



BÁO CÁO ĐỒ ÁN 3

CHỦ ĐỀ: MACHINE LEARNING



MÔN HỌC

CƠ SỞ TRÍ TUỆ NHÂN TẠO

Mục lục

1	Tìm hiểu công cụ Weka	4
1.1	Giới thiệu công cụ Weka	4
1.2	Ví dụ minh họa	11
2	Sử dụng Weka để chạy thuật toán ID3	16
2.1	Tạo tập tin Zoo.arff.....	16
2.2	Mô tả tổng quát về dữ liệu Zoo	17
2.3	Sử dụng thuật toán ID3.....	19
2.4	Kết quả dự đoán của 5 mẫu	20
3	Chạy các thuật toán khác	21
3.1	Chương trình Python cho giải thuật Naïve Bayes	21
3.2	Chạy với thuật toán J48	21
3.3	Chạy với thuật toán Naïve Bayes	26
3.4	Chạy với thuật toán IBK.....	27

Figure 1. Màn hình bắt đầu của Weka	4
Figure 2. Explorer	5
Figure 3. Cấu trúc file ARFF	5
Figure 4. Các bộ lọc filter	6
Figure 5. Classify	7
Figure 6. Cluster	8
Figure 7. Associate	9
Figure 8. Select attributes	10
Figure 9. Visualize	10
Figure 10. weather.arff	11
Figure 11. Preprocess weather.arff	11
Figure 12. Tab Classify	12
Figure 13. Test options	12
Figure 14. More options	13
Figure 15. Classify using J48	13
Figure 16. Chọn Visualize tree	14
Figure 17. Cây quyết định	15
Figure 18 Tạo tập tin Zoo.arff	16
Figure 19 Đưa dữ liệu vào phần mềm Weka	17
Figure 20 File Zoo sau khi chỉnh sửa	19
Figure 21 Cây sinh ra bởi thuật toán ID3	20
Figure 22 Kết quả dự đoán của 5 mẫu	20
Figure 23 Kết quả chạy chương trình	21
Figure 24. Chọn thuật toán J48	22
Figure 25. Kết quả phân lớp	22
Figure 26. Supplied test set	23
Figure 27. Test set	23
Figure 28. More options	24
Figure 29. Chọn PlainText	25
Figure 30. Kết quả test với thuật toán J48	25
Figure 31. Kết quả dự đoán 5 mẫu	26
Figure 32. Kết quả chạy với thuật toán Naïve Bayes	26
Figure 33. Kết quả dự đoán 5 mẫu	27
Figure 34. Kết quả chạy với thuật toán IBK	27
Figure 35. Kết quả dự đoán 5 mẫu	28

Thông tin nhóm

STT	MSSV	Tên thành viên	Email	Hoàn thành
1	1612348	Lý Vĩnh Lợi	vinhloiit1327@gmail.com	100%
2	1612756	Nguyễn Hữu Trường	ngoctruong9x.inc@gmail.com	100%

Phân công công việc

STT	Nội dung	Người thực hiện	Hoàn thành (%)
1	Tìm hiểu công cụ Weka	Trường	100
2	Sử dụng Weka để chạy thuật toán ID3	Lợi	100
3	Viết chương trình python cho giải thuật Naïve Bayes	Lợi	100
4	Sử dụng Weka chạy thêm các thuật toán khác	Trường	100

1 Tìm hiểu công cụ Weka

1.1 Giới thiệu công cụ Weka

- **Weka** (viết tắt của Waikato Environment for Knowledge Analysis) là một bộ phần mềm học máy được Đại học Waikato, New Zealand phát triển bằng Java. Weka là phần mềm tự do phát hành theo Giấy phép Công cộng GNU.
- Các tính năng chính
 - ❖ Chứa 1 tập các công cụ tiền xử lý dữ liệu, các giải thuật học máy, khai phá dữ liệu và các phương pháp thí nghiệm đánh giá
 - ❖ Giao diện đồ họa
 - ❖ Môi trường cho phép so sánh các giải thuật học máy và khai phá dữ liệu
- Môi trường làm việc:

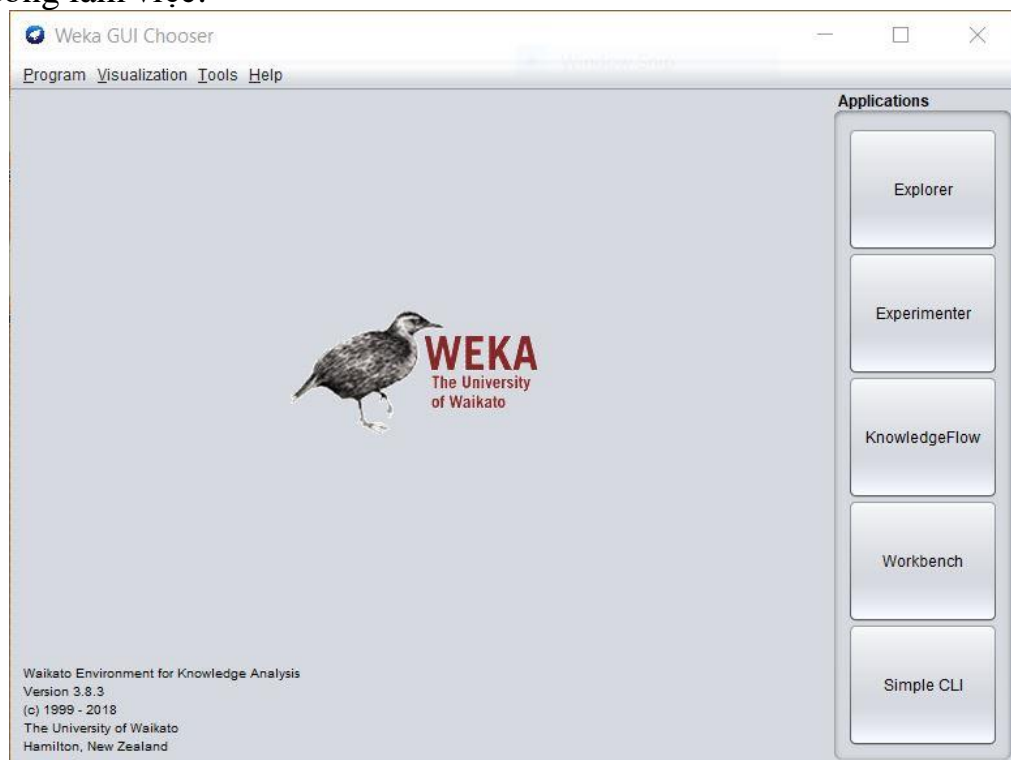


Figure 1. Màn hình bắt đầu của Weka

- ❖ **Simple CLI**: Giao diện dòng lệnh
- ❖ **Explorer**: Môi trường cho phép sử dụng tất cả các khả năng của WEKA để khám phá dữ liệu
- ❖ **Experiment**: Tiến hành các thí nghiệm và thực hiện các kiểm tra thống kê (statistical tests) giữa các mô hình học máy
- ❖ **KnowledgeFlow**: Môi trường cho phép tương tác đồ họa kéo thả để thiết kế các bước của 1 thí nghiệm

Chúng ta thao tác chủ yếu trong **Explorer**.

Click chọn **Explorer**, 1 cửa sổ mới sẽ hiện ra:

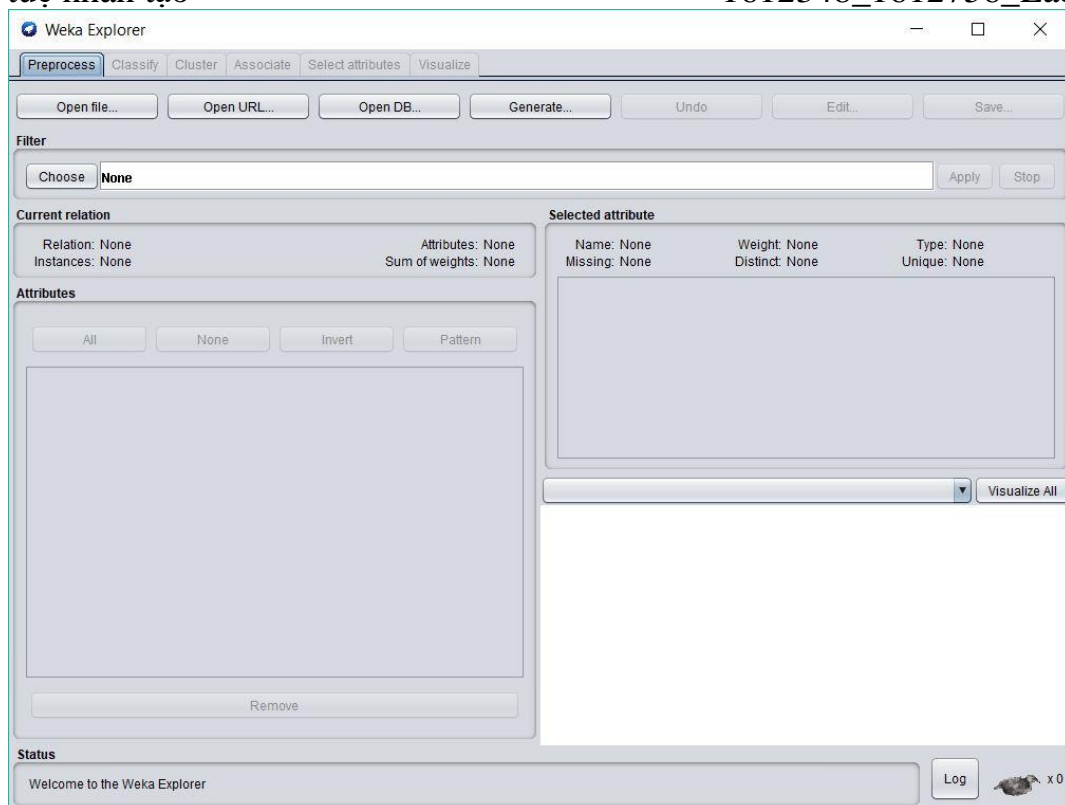


Figure 2. Explorer

Các chức năng trong **Explorer**:

- ❖ **Preprocess**: Chọn và tiền xử lý data
 - ❖ **Classify**: Huấn luyện và kiểm tra mô hình học máy
 - ❖ **Cluster**: Phân cụm (học các nhóm từ data)
 - ❖ **Associate**: Khám phá các luật từ dữ liệu
 - ❖ **Select attributes**: Lựa chọn các thuộc tính quan trọng
 - ❖ **Visualize**: Hiển thị biểu đồ tương tác 2 chiều
- Dữ liệu trong Weka
Weka chỉ làm việc với các tập tin văn bản (text) dạng ARFF. Dữ liệu được nhập vào có dạng ARFF hoặc CSV. Hoặc có thể đọc từ địa chỉ URL hay một cơ sở dữ liệu thông qua JDBC.

```

@relation weather
@attribute outlook {sunny, overcast, rainy}
@attribute temperature real
@attribute humidity real
@attribute windy {TRUE, FALSE}
@attribute play {yes, no}

@data
sunny, 85, 85, FALSE, no
overcast, 83, 86, FALSE, yes
  
```

Tên của tập dữ liệu

Thuộc tính kiểu định danh

Thuộc tính kiểu số

Thuộc tính phân lớp (mặc định là thuộc tính cuối cùng)

Các ví dụ (instances)

Figure 3. Cấu trúc file ARFF

- Tiền xử lí (filters)
 - ❖ Discretization
 - ❖ Normalization
 - ❖ Re-sampling
 - ❖ Attribute selection
 - ❖ Transforming
 - ❖ Combining

