UNIVERSIDADE FEDERAL DO PARANÁ

DANIEL PIMENTA FURTADO

PROJETO FINAL
PHISHING WEBSITES

CURITIBA PR

2021

1 INTRODUÇÃO

O projeto foi desenvolvido utilizando o dataset Phishing Websites Data Set [1]. O dataset contém atributos para identificação de phishing websites que são sites que tentam enganar usuários para obter informações confidenciais. O banco de dados possui 30 características para cada site:

1. having_IP_Address
2. URL_Length
3. Shortining_Service
4. having_At_Symbol
5. double_slash_redirecting
6. Prefix_Suffix
7. having_Sub_Domain
8. SSLfinal_State
9. Domain_registeration_length
10. Favicon
11. port
12. HTTPS_token
13. Request_URL
14. URL_of_Anchor
15. Links_in_tags
16. SFH
17. Submitting_to_email
18. Abnormal_URL
19. Redirect
20. on_mouseover
21. RightClick
22. popUpWidnow
23. Iframe

[1] https://archive.ics.uci.edu/ml/datasets/phishing+websites#

24. age_of_domain
25. DNSRecord
26. web_traffic
27. Page_Rank
28. Google_Index
29. Links_pointing_to_page
30. Statistical_report

A última coluna (Result) define a classificação do site como Phishing (-1) ou Legítimo (1). No dataset cada características se comporta como um classificador baseado em heurísticas. Cada característica classifica o site entre Phishing, Legítimo e em alguns casos como Suspeito (0). As características foram separadas em 4 grupos pelos autores do dataset: Características baseadas barra de endereço (1.1), Características baseadas em anormalidades (1.2), Características baseadas em HTML e JavaScript (1.3) e Características baseadas no domínio (1.4).

1.1 CARACTERÍSTICAS BASEADAS BARRA DE ENDEREÇO

1.1.1 having_IP_Address

Avalia se um endereço de IP é utilizado como alternativa ao nome de domínio na URL, como “http://125.98.3.123/fake.html”. Em alguns casos, o endereço é transformado em hexadecimal (http://0x58.0xCC.0xCA.0x62/2/paypal.ca/index.html). Regra:

- Se a parte de domínio possui endereço IP (Phishing)
- Caso contrário (Legítimo)

1.1.2 URL_Length

Avalia o tamanho da URL. Phishing websites podem utilizar URL longas para esconder partes duvidosas. Regra:

- Tamanho da URL < 54 caracteres (Legítimo)
- Tamanho da URL ≥ 54 e ≤ 75 caracteres (Suspeito)
- Caso contrário (Phishing)

1.1.3 Shortining_Service

Avalia se a URL é proveniente de serviçõe de encurtamento de URL, essas URL utilizando ”HTTP Redirect” para direcionar o usuário para site de destino. Por exemplo a URL "http://portal.hud.ac.uk/"pode ser encurtada para "bit.ly/19DXSk4". Regra:

- TinyURL (Phishing)
- Caso contrário (Legítimo)

1.1.4 having_At_Symbol

Avalia a utilização do símbolo "@", porque quando utilizado o navegador ignora tudo que precede o símbolo. Regra:

- Possui o símbolo "@"(Phishing)
- Caso contrário (Legítimo)

1.1.5 double_slash_redirecting

Avalia a utilização do símbolo "//", porque sua função é de redirecionar o usuário para outro site. Em URL que começam com HTTP e HTTPS o símbolo ”//” deve aparecer na posição 6 e 7, respectivamente. Regra:

- Posição do último ”//” > 7 (Phishing)
- Caso contrário (Legítimo)

1.1.6 Prefix_Suffix

Avalia a utilização do símbolo ”-”, esse símbolo é raramente utilizado em sites legítimos e pode ser usado para mascarar a URL, colocando parte de um domínio legítimo. Por exemplo: "http://www.Confirme-paypal.com/". Regra:

- URL possui o símbolo ”-” (Phishing)
- Caso contrário (Legítimo)

1.1.7 having_Sub_Domain

Avalia o número de subdomínios além do country-code top-level domain (ccTLD). Regra:

- Número de pontos na parte do domínio = 1 (Legítimo)
- Número de pontos na parte do domínio = 2 (Suspeito)
- Caso contrário (Phishing)

1.1.8 SSLfinal_State

Avalia a idade do certificado de HTTPS do site e seu emissor. Regra:

- Utiliza HTTPS e o emissor é confiável e a idade do certificado ≥ 1 ano (Legítimo)
- Utiliza HTTPS e o emissor não é confiável (Suspeito)
- Caso contrário (Phishing)

1.1.9 Domain_registeration_length

Avalia quando a o registro do domínio se expira, pois sites maliciosos duram por um curto período de tempo. Regra:

- Registro do domínio expira em $\leq$ 1 ano (Phishing)
- Caso contrário (Legítimo)

1.1.10 Favicon

Avalia se o site utiliza um favicon que não é carregado a partir do mesmo domínio. Regra:

- Favicon carregado a partir de um domínio externo (Phishing)
- Caso contrário (Legítimo)

1.1.11 port

Avalia se algum serviço possui sua porta fora do número padrão. Regra:

- Porta fora do número padrão (Phishing)
- Caso contrário (Legítimo)

1.1.12 HTTPS_token

Avalia a presença de um token HTTPS para enganar o usuário. Por exemplo: "http://https-www-paypal-it-webapps-mpp-home.soft-hair.com/". Regra:

- Utiliza token HTTPS (Phishing)
- Caso contrário (Legítimo)

1.2 CARACTERÍSTICAS BASEADAS EM ANORMALIDADES

1.2.1 Request_URL

Avalia se os objetos externos contidos no interior do site como imagens, vídeos e sons são carregados partir de domínios externos (Request URL). Regra:

- % de Request URL $<$ 22% (Legítimo)
- % de Request URL $\geq$ 22% e $\leq$ 61% (Suspeito)
- Caso contrário (Phishing)

1.2.2 URL_of_Anchor

Avalia se as tags <a> no site (Anchor) possuem links para domínios externos ou não possuem link (<a href="#">). Regra:

- % do URL of Anchor < 31% (Legítimo)
- % do URL of Anchor ≥ 31% e ≤ 67% (Suspeito)
- Caso contrário (Phishing)

1.2.3 Links_in_tags

Avalia a presença de links nas tags <meta>, <script> e <link> (Tags). Regra:

- % de links nas Tags < 17% (Legítimo)
- % de links nas Tags ≥ 17% e ≤ 81% (Suspeito)
- Caso contrário (Phishing)

1.2.4 SFH

Avalia se o Server Form Handler (SFH) possui empty string ou "about:blank"ou é referente a outro domínio da URL original. Regra:

- SFH possui empty string ou "about:blank"(Phishing)
- SFH é referente a outro domínio (Suspeito)
- Caso Contrário (Legítimo)

1.2.5 Submitting_to_email

Avalia a presença das funções que redireciona informações para algum e-mail (mail(), mailto:). Regra:

- Usando as funções "mail()"ou "mailto:"para submeter informações do usuário (Phishing)
- Caso Contrário (Legítimo)

1.2.6 Abnormal_URL

Avalia se o host name é parte da URL. Regra:

- O host name não é incluído na URL (Phishing)
- Caso contrário (Legítimo)

1.3 CARACTERÍSTICAS BASEADAS EM HTML E JAVASCRIPT

1.3.1 Redirect

Avalia o número de redirecionamento de um site. Regra:

- Número de redirecionamento ≤ 1 (Legítimo)
- Número de redirecionamento ≥ 2 e < 4 (Suspeito)
- Caso contrário (Phishing)

1.3.2 on_mouseover

Avalia a utilização do evento "onMouseOver"para esconder uma falsa URL. Regra:

- onMouseOver muda a barra de status (Phishing)
- onMouseOver não muda a barra de status (Legítimo)

1.3.3 RightClick

Avalia se o site utiliza JavaScript para desabilitar o RightClick. Regra:

- RightClick desabilitado (Phishing)
- Caso Contrário (Legítimo)

1.3.4 popUpWidnow

Avalia se janelas pop-up possuem campos de textos. Regra:

- Janela Pop-up contém campos de texto (Phishing)
- Caso Contrário (Legítimo)

1.3.5 Iframe

Avalia se o site utiliza a tag HTML ”iframe” para esconder sites adicionais. Regra:

- Usando iframe (Phishing)
- Caso Contrário (Legítimo)

1.4 CARACTERÍSTICAS BASEADAS NO DOMÍNIO

1.4.1 age_of_domain

Avalia a idade do domínio. Regra:

- Idade do domínio ≥ 6 meses (Legítimo)
- Caso Contrário (Phishing)

1.4.2 DNSRecord

Avalia se o DNS record é vazio ou não é encontrado. Regra:

- Nenhum DNS record para o domínio (Phishing)
- Caso Contrário (Legítimo)

1.4.3 web_traffic

Avalia o Website Rank (Alexa database) do site. Regra:

- Website Rank < 100000 (Legítimo)
- Website Rank > 100000 (Suspeito)
- Não possui Website Rank (Phishing)

1.4.4 Page_Rank

Avalia o PageRank do site que varia no intervalo de 0 a 1. Regra:

- PageRank < 0.2 (Phishing)
- Caso Contrário (Legítimo)

1.4.5 Google_Index

Avalia se a página possui Google Index. Regra:

- Página Indexada pelo Google (Legítimo)
- Caso Contrário (Phishing)

1.4.6 Links_pointing_to_page

Avalia o número de links apontandos para o site. Regra:

- Link apontando para o site = 0 (Phishing)
- Link apontando para o site > 0 e $\leq$ 2 (Suspeito)
- Caso Contrário (Legítimo)

1.4.7 Statistical_report

Avalia se o host e o domínio encontram-se entre os sites com piores reputações de acordo com os reports PhishTank e StopBadware. Regra:

- Possui reputação ruim no PhishTank ou StopBadware (Phishing)
- Caso Contrário (Legítimo)

1.5 MÉTRICAS

Durante a realização do trabalho utilizou-se as métricas: acurácia, erro absoluto, precisão, recall, F1-score (F1) e o coeficiente de matthews (MCC). O F1-Score é definido como:

$$F1 = 2\frac{PR}{P+R} \tag{1.1}$$

Onde, P é a precisão e R é o recall. O coeficiente de matthews (MCC) é definido como:

$$MCC = \frac{TPxTN - FPxFN}{\sqrt{(TP+FP)(TP+FN)(TN+FP)(TN+FN)}} \tag{1.2}$$

Onde TP, FP, FN e TN são os *True Positive*, *False Positive*, *False Negative* e *True Negative*, respectivamente. O Coeficiente de Matthews leva em conta toda a matriz de confusão e seus valores variam entre -1 a 1. $MCC = 1$ representa predição perfeita, $MCC = 0$ predição aleatória e $MCC = -1$ predição inversa.

2 EXPLORAÇÃO DE DADOS

O dataset possui 11055 sites que estão distribuídos da seguinte forma: 6157 (legítimos), 4898 (phishing), figura 2.1

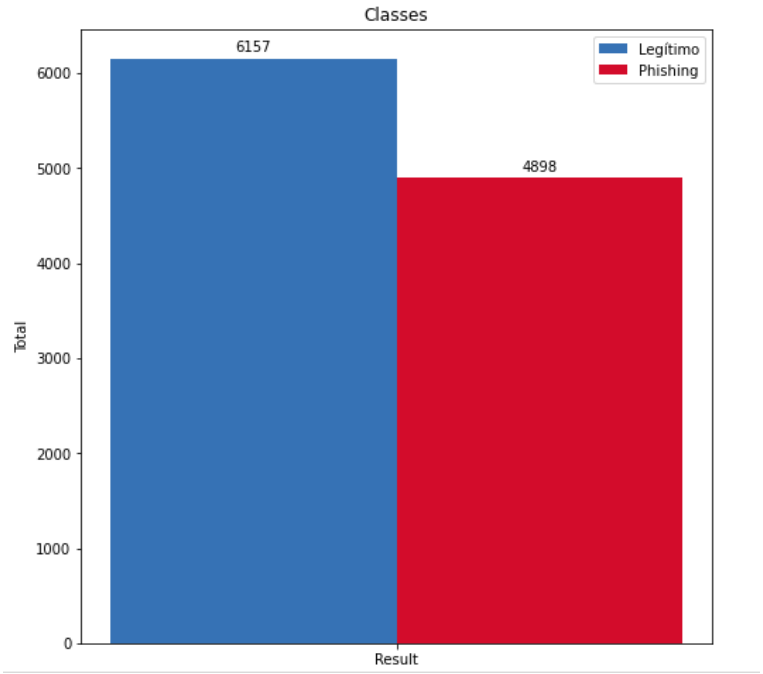

Figura 2.1: Classes

2.1 AVALIAÇÃO DOS ATRIBUTOS

Avaliou-se cada atributo em relação a sua distribuição, onde -1 (Phishing), 0 (Suspeito) e 1 (Legítimo). E como cada atributo se comporta como uma classificador avaliou-se o seu desempenho para a classificação dos sites utilizando as métricas F1, MCC e Acurácia. Os casos considerados como suspeitos não foram considerados na avaliação.

2.1.1 Características baseadas barra de endereço

A tabela 2.1, apresenta a distribuição dos dados do grupo (Características baseadas barra de endereço).

