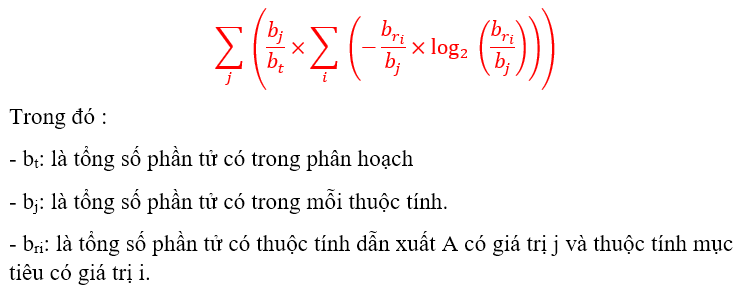
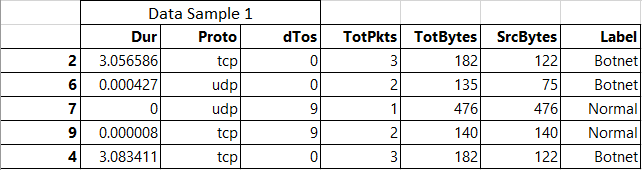
**Giải tay Random Forest**



**Bảng Data Sample 1**



|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **Dur** | | **Proto** | | **dTos** | | **TotPkts** | | | **TotBytes** | | | | **SrcBytes** | | | |
| **3** | **0** | **tcp** | **udp** | **0** | **9** | **3** | **2** | **1** | **182** | **135** | **476** | **140** | **122** | **75** | **476** | **140** |
| 2\* | 6\* | 2\* | 6\* | 2\* | 7 | 2\* | 6\* | 7 | 2\* | 6\* | 7 | 9 | 2\* | 6\* | 7 | 9 |
| 4\* | 7 | 9 | 7 | 6\* | 9 | 4\* | 9 |  | 4\* |  |  |  | 4\* |  |  |  |
|  | 9 | 4\* |  | 4\* |  |  |  |  |  |  |  |  |  |  |  |  |

Dùng \* để kí hiệu giá trị có nhãn botnet

Giải

Độ hỗn loạn TB của Dur:

= 0.551

Độ hỗn loạn TB của Proto:

0.551 + 0.4 = 0.951

Độ hỗn loạn TB của dTos:

0

Độ hỗn loạn TB của TotPkts:

= 0.4

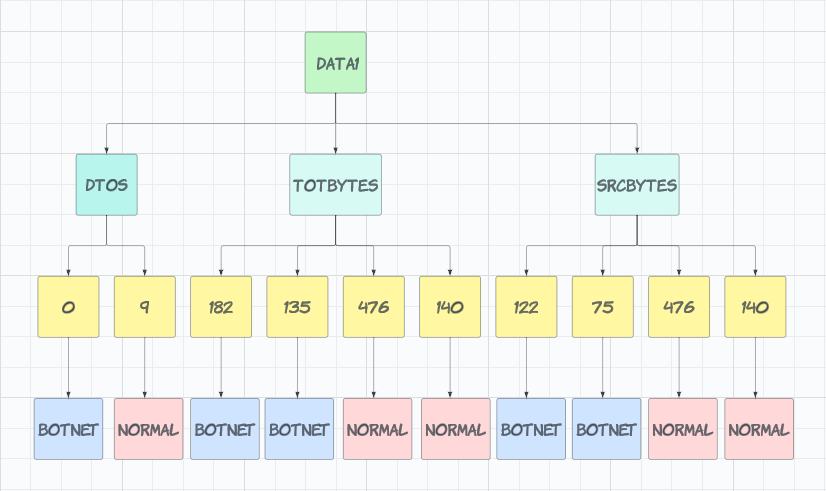
Độ hỗn loạn TB của TotBytes:

= 0

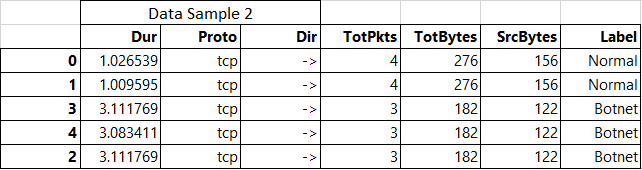
Độ hỗn loạn TB của SrcBytes:

= 0

* Chọn thuộc tính dTos, TotBytes, SrcBytes để tạo cây vì có độ hỗn loạn nhỏ nhất.



