Classificação de Tráfego Darknet

Universidade Federal do Paraná
CI1030 Ciência de Dados para Segurança

Jackson Rossi Borguezani

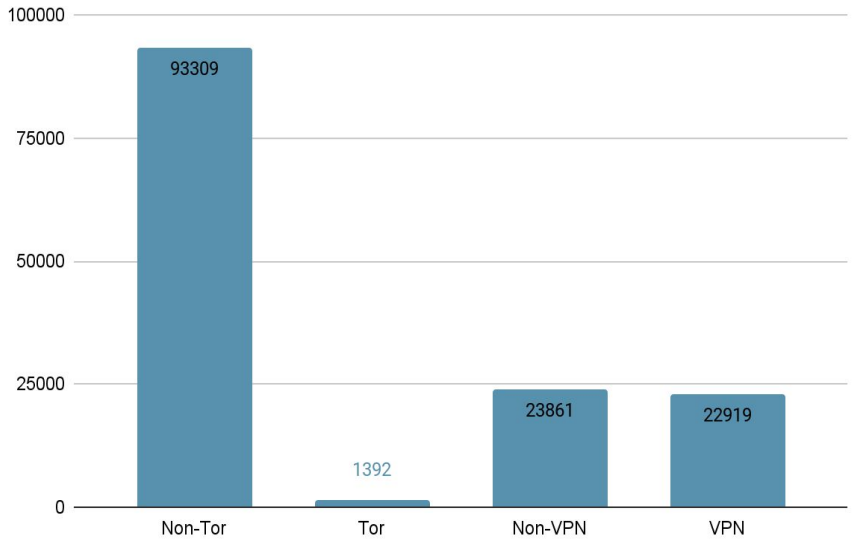
Objetivo

Detectar o tráfego darknet, combinando a categorização de dois conjuntos de dados públicos, o tráfego Tor, Non-Tor e VPN, Non-VPN.

Introdução

- CIC-Darknet2020 [1]
- Arquivo usado: Darknet.csv
- 85 colunas combinadas, contendo dados do tráfego VPN e Tor

Exploração de Dados



- Dados majoritariamente numéricos
- Má distribuição dos dados de tráfego Tor, Non-Tor
- 31 colunas pouco relevantes foram descartadas nesta análise

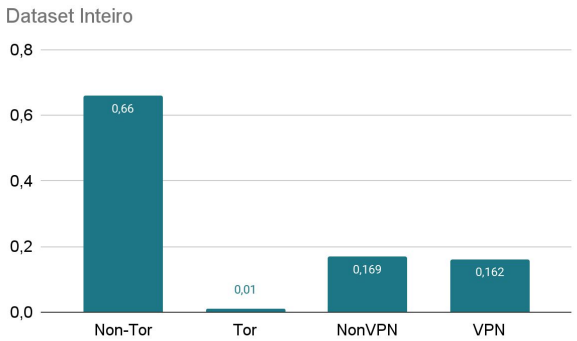
Extração de Características

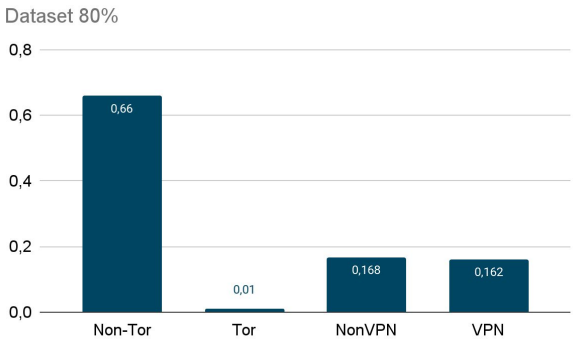
- Matriz de correlação entre 25 colunas
- Remoção de 8 colunas correlacionados que podem ser ignorados

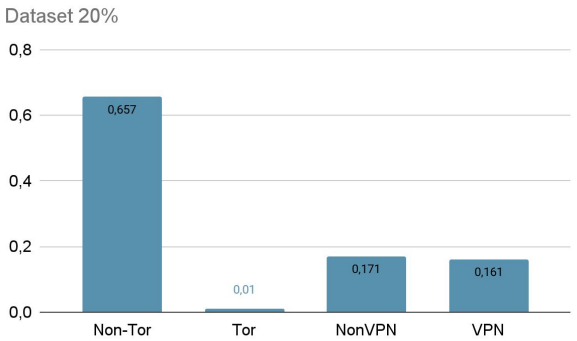
Machine Learning

1. Separar o dataset em 80% e 20%
2. Treinar com a parte de 80%
   a. Validar modelos com split 80/20
   b. Validar modelos com KFold de 5
3. Testar com os 20%