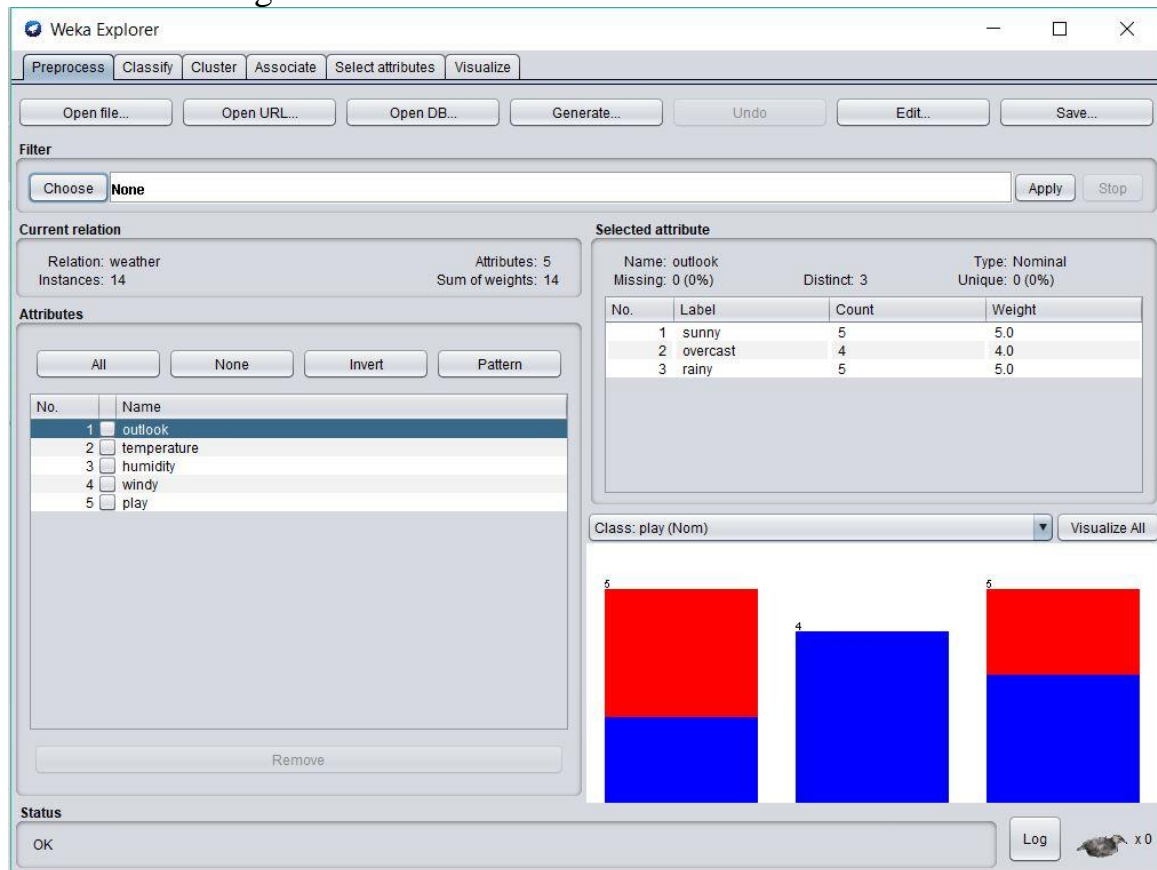


Figure 4. Các bộ lọc filter

- Phân lớp (classify)
 - ❖ Các kỹ thuật phân lớp
 - Naïve Bayesclassifier and Bayesian networks
 - Decision trees
 - Instance-based classifiers
 - Support vector machines
 - Neural networks
 - ...
 - ❖ Lựa chọn một bộ phân lớp
 - ❖ Lựa chọn các tùy chọn cho việc kiểm tra (test options)
 - Use training set: Bộ phân loại sẽ được đánh giá trên tập học
 - Supplied test set: Sử dụng một tập dữ liệu khác cho việc đánh giá
 - Percentage split: Chỉ định tỉ lệ phân chia tập dữ liệu đối với việc đánh giá
 - ❖ Các tùy chọn thêm

- 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 confusion matrix: Hiển thị thông tin về ma trận mỗi phân lớp đối với phân lớp học được
- ❖ Classifier output: Hiển thị các thông tin quan trọng
- ❖ Result list: Cung cấp một số tính năng hữu ích
 - Save model: Lưu lại mô hình tương ứng với bộ phân lớp học được vào một trong tập tin nhị phân.
 - Load model: Đọc lại một mô hình trước đó
 - Re-evaluate model on current test set: Đánh giá một mô hình học trước đó đối với tập kiểm tra hiện tại
 - Visualize classifier errors: Hiển thị cửa sổ biểu đồ thể hiện các kết quả của phân lớp.

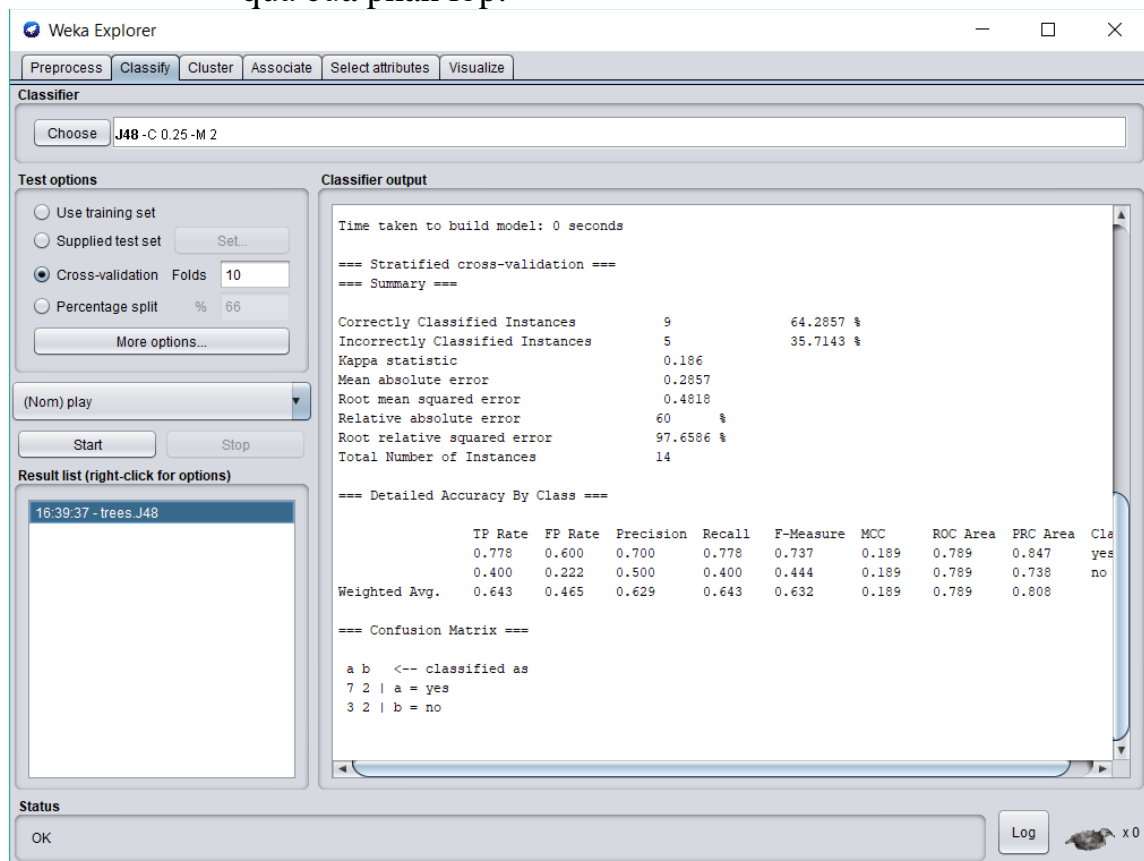


Figure 5. Classify

- Phân cụm (cluster builders)
 - ❖ Lựa chọn một bộ phân cụm (các kỹ thuật: Expectation maximization, k-Means, ...)
 - ❖ Lựa chọn chế độ phân cụm: Use training test, Supplied test set, Percentage split, Classes to clusters evaluation.
 - ❖ Store clusters for visualization: Lưu trữ các bộ phân cụm để hiển thị sau
 - ❖ Ignore attributes: Lựa chọn các thuộc tính sẽ không tham gia vào quá trình học các cụm

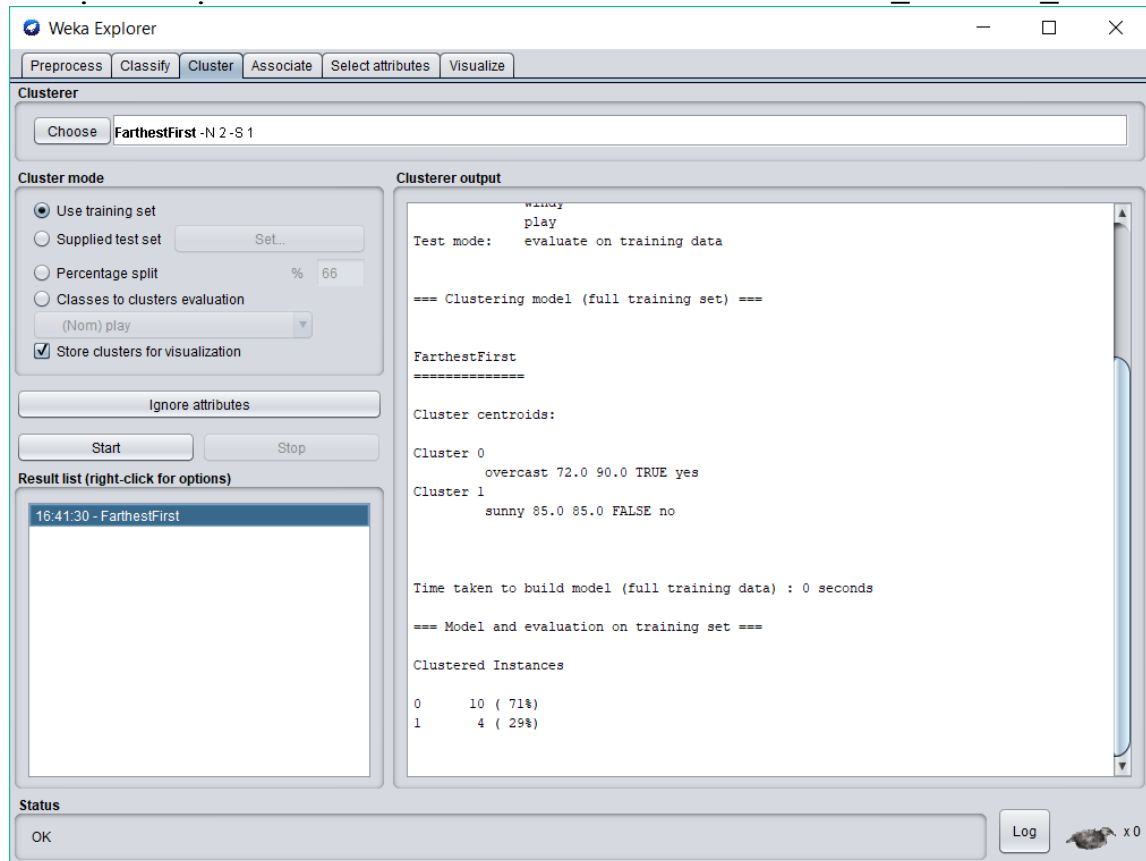


Figure 6. Cluster

- Associate
 - ❖ Lựa chọn một mô hình giải thuật phát hiện luật kết hợp
 - ❖ Associator output: hiển thị các thông tin quan trọng sau:
 - Run information: Các tùy chọn đối với mô hình phát hiện luật kết hợp, tên của tập dữ liệu, số lượng các ví dụ, thuộc tính.
 - Associator model (full training set): Biểu diễn (dạng text) của tập các luật kết hợp phát hiện được.
 - Minimum support
 - Minimum confidence
 - Large/frequent itemsets

