Treinamento DDoS Dataset

Felipe Ribeiro Quiles

Ciência de Dados para Segurança 2021/2

ÍNDICE

Apresentação do dataset

- O dataset é composto por 12794627 datapoints
- Cada datapoint corresponde a um fluxo (ida e volta).
- Os datapoints são compostos de 82 características
- Os datapoints são classificados entre:
  - DDoS
  - Benigno
- O dataset balanceado tem 51% DDoS e 49% Benigno

Processamento dos Dados para o Treinamento

- Diminuição do Dataset para 500 mil datapoints (20% oculto + 80% treino e teste)
- 50% ddos e 50% benigno
- Exclusão de 6 características
  - FlowID
  - IP Destino
  - IP Origem
  - Porta Destino
  - Porta Origem
  - Timestamp
- Troca das classes de “ddos” e “benign” para 1 e 0, respectivamente

Amostra

[6.0, 1724944.0, 3.0, 1.0, 360.0, 120.0, 120.0, 120.0, 120.0, 0.0, 120.0, 120.0, 120.0, 0.0, 278.2699032548303, 2.318915860456919, 574981.3333333334, 2771.371922592371, 578021.0, 572595.0, 1150616.0, 575308.0, 3836.7613947182067, 578021.0, 572595.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 60.0, 20.0, 1.7391868953426894, 0.5797289651142298, 120.0, 120.0, 120.0, 0.0, 0.0, 0.0, 1.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 150.0, 120.0, 120.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 3.0, 360.0, 1.0, 120.0, -1.0, 64.0, 3.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 1]

Ambiente de Treinamento

- Processador
  - Intel(R) Core(TM) i7-8750H CPU @ 2.20GHz
- Memória
  - 16 GB
- Sistema Operacional
  - Linux Mint 20

Método de Treinamento

- Os modelos a serem treinados são baseados em 3 tipos:
  - KNN
  - RandomForest
  - MLP
- Para realizar o treinamento utilizou-se o GridSearchCV
- GridSearchCV é uma ferramenta utilizada para automatizar o processo de ajuste dos parâmetros de um algoritmo
- Para início, decidiu-se que a classe positiva era a “ddos”, marcada como 1 (Problema Binário)
- Treinamentos utilizando taxas de 80/20 e 50/50
- Utilização do K-Folding para comparação de resultados
  - Foi utilizado K = 5
- Escolha do “precision_score” para a métrica do modelo

KNN - Treinamento

- Definição dos parâmetros para o GridSearch
  - 1 Vizinho
  - 3 Vizinhos
  - 5 Vizinhos
- Estatísticas do Treinamento 80/20

```
Melhores parametros do KNeighbors para o precision_score
{'n_neighbors': 1}
Melhor score do KNeighbors para o precision_score
0.9940544827265736
```

- Estatísticas Treinamento 50/50

```
Melhores parametros do KNeighbors para o precision_score
{'n_neighbors': 1}
Melhor score do KNeighbors para o precision_score
0.991343532734313
```

