Machine Learning

Sistema de Detecção de Instrusão baseado em anomalias

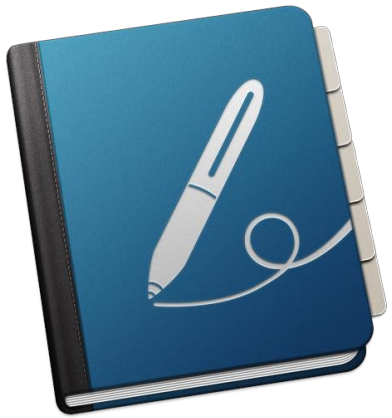
Gileno Dias dos Santos – SGP/ME
Kelson Carvalho Santos – IFPI
Warley Duarte Reis – SERPRO

Enap

coding BOOTCAMP

Machine Learning

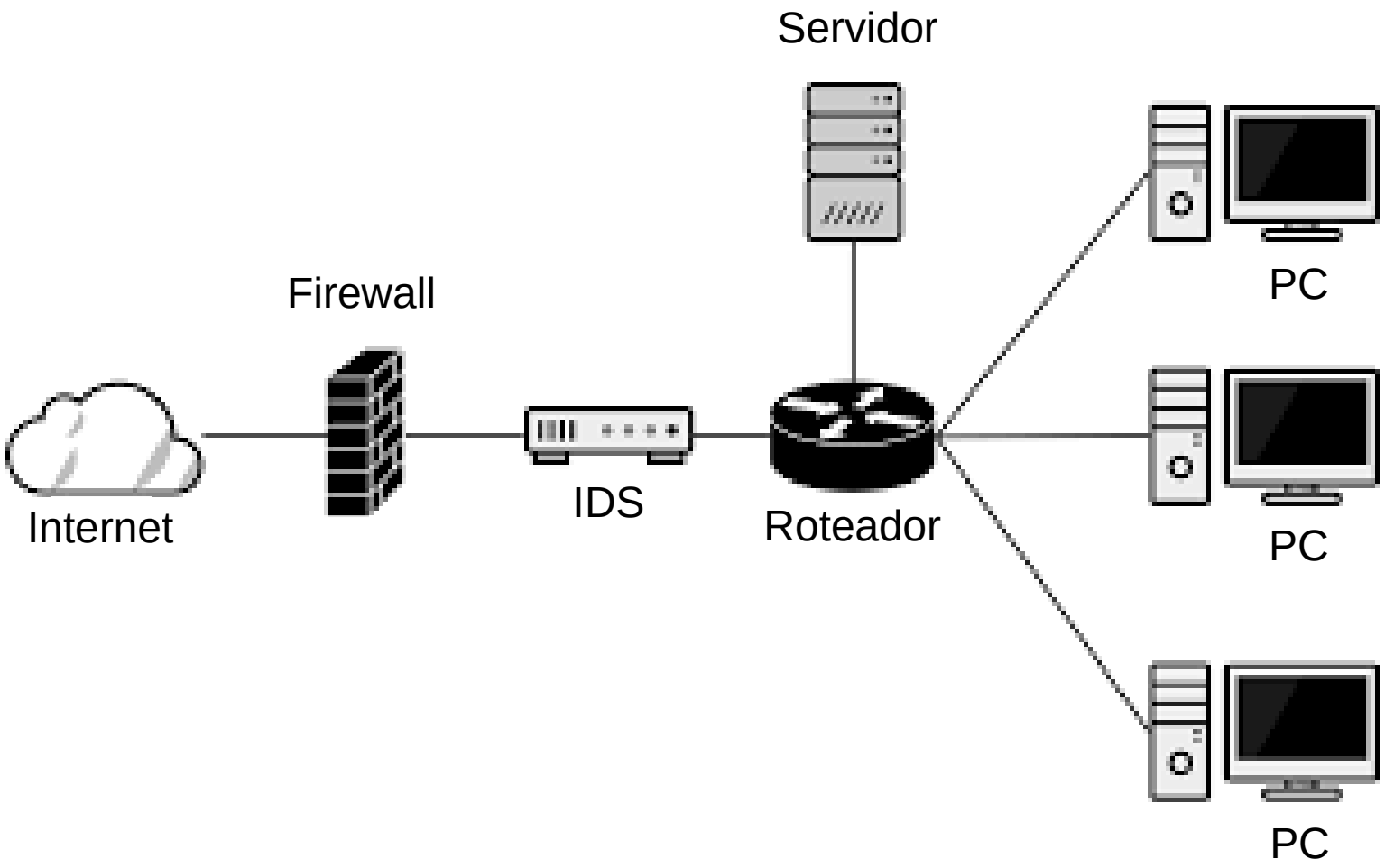
Roteiro



1. Problema
2. Materiais e Métodos
3. Resultados
4. Trabalhos Futuros

