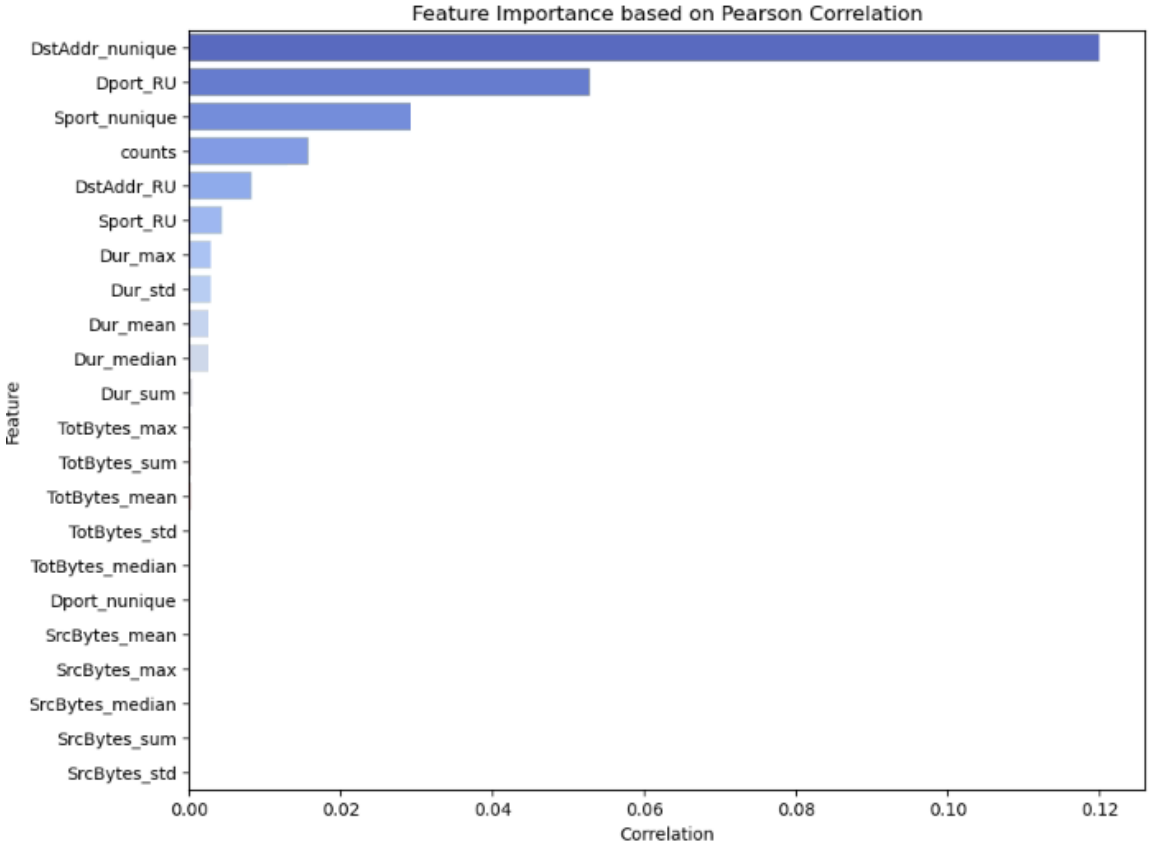
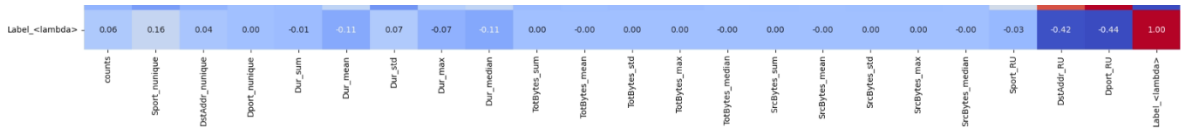
Thanawat Tejapijaya

Weekly report Week 2 (29/5/2023 - 2/6/2023)

This week I did more classifiers to find feature importance for this dataset. I divided the dataset into 3 cases to find the best feature for the dataset and since the dataset is verylarge if I combined all 3 together so I decided to use only one file which is from https://mcfp.felk.cvut.cz/publicDatasets/CTU-Malware-Capture-Botnet-44/ . First case is follow the research paper(Cyber Attack Detection thanks to Machine Learning Algorithms | Papers With Code) for numerical data, I normalized all the dataset into mean,median, std, sum and max. Moreover, for categorical data, I changed it to unique occurrence in subgroup and normalized subgroup using entropy as a result



These are the best Feature for using Pearson Correlation

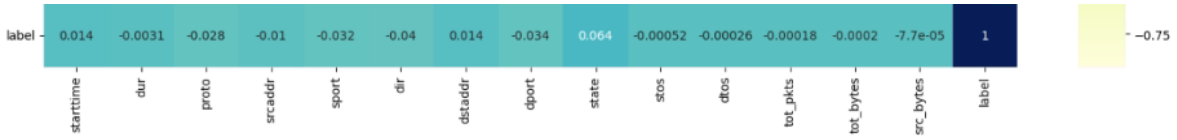


for heatmap correlation it's almost the same because DstAddr_nuique and Dport_RU have the highest rank for the feature importance

Then for the feature selection the code are still running

For case 2, I change the label into binary which are 0 and 1 and change all data in dataset into numerical and used classifier to find the feature importance and got the result as below

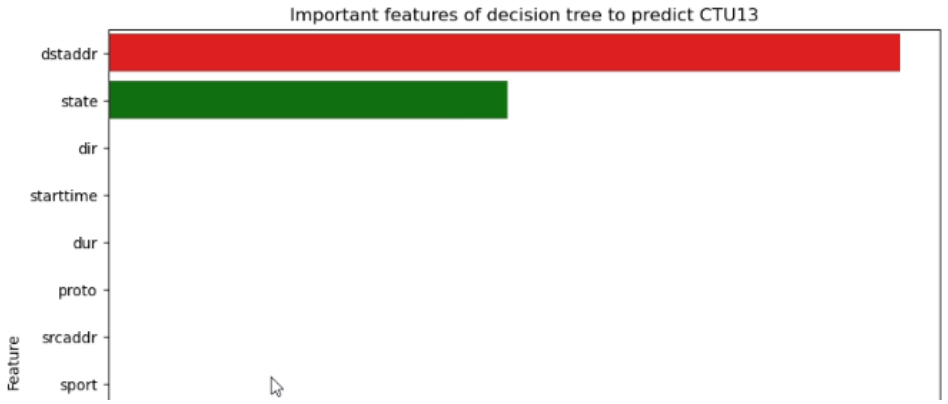
1. Correlation Heatmap



	Feature	Importance
0	starttime	0.063534
1	dur	0.039694
2	proto	0.033796
3	srcaddr	0.031795
4	sport	0.028033
5	dir	0.014040
6	dstaddr	0.013606
7	dport	0.010305
8	state	0.003099
9	stos	0.000523
10	dtos	0.000260
11	tot_pkts	0.000196
12	tot_bytes	0.000177
13	src_bytes	0.000077

