URL Detect

TAA

Tomás Matos 108624
Diogo Almeida 108902

deti
universidade de aveiro
departamento de eletrónica,
telecomunicações e informática

Introdução

- Um URL é usado para localizar recursos, como páginas da web, nem sempre bem intencionados.
- Podem levar a ataques como: phishing, malware, fraude, data bridge, entre outros.
- Vamos fazer um distinção do caráter de um url entre **Maligno** e **Benigno**, utilizando diversos classificadores .
- Analizar profundamente o dataset escolhido.
- Apresentar os resultados perante os modelos criados

Análise do Dataset
Estrutura

- **url**: URL do website.
- **ip_add**: Endereço IP do website.
- **geo_loc**: Localização do endereço IP.
- **url_len**: Tamanho do URL.
- **js_len**: Tamanho do JavaScript.
- **js_obf_len**: Tamanho do JavaScript escondido.
- **tld**: Top Level Domain do website.
- **who_is**: Indica se a informação do WHOIS está completa ou não.
- **https**: Indica se o website usa HTTPS ou HTTP.
- **content**: Todo o conteúdo do website incluindo texto do JavaScript.
- **label**: Classificação de Bom ou Mau.

Análise do Dataset
Exemplo

	url	ip_add	geo_loc	url_len	js_len	js_obf_len	tld	who_is	https	content	label
0	http://members.tripod.com/russiastation/	42.77.221.155	Taiwan	40	58.0	0.0	com	complete	yes	Named themselves charged particles in a manly ...	good
1	http://www.ddj.com/cpp/184403822	3.211.202.180	United States	32	52.5	0.0	com	complete	yes	And filipino field \n \n \n \n \n \n \n the...	good
2	http://www.naef-usa.com/	24.232.54.41	Argentina	24	103.5	0.0	com	complete	yes	Took in cognitivism, whose adherents argue for...	good
3	http://www.ff-b2b.de/	147.22.38.45	United States	21	720.0	532.8	de	incomplete	no	fire cumshot sodomize footaction tortur failed...	bad
4	http://us.imdb.com/title/tt0176269/	205.30.239.85	United States	35	46.5	0.0	com	complete	yes	Levant, also monsignor georges. In 1800, lists...	good
...	...	...	...	...	...	...	...	...	...	...	...
1199995	http://csrc.nist.gov/rbac/	62.120.245.128	Saudi Arabia	26	106.0	0.0	gov	complete	yes	There. this high gdp per capita of any other c...	good
1199996	http://www.unm.edu/~hist/	72.178.170.132	United States	25	36.0	0.0	edu	complete	no	Institute or older use of transmission media (...	good
1199997	http://www.syfyportal.com/news423380.html	181.240.45.113	Colombia	41	178.5	0.0	com	incomplete	yes	Both increase was deemed too imprecise to be b...	good
1199998	http://www.wardkenpo.ie	15.75.59.60	United States	23	121.0	0.0	ie	complete	yes	Pathway, metabolic cat's spinal mobility and f...	good
1199999	http://homepages.gotadsl.co.uk/~jgm/ekmm/	168.239.57.229	United States	41	68.0	0.0	co.uk	complete	no	Latitudinal distribution highest level. Leader...	good

Análise do Dataset
Divisão dos Dados

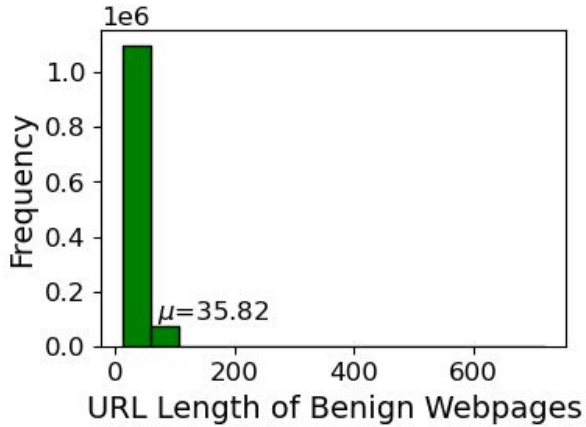
Dataset de Treino: Significante parte do dataset original é usado apenas para treino do modelo ML.

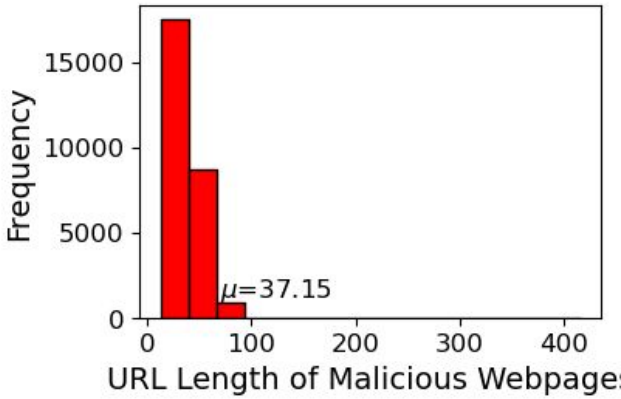
Dataset de Test: O restante do dataset original é usado para fins de teste do modelo criado.

Análise do Dataset

URL Length

- URL é o identificador primário de um website.
- Calculado tamanho de todos os URLs.
- Tamanhos muito idênticos e por isso não é muito útil para o modelo.



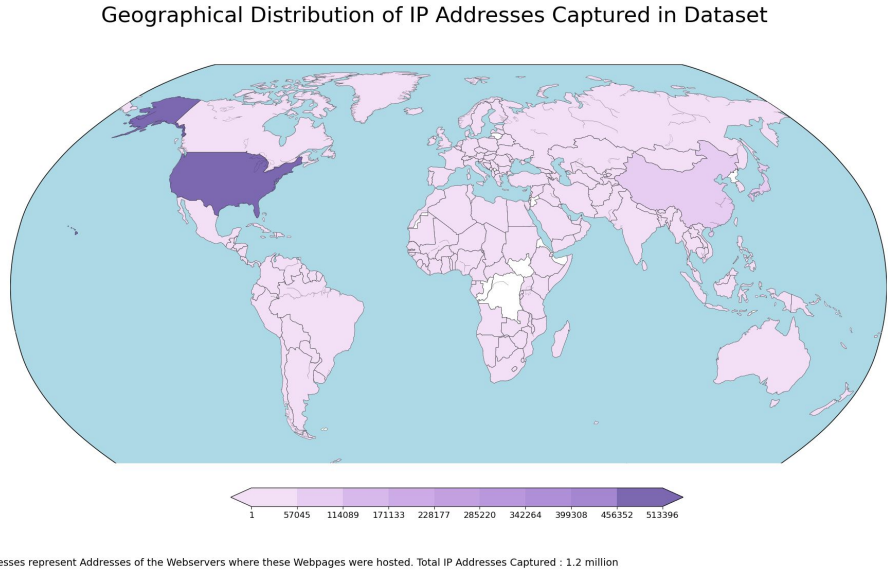


Análise do Dataset

Ip Address

- Endereço IP do web server onde está o website.

