**ĐẠI HỌC QUỐC GIA TP.HCM**

**TRƯỜNG ĐẠI HỌC CÔNG NGHỆ THÔNG TIN**



**MÔN HỌC: MẬT MÃ HỌC**

**BÁO CÁO ĐỒ ÁN CUỐI KÌ**

**Privacy-Preserving Credit Card Fraud Detection using Homomorphic Encryption**

**Giảng viên hướng dẫn: Nguyễn Ngọc Tự**

**NT219.N21.ANTN**

**Sinh viên thực hiện:**

**21522056 – Hà Thị Thu Hiền**

**21522641 – Phạm Ngọc Thơ**

**TP.HCM, ngày 10 tháng 6 năm 2023**

**Table of Contents**

[**A.** **BÁO CÁO CHI TIẾT** 2](#_Toc137379157)

[I.Mở đầu 2](#_Toc137379158)

[II. Chi tiết đồ án 3](#_Toc137379159)

[1. Ngữ cảnh 3](#_Toc137379160)

[2. Các bên liên quan 4](#_Toc137379161)

[3. Yêu cầu bảo mật 5](#_Toc137379162)

[4. Tổng quan các giải pháp 6](#_Toc137379163)

[III. Kết quả nghiên cứu 8](#_Toc137379164)

[1. Trên Jupyter Notebook: Linux d66e2834fdf9 5.15.109+ #1 SMP Wed May 31 10:45:00 UTC 2023 x86\_64 x86\_64 x86\_64 GNU/Linux (Kaggel) 9](#_Toc137379165)

[2. Trên Jupyter Notebook: Linux iec 5.15.0-73-generic #80-Ubuntu SMP Mon May 15 15:18:26 UTC 2023 x86\_64 x86\_64 x86\_64 GNU/Linux (Server) 9](#_Toc137379166)

[IV. DEMO 10](#_Toc137379167)

[1. Trên Jupyter Notebook: Linux d66e2834fdf9 5.15.109+ #1 SMP Wed May 31 10:45:00 UTC 2023 x86\_64 x86\_64 x86\_64 GNU/Linux (Kaggel) 10](#_Toc137379168)

[2. Trên Jupyter Notebook: Linux iec 5.15.0-73-generic #80-Ubuntu SMP Mon May 15 15:18:26 UTC 2023 x86\_64 x86\_64 x86\_64 GNU/Linux (Server) 17](#_Toc137379169)

[**B.** **PHÂN CÔNG CÔNG VIỆC:** 22](#_Toc137379170)

[**C.** **TÀI LIỆU THAM KHẢO:** 23](#_Toc137379171)