KNN - Treinamento

- Estatísticas do Treinamento 80/20

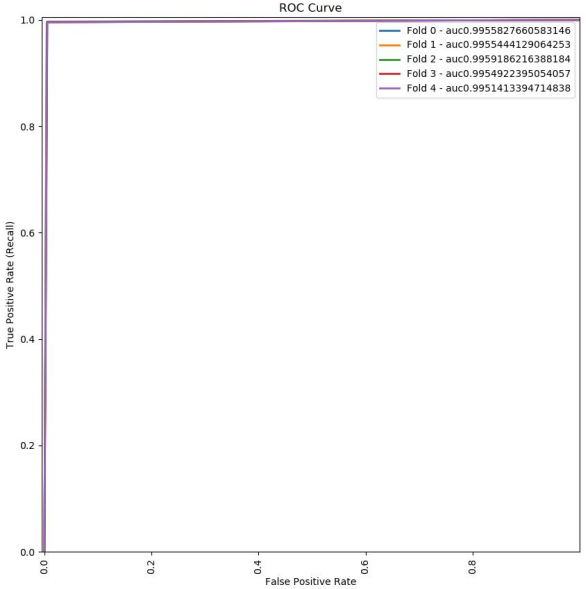
```
Tempo de treinamento para o KNeighbors-Fold0: 151.9493248462677
Auc Fold 0: 0.9955827660583146
Matriz Confusao - KNeighbors - Fold 0
       0      1
0  39519    176
1    176  39818

Tempo de treinamento para o KNeighbors-Fold1: 254.06695127487183
Auc Fold 1: 0.9955444129064253
Matriz Confusao - KNeighbors - Fold 1
       0      1
0  39510    185
1    170  39824

Tempo de treinamento para o KNeighbors-Fold2: 172.16840171813965
Auc Fold 2: 0.9959186216388184
Matriz Confusao - KNeighbors - Fold 2
       0      1
0  39501    194
1    131  39863

Tempo de treinamento para o KNeighbors-Fold3: 153.83028984069824
Auc Fold 3: 0.9954922395054057
Matriz Confusao - KNeighbors - Fold 3
       0      1
0  39487    208
1    151  39843

Tempo de treinamento para o KNeighbors-Fold4: 167.00128865242004
Auc Fold 4: 0.9951413394714838
Matriz Confusao - KNeighbors - Fold 4
       0      1
0  39478    217
1    170  39824
```

- Estatísticas do Treinamento 50/50

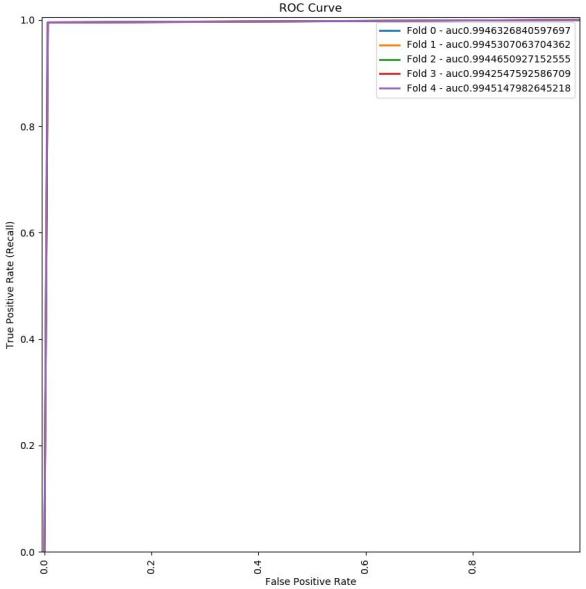
```
Tempo de treinamento para o KNeighbors-Fold0: 73.50124716758728
Auc Fold 0: 0.9946326840597697
Matriz Confusao - KNeighbors - Fold 0
       0      1
0  98666    570
1    499  99486

Tempo de treinamento para o KNeighbors-Fold1: 75.43371486663818
Auc Fold 1: 0.9945307063704362
Matriz Confusao - KNeighbors - Fold 1
       0      1
0  98614    622
1    467  99518

Tempo de treinamento para o KNeighbors-Fold2: 72.4316897392273
Auc Fold 2: 0.9944650927152555
Matriz Confusao - KNeighbors - Fold 2
       0      1
0  98598    638
1    464  99521

Tempo de treinamento para o KNeighbors-Fold3: 72.45569157600403
Auc Fold 3: 0.9942547592586709
Matriz Confusao - KNeighbors - Fold 3
       0      1
0  98590    646
1    498  99487

Tempo de treinamento para o KNeighbors-Fold4: 72.56091094017029
Auc Fold 4: 0.9945147982645218
Matriz Confusao - KNeighbors - Fold 4
       0      1
0  98590    646
1    446  99539
```