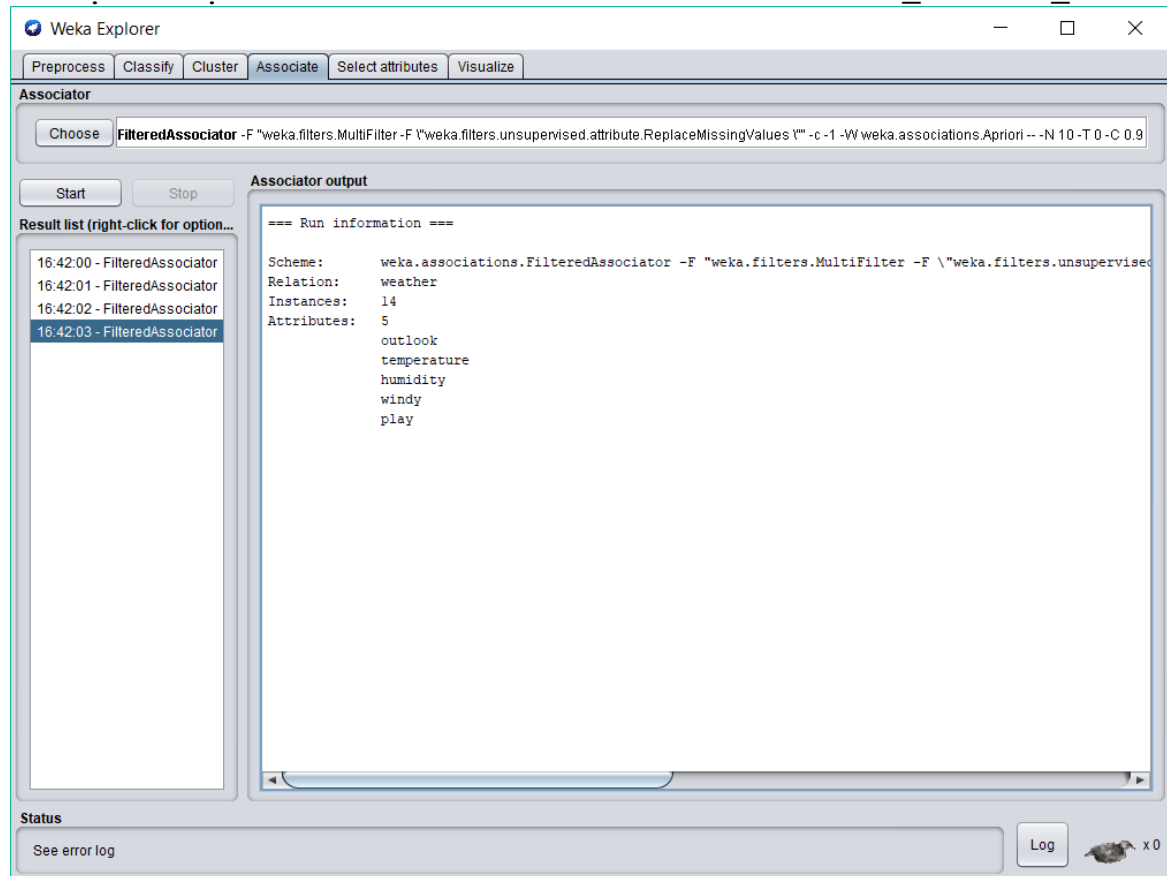


Figure 7. Associate

- Select Attributes
 - ❖ Xác định những thuộc tính quan trọng
 - ❖ Attribute Evaluator: Để xác định một phương pháp đánh giá mức độ phù hợp của thuộc tính
 - Vd: correlation-based, wrapper, information gain, chi-squared
 - ❖ Search method: Xác định phương pháp (thứ tự) xét các thuộc tính.
 - Vd: best-first, random, exhaustive, ranking, ...

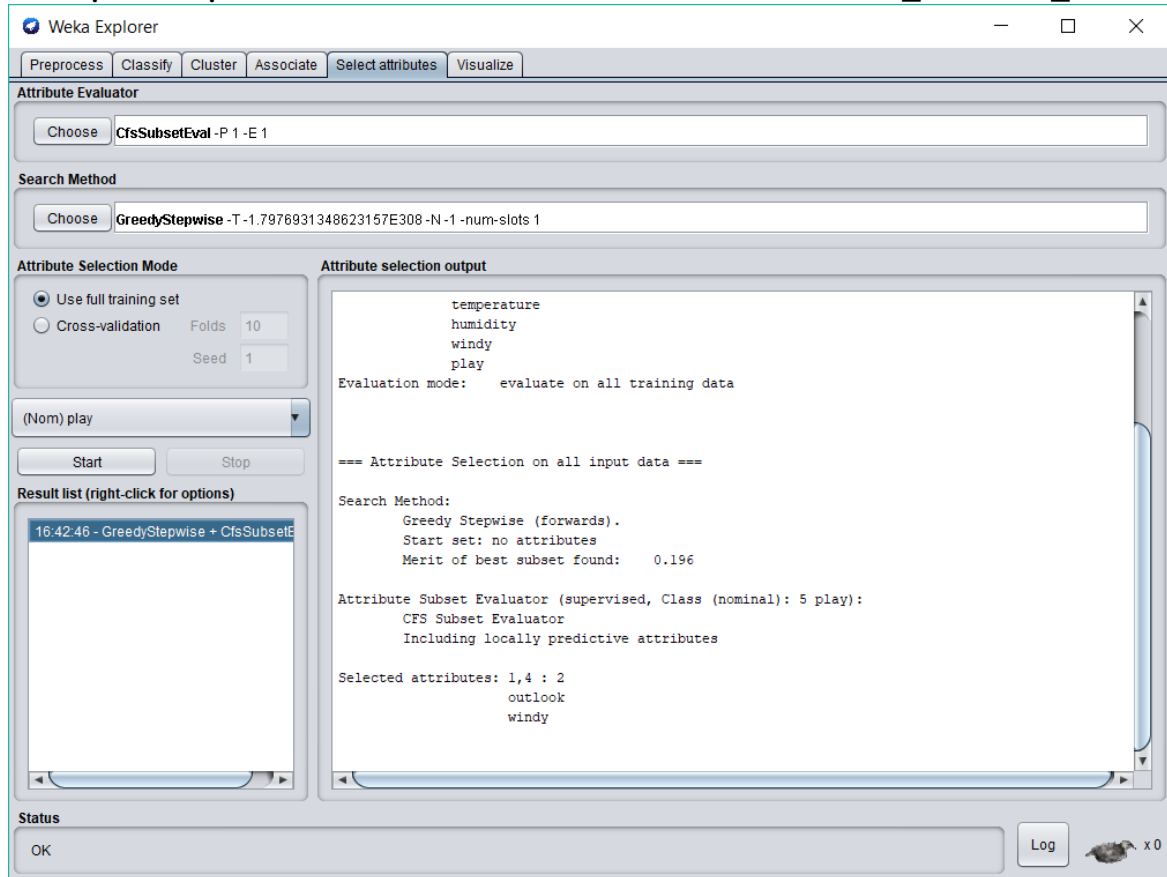


Figure 8. Select attributes

- Visualize: Hiện thị biểu đồ dữ liệu

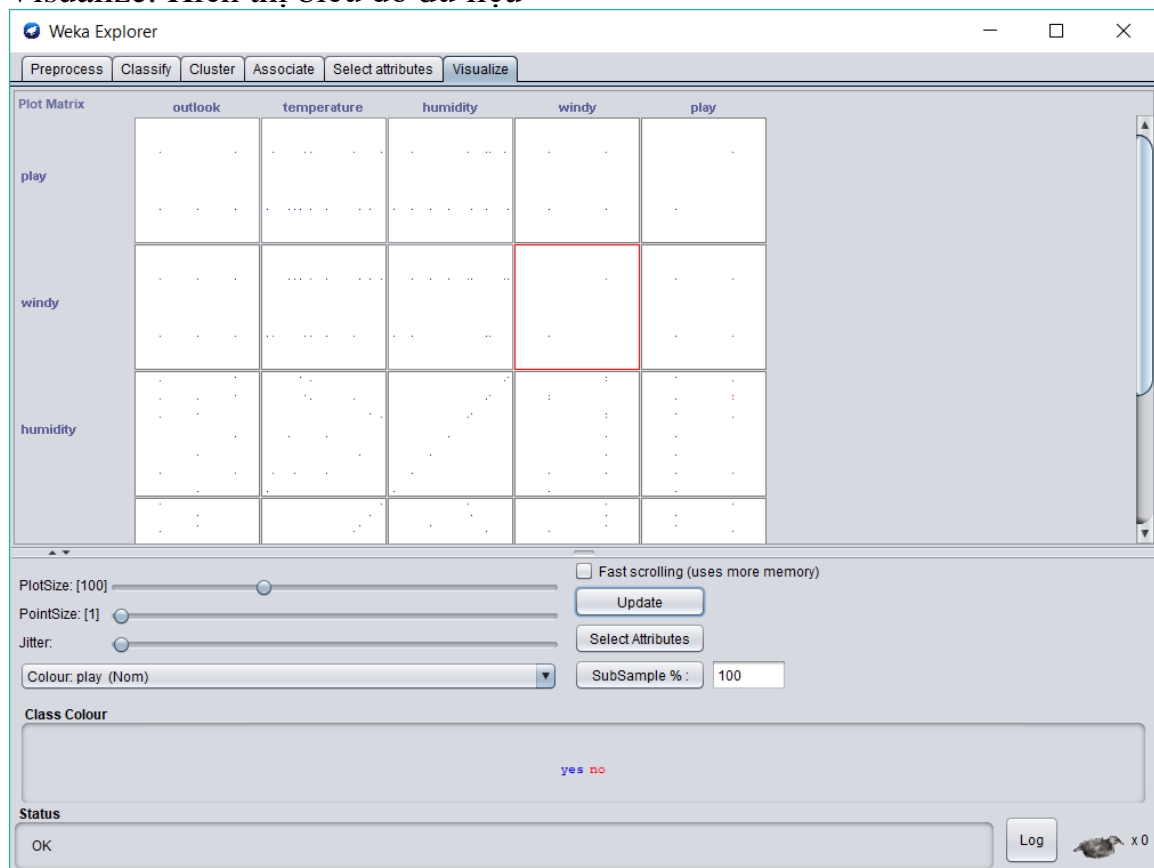


Figure 9. Visualize

1.2 Ví dụ minh họa

Xây dựng mô hình phân lớp bằng cây quyết định trong Weka cho *weather.arff*.

weather.arff có 5 thuộc tính **outlook**, **temperature**, **humidity**, **windy**, **play**. Đây là dữ liệu mô tả về khả năng có đi chơi hay không (phụ thuộc vào thời tiết).

```
weather.arff - Notepad
File Edit Format View Help
@relation weather

@attribute outlook {sunny, overcast, rainy}
@attribute temperature real
@attribute humidity real
@attribute windy {TRUE, FALSE}
@attribute play {yes, no}

@data
sunny,85,85,FALSE,no
sunny,80,90,TRUE,no
overcast,83,86,FALSE,yes
rainy,70,96,FALSE,yes
rainy,68,80,FALSE,yes
rainy,65,70,TRUE,no
overcast,64,65,TRUE,yes
sunny,72,95,FALSE,no
sunny,69,70,FALSE,yes
rainy,75,80,FALSE,yes
sunny,75,70,TRUE,yes
overcast,72,90,TRUE,yes
overcast,81,75,FALSE,yes
rainy,71,91,TRUE,no
```