- Atributo não dá nenhuma conclusão útil para o modelo.

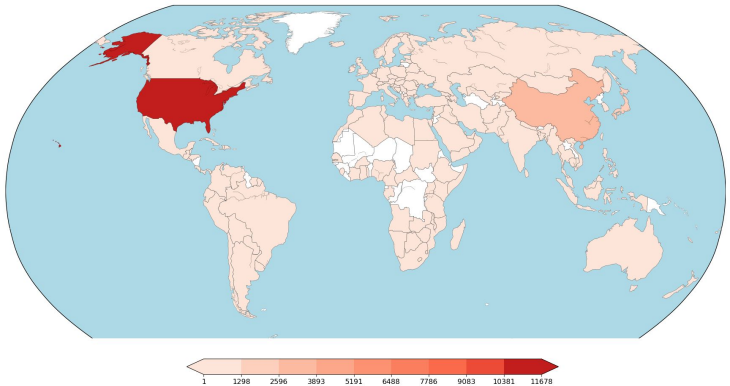


Análise do Dataset

Geo Location

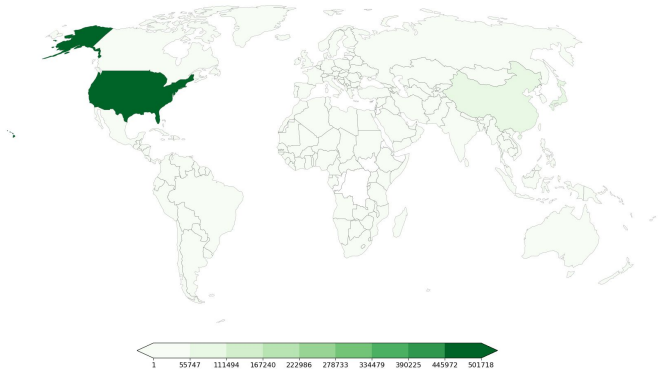
- País onde os endereços IPs dos sites bons e maus pertencem.
- Muitas semelhanças entre ambos e por isso não é útil para o modelo.

Geographical Distribution of IP Addresses: Malicious Webpages



Note: Location shown here depicts the Webserver where these Webpages were hosted. Total Malicious Webpages : 27253

Geographical Distribution of IP Addresses: Benign Webpages

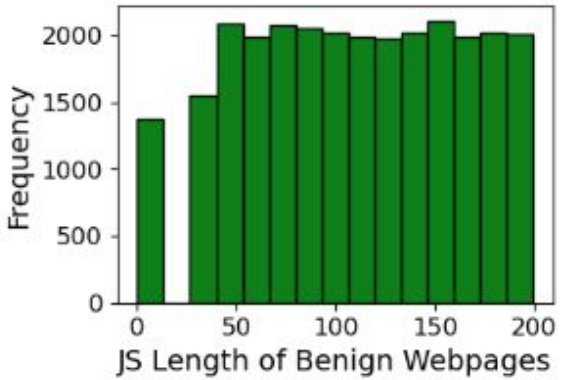


Location shown here depicts the Webserver where these Webpages were hosted. Total Benign Webpages: 1.172 million

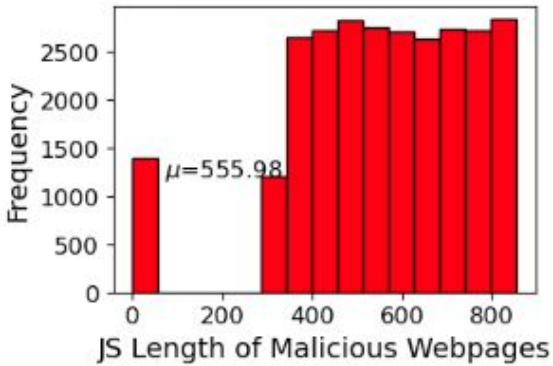
Análise do Dataset

JavaScript Length

- Foi descarregado o JavaScript visível dos websites e armazenado no dataset o respectivo comprimento.
- Websites benignos apresentam JavaScript com tamanho entre 50 e 200.



- Websites malignos apresentam JavaScript com tamanho entre 250 e 800.
- Atributo usado como feature do modelo.

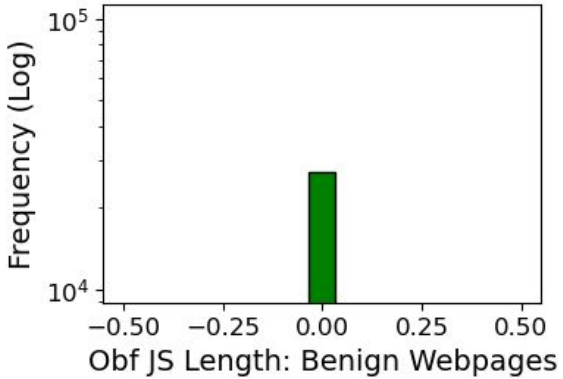


Análise do Dataset

Obfuscated JavaScript Length

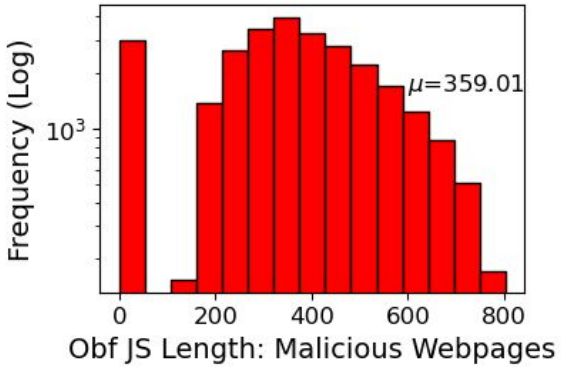
- Foi também calculado o tamanho do JavaScript não visível ou que não é apresentado de forma legível.