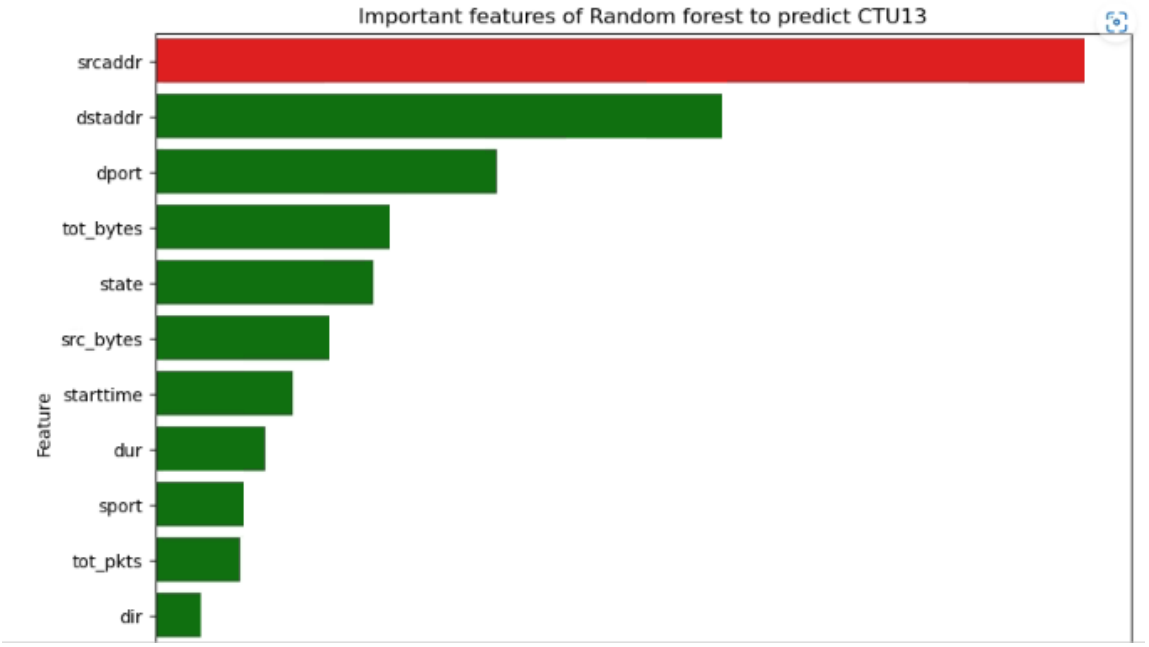
2. decision Tree

	Feature	Importance
0	dstaddr	0.663965
1	state	0.334910
2	dir	0.001125
3	starttime	0.000000
4	dur	0.000000
5	proto	0.000000
6	srcaddr	0.000000
7	sport	0.000000
8	dport	0.000000
9	stos	0.000000
10	dtos	0.000000
11	tot_pkts	0.000000
12	tot_bytes	0.000000
13	src_bytes	0.000000



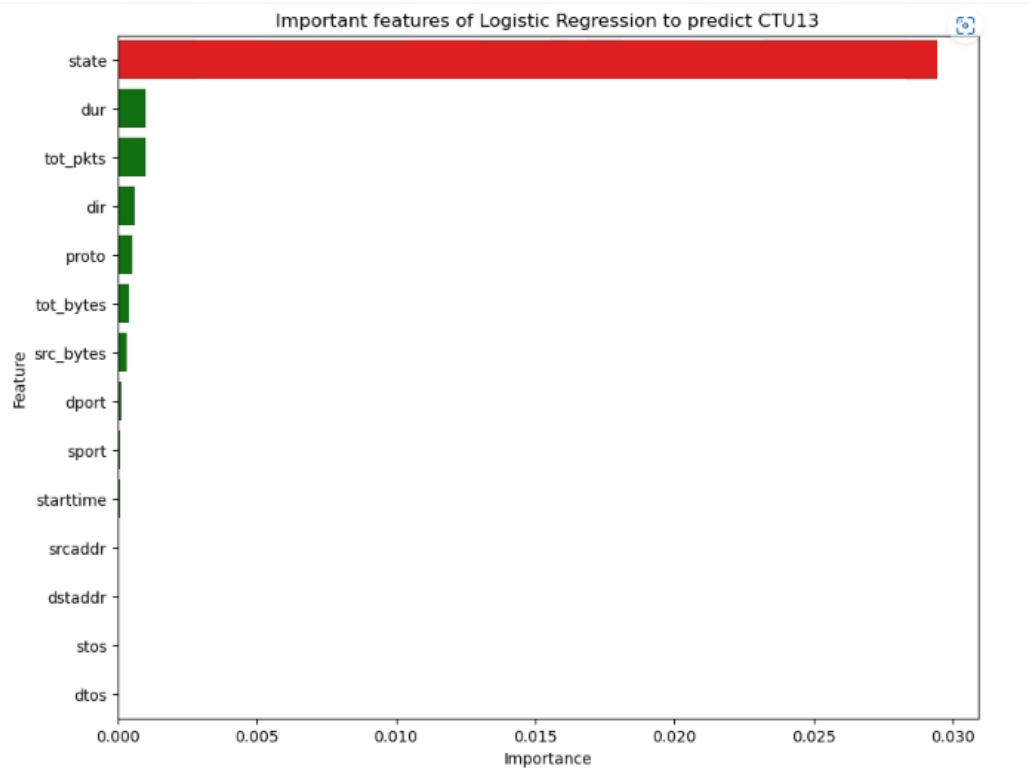
3. Random Forest

	Feature	Importance
0	srcaddr	3.133962e-01
1	dstaddr	1.912804e-01
2	dport	1.150874e-01
3	tot_bytes	7.920005e-02
4	state	7.358981e-02
5	src_bytes	5.876134e-02
6	starttime	4.640918e-02
7	dur	3.719657e-02
8	sport	2.986211e-02
9	tot_pkts	2.836823e-02
10	dir	1.523658e-02
11	proto	1.161214e-02
12	stos	1.485770e-12
13	dtos	0.000000e+00



