**ĐẠI HỌC HUẾ**

**TRƯỜNG ĐẠI HỌC KINH TẾ**

**KHOA HỆ THỐNG THÔNG TIN KINH TẾ**

**--------🙣🕮🙡--------**

****

**KHÓA LUẬN TỐT NGHIỆP**

**PHÁT HIỆN WEBSITE LỪA ĐẢO DỰA TRÊN MÔ HÌNH PHÂN LỚP RANDOM FOREST**



**Người thực hiện: Lý Thanh Tùng Giảng viên hướng dẫn:**

**Lớp: K49B Tin học kinh tế TS.Nguyễn Đình Hoa Cương**

**Niên khóa: 2015 - 2019**

***Huế, tháng 5 năm 2019***

# **LỜI CẢM ƠN**

*Luận văn tốt nghiệp được hoàn thành tại Trường Đại học Kinh tế Huế. Trong quá trình làm khóa luận em đã nhận được rất nhiều sự giúp đỡ để hoàn thành luận văn.*

*Đầu tiên em xin gửi lời cám ơn chân thành đến thầy Nguyễn Đình Hoa Cương, người đã hướng dẫn tận tình cho em trong suốt thời gian thực tập. Thầy đã không ngần ngại chỉ dẫn, truyền đạt kiến thức, định hướng đi cho em để em hoàn thành tốt nhiệm vụ. Một lần nữa em chân thành gửi lời cám ơn đến thầy và chúc thầy dồi dào sức khỏe.*

*Em xin gửi lời cám ơn đến quý thầy cô Khoa Hệ Thống Thông Tin Kinh Tế, Trường Đại Học Kinh Tế Huế, những người đã truyền đạt kiến thức quý báu cho em trong suốt thời gian học tập tại trường.*

*Ngoài ra, em xin cám ơn tất cả bạn bè, thư viện, doanh nghiệp, công ty đã giúp đỡ, dìu dắt em trong suốt thời gian qua. Tất cả mọi người đã nhiệt tình giúp đỡ, đặc biệt là Trung tâm Công nghệ Thông tin Thừa Thiên Huế - HueCIT.*

*Tuy nhiên vì kiến thức chuyên môn còn hạn chế và bản thân còn thiếu nhiều kinh nghiệm thực tiễn nên nội dung của báo cáo không tránh khỏi những thiếu xót, em rất mong nhận sự góp ý, chỉ bảo thêm của quý thầy cô cùng toàn thể cán bộ, công nhân viên tại các doanh nghiệp để báo cáo này được hoàn thiện hơn.*

*Một lần nữa xin gửi đến thầy cô, bạn bè cùng các cô chú, anh chị tại các doanh nghiệp lời cảm ơn chân thành và tốt đẹp nhất!*

*Huế, tháng 5 năm 2019*

*SV. Lý Thanh Tùng*

# **MỤC LỤC**

LỜI CẢM ƠN II

MỤC LỤC III

DANH MỤC HÌNH VẼ VIII

DANH MỤC BẢNG BIỂU IX

DANH MỤC SƠ ĐỒ X

DANH MỤC THUẬT NGỮ XI

PHẦN I. MỞ ĐẦU 1

1. Lý do chọn đề tài. 1

2. Mục tiêu nghiên cứu. 1

3. Đối tượng nghiên cứu. 1

4. Phương pháp nghiên cứu. 1

5. Cấu trúc của Khóa luận. 2

PHẦN II. NỘI DUNG NGHIÊN CỨU 3

CHƯƠNG I. CƠ SỞ LÝ THUYẾT 3

1.1. Tổng quan về khai phá dữ liệu. 3

1.1.1. Nhu cầu phát hiện tri thức từ dữ liệu. 3

1.1.2. Khai phá dữ liệu. 4

1.1.3. Các lĩnh vực ứng dụng. 5

1.2. Các phương pháp khai phá dữ liệu. 6

1.2.1. Khai phá luật kết hợp. 6

1.2.1.1. Giới thiệu về Luật kết hợp. 6

1.2.1.2. Thuật toán Apriori. 7

1.2.1.3. Thuật toán FP-Growth. 7

1.2.1.4. Ứng dụng Khai phá luật kết hợp. 8

1.2.2. Phân cụm (Clustering). 8

1.2.2.1. Phân cụm phẳng. 8

1.2.2.2. Phân cụm phân cấp. 9

1.2.3. Phân lớp (Classification). 9

1.2.3.1. Mô hình phân lớp Naïve Bayes. 9

1.2.3.2. Mô hình phân lớp K-NN. 11

1.2.3.3. Mô hình phân lớp Support Vector Machines (SVM). 11

1.2.3.4. Mô hình phân lớp Mạng Nơ-ron nhân tạo. 12

1.2.4. Hồi quy (Regression). 14

1.2.4.1. Khái niệm. 14

1.2.4.2. Phân loại. 14

1.3. Phát biểu bài toán nghiên cứu. 14

1.4. Lý thuyết về mô hình phân lớp Random Forest. 15

1.4.1. Lịch sử. 15

1.4.2. Cơ sở lý thuyết. 15

1.4.2.1. Khái niệm. 15

1.4.2.2. Cây quyết định. 15

1.4.2.3. Random Forest. 16

1.4.3. Nguyên tắc hoạt động. 17

1.4.4. Ứng dụng. 18

1.5. Các công trình liên quan. 18

CHƯƠNG II. XÂY DỰNG MÔ HÌNH 25

1.1. Mô tả bộ dữ liệu. 25

1.1.1. Giới thiệu và mô tả dữ liệu. 25

1.1.2. Thông tin thuộc tính. 26

1.1.2.1. Sử dụng địa chỉ IP (using the IP Address). 26

1.1.2.2. Độ dài của URL (URL\_Length). 26

1.1.2.3. Sử dụng dịch vụ rút ngắn URL (Using URL Shortening Services). 27

1.1.2.4. Sử dụng biểu tượng (URL using Symbol). 27

1.1.2.5. Chuyển hướng bằng cách sử dụng gạch chéo (Redirecting using double slash). 27

1.1.2.6. Tiền tố và hậu tố trong tên miền (Prefix and Suffix). 27

1.1.2.7. Tên miền phụ và nhiều tên miền phụ (Sub Domain and Multi Sub Domains). 28

1.1.2.8. Giao thức Secure Sockets Layer (HTTP with Secure Sockets Layer). 28

1.1.2.9. Độ dài đăng ký tên miền (Domain Registration Length). 28

1.1.2.10. Favicon. 29

1.1.2.11. Sử dụng cổng không chuẩn (Using Non-Standard Port). 29

1.1.2.12. HTTPS Token trong tên miền của URL (The Existence of “HTTPS” Token in the Domain Part of the URL). 30

1.1.2.13. Request URL. 30

1.1.2.14. URL of Anchor. 30

1.1.2.15. Liên kết trong các thẻ (Links in <Meta>, <Script> and <Link> tags). 30

1.1.2.16. Trình xử lý biểu mẫu máy chủ (Server Form Handler (SFH)). 31

1.1.2.17. Gửi thông tin đến Email (Submitting Information to Email). 31

1.1.2.18. URL bất thường (Abnormal URL). 31

1.1.2.19. Chuyển tiếp trang web (Website Forwarding). 32

1.1.2.20. Tùy chỉnh thanh trạng thái (Status Bar Customization). 32

1.1.2.21. Vô hiệu hóa nhấp chuột phải (Disabling Right Click). 32

1.1.2.22. Sử dụng cửa số bật lên (Using Pop-up Window). 32

1.1.2.23. Chuyển hướng IFrame (IFrame Redirection). 33

1.1.2.24. Tuổi miền (Age of domain). 33

1.1.2.25. Bản ghi DNS (DNS Record). 33

1.1.2.26. Lượng truy cập trang web (Website Traffic). 33

1.1.2.27. Xếp hạng trang (PageRank). 34

1.1.2.28. Google Index. 34

1.1.2.29. Số lượng liên kết trỏ đến trang (Number of Links Pointing to Page). 34

1.1.2.30. Tính năng dựa trên báo cáo thống kê (Statistical-Reports Based Feature). 34

1.2. Xây dựng mô hình. 35

1.3. Phương pháp đánh giá mô hình. 35

1.3.1. Tổng quan về Receiver Operating Characteristic (ROC). 35

1.3.2. Xây dựng ROC để đánh giá. 36

1.4. Xây dựng ứng dụng dự đoán Website lừa đảo. 37

CHƯƠNG III. THÍ NGHIỆM VÀ KẾT QUẢ 38

3.1. Thiết lập thí nghiệm. 38

3.1.1. Phân lớp với các bộ phân lớp cơ bản. 38

3.1.2. Phân lớp với mô hình Random Forest. 41

3.1.3. Thiết lập mô hình đánh giá. 41

3.1.4. Thiết lập ứng dụng dự đoán Website lừa đảo. 45

3.1.4.1. Chuẩn bị bộ dữ liệu Unlabel. 45

3.1.4.2. Tạo ứng dụng sử dụng Java Swing. 46

3.2. Kết quả thí nghiệm. 49

PHẦN III. KẾT LUẬN 51

1. Kết quả đạt được. 51

2. Hạn chế của đề tài. 51

3. Hướng phát triển của đề tài. 51

DANH MỤC TÀI LIỆU THAM KHẢO 52

# **DANH MỤC HÌNH VẼ**

[Hình 1. 2. Quy trình khai phá dữ liệu (Process of Data mining). 5](#_Toc4074584)

[Hình 2. 2. Đường cong ROC. 36](#_Toc4074607)

[Hình 2. 3. Mô hình đánh giá ROC trên Knowledge Flow. 37](#_Toc4074608)

[Hình 3. 1. Chia bộ dữ liệu PhishingWebsite sử dụng bộ lọc Remove Percentage. 38](#_Toc4074540)

[Hình 3. 2. Chọn bộ phân lớp Naïve Bayes để huấn luyện. 39](#_Toc4074541)

[Hình 3. 3. Tiến hành huấn luyện bộ dữ liệu Training Set sử dụng bộ phân lớp Naïve Bayes với 10-fold Cross Validation. 39](#_Toc4074542)

[Hình 3. 4. Kiểm thử lại mô hình với Test Set. 40](#_Toc4074543)

[Hình 3. 5. Kết quả kiểm thử với Test Set của mô hình phân lớp Naïve Bayes. 40](#_Toc4074544)

[Hình 3. 6. Kết quả kiểm thử với Test Set của mô hình phân lớp Random Forest. 41](#_Toc4074545)

[Hình 3. 7. Lưu mô hình huấn luyện 10-fold Cross Valication. 42](#_Toc4074546)

[Hình 3. 8. Thêm model vào mô hình phân lớp. 43](#_Toc4074547)

[Hình 3. 9. Tiến hành đánh giá mô hình. 44](#_Toc4074548)

[Hình 3. 10. Đánh giá mô hình thành công. 44](#_Toc4074549)

[Hình 3. 11. Kết quả đánh giá mô hình đường ROC trên bộ dữ liệu Phishing Website. 45](#_Toc4074550)

[Hình 3. 12. Bộ dữ liệu Unlabel. 46](#_Toc4074551)

[Hình 3. 13. Giao diện của ứng dụng. 47](#_Toc4074552)

[Hình 3. 14. Các Java Class và Jframe của ứng dụng. 47](#_Toc4074553)

[Hình 3. 15. Tạo mới thư viện trong Netbeans. 48](#_Toc4074554)

[Hình 3. 16. Thêm file .jar và folder doc của Weka vào thư viện. 48](#_Toc4074555)

[Hình 3. 17. Kết quả của chương trình sau khi dự đoán. 49](#_Toc4074556)

# **DANH MỤC BẢNG BIỂU**

Bảng 2.1. Bảng các thuộc tính của bộ dữ liệu Phishing Website. 25

Bảng 2.2. Một số cổng quan trọng của trang Web. 29

[Bảng 3.1. Bảng kết quả thí nghiệm so sánh các kết quả các mô hình. 49](#_Toc4074897)

# **DANH MỤC SƠ ĐỒ**

[Sơ đồ 1. 1: Tiến hóa của công nghệ cơ sở dữ liệu theo quan điểm của J.Han và M.Kamber. 4](#_Toc4072756)

[Sơ đồ 1. 2. Quy trình quyết định dự đoán nhiệt độ (Cây quyết định cơ bản). 16](#_Toc4072757)

# **DANH MỤC THUẬT NGỮ**

|  |  |  |  |
| --- | --- | --- | --- |
| **STT** | **Từ viết tắt** | **Tiếng việt** | **Tiếng anh** |
| 1 | RF | Rừng ngẫu nhiên | Random Forest |
| 2 | NLP | Xử lý ngôn ngữ tự nhiên | Natural Language Processing |
| 3 | URL | Định vị Tài nguyên thống nhất | Uniform Resource Locator |
| 4 | PCA-RF | Phân tích thành phần chính của Rừng ngẫu nhiên | Principal Component Analysis Random Forest |
| 5 | oRFs | Rừng ngẫu nhiên xiên | oblique Random Forests |
| 7 | HEFS | Lựa chọn tính năng đồng bộ lai | Hybrid Ensemble Feature Selection |
| 8 | CDF-g | Chức năng phân phối tích lũy gradient | Cumulative Distribution Function gradient |
| 9 | UCI | Trường đại học California, Irvine | University of California, Irvine |
| 10 | CSDL | Cơ sở dữ liệu |  |
| 11 | DT | Decision Tree | Cây quyết định |

# **PHẦN I. MỞ ĐẦU**

1. **Lý do chọn đề tài.**

Internet ngày càng phát triển nhanh chóng, người dùng bắt đầu thay đổi các sở thích của họ như giải trí, học tập, làm việc, mua sắm truyền thống sang sử dụng các dịch vụ thông qua Internet. Kéo theo đó, việc lợi dụng Internet để lừa đảo ngày càng gia tăng. Những kẻ lừa đảo sử dụng nhiều chiêu trò khác nhau để người dùng mắc bẫy từ đó thu thập các thông tin như tên người dùng, mật khẩu, thẻ ngân hàng, số điện thoại... Các thông tin này sau đó được sử dụng cho các hành vi lừa đảo chủ yếu liên quan đến chiếm đoạt tài sản. Một vấn đề đặt ra là phân biệt giữa trang web lừa đảo và hợp pháp. Các trang web lừa đảo ngày càng tinh vi hơn và rất khó có thể nhận ra. Mặc dù đã có nhiều phương pháp và kỹ thuật khác nhau được đưa ra nhưng vẫn chưa có sự đồng nhất về phương pháp tốt nhất và các tính năng tối ưu trong việc đánh giá một trang web có phải là lừa đảo hay không. Trong khóa luận này, chúng tôi tiến hành nghiên cứu về việc phát hiện Website lừa đảo dựa trên mô hình phân lớp Random Forest.

1. **Mục tiêu nghiên cứu.**

Mục tiêu của khóa luận là thu thập dữ liệu vào tiến hành huấn luyện và kiếm thử dữ liệu với các bộ phân lớp trên Weka.

Sau đó, xây dựng một ứng dụng trên Java Swing để tiến hành dự đoán từ các bộ dữ liệu không nhãn (Unlabel Set) dự đoán nhãn lớp cho bộ dữ liệu đó.

1. **Đối tượng nghiên cứu.**

Đối tượng nghiên cứu ở đây là mô hình phân lớp Random Forest, bộ dữ liệu Phishing Website và cách xây dựng mô hình phân lớp cũng như thiết lập dự đoán Website lừa đảo.

1. **Phương pháp nghiên cứu.**

Phương pháp sử dụng xuyên suốt trong quá trình thực hiện đề tài là tìm hiểu về các kĩ thuật khai phá dữ liệu và sử dụng kĩ thuật đó trong khai phá dữ liệu.

Sử dụng phần mềm Weka để tiến hành các kĩ thuật khai phá, sử dụng các thuật toán để huấn luyện, kiểm thử và sau đó đánh giá mô hình.

Sử dụng NetBeans IDE xây dựng ứng dụng dự đoán Website lừa đảo.

1. **Cấu trúc của Khóa luận.**

Chương I: Cơ sở lý thuyết. Chương này giới thiệu tổng quan về khai phá dữ liệu, các phương pháp khai phá dữ liệu và phương pháp xây dựng mô hình đánh giá. Giới thiệu các công thức toán trong khai phá dữ liệu và phát biểu bài toán nghiên cứu của đề tài.

Chương II: Xây dựng mô hình. Chương này trình bày phương pháp xây dựng mô hình phân lớp của Random Forest và các bộ phân lớp cơ bản, phương pháp đánh giá ROC, và cách xây dựng ứng dụng dự đoán Website lừa đảo.

Chương III: Thí nghiệm và kết luận. Chương này tiến hành thí nghiệm, chia tập dữ liệu đã chuẩn bị thành bộ dữ liệu huấn luyện và kiểm thử, tiếp đến tiến hành kiểm thử với mô hình phân lớp Random Forest. Sau đó thiết lập mô hình đánh giá ROC bằng ứng dụng Knowledge Flow trên Weka và thiết lập ứng dụng trên Java Swing để tiến hành dự đoán Website lừa đảo cho các bộ dữ liệu không nhãn lớp. Cuối cùng sẽ đưa ra kết quả của cuộc thí nghiệm và đánh giá kết quả.

# **PHẦN II. NỘI DUNG NGHIÊN CỨU**

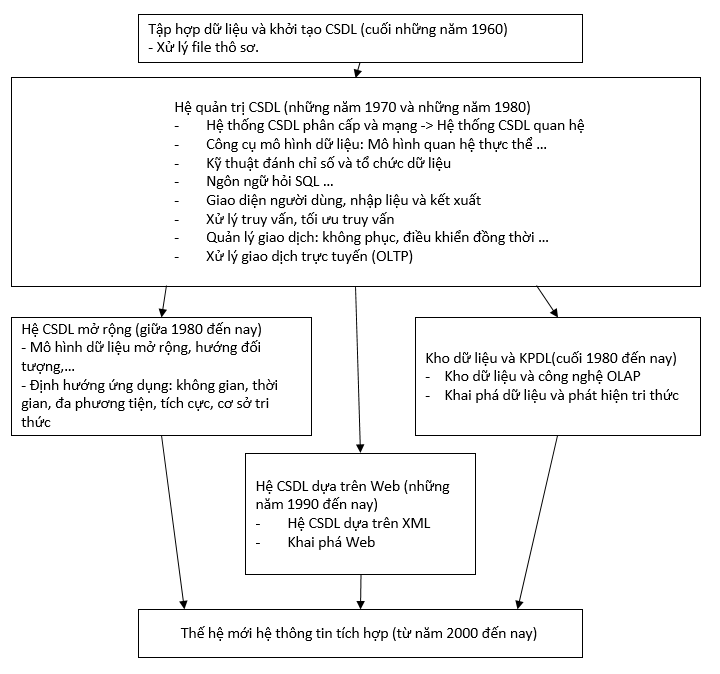
## **CHƯƠNG I. CƠ SỞ LÝ THUYẾT**

* 1. **Tổng quan về khai phá dữ liệu.**
     1. *Nhu cầu phát hiện tri thức từ dữ liệu.*

Tốc độ vượt bậc về dung lượng dữ liệu do con người khởi tạo, lưu giữ và truyền dẫn ngày càng tăng cùng với nhu cầu hoạt động về mọi mặt của đời sống xã hội và sự phát triển của công nghệ chế tạo các thiết bị xử lý, lưu trữ và truyền dẫn dữ liệu đã dẫn đến “Hiện tượng bùng nổ thông tin”.