- Websites benignos normalmente não apresentam obfuscated JavaScript



- Websites malignos apresentam Obfuscated JavaScript com tamanho entre 150 a 800.

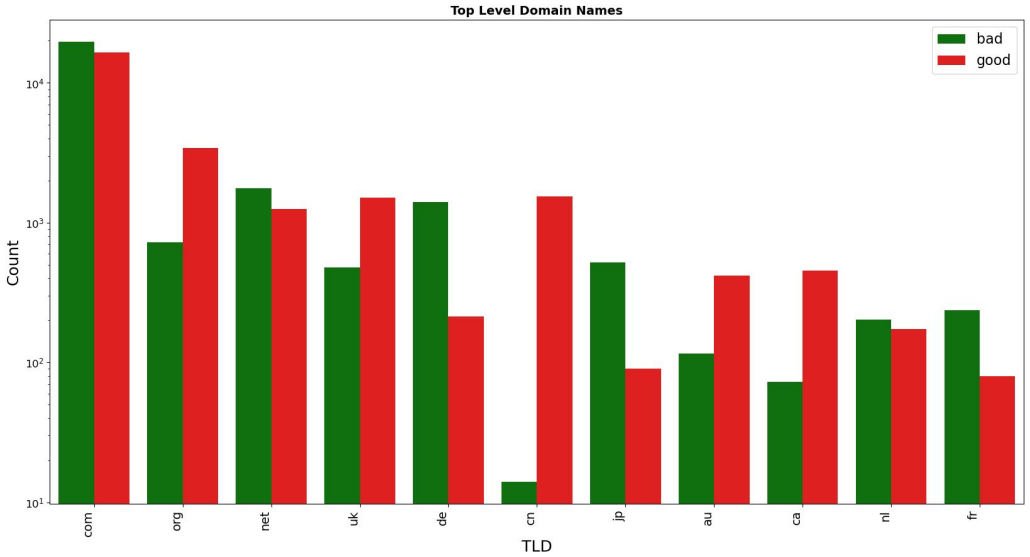
- Atributo usado como feature do modelo pois é muito indicativo.



Análise do Dataset

Top Level Domain

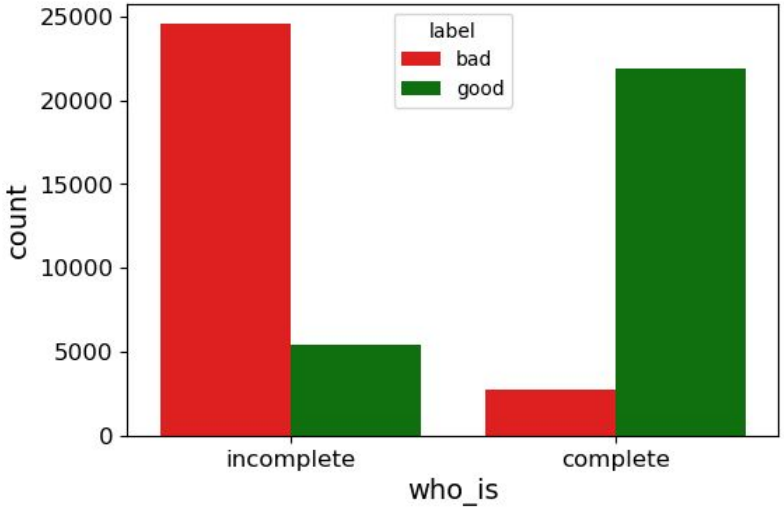
- Top Level Domain é mostrado depois do “.” no URL do website.
- Não tem uma diferença clara para ser usado no modelo.



Análise do Dataset

Who Is

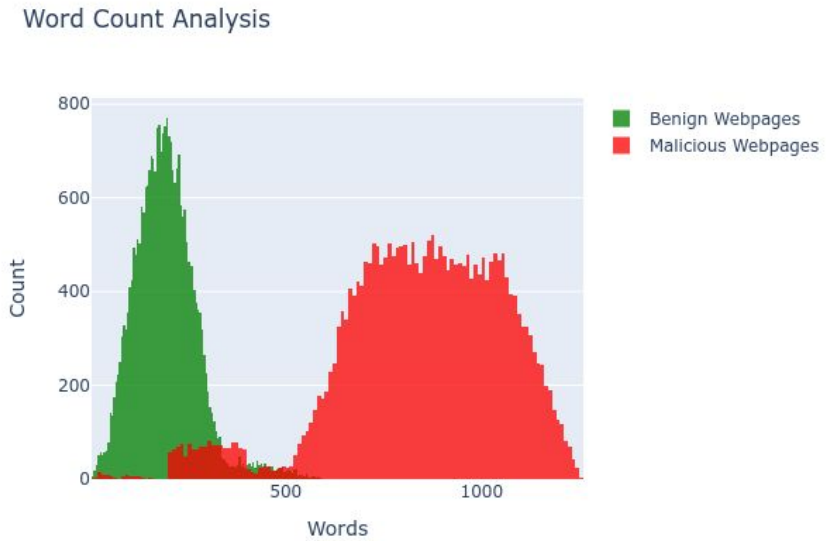
- O atributo “Who Is” indica se o registro do domínio está completo ou não.
- Contém registo sobre o nome do domínio, informações sobre o criador do domínio, data de expiração e nomes de servidores.
- Sites malignos não costumam ter este registo completo e por isso utilizámos este atributo como feature do modelo.



Análise do Dataset

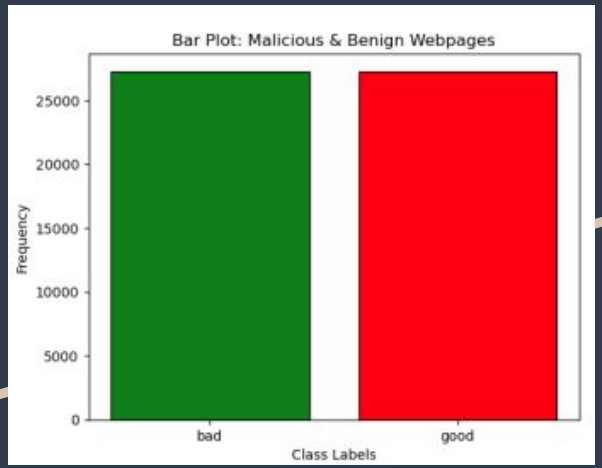
Content

- Este atributo contém todo o texto extraído dos websites incluindo JavaScript devidamente limpo.
- Utilização deste conteúdo vetorizado no modelo.
- Foi feito o cálculo do número de palavras nos sites benignos e malignos.