# **BÁO CÁO CHI TIẾT**

## I.Mở đầu

* Gian lận thẻ tín dụng là một vấn đề mà các tổ chức tài chính và khách hàng của họ liên tục phải đối mặt, vấn đề này được giảm thiểu bằng các hệ thống phát hiện gian lận. Tuy nhiên, các hệ thống này yêu cầu sử dụng dữ liệu giao dịch nhạy cảm của khách hàng, điều này dẫn đến việc thiếu quyền riêng tư cho khách hàng và lỗ hổng vi phạm dữ liệu đối với nhà cung cấp thẻ. Bài viết này đề xuất một hệ thống phát hiện gian lận riêng tư đối với các giao dịch được mã hóa bằng cách sử dụng mã hóa đồng cấu. A feedforward classifier neural network (mạng thần kinh phân loại chuyển tiếp), được đào tạo thành công cụ phát hiện gian lận trên dữ liệu văn bản gốc. Sau đó, chúng được chuyển đổi thành các mô hình sử dụng mã hóa đồng cấu để suy luận riêng tư.
* Gian lận thẻ tín dụng là việc sử dụng thẻ thanh toán của ai đó một cách gian lận, thường liên quan đến việc đánh cắp tiền thông qua chuyển khoản hoặc mua hàng hóa. Ngân hàng Trung ương Châu Âu nhận thấy rằng chỉ riêng trong năm 2019, giá trị của các giao dịch gian lận đã lên tới hơn 1,8 tỷ €. Gian lận thẻ tín dụng là một vấn đề quan trọng đối với các ngân hàng, vì họ phải thực hiện các khoản thanh toán lớn cho khách hàng mỗi năm. Đó cũng là một sự thất vọng lớn đối với khách hàng khi thẻ của họ bị đánh cắp tiền hoặc bị đóng băng do các hoạt động gian lận. Điều này làm cho việc phát hiện gian lận thẻ tín dụng trở thành một vấn đề sống còn đối với các ngân hàng và khách hàng, đồng thời là một bài toán khó giải trong thế giới tài chính. Các công cụ phát hiện gian lận hiện tại đã đạt được thành công khi sử dụng các phương pháp dựa trên quy tắc hoặc phương pháp Máy học (ML) trên dữ liệu xung quanh các giao dịch.
* Mặc dù các công cụ phát hiện đã làm giảm tương đối gian lận thẻ tín dụng, nhưng điều đó dẫn đến nhu cầu người dùng chia sẻ dữ liệu bổ sung, ở dạng văn bản thuần túy. Xu hướng vi phạm dữ liệu ngày càng tăng thúc đẩy việc giảm lượng dữ liệu nhạy cảm được thu thập từ người dùng. Ngoài ra, một số người dùng có nhu cầu cao về quyền riêng tư có thể không muốn chia sẻ dữ liệu này với ngân hàng của họ, chứ đừng nói đến các tin tặc tiềm năng. Điều này đòi hỏi một giải pháp trong đó người dùng dữ liệu có thể được giữ an toàn, đồng thời cho phép sử dụng dữ liệu đó trong các công cụ phát hiện gian lận. Bài viết này khám phá việc sử dụng HE (Homomophic Encryption) để xây dựng một công cụ phát hiện gian lận thẻ tín dụng đáp ứng nhu cầu về quyền riêng tư này. Vì trọng tâm chính là bảo vệ quyền riêng tư nên mục đích không phải là đạt được hiệu suất tối tân trong nhiệm vụ phát hiện gian lận, mà là tạo ra một mô hình hoạt động tốt như một công cụ phát hiện gian lận trong miền HE.

## II. Chi tiết đồ án

### Ngữ cảnh

A picture containing text

Description automatically generated

A diagram of a person sitting in a chair

Description automatically generated with medium confidence

### Các bên liên quan



* **Ngân hàng:** Ngân hàng cung cấp dữ liệu giao dịch được mã hóa đồng cấu, giúp bảo vệ thông tin người dùng.



* **Bên thứ ba (Nhà cung cấp dịch vụ):** Nhà cung cấp dịch vụ này sử dụng mô hình phát hiện gian lận của họ để dự đoán và phát hiện các hoạt động gian lận dựa trên dữ liệu giao dịch đã mã hóa.



* **Người dùng:** Những người sử dụng ngân hàng mà thông tin giao dịch của họ được sử dụng để phát hiện gian lận.



* **Attacker (Kẻ tấn công):** Người cố gắng tấn công mô hình hoặc dữ liệu, có thể là để thực hiện gian lận hoặc lấy cắp thông tin.
  + **Tấn công dữ liệu:** Kẻ tấn công có thể tìm cách lấy cắp dữ liệu đã được mã hóa đồng cấu từ ngân hàng hoặc từ nhà cung cấp dịch vụ. Tuy nhiên, dữ liệu đã được mã hóa đồng cấu thì sẽ rất khó để giải mã và khôi phục thông tin gốc.
  + **Tấn công mô hình:** Kẻ tấn công có thể tìm cách tấn công vào mô hình phát hiện gian lận, ví dụ như thông qua các tấn công adversarial, để làm cho mô hình không chính xác hoặc không đáng tin cậy.
  + **Tấn công hệ thống mã hóa:** Kẻ tấn công có thể cố gắng tìm ra các lỗ hổng trong hệ thống mã hóa đồng cấu để giải mã dữ liệu.