Figure 10. *weather.arff*

- Bước 1: Tiền xử lý dữ liệu

Preprocess → **Openfile** → chọn file *weather.arff*

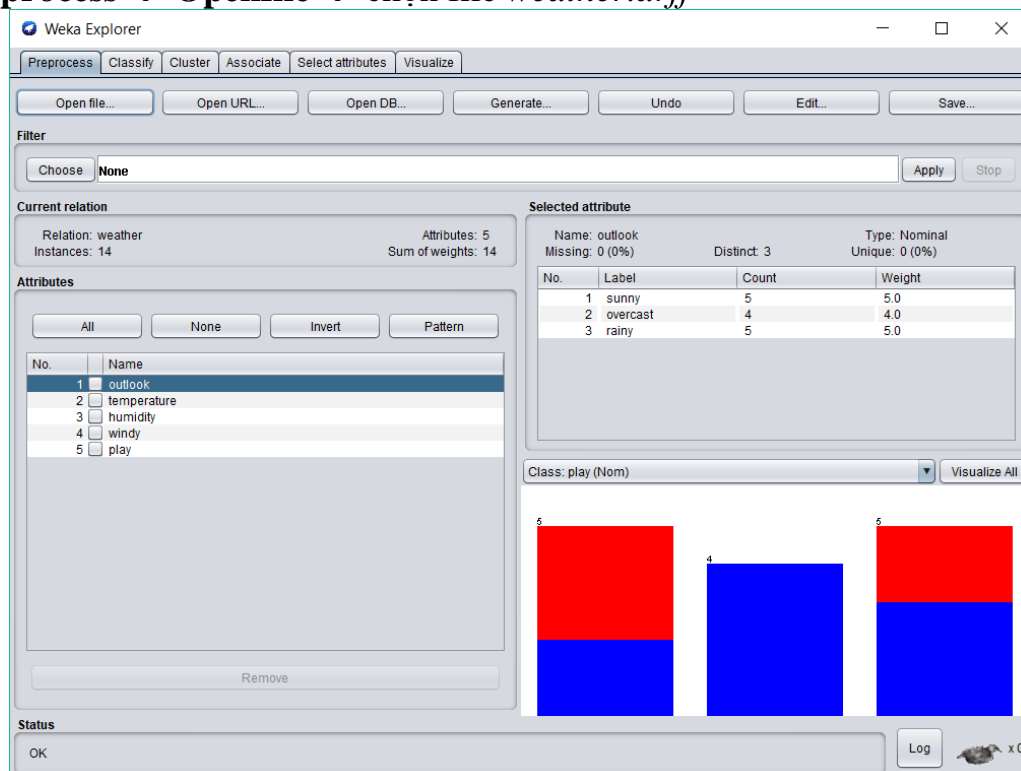


Figure 11. Preprocess *weather.arff*

- Bước 2: Tại tab **Classify**, chọn **Classifier** → **Trees** → **J48**

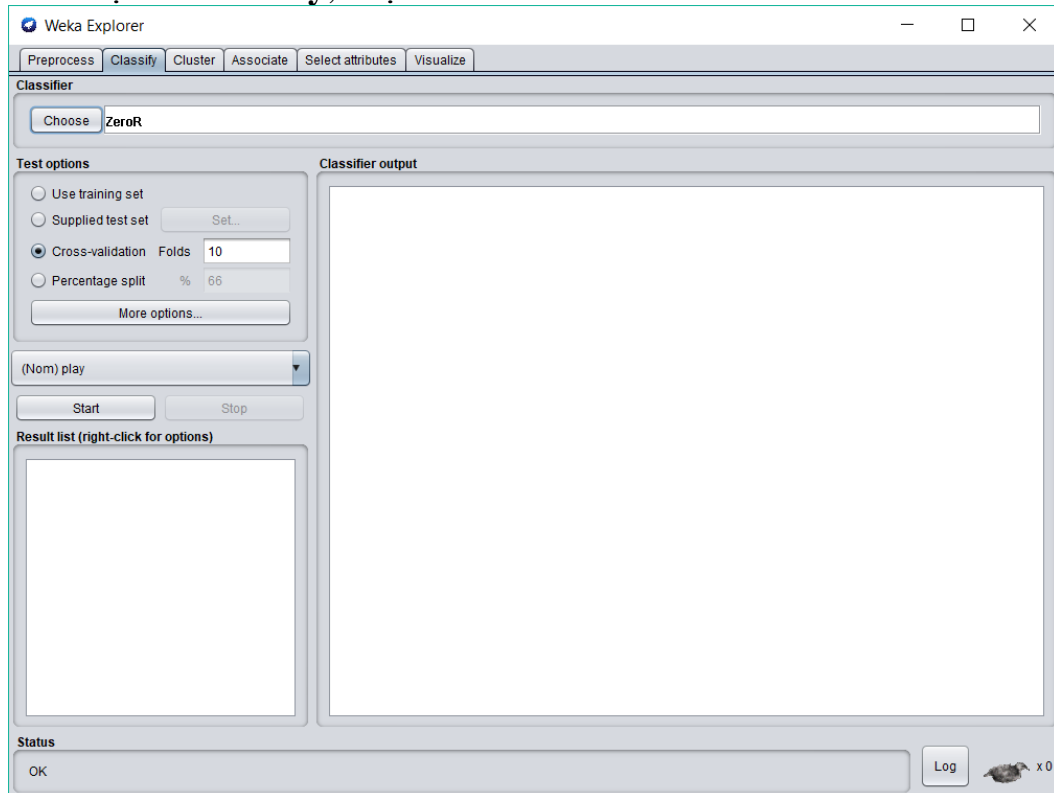


Figure 12. Tab Classify

Lựa chọn **Test options** → **Use training set**

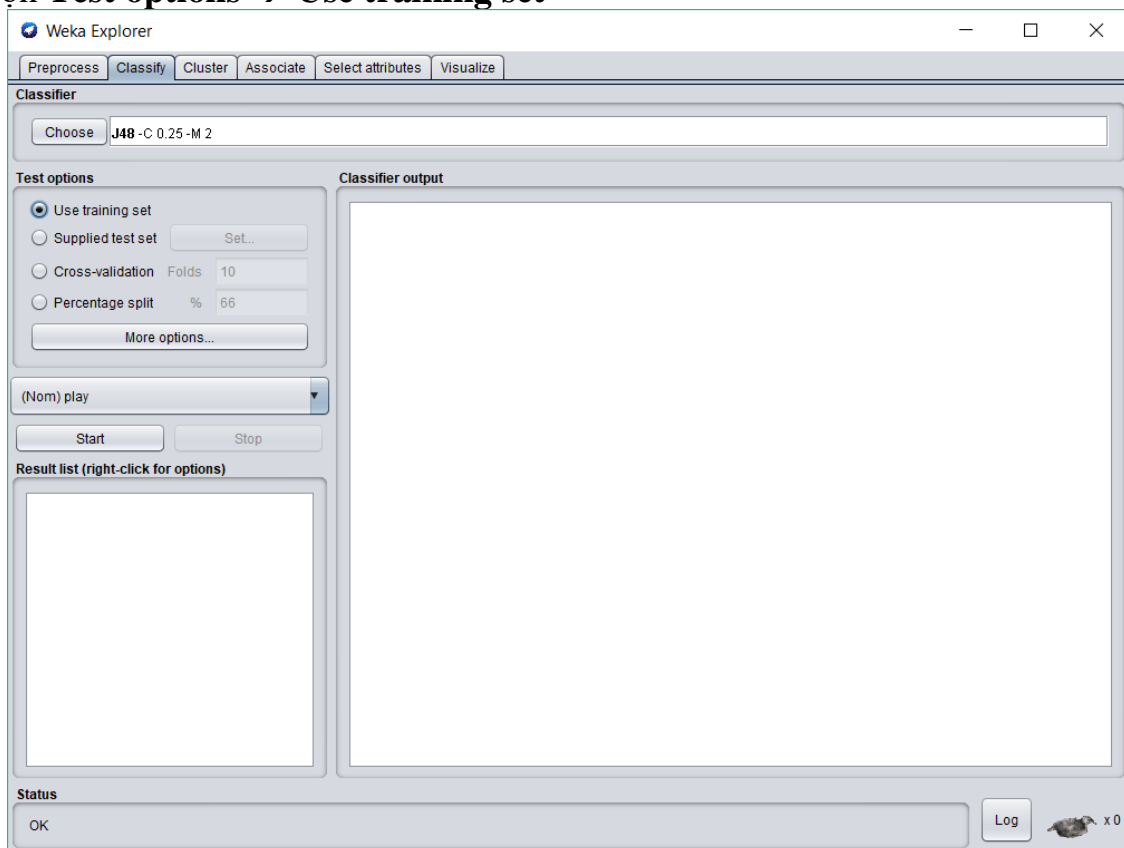


Figure 13. Test options

More options, tick vào các ô như hình

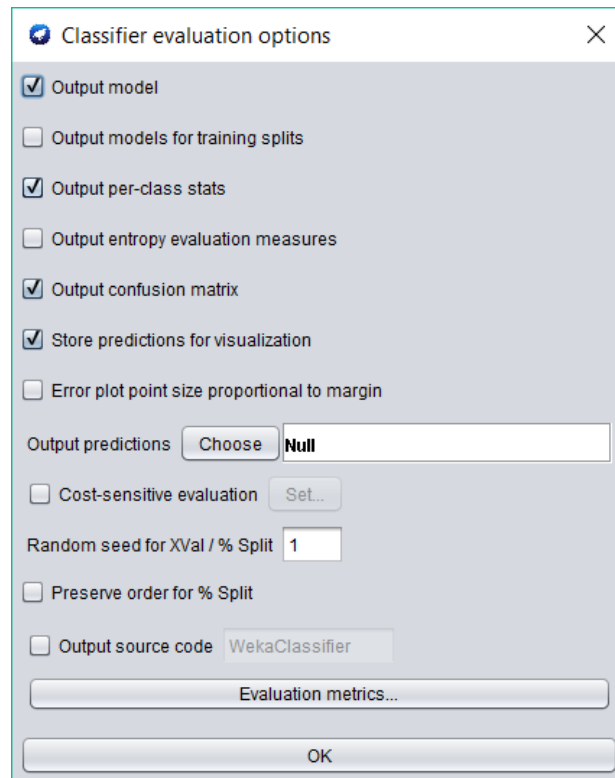


Figure 14. More options

Và cuối cùng, nhấn Start để bắt đầu phân lớp

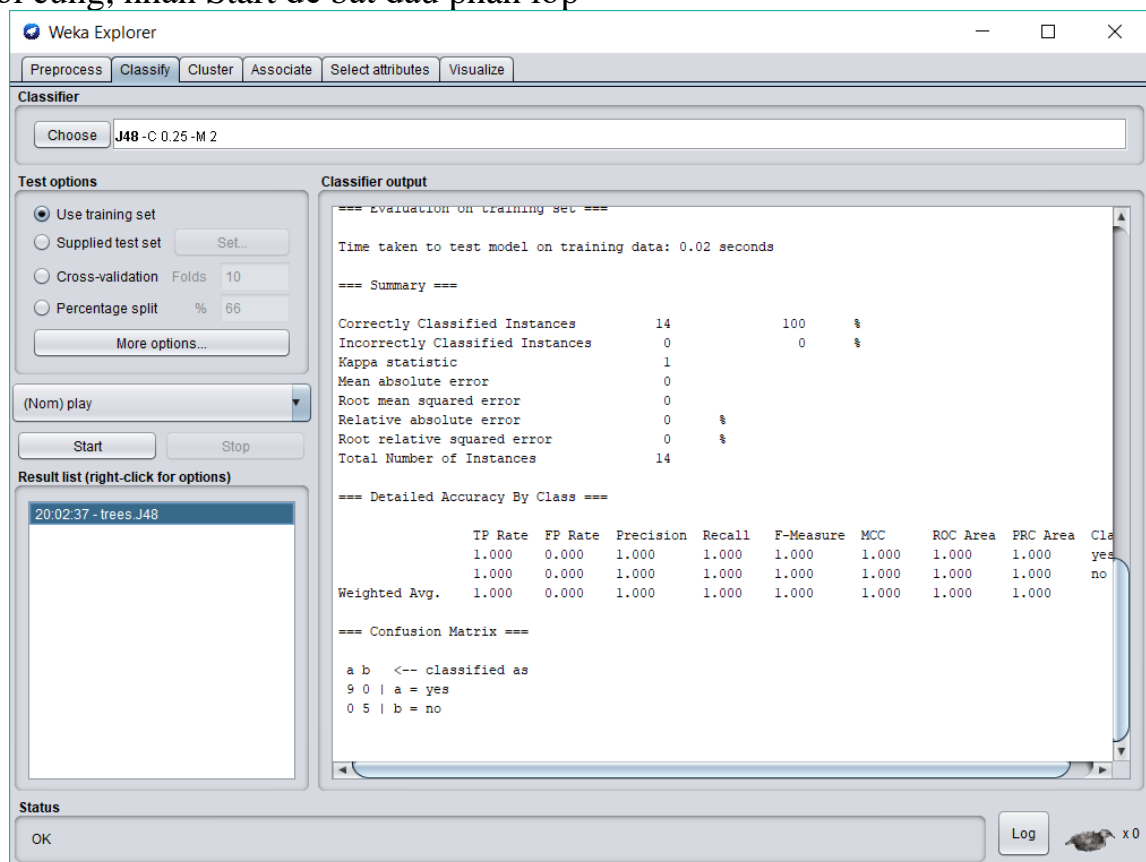


Figure 15. Classify using J48

Cơ sở trí tuệ nhân tạo
Để xem cây quyết định ta làm như sau:

1612348_1612756_Lab03

Right click vào tập tin kết quả, chọn Visualize tree

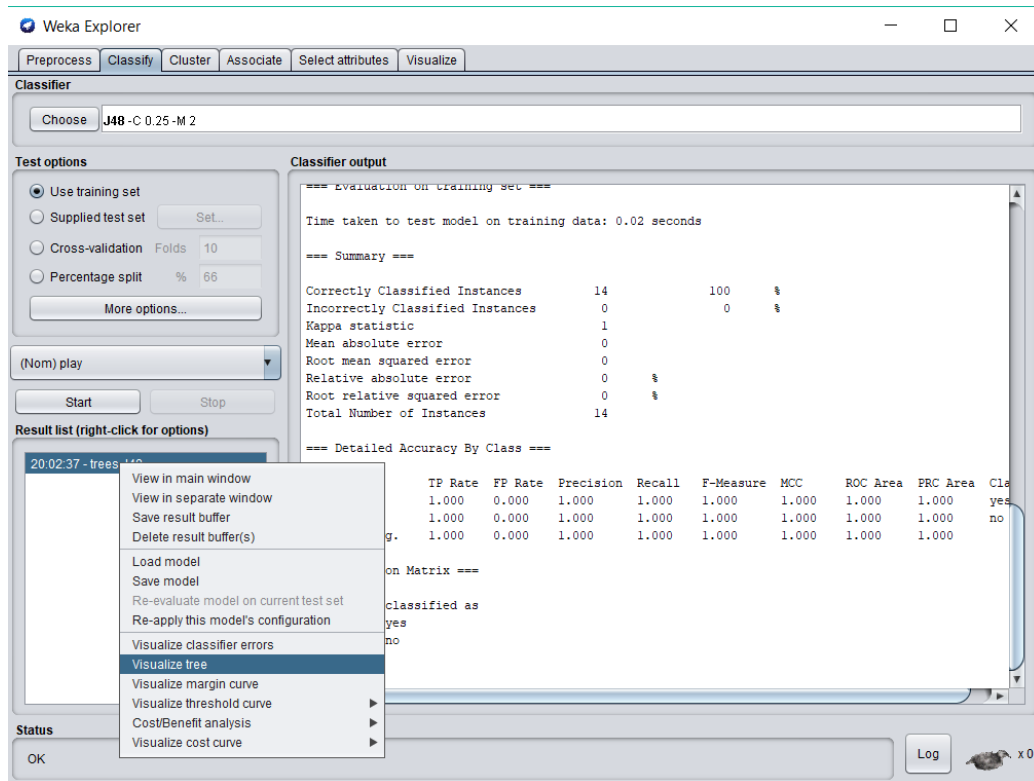


Figure 16. Chọn Visualize tree

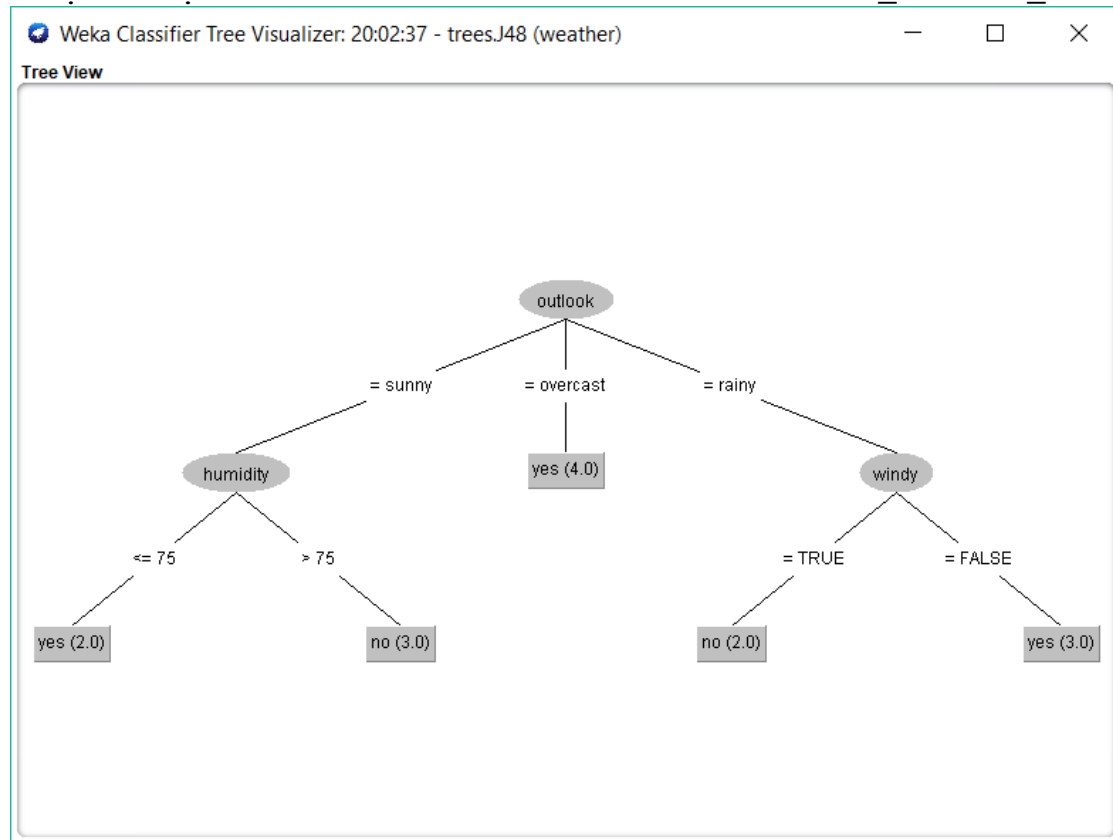
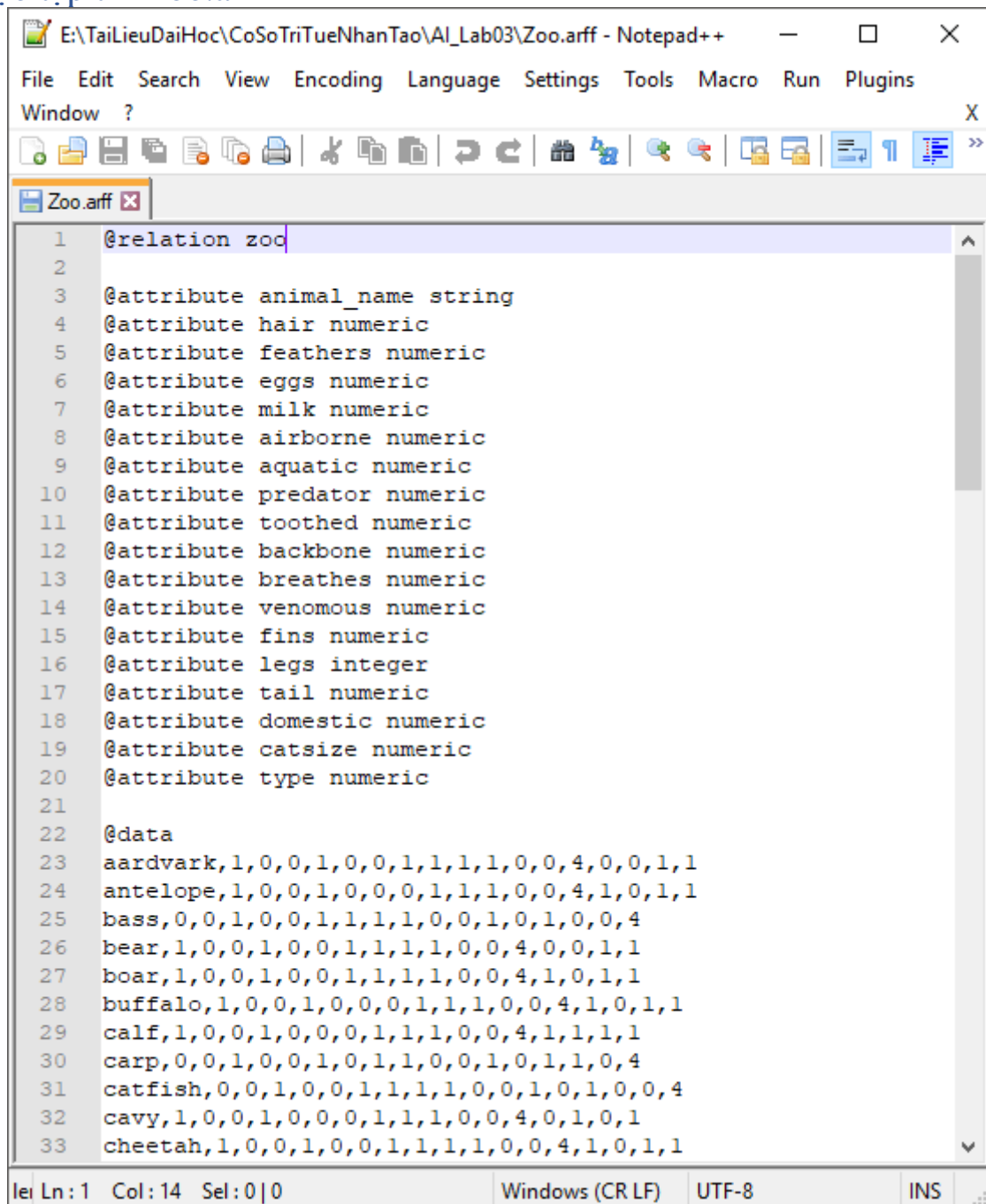


Figure 17. Cây quyết định

2 Sử dụng Weka để chạy thuật toán ID3

2.1 Tạo tập tin Zoo.arff



```

1 @relation zoo
2
3 @attribute animal_name string
4 @attribute hair numeric
5 @attribute feathers numeric
6 @attribute eggs numeric
7 @attribute milk numeric
8 @attribute airborne numeric
9 @attribute aquatic numeric
10 @attribute predator numeric
11 @attribute toothed numeric
12 @attribute backbone numeric
13 @attribute breathes numeric
14 @attribute venomous numeric
15 @attribute fins numeric
16 @attribute legs integer
17 @attribute tail numeric
18 @attribute domestic numeric
19 @attribute catsize numeric
20 @attribute type numeric
21
22 @data
23 aardvark,1,0,0,1,0,0,1,1,1,1,0,0,4,0,0,1,1
24 antelope,1,0,0,1,0,0,0,1,1,1,0,0,4,1,0,1,1
25 bass,0,0,1,0,0,1,1,1,1,0,0,1,0,1,0,0,4
26 bear,1,0,0,1,0,0,1,1,1,1,0,0,4,0,0,1,1
27 boar,1,0,0,1,0,0,1,1,1,1,0,0,4,1,0,1,1
28 buffalo,1,0,0,1,0,0,0,1,1,1,0,0,4,1,0,1,1
29 calf,1,0,0,1,0,0,0,1,1,1,0,0,4,1,1,1,1
30 carp,0,0,1,0,0,1,0,1,1,0,0,1,0,1,1,0,4
31 catfish,0,0,1,0,0,1,1,1,1,0,0,1,0,1,0,0,4
32 cavy,1,0,0,1,0,0,0,1,1,1,0,0,4,0,1,0,1
33 cheetah,1,0,0,1,0,0,1,1,1,1,0,0,4,1,0,1,1

