**PHÁT HIỆN TIN TỨC GIẢ VÀ THẬT VỚI**

**MÔ HÌNH HỌC MÁY**

Học viện Chính sách và Phát triển

|  |  |  |  |
| --- | --- | --- | --- |
| **NHÓM 8** | | | |
| **Thành viên** | **Mã sinh viên** | **Nhiệm vụ** | **Chấm điểm** |
| Dương Quốc Anh  (Nhóm trưởng) | 7123112070 | Tìm hiểu, viết Code và tài liệu tham khảo | 9 |
| Nguyễn Bùi Trung Hiếu | 7123112088 | Làm word, tìm tài liệu về các thuật toán, code | 9 |
| Vũ Đình Phúc | 7123112114 | Tìm hiểu, viết Code và tài liệu tham khảo | 9 |
| Bùi Đức Trung | 7123112122 | Slide, viết Code | 8 |
| Nguyễn Tú Anh | 7123112073 | Làm word | 8 |

1. **GIỚI THIỆU BÀI TOÁN:**

Thuật ngữ "tin giả" là một khái niệm tương đối mới và cho đến nay vẫn chưa có một định nghĩa chung được thống nhất về tin tức giả mạo hay tin giả (Fake News). Theo từ điển Oxford "*Tin giả là thông tin sai sự thật được phát sóng hoặc xuất bản dưới dạng tin tức nhằm mục đích lừa đảo hoặc có động cơ chính trị. Tin giả tạo ra sự nhầm lẫn đáng kể của công chúng về các sự kiện hiện tại. Tin giả bùng nổ trên phương tiện truyền thông xã hội, đang xâm nhập vào các kênh truyền thông chính*"[[1]](#footnote-1). Học giả về truyền thông Nolan Higdon đã định nghĩa "*Tin tức giả là nội dung sai sự thật hoặc gây hiểu lầm được trình bày dưới dạng tin tức và được truyền đạt dưới các định dạng bao gồm truyền thông nói, viết, in, điện tử và kỹ thuật số*"[[2]](#footnote-2). Tin tức giả mạo cũng đề cập đến những câu chuyện bịa đặt có rất ít hoặc không có sự thật và khó có thể xác minh được. Thậm chí rộng hơn, sau kỳ bầu cử tổng thống Mỹ năm 2020, người ta đã mở rộng ý nghĩa của "tin tức giả" để bao gồm cả các tin tức tiêu cực về niềm tin và hành động cá nhân của họ.

Trong thời đại kỹ thuật số ngày nay, sự gia tăng tin tức giả mạo trên các nền tảng truyền thông xã hội đã trở nên đáng báo động vì nó gây ra mối đe dọa cho các khía cạnh khác nhau của xã hội. Tin giả bao gồm thông tin bị làm sai lệch có chủ ý nhằm đánh lừa người đọc và thao túng nhận thức của công chúng, dẫn đến nội dung có chất lượng thấp và gây hiểu nhầm. Từ đó, sự phổ biến của thông tin sai lệch đã trở thành mối lo ngại cấp bách, tràn ngập mọi khía cạnh trong cuộc sống trực tuyến của chúng ta. Tin giả, thường được ngụy trang dưới dạng thông tin hợp pháp, có sức mạnh gây ảnh hưởng đến dư luận, kích động nỗi sợ hãi và hận thù, thậm chí ảnh hưởng đến các cuộc bầu cử. VD: Một số tin giả liên quan đến dịch COVID-19,… Vì vậy, nhu cầu về cơ chế phát hiện tin giả hiệu quả chưa bao giờ quan trọng hơn thế.

Phát hiện tin giả là quá trình xác định và vạch trần thông tin sai lệch hoặc gây hiểu lầm được phổ biến thông qua các kênh truyền thông khác nhau, đặc biệt là các nền tảng trực tuyến và mạng xã hội. Thách thức đa chiều này đòi hỏi một cách tiếp cận toàn diện, tích hợp các tiến bộ công nghệ, kỹ năng tư duy phê phán và nỗ lực hợp tác từ các cá nhân, tổ chức và chính phủ.

Đi đầu trong việc phát hiện tin tức giả là các thuật toán tiên tiến và kỹ thuật máy học. Những công nghệ này phân tích lượng dữ liệu khổng lồ, bao gồm nội dung văn bản, hình ảnh và siêu dữ liệu để xác định các mẫu biểu thị thông tin sai lệch. Ví dụ: thuật toán xử lý ngôn ngữ tự nhiên (NLP) có thể nhận ra sự mâu thuẫn về ngữ nghĩa và các tín hiệu ngôn ngữ có thể báo hiệu ý định lừa đảo. Tương tự, các công cụ phân tích hình ảnh và video sử dụng tìm kiếm hình ảnh ngược và các kỹ thuật khác để xác minh tính xác thực của nội dung trực quan, hoặc sử dụng phương pháp học máy, học sâu, sử dụng phân tích cảm tính để phát hiện các tin tức giả mạo trên các nền tảng mạng xã hội đang ngày càng gia tăng như hiện nay.

Tuy nhiên, chỉ riêng giải pháp công nghệ là không đủ để chống lại tính chất phức tạp và ngày càng phát triển của tin tức giả. Sự can thiệp của con người vẫn không thể thiếu trong quá trình xác minh. Những người xác minh sự thật và nhà báo đóng vai trò quan trọng trong việc xem xét kỹ lưỡng những tuyên bố đáng ngờ, chứng thực thông tin từ nhiều nguồn và cung cấp bối cảnh để phân biệt sự thật và sự giả dối. Hơn nữa, việc nâng cao hiểu biết về truyền thông và kỹ năng tư duy phản biện trong công chúng là điều cần thiết để trao quyền cho các cá nhân trong việc phân biệt các nguồn đáng tin cậy với những nguồn không đáng tin cậy.

Sự hợp tác giữa các bên liên quan khác nhau là điều cần thiết để chống lại sự lan truyền của tin tức giả một cách hiệu quả. Các công ty công nghệ, nền tảng truyền thông xã hội, tổ chức học thuật và cơ quan chính phủ phải hợp tác để phát triển và triển khai hệ thống phát hiện tin tức giả mạo mạnh mẽ. Điều này liên quan đến việc chia sẻ dữ liệu, tài nguyên và chuyên môn để nâng cao tính chính xác và hiệu quả của các thuật toán phát hiện đồng thời bảo vệ quyền riêng tư và quyền tự do ngôn luận của người dùng.

Hơn nữa, cần có các khuôn khổ và chính sách pháp lý để buộc những người cung cấp tin giả phải chịu trách nhiệm về hành động của họ. Các chính phủ phải ban hành luật nhằm giải quyết việc phổ biến thông tin sai lệch trong khi vẫn duy trì các nguyên tắc tự do ngôn luận và tự do báo chí. Ngoài ra, các nền tảng phải thực thi các chính sách kiểm duyệt nội dung minh bạch và chống lại các thành kiến ​​về thuật toán có thể vô tình khuếch đại thông tin sai lệch.

Tóm lại, việc phát hiện tin giả là một thách thức nhiều mặt đòi hỏi nỗ lực phối hợp từ các nhà phát triển công nghệ, người kiểm chứng sự thật, nhà báo, nhà hoạch định chính sách và công chúng nói chung. Bằng cách tận dụng các công nghệ tiên tiến, nâng cao hiểu biết về truyền thông, thúc đẩy hợp tác và thực hiện các biện pháp quản lý hiệu quả, chúng ta có thể giảm thiểu tác hại của tin giả và duy trì tính toàn vẹn của thông tin trong thời đại kỹ thuật số. Chỉ thông qua hành động tập thể, chúng ta mới có thể vén màn bức màn lừa dối và đề cao các nguyên tắc về sự thật và minh bạch trong xã hội của chúng ta.