A tabela 2.2 apresenta os resultados de classificação dos atributos do grupo (Características baseadas barra de endereço)

2.1.2 Características baseadas em anormalidades

A tabela 2.3, apresenta a distribuição dos dados do grupo (Características baseadas em anormalidades).

A tabela 2.4 apresenta os resultados de classificação dos atributos do grupo (Características baseadas em anormalidades)

Tabela 2.1: Distribuição dos dados - Grupo (Características baseadas barra de endereço)

	-1	0	1
having_IP_Address	3793	0	7262
URL_Length	8960	135	1960
Shortining_Service	1444	0	9611
having_At_Symbol	1655	0	9400
double_slash_redirecting	1429	0	9626
Prefix_Suffix	9590	0	1465
having_Sub_Domain	3363	3622	4070
SSLfinal_State	3557	1167	6331
Domain_registeration_length	7398	0	3666
Favicon	2053	0	9002
port	1502	0	9553
HTTPS_token	1796	0	9259

Tabela 2.2: Classificação - Grupo (Características baseadas barra de endereço)

	Acurácia	F1	MCC
having_IP_Address	0.5623	0.6394	0.0942
URL_Length	0.4856	0.3035	0.0616
Shortining_Service	0.5193	0.6630	-0.0680
having_At_Symbol	0.5587	0.6864	0.0529
double_slash_redirecting	0.5294	0.6703	-0.0386
Prefix_Suffix	0.5756	0.3844	0.3486
having_Sub_Domain	0.6856	0.7361	0.3606
SSLfinal_State	0.8779	0.9032	0.7388
Domain_registeration_length	0.3752	0.2969	-0.2258
Favicon	0.5357	0.6614	-0.0003
port	0.5539	0.6861	0.0364
HTTPS_token	0.5238	0.6585	-0.0399

Tabela 2.3: Distribuição dos dados - Grupo (Características baseadas em anormalidade)

	-1	0	1
Request_URL	4495	0	6560
URL_of_Anchor	3282	5337	2436
Links_in_tags	3956	4449	2650
SFH	8440	761	1854
Submitting_to_email	2014	0	9041
Abnormal_URL	1629	0	9426

Tabela 2.4: Classificação - Grupo (Características baseadas em anormalidade)

	Acurácia	F1	MCC
Request_URL	0.6343	0.6821	0.2534
URL_of_Anchor	0.9675	0.9609	0.9339
Links_in_tags	0.6464	0.6172	0.3081
SFH	0.5532	0.3879	0.2224
Submitting_to_email	0.5432	0.6677	0.0182
Abnormal_URL	0.5189	0.6587	-0.0605

2.1.3 Características baseadas em HTML e JavaScript

A tabela 2.5, apresenta a distribuição dos dados do grupo (Características baseadas em HTML e JavaScript).

Tabela 2.5: Distribuição dos dados - Grupo (Características baseadas em HTML e JavaScript)

	-1	0	1
Redirect	0	9776	1279
on_mouseover	1315	0	9740
RightClick	476	0	10579
popUpWidnow	2137	0	8918
Iframe	1012	0	10043

A tabela 2.6 apresenta os resultados de classificação dos atributos do grupo (Características baseadas em HTML e JavaScript). O MCC para o Redirect não pode ser avaliado, pois a característica não classificou nenhum site como phishing.

Tabela 2.6: Classificação - Grupo (Características baseadas em HTML e JavaScript)

	Acurácia	F1	MCC
Redirect	0.5293	0.6922	-
on_mouseover	0.5569	0.6918	0.0418
RightClick	0.5546	0.7058	0.0127
popUpWidnow	0.5350	0.6590	0.0001
Iframe	0.5455	0.6899	-0.0034

2.1.4 Características baseadas no domínio

A tabela 2.7, apresenta a distribuição dos dados do grupo (Características baseadas no domínio).

A tabela 2.8 apresenta os resultados de classificação dos atributos do grupo (Características baseadas no domínio)

Tabela 2.7: Distribuição dos dados - Grupo (Características baseadas no domínio)

	-1	0	1
age_of_domain	5189	0	5866
DNSRecord	3443	0	7612
web_traffic	2655	2569	5831
Page_Rank	8201	0	2854
Google_Index	1539	0	9516
Links_pointing_to_page	548	6156	4351
Statistical_report	1550	0	9505

Tabela 2.8: Classificação - Grupo (Características baseadas no domínio)

	Acurácia	F1	MCC
age_of_domain	0.5637	0.5989	0.1215
DNSRecord	0.5563	0.6438	0.0757
web_traffic	0.7067	0.7765	0.3560
Page_Rank	0.5180	0.4086	0.1046
Google_Index	0.5854	0.7076	0.1290
Links_pointing_to_page	0.5650	0.7073	-0.0360
Statistical_report	0.5685	0.6954	0.0799

2.1.5 Conclusão

Algumas características classificam muitos site com uma determinada classe: having_At_Symbol, double_slash_redirecting, Favicon, port, HTTPS_token, Abnormal_URL, on_mouseover, RightClick, Iframe, Google_Index e Statistical_report.

O atributo Prefix_Suffix não apresentou nenhum Falso Positivo. A figura 2.2 apresenta sua matriz de confusão.

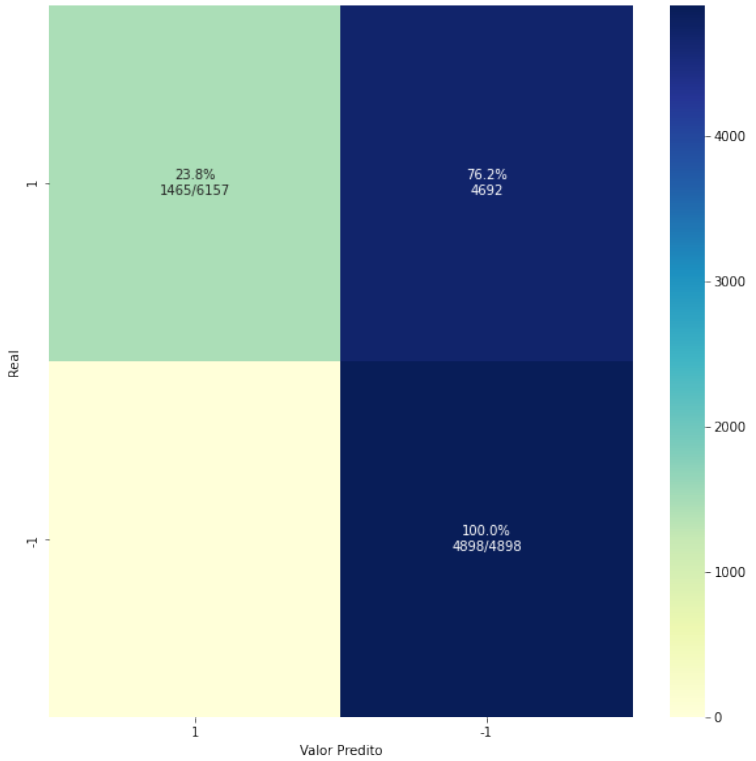

Figura 2.2: Matriz de Confusão - Prefix_Suffix

Os melhores resultados para a classificação foram obtidos pelos atributos: Prefix_Suffix, having_Sub_Domain, SSLfinal_State, Request_URL, URL_of_Anchor, Links_in_tags, SFH, web_traffic, Page_Rank, Google_Index.