Các kết quả của sự phát triển công nghệ phần cứng máy tính đã tạo điều kiện thuận lợi cho sự phát triển công nghệ CSDL (liên quan tới tổ chức và quản lý CSDL) và công nghệ mạng (liên quan tới hoạt động truyền dẫn dữ liệu). Công nghệ CSDL đã và đang phát triển không ngừng nhằm đáp ứng nhu cầu quản lý dữ liệu của con người nói chung và trong hoạt động quản lý nói riêng. Trong quá trình tiến hóa của công nghệ CSDL, nhiều hệ quản trị cơ sở dữ liệu được phát triển và năng lực của hệ quản trị cũng ngày càng được nâng cao. Một ví dụ về sự tăng trưởng nổi bật về kích thước CSDL quản lý đó chính là sự xuất hiện nhiều CSDL có kích thước hàng trăm TB (1TB = 1000 GB). Ngoài ra sự phát triển về công nghệ mạng cả về quy mô và tốc độ đã tạo ra sự tăng trưởng mạnh về năng lực truyền dẫn thông tin.

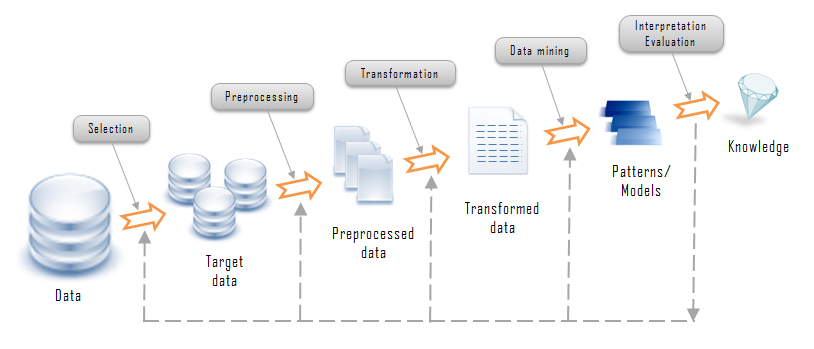
Thích ứng với hoàn cảnh đó, các phương pháp xử lý dữ liệu và các bộ công cụ tiện ích thi hành các phương pháp này ra đời. Xây dựng phát triển phương pháp và công cụ xử lý dữ liệu lớn nhằm mục đích phát hiện tri thức tiềm ẩn là nội dung của lĩnh vực phát hiện tri thức trong cơ sở dữ liệu.



Sơ đồ 1. 1: Tiến hóa của công nghệ cơ sở dữ liệu theo quan điểm của J.Han và M.Kamber (nguồn: Giáo trình khai phá dữ liệu, nhà xuất bản Đại học Quốc gia Hà Nội).

* + 1. *Khai phá dữ liệu.*

Khai phá dữ liệu (Data mining) [1] là quá trình phân loại, sắp xếp và tính toán các tập dữ liệu lớn để xác định các mẫu và thiết lập các mối liên hệ nhằm giải quyết các vấn đề từ việc phân tích dữ liệu từ đó chuyển bộ dữ liệu thành một cấu trúc dễ hiểu để sử dụng tiếp.



Hình 1. 1. Quy trình khai phá dữ liệu (Quy trình khai phá dữ liệu (Process of Data mining, nguồn: [https://bienuit.wordpress.com/khaiphadulieu](https://bienuit.wordpress.com/2013/09/07/quy-trinh-khai-pha-du-lieu-process-of-data-mining/)).

Quá trình khai phá dữ liệu bao gồm:

* Bước 1: Xác định vấn đề, lĩnh vực cần nghiên cứu và dữ liệu liên quan để giải quyết vấn đề (Problem understanding and data understanding).
* Bước 2: Chuẩn bị dữ liệu (Data preparation), bao gồm các quá trình làm sạch dữ liệu (Data cleaning), tích hợp dữ liệu (Data integration), chọn dữ liệu (Data selection), biến đổi dữ liệu (Data transformation).
* Bước 3: Khai phá dữ liệu (Data mining): xác định nhiệm vụ khai thác dữ liệu và lựa chọn kỹ thuật khai thác dữ liệu. Kết quả của bước này cho ta một nguồn tri thức thô.
* Bước 4: Đánh giá (Evaluation): dựa trên một số tiêu chí tiến hành kiểm tra và lọc nguồn tri thức thu được.
* Bước 5: Triển khai (Deployment).
  + 1. *Các lĩnh vực ứng dụng.*

Một số lĩnh vực có ứng dụng mạnh về khai phá dữ liệu bao gồm:

* [Thiên văn học](https://vi.wikipedia.org/wiki/Thi%C3%AAn_v%C4%83n_h%E1%BB%8Dc).
* [Tin sinh học](https://vi.wikipedia.org/wiki/Tin_sinh_h%E1%BB%8Dc).
* Bào chế thuốc.
* [Thương mại điện tử](https://vi.wikipedia.org/wiki/Th%C6%B0%C6%A1ng_m%E1%BA%A1i_%C4%91i%E1%BB%87n_t%E1%BB%AD).
* Phát hiện lừa đảo.
* [Quảng cáo](https://vi.wikipedia.org/wiki/Qu%E1%BA%A3ng_c%C3%A1o).
* Marketing.
* [Quản lý quan hệ khách hàng](https://vi.wikipedia.org/wiki/Qu%E1%BA%A3n_l%C3%BD_quan_h%E1%BB%87_kh%C3%A1ch_h%C3%A0ng).
* Chăm sóc sức khỏe.
* [Viễn thông](https://vi.wikipedia.org/wiki/Vi%E1%BB%85n_th%C3%B4ng).
* [Thể thao](https://vi.wikipedia.org/wiki/Th%E1%BB%83_thao), [giải trí](https://vi.wikipedia.org/wiki/Gi%E1%BA%A3i_tr%C3%AD).
* [Đầu tư](https://vi.wikipedia.org/wiki/%C4%90%E1%BA%A7u_t%C6%B0).
* [Máy tìm kiếm](https://vi.wikipedia.org/wiki/M%C3%A1y_truy_t%C3%ACm_d%E1%BB%AF_li%E1%BB%87u) (Web).
  1. **Các phương pháp khai phá dữ liệu.**
     1. *Khai phá luật kết hợp.*
        1. Giới thiệu về Luật kết hợp.

Khai phá luật kết hợp [2] là tìm ra các mẫu có tần suất cao, các mẫu kết hợp, liên quan hoặc các cấu trúc tồn tại giữa các tập hợp đối tượng trong cơ sở dữ liệu các giao dịch, cơ sở dữ liệu quan hệ hoặc các kho chứa thông tin khác. Nói các khác là chúng ta tìm tất cả các tập phổ biến từ trong dữ liệu.

Bài toán của Luật kết hợp được diễn giải như sau.

Cho biết là tập các giao dịch với là số các giao dịch có trong . Tập là một tập gồm n tập mục khác nhau xuất hiện trong . Mỗi giao dịch là một tập các mục xuất hiện đồng thời. Ta có . Với và là các tập mục.

Một luật kết hợp được biểu diễn bời công thức:

(1)

Trong đó , và

Một giao dịch thuộc chứa một tập mục nếu là tập con của .Độ hỗ trợ của trong là số giao dịch chứa trong (viết tắt là ).Để đo độ mạnh của một luật kết hợp, ta sử dụng hai độ do sau: Độ hỗ trợ và độ tin cậy.

Độ hỗ trợ của một luật là tỉ lệ phần trăm (%) các giao dịch trong mà chứa cả và . Nó giúp xác định mức độ phổ biến của các giao dịch chứa tập mục trong tổng số tất cả các giao dịch. Công thức tính độ hỗ trợ (Support):

(2)

Độ tin cậy của luật là tỉ lệ % các giao dịch trong chứa cả và trên tổng số các giao dịch trong chỉ chứa. Công thức tính độ tin cậy (confidence):

(3)

Những luật kết hợp thỏa mãn điều kiện độ hỗ trợ tối thiểu () và độ tin cậy tối thiểu () được gọi là các luật mạnh.

Nhìn chung, khai phá luật kết hợp là một quá trình gồm 2 bước sau:

* Bước 1: Tìm tất cả các tập mục thường xuyên. Tập mục thường xuyên là tập mục mà độ hỗ trợ của nó lớn hơn hoặc bằng .
* Bước 2: Tạo các luật kết hợp mạnh từ tập mục thường xuyên. Luật kết hợp mạnh là những luật có độ hỗ trợ và độ tin cậy lớn hơn và tương ứng.
  + - 1. Thuật toán Apriori.

Apriori [2] là một thuật toán được giới thiệu lần đầu vào năm 1994 bởi R.Agrawal và R.Srikant, nhằm khai phá tập mục phổ biến nhị phân.

Thuật toán này thực hiện lặp lại việc tìm kiếm theo mức, sử dụng thông tin ở mức để duyệt mức . Thuật toán này được tiến hành như sau:

* Đầu tiên, các tập mục thường xuyên có độ dài là 1 được xây dựng bằng việc duyệt qua toàn bộ dữ liệu để đếm sự xuất hiện của từng phần tử và giá trị này phải lớn hơn hoặc bằng min\_sup. Kết quả của việc đếm này được ký hiệu là .
* Tiếp theo được sử dụng để tìm kiếm tập mục thường xuyên có độ dài là 2. Việc này được thực hiện lặp đi lặp lại đến khi không tìm được tập mục thường xuyên có độ dài là k thỏa mãn điều kiện .
  + - 1. Thuật toán FP-Growth.

Thuật toán FP-Growth [2] được giới thiệu năm 2000 bời Jiawei Hai Jian Pei và Yiwen Yin nhằm khắc phục hai nhược điểm của Thuật toán Apriori đó là:

* Chi phí lớn cho một số lượng lớn các tập ứng cử.
* Đòi hỏi lặp lại nhiều lần duyệt CSDL, để kiểm tra tất cả các tập ứng cử.

Thuật toán này được tiến hành như sau:

* Duyệt CSDL lần thứ nhất để tính độ hỗ trợ của tất cả 1-itemset và loại bỏ những mục có độ hỗ trợ bé hơn min\_sup. Các mục còn lại được sắp xếp theo thứ tự giảm dần độ hỗ trợ (cũng tức là giảm dần theo số lần xuất hiện trong CSDL), sau đó ta nhận được danh sách L đã sắp xếp.
* Duyệt CSDL lần hai, với mỗi tác vụ t, loại bỏ các mục không đủ độ hỗ trợ, các mục còn lại theo thứ tự giảm dần độ hỗ trợ được đưa vào cây FP-tree.
* Tìm các tập mục phổ biến trên cây FP-tree đã xây dựng mà không duyện lại CSDL nữa.
  + - 1. Ứng dụng Khai phá luật kết hợp.

Khai phá luật kết hợp có một số ứng dụng như: phân tích luật kết hợp để hỗ trợ kinh doanh, tìm hiểu thói quen mua sắm của khách hàng, khả năng tìm ra những dấu hiệu để tìm vi phạm bản quyền tài liệu, gợi ý và hỗ trợ chuẩn đoán bệnh tật và hỗ trợ ra quyết định trong chứng khoáng. Một số kĩ thuật mới như luật kết hợp hiếm và luật kết hợp âm đã được phát triển nhằm tăng khả năng và mở rộng phạm vị ứng dụng của khai phá luật kết hợp.

* + 1. *Phân cụm (Clustering).*

Phân cụm là [2] việc thực hiện nhóm dữ liệu thành các “cụm” để có thể phát hiện được các mẫu phân bố dữ liệu trong miền ứng dụng. Phân cụm là một phương pháp học máy không giám sát, cố gắng khám phá các nhóm tự nhiên và phân phối thống kê dữ liệu.

* + - 1. Phân cụm phẳng.

Giải thuật k-means thuộc lớp phân cụm phẳng, đầu vào cho thuật toán k-means là tập dữ liệu D gồm n phần tử dữ liệu, số lượng các cụm đầu ra k. Đầu ra của giải thuật là k cụm dữ liệu.

Giải thuật k-means được trình bày như sau:

1. Chọn ngẫu nhiên k phần tử trong tập D làm trọng tâm ban đầu cho các cụm.

2. Phân các phần tử dữ liệu trong D vào các cụm dựa vào độ tương đồng của nó với trọng tâm của các cụm. Phần tử dữ liệu sẽ được phân vào cụm có độ tương đồng lớn nhất.

3. Tính lại trọng tâm của các cụm.

4. Nhảy đến bước 2 cho đến khi quá trình hội tụ (không có sự gán lại các phần tử dữ liệu giữa các cụm, hay trọng tâm của các cụm là không đổi).

* + - 1. Phân cụm phân cấp.

Khác với các giải thuật phân cụm phẳng, các thuật toán phân cụm phân cấp sẽ tạo ra một cây phân cấp các cụm dữ liệu. Các thuật toán phân cụm phân cấp bao gồm: Thuật toán phân cụm gộp, thuật toán phân cấp BIRCH, thuật toán phân cụm DIANA và thuật toán phân cụm ROCK. Ngoài phân cụm phẳng và phân cụm phân cấp, còn một số phương pháp như: Phân cụm dựa vào mật độ, dựa trên lưới, dựa trên mô hình, phân cụm đơn đinh, phân cụm dữ liệu có chiều lớn, phân cụm dựa trên ràng buộc, phân cụm theo lô và phân cụm gia tăng. Phân cụm thuộc khai phá dữ liệu mô tả.

* + 1. *Phân lớp (Classification).*

Phân lớp [2] là phương pháp dự báo, cho phép phân loại một đối tượng vào một hoặc một số lớp cho trước. Phân lớp là một hình thức phân tích dữ liệu phổ biến được dung để tạo ra các mô hình mô tả các lớp dữ liệu quan trọng để giúp hiểu rõ hơn về dữ liệu ở mức cao hơn, bao quát hơn ngoài ra còn được sử dụng để dự đoán các nhãn lớp cho dự liệu đầu vào tương ứng. Đặc trưng của mô hình phân lớp là dựa trên các tập dữ liệu có sẵn, các tập dữ liệu này sẽ được chuyển đổi thành các vector với số chiều là các trường trong đó có 1 trường là nhãn lớp.

Dựa trên tập dữ liệu vào dạng vector này, các thuật toán đặc thù sẽ được triển khai để tiến hành thiết lập một ánh xạ . Trong đó là tập các nhãn và là vector với các chiều, chính là mô hình thu được thông qua quá trình học.

Các mô hình này được xây dựng với mục tiêu hướng đến là khả năng dự báo có độ chính xác cao nhất, tin cậy nhất, hoặc nằm trong khoảng chấp nhận được. Do đó, dựa trên tập dữ liệu đầu vào, người ta rút bớt một phần để làm tập dữ liệu kiểm thử. Các mô hình được xây dựng trên phần dữ liệu còn lại (bộ dữ liệu huấn luyện) và sẽ được kiểm thử hiệu năng trên tập dữ liêu kiểm thử.

* + - 1. Mô hình phân lớp Naïve Bayes.

Phân lớp Naïve Bayes [3] được dựa trên định lý Bayes của tác giả Thomas Bayes, là một giải thuật phân lớp thống kê, nó có thể dự đoán xác suất của một phần tử dữ liệu thuộc một lớp.

Lý thuyết Bayes:

* Cho “” là một bộ dữ liệu được đo trên n thuộc tính khác nhau.
* Cho “” là một giả thuyết nào để cho “” thuộc một phân lớp “” nào đó.
* Trong bài toán phân lớp này, để xác định giá trị – là xác suất xảy ra “” khi “” đã xảy ra. là xác suất hậu nghiệm.

Ví dụ: Giả sử tập dữ liệu sinh viên được mô tả bởi các thuộc tính giới tính và sở thích. “” là giả thuyết sinh viên tham gia câu lạc bộ bóng đá. Khi đó ) biểu đạt xác suất sinh viên “” sẽ tham gia câu lạc bộ bóng đá khi biết giới tính và sở thích của sinh viên. Ngược lại được gọi là xác suất tiên nghiệm. Dựa vào đinh lý Bayes ta có:

(4)

Ưu điểm:

* Dễ sử dụng.
* Nhanh khi chuẩn đoán dữ liệu test set.
* Khi các đặc tính của dữ liệu độc lập với nhau thì Naïve bayes chạy tốt hơn so với các bộ phân lớp còn lại và cũng cần ít dữ liệu hơn.

Nhược điểm:

* Độ chính xác không được cao.
* Trong thực thế, hầu như bất khả thi khi các đặc tính của dữ liệu test độc lập với nhau.

Nguyên tắc hoạt động của mô hình phân lớp Naïve Bayes.

Gọi “” là tập dữ liệu huấn luyện với các nhãn lớp tương ứng. Ta có, mỗi bộ dữ liệu được mô tả bởi “” thuộc tính và được diễn đạt dưới dạng vector “” chiều .

Giả sử có “” nhãn lớp khác nhau gồm . Cho một bộ dữ liệu “”, bộ phân lớp sẽ dự đoán “” thuộc về phân lớp có xác xuất hậu nghiệm cao nhất.

với

(5)

Do không đổi nên ta chỉ cần cực đại hóa giá trị .

* + - 1. Mô hình phân lớp K-NN.

Phương pháp K-lân cận [4] được đề xuất từ những năm 1950. Phương pháp này dựa chủ yếu vào các phần tử lân cận trong tập dữ liệu huấn luyện.

K-NN là phương pháp để phân lớp các đối tượng dựa vào khoảng cách gần nhất giữa đối tượng cần xếp lớp và tất cả các đối tượng trong Training Data.

Một tập dữ liệu được phân lớp dựa vào K láng giềng của nó. K là một số lẻ nguyên dương được ta xác định trước khi tiến hành thuật toán.

Ưu điểm:

* Hỗ trợ xử lý dữ liệu một cách gia tăng.
* Không tốn thời gian để học.

Nhược điểm:

* Tính toán nhiều trong quá trình phân lớp.
* Chi phí phân lớp sẽ càng cao khi tập dữ liệu huấn luyện càng lớn.

Dưới đây là nguyên tắc hoạt động của mô hình phân lớp K-NN.

Xét một bộ dữ liệu chưa được gán nhãn (mỗi phần tử dữ liệu đươc xem như là 1 điểm trong không gian n chiều). Khi đó, bộ phân lớp K-NN sẽ tìm kiếm trong không gian những phần tử dữ liệu nào gần nhất với phần tử dữ liệu hiện xét.

Dựa vào các nhãn lớp của phần tử dữ liệu láng giềng, K-NN sẽ quy định các nhãn lớp của tập dữ liệu mới này thuộc lớp nào.

Những bộ dữ liệu k-lân cận này được xét dựa trên phép đo Euclide: (6)

Nhãn lớp phổ biến trong số k lân cận sẽ là nhãn lớp của bộ dữ liệu .

* + - 1. Mô hình phân lớp Support Vector Machines (SVM).

