Đại học Quốc gia Thành phố Hồ Chí Minh
Trường Đại học Công nghệ Thông tin (UIT)

BÁO CÁO BÀI TẬP

Môn học: Hệ thống tìm kiếm, phát hiện và ngăn ngừa xâm nhập

Kỳ báo cáo: Buổi 05

Tên chủ đề: Học máy trong IDS

GV: Đỗ Hoàng Hiển

Ngày báo cáo: 29/05/2023

Nhóm: 07

1. THÔNG TIN CHUNG:

Lớp: NT204.N21.ANTT.2

STT	Họ và tên	MSSV	Email
1	Phạm Phúc Đức	20520162	20520162@gm.uit.edu.vn
2	Lê Trần Thùy Trang	20520323	20520323@gm.uit.edu.vn
3	Nguyễn Đức Tấn	20520751	20520751@gm.uit.edu.vn

2. NỘI DUNG THỰC HIỆN:[1]

STT	Công việc	Kết quả tự đánh giá	Người đóng góp
1	Yêu cầu 1.1	100%	Phạm Phúc Đức
2	Yêu cầu 1.2	100%	Lê Trần Thuỳ Trang
3	Yêu cầu 1.3	100%	Nguyễn Đức Tấn

Phần bên dưới của báo cáo này là tài liệu báo cáo chi tiết của nhóm thực hiện.

[1] Ghi nội dung công việc, các kịch bản trong bài Thực hành

BÁO CÁO CHI TIẾT

1. Yêu cầu 1.1 Tìm hiểu về tập dữ liệu KDD Cup 1999.

TÌM HIỂU VỀ TẬP DỮ LIỆU KDD CUP 1999 Dữ liệu trong bộ dữ liệu KDD Cup 1999 là lưu lượng mạng đã được thu thập, phân tích, xử lý để lấy các thuộc tính và từ đó gán nhãn tương ứng với loại tấn công hoặc dữ liệu bình thường. Sinh viên tìm hiểu các phần sàu:

1. Số nhóm tấn công: **4** .Kể tên các nhóm tấn công: **dos, u2r, r21, probe.**

- DoS (Denial of Service): Đây là một loại tấn công nhằm làm cho dịch vụ hoặc hệ thống trở nên không khả dụng cho người dùng hợp lệ.

- U2R (User to Root): U2R là một loại tấn công mục tiêu vào quyền truy cập của người dùng bình thường để nâng cấp quyền truy cập thành quyền root hoặc quyền quản trị hệ thống.

- R2L (Remote to Local): R2L là một loại tấn công mà kẻ tấn công cố gắng xâm nhập vào một hệ thống từ xa để thu được quyền truy cập cục bộ.

- Probe: Probe là một loại tấn công nhằm khảo sát hoặc thăm dò hệ thống mạng hoặc các thiết bị để thu thập thông tin

2. Số kiểu tấn công: **23** .Kể tên các kiểu tấn công được gán nhãn:

- back: Tấn công back là một kiểu tấn công nhằm lợi dụng các lỗ hổng bảo mật để xâm nhập vào một hệ thống từ xa và kiểm soát nó
- buffer_overflow: Tấn công buffer overflow là một kỹ thuật khai thác lỗ hổng trong việc xử lý đệm của một ứng dụng, trong đó kẻ tấn công ghi đè lên vùng nhớ đệm và gây ra lỗi thực thi mã độc hại hoặc kiểm soát luồng chương trình.
- ftp_write: Tấn công ftp_write là việc sử dụng lỗ hổng trong dịch vụ FTP để ghi hoặc tải lên các tệp tin trái phép lên hệ thống mục tiêu.
- guess_passwd: Tấn công guess_passwd là quá trình thử đoán mật khẩu của người dùng bằng cách liên tục thử các mật khẩu khác nhau để xâm nhập vào hệ thống.

- imap: Tấn công imap là việc tìm lỗ hổng trong dịch vụ Internet Message Access Protocol (IMAP) để tấn công hệ thống hoặc lấy trộm thông tin.
- ipsweep: Tấn công ipsweep là việc quét liên tục một loạt các địa chỉ IP để thu thập thông tin về các máy chủ và thiết bị trên mạng.
- probe: Probe là việc khảo sát hoặc thăm dò hệ thống mạng hoặc các thiết bị để thu thập thông tin về các lỗ hổng bảo mật hoặc cấu trúc hệ thống.
- land: Tấn công land là việc tạo ra và gửi các gói tin mạng có địa chỉ nguồn và địa chỉ đích giống nhau, dẫn đến sự đụng độ và làm cho hệ thống bị quá tải hoặc treo.
- loadmodule: Tấn công loadmodule là việc sử dụng lỗ hổng trong một ứng dụng hoặc hệ thống để tải và thực thi các module hay mã độc hại.
- multihop: Tấn công multihop là một loại tấn công mà kẻ tấn công sử dụng nhiều nút trung gian (hop) để che giấu hoạt động xâm nhập của mình.
- neptune: Neptune là một loại tấn công DDoS (Distributed Denial of Service) mà một lượng lớn yêu cầu tới hệ thống được gửi từ nhiều nguồn khác nhau nhằm làm quá tải hệ thống và gây ra sự cố không khả dụng.
- nmap: Tấn công nmap là việc sử dụng công cụ nmap để quét mạng và thu thập thông tin về các cổng mạng, máy chủ và dịch vụ đang chạy trên hệ thống mục tiêu.
- perl: Tấn công perl là việc sử dụng mã Perl độc hại để thực thi các hoạt động xâm nhập hoặc gây hại trên hệ thống mục tiêu.
- portsweep: Tấn công portsweep là quá trình quét một loạt các cổng mạng trên một hệ thống hoặc mạng để tìm kiếm các cổng mạng mở và thu thập thông tin về hệ thống.
- rootkit: Rootkit là một loại phần mềm độc hại được cài đặt trên hệ thống mà nó che giấu các hoạt động xâm nhập và cung cấp quyền kiểm soát bất hợp pháp cho kẻ tấn công.
- satan: Tấn công satan là việc sử dụng công cụ kiểm tra bảo mật Satan để xác định các lỗ hổng bảo mật trong hệ thống mục tiêu.

- smurf: Tấn công smurf là một loại tấn công DDoS mà kẻ tấn công gửi một số lượng lớn yêu cầu ping tới một địa chỉ phủ định dịch vụ (broadcast address), dẫn đến quá tải và làm cho hệ thống bị quá tải hoặc treo.
- teardrop: Tấn công teardrop là việc gửi các gói tin mạng được tạo sao chép lỗi hoặc gây ra chồng chéo khô ng hợp lệ, dẫn đến sự xáo trộn và làm cho hệ thống hoặc mạng bị quá tải hoặc treo.
- warezclient: Warezclient là một phần mềm hoặc ứng dụng được sử dụng để truy cập và tải xuống các tệp tin hoặc nội dung không được phép phân phối, chẳng hạn như phần mềm bản quyền hoặc nội dung bảo vệ bản quyền.
- warezmaster: Warezmaster thường được sử dụng để chỉ người điều hành, quản lý hoặc điều hành các trang web, cộng đồng hoặc hệ thống có liên quan đến việc chia sẻ phần mềm hoặc tài nguyên không được phép phân phối.