3 SELEÇÃO DE ATRIBUTOS

Como o dataset possui 30 características muitas não são relevantes, portanto na parte de seleção de atributos procurou-se obter um subset dos atributos que possuem um desempenho satisfatório. O primeiro grupo foi definido a partir da exploração de dados. Foi avaliado também os 4 grupos apresentados pelos autores do trabalho: barra de endereço, anormalidades, HTML e JavaScript e Domínio.

Também avaliou-se duas ferramentas de feature selection do sklearn baseados em modelos. No grupo L1 utilizou-se um LinearSVC (C=0.001, penalty="L1") e no grupo tree utilizou-se um ExtraTreesClassifier padrão. A lista a seguir apresenta os grupos juntamente com os atributos selecionados:

- Grupo 1 (Exploração de Dados) = Prefix_Suffix, having_Sub_Domain, SSLfinal_State, Request_URL, URL_of_Anchor, Links_in_tags, SFH, web_traffic, Page_Rank, Google_Index
- Grupo 2 (Barra de Endereço) = having_IP_Address, URL_Length, Shortining_Service, having_At_Symbol, double_slash_redirecting, Prefix_Suffix, having_Sub_Domain, SSLfinal_State, Domain_registeration_length, Favicon, port, HTTPS_token
- Grupo 3 (Anormalidades) = Request_URL, URL_of_Anchor, Links_in_tags, SFH, Submitting_to_email, Abnormal_URL
- Grupo 4 (HTML e JavaScript) = Redirect, on_mouseover, RightClick, popUpWidnow, Iframe
- Grupo 5 (Domínio) = age_of_domain, DNSRecord, web_traffic, Page_Rank, Google_Index, Links_pointing_to_page, Statistical_report
- Grupo 6 (L1) = Prefix_Suffix, having_Sub_Domain, SSLfinal_State, Request_URL, URL_of_Anchor, Links_in_tags, SFH, web_traffic, Google_Index
- Grupo 7 (Tree) = Prefix_Suffix, having_Sub_Domain, SSLfinal_State, URL_of_Anchor, Links_in_tags, web_traffic

Observa-se que a seleção L1 obteve resultados parecidos com a seleção manual proveniente da exploração de dados. Nenhuma das 3 seleções (1,6 e 7) selecionaram atributos do Grupo 4 (HTML e JavaScript).

Por último, um grupo 8 (soma) foi proposto que soma os resultados dos 4 grupos apresentados pelos autores e depois realiza o rescale min-max das quatro somas resultantes. O rescale tem como objetivo nivelar os resultados entre os grupos, pois eles possuem número de atributos diferentes. O min-max foi definido no intervalo de [-1,1]. E o Grupo 9 é o grupo com todas as características.

4 CLASSIFICAÇÃO

Na parte de classificação o banco foi divido em 80% para treinamento e 20% para teste. Na etapa de treinamento (4.2) foi utilizando a metologia de Kfold Cross validation (5 Folds), para selecionar 3 grupos de atributos para as etapas posteriores juntamente com o seu modelo de melhor desempenho. Os modelos testados em todos os grupos foram: LogisticRegression (Regressão Logística), LinearSVC (SVM Linear) e Random Forest.

Na etapa de GridSearchCV (4.3) utilizou um GridSearch juntamente com Kfold Cross Validation (5 Folds) para ajustar os parâmetros dos modelos com melhores desempenho em seus respectivos grupos. Por fim, os 3 modelos foram para a etapa de teste (4.4) que consiste no treinamento do modelo na base completa de treinamento e avaliação na base de teste.

4.1 BANCO DE TREINAMENTO E TESTE

A base foi divida utilizando o split train test do sklearn que separa os dados de forma estratificada. A figura 4.1 apresenta a distribuição da base de treinamento

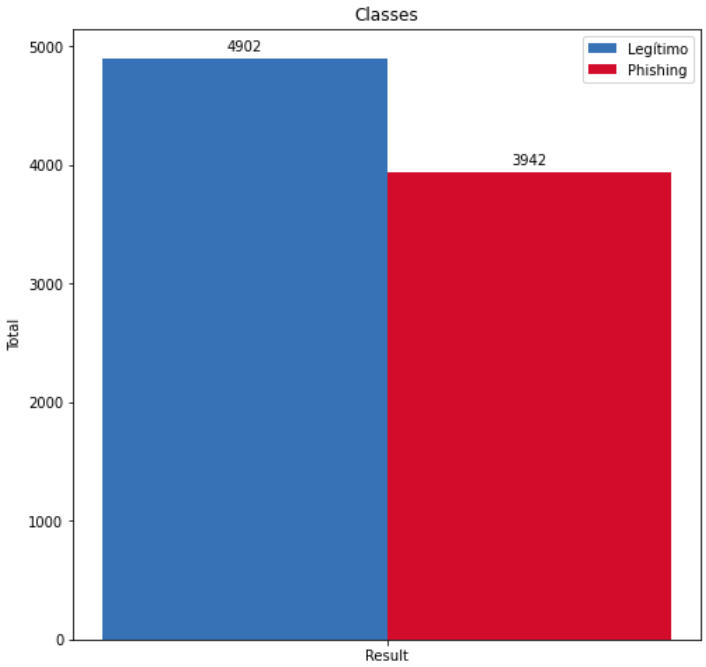

Figura 4.1: Base de Treinamento

A figura 4.2 apresenta a distribuição da base de teste.

