# UNIVERSIDAD NACIONAL DEL ALTIPLANO

Escuela Profesional del Ingeniería de Sistemas

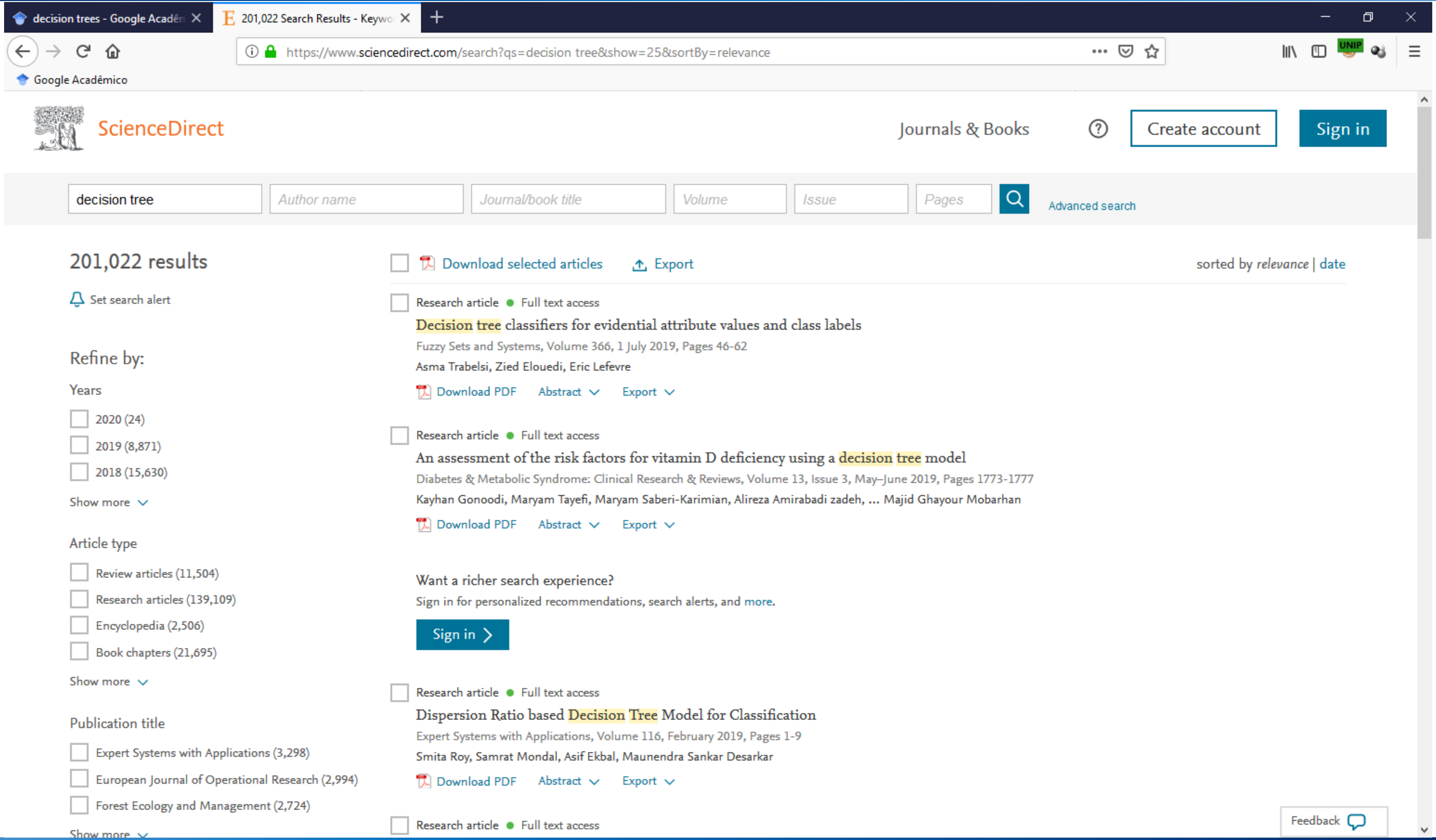
## LÓGICA PARA INTELIGENCIA ARTIFICIAL

## Árboles de Decisión

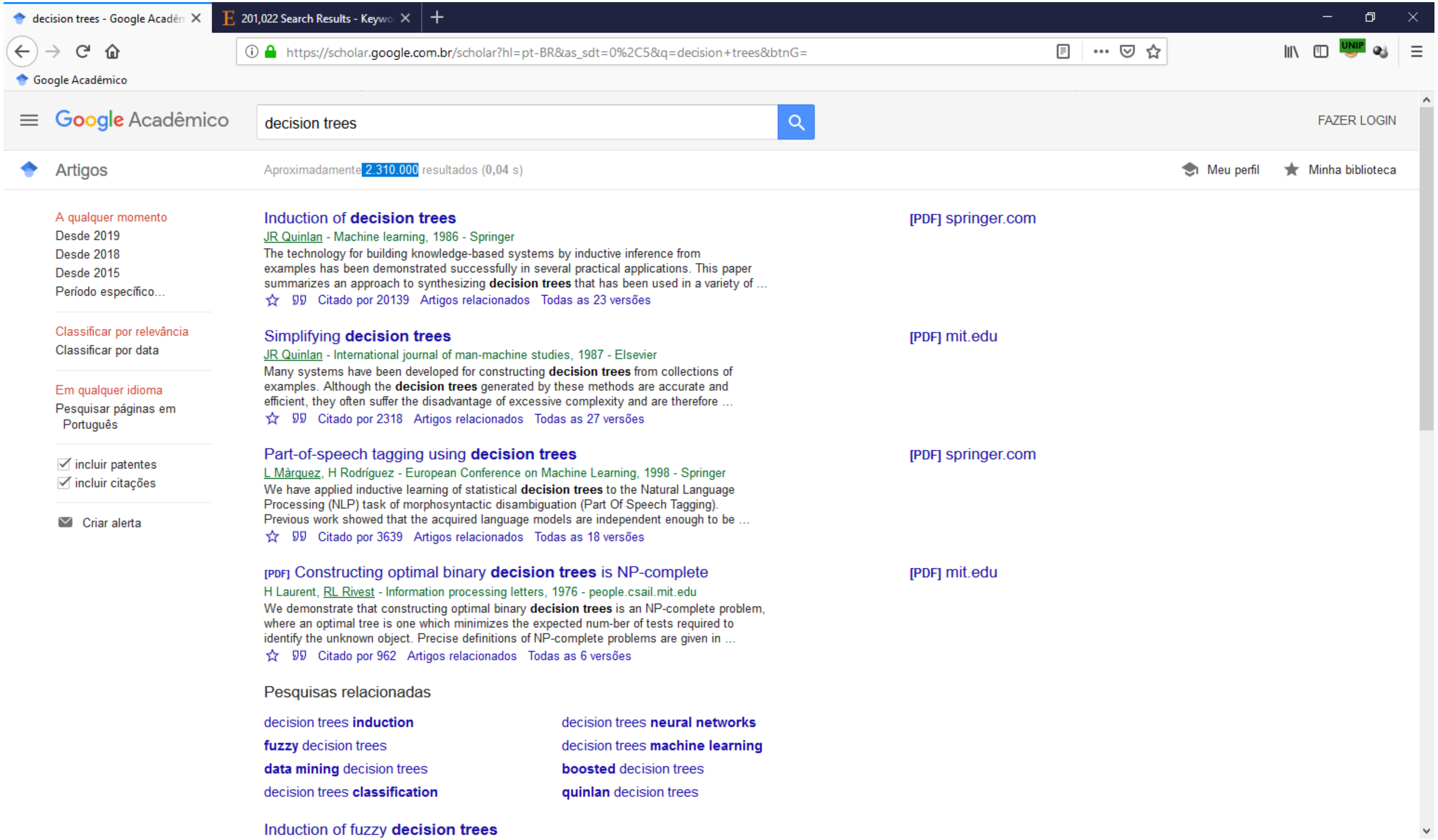
Prof. Ing. José Luis Juarez Ruelas

Puno, Mayo del 2019

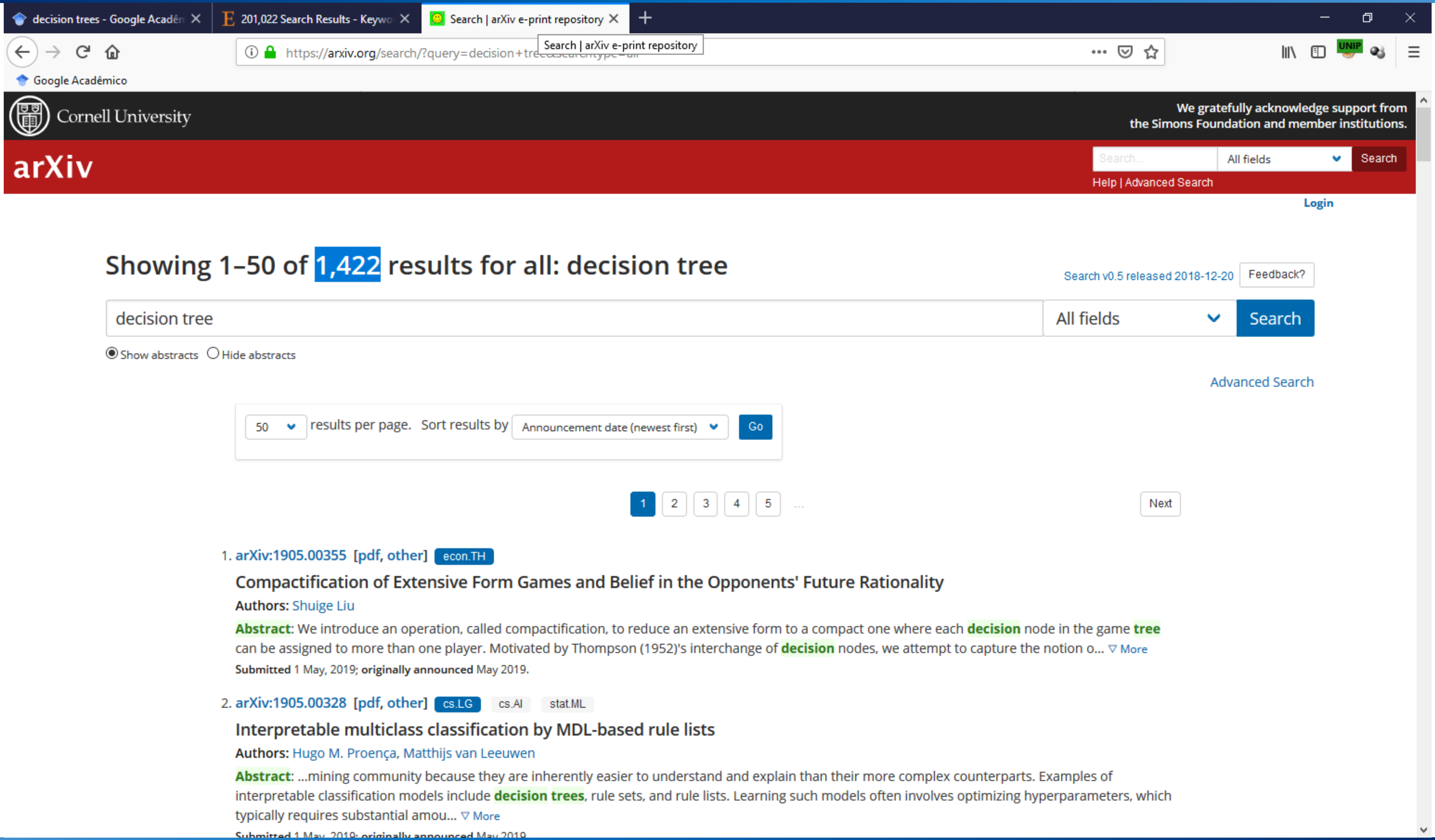
# Introducción



# Introducción



# Introducción



# Introducción

| Gender | Occupation | App |
|---|---|---|
| F | Study | Pokemon Go |
| F | Work | WhatsApp |
| M | Work | Snapchat |
| F | Work | WhatsApp |
| M | Study | Pokemon Go |
| M | Study | Pokemon Go |

**¿Cual aplicación recomendamos ?**

-Para una mujer que trabaja en oficina
-Para un hombre que trabaja en fabrica
-Para alguien que va al colegio

# Introducción

| Gender | Occupation | App |
|---|---|---|
| F | Study | Pokemon Go |
| F | Work | WhatsApp |
| M | Work | Snapchat |
| F | Work | WhatsApp |
| M | Study | Pokemon Go |
| M | Study | Pokemon Go |

# Introducción

| Gender | Occupation | App |
|---|---|---|
| F | Study | Pokemon Go |
| F | Work | WhatsApp |
| M | Work | Snapchat |
| F | Work | WhatsApp |
| M | Study | Pokemon Go |
| M | Study | Pokemon Go |

# Introducción

| Gender | Occupation | App |
|---|---|---|
| F | Study | Pokemon Go |
| F | Work | WhatsApp |
| M | Work | Snapchat |
| F | Work | WhatsApp |
| M | Study | Pokemon Go |
| M | Study | Pokemon Go |

Entre **género** y **ocupación** ¿Cuál variable le parece más decisivo para predecir la aplicación descargarán los usuarios?

