Câu 1:  
- Khi thực hiện việc học tại Trường Đại học Công Nghệ Thông Tin, ta thường phải thực hiện nghĩa vụ mua bảo hiểm y tế vào mỗi năm với giá hơn 680.000 đồng.

1. v/v Tạo Giá trị:

"Trường Đại học Công Nghệ Thông Tin" là người tạo giá trị BHYT: 680000 đồng  
Ban Giám hiệu Nhà Trường đã QLRR cho sinh viên bằng cách tạo ra giá trị là 680000đ    
QLRR ở đây là rủi ro về sức khỏe của sinh viên.

2. v/v Bảo vệ giá trị:  
Để bảo vệ giá trị 680000đ, nhà trường làm những việc sau đây cho sinh viên để khuyến khích sinh viên mua BHYT hàng năm:  
2.1 Yêu cầu các cơ quan y tế để giảm chi phí khám chữa bệnh tại các bệnh viện công;  
2.2 Thực hiện xét điều kiện để cấp danh hiệu sinh viên khá, giỏi, xuất sắc và các loại học bổng có trong chương trình đào tạo (trong đó có học bổng tài năng của chương trình tài năng mỗi kỳ học) cho các sinh viên mua BHYT với giá trị như trên.  
2.3 Chi trả một phần viện phí bởi các tổ chức liên quan khi thực hiện khám chữa bệnh tại một số bệnh viện công trên cả nước;  
2.4  Thực hiện xét điều kiện để cấp các loại danh hiệu và học bổng cho sinh viên đủ điều kiện.

Câu 2:  
Câu này em sử dụng ngữ cảnh ứng dụng mô hình học máy và mô hình học sâu để phát hiện lỗ hỏng trong hợp đồng thông minh (smart-contract) trong blockchain

a) ĐO LƯỜNG (MEASURE)  
Các chỉ số báo cáo về Accuracy, Precison, Recall và F1 Score của em đạt 96.7%, 99.9%, 87.5%, và 93.3%

Một bài báo khác (<https://arxiv.org/abs/2309.08474>) chỉ các chỉ số trên ở mức 77% đến 78%

b) PHÂN TÍCH (ANALYSE)  
Phân tích vì sao độ chính xác của mô hình của em cao hơn bài báo đó?  
- Bài của em sử dụng bộ dữ liệu (dataset) do chính em dựng lên từ các kịch bản đã có trước - CrossChainSentinel (<https://github.com/anhkiet1227/CrossChainSentinel>) trong khi đó bài báo sử dụng bộ dữ liệu (dataset) SolidiFI-Benchmark và Smartbugs Wild (nguồn public và có thể có nhiều dữ liệu bị nhiễu)

- Bài của em xây dựng model bằng phương pháp RoBERTa (phương pháp cải tiến của BERT) trong khi bài báo sử dụng phương pháp cũ hơn là BERT, Bi-LSTM và GNN

c) QUẢN LÝ (MANAGE)

c1 Về việc quản lý thì hiện tại em tự làm việc một mình, tự thực hiện quản lý tất cả các công việc

c2 Các tài liệu về quy tắc, quy trình:

|  |  |  |
| --- | --- | --- |
| Stt | Tên tài liệu | Mô tả tóm tắt nội dung tài liệu |
| 1 | Xác định các loại lỗ hỏng | Định nghĩa các loại lỗ hỏng có trong smart contract |
| 2 | Xây dựng tập dữ liệu (dataset) huấn luyện mô hình học máy | Hướng dẫn các phương pháp để xây dựng tập dữ liệu nhằm huấn luyện và đánh giá mô hình học máy |

|  |  |  |
| --- | --- | --- |
| 3 | Hướng dẫn các phương pháp, các mô hình học máy được sử dụng trong việc kiểm tra lỗ hỏng | Tài liệu nhằm hướng dẫn sử dụng các mô hình học máy nhằm phát hiện ra các lỗ hỏng một cách tự động bằng các mô hình học máy được áp dụng từ đó phát hiện ra các smart contract một cách tự động |