4.2 TREINAMENTO

Para seleção dos grupos utilizou-se a métrica F1 para a comparação dos resultados. A tabela 4.1 apresenta os resultados para os grupos e para os modelos. Todos os modelos (LinearSVC, Logistic Regression e Random Forest) foram utilizados em sua forma padrão definida pelo sklearn. Para a separação dos folds utilizou-se o StratifiedKFold(n_splits=5)

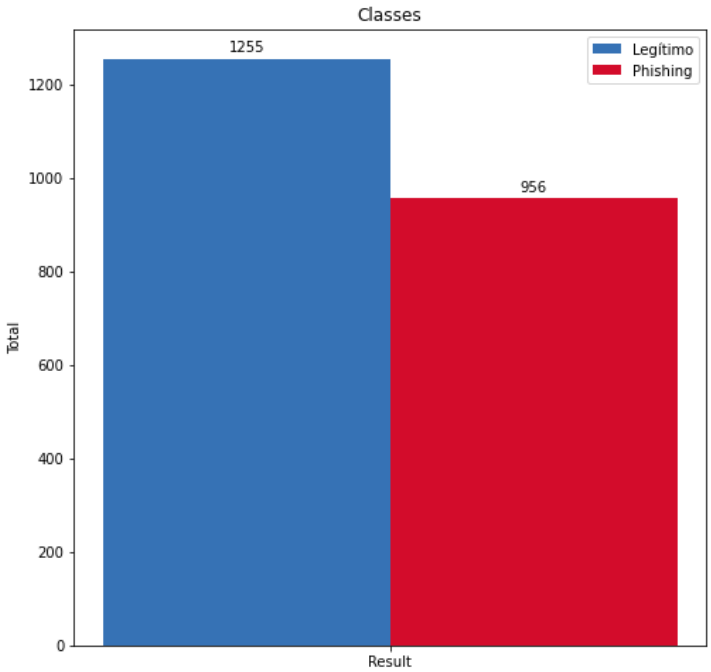

Figura 4.2: Base de Teste

Tabela 4.1: Distribuição por características

	Random Forest	LogisticRegression	LinearSVC
Grupo 1	0.952671	0.930683	0.931840
Grupo 2	0.917702	0.904317	0.905580
Grupo 3	0.890296	0.878645	0.881101
Grupo 4	0.714688	0.703560	0.703560
Grupo 5	0.773944	0.715063	0.715734
Grupo 6	0.948334	0.930729	0.932076
Grupo 7	0.942996	0.923717	0.925873
Grupo 8	0.864812	0.773261	0.773064
Grupo 9	0.972709	0.935974	0.936362

Para as etapas posteriores foram selecionados os grupos 1, 7, 9. Todos os modelos se beneficiaram da utilização de todos os atributos, entretanto a Random Forest obteve um resultado consideravelmente melhor que os outros modelos no grupo 9. A Random Forest foi superior aos demais modelos em todos os grupos. O grupo 7 com apenas 6 atributos obteve resultados bem parecidos que o grupo 6 que possuía 9, portanto selecionou-se o grupo 7 como terceira escolha.

Em relação aos grupos definidos pelos autores (2,3,4,5) o pior desempenho foi apresentado pelo Grupo 4 (HTML e JavaScript). Esses resultados explicam porque não foi selecionado nenhum atributo do Grupo 4 pelos grupos (1, 6 e 7). O melhor foi o Grupo 2 (Barra de Endereço).

4.3 GRIDSEARCHCV

Utilizando a mesma base de treinamento e o mesmo StraitifiedKFold. Utilizou-se o módulo do Sklearn GridSearchCV para selecionar os parâmetros da Random Forest para cada grupo. O universo de busca foi:

- n_estimators = [50,100,200,500]
- max_features = ['auto', 'sqrt']
- max_depth = [None, 10,25,50]
- min_samples_split = [2, 16, 64]
- min_samples_leaf = [1, 16, 64]
- bootstrap = [True, False]
- criterion = ['gini', 'entropy']

Totalizando 1152 possibilidades. Os demais parâmetros da Random Forest foram mantidos nos valores padrões. A métrica F1 foi utilizada para definir o melhor conjunto de parâmetros. Os parâmetros selecionados para cada grupo foram:

- Grupo 1 = 'bootstrap': True, 'criterion': 'gini', 'max_depth': None, 'max_features': 'auto', 'min_samples_leaf': 1, 'min_samples_split': 2, 'n_estimators': 200
- Grupo 7 = 'bootstrap': True, 'criterion': 'gini', 'max_depth': 50, 'max_features': 'sqrt', 'min_samples_leaf': 1, 'min_samples_split': 16, 'n_estimators': 50
- Grupo 9 = 'bootstrap': False, 'criterion': 'entropy', 'max_depth': None, 'max_features': 'sqrt', 'min_samples_leaf': 1, 'min_samples_split': 2, 'n_estimators': 100

E os novos resultados para o treinamento com cross validation encontrados foram:

Tabela 4.2: GridSearchCV

	F1
Grupo 1	0.95346
Grupo 7	0.944187
Grupo 9	0.974146

4.4 TESTE

A tabela 4.3 apresenta os resultados dos grupos na base de teste.

Tabela 4.3: Teste

	F1	Acurácia	MCC
Grupo 1	0.9505	0.9435	0.8847
Grupo 7	0.9491	0.9412	0.8803
Grupo 9	0.9719	0.9679	0.9346

Os resultados dos testes tiveram o mesmo comportamento que os de treinamento. A Random Forest se beneficia da utilização de todos os atributos (Grupo 9). Entretanto, o resultado do Grupo 7 foram interessantes, pois eles são próximos ao do grupo 1 utilizando menos atributos.

Portanto, o grupo 7 com somente 6 atributos atingiu resultados de F1-score de 94% que é 2.5% inferior que o grupo 9. Portanto, o grupo 7 foi selecionado para uma análise completa dos resultados de treinamento e teste.

5 VISUALIZAÇÃO DOS RESULTADOS - GRUPO (TREE)

Por último realizou uma avaliação aprofundada da Grupo (Tree) tanto para resultados na fase de treinamento como teste

5.1 TREINAMENTO

5.1.1 Percentage Split

Na fase de treino percentage split separou-se o base de treinamento em 80% para treino e 20% para teste. Manteve-se os parâmetros da Random Forest encontrados no GridSearchCV. A figura 5.1 apresenta a curva ROC do classificador e a figura 5.2 a sua matriz de confusão.

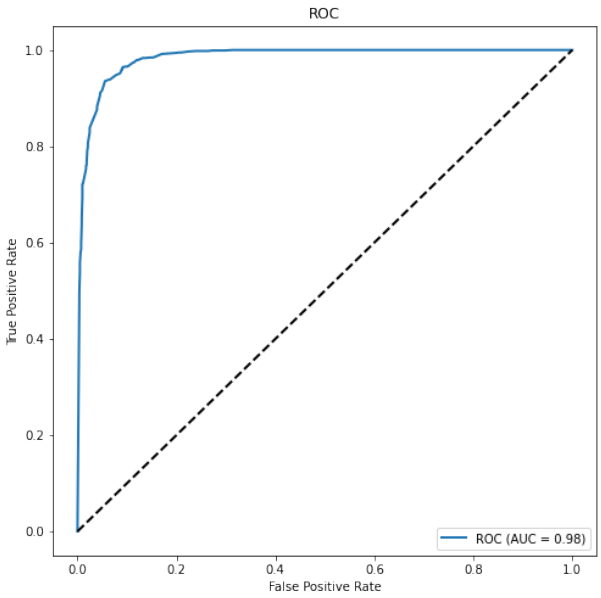

Figura 5.1: Curva Roc - Treino (Percentage Split)

A tabela 5.1 apresenta os resultados do treino percentage split.

