Universidade Estadual Paulista “Júlio de Mesquita Filho”

APLICAÇÃO DAS TÉCNICAS EXPLAINABLE ARTIFICIAL INTELLIGENCE E ENSEMBLE LEARNING PARA DETECÇÃO DE ATAQUES DDOS

João Vitor Andreossi

Orientador: Prof. Dr. Kelton Augusto Pontara da Costa

01

Introdução

01 Introdução

Segundo um estudo realizado por pesquisadores da Netscout, houveram 5.4 milhões de ataques DDoS executados no primeiro semestre de 2021, um aumento de 11% em relação a 2020.

Enquanto isso, o número de brasileiros que utilizam os serviços digitais do governo federal passou de 1.7 milhões em 2019 para 113 milhões em 2021, segundo informações da Câmara dos Deputados.

01 Introdução

- O crescimento explosivo de serviços digitais nos últimos anos aumentou a dependência das pessoas a esses serviços.
- Com essa dependência, se o serviço se tornar indisponível, muitas pessoas podem ser afetadas negativamente.
- Empresas também utilizam serviços digitais, então toda uma cadeia produtiva pode sofrer grandes prejuízos.

01 Introdução

Baseando-se no modelo proposto por Idhammad et al. (2018), desenvolvemos um sistema de detecção de ataques DDoS.

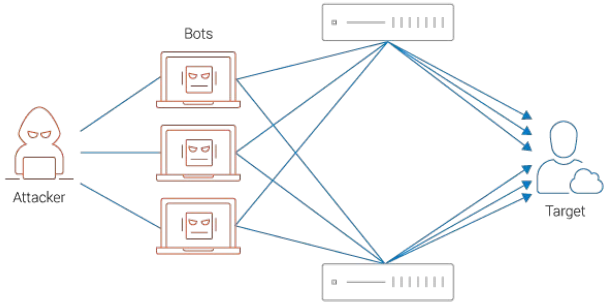
- Detecção de anomalias através do cálculo da entropia da informação.
- Classificação por modelos ensemble.
- Análise dos modelos por Explainable Artificial Intelligence.

02

Fundamentação teórica

02 Fundamentação teórica

Distributed Denial-of-Service (DDoS)



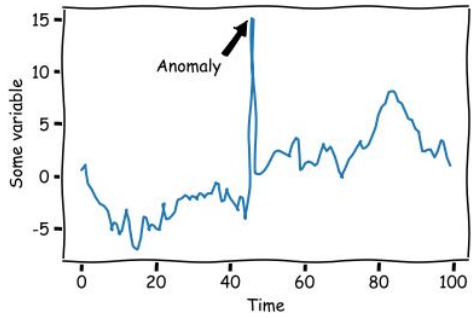
Fonte: A10 Networks

O objetivo de um ataque DDoS é inundar o servidor alvo com requisições até esgotarem todos os recursos da máquina.

- Não são necessários muitos recursos para serem executados.
- Difíceis de serem detectados.

02 Fundamentação teórica

Detecção de anomalias



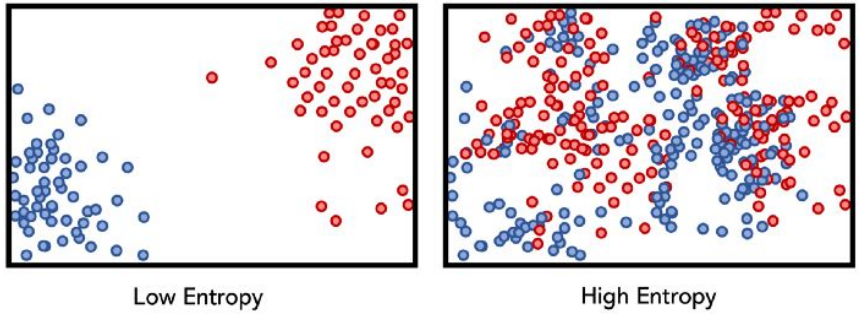
Fonte: github.com/DHI/tsod

Detecção de anomalias é o processo de encontrar padrões em um conjunto de dados que não condizem com o comportamento esperado.

- Anomalias podem ser úteis para indicar possíveis problemas
- Muito importante na área de segurança da informação.

Entropia da informação

Entropia é um conceito da termodinâmica que se refere a medida do grau de desordem ou imprevisibilidade de um sistema.



