Caracterização do tráfego na Darknet utilizando árvores de decisão

Mateus Coutinho Marim[1], **Paulo Vitor Barbosa Ramos**[1]

[1]PPGCC – Universidade Federal de Juiz de Fora (UFJF)
Caixa Postal xxxx – xxxxx – Juiz de Fora – MG – Brasil

mateus.marim@ice.ufjf.br, paulo.barbosa@estudante.ufjf.br

Abstract. *The Darknet Traffic Classification Problem, related to the encrypted traffic, has been the main origin classification and application categorization models object. Through a literature revision, it is possible to discover the use of statically inference techniques and the uses of deep learning models. With a feature analysis and dataset complementation, it is possible to apply Decision Tree and Random Forest models in a way to reach accuracy beyond 95% in classes related to the origin and application classification and categorization in the encrypted traffic. Obtained results with correction related to the normalization, features selection and dataset feature expansion.*

Resumo. *O problema de Darknet Traffic Classification, envolvida no tráfego encriptado, vêm sendo objeto para os modelos de classificação da origem do tráfego e categorização da aplicação. Por meio de uma revisão da literatura é possível descobrir o uso de técnicas de inferência estatística e aplicações de modelos usando redes de aprendizado profundo. Com uma análise de atributos e complementação do dataset é possível aplicar modelos de Decision Tree e Random Forest de forma a alcançar acurácias acima de 95% na classificação e categorização para classes relacionadas a origem e aplicação usada no tráfego encriptado. Resultados obtidos por correções relacionadas à normalização, seleção de atributos e expansão daqueles já inseridos na base de dados.*

1. Introdução

A *Internet* é um incrível nicho diversificado entre diferentes camadas de segurança. Aplicações de redes sociais, plataformas de hospedagem e tantas outas ferramentas que conhecemos, é apenas uma parcela do que seria a rede mundial de computadores. Classificadas em 3 níveis, conhecemos a *Surface Web*, configurando uma porcentagem mínima do que seria a rede mundial de comunicação [Rudesill et al. 2015]. Indo mais fundo, ainda existem a *Deep Web*, acessada por pares seguros com a hospedagem continuamente alterada, e a *Darknet*, um subconjunto da *Deep Web*.

Darknet ou *Dark web* figura uma disponibilização de serviços com conexão direta entre os pares confiáveis. A hospedagem é constantemente alterada para manter a segurança para diversos fins da rede: ativismo, cyber-segurança e atividades ilícitas. Embora haja atividades que estão no lado obscuro da legalidade, a *Darknet* possibilita a percepção de uma diversificada rede no mais alto nível em técncias de segurança da informação. Devido algumas de suas características de natureza anônima, como o uso de criptomoedas, conexão encriptadas e mercados virtuais, fazem com que a *Darkweb* seja

uma fonte segura para que qualquer indivíduo possa estabelecer atividades de qualquer natureza, tanto lícitas quanto ilícitas, sem que seja rastreado por ferramentas comuns na área *forense* de segurança da informação [Mirea et al. 2019].

O problema consiste em um estudo chamado *Traffic Classification*. O objetivo é classificar o fluxo do tráfego de dados em classes relacionadas à aplicação e caracterização do tráfego. Diferentes abordagens foram usadas para a classificação do tráfego desses dados, tais como análise de portas ou inferência estatística dos pacotes enviados e recebidos [Valenti et al. 2013].

Esses são somente exemplos de abordagens clássicas, mas há outras como uso de rede neurais profundas, recentemente usada em [Gurdip Kaur 2020] aplicando *Deep Image Learning*. O intuito desses modelos é aperfeiçoar o *Quality of Service* e detectar variações que determinam a baixa qualidade do serviço prestado em rede [Parchekani et al. 2020].

O problema de *Traffic Classification* se mostra uma tendência para as diversas abordagens de aprendizado de máquinas, principalmente quando os dados são encrpitados. Um dos principais problemas visto nos documentos de referência é a falta da abordagem da fundamentação do que seria a base empregada, ou seja, a inferência estatística antes da criação do modelo, de forma a tratar e expandir a base.

Desta forma, o presente trabalho aborda o problema de classificação de tráfego pelo *Darknet dataset* que contém registros de tráfego real à partir da *Internet* comum, que chamam de tráfego benigno, e da *Darknet*. É feita uma análise dos atributos existentes para a criação de novos que acrescentem mais informação para o *dataset* e que, posteriormente, alimentem modelos de classificação baseados em árvore de decisão para classificação da origem do tráfego, *Darknet* ou benigno, e também na caracterização do tráfego advindo da *Darknet* pelo seu serviço. Ao final, foi realizado uma seleção de atributos para determinar se o subconjunto de atributos selecionado contém os adicionados e os definidos por [Gurdip Kaur 2020] como importantes na caracterização do tráfego.