Tabela 5.1: Resultados - Treino (Percentage Split)

	Erro Absoluto	Precisão	Recall	F1	Acurácia	MCC
Resultados	0.1232	0.9259	0.9635	0.9443	0.9384	0.8763

5.1.2 KFold Cross Validation

Na fase de Cross Validation foi utilizado a mesma metologia utilizada anteriormente utilizando o StratifiedKFold(n_split=5). A figura 5.3 apresenta a curva ROC para cada fold e a curva média (mean).

As figura de 5.4 a 5.8 apresentam as matrizes de confusão para cada fold.

A tabela apresenta os resultados para cada fold e a média dos resultados.

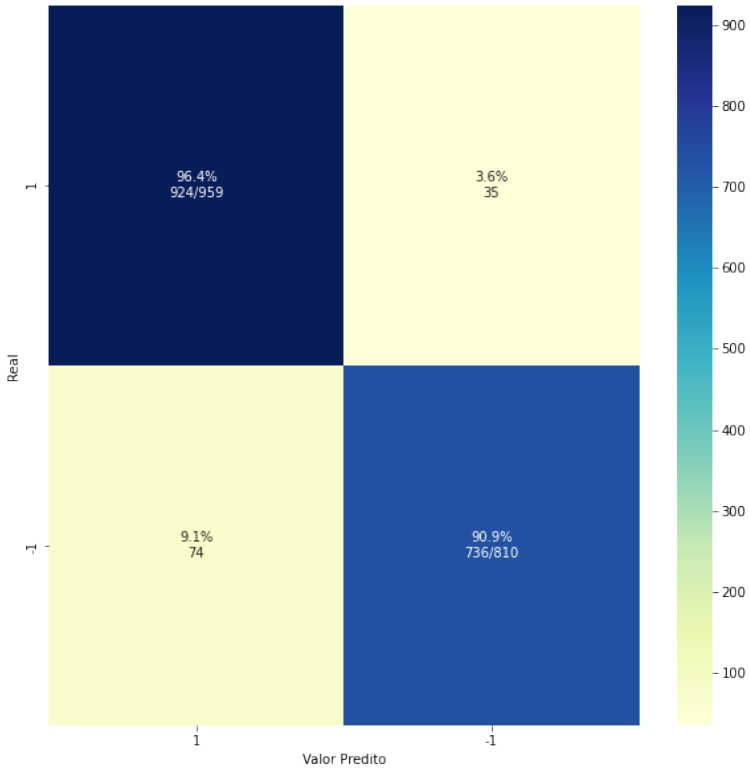

Figura 5.2: Matriz de Confusão - Treino (Percentage Split)

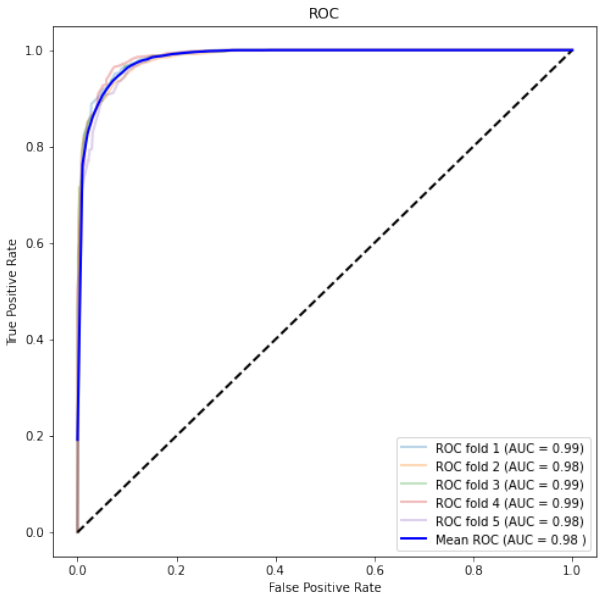

Figura 5.3: Curva Roc - Treino (CV)

5.2 TESTE

Para os resultados para a base de teste seguiu a mesma metologia da seção 4.4. A figura 5.9 apresenta a curva Roc para a fase de teste.

A figura 5.10 apresenta a matriz de confusão para fase de teste.

A tabela 5.3 apresenta os resultados para a fase de teste

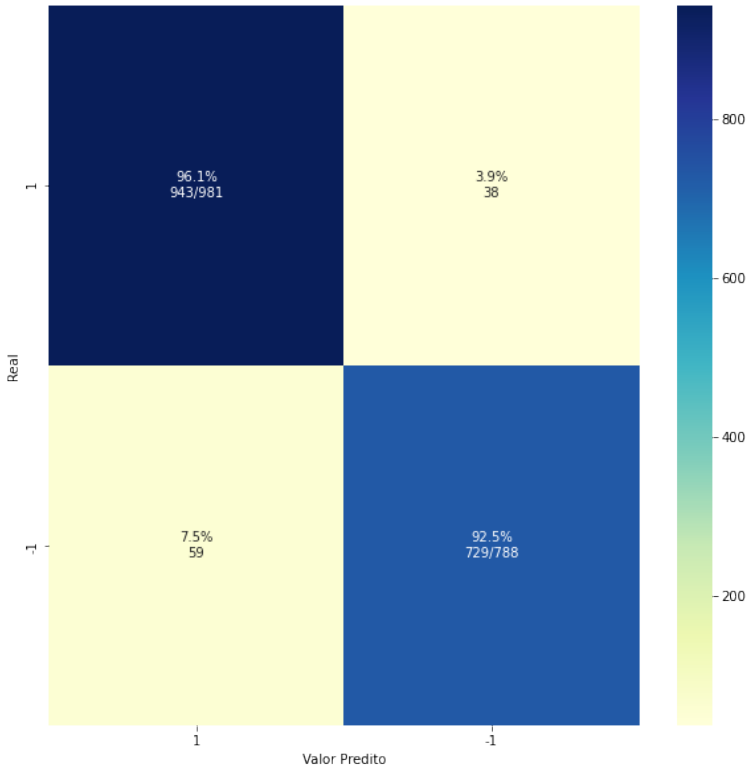

Figura 5.4: Matriz de Confusão - Treino (CV1)