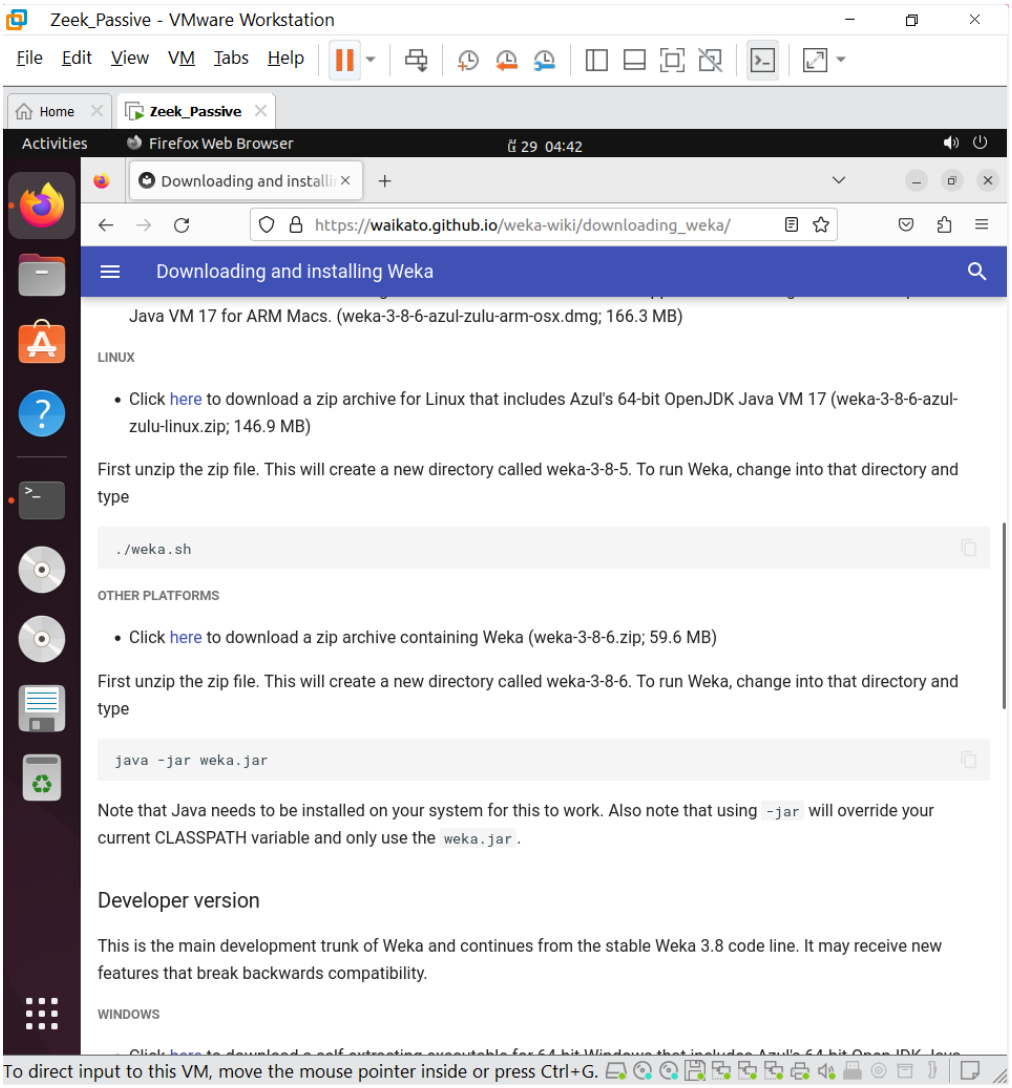
3. Mỗi instance trong tập dữ liệu KDD Cup 1999 bào gồm **41** thuộc tính, cụ thể gồm các thuộc tính:

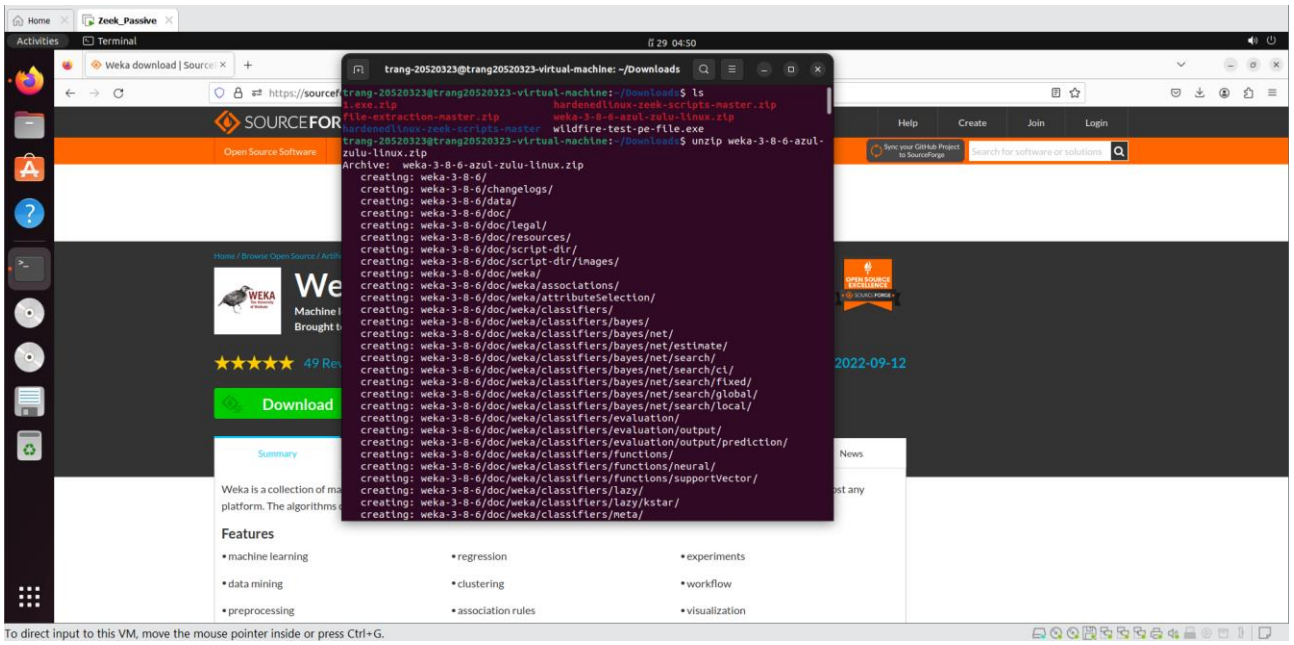
- duration
- protocol_type
- service
- flag
- src_bytes
- dst_bytes
- land
- wrong_fragment
- urgent
- hot
- num_failed_logins
- logged_in
- num_compromised
- root_shell
- su_attempted