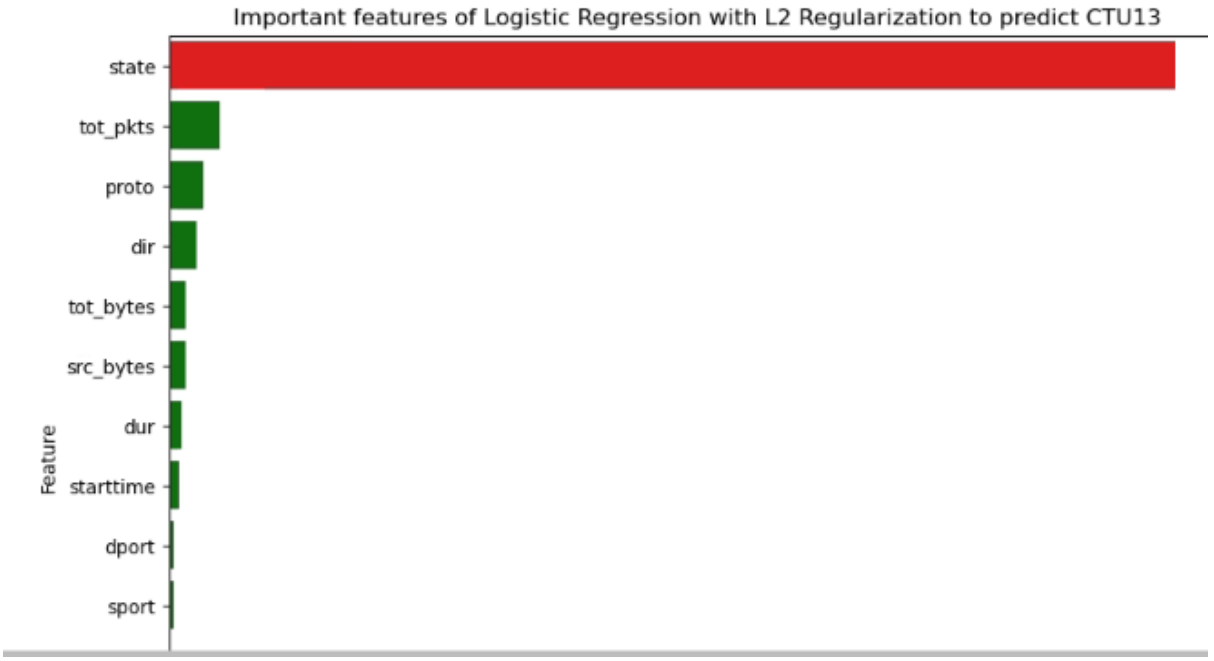
4. logistic regression

	Feature	Importance
0	state	2.943818e-02
1	dur	1.000914e-03
2	tot_pkts	1.000729e-03
3	dir	6.190468e-04
4	proto	5.409738e-04
5	tot_bytes	4.157962e-04
6	src_bytes	3.291808e-04
7	dport	1.167373e-04
8	sport	1.066408e-04
9	starttime	8.932052e-05
10	srcaddr	3.630971e-05
11	dstaddr	1.131974e-05
12	stos	6.609884e-07
13	dtos	2.255553e-07



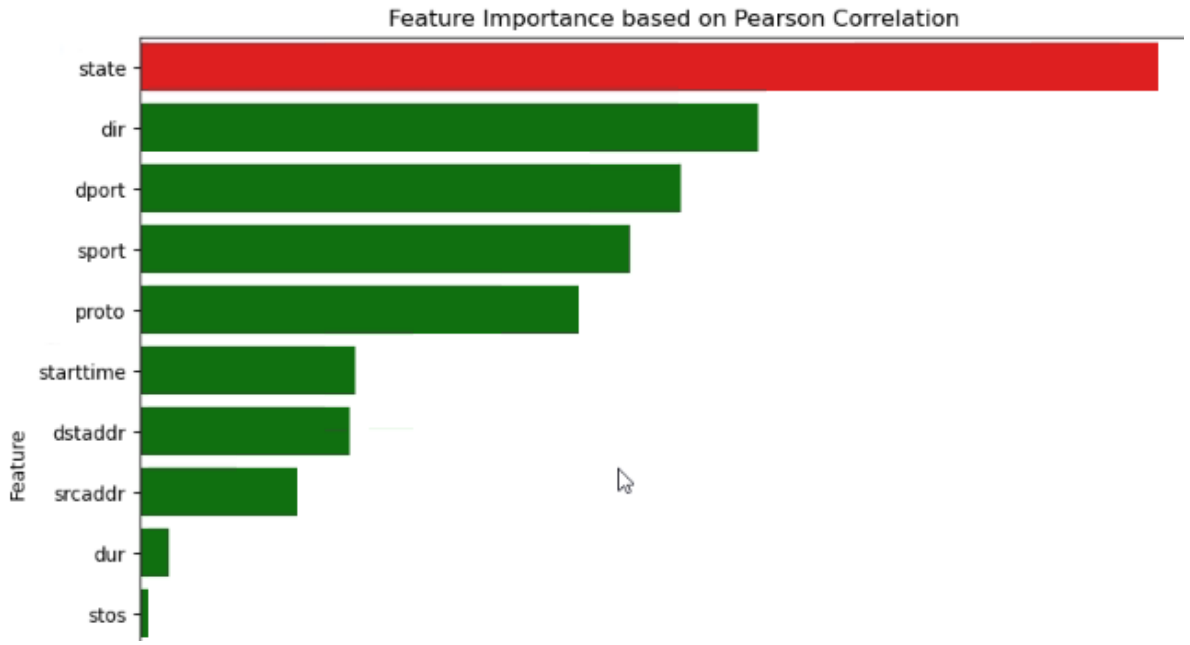
5. Lasso Logistic Regression

	Feature	Importance
0	state	2.902901e-02
1	tot_pkts	1.477215e-03
2	proto	9.856483e-04
3	dir	7.919819e-04
4	tot_bytes	4.990482e-04
5	src_bytes	4.942588e-04
6	dur	3.657647e-04
7	starttime	2.789779e-04
8	dport	1.220673e-04
9	sport	1.072745e-04
10	srcaddr	3.433103e-05
11	dstaddr	1.363853e-05
12	stos	8.140254e-07
13	dtos	3.156339e-07