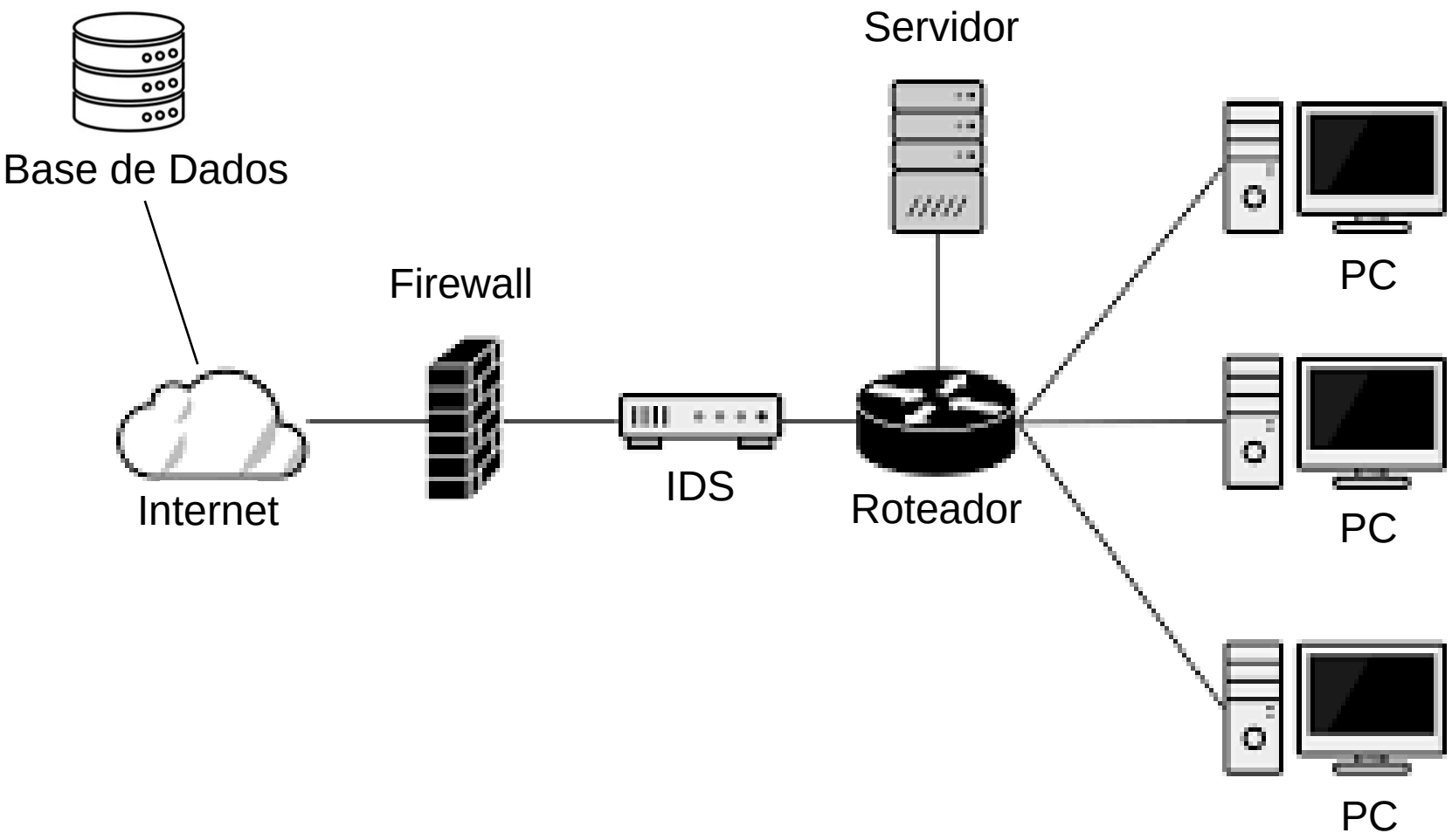
1. Problema

1 Problema



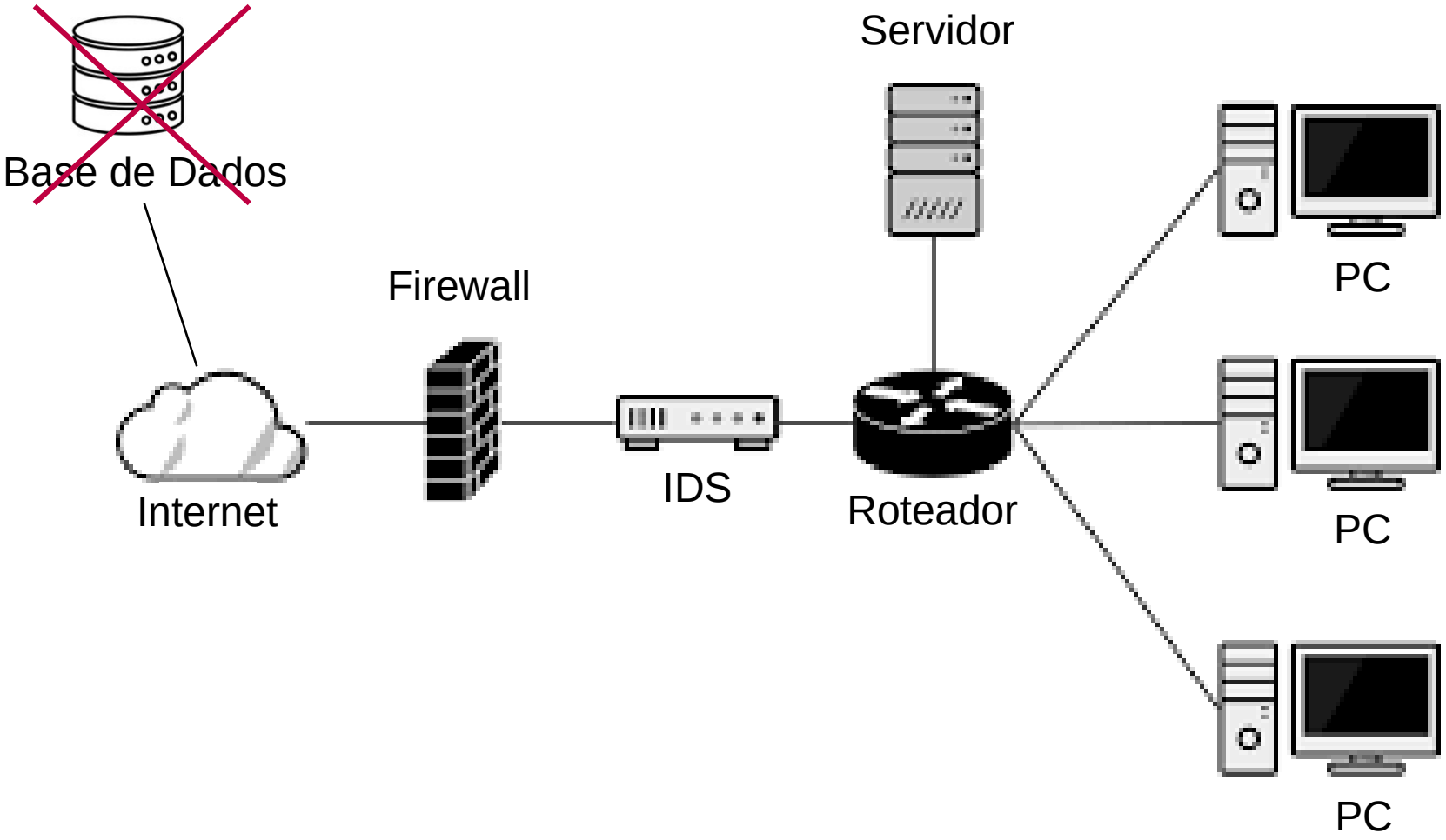
Fonte: https://bdm.unb.br/bitstream/10483/13197/1/2016_JoaoFranciscoGoncalvesSobrinhoJunior.pdf

