e--



**TRƯỜNG ĐẠI HỌC KHOA HỌC TỰ NHIÊN - TP HCM**

**KHOA CÔNG NGHỆ THÔNG TIN**

**MÔN CƠ SỞ TRÍ TUỆ NHÂN TẠO**

**LỚP CỬ NHÂN TÀI NĂNG 2016**

Machine Learning

ĐOÀN QUANG TUẤN 1612780

**LÊ HOÀNG SANG 1612554**

NHÓM SINH VIÊN THỰC HIỆN

**BÁO CÁO ĐỒ ÁN**

1. **Tìm hiểu công cụ Weka.**
   1. **Giới thiệu weka**

Weka (viết tắt của Waikato Environment for Knowledge Analysis) là một bộ phần mềm [học máy](https://vi.wikipedia.org/wiki/H%E1%BB%8Dc_m%C3%A1y) được [Đại học Waikato](https://vi.wikipedia.org/w/index.php?title=%C4%90%E1%BA%A1i_h%E1%BB%8Dc_Waikato&action=edit&redlink=1), [New Zealand](https://vi.wikipedia.org/wiki/New_Zealand) phát triển bằng [Java](https://vi.wikipedia.org/wiki/Java_(ng%C3%B4n_ng%E1%BB%AF_l%E1%BA%ADp_tr%C3%ACnh)). Weka là [phần mềm tự do](https://vi.wikipedia.org/wiki/Ph%E1%BA%A7n_m%E1%BB%81m_t%E1%BB%B1_do) phát hành theo [giấy phép công cộng GNU](https://vi.wikipedia.org/wiki/Gi%E1%BA%A5y_ph%C3%A9p_C%C3%B4ng_c%E1%BB%99ng_GNU).

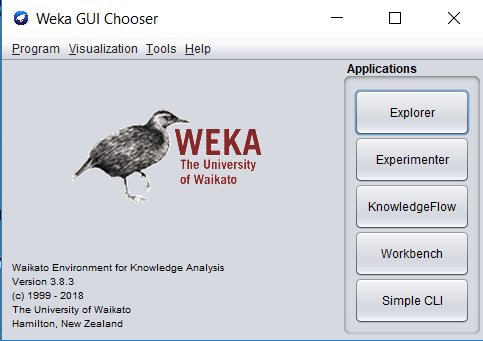
Mục đích nhằm xây dựng một công cụ hiện đại đẻ phát triển các kỹ thuật trong máy học và áp dụng chúng vào bài toán khai phá dữ liệu trong thực tế.

WEKA được xây dựng bằng ngôn ngữ Java, cấu trúc gồm hơn 600 lớp , tổ chức thành 10 packages.

Các chức năng chính của phần mềm :

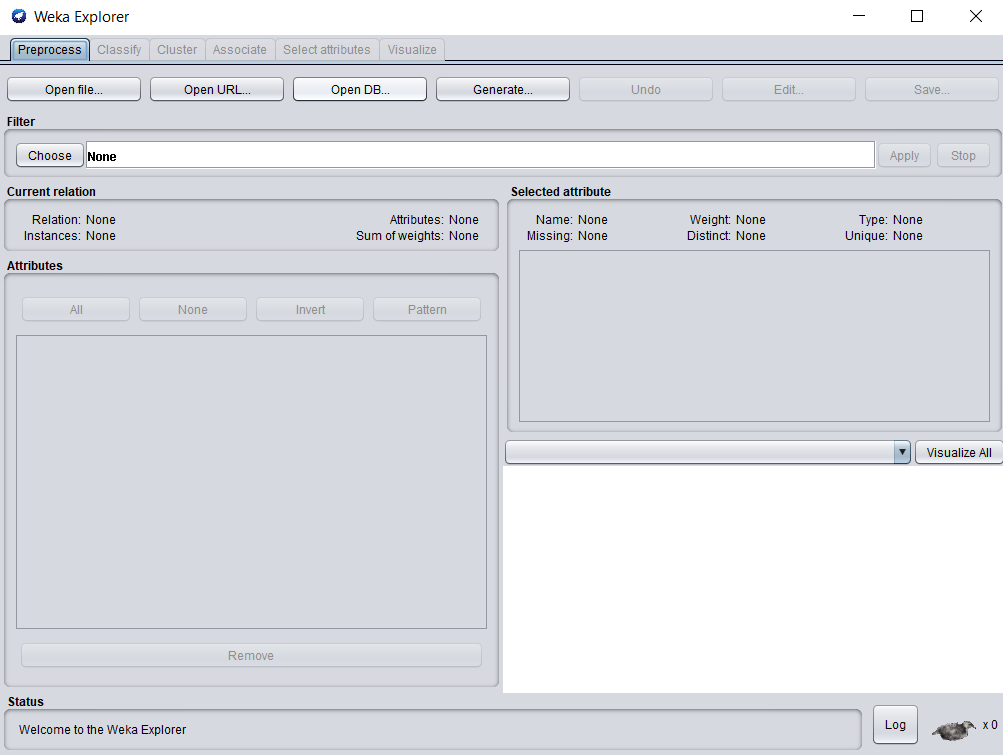
* Khảo sát dữ liệu : tiền xử lí dữ liệu, phân lớp, gom nhóm dữ liệu, và khai thác luật kết hợp .
* Thực nghiệm mô hình: cung cấp phương tiện để kiểm chứng, đánh giá các mô hình học .
* Biểu diễn trực quan dữ liệu bằng nhiều dạng đồ thị khác nhau.
  1. **Các chức năng và cách sử dụng căn bản.**