Thuật toán SVM [5] là một thuật toán giám sát, nó có thể sử dụng cho cả việc phân loại hoặc đệ quy nhưng thường chủ yếu sử dụng cho phân loại. Nó có khả năng xử lý cả dữ liệu tuyến tính và dữ liệu không tuyến tính. Bản chất của giải thuật này là nó xây dựng một siêu phẳng để phân chia dữ liệu thành 2 nửa. Trong trường hợp nếu dữ liệu là không tuyến tính thì nó sẽ sử dụng một hàm nhân (kernel function) để chuyển đổi tập dữ liệu ban đầu sang một không gian mới có số chiều lớn hơn để xử lý.

Phương pháp này có nhược điểm là thời gian huấn luyện tương đối lâu, nhưng bù lại kết quả cho độ chính xác khá cao.

Dưới đây là nguyên tắc hoạt động của SVM.

Cho tập dữ liệu “” như sau: với là tập dữ liệu huấn luyện gắn liền với các nhãn lớp tương ứng yi (nhận một trong hai giá trị là +1 hay -1.

SVM tiếp cận bài toán bằng cách tìm kiếm siêu phẳng viền tối đa. Một siêu phẳng phân tách được diễn đạt bằng ngôn ngữ toán học như sau:

(7)

Trong đó: “” là vector trọng số với , “” là số các thuộc tính và “” là một đại lượng vô hướng.

 Dựa vào phương trình toán học này, các tập dữ liệu được phân tách bằng cách xác định dấu (+/-) khi áp dụng với giá trị cụ thẻ của từng bộ dữ liệu.

(8)

Hay (9)

Tương ứng với hai nhãn lớp được xác định của bài toán phân lớp. Các trọng số có thể được điều chỉnh để xác định “phía” của siêu phẳng mà bộ dữ liệu rơi vào.

với (10)

Và -1 với (11)

Bất cứ bộ dữ liệu huấn luyện nào rơi vào phía trên siêu phẳng “” hoặc “” được gọi là support vector.

Đối với các trường hợp dữ liệu phi tuyến, chúng ta không thể vẽ đường thằng phân tách giữa các tập dữ liệu, SVM cần tìm ra những siêu phẳng phi tuyến để phân tách những tập dữ liệu phi tuyến tương ứng.

* + - 1. Mô hình phân lớp Mạng Nơ-ron nhân tạo.

Mạng nơ-ron nhân tạo [6] là mô hình xử lý thông tin được mô phỏng dựa trên hoạt động của hệ thống thần kinh của sinh vật, bao gồm số lượng lớn các nơ-ron được gắn kết để xử lý thông tin.

Mạng nơ-ron là tập các đơn vị được kết nối với nhau từ đầu với đến đầu ra và mỗi nút trong đó có một trọng số riêng.

Trong quá trình học, mạng nơ-ron điều chỉnh các trọng số của các nút trong nó để có kết quả dự đoán phù hợp với kết quả thực tế của bộ dữ liệu huấn luyện.

Ưu điểm:

* Khả năng chịu nhiễu cao cùng với khả năng phân lớp đối với những tập dữ liệu chưa từng được huấn luyện trước đó.
* Kết quả huấn luyện với mô hình phân lớp mạng nơ-ron có độ chính xác khá cao.

Nhược điểm:

* Thời gian huấn luyện thường rất lâu.
* Thường khó giải thích cách thức đưa ra quyết định của mạng nơ-ron.

Với các ưu và nhược điểm đó, mạng nơ-ron thường được sử dụng khi ta có kiến thức về mối quan hệ giữa các thuộc tính nhãn lớp.

Mạng nơ-ron đặc biệt phù hợp với dữ liệu đầu vào và đầu ra các giá trị liên tục.

Cấu trúc mạng nơ-ron bao gồm:

* Gồm nhiều tầng, mỗi tầng gồm nhiều nút.
* Ở tầng đầu vào, số nút tương ứng với số thuộc tính của tập dữ liệu.

Ở tầng ẩn giấu, dữ liệu đầu ra của tầng ẩn giấu này là dữ liệu đầu vào của tầng ẩn giấu kế tiếp. Kết quả của tầng ẩn giấu cuối cùng là đầu vào cho tầng kết quả (đầu ra). Số các tầng ẩn giấu được xác định một cách chủ quan, thông thường là 1.

Nguyên tắc hoạt động của Mạng nơ-ron được trình bày như sau.

Trước khi huấn luyện phải xác định rõ ràng các topo mạng. Các thông số định nghĩa topo của mạng nơ-ron bao gồm: số nút ở tầng vào (input), số tầng ẩn giấu, số nút ở mỗi tầng ẩn giấu, số nút ở tầng kết quả (output).

Các giá trị của mỗi thuộc tính thường sẽ được chuẩn hóa để đảm bảo nằm trong khoảng 0.0 đến 1.0, đối với các giá trị rời rạc thì có thể được mã hóa cho phù hơp.

(12)

Các hàm kích hoạt:

1. (13)
2. (14)
3. (15)

* + 1. *Hồi quy (Regression).*
       1. Khái niệm.

Hồi quy là kỹ thuật thống kê cho phép dự đoán các trị số liên tục. Theo Wikipedia, phân tích hồi quy [7] là một [phân tích thống kê](https://vi.wikipedia.org/w/index.php?title=Ph%C3%A2n_t%C3%ADch_th%E1%BB%91ng_k%C3%AA&action=edit&redlink=1) để xác định xem các [biến độc lập](https://vi.wikipedia.org/w/index.php?title=Bi%E1%BA%BFn_%C4%91%E1%BB%99c_l%E1%BA%ADp&action=edit&redlink=1) (biến thuyết minh) quy định các [biến phụ thuộc](https://vi.wikipedia.org/w/index.php?title=Bi%E1%BA%BFn_ph%E1%BB%A5_thu%E1%BB%99c&action=edit&redlink=1) (biến được thuyết minh) như thế nào.

Mô hình hổi quy khá thông dụng trong dự báo dài hạn. Trong khai phá dữ liệu, bài toán hồi quy được quy về việc học một hàm ánh xạ dữ liệu nhằm xác định giá trị thực của một biến theo một biến khác.

* + - 1. Phân loại.

Hồi qui tuyến tính và phi tuyến.

Hồi qui đơn biến và đa biến.

Hồi qui có thông số, phi thông số và thông số kết hợp.

Hồi qui đối xứng và bất đối xứng.

* 1. **Phát biểu bài toán nghiên cứu.**

Gọi là tập dữ liệu mô tả website. Mỗi website xi được mô tả bởi 30 thuộc tính. Hay nói các khác xi được diễn đạt như một vector trong không gian 30 chiều.

và là tập các nhãn lớp được áp dụng cho từng website với Phishing – lừa đảo và Legitimate – hợp pháp.

Bài toán phát hiện website lừa đảo được phát biểu như sau:

Gọi f là hàm phát hiện website lừa đảo

Xây dựng hàm (16)

Với

* 1. **Lý thuyết về mô hình phân lớp Random Forest.**
     1. *Lịch sử.*

Phương pháp chung của Random Forest [8] được tạo ra bởi [Tin Kam Ho](https://en.wikipedia.org/wiki/Tin_Kam_Ho) vào năm 1995 bằng các sử dụng phương pháp không gian con ngẫu nhiên (Random Subspace Method).

* + 1. *Cơ sở lý thuyết.*
       1. Khái niệm.

Như tên gọi của nó, Random Forest dựa trên cơ sở:

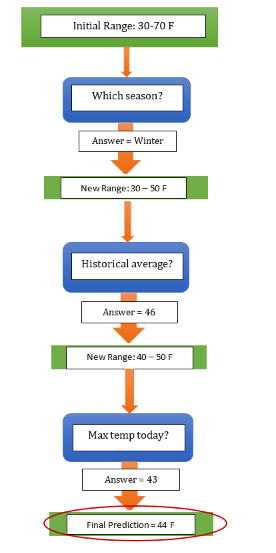
1. Random = Tính ngẫu nhiên.
2. Forest = Rừng, nhiều cây quyết định.

Random Forest là một thành viên trong họ thuật toán Decision Tree (DT), tư tưởng của RF là xây dựng nhiều cây quyết định, mỗi cây quyết định dự đoán một kết quả và kết quả nào được cây quyết định dự đoán nhiều nhất thì nó là kết quả cuối cùng. Để chắc chắn rằng tất cả các DT đều không ra một câu trả lời (nếu xây dựng cây quyết định theo đúng một cách thì nó sẽ cho ra 1 kết quả) nên trong quá trình xây dựng, RF ngẫu nhiên quan sát và chọn ngẫu nhiên các thuộc tính.

Theo Wikipedia, Random Forest [8] là một phương pháp học tập để phân loại, hồi quy và các nhiệm vụ khác bằng cách xây dựng vô số cây quyết định tại thời điểm đào tạo và đưa ra các nhãn lớp (Phân loại) hoặc dự đoán trung bình (Hồi quy) của các cây riêng lẻ.

* + - 1. Cây quyết định.

Đơn vị cơ bản của một khu rừng đó chính là cây. Tương tự như vậy, đơn vị cơ bản nhất của RF là những cây quyết định. Cây quyết định được định nghĩa là một cấu trúc luồng dạng cây với mỗi nút trung gian (không phải nút lá) đại diện cho mỗi phép thử trên một thuộc tính. Mỗi nhánh đại diện cho mỗi đầu ra (kết quả) của phép thử. Và mỗi nút lá (nút cuối cùng) đại diện cho mỗi nhãn lớp. Nút trên cùng của cây được gọi là nút gốc. Bằng cách đi theo các giá trị thuộc tính trên cây, ta sẽ biết giá trị dự đoán. Một đường đi từ nút gốc đến một nút lá biểu đạt dự đoán phân lớp cho một bộ dữ liệu tương ứng. Cây quyết định có thể dễ dàng chuyển đổi thành các luật phân lớp.



Sơ đồ 1. 2. Quy trình quyết định dự đoán nhiệt độ - Cây quyết định cơ bản (Giải thích rừng ngẫu nhiên đơn giản, nguồn: [https://medium.com/random-forest](mailto:%20https://medium.com/@williamkoehrsen/random-forest-simple-explanation-377895a60d2d)).

* + - 1. Random Forest.

Random Forest (RF) là một thuật toán học máy có giám sát. Ý tưởng cơ bản của RF là kết hợp nhiều cây quyết định thành một mô hình duy nhất. Một cây quyết định khi dự đoán có thể sẽ không chính xác nhưng khi kết hợp chúng lại với nhau sẽ cho ra một kết quả trung bình.

Ví dụ: Bạn đang muốn đi du lịch vào kì nghỉ hè sắp đến, nhưng bạn đang không biết là phải đi đâu? Đầu tiên bạn sẽ hỏi một người bạn của mình, người đó sẽ nói về những nơi cậu ấy đã đi và cậu ấy có thích hay không? Dựa trên những câu trả lời và cảm nhận của mình, cậu ấy sẽ cho bạn một lời khuyên. Sau đó, bạn bắt đầu hỏi càng nhiều bạn bè và họ lại cho bạn nhiều câu trả lời khác nhau từ đó bạn chọn những nơi được nhiều lời khuyên nhất để ra quyết định cho mình. Đây là ví dụ điển hình của rừng ngẫu nhiên.

Thuật toán của RF áp dụng kỹ thuật [Bootstrap Aggregating](https://en.wikipedia.org/wiki/Bootstrap_aggregating) hay còn gọi là Bagging, nó là một thuật toán học máy được thiết kế để cải thiện độ chính xác và ổn định của các thuật toán học máy được sử dụng trong phân lớp và hồi quy.

Ưu điểm:

* RF là một thuật toán đơn giản.
* Có thể sử dụng cho bài toán Phân lớp và Hồi quy.
* Có thể làm việc được với các dữ liệu thiếu giá trị.
* Tránh được trường hợp Overfitting tập dữ liệu.
* Có thể tạo mô hình cho các giá trị phân loại.

Nhược điểm:

* Khối lượng tính toán lớn nhưng với công nghệ ngày càng phát triển thì điều này ngày càng được khắc phục.
* Một số lượng lớn cây có thể làm cho thuật toán bị chậm và không hiệu quả đối với các dự đoán theo thời gian thực.
  + 1. *Nguyên tắc hoạt động.*

Random Forest hoạt động bằng cách đánh giá nhiều Cây quyết định ngẫu nhiên, và lấy ra kết quả được đánh giá tốt nhất trong số kết quả trả về.

Mã giải cho hoạt động của RF [9]:

* Bước 1: Chọn ngẫu nhiên “” thuộc tính từ tập “” thuộc tính. Để .
* Bước 2: Từ tập “” thuộc tính, tính toán ra nút “” là tốt nhất cho nút phân loại.
* Bước 3: Chia các nút con theo nút tốt nhất vừa tìm được.
* Bước 4: Lặp lại bước 1-3 cho đến khi đạt đến “” nút.
* Bước 5: Lặp lại bước 1-4 để tạo ra “” cây.

Sau các bước trên chúng ta có một Random Forest.

Các bước dự đoán sử dụng Random Forest đã huấn luyện:

* Bước 1: Lấy các thuộc tính kiểm thử và sử dụng các Cây quyết định đã tạo ra để dự đoán kết quả, lưu nó vào một danh sách.
* Bước 2: Tính toán số lượng trên toàn bộ rừng cho từng kết quả.
* Bước 3: Lấy kết quả có số lượng lớn nhất làm kết quả cuối cho mô hình.
  + 1. *Ứng dụng.*

Thuật toán Random Forest được sử dụng trong rất nhiều lĩnh vực khác nhau như Ngân hàng, Chứng khoán, Y học và Thương mại điện tử.

Trong Ngân hàng: nó được sử dụng để phát hiện những khách hàng sẽ sử dụng dịch vụ của ngân hàng thường xuyên hơn và trả nợ đúng hạn. Nó cũng còn được sử dụng để phát hiện những khách hàng lừa đảo và muốn lừa đảo.

Trong Chứng khoán, nó được sử dụng để xác định hành vi của một cổ phiếu trong tương lai, giúp nhà đầu tư định hướng được khả năng lên xuống của cổ phiếu và có những quyết định đầu tư đúng đắn.

Trong Y học, nó được sử dụng để xác định sự kết hợp chính xác của các thành phần trong y học và để phân tích lịch sử y tế của bệnh nhân để xác định bệnh.

Và cuối cùng, trong Thương mại điện tử, nó được sử dụng để xác định xem khách hàng có thực sự thích sản phẩm đó hay không.

* 1. **Các công trình liên quan.**

Ozgur Koray Sahingoz và công sự (2019) [10] đã giới thiệu một phương pháp phát hiện lừa đảo từ URL dựa trên học máy, một hệ thống chống lừa đảo thời gian thực, sử dụng bảy thuật toán phân loại khác nhau và các tính năng dựa trên xử lý ngôn ngữ tự nhiên (Natural Language Processing). Hệ thống này có các đặc tính phân biệt sau đây với các nghiên cứu khác trong tài liệu: độc lập ngôn ngữ, sử dụng một lượng lớn dữ liệu lừa đảo và dữ liệu hợp pháp, thực thi thời gian thực, phát hiện các trang web mới, độc lập với các dịch vụ của bên thứ ba và sử dụng các phân loại giàu tính năng. Để đo hiệu suất của hệ thống, một bộ dữ liệu mới được xây dựng và kết quả thử nghiệm được thử nghiệm trên hệ thống. Theo kết quả thử nghiệm và so sánh từ các thuật toán phân lớp đã triển khai, thuật toán RF có các tính năng dựa trên NLP cho hiệu suất tốt nhất với tỷ lệ chính xác 97,98% để phát hiện các URL lừa đảo.

Routhu Srinivasa Rao và Alwyn Roshan Pais (2018) [11] đã giới thiệu một phương pháp phát hiện lừa đảo trang web dựa trên tính năng hiện quả của học máy. Trong bài báo này, ông đã đề xuất một mô hình phân loại mới, dựa trên các tính năng heuristic được trích xuất từ ​​URL, mã nguồn và dịch vụ của bên thứ ba để khắc phục nhược điểm của các kỹ thuật chống lừa đảo hiện có. Mô hình đã được đánh giá bằng tám thuật toán học máy khác nhau và trong đó, thuật toán RF hoạt động tốt nhất với độ chính xác 99.31%. Các thí nghiệm được lặp lại với các phân loại rừng ngẫu nhiên khác nhau để tìm ra phân loại tốt nhất để phát hiện trang web lừa đảo. Phân loại PCA-RF đã thực hiện tốt nhất trong số tất cả các oRFs với độ chính xác 99.55%. Kết quả được so sánh với các mô hình cơ sở cho thấy kết quả của chúng tôi vượt trội so với phương pháp này.

Shinelle Hutchinson và cộng sự (2018) [10] đã giới thiệu phương pháp phát hiện Website lừa đảo dựa trên mô hình phân lớp Random Forest. Trong bài báo này, họ đã phân tích phát hiện lừa đảo dựa trên web bằng cách sử dụng RF bằng một số tính năng URL quan trọng, các tính năng này được xác định và nghiên cứu cho thấy hiệu quả và hiệu suất phát hiện được cải thiện.

S.Jagadeesan (2018) [13] đã giới thiệu phương pháp phân tích lừa đảo URL bằng cách sử dụng Random Forest. Trong phương pháp của Jagadeesan, ông chỉ sử dụng thông tin về URL của trang web để xác định xem trang web đó có phải là trang web lừa đảo hay hợp pháp. Do đó, không cần thực sự truy cập một trang web để xác định xem nó có lừa đảo hay không. Điều này cũng cho phép người dùng không truy cập các trang web lừa đảo và tiếp xúc với các mã độc hại mà nó có thể mang theo. Ngoài ra, Jagadeesan thảo luận về cách dữ liệu meta của các URL có thể được sử dụng để xác định xem URL như thế nào. Các thuật toán RF và SVM sau đó có thể được áp dụng cho một tập dữ liệu có các tính năng như vậy có chứa dữ liệu meta của các URL. Thuật toán RF cung cấp lợi thế của việc không cung cấp quá nhiều dữ liệu.

Ebubekir Buber, Banu Diri và Ozgur Koray Sahingoz (2017) [14] đã giới thiệu phương pháp hiện các cuộc tấn công lừa đảo từ URL bằng cách sử dụng các kỹ thuật NLP. Trong nghiên cứu này, trước tiên, các đặc điểm của loại tấn công này được giải thích, và sau đó một hệ thống dựa trên máy học được đề xuất để phát hiện ra chúng. Trong hệ thống được đề xuất, một số tính năng đã được trích xuất bằng cách sử dụng các kỹ thuật NLP. Hệ thống được triển khai bằng cách kiểm tra các URL được sử dụng trong Tấn công lừa đảo trước khi mở chúng bằng cách sử dụng một số tính năng được trích xuất. Nhiều thử nghiệm đã được áp dụng cho hệ thống được tạo và có thể thấy rằng thuật toán tốt nhất trong số các thử nghiệm là thuật toán RF với tỷ lệ thành công là 89.90%.

Shraddha Parekh và cộng sự (2018) [15] đã giới thiệu một phương pháp mới để phát hiện trang web lừa đảo bằng URL. Lừa đảo là một hoạt động bất hợp pháp khiến mọi người bị lừa vào các trang web sai bằng cách sử dụng các phương thức lừa đảo khác nhau. Mục đích của các trang web lừa đảo này là thu thập thông tin cá nhân hoặc các thông tin tài chính khác vì lợi ích cá nhân hoặc lạm dụng. Khi công nghệ tiến bộ, các phương pháp lừa đảo được sử dụng cần phải được tiến bộ và cần có sự bảo mật tốt hơn và các cơ chế tốt hơn để ngăn chặn cũng như phát hiện các phương thức lừa đảo này. Trọng tâm chính của bài viết này là đưa ra một mô hình như một giải pháp để phát hiện các trang web lừa đảo bằng cách sử dụng phương pháp phát hiện URL bằng thuật toán RF.