1. **NGHIÊN CỨU LIÊN QUAN:**

Trong quá khứ đã có nhiều bài nghiên cứu đưa ra các đề xuất phương pháp, mô hình nhằm phát hiện tin giả. Để xác định được đâu là tin giả, tin thật cần phải sử dụng nhiều phương pháp cả truyền thống lẫn sử dụng các công nghệ hiện đại hiện nay. Một số kết quả từ các bài nghiên cứu trong quá khứ như sau:

* Các nhà nghiên cứu đang khám phá cách thực hiện tự động hóa quá trình xác minh thông tin, sử dụng các công nghệ dựa trên xử lý ngôn ngữ tự nhiên (NLP) và học máy để tự động dự đoán tính xác thực của báo cáo Mayank và cộng sự: “DEAP-FAKED: Phương pháp tiếp cận dựa trên Sơ đồ tri thức để phát hiện tin tức giả mạo” [[3]](#footnote-3)[3] đã đề xuất một khung dựa trên biểu đồ tri thức (KG) bao gồm bộ mã hóa tin tức, bộ mã hóa thực thể và lớp phân loại cuối cùng để phát hiện các bài báo giả mạo. Đầu tiên, công nghệ dựa trên NLP được sử dụng để mã hóa tiêu đề của tin tức. Sau đó, nhận dạng thực thể được đặt tên (NER) được áp dụng để nhận dạng và trích xuất các thực thể được đặt tên từ văn bản, sau đó, định hướng thực thể được đặt tên (NED) được áp dụng để ánh xạ các thực thể tới KG  và Tin tức giả trên các nền tảng mạng xã hội đã thu hút rất nhiều sự chú ý đó làcác sự kiện liên quan đến chính trị (bầu cử Tổng thống Hoa Kỳ năm 2016) và chăm sóc sức khỏe (đại dịch thông tin trong thời kỳ COVID-19), DEAP-FAKED đạt được điểm F1 là 88% và 78% cho hai tập dữ liệu, mức cải thiện tương ứng là ~ 21% và ~ 3%, cho thấy tính hiệu quả của phương pháp này.
* Hu và cộng sự trong bài nghiên cứu: “Compare to The knowledge: Graph Neural Fake News Detection with External Knowledge”[[4]](#footnote-4) [4]. Mô hình ComplEx được áp dụng để thu được sự biểu diễn của các thực thể. Cuối cùng, việc nhúng hai phần được kết hợp để phát hiện tin giả. Hu và cộng sự đã thiết kế một mô hình thần kinh đồ thị mới, so sánh tin tức với cơ sở kiến ​​thức bên ngoài. Đầu tiên, một biểu đồ tài liệu không đồng nhất có hướng chứa các chủ đề và thực thể cho từng loại tin tức được xây dựng. Sau đó, thông qua mạng so sánh thực thể được thiết kế cẩn thận, các biểu diễn thực thể khác nhau sẽ được so sánh.
* Pan và cộng sự đề xuất phương pháp phát hiện tin giả bằng phương pháp TransE và B-TransE trong bài nghiên cứu: “Phát hiện tri thức giả dựa trên nội dung bằng sơ đồ tri thức”[[5]](#footnote-5). Đầu tiên, ba biểu đồ tri thức khác nhau được tạo ra để tạo ra kiến ​​thức nền tảng. Sau đó, phương pháp B-TransE được sử dụng để thiết lập việc nhúng thực thể và mối quan hệ cũng như kiểm tra xem các bài báo có xác thực hay không. Phương pháp tiếp cận của họ được đánh giá bằng tập dữ liệu 'Nhận biết sự thật về tin tức giả' của Kaggle và một số bài viết xác thực từ các phương tiện truyền thông chính thống. Các đánh giá cho thấy một số phương pháp tiếp cận của họ có điểm F1 trên 0,80.
* Trong nghiên cứu của **Mohammad Javad Hosseini: “Đánh giá các kỹ thuật học máy và học sâu để phát hiện tin tức giả”**[[6]](#footnote-6)**:**

Bài báo này trình bày một phương pháp dựa trên học máy để phát hiện tin giả bằng cách sử dụng:

1. Kiến thức về Sentiment (Khuynh hướng): tính toán điểm số khuynh hướng của một câu làm một trong những đặc điểm cho phân loại cấp câu.
2. Kiến thức về thống kê và ngôn ngữ: để trích xuất các đặc điểm cần thiết.

Phương pháp được đề xuất kết hợp nhiều loại đặc điểm thành một bộ đặc điểm thống nhất để thiết kế một hệ thống phân loại chính xác hơn (“True”: tin giả; “False”: không phải tin giả). Do đó, để đạt được điểm hiệu suất tốt hơn họ đã tiến hành nghiên cứu hiệu suất của năm kỹ thuật chọn lọc đặc điểm nổi tiếng và một số bộ phân loại nổi tiếng nhất để lựa chọn bộ đặc điểm có liên quan nhất và tìm ra bộ phân loại Học máy (ML) hiệu quả.

Cùng với các mô hình học máy truyền thống cũng áp dụng các mô hình Học sâu (DL) khác nhau. Các phương pháp học máy và học sâu khác nhau được áp dụng cho ba bộ dữ liệu khác nhau bao gồm Liar, GossipCop và PolitiFact, và kết quả cho thấy việc tích hợp phương pháp phân loại dựa trên học máy và kỹ thuật chọn lọc đặc điểm có thể cải thiện đáng kể hiệu suất và khiến phương pháp này có thể sánh ngang với các phương pháp hiện có khác. Hơn nữa, các mô hình DL, đặc biệt là LSTM, đã thể hiện ngưỡng độ chính xác đáng chú ý lên đến 88% trong việc nắm bắt các mẫu phức tạp trong dữ liệu văn bản.

* Rathika Natarajan và cộng sự trong bài nghiên cứu: “Intelligent gravitational search randoom forest algorithm for fake new detection”[[7]](#footnote-7). Bài báo này đề xuất thuật toán Randoom forest tìm kiếm hấp dẫn (IGSRF) thông minh được sử dụng để phát hiện tin giả. Thuật toán IGSRF kết hợp Thuật toán tìm kiếm hấp dẫn thông minh (IGSA) và thuật toán Randoom forest (RF). IGSA là một biến thể thông minh cải tiến của thuật toán tìm kiếm hấp dẫn cổ điển (GSA), bổ sung thông tin về các tác nhân khối lượng hấp dẫn tốt nhất và tồi tệ nhất nhằm duy trì khả năng khai thác của các tác nhân ở các lần lặp sau và do đó tránh được bẫy của GSA cổ điển trong cục bộ. tối ưu. Trong thuật toán IGSRF được đề xuất, tất cả các tác nhân đại chúng thông minh xác định giải pháp bằng cách tạo ra cây quyết định (DT) với một tập hợp con các thuộc tính ngẫu nhiên theo giả thuyết về Randoom forest. Các tác nhân đại chúng tạo ra tập hợp các nghiệm từ không gian nghiệm bằng cách sử dụng các quy tắc tỷ lệ ngẫu nhiên. Dự đoán toàn diện để quyết định loại tin tức (giả hay thật) được xác định bởi tất cả các tác nhân theo thuộc tính của Randoom forest. Hiệu suất của thuật toán đề xuất được xác định cho tập dữ liệu FakeNewsNet, có các danh mục phụ là danh mục tin tức BuzzFeed và PolitiFact. Để phân tích tính hiệu quả của thuật toán đề xuất, kết quả cũng được đánh giá bằng thuật toán cây quyết định và Randoom forest. Thuật toán IGSRF đề xuất đã đạt được kết quả vượt trội so với các kỹ thuật DT, RF và tiên tiến nhất.

1. **PHƯƠNG PHÁP NGHIÊN CỨU:**
2. **Gradient Boosting Classifier (GBC):**

* Gradient Boosting Classifier (GBC) là một thuật toán trong học máy được sử dụng để phân loại. Nó thuộc nhóm các thuật toán "boosting", tức là kết hợp nhiều mô hình yếu (weak learner) để tạo thành một mô hình mạnh (strong learner) có khả năng dự đoán chính xác hơn.
* Cách thức hoạt động của GBC:
* Khởi tạo: Bắt đầu với một mô hình yếu, thường là một cây quyết định đơn giản. Mô hình này được huấn luyện trên dữ liệu và dự đoán kết quả ban đầu.
* Tính toán gradient : Sau đó, tính toán gradient (độ dốc) của hàm mất mát (loss function) - hàm đo lường mức độ sai lệch của dự đoán. Gradient cho biết hướng cần điều chỉnh để cải thiện độ chính xác của mô hình.
* Tạo mô hình yếu tiếp theo: Dựa trên gradient, tạo ra một mô hình yếu mới nhằm học hỏi chính xác những lỗi sai của mô hình trước. Mô hình yếu mới này tập trung vào các mẫu dữ liệu mà mô hình trước dự đoán sai.
* Kết hợp các mô hình): Kết hợp dự đoán của tất cả các mô hình yếu theo một trọng số nhất định. Trọng số được tính toán để ưu tiên các mô hình yếu có hiệu suất tốt hơn.
* Lặp lại: Lặp lại các bước 2-4 cho đến khi đạt được độ chính xác mong muốn hoặc đạt đến số lượng mô hình yếu tối đa.
* Ưu điểm của Gradient Boosting Classifier:
* Hiệu quả cao: GBC có thể đạt được độ chính xác cao trên nhiều tập dữ liệu khác nhau.
* Linh hoạt: Có thể sử dụng với nhiều loại dữ liệu đầu vào khác nhau, không chỉ dữ liệu văn bản.
* Khả năng chống quá khớp (overfitting): Bằng cách tập trung vào các lỗi sai của các mô hình trước, GBC có thể giúp giảm thiểu overfitting.
* Nhược điểm của Gradient Boosting Classifier:
* Tính toán phức tạp: Việc huấn luyện nhiều mô hình yếu theo chuỗi có thể tốn nhiều thời gian tính toán.
* Tính dễ diễn giải: Kết quả của GBC có thể khó diễn giải do là sự kết hợp của nhiều mô hình yếu.

