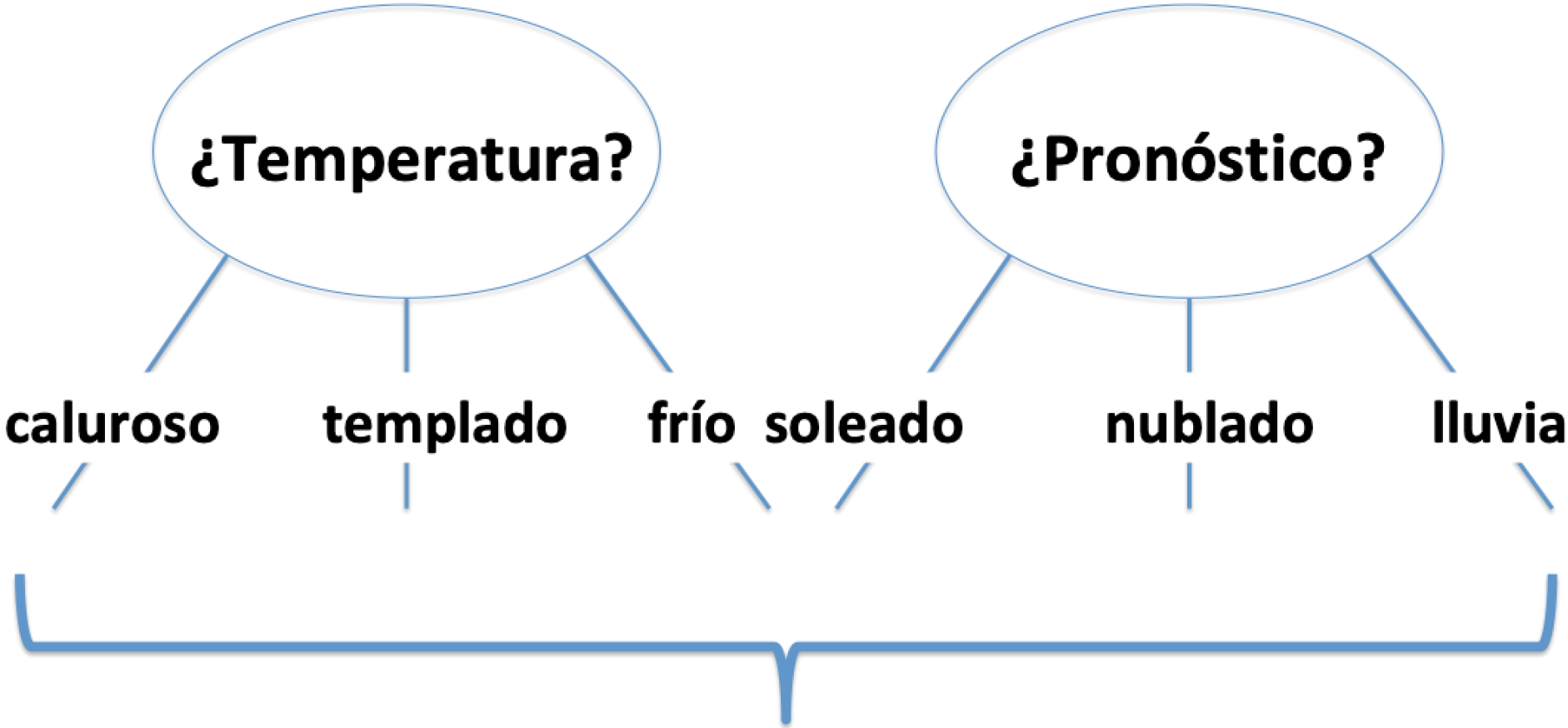

MiDaS


¿Temperatura?
caluroso
templado
frío
¿Pronóstico?
soleado
nublado
lluvia

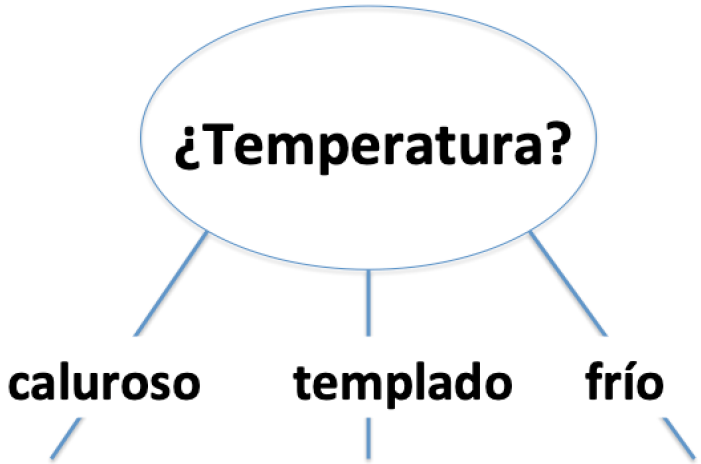
MiDaS


¿Temperatura?
caluroso
templado
frío
¿Pronóstico?
soleado
nublado
lluvia
¿Humedad?
alta
normal
¿Viento?
sí
no

