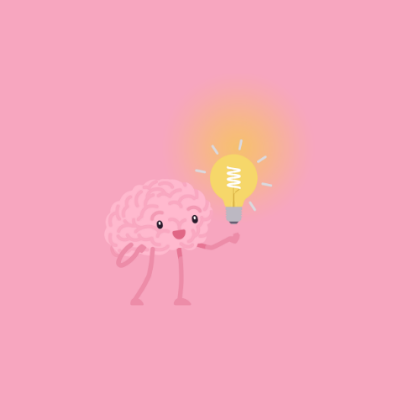
|  |  |
| --- | --- |
| **bỘ GIÁO DỤC VÀ ĐÀO TẠO** | **BỘ NÔNG NGHIỆP VÀ PTNT** |

TRƯỜNG ĐẠI HỌC



NGUYỄN TIẾN SĨ

ỨNG DỤNG CÁC MÔ HÌNH HỌC MÁY TRONG DỰ ĐOÁN CƠN ĐỘNG KINH

ĐỒ ÁN TỐT NGHIỆP

HÀ NỘI, NĂM 2024

|  |  |
| --- | --- |
| **bỘ GIÁO DỤC VÀ ĐÀO TẠO** | **BỘ NÔNG NGHIỆP VÀ PTNT** |

TRƯỜNG ĐẠI HỌC

NGUYỄN TIẾN SĨ

ỨNG DỤNG CÁC MÔ HÌNH HỌC MÁY TRONG DỰ ĐOÁN CƠN ĐỘNG KINH

|  |  |
| --- | --- |
| Ngành: | Công nghệ thông tin |
| Mã số: | 7480201 |

|  |  |
| --- | --- |
| NGƯỜI HƯỚNG DẪN | TS Lê Nguyễn Tuấn Thành |

HÀ NỘI, NĂM 2024

**GÁY BÌA ĐỒ ÁN TỐT NGHIỆP, KHÓA LUẬN TỐT NGHIỆP**

**NGUYỄN XUÂN TÂN ĐỒ ÁN TỐT NGHIỆP HÀ NỘI, NĂM 2024**

|  |  |
| --- | --- |
|  | **CỘNG HOÀ XÃ HỘI CHỦ NGHĨA VIỆT NAM**  **Độc lập - Tự do - Hạnh phúc**  ----------★----------  **NHIỆM VỤ ĐỒ ÁN TỐT NGHIỆP** |

**Họ tên sinh viên:**Nguyễn Tiến Sĩ **Hệ đào tạo:** Đại học chính quy

**Lớp:** 61TH6  **Ngành:** Công nghệ thông tin

**Khoa:** Công nghệ thông tin

1- TÊN ĐỀ TÀI:

**Ứng Dụng Các Mô Hình Học Máy Trong Dự Đoán Cơn Động Kinh**

2- CÁC TÀI LIỆU CƠ BẢN:

|  |  |
| --- | --- |
| **[1]** | **https://www.who.int/news-room/fact-sheets/detail/epilepsy** |
| **[2]** | **https://www.elcom.com.vn/12-ung-dung-may-hoc-machine-learning-hang-dau-trong-thuc-tien-1693998666** |
| **[3]** | **https://viettelidc.com.vn/tin-tuc/machine-learning-la-gi** |
| **[4]** | **https://www.ibm.com/topics/random-forest#Random+forest+applications** |
| **[5]** | **https://viblo.asia/p/support-vector-machine-trong-hoc-may-mot-cai-nhin-don-gian-hon-XQZkxoQmewA** |
| **[6]** | **https://www.who.int/news-room/fact-sheets/detail/epilepsy** |

3 - NỘI DUNG CÁC PHẦN THUYẾT MINH VÀ TÍNH TOÁN

|  |  |
| --- | --- |
| **Nội dung các phần** | **Tỷ lệ (%)** |
| Chương 1: Tổng Quan | **10%** |
| Chương 2: Các Kỹ Thuật Sử Dụng Trong Bài Toán | **20%** |
| Chương 3: Ứng Dụng Vào Dự Đoán Cơn Động Kinh | **50%** |

4 - GIÁO VIÊN HƯỚNG DẪN TỪNG PHẦN

|  |  |
| --- | --- |
| **Nội dung các phần** | **Giáo viên hướng dẫn** |
| Chương 1: Tổng Quan | **TS Lê Nguyễn Tuấn Thành** |
| Chương 2: Các Kỹ Thuật Sử Dụng Trong Bài Toán | **TS Lê Nguyễn Tuấn Thành** |
| Chương 3: Ứng Dụng Vào Dự Đoán Cơn Động Kinh | **TS Lê Nguyễn Tuấn Thành** |

5. NGÀY GIAO NHIỆM VỤ ĐỒ ÁN TỐT NGHIỆP

Ngày 23 tháng 9 năm 2024

|  |  |
| --- | --- |
| **Trưởng Bộ môn**  *(Ký và ghi rõ Họ tên)* | **Giáo viên hướng dẫn chính**  *(Ký và ghi rõ Họ tên)* |

Nhiệm vụ Đồ án tốt nghiệp đã được Hội đồng thi tốt nghiệp của Khoa thông qua.