Tóm lại, Gradient Boosting Classifier là một thuật toán phân loại mạnh mẽ và linh hoạt, được sử dụng rộng rãi trong học máy.

1. **Decision Tree Classification:**

* Decision Tree Classification (DTC) là một thuật toán học máy được sử dụng để giải quyết các bài toán phân loại. Nó hoạt động bằng cách xây dựng một mô hình phân cấp giống như cây, với các nhánh đại diện cho các quyết định và lá cây đại diện cho các phân loại cuối cùng.

- Cấu trúc của Decision Tree:

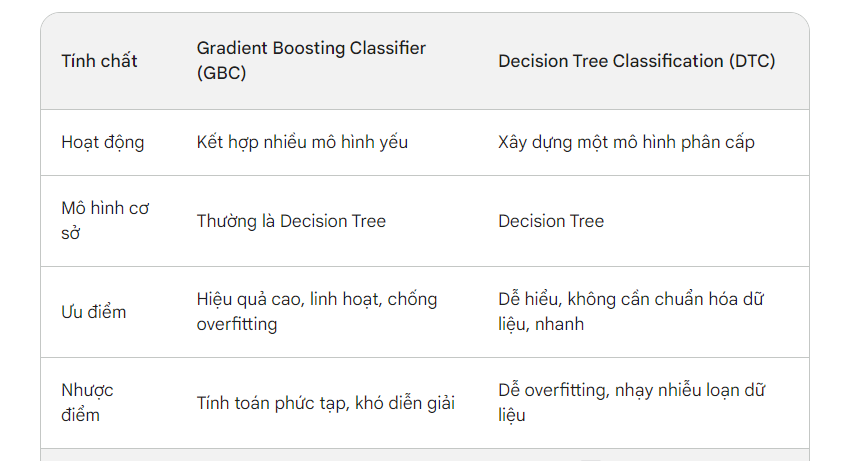
* Root Node (Nút gốc): Là nút đầu tiên của cây, chứa toàn bộ dữ liệu huấn luyện.
* Internal Nodes (Nút trong): Đại diện cho các thuộc tính (feature) của dữ liệu. Mỗi nút đặt ra một câu hỏi về một thuộc tính cụ thể.
* Leaf Nodes (Nút lá): Đại diện cho các lớp (class) phân loại cuối cùng.
* Quyết định: Tại mỗi nút trong, thuật toán sẽ chọn thuộc tính tốt nhất để phân chia dữ liệu thành các nhóm con. Thuộc tính tốt nhất thường là thuộc tính giúp phân tách dữ liệu rõ ràng nhất theo các lớp phân loại.
* Cách thức hoạt động của DTC:
* Huấn luyện: Thuật toán bắt đầu với toàn bộ dữ liệu huấn luyện tại nút gốc.
* Tách dữ liệu: Tại mỗi nút trong, thuật toán chọn thuộc tính tách dữ liệu thành các nhóm con đồng nhất về mặt phân loại nhất có thể.
* Lựa chọn thuộc tính tách: Thuật toán sử dụng các hàm đo lường như Entropy hoặc Gini Impurity (Độ không tinh khiết Gini) để chọn thuộc tính giúp phân chia dữ liệu hiệu quả nhất.
* Dừng: Quá trình phân chia dừng lại khi đạt đến một trong các điều kiện:
  + Tất cả các mẫu dữ liệu tại một nút đều thuộc cùng một lớp (lá).
  + Không thể tìm thấy thuộc tính nào để phân chia dữ liệu hiệu quả hơn nữa.
* Phân loại: Khi gặp một mẫu dữ liệu mới (dữ liệu cần dự đoán), thuật toán sẽ đưa mẫu đó đi qua các nhánh của cây theo các câu hỏi được đặt ra tại mỗi nút. Dựa vào các câu trả lời (thuộc tính của mẫu dữ liệu), đi đến lá cây cuối cùng và gán nhãn phân loại cho mẫu dữ liệu đó.
* Ưu điểm của Decision Tree Classification:
* Dễ hiểu và minh bạch: Kết quả phân loại của DTC dễ dàng diễn giải bằng cách theo dõi các nhánh của cây.
* Không cần chuẩn hóa dữ liệu: DTC có thể hoạt động tốt với nhiều loại dữ liệu thô khác nhau.
* Hiệu quả tính toán: Huấn luyện và dự đoán với DTC thường nhanh chóng.

- Nhược điểm của Decision Tree Classification:

* Dễ bị quá khớp (overfitting): Nếu cây quá phức tạp, nó có thể học thuộc lòng dữ liệu huấn luyện thay vì học được các quy tắc phân loại tổng quát.
* Nhạy với nhiễu loạn dữ liệu: Một thay đổi nhỏ trong dữ liệu có thể dẫn đến thay đổi lớn trong cấu trúc của cây.

Tóm lại, Decision Tree Classification là một thuật toán phân loại đơn giản, dễ hiểu và hiệu quả tính toán. Tuy nhiên, cần lưu ý đến khả năng overfitting và nhạy cảm với nhiễu loạn dữ liệu khi sử dụng DTC.

So sánh Decision Tree Classification (DTC) với Gradient Boosting Classifier (GBC):

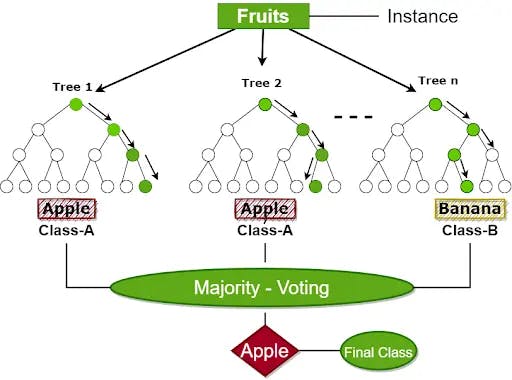


* Môi trường sử dụng:
* GBC thường được sử dụng khi cần độ chính xác cao và tính linh hoạt với nhiều loại dữ liệu.
* DTC được sử dụng khi cần mô hình dễ hiểu và nhanh chóng, nhưng cần lưu ý đến trang bị quá mức (overfitting) và nhiễu loạn dữ liệu.

1. **Randoom Forest Classifier (RFC):**

* Random Forest Classifier (RFC) là một thuật toán phân loại được sử dụng rộng rãi trong học máy. Nó thuộc nhóm thuật toán ensemble learning, tức là kết hợp nhiều mô hình yếu (weak learner) để tạo thành một mô hình mạnh (strong learner) có khả năng dự đoán chính xác hơn. Trong trường hợp của RFC, các mô hình yếu được sử dụng là Decision Tree (Cây quyết định).
* Cách thức hoạt động của RFC:
* Tạo rừng cây (Forest of Trees): Thuật toán bắt đầu bằng việc tạo ra một tập hợp ngẫu nhiên gồm nhiều cây quyết định (thường là hàng trăm hoặc hàng nghìn cây).
* Bootstrap Aggregation (Bagging): Mỗi cây quyết định được huấn luyện trên một tập hợp con (subset) được lấy có thay thế (with replacement) từ dữ liệu gốc (bagging). Đây là điểm khác biệt chính giữa Random Forest và Gradient Boosting.
* Random Feature Selection: Tại mỗi nút phân chia trong mỗi cây, chỉ xem xét một tập hợp con ngẫu nhiên của các thuộc tính (feature) để chọn thuộc tính phân chia tốt nhất. Điều này giúp tăng thêm tính đa dạng cho các cây trong rừng.
* Dự đoán: Khi gặp một mẫu dữ liệu mới (dữ liệu cần dự đoán), RFC đưa mẫu đó qua tất cả các cây trong rừng. Mỗi cây sẽ dự đoán một nhãn phân loại.
* Lập kết quả: Cuối cùng, RFC kết hợp các dự đoán của tất cả các cây theo phương thức majority vote (bỏ phiếu đa số). Lớp nhận được nhiều phiếu nhất sẽ được chọn làm nhãn phân loại cuối cùng cho mẫu dữ liệu mới.
* Ưu điểm của Random Forest Classifier:
* Hiệu quả cao: RFC thường đạt được độ chính xác cao trên nhiều tập dữ liệu khác nhau.
* Khả năng chống quá khớp (overfitting): Do sử dụng nhiều cây được huấn luyện trên các tập con dữ liệu khác nhau và lựa chọn ngẫu nhiên các thuộc tính để phân chia, RFC có khả năng chống overfitting tốt hơn so với Decision Tree đơn lẻ.
* Không cần chuẩn hóa dữ liệu: RFC có thể hoạt động tốt với nhiều loại dữ liệu thô khác nhau.
* Ước lượng lỗi ngoài mẫu (Out-of-Bag error): Dữ liệu không được chọn trong quá trình bagging có thể được sử dụng để ước tính lỗi ngoài mẫu (generalization error) của RFC.
* Nhược điểm của Random Forest Classifier:
* Tính toán phức tạp: Việc huấn luyện nhiều cây quyết định có thể tốn nhiều thời gian tính toán.
* Tính dễ diễn giải: Kết quả của RFC có thể khó diễn giải chi tiết do là sự kết hợp của nhiều cây quyết định.