O trabalho está dividido da seguinte forma: a Seção 2 traz alguns trabalhos relacionados ao tema, fundamentando a base estudada e o tema relacionado. A Seção 3, além de descrever as tratativas feitas, descreve seus principais atributos e as abordagens que podem ser realizados com seus dados. A Seção 4 demonstrará os resultados obtidos da inferência estatística e interpretações gráficas. A Seção **??** decreverá as seleções de características, demonstrando os atributos mais importantes para o modelos de classificação. A Seção 5 traz as considerações finais sobre as interpretações do trabalho base, da base e da análise, disponibilizando sugestões para trabalhos futuros a serem feitos usando a base estudada.

2. Trabalhos Relacionados

Antes de demonstrar diferentes soluções para o problema de *Traffic Classification*, é necessário entender a definição do problema relacionado, na qual consiste em usar os dados de tráfego entre remetente e destinatário para classificar e categorizar a aplicação usada para o tráfego gerado. Um dos principais desafios é realizar essa tarefa usando dados encriptados, abordagem feita em duas bases disponibilizadas pela *University of New Brunswick*, a *ISCXVPN2016* [Draper-Gil et al. 2016] e *ISCXTor2016*

[Lashkari et al. 2017], que, respectivamente, fornecem o tráfego em redes usando VPN e Tor.

Recentemente, num trabalho publicado por [Gurdip Kaur 2020], houve a disponibilização de uma base de dados que é a união dessas outras duas, a chamada *CIC-Darknet2020*. Tal trabalho realiza a classificação das aplicações provenientes da *Internet* provinda da *Darknet* , tendo o tráfego no uso da VPN e do Tor. Além de sua contribuição com a publicação da base de dados, o autor demonstra uma acurácia de 92% para identificação da origem do tráfego e 86% para a categorização do tráfego na apresentação dos resultados de seu modelo na classificação da base usando uma técnica chamada de *Deep Image Learning*.

[Draper-Gil et al. 2016] aborda a classificação da comunicação via VPN, usando redes neurais e a base de dados *ISCXVPN2016* para a classificação do tráfego em dois estágios. O primeiro, usando *Multi-Layer Perceptron* como função de ativação para o segundo estágio, uma *Recorrent Neural Network* para identificar as 6 classes empregadas pelo *dataset*. Com esse modelo de classificação, os resultados alcançados chegam a uma acurácia de 75%.

[Lotfollahi et al. 2017] usando a mesma base, demonstra o desenvolvimento do *framework Deep Packet* para a solução do problema. O *framework* compreende em dois métodos de rede de aprendizado profundo, uma rede neural convolucional e um autoencoder, ambos para a tarefa de classificação e caracterização. Esse trabalho ainda ressalta a existência de outros três métodos:

- A classificação por meio das portas, na qual não é tão usada pelo fato de o método não classificar mais do que 70% do tráfego da *Internet* atual. A porcentagem de classificação diminui quando há o uso de dados encriptados ou uso de regras *proxies* [Moore and Papagiannaki 2005].
- Classificação por meio da inspeção do *payload*, disponível na camada de aplicação dos pacotes. Essa análise, chamada de *Deep Packet Inspection* (DPI) usa padrões pré-definidos, tendo suas derivações para distinguir os protocolos usados para a transmissão do pacote.
- Por fim, a classificação usando uma abordagem estatística. Função de densidade de probabilidade (PDF) utilizada na análise de pacotes entre hora de chegada e limites normalizados possibilitou uma acurácia de 91% para a classificação de protocolos HTTP, POP3 e SMTP no trabalho de [Crotti et al. 2007]. Em outro trabalho, considerando o tamanho do pacote, também usando PDF, houve uma acurácia de pelo menos 87% para identificar protocolos FTP, IMAP, SSH e TELNET [Wang and Parish 2010].

[Lotfollahi et al. 2017] e [Draper-Gil et al. 2016], embora tenham contribuído para a categorização da aplicação pelo tráfego, não focaram nos dados provenientes da *Darknet*. [Gurdip Kaur 2020] aborda a categorização em duas camadas, a primeira relacionada a classificação da origem e a segunda em relação ao tráfego proveniente da *Darknet*, verificando os atributos do *dataset* mais importantes para a classificação. Mesmo com uma acurácia de 86% para o problema, os autores não realizam qualquer tipo de comparação com modelos de classificação mais simples.

