BÁO CÁO BÀI TẬP

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

Tên chủ đề: Lab 5

GVHD: ThS. Đỗ Hoàng Hiển

1. THÔNG TIN CHUNG:

(Liệt kê tất cả các thành viên trong nhóm)

Lớp: NT204.O21.ATCL.1

STT	Họ và tên	MSSV	Email
1	Nguyễn Đại Nghĩa	21521182	21521182@gm.uit.edu.vn
2	Hoàng Gia Bảo	21521848	21521848@gm.uit.edu.vn

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

BÁO CÁO CHI TIẾT

Yêu cầu 1.1 Sinh viên tìm hiểu về tập dữ liệu KDD Cup 1999 và điền các kết quả tìm hiểu được vào form bên dưới.

Em thực hiện tìm hiểu về tập dữ liệu KDD Cup 1999 thông qua bài báo ở đường link sau:
https://www.researchgate.net/publication/326000849_Review_of_KDD_Cup_'99_NSL-KDD_and_Kyoto_2006_datasets

Trong bài báo có chỉ ra các thông tin như sau:

features labeled as normal or attacks (See Table 1).

Table 1 – Features in the KDD Cup '99 dataset
Таблица 1 – Атрибуты в KDD Cup '99 базе данных
Табела 1 – Атрибути у KDD Cup '99 бази података

Index	Feature name	Description
1	duration	Length of connection
2	protocol type	Type of protocol (TCP, UDP...)
3	service	Destination service (ftp, telnet...)
4	flag	Status of connection
5	source bytes	No. of B from source to destination
6	destination bytes	No. of B from destination to source
7	land	If the source and destination address are the same land=1/if not, then 0
8	wrong fragments	No. of wrong fragments
9	urgent	No. of urgent packets
10	hot	No. of hot indicators
11	failed logins	No. of unsuccessful attempts at login

Index	Feature name	Description
12	logged in	If logged in=1/if login failed 0
13	# compromised	No. of compromised states
14	root shell	If a command interpreter with a root account is running root shell=1/if not, then 0
15	su attempted	If an su command was attempted su attempted=1/if not, then 0 (temporary login to the system with other user credentials)
16	# root	No. of root accesses
17	# file creations	No. of operations that create new files
18	# shells	No. of active command interpreters
19	# access files	No. of file creation operations
20	# outbound cmds	No. of outbound commands in an ftp session
21	is hot login	is host login=1 if the login is on the host login list/if not, then 0
22	is guest login	If a guest is logged into the system, is guest login=1/if not, then 0
23	count	No. of connections to the same host as the current connection at a given interval
24	srv count	No. of connections to the same service as the current connection at a given interval
25	serror rate	% of connections with SYN errors
26	srv error rate	% of connections with SYN errors
27	rerror rate	% of connections with REJ errors
28	srv rerror rate	% of connections with REJ errors
29	same srv rate	% of connections to the same service
30	diff srv rate	% of connections to different services
31	srv diff host rate	% of connections to different hosts
32	dst host count	No. of connections to the same destination
33	dst host srv count	No. of connections to the same destination that use the same service
34	dst host same src rate	% of connections to the same destination that use the same service
35	dst host srv rate	% of connections to different hosts on the same system
36	dst host same srv port rate	% of connections to a system with the same source port
37	dst host srv diff host rate	% of connections to the same service coming from different hosts
38	dst host serror rate	% of connections to a host with an S0 error
39	dst host srv serror rate	% of connections to a host and specified service with an S0 error
40	dst host serror rate	% of connections to a host with an RST error
41	dst host srv serror rate	% of connections to a host and specified service with an RST error

Table 2 – Categories of attacks
Таблица 2 – Категория атак
Табела 2 – Категорије напада

Category of Attack	Attack name
Probe	ipsweep, nmap, portsweep, satan
DoS (Denial of Service)	back, land, neptune, pod, smurf, teardrop
U2R (User to Root)	buffer_overflow, loadmodule, perl, rootkit
R2L (Remote to Local)	ftp_write, guesspasswd, imap, multihop, phf, spy, warezlient, warezmaster

Từ những thông tin đã tìm hiểu được thì em sẽ điền vào chỗ trống như sau:

1. Số nhóm tấn công: 4.
   Kể tên các nhóm tấn công: Probe, DoS, U2R, R2L.

2. Số kiểu tấn công: 22.
   Kể tên các kiểu tấn công được gán nhãn: ipsweep, nmap, portsweep, satan, back, land, neptune, pod, smurf, teardrop, buffer_overflow, loadmodule, perl, rootkit, ftp_write, imap, multihop, phf, spy, warezmaste.

   Đối với các kiểu tấn công được gán nhãn thì em chỉ thực hiện việc tìm kiếm trên tập kddcup.data_10_percent, thì cho được kết quả như trên, nó không xuất hiện guesspasswd và warezlient.

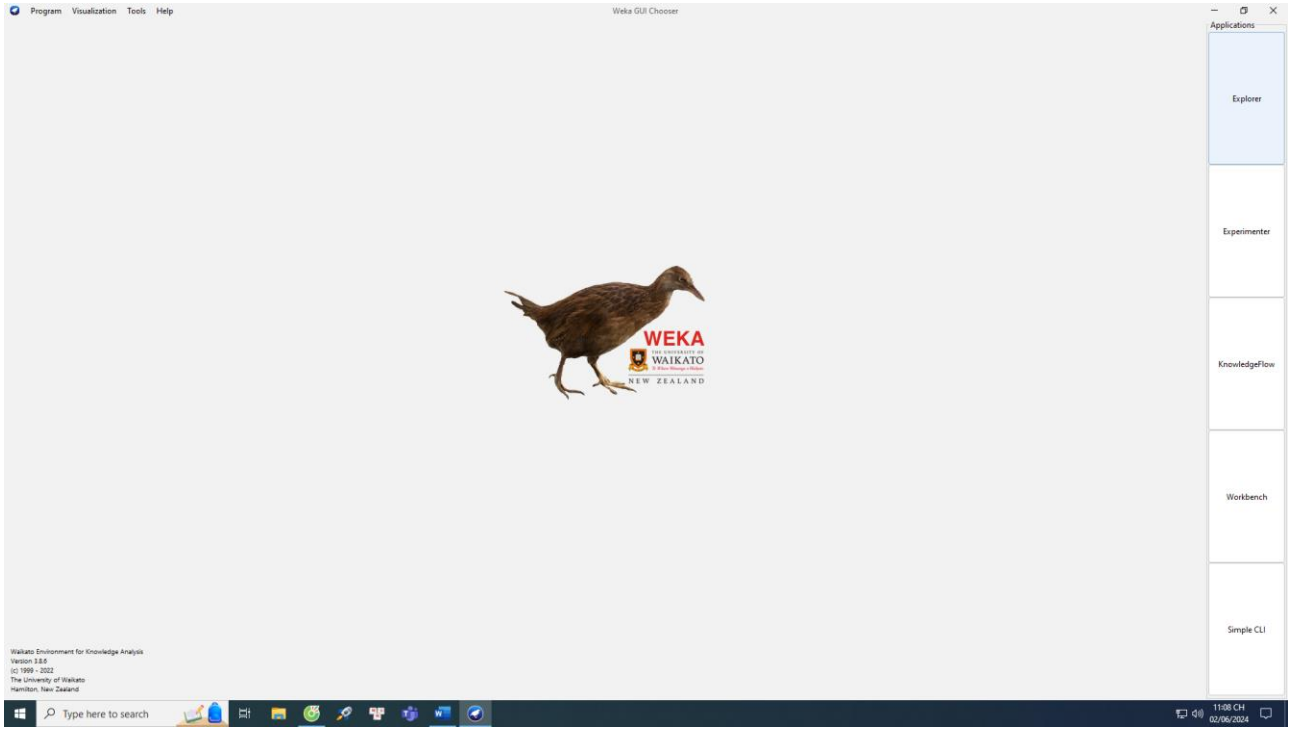
3. Mỗi instance trong tập dữ liệu KDD Cup 1999 bao gồm 41 thuộc tính, cụ thể gồm các thuộc tính:

   1. duration
   2. protocol type
   3. service
   4. flag
   5. source bytes
   6. destination bytes
   7. land
   8. wrong fragments
   9. urgent
   10. hot
   11. failed logins
   12. logged in
   13. # compromised
   14. root shell
   15. su attempted