Abdulhamit Subasi và cộng sự (2017) [16] đã giới thiệu phương pháp phát hiện lừa đảo thông minh bằng cách sử dụng phân loại Random Forest. Trang web lừa đảo được định nghĩa là bắt chước một trang web của công ty có uy tín, nhằm mục đích lấy thông tin cá nhân của người dùng. Đã có nhiều giải pháp khác nhau được để xuất để loại bỏ việc lừa đảo. Tuy nhiên, không thể loại bỏ hoàn toàn mối đe dọa này. Khai phá dữ liệu là một kỹ thuật đầy hứa hẹn được sử dụng để phát hiện các cuộc tấn công lừa đảo. Trong bài báo này, một hệ thống thông minh để phát hiện các cuộc tấn công lừa đảo được trình bày. Abdulhamit Subasi và cộng sự đã sử dụng các kỹ thuật khai thác dữ liệu khác nhau để quyết định các loại trang web: hợp pháp hoặc lừa đảo. Các phân loại khác nhau đã được sử dụng để xây dựng hệ thống thông minh chính xác để phát hiện trang web lừa đảo. Độ chính xác phân loại, diện tích dưới các đường cong ROC và F-measure được sử dụng để đánh giá hiệu suất của các kỹ thuật khai phá dữ liệu. Kết quả cho thấy RF đã vượt trội nhất trong số các phương pháp phân loại bằng cách đạt được độ chính xác cao nhất 97.36%. Thời gian RF khá nhanh và có thể xử lý các trang web khác nhau để phát hiện lừa đảo.

Lee Jin Lee, Doo Ho Park và Chang Hoon Lee (2015) [17] đã giới thiệu phương pháp phát hiện lừa đảo bằng cách sử dụng trang Web Heuristic. Trong năm gần đây, các cuộc tấn công lừa đảo đang tràn ngập các dịch vụ dựa trên công nghệ Web. Lừa đảo đang ảnh hưởng đáng kể đến bảo mật trực tuyến từng ngày với lỗ hổng của các trang Web. Để ngăn chặn các cuộc tấn công lừa đảo, rất nhiều kỹ thuật chống lừa đảo đã được thực hiện với lợi thế riêng nhưng cuộc tấn công lừa đảo vẫn chưa bị xóa bỏ hoàn toàn. Trong bài báo này, họ đã nghiên cứu chi tiết về lừa đảo và phân loại một quá trình tấn công lừa đảo theo hai phần - Giai đoạn hạ cánh (Landing-phase), Giai đoạn tấn công (Attack-phase). Ngoài ra, họ đề xuất một phương pháp phát hiện lừa đảo dựa trên các trang Web Heuristic. Phương pháp này cho thấy độ chính xác phát hiện là 98.90% với thuật toán RF.

Kang Leng Chiew và cộng sự (2019) [18] đã giới thiệu một khung lựa chọn tính năng kết hợp mới cho hệ thống phát hiện lừa đảo dựa trên học máy. Bài viết này đề xuất một tính năng mới cho hệ thống phát hiện lừa đảo dựa trên học máy, được gọi là Hybrid Ensemble Feature Selection (HEFS). Trong giai đoạn đầu tiên của HEFS, thuật toán Cumulative Distribution Function gradient (CDF-g) mới được khai thác để tạo ra các tập hợp tính năng chính, sau đó được đưa vào tập hợp nhiễu dữ liệu để tạo ra các tập hợp tính năng thứ cấp. Giai đoạn thứ hai xuất phát một tập hợp các tính năng cơ bản từ các tập hợp tính năng thứ cấp bằng cách sử dụng tập hợp nhiễu loạn chức năng. Các kết quả thí nghiệm tổng thể cho thấy HEFS hoạt động tốt nhất khi được tích hợp với mô hình phân lớp RF, trong đó các tính năng cơ bản phân biệt chính xác 94.6% trang web lừa đảo và hợp pháp chỉ sử dụng 20.8% các tính năng ban đầu. Trong một thử nghiệm khác, các tính năng cơ bản (tổng cộng 10) được sử dụng trên Random Forest vượt trội hơn so với tập hợp tất cả các tính năng (tổng cộng 48) được sử dụng trên các phân loại SVM, Naive Bayes, C4.5, JRip và Part. HEFS cũng cho thấy kết quả đầy hứa hẹn khi được điểm chuẩn bằng cách sử dụng một bộ dữ liệu lừa đảo nổi tiếng khác từ kho lưu trữ của UCI. Do đó, HEFS là một kỹ thuật lựa chọn tính năng thực tế và rất phù hợp cho các hệ thống phát hiện lừa đảo dựa trên học máy.

Ebubekir Buber, Banu Diri và Ozgur Koray Sahingoz (2018) [19] đã giới thiệu phương pháp phát hiện các cuộc tấn công lừa đảo dựa vào NLP từ URL. Trong bài báo này, một hệ thống phát hiện lừa đảo có thể phát hiện loại tấn công này bằng cách sử dụng một số thuật toán học máy và phát hiện một số điểm tương đồng trực quan với sự trợ giúp của một số kỹ thuật NLP. Nhiều thử nghiệm đã được áp dụng trên hệ thống được đề xuất và kết quả thử nghiệm cho thấy thuật toán Rừng ngẫu nhiên có hiệu suất rất tốt với tỷ lệ thành công là 97.2%.

Martyn Weedon, Dimitris Tsaptsinos và James Denholm-Price (2017) [20] đã giới thiệu một phương pháp khai khá rừng ngẫu nhiên để phân loại URL. Mục tiêu của báo cáo này là đánh giá hiệu năng của thuật toán RF bằng cách sử dụng bộ dữ liệu Lexical Only Dataset. Hiệu suất được kiếm chứng so với các thuật toán học máy khác và bổ sung so với các thuật toán được báo cáo trong tài liệu. Kết quả ban đầu từ các thí nghiệm cho thấy thuật toán RF thực hiện tốt nhất với độ chính xác 86.9%.

Mazharul Islam và Nihad Karim Chowdhury (2016) [21] đã giới thiệu phương pháp phát hiện các trang. Báo cáo này đã áp dụng các loại thuật toán phân loại dựa trên máy học khác nhau, bao gồm Naïve Bayes, [Support Vector Machine](https://viblo.asia/p/support-vector-machine-trong-hoc-may-mot-cai-nhin-don-gian-hon-XQZkxoQmewA), Nearest Neighbour, RF, phân loại IBK và J48. Cuối cùng, so sánh hiệu suất của bộ phân loại về độ chính xác. Kết quả cho thấy RF có độ chính xác cao nhất với 97.47%

Sami Smadi và cộng sự (2015) [22] đã giới thiệu một phương pháp phát hiện các email lừa đảo bằng các thuật toán khai phá dữ liệu. Bài viết này đề xuất một mô hình thông minh để phát hiện các email lừa đảo phụ thuộc vào giai đoạn tiền xử lý trích xuất một tập hợp các tính năng liên quan đến các phần email khác nhau. Các tính năng được trích xuất được phân loại bằng thuật toán phân loại J48. Chúng tôi đã thử nghiệm với tổng số 23 tính năng đã được sử dụng trong tài liệu. Ten-fold cross-validation đã được áp dụng cho đào tạo, kiểm tra và xác nhận. Trọng tâm chính của bài viết này là nâng cao các giá trị số liệu tổng thể của phân loại email bằng cách tập trung vào giai đoạn tiền xử lý và xác định thuật toán tốt nhất có thể được sử dụng trong lĩnh vực này. Kết quả cho thấy mô hình đạt được độ chính xác 98.87% cho thuật toán RF.

Rabail Kazi (2016) [23] đã giới thiệu môt hệ thống đánh giá các phân lớp học máy để phát hiện lừa đảo. Hệ thống phát hiện lừa đảo được sử dụng trong bài viết này là một bộ lọc trang web dựa trên Simple Logistic Heuristic là một thuật toán học máy. Weka là một công cụ được sử dụng để thực hiện các thuật toán học máy. Trong báo cáo này, một số phân lớp có trong Weka được kiểm tra dựa trên tập dữ liệu cố định. Mục đích là để kiểm tra các phân loại học máy để phát hiện lừa đảo. Kết quả thí nghiệm được trình bày chứng minh rằng RF vượt trội hơn tất cả các phân loại khác với độ chính xác 93%. Độ chính xác được cải thiện hơn nữa cho RF bằng cách sử dụng trình phân lớp Auto-WEKA. Trình phân lớp này có thể phát hiện tới 99% trang web lừa đảo, với tỷ lệ sai chỉ 1%. Do đó, độ chính xác của hệ thống phát hiện lừa đảo có thể được cải thiện bằng cách sử dụng trình phân lớp RF và Auto-WEKA.

Yang Yaokai (2019) [24] đã giới thiệu phát hiện lừa đảo bằng các phương pháp của học máy. Luận án này đề xuất một hệ thống phát hiện URL lừa đảo linh hoạt và hiệu quả với bộ tính năng phong phú phản ánh các đặc điểm đa dạng của các trang web lừa đảo và nền tảng lưu trữ của chúng, bao gồm các tính năng khó bị giả mạo. Sử dụng thuật toán RF, hệ thống này có khả năng phát hiện cao và tỷ lệ lỗi thấp.

Mohammed Zakariah (2014) [25] đã giới thiệu phương pháp phân loại các bộ dữ liệu lớn bằng thuật toán RF trong các ứng dụng khác nhau. Bài viết này thảo luận về nhiều ứng dụng sử dụng RF để phân loại dữ liệu như phát hiện xâm nhập mạng, phát hiện spam email, phân loại gen, phát hiện gian lận thẻ tín dụng và phân loại văn bản. Trong bài báo này, mỗi ứng dụng được giới thiệu ngắn gọn và sau đó bộ dữ liệu được sử dụng để thực hiện sẽ được thảo luận và cuối cùng là việc thực hiện thuật toán RF để thu kết quả và sau đó được thảo luận.

Raj Kumar Nepali và Yong Wang (2016) [26] đã giới thiệu phương pháp phân loại các URL ngắn độc hại trên Twitter. Twitter là một trong những Mạng xã hội trực tuyến phổ biến nhất. Nó được sử dụng bởi hàng triệu người dùng trên toàn thế giới mỗi ngày. Do giới hạn văn bản trên Twitter (140 ký tự mỗi tweet, các mẫu tin nhỏ), các dịch vụ rút ngắn URL được sử dụng rộng rãi; tuy nhiên họ không thoát khỏi rủi ro. Các URL rút ngắn hoàn toàn khác với các URL gốc và do đó người dùng không biết các URL ngắn sẽ hướng họ đến đâu. Những kẻ tấn công tận dụng kiến ​​thức này là lợi thế để truyền bá các URL độc hại. Hầu hết các cách tiếp cận được đề xuất để phân loại URL độc hại sử dụng thông tin từ cả mạng xã hội và nhà cung cấp dịch vụ rút ngắn URL. Bài viết này đề xuất một phương pháp mới để phát hiện các URL ngắn độc hại chỉ sử dụng các tính năng hiển thị của tweet và hồ sơ người dùng. Chúng tôi kiểm tra bốn thuật toán học máy, tức là Naïve Bayes, RF, Support Vector Machines và hồi quy Logistic. Kết quả có được độ chính xác lên tới 97% khi sử dụng phân lớp RF. Kết quả thử nghiệm của chúng tôi cho thấy rằng phương pháp sử dụng các tính năng hiển thị từ các mạng xã hội để phát hiện các URL độc hại là có thể thực hiện được.

Adnan Hodžić, Jasmin Kevrić và Adem Karadag (2016) [27] đã có công trình so sánh các kỹ thuật học máy trong phân loại trang Web. Bài viết này mô tả so sánh trong phân lớp các trang web lừa đảo bằng các thuật toán học máy khác nhau. RF, C4.5, REP tree, Decision Stump, Hoeffding Tree, Rotation Forest và Multi-layer Perceptron đã được sử dụng để xác định phương pháp nào cung cấp kết quả tốt nhất trong phân loại trang web lừa đảo. Kết quả cho thấy RF với REP Tree cho thấy hiệu suất tốt nhất trên bộ dữ liệu này để phân loại các trang web lừa đảo.

Adwan Yasin và Abdelmunem Abuhasan (2016) [28] đã giới thiệu một mô hình phân loại thông minh để phát hiện trang web. Bài viết này trình bày một mô hình phân loại thông minh để phát hiện email lừa đảo bằng cách sử dụng khám phá kiến ​​thức, khai thác dữ liệu và kỹ thuật xử lý văn bản. Mô hình đã áp dụng các quy trình khám phá tri thức bằng năm thuật toán phân loại phổ biến và đạt được sự nâng cao đáng chú ý về độ chính xác phân loại, độ chính xác 99.10% đã đạt được bằng thuật toán RF và 98.40% bằng J48.

Andronicus A. Akinyelu và Aderemi O. Adewumi (2014) [29] đã giới thiệu phương pháp phân loại email lừa đảo bằng kỹ thuật học máy rừng ngẫu nhiên. Bài viết này điều tra và báo cáo việc sử dụng thuật toán RF trong phân loại các cuộc tấn công lừa đảo, với mục tiêu chính là phát triển một mô hình loại email lừa đảo cải tiến với độ chính xác dự đoán tốt hơn và số lượng tính năng ít hơn. Từ bộ dữ liệu bao gồm 2000 email lừa đảo, một tập hợp các tính năng email lừa đảo nổi bật (được xác định từ tài liệu) đã được thuật toán học máy trích xuất và sử dụng với độ chính xác phân loại là 99.70%.

## 

## **CHƯƠNG II. XÂY DỰNG MÔ HÌNH**

* 1. **Mô tả bộ dữ liệu.**
     1. *Giới thiệu và mô tả dữ liệu.*

Bộ dữ liệu Phishing Website [[1]](#footnote-1) lấy từ kho lưu trữ tài liệu UCI, được thu thập chủ yếu từ các kho lưu trữ: PhishTank, MillerSmiles và các kho tìm kiếm từ Google bởi Rami Mustafa A Mohammad (Đại học Huddersfield), Lee McCluskey (Đại học Huddersfield) và Fadi Thabtah (Đại học Canada Dubai). Nó liên quan đến các tính năng để nhận biết một Website lừa đảo hay là hợp pháp bao gồm 31 thuộc tính (1 thuộc tính nhãn lớp) và 11055 trường hợp.

Bảng 2.1. Bảng các thuộc tính của bộ dữ liệu Phishing Website.

|  |  |
| --- | --- |
| **Thuộc tính** | **Nhãn** |
| 1. having\_IP\_Address | { -1,1 } |
| 2. URL\_Length | { 1,0,-1 } |
| 3. Shortining\_Service | { 1,-1 } |
| 4. having\_At\_Symbol | { 1,-1 } |
| 5. double\_slash\_redirecting | { -1,1 } |
| 6. Prefix\_Suffix | { -1,1 } |
| 7. having\_Sub\_Domain | { -1,0,1 } |
| 8. SSLfinal\_State | { -1,1,0 } |
| 9. Domain\_registeration\_length | { -1,1 } |
| 10. Favicon | { 1,-1 } |
| 11. port | { 1,-1 } |
| 12. HTTPS\_token | { -1,1 } |
| 13. Request\_URL | { 1,-1 } |
| 14. URL\_of\_Anchor | { -1,0,1 } |
| 15. Links\_in\_tags | { 1,-1,0 } |
| 16. SFH | { -1,1,0 } |
| 17. Submitting\_to\_email | { -1,1 } |
| 18. Abnormal\_URL | { -1,1 } |
| 19. Redirect | { 0,1 } |
| 20. on\_mouseover | { 1,-1 } |
| 21. RightClick | { 1,-1 } |
| 22. popUpWidnow | { 1,-1 } |
| 23. Iframe | { 1,-1 } |
| 24. age\_of\_domain | { -1,1 } |
| 25. DNSRecord | { -1,1 } |
| 26. web\_traffic | { -1,0,1 } |
| 27. Page\_Rank | { -1,1 } |
| 28. Google\_Index | { 1,-1 } |
| 29. Links\_pointing\_to\_page | { 1,0,-1 } |
| 30. Statistical\_report | { -1,1 } |
| **31. Result (Nhãn lớp)** | **{ -1,1 }** |

Chú giải:

“1” có nghĩa là lừa đảo.

“0” có nghĩa là nghi ngờ.

“-1” có nghĩa là hợp pháp.

* + 1. *Thông tin thuộc tính.*

Các thuộc tính trong bộ dữ liệu này là những tính năng được chứng minh là hợp lý và hiệu quả trong việc dự đoán các trang web lừa đảo.

* + - 1. Sử dụng địa chỉ IP (using the IP Address).

Nếu một địa chỉ IP được sử dụng thay thế cho tên miền URL, người dùng có thể chắc chắn là ai đó đang cố lấy cắp thông tin của họ. Đôi khi địa chỉ IP còn được đổi thành mã thập lục phân.

Luật:

* Nếu tên miền có địa chỉ IP là lừa đảo.
* Trường hợp còn lại là hợp pháp.
  + - 1. Độ dài của URL (URL\_Length).

Những kẻ lừa đảo có thể sử dụng URL dài để ẩn phần đáng ngờ trong thanh địa chỉ. Để đảm bảo tính chính xác, chúng tôi đã tính toán độ dài của URL trong bộ dữ liệu và tạo ra độ dài URL trung bình. Kết quả cho thấy, nếu độ dài của URL lớn hơn hoặc bằng 54 ký tự thì URL có thể là Phishing.

Luật:

* Độ dài URL < 54 là hợp pháp.
* Độ dài URL nằm trong khoảng từ 54 đến 75 là nghi ngờ.
* Trường hợp còn lại là lừa đảo.
  + - 1. Sử dụng dịch vụ rút ngắn URL (Using URL Shortening Services).

Rút ngắn URL là một phương pháp trên World Wide Web, trong đó một URL có thể được làm nhỏ hơn đáng kể về chiều dài và vẫn dẫn đến trang web bắt buộc. Điều này được thực hiện bằng phương tiện Chuyển hướng HTTP trên một tên miền ngắn, liên kết đến trang web có URL dài.

Luật:

* Có sử dụng dịch vụ rút ngắn URL là lừa đảo.
* Trường hợp còn lại là hợp pháp.
  + - 1. Sử dụng biểu tượng (URL using Symbol).

Việc sử dụng biểu tượng “@” khiến trình duyệt bỏ qua mọi thứ trước “@” và địa chỉ thực thường theo sau “@”.

Luật:

* URL có biểu tượng “@” => lừa đảo
* Trường hợp còn lại là hợp pháp.
  + - 1. Chuyển hướng bằng cách sử dụng gạch chéo (Redirecting using double slash).

Sự tồn tại của “//” trong đường dẫn URL có nghĩa là người dùng sẽ được chuyển hướng đến một trang web khác. Chúng tôi kiểm tra vị trí nơi “//” xuất hiện thấy rằng nết URL bắt đầu bằng “http” thì “//” sẽ xuất hiện ở vị trị thứ 6 và nếu bắt đầu bằng “https” thì sẽ xuất hiện ở vị trí thứ 7.