Fonte: towardsdatascience.com

02 Fundamentação teórica

Entropia da informação

Segundo Shannon (1948), para um conjunto de eventos possíveis com probabilidades p_i, existe uma medida H que representa a incerteza do resultado, tal que:

- H é contínua em p_i
- Se todos os p_i são igualmente prováveis, então $p_i = 1/n$. Sendo H uma função monótona crescente de n
- H é a soma ponderada dos valores anteriores.

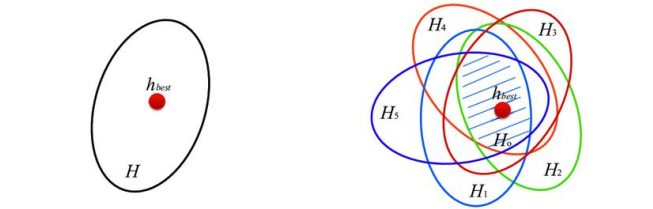
$$H = -\sum_{i=1}^{n} p_i log p_i$$ (Entropia de Shannon)

02 Fundamentação teórica

Modelos ensemble

Segundo Dietterich (2002), algoritmos de machine learning tradicionais procuram todo o espaço de funções possíveis, chamadas hipóteses, buscando por uma função h que melhor aproxima uma função desconhecida f.

Já modelos ensemble, ao invés de procurar apenas uma hipótese que melhor representa os dados, eles estabelecem um conjunto de hipóteses e "votam" na que mais se adequa ao problema.



(a) Hypothesis space of a single classifier (b) Hypothesis space of an ensemble classifier

Fonte: Yang et al. (2010)

02 Fundamentação teórica

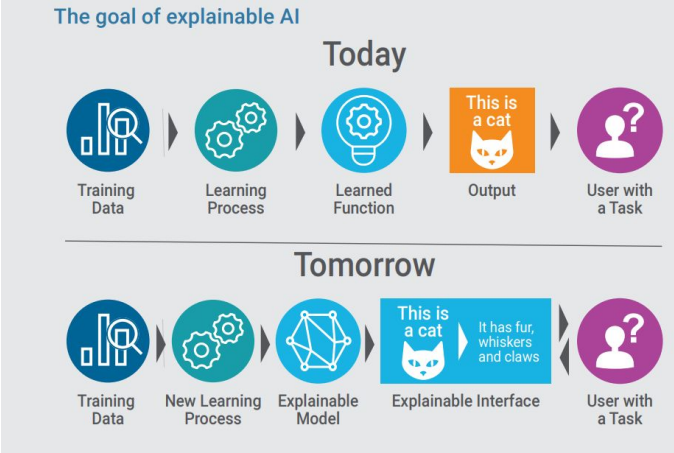
Explainable Artificial Intelligence

Modelos de machine learning tendem a agir como uma caixa preta, onde os dados são alimentados ao modelo e um resultado é retornado, mas o processo de obtenção desse resultado não costuma ser claro.

Fonte: expoundai.wordpress.com

02 Fundamentação teórica

Explainable Artificial Intelligence

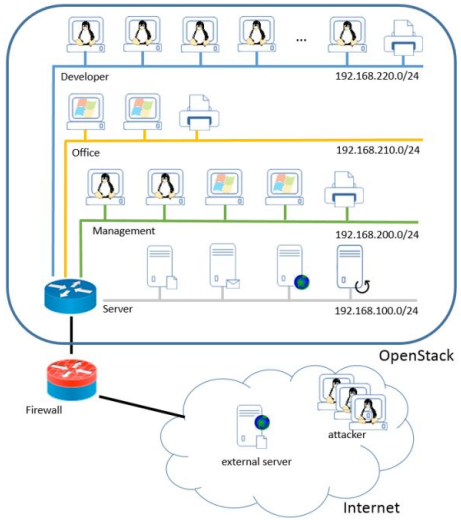


Fonte: AI and Machine Learning: Key FICO Innovations (2017)

Explainable AI se refere a métodos de análise de modelos de machine learning, que tem o objetivo de trazer mais transparência para o processo de classificação de dados.

