Thanawat Tejapijaya

Weekly report Week 1 (22/5/2023 - 26/5/2023)

This week I start with selecting a dataset which is CTU13 dataset that is from CTU university 2011 (https://www.stratosphereips.org/datasets-ctu13), this CTU13 dataset contain 13 .binetflow files and this dataset contain 15 Features which are

1.1. StartTime: Start Time of the recorded traffic flow
1.2. Dur: Duration of the flow/ How long the flow connect
1.3. Proto: protocol used
1.4. SrcAddr: IP Address of Source
1.5. Sport: Port of Source
1.6. Dir: Direction of the flow
1.7. DstAddr: IP Address of Destination
1.8. Dport: Port of Destination
1.9. State: The state is protocol dependent and _ is a separator for one end of the connection.
1.10. sTOS: Source TOS byte value - use to tell priority of packet
1.11. dTOS: Destination TOS byte value
1.12. TotPkts: Total numbers of transaction of each Packet
1.13. TotBytes: total numbers of transaction Bytes
1.14. SrcBytes: Total number of transaction Bytes from the Source
1.15. Label: Label made of “flow=” followed by a short description

So which label do I consider as an attack?

the attack label will contain flow=From-Botnet-.....

First of all I do try run code that is attach from this paper

(https://paperswithcode.com/paper/cyber-attack-detection-thanks-to-machine)

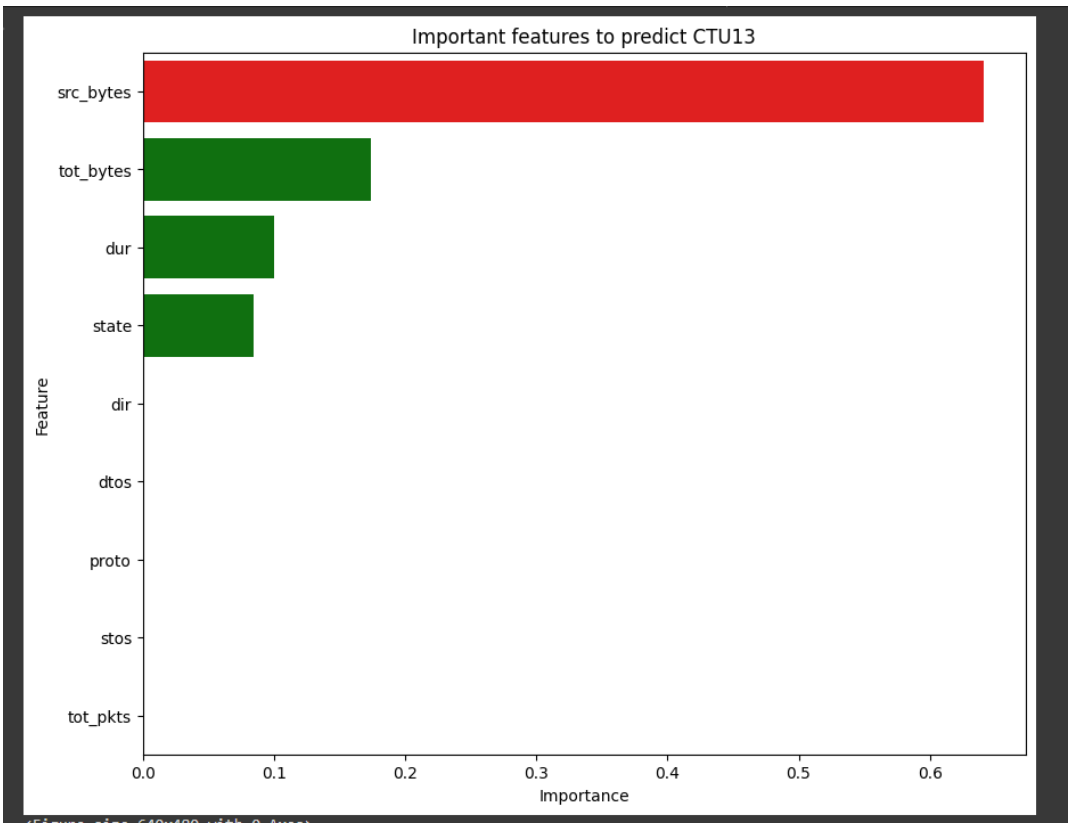
I used 3 days to fix bugs and try run the code as an result it take quite a long time to run

So I use clean up code for this dataset using the code from Kaggle

(https://www.kaggle.com/code/dhoogla/ctu-13-00-cleaning)

I will follow the code that it will dropped 5 features for this dataset which are StartTime, SrcAddr, Sport, DstAddr and Dport and change category features into numerical features ,after that I used feature important technique to find the feature that is important to the dataset which will be based on attack label (as an y) and I did tried 2 methods which are