# Introducción

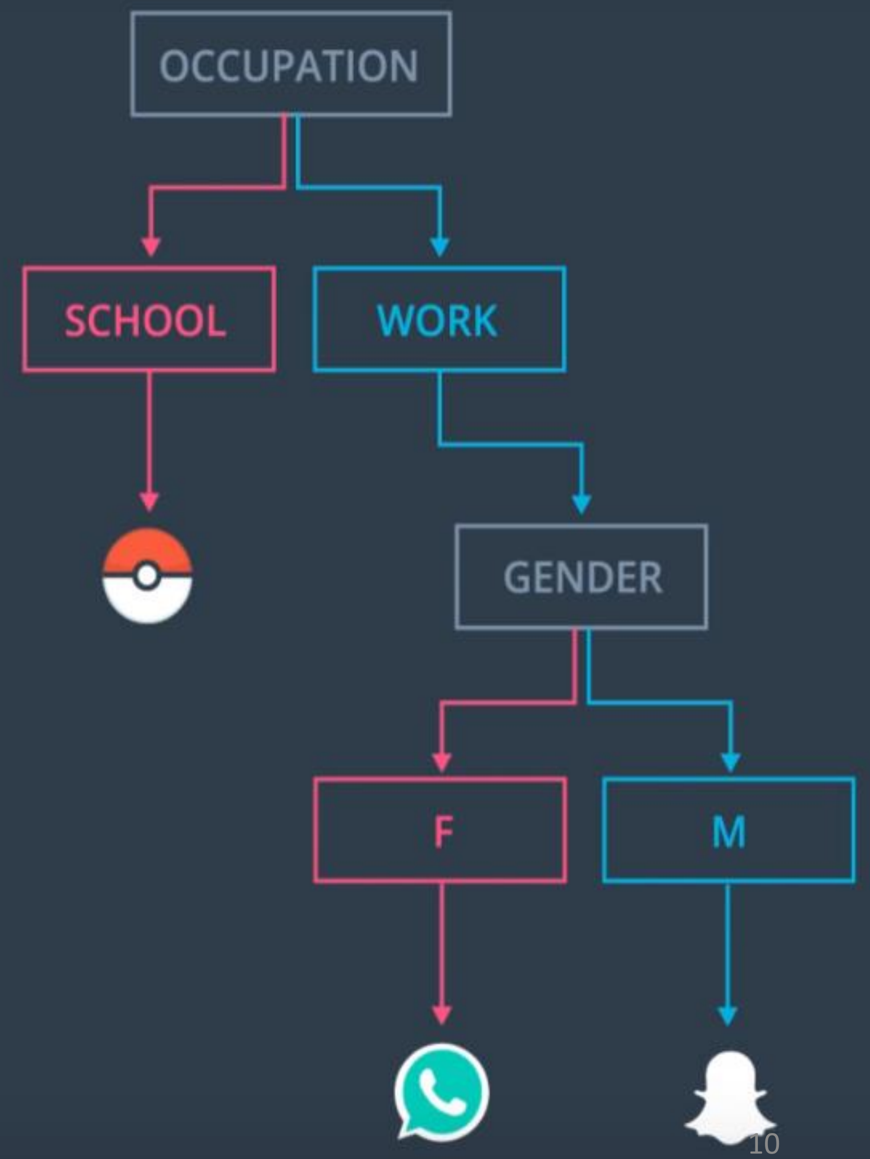
| Gender | Occupation | App |
|---|---|---|
| F | Study | Pokemon Go |
| F | Work | WhatsApp |
| M | Work | Snapchat |
| F | Work | WhatsApp |
| M | Study | Pokemon Go |
| M | Study | Pokemon Go |

Entre **género** y **ocupación**
¿Cuál variable le parece más decisivo para predecir qué aplicación descargarán los usuarios?

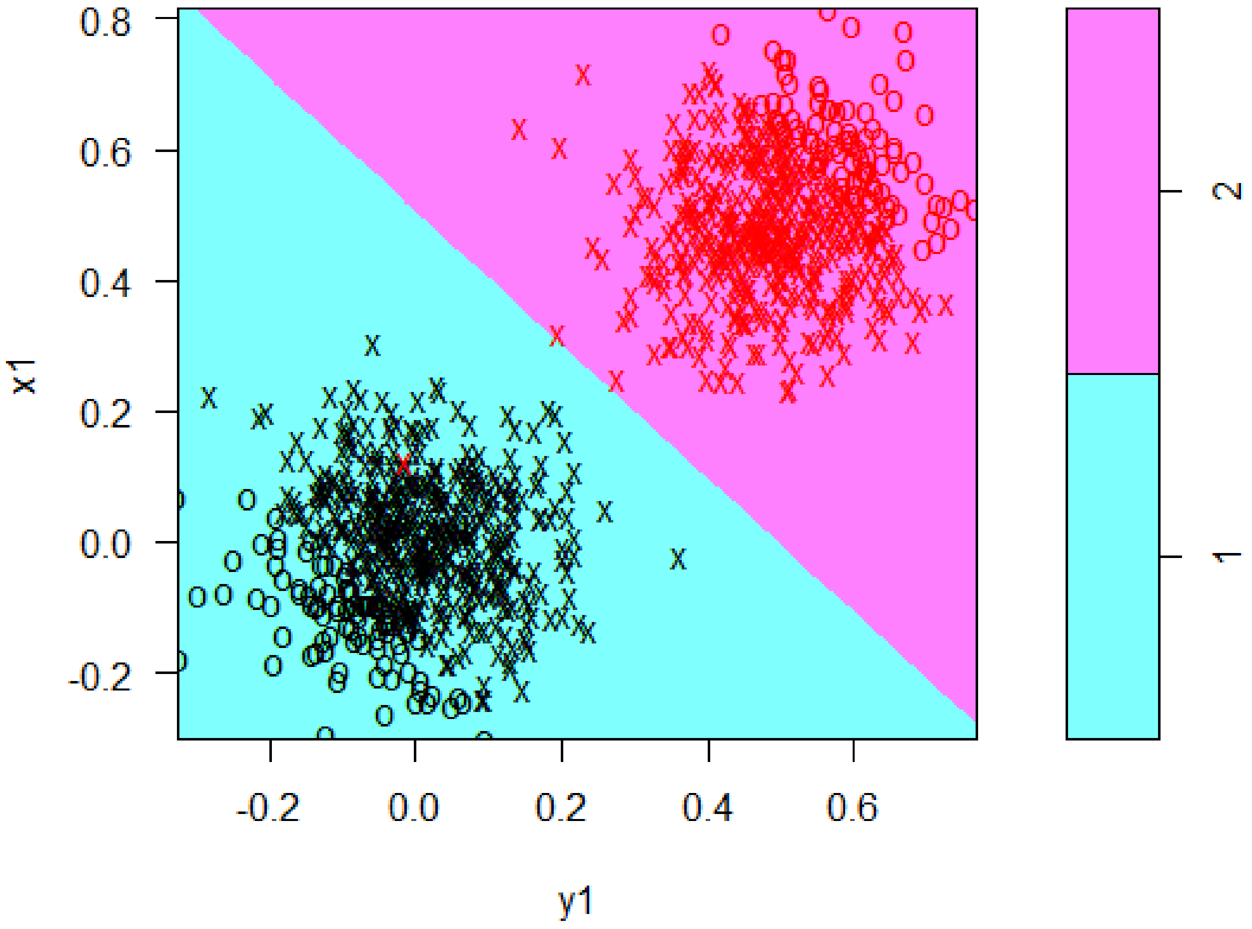
# Introducción

## Recommending Apps

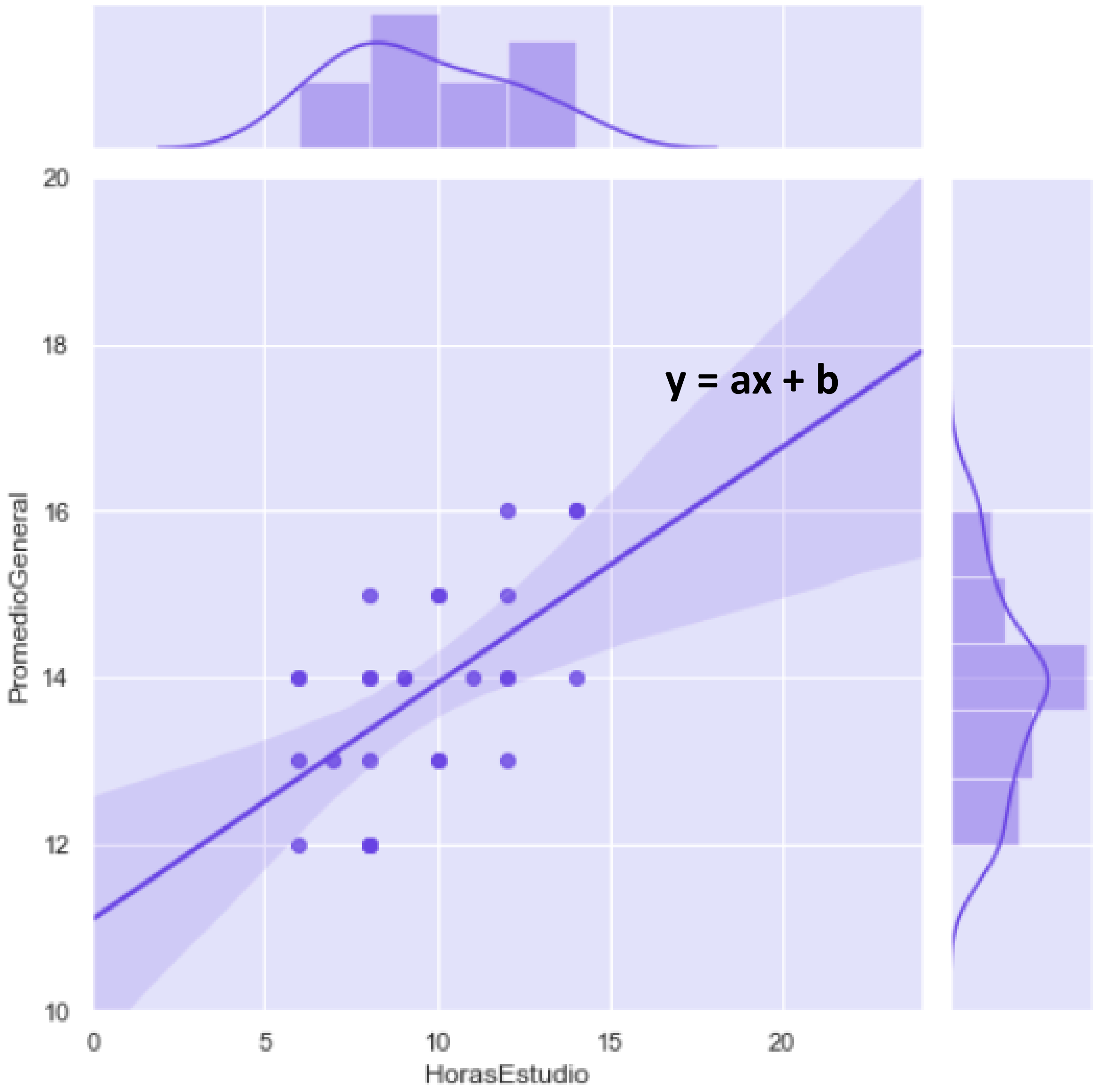
| Gender | Occupation | App |
|---|---|---|
| F | Study | |
| F | Work | |
| M | Work | |
| F | Work | |
| M | Study | |
| M | Study | |