1. Separação do dataset







2. Treinamento

- Normalização usando MinMaxScaler
- Foram treinados 4 modelos

SVM

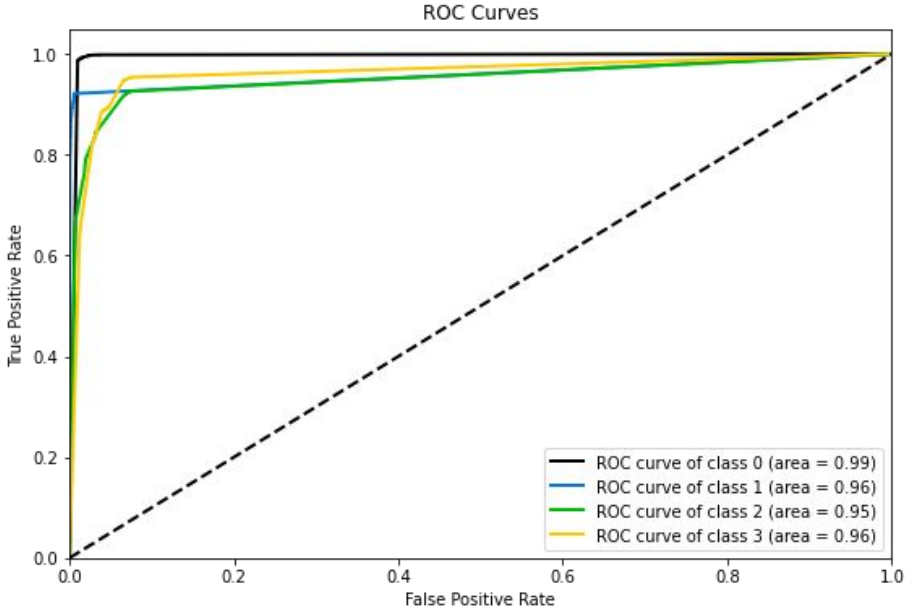
class_weight	‘balanced’
random_state	21
cache_size	500
probability	True
max_iter	1000

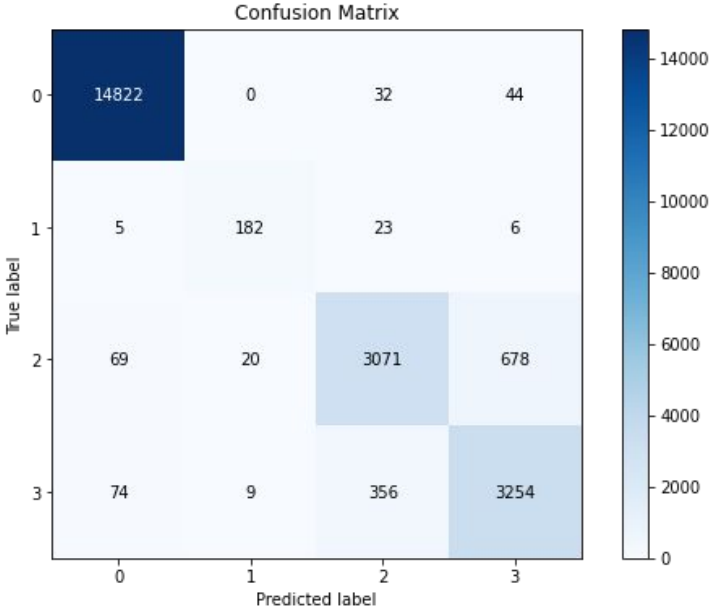
Random Forest

n_estimators	75
random_state	21

KNN

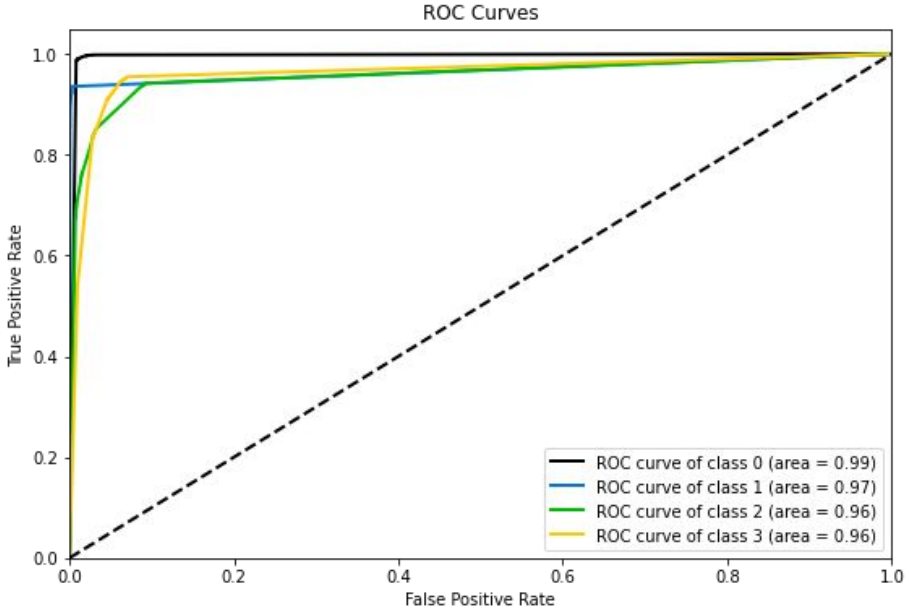
n_neighbors	4
weights	‘distance’