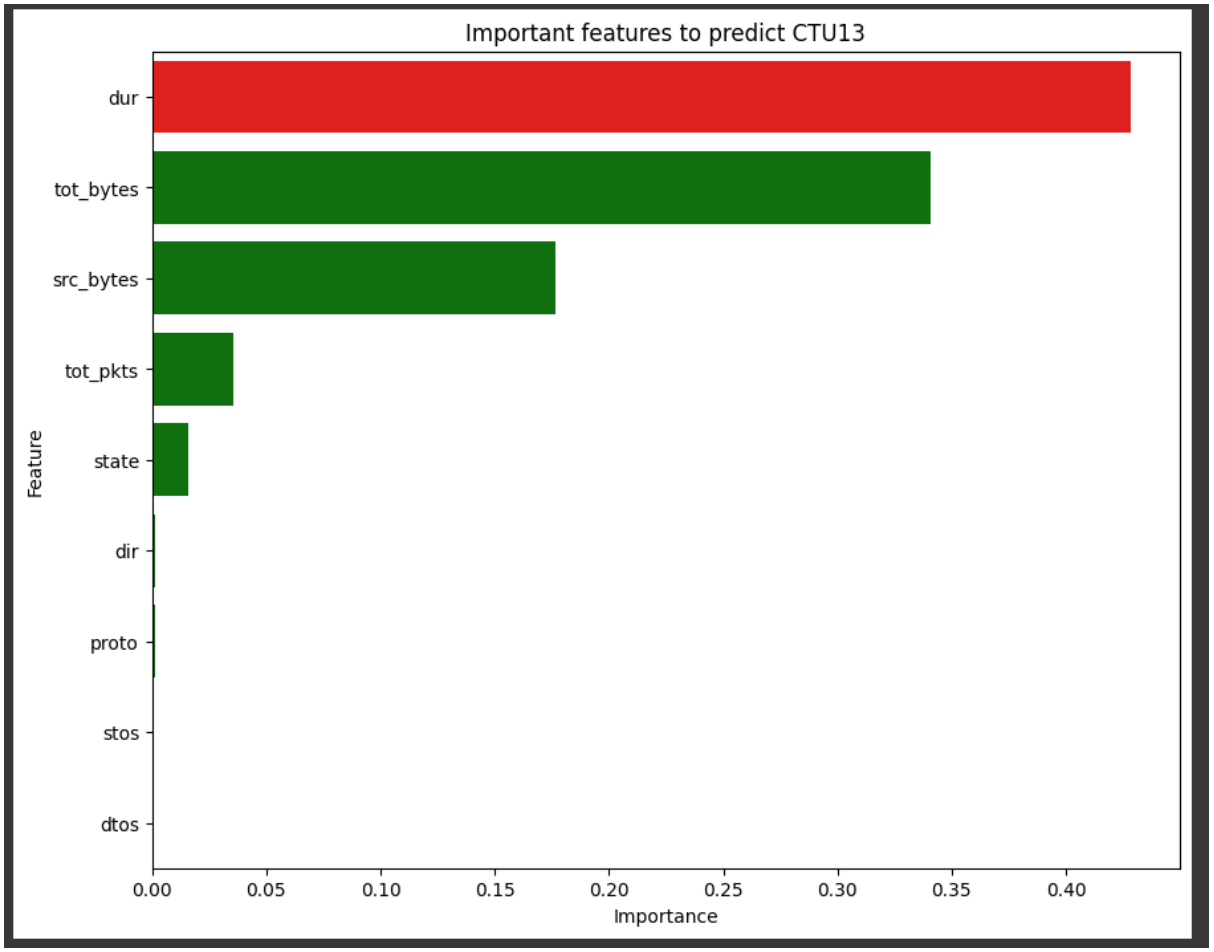
1. decision tree



	Feature	Importance
0	src_bytes	0.640989
1	tot_bytes	0.174150
2	dur	0.100360
3	state	0.084500
4	dir	0.000000
5	dtos	0.000000
6	proto	0.000000
7	stos	0.000000
8	tot_pkts	0.000000

As a result src_bytes is the most important feature for this classifier

2. Random Forest



	Feature	Importance
0	dur	0.428340
1	tot_bytes	0.341060
2	src_bytes	0.176350
3	tot_pkts	0.035631
4	state	0.015801
5	dir	0.001371
6	proto	0.001354
7	stos	0.000056
8	dtos	0.000037

As a result dur is the most important feature for this classifier

I still need to try more of difference classifier next week and will be focusing more on analyzing dataset before moving on to the next step so I will have a clear perspective and can fully understand the dataset