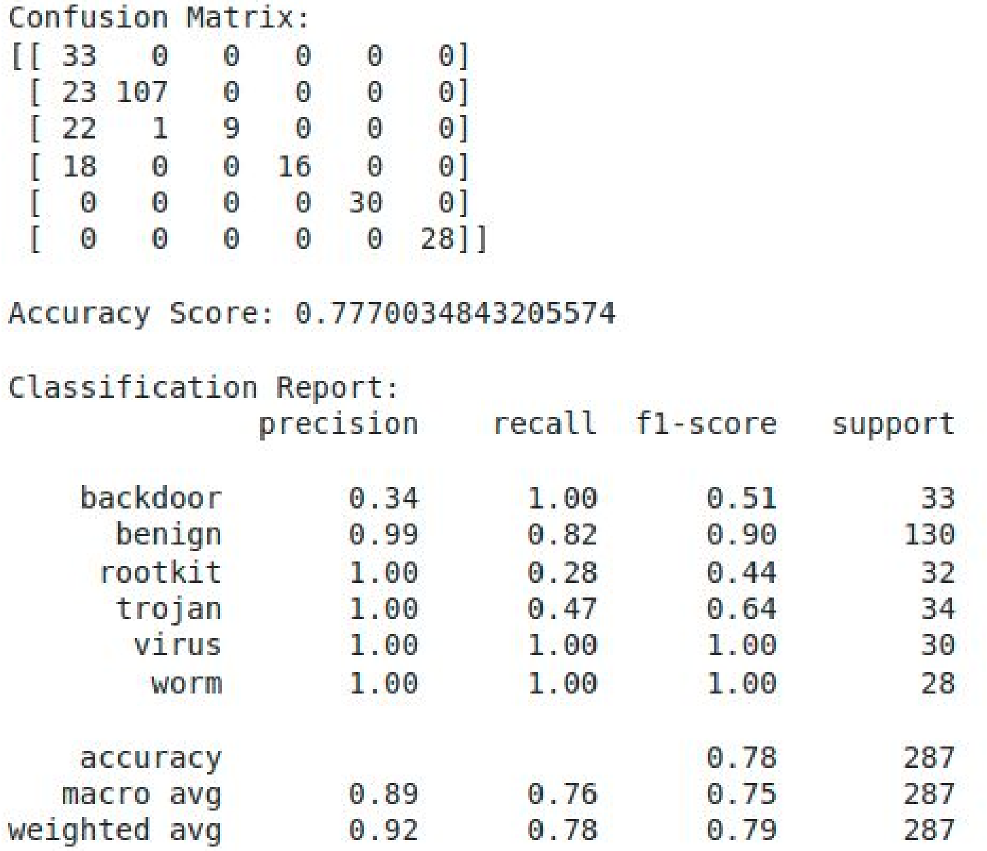
Assignment 1: Support Vector Machine for Multi-class Classification

- Description: This assignment uses the multiclass sample, all 16 features in the hardware performance collect sample were used (Features: bus-cycles, branch-instructions, cache-referencess, node-loads, branch-missed, node-stores, chase-misses, instructions, L1-icache-load-misses, branch-loads, LLC-load-misses, L1-dcache loads, LLC-loads, L1-dchace-stores, L1-dchace-load misses, iTLB-load-misses)
- Results:

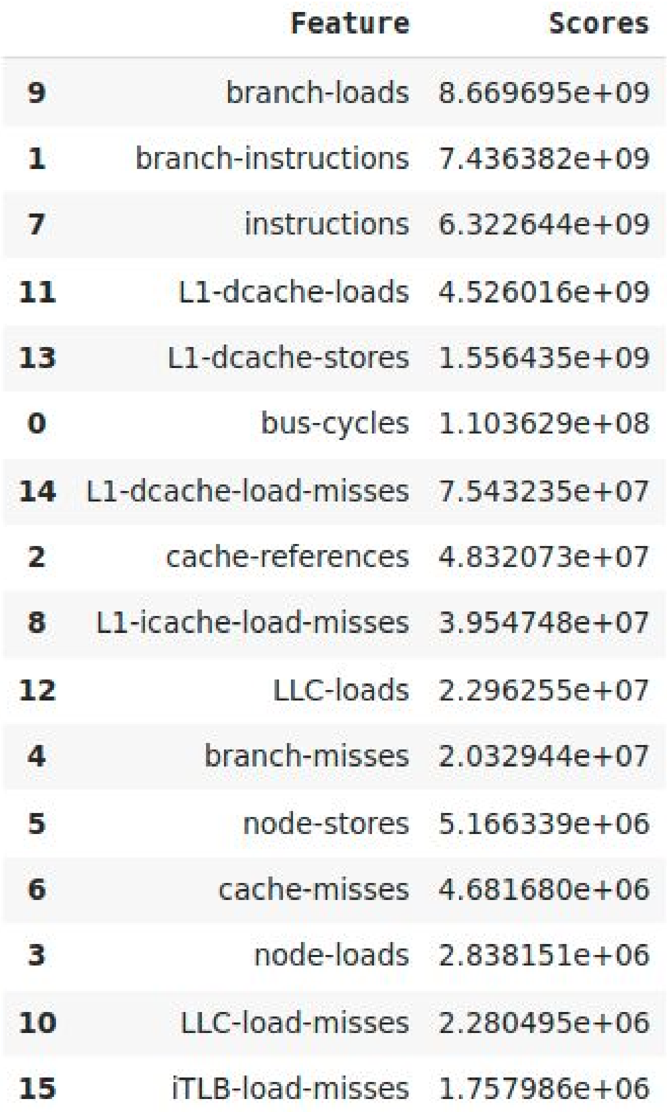


```
Confusion Matrix:
[[ 33   0   0   0   0   0]
 [ 23 107   0   0   0   0]
 [ 22   1   9   0   0   0]
 [ 18   0   0  16   0   0]
 [  0   0   0   0  30   0]
 [  0   0   0   0   0  28]]

Accuracy Score: 0.7770034843205574

Classification Report:
              precision    recall  f1-score   support

    backdoor       0.34      1.00      0.51        33
      benign       0.99      0.82      0.90       130
     rootkit       1.00      0.28      0.44        32
      trojan       1.00      0.47      0.64        34
       virus       1.00      1.00      1.00        30
        worm       1.00      1.00      1.00        28

    accuracy                           0.78       287
   macro avg       0.89      0.76      0.75       287
weighted avg       0.92      0.78      0.79       287
```

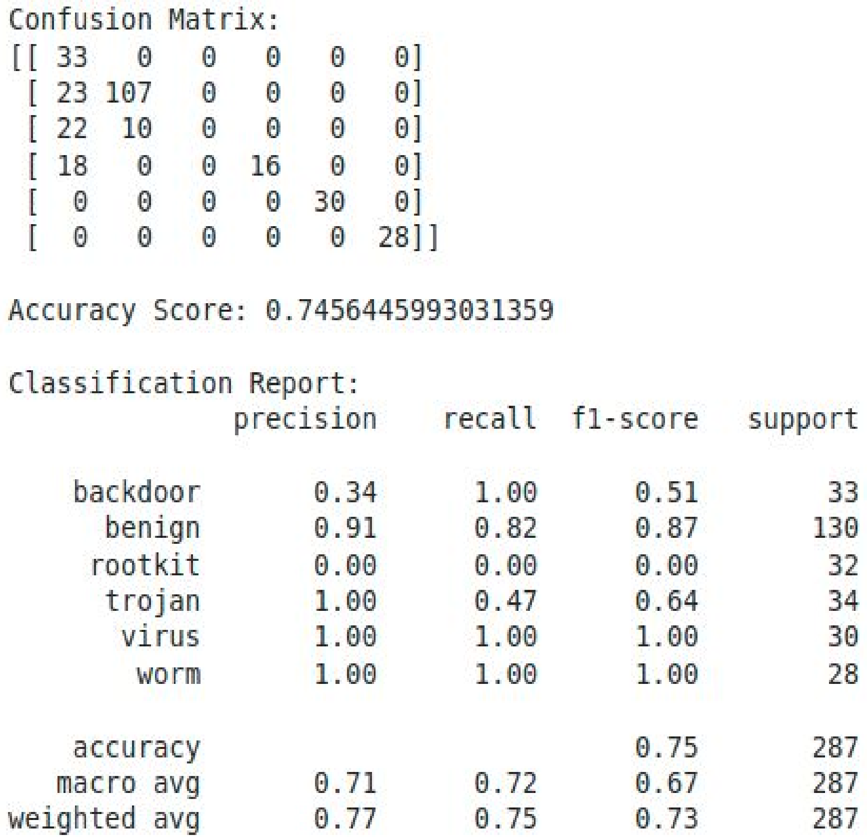
Assignment 2: Support Vector Machine for Multi-class Classification