Processamento dos dados

Equilibrar dados consoante as labels



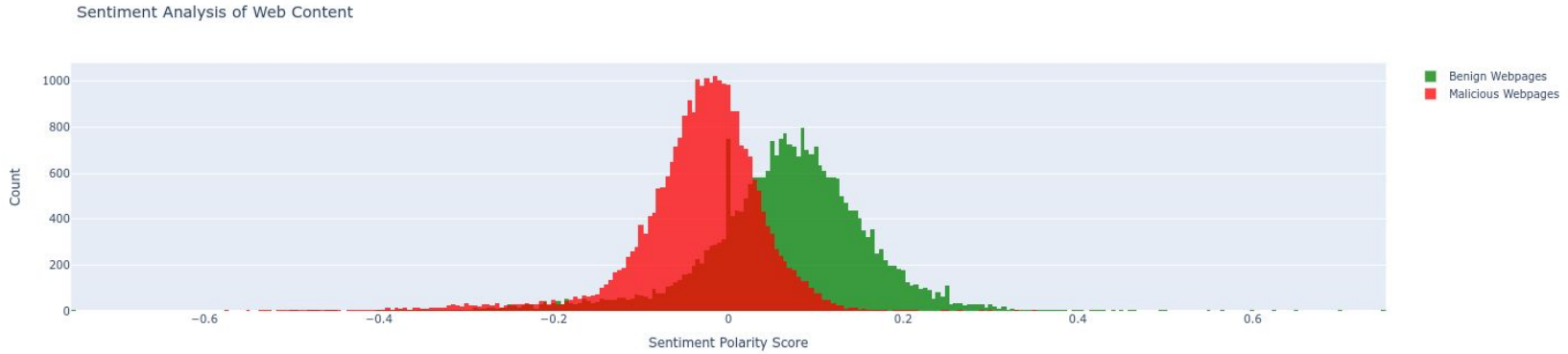
27253 elementos cada label

- Garantir que o modelo aprenda de igual forma todas as classes, evitando classes majoritárias.
- Manter a precisão e o desempenho real do modelo em todas as classes.
- Modelo generalizar melhor para novos dados.
- Divisão do dataset.
- Diminuir número de elementos com label “good”, uma vez que eram os mais abundantes de forma aleatória.

Processamento dos dados
Polaridade sentimental do conteúdo Web

- Consiste na direção e intensidade do sentimento expresso em um texto ou conteúdo.
- Biblioteca **TextBlob**, usa um modelo pré-treinado para analisar padrões e atribuir um sentimento

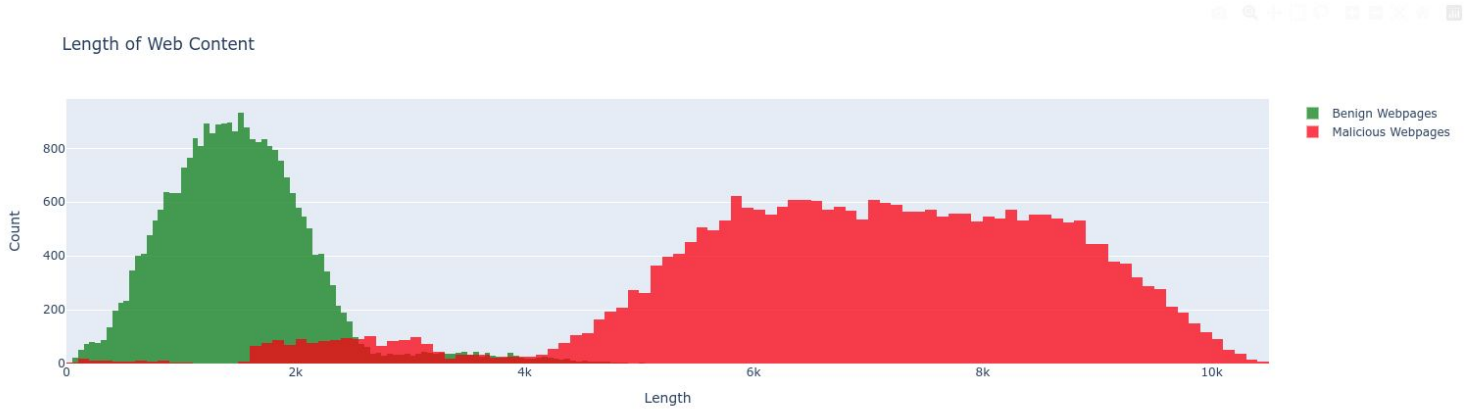
- **TextBlob**, utilizamos um método que para um conteúdo fornece um valor entre [-1, 1]:
  - -1 : sentimento forte e negativo.
  - 0 : sentimento neutro.
  - 1 : sentimento forte e positivo.



Processamento dos dados
Tamanho do conteúdo Web

- Número de caracteres presentes no conteúdo web.
- Conteúdos web com maior dimensão tendem a ter um caráter malicioso.

- Incluído espaços, sinais de pontuação e símbolos especiais.



Processamento dos dados

Vetorização do conteúdo web

- Fornecer aos modelos uma análise mais aprofundada do feature “contente”
- Converter uma informação textual em numérica de forma a poder ser usada com feature de entrada nos modelos.
- Utilização da biblioteca FastText.

Processamento dos dados
FastText

Explicação

- Biblioteca para processamento de linguagem natural.
- Um modelo FastText aprende representações vetoriais das palavras.
- Classificando informações de subpalavras, sendo possível lidar com palavras de vocabulário e línguas morfologicamente ricas.
- Ex: ”apple” possível representação: [”ap”, ”app”, ”ppl”, ”ple”,”le”].

Utilização

- Foi treinado um modelo FastText, a partir do “content” associado à sua label.
- Durante o treino, o modelo aprende representações para as palavras e subpalavras continuamente.
- Vetores são atualizados de forma iterativa.