### Yêu cầu bảo mật

* **Ngân hàng:** Ngân hàng cần đảm bảo rằng dữ liệu giao dịch và thông tin khách hàng được bảo mật một cách tuyệt đối. Điều này có thể đòi hỏi việc sử dụng các công nghệ bảo mật như mã hóa đồng cấu, cũng như quản lý truy cập chặt chẽ và các biện pháp bảo mật dữ liệu khác. Ngân hàng cũng cần tuân thủ các quy định về bảo vệ dữ liệu cá nhân.
* **Bên thứ ba (Nhà cung cấp dịch vụ):** Nhà cung cấp dịch vụ cần bảo vệ mô hình của họ khỏi các tấn công và đảm bảo rằng họ chỉ xử lý dữ liệu đã được mã hóa, không tiếp xúc với dữ liệu gốc. Họ cũng cần đảm bảo rằng hệ thống và cơ sở hạ tầng của họ đủ mạnh để chống lại các cuộc tấn công và rằng họ có các kế hoạch phục hồi sau khi bị tấn công.
* **Người dùng (khách hàng của ngân hàng):** Người dùng cần được bảo vệ khỏi việc thông tin cá nhân và dữ liệu giao dịch bị tiết lộ. Họ cũng cần được giáo dục về các biện pháp bảo vệ bản thân trên mạng và cách thức sử dụng ngân hàng trực tuyến một cách an toàn.

### Tổng quan các giải pháp

#### **HE Feedforward Classifer**

* A Feedforward Classifier NN (Neural Network) model đã được chọn do thành công trước đó của nó đối với việc phát hiện gian lận thẻ tín dụng và kiến trúc đơn giản của nó. Việc triển khai HE được sử dụng là lược đồ CKKS, vì nó được phát hiện là có thời gian suy luận nhanh nhất đối với một kiến trúc tương tự và các triển khai nguồn mở được hỗ trợ bằng các ngôn ngữ lập trình khác nhau. Lược đồ CKKS hoạt động tốt đối với NN, vì nó có độ trễ nhanh bằng cách sử dụng số học gần đúng. Vì Feedforward Classifer NNs tạo ra đầu ra mờ, nên việc sử dụng phép tính gần đúng có thể dẫn đến đầu ra hơi khác, nhưng hiển thị cùng một dự đoán sau khi chuyển đổi thành nhị phân.

A picture containing diagram, text, line, font

Description automatically generated

#### **Kiến trúc hệ thống (System Architecture)**

A screenshot of a computer

Description automatically generated with medium confidence

cA picture containing text, screenshot, diagram, font

Description automatically generated

#### **Bộ dữ liệu được chọn (Datasets Chosen)**

* **"IEEE CIS Fraud Detection" (Vesta)**
* Bộ dữ liệu Vesta đã được phát hành để sử dụng trong cuộc thi Phát hiện gian lận của IEEE CIS3.
* Nó chứa 590.540 giao dịch, 20.663 trong số đó là gian lận (3,5%). Mỗi giao dịch có 431 đặc điểm (400 số, 31 phân loại), cùng với tương đối dấu thời gian và nhãn cho dù đó là gian lận hay hợp pháp. Mục đích để ẩn danh, tên của các tính năng nhận dạng đã được che dấu, cùng với tên của các tính năng bổ sung được thiết kế bởi Vesta. Một số tính năng thưa thớt, có nghĩa là tính năng này trống rỗng hoặc Không phải là Số (NaN) trong nhiều giao dịch.