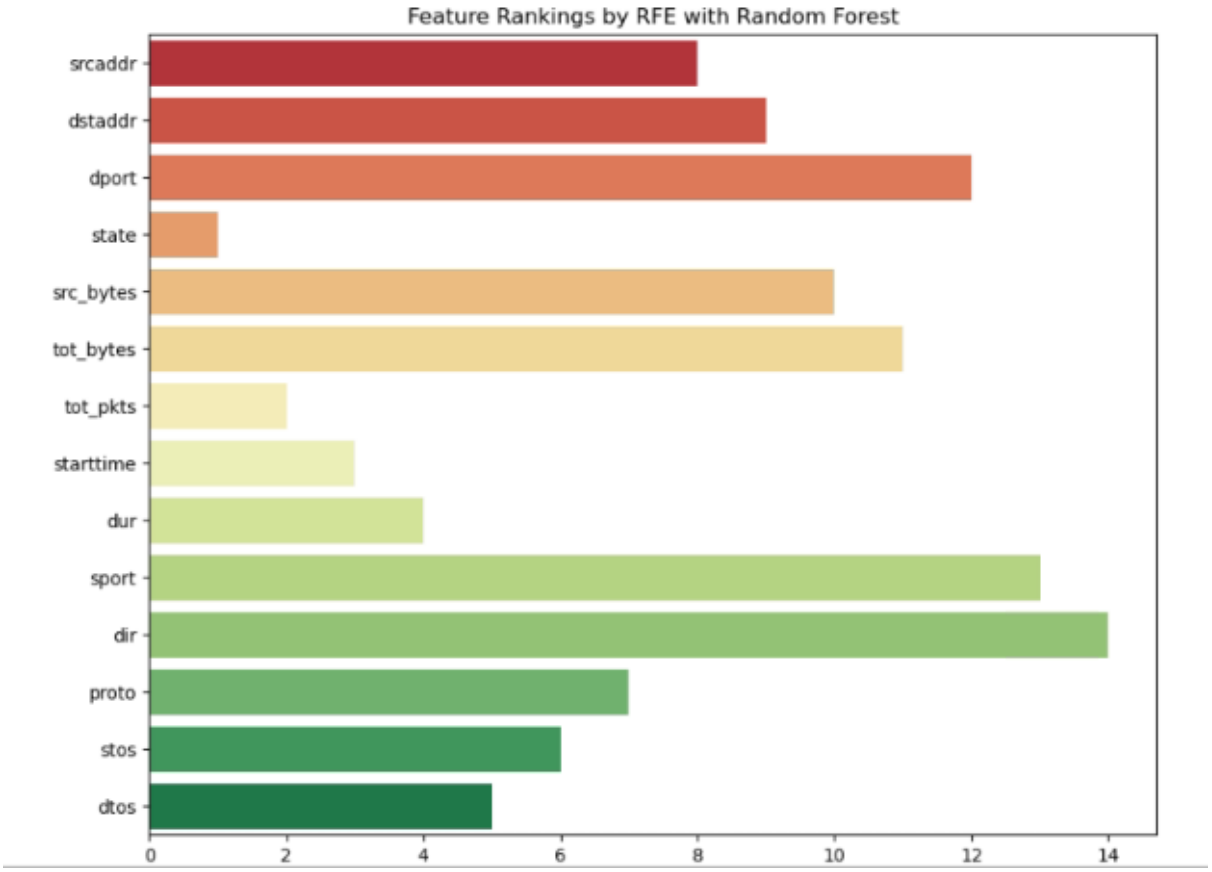
6. Pearson Correlation

	Feature	Correlation
8	state	0.064218
5	dir	0.038969
7	dport	0.034142
4	sport	0.030876
2	proto	0.027693
0	starttime	0.013588
6	dstaddr	0.013211
3	srcaddr	0.009935
1	dur	0.001815
9	stos	0.000542
10	dtos	0.000254
12	tot_bytes	0.000155
11	tot_pkts	0.000151
13	src_bytes	0.000057



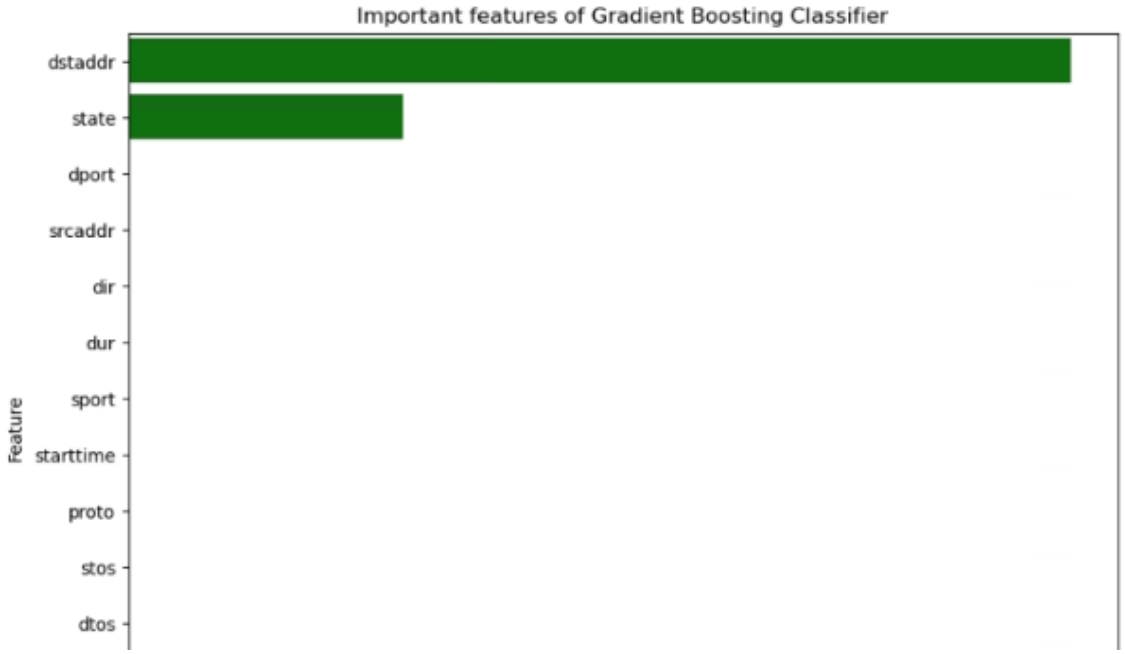
7. Random Forest RFE

	Feature	Importance
0	dir	14
1	sport	13
2	dport	12
3	tot_bytes	11
4	src_bytes	10
5	dstaddr	9
6	srcaddr	8
7	proto	7
8	stos	6
9	dtos	5
10	dur	4
11	starttime	3
12	tot_pkts	2
13	state	1