Tóm lại, Random Forest Classifier là một thuật toán phân loại mạnh mẽ và linh hoạt, thường được sử dụng khi cần độ chính xác cao và khả năng chống overfitting. So với Decision Tree đơn lẻ, RFC có ưu điểm về độ chính xác và khả năng chống overfitting do tận dụng tính đa dạng của nhiều cây được huấn luyện trên các tập dữ liệu con khác nhau.



1. **Hồi quy Logistic (LR):**

* Hồi quy Logistic (Logistic Regression) là một thuật toán học máy được sử dụng để dự đoán xác suất của một sự kiện xảy ra. Thuật toán này sử dụng một hàm logistic (hay hàm logit) để mô tả mối quan hệ giữa biến phụ thuộc (thường là nhị phân) và các biến độc lập.
* Công thức:

Mô hình hồi quy Logistic sử dụng hàm logistic để mô tả mối quan hệ giữa biến phụ thuộc (thường là nhị phân) và các biến độc lập. Hàm logistic có dạng:

P(Y = 1 | X) = exp(β0 + β1X1 + β2X2 + ... + βnXn) / (1 + exp(β0 + β1X1 + β2X2 + ... + βnXn))

* P(Y = 1 | X): Xác suất tin tức là thật (Y = 1)
* X: Biến độc lập (ví dụ: độ dài bài viết, số lượng từ khóa)
* β0: Hệ số chặn
* β1, β2, ..., βn: Hệ số của các biến độc lập
* Cách thức hoạt động:
* Hàm Logistic: Hàm logistic có dạng hình chữ S, với giá trị đầu ra nằm trong khoảng từ 0 đến 1. Giá trị đầu ra thể hiện xác suất của sự kiện xảy ra.
* Biến phụ thuộc: Biến phụ thuộc trong hồi quy logistic thường là một biến nhị phân, ví dụ:
* Tin tức: Thật hoặc Giả
* Email: Spam hoặc Hợp lệ
* Khách hàng: Mua hàng hoặc Không mua hàng
* Biến độc lập: Biến độc lập có thể là số hoặc văn bản. Ví dụ:
* Tin tức: Độ dài bài viết, số lượng từ khóa
* Email: Địa chỉ email, nội dung email
* Khách hàng: Tuổi tác, thu nhập, lịch sử mua hàng

Hồi quy Logistic sử dụng phương pháp tối ưu hóa để tìm ra các tham số của hàm logistic, giúp dự đoán xác suất của sự kiện xảy ra một cách chính xác nhất.

* Ưu điểm:
* Đơn giản và dễ hiểu.
* Hiệu quả với các tập dữ liệu lớn.
* Có thể xử lý các biến độc lập số và văn bản.
* Nhược điểm:
* Giả định mối quan hệ tuyến tính giữa các biến.
* Có thể không hiệu quả với các tập dữ liệu nhỏ hoặc không cân bằng.
* Ứng dụng:
* Hồi quy Logistic được ứng dụng trong nhiều lĩnh vực, bao gồm:
* Phân loại văn bản (như phân loại email spam)
* Phân loại hình ảnh (như nhận diện khuôn mặt)
* Chẩn đoán y tế (như dự đoán nguy cơ mắc bệnh)
* Marketing (như dự đoán khả năng mua hàng của khách hàng)
* Phát hiện tin tức giả:
* Thu thập dữ liệu: Bao gồm tin tức thật và tin tức giả, cùng với các đặc điểm của chúng (ví dụ: độ dài bài viết, số lượng từ khóa).
* Huấn luyện mô hình: Sử dụng dữ liệu thu thập được để huấn luyện mô hình hồi quy Logistic.
* Dự đoán: Sử dụng mô hình đã huấn luyện để dự đoán xác suất một tin tức mới là thật hoặc giả.

Ví dụ:

* Giả sử bạn có một mô hình hồi quy Logistic đã được huấn luyện với các đặc điểm sau:
* Độ dài bài viết: Tin tức thật thường dài hơn tin tức giả.
* Số lượng từ khóa: Tin tức giả thường sử dụng nhiều từ khóa hơn tin tức thật.
* Mô hình dự đoán xác suất một tin tức mới là thật là 0,75. Điều này có nghĩa là mô hình tin rằng tin tức mới có 75% khả năng là thật.

Hồi quy Logistic là một thuật toán học máy mạnh mẽ và linh hoạt được sử dụng để dự đoán xác suất của một sự kiện xảy ra. Thuật toán này được ứng dụng rộng rãi trong nhiều lĩnh vực và có thể mang lại hiệu quả cao khi được sử dụng đúng cách.

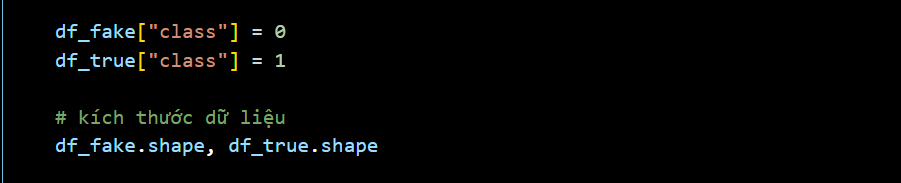
1. **THỰC NGHIỆM:**
   1. **Miêu tả dữ liệu:**

* Ngữ cảnh: Bao gồm 2 tập dữ liệu: Một tập là tin tức đúng, 1 tập là tin tức sai. Chủ đề của 2 tập dữ liệu là về tin tức chính trị, xã hội của Hợp chúng quốc Hoa Kỳ.
* Mục tiêu: Dự đoán xem tin tức là thật hay giả mạo. Điều này giúp chúng ta có thể phát hiện và ngăn chặn sự lan truyền của tin tức giả, bảo vệ người đọc khỏi thông tin sai lệch và đảm bảo tính chính xác của thông tin mà họ tiếp nhận.
* Bộ dữ liệu bao gồm thông tin:
* title (Tiêu đề): Đây là cột chứa tiêu đề của các bài báo hoặc các mục tin tức. Tiêu đề thường là một phần quan trọng của bài viết, thường ngắn gọn nhưng phản ánh nội dung chính của tin tức.
* text (Nội dung): Đây là cột chứa nội dung chi tiết của các bài báo hoặc các mục tin tức. Nội dung có thể bao gồm thông tin cụ thể, sự kiện, ý kiến, hoặc bất kỳ loại thông tin nào khác mà bài báo muốn truyền đạt.
* subject (Chủ đề): Đây là cột chứa thông tin về chủ đề của các bài báo hoặc các mục tin tức. Chủ đề có thể bao gồm các lĩnh vực như chính trị, kinh doanh, giáo dục, y tế, văn hóa, khoa học và công nghệ, v.v.
* date (Ngày): Đây là cột chứa thông tin về ngày xuất bản hoặc ngày phát hành của các bài báo hoặc các mục tin tức. Thông tin về ngày có thể giúp phân tích xu hướng và thay đổi về thời gian của các sự kiện hoặc tin tức.

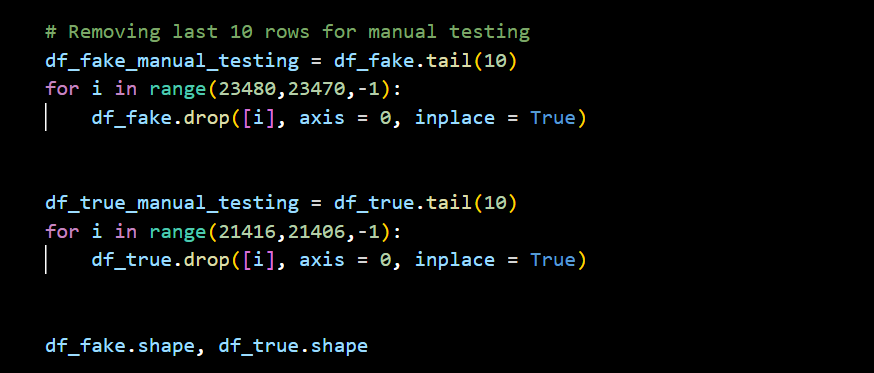
**Bảng 1. Miêu tả dữ liệu**

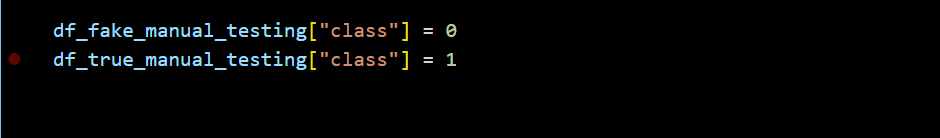
|  |  |  |
| --- | --- | --- |
| Trường dữ liệu |  | Mô tả |
| Title |  | Tiêu đề |
| Text |  | Nội dung |
| Subject |  | Chủ đề |
| Date |  | Ngày |
| Class |  | Dữ liệu là True(1) hay Fake(0) |