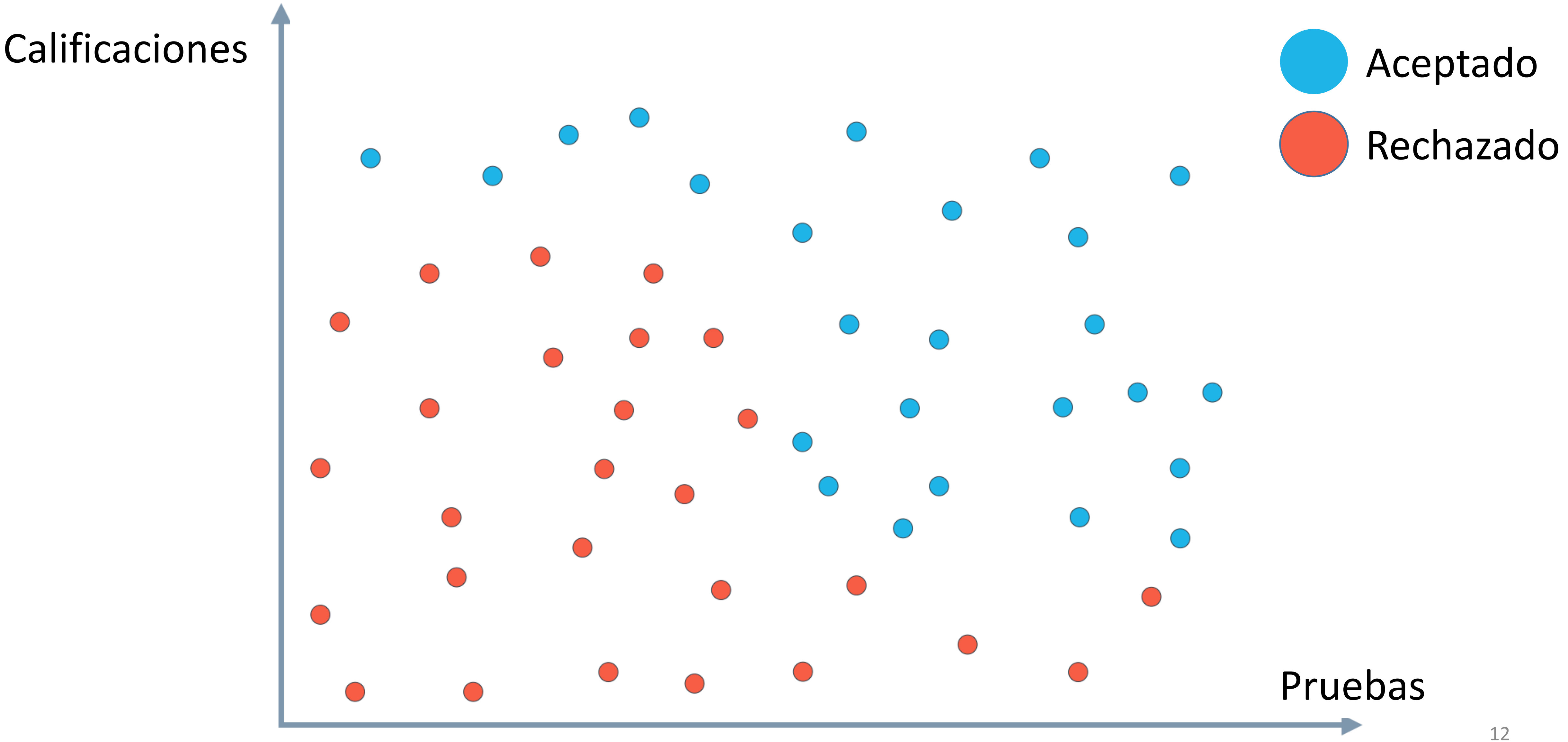
# Clasificación



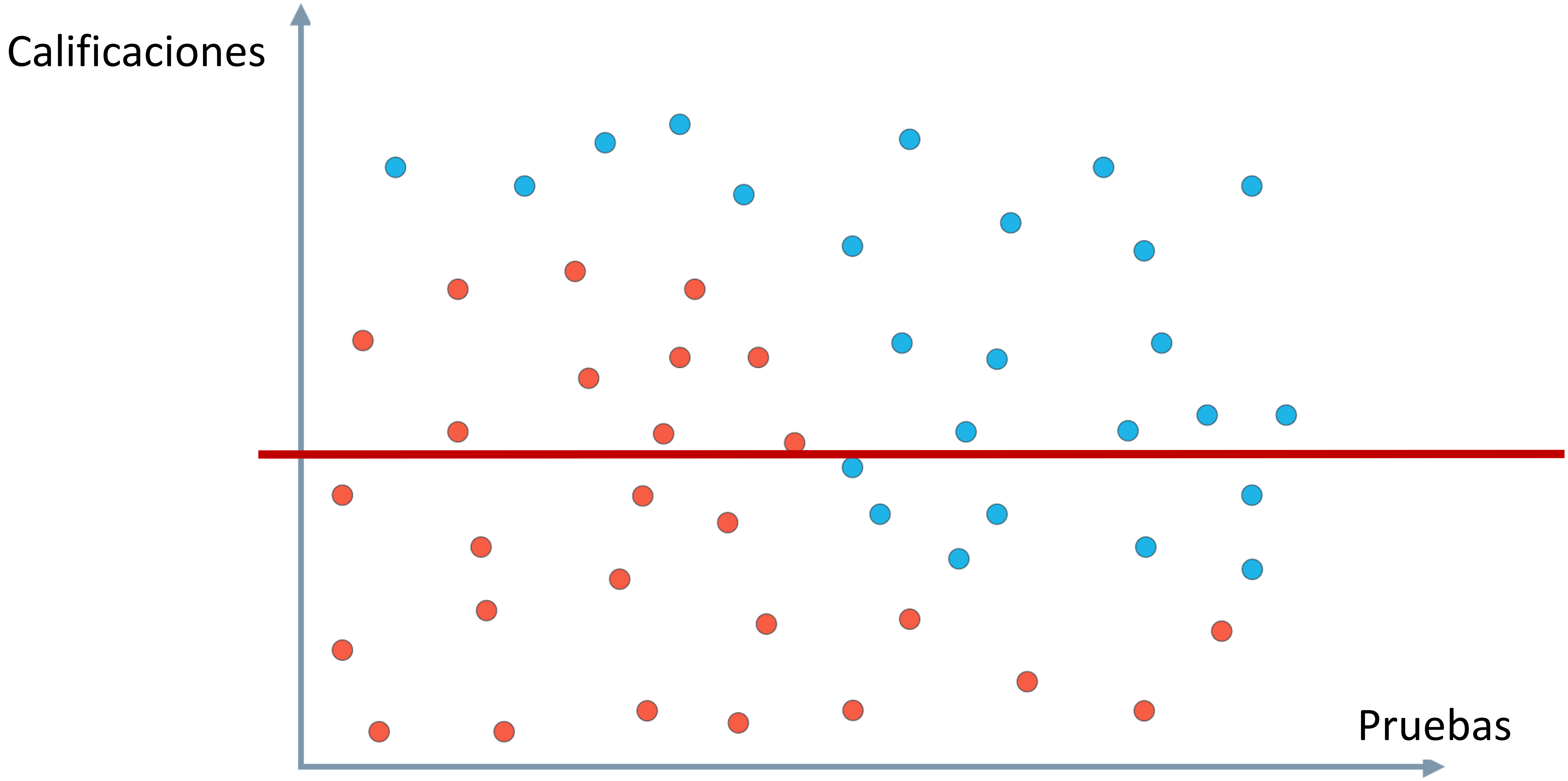

# Regresión



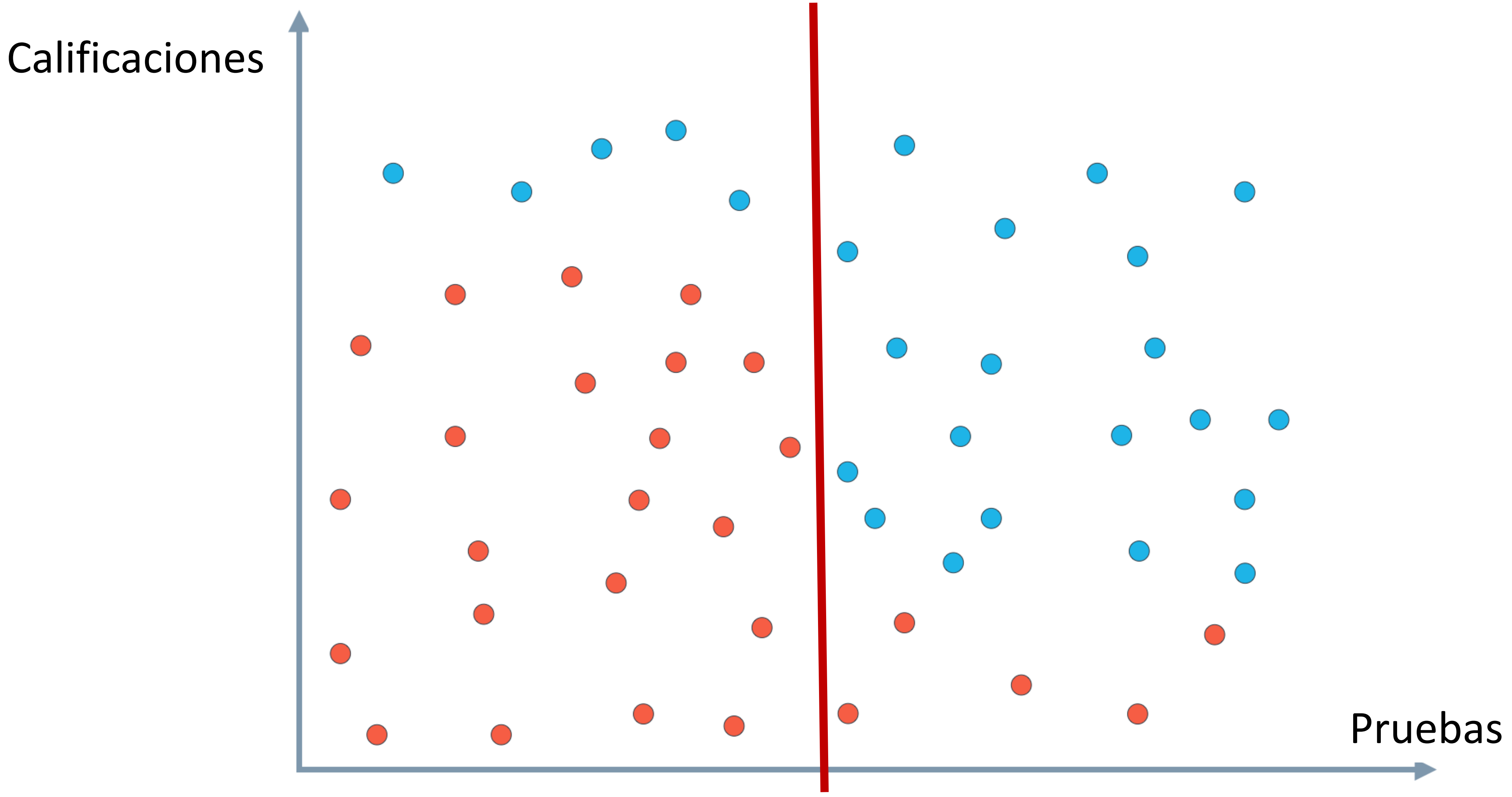
# La mejor recta de división para ser aceptado en un programa de estudios



# ¿Cuál es la mejor recta de división?

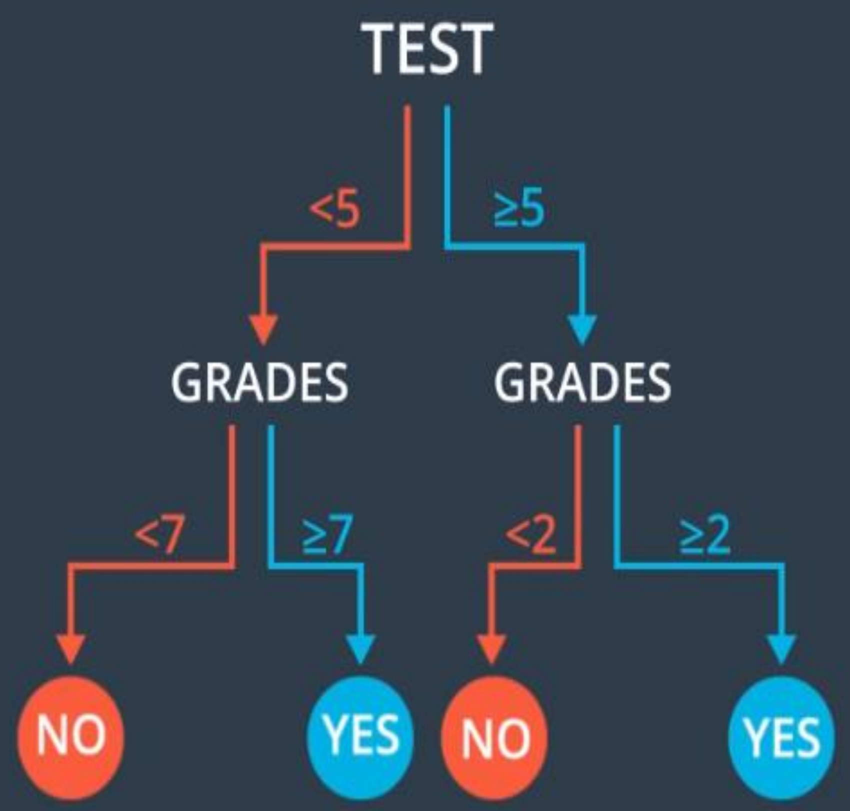


# La mejor recta de división



# Creando el árbol de decisión

# Creando el **árbol de decisión**

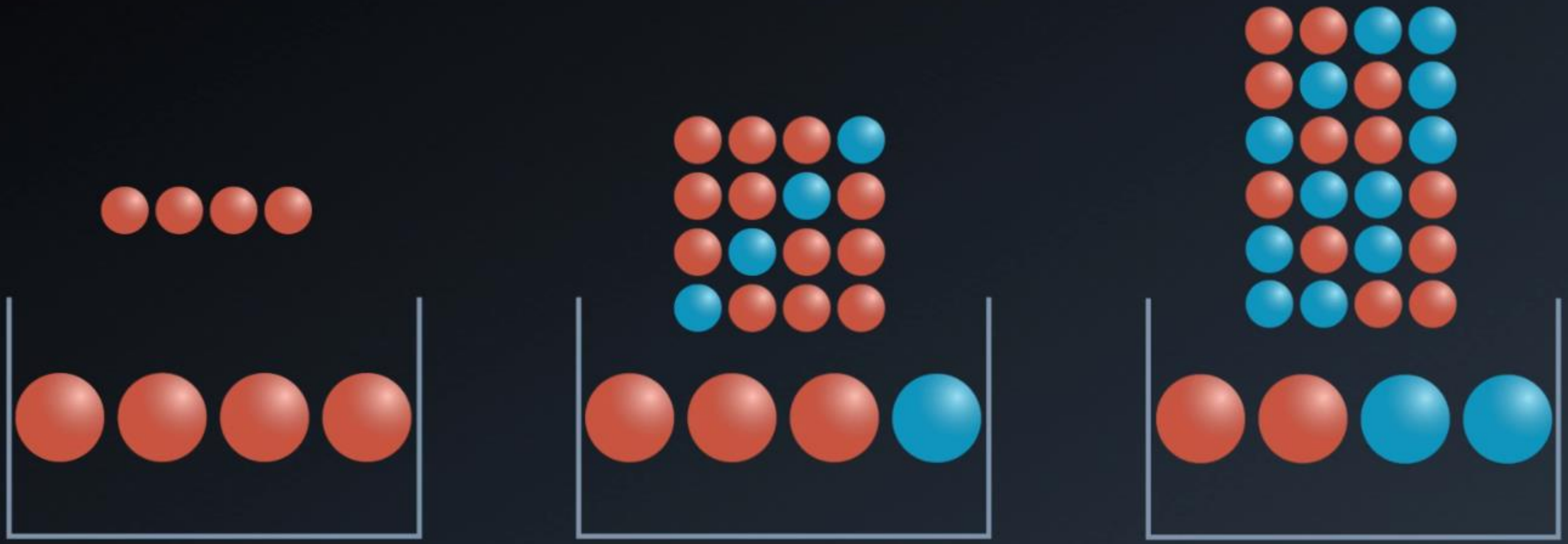



# Entropía


| Solido | Liquido | Vapor |
| --- | --- | --- |
| Entropía baja | Entropía media | Entropía alta |