Phần mềm gồm 5 môi trường: Explorer, Experimenter, KnowledgeFlow, Workbench và Simple CLI.



1. **Môi trường Explorer.**

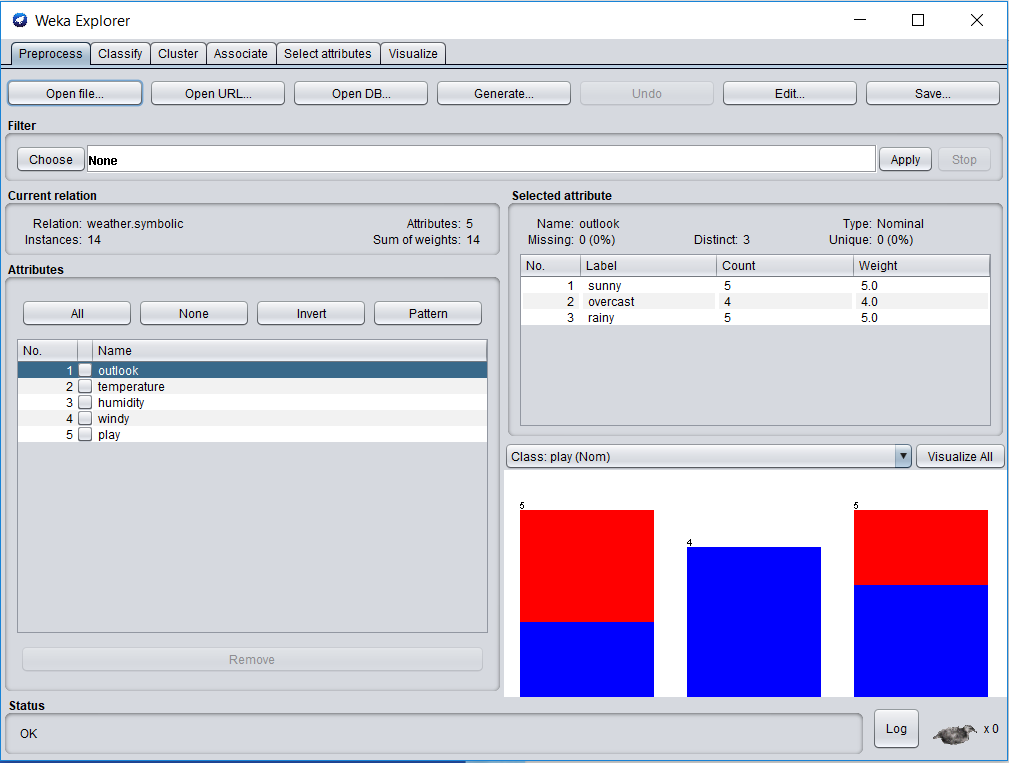
Giao diện môi trường Explorer gồm: preprocess, classify, cluster, associate, select attributes, visualize.

