Ciência de Dados para Segurança

Maria Teresa Kravetz Andrioli[1]

[1]Universidade Federal do Paraná (UFPR)
Curitiba – PR – Brazil

mtka17@inf.ufpr.br

1. Dataset

O dataset escolhido para este trabalho foi o Open Source Cluster IOTs for RE Malwares. Ele contém análises manuais e automáticas de malwares num período de aproximadamente 2 anos (Outubro 2014 a Dezembro 2016) com 143332 registros de malwares. As entradas numéricas captadas passaram por uma transformação PCA devido a problemas de confidencialidade. Além disso, utiliza o MISP (Malware Information Sharing Platform) para rotular os dados. O dataset em si consiste num documento CSV organizado da seguinte maneira:

Nome	**Descrição**	**Tipo de Dado**
uuid	Identificador único	String
event_id	ID sequencial do evento do MISP	Integer
category	Categoria segundo o MISP	String
type	Tipo segundo o MISP	String
value	Valor segundo o MISP	String
to_ids	ID designado para o IDS (Sistema de detecção de intrusão)	Integer
date	Data e hora da análise	Integer

2. Exploração de dados

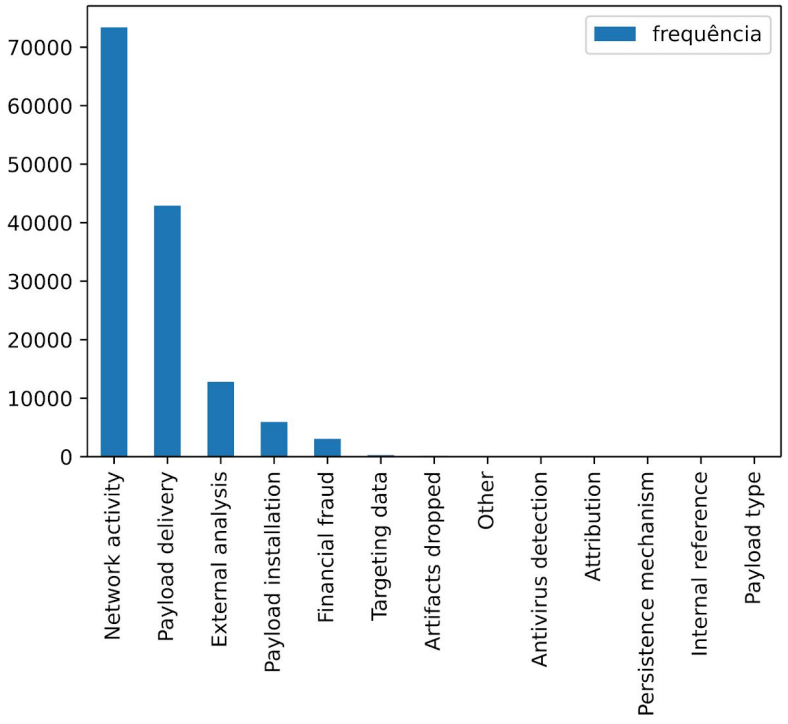
No arquivo exploracao.ipynb está a fase de exploração de dados do dataset. Após a análise inicial, foi escolhidos manter as colunas “category”, “type” e “value” e retirar “uuid”, “event_id”, “to_ids” e “date” de modo a cumprir o objetivo de usar esse dataset para analisar quais malwares são mais comuns de acordo com sua categoria, assim possibilitando, em uma solução hipotética para além dessa disciplina, estudar esses malwares e combater esses ataques.

Nesse arquivo, primeiramente é feita a leitura do dataset, observando todas as colunas e campos, tipo de cada dado e imaginando como esses dados podem ser usados para uma análise crítica.