KNN- Resultados

- Curva ROC - Treinamento 80/20

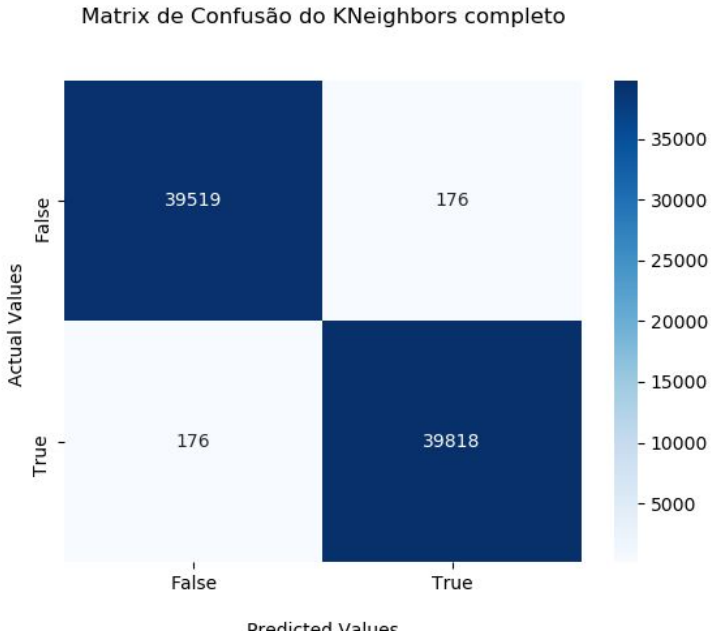


- Curva ROC - Treinamento 50/50

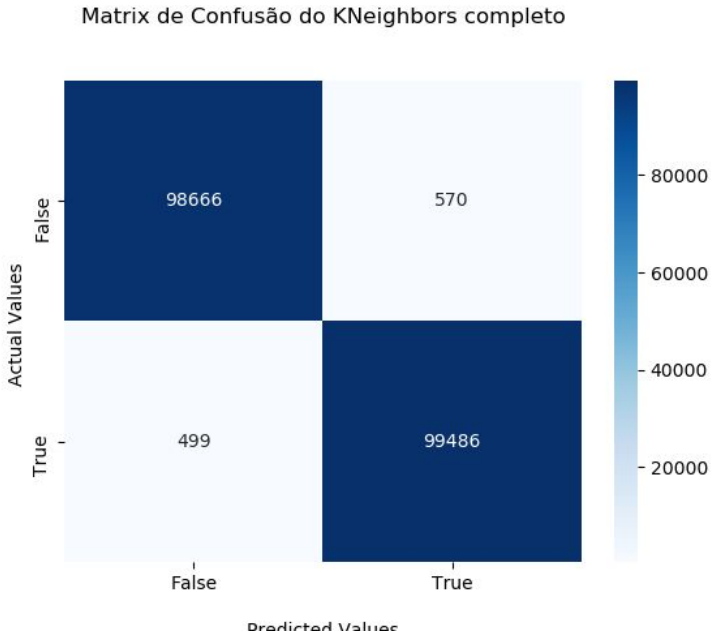


KNN-Resultados

- Matriz de Confusão 80/20



- Matriz de Confusão 50/50



KNN – Comparação

- Autores
  - KNN com K = 6
  - Precisão de 99.95%
  - Misclassification = 2295 de 4222227

- Replicação
  - KNN com K = 1
  - Precisão de 99.12% e 98.93%, 80/20 e 50/50 respectivamente

RF - Treinamento

- Definição dos parâmetros para o GridSearch
  - 50 Árvores
  - 100 Árvores
- Estatísticas do Treinamento 80/20

```
Melhores parametros do RandomForest para o precision_score
{'n_estimators': 100}
Melhor score do RandomForest para o precision_score
0.9988193395235149
```

- Estatísticas Treinamento 50/50

```
Melhores parametros do RandomForest para o precision_score
{'n_estimators': 50}
Melhor score do RandomForest para o precision_score
0.998449013059988
```

