

LAPORAN RESMI
PRAKTIKUM KEAMANAN JARINGAN
DATA MINING FOR NETWORK SECURITY



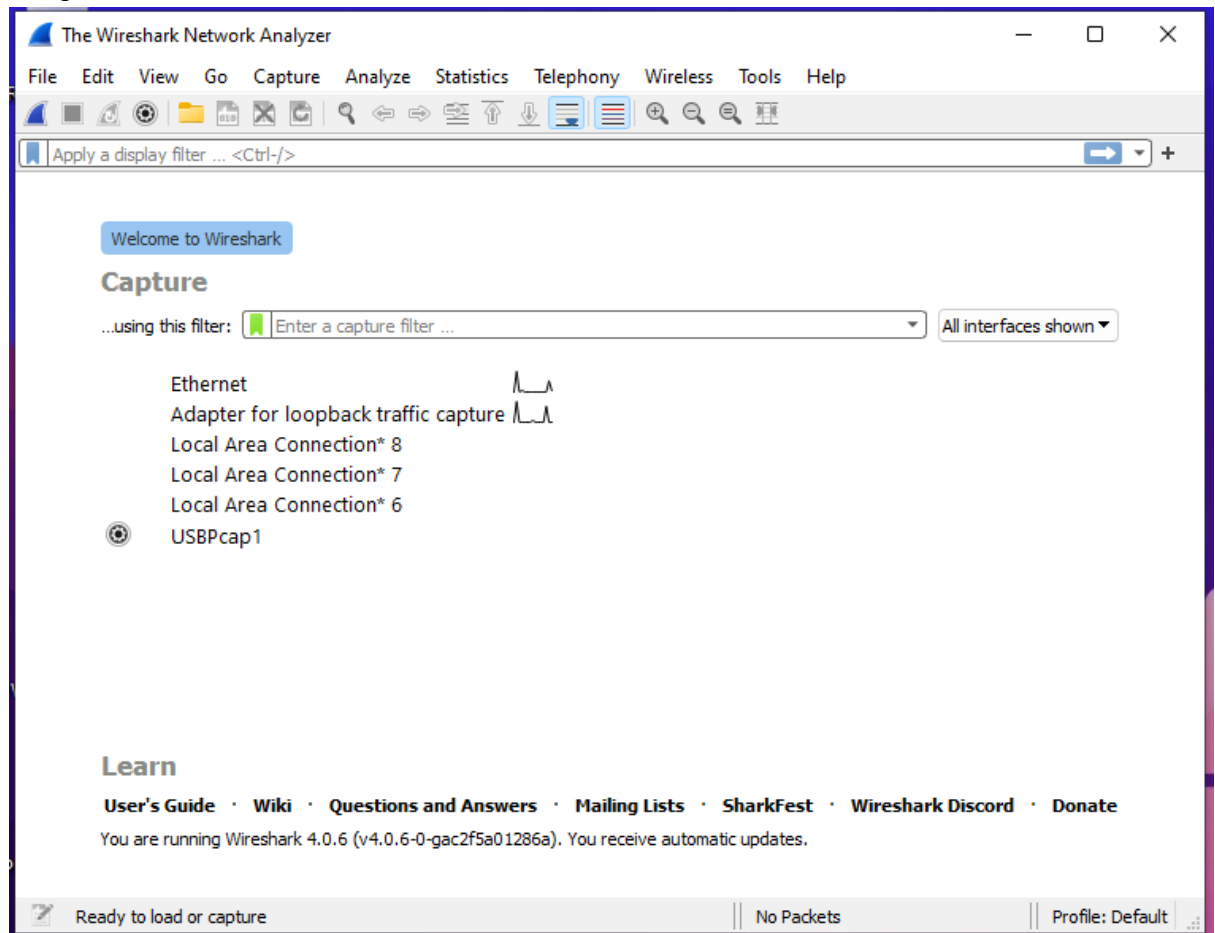
Oleh :

Fisabili Maghfirona Firdaus 3122640051
D4 LJ Teknik Informatika B

POLITEKNIK ELEKTRONIKA NEGERI SURABAYA
TAHUN AJARAN 2022/2023

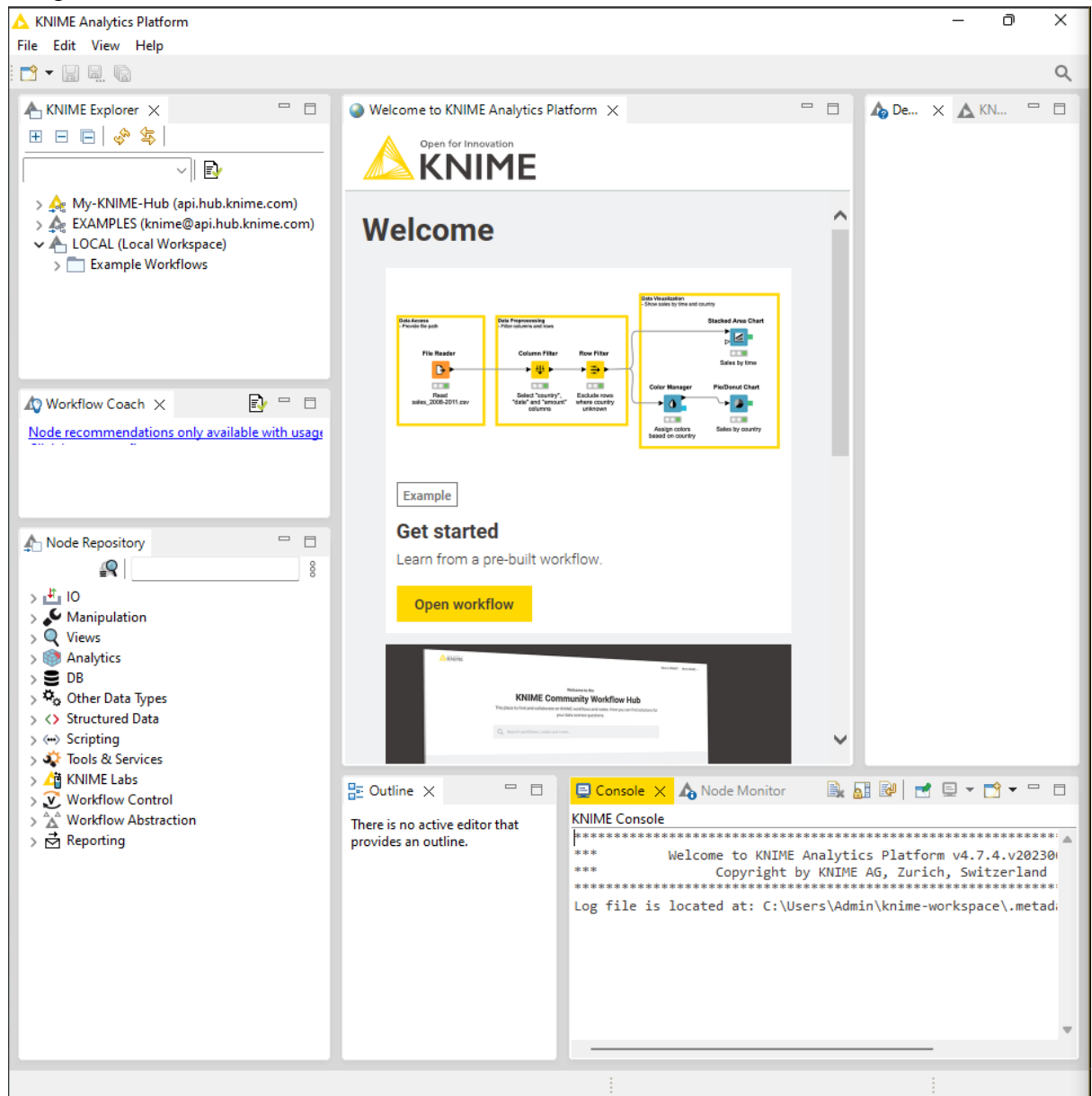
A. INSTALL WIRESHARK

1. Download software Wireshark pada website resmi mereka dan pilih OS yang sesuai.
2. Install dengan pengaturan default apabila telah memahami langkah install
3. Tampilan selesai install








B. INSTALL KNIME

1. Download software KNIME pada website resmi mereka dan pilih OS yang sesuai.
2. Install dengan pengaturan default apabila telah memahami langkah install
3. Tampilan selesai install

