Weekly report Week 4 (12/6/2023 - 16/6/2023)

As I said last week that I will implement feature extraction and selection this week so on this is my result from feature extraction from the paper, as I said on the past few weeks this CTU13 dataset contain 15 features in total, 1 is the label column and other 14 features need to be clean and normalized so from the paper they are using window width 120 seconds and window stride 60 seconds.

In each time window data are grouped by the source address

For categorical features which are Source IP addresses, Destination IP addresses and Destination ports, there are 2 features as an result for feature extraction which are

- the number of unique occurrences for each sub group
- normalized each subgroup with entropy by

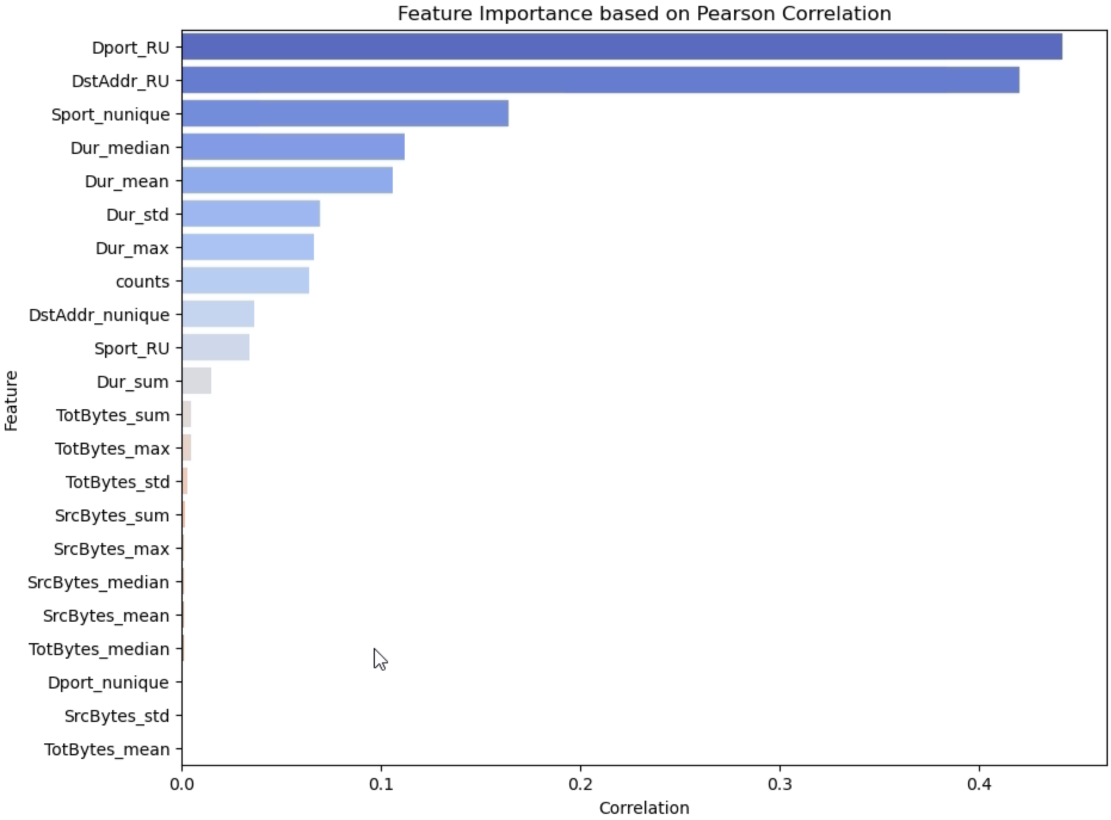
$$E = -\sum_{x_i \in X} p(x_i) \log p(x_i) \quad \text{with} \quad p(x_i) = \frac{\#x_i}{\#X} \quad \text{and} \quad \text{X, the subgroup of the source address}$$

For numerical features which are duration of the communication, total number of exchanged bytes, number of bytes sent by the source, 5 features are extracted which are