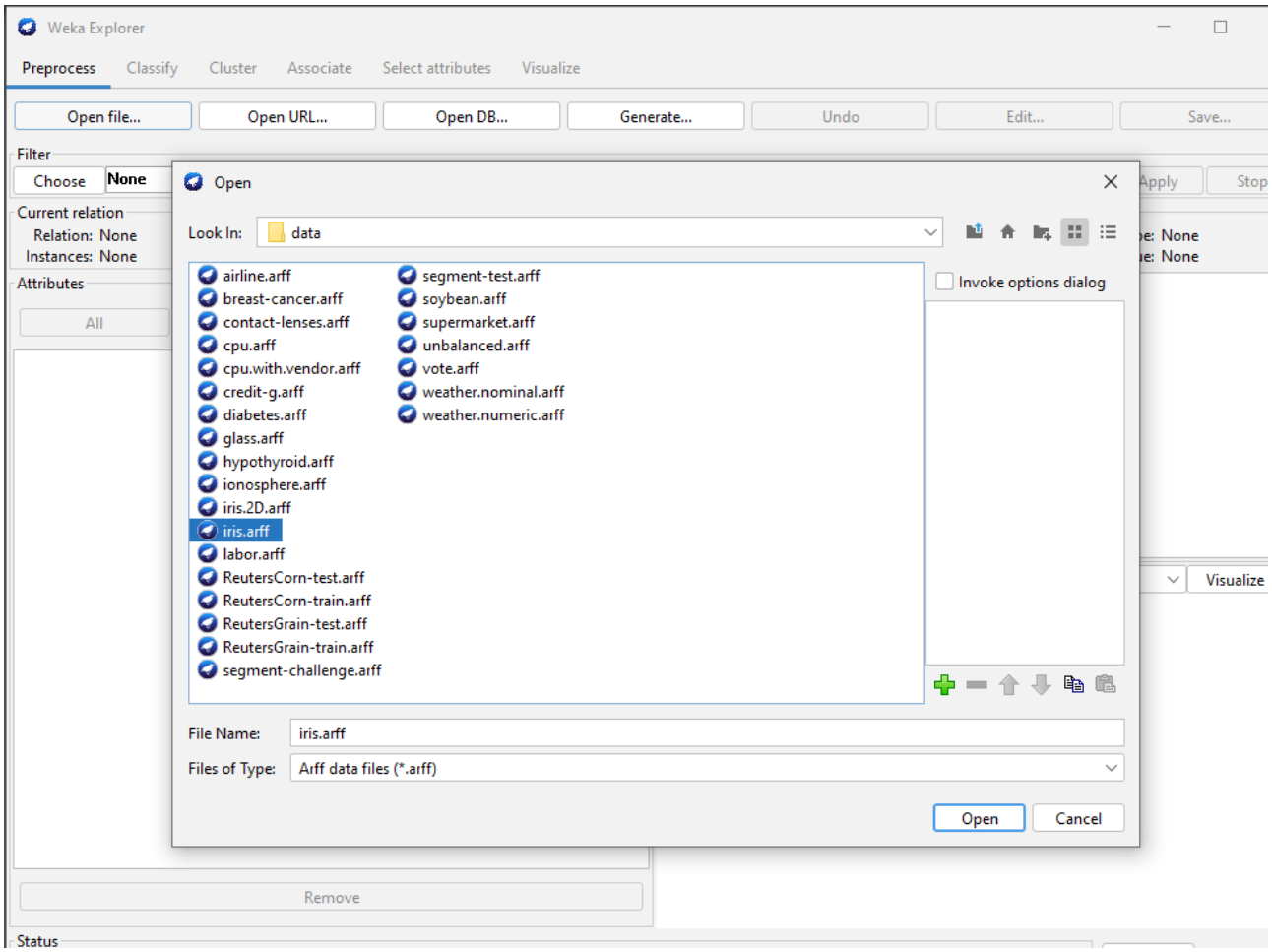
16.# root
17.# file creations
18.# shells
19.# access files
20.# outbound cmds
21.is hot login
22.is guest login
23.count
24.srv count
25.serror rate
26.srv error rate
27.rerror rate
28.srv rerror rate
29.same srv rate
30.diff srv rate
31.srv diff host rate
32.dst host count
33.dst host srv count
34.dst host same src rate
35.dst host srv rate
36.dst host same srv port rate
37.dst host srv diff host rate
38.dst host serror rate
39.dst host srv serror rate
40.dst host serror rate
41.dst host srv serror rate

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.

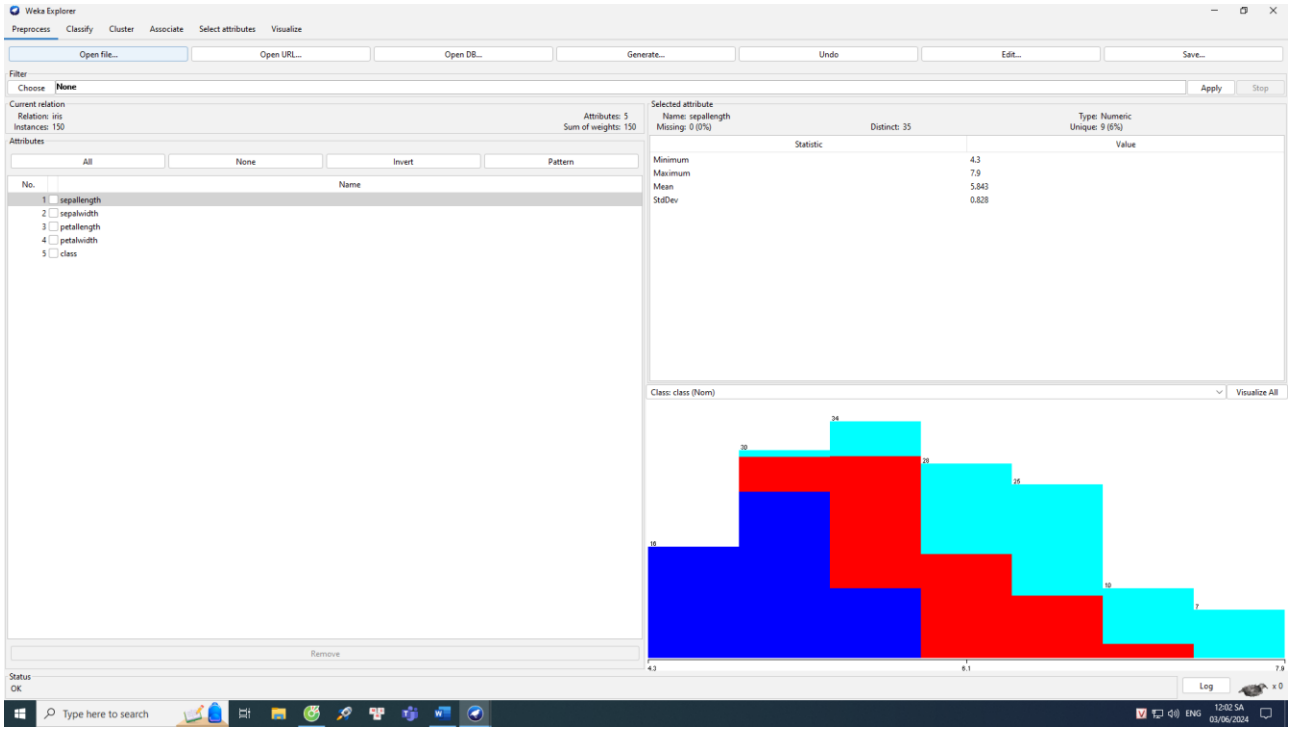
Sau khi tải về và cài đặt thành công thì em mở WEKA lên và xuất hiện giao diện như sau:



Tiến hành chọn iris.arff:



Sau khi mở file lên nhận được kết quả như sau:



Nhìn qua những gì hiển thị trong tab Preprocess, em thấy được các giá trị sau:

- Relation: iris - Đây là tên của tập dữ liệu đang được sử dụng.

- Attributes: 5 - Có 5 thuộc tính được liệt kê:
  - sepallength: Chiều dài đài hoa.
  - sepalwidth: Chiều rộng đài hoa.
  - petallength: Chiều dài cánh hoa.
  - petalwidth: Chiều rộng cánh hoa.
  - class: Phân loại các loại hoa Iris.

- Instances: 150 - Bộ dữ liệu có 150 mẫu dữ liệu.

- Sum of weights: 150 - Tổng số trọng số của các instances, trong hầu hết các trường hợp, mỗi instance có trọng số là 1, do đó tổng số trọng số sẽ bằng số lượng instances.

- Selected attribute: Thông tin chi tiết về thuộc tính được chọn ("sepallength" trong trường hợp này thông qua Name):
  - Missing (0%): Phần trăm dữ liệu bị thiếu cho thuộc tính đang được xem xét. Trong trường hợp này là 0%, điều này có nghĩa là không có giá trị nào bị thiếu cho thuộc tính "sepallength".

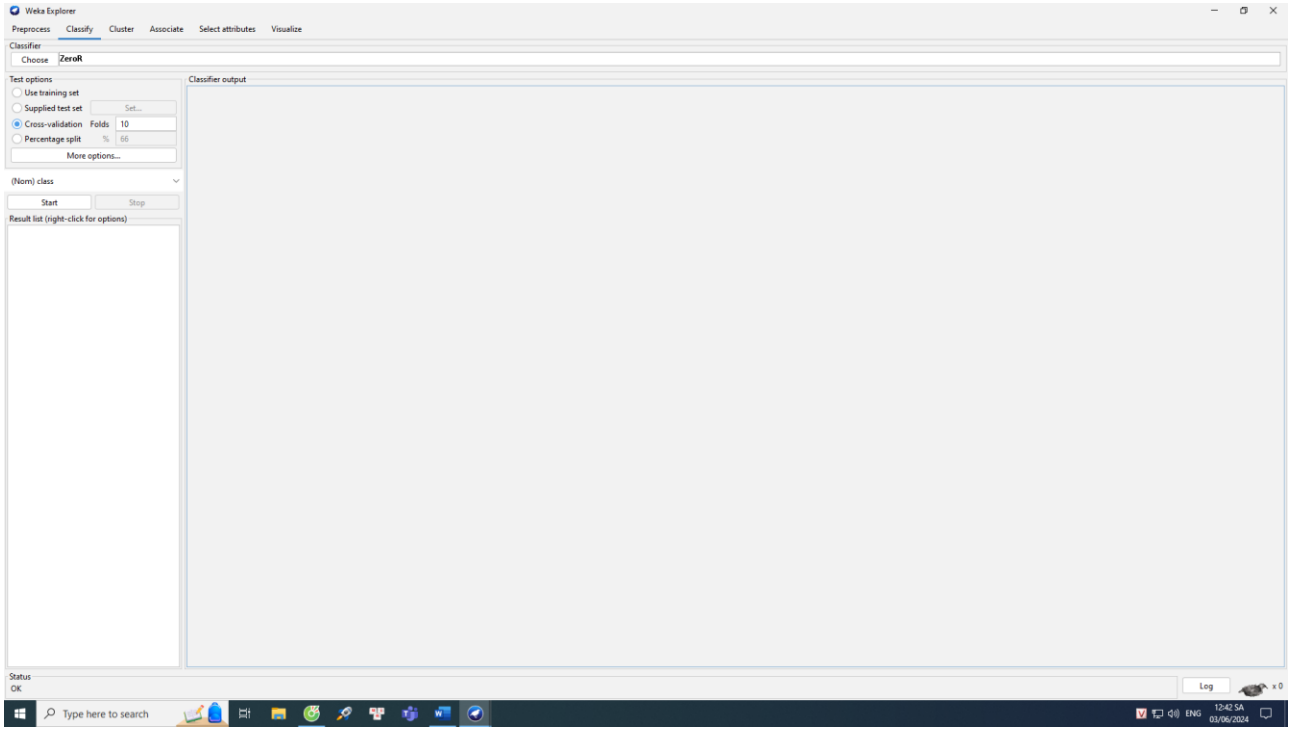
- Distinct (35): Số lượng giá trị riêng biệt được tìm thấy trong thuộc tính này. Có 35 giá trị khác nhau được ghi nhận cho "sepallength".
- Type (Numeric): Kiểu dữ liệu của thuộc tính, trong trường hợp này là số.
- Unique (9%): Phần trăm giá trị trong thuộc tính này là duy nhất. Điều này cho thấy 9% các giá trị của thuộc tính "sepallength" là duy nhất.
- Ngoài ra ở phía dưới còn có thêm các giá trị sau:
  - Minimum: Giá trị nhỏ nhất của sepallength là 4.3.
  - Maximum: Giá trị lớn nhất của sepallength là 7.9.
  - Mean: Giá trị trung bình là 5.843.
  - StdDev: Độ lệch chuẩn là 0.828.