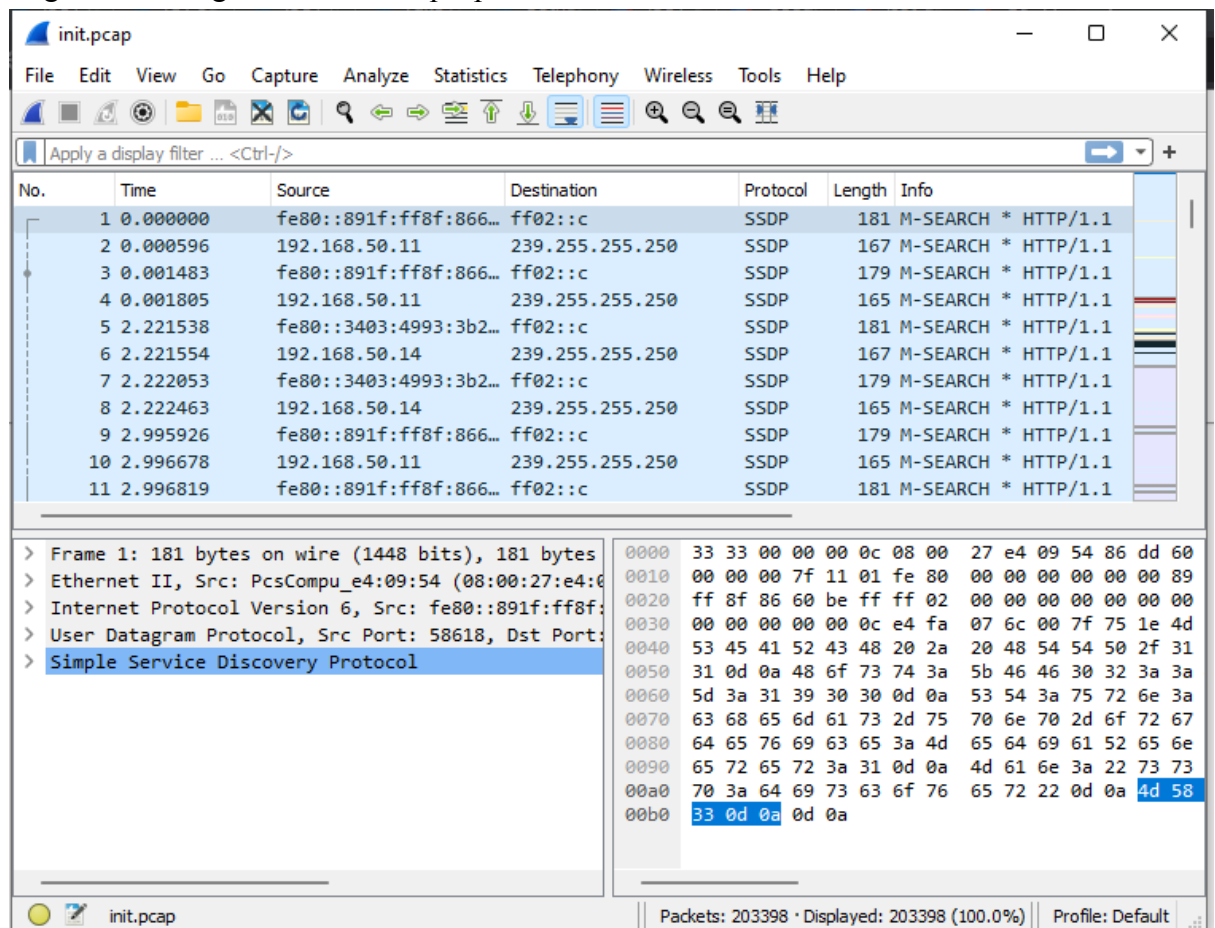


WIRESHARK

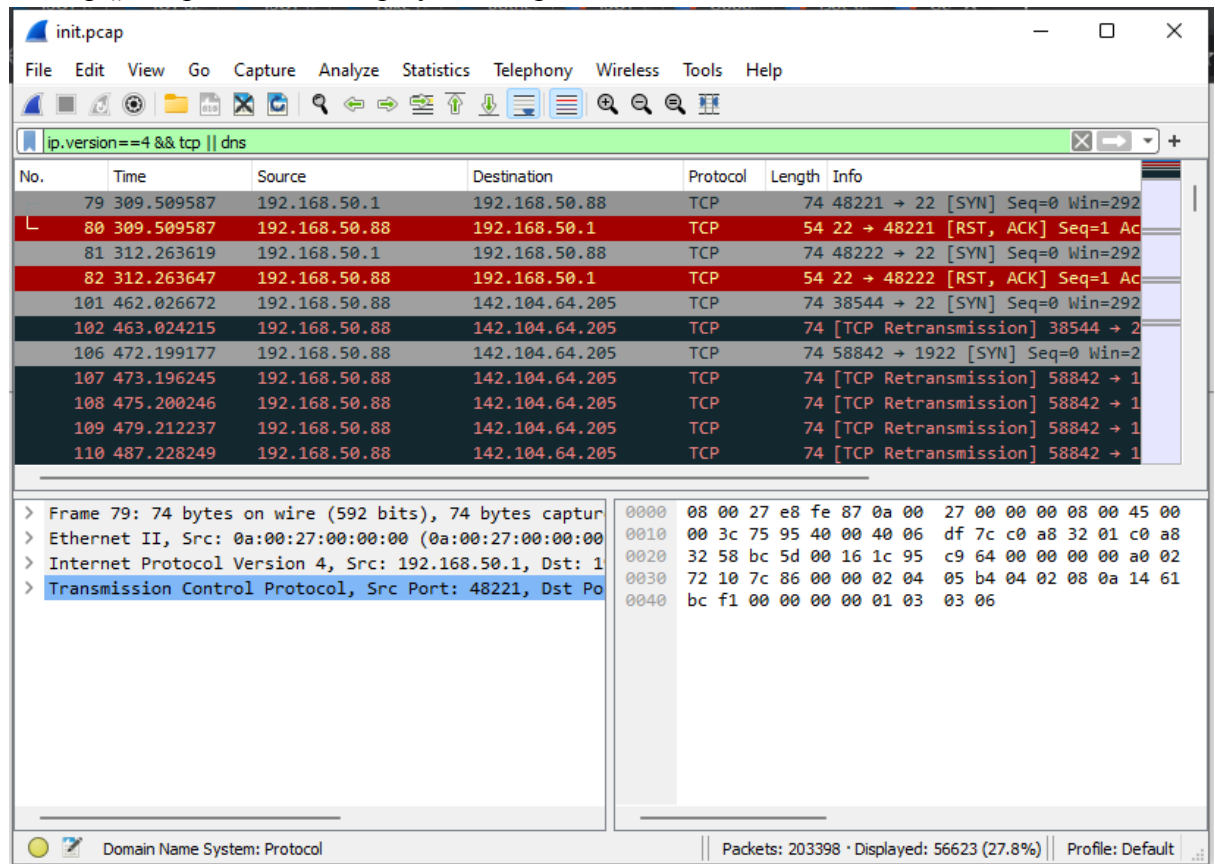
1. Proses ini digunakan untuk mengamati file Packet Capture (.pcap). File tersebut berisi lalu lintas jaringan yang ditangkap oleh komputer. Pada kasus kali ini akan digunakan dataset traffic DNS ISOT yang berasal dari University of Victoria karena terdapat simulasi serangan Botnet pada traffic DNS. Untuk memperoleh dataset ini dapat mengunjungi link: <https://www.uvic.ca/engineering/ece/isot/datasets/> File tersebut dibagi menjadi 5 yaitu : init.pcap, init2.pcap, init3.pcap, init4.pcap, init5.pcap

Name	Size
 init.pcap	163,160 KB
 init2.pcap	576,966 KB
 init3.pcap	159,862 KB
 init4.pcap	656,556 KB
 init5.pcap	2 KB

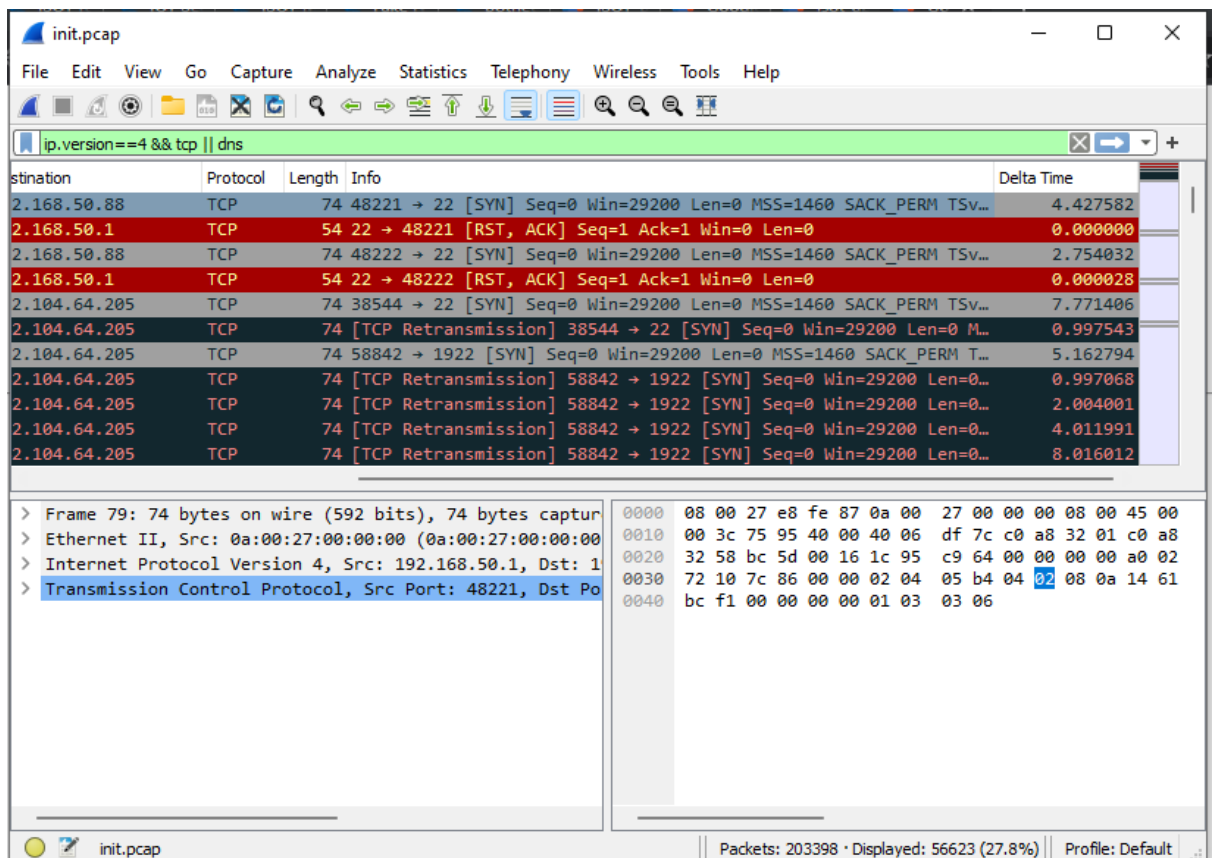
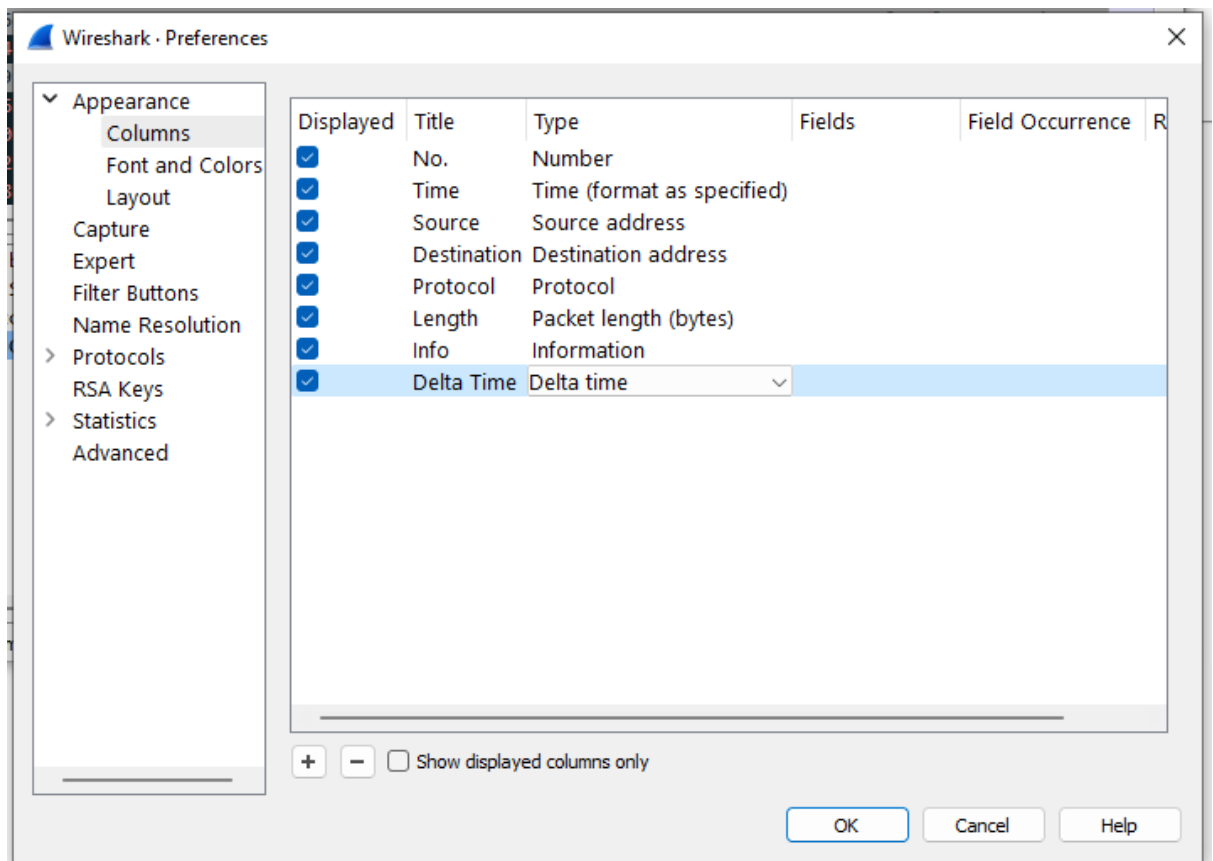
2. Kemudian buka file tersebut secara bergantian menggunakan Wireshark. Pada langkah ini kita gunakan file init.pcap