****

* **Preprocess** – **là bước tiền xử lý dữ liệu.**

Trong qui trình khai phá dữ liệu, công việc xử lý dữ liệu trước khi đưa vào các mô hình là rất cần thiết, bước này làm cho dữ liệu có được ban đầu qua thu thập dữ liệu (gọi là dữ liệu gốc ordinal data) có thể áp dụng được (thích hợp) với các mô hình khai phá dữ liệu  
(data mining model) cụ thể. Các công việc cụ thể của tiền xử lý dữ liệu bao gồm những công việc như:

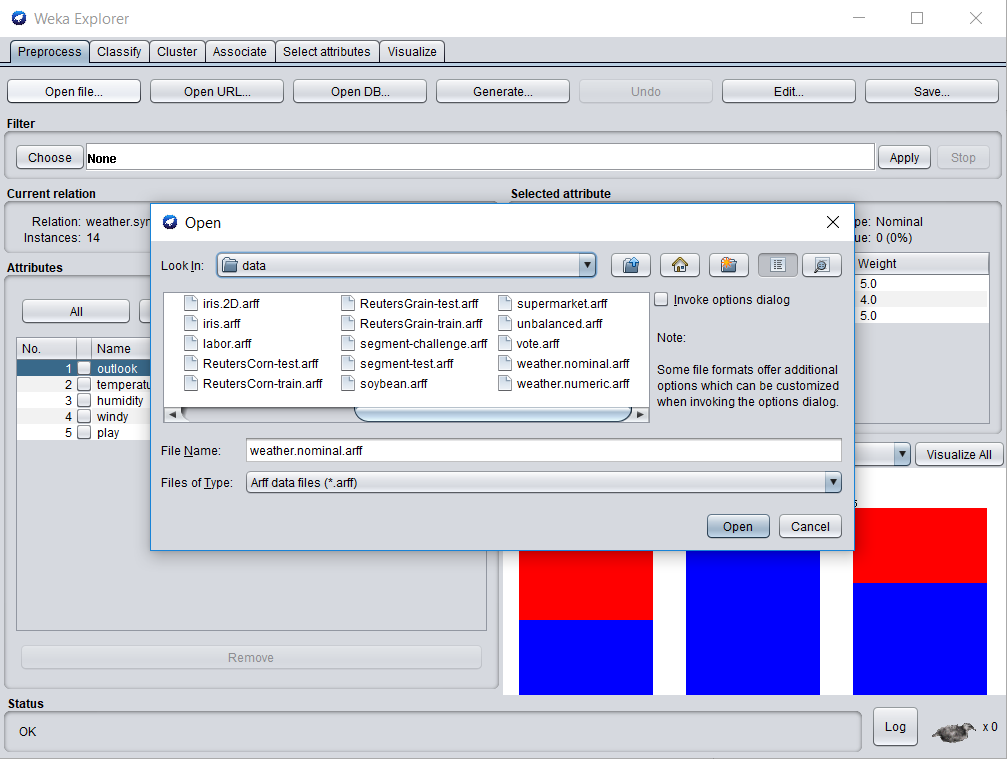
* *Filtering Attributes:* Chọn các thuộc tính phù hợp với mô hình.
* *Filtering samples:* Lọc các mẫu (instances, patterns) dữ liệu cho mô hình.
* *Clean data:* Làm sạch dữ liệu như xóa bỏ các dữ liệu bất thường (Outlier).
* *Transformation:* Chuyển đổi dữ liệu cho phù hợp với các mô hình như chuyển đổi dữ liệu từ numeric qua nomial hay ordinal.
* *Discetization (rời rạc hóa dữ liệu):* Nếu bạn có dữ liệu liên tục nhưng một vài mô hình chỉ áp dụng cho các dữ liệu rời rạc (luật kết hợp chẳng hạn) thì bạn phải thực hiện việc rời rạc hóa dữ liệu.