```

Figure 18 Tạo tập tin Zoo.arff

2.2 Mô tả tổng quát về dữ liệu Zoo

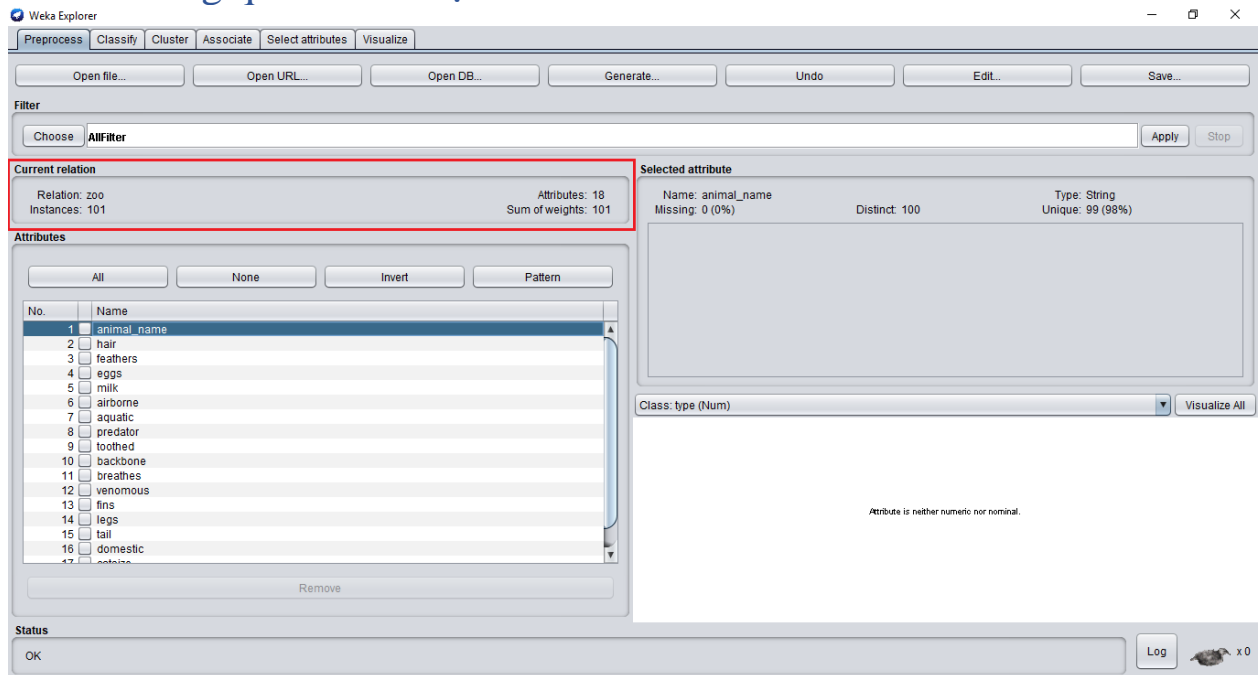


Figure 19 Đưa dữ liệu vào phần mềm Weka

- Số mẫu: 101
- Tên và ý nghĩa các thuộc tính

STT	Tên	Ý nghĩa
1	animal_name	Tên động vật
2	hair	Có lông hay không
3	feathers	Có lông vũ hay không
4	eggs	Có đẻ trứng hay không
5	milk	Có sữa hay không
6	airborne	Có xương sống hay không
7	aquatic	Có sống dưới nước hay không
8	predator	Là động vật ăn thịt hay không
9	toothed	Có răng hay không
10	backbone	Có xương sống hay không
11	breathes	Có thở hay không
12	venomous	Có độc hay không
13	fins	Có vây hay không
14	legs	Số chân
15	tail	Có đuôi hay không
16	domestic	Có ở trong nước khay không
17	catsize	
18	type	Loài

- Đặt tên cho các lớp

Lớp	Tên	Ý nghĩa
1	Mammals	Động vật có vú
2	Birds	Chim
3	Reptile	Bò sát
4	Fish	Cá
5	Amphibians	Lưỡng cư
6	Insects	Côn trùng
7	Invertebrate	Động vật không xương sống

- File Zoo sau khi chỉnh sửa

```

1 @relation Zoo
2
3 @attribute animal_name string
4 @attribute hair {0,1}
5 @attribute feathers {0,1}
6 @attribute eggs {0,1}
7 @attribute milk {0,1}
8 @attribute airborne {0,1}
9 @attribute aquatic {0,1}
10 @attribute predator {0,1}
11 @attribute toothed {0,1}
12 @attribute backbone {0,1}
13 @attribute breathes {0,1}
14 @attribute venomous {0,1}
15 @attribute fins {0,1}
16 @attribute legs {0,2,4,5,6,8}
17 @attribute tail {0,1}
18 @attribute domestic {0,1}
19 @attribute catsize {0,1}
20 @attribute type {Mammals, Birds, Reptile, Fish, Amphibians, Insects, Invertebrate}
21
22 @data
23 aardvark,1,0,0,1,0,0,1,1,1,1,0,0,4,0,0,1,Mammals
24 antelope,1,0,0,1,0,0,0,1,1,1,0,0,4,1,0,1,Mammals
25 bass,0,0,1,0,0,1,1,1,1,0,0,1,0,1,0,0,Fish
26 bear,1,0,0,1,0,0,1,1,1,1,0,0,4,0,0,1,Mammals
27 boar,1,0,0,1,0,0,1,1,1,1,0,0,4,1,0,1,Mammals
28 buffalo,1,0,0,1,0,0,0,1,1,1,0,0,4,1,0,1,Mammals
29 calf,1,0,0,1,0,0,0,1,1,1,0,0,4,1,1,1,Mammals
30 carp,0,0,1,0,0,1,0,1,1,0,0,1,0,1,1,0,Fish
31 catfish,0,0,1,0,0,1,1,1,1,0,0,1,0,1,0,0,Fish
32 cavy,1,0,0,1,0,0,0,1,1,1,0,0,4,0,1,0,Mammals
33 cheetah,1,0,0,1,0,0,1,1,1,1,0,0,4,1,0,1,Mammals
34 chicken,0,1,1,0,1,0,0,0,1,1,0,0,2,1,1,0,Birds
  
```

Figure 20 File Zoo sau khi chỉnh sửa

2.3 Sử dụng thuật toán ID3

Cây sinh ra bởi thuật toán ID3

```

legs = 0
| fins = 0
| | toothed = 0: Invertebrate
| | toothed = 1: Reptile
| fins = 1
| | eggs = 0: Mammals
| | eggs = 1: Fish
legs = 2
| hair = 0: Birds
| hair = 1: Mammals
legs = 4
| hair = 0
| | aquatic = 0: Reptile
| | aquatic = 1
| | | toothed = 0: Invertebrate
| | | toothed = 1: Amphibians
| hair = 1: Mammals
legs = 5: Invertebrate
legs = 6
| aquatic = 0: Insects
| aquatic = 1: Invertebrate
legs = 8: Invertebrate

```

Figure 21 Cây sinh ra bởi thuật toán ID3

2.4 Kết quả dự đoán của 5 mẫu

inst#	actual	predicted	error	prediction
1	1:?	1:Mammals	1	
2	1:?	2:Birds	1	
3	1:?	3:Reptile	1	
4	1:?	4:Fish	1	
5	1:?	5:Amphibians	1	

Figure 22 Kết quả dự đoán của 5 mẫu

STT	Mẫu	Dự đoán
1	NameIsSecret,1,0,0,1,0,0,0,1,1,1,0,0,4,1,0,1,?	Mammals
2	NameIsSecret,0,1,1,0,1,0,0,0,1,1,0,0,2,1,1,0,?	Birds
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,?	Amphibians

3 Chạy các thuật toán khác

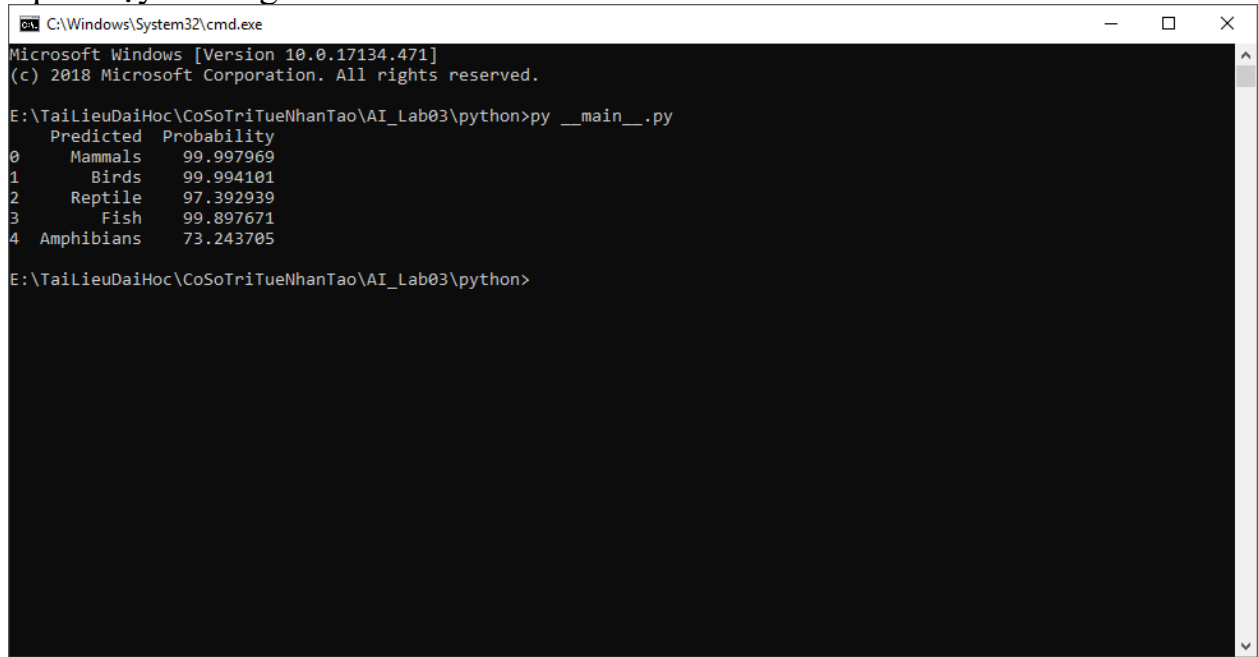
3.1 Chương trình Python cho giải thuật Naïve Bayes

a) Yêu cầu

- Chương trình được viết bằng ngôn ngữ lập trình Python (phiên bản 3)
- Các thư viện kèm theo để chạy chương trình gồm pandas, numpy, liac-arff

b) Kết quả

Kết quả chạy chương trình với 5 mẫu cho trước



```
C:\Windows\System32\cmd.exe
Microsoft Windows [Version 10.0.17134.471]
(c) 2018 Microsoft Corporation. All rights reserved.

E:\TaiLieuDaiHoc\CoSoTriTueNhanTao\AI_Lab03\python>py __main__.py
Predicted Probability
0 Mammals 99.997969
1 Birds 99.994101
2 Reptile 97.392939
3 Fish 99.897671
4 Amphibians 73.243705

E:\TaiLieuDaiHoc\CoSoTriTueNhanTao\AI_Lab03\python>
```