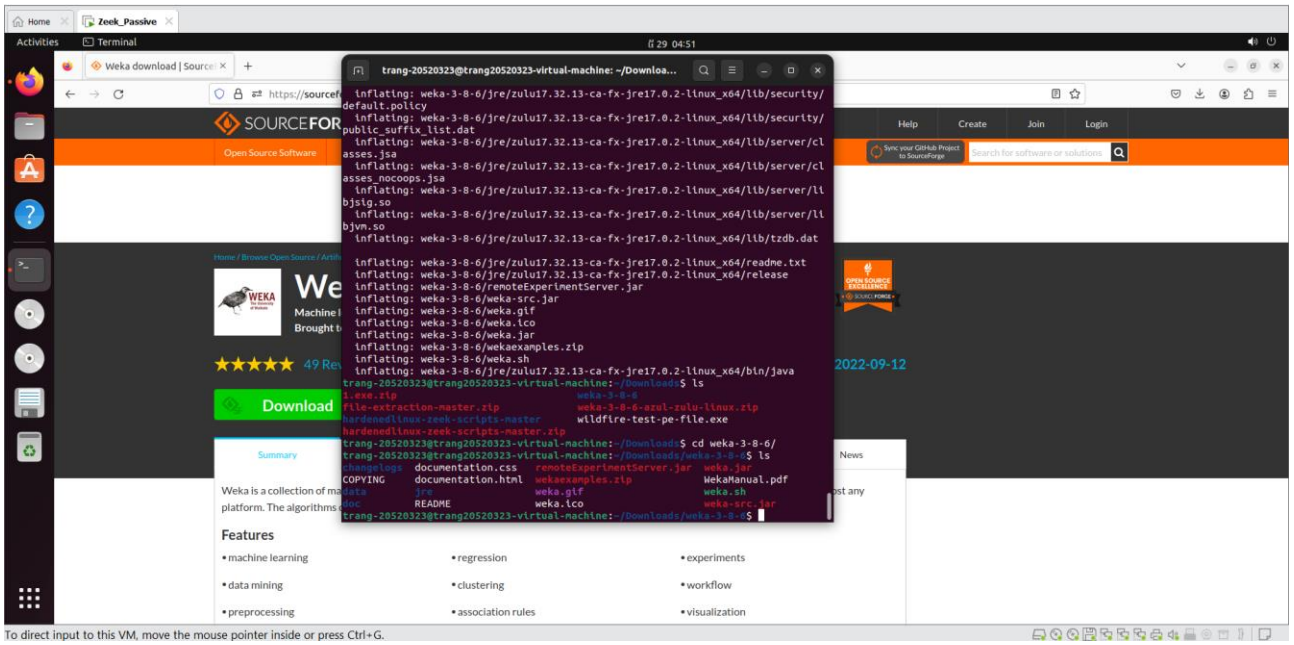
- num_root
- num_file_creations
- num_shells
- num_access_files
- num_outbound_cmds
- is_host_login
- is_guest_login
- count
- srv_count
- serror_rate
- srv_serror_rate
- rerror_rate
- srv_rerror_rate
- same_srv_rate
- diff_srv_rate
- srv_diff_host_rate
- dst_host_count
- dst_host_srv_count
- dst_host_same_srv_rate
- dst_host_diff_srv_rate
- dst_host_same_src_port_rate
- dst_host_srv_diff_host_rate
- dst_host_serror_rate
- dst_host_srv_serror_rate
- dst_host_rerror_rate
- dst_host_srv_rerror_rate

2. **Yêu cầu 2.1 Sinh viên cài đặt WEKA, tìm hiểu và load một tập dữ liệu có định dạng .arff đơn giản có sẵn của WEKA.**