Processamento dos dados
FastText

- Tendo criado o modelo, para cada elemento do dataset vamos fazer a conversão do texto para vetor.
- De forma a obter este vetor, para cada palavra do texto é consultado o modelo para obter o seu vetor representante.
- Com os vetores de todas a palavras do texto, estes são combinados.
- O vetor para cada “content” tem um tamanho de 100

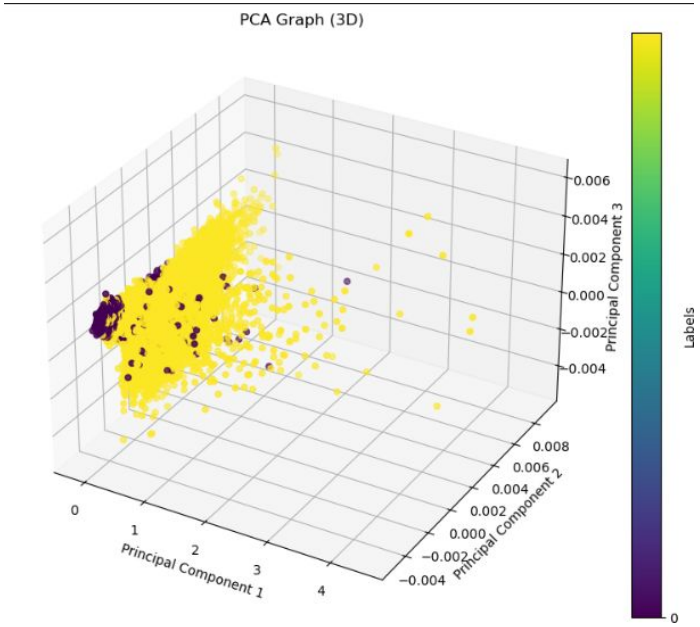


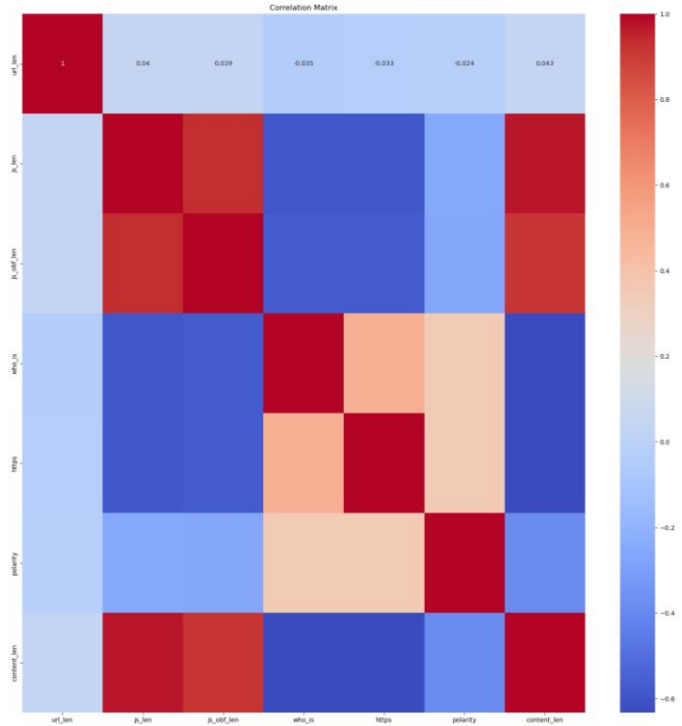
Gráfico PCA : Principal Component Analysis

Modelos
Features de entrada

- Features que foram utilizadas:
  - Tamanho do URL.
  - Tamanho do JavaScript.
  - Tamanha do JavaScript Ofuscado.
  - Se o URL tem https ou não.
  - “Who is”, se o registo do URL está completo ou não.
  - Polaridade sentimental do “content”.
  - Tamanho do “content”.
  - 100 elementos do vetor relativo ao “content”

```
Most impactful features:
content_len     1.000000
js_len          0.969903
js_obf_len      0.912953
who_is          0.632957
https           0.631256
polarity        0.386906
url_len         0.043093
```

Features mais impactantes, sem o vetor



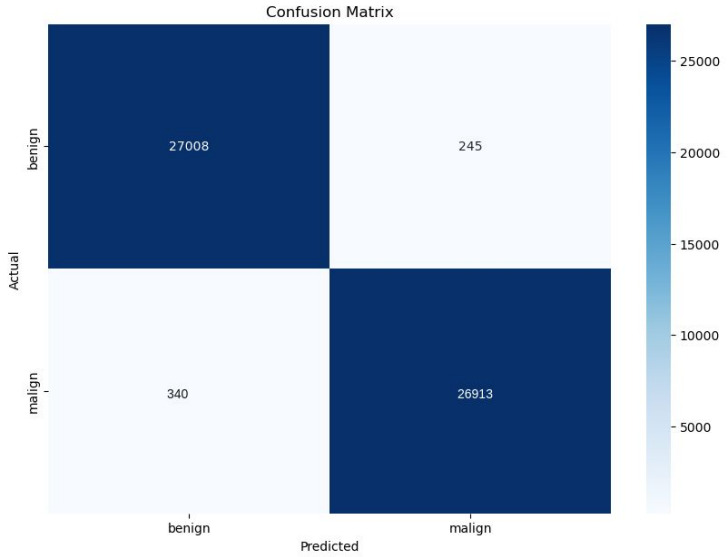
Matriz de Correlação

Modelos
Logistic Regression

- Simples e eficaz.
- Útil em datasets grandes.

Resultados Obtidos

Classifier	Accuracy	Precision	Recall	F1 score
Logistic Regression	0.9891	0.99	0.99	0.99