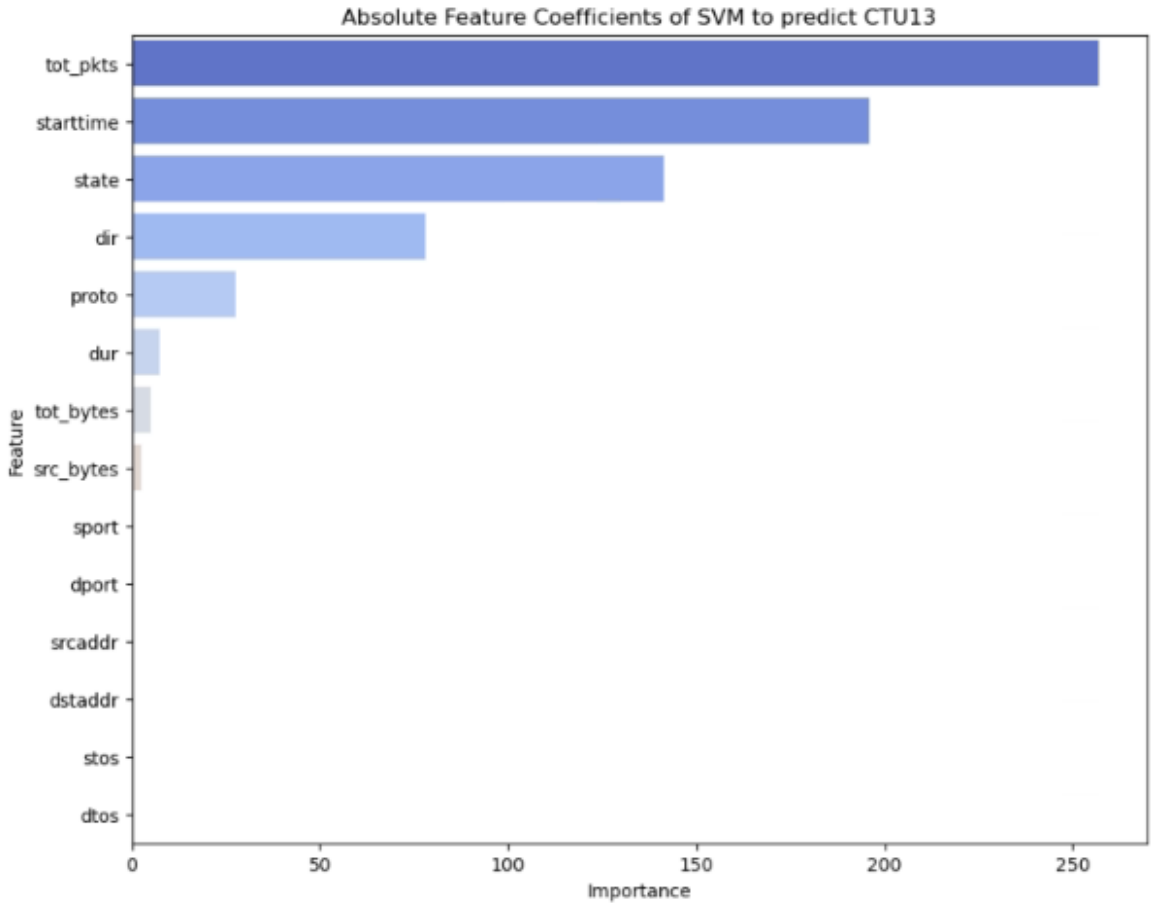


8. Gradient Boosting

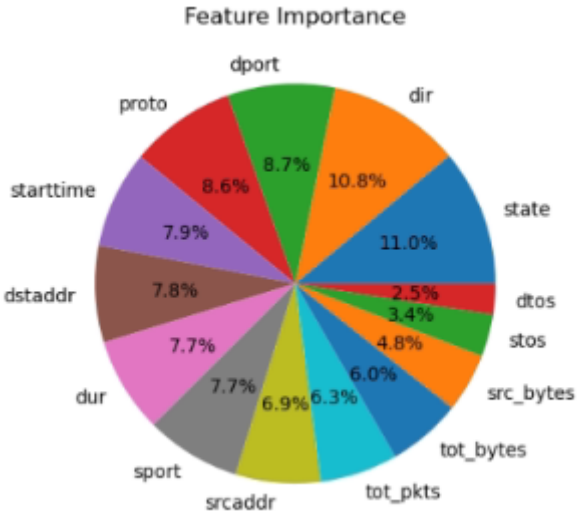
	Feature	Importance
0	dstaddr	7.730102e-01
1	state	2.255046e-01
2	dport	1.263104e-03
3	srcaddr	1.650472e-04
4	dir	5.685737e-05
5	dur	2.339720e-07
6	sport	3.578923e-09
7	starttime	0.000000e+00
8	proto	0.000000e+00
9	stos	0.000000e+00
10	dtos	0.000000e+00
11	tot_pkts	0.000000e+00
12	tot_bytes	0.000000e+00
13	src_bytes	0.000000e+00



9. SVM



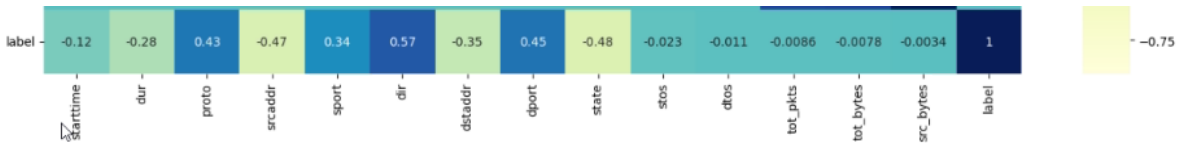
	Feature	Importance
0	tot_pkts	256.810805
1	starttime	196.007531
2	state	141.167798
3	dir	78.070944
4	proto	27.559701
5	dur	7.311818
6	tot_bytes	4.978978
7	src_bytes	2.362930
8	sport	0.493921
9	dport	0.446116
10	srcaddr	0.296564
11	dstaddr	0.075567
12	stos	0.000000
13	dtos	0.000000



