(Hitz Research Center), Departamento de LSI, EHU

# Sist. de Ayuda a la Decisión:

## Clasificación: Arboles de Decisión

**AitZiber AtutXa**
aitziber.atucha@ehu.eus

marzo de 2023

# Índice

# Objetivos de Aprendizaje

- Ser capaces de explicar la intuición tras el algoritmo de *Random Forest*.
- Ser capaces de formalizar matemáticamente el algoritmo.
- Ser capaces de explicar el rol del *bagging* en el algoritmo.
- Entender las ventajas que aporta frente a un Arbol de Decisión Simple

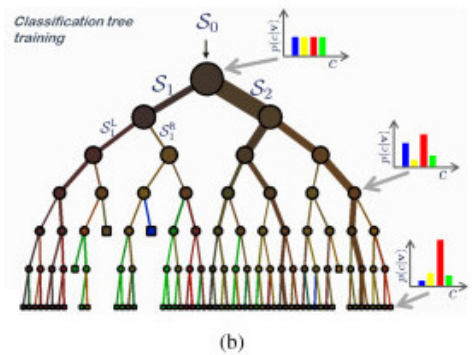
# Motivación

Mejorar los árboles de decisión:

- Mejorar la calidad de las predicciones.
- Mejorar la eficiencia a través de la paralelización.

# Recapitulando

| | Alt | ... | Lleno | ... | Tipo | Estim | Esp? |
|---|---|---|---|---|---|---|---|
| $x^{\{1\}}$ | Si | ... | Algo | ... | Frances | 0-10 | Si |
| $x^{\{2\}}$ | Si | ... | Lleno | ... | Thai | 30-60 | No |
| $x^{\{3\}}$ | No | ... | Algo | ... | Burger | 0-10 | Si |
| $x^{\{4\}}$ | Si | ... | Lleno | ... | Thai | 10-30 | Si |
| $x^{\{5\}}$ | Si | ... | Lleno | ... | Frances | >60 | No |
| $x^{\{6\}}$ | No | ... | Algo | ... | Italiano | 0-10 | Si |
| $x^{\{7\}}$ | No | ... | Nada | ... | Burger | 0-10 | No |
| $x^{\{8\}}$ | No | ... | Algo | ... | Thai | 0-10 | Si |
| $x^{\{9\}}$ | No | ... | Lleno | ... | Burger | >60 | No |
| $x^{\{10\}}$ | Si | ... | Lleno | ... | Italiano | 10-30 | No |
| $x^{\{11\}}$ | No | ... | Nada | ... | Thai | 0-10 | No |
| $x^{\{12\}}$ | Si | ... | Lleno | ... | Burger | 30-60 | Si |

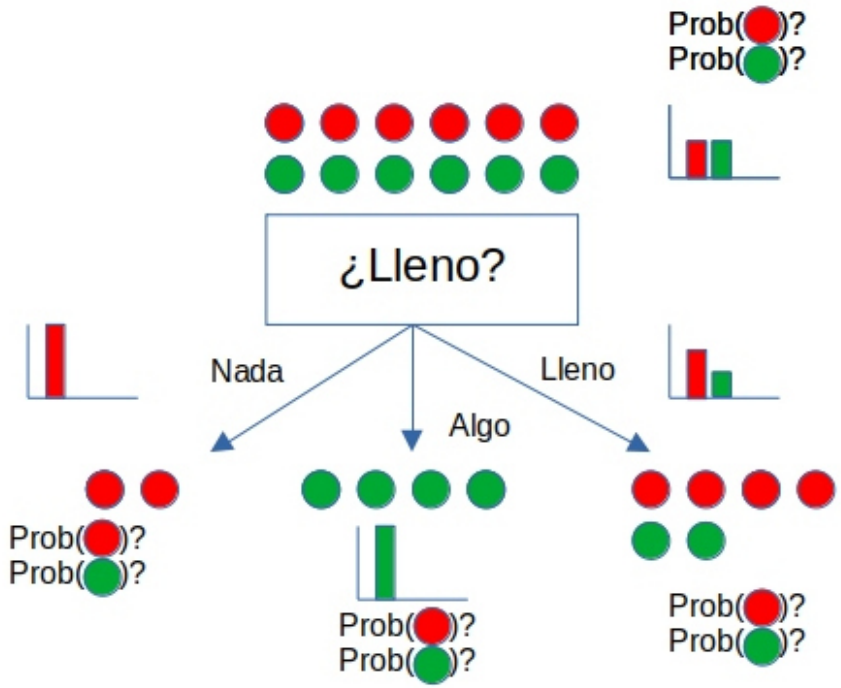


(b)