No trabalho atual utilizamos os modelos *Decision Tree* e *Random Forest* para realizar a classificação da origem do tráfego e a categorização da aplicação dos dados

provenientes da Darknet. Ampliamos a base com novas características para os registros, inserindo informação dos endereços IP de origem e destino, realizando a divisão *n-grams* e manipulações nos registros originais.

3. Dataset

Disponibilizado por [Gurdip Kaur 2020], a base de dados em apreço, *CICDarknet2020*, é uma junção de dois trabalhos que publicaram *ISCXTor2016* [Lashkari et al. 2017] e *ISCXVPN2016* [Draper-Gil et al. 2016], que analisam o tráfego da rede Tor e VPN, respectivamente. O dataset compreende dois tipos de tarefa de classificação, a primeira classifica o tráfego vinda da internet benigna, que não utilizam o Tor ou VPN, e da *Darknet*, a segunda classificação diz respeito à aplicação a qual o tráfego corresponde, dividindo nas segunites classes: *browsing*, *email*, *chat*, *audio-streaming*, *video-streaming*, transferência de arquivos, VOIP e P2P. A Figura 1 demonstra a divisão da origem dos dados do *dataset*.

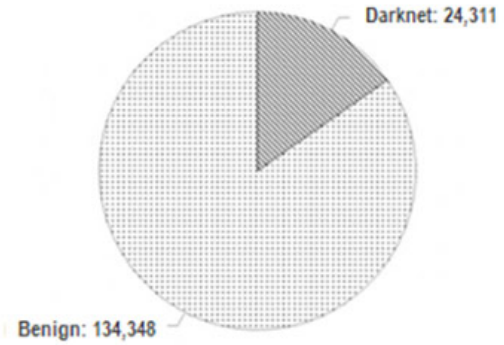


Figura 1. Distribuição pela origem dos dados [Gurdip Kaur 2020]

O *CICDarknet2020* possui uma variedade de campos para análise, trazendo informações sobre IP, porta, classificador, duração do tráfego dos pacotes entre outros. São no total 158.659 registros, dos quais 24.311 provenientes da *Darknet* e 134.348 da rede benigna.

Alguns dados não são úteis e outros precisaram ser complementados. A base fornece endereços de IP de origem e destino, mas não disponibiliza qualquer informação sobre esses endereços, tais como localização ou organização na qual está registrado. Para suprir essa lacuna, foi implementado um *script* para usar os dados da base como parâmetros para a biblioteca *IpInfo* do *Python*.

Feito essa alteração, analisando a distribuição dos dados contidos na base, é possível verificar a união das outras duas bases por meio da verificação de que os dados estão classificados pela origem benigna (não VPN e não Tor) e da *Darknet* (VPN e Tor). A Figura 2 mostra que 80% dos dados da base são provenientes da internet benigna.

Como a base possui as classes de aplicação, é importante verificar a distribuição de acordo com a origem. A Figura 3 possbilita a percepção de que aplicações voltadas para *Streaming* de áudio e *Chat* são mais comuns para a os dados provenientes da *Darknet*.

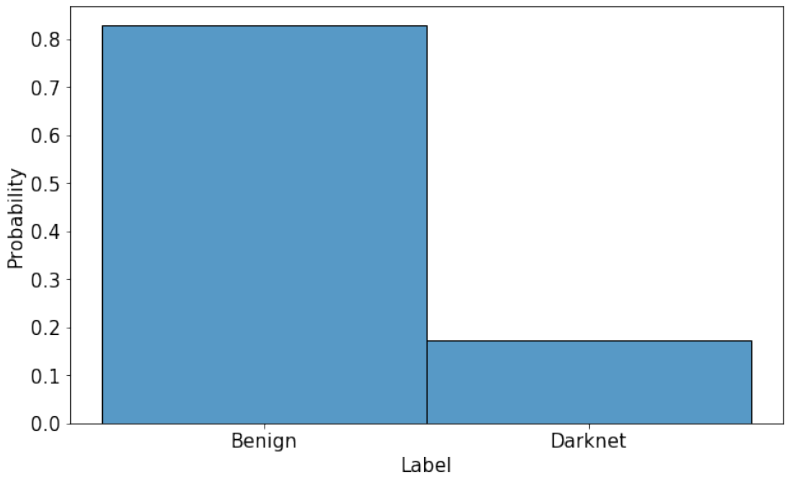


Figura 2. Probabilidade de ocorrência de um tipo de tráfego

Por outro lado, quando verificado os dados provenientes da rede benigna, figuram os as classes menos comuns. Aplicações voltadas para navegação e P2P são as mais comuns no tipo benigno.