- Cuối cùng là hình vẽ biểu đồ ở góc cuối bên phải hiển thị sự phân bố của thuộc tính "sepallength" với các lớp khác nhau được đánh dấu bằng các màu khác nhau, cho phép người dùng nhìn thấy mối quan hệ giữa chiều dài đài hoa và các loài hoa Iris khác nhau.

Tương tự vậy 4 thuộc tính còn lại cũng có các thông tin giống như “sepallength”, chỉ có là khác về chỉ số.

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

Đây là tab Classify sau khi mở lên:



Nhìn vào Test option thì em thấy có 4 options sau:

- Use training set: Khi lựa chọn này được chọn, toàn bộ tập dữ liệu huấn luyện sẽ được sử dụng để đào tạo mô hình và cũng được sử dụng để kiểm thử mô hình. Tuy nhiên, phương pháp này thường không khách quan vì có thể dẫn đến hiện tượng overfitting.

- Supplied test set: Tùy chọn này cho phép sử dụng một tập dữ liệu thử nghiệm riêng biệt, không phải là phần của tập dữ liệu huấn luyện, để kiểm thử mô hình. Điều này đảm bảo rằng mô hình được đánh giá một cách khách quan hơn.

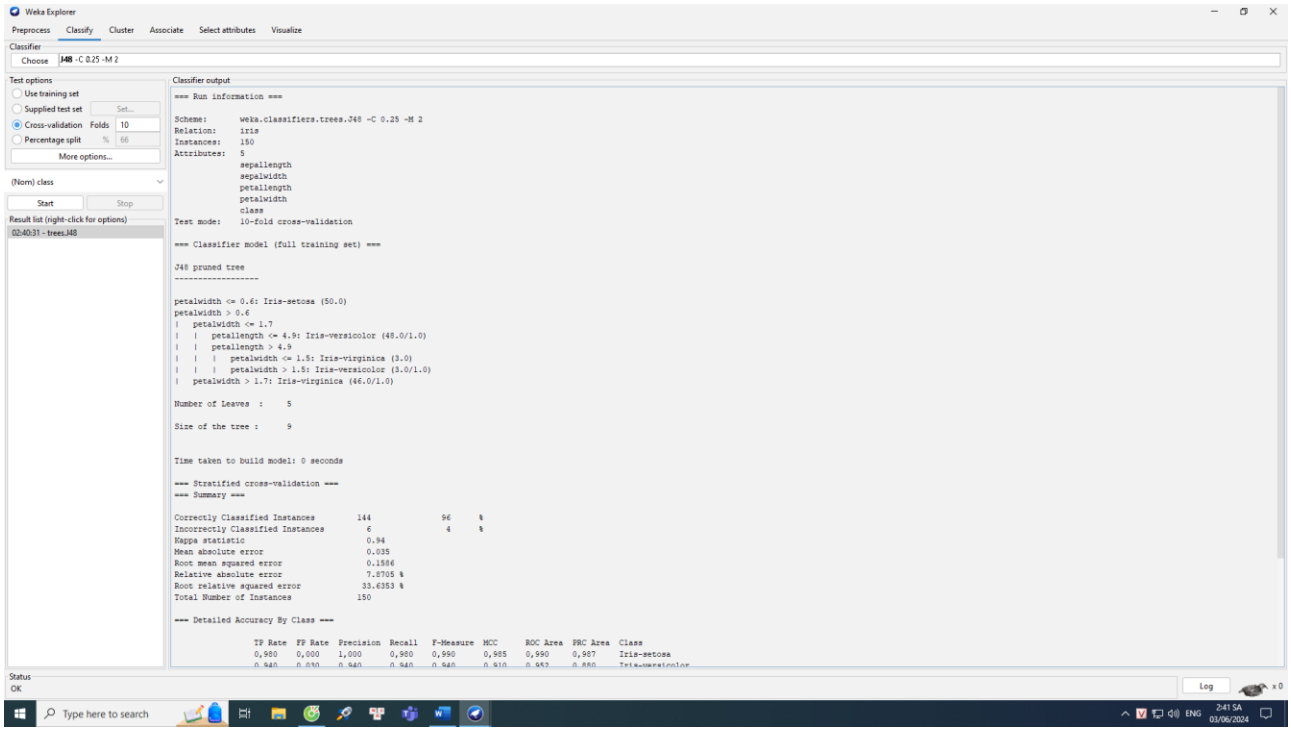
- Percentage split: Phương pháp này chia tập dữ liệu thành hai phần dựa trên tỷ lệ phần trăm được nhập vào, ví dụ là 66% như trong hình ảnh. Phần này của tập dữ liệu (66%) sẽ được sử dụng để đào tạo mô hình, và phần còn lại (34%) sẽ được dùng để kiểm thử hiệu suất.

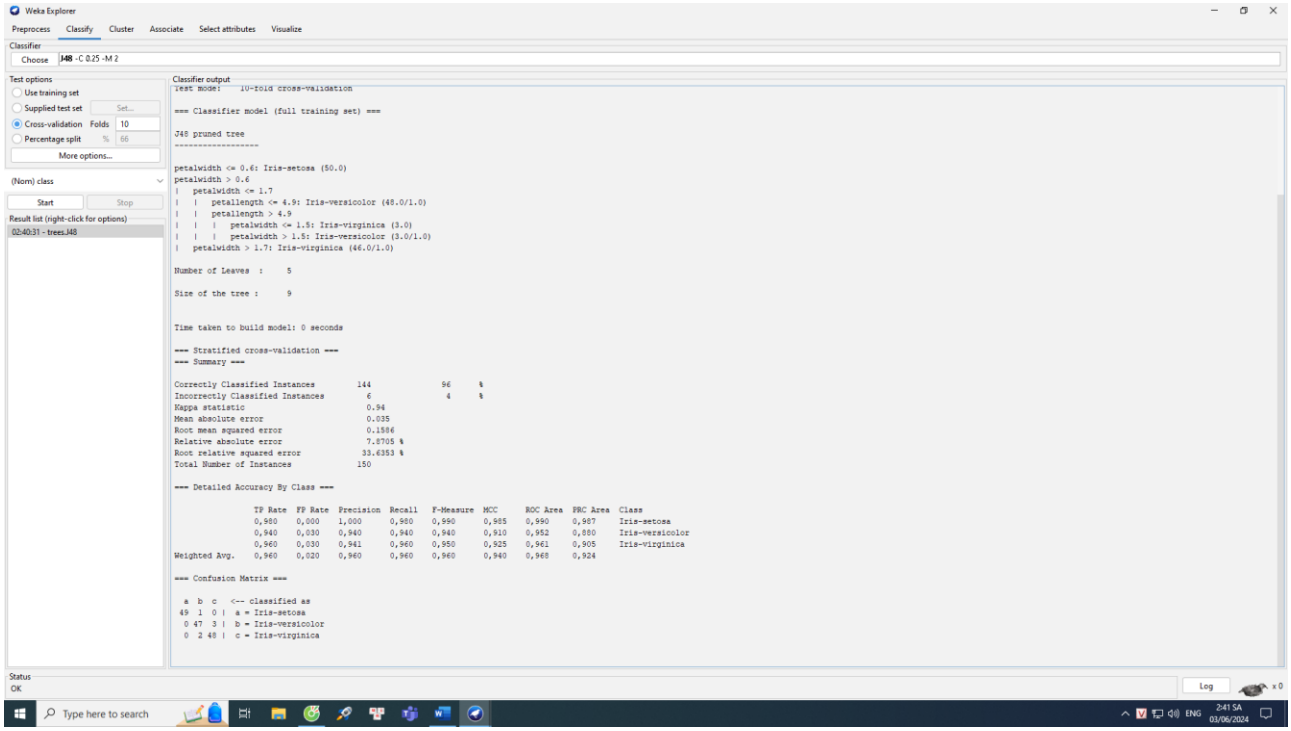
- Cross-validation: Đây chính là option mà em lựa chọn. Trong 10-fold cross-validation, tập dữ liệu được chia ngẫu nhiên thành 10 folds gần như bằng nhau về kích thước. Mỗi phần này chứa một tỷ lệ đại diện của toàn bộ tập dữ liệu. Quá trình kiểm thử diễn ra qua 10 lần lặp. Trong mỗi lần lặp, một trong số 10 folds được chọn làm tập kiểm thử, và 9 folds còn lại được dùng để huấn luyện mô hình. Sau mỗi lần lặp, mô hình được đánh giá dựa trên tập kiểm thử và hiệu

suất được ghi lại. Cuối cùng, hiệu suất của mô hình từ 10 lần lặp này được tính trung bình để đưa ra một ước tính tổng thể về hiệu suất của mô hình.