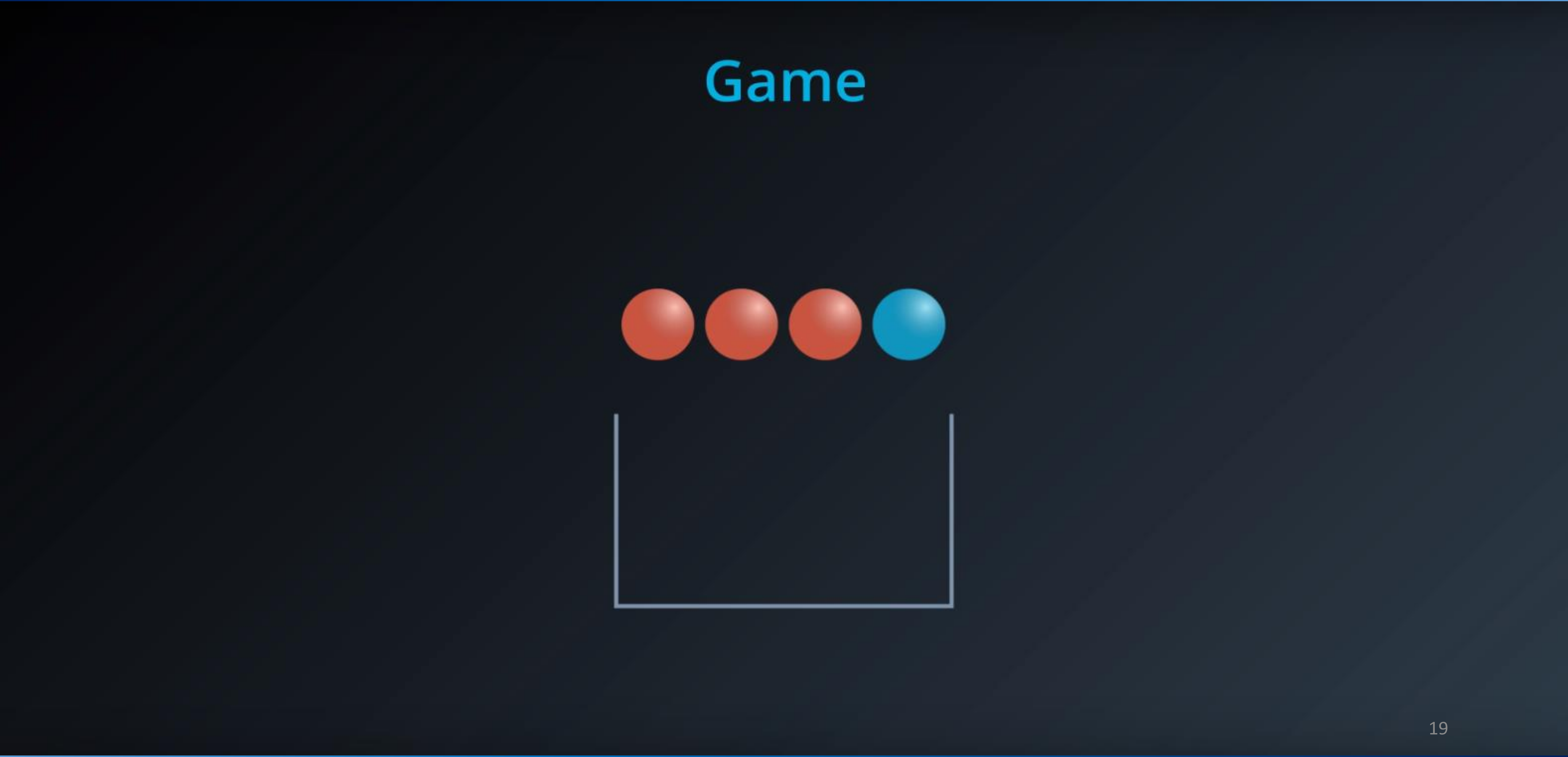
# Entropía

Def. En termodinámica, la entropía (simbolizada como S) es una magnitud física para un sistema termodinámico en equilibrio. Mide el número de microestados compatibles con el macroestado de equilibrio, también se puede decir que **<u>mide el grado de organización del sistema.</u>**

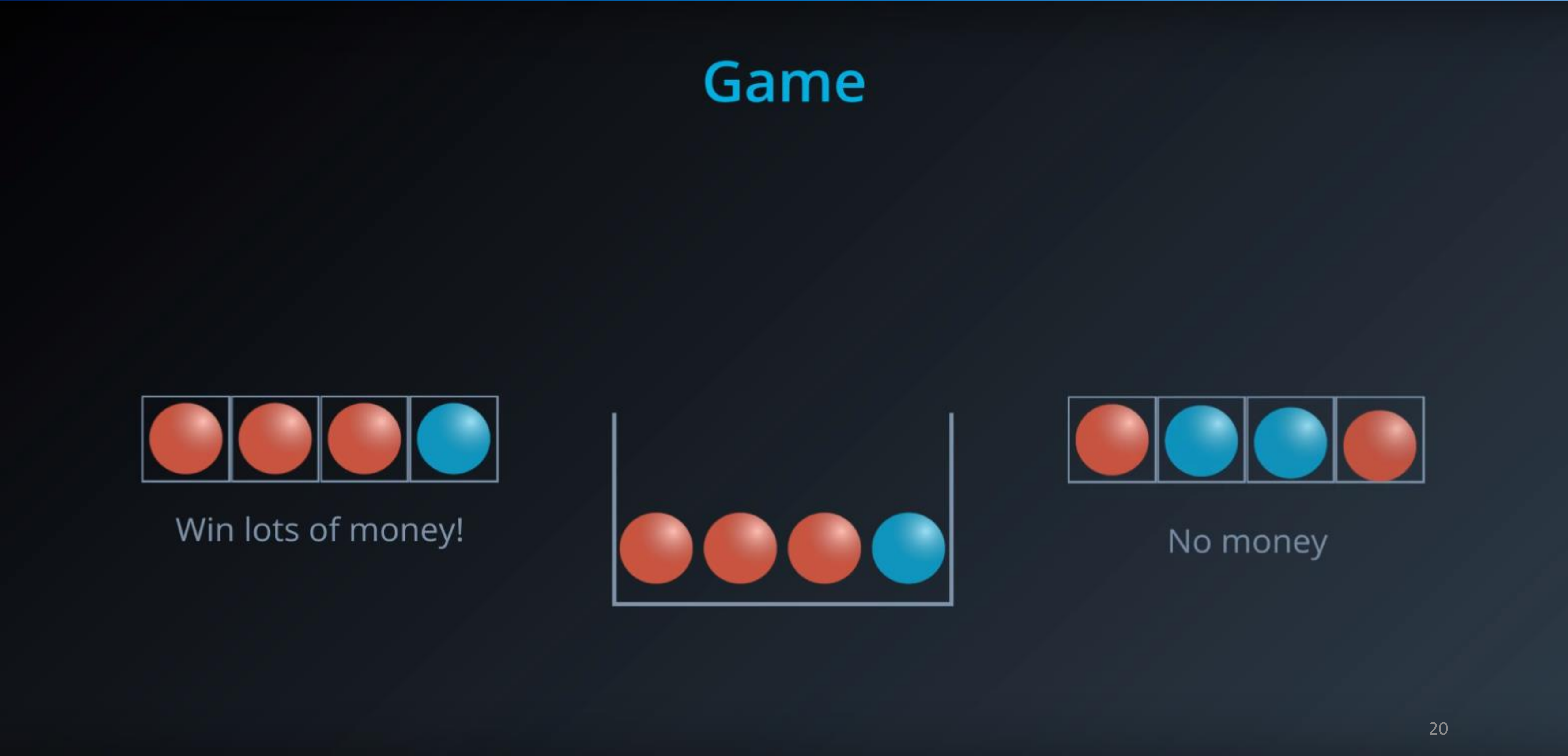


Entropía baja Entropía media Entropía alta

# Entropía



# Entropía



# Entropía

## Probability of Winning

| | $P$ $(red)$ | $P$ $(blue)$ | $P$ $(winning)$ |
|---|---|---|---|
|  | 1 | 1 | $1 \times 1 \times 1 \times 1 \boxed{= 1}$ |
|  | 0.75 | 0.25 | $0.75 \times 0.75 \times 0.75 \times 0.25 \boxed{= 0.105}$ |
|  | 0.5 | 0.5 | $0.5 \times 0.5 \times 0.5 \times 0.5 \boxed{= 0.0625}$ |

# ¿Qué función nos ayudará a convertir los productos en sumas?

## Products

$$0.75 * 0.75 * 0.75 * 0.25 = 0.105$$


- Sin
- Cos
- Log
- Exp

# ¿Qué función nos ayudará a convertir los productos en sumas?

$$0.75 * 0.75 * 0.75 * 0.25 = 0.105$$

$$log(0.75) + log(0.75) + log(0.75) + log(0.75) = -3.245$$

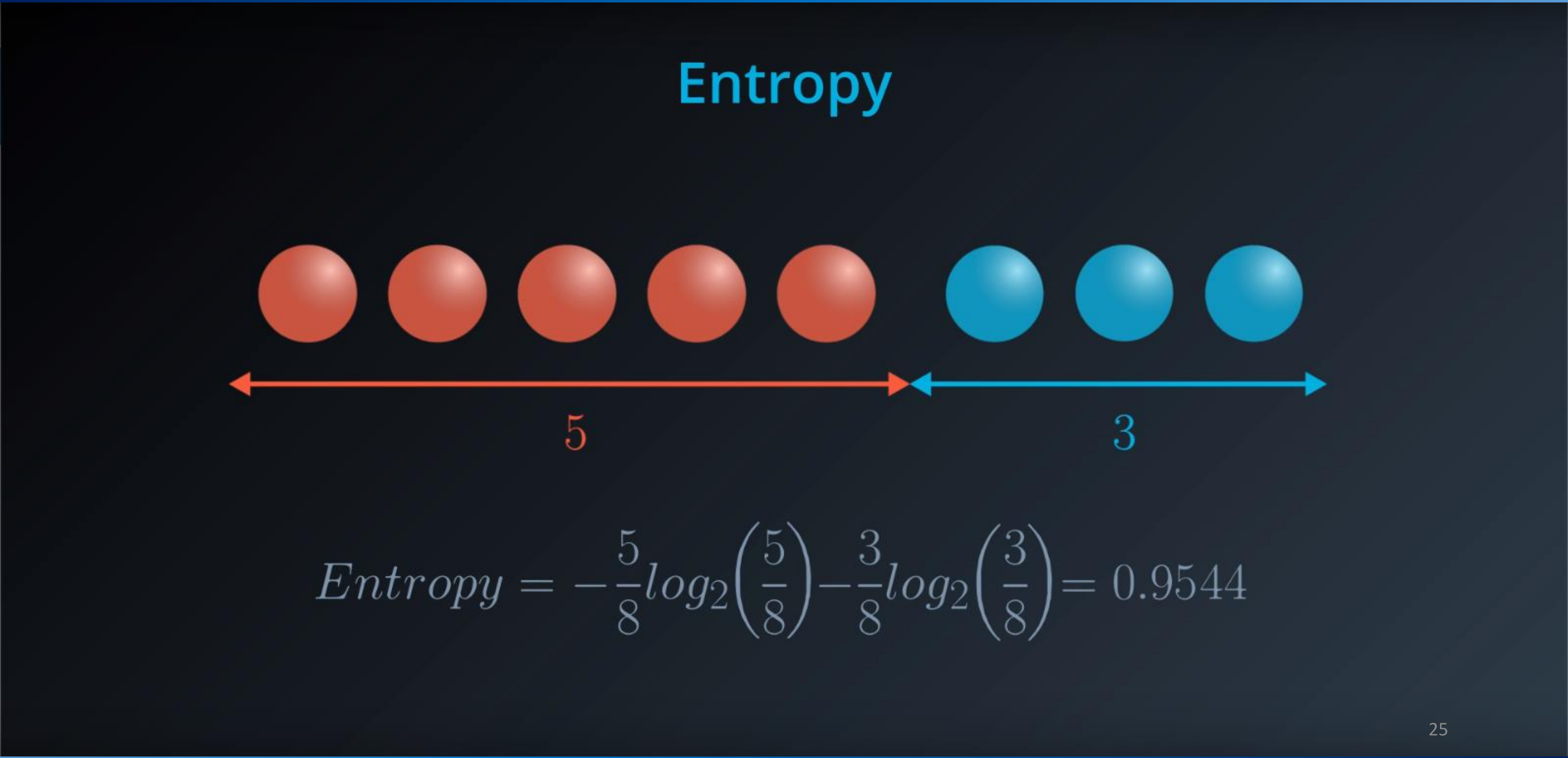

Log(ab) = Log (a) + Log (b)

# -$Log_2$, promedio de entropía

## Entropy

| | $P$ $(red)$ | $P$ $(blue)$ | $P$ $(winning)$ | -$\log_2$ $(P(winning))$ | $Entropy$ $(average)$ |
|---|---|---|---|---|---|
|  | 1 | 1 | $1 \times 1 \times 1 \times 1 = 1$ | $0 + 0 + 0 + 0 = 0$ | 0 |
|  | 0.75 | 0.25 | $0.75 \times 0.75 \times 0.75 \times 0.25 = 0.105$ | $0.415 + 0.415 + 0.415 + 2 = 3.245$ | 0.81 |
|  | 0.5 | 0.5 | $0.5 \times 0.5 \times 0.5 \times 0.5 = 0.0625$ | $1 + 1 + 1 + 1 = 4$ | 1 |

# Resumen


Entropy
5
3
$Entropy = -\frac{5}{8}log_2\left(\frac{5}{8}\right)-\frac{3}{8}log_2\left(\frac{3}{8}\right)= 0.9544$

# Ecuación: Entropía

## Entropy


$$Entropy = -\frac{m}{m-n}log_2\left(\frac{m}{m+n}\right) - \frac{n}{m+n}log_2\left(\frac{n}{m+n}\right)$$

