Đại học Quốc gia Thành phố Hồ Chí Minh
Trường Đại học Công nghệ Thông tin (UIT)

BÁO CÁO LAB

Môn học: Phương pháp học máy trong an toàn thông tin

Ex 1: Anomaly Detection in Cybersecurity

Nhóm: 06

THÔNG TIN CHUNG:

Lớp: NT522.N21.ATCL

STT	Họ và tên	MSSV
1	Đỗ Quang Thắng	20521893
2	Nguyễn Đoàn Thiên Cung	20521146
3	Vũ Trọng Nghĩa	20520651

1. **NỘI DUNG THỰC HIỆN:**

BÁO CÁO CHI TIẾT

Colab :
https://colab.research.google.com/drive/1UrVQQK37B6kIV_pZusSkoIS4KxnSeGSn?usp=sharing

A. **Anomaly Detection**

Đánh giá :

- Chiến lược huấn luyện của IF và OC-SVM đều là học không giám sát, tức là không có sự can thiệp của người giám sát trong quá trình huấn luyện. Cả hai thuật toán đều sử dụng các đặc trưng của lưu lượng mạng để xác định sự khác biệt giữa các lưu lượng bình thường và các lưu lượng dị thường.
- Về khả năng nhận diện các lưu lượng mạng tấn công, xâm nhập, kết quả đánh giá của các mô hình IF và OC-SVM phụ thuộc vào tính chất của các lưu lượng mạng trong tập dữ liệu và cách thức xử lý dữ liệu. Tuy nhiên, các mô hình học máy này thường có khả năng phát hiện các lưu lượng mạng dị thường với độ chính xác tương đối cao, đặc biệt là trong trường hợp các lưu lượng này có tính chất đặc trưng và khác biệt so với các lưu lượng bình thường

B. **Attack Classification**

- Kết quả đánh giá của các mô hình trên tập kiểm tra được trình bày trong bảng dưới đây:

Mô hình	Accuracy	Precision	Recall	F1-score
Decision Tree	0.9998	0.9997	0.9999	0.9998
Logistic	0.9358	0.9364	0.9333	0.9348
SVM (rbf)	0.9979	0.9981	0.9975	0.9978

- Ta thấy rằng mô hình Decision Tree đạt được kết quả tốt nhất trên tập kiểm tra với các tiêu chí đánh giá. Mô hình Logistic Regression đạt hiệu quả trung bình, trong khi mô hình SVM đạt hiệu quả tương đối tốt.

- Để đánh giá khả năng nhận diện các lưu lượng mạng tấn công, xâm nhập trong hệ thống mạng ở điều kiện thực tế, chúng ta cần xem xét các tiêu chí đánh giá khác như ROC, AUC, Precision-Recall curve, F1-score weighted.

```
from sklearn.metrics import plot_roc_curve

plot_roc_curve(dt, X_test, y_test)
plot_roc_curve(lr, X_test, y_test)
plot_roc_curve(svm, X_test, y_test)
```

- Biểu đồ ROC cho các mô hình trên tập kiểm tra :

- Ta thấy rằng mô hình Decision Tree lại cho kết quả tốt nhất với AUC = 0.9998, mô hình Logistic Regression đạt AUC = 0.9022, trong khi mô hình SVM đạt AUC = 0.9925. Do đó, mô hình Decision Tree có khả năng nhận diện các lưu lượng mạng tấn công, xâm nhập trong hệ thống mạng tốt nhất.

- Tỉ lệ giữa nhãn dữ liệu Normal-Attack trong tập dữ liệu ảnh hưởng đến hiệu năng của mô hình học máy.

- Khi tập dữ liệu bị mất cân bằng tỉ lệ giữa các nhãn, thì thông số/tiêu chí nào để đánh giá mô hình học máy được xem là phù hợp?

  Trong trường hợp này, khi tập dữ liệu bị mất cân bằng, chỉ sử dụng Accuracy là không đủ để đánh giá hiệu quả của mô hình. Thay vào đó, chúng ta cần sử dụng các tiêu chí đánh giá khác như Precision, Recall, F1-score weighted. Các tiêu chí

này đánh giá hiệu quả của mô hình trên từng lớp riêng biệt và có thể hiển thị được khả năng phát hiện tấn công trong các lớp thiểu số.