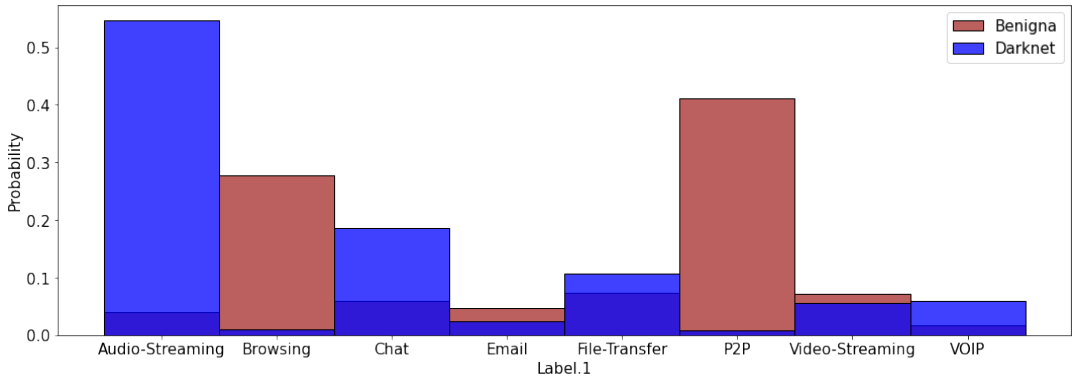


Figura 3. Probabilidade de ocorrência do tráfego de um tipo de serviço

3.1. Pré-processamento do *dataset*

A primeira correção consiste em normalização das duas classes disponibilizadas pela base de dados, a de origem dos dados e do tipo de aplicação do tráfego. Através da inspeção das classes do *dataset* é possível verificar falta de padronização na nomeclatura das classes, assim como a sua redundância. Para corrigir esse problema, realizamos a padronização dos nomes, retirando as redundâncias.

A base nos fornece IP's de origem e destino, o que possibilita extrair mais atributos relacionados aos endereços de rede. Uma das possbilidades é o uso de *One Hot Encoder*, mas como visto em [Baiardi et al. 2017] o uso de *n-grams* pode ajudar na diminuição dos erros do modelo de classificação gerado, consequentemente, reduzindo a percentagem de previsões falsas positivas. Dessa forma, a base foi expandida usando os IP's fragmentando-os em *n-grams* (Unigram, Bigram e Trigram), além de trazer as informações de hospedagem, geolocalização, *bogons* (endereços falsos), entre outros.

Com essas divisões em *n-grams* dos endereços de origem e destino, usamos a técnica de *Hashing Encoding* para criação de 100 novos atributos nomeados de col_i onde i representa o id do novo atributo pois, como visto nos resultados de [Weinberger et al. 2009], possibilita a compressão dos atributos em relação ao, por exemplo, *One Hot Encoding* na qual poderia gerar milhares de novos atributos dependendo do número de categorias únicas de um atributo sendo processado, sendo bastante útil para registros grandes e que são aplicados em modelos de aprendizado de máquinas, mas tendo a desvantagem de que os novos atributos criados perdem no quesito de interpretabilidade. Além disso, como a pesquisa dos endereços de rede trouxe o país de origem do IP, criamos um novo atributo com essa informação e convertemos os países em números ordinais.

Uma outra característica que foi extraída foi a hora em que ocorreu a captura dos dados pelo campo *TimeStamp* do *dataset*. A Figura 4 demonstra a relação entre as horas de captura para as duas classes de origem do tráfego. É possível verificar dois diferentes padrões, um para cada classe, sendo possível observar que pode ser importante para a classificação do tráfego, consequentemente, na sua utilização no modelo.

Além de ser possível a distinção dos horários, essa relação de tempo permite dizer os horários em que há uma maior probabilidade de utilização. Para a rede benigna, com uma volume alto do tráfego contido entre as 7 horas e 12 horas, tendo alguns picos durante a madrugada. Para a rede *Darknet*, a distribuição é mais esparsa, não possuindo picos de utilização além da normal, contidas entre às 24 horas e 7 horas. Essa relação demonstra uma disjunção exclusiva, ou seja, há uma probabilidade considerável de não haver tráfego da *Darknet* e benigna concomitantemente.