# Entropía multiclase

$$entropy = -\frac{m}{m+n}\log_2(\frac{m}{m+n}) - \frac{n}{m+n}\log_2(\frac{n}{m+n})$$

$$p_1 = \frac{m}{m+n} \qquad p_2 = \frac{n}{m+n}$$

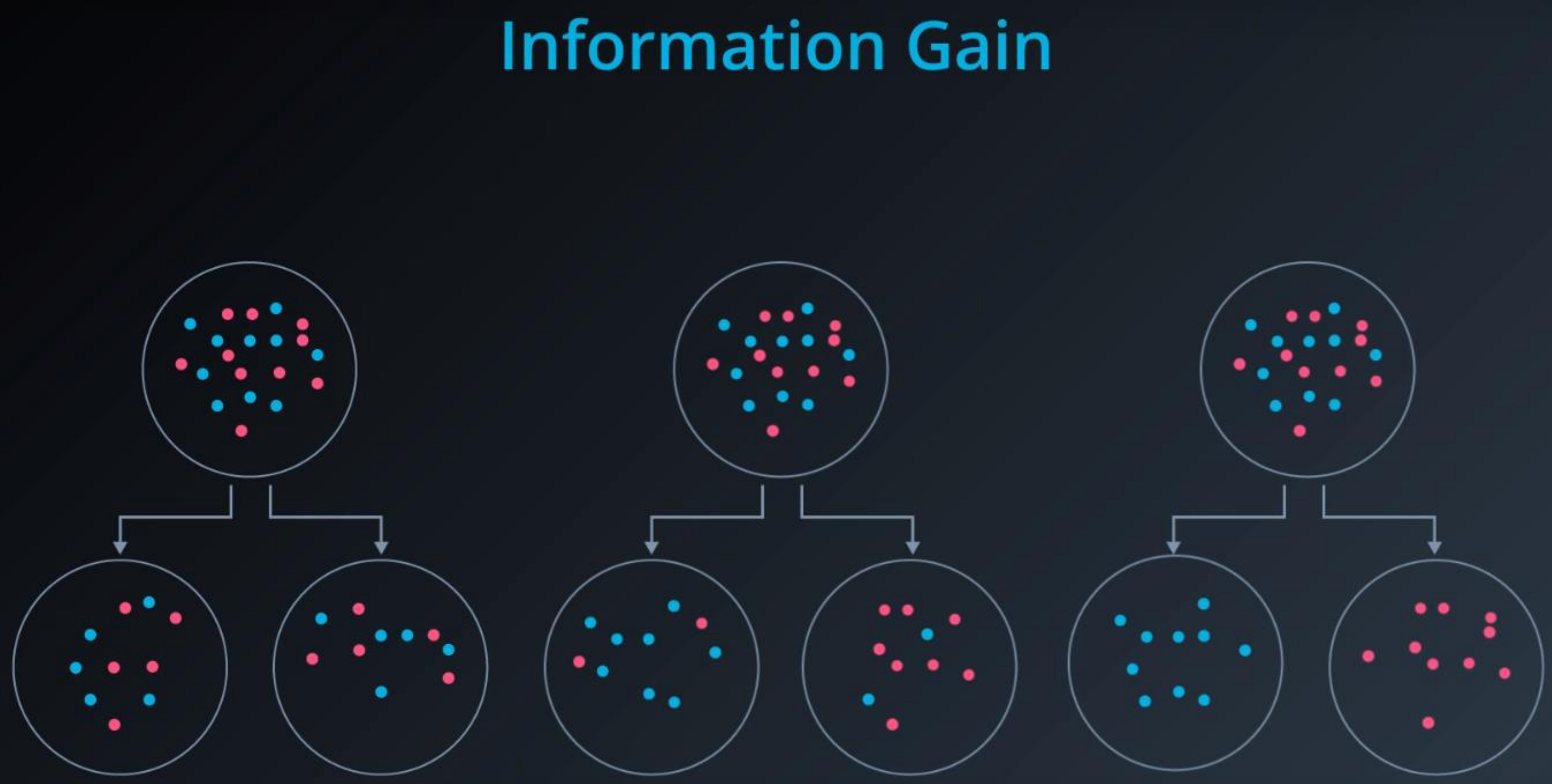
$$entropy = -p1\ log_2(p1) - p2\ log_2(p2)$$

Esta ecuación de entropía se puede extender al caso de varias clases, donde tenemos tres o más valores posibles:

$$entropy = -p1\ log_2(p1) - p2\ log_2(p2) - \ldots - pn\ log_2(pn)$$

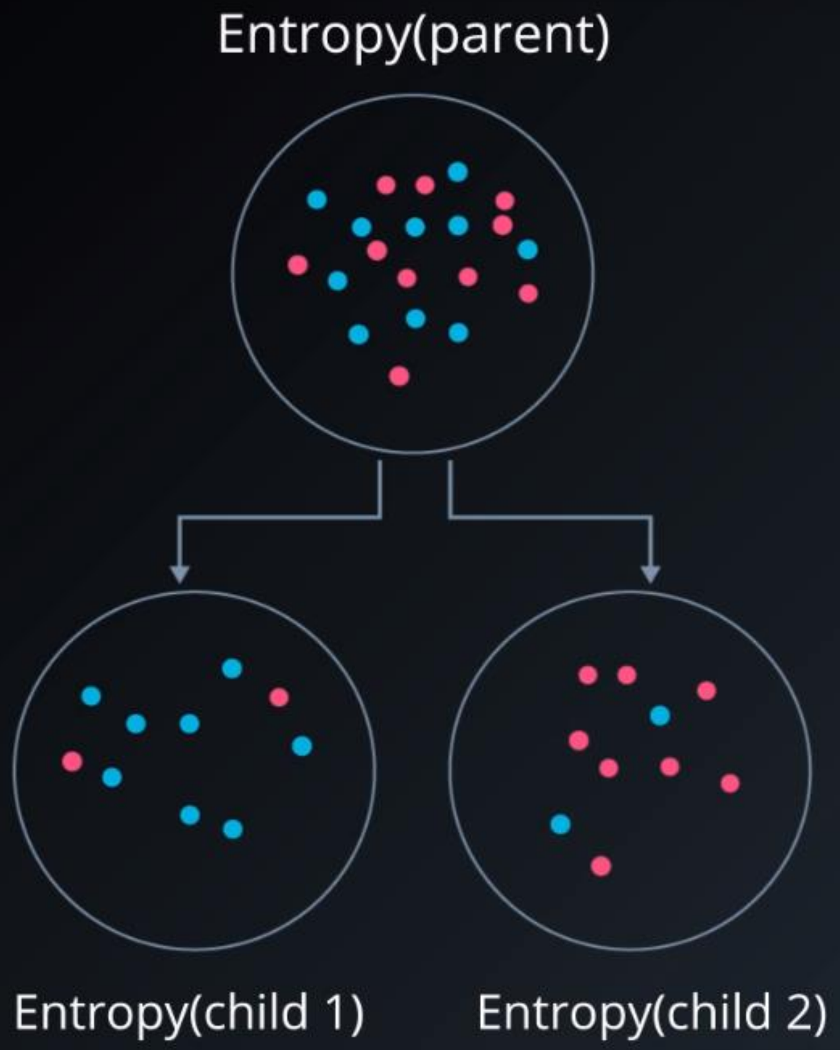
$$entropy = -\sum_{i=1}^{n} pi\ log_2(pi)$$

# Ganancia de información (Information Gain)



# Ganancia de información (Information Gain)

# Ganancia de información (Information Gain)




Information Gain =

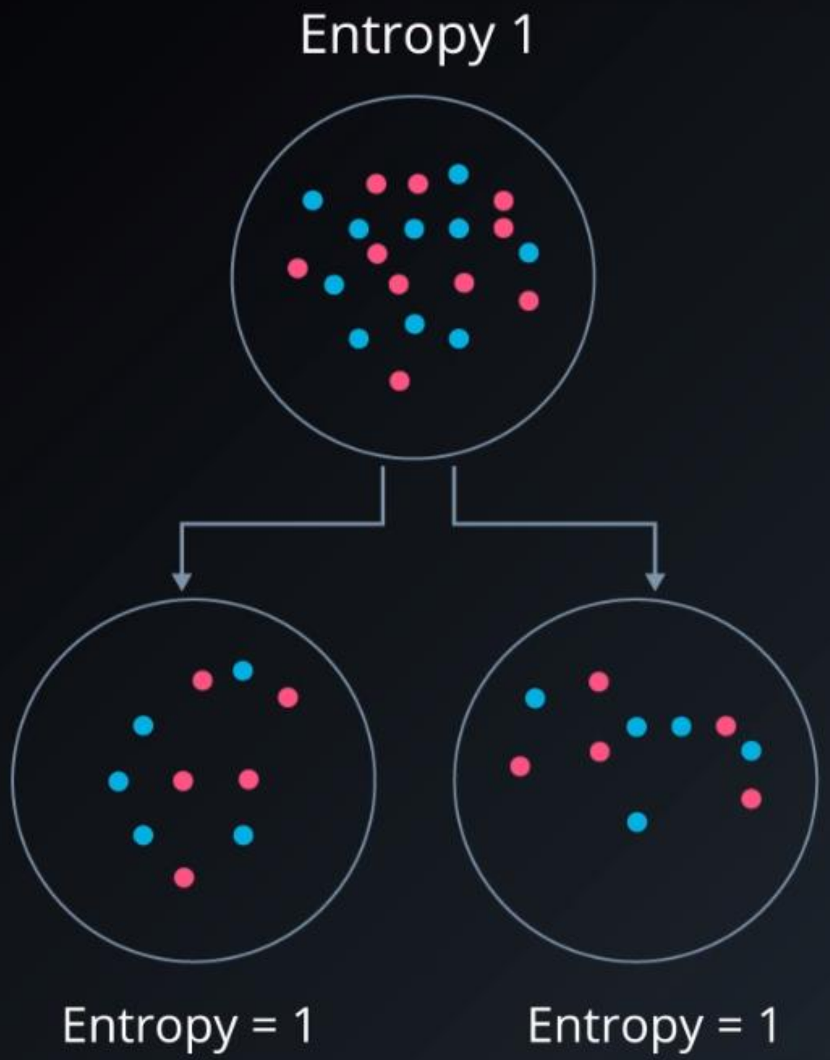
Entropy(parent)

- 0.5 [Entropy(child 1) + Entropy(child 2)]