1 Problema



Fonte: https://bdm.unb.br/bitstream/10483/13197/1/2016_JoaoFranciscoGoncalvesSobrinhoJunior.pdf

1 Problema



Fonte: https://bdm.unb.br/bitstream/10483/13197/1/2016_JoaoFranciscoGoncalvesSobrinhoJunior.pdf

2. Materiais e Métodos

BASE DE DADOS

Conjunto de dados abertos para avaliação de detecção de intrusão (CIC-IDS2017).

Canadian Institute for Cybersecurity (CIC)

University New Brunswick (UNB).

Fonte: https://www.unb.ca/cic/datasets/ids-2017.html

Dataset

2.830.743 amostras de fluxos

78 features
1 target

DIVISÃO DO DATASET

Dataset

2.830.743 amostras de fluxos

78 features
1 target

NaN e Infinity (2.867)

Dataset

2.827.876 amostras de fluxos

78 features
1 target

DIVISÃO DO DATASET

Dataset

2.827.876 amostras
de fluxos

78 features
1 target

Extração

Dataset
Treino / Validação

100.000 amostras

60.000 Normal
40.000 Malicioso

Restante

Dataset
Teste

2.727.876 amostras

2.271.320 Normal
556.556 Malicioso

CRIAÇÃO E ANÁLISE DOS MODELOS

Dataset
Treino (70) / Validação (30)

100.000
amostras

60.000 Normal
40.000 Malicioso

Execução

Logistic Regression

KNN

Decision Tree

Random Forest

Execução

Dataset
Teste

2.727.876
amostras

2.271.320 Normal
556.556 Malicioso

OTIMIZAÇÃO DOS MODELOS

78 features
1 target

Execução

Hiperparâmetros
Normalização (Scaler)
Codificação
Feature Importance

OTIMIZAÇÃO DOS MODELOS

78 features
1 target

Execução

Hiperparâmetros
Normalização (Scaler)
Codificação
Feature Importance

OTIMIZAÇÃO DOS MODELOS

78 features
1 target

Execução

Hiperparâmetros
Normalização (Scaler)
Codificação
Feature Importance

```
80       33852
53       25296
443      13329
21         724
22         716
         ...
2342         1
39220        1
63800        1
51518        1
34319        1
Name: Destination_Port, Length: 10303, dtype: int64
```

Codificação

```
0    66148
1    33852
Name: Destination_Port_80, dtype: int64
0    74704
1    25296
Name: Destination_Port_53, dtype: int64
0    86671
1    13329
Name: Destination_Port_443, dtype: int64
0    99276
1      724
Name: Destination_Port_21, dtype: int64
0    99284
1      716
Name: Destination_Port_22, dtype: int64
0    99364
1      636
Name: Destination_Port_123, dtype: int64
0    97219
1     2781
Name: Destination_Port_Demais_Portas_Reservadas, dtype: int64
0    77334
1    22666
Name: Destination_Port_Portas_Nao_Reservadas, dtype: int64
```

