

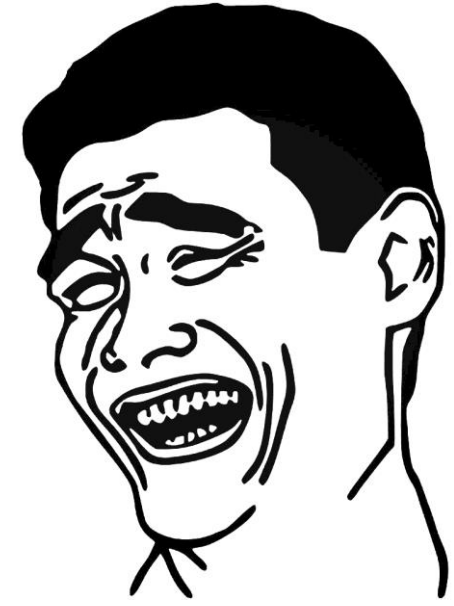
THE BRIDGE

Isolation Forest

Introducción

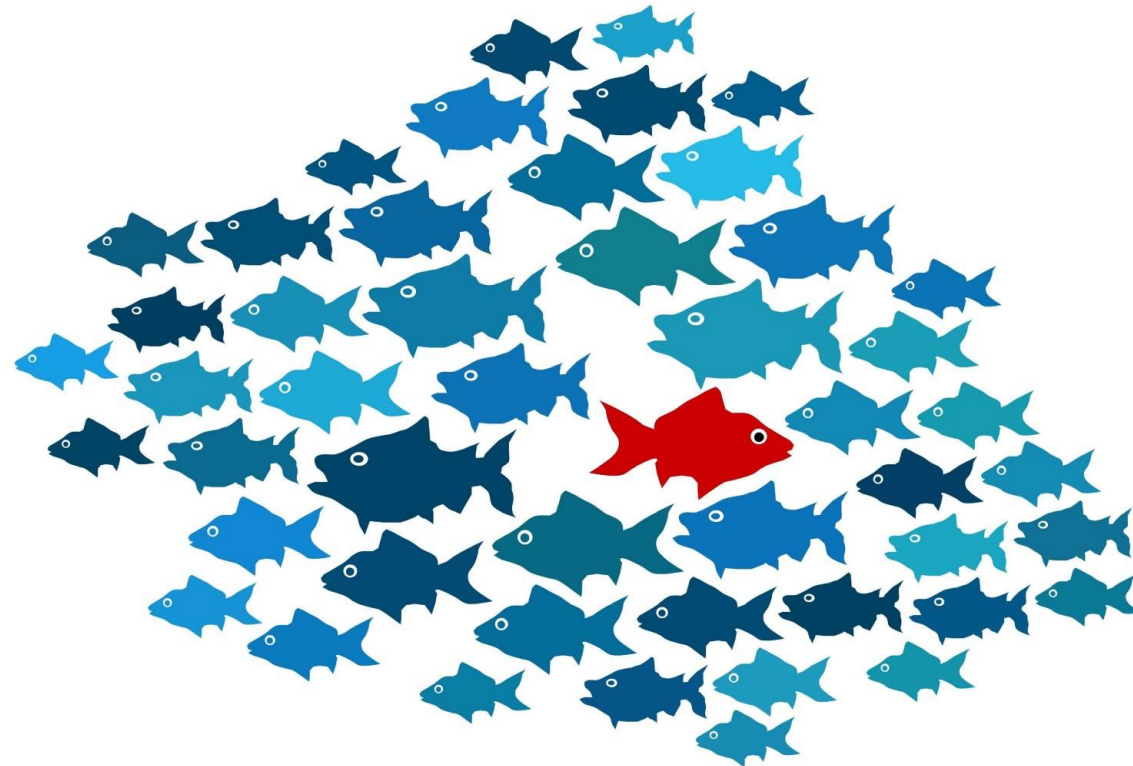
$$\begin{aligned}
 iG^{\mu\nu}(p) = & \frac{1}{1-J_2} \left\{ \frac{\bar{m} + \not{p}}{\bar{m}^2 - p^2} (\mathcal{P}^{3/2})^{\mu\nu} \right. \\
 & + \frac{1}{2} \left[\frac{2\bar{m} - 2\sqrt{p^2} + A_+}{-\bar{m}^2 + X_+} + \frac{2\bar{m} + 2\sqrt{p^2} + A_-}{-\bar{m}^2 + X_-} \right] (\mathcal{P}_{11}^{1/2})^{\mu\nu} \\
 & + \frac{1}{2\sqrt{p^2}} \left[-\frac{2\bar{m} - 2\sqrt{p^2} + A_+}{-\bar{m}^2 + X_+} + \frac{2\bar{m} + 2\sqrt{p^2} + A_-}{-\bar{m}^2 + X_-} \right] \not{p} (\mathcal{P}_{11}^{1/2})^{\mu\nu} \\
 & + \frac{1}{2} \left[\frac{3 \frac{J_3 - \sqrt{p^2} J_4}{1-J_2}}{-\bar{m}^2 + X_+} + \frac{3 \frac{J_3 + \sqrt{p^2} J_4}{1-J_2}}{-\bar{m}^2 + X_-} \right] (\mathcal{P}_{22}^{1/2})^{\mu\nu} \\
 & + \frac{1}{2\sqrt{p^2}} \left[\frac{3 \frac{J_3 - \sqrt{p^2} J_4}{1-J_2}}{-\bar{m}^2 + X_+} - \frac{3 \frac{J_3 + \sqrt{p^2} J_4}{1-J_2}}{-\bar{m}^2 + X_-} \right] \not{p} (\mathcal{P}_{22}^{1/2})^{\mu\nu} \\
 & + \frac{\sqrt{3}}{2} \left[\frac{\bar{m} - \left(\frac{J_1 + \sqrt{3} J_7}{1-J_2} \right)}{-\bar{m}^2 + X_+} - \frac{\bar{m} - \left(\frac{J_1 - \sqrt{3} J_7}{1-J_2} \right)}{-\bar{m}^2 + X_-} \right] [(\mathcal{P}_{21}^{1/2})^{\mu\nu} + (\mathcal{P}_{12}^{1/2})^{\mu\nu}] \\
 & - \frac{\sqrt{3}}{2\sqrt{p^2}} \left[\frac{\bar{m} - \left(\frac{J_1 + \sqrt{3} J_7}{1-J_2} \right)}{-\bar{m}^2 + X_+} + \frac{\bar{m} - \left(\frac{J_1 - \sqrt{3} J_7}{1-J_2} \right)}{-\bar{m}^2 + X_-} \right] \not{p} [(\mathcal{P}_{21}^{1/2})^{\mu\nu} - (\mathcal{P}_{12}^{1/2})^{\mu\nu}] \left. \right\}, \quad (11)
 \end{aligned}$$