Ngày . . . tháng . . . năm 2024  **Chủ tịch Hội đồng**

*(Ký và ghi rõ Họ tên)*

Sinh viên đã hoàn thành và nộp bản Đồ án tốt nghiệp cho Hội đồng thi ngày ... tháng ... năm ......

**Sinh viên làm Đồ án tốt nghiệp**

*(Ký và ghi rõ Họ tên)*

Nguyễn Tiến Sĩ

|  |  |
| --- | --- |
| **logo** | **TRƯỜNG ĐẠI HỌC THUỶ LỢI**  **KHOA CÔNG NGHỆ THÔNG TIN**  **BẢN TÓM TẮT ĐỀ CƯƠNG ĐỒ ÁN TỐT NGHIỆP** |

**Tên đề tài: Ứng dụng các mô hình học máy trong dự đoán cơn động kinh**

*Sinh viên thực hiện*: Nguyễn Tiến Sĩ

*Lớp*: 61TH6

*Mã sinh viên :*1951060998

*Email*: nxt.tattoo@gmail.com

*Số điện thoại*: 0876913397

*Giáo viên hướng dẫn*:  TS. Lê Nguyễn Tuấn Thành

**TÓM TẮT ĐỀ TÀI**

Động kinh là một rối loạn thần kinh mãn tính, gây ra các cơn co giật đột ngột và không thể dự đoán trước. Việc dự đoán chính xác các cơn động kinh có vai trò quan trọng trong việc cải thiện chất lượng cuộc sống của bệnh nhân, giúp họ có thể chuẩn bị và phòng tránh những nguy hiểm tiềm ẩn. Với sự phát triển của công nghệ và dữ liệu y tế, học máy (machine learning) trở thành công cụ mạnh mẽ trong việc phân tích các tín hiệu sinh học như điện não đồ (EEG) để dự đoán trước cơn động kinh.

Đề tài này tập trung vào việc nghiên cứu và ứng dụng các phương pháp học máy khác nhau như SVM, hồi quy logistic và ANN, bao gồm nhưng không giới hạn ở các mô hình học sâu (deep learning), mạng nơ-ron nhân tạo (artificial neural networks), và các kỹ thuật xử lý tín hiệu. Các thuật toán này sẽ được sử dụng để phân tích dữ liệu EEG, phát hiện các mẫu hoặc dấu hiệu tiền triệu có thể dự báo được cơn động kinh.

Mục tiêu chính của đề tài là xây dựng một hệ thống tự động có khả năng dự đoán cơn động kinh với độ chính xác cao, thời gian cảnh báo đủ sớm để bệnh nhân có thể kịp thời chuẩn bị. Kết quả nghiên cứu không chỉ góp phần giảm thiểu tác động tiêu cực của bệnh động kinh mà còn mở ra tiềm năng ứng dụng trong nhiều lĩnh vực y tế khác.

Với tầm quan trọng của việc dự đoán được cơn động kinh, em đã chọn và làm về đề tài: “**Ứng dụng các mô hình học máy trong dự đoán cơn động kinh**” em dự kiến sẽ sử dụng ngôn ngữ lập trình Python và kết hợp các thư viện hỗ trợ để trực quan hóa, cài đặt một số mô hình học máy giải quyết bài toán.

**CÁC MỤC TIÊU CHÍNH**

* Mục tiêu 1: Tìm hiểu bài toán dự đoán cơn động kinh
* Mục tiêu 2: Tìm hiểu, nghiên cứu một số mô hình học máy ví dụ như vector hỗ trợ SVM (Support Vector Machine)
* Mục tiêu 3: Huấn luyện mô hình với bộ dữ liệu
* Mục tiêu 4: Đánh giá mô hình
* Mục tiêu 5: Kết luận

**KẾT QUẢ DỰ KIẾN**

* Hoàn thành các mục tiêu đề ra ở trên
* Hiểu, áp dụng, xây dựng được mô hình học máy cho bài toán dự đoán cơn động kinh
* Báo cáo tổng kết ĐATN