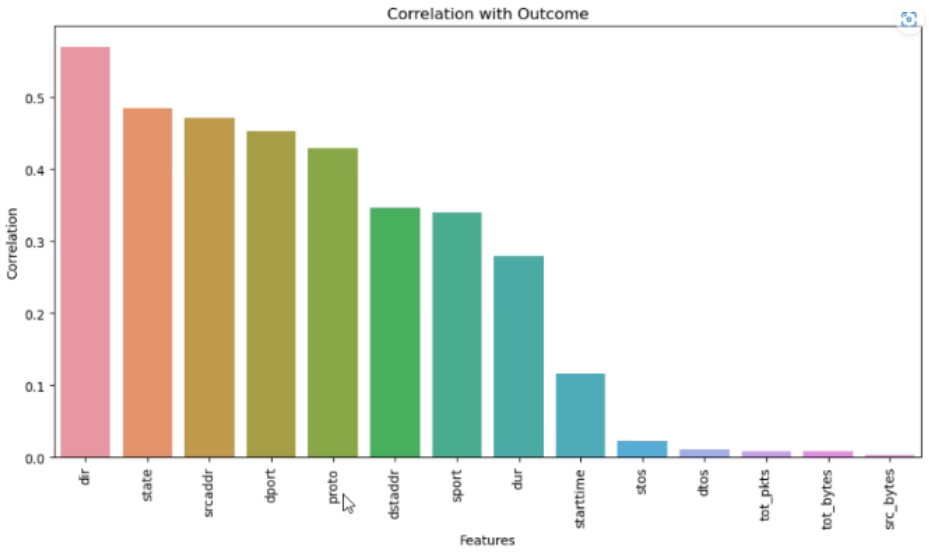
As a result for this case state and dir have the highest rank

For the final case, case 3 I change label into polynomial so I can see that if I make data more complex will it affect anything, I used the same classifier as case 2 which are

1. Correlation Heatmap

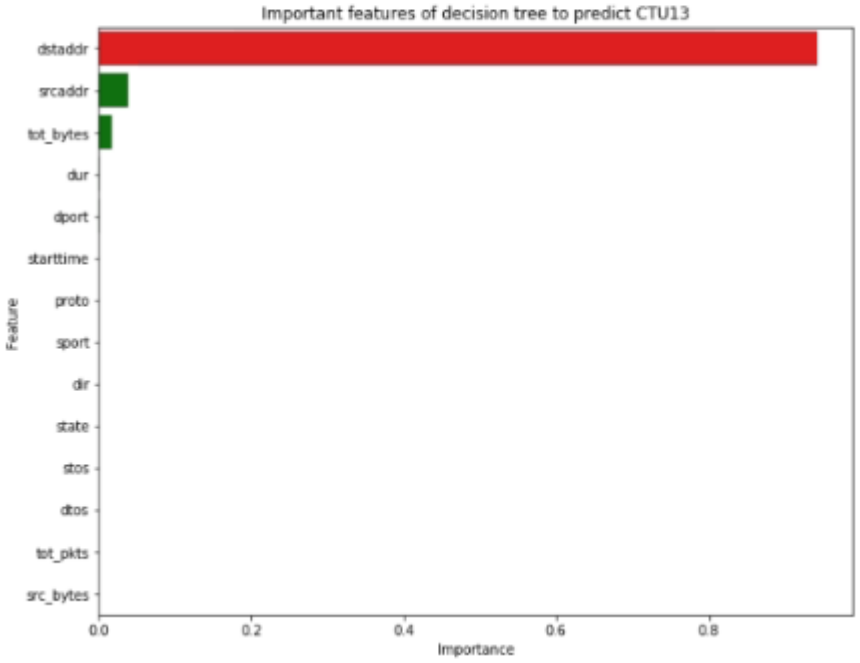


dir	0.570130
state	0.484801
srcaddr	0.471480
dport	0.453442
proto	0.429283
dstaddr	0.346466
sport	0.340477
dur	0.279530
starttime	0.115607
stos	0.023043
dtos	0.011431
tot_pkts	0.008630
tot_bytes	0.007850
src_bytes	0.003408