Đối với bộ phân lớp thì em chọn lựa J48, đây là một biến thể của thuật toán C4.5 và được sử dụng rộng rãi trong việc xây dựng cây quyết định.

Sau khi thực hiện chạy, em nhận được kết quả sau:





Giải thích qua kết quả mà em nhận được sẽ là như sau:

Trước tiên là với mục này:

```
=== Classifier model (full training set) ===

J48 pruned tree
------------------

petalwidth <= 0.6: Iris-setosa (50.0)
petalwidth > 0.6
|   petalwidth <= 1.7
|   |   petallength <= 4.9: Iris-versicolor (48.0/1.0)
|   |   petallength > 4.9
|   |   |   petalwidth <= 1.5: Iris-virginica (3.0)
|   |   |   petalwidth > 1.5: Iris-versicolor (3.0/1.0)
|   petalwidth > 1.7: Iris-virginica (46.0/1.0)

Number of Leaves  :     5

Size of the tree :      9


Time taken to build model: 0 seconds
```

Các quyết định:

- Nếu petalwidth ≤ 0.6, thì hoa được phân loại là Iris-setosa (có 50 mẫu đều chính xác).

- Nếu petalwidth > 0.6 và ≤ 1.7:

    - Nếu petallength ≤ 4.9, hoa là Iris-versicolor (có 48 mẫu phân đúng và 1 mẫu phân nhầm).
    - Nếu petallength > 4.9:

        - Nếu petalwidth ≤ 1.5, hoa là Iris-virginica (có 3 mẫu phân đúng).
        - Nếu petalwidth > 1.5, hoa là Iris-versicolor (có 3 mẫu phân đúng và 1 mẫu phân nhầm).

- Nếu petalwidth > 1.7, hoa là Iris-virginica (có 46 mẫu phân đúng và 1 mẫu phân nhầm).

Mô hình đã xây dựng một cây quyết định dựa trên độ rộng của petal (petalwidth) và độ dài của petal (petallength) để phân biệt giữa ba loại hoa Iris. Cây có 5 lá và tổng cộng 9 nút.

Tiếp đến là mục này:

```
=== Stratified cross-validation ===
=== Summary ===

Correctly Classified Instances         144               96      %
Incorrectly Classified Instances         6                4      %
Kappa statistic                          0.94
Mean absolute error                      0.035
Root mean squared error                  0.1586
Relative absolute error                  7.8705 %
Root relative squared error             33.6353 %
Total Number of Instances              150
```

Hiệu suất tổng thể:

- Tỉ lệ phân loại chính xác: 96% (144/150 mẫu đúng).
- Tỉ lệ phân loại sai: 4% (6/150 mẫu sai).
- Kappa statistic: 0.94, cho biết mức độ thỏa thuận cao giữa các nhãn dự đoán và nhãn thực tế, phản ánh độ chính xác cao của mô hình.

Đối với lỗi và Sai Số:

- Mean absolute error (MAE): 0.035, biểu thị sai số trung bình thấp.
- Root mean squared error (RMSE): 0.1586, cho thấy phương sai của các dự đoán, cũng khá thấp.

- Relative absolute error: 7.8705%, chỉ ra rằng sai số tuyệt đối của mô hình thấp so với một mô hình ngẫu nhiên.
- Root relative squared error: 33.6353%, phản ánh sai số bình phương gốc tương đối cũng thấp.

Tiếp theo là mục này:

```
=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 0,980    0,000    1,000      0,980    0,990      0,985    0,990     0,987     Iris-setosa
                 0,940    0,030    0,940      0,940    0,940      0,910    0,952     0,880     Iris-versicolor
                 0,960    0,030    0,941      0,960    0,950      0,925    0,961     0,905     Iris-virginica
Weighted Avg.    0,960    0,020    0,960      0,960    0,960      0,940    0,968     0,924
```

Nhìn qua thì em có thể đánh giá chi tiết theo lớp như sau:

- Iris-setosa: Có tỷ lệ phân loại chính xác cao nhất với Precision và Recall lần lượt là 100% và 98%, đạt F-Measure là 0.990.
- Iris-versicolor: Precision, Recall và F-Measure đều ở mức 0.940, cho thấy mô hình cũng phân loại khá chính xác loài này.
- Iris-virginica: Tuy có một vài trường hợp nhầm lẫn với Iris-versicolor, nhưng tỷ lệ phân loại chính xác vẫn cao với Precision là 94.1% và Recall là 96%.

Cuối cùng là mục này:

```
=== Confusion Matrix ===

  a  b  c   <-- classified as
 49  1  0 |  a = Iris-setosa
  0 47  3 |  b = Iris-versicolor
  0  2 48 |  c = Iris-virginica
```

Đây là ma trận nhầm lẫn cung cấp cái nhìn trực quan về các trường hợp được phân loại chính xác và sai.

Với kết quả mà em nhận được như trên thì có thể hiểu được là:

Hàng đầu tiên (Iris-setosa):

- 49 (a): 49 mẫu Iris-setosa được phân loại chính xác.
- 1 (b): 1 mẫu Iris-setosa bị nhầm là Iris-versicolor.
- 0 (c): Không có mẫu Iris-setosa nào bị nhầm là Iris-virginica.

Hàng thứ hai (Iris-versicolor):

- 0 (d): Không có mẫu Iris-versicolor nào bị nhầm là Iris-setosa.
- 47 (e): 47 mẫu Iris-versicolor được phân loại chính xác.
- 3 (f): 3 mẫu Iris-versicolor bị nhầm là Iris-virginica.

Hàng thứ ba (Iris-virginica):

- 0 (g): Không có mẫu Iris-virginica nào bị nhầm là Iris-setosa.
- 2 (h): 2 mẫu Iris-virginica bị nhầm là Iris-versicolor.
- 48 (i): 48 mẫu Iris-virginica được phân loại chính xác.

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 giá kết quả.

Chuẩn bị dữ liệu theo yêu cầu

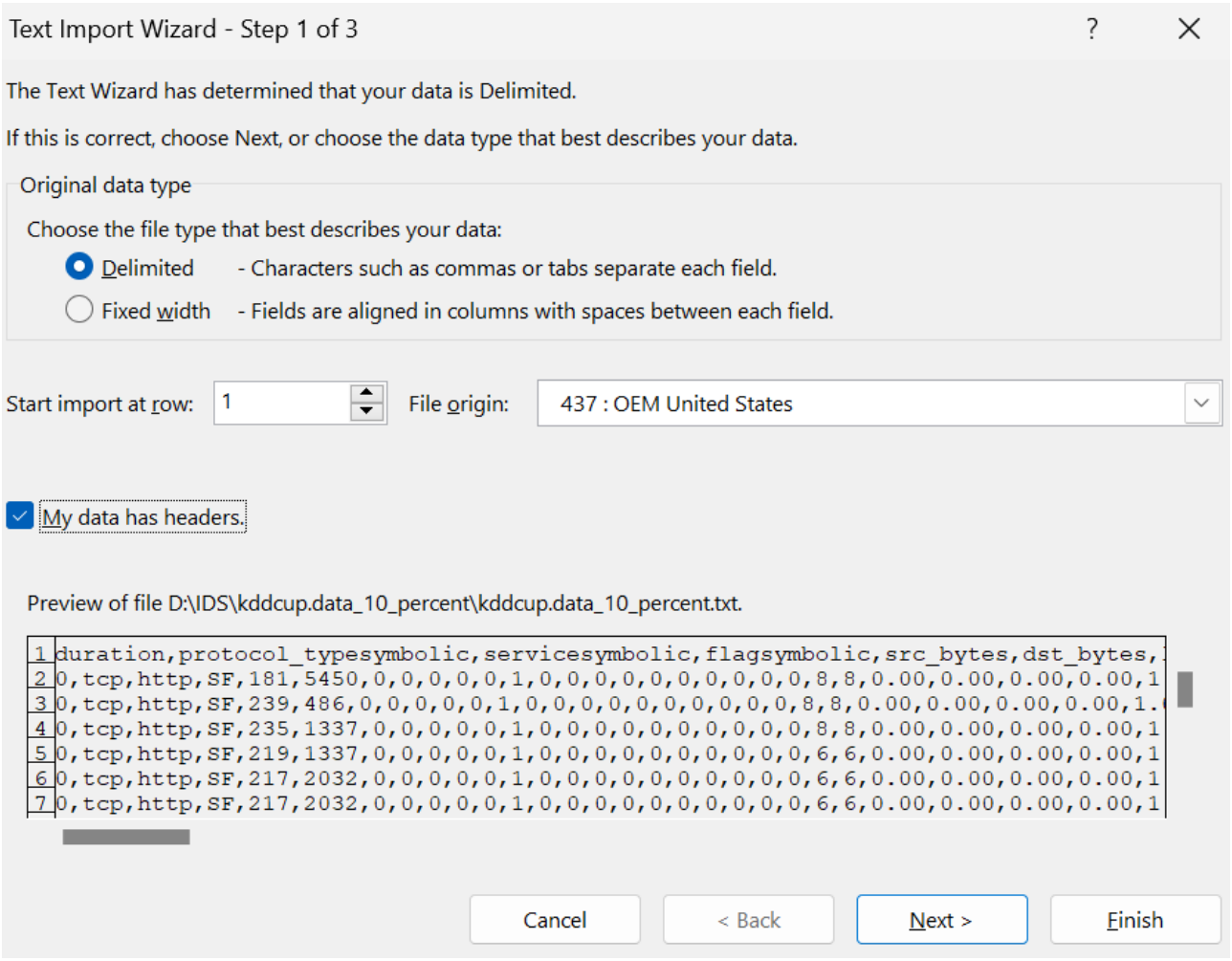
Viết 1 đoạn code python đơn giản để chuyển đổi file kddname thành file cvf.txt có kiểu như dưới