MiDaS

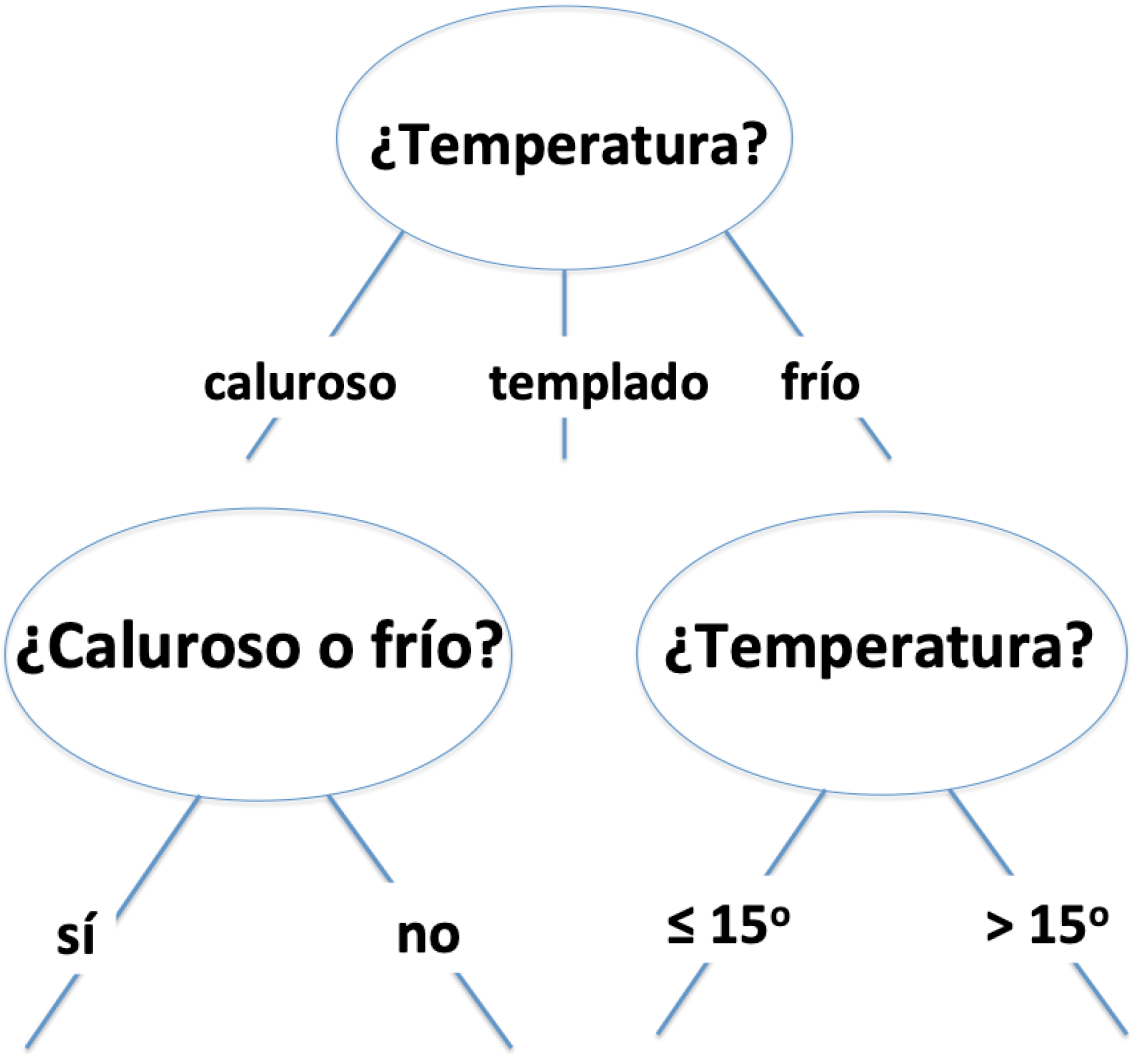

¿Temperatura?
¿Pronóstico?
caluroso
templado
frío
soleado
nublado
lluvia

MiDaS




MiDaS


MiDaS


¿Temperatura?
caluroso
templado
frío
¿Caluroso o frío?
¿Temperatura?
sí
no
≤ 15º
> 15º

MiDaS

# Árboles de decisión

## Definición

Un árbol de decisión es un modelo de predicción representado a través de un diagrama de flujo, que explora las alternativas de decisión y los posibles resultados.

- Se comienza con un conjunto de datos y sus clases asociadas
- Estos datos son divididos recursivamente en grupos menores
- Medidas de selección de variables: seleccionar aquella que mejor particiona los registros en las diferentes clases.