- Untuk mempermudah pada saat proses analisa yang akan dilakukan nantinya, kita akan mengambil data dengan ip versi 4 (ipv4) dan protocol TCP, DNS saja. Untuk proses tersebut dapat dilakukan pada wireshark menggunakan perintah `ip.version==4 && tcp || dns` pada kolom display filter tepat dibawah toolbar.



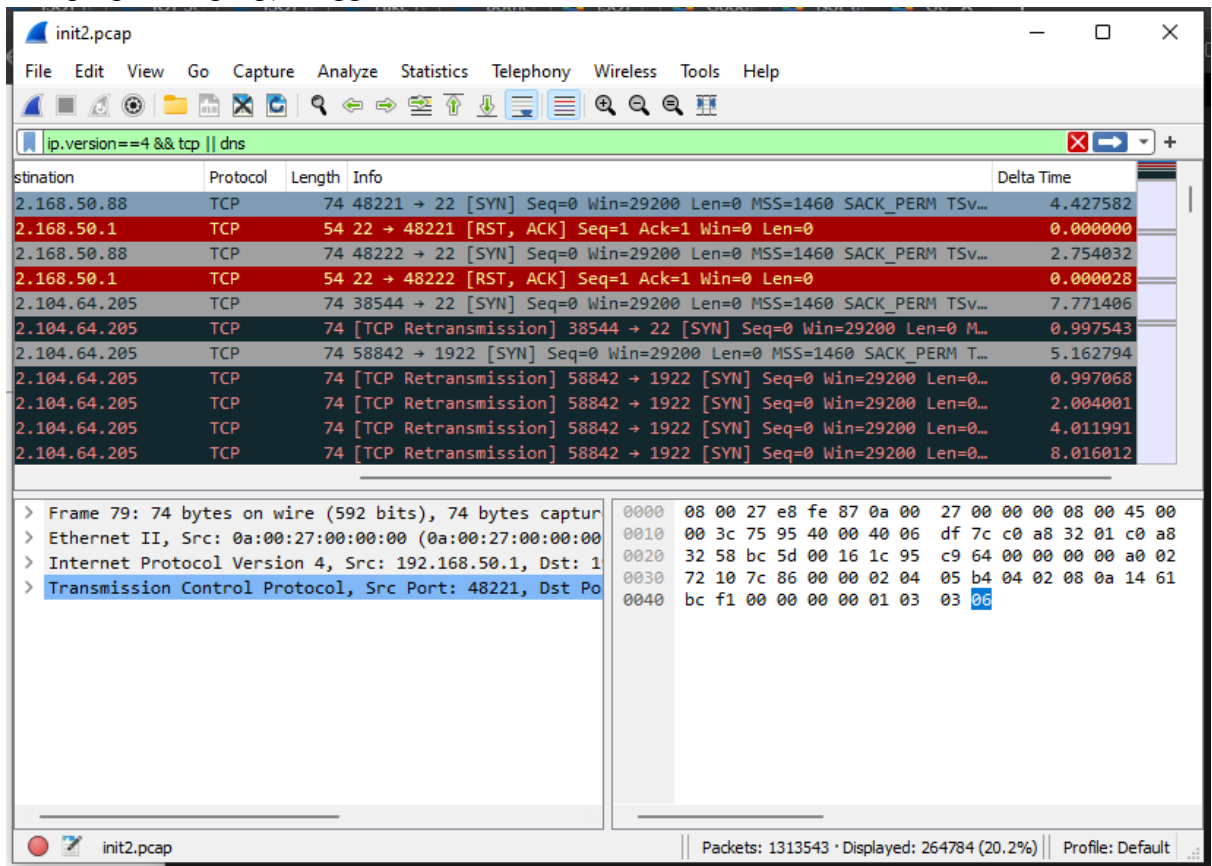
- Kemudian kita membutuhkan kolom tambahan yaitu delta time. Untuk mendapatkan delta time dan delta time dan delta time display, klik Edit – Preferences – Column
- Kemudian klik pada tanda + untuk menambah kolom baru. Kemudian pada Type, pilih Delta Time. Kemudian lakukan hal yang sama untuk kolom delta time display. Kemudian Klik OK. Berikut hasilnya.



- Langkah terakhir yaitu export file pcap tersebut keformat Comma-separated Value (.csv) dengan cara klik File – Export Packet Dissections – As CSV. Yang perlu

diperhatikan yaitu pada Pacet Range, pastikan yang terpilih yaitu Displayed, karena data pada Displayed ini sudah terfilter dengan nip version 4.

7. Lakukan semua proses diatas pada dataset berikutnya (init2.pcap, init3.pcap, init4.pcap, init5.pcap) hingga seluruh data sudah terkonversi ke dalam format .csv



init3.pcap

File Edit View Go Capture Analyze Statistics Telephony Wireless Tools Help

ip.version==4 && tcp || dns

Destination	Protocol	Length	Info	Delta Time
192.168.50.88	DNS	83	Standard query 0xed53 A ns1.random.dns.d0wn.biz	0.456482
192.168.50.31	DNS	310	Standard query response 0xed53 A ns1.random.dns.d0wn.biz A 17...	0.000252
192.168.50.88	DNS	83	Standard query 0xe364 A ns2.random.dns.d0wn.biz	0.028779
192.168.50.31	DNS	310	Standard query response 0xe364 A ns2.random.dns.d0wn.biz A 18...	0.004263
192.168.50.88	DNS	76	Standard query 0x3591 A zeus.botnet.isot	0.443257
192.168.50.34	DNS	160	Standard query response 0x3591 A zeus.botnet.isot A 192.168.5...	0.000257
192.168.50.88	DNS	76	Standard query 0xc56e A blue.botnet.isot	0.000161
192.168.50.15	DNS	160	Standard query response 0xc56e A blue.botnet.isot A 192.168.5...	0.000214
192.168.50.88	DNS	79	Standard query 0x821c A citadel.botnet.isot	0.401343
192.168.50.30	DNS	163	Standard query response 0x821c A citadel.botnet.isot A 192.16...	0.000236
192.168.50.88	DNS	80	Standard query 0x06d1 A anyone.dnsrec.meo.ws	0.221481

> Frame 2: 83 bytes on wire (664 bits), 83 bytes captured
 > Ethernet II, Src: PcsCompu_bc:f8:e2 (08:00:27:bc:f8:e2)
 > Internet Protocol Version 4, Src: 192.168.50.31, Dst: 1
 > User Datagram Protocol, Src Port: 54269, Dst Port: 53
 > Domain Name System (query)

```

0000 08 00 27 e8 fe 87 08 00 27 bc f8 e2 08 00 45 00
0010 00 45 07 99 00 00 80 11 4d 47 c0 a8 32 1f c0 a8
0020 32 58 d3 fd 00 35 00 31 85 5b ed 53 01 00 00 01
0030 00 00 00 00 00 00 03 6e 73 31 06 72 61 6e 64 6f
0040 6d 03 64 6e 73 04 64 30 77 6e 03 62 69 7a 00 00
0050 01 00 01
  
```

Domain Name System: Protocol | Packets: 1263913 · Displayed: 913944 (72.3%) | Profile: Default

init4.pcap

File Edit View Go Capture Analyze Statistics Telephony Wireless Tools Help

ip.version==4 && tcp || dns

No.	Time	Source	Destination	Protocol	Length	Info
1	0.000000	192.168.50.88	192.168.50.34	DNS	73	Standard query response 0x8cd6 S
3	0.690849	192.168.50.17	192.168.50.88	DNS	83	Standard query 0x3bb0 A ns2.rand
5	0.944173	192.168.50.88	8.8.8.8	DNS	94	Standard query 0x0980 A ns2.rand
6	0.944248	192.168.50.88	8.8.8.8	DNS	70	Standard query 0xbd7e NS <Root>
11	3.831927	192.168.50.31	192.168.50.88	DNS	79	Standard query 0xc0f3 A citadel.
12	3.832168	192.168.50.88	192.168.50.31	DNS	163	Standard query response 0xc0f3 A
14	4.690303	192.168.50.17	192.168.50.88	DNS	83	Standard query 0x3bb0 A ns2.rand
16	4.706928	192.168.50.31	192.168.50.88	DNS	89	Standard query 0xd890 A alors.de
17	4.707297	192.168.50.88	8.8.8.8	DNS	100	Standard query 0x8ec1 A alors.de
22	5.194788	192.168.50.88	8.8.4.4	DNS	94	Standard query 0x113c A ns2.rand
23	5.194924	192.168.50.88	8.8.4.4	DNS	70	Standard query 0xa04f NS <Root>