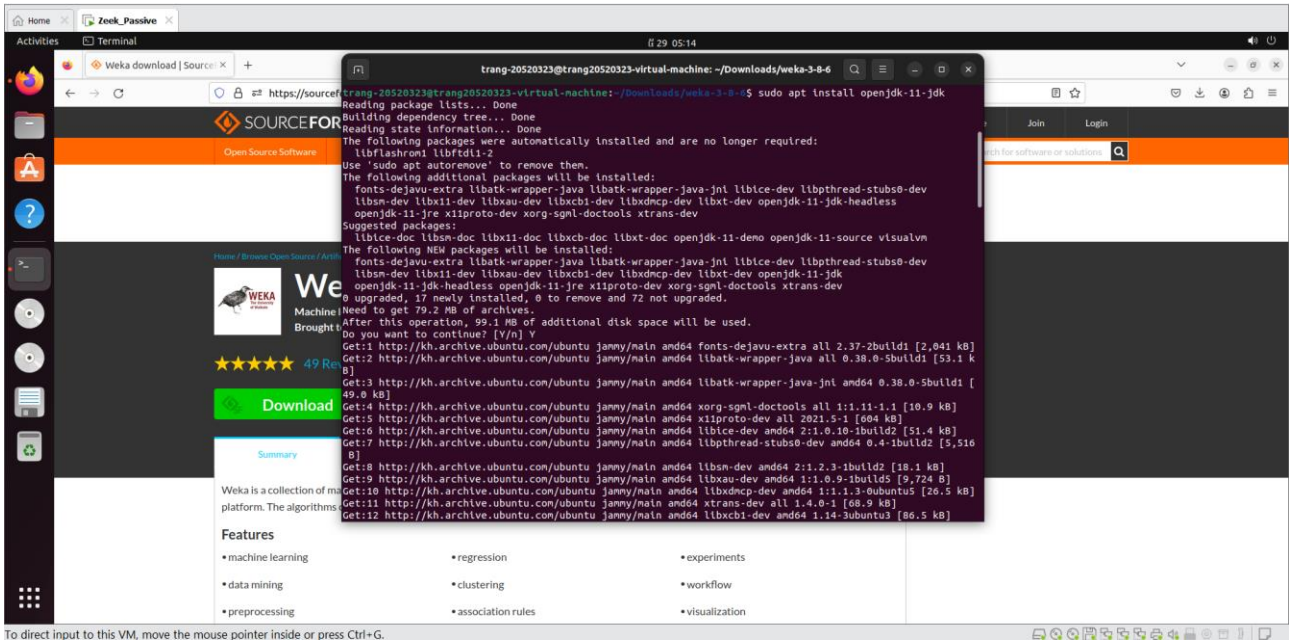
   • **Bước 1:** Cài đặt WEKA

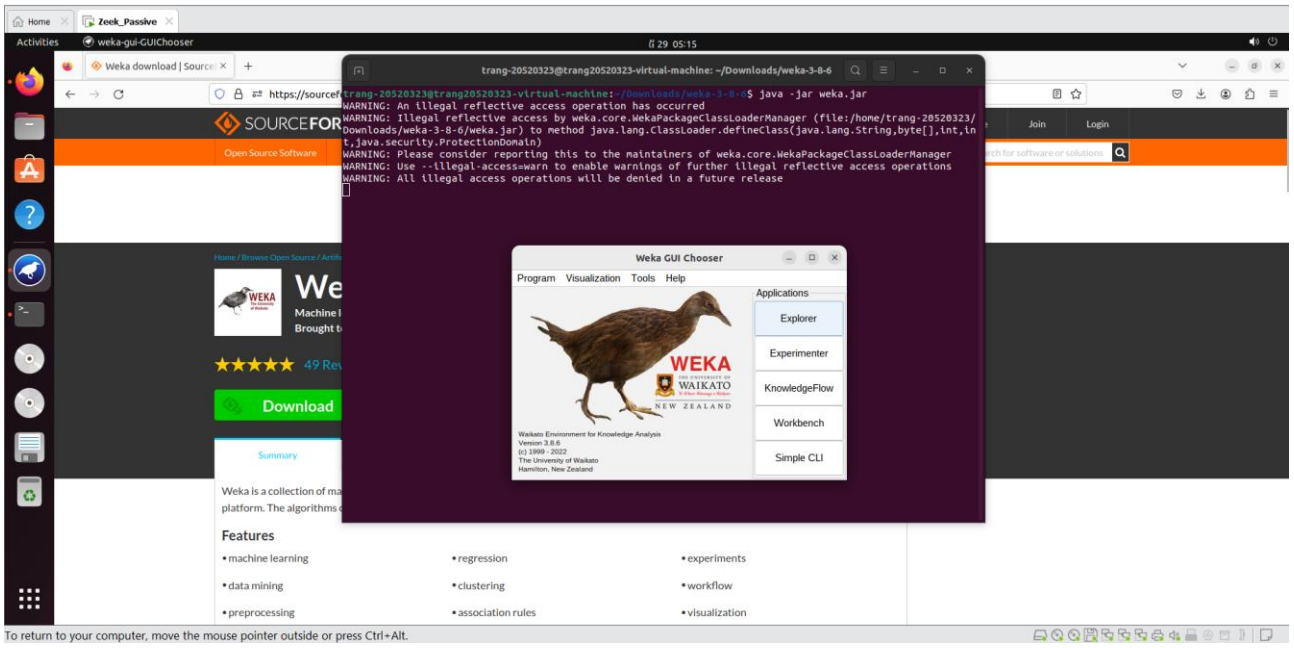

Zeek_Passive - VMware Workstation
File Edit View VM Tabs Help
Home
Zeek_Passive
Activities
Firefox Web Browser
29 04:42
Downloading and installi
https://waikato.github.io/weka-wiki/downloading_weka/
Downloading and installing Weka
Java VM 17 for ARM Macs. (weka-3-8-6-azul-zulu-arm-osx.dmg; 166.3 MB)
LINUX
Click here to download a zip archive for Linux that includes Azul's 64-bit OpenJDK Java VM 17 (weka-3-8-6-azul-zulu-linux.zip; 146.9 MB)
First unzip the zip file. This will create a new directory called weka-3-8-5. To run Weka, change into that directory and type
./weka.sh
OTHER PLATFORMS
Click here to download a zip archive containing Weka (weka-3-8-6.zip; 59.6 MB)
First unzip the zip file. This will create a new directory called weka-3-8-6. To run Weka, change into that directory and type
java -jar weka.jar
Note that Java needs to be installed on your system for this to work. Also note that using -jar will override your current CLASSPATH variable and only use the weka.jar .
Developer version
This is the main development trunk of Weka and continues from the stable Weka 3.8 code line. It may receive new features that break backwards compatibility.
WINDOWS
To direct input to this VM, move the mouse pointer inside or press Ctrl+G.


trang-20520323@trang20520323-virtual-machine: ~/Downloads
trang-20520323@trang20520323-virtual-machine:~/Downloads$ ls
trang-20520323@trang20520323-virtual-machine:~/Downloads$ unzip weka-3-8-6-azul-zulu-linux.zip
Archive: weka-3-8-6-azul-zulu-linux.zip
creating: weka-3-8-6/
creating: weka-3-8-6/changelogs/
creating: weka-3-8-6/data/
creating: weka-3-8-6/doc/
creating: weka-3-8-6/doc/legal/
creating: weka-3-8-6/doc/resources/
creating: weka-3-8-6/doc/script-dir/
creating: weka-3-8-6/doc/script-dir/images/
creating: weka-3-8-6/doc/weka/
creating: weka-3-8-6/doc/weka/associations/
creating: weka-3-8-6/doc/weka/attributeSelection/
creating: weka-3-8-6/doc/weka/classifiers/
creating: weka-3-8-6/doc/weka/classifiers/bayes/
creating: weka-3-8-6/doc/weka/classifiers/bayes/net/
creating: weka-3-8-6/doc/weka/classifiers/bayes/net/estimate/
creating: weka-3-8-6/doc/weka/classifiers/bayes/net/search/
creating: weka-3-8-6/doc/weka/classifiers/bayes/net/search/ci/
creating: weka-3-8-6/doc/weka/classifiers/bayes/net/search/fixed/
creating: weka-3-8-6/doc/weka/classifiers/bayes/net/search/global/
creating: weka-3-8-6/doc/weka/classifiers/bayes/net/search/local/
creating: weka-3-8-6/doc/weka/classifiers/evaluation/
creating: weka-3-8-6/doc/weka/classifiers/evaluation/output/
creating: weka-3-8-6/doc/weka/classifiers/evaluation/output/prediction/
creating: weka-3-8-6/doc/weka/classifiers/functions/
creating: weka-3-8-6/doc/weka/classifiers/functions/neural/
creating: weka-3-8-6/doc/weka/classifiers/functions/supportVector/
creating: weka-3-8-6/doc/weka/classifiers/lazy/
creating: weka-3-8-6/doc/weka/classifiers/lazy/kstar/
creating: weka-3-8-6/doc/weka/classifiers/meta/
Download
Features
machine learning
regression
experiments
data mining
clustering
workflow
preprocessing
association rules
visualization
To direct input to this VM, move the mouse pointer inside or press Ctrl+G.



