Classificando ataques em honeypots com inteligência artificial

Pamela Moura Gonçalves[1], Roben Castagna Lunardi[1*]
Orientador(a)[*]
[1]Instituto Federal de Educação, Ciência e Tecnologia do Rio Grande do Sul (IFRS) - Campus Restinga. Porto Alegre, RS

Os honeypots simulam um sistema com vulnerabilidade para atrair pessoas mal-intencionadas, ajudando a aprender mais sobre seus comportamentos com as informações presentes nos logs. Esses registros são diversificados, conforme o tipo de serviço que possui e seu objetivo principal. Como podem ser feitas muitas tentativas, às vezes até repetitivas, acabam sendo uma tarefa massiva para verificar e validar manualmente possíveis dados relevantes. Entretanto, nos últimos anos as áreas de ciência de dados e tratamento de dados evoluíram, permitindo a utilização de algoritmos de inteligência artificial, em especial, aprendizado de máquina, como ferramenta. Desta forma, além de ser capaz de lidar com grandes quantidades de dados, ajuda na identificação e predição de comportamentos fora do comum e possibilita a classificação do nível de risco de um eventual ataque. Em cenários maliciosos, poderia disparar gatilhos para o sistema lidar de maneira defensiva e de formas a enganar o adversário, dependendo do nível de interação e do que foi realizado. Por exemplo, o bloqueio do IP suspeito se tiver uma frequência grande de tentativas ou o redirecionamento do atacante para uma armadilha como forma de contra-ataque. Como citado anteriormente, os objetivos dos honeypots são diferentes. Desta forma, este trabalho tem por objetivo avaliar os serviços SSH e Telnet, utilizando o honeypot Cowrie, utilizando os respectivos logs através de algoritmos de aprendizado de máquina. Após uma análise, foram extraídos do total 16 features, incluindo tentativas de login falhas, comandos privilegiados, operações de arquivos, downloads e entre outras formas de poder agir. Para avaliar diferentes algoritmos, foram aplicados em modelos supervisionados: Random forest, KNN, SVR e Decision tree. Os algoritmos são do tipo regressão, devido a escolher um valor contínuo entre 0 e 1 como label para ser flexível. Ainda, como modelos não supervisionados foram avaliados: Isolation forest, K-means junto com PCA, que identificam padrões de comportamento, testados no ambiente do Google Colab. Após as análises, observou-se que Random Forest e KNN obtiveram os melhores resultados, ambos com R^2 treino e teste=0.99 sem variância, comparando com os modelos SVR com R^2 treino=0.22 e R^2 teste=0.24, e Decision Tree com R^2 treino=1.0 perfeito demais, possível overfitting e R^2 teste=0.99. Para está comparação, foi feita a avaliação entre seus dados com R^2, que determina o quanto está próximo das medidas, MAE, o quanto o modelo está, em média, distante do valor e RMSE, mais sensível a erros maiores. Foi observado que seria de grande ajuda utilizar a inteligência artificial, tanto para lidar com grande quantidade de dados como para classificar o nível de risco que iria oferecer conforme as atividades observadas e, a partir disso, em conjunto poder fazer uma automatização defensiva, diversificando com o grau de ameaça.

Palavras-chave: IA; Logs; Honeypots.