imagen: Crisminisi et al. 2011

# Construyendo el Arbol

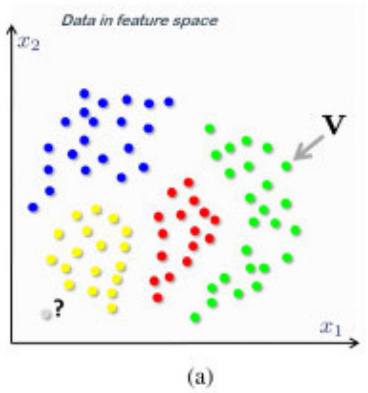
| | Lleno | HacerEsp |
|---|---|---|
| $x^{\{1\}}$ | Algo | Si |
| $x^{\{2\}}$ | Lleno | No |
| $x^{\{3\}}$ | Algo | Si |
| $x^{\{4\}}$ | Lleno | Si |
| $x^{\{5\}}$ | Lleno | No |
| $x^{\{6\}}$ | Algo | Si |
| $x^{\{7\}}$ | Nada | No |
| $x^{\{8\}}$ | Algo | Si |
| $x^{\{9\}}$ | Lleno | No |
| $x^{\{10\}}$ | Lleno | No |
| $x^{\{11\}}$ | Nada | No |
| $x^{\{12\}}$ | Lleno | Si |




Ganancia de Información o Indice Gini

# ¿Cómo tratar los atributos continuos en DT?

- Preprocesado (no supervisado):
  - binning por amplitud: (max-min)/numBins
  - bining por frecuencia: todos las alternativas mismo número de instancias
- Supervisado:



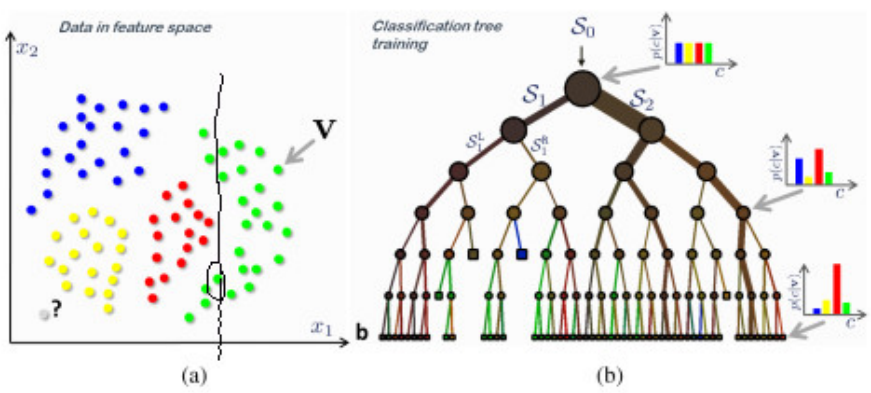




imagen: Crisminisi et al. 2011

# Construyendo el Arbol

## ¿Cómo mejoramos los DTs?

$\mathcal{X}^{train} = \{x^{\{1\}}, x^{\{2\}}, ...x^{\{n\}}\}\ n = 5(5^{instancias})$

$F = \{f_1, f_2, ..., f_j\}\ j = 3(3_{atributos})$

¿Cuál es el valor $x_3^{\{2\}}$?

| $\mathcal{X}^{train}$ | $f_1$ | $f_2$ | $f_3$ |
|---|---|---|---|
| $x^{\{1\}}$ | 1 | 0 | 2 |
| $x^{\{2\}}$ | 3 | 6 | 1 |
| $x^{\{3\}}$ | 0 | 2 | 4 |
| $x^{\{4\}}$ | 8 | 9 | 0 |
| $x^{\{5\}}$ | 5 | 5 | 1 |