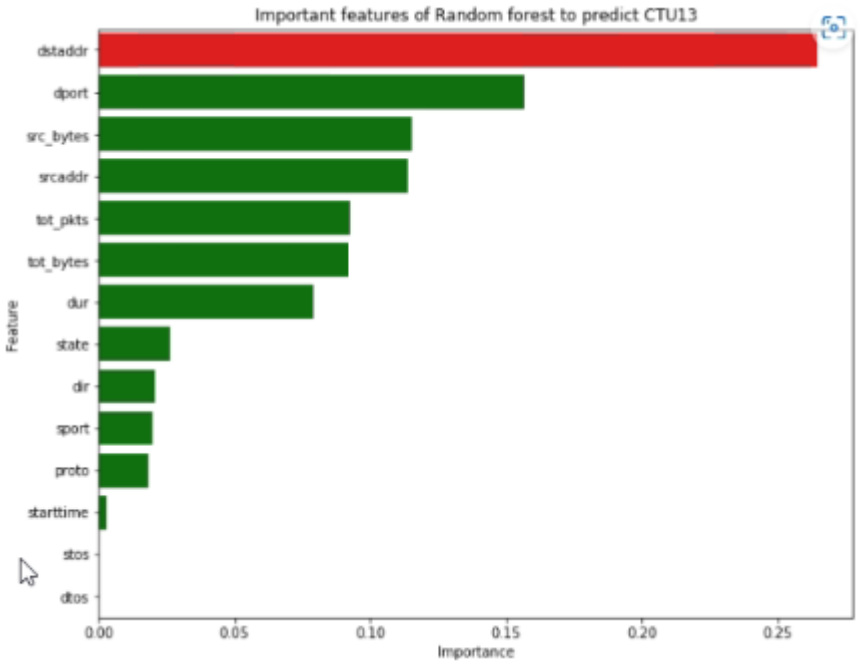


2. Decision Tree

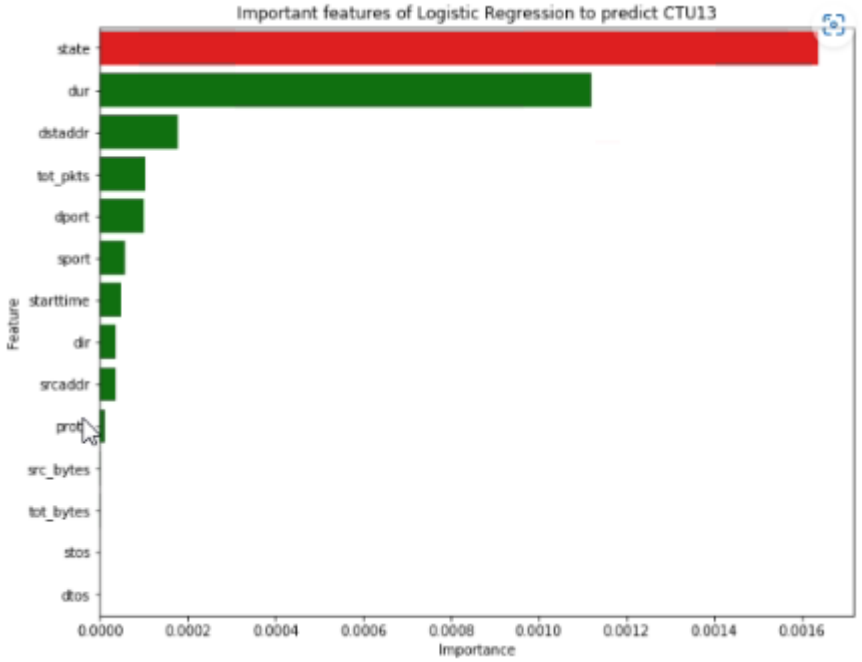
	Feature	Importance
0	dstaddr	0.942468
1	srcaddr	0.037736
2	tot_bytes	0.017535
3	dur	0.001610
4	dport	0.000651
5	starttime	0.000000
6	proto	0.000000
7	sport	0.000000
8	dir	0.000000
9	state	0.000000
10	stos	0.000000
11	dtos	0.000000
12	tot_pkts	0.000000
13	src_bytes	0.000000