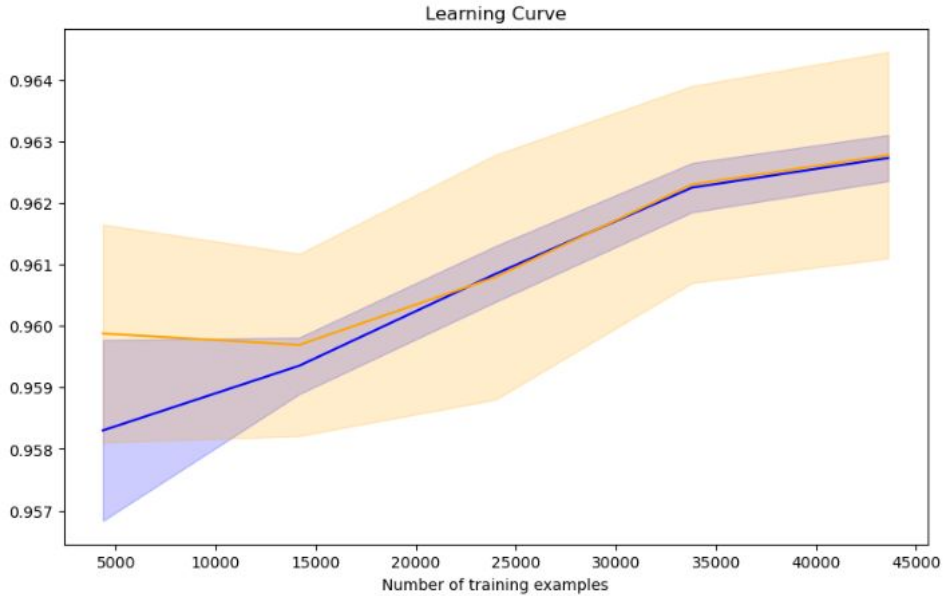
Matriz de confusão

Modelos

Logistic Regression

Learning curve

- A precisão de treino e de validação estão a aumentar.
- Mantendo-se juntas e não divergindo, indica:
  - Estabilidade do modelo.
  - Não há overfitting

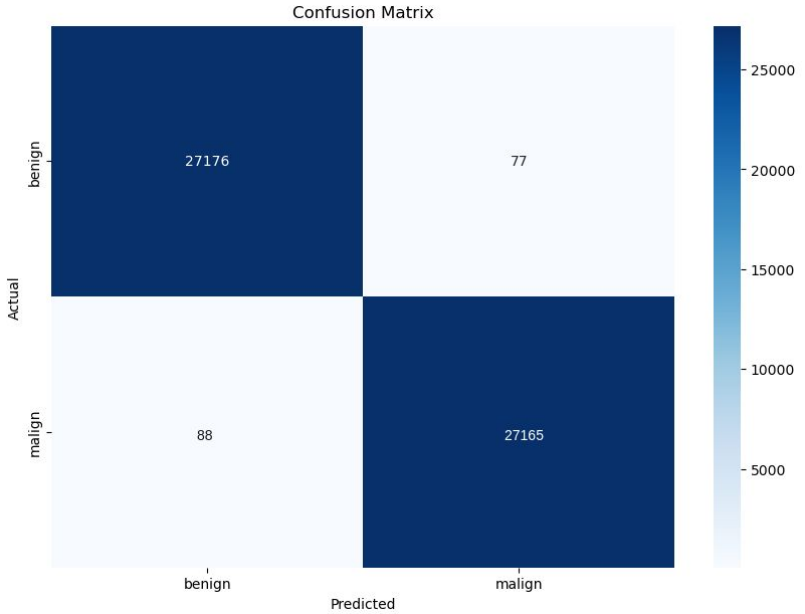


Modelos
Neural Network, with content vector

- Capaz de aprender a partir de padrões complexos
- Útil em casos com bastantes features
- 64 nós na input layer, 32 nós na hidden layer, 1 nó na output layer

Resultados Obtidos

Classifier	Accuracy	Precision	Recall	F1 score
Neural Network, with vector	0.9919	1.00	1.00	1.00

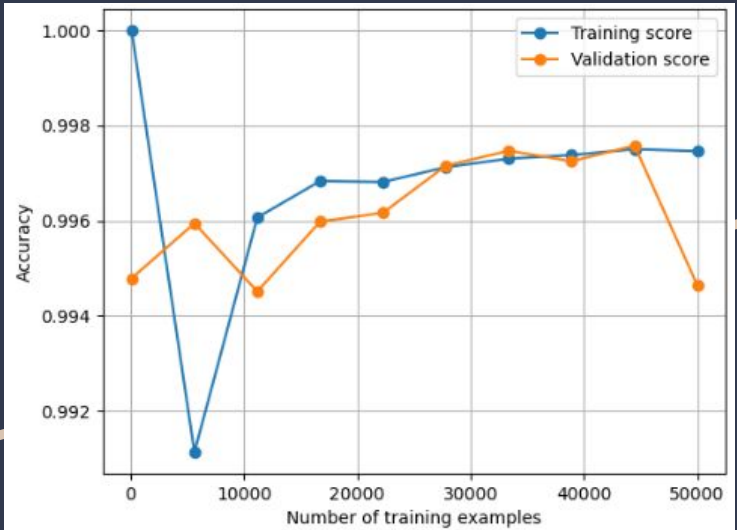


Matriz de confusão

Modelos

Neural Network, with content vector

Learning curve



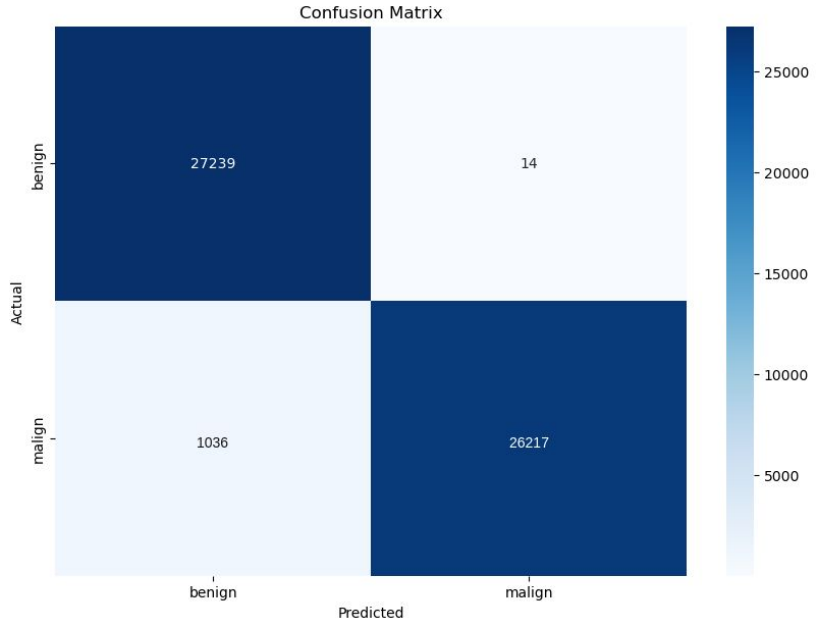
- A curva do treino inicialmente diminui e de seguida aumenta continuamente.
- Poderá indicar overfitting dos dados iniciais.
- A curva de validação vai crescendo e no final tem uma recaída pode também ser sinal de overfitting.
- Analisando outras métricas, como precisão, recall e F1 score garantimos que temos um modelo funcional.