- Description: The assignment uses the same multiclass sample, all 16 features are taken into consideration, but feature selection is employed by selecting the k best features based on k highest-scores, the features are compared using the chi-squared statistic. In this model the k = 8 features were used, but k can be modified.
- Results:

Feature rank:



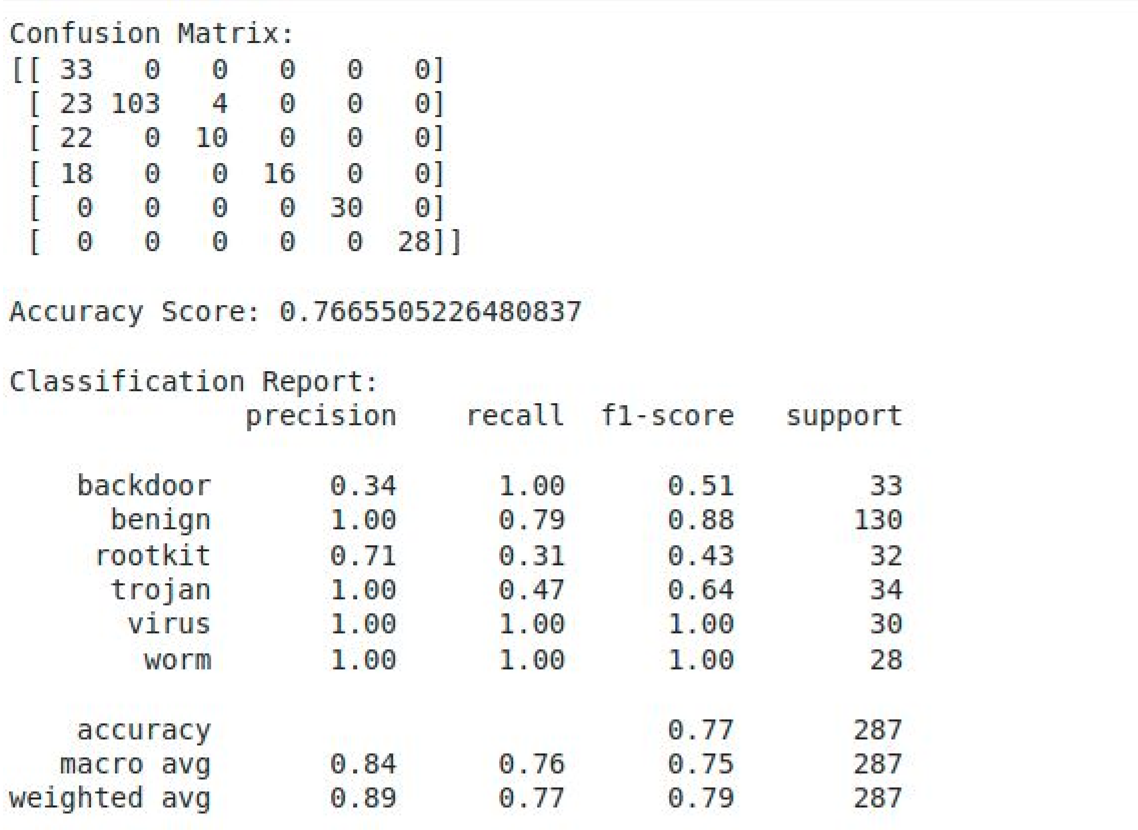
	Feature	Scores
9	branch-loads	8.669695e+09
1	branch-instructions	7.436382e+09
7	instructions	6.322644e+09
11	L1-dcache-loads	4.526016e+09
13	L1-dcache-stores	1.556435e+09
0	bus-cycles	1.103629e+08
14	L1-dcache-load-misses	7.543235e+07
2	cache-references	4.832073e+07
8	L1-icache-load-misses	3.954748e+07
12	LLC-loads	2.296255e+07
4	branch-misses	2.032944e+07
5	node-stores	5.166339e+06
6	cache-misses	4.681680e+06
3	node-loads	2.838151e+06
10	LLC-load-misses	2.280495e+06
15	iTLB-load-misses	1.757986e+06



```
Confusion Matrix:
[[ 33   0   0   0   0   0]
 [ 23 107   0   0   0   0]
 [ 22  10   0   0   0   0]
 [ 18   0   0  16   0   0]
 [  0   0   0   0  30   0]
 [  0   0   0   0   0  28]]

Accuracy Score: 0.7456445993031359

Classification Report:
              precision    recall  f1-score   support

    backdoor       0.34      1.00      0.51        33
      benign       0.91      0.82      0.87       130
     rootkit       0.00      0.00      0.00        32
      trojan       1.00      0.47      0.64        34
       virus       1.00      1.00      1.00        30
        worm       1.00      1.00      1.00        28

    accuracy                           0.75       287
   macro avg       0.71      0.72      0.67       287
weighted avg       0.77      0.75      0.73       287
```