| $ytrain$ |
|---|
| 0 |
| 1 |
| 0 |
| 1 |
| 0 |

Seleccionamos aleatoriamente 2 atributos. P.e. $f_1, f_3$ □

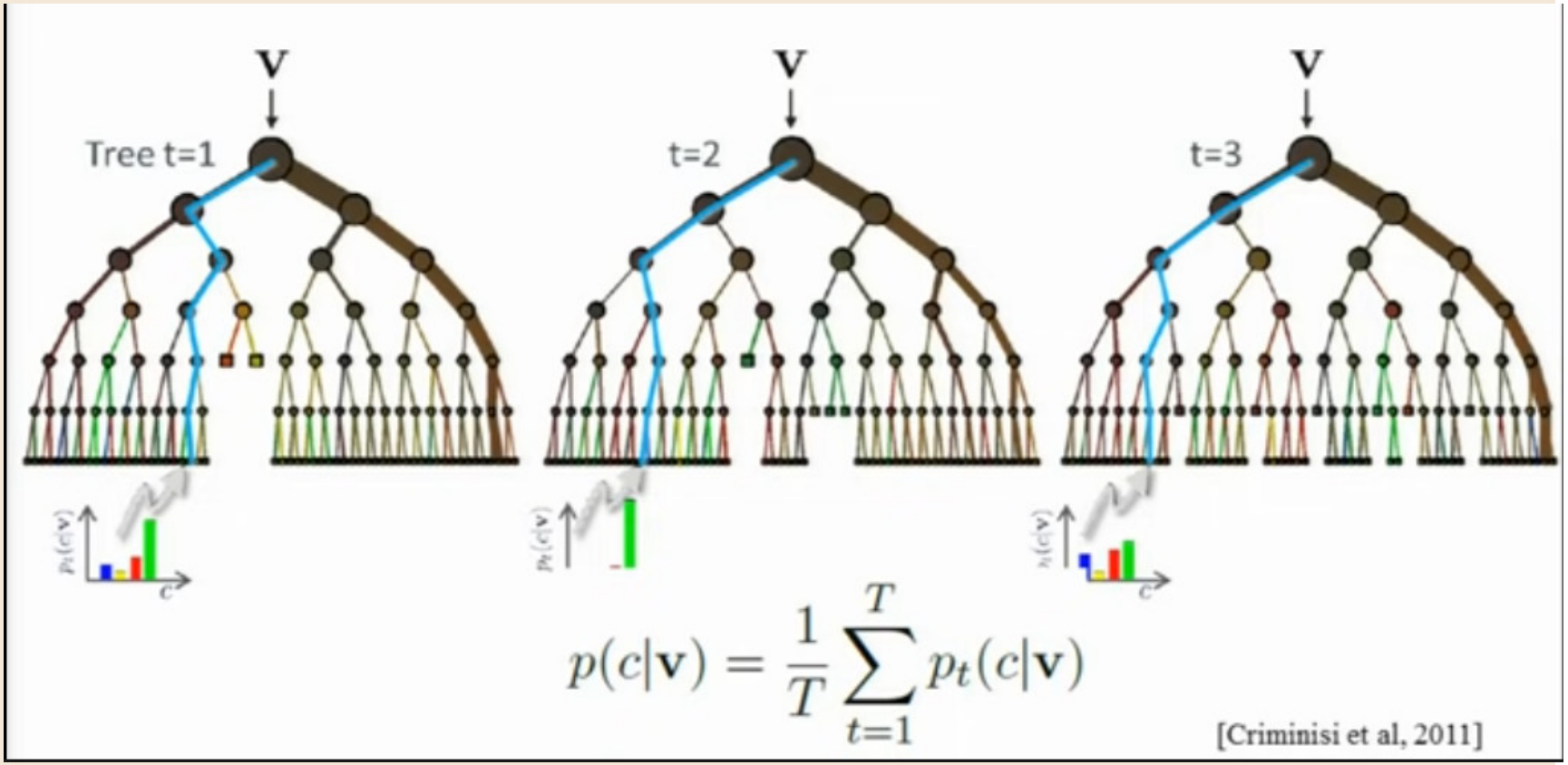
# Random Forest

**Algorithm 1** Random Forrest

**for** $t \in \mathcal{T}$ **do**
Selecciona una muestra $Z^*$ de tamaño N $\in$ $\mathcal{X}^{train}$ (**booststraping|bagging**)
Genera un Arbol de Decisión a partir de $Z^*$ hasta maxProfundidad o minTamañoNodo
Selecciona j atributos $f \in F$ aleatoriamente
Elige el mejor umbral (empleando IG o Gini)
Divide en las ramas
**end for**
Genera un modelo conjunto (ensemble) combinando los rdos a través de la media $prob(c|v) = \frac{1}{T} \sum_{t=1}^{T} p_t(c|v)$

# Random Forest

## ¿Puedo paralelizar con Bagging?



# Mejoras

¿Cuáles crees que son las mejoras frente a DT?.

¿Crees que es menos eficiente?

¿Crees que reduce el *sobreajuste (overfitting)*? Es decir, ¿generaliza mejor?

□

# ¿Y si las instancias son textos?

**¿Cómo construríais los árboles para un detector de Spam?**

- ¿Cómo representaríais vuestros emails? (tf-idf o one-hot BOW)
- ¿qué repercusiones tendría cada representación?

□

# Bibliografía


- ”Computer Graphics and Vision. Vol. 7, Nos. 2–3 (2011) 81–227”, A. Criminisi, J. Shotton and E. Konukoglu
- Vídeo con un ejemplo de cómo aplicar el índice Gini
  https://es.coursera.org/lecture/build-decision-trees-svms-neural-networks/gini-index-example-rPvWM