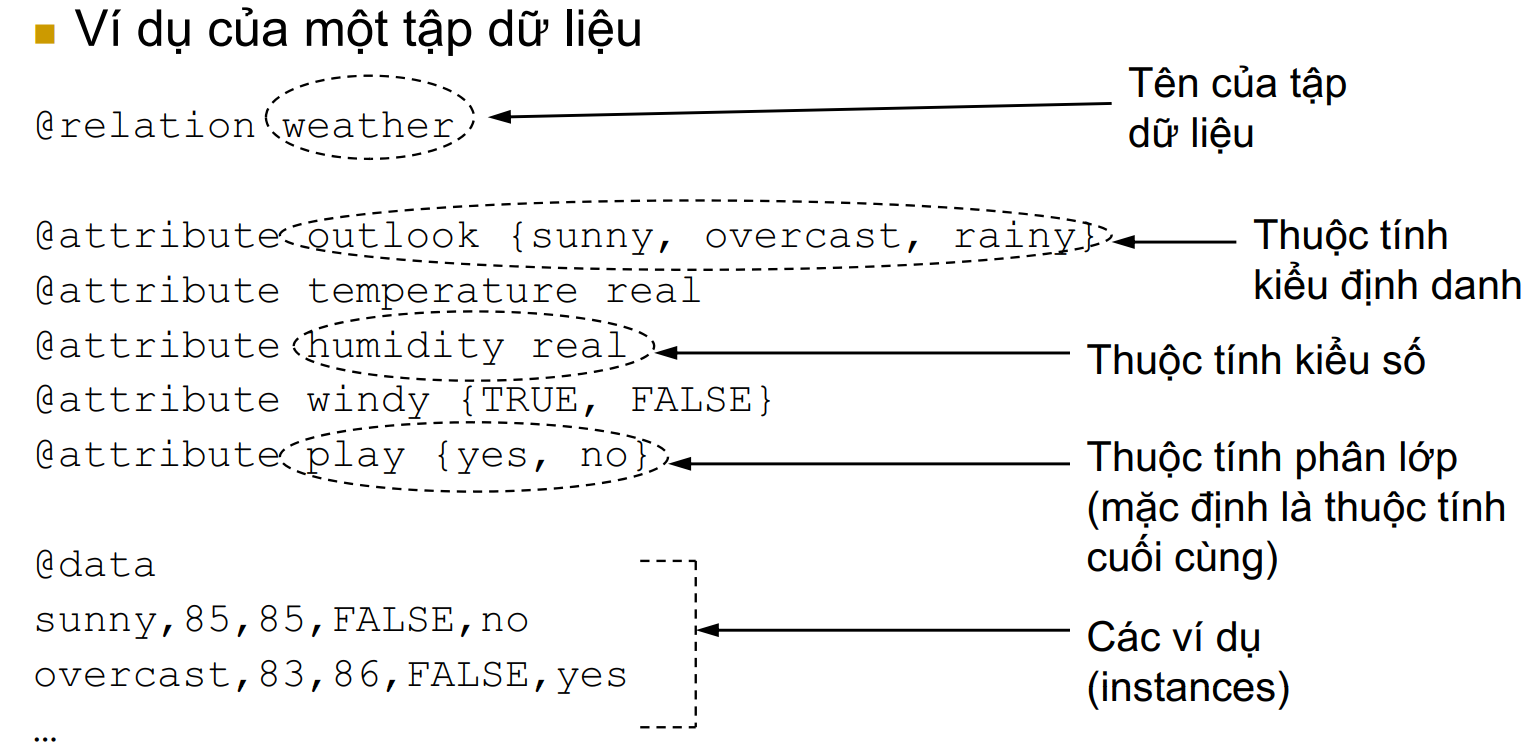


Dữ liệu có thể được nhập vào từ một tập tin có khuôn dạng ARFF, CSV.