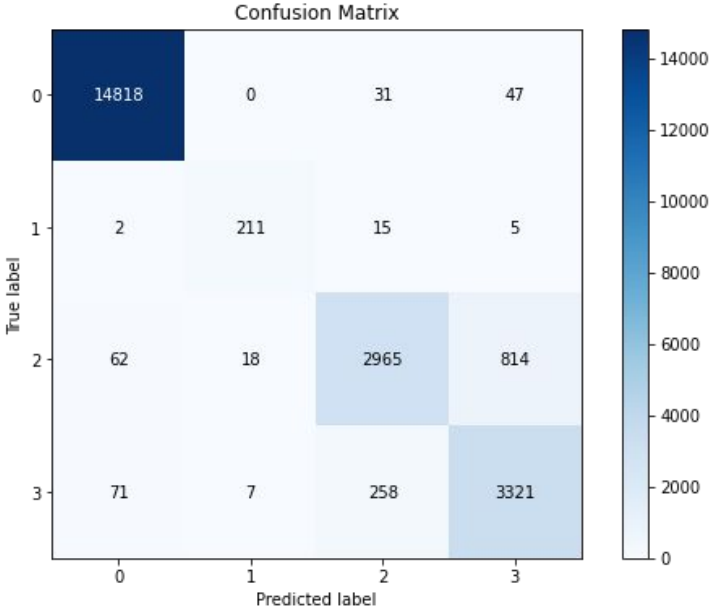

ROC Curves
True Positive Rate
False Positive Rate
0.0
0.2
0.4
0.6
0.8
1.0
ROC curve of class 0 (area = 0.99)
ROC curve of class 1 (area = 0.96)
ROC curve of class 2 (area = 0.95)
ROC curve of class 3 (area = 0.96)


Confusion Matrix
True label
Predicted label
14822
0
32
44
5
182
23
6
69
20
3071
678
74
9
356
3254
14000
12000
10000
8000
6000
4000
2000
0

KNN

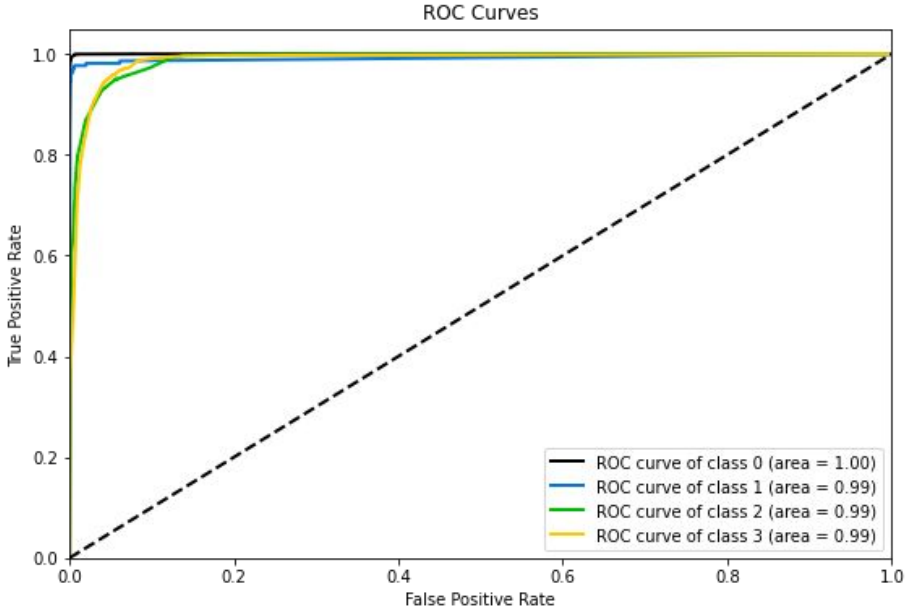
80/20

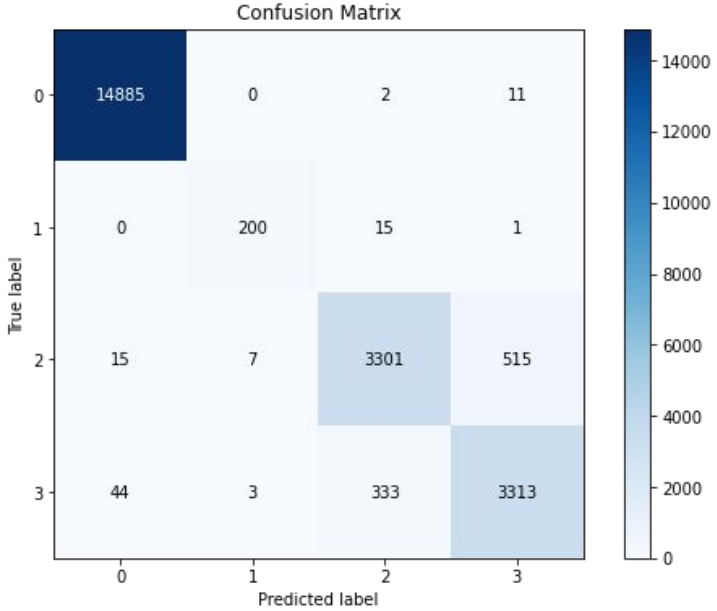

ROC Curves
True Positive Rate
False Positive Rate
0.0
0.2
0.4
0.6
0.8
1.0
ROC curve of class 0 (area = 0.99)
ROC curve of class 1 (area = 0.97)
ROC curve of class 2 (area = 0.96)
ROC curve of class 3 (area = 0.96)