# Ganancia de información (Information Gain)

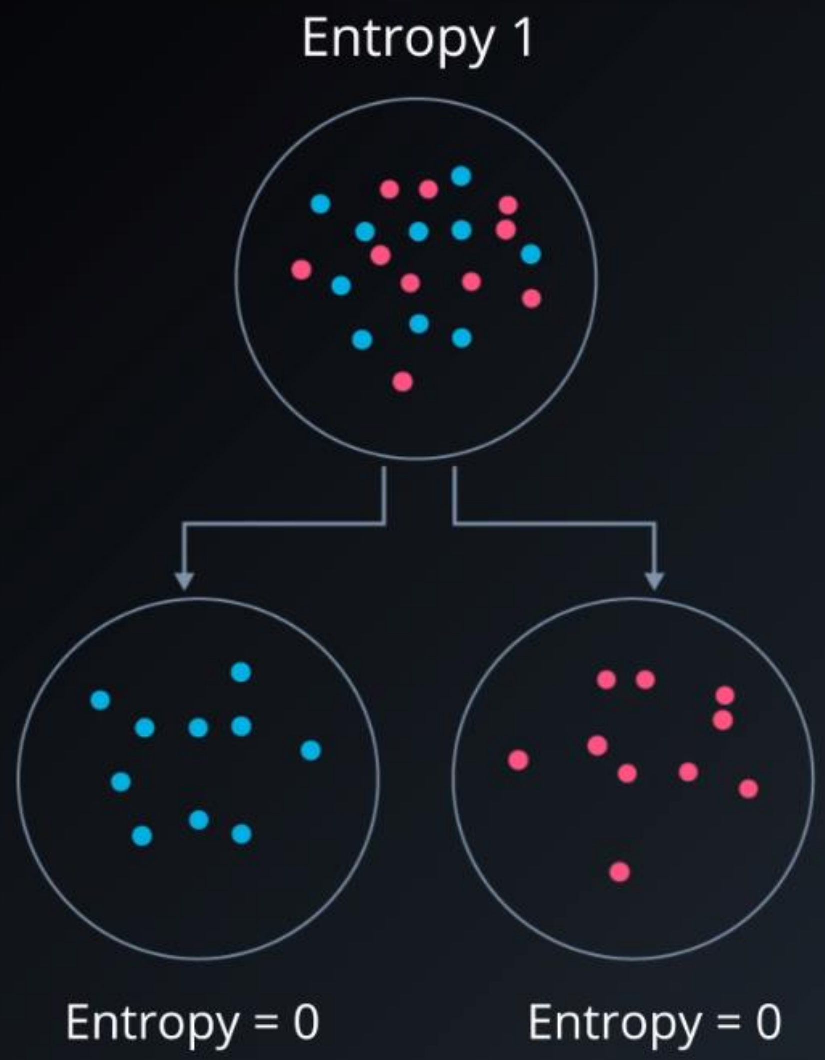

Information gain
1-0.72 = 0.28

# Ganancia de información (Information Gain)




Information gain
1 - 1 = 0

# Ganancia de información (Information Gain)




Information gain
1 - 0 = 1

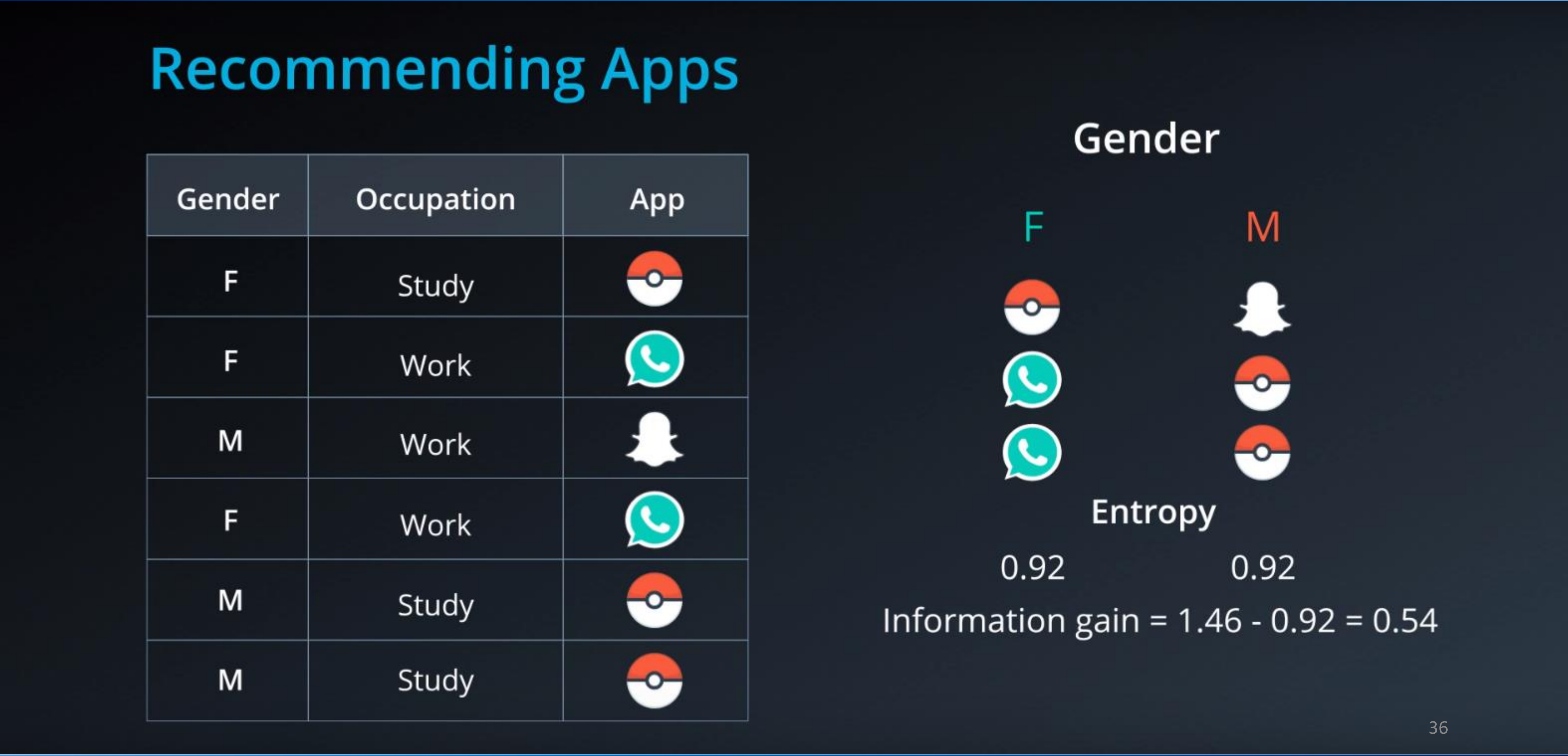
# Ganancia de información (Information Gain)

# Maximizando la Ganancia de información

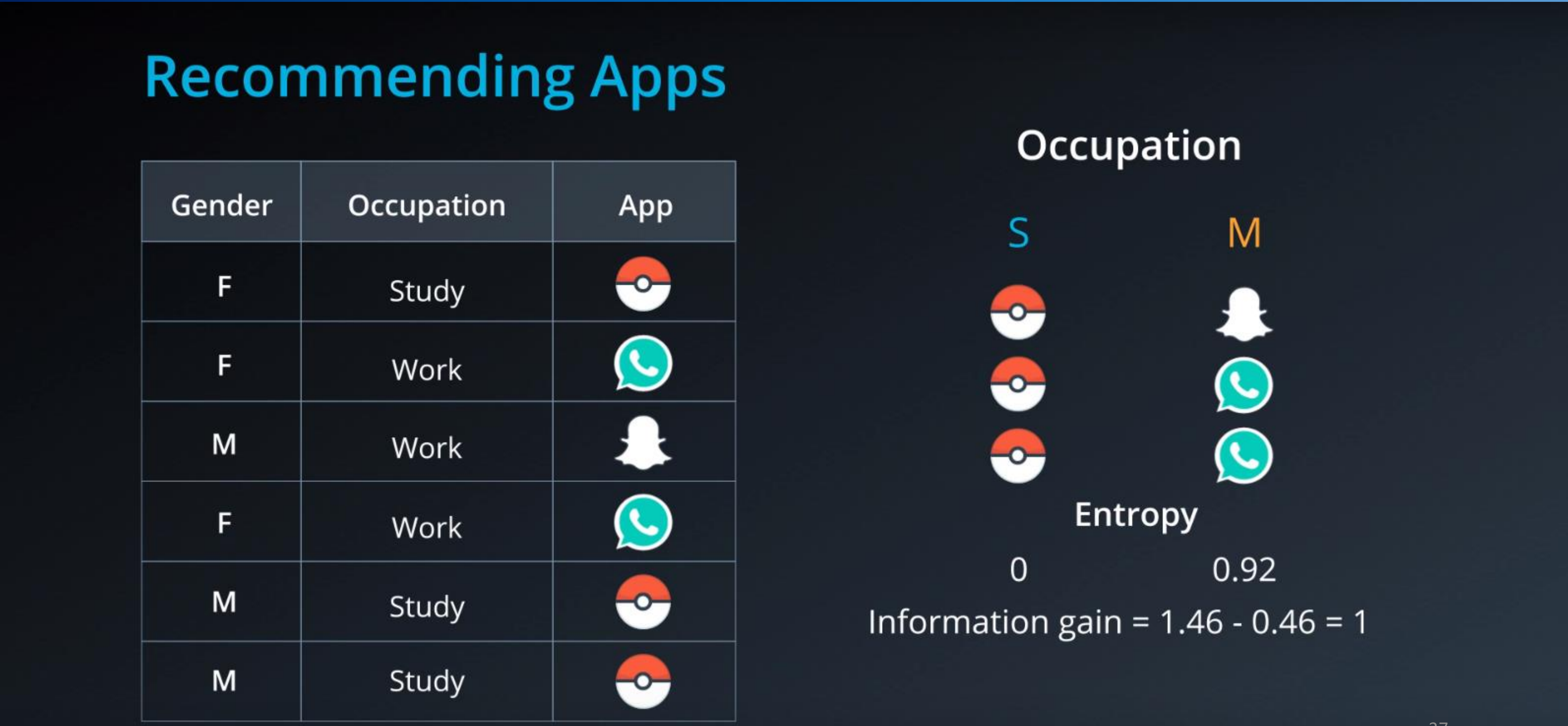
| Gender | Occupation | App |
|---|---|---|
| F | Study | |
| F | Work | |
| M | Work | |
| F | Work | |
| M | Study | |
| M | Study | |

$$\text{Entropy} = -\frac{3}{6}\log_2\left(\frac{3}{6}\right) - \frac{2}{6}\log_2\left(\frac{2}{6}\right) - \frac{1}{6}\log_2\left(\frac{1}{6}\right)$$
$$= 1.46$$

# Maximizando la Ganancia de información



# Maximizando la Ganancia de información



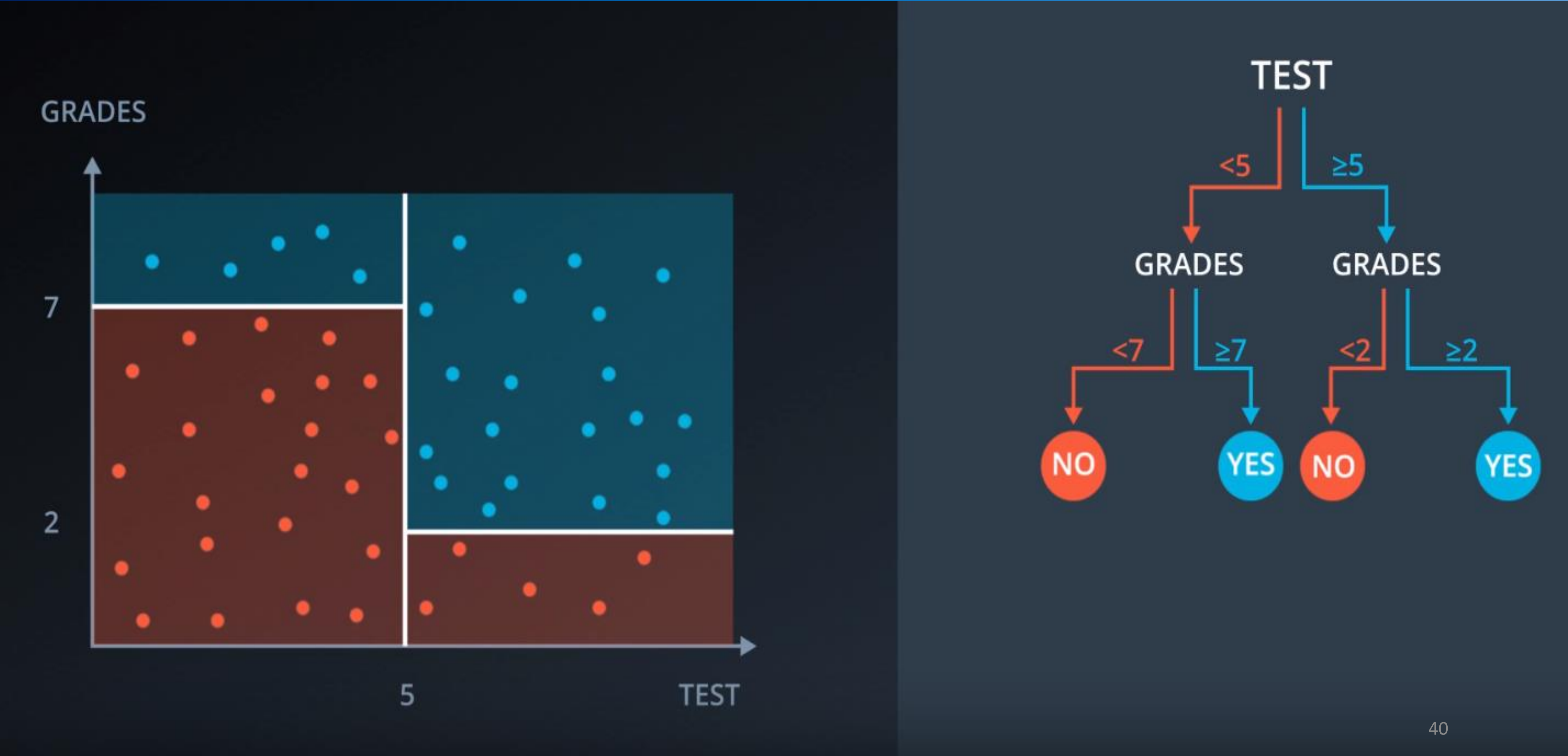
# Maximizando la Ganancia de información

## Recommending Apps

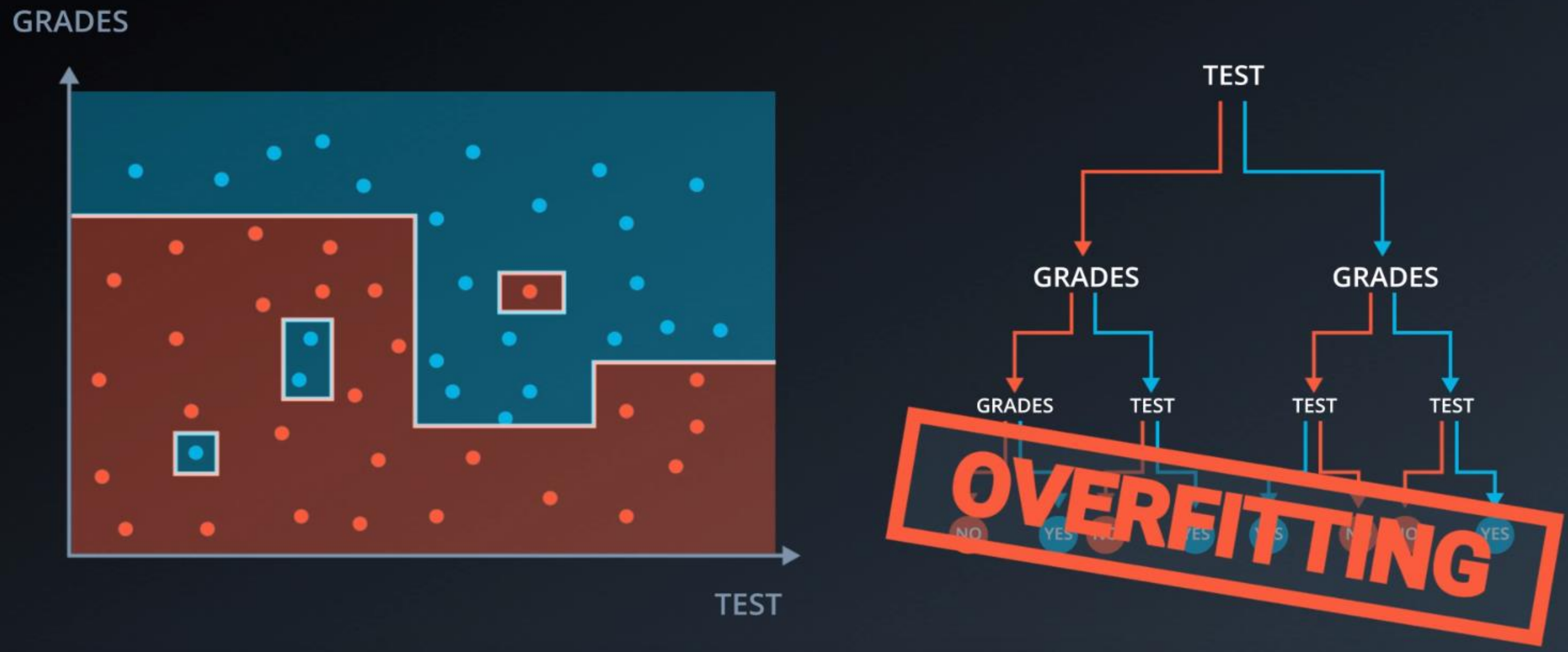
| 0.54 | 1 | |
|---|---|---|
| **Gender** | **Occupation** | **App** |
| F | Study | |
| F | Work | |
| M | Work | |
| F | Work | |
| M | Study | |
| M | Study | |