**TIẾN ĐỘ THỰC HIỆN**

|  |  |  |  |
| --- | --- | --- | --- |
| **TT** | **Thời gian** | **Nội dung công việc** | **Kết quả dự kiến đạt được** |
| 1 | 24/9/2024-28/9/2024 | - Tìm hiểu, lựa chọn đề tài.  - Tìm tài liệu tham khảo.  - Tìm kiếm nguồn dữ liệu cần dùng cho đề tài. | - Tìm ra được đề tài thực hiện.  -Tìm được các bài báo liên quan đến đề tài thực hiện.  - Có dữ liệu thực hiện đề tài. |
| 2 | 29/9/2024-  15/11/2024 | - Phân tích đề tài.  - Tìm hiểu và phân tích các tài liệu liên quan.  - Phân tích và khai thác các bài toán sử dụng vào đề tài. | - Hiểu hết về đề tài.  - Hiểu hết các thuật toán muốn sử dụng vào đề tài. |
| 3 | 15/11/2024-5/1/2025 | - Áp dụng các thuật toán vào đề tài.  - Chạy kết quả và đánh giá | - Áp dụng được các thuật toán vào đề tài.  - Cho ra kết quả nhanh chóng và có độ chính xác trên 95%. |

**TÀI LIỆU THAM KHẢO**

- Ứng dụng máy học vector hỗ trợ SVM trong dự đoán cơn động kinh

https://vjol.info.vn/index.php/TCDHDL/article/view/46662

- Machine Learning Algorithms for Epileptic Seizures

https://www.kaggle.com/code/harunshimanto/machine-learning-algorithms-for-epileptic-seizures

LỜI CAM ĐOAN

Em là Nguyễn Tiến Sĩ, em xin cam đoan đây là Đồ án tốt nghiệp là thành quả của bản thân nghiên cứu. Các kết quả trong Đồ án tốt nghiệp này là trung thực, và không sao chép từ bất kỳ một nguồn nào và dưới bất kỳ hình thức nào. Việc tham khảo các nguồn tài liệu (nếu có) đã được thực hiện trích dẫn và ghi nguồn tài liệu tham khảo đúng quy định.

Em xin chân thành cảm ơn!

|  |  |
| --- | --- |
|  | **Tác giả ĐATN/KLTN**  *Chữ ký*  **Nguyễn Tiến Sĩ** |

LỜI CÁM ƠN

Em xin gửi lời cảm ơn sâu sắc đến TS Lê Nguyễn Tuấn Thành, người đã dành thời gian quý báu để hướng dẫn, hỗ trợ và chia sẻ những kiến thức cũng như kinh nghiệm trong suốt quá trình thực hiện đề tài. Những lời khuyên và sự động viên của thầy đã là nguồn động lực lớn giúp em hoàn thiện luận văn này.

Em cũng xin chân thành cảm ơn Khoa Công nghệ thông tin - Trường Đại học , nơi đã tạo ra một môi trường học tập lý tưởng với đầy đủ trang thiết bị và tài liệu, giúp em có thể phát huy tối đa khả năng của mình trong việc nghiên cứu và hoàn thành đề tài.

Em đặc biệt biết ơn gia đình và bạn bè, những người luôn ở bên, ủng hộ và chia sẻ những khó khăn, thử thách cùng em trong suốt quá trình học tập.

Em xin trân trọng cảm ơn!

MỤC LỤC

[DANH MỤC CÁC HÌNH ẢNH x](#_Toc186823862)

[DANH MỤC BẢNG BIỂU xi](#_Toc186823863)

[DANH MỤC CÁC TỪ VIẾT TẮT VÀ GIẢI THÍCH CÁC THUẬT NGỮ xii](#_Toc186823864)

[CHƯƠNG 1 TỔNG QUAN 1](#_Toc186823865)

[1.1 Đặt vấn đề 1](#_Toc186823866)

[1.2 Mục tiêu nghiên cứu 2](#_Toc186823867)

[1.2.1 Mục tiêu tổng quát 2](#_Toc186823868)

[1.2.2 Mục tiêu cụ thể 2](#_Toc186823869)

[1.3 Đối tượng nghiên cứu và phạm vi nghiên cứu 2](#_Toc186823870)

[1.3.1 Đối tượng nghiên cứu 2](#_Toc186823871)

[1.3.2 Phạm vi nghiên cứu 3](#_Toc186823872)

[CHƯƠNG 2 CÁC KỸ THUẬT SỬ DỤNG TRONG BÀI TOÁN 4](#_Toc186823873)