A picture containing text, screenshot, font, receipt

Description automatically generated

Link: [IEEE-CIS Fraud Detection | Kaggle](https://www.kaggle.com/competitions/ieee-fraud-detection/data)

#### **Thực hiện (Implementation)**

**Core Language & Tools:**

* **Python 3:** được chọn làm ngôn ngữ lập trình để triển khai, sơ đồ CKKS cũng có các triển khai Python.
* **Weights & Biases:** Trọng số & độ lệch (W&B) là một công cụ hỗ trợ đào tạo các mô hình ML, thông qua ghi nhật ký hiệu quả và trực quan hóa các lần chạy đào tạo. Việc đưa W&B vào quy trình đào tạo cho phép ghi nhật ký các số liệu để đảm bảo các mô hình đang đào tạo đúng cách và để theo dõi hiệu suất của từng mô hình.
* **Jupyter Notebook:** 
  + Linux d66e2834fdf9 5.15.109+ #1 SMP Wed May 31 10:45:00 UTC 2023 x86\_64 x86\_64 x86\_64 GNU/Linux **(Kaggel)**
  + Linux iec 5.15.0-73-generic #80-Ubuntu SMP Mon May 15 15:18:26 UTC 2023 x86\_64 x86\_64 x86\_64 GNU/Linux **(Server)**
* **Model Training Packages:** PyTorch được cho là phù hợp để chuyển đổi mô hình sang HE, vì nó sử dụng cách tiếp cận hướng đối tượng.
* **TenSEAL:** TenSEAL là một thư viện chứa các liên kết Python để triển khai lược đồ C++ CKKS trong Microsoft SEAL. TenSEAL cũng đã cung cấp các hướng dẫn về việc chuyển đổi các NN văn bản gốc sang các mô hình HE, điều này rất hữu ích cho việc viết triển khai này.
* **Model Training:** Logistic Regression model.

## III. Kết quả nghiên cứu

### Trên Jupyter Notebook: Linux d66e2834fdf9 5.15.109+ #1 SMP Wed May 31 10:45:00 UTC 2023 x86\_64 x86\_64 x86\_64 GNU/Linux (Kaggel)

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Model** | **Performance** | **Latency Results** | | | **Storage Size Results** |
| **Accuracy** | **Encryption of the test-set** | **Encryption of the training-set** | **Average time per epoch** | **The size of the model** |
| **Logistic Regression (LR)** | 0.96 | - | - | Vài ms | 2.2 KB |
| **Encrypted Logistic Regression (EncryptedLR)** | 0.96 | 3 s | 31 s | 435 s | 2 KB |

### Trên Jupyter Notebook: Linux iec 5.15.0-73-generic #80-Ubuntu SMP Mon May 15 15:18:26 UTC 2023 x86\_64 x86\_64 x86\_64 GNU/Linux (Server)

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Model** | **Performance** | **Latency Results** | | | **Storage Size Results** |
| **Accuracy** | **Encryption of the test-set** | **Encryption of the training-set** | **Average time per epoch** | **The size of the model** |
| **Logistic Regression (LR)** | 0.94 | - | - | Vài ms | 2.2 KB |
| **Encrypted Logistic Regression (EncryptedLR)** | 0.95 | 1 s | 49 s | 699 s | 2 KB |

## IV. DEMO

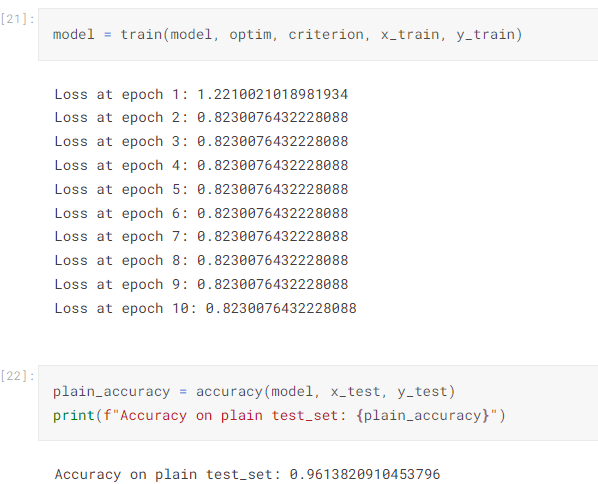
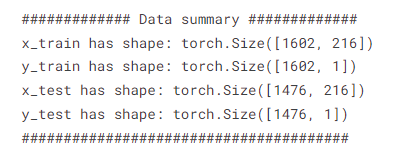
### Trên Jupyter Notebook: Linux d66e2834fdf9 5.15.109+ #1 SMP Wed May 31 10:45:00 UTC 2023 x86\_64 x86\_64 x86\_64 GNU/Linux (Kaggel)