Luật:

* Vị trí xuất hiện cuối cùng của “//” trong URL >7 là lừa đảo.
* Trường hợp còn lại còn hợp pháp.
  + - 1. Tiền tố và hậu tố trong tên miền (Prefix and Suffix).

Biểu tượng dấu (-) hiếm khi được sử dụng trong các URL hợp pháp. Những kẻ lừa đảo có xu hướng thêm tiền tố hoặc hậu tố được phân tách bằng (-) vào tên miền để người dùng cảm thấy rằng họ đang xử lý một trang web hợp pháp.

Luật:

* Tên miền có sử dụng (-) là lừa đảo.
* Trường hợp còn lại là hợp pháp.
  + - 1. Tên miền phụ và nhiều tên miền phụ (Sub Domain and Multi Sub Domains).

Ví dụ: [www.hce.edu.vn](http://www.hce.edu.vn)

Một tên miền bao gồm các tên miền cấp cao theo mã quốc gia, ở ví dụ trên là “vn”. Phần “edu” là tên miền cấp 2. Để tạo quy tắc cho chức năng này, trước tiên chúng tôi bỏ qua [www.](http://www.) Sau đó bỏ qua “.vn” nếu nó tồn tại. Cuối cùng nếu số “.” lớn hơn 1 thì URL được phân loại là Suspicious, lớn hơn 2 sẽ được phân loại là Phishing. Và nếu URL không có tên miền phụ thì sẽ được gán cho Legitimate.

Luật:

* Số dấu chấm trong tên miền bằng 1 là hợp pháp.
* Số dấu chấm trong tên miền bằng 2 là nghi ngờ.
* Trường hợp còn lại là Lừa đảo.
  + - 1. Giao thức Secure Sockets Layer (HTTP with Secure Sockets Layer).

Sự tồn tại của HTTP là rất quan trọng trong việc tạo ấn tượng về tính hợp pháp của trang web nhưng điều này rõ ràng là không đủ. Chứng chỉ HTTP bao gồm phạm vi của tổ chức phát hành và tuổi chứng nhận. Bằng cách kiểm tra các bộ dữ liệu của chúng tôi, chúng tôi thấy rằng độ tuổi tối thiểu của chứng chỉ có uy tín là hai năm.

Luật:

* Sử dụng HTTP nhà phát hành đáng tin cậy và độ tuổi chứng nhận từ 1 năm trở lên là hợp pháp.
* Sử dụng HTTP và nhà phát hàng không đáng tin cậy là nghi ngờ.
* Trường hợp còn lại là lừa đảo.
  + - 1. Độ dài đăng ký tên miền (Domain Registration Length).

Dựa trên thực tế, một trang web lừa đảo tồn tại trong một khoảng thời gian ngắn. Các tên miền đáng tin cậy thường được trả trước vài năm. Trong tập dữ liệu, chúng tôi thấy rằng các tên miền lừa đảo dài nhất chỉ được sử dụng 1 năm.

Luật:

* Tên miền hết hạn sau 1 năm là lừa đảo.
* Còn không là hợp pháp.
  + - 1. Favicon.

Favicon là 1 biểu tượng được liên kết với 1 trang web cụ thể. Nhiều người dùng hiện có đọc tin hiển thị Favicon như một lời nhắc nhở trực quan về danh tính trang web trong thanh địa chỉ. Nếu favicon được tải lên từ một tên miền khác với tên miền hiển thị trong thanh địa chỉ thì trang web được coi là Phishing.

Luật:

* Favcon được tải từ tên miền bên ngoài là hợp pháp.
* Còn không là nghi ngờ.
  + - 1. Sử dụng cổng không chuẩn (Using Non-Standard Port).

Tính năng này rất hữu ích trong việc xác thực nếu một dịch vụ cụ thể tăng hay giảm trên một máy chủ cụ thể. Trong mục đích kiểm soát sự xâm nhập, nên mở những cổng mà bạn cần. Mặc định, một số tường lửa, máy chủ Proxy và Network Address Translation (NAT) sẽ chặn tất cả các hầu hết các cổng và chỉ mở các cổng được chọn. Nếu tất cả các cổng điều mở, những kẻ lừa đảo có thể chạy hầu hết mọi dịch vụ mà họ muốn và kết quả thông tin người dùng bị đe dọa.

Dưới đây là một số cổng quan trọng nhất và trạng thái thường xuyên của chúng.

Bảng 2.2. Một số cổng quan trọng của trang Web (Tính năng Website lừa đảo, nguồn: [https://archive.ics.uci.edu/phishing-website-features](https://archive.ics.uci.edu/ml/machine-learning-databases/00327/)).

|  |  |  |  |
| --- | --- | --- | --- |
| **PORT** | **Service** | **Ý nghĩa** | **Trạng thái thường xuyên** |
| 21 | FTP | Chuyển tập tin từ máy chủ này sang máy chủ khác. | Đóng |
| 22 | SSH | Giao thức truyền tâp tin an toàn. | Đóng |
| 23 | Telnet | Tương tác hai chiều bằng văn bản. | Đóng |
| 80 | HTTP | Giao thức truyền tải siêu văn bản. | Mở |
| 443 | HTTPS | Giao thức truyền tải siêu văn bản được bảo mật. | Mở |
| 445 | SMB | Cung cấp quyền truy cập chung vào các tệp, máy in, cổng kết nối. | Đóng |
| 1433 | MSSQL | Lưu trữ và truy xuất dữ liệu theo yêu cầu của các ứng dụng phần mềm khác. | Đóng |
| 1521 | ORACLE | Truy cập cơ sở dữ liệu oracle từ trang web. | Đóng |
| 3306 | MySQL | Truy cập cơ sở dữ liệu MySQL từ trang web. | Đóng |
| 3389 | Remote Desktop | Cho phép truy cập từ xa và cộng tác từ xa. | Đóng |

Luật:

* Cổng được sử dụng khác với trạng thái thường xuyên là lừa đảo.
* Còn không là lợp pháp.
  + - 1. HTTPS Token trong tên miền của URL (The Existence of “HTTPS” Token in the Domain Part of the URL).

Các kẻ lừa đảo có thể thêm phần “HTTPS” vào phần tên miền của URL để đánh lừa người dùng.

Luật:

* Sử dụng “HTTPS” trong phần tên miền của URL là lừa đảo.
* Còn không là hợp pháp.
  + - 1. Request URL.

URL yêu cầu kiểm tra các đối tượng bên ngoài có trong 1 trang web như hình ảnh, video, âm thanh có được tải lên từ một tên miền khác hay không. Trong các trang web hợp pháp, hầu hết các đối tượng đều chia sẻ cùng một tên miền.

Luật:

* Tỉ lệ phần trăm các đối tượng được tải lên từ tên miền khác < 22% là hợp pháp.
* Tỉ lệ phần trăm các đối tượng được tải lên từ tên miền khác từ 22% đến 61% là nghi ngờ.
* Trường hợp còn lại là lừa đảo.
  + - 1. URL of Anchor.

Anchor được xác định bởi thẻ <a>. Tính năng này được xử lý tương tự Request URL.

Luật:

* Tỉ lệ phần trăm anchor trong URL <31% là hợp pháp.
* Tỉ lệ phần trăm anchor từ 31% đến 67% là nghi ngờ.
* Trường hợp còn lại là lừa đảo.
  + - 1. Liên kết trong các thẻ (Links in <Meta>, <Script> and <Link> tags).

Các trang web hợp pháp thường dùng các thẻ với chức năng sau:

* Thẻ <Meta> cung cấp siêu dữ liệu về HTML Document.
* Thẻ <Scipt> tạo lập lệnh cho máy khách.
* Thẻ <Link> lấy tài nguyên từ trang web khác.

Các thẻ này thường được liên kết đến cùng 1 tên miền của trang web.

Luật:

* Tỉ lệ phần trăm liên kế trong thẻ <Meta>, <Scipt> và <Link> nhỏ hơn 17% là Hợp pháp.
* Tỉ lệ phần trăm liên kế trong thẻ <Meta>, <Scipt> và <Link> từ 17% đến 81% là Nghi ngờ.
* Trường hợp còn lại là Lừa đảo.
  + - 1. Trình xử lý biểu mẫu máy chủ (Server Form Handler (SFH)).

Form Handler là 1 trang chạy ở phía server cho phép xử lý dữ liệu nhập. Các SFH có chứa một chuỗi trống hoặc khoảng trống được coi là đáng ngờ vì một hành động nên được thực hiện dựa trên thông tin đã gửi. Ngoài ra, nếu tên miền trong SFH khác với tên miền của trang web, điều này cho thấy rằng trang web đáng ngờ vì thông tin đã gửi hiếm khi được xử lý bởi các tên miền bên ngoài.

Luật:

* SFH chứa 1 chuỗi trống hoặc khoảng trống là lừa đảo.
* SFH đề cập đến tên 1 miền khác là nghi ngờ.
* Trường hợp còn lại là hợp pháp.
  + - 1. Gửi thông tin đến Email (Submitting Information to Email).

Web form cho phép người dùng gửi thông tin cá nhân của mình đến máy chủ để được xử lý. Một kẻ lừa đảo có thể chuyển hướng thông tin của người dùng đến email của mình. Với mục đích đó, những kẻ lừa đảo thường sử dụng các hàm “mail()” và “mailto” trong PHP để đánh cắp thông tin.

Luật:

* Sử dụng “mail()” hoặc “mailto” là lửa đảo
* Trường hợp còn lại là hợp pháp.
  + - 1. URL bất thường (Abnormal URL).

Whois là một trang web giúp bạn tìm thông tin của người chủ sở hữu domain. Tính năng này có thể được trích từ cơ sở dữ liệu WHOIS. Ở các trang web hợp pháp, danh tính là 1 điều không thể thiếu của URL.

Luật:

* Tên máy chủ không có danh tính là lừa đảo.
* Trường hợp có danh tính là hợp pháp.
  + - 1. Chuyển tiếp trang web (Website Forwarding).

Một cách tốt để phân biệt trang web hợp pháp và lừa đảo là xem số lần chuyển tiếp của trang web đó. Trong bộ dữ liệu, chúng tôi tìm thấy rằng, trang web hợp pháp có số lần chuyển tiếp tối đa một lần. Đối với trang web lừa đảo thì ít nhất là 4 lần.

Luật:

* Số lần chuyển tiếp nhỏ hơn hoặc bằng 1 là hợp pháp.
* Số lần chuyển tiếp từ 2 đến nhỏ hơn 4 là nghi ngờ.
* Số lần chuyển tiếp lớn hơn 4 là lửa đảo.
  + - 1. Tùy chỉnh thanh trạng thái (Status Bar Customization).

Kẻ lừa đảo có thể sử dụng JavaScript để hiển thị URL giả trong thanh trạng thái cho người dùng. Để kiểm tra chức năng này, chúng tôi phải khai thác mã nguồn trang web, đặc biệt là sự kiện trên onMouseOver và kiểm tra xem có bất kì sự thay đổi nào trên thanh trạng thái.

Luật:

* Có sự thay đổi trên thanh trạng thái là lừa đảo.
* Nếu không là hợp pháp.
  + - 1. Vô hiệu hóa nhấp chuột phải (Disabling Right Click).

Những kẻ lừa đảo sử dụng JavaScript để vô hiệu hóa nhấp chuột phải, vì vậy người dùng không thể xem và lưu mã nguồn của trang web. Đối với tính năng này, chúng tôi sẽ tìm kiếm sự kiện “event button ==2” trong mã nguồn và kiểm tra chuột phải có bị vô hiệu hóa.

Luật:

* Nhấp chuột phải bị vô hiệu hóa là lừa đảo
* Còn không là hợp pháp.
  + - 1. Sử dụng cửa số bật lên (Using Pop-up Window).

Thật bất thường khi thấy một trang web hợp pháp yêu cầu người dùng gửi thông tin cá nhân của họ thông qua cửa sổ bật lên. Mặt khác, tính năng này đã được sử dụng trong 1 số trang web hợp pháp và mục tiêu của nó là cảnh báo người dùng về các hoạt động lừa đảo hoặc phát thông báo chào mừng, mặc dù không có thông tin cá nhân nào được yêu cầu điền vào thông qua các cửa sổ bật lên.

Luật:

* Cửa sổ bật lên chứa các trường điền thông tin là lừa đảo.
* Còn không là hợp pháp.
  + - 1. Chuyển hướng IFrame (IFrame Redirection).

IFrame là thẻ HTML được sử dụng để hiển thị một trang web bổ sung thành một trang web hiện đang được hiển thị. Những kẻ lừa đảo có thể sử dụng thẻ “iframe” và làm cho nó vô hình, tức là không có viền khung.

Luật:

* Sử dụng iframe là lừa đảo.
* Còn không là hợp pháp.
  + - 1. Tuổi miền (Age of domain).

Hầu hết các trang web lừa đảo có tuổi đời ngắn. Bằng cách xem xét dữ liệu của chúng tôi, chúng tôi thấy rằng độ tuổi tuối thiểu của 1 trang web hợp pháp là 6 tháng.

Luật:

* Độ tuổi của miền từ 6 tháng trở lên là hợp pháp.
* Trường hợp còn lại là lừa đảo.
  + - 1. Bản ghi DNS (DNS Record).

Đối với các trang web lừa đảo, danh tính được xác nhận không được cơ sở dữ liệu WHOIS công nhận hoặc không có hồ sơ nào được thiết lập cho tên máy chủ.

Luật:

* Bản ghi DNS trống là lừa đảo.
* Còn không là hợp pháp.
  + - 1. Lượng truy cập trang web (Website Traffic).

Tính năng này đo lượng mức độ phổ biến của trang web bằng cách xác định số lượng khách truy cập và số lượng trang họ truy cập. Tuy nhiên, do các trang web lừa đảo tồn tại trong thời gian ngắn và chúng có thể không được cơ sở dữ liệu của Alexa nhận ra. Bằng cách xem xét cơ sở dữ liệu của chúng tôi, chúng tôi thấy rằng các trang web hợp pháp thường được xếp hạng trong số 100.000 trang web hàng đầu. Còn nếu tên miền không có lượng truy cập hoặc không được trang web Alexa nhận ra thì nó là Phishing.

Luật:

* Thuộc danh sách 100.000 trang web có lượng truy cập cao nhất là hợp pháp.
* Không thuộc danh sách 100.000 nhưng Alexa nhận ra là nghi ngờ.
* Còn lại là lừa đảo.
  + - 1. Xếp hạng trang (PageRank).

PageRank là một giá trị nằm trong khoảng từ "0" đến "1". PageRank nhằm mục đích đo lường tầm quan trọng của một trang web trên Internet.

Luật:

* PageRank < 0.2 là lừa đảo.
* Trường hợp còn lại là hợp pháp.
  + - 1. Google Index.

Tính năng này kiểm tra xem một trang web có nằm trong chỉ mục của Google hay không. Khi một trang web được Google lập chỉ mục, nó sẽ được hiển thị trên kết quả tìm kiếm. Thông thường, các trang web lừa đảo chỉ có thể truy cập trong một thời gian ngắn và do đó, nhiều trang web lừa đảo có thể không được tìm thấy trên chỉ mục của Google.

Luật:

* Được google lập chỉ mục là hợp pháp.
* Còn không là lừa đảo.
  + - 1. Số lượng liên kết trỏ đến trang (Number of Links Pointing to Page).

Số lượng liên kết trỏ đến trang web cho biết mức độ hợp pháp của nó, ngay cả khi một số liên kết có cùng tên miền (Dean, 2014). Trong bộ dữ liệu của chúng tôi và do tuổi thọ ngắn, chúng tôi thấy rằng 98% các mục dữ liệu lừa đảo không có liên kết trỏ đến chúng. Mặt khác, các trang web hợp pháp có ít nhất 2 liên kết bên ngoài trỏ đến chúng.

Luật:

* Số liên kết trỏ đến trang Web bằng 0 là lừa đảo.
* Số liên kết trỏ đến trang web từ không đến 2 là nghi ngờ.
* Trường hợp còn lại là hợp pháp.
  + - 1. Tính năng dựa trên báo cáo thống kê (Statistical-Reports Based Feature).

Một số bên như PhishTank (PhishTank Stats, 2010-2012) và StopBadware (StopBadware, 2010-2012) tạo ra nhiều báo cáo thống kê trên các trang web lừa đảo tại mọi thời điểm nhất định; một số là hàng tháng và những người khác là hàng quý. Trong nghiên cứu của chúng tôi, chúng tôi đã sử dụng 2 hình thức trong số mười thống kê hàng đầu từ PhishTank: 10 tên miền hàng đầu và 10 IP hàng đầu theo các báo cáo thống kê được công bố trong ba năm qua, bắt đầu từ tháng 1 năm 2010 đến tháng 11 năm 2012. Mùi, chúng tôi đã sử dụng 50 địa chỉ IP hàng đầu.

Luật:

* Lưu trữ thuộc về IP lừa đảo hàng đầu hoặc Tên miền lừa đảo hàng đầu là lừa đảo.
* Mặt khác là hợp pháp.
  1. **Xây dựng mô hình.**

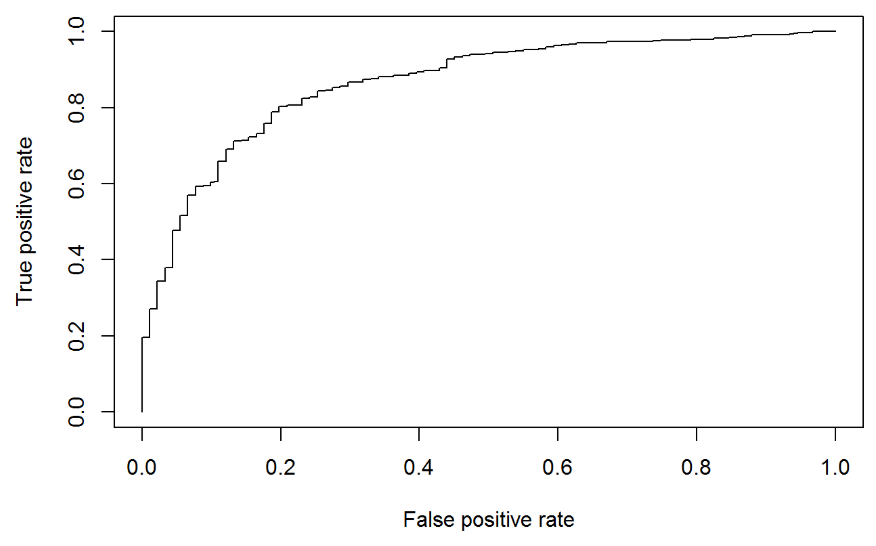
Để kiểm tra hiệu quả của mô hình phân lớp Random Forest trong việc dự đoán các trang Web lừa đảo, chung tôi sử dụng một số mô hình phân lớp cơ sở để so sánh, bao gồm: (i) Naïve Bayes; (ii) K-nn (IBk); (iii) SVM; (iv) MultilayerPerceptron.