[2.1 Tổng quan về học máy 4](#_Toc186823874)

[2.1.1 Khái niệm học máy 4](#_Toc186823875)

[2.1.2 Ứng dụng học máy 5](#_Toc186823876)

[2.1.3 Cách thức hoạt động 6](#_Toc186823877)

[2.1.4 Phân loại nhóm học máy 7](#_Toc186823878)

[2.2 Các mô hình phân loại 13](#_Toc186823879)

[2.2.1 Thuật toán Iterative Dichotomiser 3 13](#_Toc186823880)

[2.2.2 Thuật toán Support Vector Machine 15](#_Toc186823881)

[2.2.3 Thuật toán Artificial Neural Network 17](#_Toc186823882)

[2.3 Quy trình tổng quan 19](#_Toc186823883)

[2.3.1 Xác định vấn đề 19](#_Toc186823884)

[2.3.2 Thu thập dữ liệu 19](#_Toc186823885)

[2.3.3 Tiền xử lý dữ liệu 20](#_Toc186823886)

[2.3.4 Lựa chọn mô hình 20](#_Toc186823887)

[2.3.5 Huấn luyện mô hình 21](#_Toc186823888)

[2.3.6 Đánh giá mô hình 22](#_Toc186823889)

[2.3.7 Tinh chỉnh mô hình 23](#_Toc186823890)

[2.3.8 Triển khai mô hình 24](#_Toc186823891)

[CHƯƠNG 3 ỨNG DỤNG VÀO DƯ ĐOÁN CƠN ĐỘNG KINH 25](#_Toc186823892)

[3.1 Công nghệ sử dụng 25](#_Toc186823893)

[3.1.1 Ngôn ngữ 25](#_Toc186823894)

[3.1.2 Các thư viện sử dụng 26](#_Toc186823895)

[3.2 Phát biểu bài toán 28](#_Toc186823896)

[3.3 Quy trình tổng quan bài toán 28](#_Toc186823897)

[3.4 Thu thập dữ liệu 28](#_Toc186823898)

[3.5 Làm sạch dữ liệu 30](#_Toc186823899)

[3.5.1 Xóa các thông tin bị thiếu 30](#_Toc186823900)

[3.5.2 Xóa dữ liệu trùng lặp 30](#_Toc186823901)

[3.6 Phân tích dữ liệu 30](#_Toc186823902)

[3.7 Tiền xử lý dữ liệu 31](#_Toc186823903)

[3.7.1 Chuẩn hóa dữ liệu 31](#_Toc186823904)

[3.7.2 Cân bằng dữ liệu 31](#_Toc186823905)

[3.8 Lựa chọn mô hình 32](#_Toc186823906)

[3.9 Huấn luyện mô hình 32](#_Toc186823907)

[3.10 Đánh giá mô hình 34](#_Toc186823908)

[3.10.1 Iterative Dichotomiser 3 34](#_Toc186823909)

[3.10.2 SVM 35](#_Toc186823910)

[3.10.3 ANN 36](#_Toc186823911)

[3.10.4 Nhận xét chung 36](#_Toc186823912)

[3.11 Triển khai mô hình 37](#_Toc186823913)

[3.11.1 Lưu mô hình 37](#_Toc186823914)

[3.11.2 Triển khai mô hình 38](#_Toc186823915)

[KẾT LUẬN 40](#_Toc186823916)

[TÀI LIỆU THAM KHẢO 41](#_Toc186823917)

[PHỤ LỤC 42](#_Toc186823918)

DANH MỤC CÁC HÌNH ẢNH

[Hình 2.1 Học máy 4](#_Toc186823842)

[Hình 2.2 Phân loại học máy 7](#_Toc186823843)

[Hình 2.3 Xác định vấn đề 19](#_Toc186823844)

[Hình 2.4 Thu thập dữ liệu 20](#_Toc186823845)

[Hình 2.5 Tiền xử lý dữ liệu 20](#_Toc186823846)

[Hình 2.6 Lựa chọn mô hình 21](#_Toc186823847)

[Hình 2.7 Huấn luyện mô hình 22](#_Toc186823848)

[Hình 2.8 Đánh giá mô hình 23](#_Toc186823849)

[Hình 2.9 Tinh chỉnh mô hình 24](#_Toc186823850)

[Hình 2.10 Triển khai mô hình 24](#_Toc186823851)

[Hình 3.1 Quy trình tổng quan bài toán 28](#_Toc186823852)

[Hình 3.2 Thông tin dữ liệu thu thập 29](#_Toc186823853)

