Data Mining simple: ID3 tree implementation

1. First step was pre-processing the Original Data with accordance to KDD
2. This process is out of the current scope. The purpose here is to get close understanding of ID3 tree, by processing very simple data-base
3. Thus the input for this tutorial is a "ML ready Table" like shown at the bottom:

D=12 records, 4 attributes, 5 classes (reptiles, fishes, mammals, amphibian, birds

birth	fly	water	legs	**family**
n	n	n	n	reptiles
n	n	y	n	fishes
y	n	y	n	mammals
n	n	s	y	amphibian
y	n	n	y	mammals
n	n	s	y	amphibian
y	n	y	n	mammals
n	y	n	y	birds
n	n	s	y	birds
n	n	s	y	reptiles
y	n	y	n	fishes
y	n	n	y	mammals

Data Mining simple: ID3 tree implementation

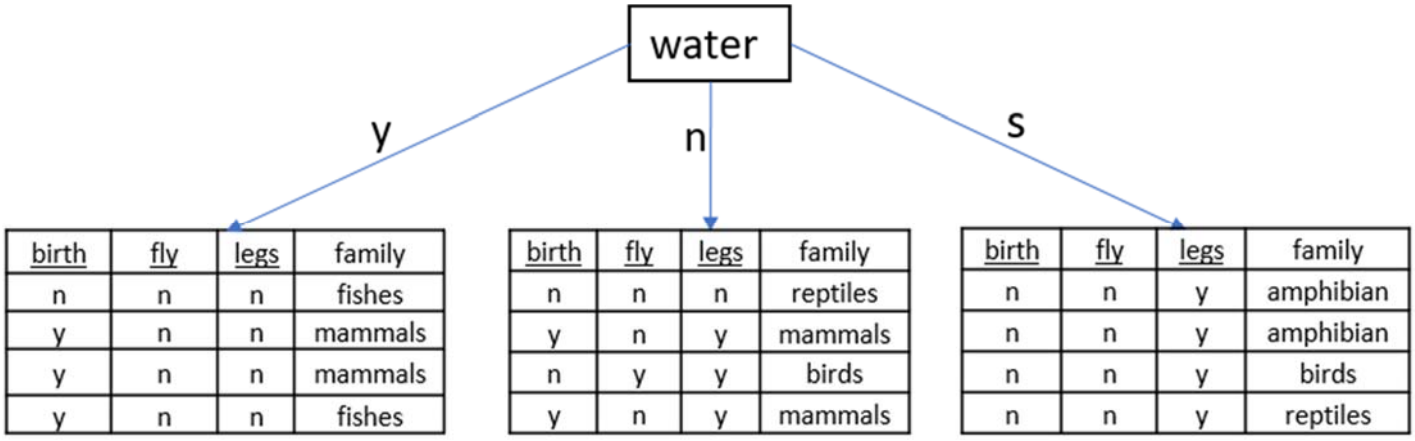
Algorithm: classic ID3, no pruning, Information Gain

$$Info(D) = -\sum_{i=1}^{m} p_i \cdot \log_2(p_i) = -(\frac{1}{6}\log(\frac{1}{6}) \cdot 4 + \frac{1}{3}\log_2(\frac{1}{3})) = \boxed{2.251}$$

$$\begin{cases} Info_A(D) = -\sum_{j=1}^{v} \frac{|D_j|}{|D|} \cdot Info(D_j) \\ Gain = Info(D) - Info_A(D) \end{cases}$$

Gain 'birth': 0.813
Gain 'fly': 0.247
Gain 'water': 0.918 ➔ Max Info-Gain
Gain 'legs': 0.479

The tree - After first split(By "water")



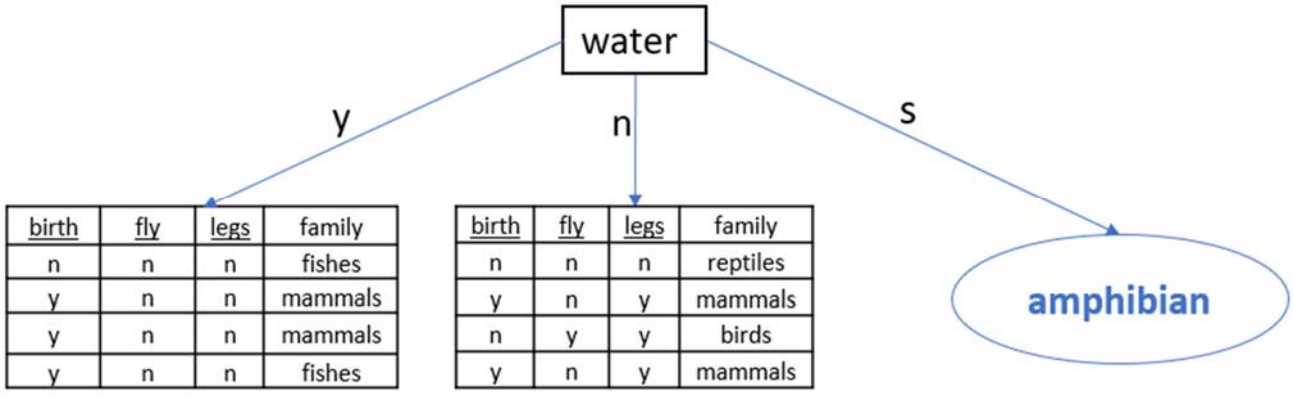
birth	fly	legs	family
n	n	n	fishes
y	n	n	mammals
y	n	n	mammals
y	n	n	fishes