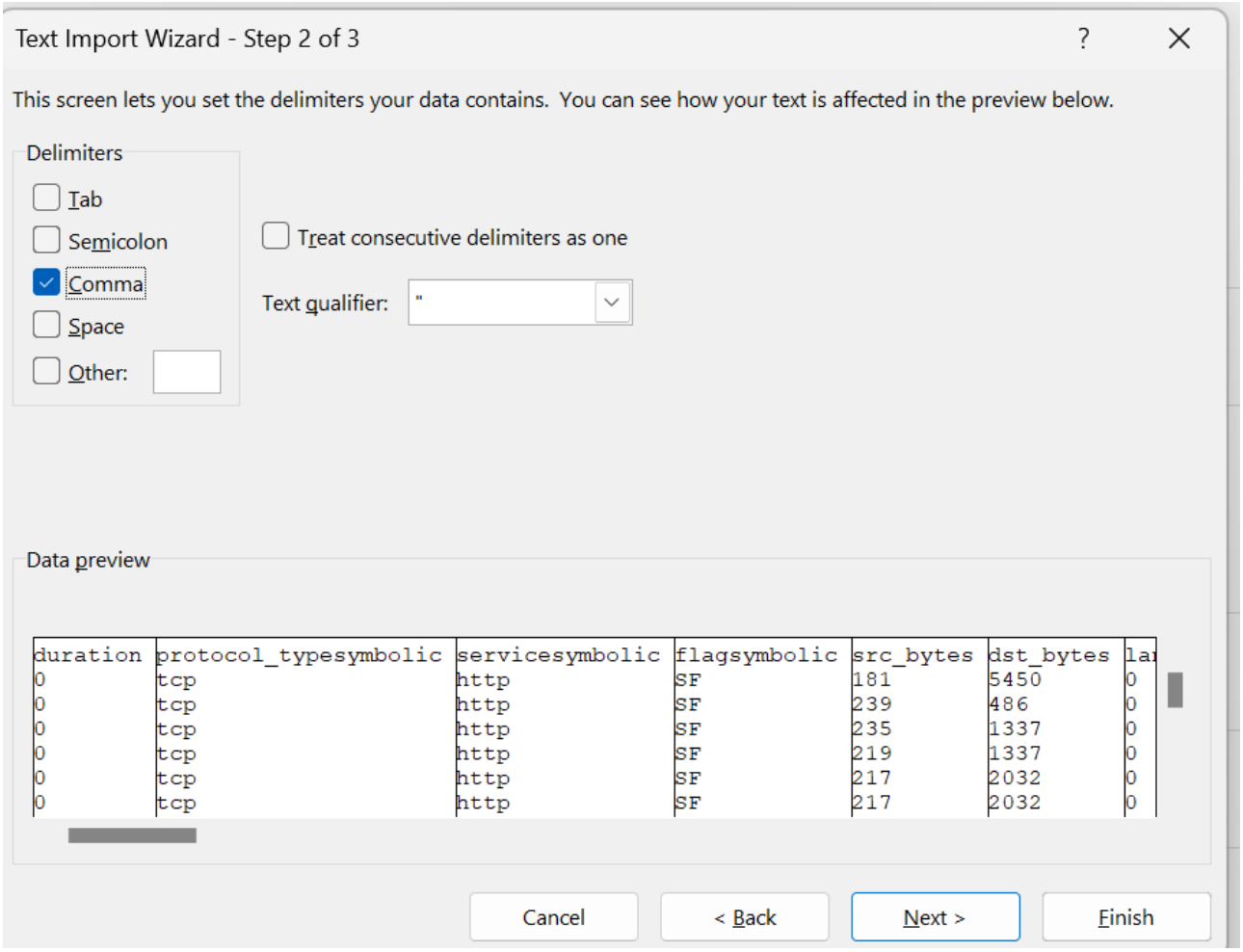
```python
# Convert.py
def convert_and_write_file(input_file, output_file):
    print(f"Đường dẫn tệp đầu vào: {input_file}")
    print(f"Đường dẫn tệp đầu ra: {output_file}")

    try:
        with open(input_file, 'r') as infile, open(output_file, 'w') as outfile:
            for line in infile:
                # Replace dots with commas, remove "continuous", colons, and whitespace
                new_line = line.replace('.', ',').replace('continuous', '').replace(':', '').replace(' ', '').strip()
                # Write the new line to the output file without newline character
                outfile.write(new_line)
    except FileNotFoundError as e:
        print(f"Error: {e}")

# Đường dẫn tới file đầu vào và file đầu ra
input_file = 'D:/IDS/kddcup.names'
output_file = 'D:/IDS/cvf.txt'

# Gọi hàm để chuyển đổi và xuất ra file mới
convert_and_write_file(input_file, output_file)
```

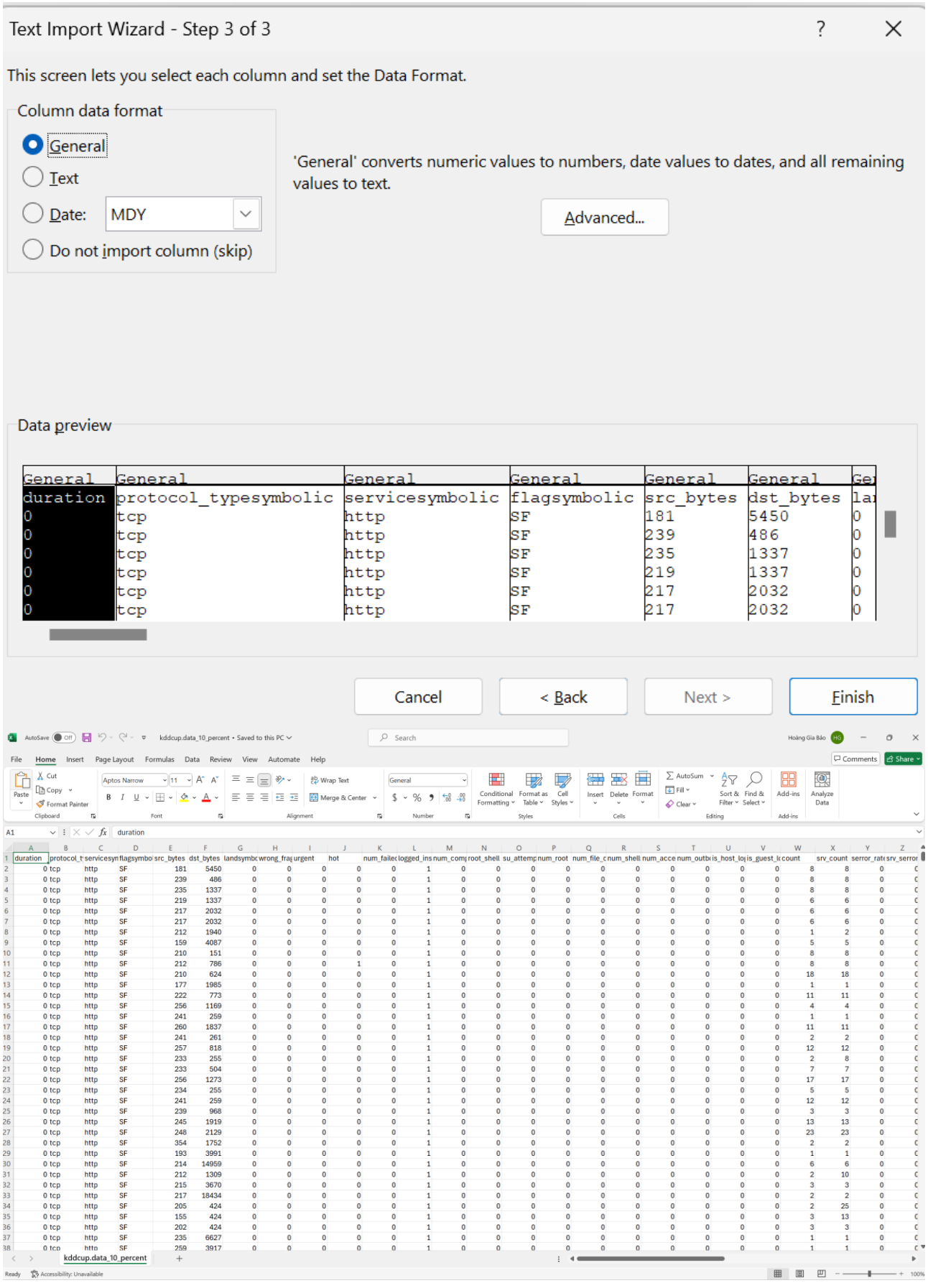
```
back,buffer_overflow,ftp_write,guess_passwd,imap,ipsweep,land,loadmodule,multihop,neptune,nmap,normal,perl,phf,pod,portsweep,rootkit,satan,smurf,spy
,teardrop,warezclient,warezmaster,duration,protocol_typesymbolic,servicesymbolic,flagsymbolic,src_bytes,dst_bytes,landsymbolic,wrong_fragment,urgent
,hot,num_failed_logins,logged_insymbolic,num_compromised,root_shell,su_attempted,num_root,num_file_creations,num_shells,num_access_files,num_outboun
d_cmds,is_host_loginsymbolic,is_guest_loginsymbolic,count,srv_count,serror_rate,srv_serror_rate,rerror_rate,srv_rerror_rate,same_srv_rate,diff_srv_r
ate,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,
```