> Frame 1: 73 bytes on wire (584 bits), 73 bytes captured
 > Ethernet II, Src: PcsCompu_e8:fe:87 (08:00:27:e8:fe:87)
 > Internet Protocol Version 4, Src: 192.168.50.88, Dst: 1
 > User Datagram Protocol, Src Port: 53, Dst Port: 65335
 > Domain Name System (response)

```

0000 08 00 27 ee d4 3c 08 00 27 e8 fe 87 08 00 45 00
0010 00 3b 03 c3 00 00 40 11 91 24 c0 a8 32 58 c0 a8
0020 32 22 00 35 ff 37 00 27 e6 03 8c d6 81 82 00 01
0030 00 00 00 00 00 00 03 77 77 77 06 67 6f 6f 67 6c
0040 65 02 63 61 00 00 01 00 01
  
```

init4.pcap | Packets: 7266524 · Displayed: 2952076 (40.6%) | Profile: Default

init5.pcap

File Edit View Go Capture Analyze Statistics Telephony Wireless Tools Help

ip.version==4 && tcp || dns

No.	Time	Source	Destination	Protocol	Length	Info
1	0.000000	192.168.50.50	192.168.50.88	DNS	79	Standard query 0x5e82 A clients2.googl
4	0.802921	192.168.50.19	192.168.50.88	DNS	81	Standard query 0xf4f4 A client-cf.drop
9	1.848537	192.168.50.88	8.8.4.4	DNS	97	Standard query 0xfa0b A updatekeepaliv
11	2.096566	192.168.50.51	192.168.50.88	DNS	84	Standard query 0x4fe5 A www.google-ana
15	2.848113	192.168.50.88	192.168.50.51	DNS	86	Standard query response 0x4234 Server

> Frame 1: 79 bytes on wire (632 bits), 79 bytes captured
 > Ethernet II, Src: PcsCompu_96:32:51 (08:00:27:96:32:51)
 > Internet Protocol Version 4, Src: 192.168.50.50, Dst: 192.168.50.88
 > User Datagram Protocol, Src Port: 56905, Dst Port: 53
 > Domain Name System (query)

0000 08 00 27 e8 fe 87 08 00 27 96 32 51 08 00 45 00
 0010 00 41 07 80 00 00 80 11 4d 51 c0 a8 32 32 c0 a8
 0020 32 58 de 49 00 35 00 2d a2 f5 5e 82 01 00 00 01
 0030 00 00 00 00 00 00 08 63 6c 69 65 6e 74 73 32 06
 0040 67 6f 6f 67 6c 65 03 63 6f 6d 00 00 01 00 01

init5.pcap | Packets: 15 · Displayed: 5 (33.3%) | Profile: Default

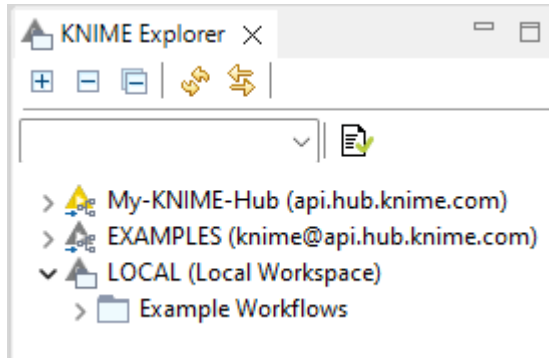
init.csv	9,099 KB	CSV File
init2.csv	46,326 KB	CSV File
init3.csv	159,343 KB	CSV File
init4.csv	385,712 KB	CSV File
init5.csv	1 KB	CSV File

KNIME

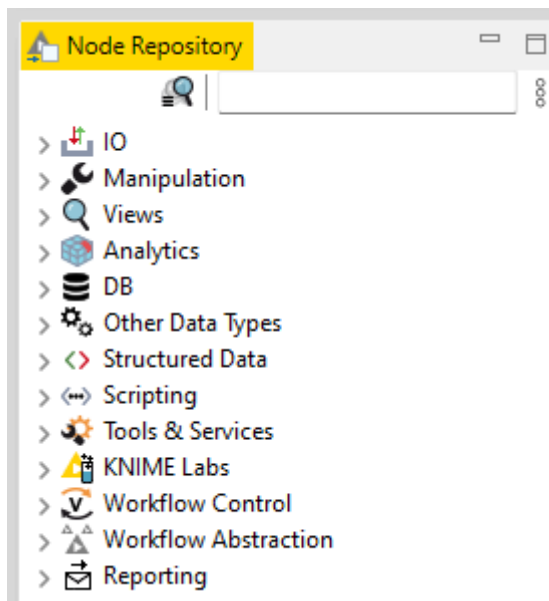
A. Penggabungan Data

1. Setelah semua file tadi telah diexport menjadi file .csv. Buka software Knime Analytics Platform untuk melakukan proses analisa pada traffic DNS. Berikut adalah file yang telah terexport menjadi csv

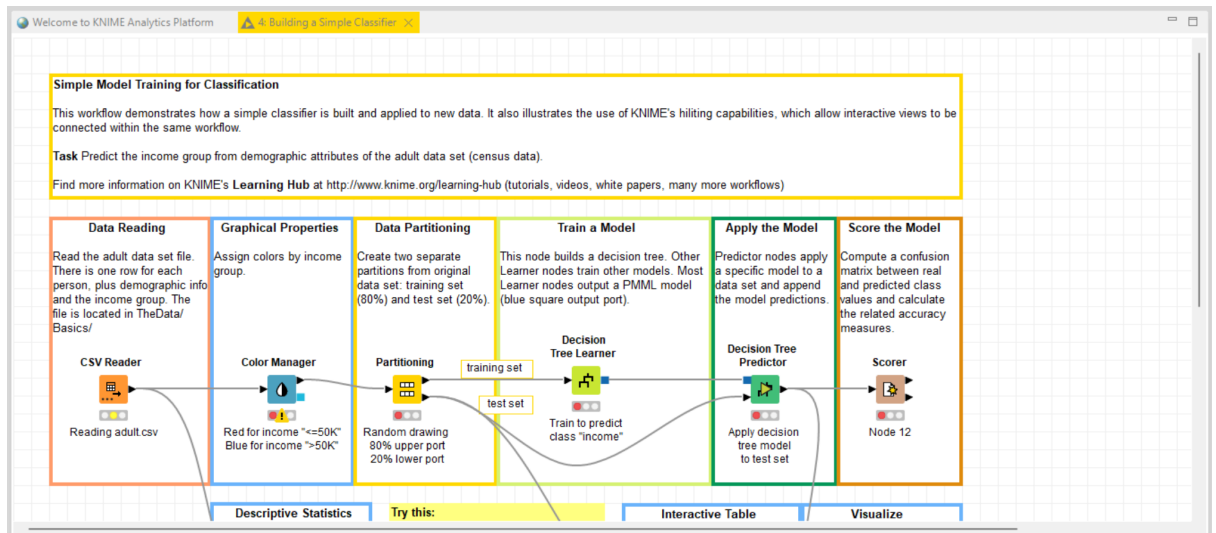
Setelah software Knime telah terbuka. Terdapat 3 bagian utama dari software ini. Yang pertama yaitu Knime Explorer yang isinya adalah project project yang kita buat pada software ini.



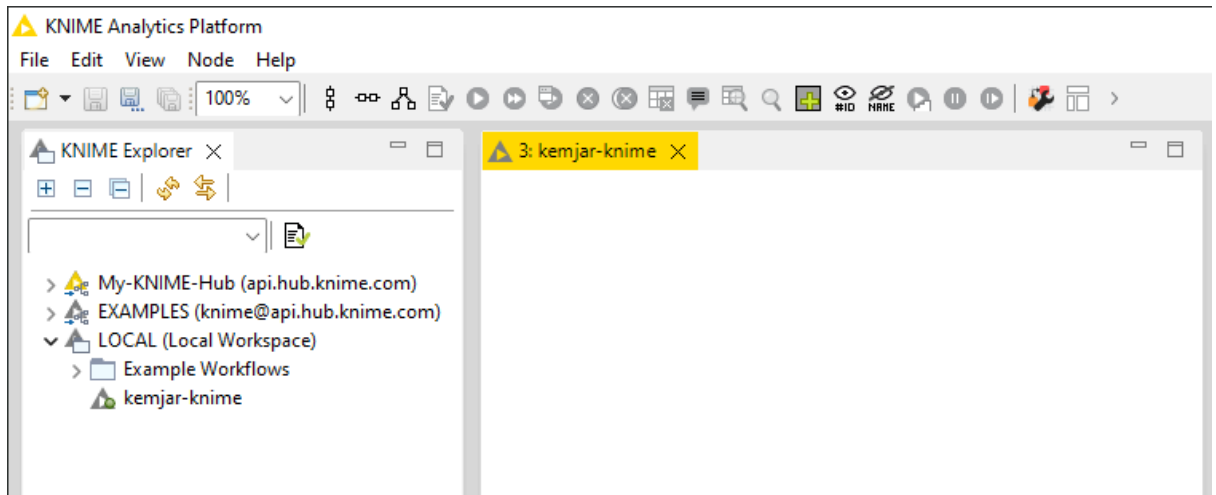
Kemudian terdapat Node Repository, bagian ini merupakan bagian yang sangat penting, karena berisi seluruh fungsi tools dari software ini yang dinamakan dengan Node.