Figure 23 Kết quả chạy chương trình

3.2 Chạy với thuật toán J48

Mở file zoo.arff, chọn tab Classify

Classifier → Choose → Trees → J48

Test options → Use training set

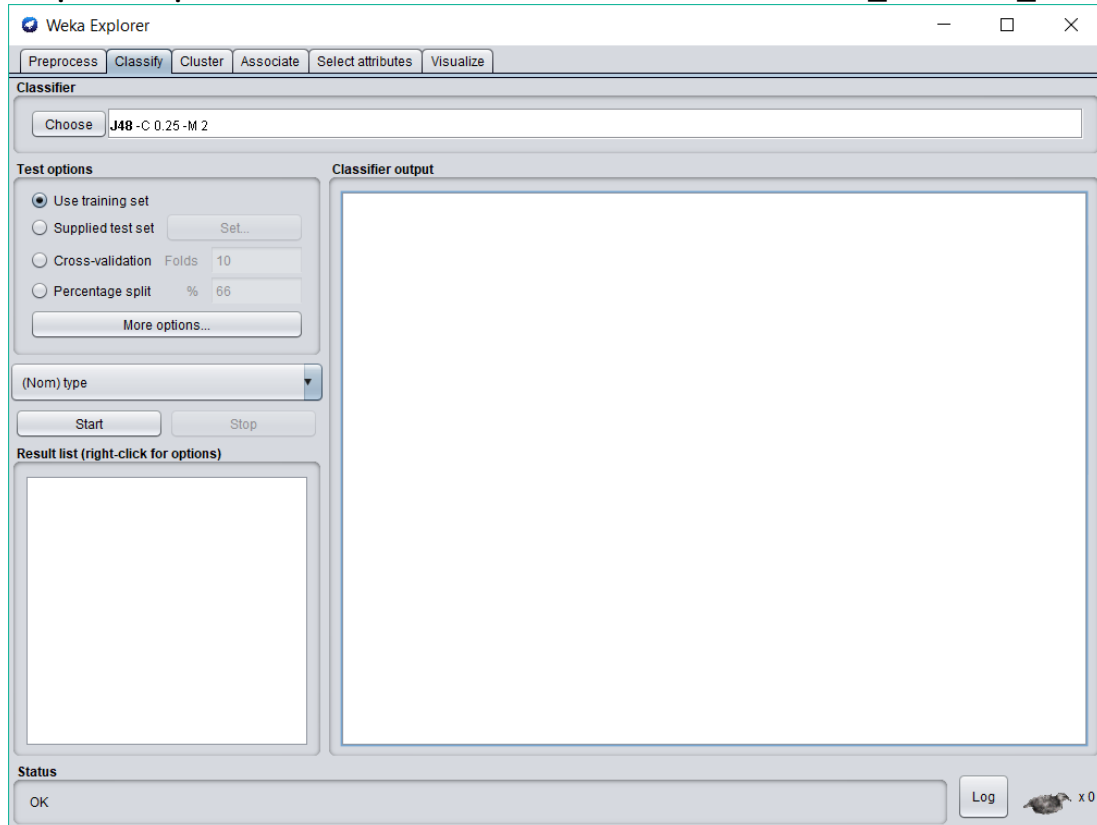


Figure 24. Chọn thuật toán J48

Chọn Start, và kết quả hiện ra

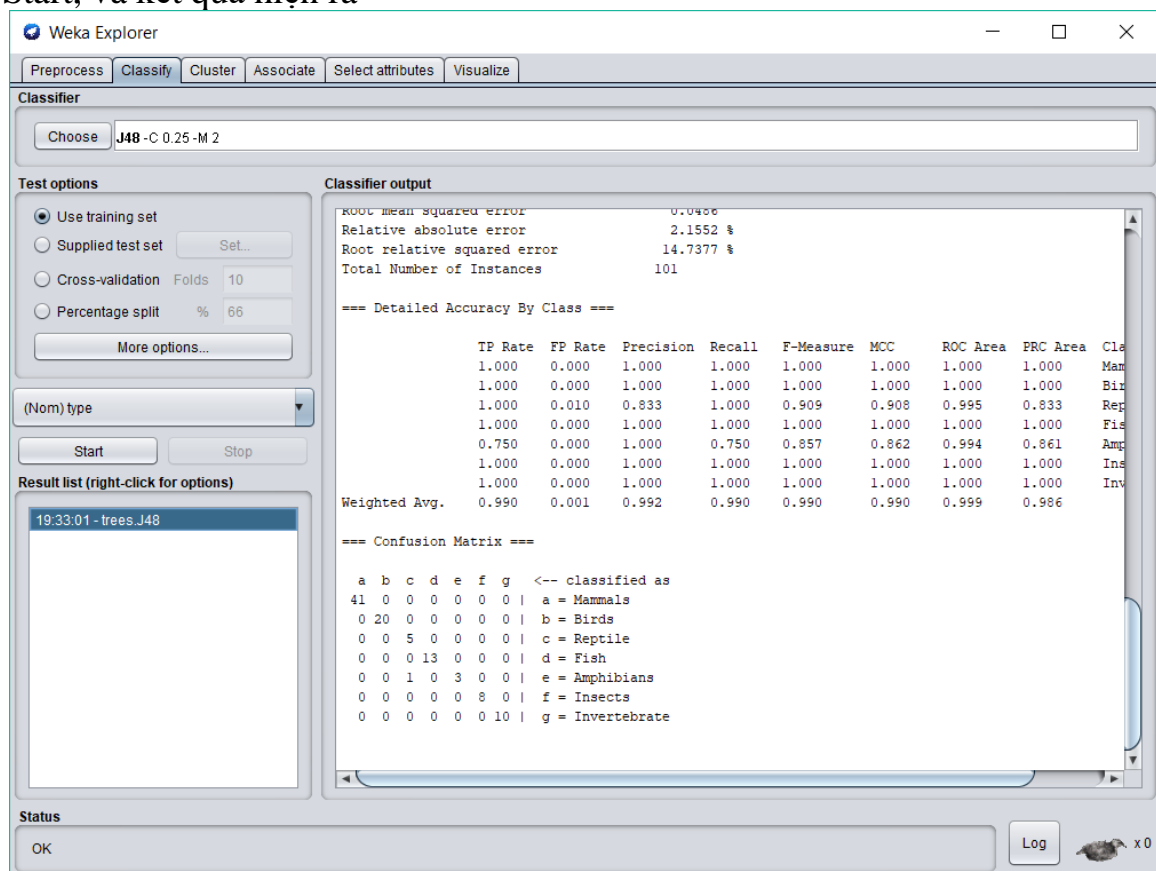


Figure 25. Kết quả phân lớp

Bây giờ, chúng ta đi test: Click chọn Supplied test set

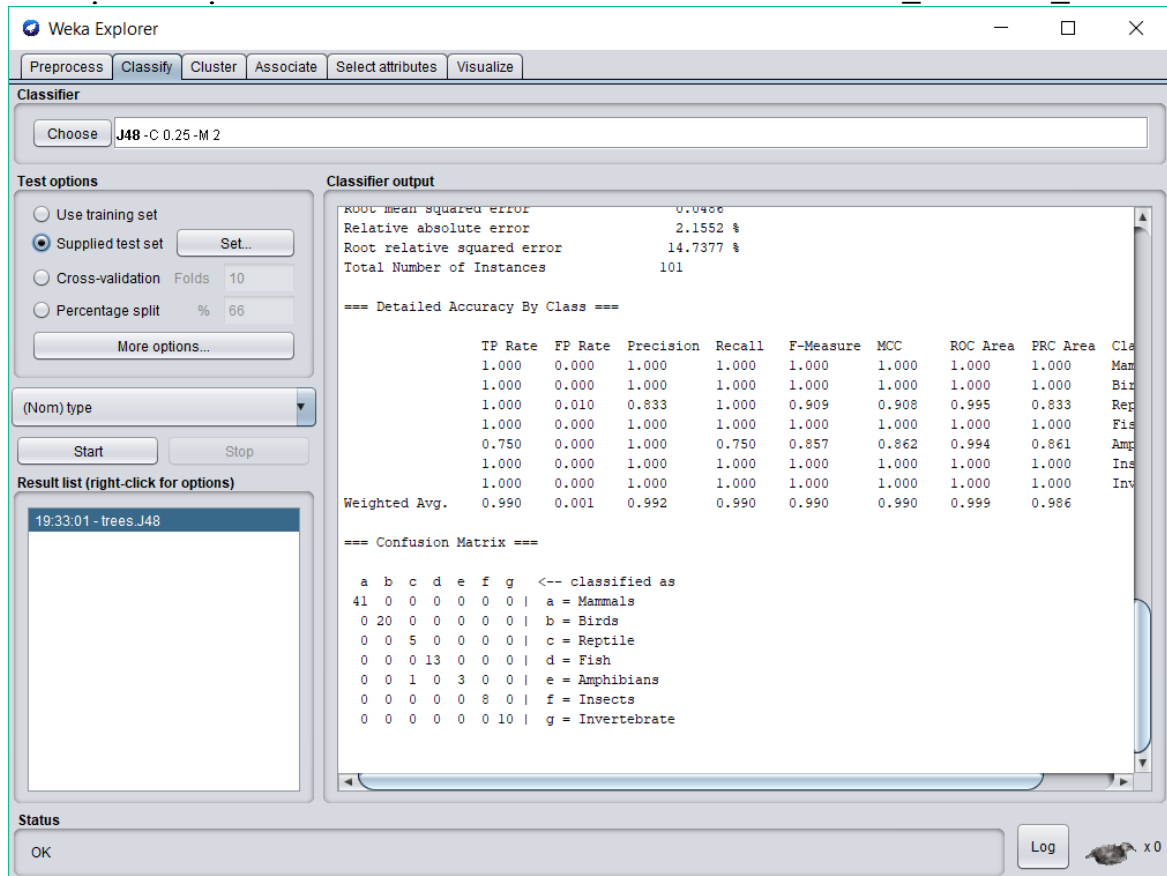


Figure 26. Supplied test set

Bấm chọn Set, 1 dialog hiện ra

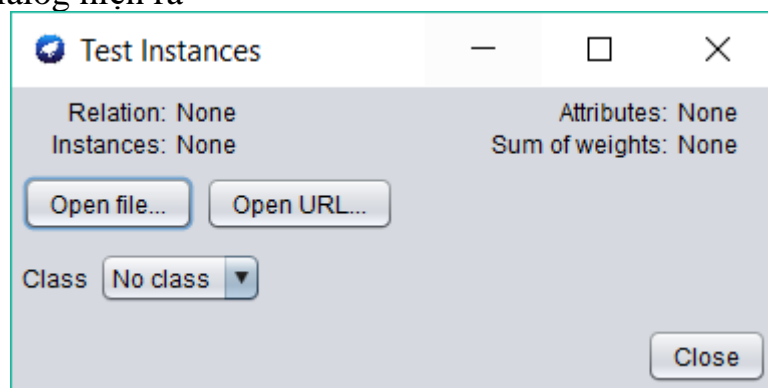


Figure 27. Test set

Chúng ta bấm vào Open file rồi chọn file test đã chuẩn bị sẵn, bấm close
Tiếp tới chúng ta chọn More options, 1 dialog hiện ra