MiDaS

# Algoritmo básico de un árbol de decisión

Paso 1: El árbol parte con un solo grupo que contiene todos los datos.

Paso 2: Si los datos pertenecen todos a una misma clase, entonces no se hace ninguna división, y se etiqueta según su clase.

Paso 3: Si no, se recurre a un método de selección de variables para determinar el criterio de separación: variable de división y punto de corte o subconjunto de división.

Paso 4: Se construyen las ramas desde el grupo de acuerdo a los criterios de separación y se separan los datos en subgrupos.

Paso 5: El algoritmo usa el mismo proceso recursivamente.

Paso 6: El proceso para cuando se cumple alguno de los siguientes criterios:

1. todas los datos pertenecen a la misma clase
2. no hay más variables para particionar (clase mayoritaria)

MiDaS

# Construcción de un árbol de decisión

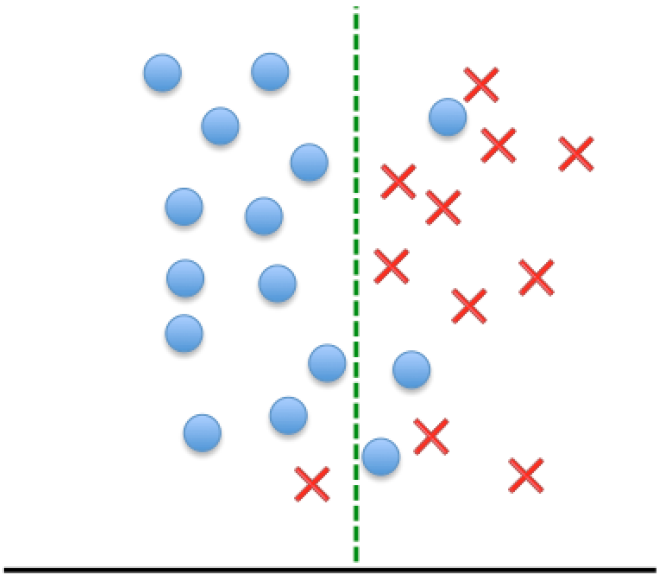
## Medidas de selección de variables

- Técnicas o métodos para seleccionar la variable que mejor separa un conjunto de datos en categorías o clases
- El ideal es que cada partición se pura $\Longrightarrow$ todos $\in$ misma clase
- También se conocen como reglas de división
- Se busca seleccionar la variable que logra “ordenar" de manera más eficiente los datos con respecto a las categorías de clasificación

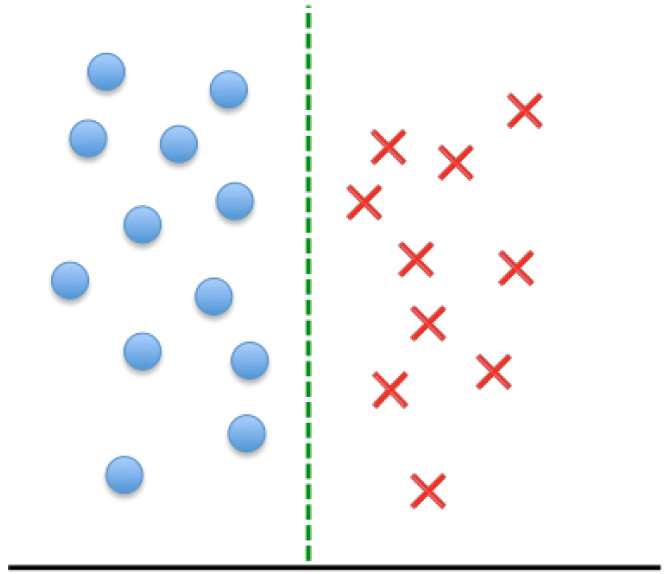
MiDaS

# Pureza / Impureza

¿Qué variable ordena mejor los datos (es más informativa)?