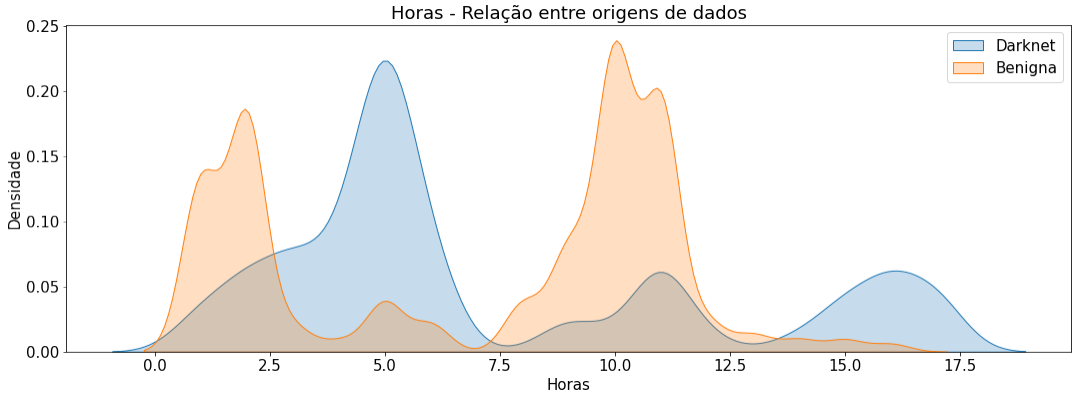


Figura 4. Densidade do tráfego em uma dada hora

Uma das transformações mais importantes a serem aplicadas nos dados é o escalonamento dos atributos. Geralmente, poucos algoritmos de aprendizagem de máquina tem um bom desempenho com atributos de escalas muito diferentes [Géron 2019]. Realizamos o escalonamento dos atributos que são representados por números reais através da padronização dos seus intervalos.

4. Experimentos

Nesse trabalho abordamos as duas tarefas de classificação propostas pelo *dataset*, a da predição da origem do tráfego da rede, com classificações entre *Benign* e *Darknet* e da caracterização dos serviços do tráfego na Darknet. Os modelos escolhidos são baseados

em árvores e foram selecionados devido a sua simplicidade e facilidade na interpretação, além disso, em conjunto com uma seleção de características é possível estimar a importância dos atributos de acordo com a sua influência na classificação. Abaixo são brevemente descritos os modelos selecionados [Géron 2019].

- *Decision Trees* (DTs): modelos de aprendizado supervisionado não paramétrico que podem ser usados para classificação e regressão. Funciona pelo aprendizado de regras de decisão simples inferidas dos dados para a predição da variável alvo. São modelos simples de entender e interpretar e as árvores geradas podem ser visualizadas. A implementação do *sklearn* é uma CART otimizada.
- *Random Forest* (RFs): é um modelo *ensemble* que usa DTs como classificadores fracos com o objetivo de gerar um classificador forte, a RF treina cada uma das DTs com a técnica de *Bagging* com o objetivo de gerar um cassificador co uma performance melhor que a de seus componentes individuais.

4.1. Metodologia

Cada um dos modelos foi treinado, mantendo os parâmetros padrões definidos pelo *sklearn*, com o *Darknet dataset* sendo separado em dois conjuntos de treino e de teste, o conjunto de teste tem 33% da quantidade de amostras, onde os rótulos de um correspondem à origem do tráfego e outro em que representam o serviço que está gerando o tráfego na *Darknet*.

Em ambas tarefas a performance do modelo foi estimada com uma validação cruzada estratificada de 10-fold, também é feita uma última validação do modelo com o conjunto de teste que foi mantido separado para mensurar a acurácia de teste dos modelos. Por final, também são consideradas para análise da qualidade dos modelos as métricas:

- Precisão: acurácia das predições positivas ou corretas.

$$precisao = \frac{TP}{TP + FP}$$

- *Recall*: a precisão é geralmente encontrada com a medida *recall* que representa taxa de positivos verdadeiros.