Chèn dòng tên các thuộc tính trên vào đầu file kddcup.data_10_percent

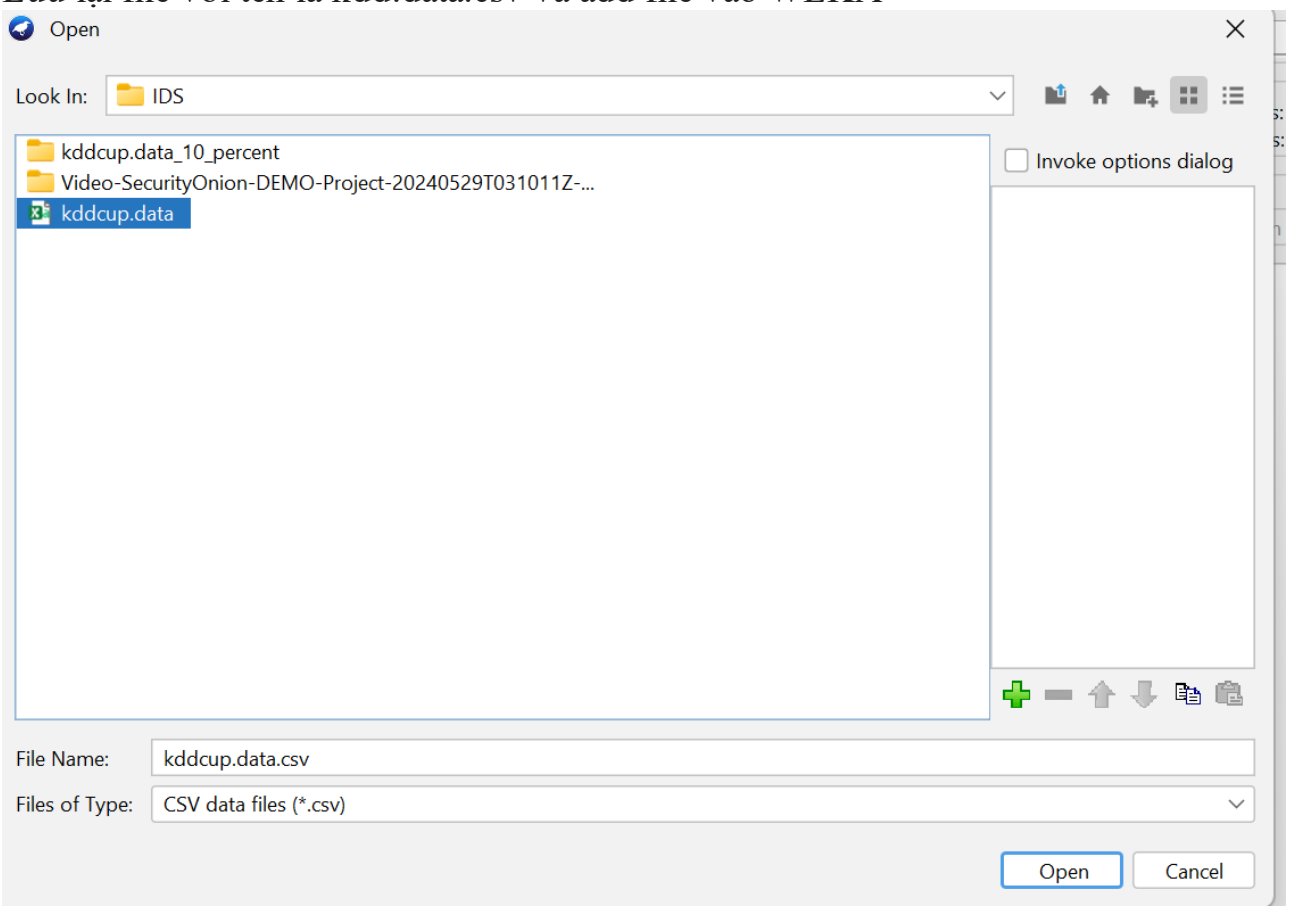
```
duration,protocol_typesymbolic,servicesymbolic,flagsymbolic,src_bytes,dst_bytes,landsymbolic,wrong_fragment,urgent,hot,num_failed_logins,logged_insymbol
ic,num_compromised,root_shell,su_attempted,num_root,num_file_creations,num_shells,num_access_files,num_outbound_cmds,is_host_loginsymbolic,is_guest_logi
nsymbolic,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_serro
r_rate,dst_host_rerror_rate,dst_host_srv_rerror_rate,result
0,tcp,http,SF,181,5450,0,0,0,0,0,1,0,0,0,0,0,0,0,0,0,0,8,8,0.00,0.00,0.00,0.00,1.00,0.00,0.00,9,9,1.00,0.00,0.11,0.00,0.00,0.00,0.00,0.00,normal.
0,tcp,http,SF,239,486,0,0,0,0,0,1,0,0,0,0,0,0,0,0,0,0,8,8,0.00,0.00,0.00,0.00,1.00,0.00,0.00,19,19,1.00,0.00,0.05,0.00,0.00,0.00,0.00,0.00,normal.
0,tcp,http,SF,235,1337,0,0,0,0,0,1,0,0,0,0,0,0,0,0,0,0,8,8,0.00,0.00,0.00,0.00,1.00,0.00,0.00,29,29,1.00,0.00,0.03,0.00,0.00,0.00,0.00,0.00,normal.
0,tcp,http,SF,219,1337,0,0,0,0,0,1,0,0,0,0,0,0,0,0,0,0,6,6,0.00,0.00,0.00,0.00,1.00,0.00,0.00,39,39,1.00,0.00,0.03,0.00,0.00,0.00,0.00,0.00,normal.
0,tcp,http,SF,217,2032,0,0,0,0,0,1,0,0,0,0,0,0,0,0,0,0,6,6,0.00,0.00,0.00,0.00,1.00,0.00,0.00,49,49,1.00,0.00,0.02,0.00,0.00,0.00,0.00,0.00,normal.
0,tcp,http,SF,217,2032,0,0,0,0,0,1,0,0,0,0,0,0,0,0,0,0,6,6,0.00,0.00,0.00,0.00,1.00,0.00,0.00,59,59,1.00,0.00,0.02,0.00,0.00,0.00,0.00,0.00,normal.
0,tcp,http,SF,212,1940,0,0,0,0,0,1,0,0,0,0,0,0,0,0,0,0,1,2,0.00,0.00,0.00,0.00,1.00,0.00,1.00,1,69,1.00,0.00,1.00,0.04,0.00,0.00,0.00,0.00,normal.
0,tcp,http,SF,159,4087,0,0,0,0,0,1,0,0,0,0,0,0,0,0,0,0,5,5,0.00,0.00,0.00,0.00,1.00,0.00,0.00,11,79,1.00,0.00,0.09,0.04,0.00,0.00,0.00,0.00,normal.
0,tcp,http,SF,210,151,0,0,0,0,0,1,0,0,0,0,0,0,0,0,0,0,8,8,0.00,0.00,0.00,0.00,1.00,0.00,0.00,8,89,1.00,0.00,0.12,0.04,0.00,0.00,0.00,0.00,normal.
0,tcp,http,SF,212,786,0,0,0,1,0,1,0,0,0,0,0,0,0,0,0,0,8,8,0.00,0.00,0.00,0.00,1.00,0.00,0.00,8,99,1.00,0.00,0.12,0.05,0.00,0.00,0.00,0.00,normal.
0,tcp,http,SF,210,624,0,0,0,0,0,1,0,0,0,0,0,0,0,0,0,0,18,18,0.00,0.00,0.00,0.00,1.00,0.00,0.00,18,109,1.00,0.00,0.06,0.05,0.00,0.00,0.00,0.00,normal.
0,tcp,http,SF,177,1985,0,0,0,0,0,1,0,0,0,0,0,0,0,0,0,0,1,1,0.00,0.00,0.00,0.00,1.00,0.00,0.00,28,119,1.00,0.00,0.04,0.04,0.00,0.00,0.00,0.00,normal.
0,tcp,http,SF,222,773,0,0,0,0,0,1,0,0,0,0,0,0,0,0,0,0,11,11,0.00,0.00,0.00,0.00,1.00,0.00,0.00,38,129,1.00,0.00,0.03,0.04,0.00,0.00,0.00,0.00,normal.
0,tcp,http,SF,256,1169,0,0,0,0,0,1,0,0,0,0,0,0,0,0,0,0,4,4,0.00,0.00,0.00,0.00,1.00,0.00,0.00,4,139,1.00,0.00,0.25,0.04,0.00,0.00,0.00,0.00,normal.
0,tcp,http,SF,241,259,0,0,0,0,0,1,0,0,0,0,0,0,0,0,0,0,1,1,0.00,0.00,0.00,0.00,1.00,0.00,0.00,14,149,1.00,0.00,0.07,0.04,0.00,0.00,0.00,0.00,normal.
0,tcp,http,SF,260,1837,0,0,0,0,0,1,0,0,0,0,0,0,0,0,0,0,11,11,0.00,0.00,0.00,0.00,1.00,0.00,0.00,24,159,1.00,0.00,0.04,0.04,0.00,0.00,0.00,0.00,normal.
0,tcp,http,SF,241,261,0,0,0,0,0,1,0,0,0,0,0,0,0,0,0,0,2,2,0.00,0.00,0.00,0.00,1.00,0.00,0.00,34,169,1.00,0.00,0.03,0.04,0.00,0.00,0.00,0.00,normal.
0,tcp,http,SF,257,818,0,0,0,0,0,1,0,0,0,0,0,0,0,0,0,0,12,12,0.00,0.00,0.00,0.00,1.00,0.00,0.00,44,179,1.00,0.00,0.02,0.03,0.00,0.00,0.00,0.00,normal.
0,tcp,http,SF,233,255,0,0,0,0,0,1,0,0,0,0,0,0,0,0,0,0,2,8,0.00,0.00,0.00,0.00,1.00,0.00,0.25,54,189,1.00,0.00,0.02,0.03,0.00,0.00,0.00,0.00,normal.
0,tcp,http,SF,233,504,0,0,0,0,0,1,0,0,0,0,0,0,0,0,0,0,7,7,0.00,0.00,0.00,0.00,1.00,0.00,0.00,64,199,1.00,0.00,0.02,0.03,0.00,0.00,0.00,0.00,normal.
```