KNN - Treinamento

- Estatísticas do Treinamento 80/20

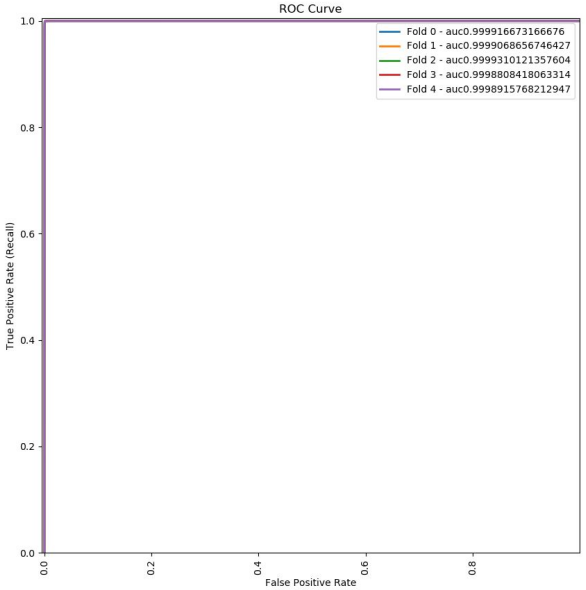
```
Tempo de treinamento para o RandomForest-Fold0: 208.64189100265503
Auc Fold 0: 0.9998923040370655
Matriz Confusao - RandomForest - Fold 0
       0      1
0  39654     41
1     34  39960

Tempo de treinamento para o RandomForest-Fold1: 227.01938128471375
Auc Fold 1: 0.9999321157148254
Matriz Confusao - RandomForest - Fold 1
       0      1
0  39658     37
1     18  39976

Tempo de treinamento para o RandomForest-Fold2: 209.87416791915894
Auc Fold 2: 0.9998685925196376
Matriz Confusao - RandomForest - Fold 2
       0      1
0  39661     34
1     26  39968

Tempo de treinamento para o RandomForest-Fold3: 208.07083320617676
Auc Fold 3: 0.9998932410084462
Matriz Confusao - RandomForest - Fold 3
       0      1
0  39656     39
1     23  39971

Tempo de treinamento para o RandomForest-Fold4: 222.34196162223816
Auc Fold 4: 0.9999033483313213
Matriz Confusao - RandomForest - Fold 4
       0      1
0  39662     33
1     28  39966
```

- Estatísticas do Treinamento 50/50

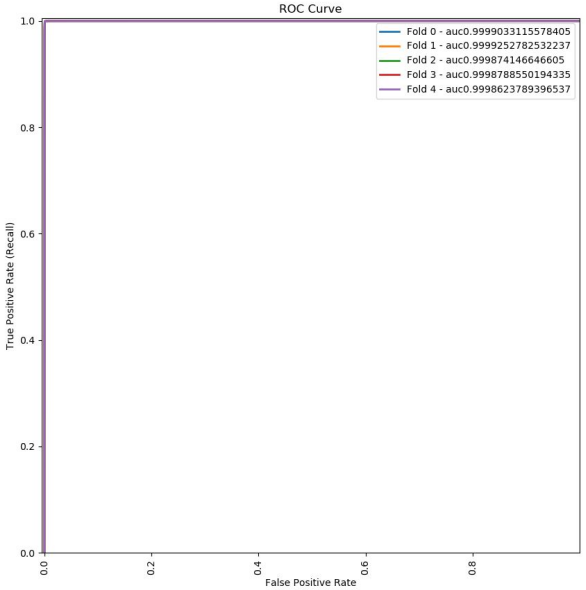
```
Tempo de treinamento para o RandomForest-Fold0: 67.08633255958557
Auc Fold 0: 0.9998997645809554
Matriz Confusao - RandomForest - Fold 0
       0      1
0  99099    137
1     60  99925

Tempo de treinamento para o RandomForest-Fold1: 71.5131003856659
Auc Fold 1: 0.9999062178948754
Matriz Confusao - RandomForest - Fold 1
       0      1
0  99110    126
1     50  99935

Tempo de treinamento para o RandomForest-Fold2: 74.37931418418884
Auc Fold 2: 0.999904480512659
Matriz Confusao - RandomForest - Fold 2
       0      1
0  99115    121
1     56  99929

Tempo de treinamento para o RandomForest-Fold3: 73.65085124969482
Auc Fold 3: 0.9998984963025199
Matriz Confusao - RandomForest - Fold 3
       0      1
0  99115    121
1     67  99918

Tempo de treinamento para o RandomForest-Fold4: 66.19434905052185
Auc Fold 4: 0.9998486774205216
Matriz Confusao - RandomForest - Fold 4
       0      1
0  99112    124
1     72  99913
```