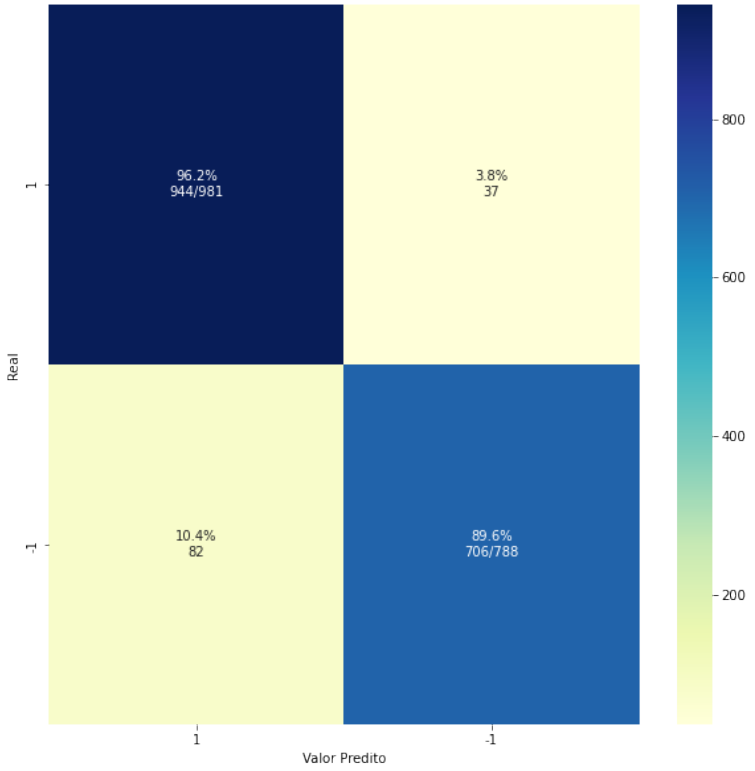

Figura 5.5: Matriz de Confusão - Treino (CV2)

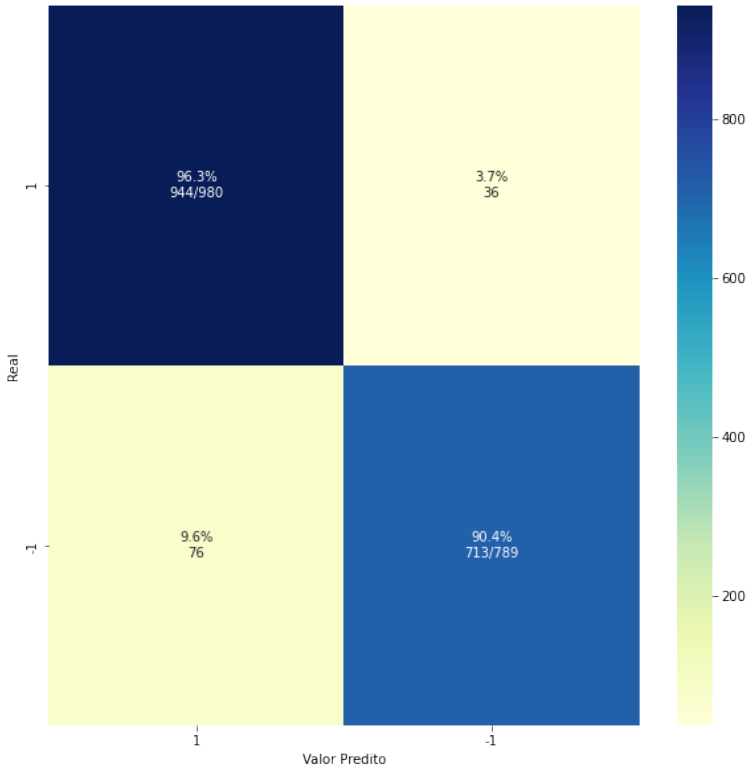

Figura 5.6: Matriz de Confusão - Treino (CV3)

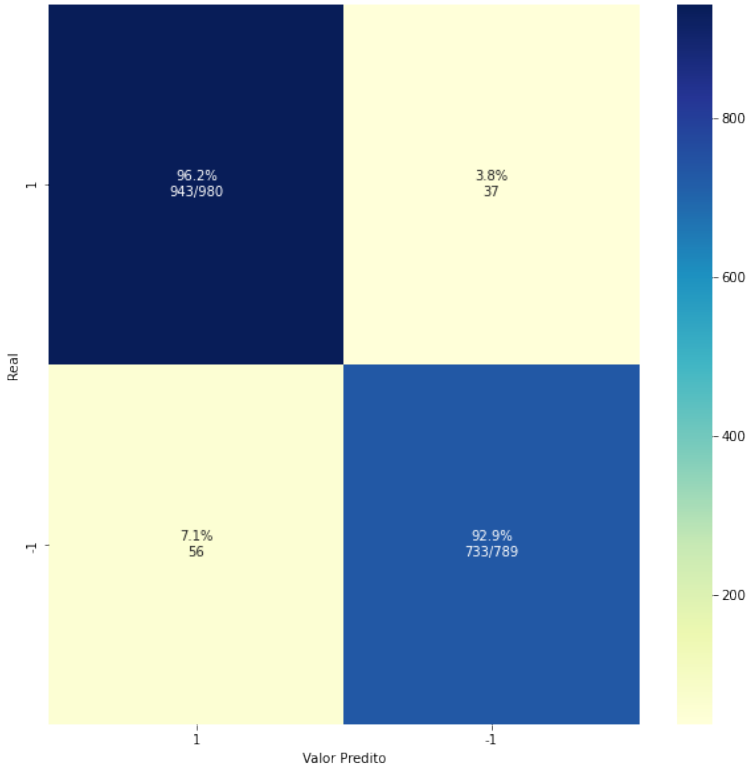

Figura 5.7: Matriz de Confusão - Treino (CV4)