* 1. **Tiền xử lý dữ liệu:**
* Chèn thêm cột “Class” làm đặc trưng cho mục tiêu

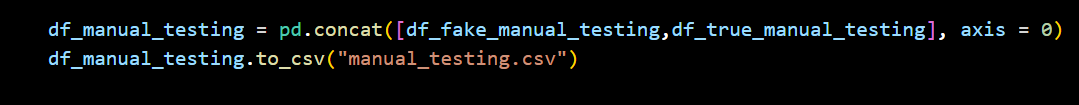
****

* Tách 10 dòng cuối cùng để tạo 1 bộ thử nghiệm riêng

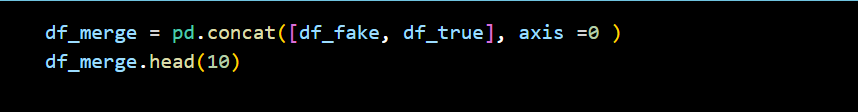


****

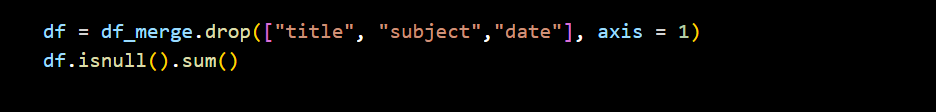
* Ghép hai DataFrame df\_fake\_manual\_testing và df\_true\_manual\_testing theo chiều dọc (axis=0), tạo thành một DataFrame mới là df\_manual\_testing. Sau đó, DataFrame df\_manual\_testing này được lưu vào file CSV có tên là "manual\_testing.csv".

****

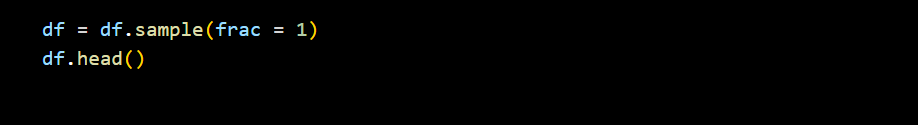
* Hợp nhất các DataFrame True và Fake

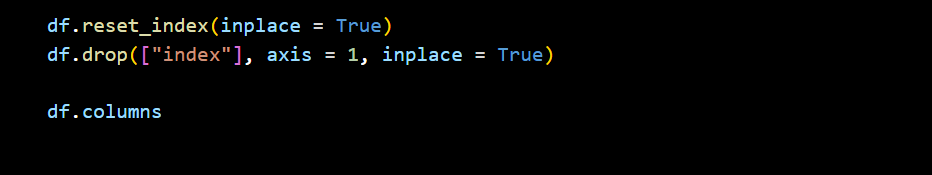
****

* Loại bỏ các cột không cần thiết

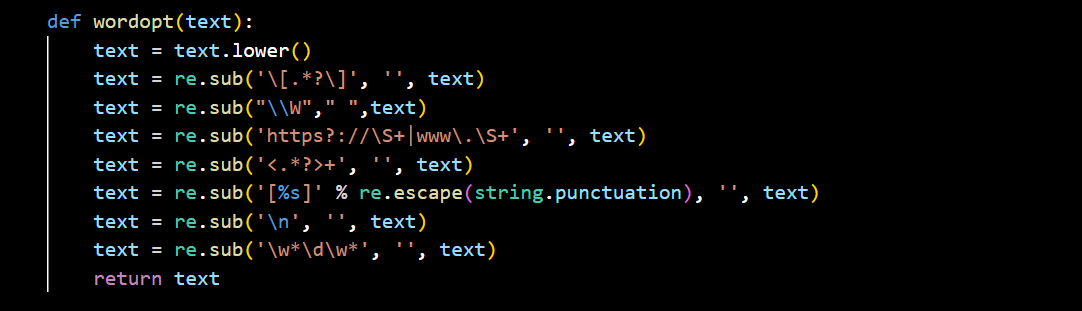
****

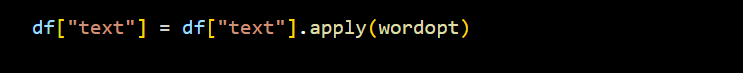
* Xáo trộn ngẫu nhiên các dòng trong dataframe

****

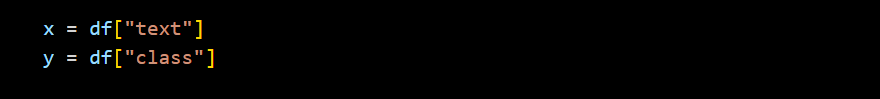
****

* Tạo một hàm để xử lý các văn bản

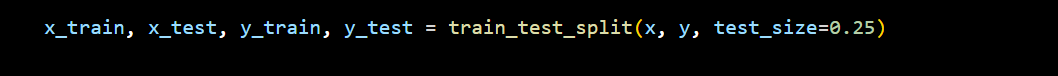


****

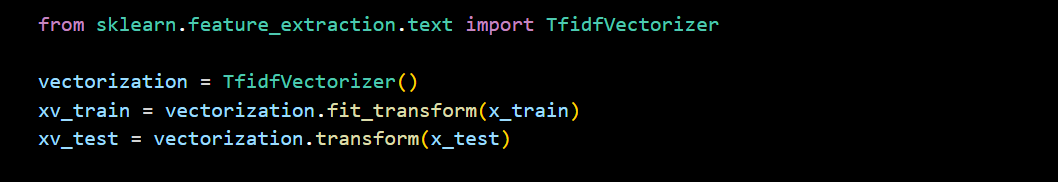
* Xác định các biến phụ thuộc và độc lập

****

* Phân chia dữ liệu huấn luyện và dữ liệu kiểm tra:
* 75% cho tập huấn luyện và 25% cho thử nghiệm



* Vector hóa

****

* 1. Các độ đo đánh giá hiệu năng:

Trong nghiên cứu này chúng tôi sử dụng bốn chỉ số bao gồm Accuracy, Precision, Recall và F1-score để đánh giá mô hình:

* Precision là tỷ lệ giữa số lượng mẫu được dự đoán khớp mẫu trên tổng dự đoán khớp mẫu và dự đoán đúng nhưng sai.

https://lh7-us.googleusercontent.com/nJV3_MlCCuuJxvLnRzvICLp4gjaaonSae8gVDqTPxa7tEAvPX3BlzS63iIolLNeFazoU3TF1xLMykrloLkGPjC9JJZkkDar7WgdCsrg3nHVVZQu1yNO-jYIWOnBl591Zj01ijh-7D99tH0j5yQzSNbk

* Recall là tỷ lệ giữa số lượng mẫu được dự đoán khớp mẫu với tổng dự đoán khớp mẫu và dự đoán sai nhưng đúng.

https://lh7-us.googleusercontent.com/ob4mR5fjCQJdm4wnD9JkP0B-0IzpqNkjjYQ2YLXIkwwA2KUe235ciYW1TNWh5l4lgoazZBwPP_wCMVA7mbAcxirVEoorHitrFwRiccuPHvl8kKC_prHRSvedvRUrD53ECVxrN5j7NXckhSf6f2P2KYk

* Accuracy là độ chính xác của mô hình, là tỷ lệ giữa số lượng mẫu được phân loại chính xác trên cho tổng số mẫu của tập dữ liệu huấn luyện.

https://lh7-us.googleusercontent.com/Mr7PsXPOSG5mt-MU2A06xbS1enROZRnI83FsycTRd7Fp-JoYoMBjKRG-I3QPmg1F6roSjZn79g61FLNgBZWQUPyk6oaqhcvBsEGJm8vgcOjdHOTGNAmGDknFhuc8kH9NhvtKZLXQKUsbCeOvj54Ta0k

* F1-score là trung bình trọng số của Precision và Recall.