RF- Resultados

- Curva ROC - Treinamento 80/20

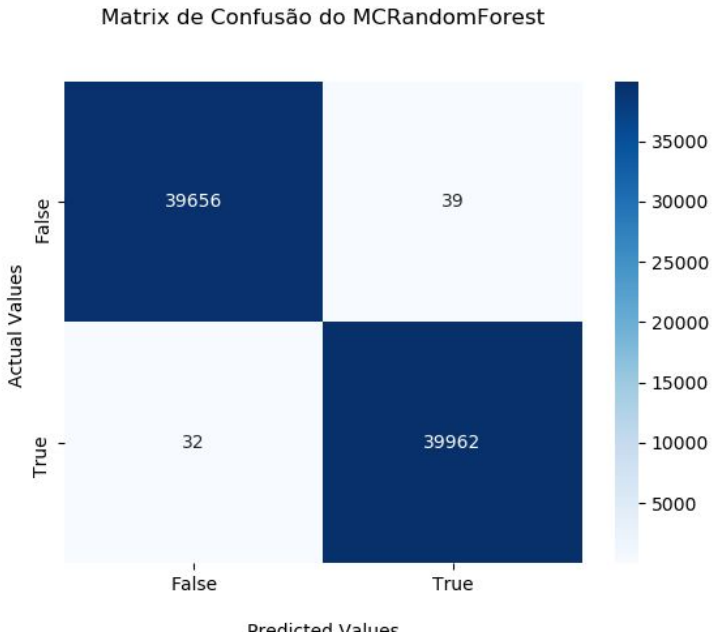


- Curva ROC - Treinamento 50/50

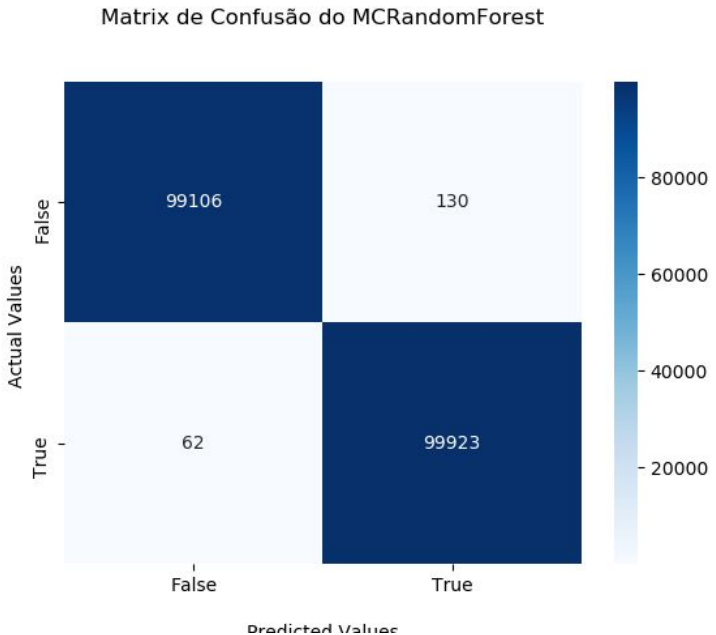


RF-Resultados

- Matriz de Confusão 80/20



- Matriz de Confusão 50/50



RF - Comparação

- Autores
  - RF com n_estimators = 200
  - RF com max_depth = 5
  - Precisão de 99.95%
  - Misclassification = 2315 de 4222227

- Replicação
  - RF com n_estimators = 100
  - RF com max_depth = None
  - Precisão de 99.82% para ambos treinamentos

MLP - Treinamento

- Definição dos parâmetros para o GridSearch
  - 1000 Épocas, erro 0.01
  - 5000 Épocas, erro 0.01
- Estatísticas do Treinamento 80/20

```
Melhores parametros do MLP para o precision_score
{'hidden_layer_sizes': 7, 'max_iter': 1000, 'tol': 0.01}
Melhor score do MLP para o precision_score
0.703568677745978
```

- Estatísticas Treinamento 50/50

```
Melhores parametros do MLP para o precision_score
{'hidden_layer_sizes': 7, 'max_iter': 1000, 'tol': 0.01}
Melhor score do MLP para o precision_score
0.7045403517447719
```