Confusion Matrix
True label
Predicted label
14818
0
31
47
2
211
15
5
62
18
2965
814
71
7
258
3321
14000
12000
10000
8000
6000
4000
2000
0

KNN

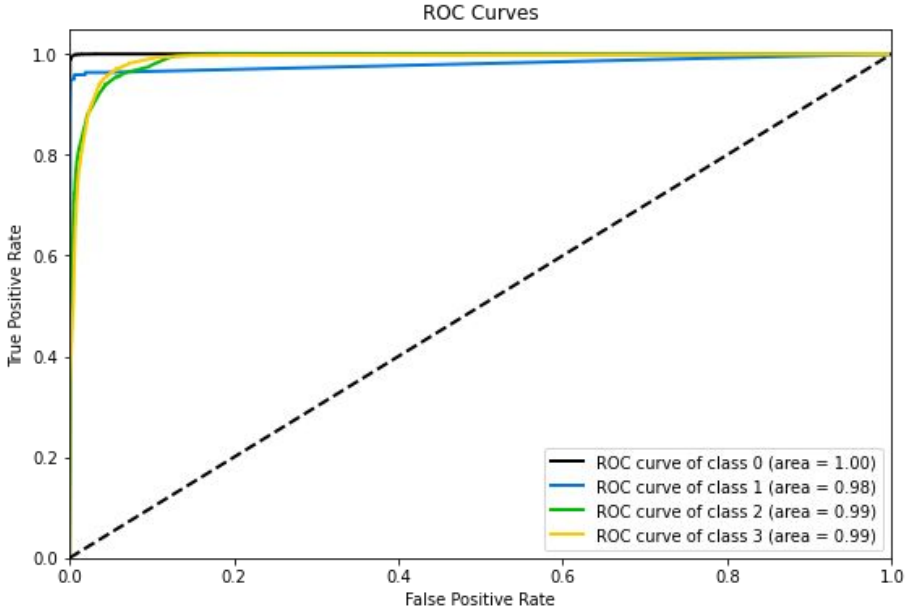
KFold5

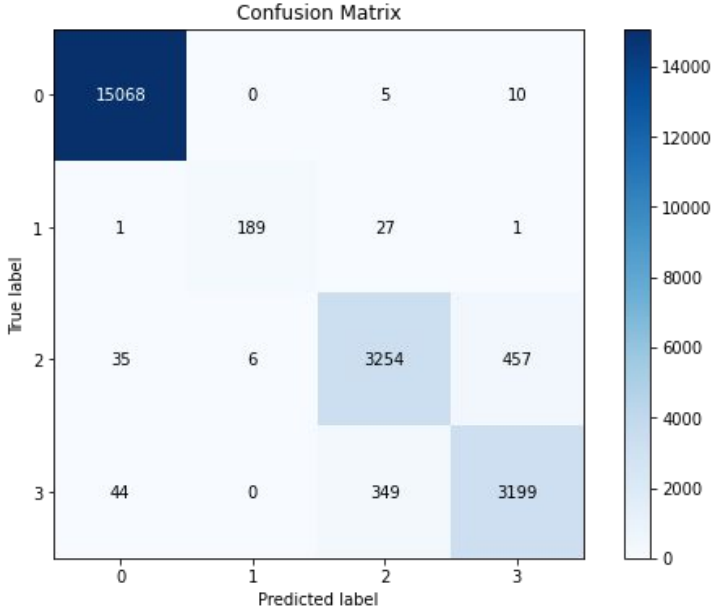




Random Forest

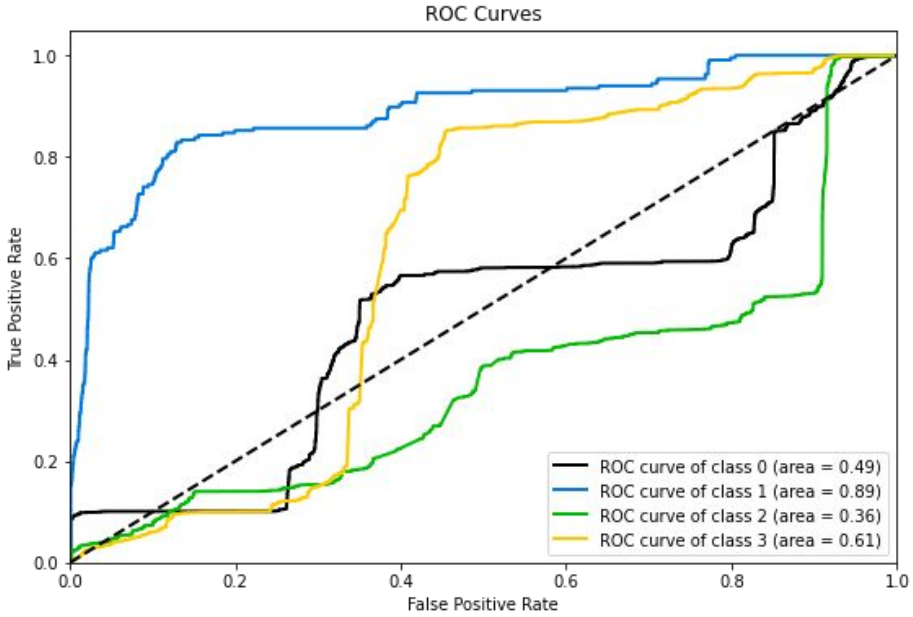
80/20

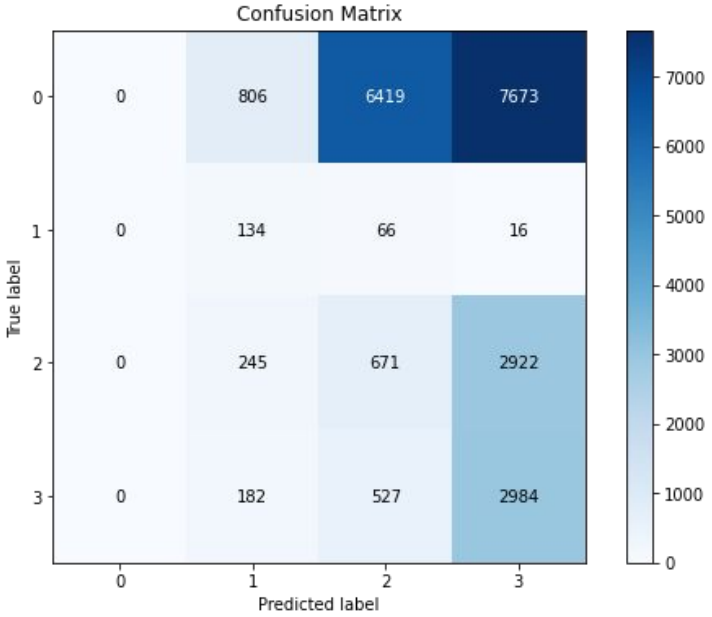