# Maximizando la Ganancia de información

# Maximizando la Ganancia de información



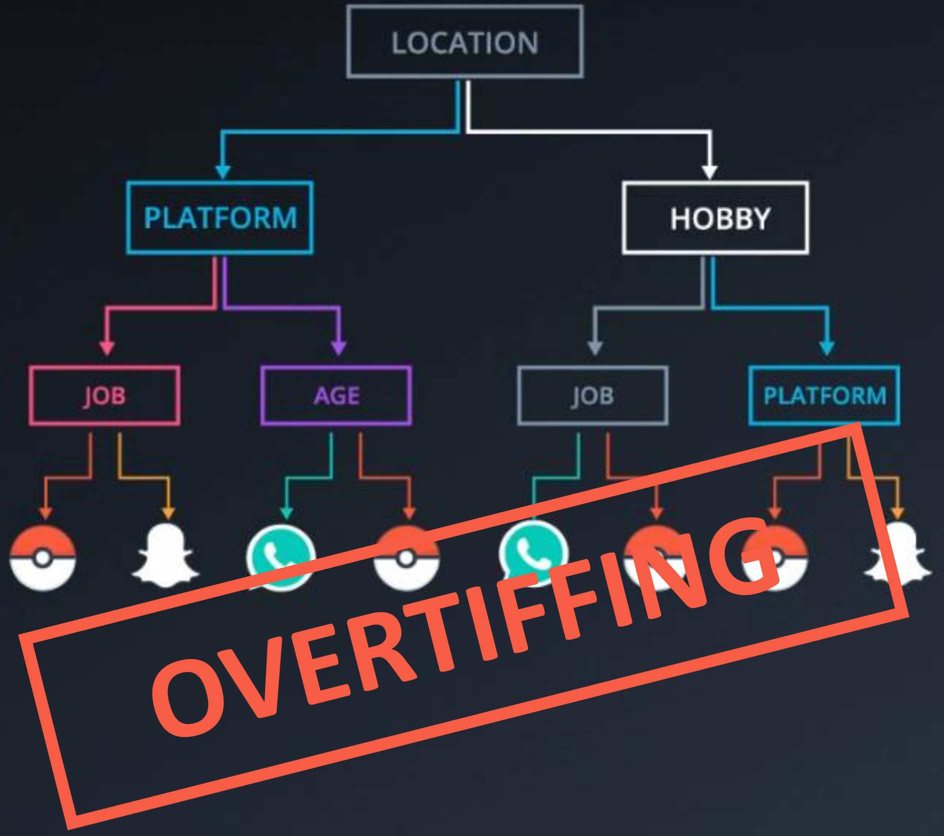
# Overfitting



# Overfitting

## Large Tables

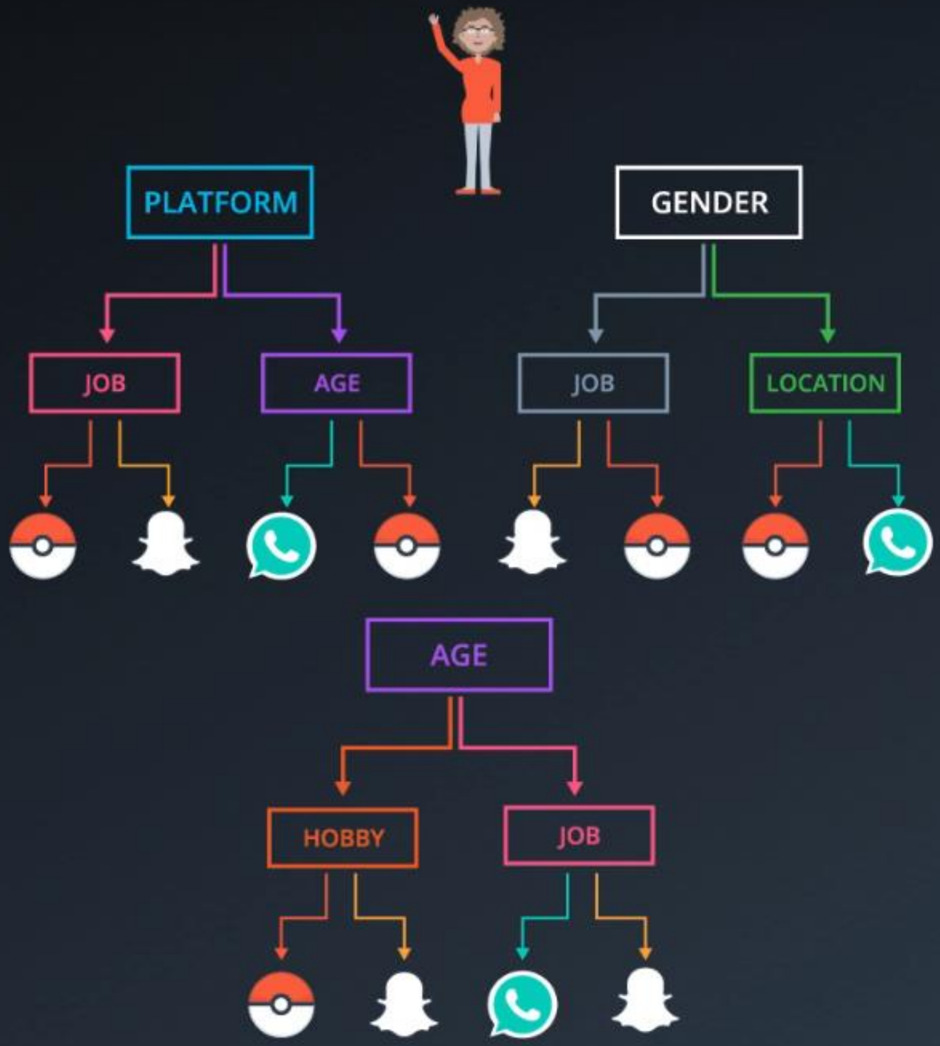
| Gender | Age | Location | Platform | Job | Hobby | App |
|---|---|---|---|---|---|---|
| F | 15 | US | iOS | School | Videogames | |
| F | 25 | France | Android | Work | Tennis | |
| M | 32 | Chile | iOS | Temp | Tennis | |
| F | 40 | China | iOS | Retired | Chess | |
| M | 12 | US | Android | School | Tennis | |
| M | 14 | Australia | Android | School | Videogames | |



# Random Forests

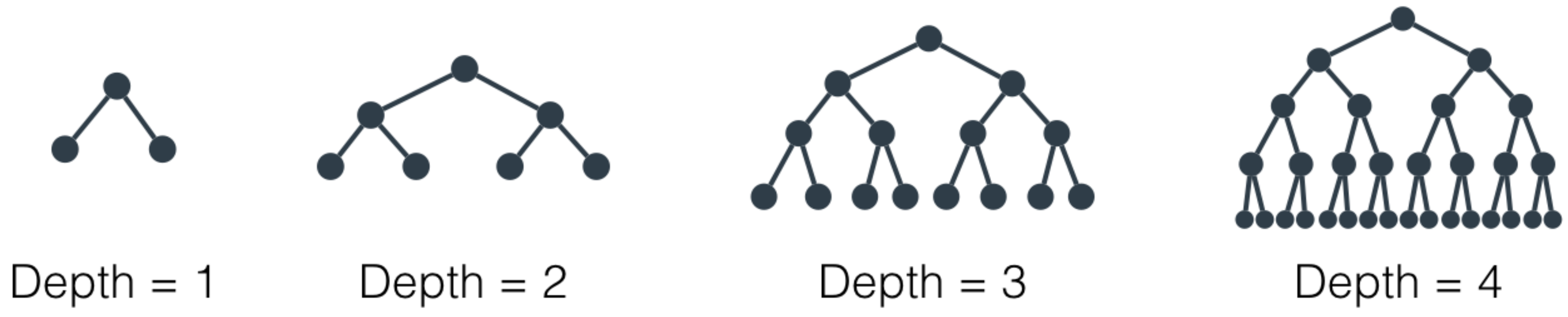
## Random Forests

| Gender | Age | Location | Platform | Job | Hobby | App |
|---|---|---|---|---|---|---|
| F | 15 | US | iOS | School | Videogames | |
| F | 25 | France | Android | Work | Tennis | |
| M | 32 | Chile | iOS | Temp | Tennis | |
| F | 40 | China | iOS | Retired | Chess | |
| M | 12 | US | Android | School | Tennis | |
| M | 14 | Australia | Android | School | Videogames | |



# Hiperparametros para Árboles de Decisión

**Maximum Depth (**Profundidad máxima)
La profundidad máxima de un árbol de decisión es simplemente la longitud más grande entre la raíz y una hoja. Un árbol de longitud máxima k puede tener como máximo 2^k hojas.

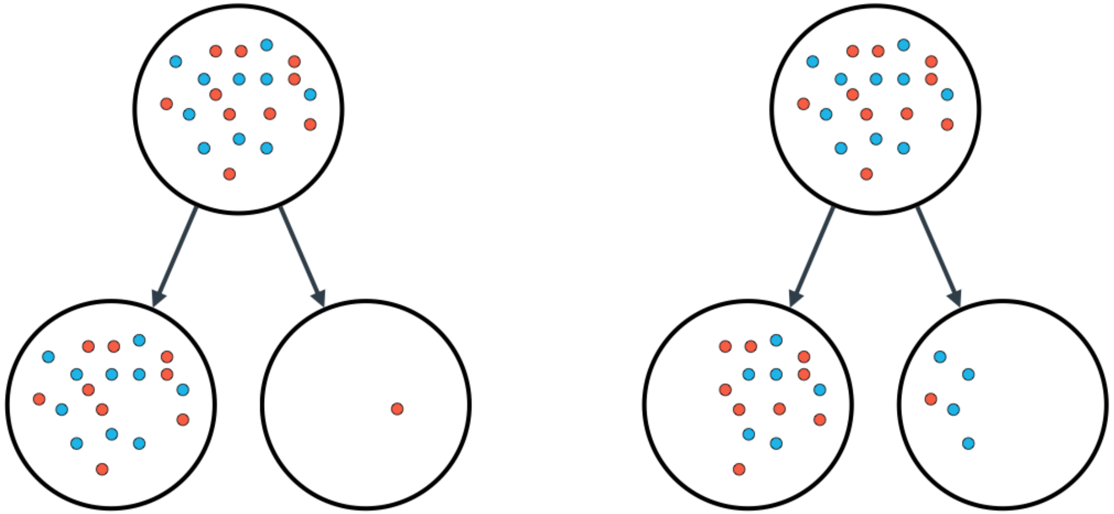


# Hiperparametros para Árboles de Decisión

**Minimum number of samples per leaf (**Número mínimo de muestras por hoja)

Cuando se divide un nodo, uno podría tener el problema de tener 99 muestras en una de ellas y 1 en la otra. Esto no nos llevará demasiado lejos en nuestro proceso, y sería una pérdida de tiempo y recursos.

Si queremos evitar esto, podemos establecer un mínimo para el número de muestras que permitimos en cada hoja.



Minimum samples per leaf = 1    Minimum samples per leaf = 5

# Hiperparametros para Árboles de Decisión

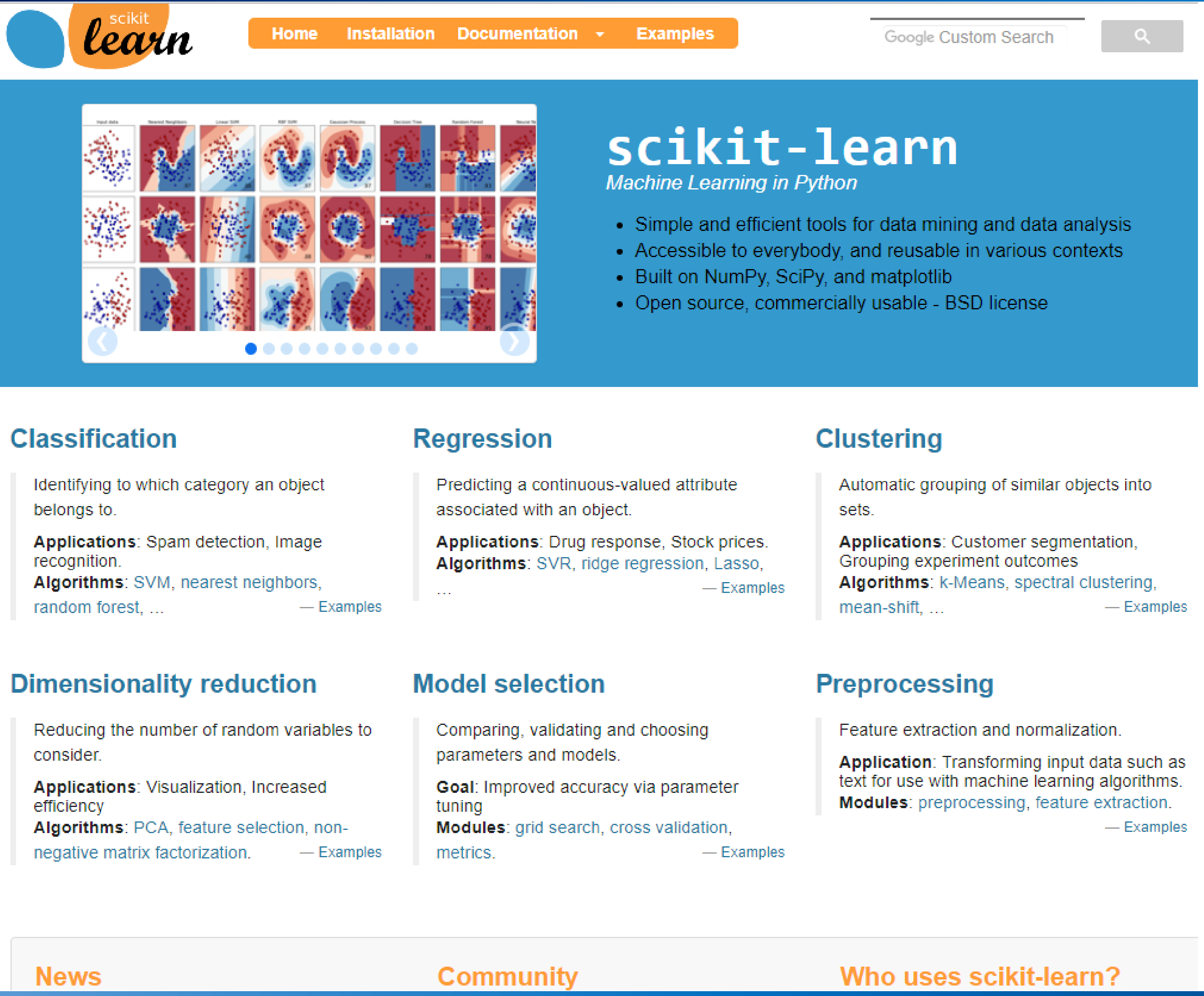
**Minimum number of samples per split (Número mínimo de muestras por división)**