Việc xây dựng mô hình phân lớp bằng Random Forest cũng như các bộ phân lớp cơ bản gồm các bước sau:

* Bước 1: Tiền hành chia tập dữ liệu Phishing Data thành hai bộ: Bộ dữ liệu huấn luyện (Training Set) và Bộ dữ liệu kiểm thử (Test Set) theo tỉ lệ 70 và 30.
* Bước 2: Tải bộ dữ liệu huấn luyện vào bộ nhớ và tiến hành huấn luyện mô hình với 10-fold Cross Validation.
* Bước 3: Kiểm thử mô hình bằng bộ dữ liệu kiểm thử.
  1. **Phương pháp đánh giá mô hình.**
     1. *Tổng quan về Receiver Operating Characteristic (ROC).*

ROC được sử dụng khá phổ biến để đánh giá hiệu năng của các mô hình phân lớp. Nó là một đồ thị đường cong minh họa cho một hệ thống phân loại nhị phân khi các ngưỡng phân loại bị thay đổi.

ROC được tạo ra bằng cách biểu diễn tỷ lệ dự báo true positive rate (TPR) dựa trên tỷ lệ dự báo failse positive rate (FPR) tại các ngưỡng khác nhau.



Hình 2. 1. Đường cong ROC (Giới thiệu đường cong ROC, nguồn: [https://rstudio-pubs-static.s3.amazonaws.com](https://rstudio-pubs-static.s3.amazonaws.com/267441_5459af9d83ae44f18a13aea4a479f31f.html)).

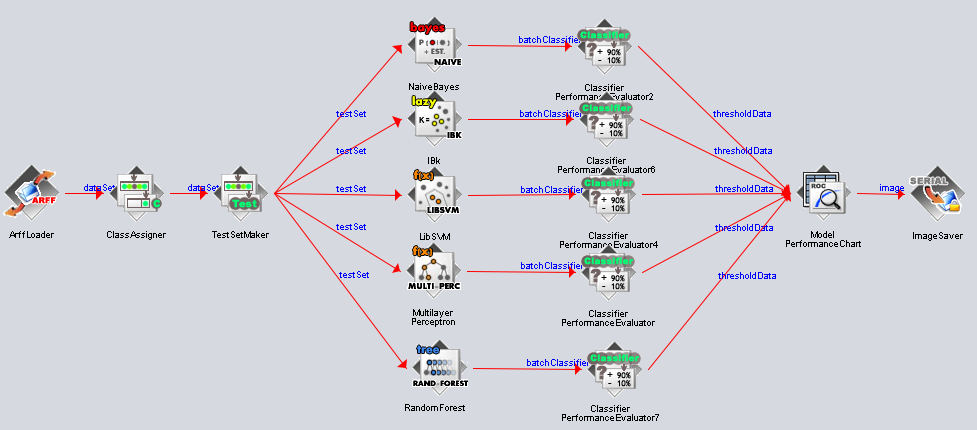
Đường ROC của mô hình nào càng gần với đỉnh trên cùng bên trái thì cho độ hiệu năng của mô hình càng cao và ngược lại những đường ROC càng gần đường chéo thì hiệu năng càng thấp.

* + 1. *Xây dựng ROC để đánh giá.*

Để so sánh hiệu năng của các mô hình phân lớp cơ bản như Naïve Bayes; K-nn (IBk); SVM; MultilayerPerceptron đối với mô hình Random Forest, chúng tôi sử dụng công cụ Knowledge Flow của Weka.

Quy trình xây dựng ROC trong Knowledge Flow bao gồm:

* Bước 1: Tải bộ dữ liệu kiểm thử vào ArffLoader trong DataSource.
* Bước 2: Chọn thuộc tính làm nhãn lớp ở ClassAssigner và nối dataset tới TestSetMaker trong mục Evaluation.
* Bước 3: Chọn các bộ phân lớp Random Forest; Naïve Bayes; K-nn (IBk); SVM (SMO); MultilayerPerceptron và thiết lập các thông số của bộ phân lớp đó trong Classifier. Mỗi bộ phân lớp sẽ ứng với một trình hiệu suất đánh giá ClassifierPerformanceEvaluator.
* Bước 4: Đưa kết quả trên vào ModelPerfomanceChart trong Visualization và lưu kết quả vào ImageSaver để lấy hình ảnh kết quả.



Hình 2. 2. Mô hình đánh giá ROC trên Knowledge Flow.

* 1. **Xây dựng ứng dụng dự đoán Website lừa đảo.**

Trong khóa luận này, chúng tôi xây dựng ứng dụng với mục tiêu là tiến hành dự đoán nhãn lớp của các bộ dữ liệu không nhãn sử dụng mô hình phân lớp Random Forest. Bộ dữ liệu chúng tôi sử dụng trong xuyên suốt khóa luận là bộ dữ liệu Phishing Website với nhãn lớp là Kết quả (Result) có hai giá trị là Lừa đảo (Phishing) tương ứng với “1”, và Hợp pháp (Legitimate) tương ứng với “-1”.

Đầu tiên, chúng tôi tiến hành tạo một số bộ dữ liệu không nhãn lớp (Unlabel Set) tách ra từ bộ dữ liệu kiểm thử (Test Set), các trường hợp của bộ Unlabel sẽ bị xóa đi nhãn lớp.

Sau đó, tạo giao diện và viết code có ứng dụng dự đoán với các chức năng sau:

* Chọn bộ dữ liệu từ thư mục mình đã lưu, sau đó ứng dụng sẽ chia bộ dữ liệu theo tỉ lệ 70% Training Set và 30% Test Set.
* Chọn bộ dữ liệu Unlabel Set muốn dự đoán nhãn lớp.
* Sau đó tiến hành huấn luyện mô hình với 10-fold Cross Valication, kiểm thử lại mô hình bằng bộ Test Set đã chia ở trên sử dụng mô hình phân lớp Random Forest.
* Cuối cùng là tiến hành dự đoán nhãn lớp cho bộ dữ liệu Unlabel Set và xuất kết quả ra bảng kết quả.

## 

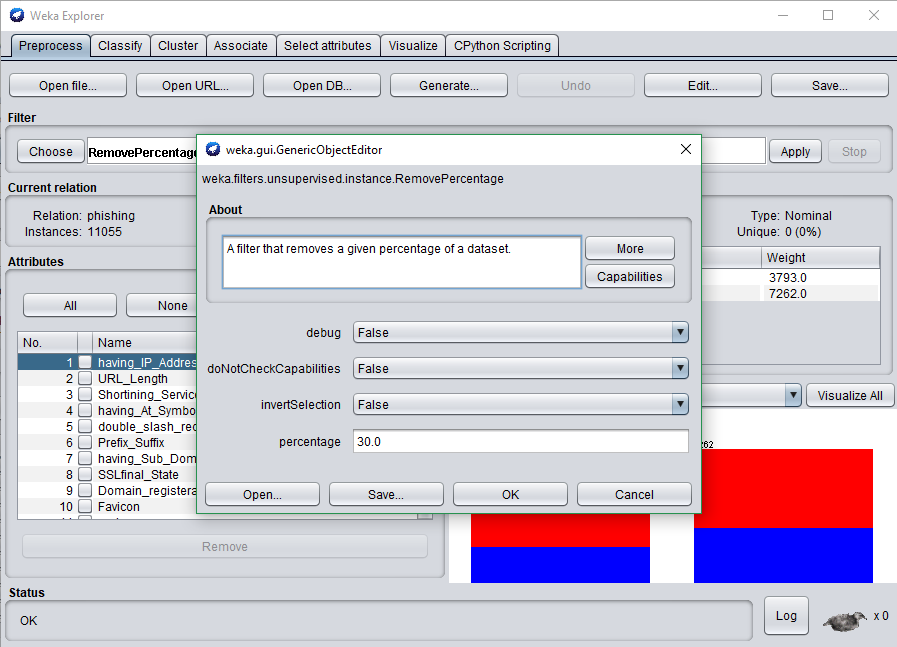
## **CHƯƠNG III. THÍ NGHIỆM VÀ KẾT QUẢ**

### **3.1. Thiết lập thí nghiệm.**

#### 3.1.1. Phân lớp với các bộ phân lớp cơ bản.

Ta tiến hành thiết lập thí nghiệm phân lớp với các bộ phân lớp trên như sau:

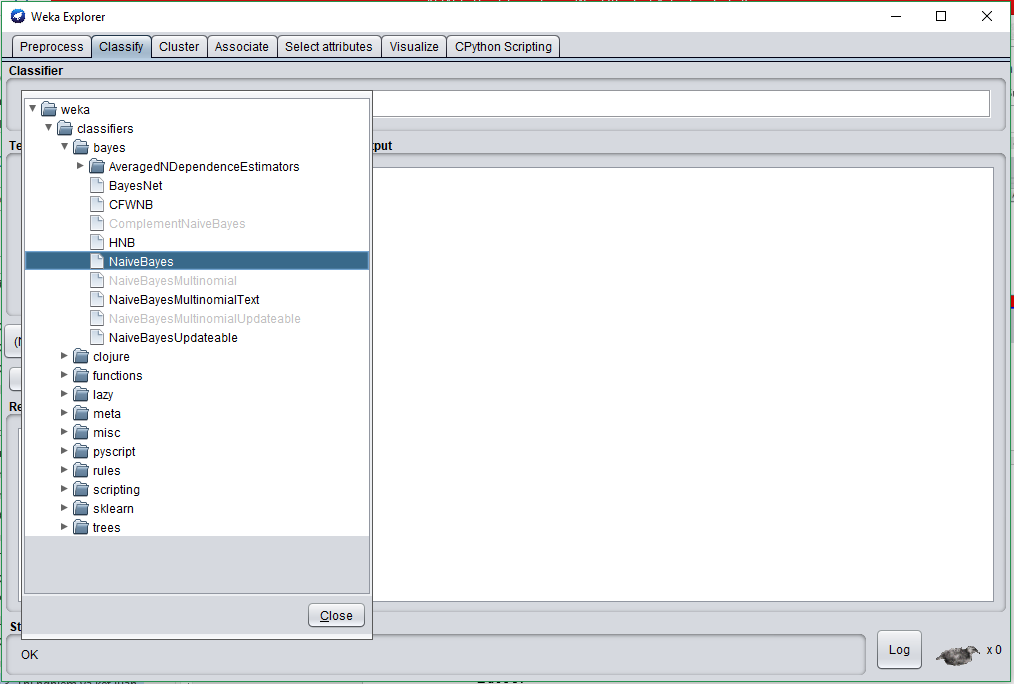
* Bước 1: Tải bộ dữ liệu PhishingWebsiteDataSet.arff vào bộ nhớ. Tiến hành chia bộ dữ liệu đã tải thành 2 bộ Training Set và Data Set với tỷ lệ 70/30 bằng bộ lọc Remove Percentage.



Hình 3. 1. Chia bộ dữ liệu PhishingWebsite sử dụng bộ lọc Remove Percentage.

* Bước 2: Tải bộ Training Set vào bộ nhớ. Lựa chọn classifier và chọn bộ phân lớp bạn muốn sử dụng để huấn luyện.

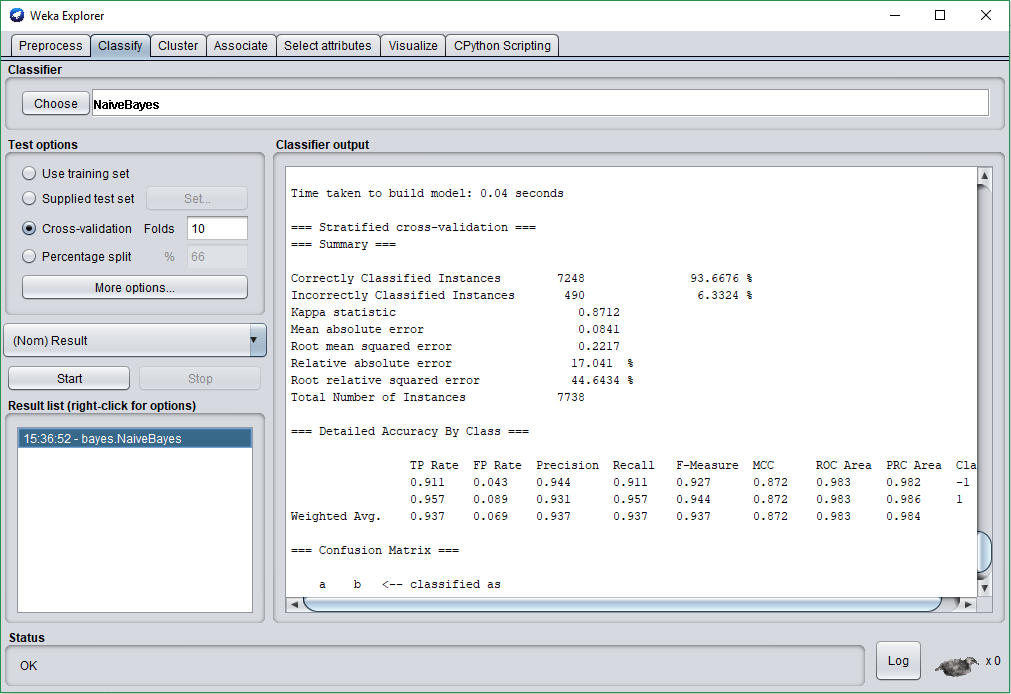
Ví dụ: Chọn mô hình Naïve Bayes từ Bayes.



Hình 3. 2. Chọn bộ phân lớp Naïve Bayes để huấn luyện.

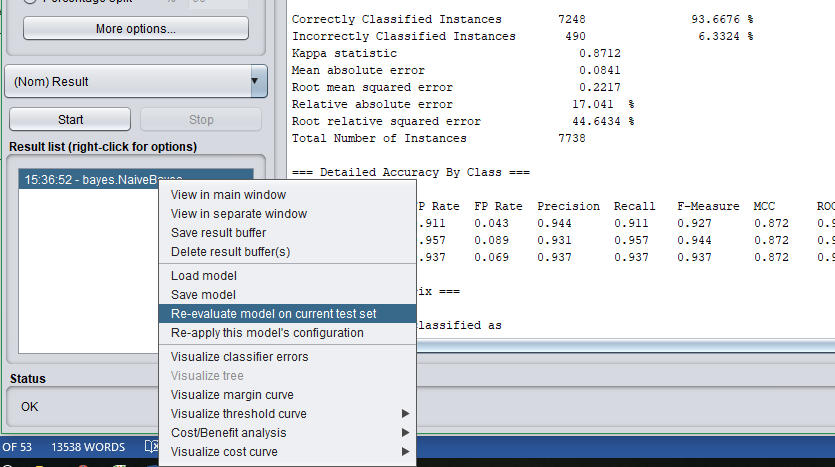
* Bước 3: Thiết lập chế độ huấn luyện. Và ấn start để tiến hành huấn luyện.

Ví dụ: 10-fold Cross Validation hoặc Percentage split.

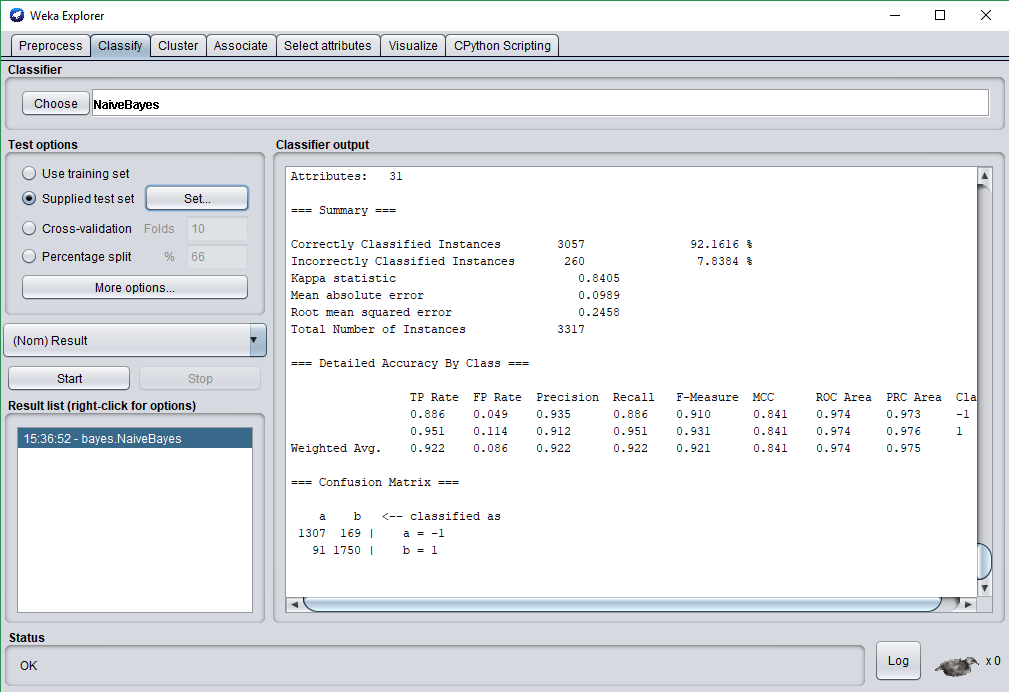


Hình 3. 3. Tiến hành huấn luyện bộ dữ liệu Training Set sử dụng bộ phân lớp Naïve Bayes với 10-fold Cross Validation.

* Bước 4: Tiến hành kiểm thử mô hình bằng cách chọn Supplied test set và chọn bộ dữ liệu Test Set đã chia ở bước 1. Sau đó kích chuột phải vào mô hình đã huấn luyện chọn Re-evaluate model on current test set để kiểm thử lại với bộ dữ liệu Test Set.



Hình 3. 4. Kiểm thử lại mô hình với Test Set.

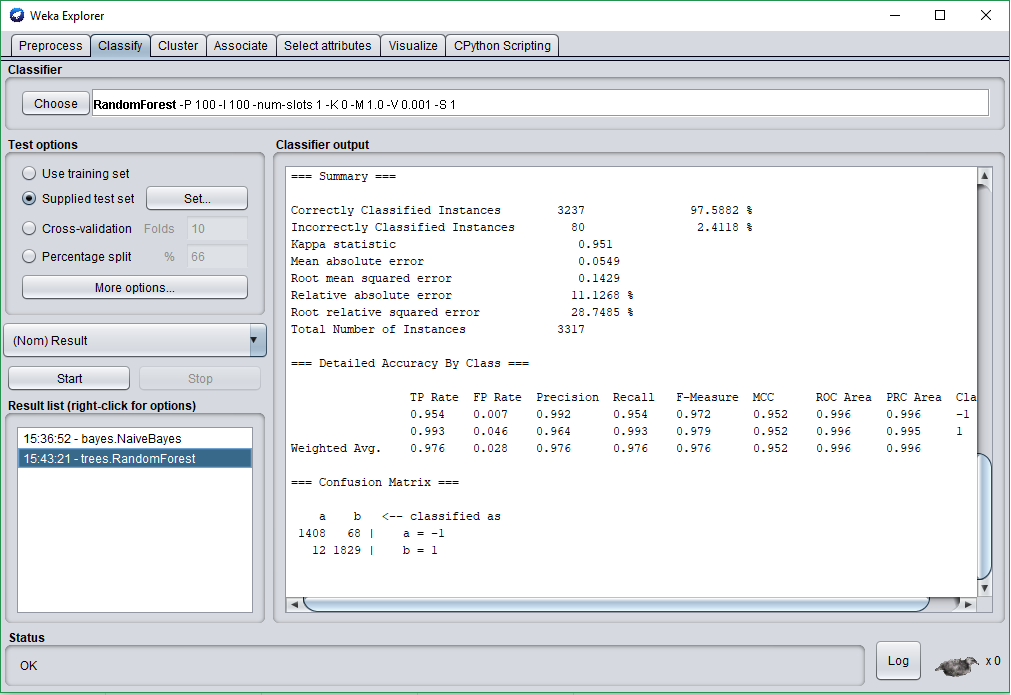


Hình 3. 5. Kết quả kiểm thử với Test Set của mô hình phân lớp Naïve Bayes.