Random Forest

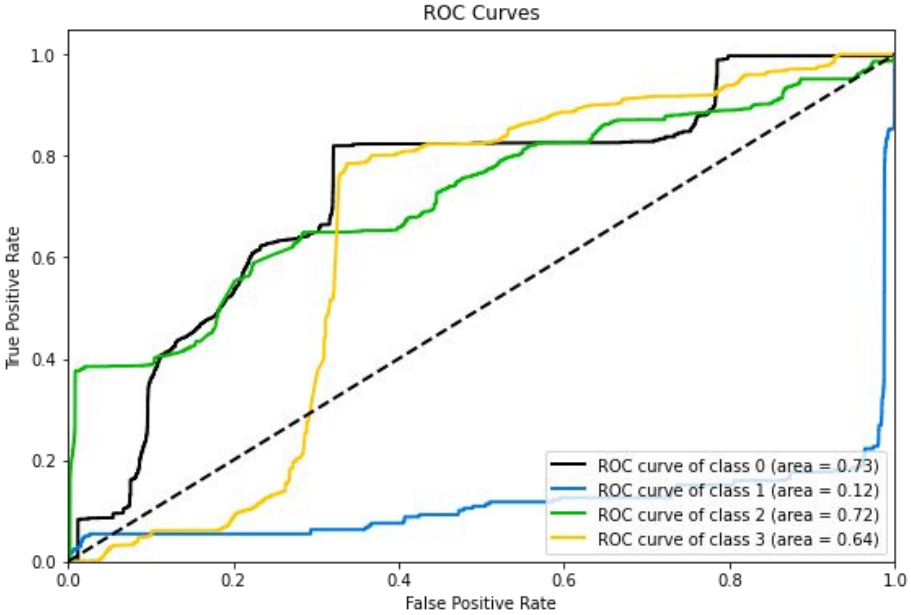
KFold5

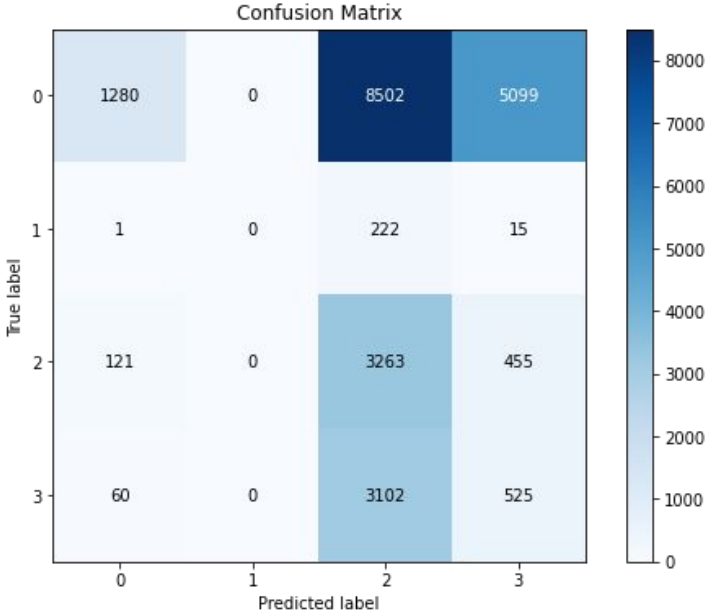

ROC Curves
True Positive Rate
False Positive Rate
0.0
0.2
0.4
0.6
0.8
1.0
ROC curve of class 0 (area = 0.49)
ROC curve of class 1 (area = 0.89)
ROC curve of class 2 (area = 0.36)
ROC curve of class 3 (area = 0.61)