birth	fly	legs	family
n	n	n	reptiles
y	n	y	mammals
n	y	y	birds
y	n	y	mammals

birth	fly	legs	family
n	n	y	amphibian
n	n	y	amphibian
n	n	y	birds
n	n	y	reptiles

In S branch, Info gains for all of the attributes equals 0, thus will choose the most common class: '**amphibian**'

Data Mining simple: ID3 tree implementation



birth	fly	legs	family
n	n	n	fishes
y	n	n	mammals
y	n	n	mammals
y	n	n	fishes

birth	fly	legs	family
n	n	n	reptiles
y	n	y	mammals
n	y	y	birds
y	n	y	mammals

Recursively calculating the measures for all of the brunches don't ends with a leaf.

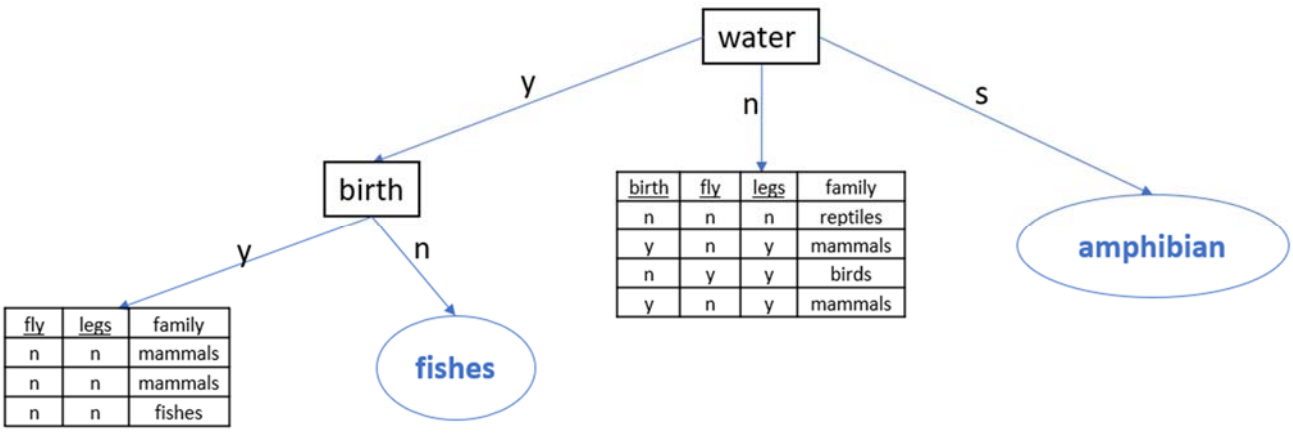
Y - branch:

```
Entropy=Info(D): 1.0

Info(Attribute): {'birth': 0.689, 'fly':1.0, 'legs': 1.0 }
Info-Gains: {'birth': 0.311, 'fly': 0.0, 'legs': 0.0}

MAX Information Gain = 'birth':0.311
```