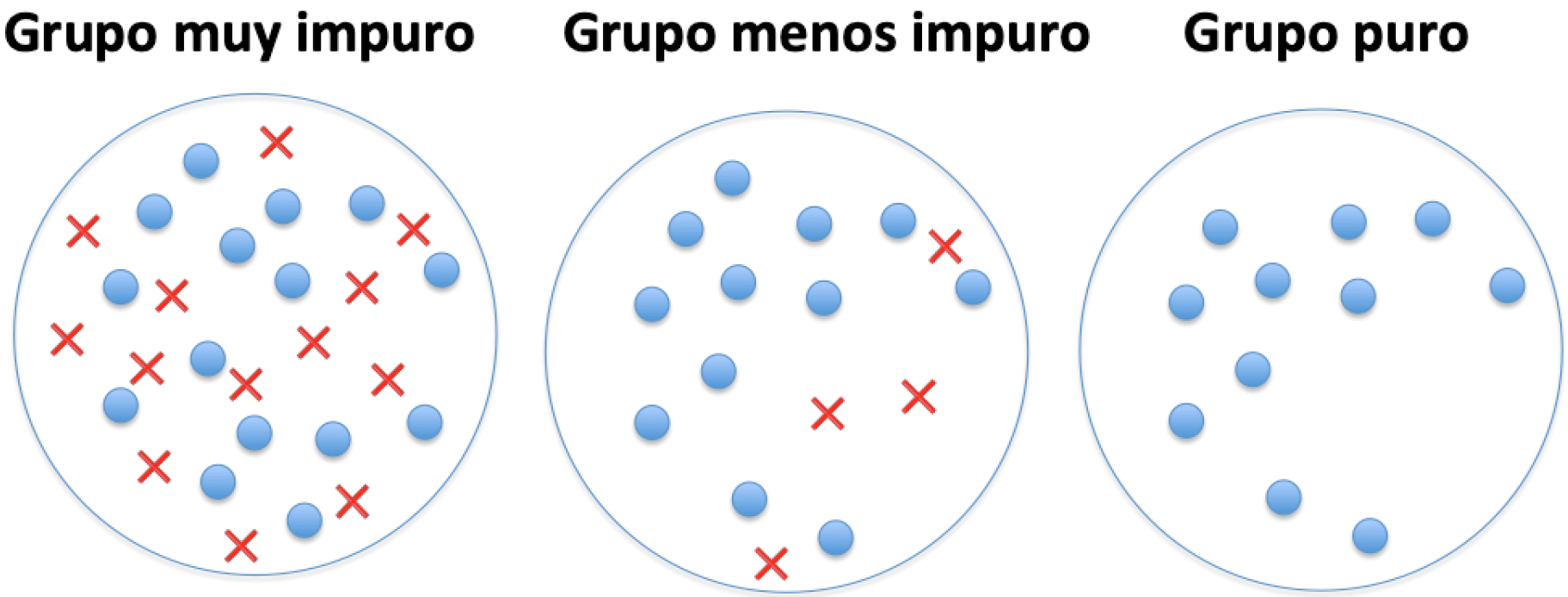
Variable A



Variable B

MiDaS

Necesitamos una medida que cuantifique el nivel de impureza en un grupo.




Interesa determinar qué atributo es el más eficiente para discriminar entre las clases.

MiDaS

## Ganancia de información

Selecciona como variable de división, aquella que logra “ordenar" mejor los datos en las categorías.

## Variables de control en la construcción de un árbol

- Profundidad máxima: número máximo de niveles de crecimiento del árbol.
- Tamaño mínimo de nodo: cantidad mínima de observaciones en un subgrupo para intentar subdividirlo nuevamente.
- Costo de complejidad: medida que balancea el tamaño de un árbol con la exactitud que logra en la clasificación.

MiDaS

## En R

Se necesitan los siguientes paquetes:

- `library(rpart)`
- `library(rpart.plot)`

| Nombre de la función | Utilidad |
|---|---|
| `rpart()` | ajusta el modelo de clasificación según el criterio de ganancia de información. |
| `rpart.plot()` | grafica el árbol obtenido con `rpart`. |
| `rpart.control()` | ajuste de parámetros de control del árbol a construir. |