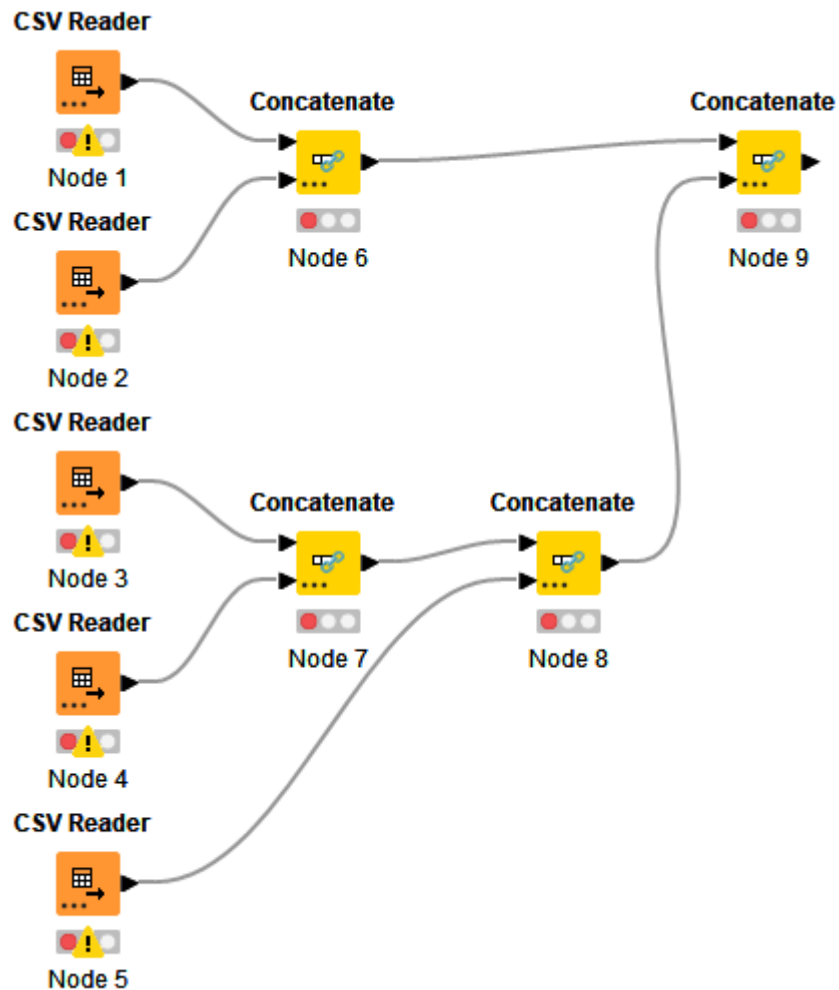
Terakhir yaitu Knime Workflow, bagian ini adalah bagian visual pada Knime, seluruh fungsi yang digunakan akan ditampilkan pada bagian ini. Berikut adalah contoh tampilan pada Knime Workflow



- Setelah mengenal semua bagian dari Knime. Setelah itu kita akan membuat workflow/project baru. Dengan cara klik File – New – New Knime Workflow – Tulis Nama workflow dan Lokasi workflow tersebut – Klik Finish



3. Selanjutnya yaitu menggabungkan seluruh data tadi menjadi 1 data. Node yang dibutuhkan untuk proses ini yaitu :
- File Reader : untuk membaca data
 - Concatenate : untuk menggabungkan data



Karena node Concatenate hanya dapat menerima input dari 2 data, maka diperlukan lebih dari 1 node Concatenate.

Untuk melihat konfigurasi dari File Reader, dapat digunakan cara klik kanan pada Node, lalu configure

Settings Transformation Advanced Settings Limit Rows Encoding Flow Variables Job Manager Selection Memory Policy

Input location

Read from Local File System

Mode ☒ File ☐ Files in folder

File C:\botnet_data\init.csv Brow

Reader options

Format

Autodetect format ⚙

Column delimiter Row delimiter ☒ Line break ☐ Custom

Quote char Quote escape char

Comment char

☒ Has column header ☐ Has row ID

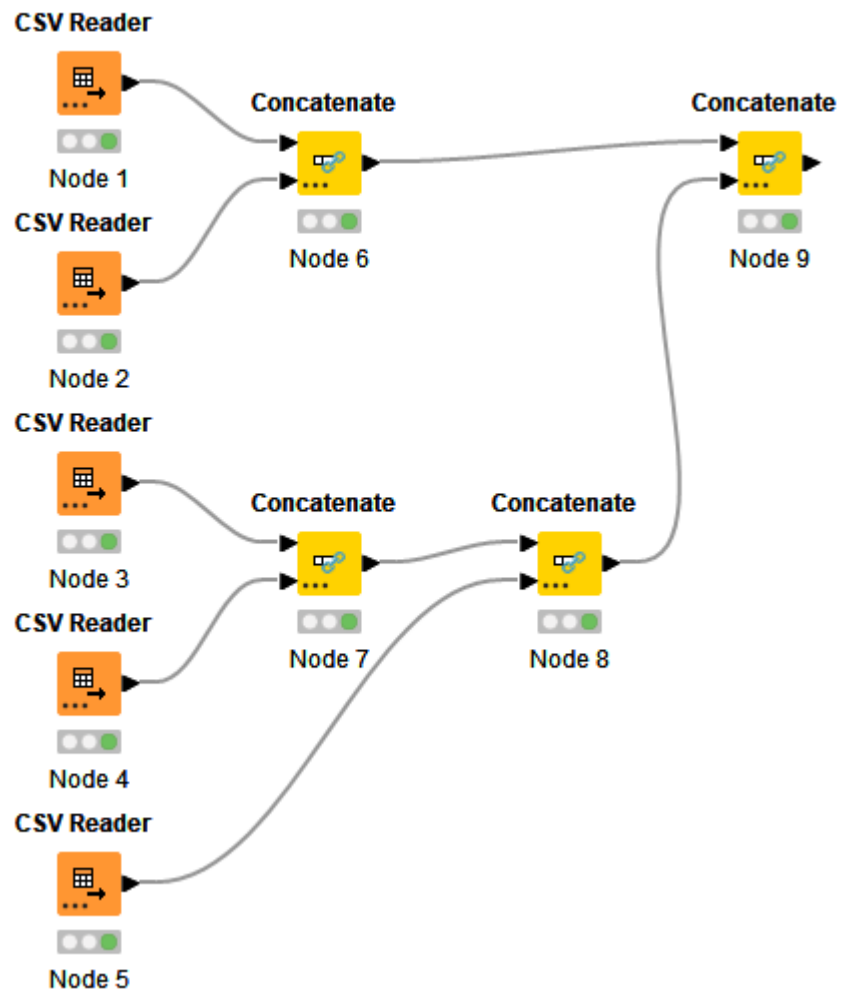
☐ Support short data rows ☐ Prepend file index to row ID

Preview

i The suggested column types are based on the first 10000 rows only. See 'Advanced Settings' tab.

Row ID	I No.	D Time	S Source	S Destin...	S Protocol	I Length	S Info
Row0	79	309.51	192.168.50.1	192.168.50.88	TCP	74	48221 > 22 [SYN] Seq=0 Win=29200 Len=0 MSS=1460 SACK_PERM TS
Row1	80	309.51	192.168.50.88	192.168.50.1	TCP	54	22 > 48221 [RST, ACK] Seq=1 Ack=1 Win=0 Len=0
Row2	81	312.264	192.168.50.1	192.168.50.88	TCP	74	48222 > 22 [SYN] Seq=0 Win=29200 Len=0 MSS=1460 SACK_PERM TS
Row3	82	312.264	192.168.50.88	192.168.50.1	TCP	54	22 > 48222 [RST, ACK] Seq=1 Ack=1 Win=0 Len=0
Row4	101	462.027	192.168.50.88	142.104.64....	TCP	74	38544 > 22 [SYN] Seq=0 Win=29200 Len=0 MSS=1460 SACK_PERM TS
Row5	102	463.024	192.168.50.88	142.104.64....	TCP	74	[TCP Retransmission] 38544 > 22 [SYN] Seq=0 Win=29200 Len=0 MSS=
Row6	106	472.199	192.168.50.88	142.104.64....	TCP	74	58842 > 1922 [SYN] Seq=0 Win=29200 Len=0 MSS=1460 SACK_PERM
Row7	107	473.196	192.168.50.88	142.104.64....	TCP	74	[TCP Retransmission] 58842 > 1922 [SYN] Seq=0 Win=29200 Len=0 MS
Row8	108	475.2	192.168.50.88	142.104.64....	TCP	74	[TCP Retransmission] 58842 > 1922 [SYN] Seq=0 Win=29200 Len=0 MS
Row9	109	479.212	192.168.50.88	142.104.64....	TCP	74	[TCP Retransmission] 58842 > 1922 [SYN] Seq=0 Win=29200 Len=0 MS
Row10	110	487.228	192.168.50.88	142.104.64....	TCP	74	[TCP Retransmission] 58842 > 1922 [SYN] Seq=0 Win=29200 Len=0 MS
Row11	114	503.28	192.168.50.88	142.104.64....	TCP	74	[TCP Retransmission] 58842 > 1922 [SYN] Seq=0 Win=29200 Len=0 MS
Row12	122	521.046	192.168.50.88	142.104.64....	TCP	74	37890 > 1922 [SYN] Seq=0 Win=29200 Len=0 MSS=1460 SACK_PERM
Row13	123	521.046	142.104.64....	192.168.50.88	TCP	74	1922 > 37890 [SYN, ACK] Seq=0 Ack=1 Win=28960 Len=0 MSS=1460
Row14	124	521.046	192.168.50.88	142.104.64....	TCP	66	37890 > 1922 [ACK] Seq=1 Ack=1 Win=29312 Len=0 TSval=19514756
Row15	125	521.047	192.168.50.88	142.104.64....	TCP	107	37890 > 1922 [PSH, ACK] Seq=1 Ack=1 Win=29312 Len=41 TSval=195
Row16	126	521.047	142.104.64....	192.168.50.88	TCP	66	1922 > 37890 [ACK] Seq=1 Ack=42 Win=28992 Len=0 TSval=3420025

Untuk konfigurasi file reader hanya tinggal memasukkan file csv yang telah diexport pada langkah sebelumnya. Klik Apply – OK. Proses ini belum selesai, karena Node belum di jalankan, untuk menjalankan Node bisa dengan cara klik kanan pada Node – Execute. Bila berhasil dijalankan, status Node yang berada dibawah Node akan berubah berwarna Hijau.



B. Pelabelan Data

1. Selanjutnya yaitu proses melabeli data. Pada data ISOT terdapat 2 tipe data, yaitu malicious dan normal. Pertama kita harus menyediakan tabel dengan label malicious yang sudah ditentukan seperti berikut.