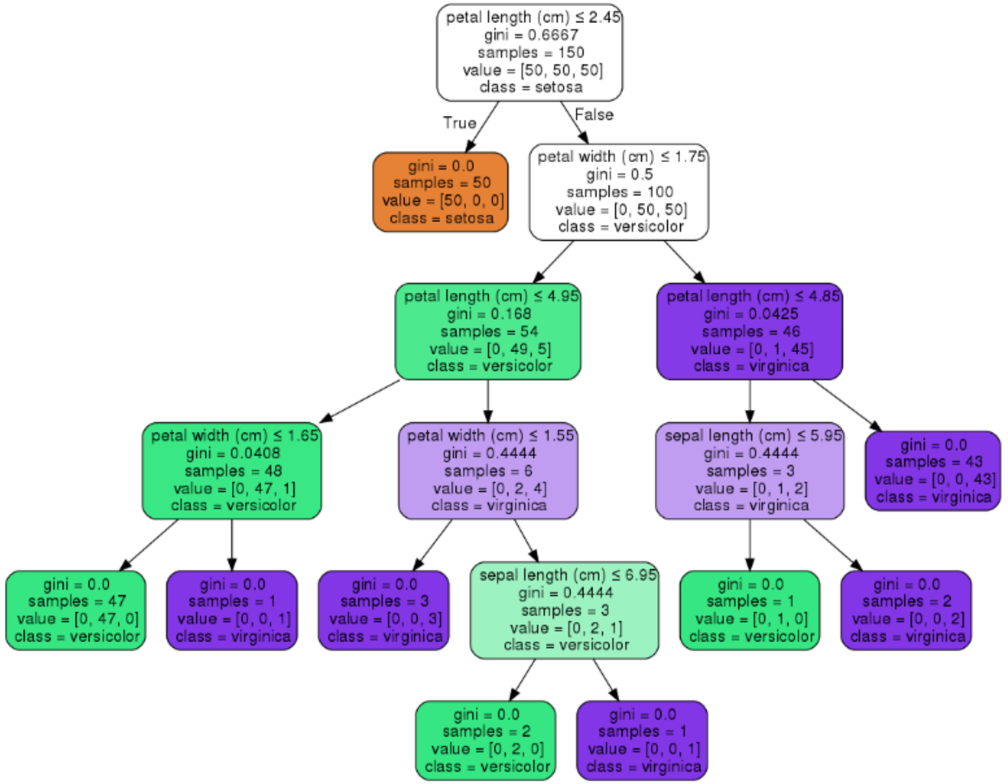
Este es el mismo que el número mínimo de muestras por hoja, pero se aplica en cualquier división de un nodo.

**Maximum number of features (Número máximo de características)**

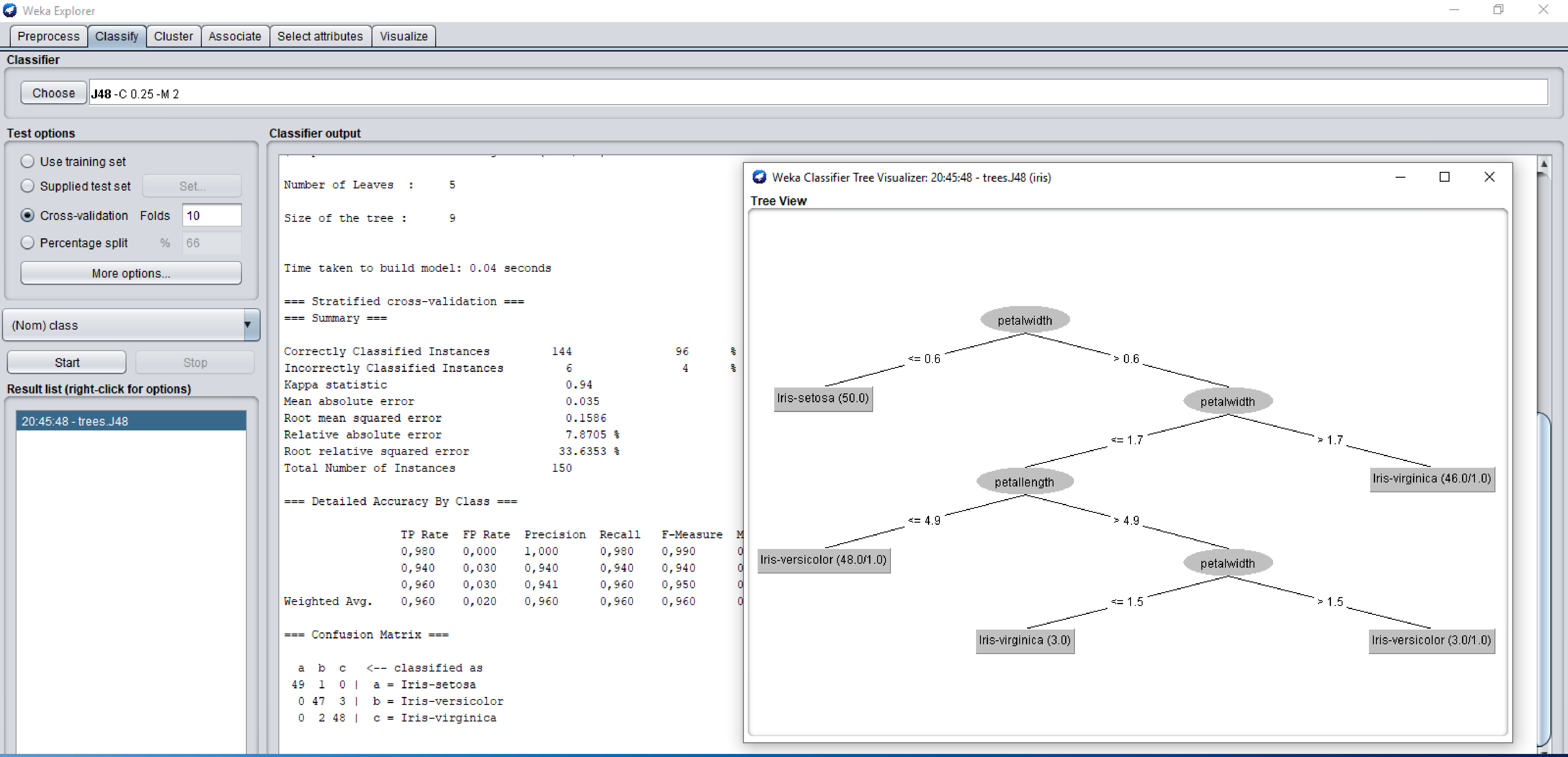
A menudo, tendremos demasiadas características para construir un árbol. Si este es el caso, en cada división, tenemos que verificar el conjunto de datos completo en cada una de las funciones. Esto puede ser muy costoso. Una solución para esto es limitar la cantidad de funciones que uno busca en cada división. Si este número es lo suficientemente grande, es muy probable que encontremos una buena característica entre las que buscamos

# Árboles de decisión en Sklearn

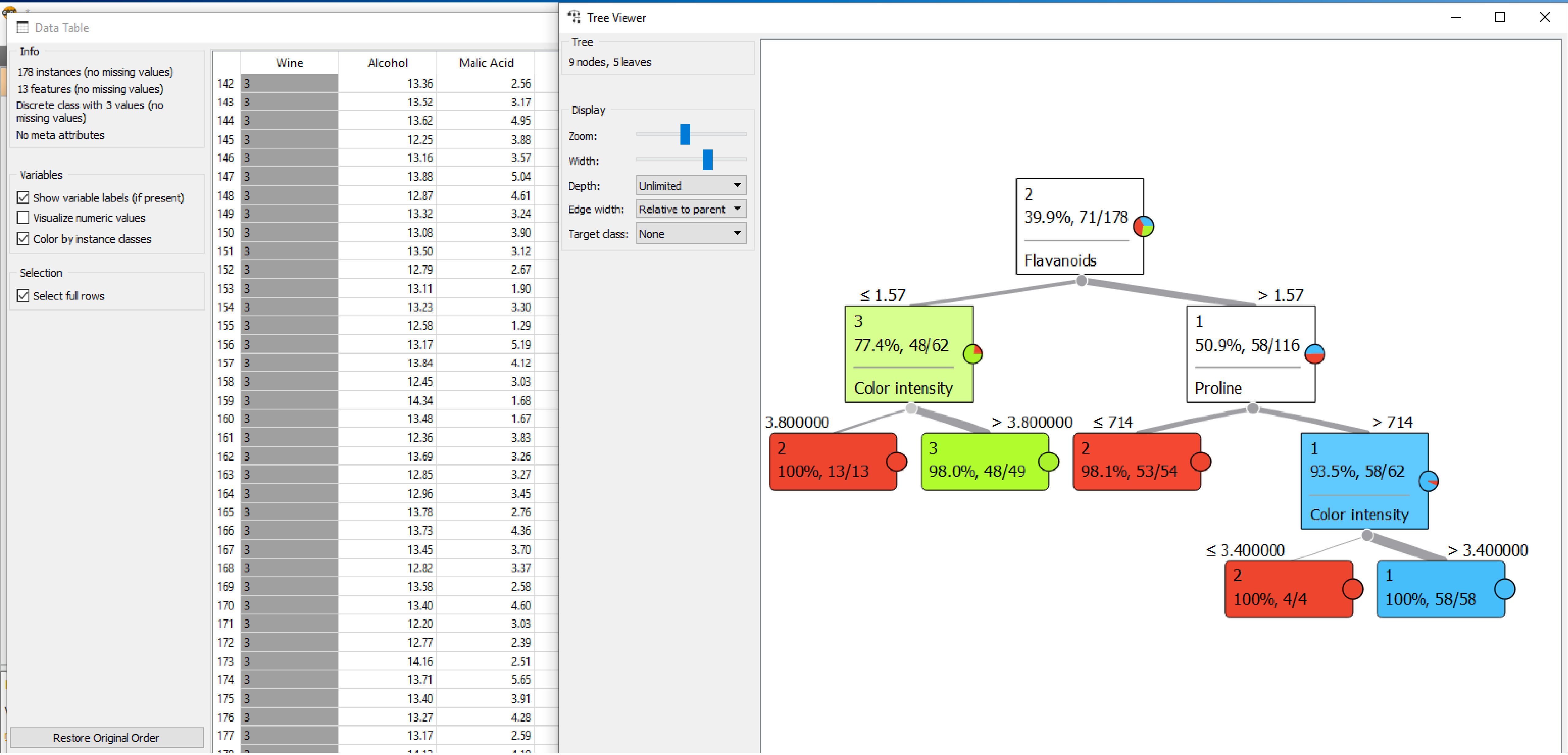


# Árboles de decisión en Weka



# Árboles de decisión en Orange




| | Wine | Alcohol | Malic Acid |
|---|---|---|---|
| 142 | 3 | 13.36 | 2.56 |
| 143 | 3 | 13.52 | 3.17 |
| 144 | 3 | 13.62 | 4.95 |
| 145 | 3 | 12.25 | 3.88 |
| 146 | 3 | 13.16 | 3.57 |
| 147 | 3 | 13.88 | 5.04 |
| 148 | 3 | 12.87 | 4.61 |
| 149 | 3 | 13.32 | 3.24 |
| 150 | 3 | 13.08 | 3.90 |
| 151 | 3 | 13.50 | 3.12 |
| 152 | 3 | 12.79 | 2.67 |
| 153 | 3 | 13.11 | 1.90 |
| 154 | 3 | 13.23 | 3.30 |
| 155 | 3 | 12.58 | 1.29 |
| 156 | 3 | 13.17 | 5.19 |
| 157 | 3 | 13.84 | 4.12 |
| 158 | 3 | 12.45 | 3.03 |
| 159 | 3 | 14.34 | 1.68 |
| 160 | 3 | 13.48 | 1.67 |
| 161 | 3 | 12.36 | 3.83 |
| 162 | 3 | 13.69 | 3.26 |
| 163 | 3 | 12.85 | 3.27 |
| 164 | 3 | 12.96 | 3.45 |
| 165 | 3 | 13.78 | 2.76 |
| 166 | 3 | 13.73 | 4.36 |
| 167 | 3 | 13.45 | 3.70 |
| 168 | 3 | 12.82 | 3.37 |
| 169 | 3 | 13.58 | 2.58 |
| 170 | 3 | 13.40 | 4.60 |
| 171 | 3 | 12.20 | 3.03 |
| 172 | 3 | 12.77 | 2.39 |
| 173 | 3 | 14.16 | 2.51 |
| 174 | 3 | 13.71 | 5.65 |
| 175 | 3 | 13.40 | 3.91 |
| 176 | 3 | 13.27 | 4.28 |
| 177 | 3 | 13.17 | 2.59 |