$$recall = \frac{TP}{TP + FN}$$

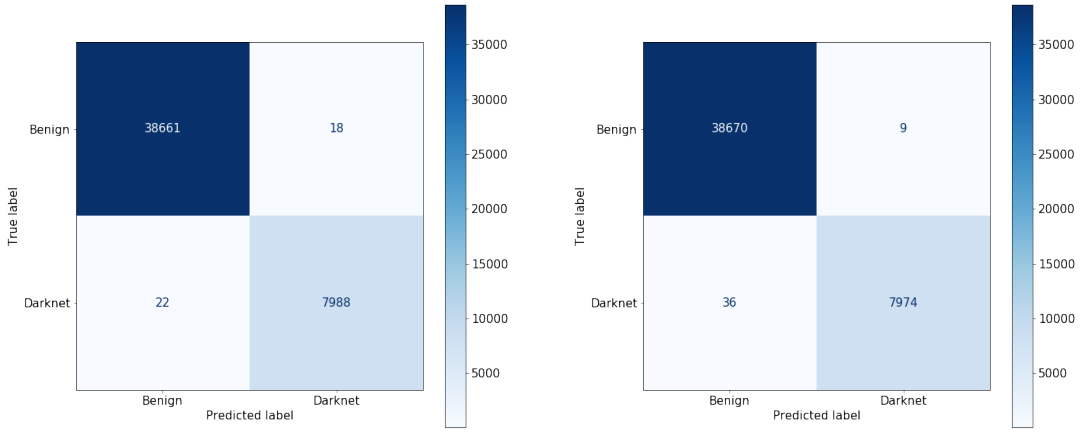
- *F-score*: o *F-score* combina a precisão e o *recall* com a média harmônica dos dois, enquanto a média comum trata todos valores igualmente a média harmônica coloca mais peso em valores mais baixos. O *F-score* assume valores próximos de 1 quando ambas a precisão e o *recall* estão altos.

$$F\text{-}score = \frac{TP}{TP + \frac{FN+TP}{2}}$$

Todos experimentos foram feitos com a biblioteca *sklearn* [Pedregosa et al. 2011] do Python em um computador com processador *Intel Core i5-7200U* com 4 núcleos de 2.5GHz, 20GB de RAM e sistema operacional *Ubuntu* 20.04. Além disso, foi usada a semente 42 nos algoritmos com alguma aleatoriedade para permitir a reprodutibilidade dos resultados.

4.2. Detecção do tráfego da *Darknet*

Nas matrizes de confusão das Figuras 5a e 5b, correspondentes aos modelos de *decision tree* e *random forest*, é possível observar que mesmo com um dos modelos tendo esse padrão atênuado, existe uma tendência, provavelmente causada pelo baixo número de exemplos de tráfego da darknet, de que uma classificação errônea provavelmente seja de um tráfego provindo da *Darknet* sendo classificado como tráfego benigno, apesar disso, a ocorrência desse erro é muito pequena em comparação com o desempenho geral do modelo treinado.



(a) Decision tree **(b) Random forest**

Figura 5. Matrizes de confusão da detecção de tráfego da *Darknet*

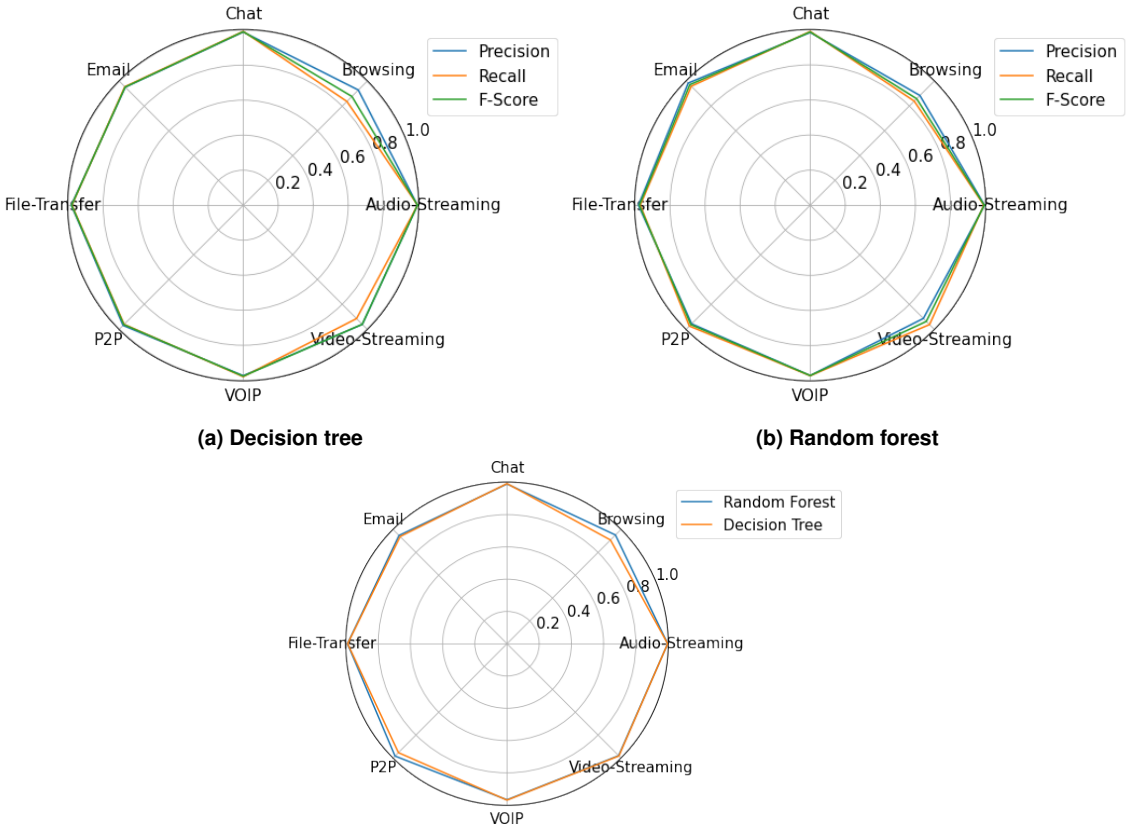
A Tabela 1 sumariza os valores das métricas de cada modelo na classificação entre os rótulos *Benign* e *Darknet*, além disso, o modelo conseguiu uma acurácia de 99.89% no 10-fold e uma boa capacidade de generalização dado que teve um resultado similar na acurácia de teste de cada rótulo.