Dialog - 3:11 - CSV Reader

File

Settings Transformation Advanced Settings Limit Rows Encoding Flow Variables Job Manager Selection Memory Policy

Input location

Read from: Local File System

Mode: ☒ File ☐ Files in folder

File: C:\Users\Admin\Downloads\label.csv

Reader options

Format

Autodetect format

Column delimiter: , Row delimiter: ☒ Line break ☐ Custom \r\n

Quote char: " Quote escape char: \"

Comment char: #

Has column header: ☒ Has row ID: ☐

Support short data rows: ☐ Prepend file index to row ID: ☐

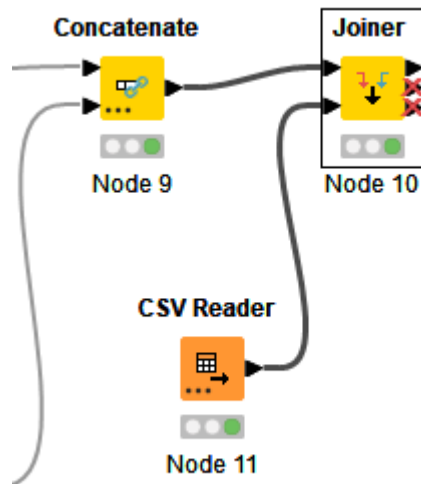
Preview

The suggested column types are based on the first 10000 rows only. See 'Advanced Settings' tab.

Row ID	Source	label
Row0	192.168.50.14	malicious
Row1	192.168.50.15	malicious
Row2	192.168.50.16	malicious
Row3	192.168.50.17	malicious
Row4	192.168.50.18	malicious
Row5	192.168.50.30	malicious
Row6	192.168.50.31	malicious
Row7	192.168.50.32	malicious
Row8	192.168.50.34	malicious

OK Apply Cancel ?

2. Kemudian selanjutnya akan menggabungkan 2 tabel menggunakan node Joiner. Node ini membutuhkan 2 input data.



3. Untuk konfigurasi Joiner, dapat menggunakan konfigurasi berikut
Lalu klik OK dan jalankan. Hasilnya akan seperti berikut.

Dialog - 3:10 - Joiner

File

Joiner Settings

Column Selection

Performance

Flow Variables

Job Manager Selection

Memory Policy

Join columns

Match

☒ all of the following

☐ any of the following

Top Input ('left' table)

Bottom Input ('right' table)

S

Source

S

i»zSource

+

-

+

Compare values in join columns by

☒ value and type

☐ string representation

☐ making integer types compatible


Include in output

☒ Matching rows

☒ Left unmatched rows

☐ Right unmatched rows

Left outer join



Output options

☐ Split join result into multiple tables (top = matching rows, middle = left unmatched rows, bottom = right unmatched rows)

☐ Merge join columns

☐ Hlling enabled

Row Keys

☒ Concatenate original row keys with separator


☐ Assign new row keys sequentially

☐ Keep row keys

OK

Apply

Cancel



Dialog - 3:10 - Joiner

File

Joiner Settings

Column Selection

Performance

Flow Variables

Job Manager Selection

Memory Policy

Top Input (left table)

Manual Selection

Wildcard/Regex Selection

Type Selection

Exclude

Filter

No columns in this list

Enforce exclusion

Include

Filter

I

No.

D

Time

S

Source

S

Destination

S

Protocol

I

Length

S

Info

D

Delta Time

Enforce inclusion

>

>>

<

<<

Bottom Input (right table)

Manual Selection

Wildcard/Regex Selection

Type Selection

Exclude

Filter

S

Source

Enforce exclusion

Include

Filter

S

label

Enforce inclusion

>

>>

<

<<

Duplicate column names

Do not execute

Append custom suffix

(right)

OK

Apply

Cancel

?

Console Node Monitor

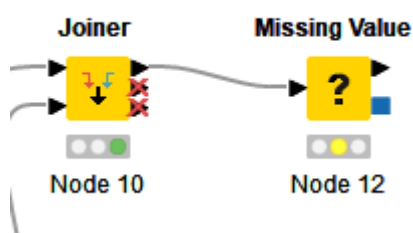
Node: Joiner (3:10)

State: EXECUTED

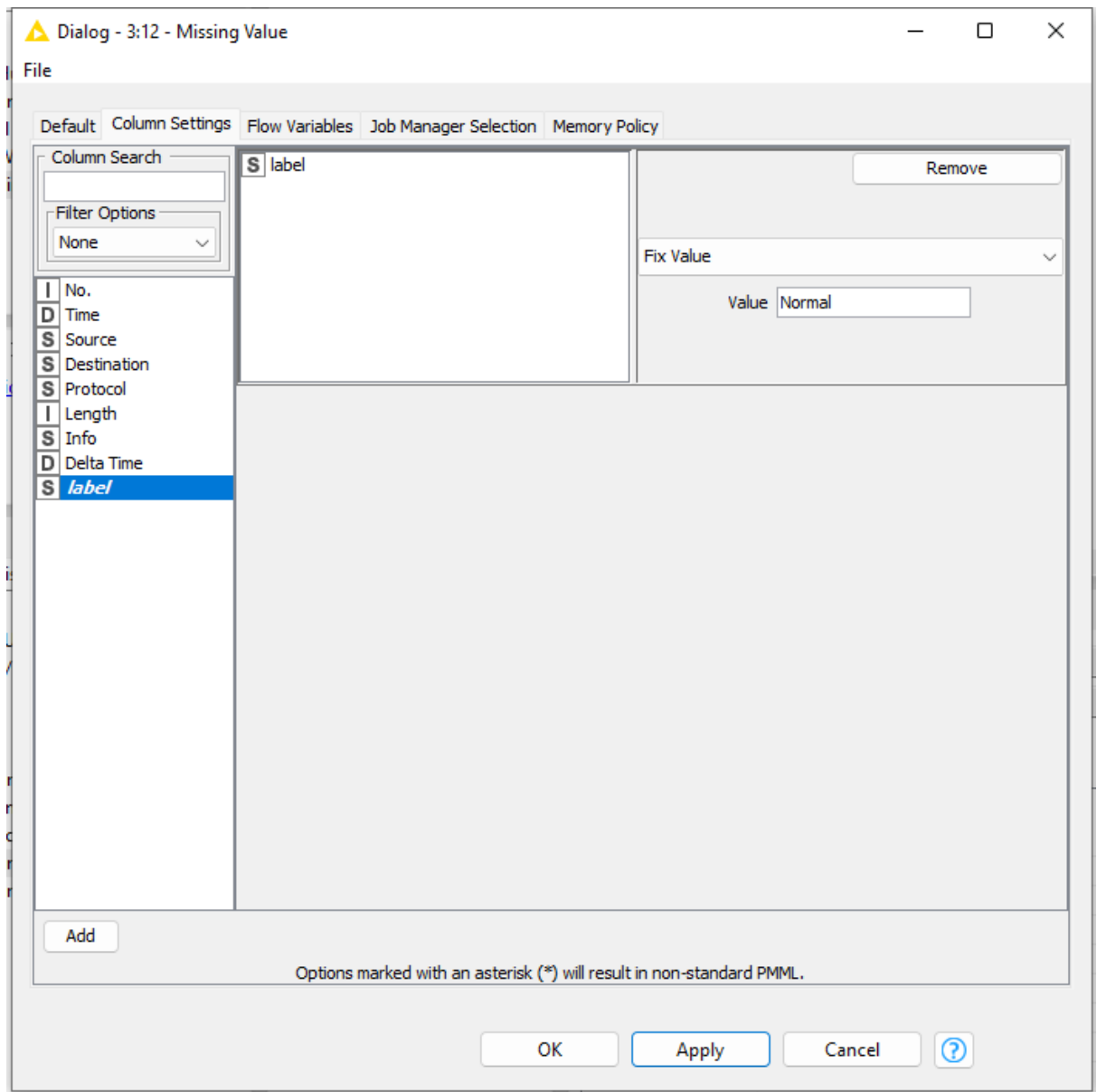
Port Output Port 0 Load data Rows: 4187432, Columns: 9

ID	No.	Time	Source	Destination	Protocol	Len
Row0_?	79	309.509587	192.168.50.1	192.168.50.88	TCP	74
Row1_?	80	309.509587	192.168.50.88	192.168.50.1	TCP	54
Row2_?	81	312.263619	192.168.50.1	192.168.50.88	TCP	74
Row3_?	82	312.263647	192.168.50.88	192.168.50.1	TCP	54
Row4_?	101	462.026672	192.168.50.88	142.104.64.205	TCP	74
Row5_?	102	463.024215	192.168.50.88	142.104.64.205	TCP	74
Row6_?	106	472.199177	192.168.50.88	142.104.64.205	TCP	74
Row7_?	107	473.196245	192.168.50.88	142.104.64.205	TCP	74
Row8_?	108	475.200246	192.168.50.88	142.104.64.205	TCP	74
Row9_?	109	479.212237	192.168.50.88	142.104.64.205	TCP	74
Row10_?	110	487.228249	192.168.50.88	142.104.64.205	TCP	74
Row11_?	114	503.279506	192.168.50.88	142.104.64.205	TCP	74
Row12_?	122	521.04626	192.168.50.88	142.104.64.203	TCP	74

4. Dari langkah 7 menghasilkan output data dengan label malicious tetapi masih terdapat data dengan berlabel “?”. Hal ini dikarenakan kita hanya labeling untuk data malicious saja. Untuk melakukan labeling data normal kita akan menggunakan Node Missing Value. Node ini digunakan untuk mengisi data kosong.



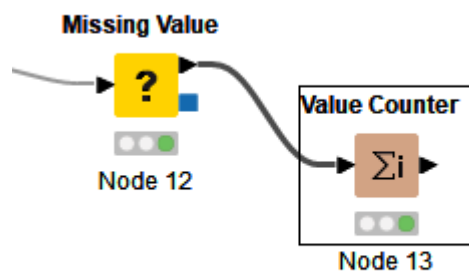
Untuk konfigurasi, dapat menggunakan konfigurasi berikut.