MLP - Treinamento

- Estatísticas do Treinamento 80/20

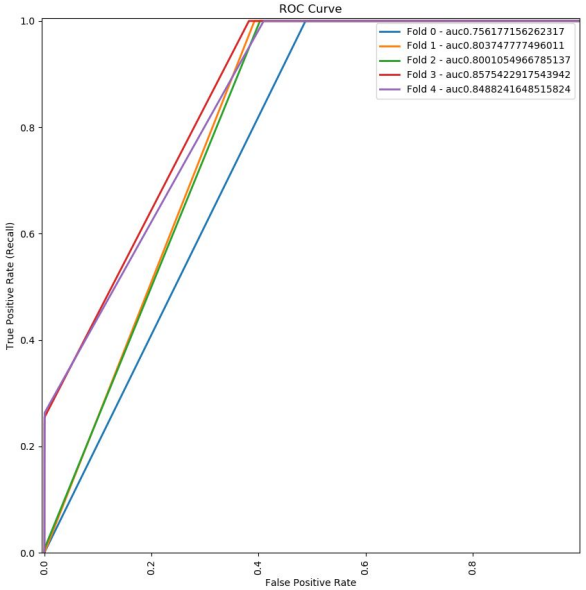
```
Tempo de treinamento para o MLP-Fold0: 127.84013605117798
Auc Fold 0: 0.756177156262317
Matriz Confusao - MLP - Fold 0
       0      1
0  20283  19412
1      0  39994

Tempo de treinamento para o MLP-Fold1: 107.79522180557251
Auc Fold 1: 0.803747777496011
Matriz Confusao - MLP - Fold 1
       0      1
0  24103  15592
1      9  39985

Tempo de treinamento para o MLP-Fold2: 125.30146956443787
Auc Fold 2: 0.8001054966785137
Matriz Confusao - MLP - Fold 2
       0      1
0  23699  15996
1      7  39987

Tempo de treinamento para o MLP-Fold3: 93.51204419136047
Auc Fold 3: 0.8575422917543942
Matriz Confusao - MLP - Fold 3
       0      1
0  24533  15162
1     10  39984

Tempo de treinamento para o MLP-Fold4: 122.27955436706543
Auc Fold 4: 0.8488241648515824
Matriz Confusao - MLP - Fold 4
       0      1
0  23420  16275
1      3  39991
```

- Estatísticas do Treinamento 50/50

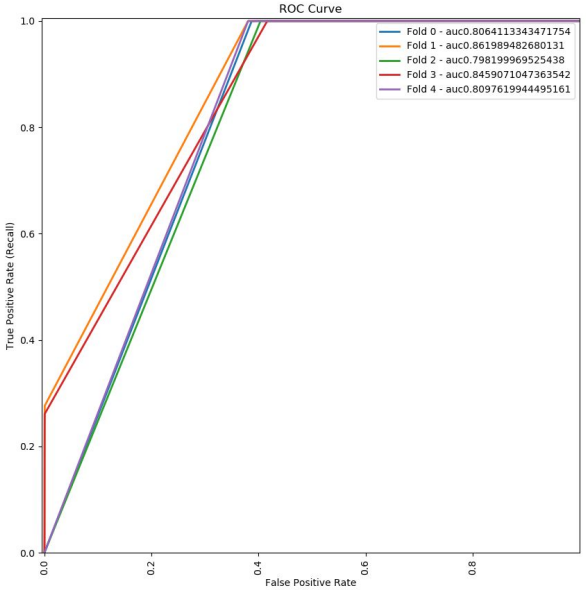
```
Tempo de treinamento para o MLP-Fold0: 88.75690603256226
Auc Fold 0: 0.8064113343471754
Matriz Confusao - MLP - Fold 0
       0      1
0  60704  38532
1     51  99934

Tempo de treinamento para o MLP-Fold1: 91.98226761817932
Auc Fold 1: 0.861989482680131
Matriz Confusao - MLP - Fold 1
       0      1
0  61470  37766
1      9  99976

Tempo de treinamento para o MLP-Fold2: 75.76618003845215
Auc Fold 2: 0.798199969525438
Matriz Confusao - MLP - Fold 2
       0      1
0  59129  40107
1      9  99976

Tempo de treinamento para o MLP-Fold3: 96.3425784111023
Auc Fold 3: 0.8459071047363542
Matriz Confusao - MLP - Fold 3
       0      1
0  57846  41390
1      8  99977

Tempo de treinamento para o MLP-Fold4: 86.46627712249756
Auc Fold 4: 0.8097619944495161
Matriz Confusao - MLP - Fold 4
       0      1
0  61401  37835
1     16  99969
```