OTIMIZAÇÃO DOS MODELOS

78 features
1 target

Execução →

Hiperparâmetros
Normalização (Scaler)
Codificação
Feature Importance

	features	valor
52	Average_Packet_Size	0.383439
13	Bwd_Packet_Length_Std	0.372066
35	Bwd_Header_Length	0.143762
0	Destination_Port	0.047371
39	Max_Packet_Length	0.042413
69	min_seg_size_forward	0.005935
71	Active_Std	0.004734
65	Subflow_Bwd_Bytes	0.000141
67	Init_Win_bytes_backward	0.000070
..	...	...
32	Fwd_URG_Flags	0.000000
25	Bwd_IAT_Total	0.000000
31	Bwd_PSH_Flags	0.000000
30	Fwd_PSH_Flags	0.000000
29	Bwd_IAT_Min	0.000000
28	Bwd_IAT_Max	0.000000
27	Bwd_IAT_Std	0.000000
26	Bwd_IAT_Mean	0.000000
77	Idle_Min	0.000000

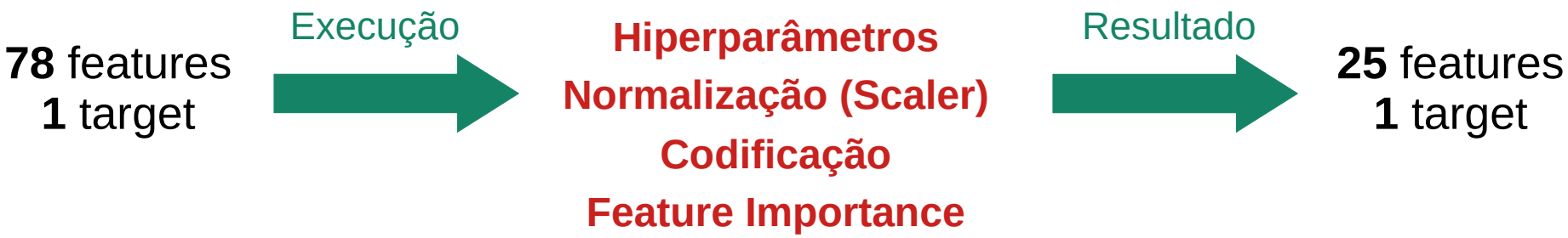
[78 rows x 2 columns]

Feature Importance →

```
---  ------
 0   Destination_Port
 1   Total_Length_of_Fwd_Packets
 2   Flow_Bytes_s
 3   Init_Win_bytes_backward
 4   Fwd_Header_Length
 5   Fwd_Packet_Length_Max
 6   Packet_Length_Variance
 7   Subflow_Bwd_Packets
 8   Bwd_Packet_Length_Std
 9   PSH_Flag_Count
 10  Fwd_IAT_Min
 11  Total_Length_of_Bwd_Packets
 12  Fwd_Packet_Length_Std
 13  Bwd_IAT_Min
 14  Bwd_Packets_s
 15  Avg_Fwd_Segment_Size
 16  Flow_IAT_Min
 17  Target
 18  Destination_Port_80
 19  Destination_Port_53
 20  Destination_Port_443
 21  Destination_Port_21
 22  Destination_Port_22
 23  Destination_Port_123
 24  Destination_Port_Demais_Portas_Reservadas
 25  Destination_Port_Portas_Nao_Reservadas
dtypes: float64(14), int64(11), object(1)
memory usage: 20.6+ MB
```

Materiais e Métodos

OTIMIZAÇÃO DOS MODELOS



2 Materiais e Métodos

OTIMIZAÇÃO DOS MODELOS

78 features
1 target

Execução

Hiperparâmetros
Normalização (Scaler)
Codificação
Feature Importance

Resultado

25 features
1 target

Dataset
Treino (70) / Validação (30)