MiDaS

# Ejemplo del tenis en R

## Lectura de datos

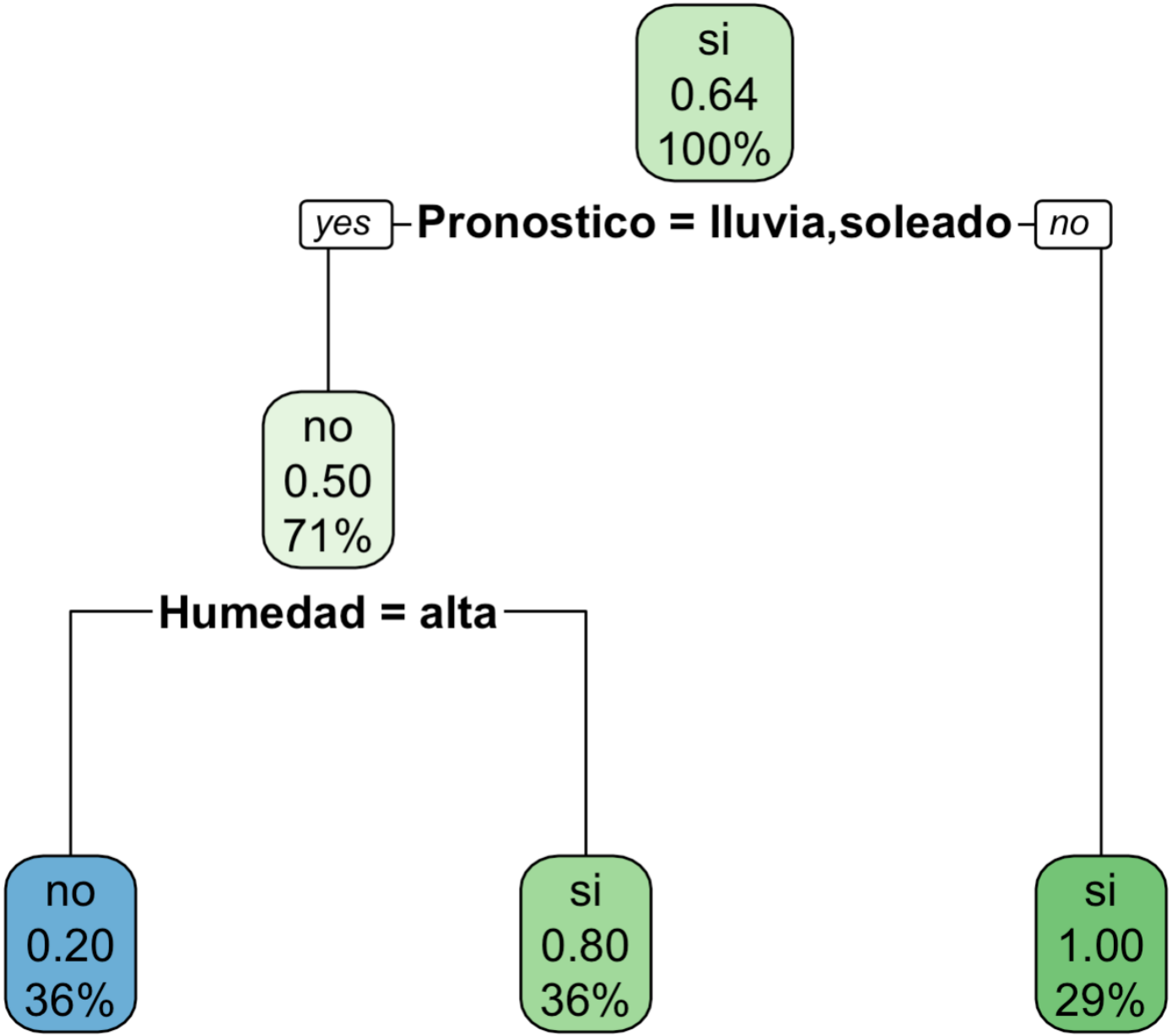
```
> datos <- read.csv("datos.csv", header = T, sep = ";")

> head(datos)
  Temperatura Pronostico Humedad Viento Tenis
1    caluroso    nublado    alta     no    si
2    caluroso    nublado  normal     no    si
3    caluroso    soleado    alta     no    no
4    caluroso    soleado    alta     si    no
5        frio     lluvia  normal     no    si
6        frio     lluvia  normal     si    no
```

MiDaS

## Construcción del árbol

```
> Formula <- Tenis ~ Temperatura + Pronostico +
+             Humedad + Viento


> arbol1 <- rpart(Formula, data = datos, method = 'class',
+                 parms = list(split = "information"),
+                 minsplit = 5)


> rpart.plot(arbol1)
```

MiDaS


si
0.64
100%
yes
Pronostico = lluvia,soleado
no
no
0.50
71%
Humedad = alta
no
0.20
36%
si
0.80
36%
si
1.00
29%