Confusion Matrix
True label
Predicted label
0 806 6419 7673
1 0 134 66 16
2 0 245 671 2922
3 0 182 527 2984
0
1
2
3
7000
6000
5000
4000
3000
2000
1000
0

SVM

80/20


ROC Curves
True Positive Rate
False Positive Rate
0.0
0.2
0.4
0.6
0.8
1.0
ROC curve of class 0 (area = 0.73)
ROC curve of class 1 (area = 0.12)
ROC curve of class 2 (area = 0.72)
ROC curve of class 3 (area = 0.64)

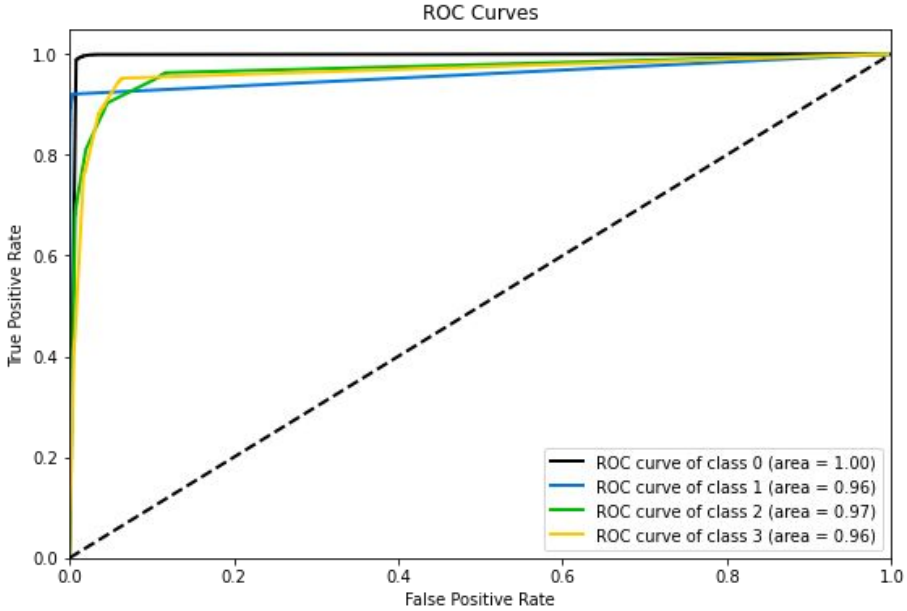

Confusion Matrix
True label
Predicted label
1280
0
8502
5099
1
0
222
15
121
0
3263
455
60
0
3102
525
8000
7000
6000
5000
4000
3000
2000
1000
0