$$\begin{aligned}
 \hat{H} &= \sum_{n=1}^N \frac{\hat{p}_n^2}{2m_n} + V(x_1, x_2, \dots, x_N) \\
 &= -\frac{\hbar^2}{2} \sum_{n=1}^N \frac{1}{m_n} \frac{\partial^2}{\partial x_n^2} + V(x_1, x_2, \dots, x_N)
 \end{aligned}$$



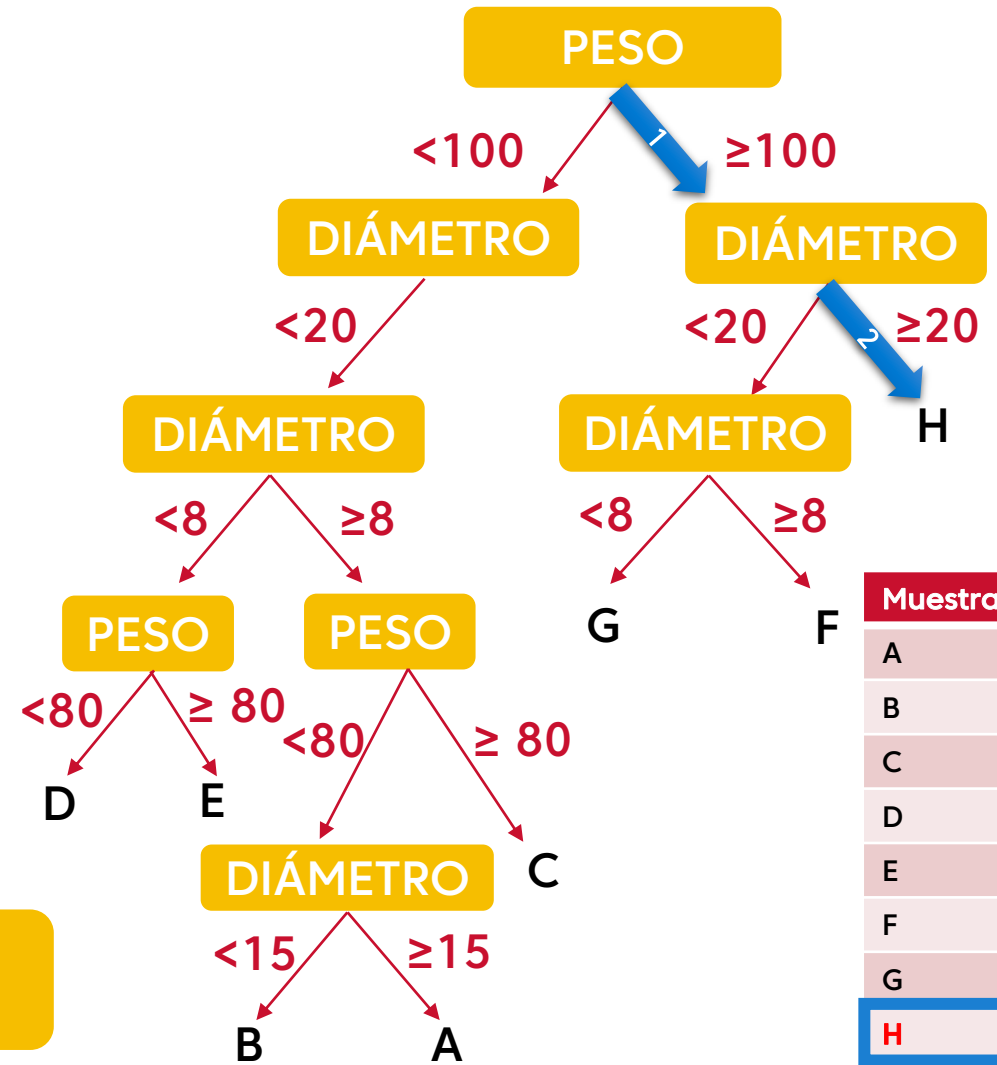
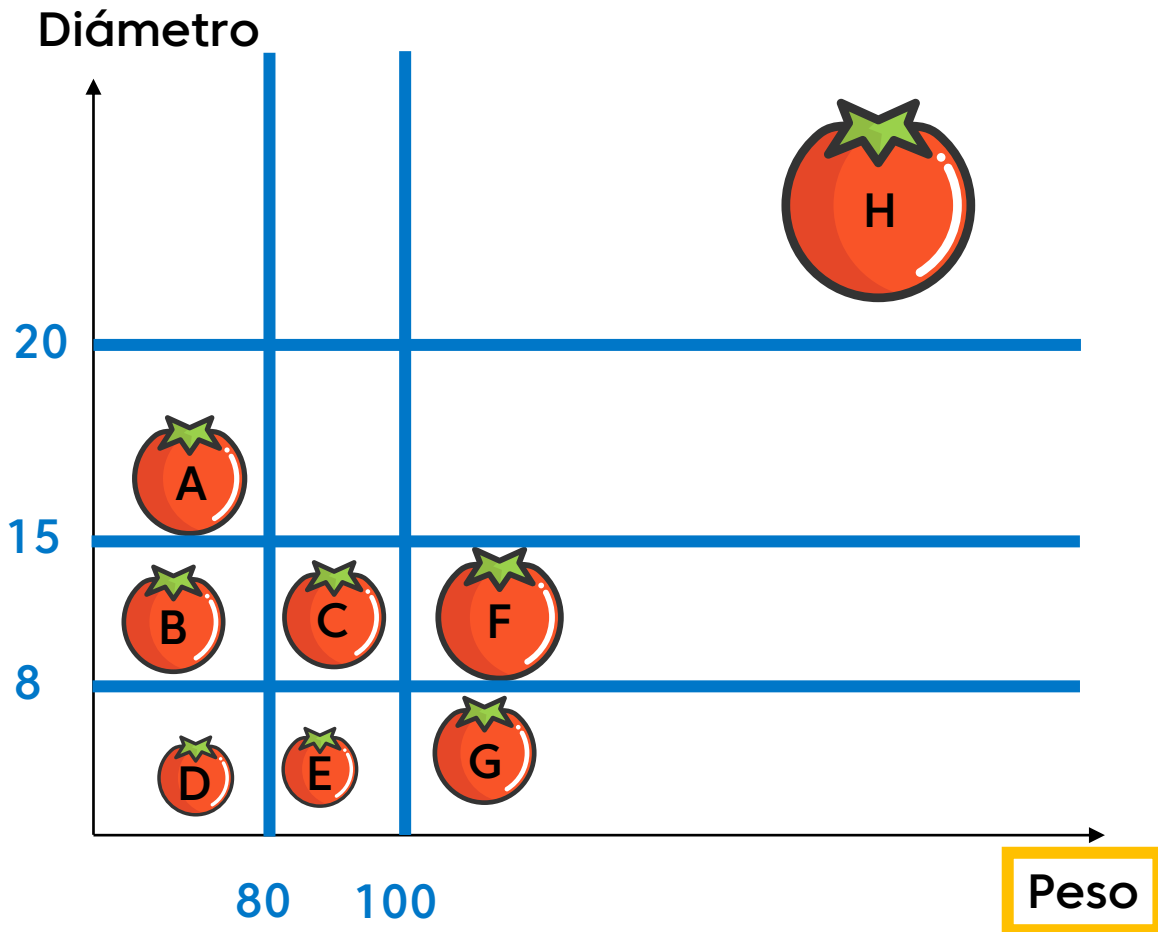
Introducción