MLP- Resultados

- Curva ROC - Treinamento 80/20

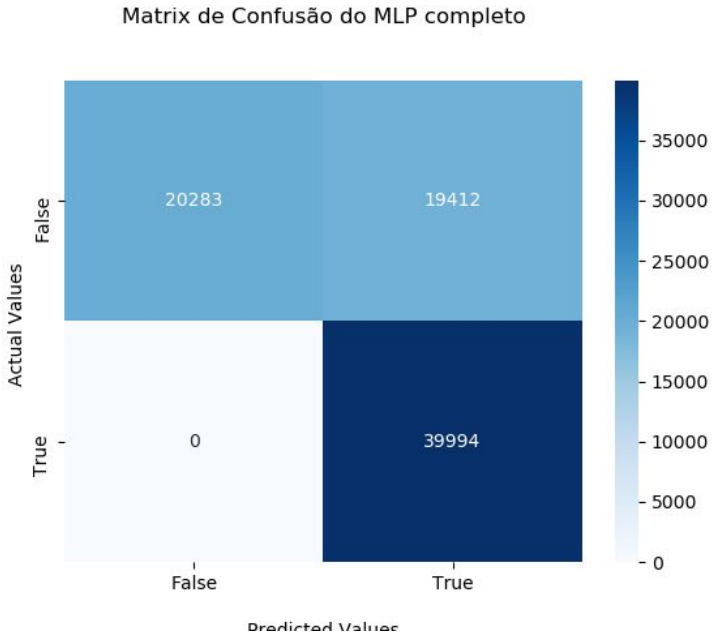


- Curva ROC - Treinamento 50/50

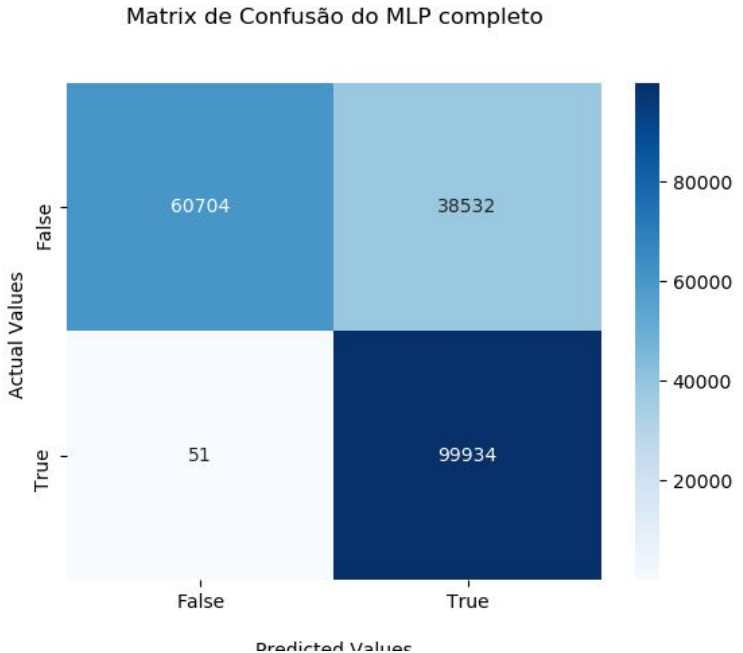


MLP-Resultados

- Matriz de Confusão 80/20



- Matriz de Confusão 50/50



Experimentação dos Modelos

- Experimentação e obtenção de resultados para os 6 seguintes modelos:
  - KNN com 80/20
  - KNN com 50/50
  - RF com 80/20
  - RF com 50/50
  - MLP com 80/20
  - MLP com 50/50
- Será passado os 20% restante do dataset original + arquivo pcap de DDOS

Conclusão

- MLP ficou muito específico em acertar o DDoS, mas resulta em muitos Falso Positivos
- Bons resultados obtidos para o dataset em questão
- Dificuldade de colocar em prática
- KNN e RF não atingem resultados muito expressivos quando testados com outros datasets

Obrigado!