100.000
amostras

60.000 Normal
40.000 Malicioso

Execução

Logistic Regression

KNN

Decision Tree

Random Forest

Execução

Dataset
Teste

2.727.876
amostras

2.271.320 Normal
556.556 Malicioso

3. Resultados

TABELA RESUMO DOS RESULTADOS

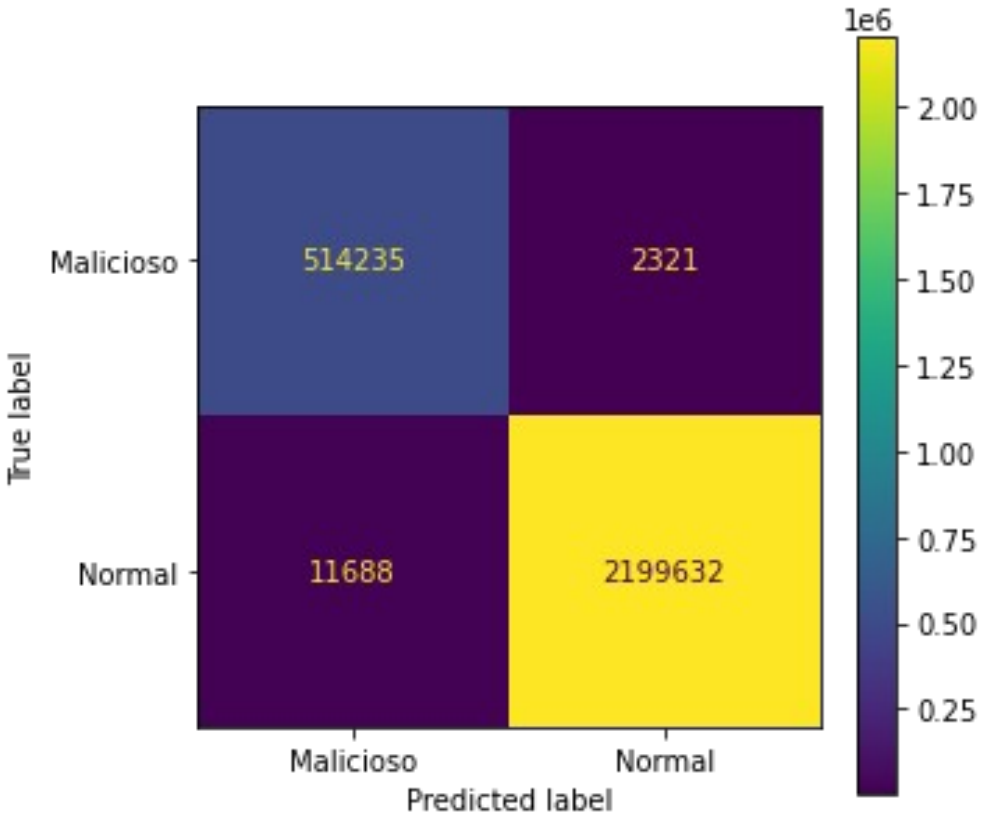
Classificadores	Aplicação Modelo	
	Treino/Validação	Teste
Logistic Regression	0.93943	0.94253
KNN	0.98070	0.97920
Decision Tree	0.98300	0.98384
Random Forest	0.98843	0.98989

