Project-2 Report

1. What model or algorithm you use?

   Ans: 我使用 scikit-learn 這個 python 套件內的 decision tree model。

2. What features/rules you used for your model?

   Ans: 在每個檔案內，我隨機挑選 100 個 data 當作一個單一的 entry，去統計這些 data 的_score 值、source ip/port, destination ip/port 的種類有沒有超過指定的 threshold，若有超過 threshold 則標記成 1，反之則為 0，而他們的 label 則根據他們所在的檔案決定。以下提供一個例子：
   在檔案 DDoS.json 中我挑選了 100 個 data，這些 data 的_score 值以 1 為最多，且 source ip 的種類超過 20 種，source port 的種類超過 10 種，destination ip 的種類沒有超過 20 種，destination port 的種類沒有超過 10 種，則這 100 筆 data 會形成 1 個 feature_entry = [1,1,1,0,0], label = “DDoS”。
   以上這個步驟皆會在每個檔案中做至少 50 次來形成足夠多的 dataset。

3. Why do you select them?

   Ans: 檔案裡面有很多 feature 可以選，但是我發現_score=1 的只有其中兩個檔案，其他都是=0，所以我把它涵蓋進來。另外我觀察在這些網路攻擊中，source ip/port, destination ip/port 似乎是最重要的，例如在 DDoS 中 destination port 就一定是 22，而在 RDP_bruteforce 中，destination port 就一定是 3389。 而套用在 Test 集上的效果也不錯，所以我就這樣選了。

4. Anything interesting things you find or problems you encounter.

   Ans: 在這個問題中比較困難的地方可能是我模型中 feature_entry 的產生會需要自訂 threshold 來表示說可能 destination port 的種類過多或是過少，這個 threshold 有點難選擇出來，不過我實驗了幾個值後目前對 IP 類型使用總數 20%的 threshold 和 port 類型使用總數 20%的 threshold，應該是比較好的一個值。

5. Result/Accuracy

   0: IP_scan, **1**: port_scan, **2**: DDoS, **3**: RDP_bruteforece, **4**: C&C

	precision	recall	f1-score	support
0	1.00	1.00	1.00	1278
1	1.00	1.00	1.00	2231
2	0.99	1.00	0.99	2305
3	1.00	0.95	0.98	591
4	1.00	1.00	1.00	42
accuracy			1.00	6447
macro avg	1.00	0.99	0.99	6447
weighted avg	1.00	1.00	1.00	6447