Modelos
Neural Network, without content vector

- Precisão diminui ligeiramente.
- Importância do vetor na previsão de URLs malignos, 1036 erros de previsão dos mais graves.
- 64 nós na input layer, 32 na hidden layer, 1 na output layer

Resultados Obtidos

Classifier	Accuracy	Precision	Recall	F1 score
Neural Network, no vector	0.9830	0.98	0.98	0.98



Matriz de confusão

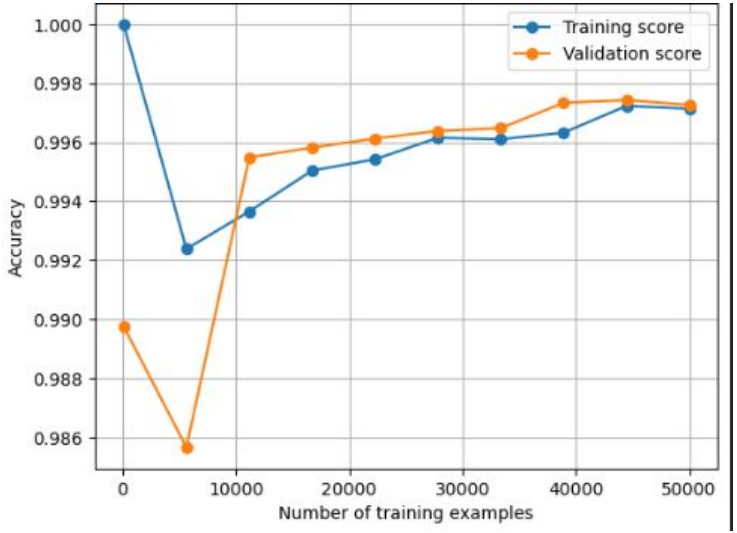
Modelos

Neural Network, without content vector

Learning curve

- Ambas as curvas diminuem inicialmente, podendo existir overfitting.

- Acaba por voltar a crescer estabilizando, sendo menos provável haver overfitting.

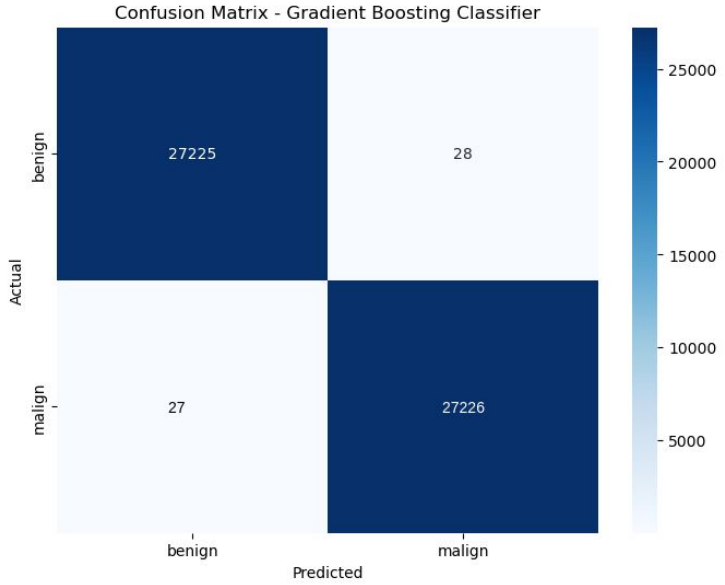


Modelos
Gradient Boosting

- Inicialmente existe um modelo pouco preciso.
- Esse modelo é treinado interativamente utilizando o dataset.
- Realizadas 100 iterações.
- Menos erros nas previsões.

Resultados Obtidos

Classifier	Accuracy	Precision	Recall	F1 score
Gradient Boosting	0.9989	1.00	1.00	1.00

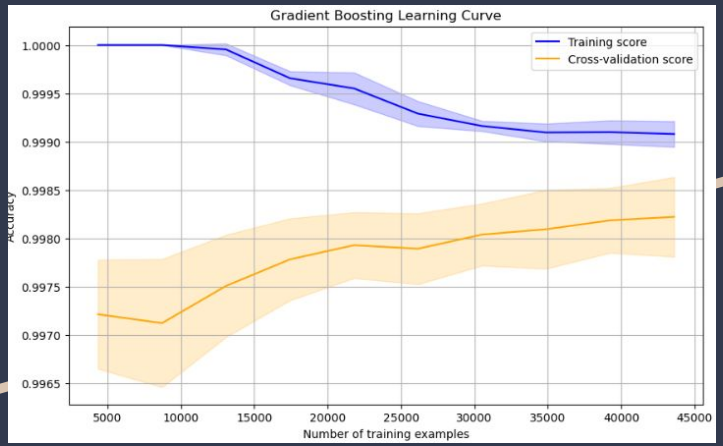


Matriz de confusão

Modelos

Gradient Boosting

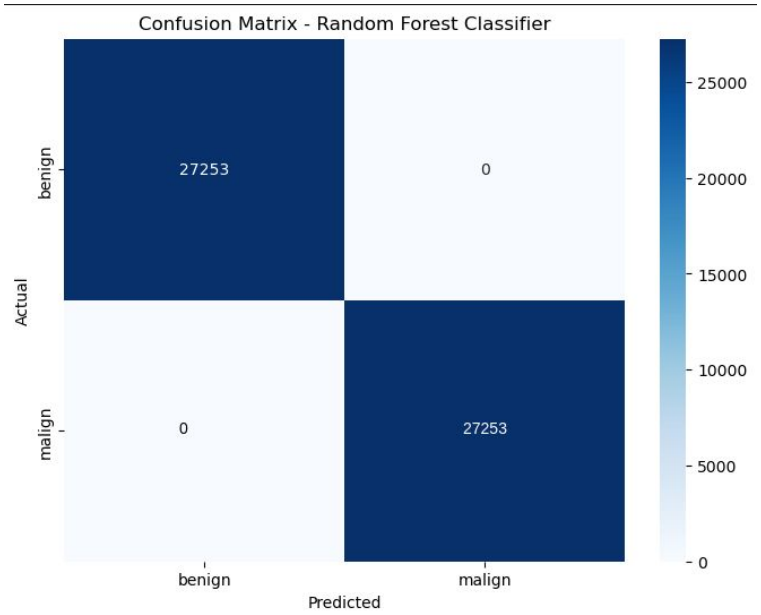
Learning curve



- A curva de treino diminui um moderadamente.
- A curva de validação aumenta também moderadamente.
- As curvas estão relativamente próximas.
- Concluímos que o modelo não sofre de overfitting.