Text Import Wizard - Step 1 of 3
The Text Wizard has determined that your data is Delimited.
If this is correct, choose Next, or choose the data type that best describes your data.
Original data type
Choose the file type that best describes your data:
Delimited - Characters such as commas or tabs separate each field.
Fixed width - Fields are aligned in columns with spaces between each field.
Start import at row: 1
File origin: 437 : OEM United States
My data has headers.
Preview of file D:\IDS\kddcup.data_10_percent\kddcup.data_10_percent.txt.
1 duration,protocol_typesymbolic,servicesymbolic,flagsymbolic,src_bytes,dst_bytes,
2 0,tcp,http,SF,181,5450,0,0,0,0,0,1,0,0,0,0,0,0,0,0,0,0,8,8,0.00,0.00,0.00,0.00,1
3 0,tcp,http,SF,239,486,0,0,0,0,0,1,0,0,0,0,0,0,0,0,0,0,8,8,0.00,0.00,0.00,0.00,1.
4 0,tcp,http,SF,235,1337,0,0,0,0,0,1,0,0,0,0,0,0,0,0,0,0,8,8,0.00,0.00,0.00,0.00,1
5 0,tcp,http,SF,219,1337,0,0,0,0,0,1,0,0,0,0,0,0,0,0,0,0,6,6,0.00,0.00,0.00,0.00,1
6 0,tcp,http,SF,217,2032,0,0,0,0,0,1,0,0,0,0,0,0,0,0,0,0,6,6,0.00,0.00,0.00,0.00,1
7 0,tcp,http,SF,217,2032,0,0,0,0,0,1,0,0,0,0,0,0,0,0,0,0,6,6,0.00,0.00,0.00,0.00,1
Cancel
< Back
Next >
Finish

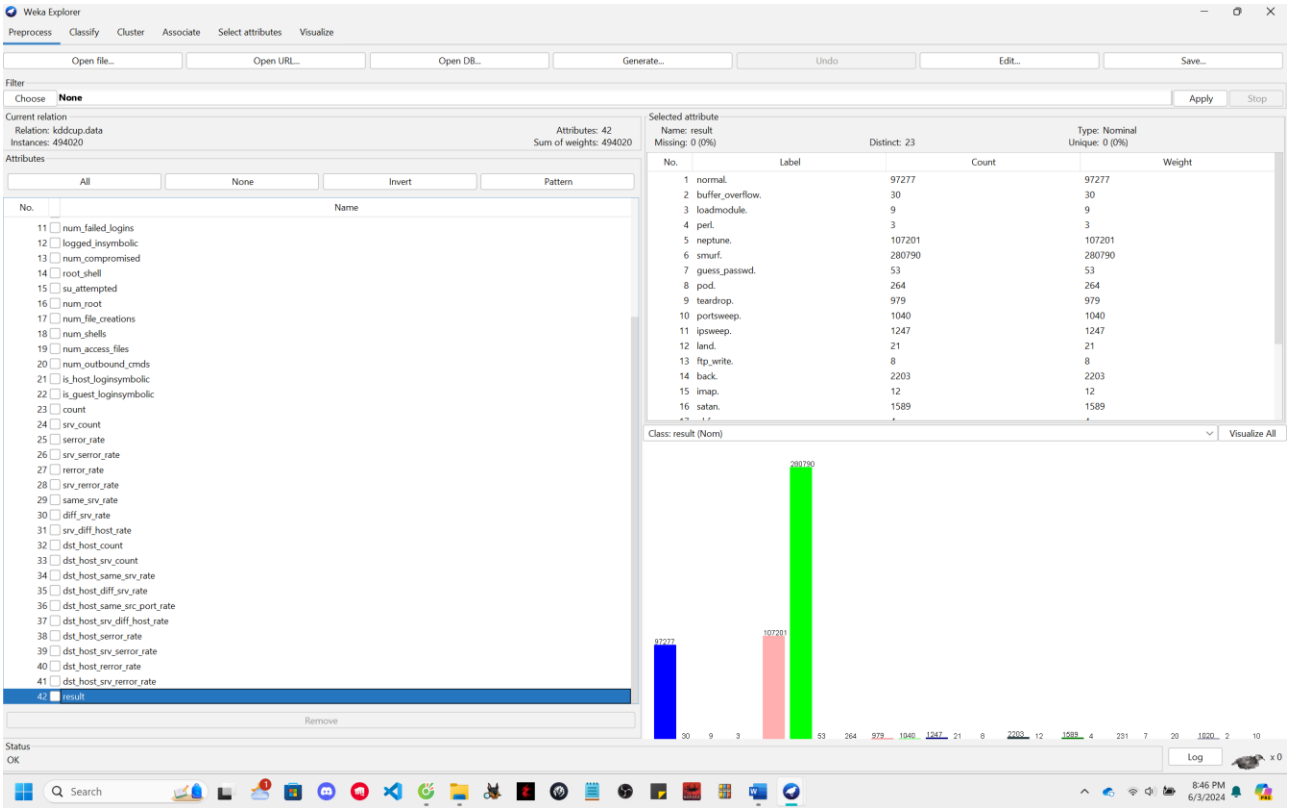

Text Import Wizard - Step 2 of 3
This screen lets you set the delimiters your data contains. You can see how your text is affected in the preview below.
Delimiters
Tab
Semicolon
Comma
Space
Other:
Treat consecutive delimiters as one
Text qualifier:
"
Data preview
duration protocol_typesymbolic servicesymbolic flagsymbolic src_bytes dst_bytes
0 tcp http SF 181 5450 0
0 tcp http SF 239 486 0
0 tcp http SF 235 1337 0
0 tcp http SF 219 1337 0
0 tcp http SF 217 2032 0
0 tcp http SF 217 2032 0
Cancel
< Back
Next >
Finish


Text Import Wizard - Step 3 of 3
This screen lets you select each column and set the Data Format.
Column data format
General
Text
Date: MDY
Do not import column (skip)
'General' converts numeric values to numbers, date values to dates, and all remaining values to text.
Advanced...
Data preview
General General General General General General
duration protocol_typesymbolic servicesymbolic flagsymbolic src_bytes dst_bytes
0 tcp http SF 181 5450 0
0 tcp http SF 239 486 0
0 tcp http SF 235 1337 0
0 tcp http SF 219 1337 0
0 tcp http SF 217 2032 0
0 tcp http SF 217 2032 0
Cancel
< Back
Next >
Finish

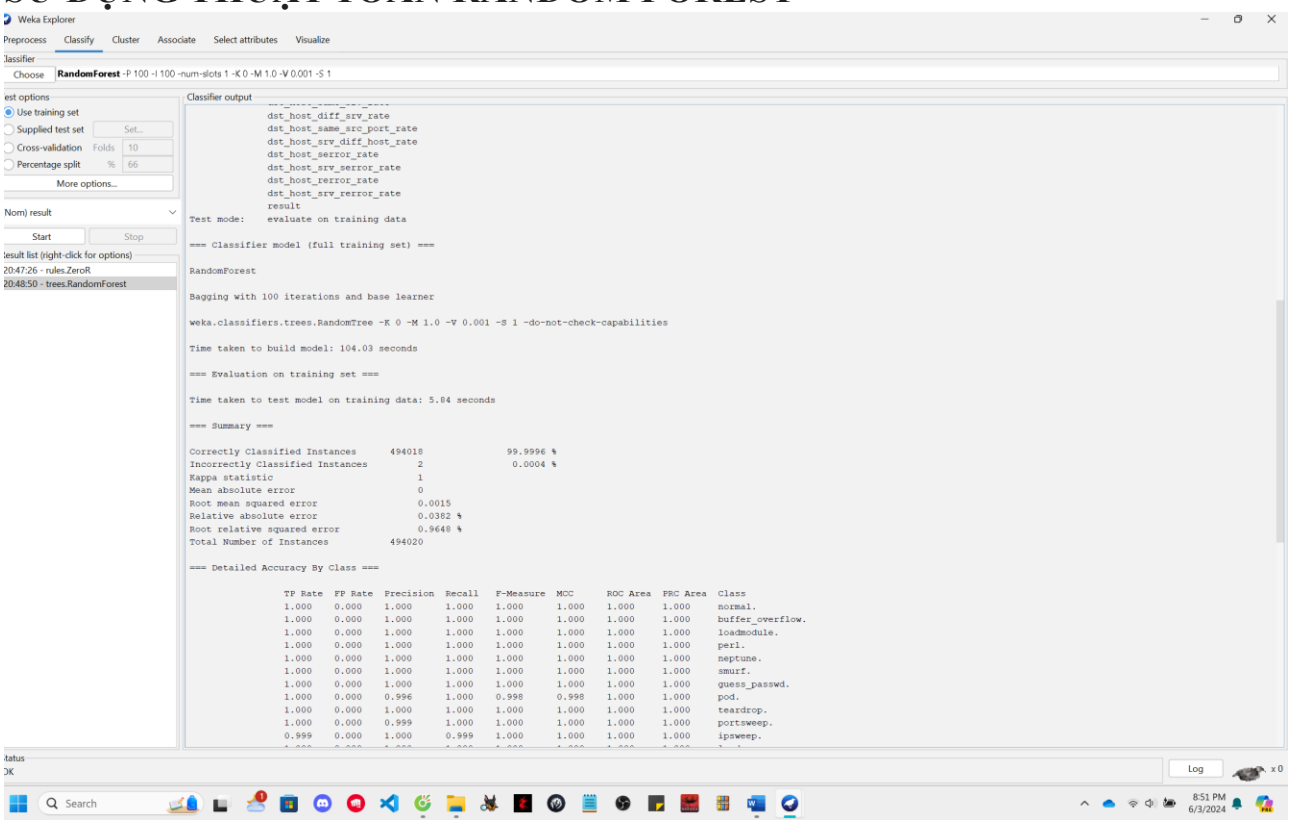
Lưu lại file với tên là kdd.data.csv và add file vào WEKA