**LINK NOTEBOOK:** [**https://www.kaggle.com/code/skn443/he-ml-ieee-fraud-detection**](https://www.kaggle.com/code/skn443/he-ml-ieee-fraud-detection)

#### **Training a Logistic Regression Model**

A screenshot of a computer

Description automatically generatedA screenshot of a computer code

Description automatically generated with low confidenceA screenshot of a computer program

Description automatically generated with medium confidence

#### **Encrypted Evaluation**

A screen shot of a computer code

Description automatically generated with low confidenceA screenshot of a computer code

Description automatically generated with low confidence

#### **Training an Encrypted Logistic Regression Model on Encrypted Data**A screenshot of a computer code Description automatically generated with medium confidenceA screenshot of a computer Description automatically generatedA screenshot of a computer Description automatically generated with medium confidence

### Trên Jupyter Notebook: Linux iec 5.15.0-73-generic #80-Ubuntu SMP Mon May 15 15:18:26 UTC 2023 x86\_64 x86\_64 x86\_64 GNU/Linux (Server)

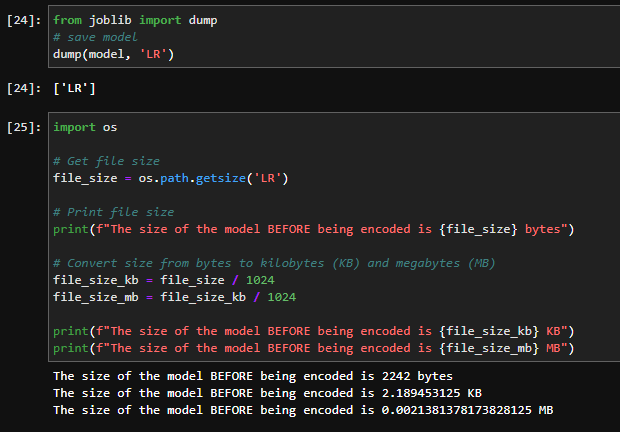
#### **2.1 Training a Logistic Regression Model**

A screen shot of a computer

Description automatically generated with medium confidenceA screenshot of a computer screen

Description automatically generated with low confidenceA screen shot of a computer

Description automatically generated with medium confidenceA screenshot of a computer program

Description automatically generated with medium confidence

#### **2.2 Encrypted Evaluation**

A screen shot of a computer program

Description automatically generated with low confidenceA picture containing text, screenshot, display, software

Description automatically generated

#### **Training an Encrypted Logistic Regression Model on Encrypted Data**

A picture containing text, screenshot, software, multimedia software

Description automatically generated

A picture containing text, screenshot, font

Description automatically generatedA screenshot of a computer program

Description automatically generated with medium confidence

# **PHÂN CÔNG CÔNG VIỆC:**

|  |  |  |
| --- | --- | --- |
| **MSSV** | **Họ và tên** | **Công việc** |
| **21522056** | **Hà Thị Thu Hiền** | **Đọc báo**  **Nghiên cứu và phân tích lí thuyết**  **Thu thập tài liệu**  **Làm slide**  **Làm demo**  **Thuyết trình**  **Viết báo cáo cuối kỳ** |
| **21522641** | **Phạm Ngọc Thơ** | **Đọc báo**  **Nghiên cứu và phân tích lí thuyết**  **Thu thập tài liệu**  **Làm demo** |

# **TÀI LIỆU THAM KHẢO:**

**Nugent, D. (2022). Privacy-Preserving Credit Card Fraud Detection using Homomorphic Encryption. *arXiv preprint arXiv:2211.06675*.**