		Precisão	Recall	F-score	Acc. de teste
Decision tree	**Benign**	0.9994	0.9992	0.9993	99.94%
	Darknet	0.9964	0.9971	0.9967	99.78%
Random forest	**Benign**	0.9989	0.9997	0.9993	99.91%
	Darknet	0.9987	0.9947	0.9967	99.89%

Tabela 1. Sumário das métricas de avaliação dos modelos

4.3. Caracterização do tráfego da *Darknet*

A Figura 6a e 6b relaciona em coordenadas polares os valores das métricas de precisão, *recall* e *F-score* para os modelos de *decision tree* e *random forest*, respectivamente. Também é fácil de verificar que houve pouca variação dos valores das métricas entre os dois modelos, apesar disso, podemos ver na Figura 6c que a *random forest* conseguiu melhorar a acurácia nos rótulos *Browsing* e *P2P* em relação a *decision tree*. Em comparação com o trabalho de [Gurdip Kaur 2020], todas as métricas obtiveram valores bem mais altos, até mesmo nas classes que haviam muitos erros, podendo indicar que esse problema é resolvido sem a necessidade de um modelo complexo.



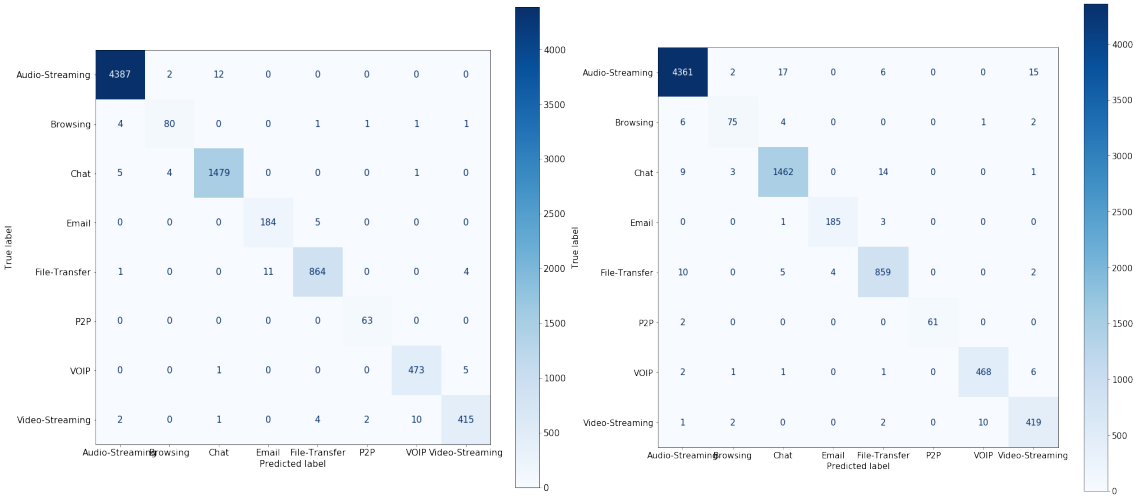
(a) Decision tree

(b) Random forest

(c) Acurácia de teste dos modelos

Figura 6. Resultados das métricas para detecção de tráfego da *Darknet*

Nas matrizes de confusão das Figuras 7a e 7b, que correspondem a *decision tree* e *random forest* fica evidente que os erros mais cometidos são entre os tráfegos com rótulos de *Chat* e *Audio-Streaming* e vice-versa, podendo indicar que existe alguma similaridade nos rótulos que pode causar confusão no modelo.