Assignment 3: Naive Bayes for Multi-class Classification with Feature Selection

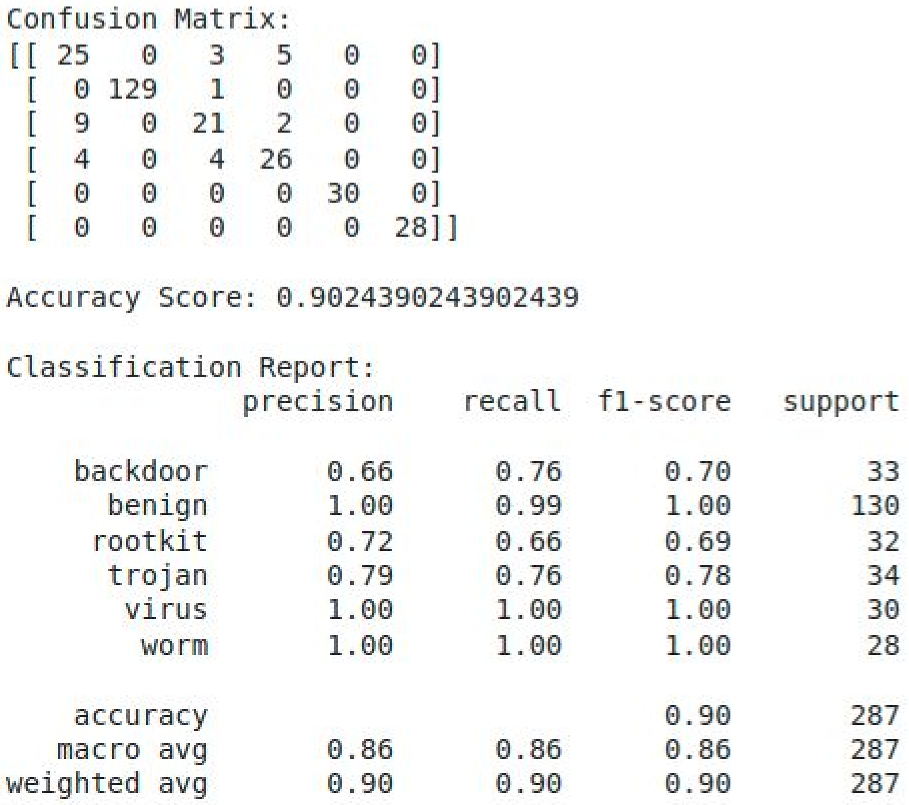
- Description: Again using the hardware performance collect sample with 16 features, we use another algorithm, the Naive Bayes classifier, we employ the same feature selection method as we did above to rank the features accordingly, and use 14 features.
- Results:
  Feature ranks: **the same as above**



```
Confusion Matrix:
[[ 33   0   0   0   0   0]
 [ 23 103   4   0   0   0]
 [ 22   0  10   0   0   0]
 [ 18   0   0  16   0   0]
 [  0   0   0   0  30   0]
 [  0   0   0   0   0  28]]

Accuracy Score: 0.7665505226480837

Classification Report:
              precision    recall  f1-score   support

    backdoor       0.34      1.00      0.51        33
      benign       1.00      0.79      0.88       130
     rootkit       0.71      0.31      0.43        32
      trojan       1.00      0.47      0.64        34
       virus       1.00      1.00      1.00        30
        worm       1.00      1.00      1.00        28

    accuracy                           0.77       287
   macro avg       0.84      0.76      0.75       287
weighted avg       0.89      0.77      0.79       287
```