Konfigurasi ini nantinya akan mengisi value yang kosong dengan value Normal. Berikut hasil dari proses Missing Value

Output table - 3:12 - Missing Value						
File Edit Hilite Navigation View						
Table "default" - Rows: 4187432 Spec - Columns: 9 Properties Flow Variables						
Row ID		[D] Delta Ti...	[S] label			
Row0_?	200 Len=0 MSS=1460 SACK_PERM TSval=341949681 TSecr=0 WS=64	4.428	Normal			
Row1_?	k=1 Win=0 Len=0	0	Normal			
Row2_?	200 Len=0 MSS=1460 SACK_PERM TSval=341950369 TSecr=0 WS=64	2.754	Normal			
Row3_?	k=1 Win=0 Len=0	0	Normal			
Row4_?	200 Len=0 MSS=1460 SACK_PERM TSval=195132807 TSecr=0 WS=...	7.771	Normal			
Row5_?	[SYN] Seq=0 Win=29200 Len=0 MSS=1460 SACK_PERM TSval=1951...	0.998	Normal			
Row6_?	29200 Len=0 MSS=1460 SACK_PERM TSval=195135350 TSecr=0 WS...	5.163	Normal			
Row7_?	22 [SYN] Seq=0 Win=29200 Len=0 MSS=1460 SACK_PERM TSval=19...	0.997	Normal			
Row8_?	22 [SYN] Seq=0 Win=29200 Len=0 MSS=1460 SACK_PERM TSval=19...	2.004	Normal			
Row9_?	22 [SYN] Seq=0 Win=29200 Len=0 MSS=1460 SACK_PERM TSval=19...	4.012	Normal			
Row10_?	22 [SYN] Seq=0 Win=29200 Len=0 MSS=1460 SACK_PERM TSval=19...	8.016	Normal			
Row11_?	22 [SYN] Seq=0 Win=29200 Len=0 MSS=1460 SACK_PERM TSval=19...	1.888	Normal			
Row12_?	29200 Len=0 MSS=1460 SACK_PERM TSval=195147562 TSecr=0 WS...	3.362	Normal			
Row13_?	Ack=1 Win=28960 Len=0 MSS=1460 SACK_PERM TSval=342002564...	0	Normal			
Row14_?	1 Win=29312 Len=0 TSval=195147562 TSecr=342002564	0	Normal			

5. Untuk memastikan bahwa kolom label sudah terisi dengan value Malicious atau Normal, dapat menggunakan node Value Counter. Node ini berfungsi untuk menghitung jumlah seluruh value pada kolom terpilih.

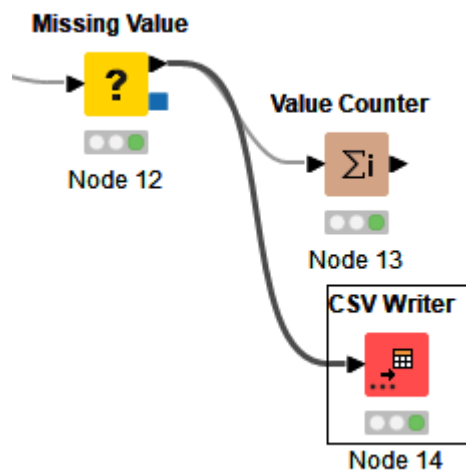


Berikut adalah hasil dari perhitungan value dengan Value Counter. Dalam konfigurasinya tinggal memilih kolom yang ingin dihitung yaitu kolom label.

ID	count
Normal	3135405
malicious	1052027

Dari gambar tersebut bisa dilihat jika sudah tidak ada data yang memiliki label kosong, hanya terdapat 2 label yaitu Normal dan malicious.

6. Export file ke dalam format .csv dengan menggunakan node CSV Writer.



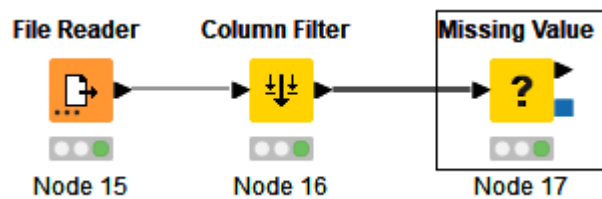
KNIME FOR DATA MINING

Pada langkah ini, kita akan melakukan sebuah analisa pada data ISOT dengan menggunakan platform Knime. Pada proses sebelumnya kita sudah melakukan penggabungan data dan pelabelan data. Pada proses ini kita akan melakukan analisa data dengan menggunakan Teknik Data Mining. Data mining adalah proses penggalian data atau penemuan informasi baru dengan mencari pola atau aturan tertentu dari sebuah data. Di dalam Data Mining terdapat beberapa proses yang harus dilakukan terlebih dahulu sebelum dilakukannya data mining. Berikut tahapan-tahapan untuk melakukan data mining, yaitu

1. Data Pre-Processing
2. Data Transformation
3. Data Mining
4. Evaluation

A. Data Pre-processing

1. Proses ini adalah proses dimana data akan dibersihkan (cleaning) karena biasanya didalam suatu data terdapat nilai-nilai yang tidak sempurna atau bahkan terdapat nilai-nilai yang hilang atau kosong yang nantinya akan dapat mempengaruhi proses kedepannya. Pada proses ini kita membutuhkan Node-node berikut : File Reader, Column Filter, Missing Value.



2. Sebelum memulai data pre-processing, langkah pertama yaitu membaca file data csv yang sudah diexport pada proses sebelumnya menggunakan Node File Reader. Untuk konfigurasinya cari lokasi file data yang diexport tadi.

Dialog - 3:15 - File Reader

File

Settings Transformation Advanced Settings Limit Rows Encoding Flow Variables Job Manager Selection Memory Policy

Input location

Read from Local File System

Mode ☒ File ☐ Files in folder

File C:\botnet_data\init_result.csv

Reader options

Format

Autodetect format

Column delimiter , Row delimiter ☒ Line break ☐ Custom \r\n

Quote char " Quote escape char \"

Comment char #

☒ Has column header ☐ Has row ID

☐ Support short data rows ☐ Prepend file index to row ID

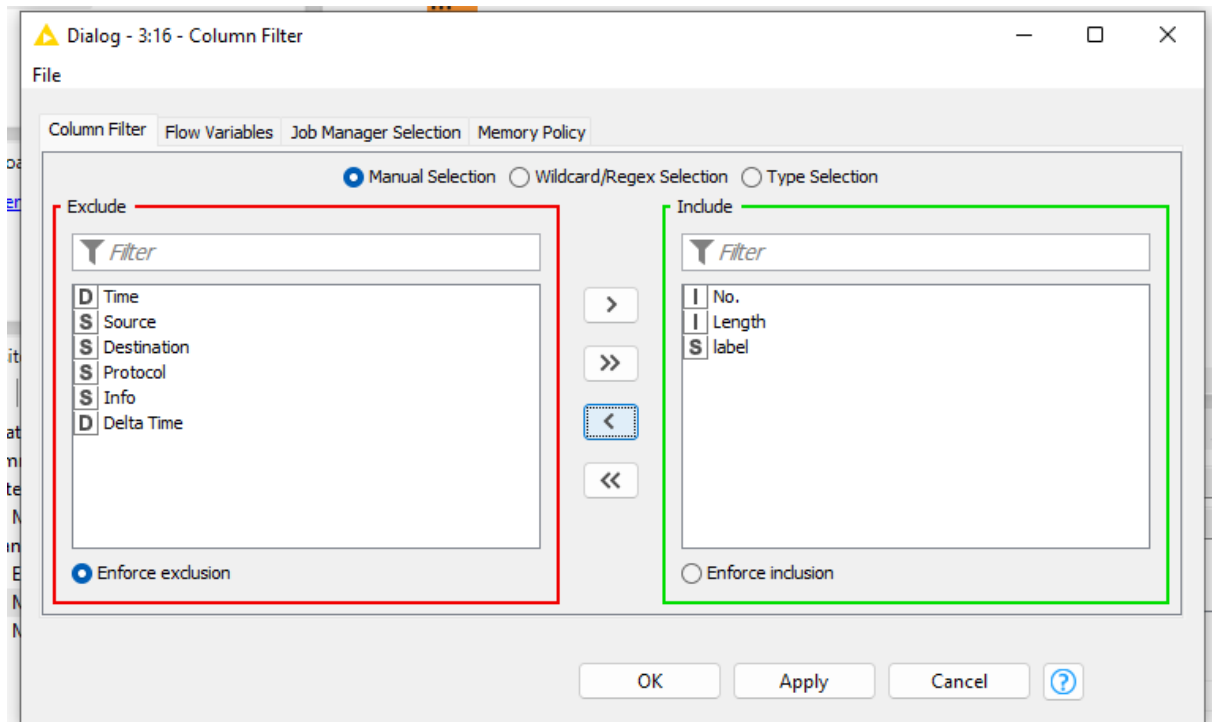
Preview

The suggested column types are based on the first 10000 rows only. See 'Advanced Settings' tab.

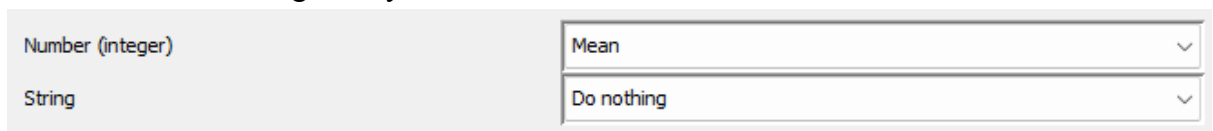
Row ID	I No.	D Time	S Source	S Destinat...	S Protocol	I Length	S Info
Row0	79	309.51	192.168.50.1	192.168.50.88	TCP	74	48221 > 22 [SYN] Seq=0 Win=29200 Len=0 MSS=1460 SACK_PERM TS
Row1	80	309.51	192.168.50.88	192.168.50.1	TCP	54	22 > 48221 [RST, ACK] Seq=1 Ack=1 Win=0 Len=0
Row2	81	312.264	192.168.50.1	192.168.50.88	TCP	74	48222 > 22 [SYN] Seq=0 Win=29200 Len=0 MSS=1460 SACK_PERM TS
Row3	82	312.264	192.168.50.88	192.168.50.1	TCP	54	22 > 48222 [RST, ACK] Seq=1 Ack=1 Win=0 Len=0
Row4	101	462.027	192.168.50.88	142.104.64...	TCP	74	38544 > 22 [SYN] Seq=0 Win=29200 Len=0 MSS=1460 SACK_PERM TS
Row5	102	463.024	192.168.50.88	142.104.64...	TCP	74	[TCP Retransmission] 38544 > 22 [SYN] Seq=0 Win=29200 Len=0 MSS=
Row6	106	472.199	192.168.50.88	142.104.64...	TCP	74	58842 > 1922 [SYN] Seq=0 Win=29200 Len=0 MSS=1460 SACK_PERM
Row7	107	473.196	192.168.50.88	142.104.64...	TCP	74	[TCP Retransmission] 58842 > 1922 [SYN] Seq=0 Win=29200 Len=0 MS
Row8	108	475.2	192.168.50.88	142.104.64...	TCP	74	[TCP Retransmission] 58842 > 1922 [SYN] Seq=0 Win=29200 Len=0 MS
Row9	109	479.212	192.168.50.88	142.104.64...	TCP	74	[TCP Retransmission] 58842 > 1922 [SYN] Seq=0 Win=29200 Len=0 MS
Row10	110	487.228	192.168.50.88	142.104.64...	TCP	74	[TCP Retransmission] 58842 > 1922 [SYN] Seq=0 Win=29200 Len=0 MS
Row11	114	503.28	192.168.50.88	142.104.64...	TCP	74	[TCP Retransmission] 58842 > 1922 [SYN] Seq=0 Win=29200 Len=0 MS
Row12	122	521.046	192.168.50.88	142.104.64...	TCP	74	37890 > 1922 [SYN] Seq=0 Win=29200 Len=0 MSS=1460 SACK_PERM
Row13	123	521.046	142.104.64...	192.168.50.88	TCP	74	1922 > 37890 [SYN, ACK] Seq=0 Ack=1 Win=28960 Len=0 MSS=1460
Row14	124	521.046	192.168.50.88	142.104.64...	TCP	66	37890 > 1922 [ACK] Seq=1 Ack=1 Win=29312 Len=0 TSval=19514756
Row15	125	521.047	192.168.50.88	142.104.64...	TCP	107	37890 > 1922 [PSH, ACK] Seq=1 Ack=1 Win=29312 Len=41 TSval=195
Row16	126	521.047	142.104.64...	192.168.50.88	TCP	66	1922 > 37890 [ACK] Seq=1 Ack=42 Win=28992 Len=0 TSval=3420025

OK Apply Cancel ?