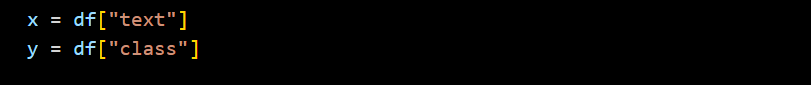
https://lh7-us.googleusercontent.com/FaDEdRR3gVxyqphZq07rHeTwaUhp6gBJGAJXQ5zc5I0e_G1bsBUlTa7ll2ZP3Slt9uqjWY6OSxEbt4M1i9S_bI1c0K626fIF8OUBAJkDeY1JF6_qfSRN6NJicA-w1Eavtc5Zwi-1AgvPV-8H9XMvk5I

Trong đó:

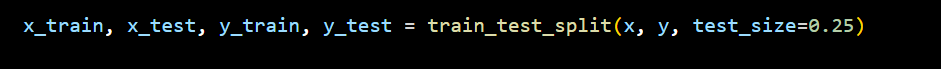
* + TP (True Positive): tổng số trường hợp dự báo khớp mẫu đúng.
  + TN (True Negative): tổng số trường hợp dự báo khớp mẫu sai.
  + FP (False Positive): tổng số trường hợp dự báo các quan sát thuộc mẫu đúng nhưng tính thành sai.
  + FN (False Negative): tổng số trường hợp dự báo các quan sát thuộc mẫu sai nhưng tính thành đúng.
  1. **Các tham số và môi trường cài đặt:**
     1. **Tham số:**
* Sử dụng tham số: criterion, n\_estimantors, min\_samples\_leaf, learning\_rate, subsample

**4.4.2. Môi trường cài đặt:**

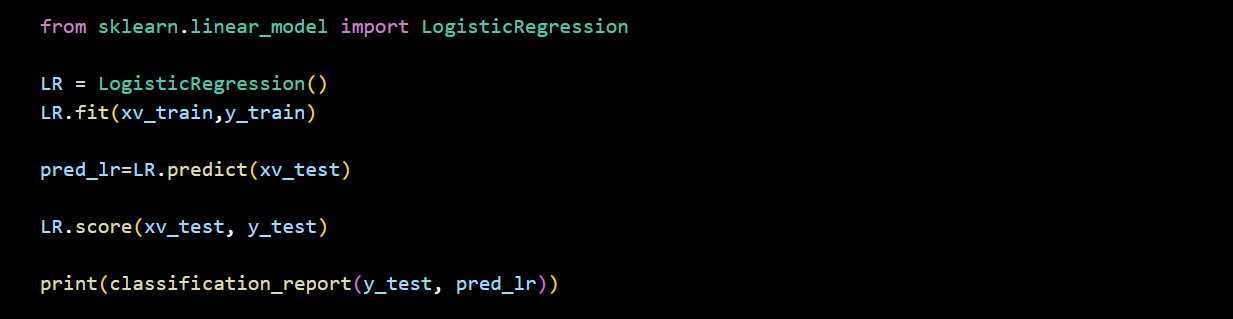
* Ngôn ngữ lập trình Python.
* Thư viện scikit-learn: cung cấp công cụ xây dựng và huấn luyện mô hình.
* Thư viện pandas: cung cấp cấu trúc dữ liệu và công cụ phân tích dữ liệu
* Thư viện numpy: hỗ trợ các phép tính số học.
* Thư viện matplotlib và seaborn: trực quan hóa dữ liệu bằng các biểu đồ.
  1. **Các phương pháp cơ sở:**
* Phương pháp phân loại:  Logistic Regression
* Sử dụng các mô hình học máy:
* Trình phân loại cây quyết định (DT)
* Trình phân loại tăng cường độ dốc (GBC)
* Phân loại rừng ngẫu nhiên (RFC)
  1. **Phân tích, so sánh các kết quả**
     1. **Mô hình phân loại**
* Logistic Regression
* Phân chia biến độc lập và phụ thuộc:

****

* Xây dựng và huấn luyện mô hình:

****

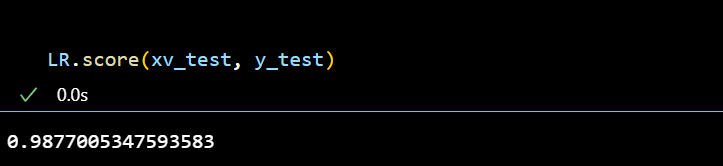
* Trong trường hợp này, bộ dữ liệu sẽ được chia theo tỷ lệ, 25% của dữ liệu sẽ được sử dụng cho tập kiểm tra và 75% còn lại sẽ được sử dụng cho tập huấn luyện (training set). Việc chia dữ liệu thành tập huấn luyện và tập kiểm tra giúp đánh giá hiệu suất của mô hình trên dữ liệu mà nó chưa từng thấy trước đó, giúp đánh giá tính tổng quát của mô hình và tránh overfitting.



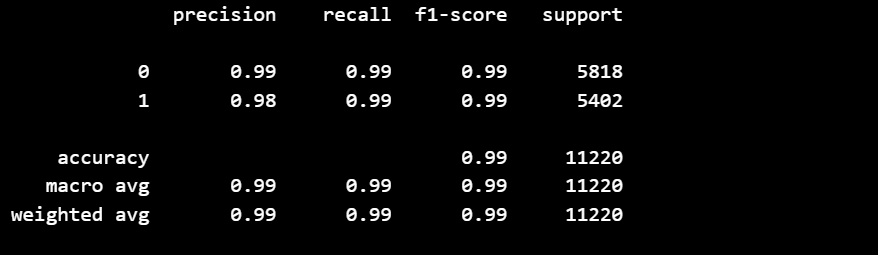
* from sklearn.linear\_model import LogisticRegression**):**
* Dòng này nhập LogisticRegression từ linear\_model mô-đun con của thư viện scikit-learn. Lớp này cho phép tạo và huấn luyện mô hình Hồi quy logistic, đây là một thuật toán phổ biến cho các nhiệm vụ phân loại nhị phân (trong đó có hai kết quả có thể xảy ra).
* LR = LogisticRegression()**):**
* Dòng code LR = LogisticRegression() tạo một mô hình Logistic Regression mới và gán nó vào biến LR. Sau đó, có thể sử dụng LR để huấn luyện và sử dụng mô hình cho các dự đoán.
* LR.fit(xv\_train, y\_train)
* Dòng này phù hợp với mô hình hồi quy logistic trên dữ liệu đào tạo được cung cấp. Trong quá trình này, mô hình tìm hiểu mối quan hệ giữa các tính năng

( xv\_train) và các nhãn tương ứng ( y\_train) để phân loại chính xác cho dữ liệu chưa nhìn thấy.

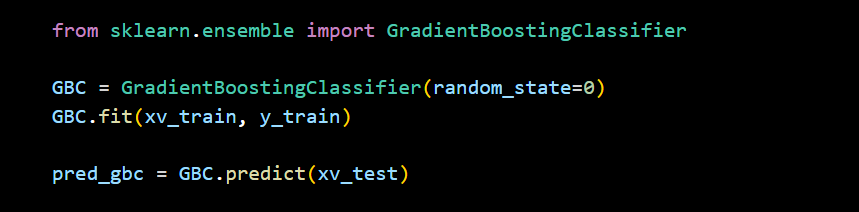
Sau khi hoàn thành các bước này, đối LR tượng sẽ là mô hình hồi quy logistic đã được huấn luyện mà bạn có thể sử dụng để đưa ra dự đoán về dữ liệu mới**.**



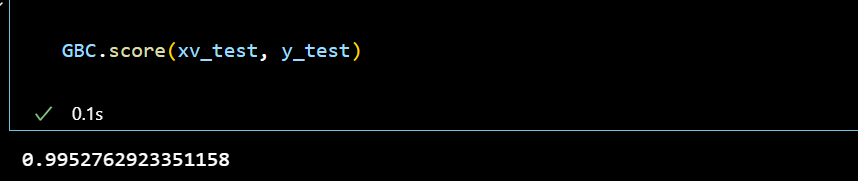
* Đánh giá hiệu suất mô hình: Ở đây, mô hình đạt 98,77% một con số cao. Tuy nhiên, nó chỉ là một thước đo hiệu suất và không nên được xem là thước đo duy nhất về độ tin cậy của mô hình.Cần xem xét các yếu tố khác như độ phức tạp của mô hình và khả năng giải thích để đưa ra quyết định sử dụng mô hình.



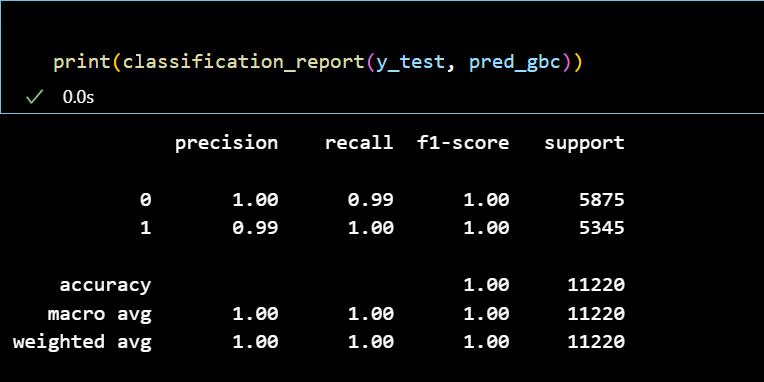
* Mô hình Logistic Regression dự đoán đúng 99% các trường hợp tin tức là giả với recall khoảng 99% nghĩa là mô hình có khả năng tốt trong việc nhận diện những tin tức giả. Tuy nhiên tỷ lệ dự đoán tin thật lại thấp hơn ( khoảng 98%) với recall 99% điều này cho thấy mô hình có hiệu quả. Độ chính xác của mô hình là 0,99 tương đương khoảng 99% cho thấy tỷ lệ dự đoán đúng của mô hình rất cao.
  + 1. Gradient Boosting Classifier