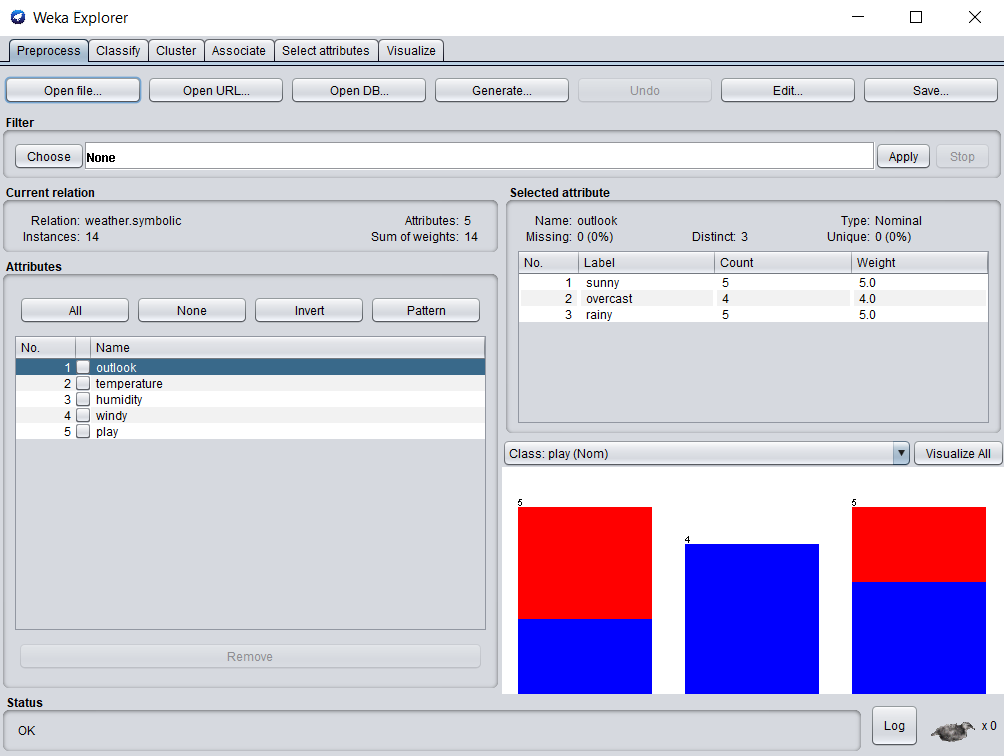
Dữ liệu có thể được đọc từ một địa chỉ URL, hoặc từ một cơ sở dữ liệu thông qua JDBC.

* Chọn kiểu open ứng với nguồn dữ liệu để import một bộ dữ liệu vào weka.



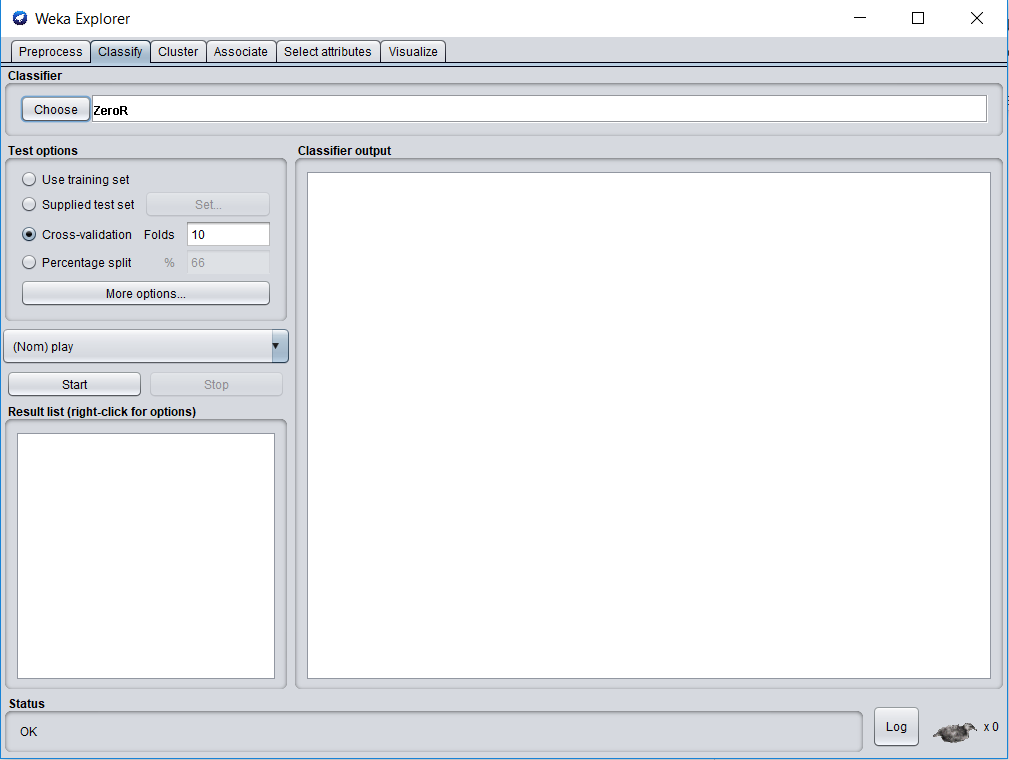


Sau khi mở tập dữ liệu ta có giao diện như sau:



* Phần **filter** thiết lập kiểu lọc (choose) và apply.
* Phần **Current relation** nêu các mô tả cơ bản của tập dữ liệu.
* Phần **Attributes** chứa các thuộc tính của đối tượng trong dataset.
* Phần **Selected attribute** cho biết một số đặc tính cơ bản của thuộc tính đang được chọn.
* Phần biểu đồ mô tả phân bố của thuộc tính đang chọn. Mỗi màu trên biểu đồ đại diện cho 1 class.
* **Classify.**

Giao diện classify:



Các bộ phân lớp (Classifiers) của WEKA tương ứng với  
các mô hình dự đoán các đại lượng kiểu định danh (phân lớp) hoặc các đại lượng kiểu số (hồi quy/dự đoán).

Các kỹ thuật phân lớp được hỗ trợ bởi WEKA:

* Naïve Bayes classifier and Bayesian networks.
* Decision trees.
* Instance-based classifiers.
* Support vector machines.
* Neural networks.
* …
* Mô tả về giao diện Classify:
* **Classifiers** để chọn bộ phân lớp (gồm các bộ ở trên).
* **Test options** lựa chọn các tùy chọn cho việc kiểm tra.
* **Use training set**. Bộ phân loại học được sẽ được đánh giá trên tập học (training).
* **Supplied test set**. Sử dụng một tập dữ liệu khác (với tập học) để cho việc đánh giá.
* **Cross-validation**. Tập dữ liệu sẽ được chia đều thành k tập (folds) có kích thước xấp xỉ nhau, và bộ phân loại học được sẽ được đánh giá bởi phương pháp *cross-validation.*
* **Percentage split**. Chỉ định tỷ lệ phân chia tập dữ liệu đối với việc đánh giá.
* **More options…**- **Output model**. Hiển thị bộ phân lớp học được  
  - **Output per-class stats**. Hiển thị các thông tin thống kê về precision/recall đối với mỗi lớp  
  - **Output entropy evaluation measures**. Hiển thị đánh giá độ hỗn tạp (entropy) của tập dữ liệu.  
  - **Output confusion matrix**. Hiển thị thông tin về ma trận lỗi phân lớp (confusion matrix) đối với phân lớp học được.  
  - **Store predictions for visualization**. Các dự đoán của bộ phân lớp được lưu lại trong bộ nhớ, để có thể được hiển thị sau đó.  
  - **Output predictions**. Hiển thị chi tiết các dự đoán đối với tập kiểm tra  
  - **Cost-sensitive evaluation**. Các lỗi (của bộ phân lớp) được xác định dựa trên ma trận chi phí (cost matrix) chỉ định.  
  - **Random seed for XVal / % Split**. Chỉ định giá trị *random seed* được sử dụng cho quá trình lựa chọn ngẫu nhiên các ví dụ cho tập kiểm tra.
* **Cluster.**
* **Associate.**
* **Select attributes.**
* **Visualize.**

1. **Sử dụng Weka để chạy thuật toán ID3.**
2. **Mô tả dữ liệu Zoo.**

* Số mẫu trong tập dữ liệu là 101 loại động vật.
* Tên và ý nghĩa các thuộc tính (18):

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **STT** | **Thuộc tính** | **Kiểu dữ liệu** | **Ý nghĩa** | **Miền giá trị** |
| 1 | name | String | Tên động vật | Chuỗi |
| 2 | hair | Boolean | Tóc | 1,0 |
| 3 | feathers | Boolean | Lông | 1,0 |
| 4 | eggs | Boolean | Đẻ trứng | 1,0 |
| 5 | milk | Boolean | Có sữa | 1,0 |
| 6 | airbone | Boolean | Biết bay | 1,0 |
| 7 | aquatic | Boolean | Đông vật sống dưới nước | 1,0 |
| 8 | predator | Boolean | Động vật ăn thịt | 1,0 |
| 9 | toothed | Boolean | Có răng | 1,0 |
| 10 | backbone | Boolean | Có xương sống | 1,0 |
| 11 | breathes | Boolean | Có thở | 1,0 |
| 12 | venomous | Boolean | Có độc | 1,0 |
| 13 | fins | Boolean | Có vây | 1,0 |
| 14 | legs | Numeric | Số chân | 0,2,4,5,6,8 |
| 15 | tail | Boolean | Có đuôi | 1.0 |
| 16 | domestic | Boolean | Động vật được nuôi | 1,0 |
| 17 | type | Numeric | Lớp động vật | 1,2,3,4,5,6,7 |