- Selanjutnya kita akan menggunakan Node Column Filter. Node ini berfungsi untuk mem-filter kolom atau atribut yang tidak digunakan. Disini kita hanya menggunakan 3 atribut, yaitu : Delta_time, length dan label.

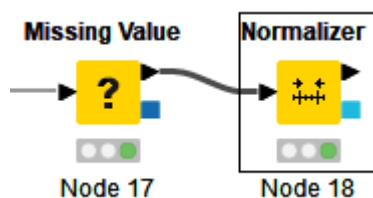


4. Setelah langkah 2 dijalankan, data tersebut menjadi memiliki 3 kolom atau atribut yang sebelumnya terdapat 9 kolom. Selanjutnya kita akan menjalankan Node Missing Value. Node ini sudah pernah kita pakai pada proses labeling data. Tetapi pada proses ini kita akan melakukan pembersihan data, karena biasanya didalam suatu data terdapat kolom yang tidak sempurna seperti data yang hilang atau atribut yang tidak relevan, untuk itu Node ini diperlukan untuk mengatasi hal tersebut. Berikut konfigurasinya.



B. Data Transformation

1. Setelah melakukan data pre-processing, selanjutnya akan menuju ke proses data transformation, pada proses ini data akan diubah ke format yang sesuai untuk proses data mining. Node yang digunakan pada tahap ini yaitu Normalizer. Berikut konfigurasinya

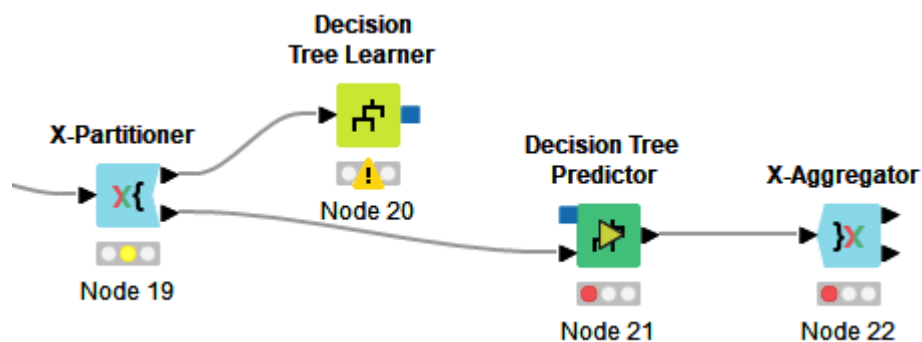


2. Setelah dijalankan, kolom dari data tersebut akan berubah menjadi bentuk range. Data inilah yang nantinya akan digunakan dalam pengenalan pola. Berikut adalah hasil dari Node Normalizer
Setelah mendapatkan data ini, baru kita dapat menjalankan proses Data Mining

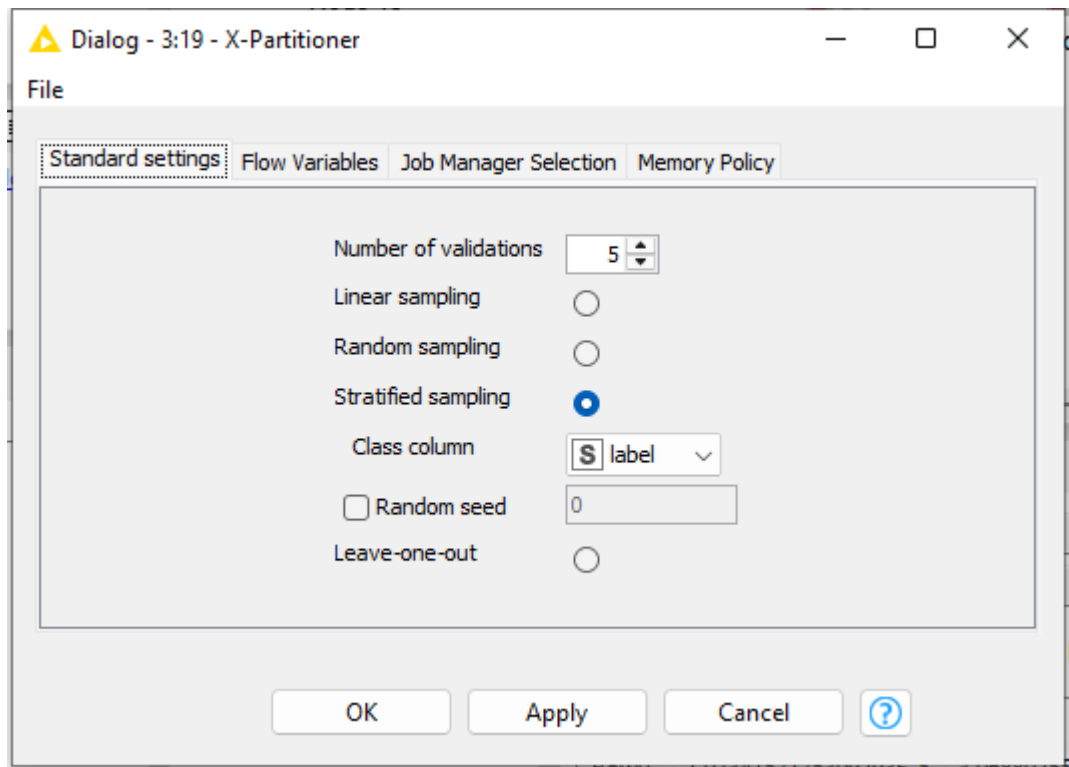
ID	Length	label
Row0	3.068802553243725E-4	Normal
Row1	0.0	Normal
Row2	3.068802553243725E-4	Normal
Row3	0.0	Normal
Row4	3.068802553243725E-4	Normal
Row5	3.068802553243725E-4	Normal
Row6	3.068802553243725E-4	Normal
Row7	3.068802553243725E-4	Normal
Row8	3.068802553243725E-4	Normal
Row9	3.068802553243725E-4	Normal
Row10	3.068802553243725E-4	Normal
Row11	3.068802553243725E-4	Normal
Row12	3.068802553243725E-4	Normal
Row13	3.068802553243725E-4	Normal

C. Data Mining

1. Setelah menyelesaikan tahap data transformation, kita akan menjalankan proses Data Mining, dalam proses ini kita akan menggunakan Metode Klasifikasi Decision Tree dengan teknik Cross Validation. Pada proses ini kita membutuhkan Node-node berikut: X-Partitioner, Decision Tree Learner, Decision Tree Predictor, X-Aggregator. Sehingga akan membentuk flow seperti ini



2. X-Partitioner berfungsi untuk menentukan jumlah iterasi atau pengulangan pada teknik cross validation, data ini nantinya akan terbagi menjadi 2 yaitu data training dan data testing. Berikut konfigurasinya



3. Decision Tree Learner berfungsi sebagai data training, karena metode Decision Tree merupakan supervised learning, sehingga membutuhkan data training untuk mengenali pola dari setiap data. Berikut konfigurasi dari Decision Tree Learner

Dialog - 3:20 - Decision Tree Learner

File

Options PMMLSettings Flow Variables Job Manager Selection

General

Class column **S** label

Quality measure Gini index

Pruning method No pruning

☒ Reduced Error Pruning

Min number records per node 2

Number records to store for view 10,000

☒ Average split point

Number threads 8

☒ Skip nominal columns without domain information

Root split

☐ Force root split column

Root split column **D** Length

Binary nominal splits

☐ Binary nominal splits

Max #nominal 10

☐ Filter invalid attribute values in child nodes

OK Apply Cancel ?

4. Setelah menjalankan Decision Tree Learner, lalu dilanjutkan dengan Decision Tree Predictor. Node ini berfungsi untuk menklasifikasi data dengan cara menguji data testing dengan hasil dari proses Decision Tree Learner. Berikut konfigurasinya.

Options Flow Variables Job Manager Selection Memory Policy

Maximum number of stored patterns for HiLite-ing: 5,000,000

☐ Change prediction column name
Prediction (label)

☐ Append columns with normalized class distribution
Suffix for probability columns

5. Node X-Aggregator berfungsi sebagai akhir dari proses cross validation. Node ini akan mengumpulkan hasil dari Node Predictor yang akan menampilkan hasil dari prediksi dari beberapa iterasi yang dilakukan. Tidak ada konfigurasi khusus dari node ini, sehingga bisa langsung dijalankan. Berikut adalah hasil dari node X-Aggregator. Dari hasil ini akan mendapatkan kolom baru yaitu kolom prediksi.

Row ID	D Delta_ti...	D Length	S label	S Predicti...
Row4	18.839	-0.094	normal	normal
Row12	7.998	-0.094	normal	normal
Row16	-0.266	-0.099	normal	normal
Row27	-0.167	-0.099	normal	normal
Row32	-0.25	-0.067	normal	normal
Row46	-0.265	-0.099	normal	normal
Row47	-0.263	-0.099	normal	normal
Row49	2.263	-0.094	normal	malicious
Row57	-0.266	-0.099	normal	normal
Row58	-0.265	0.722	normal	normal
Row63	-0.259	-0.09	normal	normal
Row65	-0.266	-0.072	normal	normal
Row67	-0.265	-0.072	normal	normal