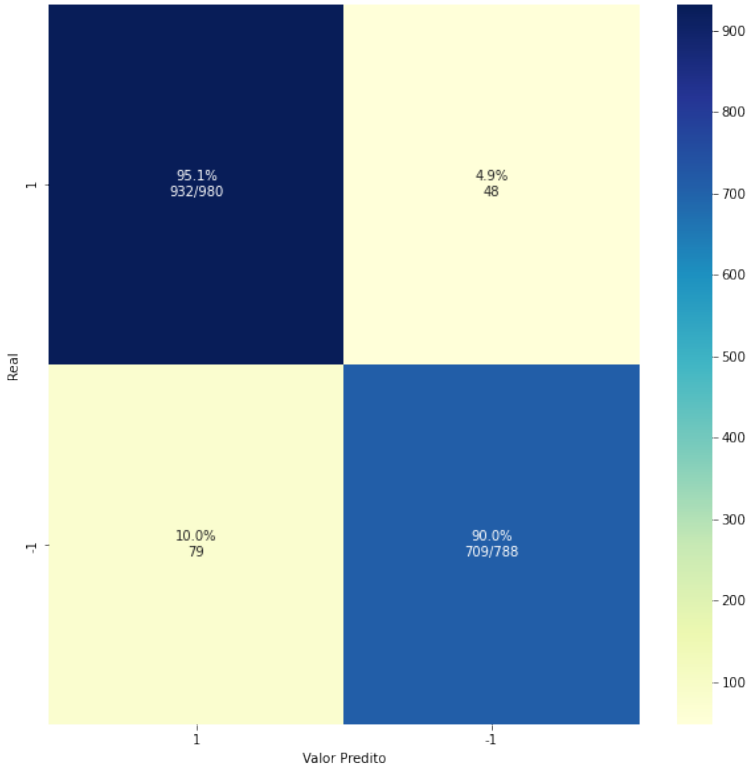

Figura 5.8: Matriz de Confusão - Treino (CV5)

Tabela 5.2: Resultados - Treino (CV)

	Erro Absoluto	Precisão	Recall	F1	Acurácia	MCC
Fold 1	0.1221	0.9405	0.9500	0.9452	0.9389	0.8763
Fold 2	0.1436	0.9138	0.9613	0.9369	0.9282	0.8552
Fold 3	0.1221	0.9266	0.9663	0.9460	0.9389	0.8768
Fold 4	0.1051	0.9422	0.9643	0.9531	0.9474	0.8936
Fold 5	0.1437	0.9219	0.9510	0.9362	0.9282	0.8546
Média	0.1273	0.9290	0.9586	0.9435	0.9363	0.8713

Tabela 5.3: Resultados - Teste

	Erro Absoluto	Precisão	Recall	F1	Acurácia	MCC
Resultados	0.1185	0.9343	0.9633	0.9486	0.9408	0.8793

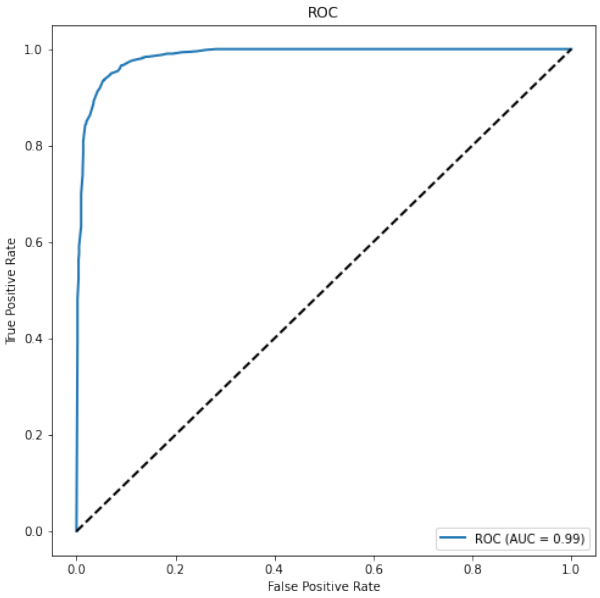

Figura 5.9: Curva Roc - Teste

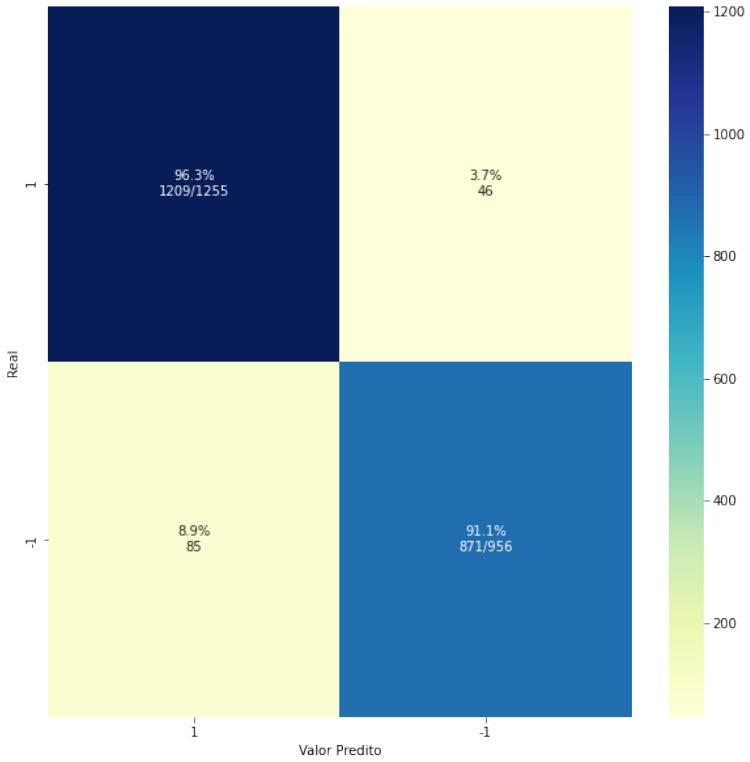

Figura 5.10: Matriz de Confusão - Teste

6 CONCLUSÃO

O trabalho conseguiu desenvolver um classificador utilizando somente alguns atributos do banco de dados original que consegue obter resultados satisfatórios. Os atributos escolhidos foram: Prefix_Suffix, having_Sub_Domain, SSLfinal_State, URL_of_Anchor, Links_in_tags, web_traffic. Mostrando que sua analise juntamente com as heurísticas apresentadas são eficientes em reconhecer phishing websites. Não foi escolhido nenhum atributo pertencente a grupo de atributos baseados em HTML e JavaScript, esse comportamento também foi observado nos grupos 1 (Exploração de dados) e Grupo 6 (L1).

Analisando os atributos por seus respectivos grupos, os atributos do grupo que possui características baseadas em HTML e JavaScipt foram os com piores resultados e o grupo baseado na barra de endereço foi o com melhores resultados.

O banco utilizados foi entregue ja tratado e anonimizados, portanto os resultados encontrados podem distorcer de uma análise real que provavelmente possuirá dados com comportamentos mais anormais. Esses comportamentos podem influenciar a avaliação das heurísticas e consequentemente o desempenho do classificador.

Em relação aos modelos a Random Forest apresentou resultados superiores que o LinearSVC e a RegressionLogistic, entretanto esse comportamento pode ser devido aos valores padrões dos outros classificadores não serem adequados que foram os parâmetros utilizados na etapa de treinamento. A Random Forest apresentou os melhores resultados com número de atributos maiores, obtendo os melhores resultados com todos os atributos (Grupo 9). Esse comportamento também foi apresentado nos outros classificadores (LinearSVC e RegressionLogistic).