Sau khi load



SỬ DỤNG THUẬT TOÁN RANDOM FOREST




```
Time taken to build model: 104.03 seconds

=== Evaluation on training set ===

Time taken to test model on training data: 5.84 seconds

=== Summary ===

Correctly Classified Instances      494018               99.9996 %
Incorrectly Classified Instances         2                0.0004 %
Kappa statistic                          1
Mean absolute error                      0
Root mean squared error                  0.0015
Relative absolute error                  0.0382 %
Root relative squared error              0.9648 %
Total Number of Instances           494020
```

Thời gian thực hiện

Thời gian xây dựng mô hình: 104.03 giây: Đây là thời gian cần thiết để huấn luyện mô hình trên tập dữ liệu huấn luyện.
Thời gian kiểm tra mô hình trên dữ liệu huấn luyện: 5.84 giây: Đây là thời gian cần thiết để đánh giá hiệu suất của mô hình trên cùng tập dữ liệu huấn luyện.

Các chỉ số đánh giá

Các chỉ số phân loại
Số lượng mẫu được phân loại đúng: 494,018 (99.9996%): Trong tổng số 494,020 mẫu, mô hình đã phân loại đúng 494,018 mẫu. Điều này có nghĩa là độ chính xác của mô hình trên tập huấn luyện rất cao, gần như đạt 100%.
Số lượng mẫu được phân loại sai: 2 (0.0004%): Chỉ có 2 mẫu trong tổng số 494,020 mẫu bị phân loại sai.
Các chỉ số khác
Kappa statistic: 1: Chỉ số Kappa cho biết sự thỏa thuận giữa các dự đoán của mô hình và giá trị thực tế, giá trị 1 cho thấy sự thỏa thuận hoàn hảo.
Mean absolute error (MAE - Lỗi tuyệt đối trung bình): 0: Lỗi tuyệt đối trung bình giữa dự đoán của mô hình và giá trị thực tế là 0, cho thấy mô hình dự đoán rất chính xác.
Root mean squared error (RMSE - Căn bậc hai của lỗi bình phương trung bình): 0.0015: Lỗi bình phương trung bình giữa dự đoán của mô hình và giá trị thực tế là rất nhỏ.
Relative absolute error (RAE - Lỗi tuyệt đối tương đối): 0.0382%: Tỷ lệ lỗi tuyệt đối trung bình so với tổng số lỗi tuyệt đối trung bình nếu dùng giá trị trung bình của dữ liệu là rất nhỏ.
Root relative squared error (RRSE - Căn bậc hai của lỗi bình phương tương đối): 0.9648%: Tỷ lệ lỗi bình phương trung bình so với tổng số lỗi bình phương trung bình nếu dùng giá trị trung bình của dữ liệu cũng rất nhỏ.
Tổng số mẫu: 494,020

=>Mô hình này có hiệu suất rất cao trên tập dữ liệu huấn luyện với độ chính xác gần như tuyệt đối và các chỉ số lỗi rất nhỏ.

Classifier output

```
=== Detailed Accuracy By Class ===

               TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
               1.000    0.000    1.000      1.000    1.000      1.000    1.000     1.000     normal.
               1.000    0.000    1.000      1.000    1.000      1.000    1.000     1.000     buffer_overflow.
               1.000    0.000    1.000      1.000    1.000      1.000    1.000     1.000     loadmodule.
               1.000    0.000    1.000      1.000    1.000      1.000    1.000     1.000     perl.
               1.000    0.000    1.000      1.000    1.000      1.000    1.000     1.000     neptune.
               1.000    0.000    1.000      1.000    1.000      1.000    1.000     1.000     smurf.
               1.000    0.000    1.000      1.000    1.000      1.000    1.000     1.000     guess_passwd.
               1.000    0.000    0.996      1.000    0.998      0.998    1.000     1.000     pod.
               1.000    0.000    1.000      1.000    1.000      1.000    1.000     1.000     teardrop.
               1.000    0.000    0.999      1.000    1.000      1.000    1.000     1.000     portsweep.
               0.999    0.000    1.000      0.999    1.000      1.000    1.000     1.000     ipsweep.
               1.000    0.000    1.000      1.000    1.000      1.000    1.000     1.000     land.
               1.000    0.000    1.000      1.000    1.000      1.000    1.000     1.000     ftp_write.
               1.000    0.000    1.000      1.000    1.000      1.000    1.000     1.000     back.
               1.000    0.000    1.000      1.000    1.000      1.000    1.000     1.000     imap.
               1.000    0.000    1.000      1.000    1.000      1.000    1.000     1.000     satan.
               1.000    0.000    1.000      1.000    1.000      1.000    1.000     1.000     phf.
               1.000    0.000    1.000      1.000    1.000      1.000    1.000     1.000     nmap.
               1.000    0.000    1.000      1.000    1.000      1.000    1.000     1.000     multihop.
               1.000    0.000    1.000      1.000    1.000      1.000    1.000     1.000     warezmaster.
               1.000    0.000    1.000      1.000    1.000      1.000    1.000     1.000     warezclient.
               1.000    0.000    1.000      1.000    1.000      1.000    1.000     1.000     spy.
               1.000    0.000    1.000      1.000    1.000      1.000    1.000     1.000     rootkit.
Weighted Avg.  1.000    0.000    1.000      1.000    1.000      1.000    1.000     1.000

=== Confusion Matrix ===

     a      b     c     d     e      f     g     h     i     j     k     l     m     n     o     p     q     r     s     t     u     v     w   <-- classified as
 97276      0     0     0     0      0     0     1     0     0     0     0     0     0     0     0     0     0     0     0     0     0     0 |     a = normal.
     0     30     0     0     0      0     0     0     0     0     0     0     0     0     0     0     0     0     0     0     0     0     0 |     b = buffer_overflow.
     0      0     9     0     0      0     0     0     0     0     0     0     0     0     0     0     0     0     0     0     0     0     0 |     c = loadmodule.
     0      0     0     3     0      0     0     0     0     0     0     0     0     0     0     0     0     0     0     0     0     0     0 |     d = perl.
     0      0     0     0 107201      0     0     0     0     0     0     0     0     0     0     0     0     0     0     0     0     0     0 |     e = neptune.
     0      0     0     0     0 280790     0     0     0     0     0     0     0     0     0     0     0     0     0     0     0     0     0 |     f = smurf.
     0      0     0     0     0      0    53     0     0     0     0     0     0     0     0     0     0     0     0     0     0     0     0 |     g = guess_passwd.
     0      0     0     0     0      0     0   264     0     0     0     0     0     0     0     0     0     0     0     0     0     0     0 |     h = pod.
     0      0     0     0     0      0     0     0   979     0     0     0     0     0     0     0     0     0     0     0     0     0     0 |     i = teardrop.
     0      0     0     0     0      0     0     0     0  1040     0     0     0     0     0     0     0     0     0     0     0     0     0 |     j = portsweep.
     0      0     0     0     0      0     0     0     0     1  1246     0     0     0     0     0     0     0     0     0     0     0     0 |     k = ipsweep.
     0      0     0     0     0      0     0     0     0     0     0    21     0     0     0     0     0     0     0     0     0     0     0 |     l = land.
     0      0     0     0     0      0     0     0     0     0     0     0     8     0     0     0     0     0     0     0     0     0     0 |     m = ftp_write.
     0      0     0     0     0      0     0     0     0     0     0     0     0  2203     0     0     0     0     0     0     0     0     0 |     n = back.
     0      0     0     0     0      0     0     0     0     0     0     0     0     0    12     0     0     0     0     0     0     0     0 |     o = imap.
     0      0     0     0     0      0     0     0     0     0     0     0     0     0     0  1589     0     0     0     0     0     0     0 |     p = satan.
     0      0     0     0     0      0     0     0     0     0     0     0     0     0     0     0     4     0     0     0     0     0     0 |     q = phf.
     0      0     0     0     0      0     0     0     0     0     0     0     0     0     0     0     0   231     0     0     0     0     0 |     r = nmap.
     0      0     0     0     0      0     0     0     0     0     0     0     0     0     0     0     0     0     7     0     0     0     0 |     s = multihop.
```

Mô hình có hiệu suất rất cao với các chỉ số Precision, Recall, và F-Measure đều đạt 1.000 cho hầu hết các lớp. Chỉ số TP Rate và FP Rate cho thấy mô hình phân loại chính xác các mẫu gần như hoàn toàn, với số lượng lỗi rất nhỏ. Ma trận lỗi cũng cho thấy số lượng mẫu bị phân loại nhầm là rất ít.