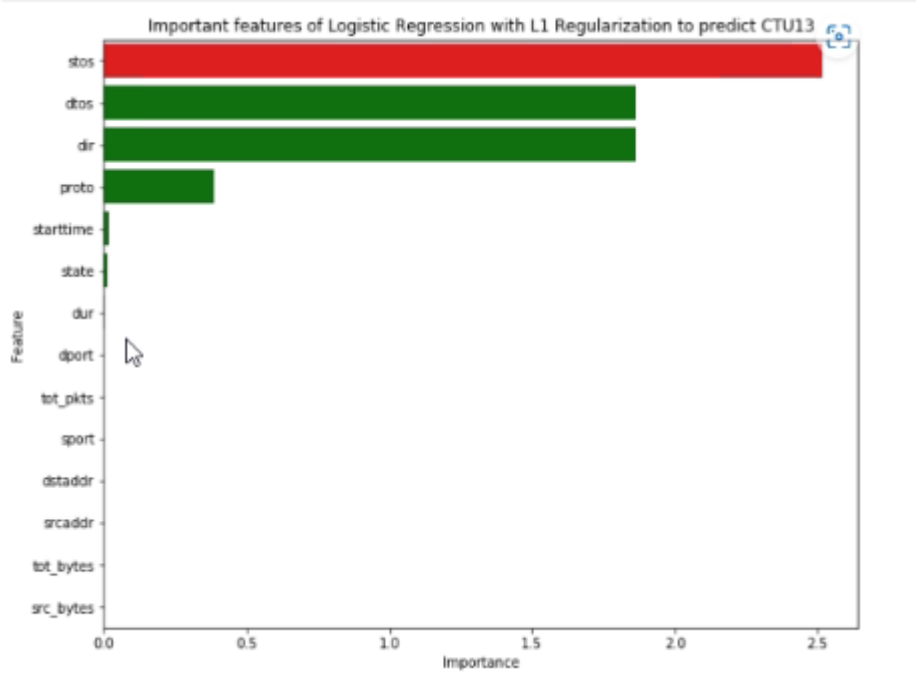
3. Random Forest



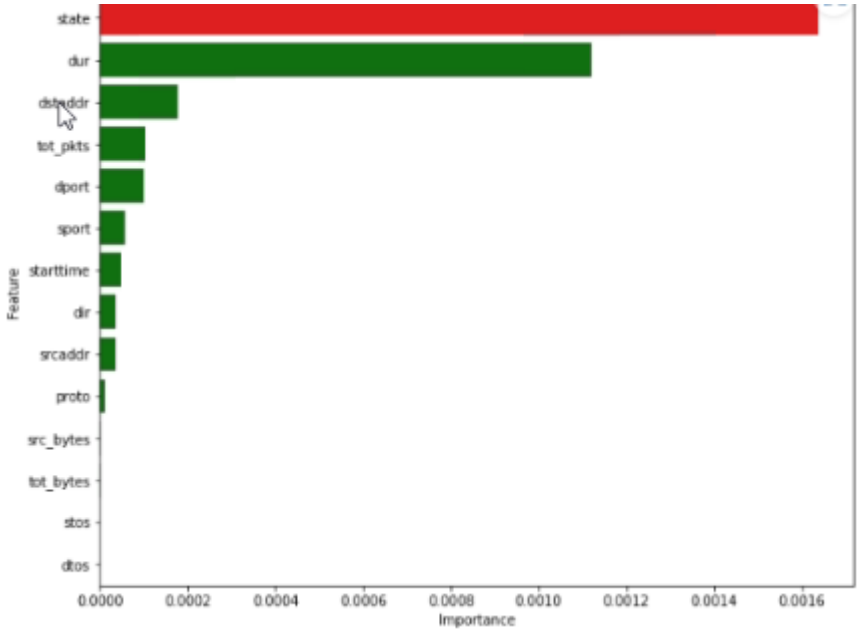
4. Logistic Regression



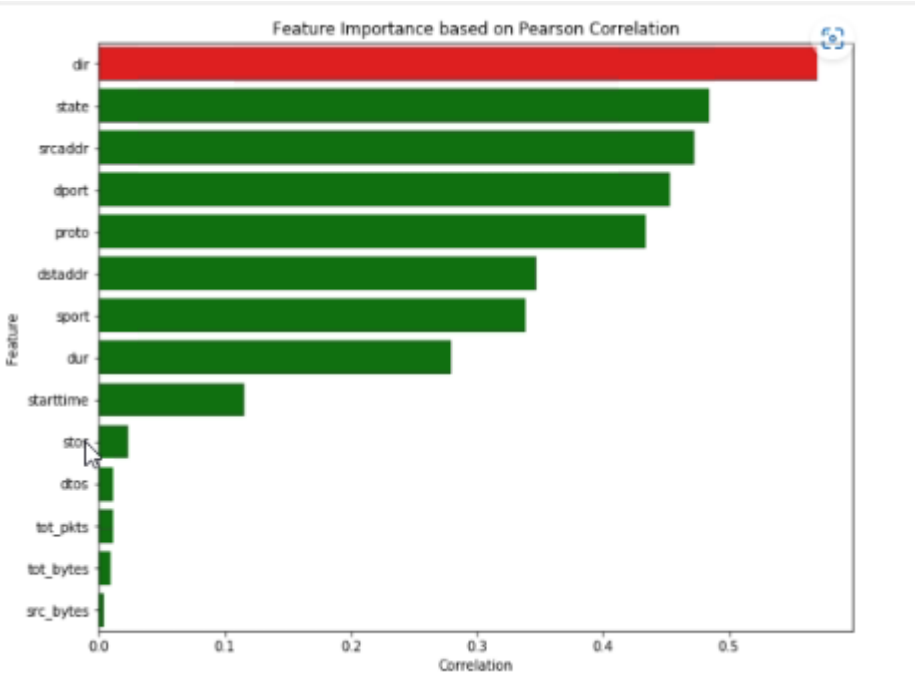
5. Lasso Logistic Regression



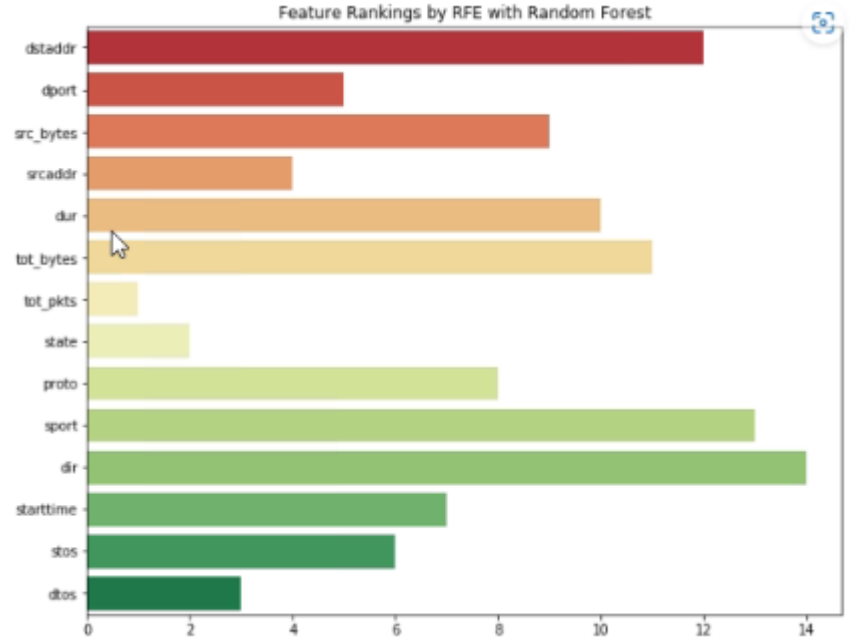
6. Ridge Logistic Regression



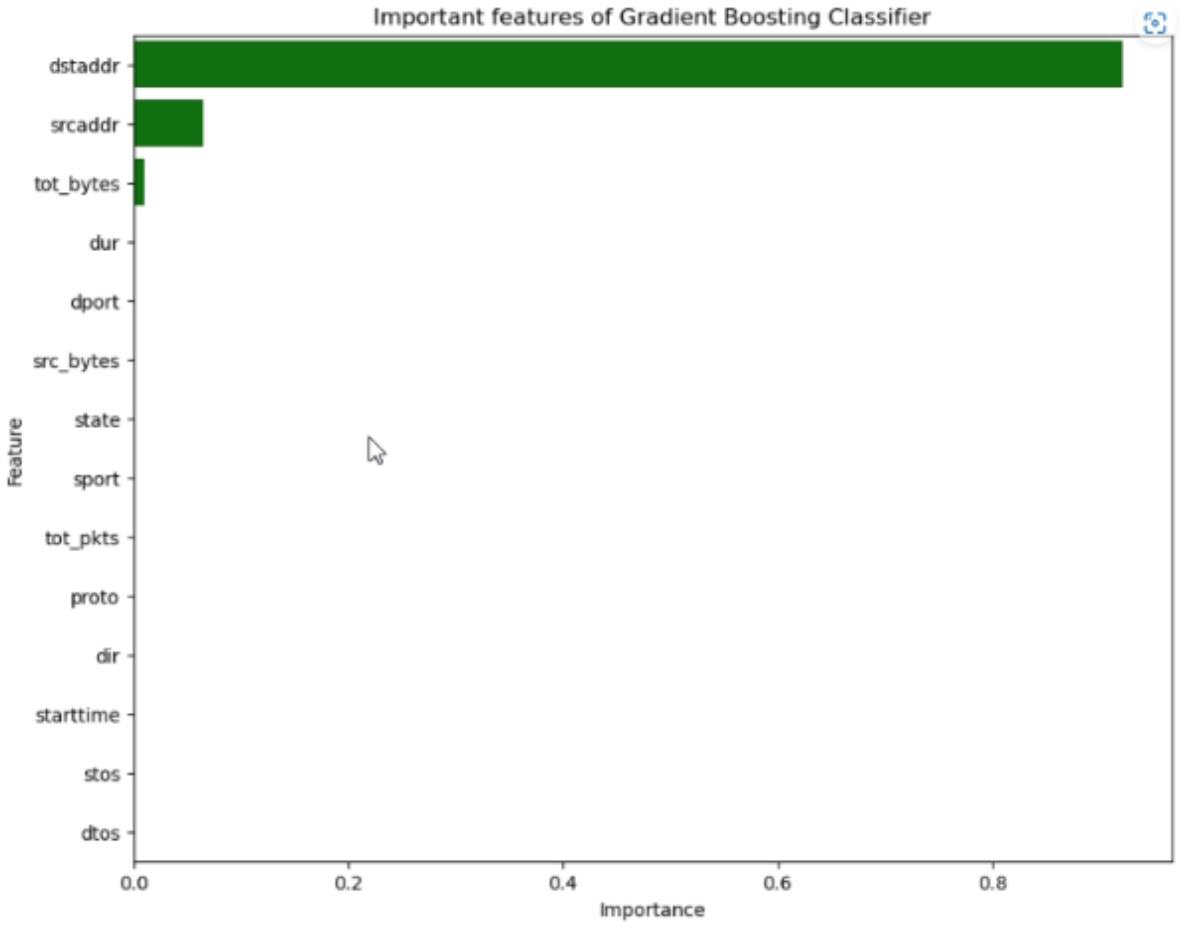
7. Pearson Correlation



8. Random Forest RFE



9. Gradient Boosting



For SVM for case 3 ,the code is still running(Session died for 2 times this will be the last time to run it if it not complete I will cut the svm off) So this is what I have done so far in this week, on next week I will go back to read where the dataset came from to have a fully clear visual of this dataset as Prof. Parinya assign me because I still have not get all the idea from this dataset and it will be hard to move on to next step, also will be prepare for the upcoming presentation.