- Algoritmo de detección de anomalías
- Basado en árboles de decisión binarios
- Eficiente y rápido



Funcionamiento

1. Selecciona un atributo aleatoriamente
2. Elige un valor aleatorio entre el máximo y el mínimo
3. Repetimos hasta aislar todas las muestras

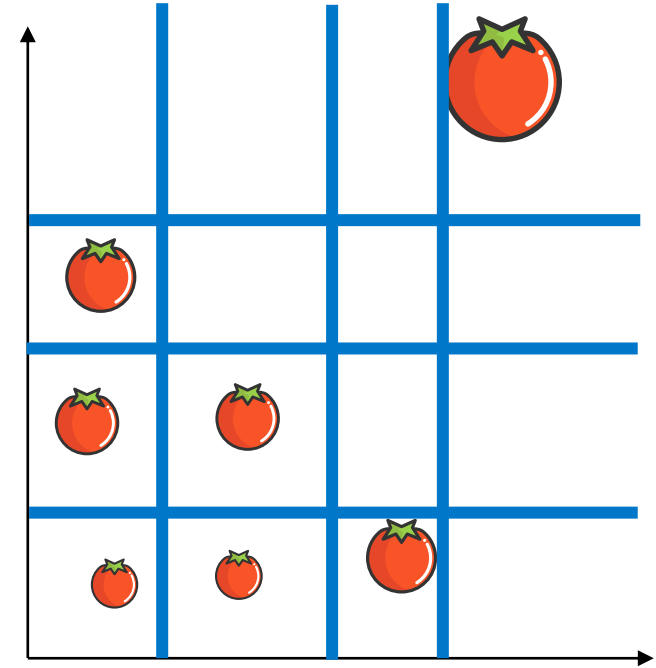
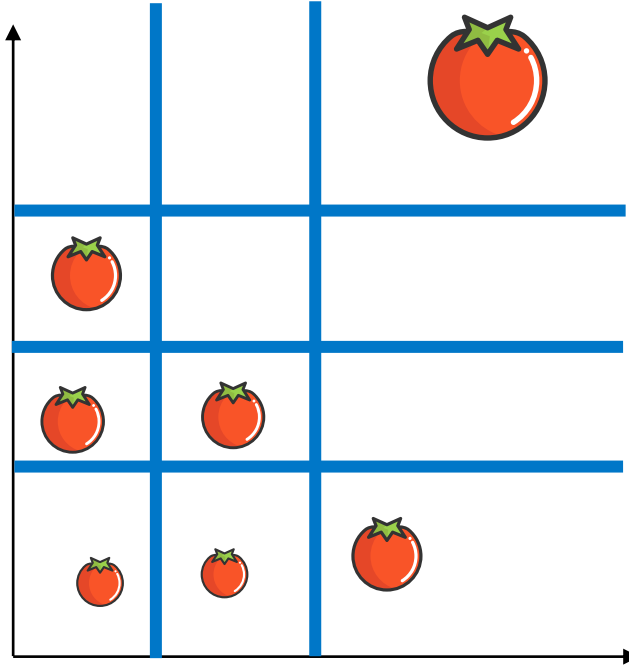
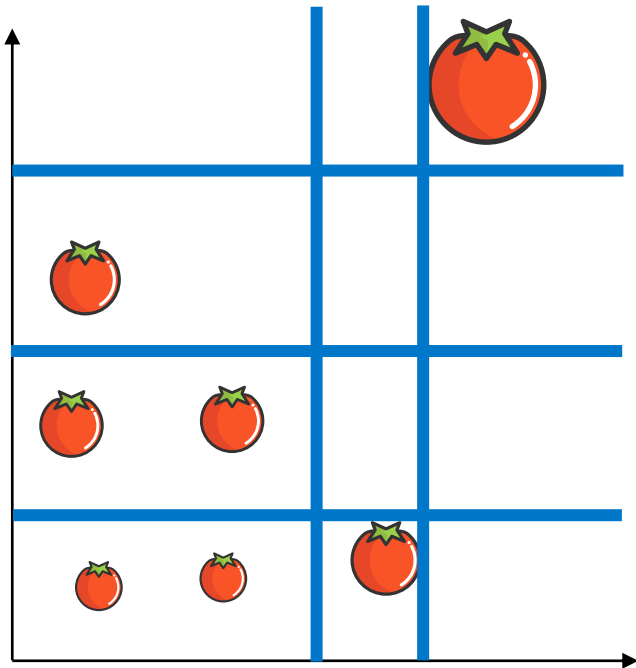