02 Fundamentação teórica

Dataset CIDDS-001



Fonte: Ring et al. (2017)

CIDDS-001 é um conjunto de dados criado por pesquisadores da Universidade de Coburg para testes em sistemas de detecção de intrusão. Os dados foram criados simulando uma rede corporativa conectada a internet.

02 Fundamentação teórica

Dataset CIDDS-001

Nr.	Name	Description
1	Src IP	Source IP Address
2	Src Port	Source Port
3	Dest IP	Destination IP Address
4	Dest Port	Destination Port
5	Proto	Transport Protocol (e.g. ICMP, TCP, or UDP)
6	Date first seen	Start time flow first seen
7	Duration	Duration of the flow
8	Bytes	Number of transmitted bytes
9	Packets	Number of transmitted packets
10	Flags	*OR* concatenation of all TCP Flags
11	Class	Class label (normal, attacker, victim, suspicious or unknown)
12	AttackType	Type of Attack (portScan, dos, bruteForce, —)
13	AttackID	Unique attack id. All flows which belong to the same attack carry the same attack id.
14	AttackDescription	Provides additional information about the set attack parameters (e.g. the number of attempted password guesses for SSH-Brute-Force attacks)

Fonte: Ring et al. (2017)

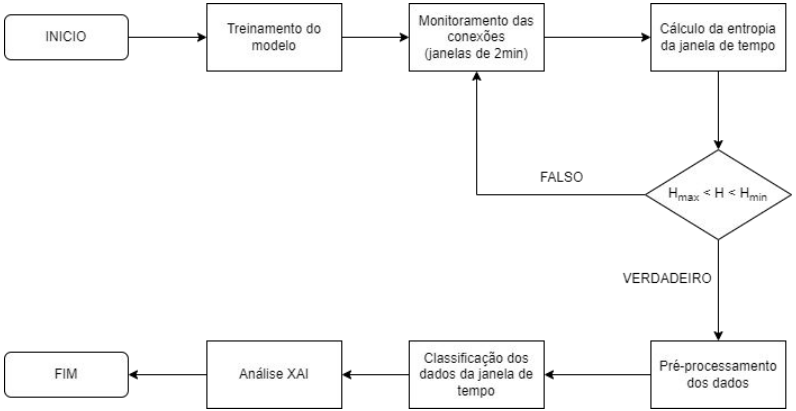
Os atributos de número 1 a 10 na Figura 2 são atributos padrões da conexão, enquanto os atributos 11 a 14 foram adicionados pelos pesquisadores como forma de categorização do dataset. O atributo *class* determina a qual categoria cada conexão pertence (*attacker*, *victim*, *normal*)

03

Metodologia

03 Metodologia

O trabalho segue o modelo proposto por Idhammad et al. (2018), que estabelece um sistema de detecção de anomalias por entropia da informação e classificação dos dados selecionados por um modelo de aprendizado de máquina, com o objetivo de determinar se as conexões analisadas podem indicar um ataque DDoS.



Fonte: Elaborado pelo autor

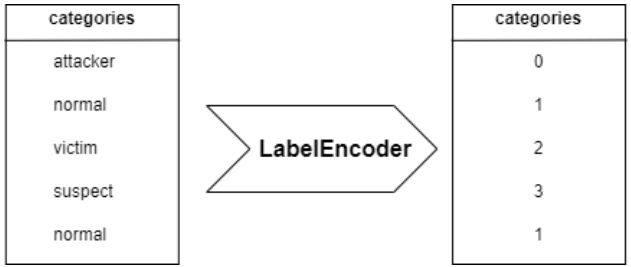
03 Metodologia

Ferramentas

03 Metodologia

Pré-processamento dos dados

Para os dados serem interpretados corretamente, precisamos realizar algumas operações para converter os dados categóricos em numéricos. Para isso utilizamos o método *LabelEncoder*.



Fonte: Elaborado pelo autor

03 Metodologia

Pré-processamento dos dados

As colunas adicionais *Flags*, *Tos*, *attackType*, *attackID* e *attackDescription* são descartadas nesta etapa, enquanto a coluna *class* é separada para servir como o atributo label nos modelos.