* Danh sách phân lớp, đặt tên cho mỗi phân lớp:

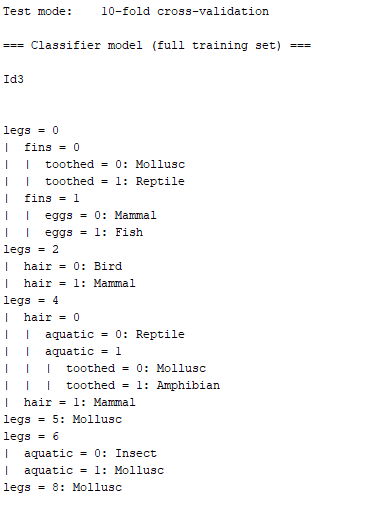
|  |  |  |  |
| --- | --- | --- | --- |
| **Type** | **Đặt tên** | **Số lượng** | **Động vật thuộc lớp** |
| 1 | Mammal | 44 | aardvark, antelope, bear, boar, buffalo, calf, cavy, cheetah, deer, dolphin, elephant, fruitbat, giraffe, girl, goat, gorilla, hamster, hare, leopard, lion, lynx, mink, mole, mongoose, opossum, oryx, platypus, polecat, pony, porpoise, puma, pussycat, raccoon, reindeer, seal, sealion, squirrel, vampire, vole, wallaby,wolf |
| 2 | Bird | 20 | chicken, crow, dove, duck, flamingo, gull, hawk, kiwi, lark, ostrich, parakeet, penguin, pheasant, rhea, skimmer, skua, sparrow, swan, vulture, wren. |
| 3 | Reptile | 5 | pitviper, seasnake, slowworm, tortoise, tuatara. |
| 4 | Fish | 13 | bass, carp, catfish, chub, dogfish, haddock, herring, pike, piranha, seahorse, sole, stingray, tuna. |
| 5 | Amphibian | 3 | frog, newt, toad |
| 6 | Insect | 8 | flea, gnat, honeybee, housefly, ladybird, moth, termite, wasp. |
| 7 | Mollusc | 10 | clam, crab, crayfish, lobster, octopus, scorpion, seawasp, slug, starfish, worm |

1. **Sử dụng thuật toán ID3 để học ra cây quyết định.**

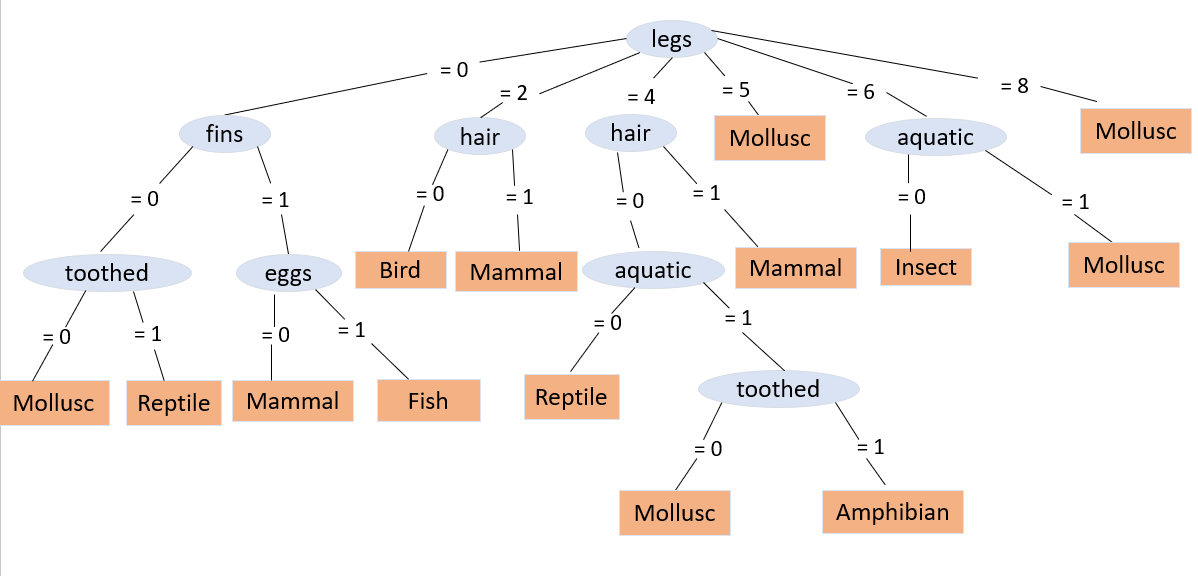
Sử dụng thuật toán id3 cho thuộc tính type. Tên của động vật chúng ta không cần xét đến nên ta sử dụng bộ dữ liệu không có thuộc tính name: **Zoo-train.arff**