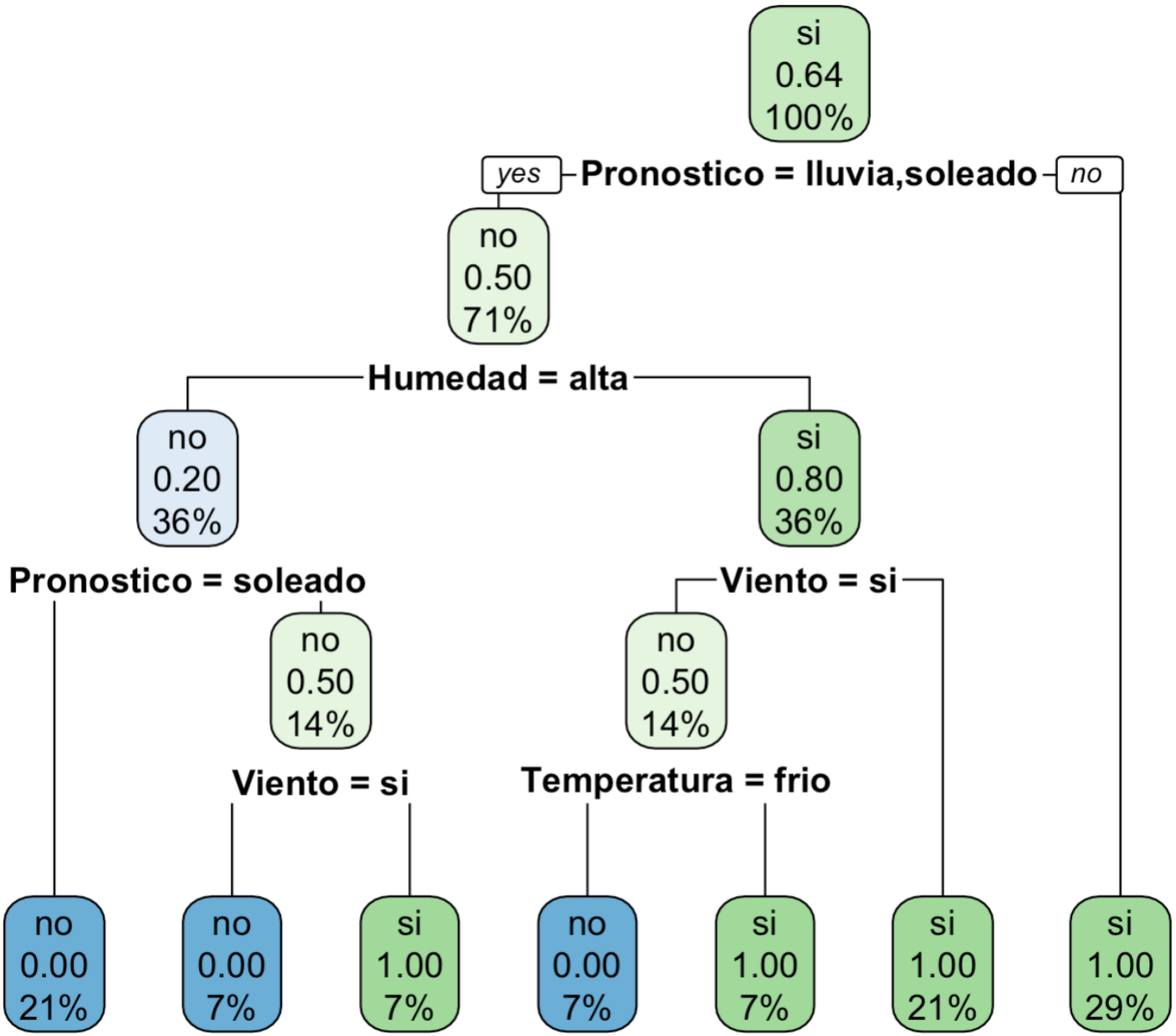
MiDaS

## ¿Cómo podemos obtener un árbol de diferente tamaño?

La alternativa que usaremos para esto es permitir que el árbol siga subdividiéndose, modificando el valor que regula el mínimo de observaciones que debe tener un nodo para intentar subdividirlo.

```
> arbol2 <- rpart(myFormula1, data = datos,
+                 method = 'class',
+                 parms = list(split = "information"),
+                 minsplit = 1)

> rpart.plot(arbol2)
```

MiDaS


si
0.64
100%
yes
Pronostico = lluvia,soleado
no
no
0.50
71%
Humedad = alta
no
0.20
36%
si
0.80
36%
Pronostico = soleado
Viento = si
no
0.50
14%
no
0.50
14%
Viento = si
Temperatura = frio
no
0.00
21%
no
0.00
7%
si
1.00
7%
no
0.00
7%
si
1.00
7%
si
1.00
21%
si
1.00
29%

MiDaS

# Atributos continuos

El algoritmo determina el mejor punto de corte para un atributo continuo, *A*, de la siguiente manera:

1. Ordena crecientemente los valores de la variable *A*.
2. Evalúa todos los puntos medios entre valores adyacentes.
3. En cada uno de ellos, evalúa el “orden", donde el número de particiones es 2.
4. Una vez que los evaluó todos, selecciona como punto de corte aquel umbral que logra “ordenar" de mejor manera los datos.
5. Así, el conjunto de datos se divide en 2 tal que $A \leq$ *punto corte*, y $A >$ *punto corte*.

MiDaS

# Evaluación y Selección de Modelos

- Una vez construido un modelo de clasificación, queremos evaluar su comportamiento con nuevos datos
- Necesario para comparar diferentes clasificadores

MiDaS

# Métricas para evaluar un clasificador

Debemos evaluar el rendimiento del clasificador, idealmente, con un conjunto de datos diferente al de trabajo, para evitar una sobreevaluación del modelo. Tomaremos un escenario con solo 2 clases.

- Se denominan *Verdaderos Positivos (VP)* a aquellos registos que han sido clasificados en la categoría de interés (Sí) y que realmente lo son.
- Se denominan *Verdaderos Negativos (VN)* a aquellos registros que **no** han sido clasificados en la categoría de interés y que realmente **no** lo son.