Como os valores podem ter grandezas muito diferentes, uma última transformação é necessária utilizando o método *MinMaxScaler*, que normaliza todos os valores entre 0 e 1.

$$x_{scaled} = \frac{x - \min(x)}{\max(x) - \min(x)}$$

Fonte: Loukas (2020)

03 Metodologia

Entropia da janela de tempo

Todas as conexões são capturadas em janelas de 2 minutos, onde ao fim do intervalo de tempo, a entropia do atributo escolhido é calculada utilizando a fórmula da entropia de Shannon. Caso o valor exceder o intervalo de entropia estabelecido, a janela de tempo passa para a etapa de classificação.

ENTRADA: distA = Distribuição do atributo A na janela de tempo

1. ***Para cada*** x ***em*** $distA$, ***faça***
2. $\quad p(x) \leftarrow quantidade(x)/tamanho(distA)$
3. $\quad h(x) \leftarrow p(x)log(p(x))$
4. $\quad H(x) \leftarrow H(x) + h(x)$
5. $\quad$ ***Fim***
6. ***Retorne*** $-H(x)$

Fonte: Elaborado pelo autor

03 Metodologia

Classificação

Quando uma janela suspeita é detectada, os dados são classificados por modelos ensemble, onde a performance do modelo é analisada durante o processo. Os classificadores escolhidos foram:

- Bagging
- AdaBoost
- Random Forest

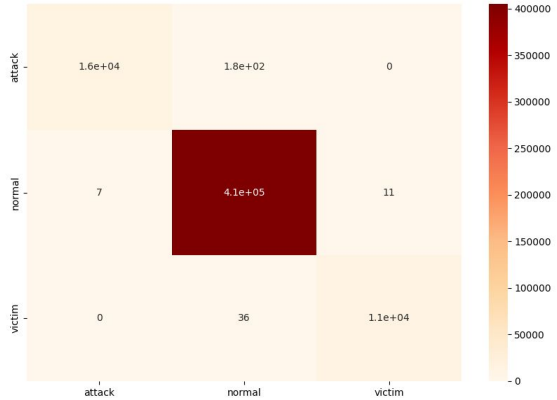
04

Resultados

04 Resultados

Treinamento do modelo

Inicialmente, o modelo foi treinado com todos os dados do 2º dia do dataset CIDDS-001, mas por conta do desbalanceamento do modelo (93.8% normal, 3.7% attacker e 2.5% victim) os resultados ficaram aquém do esperado.

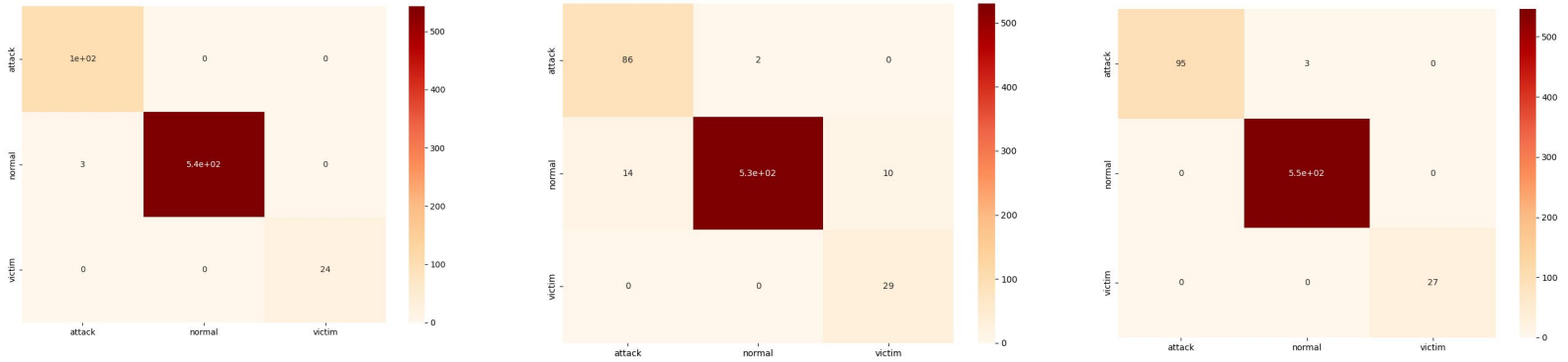


Fonte: Elaborado pelo autor

04 Resultados

Treinamento do modelo

Extraindo um subset do conjunto de dados com as categorias melhor distribuídas (82.6% normal, 13.8% attacker e 3.6% victim), observamos um grande salto na precisão do treino e das classificações. Abaixo observamos as matrizes de confusão dos modelos Bagging, AdaBoost e Random Forest, respectivamente.