TABELA RESUMO DOS RESULTADOS

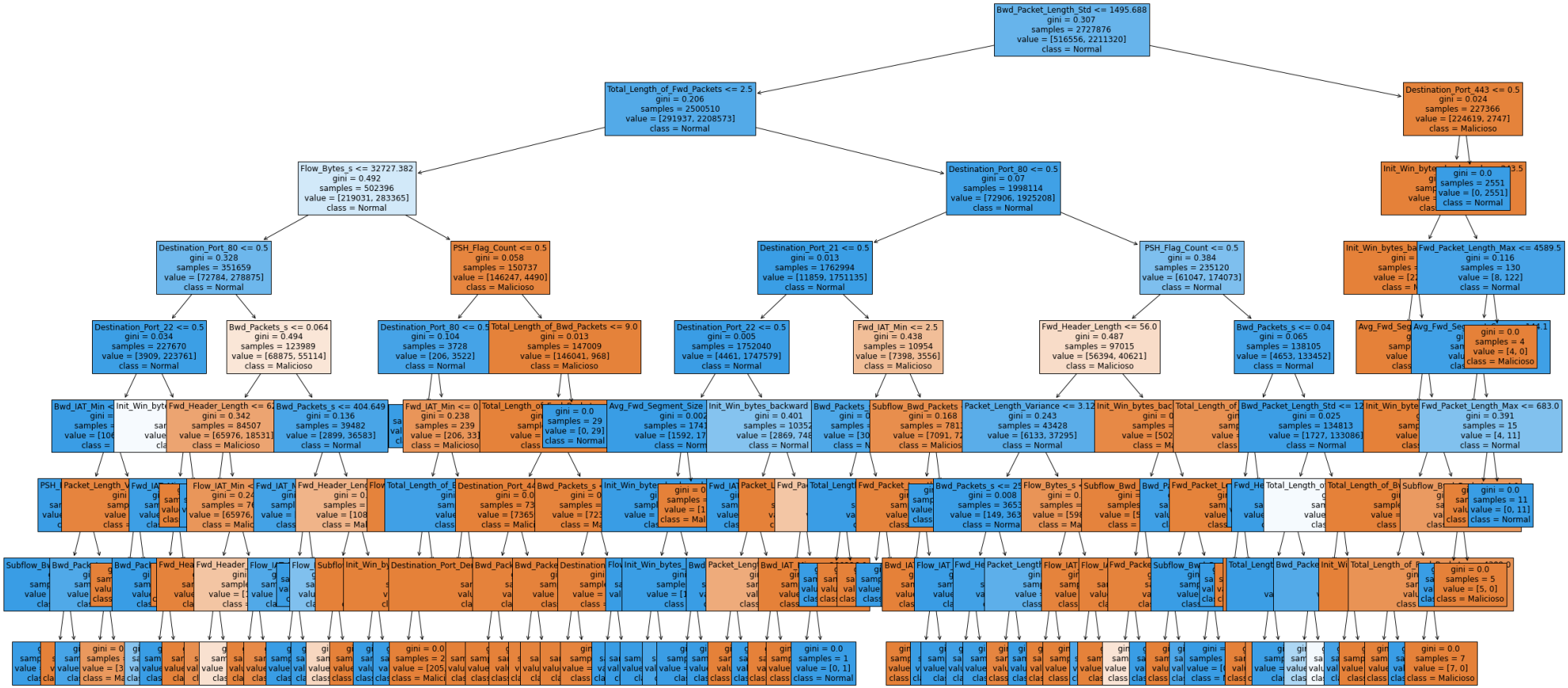
Classificadores	Aplicação Modelo		Aplicação Modelo Otimizado	
	Treino/Validação	Teste	Treino/Validação	Teste
Logistic Regression	0.93943	0.94253	0.94303	0.94452
KNN	0.98070	0.97920	0.98383	0.98829
Decision Tree	0.98300	0.98384	0.99593	0.99579
Random Forest	0.98843	0.98989	0.99486	0.99479

3 Resultados

TABELA VERDADE



ÁRVORE DE DECISÃO



4. Trabalhos Futuros

4 Trabalho Futuros

Criação de modelos para subclassificação do fluxo malicioso.

Testes com novos modelos e hiperparâmentros.

Geração de um paper para publicação e tornar os dados públicos e acessíveis.

DESAFIO

Criação de um modelo para execução em ambiente real.

Obrigado!

Machine Learning

Gileno Dias dos Santos – SGP/ME
gileno.santos@economia.gov.br

Kelson Carvalho Santos – IFPI
kelson@ifpi.edu.br

Warley Duarte Reis – SERPRO
warley.reis@serpro.gov.br