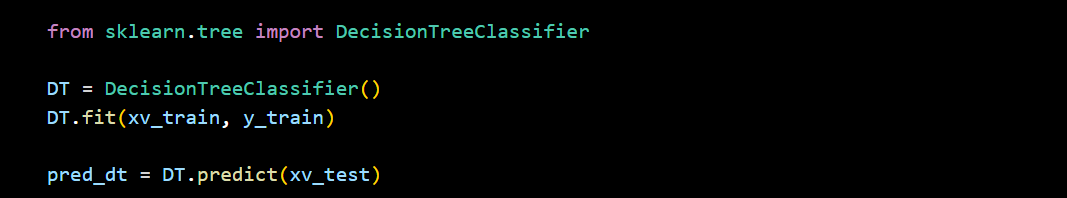
* from sklearn.ensemble import GradientBoostingClassifier: Import thư viện GradientBoostingClassifier từ scikit-learn.
* GBC = GradientBoostingClassifier(random\_state=0): Khởi tạo mô hình GBC và lưu trữ vào biến GBC.
* random\_state=0: Thiết lập giá trị ngẫu nhiên để đảm bảo tính nhất quán khi đào tạo mô hình.
* GBC.fit(xv\_train, y\_train): Dòng này phù hợp với mô hình GBC trên dữ liệu đào tạo được cung cấp. Trong quá trình này, mô hình tìm hiểu các mẫu và mối quan hệ giữa các tính năng ( xv\_train) và các nhãn tương ứng ( y\_train) để phân loại chính xác cho dữ liệu chưa nhìn thấy.
* pred\_gbc = GBC.predict(xv\_test): dòng mã này hướng dẫn mô hình GBC được đào tạo phân tích các tính năng trong xv\_test và dự đoán nhãn lớp cho từng điểm dữ liệu. Các nhãn được dự đoán sẽ được lưu trữ trong biến pred\_gbc.



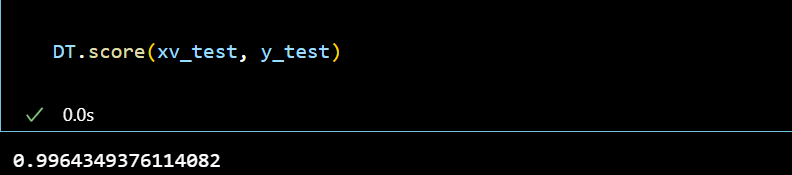
* Mô hình đạt hiệu suất 99. 53% trên tập dữ liệu thử nghiệm



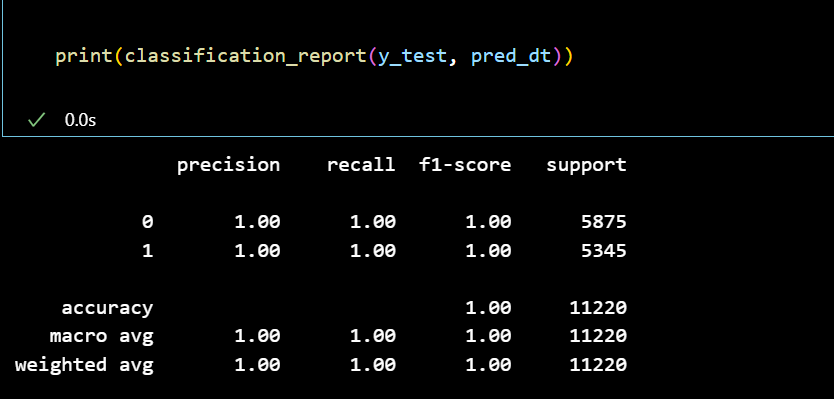
* Bảng kết quả cho thấy mô hình phân loại đã đạt được hiệu suất gần như hoàn hảo trên tập dữ liệu thử nghiệm. Mô hình có khả năng nhận diện chính xác tất cả các tin thật (recall = 1.00 cho class 0). Tuy nhiên, có một tỷ lệ nhỏ 1% tin giả có thể bị bỏ sót (recall = 0.99 cho class 1).
  + 1. Decision Tree

****

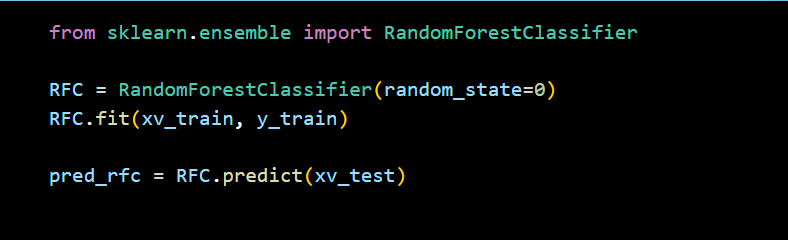
* Có thể thấy rằng như GBC, Decision Tree cũng có bước đầu như GBC nhằm mục đích phân loại chính xác cho dữ liệu chưa nhìn thấy và dự đoán nhãn lớp cho từng điểm dữ liệu. Các nhãn được dự đoán sẽ được lưu trữ trong biến pred\_gbc.

****

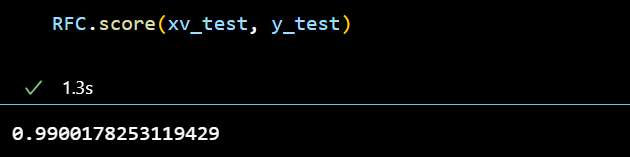
* Mô hình đạt hiệu suất 99.64% trên tập dữ liệu thử nghiệm.

****

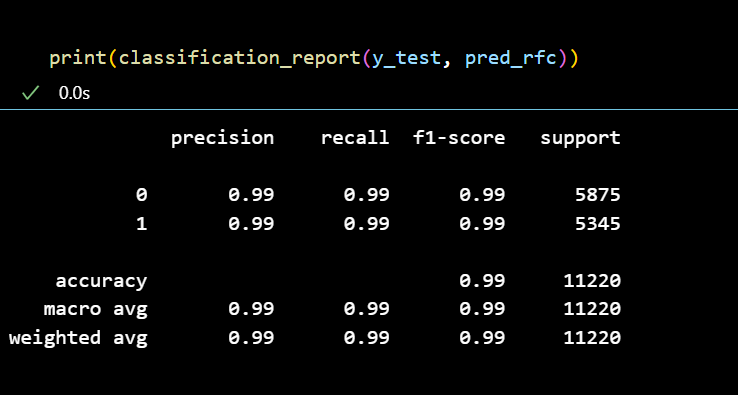
* Bảng kết quả này cho thấy mô hình phân loại đã đạt được hiệu suất gần như hoàn hảo trên tập dữ liệu thử nghiệm. Cả độ chính xác tổng thể (accuracy) lẫn các metric (precision, recall, f1-score) cho từng loại tin (tin thật, tin giả) đều đạt 1.00. Điều này cho thấy mô hình có khả năng phân biệt rất tốt giữa các loại tin tức.
  + 1. Random Forest Classifier



* Giống như GBC và Decision Tree cũng có bước đầu như GBC nhằm mục đích phân loại chính xác cho dữ liệu chưa nhìn thấy và dự đoán nhãn lớp cho từng điểm dữ liệu. Các nhãn được dự đoán sẽ được lưu trữ trong biến pred\_gbc thì Random Forest Classifier (RFC) cũng tương tự như vậy.



* Mô hình đạt hiệu suất là 99% trên tập dữ liệu thử nghiệm.



* Bảng kết quả này cho thấy mô hình phân loại đã đạt được **hiệu suất cao** trên tập dữ liệu thử nghiệm, với độ chính xác tổng thể là **99%**. Các metric (precision,nhớ lại, f1-score) cho từng loại tin (tin thật,tin giả) đều đạt **gần 0.99**, cho thấy mô hình có khả năng phân biệt tốt giữa các loại tin tức.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | **Decision Tree** | **Random Forest** | **Gradient Boosting** | **Logistic Regression** |
| **Accuracy** | 1.00 | 0.99 | 1.00 | 0.99 |
| **Recall** | 1.00 | 0.99 | 0.99 | 0.99 |
| **Precision** | 1.00 | 0.99 | 1.00 | 0.99 |
| **F1 score** | 1.00 | 0.99 | 1.00 | 0.99 |

* **Accuracy:** Tỷ lệ dự đoán đúng do mô hình đưa ra. Trong trường hợp này, tất cả các mô hình đều đạt được độ chính xác 1,00, cho thấy chúng đã phân loại hoàn hảo tất cả các điểm dữ liệu trong bộ kiểm tra. Tuy nhiên, điều quan trọng là phải xem xét cỡ mẫu và khả năng khớp quá mức khi diễn giải độ chính xác cao.
* **Recall:** Tỷ lệ các trường hợp dương tính thực tế được mô hình xác định chính xác (khả năng xác định các trường hợp liên quan). Tất cả các mẫu đều có mức thu hồi là 0,99, biểu thị rằng chúng đã nắm bắt được hầu hết các trường hợp liên quan một cách hiệu quả.
* **Precision:** Một lần nữa, tất cả các mô hình đều đạt được độ chính xác 0,99, cho thấy dự đoán của chúng hầu hết là các trường hợp dương tính chính xác.
* **F1 - Score:** Giá trị trung bình hài hòa giữa độ chính xác và khả năng thu hồi, mang lại cái nhìn cân bằng về cả hai chỉ số. Tất cả các mẫu đều có điểm F1 là 0,99, cho thấy sự cân bằng tốt giữa độ chính xác và khả năng thu hồi.