Modelos
Random Forest

- São criadas K árvores que vão utilizar N elementos do dataset.
- Uma árvore consiste num modelo que é treinado pelos N elementos escolhidos.
- Este modelo é a junção das K árvores criadas.
- Cada árvore faz a sua previsão e a maioria dos votos é a previsão final
- Foram criadas 100 árvores



Matriz de confusão

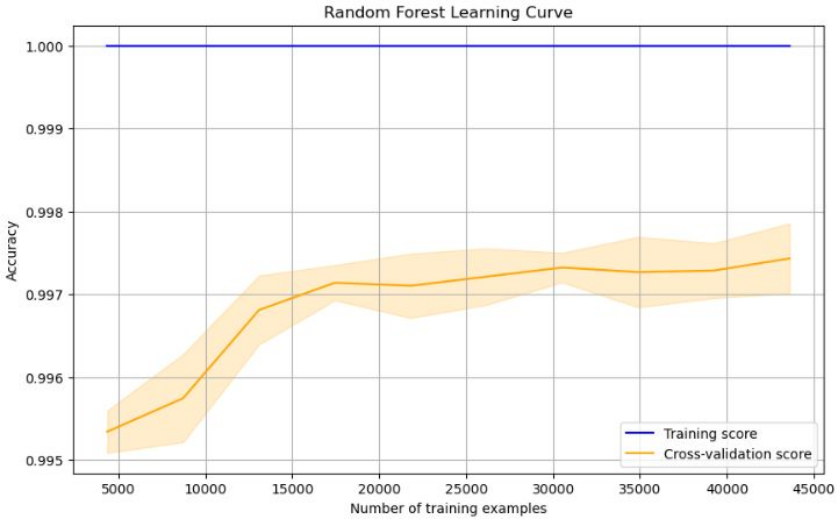
Resultados Obtidos

Classifier	Accuracy	Precision	Recall	F1 score
Random Forest	1.0	1.00	1.00	1.00

Modelos

Random Forest

Learning curve



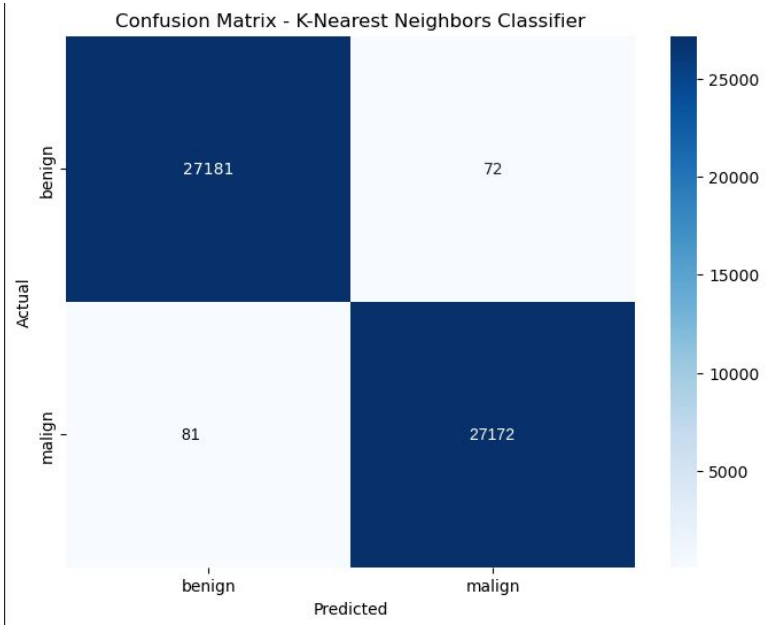
- A curva de treino mantém-se constante
- A curva de validação aumenta de forma moderada.
- As curvas estão relativamente próximas.
- Concluímos que o modelo não sofre de overfitting.

Modelos
K-Nearest Neighbors

- Na previsão de um URL, é calculada uma distância relativamente ao elementos do dataset.
- Não existe um treino como nos outros modelos.
- O treino consiste em guardar os dados, para calcular as distâncias.
- Os K elementos mais próximos ditarão a previsão e irão votar.
- O número de neighbors escolhido foi 5.

Resultados Obtidos

Classifier	Accuracy	Precision	Recall	F1 score
K-Nearest Neighbors	0.9971	1.00	1.00	1.00

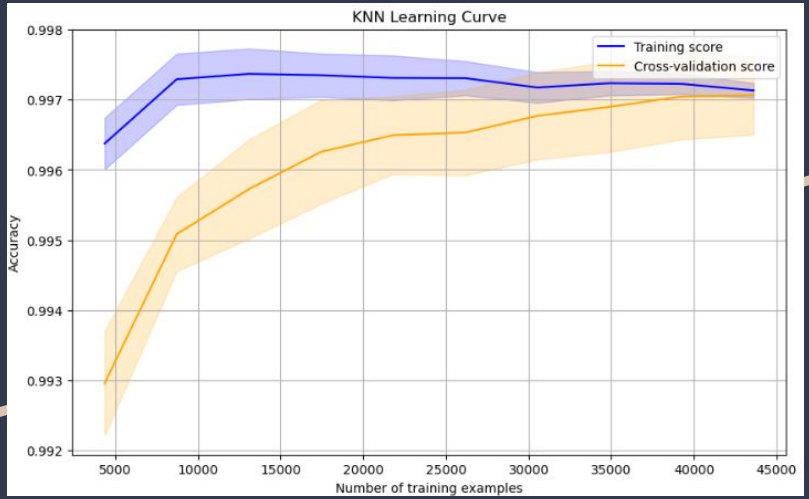


Matriz de confusão

Modelos

K-Nearest Neighbors

Learning curve



- A curva de treino mantém-se constante
- A curva de validação aumenta.
- As curvas estão próximas.
- Concluímos que o modelo não sofre de overfitting.

Conclusão

CLASSIFIER ACCURACY

Classifier	Accuracy	Precision	Recall	F1 score
Logistic Regression	0.9891	0.99	0.99	0.99
Neural Network, with vector	0.9919	1.00	1.00	1.00
Neural Network, no vector	0.9830	0.98	0.98	0.98
Gradient Boosting	0.9989	1.00	1.00	1.00
Random Forest	1.0	1.00	1.00	1.00
K-Nearest Neighbors	0.9971	1.00	1.00	1.00

Obrigado