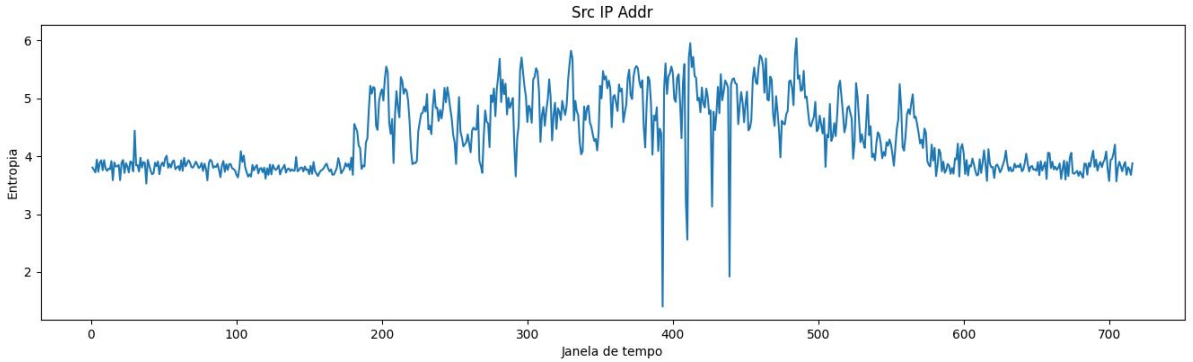


Fonte: Elaborado pelo autor

04 Resultados

Entropia

Calculando a entropia para o atributo *Src IP Addr*, é possível ver claramente o contraste entre as primeiras janelas de tempo com valores de entropia estáveis e as janelas ao longo do dia com diversos pontos de anomalia. Também observamos que os valores são consistentes com os observados por Idhammad et al. (2018).



Fonte: Elaborado pelo autor

04 Resultados

Classificação

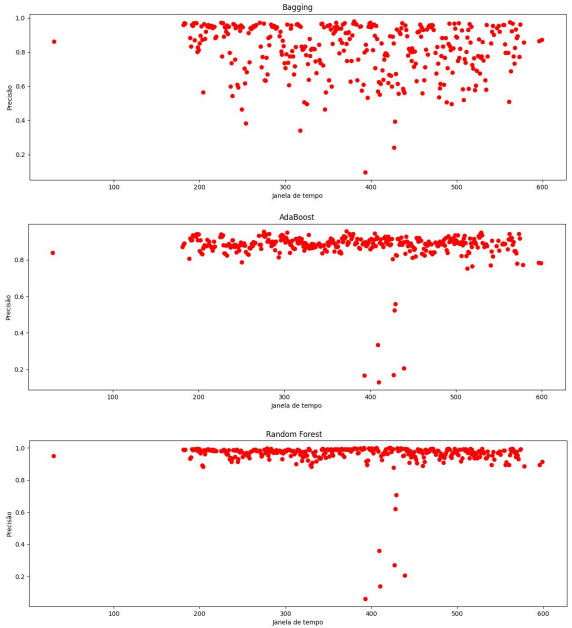
Cada janela de tempo marcada como suspeita passa pela classificação pelos modelos escolhidos. A precisão média da classificação de todas as janelas de tempo para cada modelo está na tabela abaixo.