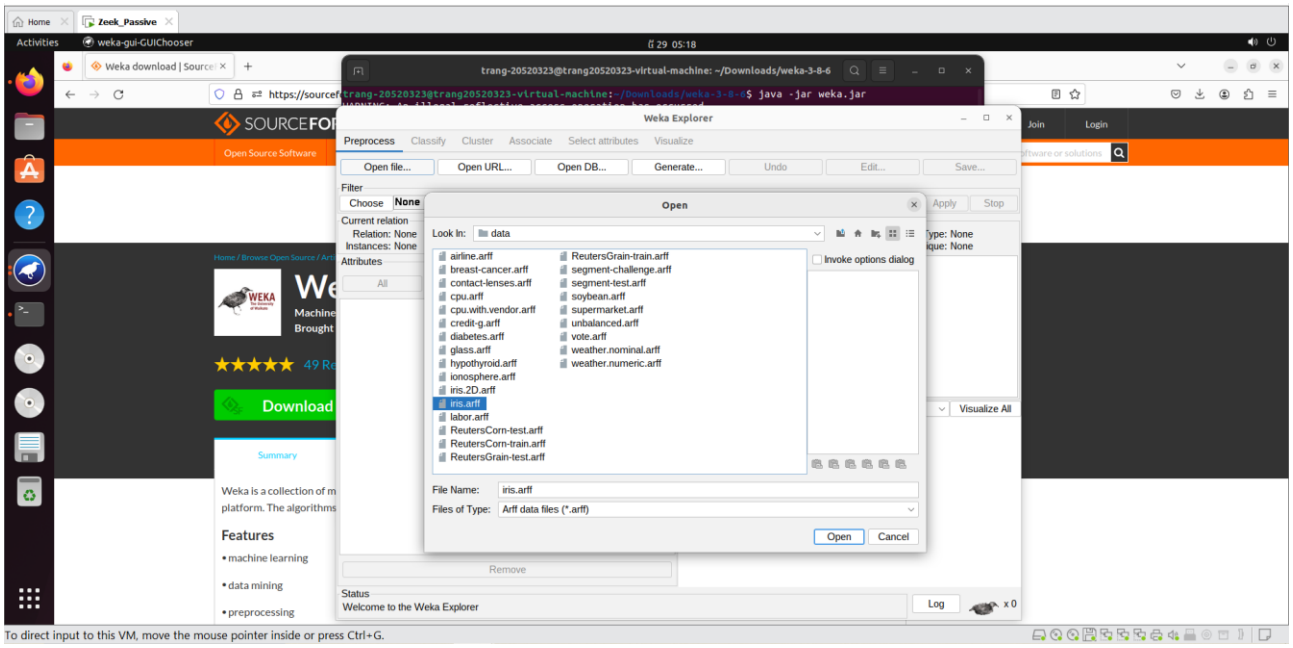
- **Bước 2:** Chạy WEKA



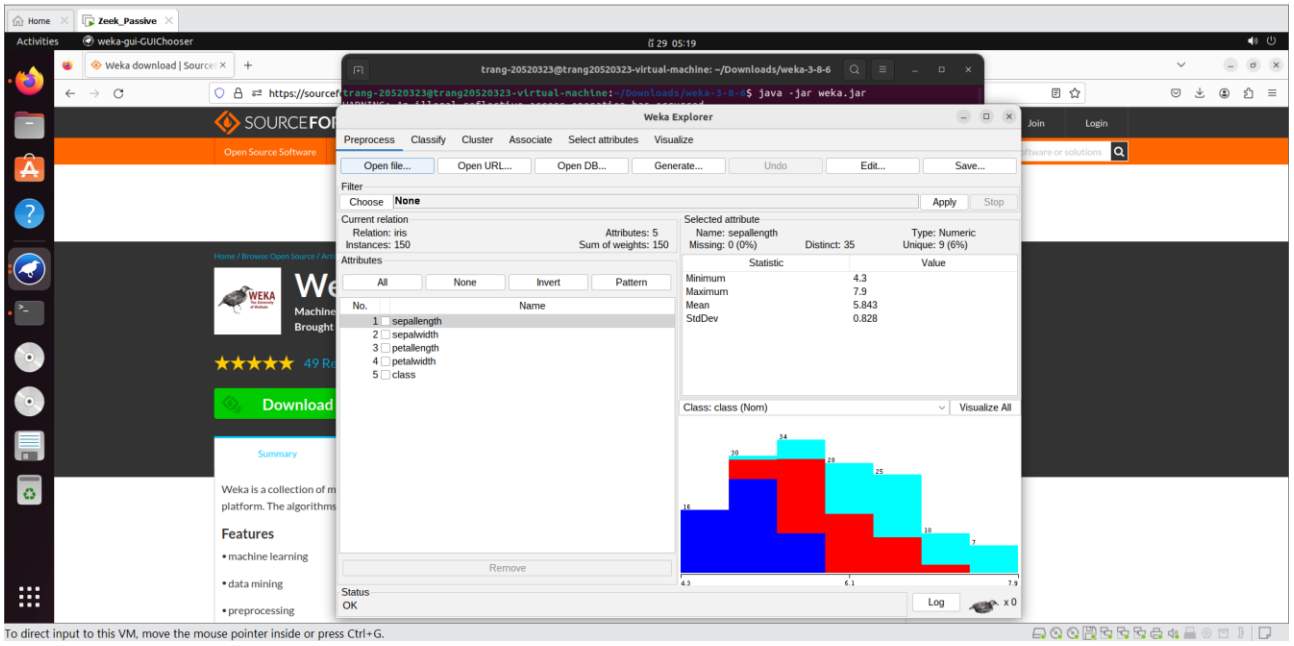


- **Bước 3:** Trong cửa sổ GUI Chooser, chọn Explorer

- **Bước 4:** mkj



- **Bước 5:** nkj



Giải thích các giá trị:

- Relation: Là tên của mối quan hệ dữ liệu. Nó thường được xác định trong tệp dữ liệu và thể hiện tên của tập dữ liệu hoặc bài toán cụ thể mà dữ liệu đại diện cho.

```
64 @RELATION iris
```

- Instances: đại diện cho số lượng mẫu hoa iris
- Attribute: Đây là các thuộc tính hoặc đặc trưng của dữ liệu, có 4 thuộc tính là "sepal length" (độ dài đài hoa), "sepal width" (độ rộng đài hoa), "petal length" (độ dài cánh hoa), và "petal width" (độ rộng cánh hoa).
- Sum of weights: Mỗi mẫu có một trọng số riêng, và tổng trọng số cho biết tổng số lượng mẫu được đại diện trong tập dữ liệu khi tính đến trọng số của chúng.
- Maximum (Giá trị tối đa): Là giá trị lớn nhất trong tập dữ liệu. Nó đại diện cho giá trị cao nhất có thể có trong tập dữ liệu.
- Minimum (Giá trị tối thiểu): Là giá trị nhỏ nhất trong tập dữ liệu. Nó đại diện cho giá trị thấp nhất có thể có trong tập dữ liệu.
- Mean (Trung bình): Là giá trị trung bình của tất cả các giá trị trong tập dữ liệu. Để tính trung bình, ta cộng tổng tất cả các giá trị và chia cho số lượng các giá trị trong tập dữ liệu.

- StdDev (Độ lệch chuẩn): Là một đại lượng thống kê mô tả mức độ phân tán của dữ liệu xung quanh giá trị trung bình.

$$\text{Độ lệch chuẩn} = \sqrt{[(\Sigma(xi - mean)^2) / N]}$$

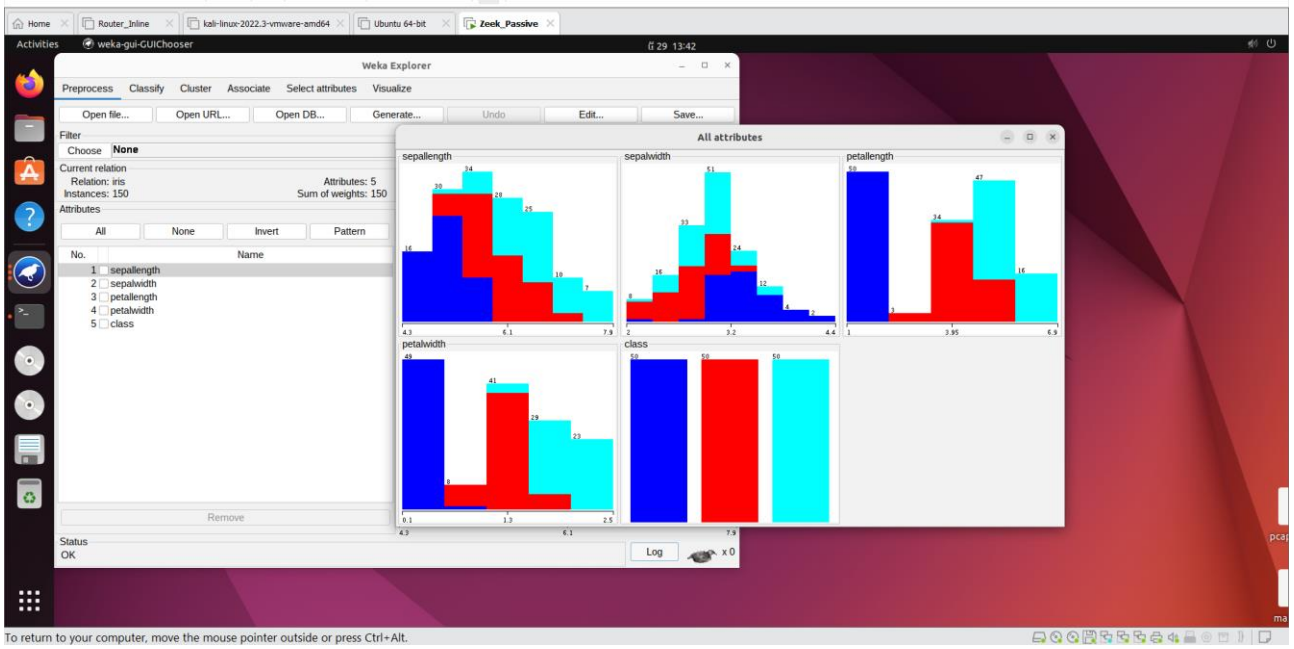
xi là giá trị của từng điểm dữ liệu.

mean là giá trị trung bình của tập dữ liệu.