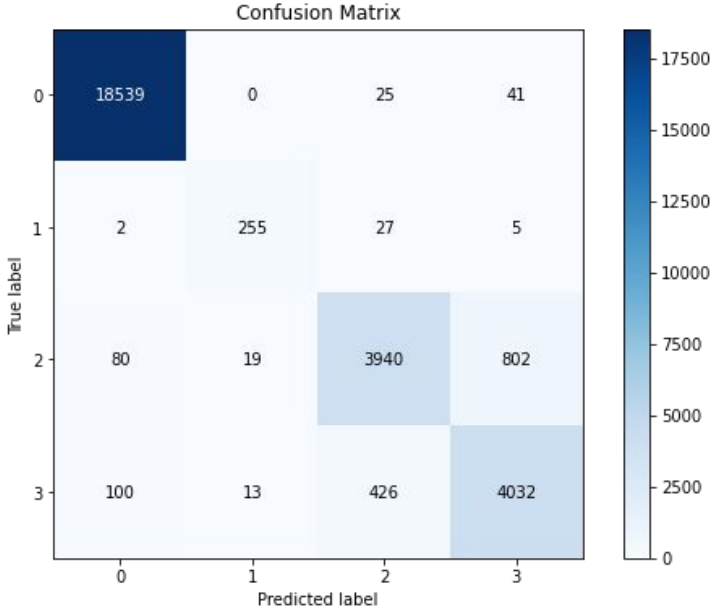
SVM

KFold5

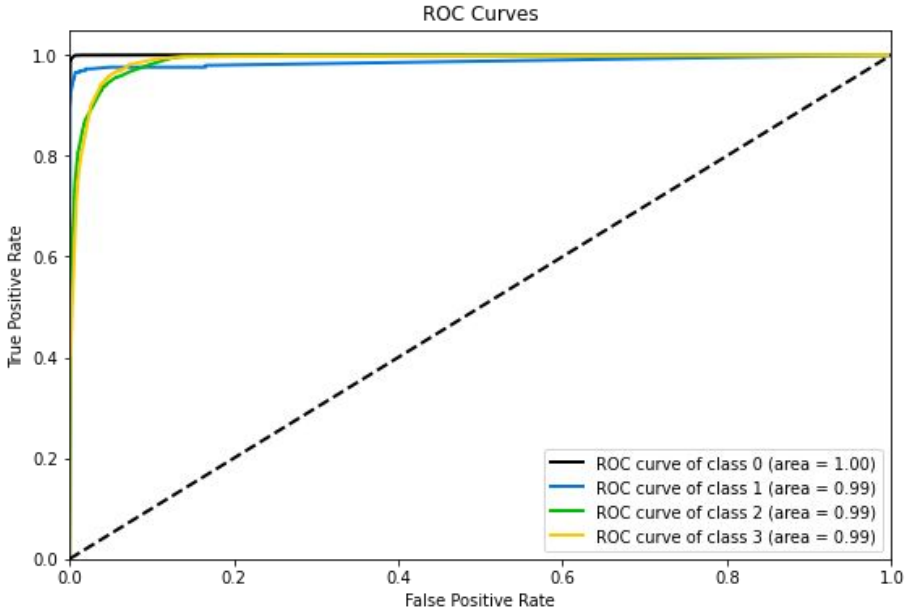
3. Teste

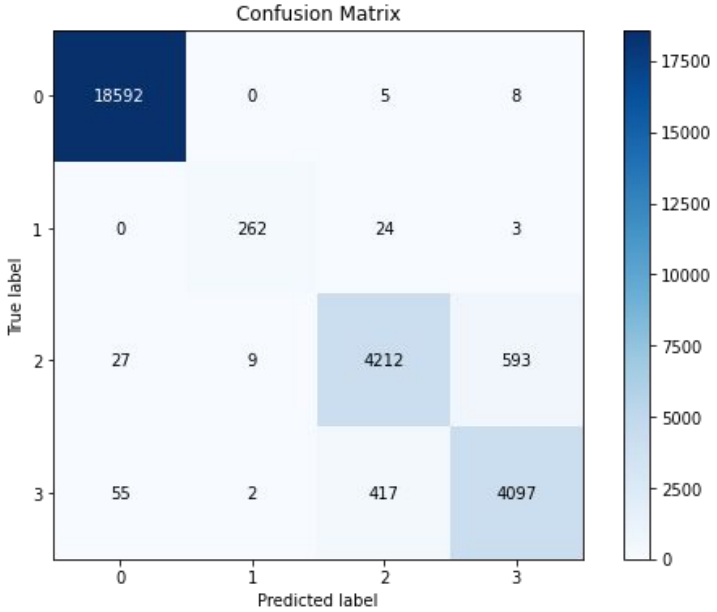
- Usando a parte de 20% como dados de produção
- Aplicado aos modelos treinados com a porção inteira de 80%


ROC Curves
True Positive Rate
False Positive Rate
0.0
0.2
0.4
0.6
0.8
1.0
ROC curve of class 0 (area = 1.00)
ROC curve of class 1 (area = 0.96)
ROC curve of class 2 (area = 0.97)
ROC curve of class 3 (area = 0.96)

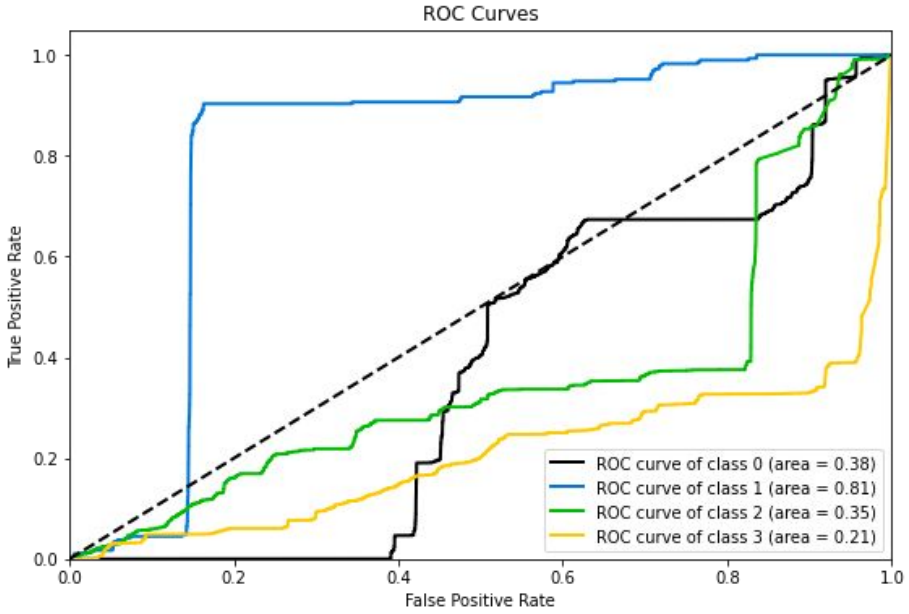

Confusion Matrix
True label
Predicted label
18539
0
25
41
2
255
27
5
80
19
3940
802
100
13
426
4032
17500
15000
12500
10000
7500
5000
2500
0