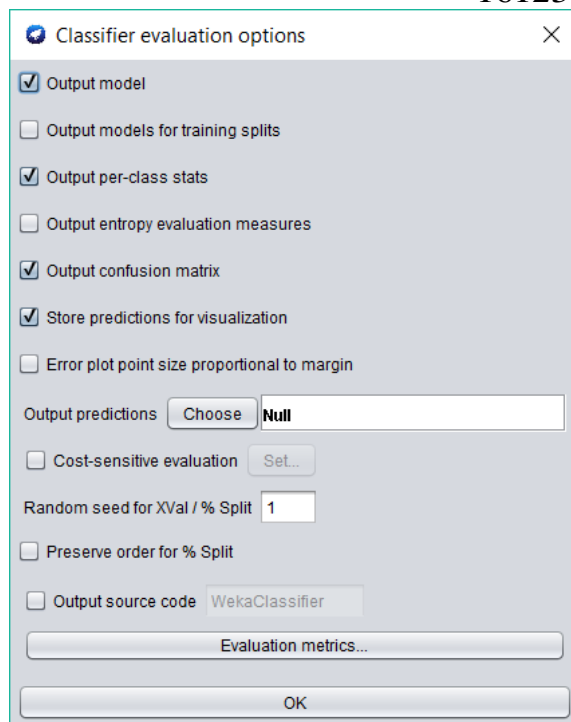


Figure 28. More options

Cơ sở trí tuệ nhân tạo
Chúng ta chọn Choose và chọn PlainText → OK

1612348_1612756_Lab03

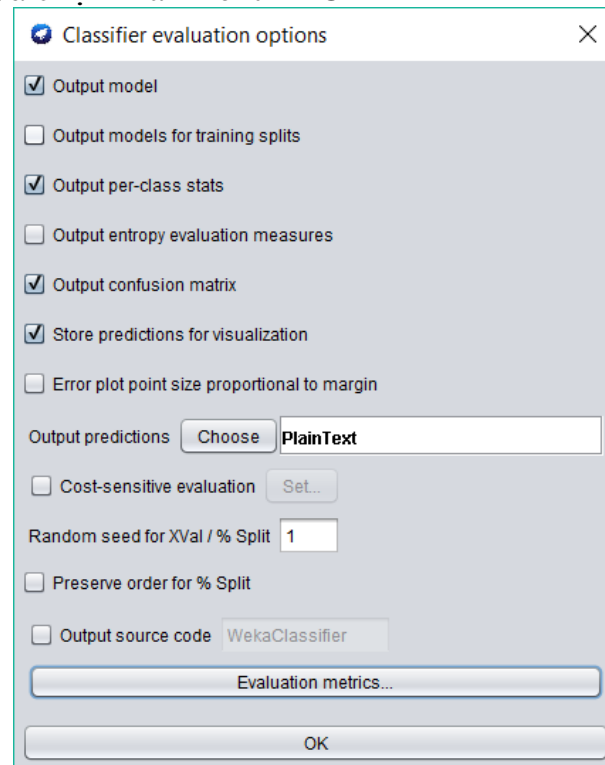


Figure 29. Chọn PlainText

Bây giờ bấm Start để bắt đầu, kết quả hiện ra

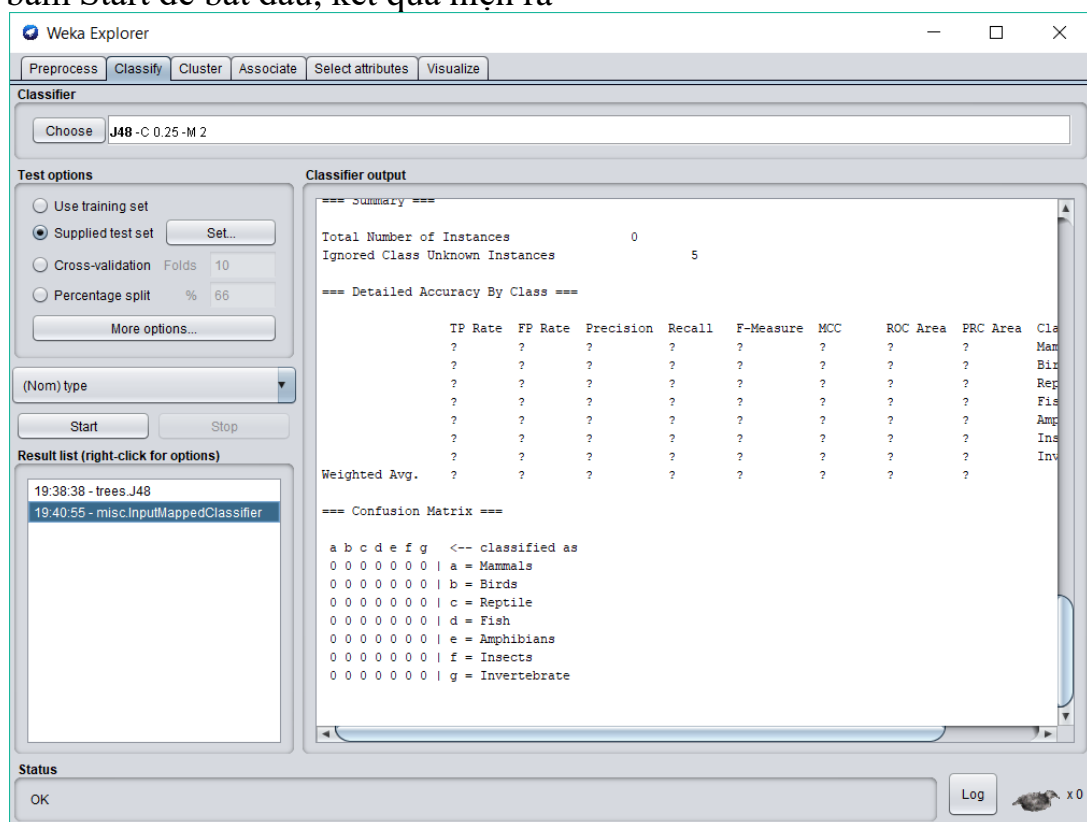


Figure 30. Kết quả test với thuật toán J48

=== Predictions on test set ===

inst#	actual	predicted	error	prediction
1	1:?	1:Mammals		1
2	1:?	2:Birds		1
3	1:?	3:Reptile	0.833	
4	1:?	4:Fish		1
5	1:?	3:Reptile	0.833	

Figure 31. Kết quả dự đoán 5 mẫu

STT	Mẫu	Dự đoán	Xác suất
1	NameIsSecret,1,0,0,1,0,0,0,1,1,1,0,0,4,1,0,1,?	Mammals	100%
2	NameIsSecret,0,1,1,0,1,0,0,0,1,1,0,0,2,1,1,0,?	Birds	100%
3	NameIsSecret,0,0,1,0,0,0,1,1,1,1,1,0,0,1,0,0,?	Reptile	83.3%
4	NameIsSecret,0,0,1,0,0,1,1,1,1,0,0,1,0,1,0,0,?	Fish	100%
5	NameIsSecret,0,0,1,0,0,1,1,1,1,1,0,0,4,1,0,0,?	Amphibians	83.3%

3.3 Chạy với thuật toán Naïve Bayes

Áp dụng tương tự các bước ở phần trên, ta thu được kết quả sau:

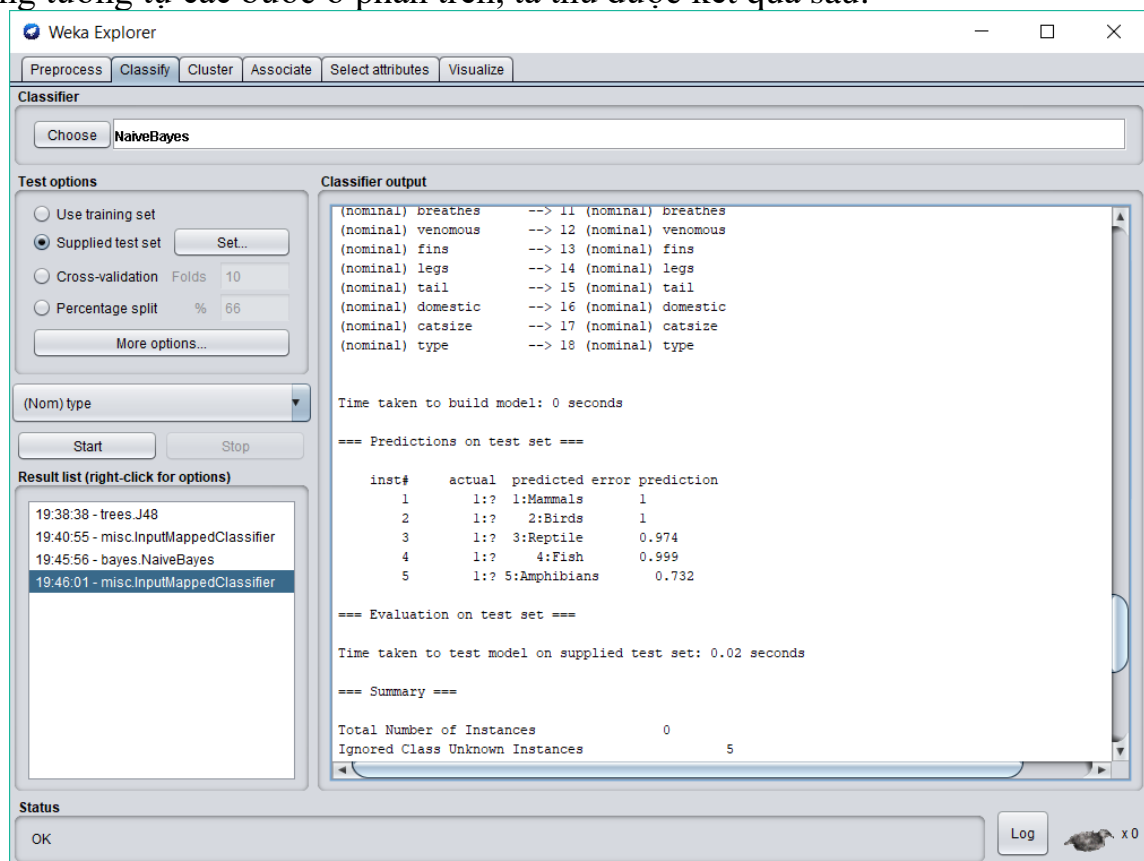


Figure 32. Kết quả chạy với thuật toán Naïve Bayes

=== Predictions on test set ===

inst#	actual	predicted	error	prediction
1	1:?	1:Mammals		1
2	1:?	2:Birds		1
3	1:?	3:Reptile	0.974	
4	1:?	4:Fish	0.999	
5	1:?	5:Amphibians	0.732	

Figure 33. Kết quả dự đoán 5 mẫu

STT	Mẫu	Dự đoán	Xác suất
1	NameIsSecret,1,0,0,1,0,0,0,1,1,1,0,0,4,1,0,1,?	Mammals	100%
2	NameIsSecret,0,1,1,0,1,0,0,0,1,1,0,0,2,1,1,0,?	Birds	100%
3	NameIsSecret,0,0,1,0,0,0,1,1,1,1,1,0,0,1,0,0,?	Reptile	97.4%
4	NameIsSecret,0,0,1,0,0,1,1,1,1,0,0,1,0,1,0,0,?	Fish	99.9%
5	NameIsSecret,0,0,1,0,0,1,1,1,1,1,0,0,4,1,0,0,?	Amphibians	73.2%

3.4 Chạy với thuật toán IBK

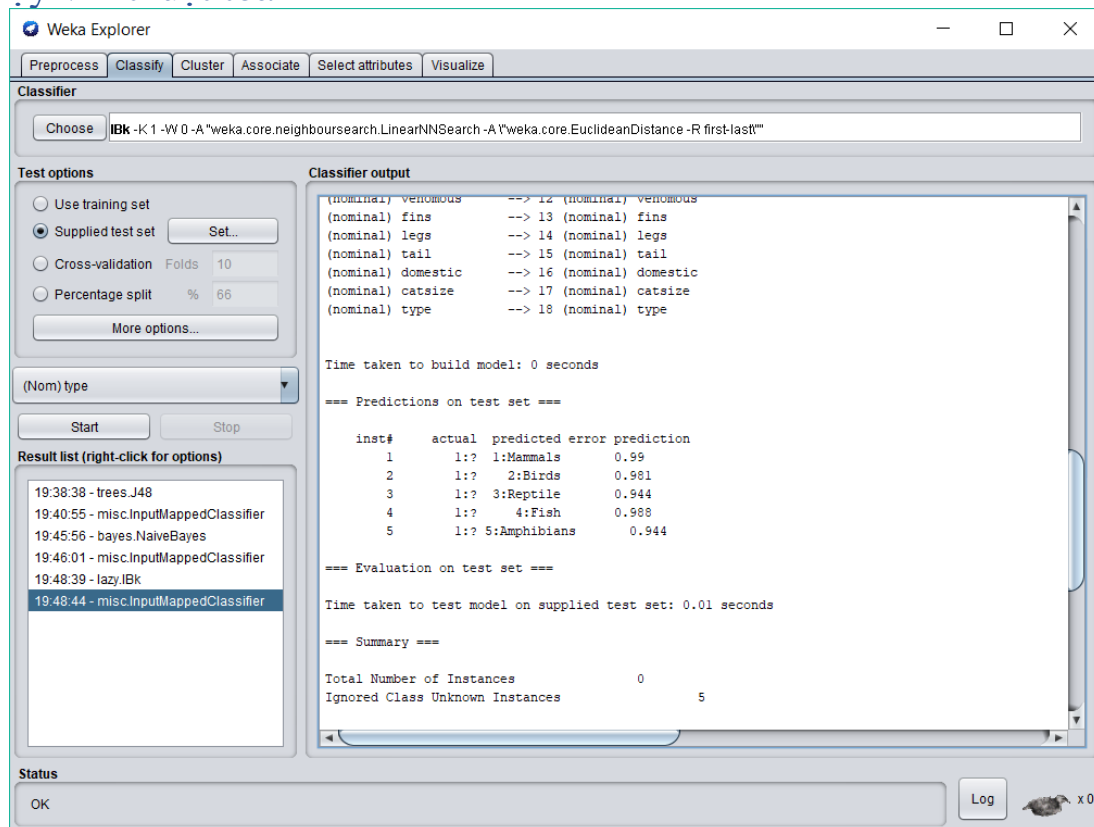


Figure 34. Kết quả chạy với thuật toán IBK

```
=== Predictions on test set ===
```

```

inst#    actual  predicted error prediction
  1      1:?    1:Mammals  0.99
  2      1:?    2:Birds   0.981
  3      1:?    3:Reptile  0.944
  4      1:?    4:Fish    0.988
  5      1:?    5:Amphibians 0.944

```

Figure 35. Kết quả dự đoán 5 mẫu

STT	Mẫu	Dự đoán	Xác suất
1	NameIsSecret,1,0,0,1,0,0,0,1,1,1,0,0,4,1,0,1,?	Mammals	99%
2	NameIsSecret,0,1,1,0,1,0,0,0,1,1,0,0,2,1,1,0,?	Birds	98.1%
3	NameIsSecret,0,0,1,0,0,0,1,1,1,1,1,0,0,1,0,0,?	Reptile	94.4%
4	NameIsSecret,0,0,1,0,0,1,1,1,1,0,0,1,0,1,0,0,?	Fish	98.8%
5	NameIsSecret,0,0,1,0,0,1,1,1,1,1,0,0,4,1,0,0,?	Amphibians	94.4%