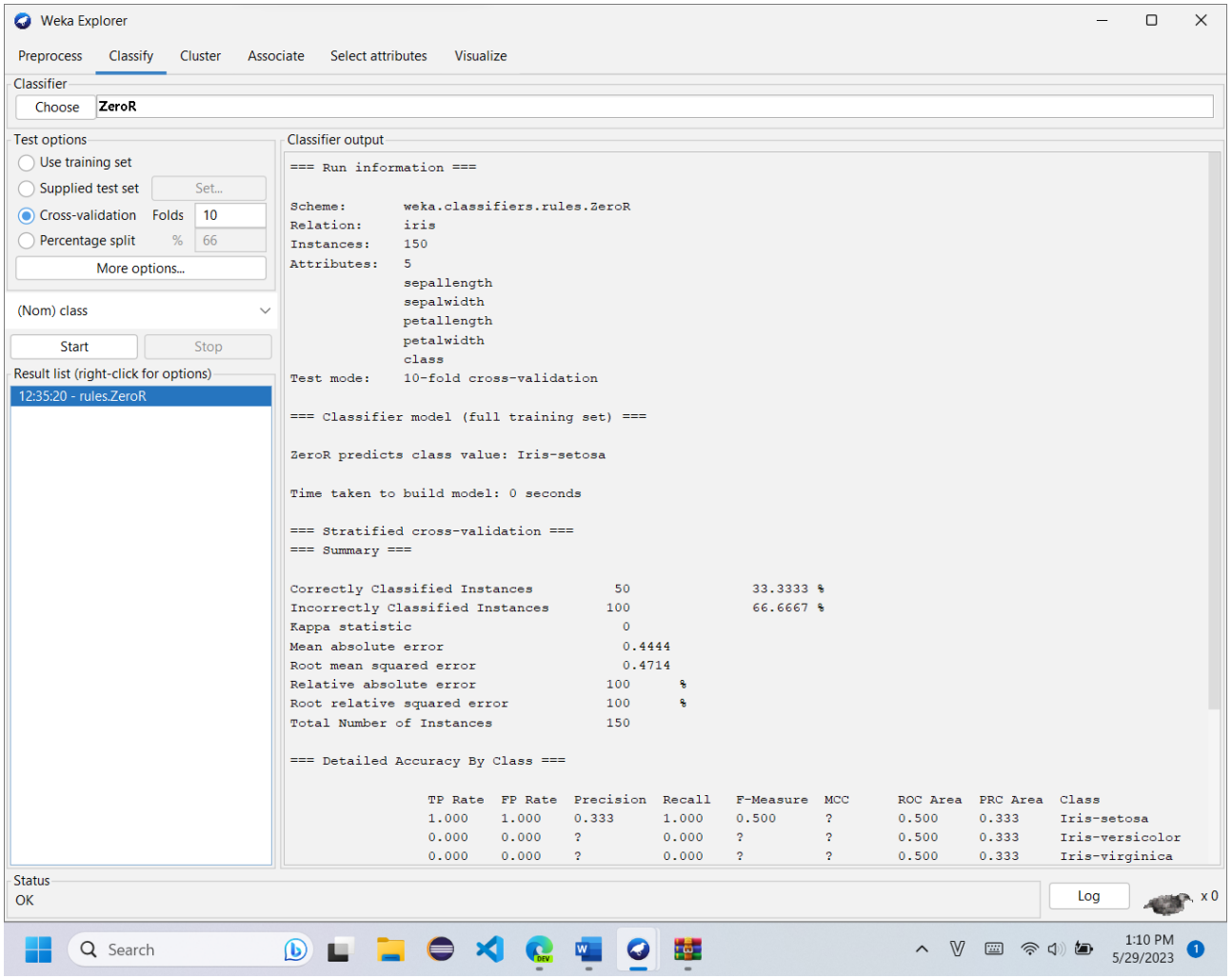
Σ là ký hiệu tổng của tất cả các giá trị.

N là số lượng các điểm dữ liệu trong tập.

- Name (Tên): Đây là tên của thuộc tính.
- Missing (Thiếu): Đây là số lượng giá trị thiếu (missing values) trong thuộc tính. Giá trị "Missing" cho biết có bao nhiêu mẫu trong tập dữ liệu mà thuộc tính này không có giá trị được gán.
- Distinct: hiển thị số lượng giá trị có trong thuộc tính đó.
- Unique (Duy nhất): Đây là số lượng giá trị duy nhất (unique values) trong thuộc tính. Giá trị "Unique" cho biết có bao nhiêu giá trị khác nhau xuất hiện trong thuộc tính, đó là một chỉ số về sự đa dạng của thuộc tính.



3. Yêu cầu 2.2 Sinh viên lựa chọn 01 bộ phân lớp (classifier) bất kỳ và thực hiện khài thác trên tập dữ liệu đã chọn ở trên. Trình bày và giải thích kết quả.



Các thuật toán:

- Naive Bayes: Thuật toán Naive Bayes dựa trên nguyên lý Bayes và giả định "naive" rằng các đặc trưng đầu vào là độc lập nhau. Nó tính toán xác suất xảy ra của một lớp dựa trên xác suất của các đặc trưng đầu vào và sử dụng ngưỡng để phân loại mẫu.

- Decision Tree: Decision Tree sử dụng cây quyết định để tạo ra một loạt các quy tắc phân loại dựa trên giá trị của các đặc trưng đầu vào. Thuật toán tạo ra cây bằng cách chọn các đặc trưng quan trọng nhất và xác định các điểm chia để phân loại dữ liệu.

- Random Forest: Random Forest là một phương pháp kết hợp nhiều cây quyết định (decision trees). Nó tạo ra nhiều cây quyết định riêng biệt và sử dụng sự phiếu bầu (voting) để đưa ra quyết định phân loại cuối cùng.

- Support Vector Machines (SVM): SVM tìm ra một siêu phẳng (hyperplane) trong không gian đặc trưng, tốt nhất phân chia giữa các lớp dữ liệu. Nó cố gắng tối đa hóa khoảng cách từ các điểm dữ liệu đến siêu phẳng để đảm bảo tính tổng quát của mô hình.

- K-Nearest Neighbors (KNN): KNN phân loại một mẫu dựa trên các mẫu gần nhất trong tập dữ liệu huấn luyện. Thuật toán tính toán khoảng cách giữa các mẫu và chọn k mẫu gần nhất để đưa ra quyết định phân loại.

Dùng:

- Lý do: Thuật toán phân loại đơn giản nhất của weka.
- Lớp được dự đoán với ZeroR: Iris-setosa
- Time taken to build model: 0 seconds -> thời gian xây dựng mô hình: 0s
- Số mẫu phân loại đúng: 50 – 33.33%
- Phân loại sai: 100 – 66.67%
- Kappa statistic: giá trị sẽ nằm trong khoản từ -1 đến 1, càng gần 1 thì khả năng phân loại càng tốt, trong trường hợp này = 0 -> không có khả năng phân loại chính xác.
- Cross-Validation 10: 1 tập dữ liệu làm traning set, tập kiểm tra (test set) 9 tập (quá trình trên lặp lại với tất cả các tập).
- Mean absolute error: độ lỗi trung bình giữa các mô hình, giá trị càng nhỏ thì mô hình càng có khả năng dự đoán chính xác.
- Root Mean absolute error: cũng là độ lỗi trung bình nhưng tính bằng căn bậc hai tổng bình phương độ lỗi đề cập ở trên và càng nhỏ thì mô hình cũng dự đoán càng chính xác.