- Test option: 10-fold cross validation

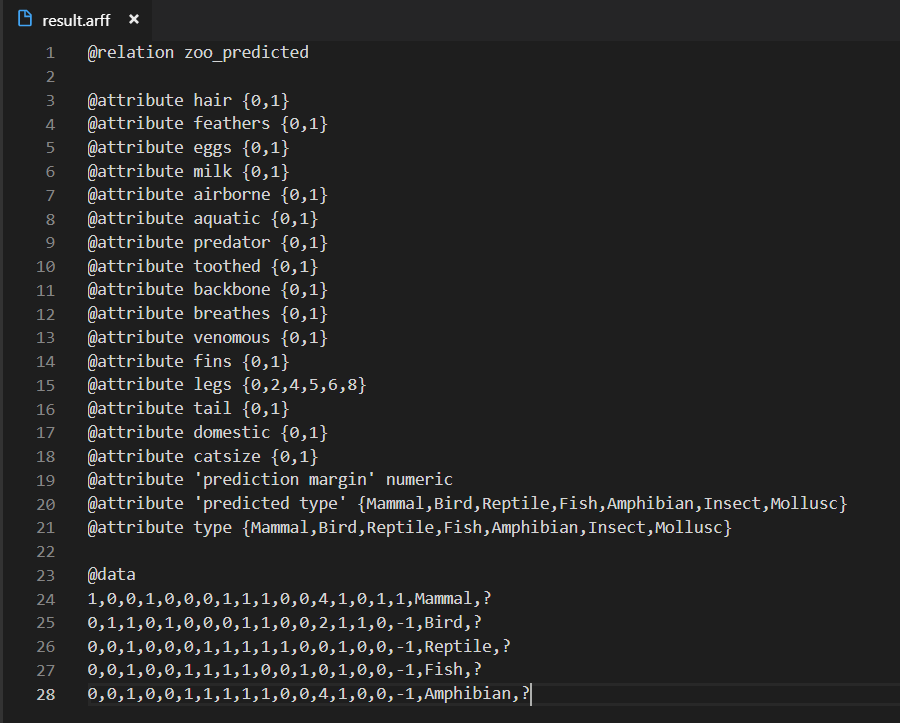
Ta được cây quyết định:



Cây được vẽ lại:



* Kết quả 5 mẫu đã cho dự đoán từ cây quyết định đã sinh ra là:



1. NameIsSecret,1,0,0,1,0,0,0,1,1,1,0,0,4,1,0,1, ? => Mamal
2. NameIsSecret,0,1,1,0,1,0,0,0,1,1,0,0,2,1,1,0, ? => Bird
3. NameIsSecret,0,0,1,0,0,0,1,1,1,1,1,0,0,1,0,0, ? => Reptile
4. NameIsSecret,0,0,1,0,0,1,1,1,1,0,0,1,0,1,0,0, ? => Fish
5. NameIsSecret,0,0,1,0,0,1,1,1,1,1,0,0,4,1,0,0, ? => Amphibian
6. **Chạy các thuật toán khác.**
   1. **Chương trình python cho giải thuật Naïve Bayes.**
   2. **Dùng weka để chạy thêm các thuật toán khác.**
7. **Naïve Bayes.**
8. **Tham khảo.**