**Bảng Data Sample 2**



|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **Dur** | | **Proto** | **Dir** | **TotPkts** | | **TotBytes** | | **SrcBytes** | |
| **1** | **3** | **tcp** | **->** | **4** | **3** | **276** | **182** | **156** | **122** |
| 0 | 3\* | 0 | 0 | 0 | 3\* | 0 | 3\* | 0 | 3\* |
| 1 | 4\* | 1 | 1 | 1 | 4\* | 1 | 4\* | 1 | 4\* |
|  | 2\* | 3\* | 3\* |  | 2\* |  | 2\* |  | 2\* |
|  |  | 4\* | 4\* |  |  |  |  |  |  |
|  |  | 2\* | 2\* |  |  |  |  |  |  |

Dùng \* để kí hiệu giá trị có nhãn botnet

Giải

Độ hỗn loạn TB của Dur:

= 0

Độ hỗn loạn TB của Proto:

= 0.971

Độ hỗn loạn TB của Dir:

= 0.971

Độ hỗn loạn TB của TotPkts:

= 0

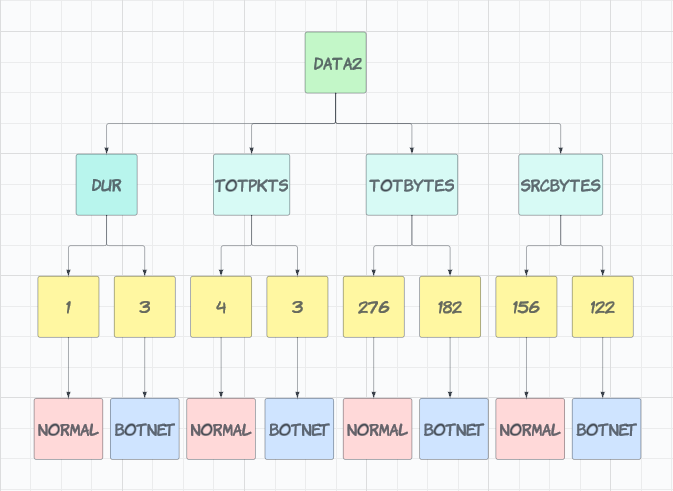
Độ hỗn loạn TB của TotBytes:

= 0

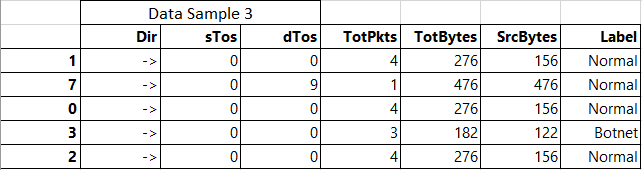
Độ hỗn loạn TB của SrcBytes:

= 0

* Chọn thuộc tính Dur, TotPkts, TotBytes, SrcBytes để tạo cây vì có độ hỗn loạn nhỏ nhất.



**Bảng Data Sample 3**



|  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **Dir** | **sTos** | **dTos** | | **TotPkts** | | | **TotBytes** | | | **SrcBytes** | | |
| **->** | **0** | **0** | **9** | **1** | **3** | **4** | **182** | **276** | **476** | **122** | **156** | **476** |
| 1 | 1 | 1 | 7 | 7 | 3\* | 1 | 3\* | 1 | 7 | 3\* | 1 | 7 |
| 7 | 7 | 0 |  |  |  | 0 |  | 0 |  |  | 0 |  |
| 0 | 0 | 3\* |  |  |  | 2 |  | 2 |  |  | 2 |  |
| 3\* | 3\* | 2 |  |  |  |  |  |  |  |  |  |  |
| 2 | 2 |  |  |  |  |  |  |  |  |  |  |  |

Dùng \* để kí hiệu giá trị có nhãn botnet

Giải

Độ hỗn loạn TB của Dir:

= 0.722

Độ hỗn loạn TB của sTos:

= 0.722

Độ hỗn loạn TB của dTos:

0.649

Độ hỗn loạn TB của TotPkts:

= 0

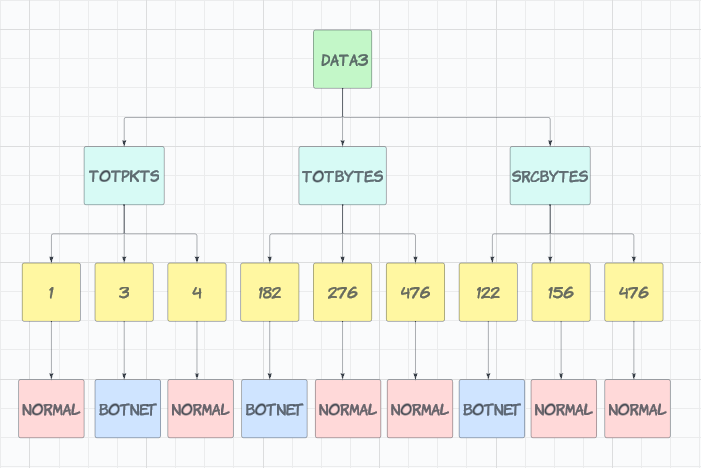
Độ hỗn loạn TB của TotBytes:

= 0

Độ hỗn loạn TB của SrctBytes:

= 0

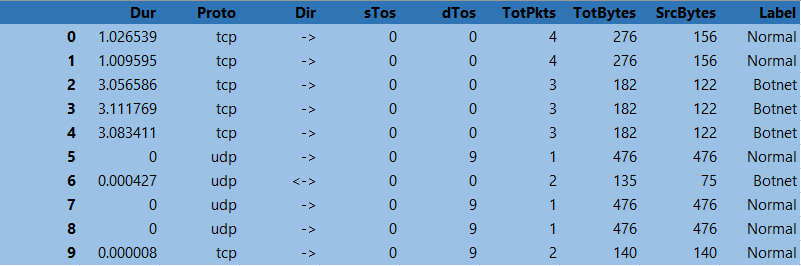
* Chọn thuộc tính TotPkts, TotBytes, SrcBytes để tạo cây vì có độ hỗn loạn nhỏ nhất.



**Giải tay thuật toán Naïve Bayes**

Công thức định lý Bayes: P(Ci | X) =

* P(Ci|X): xác suất của mục tiêu Ci với điều kiện có đặc trưng X
* P(X|Ci): xác suất của đặc trưng X khi đã biết mục tiêu y
* P(Ci) gọi là xác suất của mục tiêu của đặc trưng Ci
* P(X) gọi là xác suất của mục tiêu của đặc trưng X



Với bảng dữ liệu trên, xác định lớp của mẫu: X = (Dur = 0, Proto = udp, Dir = <->, sTos = 0, dTos = 0, TotPkts = 2, TotBytes = 135, SrcBytes = 75) sẽ rơi vào lớp nào ?

Các lớp:

* **C1: Normal**
* **C2: Botnet**

Tính giá trị xác suất trước cho mỗi phân lớp:

* **P(C1)** = 6/10 = 0.6
* **P(C2)** = 4/10 = 0.4

Tính giá trị xác suất của mỗi giá trị thuộc tính đối với mỗi phân lớp:

* P(Dur = 0 | C1) = 4/6
* P(Proto = udp | C1) = 3/6
* P(Dir = <-> | C1) = 0
* P(sTos = 0 | C1) = 6/6
* P(dTos = 0 | C1) = 2/6
* P(TotPkts = 2 | C1) = 1/6
* P(TotBytes = 135 | C1) = 0
* P(SrcBytes = 75 | C1) = 0
* P(Dur = 0 | C2) = 1/4
* P(Proto = udp | C2) = 1/4
* P(Dir = <-> | C2) = 1/4
* P(sTos = 0 | C2) = 4/4
* P(dTos = 0 | C2) = 4/4
* P(TotPkts = 2 | C2) = 1/4
* P(TotBytes = 135 | C2) = 1/4
* P(SrcBytes = 75 | C2) = 1/4

Tính xác suất có thể xảy ra của mẫu X đối với mỗi phân lớp:

Đối với phân lớp C1:

**P(X | C1)** = P(Dur = 0 | C1) x P(Proto = udp | C1) x P(Dir = <-> | C1) x P(sTos = 0 | C1) x P(dTos = 0 | C1) x P(TotPkts = 2 | C1) x P(TotBytes = 135 | C1) x P(SrcBytes = 75 | C1) = 4/6 x 3/6 x 0 x 6/6 x 2/6 x 1/6 x 0 x 0 = 0

Đối với phân lớp C2:

**P(X | C2)** = P(Dur = 0 | C2) x P(Proto = udp | C2) x P(Dir = <-> | C2) x P(sTos = 0 | C2) x P(dTos = 0 | C2) x P(TotPkts = 2 | C2) x P(TotBytes = 135 | C2) x P(SrcBytes = 75 | C2) = 1/4 x 1/4 x 1/4 x 4/4 x 4/4 x 1/4 x 1/4 x 1/4 = 0.0002

Xác định phân lớp có thể nhất:

Đối với phân lớp C1: P(C1) x P(X | C1) = 0.6 x 0 = 0

Đối với phân lớp C2: P(C2) x P(X | C2) = 0.4 x 0.0002 = 0.00008

* X sẽ có label thuộc phân lớp C2.