. MAE : Trung bình của sai biệt tuyệt đối

$$MAE = \frac{\sum abs(f_i - y_i)}{n}$$

RMSE : Căn bậc 2 của trung bình bình phương sai số

$$RMSE = \sqrt{\frac{\sum_i^n (f_i - y_i)^2}{n}}$$

- Relative absolute error: tỉ lệ giữa (độ lỗi tuyệt đối và độ lỗi trung bình của mô hình) / độ lỗi trung bình của mô hình đơn giản nhất (trong bài sẽ là mô hình Iris-setosa) -> 100% nghĩa là nó không có khả năng dự đoán tốt hơn so với mô hình Iris-setosa.
  Công thức tính (MAE / phạm vi mục tiêu) * 100%
- Root … (RMSE): tỉ lệ giữa (độ lỗi bình phương tương đối và độ lỗi bình phương trung bình) / độ lỗi bình phương trung bình của mô hình đơn giản nhất Iris-setosa.
  Tương tự, công thức tính là (RMAE/phạm vi mục tiêu) *100%

11. RRSE Root relative squared error

$$RRSE = \sqrt{\frac{\sum (f_i - y_i)^2}{\sum (y_i - \bar{y})^2}}$$

12. RAE : Relative absolute error

$$RAE = \frac{\sum (abs(f_i - y_i))}{\sum (abs(y_i - \bar{y}))}$$

- fi là giá trị dự đoán
- yi là giá trị thực tế
- ȳ là giá trị trung bình của giá trị thực tế

Precision: 50/150 = 1/3 = 0.333

Recall:

- TP: 50 mẫu phân loại đúng
- FN: xét 50 mẫu thuộc Iris-setosa và không có mẫu nào bị phân loại sai

  ➔ Recall = 50/(50+0) = 1

F-measure = F1-score = 2* (precision*recall/(precision+recall)) = 2*(0.333/1.333) = 0.5

MCC (Matthews correlation coefficient): một độ đo đánh giá hiệu suất của mô hình phân loại, giá trị từ -1 đến 1, càng gần 1 càng tốt

ROC Area: diện tích dưới đường cong ROC (Receiver Operating Characteristic)

PRC Area: diện tích dưới đường cong Precision-Recall

Kết quả lấy được từ ma trận hỗn loạn:

Lớp a (Iris-setosa): 50 mẫu được phân loại đúng, không có mẫu nào bị phân loại sai.

Lớp b (Iris-versicolor): 50 mẫu bị phân loại sai và không có mẫu nào được phân loại đúng.

Lớp c (Iris-virginica): 50 mẫu bị phân loại sai và không có mẫu nào được phân loại đúng.

ZeroR là một thuật toán phân loại đơn giản trong machine learning. Nó được gọi là ZeroR vì nó dựa trên quy tắc đơn giản nhất: chọn lớp phổ biến nhất trong tập dữ liệu huấn luyện và dự đoán mọi mẫu đầu vào thuộc lớp đó.

4. **Yêu cầu 3.1 Sinh viên lựa chọn 01 bộ phân lớp bất kỳ và thực hiện khài thác trên tập dữ liệu KDD Cup 1999. Giải thích và đánh kết quả.**

   Giải thích các test options:

   - **Use training set:** Dataset dùng để train được dùng test model
   - **Supplied test set:** Dùng dataset độc lập với data train để test model.
   - **Cross-validation:** Với tham số K fold, nó sẽ chia dataset thành K set con, dùng K-1 set để train và dùng set còn lại để test. Lặp lại quá trình này cho từng set còn lại cho đến khi hết K set đó.

- **Percentage split:** Với tham số K%, nghĩa là nó sẽ dùng K% dataset để train, số còn lại dùng để test.

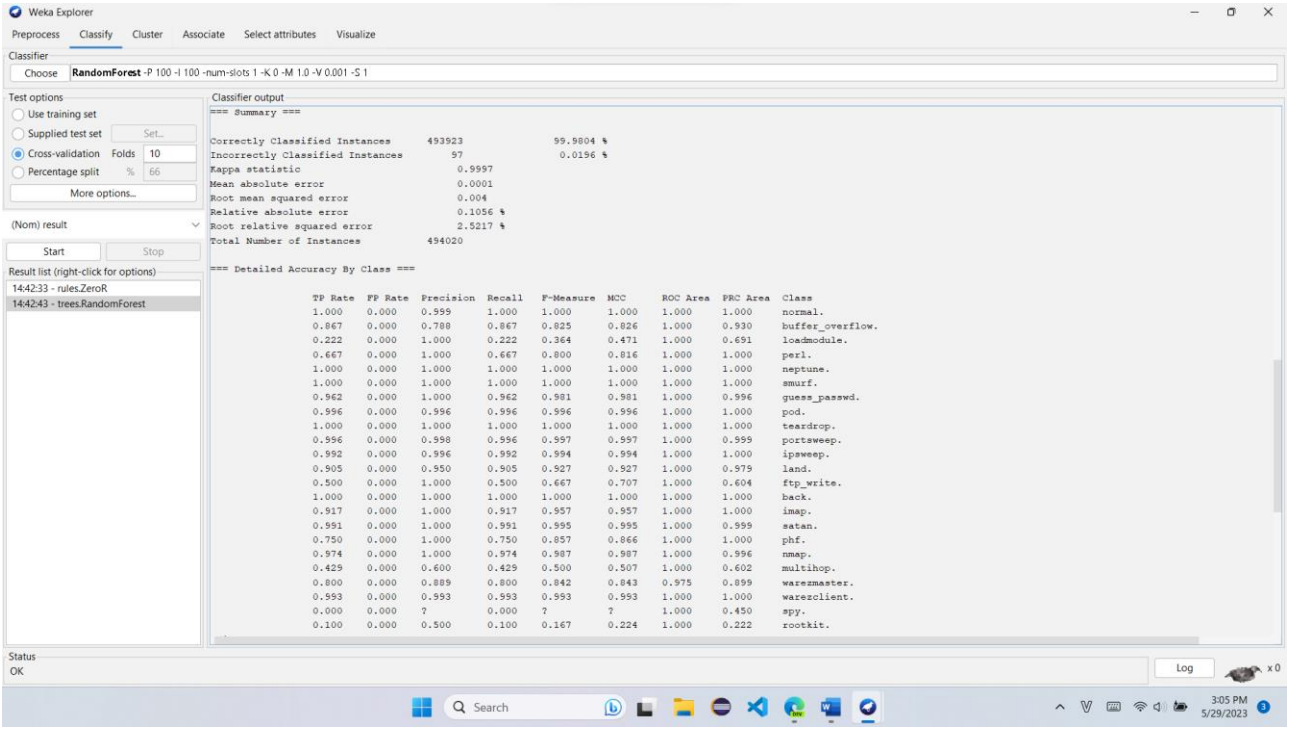
Classifier được chọn: RandomForest. Nguyên lý hoạt động là bằng cách kết hợp các dự đoán của cây quyết định nhỏ hơn để tổng hợp vào tạo ra dự đoán cuối cùng. Và test model bằng option Cross-validation với folds là 10.