1. **KẾT LUẬN:**

Bài báo cáo này đã trình bày một hệ thống phát hiện tin giả dựa trên phương pháp học máy. Hệ thống sử dụng 4 thuật toán học máy: Random Forest (RFC), Gradient Boosting Classifier (GBC), Decision Tree (DT) và Logistic Regression để phân loại và phát hiện tin tức thật và tin giả. Các kết quả thực nghiệm qua các mô hình cho thấy mô hình Decision Tree đạt cao nhất về các chỉ số: Accuracy: 1.00, Recall: 1.00, F1 - Score: 1.00, Precision: 1.00. Nên có thể nói mô hình phân loại đã đạt được hiệu suất gần như hoàn hảo trên tập dữ liệu thử nghiệm. Cả độ chính xác tổng thể (accuracy) lẫn các metric (precision, recall, f1-score) cho từng loại tin (tin thật, tin giả) đều đạt 1.00. Điều này cho thấy mô hình có khả năng phân biệt rất tốt giữa các loại tin tức. Còn Random Forest (RFC), Logistic Regression (LG), Gradient Boosting Classifier (GBC) cũng đều đạt các chỉ số từ 99% trở lên. Do đó, độ chính xác cho từng loại tin là rất cao nên có thể sử dụng kết hợp 3 mô hình thuật toán trên với nhau để phát hiện tin tức giả. Tuy nhiên, vẫn nên cần lưu ý: Hiệu suất của mô hình có thể phụ thuộc vào chất lượng và sự cân bằng của tập dữ liệu thử nghiệm. Kiểm tra hiệu suất trên các tập dữ liệu độc lập khác nhau là cần thiết để đánh giá tổng thể về tính tin cậy của mô hình và các metric khác nhau có thể hữu ích tùy thuộc vào mục đích cụ thể của việc phân loại tin tức giả. Ví dụ, nếu việc bỏ sót tin giả (low recall) có hậu quả nghiêm trọng hơn so với dự đoán nhầm tin thật thành tin giả (high precision), thì có thể cần điều chỉnh mô hình để cải thiện recall cho lớp tin giả.

Mặc dù kết quả hiện tại rất khả quan, nhưng việc theo dõi và cải thiện hiệu suất của mô hình theo thời gian là điều cần thiết. Điều này có thể đạt được bằng cách, cập nhật dữ liệu huấn luyện với các mẫu tin mới, chỉnh các siêu tham số của mô hình, thử nghiệm các kỹ thuật học máy khác nhau.

Những năm gần đây sự phát triển của mạng Internet và đặc biệt là mạng xã hội trực tuyến đã tạo điều kiện thuận lợi hơn rất nhiều để mọi người giao tiếp với nhau qua mạng. Người dùng dễ dàng chia sẻ thông tin, kết nối với những người khác và cập nhật thông tin về các sự kiện diễn ra hàng ngày. Tuy nhiên, cùng với những tiện ích to lớn thì song hành với nó là sự xuất hiện nhiều vấn nạn mới, đặc biệt là vấn nạn về tin giả. Một lượng lớn tin tức giả trên mạng tiềm ẩn nguy cơ gây ra nhiều vấn đề nghiêm trọng trong xã hội. Việc xử lý vấn đề tin giả đã và đang thu hút sự chú ý của ngành công nghiệp, công nghệ thông tin và giới học thuật nhằm tìm hiểu về nguồn gốc, sự phân bố, phát hiện, tác hại và ngăn chặn chúng.

Báo cáo mang tính chất nghiên cứu tổng quan, nhóm đã trình bày các vấn đề liên quan đến khái niệm, phân loại, các phương pháp và mô hình học máy để tìm ra cách xác định thủ công và xác định tin giả tự động. Đặc biệt, nhóm tác giả đã trình bày hai kỹ thuật ứng dụng rộng rãi hiện nay đó là kỹ thuật học máy dựa trên các phương pháp truyền thống và học sâu. Hai kỹ thuật này đều dựa trên học máy để phân tích nội dung và bước đầu đã mang lại những kết quả tích cực. Tuy nhiên, để xác định chính xác và nhanh một tin tức có phải là giả hay không vẫn còn rất nhiều thách thức.

**TÀI LIỆU THAM KHẢO**

[1],[2]. ISSN 1859-1531 - TẠP CHÍ KHOA HỌC VÀ CÔNG NGHỆ - ĐẠI HỌC ĐÀ NẴNG, VOL. 20, NO. 3, 2022

[3].https://ieeexplore.ieee.org/abstract/document/10068653?fbclid=IwAR1Tbjv0cqodBt4iZ6wuQww2oadTkFdF8YdhoOAVM1DHT0WMckimL4FLU4\_aem\_AZx1EVFDSvd2Iz8Fw\_SQgbQYuXWeXH3O3M0KraeFvnyMiDIr0H9SRaEud36arxFWIyxrz1c5W6Iw1FPD6pM0uAB2

[4].<https://aclanthology.org/2021.acllong.62/?fbclid=IwAR1MMSAMnYit8BfvuXrS4sH0G4f_RpJmbE3lrnChU>479nGd2Ji5rHRyda8E\_aem\_AZzpqEtwHXmjtzeDvJE4wAZymS

31lDV\_mHwtEzbsUHA3rmBaDCjp6HrU2h0v8zitj4c8J5NXXoyCuyGnLkgzpM6

[5]. https://link.springer.com/chapter/10.1007/978-3-030-00671-6\_39

[6]. <https://link.springer.com/chapter/10.1007/978-3-031-53082->1\_22?fbclid=IwAR1ck6XtFZmkeaKO2iMTKU7ABEiPjlnvceGfIdQZunN0KOoyFsVgN0lsXRQ\_aem\_AZxX6Y2WQTtcUxsF4eo2D68dFo5AxZFnTDVnaktFvFLcwwiVbHU83kvn9\_PUsGRBhKam76rqsDmYDmlb72WRTRFR

[7]. https://www.worldscientific.com/doi/10.1142/S012918312250084X

1. ISSN 1859-1531 - TẠP CHÍ KHOA HỌC VÀ CÔNG NGHỆ - ĐẠI HỌC ĐÀ NẴNG, VOL. 20, NO. 3, 2022 [↑](#footnote-ref-1)
2. ISSN 1859-1531 - TẠP CHÍ KHOA HỌC VÀ CÔNG NGHỆ - ĐẠI HỌC ĐÀ NẴNG, VOL. 20, NO. 3, 2022 [↑](#footnote-ref-2)
3. https://ieeexplore.ieee.org/abstract/document/10068653?fbclid=IwAR1Tbjv0cqodBt4iZ6wuiQww2oadTkFdF8YdhoOAVM1DHT0WMckimL4FLU4\_aem\_AZx1EVFDSvd2Iz8Fw\_SQgbQYuXWeXH3O3M0KraeFvnyMiDIr0H9SRaEud36arxFWIyxrz1c5W6Iw1FPD6pM0uAB2 [↑](#footnote-ref-3)
4. <https://aclanthology.org/2021.acllong.62/?fbclid=IwAR1MMSAMnYit8BfvuXrS4sH0G4f_RpJmbE3lrnChU>

   479nGd2Ji5rHRyda8E\_aem\_AZzpqEtwHXmjtzeDvJE4wAZymS31lDV\_mHwtEzbsUHA3rmBaDCjp6HrU2h0v8zitj4c8J5NXXoyCuyGnLkgzpM6 [↑](#footnote-ref-4)
5. https://link.springer.com/chapter/10.1007/978-3-030-00671-6\_39 [↑](#footnote-ref-5)
6. https://link.springer.com/chapter/10.1007/978-3-031-53082-1\_22?fbclid=IwAR1ck6XtFZmkeaKO2iMTKU7ABEiPjlnvceGfIdQZunN0KOoyFsVgN0lsXRQ\_aem\_AZxX6Y2WQTtcUxsF4eo2D68dFo5AxZFnTDVnaktFvFLcwwiVbHU83kvn9\_PUsGRBhKam76rqsDmYDmlb72WRTRFR [↑](#footnote-ref-6)
7. https://www.worldscientific.com/doi/10.1142/S012918312250084X [↑](#footnote-ref-7)