MiDaS

# Matriz de confusión

| | | Clase Predicha | | |
|---|---|---|---|---|
| | Clases | Sí | No | Total |
| Clase Real | Sí | VP | FN | P |
| | No | FP | VN | N |
| | Total | P' | N' | P + N |

MiDaS

## Tasa de clasificación correcta

Corresponde al porcentaje de registros que son correctamente clasificados.

$$\text{tasa de clasificación correcta} = \frac{VP + VN}{P + N}$$

## Tasa de clasificación incorrecta o de error

Corresponde al porcentaje de registros que están mal clasificados.

$$\text{tasa de error} = \frac{FP + FN}{P + N}$$

MiDaS

# Efecto de un desequilibrio en las clases

- La clase de interés puede ser rara o poco frecuente.
- Ejemplos: detección de fraudes, algunas enfermedades, etc.
- Un clasificador puede ser preciso para clasificar registros libres de cáncer, pero erróneo en clasificar aquellos que sí lo presentan.
- Necesitamos medidas que indiquen la bondad del clasificador en registros positivos y negativos, por separado.

MiDaS

## Sensibilidad

También se conoce como tasa de verdaderos positivos y corresponde a la proporción de registros positivos que están correctamente identificados.

$$\text{sensibilidad} = \frac{VP}{P}$$

## Especificidad

También se conoce como tasa de verdaderos negativos y corresponde a la proporción de registros negativos que están correctamente identificados.

$$\text{especificidad} = \frac{VN}{N}$$

## Nota

La tasa de reconocimiento es una función de la sensibilidad y la especificidad:

$$\text{tasa de clasificación correcta} = \left(\frac{P}{P+N}\right) \text{ sensibilidad} + \left(\frac{N}{P+N}\right) \text{ especificidad}$$

MiDaS

## En R

```
> (pred1 <- predict(arbol1, type = "class"))
 1  2  3  4  5  6  7  8  9 10 11 12 13 14
si si no no si si si si si no no si no si
Levels: no si
> table(pred1, datos$Tenis)

pred1 no si
   no  4  1
   si  1  8
> (pred2 <- predict(arbol2, type = "class"))
 1  2  3  4  5  6  7  8  9 10 11 12 13 14
si si no no si no si si si no si si no si
Levels: no si
> table(pred2, datos$Tenis)

pred2 no si
   no  5  0
   si  0  9
```

# Ejemplo

La base de datos **`titanic.csv`** contiene información de:

| Variable | Descripción | |
|---|---|---|
| `Sobrevive` | Indicador de sobrevivencia | 1 = Sí, 0 = No |
| `Clase` | Clase en que viajaba | 1, 2, 3 |
| `Sexo` | Hombre, Mujer | |
| `Edad` | Edad en años | |
| `Hnos_Conyuges` | Cant. de hnos. o cónyuges en el barco | |
| `Padres_Hijos` | Cant. de padres o hijos en el barco | |
| `Tarifa` | En libras | |

El objetivo es poder predecir si un pasajero sobrevive o no en base a sus características.