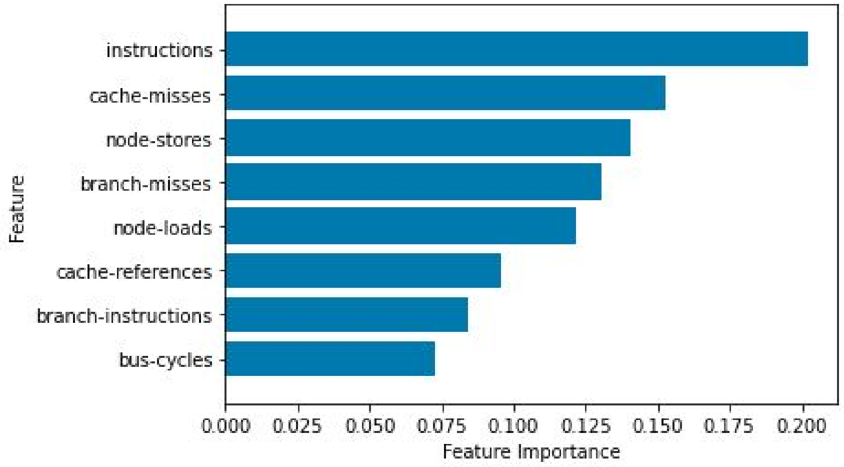
Assignment 4: Gradient Boosted Decision Tree for Multi-class Classification

- Description: Utilizing the same data set that has been used above, and the same method of feature selection, we now employ an ensemble machine learning algorithm to see if we can get any better results using the top 8 features based on the ranking above.
- Results:



```
Confusion Matrix:
[[ 25   0   3   5   0   0]
 [  0 129   1   0   0   0]
 [  9   0  21   2   0   0]
 [  4   0   4  26   0   0]
 [  0   0   0   0  30   0]
 [  0   0   0   0   0  28]]

Accuracy Score: 0.9024390243902439

Classification Report:
              precision    recall  f1-score   support

    backdoor       0.66      0.76      0.70        33
      benign       1.00      0.99      1.00       130
     rootkit       0.72      0.66      0.69        32
      trojan       0.79      0.76      0.78        34
       virus       1.00      1.00      1.00        30
        worm       1.00      1.00      1.00        28

    accuracy                           0.90       287
   macro avg       0.86      0.86      0.86       287
weighted avg       0.90      0.90      0.90       287
```

Feature importances for the 8 features we got from chi-square:


instructions
cache-misses
node-stores
branch-misses
node-loads
cache-references
branch-instructions
bus-cycles
Feature
Feature Importance