Modelo	Precisão média (%)
Bagging	81.9
AdaBoost	87.4
Random Forest	95.7

Fonte: Elaborado pelo autor

04 Resultados

Classificação

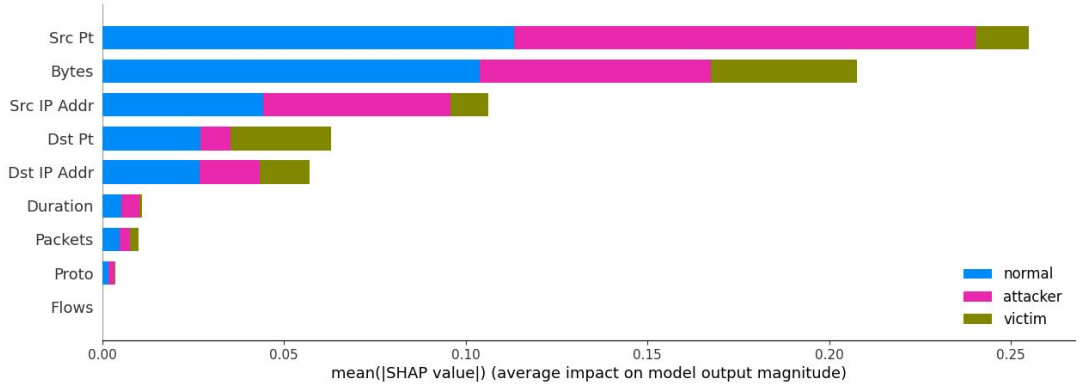


Fonte: Elaborado pelo autor

04 Resultados

Análise Explainable AI

Durante o treinamento, geramos um gráfico de impacto dos atributos com o auxílio da biblioteca SHAP.

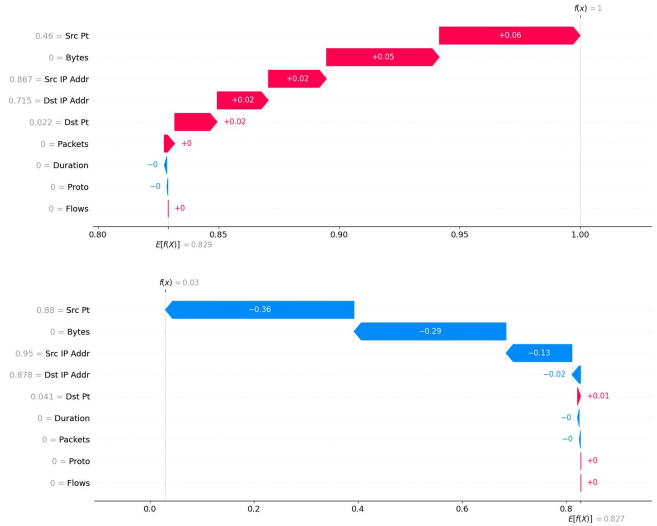


Fonte: Elaborado pelo autor

04 Resultados

Análise Explainable AI

Também é possível observar através do gráfico cascata a influência de cada atributo para a classificação das janelas de tempo.



Fonte: Elaborado pelo autor

05

Conclusão

05 Conclusão

Os resultados do projeto foram bem promissores, e a partir dos resultados podemos tirar algumas conclusões:

- Os modelos, em geral, apresentaram boas taxas de precisão.
- O balanceamento do dataset foi fundamental para um bom resultado.
- O processo de treinamento e classificação foram bem rápidos, o que permite uma resposta mais veloz ao ataque.
- A analise XAI trouxe insights importantes sobre o funcionamento interno do modelo.

Referências

IDHAMMAD, M. et al. Detection system of http ddos attacks in a cloud environment based on information theoretic entropy and random forest. Sec. and Commun. Netw., John Wiley and Sons, Inc., USA, v. 2018, jan 2018. ISSN 1939-0114. Disponível em: .

LOUKAS, S. Everything you need to know about Min-Max normalization: A Python tutorial. [*S. l.*], 28 maio 2020. Disponível em: https://towardsdatascience.com/everything-you-need-to-know-about-min-max-normalization-in-python-b79592732b79. Acesso em: 4 mar. 2022.

RING, M. et al. Flow-based benchmark data sets for intrusion detection. In: Proceedings of the 16th European Conference on Cyber Warfare and Security. ACPI. [S.l.: s.n.], 2017. p. 361–369.

YANG, Pengyi *et al*. A Review of Ensemble Methods in Bioinformatics. Current Bioinformatics, [*S. l.*], v. 5, p. 1-33, 1 dez. 2010.

NETSCOUT. NETSCOUT THREAT INTELLIGENCE REPORT. [*S. l.*], 1 set. 2021. Disponível em: https://www.netscout.com/threatreport. Acesso em: 4 mar. 2022.

AGÊNCIA CÂMARA DE NOTÍCIAS; OLIVEIRA, José Carlos. Pandemia acelera o uso de serviços públicos digitais. Câmara dos Deputados, [*S. l.*], 21 set. 2021. Disponível em: https://www.camara.leg.br/noticias/809660-pandemia-acelera-o-uso-de-servicos-publicos-digitais/. Acesso em: 4 mar. 2022.

Obrigado!