MiDaS

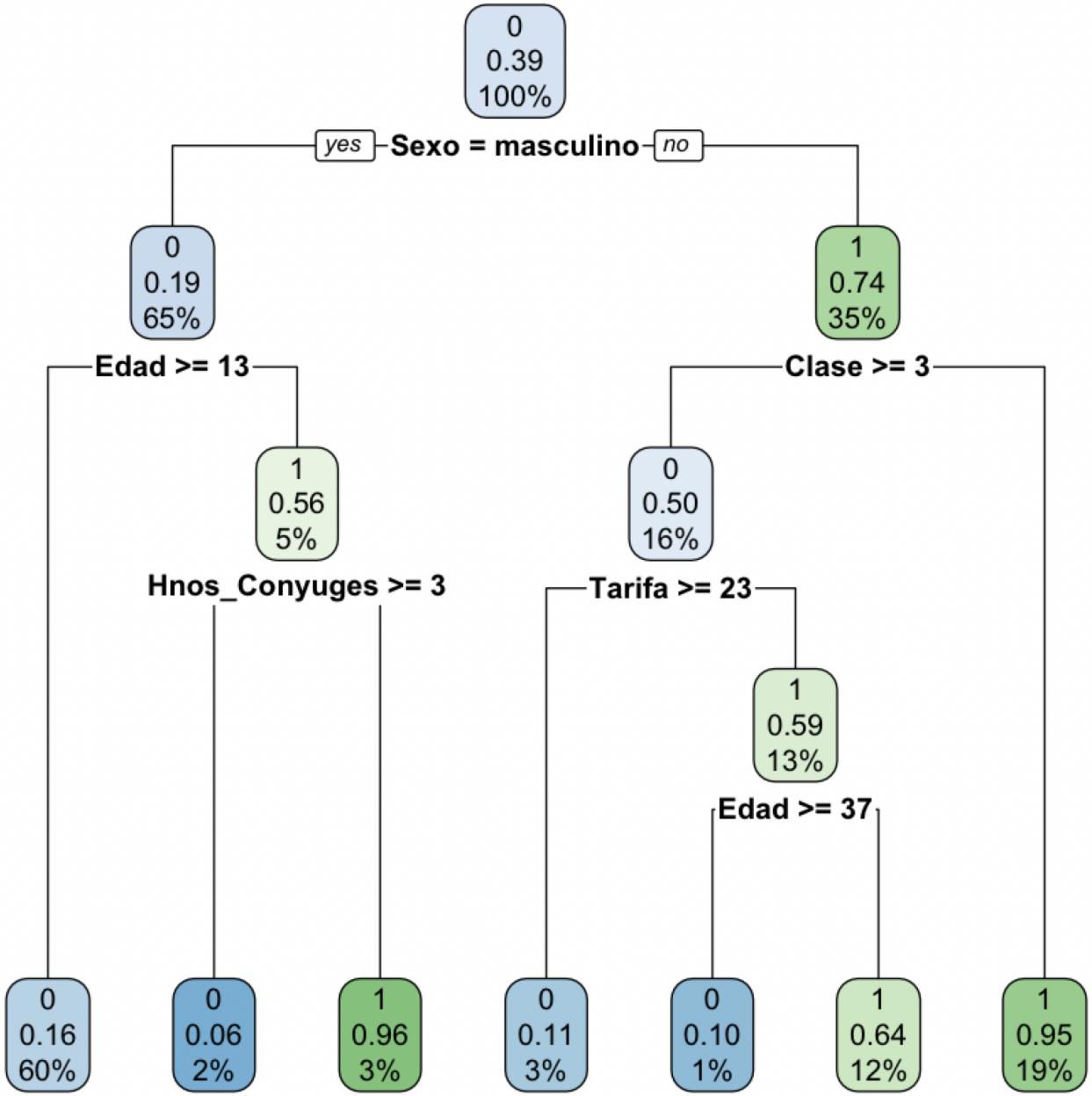
# Lectura de datos

```
> titanic <- read.csv("titanic.csv" , header = T,
+                     sep = ";", dec = ",")
>
> head(titanic)
  Sobrevive Clase      Sexo Edad Hnos_Conyuges
1         0     3 masculino   22             1
2         1     1  femenino   38             1
3         1     3  femenino   26             0
4         1     1  femenino   35             1
5         0     3 masculino   35             0
6         0     3 masculino   27             0

  Padres_Hijos  Tarifa
1            0  7.2500
2            0 71.2833
3            0  7.9250
4            0 53.1000
5            0  8.0500
6            0  8.4583
```

## Construcción del árbol

```
> Formula <- Sobrevive ~ Clase + Sexo + Edad + Hnos_Conyuges +
+                        Padres_Hijos + Tarifa

> arbol.t <- rpart(Formula, data = titanic, method = 'class',
+                  parms = list(split = "information"),
+                  minsplit = 20 )
```

MiDaS


0
0.39
100%
yes
Sexo = masculino
no
0
0.19
65%
1
0.74
35%
Edad >= 13
Clase >= 3
1
0.56
5%
0
0.50
16%
Hnos_Conyuges >= 3
Tarifa >= 23
1
0.59
13%
Edad >= 37
0
0.16
60%
0
0.06
2%
1
0.96
3%
0
0.11
3%
0
0.10
1%
1
0.64
12%
1
0.95
19%

MiDaS

## Matriz de confusión

```
> pred <- predict(arbol.t, type = "class")

> conf <- table(Predicho = pred,
+               Real = titanic$Sobrevive)

> conf
        Real
Predicho   0   1
       0 496  91
       1  49 251
```

MiDaS

# Evaluación del rendimiento del modelo

```
> conf
         Real
Predicho   0   1
       0 496  91
       1  49 251
> (vn <- conf[1,1])
[1] 496
> (vp <- conf[2,2])
[1] 251
> (fp <- conf[2,1])
[1] 49
> (fn <- conf[1,2])
[1] 91

> (tcc <- (vp + vn) / sum(conf))
[1] 0.8421646
> (tci <- (fp + fn) / sum(conf))
[1] 0.1578354
> tcc + tci
[1] 1
```

# Evaluación del rendimiento del modelo (cont.)

```
> conf
         Real
Predicho   0   1
       0 496  91
       1  49 251
> (sens <- vp / (vp + fn))
[1] 0.7339181

> (esp <- vn / (vn + fp))
[1] 0.9100917

> (vp+fn)/sum(conf)*sens + (fp+vn)/sum(conf)*esp
[1] 0.8421646
> tcc
[1] 0.8421646

> table(titanic$Sobrevive)
  0   1
545 342
```