#### 3.1.2. Phân lớp với mô hình Random Forest.

Tương tự với các bộ phân lớp cơ bản quy trình phân lớp với mô hình Random Forest cũng bao gồm các bước sau:

* Bước 1: Tải bộ Training Set vào bộ nhớ. Lựa chọn classifier là Random Forest.
* Bước 2: Thiết lập chế độ huấn luyện. Và ấn start để tiến hành huấn luyện.
* Bước 4: Tiến hành kiểm thử mô hình bằng cách chọn Supplied test set và chọn bộ dữ liệu Test Set. Sau đó kích chuột phải vào mô hình đã huấn luyện chọn Re-evaluate model on current test set để kiểm thử lại với bộ dữ liệu Test Set.

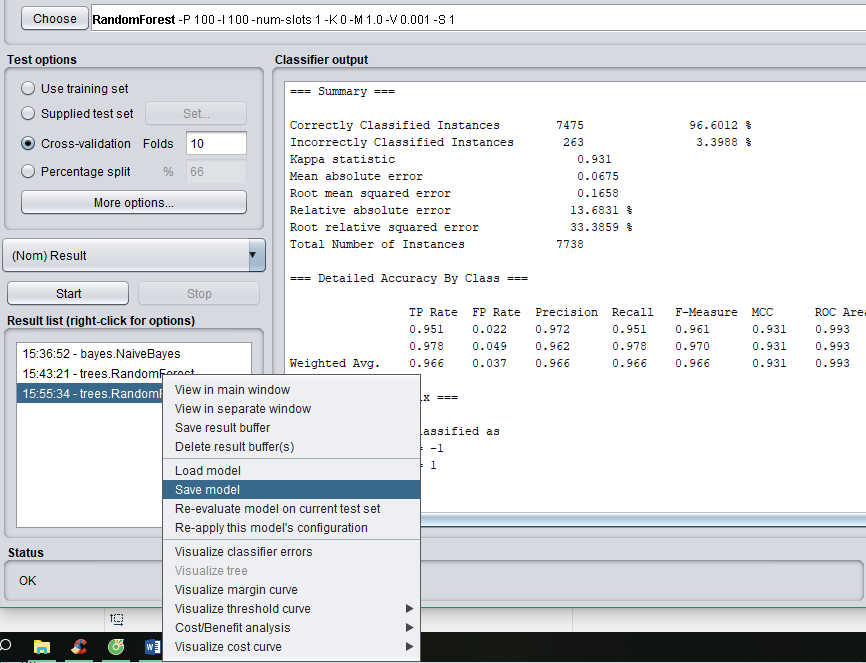


Hình 3. 6. Kết quả kiểm thử với Test Set của mô hình phân lớp Random Forest.

#### 3.1.3. Thiết lập mô hình đánh giá.

Trước khi đi vào thiết lập thí nghiệm mô hình đánh giá. Chúng ta cần chuẩn bị các mô hình (model) huấn luyện với 10-fold Cross Valication của các bộ phân lớp cơ bản và Random Forest. Các bước chuẩn bị model như sau:

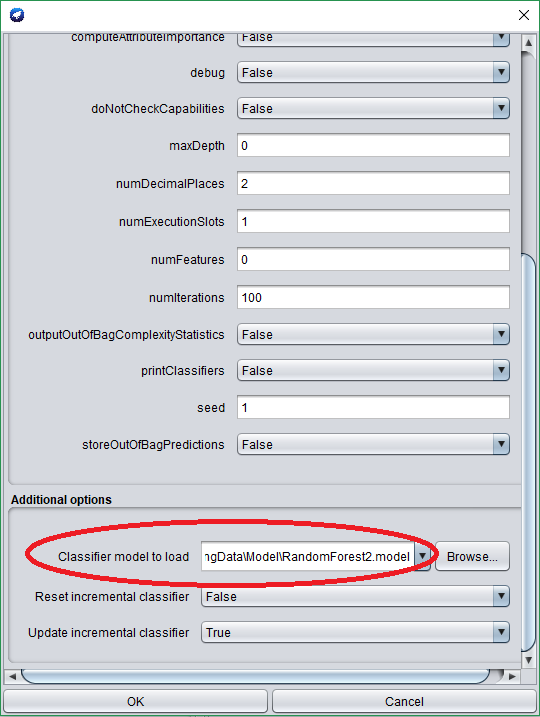
* Bước 1: Tiến hành huấn luyện Training Set với các mô hình cơ bản và Random Forest.
* Bước 2: Sau khi huấn luyện xong mỗi mô hình, chúng ta tiến hành kích chuột phải vào mô hình và chọn Save model.



Hình 3. 7. Lưu mô hình huấn luyện 10-fold Cross Valication.

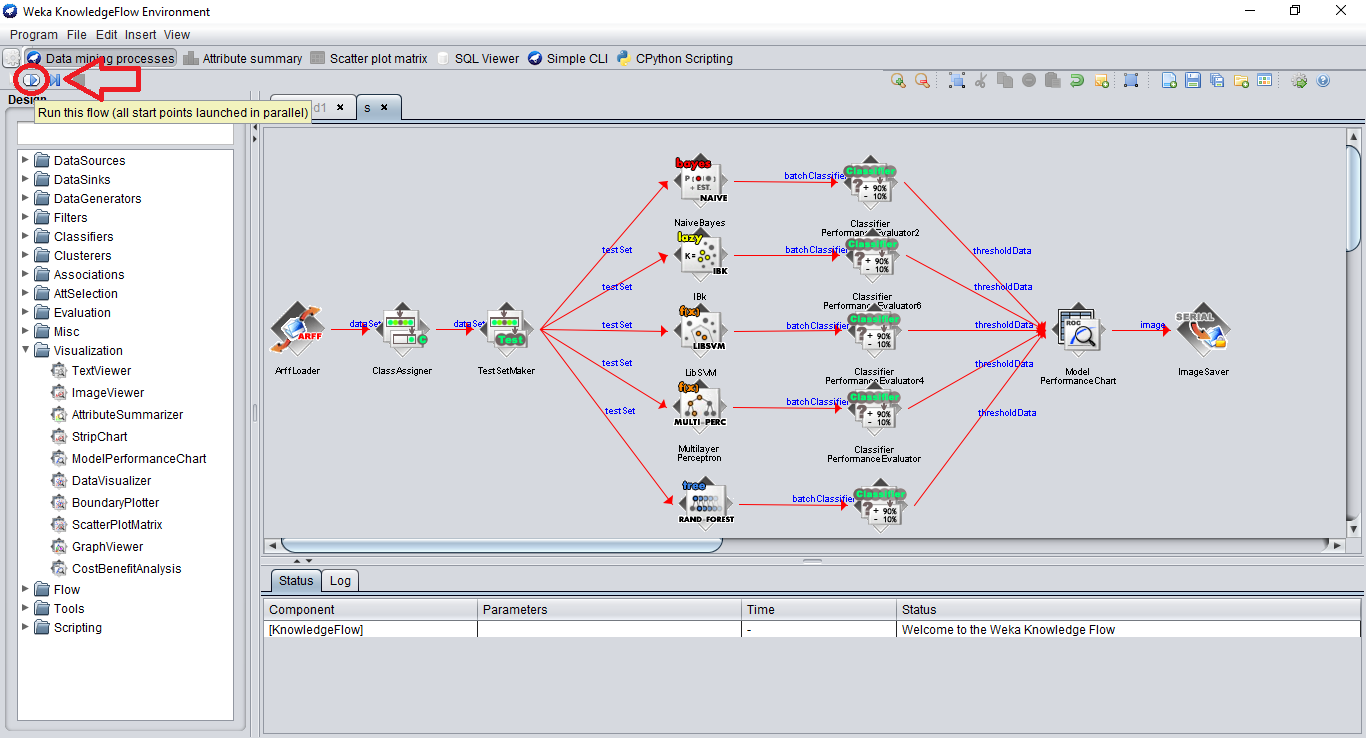
Sau khi đã lưu xong các model của các bộ phân lớp, chúng ta sử dụng công cụ Knowledge Flow của Weka để tiến hành vẽ đường ROC.

Các bước để vẽ đường ROC là các bước ở mục 2.3.2. Xây dựng ROC để đánh giá. Ở đây, chúng ta thêm vào các model ở trên vào các bộ phân lớp bằng cách kích đúp chuột vào mô hình phân lớp, tìm đến mục Classifier model to load nhấn Browse và tìm đến mục chứa model ứng với mô hình đó.



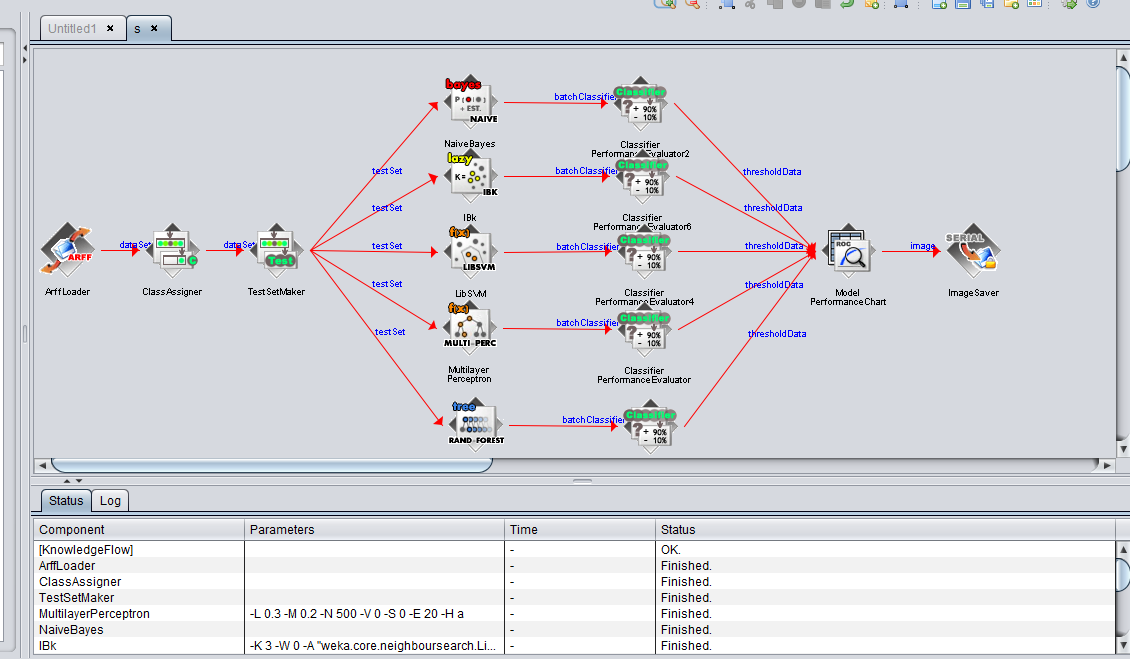
Hình 3. 8. Thêm model vào mô hình phân lớp.

Cuối cùng, chúng ta chọn Run this flow để tiến hành đánh giá.

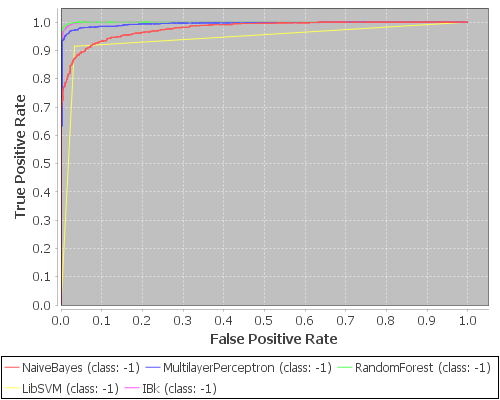


Hình 3. 9. Tiến hành đánh giá mô hình.

Sau thời gian tiến hành đánh giá, nếu dòng Status của Knowledge Flow hiện “OK”, tức quá trình đánh giá đã xong, chúng ta tìm đến thư mục lưu ảnh ở ImageSaver để có thể xem đường ROC.



Hình 3. 10. Đánh giá mô hình thành công.



Hình 3. 11. Kết quả đánh giá mô hình đường ROC trên bộ dữ liệu Phishing Website.

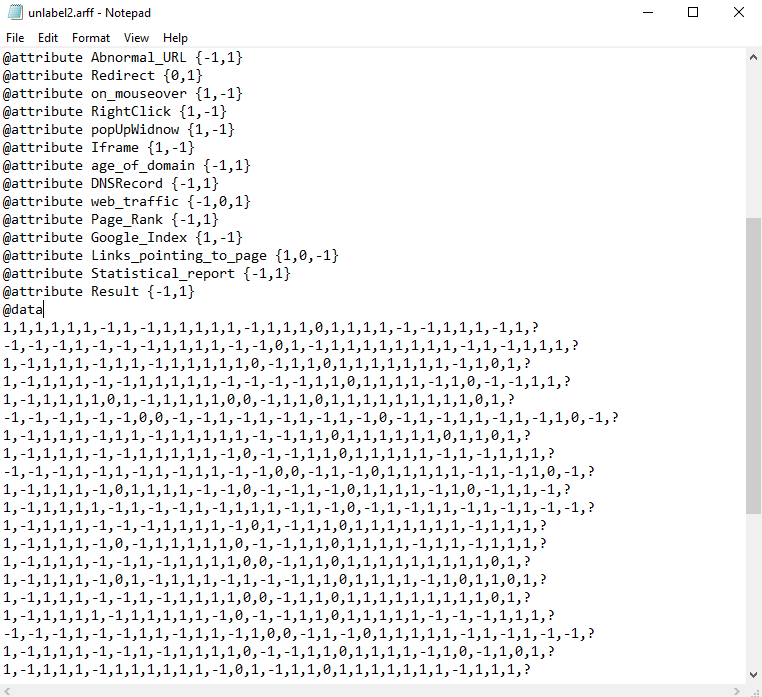
#### 3.1.4. Thiết lập ứng dụng dự đoán Website lừa đảo.

* + - 1. Chuẩn bị bộ dữ liệu Unlabel.

Unlabel Set là bộ dữ liệu được tách ra từ Test Set và xóa đi nhãn lớp của các trường hợp.

Các bước chuẩn bị bộ dữ liệu Unlabel khá đơn giản.

* Bước 1: Mở file TestSet.arff bằng Notepad. Copy từ dòng đầu tiên đến @data vào một tệp Notepad khác.
* Bước 2: Chọn ngẫu nhiên các trường hợp ở bộ Test Set và copy sang Notepad còn lại.
* Bước 3: Tiến hành chuyển giá trị nhãn lớp (thường nằm ở cuối cùng) ở đây là Result về dấu “?” và lưu lại.

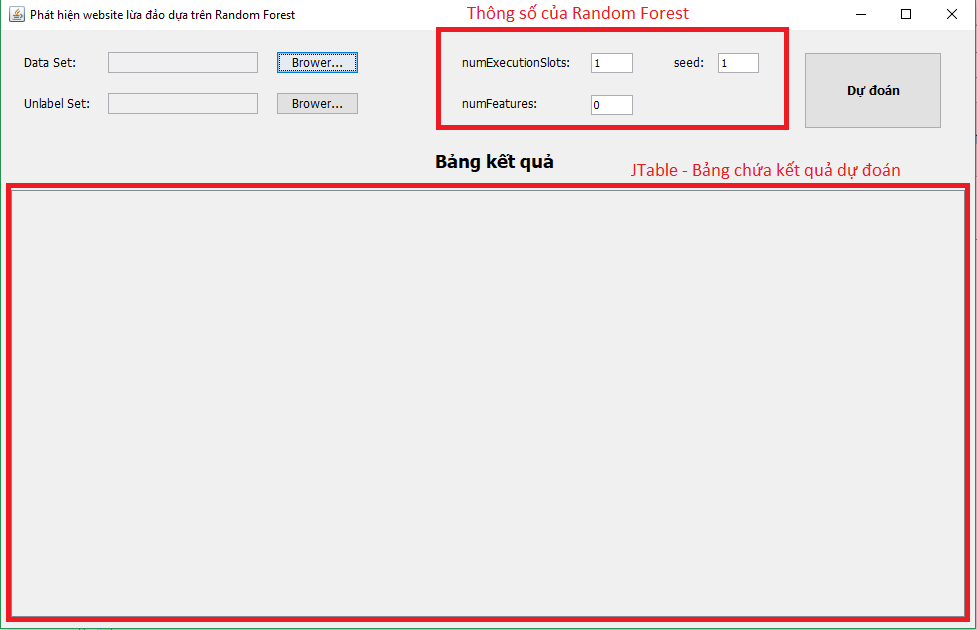


Hình 3. 12. Bộ dữ liệu Unlabel.

* + - 1. Tạo ứng dụng sử dụng Java Swing.

Tạo giao diện ứng dụng sử dụng Java Swing (Neatbeans IDE).

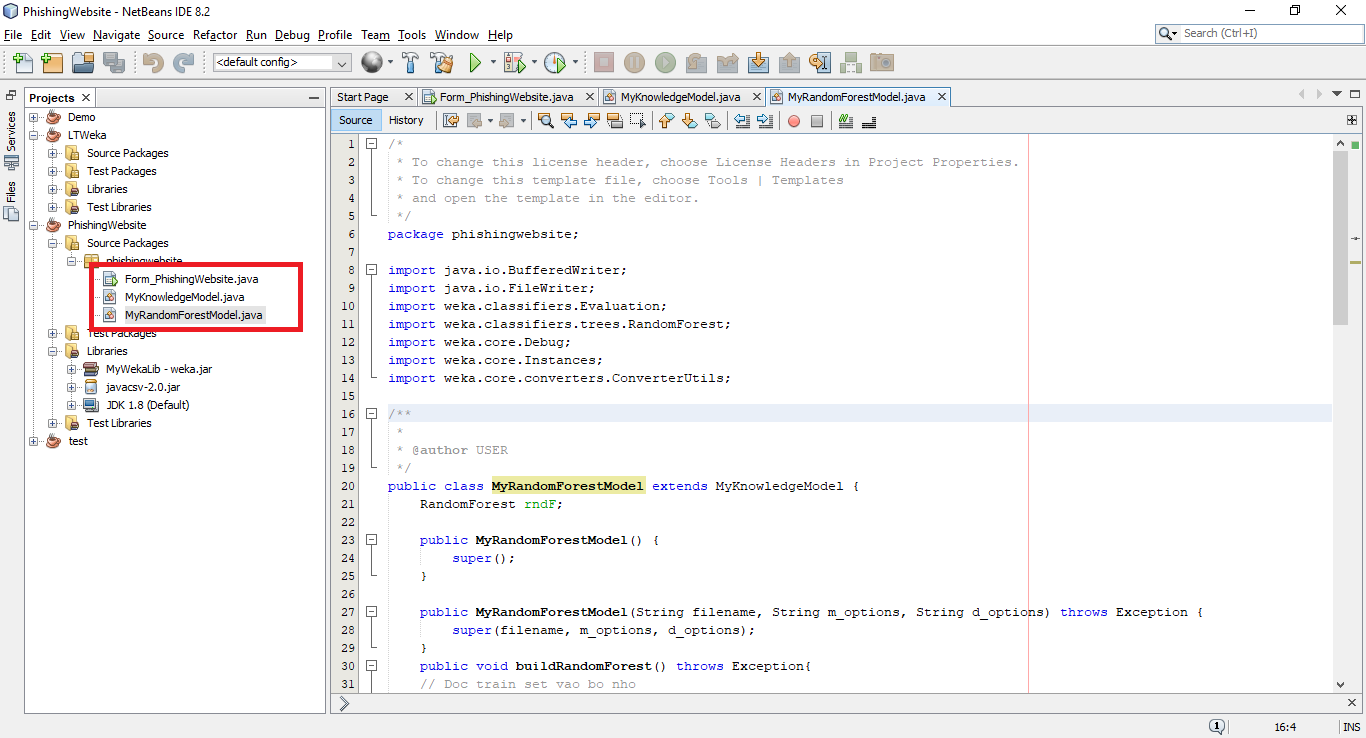
Dưới đây là hình ảnh giao diện của ứng dụng.