(a) Decision tree

(b) Random forest

Figura 7. Matrizes de confusão da detecção de tráfego da *Darknet*

5. Conclusão

Neste trabalho abordamos os problemas da detecção e caracterização do tráfego proveniente da *Darknet* através da utilização de modelos de aprendizagem baseados em árvores de decisão, sendo eles a *decision tree* e a *random forest*, que se mostraram capazes de classificar novos registros de tráfego com uma acurácia maior que 98% para cada uma das tarefas de classificação. Também foram extraidas novos atributos do *dataset* original pela busca de informações dos IPs de origem e destino do tráfego e pela codificação dos mesmos com o *hashing encoding*, outro atributo gerado foi o horário em que o tráfego ocorreu pelo *timestamp* incluído no *dataset*, que pelas nossas análises iniciais mostraram potencial para contribuir na eficiência dos modelos treinados por mostrarem que os tráfegos da internet comum e da *Darknet* costumam ocorrer em horários distintos.

Fica evidente que algoritmos de aprendizagem de máquina simples como os baseados em árvore decisória são bons candidatos para conseguir resultados competitivos para problemas do mundo real e que podem ter a sua eficiência melhorada com um processamento mais cuidadoso dos atributos já existentes, apesar disso, uma desvantagem dos modelos utilizados é que eles não podem ser treinados de forma *online*, ou seja, não é capaz de aprender com um novo exemplo a não ser que o modelo seja treinado novamente com todos os dados, uma proposta de trabalho futuro é uma pesquisa com modelos de aprendizagem *online*.

Referências

Baiardi, F., Lipilini, J., and Tonelli, F. (2017). Using s-rules to fire dynamic countermeasures. In *2017 25th Euromicro International Conference on Parallel, Distributed and Network-based Processing (PDP)*, pages 371–375. IEEE.

Crotti, M., Dusi, M., Gringoli, F., and Salgarelli, L. (2007). Traffic classification through simple statistical fingerprinting. *ACM SIGCOMM Computer Communication Review*, 37(1):5–16.

Draper-Gil, G., Lashkari, A. H., Mamun, M. S. I., and Ghorbani, A. A. (2016). Characterization of encrypted and vpn traffic using time-related. In *Proceedings of the 2nd international conference on information systems security and privacy (ICISSP)*, pages 407–414.

Géron, A. (2019). *Hands-on machine learning with Scikit-Learn, Keras, and TensorFlow: Concepts, tools, and techniques to build intelligent systems*. O'Reilly Media.

Gurdip Kaur, Arash Habibi Lashkari, A. R. (2020). Didarknet: A contemporary approach to detect and characterize the darknet traffic using deep image learning.

Lashkari, A. H., Draper-Gil, G., Mamun, M. S. I., and Ghorbani, A. A. (2017). Characterization of tor traffic using time based features. In *ICISSp*, pages 253–262.

Lotfollahi, M., Zade, R. S. H., Siavoshani, M. J., and Saberian, M. (2017). Deep packet: A novel approach for encrypted traffic classification using deep learning. *CoRR*, abs/1709.02656.

Mirea, M., Wang, V., and Jung, J. (2019). The not so dark side of the darknet: a qualitative study. *Security Journal*, 32(2):102–118.

Moore, A. W. and Papagiannaki, K. (2005). Toward the accurate identification of network applications. In Dovrolis, C., editor, *Passive and Active Network Measurement*, pages 41–54, Berlin, Heidelberg. Springer Berlin Heidelberg.

Parchekani, A., Naghadeh, S. N., and Shah-Mansouri, V. (2020). Classification of traffic using neural networks by rejecting: a novel approach in classifying vpn traffic. *arXiv preprint arXiv:2001.03665*.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., and Duchesnay, E. (2011). Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830.

Rudesill, D. S., Caverlee, J., and Sui, D. (2015). The deep web and the darknet: A look inside the internet's massive black box. *Woodrow Wilson International Center for Scholars, STIP*, 3.

Valenti, S., Rossi, D., Dainotti, A., Pescapè, A., Finamore, A., and Mellia, M. (2013). *Reviewing Traffic Classification*, pages 123–147. Springer Berlin Heidelberg, Berlin, Heidelberg.

Wang, X. and Parish, D. (2010). Optimised multi-stage tcp traffic classifier based on packet size distributions. *2010 Third International Conference on Communication Theory, Reliability, and Quality of Service*, 0:98–103.

Weinberger, K., Dasgupta, A., Langford, J., Smola, A., and Attenberg, J. (2009). Feature hashing for large scale multitask learning. In *Proceedings of the 26th annual international conference on machine learning*, pages 1113–1120.