c3 Thời gian quản lý công việc

1/9/2023 – 10/9/2023: Hoàn thành dàn ý cho ý tưởng, nghiên cứu các công trình liên quan, đưa ra mô hình tổng quan  
11/9/2023 – 1/10/2023: Hoàn thành thực nghiệm triển khai cầu nối và tấn công trên cầu nối  
2/10/2023 – 20/10/2023: Hoàn thành thu thập dữ liệu nhằm dựng file dataset  
21/10/2023 – 14/11/2023: Hoàn thành prediction model, đưa ra số liệu đánh giá, tối ưu hoá mô hình  
15/11/2023 – 31/12/2023: Hoàn thành báo cáo Khoá luận tốt nghiệp, hiệu chỉnh những thông tin cần thiết

Các chỉ số hiện tại là chỉ số cố định do tập dữ liệu đang ở dạng static, do em chưa public thành dymamic dataset nên việc các chỉ số đo về Accuracy, Precison, Recall và F1 Score là cố định

c4 Cách quản lý tập dữ liệu

Thông tin quản lý, đặc điểm của file dataset đã được lưu tại:

<https://github.com/anhkiet1227/CrossChainSentinel/blob/main/README.md>

Về phiên bản, thời gian hình thành bộ dữ liệu: Đã được lưu thông tin tại:

<https://github.com/anhkiet1227/CrossChainSentinel/graphs/code-frequency>

<https://github.com/anhkiet1227/CrossChainSentinel/graphs/commit-activity>

Và hiện tại do em đang làm một mình nên em là người quản lý toàn bộ về bộ dữ liệu.

d) KIỂM SOÁT (CONTROL)

- Thực hiện kiểm soát tài liệu để quản lý tốt hơn về bộ dữ liệu (<https://github.com/anhkiet1227/CrossChainSentinel/blob/main/README.md>)

- Thực hiện kiểm soát về giấy phép sử dụng bộ dữ liệu (<https://github.com/anhkiet1227/CrossChainSentinel/blob/main/LICENSE>)

- Thực hiện kiểm soát tài liệu để có thể chia sẻ việc cải thiện mô hình

- Mã hoá mô hình để chống những kẻ tấn công can thiệp vào mô hình, làm sai lệch đi khả năng nhận diện

- Thực hiện duy trì thành tích các chỉ số về Accuracy, Precison, Recall và F1 Score cần phải duy trì về môi trường lý tưởng cho hệ thống, hiện tại em đang sử dụng môi trường Google Collab Pro với giá 10.52 USD bao gồm (9.99 USD tiền phí duy trì và 0.53 USD tiền thuế)

- Hiện tại đối với bộ dữ liệu việc kiểm soát phiên bản bao gồm các phương pháp recover/restore/roll back được thực hiện trên môi trường Github đảm bảo cho chủ (master) của bộ dữ liệu có thểm kiểm soát được bộ dữ liệu từ cộng đồng và bộ dữ liệu của chủ. Và phương pháp phát hiện nhiễu dữ liệu thì em đang thực hiện bằng phương pháp LIME (<https://github.com/marcotcr/lime>) để phát hiện và lọc các giá trị gây nhiễu trong bộ dữ liệu

e) CẢI THIỆN (IMPROVEMENT)

Dựa theo kết quả đo lường (MEASURE) đã đạt được, em sẽ cải thiện trong tương lại như sau:

- Thực hiện nghiên cứu các kĩ thuật kết hợp các mô hình khác và tích hợp mô hình mới (GPT-2, GPT-3, GPT-4, OpenAI) để có thể có được các chỉ số phát hiện và dự đoán tốt hơn ở bốn chỉ số  Accuracy, Precison, Recall và F1 Score

- Cải thiện bộ dữ liệu để mô hình có thể học được tốt hơn (từ 300 file lên 1000 hoặc 2000 file).