Muestra	Long.
A	5
B	5
C	4
D	4
E	4
F	3
G	3
H	2

⇒ Las anomalías se aíslan más cerca de la raíz del árbol

Funcionamiento

- El algoritmo crea múltiples árboles y calcula la media de las longitudes a cada punto



Funcionamiento

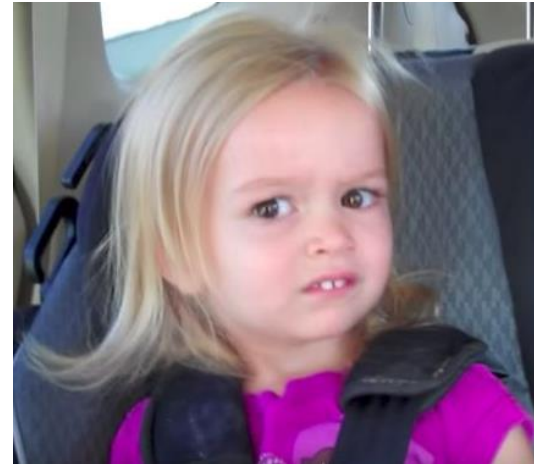
- ⦿ Por último, determina un score para cada punto x :

$$s(x, n) = 2^{-\frac{E[h(x)]}{c(n)}}$$

donde:

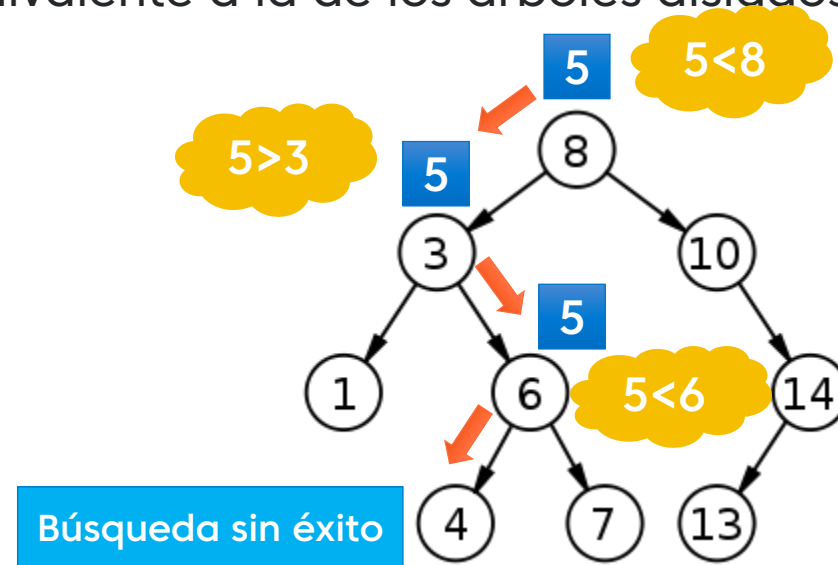
- $E[h(x)]$ es la media de las longitudes de los caminos a cada nodo en todos los árboles
- $c(n)$ es la longitud media de las búsquedas sin éxito en los árboles binarios de búsqueda:

$$c(n) = 2(\ln(n - 1) + 0.577) - \left(\frac{2(n-1)}{n}\right)$$



Inciso: Árboles binarios de búsqueda (BST)

- Tipo particular de árbol binario usado en informática que permite búsquedas eficientes
- Su estructura es equivalente a la de los árboles aislados



- Búsqueda sin éxito: por ejemplo, el elemento 5
- Matemáticamente, es posible determinar la longitud media de las búsquedas sin éxito en función del número de elementos

Funcionamiento

- Por último, determina un score para cada punto x :

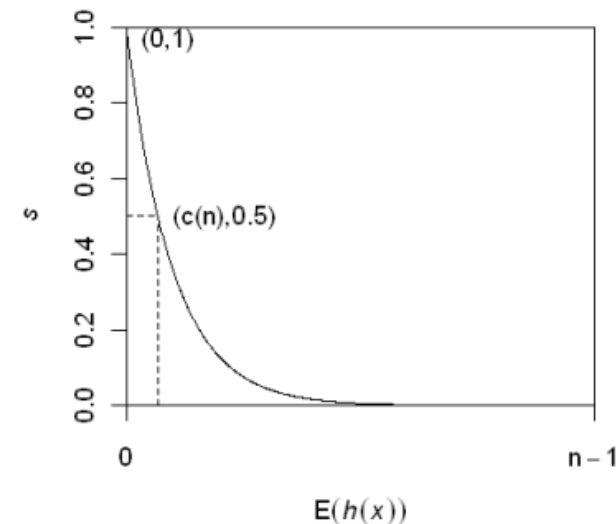
$$s(x, n) = 2^{-\frac{E[h(x)]}{c(n)}}$$

donde:

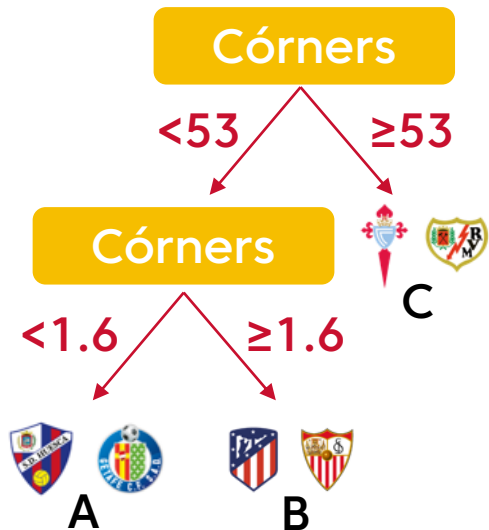
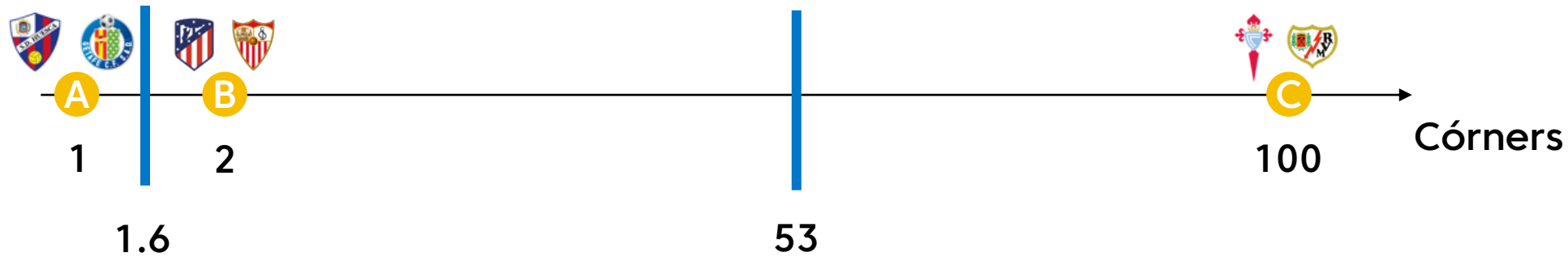
- $E[h(x)]$ es la media de las longitudes de los caminos a cada nodo
- $c(n)$ es la longitud media de las búsquedas sin éxito en los árboles binarios de búsqueda:

$$c(n) = 2(\ln(n-1) + 0.577) - \left(\frac{2(n-1)}{n}\right)$$

- $s \rightarrow 1 \Rightarrow$ Anomalías
- $s \leq 0.5 \Rightarrow$ Normal



Ejemplo



$$\text{Score: } s(x, 3) = 2^{-\frac{E[h(x)]}{c(3)}}$$

$$\text{Longitudes de los paths: } \begin{cases} h(A) = 2 \\ h(B) = 2 \\ h(C) = 1 \end{cases}$$

$$\begin{aligned} s(A) &= 2^{-2/1.2} = 0.31 \\ s(B) &= 2^{-2/1.2} = 0.31 \\ s(C) &= 2^{-1/1.2} = 0.56 \end{aligned}$$

$$c(3) = 2(\ln(3 - 1) + 0.577) - \left(\frac{2(3-1)}{3}\right) = 1.2$$

Ejemplo (en Python)



A

B

C

1

2

100

Córners

```
from sklearn.ensemble import IsolationForest
import pandas as pd

df = pd.DataFrame({'q':[1,2,100], 'Id':['A','B','C']})
X = df[['q']]

x = IsolationForest(random_state=12345, contamination='auto',
                    n_estimators=100, behaviour="new").fit(X)

iso_predictions = x.predict(X)
iso_score = x.score_samples(X)

sk_predictions = pd.DataFrame({
    "Id": df.Id,
    "Anomalía": list(map(lambda x: 1*(x == -1), iso_predictions)),
    "Score": -iso_score
})

display(sk_predictions.sort_values('Score', ascending=False).reset_index(drop=True))
```