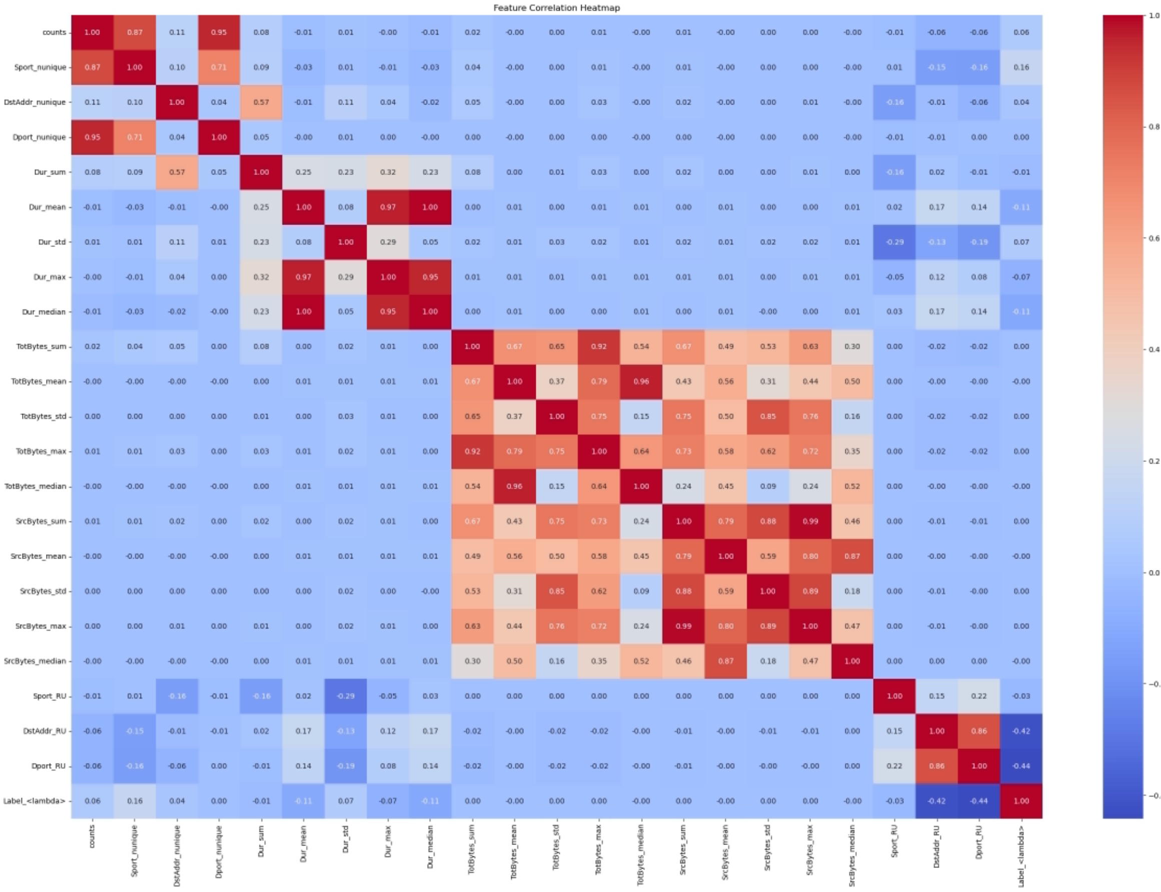
- sum
- mean
- std
- max
- median

For other features apart from these 8 features are dropped

For the feature selection I did the filter method which is pearson correlation



As you can see that DstAddr_RU and Dport_RU rank the highest so the feature that above threshold 0.1 will be select which are Dport_RU, DstAddr_RU, Sport_unique,Dur_median and Dur_mean



and as this heatmap show (Dport_RU,DstAddr_RU) and (Dur_mean,Dur median) have high correlation with each other so the feature is selected based on Label as the result of the filter method the final subset of best features contain Dport_RU, Sport_nunique and Dur_median
So for the next week I will start implement the model section