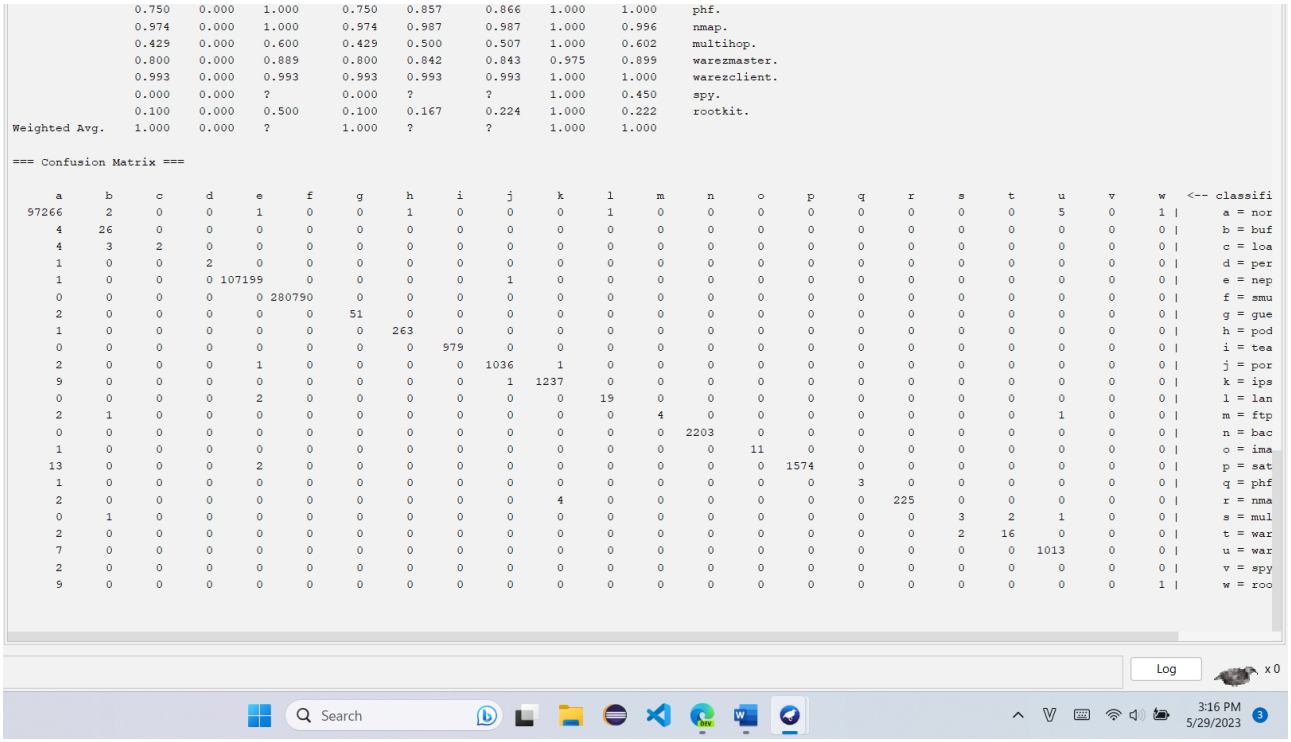
Lý do chọn:

- Kết quả output có độ chính xác cao.
- Không yêu cầu quá nhiều việc tiền xử lý data.
- Xử lý được đa dạng dữ liệu như dạng số, dạng danh mục,...

Kết quả đánh giá:

Với bộ dataset đưa vào có tất cả 494020 mẫu, sau khi train, kết quả thu được khi test model cho ra kết quả như sau:



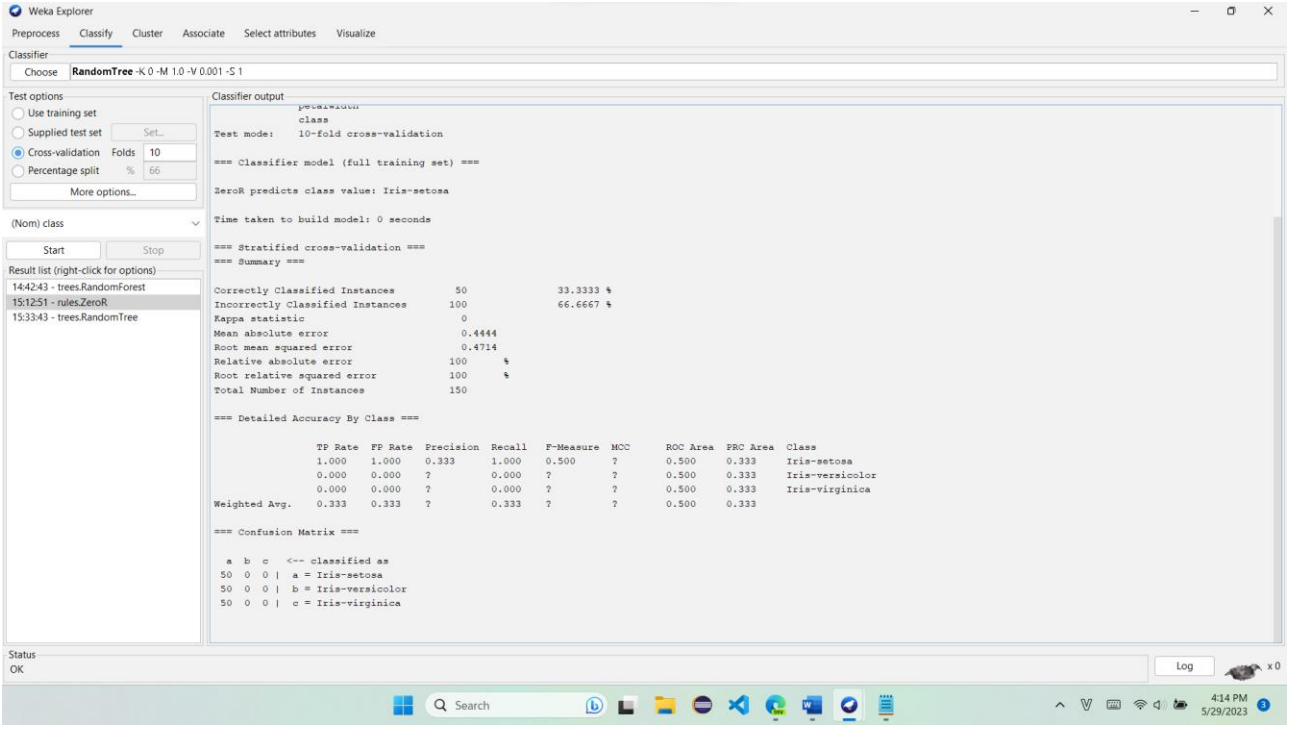


Hình cho thấy model có mức độ phát hiện chính xác rất cao với khoảng ~99%, với các chỉ số khá tốt

- Kappa statistic: cho biết mức độ chính xác của lớp phân loại, từ -1 đến 1, càng gần 1 thì càng tốt
- Mean Absolute Error: độ lỗi trung bình giữa các dự đoán và giá trị thực tế.
- Root Mean Absolute Error: độ lỗi trung bình giữa các dự đoán và giá trị thực tế.
- Relative absolute error: độ lỗi tương đối giữa dự đoán và giá trị thực tế

Note, viết cho:

- Accurancy
- Precision



YÊU CẦU CHUNG

- Sinh viên tìm hiểu và thực hành theo hướng dẫn.
- Nộp báo cáo kết quả chi tiết những việc (**Report**) bạn đã thực hiện, quan sát thấy và kèm ảnh chst setụp màn hình kết quả (nếu có); giải thích cho quan sát (nếu có).
- Sinh viên báo cáo kết quả thực hiện và nộp bài.

Báo cáo:

- File .PDF. Tập trung vào nội dung, không mô tả lý thuyết.
- Nội dung trình bày bằng Font chữ Times New Romans/ hoặc font chữ của mẫu báo cáo này (UTM Neo Sans Intel/UTM Viet Sach)– cỡ chữ 13. Canh đều (Justify) cho văn bản. Canh giữa (Center) cho ảnh chụp.
- Đặt tên theo định dạng: [Mã lớp]-SessionX_GroupY. (trong đó X là Thứ tự buổi Thực hành, Y là số thứ tự Nhóm Thực hành/Tên Cá nhân đã đăng ký với GV).

  *Ví dụ: [*NT101.K11.ANTT*]-Session1_Group3.*
- Nếu báo cáo có nhiều file, nén tất cả file vào file .ZIP với cùng tên file báo cáo.
- Không đặt tên đúng định dạng – yêu cầu, sẽ **KHÔNG** chấm điểm.

– Nộp file báo cáo trên theo thời gian đã thống nhất tại courses.uit.edu.vn.

Đánh giá: Sinh viên hiểu và tự thực hiện. Khuyến khích:

– Chuẩn bị tốt.

– Có nội dung mở rộng, ứng dụng trong kịch bản/câu hỏi phức tạp hơn, có đóng góp xây dựng.

Bài sao chép, trễ, … sẽ được xử lý tùy mức độ vi phạm.

HẾT