	Id	Anomalía	Score
0	C	1	0.563219
1	A	0	0.317216
2	B	0	0.317216

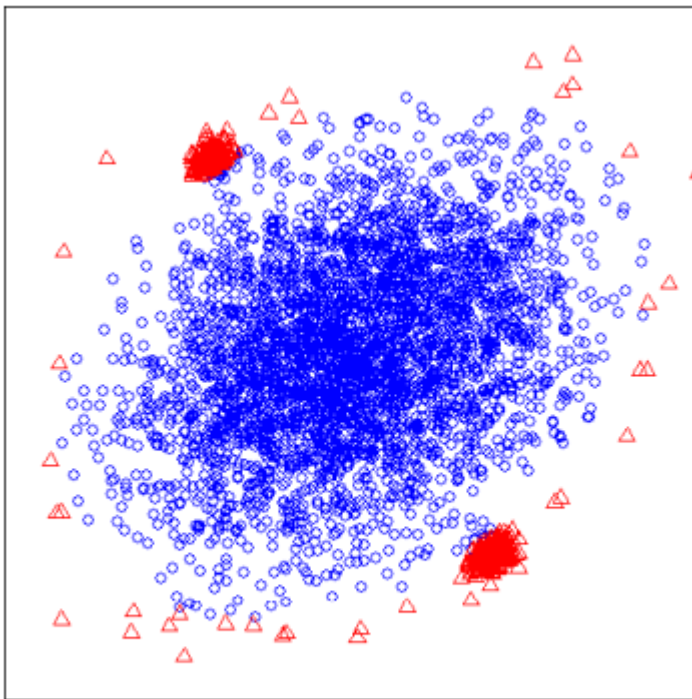
$$s(A) = 2^{-2/1.2} = 0.31$$

$$s(B) = 2^{-2/1.2} = 0.31$$

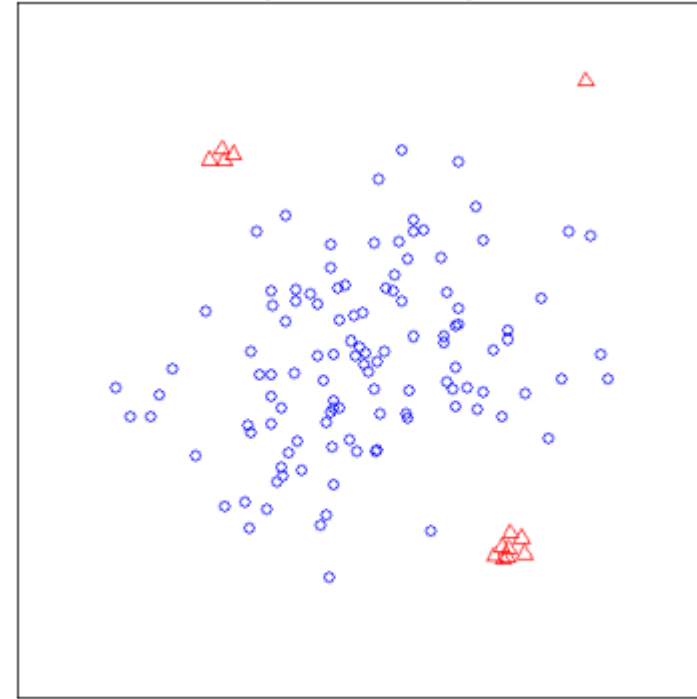
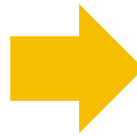
$$s(C) = 2^{-1/1.2} = 0.56$$

Submuestreo

- El algoritmo tiene dos parámetros: el número de árboles y el tamaño del submuestreo
- El submuestreo alivia los efectos del *swamping* (puntos anómalos próximos a los normales) y el *masking* (puntos anómalos muy concentrados)



(a) Original sample
(4096 instances)



(b) Sub-sample
(128 instances)

Detector de amaños



AMAÑOS

**La jueza lleva a juicio a los 37
jugadores del Levante-Zaragoza**

Detector de amaños

● El algoritmo se utilizó en el caso del posible amañó del Levante-Zaragoza (2010-2011)

● Utilizando las siguientes variables:

- Tiros a favor
- Tiros en contra
- Paradas portero propio
- Paradas portero rival
- Córners
- Córners en contra



Partido	Anomalía	Score
Levante vs Real Zaragoza	1	0.621102
Levante vs Real Madrid	0	0.578562
Levante vs Getafe	0	0.516863
Levante vs Racing de Santander	0	0.509431
Levante vs Sporting de Gijón	0	0.507827
Levante vs Málaga	0	0.495437
Levante vs Almería	0	0.489195
Levante vs Osasuna	0	0.487978
Levante vs Deportivo de La Coruña	0	0.482814
Levante vs Mallorca	0	0.474890
Levante vs Atlético de Madrid	0	0.474801
Levante vs Athletic Club	0	0.468627
Levante vs Sevilla	0	0.461898
Levante vs Barcelona	0	0.456357
Levante vs Espanyol	0	0.456129
Levante vs Hércules	0	0.443379
Levante vs Valencia CF	0	0.440604
Levante vs Villarreal	0	0.424657
Levante vs Real Sociedad	0	0.414185

Partido	Anomalía	Score
Levante vs Real Zaragoza	1	0.660592
Barcelona vs Real Zaragoza	0	0.650594
Sevilla vs Real Zaragoza	0	0.551930
Villarreal vs Real Zaragoza	0	0.498179
Getafe vs Real Zaragoza	0	0.497299
Osasuna vs Real Zaragoza	0	0.494502
Deportivo de La Coruña vs Real Zaragoza	0	0.489958
Almería vs Real Zaragoza	0	0.483145
Athletic Club vs Real Zaragoza	0	0.465174
Real Madrid vs Real Zaragoza	0	0.464799
Racing de Santander vs Real Zaragoza	0	0.461462
Mallorca vs Real Zaragoza	0	0.456427
Real Sociedad vs Real Zaragoza	0	0.454862
Atlético de Madrid vs Real Zaragoza	0	0.448973
Hércules vs Real Zaragoza	0	0.443293
Málaga vs Real Zaragoza	0	0.438020
Sporting de Gijón vs Real Zaragoza	0	0.435795
Valencia CF vs Real Zaragoza	0	0.425035
Espanyol vs Real Zaragoza	0	0.414364