Depois dessas observações, os campos escolhidos para serem removidos são retirados, criando-se um novo dataframe com apenas as categorias, tipos de valores. Nesse novo dataset, é possível escolher então, como serão rotulados os dados. Devido a escolha de usar as categorias para analisar, é feito uma extração de todas as categorias que existem, sendo elas:

Network activity, Payload delivery, External analysis, Payload installation, Financial fraud, Targeting data, Artifacts dropped, Other, Antivirus detection, Attribution, Persistence mechanism, Internal reference, Payload type

Depois de identificar as diferentes categorias, é feita a análise de frequência delas, com o seguinte gráfico demonstrando o resultado:



Total de amostras: 138724

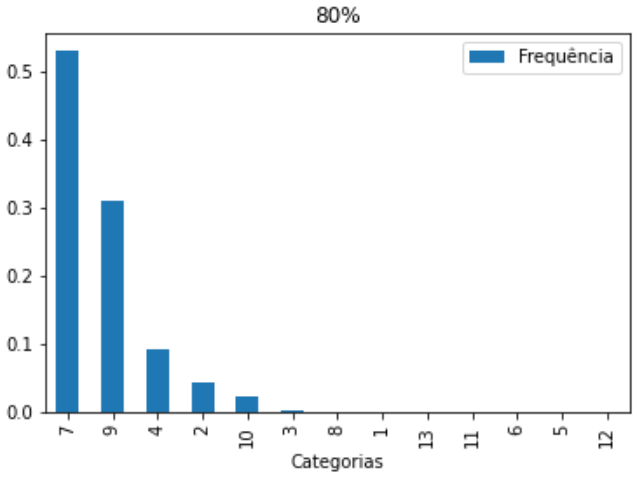
Por curiosidade, também foi feita a análise quantitativas usando dos tipos, porém não acredito que eles sejam um dado interessante para serem usados sozinhos, sem ser em relação à sua categoria, devido à variedade muito grande.

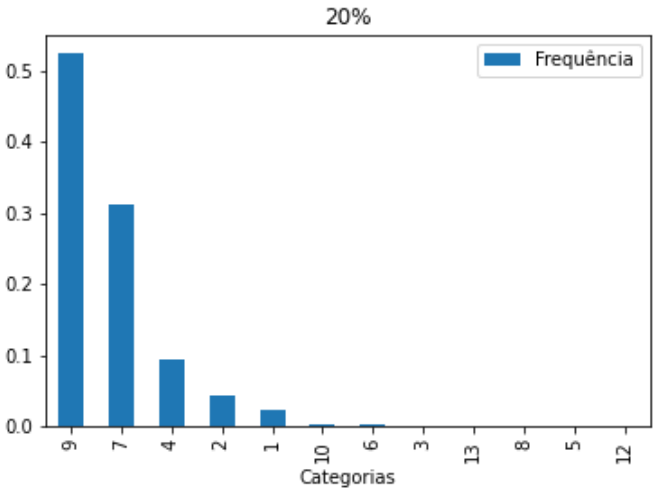
3. Project Final

Para o projeto final, foi utilizado o notebook final.ipynb. Primeiro foi criado um dataset com apenas os atributos escolhidos. Como esses eram textuais, foi feito um mapeamento de cada valor de string para um valor de inteiro e assim, gerado um novo dataset apenas com os valores inteiros a partir dos atributos “category” e “type”.

Em seguida, foi feita a separação do dataset em porções de 80% e 20%. Seguem o mapeamento das categorias e suas distribuições em cada separação:

'Antivirus detection': 1,
'Artifacts dropped': 2,
'Attribution': 3,
'External analysis': 4,
'Financial fraud': 5,
'Internal reference': 6,
'Network activity': 7,
'Other': 8,
'Payload delivery': 9,
'Payload installation': 10,
'Payload type': 11,
'Persistence mechanism': 12,
'Targeting data': 13





Para o treinamento foram usados os modelos KNN, Random Forest e Perceptron. Os caminhos e resultados específicos estão exemplificados no arquivo final.ipynb.

4. Conclusão

Observando todas as métricas obtidas, o algoritmo KNN é, num geral, a melhor opção para o dataset escolhido, visto que possui menor taxa de erro e maior taxa de score. Sobre o dataset em geral, acredito não ter sido a melhor escolha, visto que a pouca quantidade de atributos faz com que não possam ser feitas muitas análises e os algoritmos de Machine Learning parecem quase uma técnica mais avançada que o necessário para estes dados

5. Referências

- https://minerandodados.com.br/prevendo-a-demanda-de-alugueis-de-bicicletas-com-machine-learning/
- https://minerandodados.com.br/machine-learning-na-pratica-knn-python/
- https://www.kaggle.com/firebits/vulcanoio-org-misp2-4-54-initial-20161127-07h35m