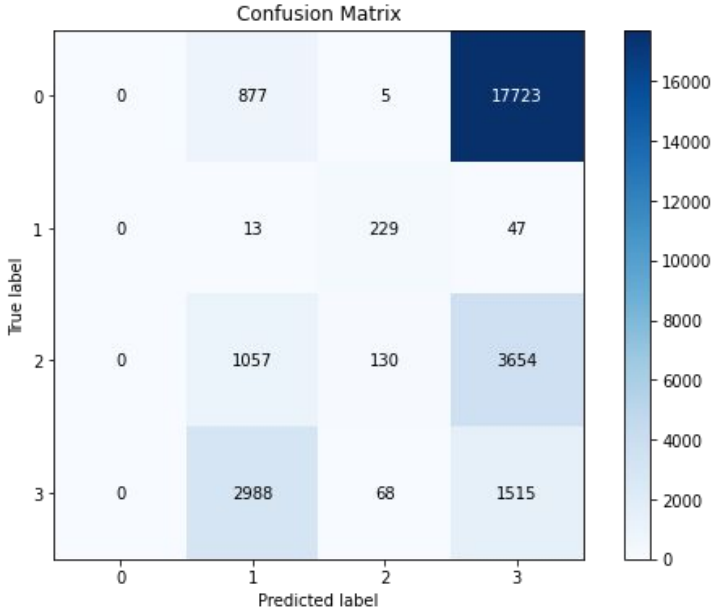
KNN





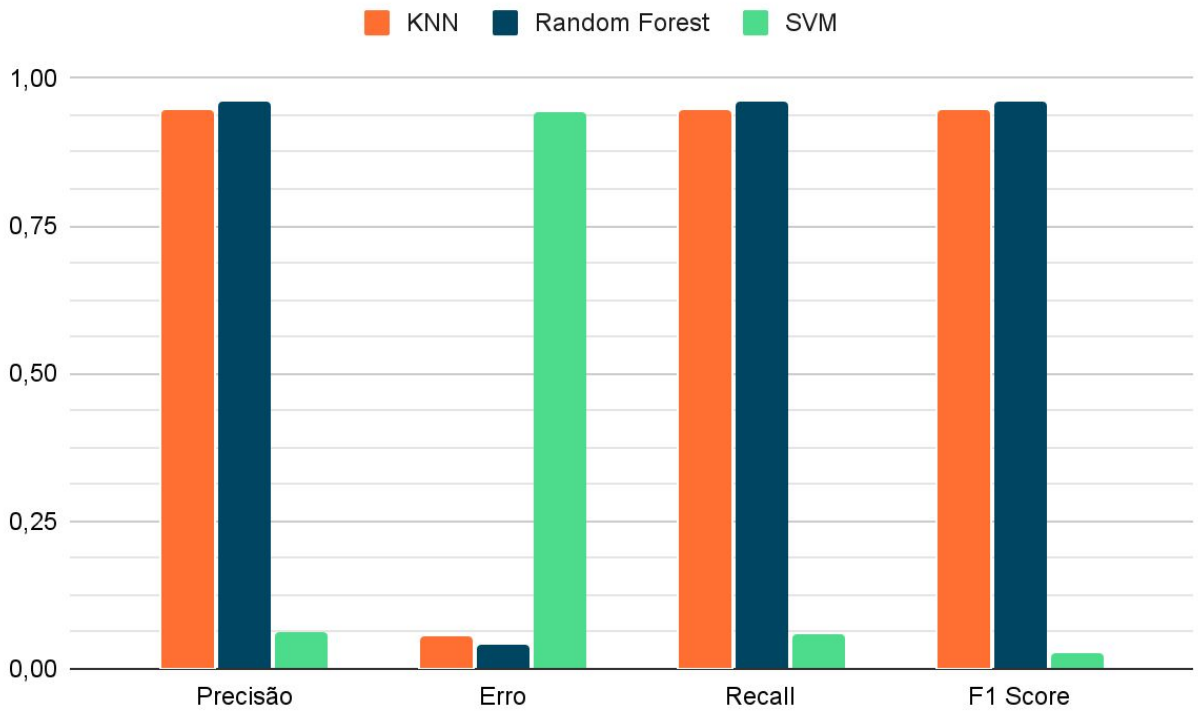
Random Forest


ROC Curves
True Positive Rate
False Positive Rate
0.0
0.2
0.4
0.6
0.8
1.0
ROC curve of class 0 (area = 0.38)
ROC curve of class 1 (area = 0.81)
ROC curve of class 2 (area = 0.35)
ROC curve of class 3 (area = 0.21)


Confusion Matrix
True label
Predicted label
0 877 5 17723
0 13 229 47
0 1057 130 3654
0 2988 68 1515
0 1 2 3
16000
14000
12000
10000
8000
6000
4000
2000
0

SVM

Resultado dos Testes



Conclusão

De modo geral, o Random Forest parece a melhor solução para classificar o tráfego:

- Tem a melhor precisão
- Menor erro
- Menor tempo de predição

O SVM teve um desempenho baixíssimo, já que seu tempo de treino foi muito grande e precisou de limitação de iterações. O fato do dataset ser multiclasse pode ser um agravante no tempo de treino, além do tamanho do dataset. Então o SVM não é viável de forma alguma para esses dados.

Já o KNN teve um tempo de treino melhor que o Random Forest. No entanto, a penalidade acontece no tempo de predição, que é muito maior ao Random Forest.

Assim, o Random Forest é o melhor algoritmo de machine learning entre os três estudados.

Links e Referências

- GitHub: https://github.com/jacksonrossi/ciencia-dados-darknet
- [1] Arash Habibi Lashkari, Gurdip Kaur, and Abir Rahali, “DIDarknet: A Contemporary Approach to Detect and Characterize the Darknet Traffic using Deep Image Learning”, 10th International Conference on Communication and Network Security, Tokyo, Japan, November 2020. https://www.unb.ca/cic/datasets/darknet2020.html