Hình 3. 13. Giao diện của ứng dụng.

Sau khi tạo xong giao diện, chúng ta tiến hành viết code cho ứng dụng.

Ở đây chúng tôi có hai Java Class: MyKnowledgeModel (lớp cha), MyRandomForestModel (lớp con) và một JFrame.

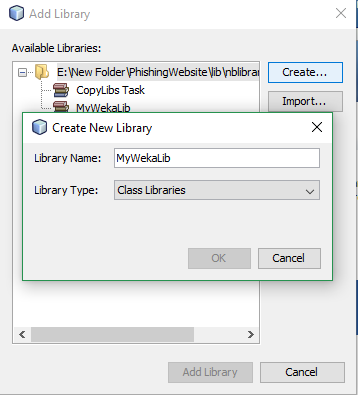


Hình 3. 14. Các Java Class và Jframe của ứng dụng.

Ở đây có một lưu ý là để tiến hành phân lớp cũng như dự đoán bộ dữ liệu chúng ta cần thêm vào thư viện của Weka trong Libraries của chương trình.

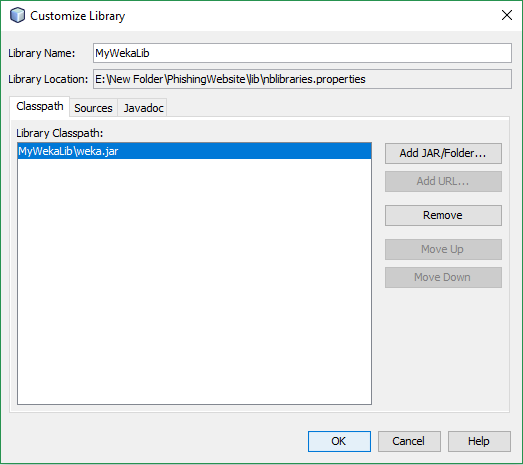
Các bước thêm thư viện như sau:

* Bước 1: Chọn Libraries, nhấn chuột phải vào chọn Add Library, tiến hành tạo mới một thư viện.



Hình 3. 15. Tạo mới thư viện trong Netbeans.

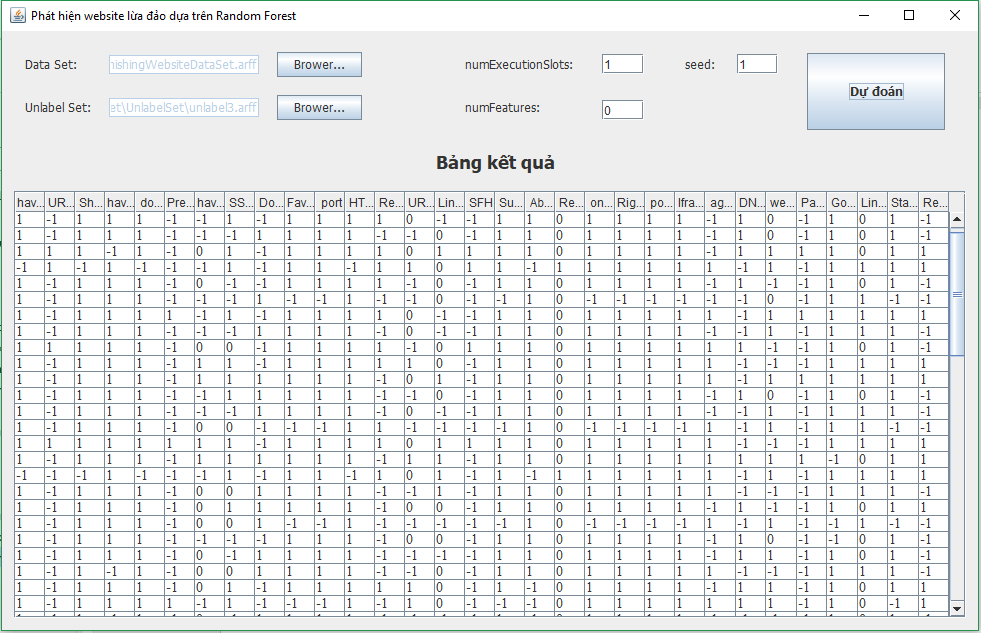
* Bước 2: Ấn Add JAR/Folder vào chọn các thư việc cần thiết để thêm vào. Ở mục Classpath chúng ta thêm file weka.jar, mục Sources thêm file weka-src.jar tại Javadoc thêm thư mục doc và cuối cùng ấn OK để lưu lại.



Hình 3. . Thêm file .jar và folder doc của Weka vào thư viện.

Cuối cùng chúng ta tiến hành chạy chương trình để tiến hành dự đoán.

Ở mục Data Set chúng ta thêm vào bộ dữ liệu Phishing Website, mục Unlabel Set thêm vào bộ dữ liệu Unlabel. Sau đó thiết lập các thông số của mô hình phân lớp RF và ấn Dự đoán để tiến hành chạy chương trình. Sau chạy xong, chương trình sẽ hiện thông báo “Dự đoán thành công” sau đó ấn OK để xem kết quả dự đoán ở Bảng kết quả.



Hình 3. 17. Kết quả của chương trình sau khi dự đoán.

### **3.2. Kết quả thí nghiệm.**

Bảng 3.1. Bảng kết quả thí nghiệm so sánh các kết quả các mô hình.

|  |  |  |
| --- | --- | --- |
| **Mô hình** | **Thành công** | **Thất bại** |
| Naive Bayes | 92.1616 % | 7.8384% |
| IBk (k =3) | 95.2367 % | 4.7633% |
| LibSVM | 93.4278 % | 6.5722% |
| MultilayerPerceptron | 97.2867 % | 2.7133% |
| **RandomForest** | **97.5882 %** | **2.4118%** |

# **PHẦN III. KẾT LUẬN**

1. **Kết quả đạt được.**

Đề tài đã đạt được mục tiêu đề ra ban đầu là thu thập dữ liệu, tiến hành chia bộ dữ liệu thành hai bộ huấn luyện và kiểm thử, sau đó huấn luyện, kiếm thử dữ liệu với các bộ phân lớp và đánh giá mô hình sử dụng ROC trên Weka.

Tiếp theo, tạo các bộ dữ liệu không nhãn (Unlabel set), xây dựng một ứng dụng trên Java Swing với chức năng dự đoán nhãn lớp cho các bộ dữ liệu đó.

Qua kết quả thí nghiệm, cho thấy rằng mô hình phân lớp Random Forest cho kết quả cao hơn các mô hình phân lớp cơ sở được đem ra so sánh là Naïve Bayes, K-nn, SVM và MultilayerPerceptron.

Từ đó phân tích được tính khả thi của việc sử dụng mô hình phân lớp Random Forest trong việc dự đoán Website lừa đảo.

1. **Hạn chế của đề tài.**

Trong quá trình thực hiện đề tài đến lúc kết thúc, do thời gian và kiến thức của bản thân có hạn nên đề tài vẫn chưa hoàn thiện hết các tính năng cũng như khai thác hết lĩnh vực khai phá dữ liệu. Những kiến thức sử dụng trong đề tài là những kiến thức cơ bản về cách huấn luyện và kiểm thử mô hình trên Weka cũng như thiết lập ứng dụng trên Java Swing.

Bộ dữ liệu nghiên cứu cũng chưa đầy đủ hết những tính năng của Website để phát hiện những Website lừa đảo.

1. **Hướng phát triển của đề tài.**

Đầu tiên xây dựng một bộ dữ liệu bao gồm các tính năng tối ưu nhất trong việc dự đoán trang web lừa đảo. Sau đó phát triển ứng dụng sử dụng mô hình phân lớp Random Forest để tiến hành dự đoán các trang web lừa đảo. Ngoài ra có thể tiến hành nghiên cứu thêm một số mô hình phân lớp khác như Bagging, Voting, Stacking hay Boosting. Bên cạnh việc phát hiện trang web lừa đảo, có thể mở rộng ra các lĩnh vực khác như thương mại điện tử, ngân hàng, chứng khoáng và y học.

# **DANH MỤC TÀI LIỆU THAM KHẢO**

1. Data mining là gì? Tìm hiểu về các công cụ khai phá dữ liệu phổ biến. Techblog của VCCloud, <https://tech.vccloud.vn/news-20181117104818209.htm>, accessed: 18/02/2019.

2. Nguyễn H.N., Nguyễn T.T., và Hà Q.H. Giáo trình KHAI PHÁ DỮ LIỆU. .

3. iamvtn (2016). [ML] Tìm hiểu về Naive Bayes Classification (phân loại Bayes đơn giản). Thành Nam’s Blog, <https://batnamv.wordpress.com/2016/08/03/ml-tim-hieu-ve-naive-bayes-classification-phan-loai-bayes-don-gian/>, accessed: 10/12/2017.

4. Thuật toán K láng giềng gần nhất - BIS. <http://bis.net.vn/forums/t/370.aspx>, accessed: 10/12/2017.

5. Một chút về thuật toán SVM (Support Vector Machine algorithm). Viblo, <https://viblo.asia/p/mot-chut-ve-thuat-toan-svm-support-vector-machine-algorithm-OeVKBgGAZkW>, accessed: 27/03/2019.

6. Giới thiệu tổng quan về Mạng Nơron nhân tạo (Artificial Neural Network- ANN). <http://nawapi.gov.vn/index.php?option=com\_content&view=article&id=3238%3Agii-thiu-tng-quan-v-mng-nron-nhan-to-artificial-neural-network-ann&catid=70%3Anhim-v-chuyen-mon-ang-thc-hin&Itemid=135&lang=vi>, accessed: 21/03/2019.

7. (2018). Phân tích hồi quy. Wikipedia tiếng Việt, <https://vi.wikipedia.org/w/index.php?title=Ph%C3%A2n\_t%C3%ADch\_h%E1%BB%93i\_quy&oldid=40342346>, accessed: 21/03/2019.

8. (2019). Random forest WiKi. Wikipedia, <https://en.wikipedia.org/w/index.php?title=Random\_forest&oldid=883753888>, accessed: 20/02/2019.

9. Couhp (2018). Random Forest, thế nào là một rừng ngẫu nhiên. Finding Data, <https://couhpcode.wordpress.com/2018/01/24/random-forest-the-nao-la-mot-rung-ngau-nhien/>, accessed: 18/02/2019.

10. Sahingoz O.K., Buber E., Demir O. và cộng sự. (2019). Machine learning based phishing detection from URLs. *Expert Systems with Applications*, **117**, 345–357.

11. Rao R.S. và Pais A.R. (2018). Detection of phishing websites using an efficient feature-based machine learning framework. *Neural Computing and Applications*.

12. Hutchinson S., Zhang Z., và Liu Q. (2018). Detecting Phishing Websites with Random Forest. *Machine Learning and Intelligent Communications*. Springer International Publishing, Cham, 470–479.

13. Jagadeesan S. (2018). URL Phishing Analysis using Random Forest. 6.

14. Buber E., Diri B., và Sahingoz O.K. (2017). Detecting phishing attacks from URL by using NLP techniques. *2017 International Conference on Computer Science and Engineering (UBMK)*, Antalya, IEEE, 337–342.

15. Parekh S., Parikh D., Kotak S. và cộng sự. (2018). A New Method for Detection of Phishing Websites: URL Detection. *2018 Second International Conference on Inventive Communication and Computational Technologies (ICICCT)*, Coimbatore, IEEE, 949–952.

16. Subasi A., Molah E., Almkallawi F. và cộng sự. (2017). Intelligent phishing website detection using random forest classifier. *2017 International Conference on Electrical and Computing Technologies and Applications (ICECTA)*, 1–5.

17. Lee J.L., Park D.H., và Lee C.H. (2015). Phishing Detection Methodology Using Web Sites Heuristic. *KIPS Transactions on Computer and Communication Systems*, **4**(**10**), 349–360.

18. Chiew K.L., Tan C.L., Wong K. và cộng sự. (2019). A new hybrid ensemble feature selection framework for machine learning-based phishing detection system. *Information Sciences*, **484**, 153–166.

19. Buber E., Diri B., và Sahingoz O.K. (2018). NLP Based Phishing Attack Detection from URLs. *Intelligent Systems Design and Applications*. Springer International Publishing, Cham, 608–618.

20. Weedon M., Tsaptsinos D., và Denholm-Price J. (2017). Random forest explorations for URL classification. *2017 International Conference On Cyber Situational Awareness, Data Analytics And Assessment (Cyber SA)*, London, United Kingdom, IEEE, 1–4.

21. Islam M. và Chowdhury N.K. (2016). Phishing Websites Detection Using Machine Learning Based Classification Techniques. 4.

22. Smadi S., Aslam N., Zhang L. và cộng sự. (2015). Detection of phishing emails using data mining algorithms. *2015 9th International Conference on Software, Knowledge, Information Management and Applications (SKIMA)*, Kathmandu, Nepal, IEEE, 1–8.

23. Kazi R. (2016). Evaluation of Machine Learning Classifiers for Phishing Detection. .

24. Yaokai Y. (2019), *Effective Phishing Detection Using Machine Learning Approach*, Case Western Reserve University.

25. Zakariah M. (2014). Classification of large datasets using Random Forest Algorithm in various applications: Survey. **4**(**3**), 11.

26. Kumar Nepali R. và Wang Y. (2016). You Look Suspicious!! : Leveraging Visible Attributes to Classify Malicious Short URLs on Twitter. <https://ieeexplore.ieee.org/abstract/document/7427515>.

27. Hodžić A. và Kevrić J. (2016). COMPARISON OF MACHINE LEARNING TECHNIQUES IN PHISHING WEBSITE CLASSIFICATION. *Regional Economic Development*, 8.

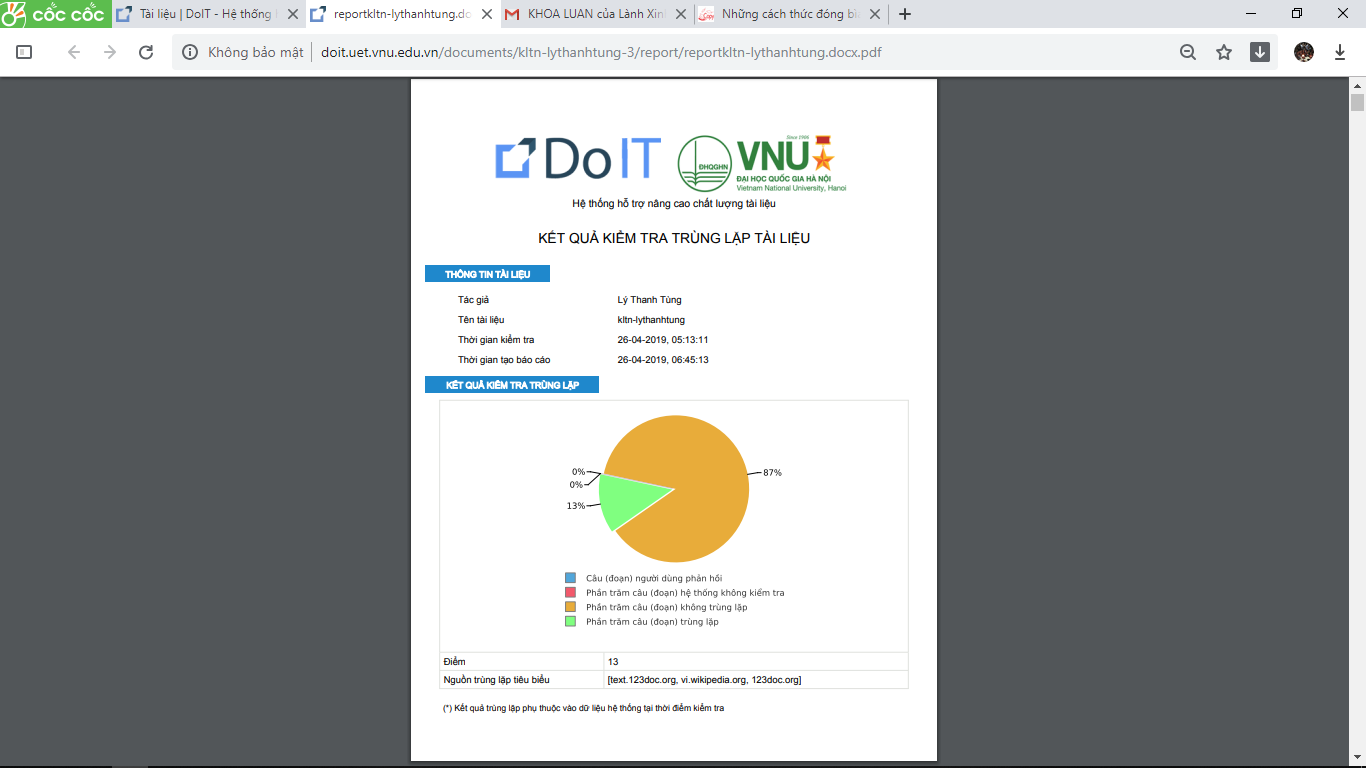
28. Yasin A. và Abuhasan A. (2016). An Intelligent Classification Model for Phishing Email Detection. *International Journal of Network Security & Its Applications*, **8**(**4**), 55–72.

29. Akinyelu A. A. và Adewumi O. A. (2014). Classification of Phishing Email Using Random Forest Machine Learning Technique. Journal of Applied Mathematics, 2014, 1–6 | 10.1155/2014/425731. <http://sci-hub.tw/https://www.hindawi.com/journals/jam/2014/425731/abs/>, accessed: 01/03/2019.

**KẾT QUẢ KIỂM TRA ĐẠO VĂN**

****

**KẾT QUẢ KIỂM TRA LỖI CHÍNH TẢ [[2]](#footnote-2)**



1. 1 Nguồn: https://archive.ics.uci.edu/ml/datasets/Phishing+Websites#. [↑](#footnote-ref-1)
2. Nguồn: http://doit.uet.vnu.edu.vn/documents/kltn-lythanhtung-3/report/reportkltn-lythanhtung.docx.pdf [↑](#footnote-ref-2)