birth	fly	legs	family
n	n	n	fishes
y	n	n	mammals
y	n	n	mammals
y	n	n	fishes



fly	legs	family
n	n	mammals
n	n	mammals
n	n	fishes

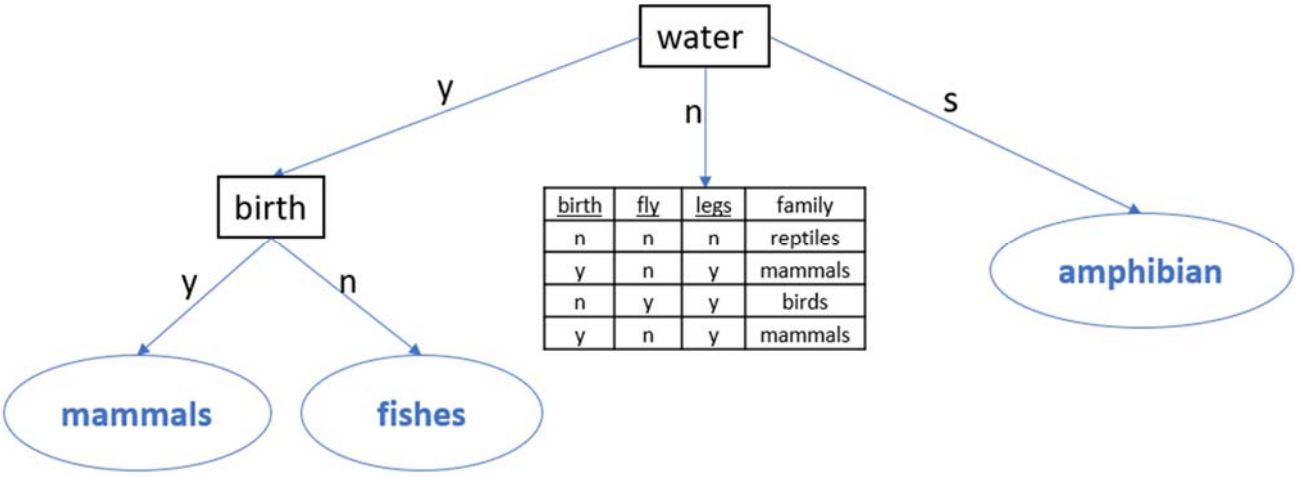
birth	fly	legs	family
n	n	n	reptiles
y	n	y	mammals
n	y	y	birds
y	n	y	mammals

Stop condition has reached in: water->y->birth->y, **Information Gain=0 for the atrr.**

Classification according most common: **"amphibian"**

Data Mining simple: ID3 tree implementation

In the same way: "mammals" for the birth->y branch



birth	fly	legs	family
n	n	n	reptiles
y	n	y	mammals
n	y	y	birds
y	n	y	mammals

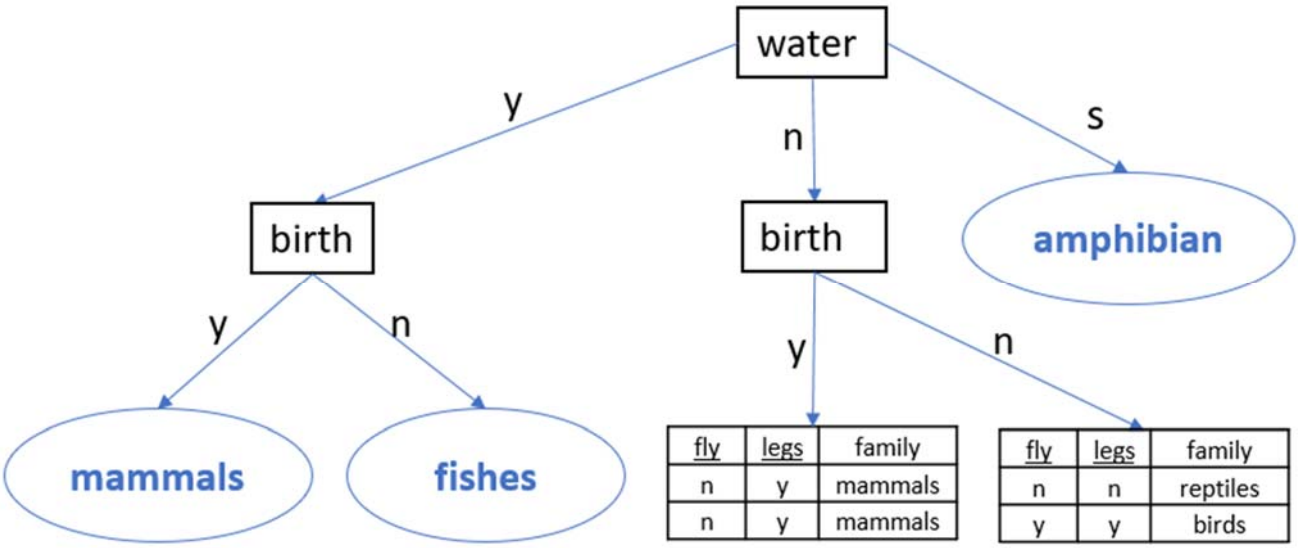
Water->n :

```
Entropy=info(D): 1.5

Info(Attribute): {'birth': 0.5, 'fly':0.689, 'legs': 0.689 }
Info - Gains: {'birth': 1.0, 'fly': 0.811, 'legs': 0.811}


MAX Information Gain = 'birth':1.0
```

birth	fly	legs	family
n	n	n	reptiles
y	n	y	mammals
n	y	y	birds
y	n	y	mammals



fly	legs	family
n	y	mammals
n	y	mammals

fly	legs	family
n	n	reptiles
y	y	birds

Stop condition in : water->n->birth->y, homogeneous records class: "mammals"

in water->n->birth->n, split randomly by "fly" (the same inf.Gain)

Data Mining simple: ID3 tree implementation

The Final Tree