[Hình 3.3 Hình ảnh thông tin dữ liệu 30](#_Toc186823854)

[Hình 3.4 Giá trị sau khi cân bằng 32](#_Toc186823855)

[Hình 3.5 Ma trận matrix thuật toánIterative Dichotomiser 3 34](#_Toc186823856)

[Hình 3.6 Kết quả báo cáo phân loại thuật toánIterative Dichotomiser 3 34](#_Toc186823857)

[Hình 3.7 Ma trận matrix thuật toán SVM 35](#_Toc186823858)

[Hình 3.8 Kết quả báo cáo phân loại thuật toán SVM 35](#_Toc186823859)

[Hình 3.9 Ma trận matrix thuật toán ANN 36](#_Toc186823860)

[Hình 3.10 Giao diện người dùng 39](#_Toc186823861)

DANH MỤC BẢNG BIỂU

[Bảng 5.1 Bảng dự đoán của 3 mô hình 37](#_Toc186359404)

DANH MỤC CÁC TỪ VIẾT TẮT VÀ GIẢI THÍCH CÁC THUẬT NGỮ

**ĐATN** Đồ án tốt nghiệp

**SVM** Support Vector Machine

# TỔNG QUAN

## Đặt vấn đề

Động kinh là một bệnh lý thần kinh phổ biến, tác động đến khoảng 50 triệu người trên toàn cầu, theo thông tin từ Tổ chức Y tế Thế giới (WHO) [1].Bệnh đặc trưng bởi các cơn co giật tái diễn do sự bất thường trong hoạt động điện của não, gây ra nhiều ảnh hưởng nghiêm trọng đến sức khỏe và chất lượng cuộc sống của người bệnh. Hiện nay, chẩn đoán động kinh chủ yếu dựa vào phân tích điện não đồ (EEG), nhưng phương pháp này phụ thuộc nhiều vào trình độ chuyên môn của các bác sĩ và có thể gặp khó khăn trong việc phát hiện các tín hiệu phức tạp. Trước thực trạng đó, học máy (machine learning) đã mở ra hướng đi mới, mang lại tiềm năng cải thiện độ chính xác và tự động hóa quá trình chẩn đoán.

Học máy có khả năng xử lý khối lượng dữ liệu lớn và phát hiện những mẫu tín hiệu phức tạp mà con người khó nhận ra. Nhờ đó, các mô hình học máy không chỉ giúp xác định chính xác cơn co giật mà còn dự đoán trước thời điểm xảy ra, hỗ trợ bệnh nhân chuẩn bị tốt hơn và giảm rủi ro. Tuy nhiên, dữ liệu EEG phức tạp và dễ nhiễu, đòi hỏi các phương pháp xử lý tín hiệu tiên tiến để giữ lại thông tin quan trọng. Lựa chọn mô hình học máy phù hợp, như SVM, ANN hay CNN, cũng đóng vai trò quyết định và cần tùy chỉnh theo yêu cầu cụ thể của bài toán.

Dù còn nhiều thách thức, lợi ích mà học máy mang lại là rất lớn. Các mô hình học máy đã được chứng minh là có khả năng nâng cao độ chính xác và độ nhạy trong việc phát hiện động kinh. Nhiều nghiên cứu cho thấy mạng nơ-ron tích chập (CNN) đạt độ chính xác vượt trội, thậm chí trên 90% khi phân tích dữ liệu EEG, giúp nhận diện và dự đoán các cơn động kinh hiệu quả hơn. Một số hệ thống còn có thể dự đoán trước các cơn co giật vài phút, giúp bệnh nhân chuẩn bị và giảm nguy cơ gặp chấn thương hoặc biến chứng nguy hiểm.

Vì vậy, việc ứng dụng học máy vào bài toán dự đoán động kinh là một bước tiến quan trọng, không chỉ khắc phục những hạn chế của phương pháp truyền thống mà còn mang lại nhiều lợi ích vượt trội về hiệu quả, chi phí và tính cá nhân hóa. Từ những vấn đề thực tế, nhu cầu cần thiết của ứng dụng Ứng dụng các mô hình học máy trong dự đoán cơn động kinh em lựa chọn đề tài “**Xây dựng ứng dụng Ứng dụng các mô hình học máy trong dự đoán cơn động kinh**”

## Mục tiêu nghiên cứu

### Mục tiêu tổng quát

Xây dựng ứng dụng học máy Ứng dụng các mô hình học máy trong dự đoán cơn động kinh

### Mục tiêu cụ thể

Đề tài "Ứng dụng học máy trong Ứng dụng các mô hình học máy trong dự đoán cơn động kinh" nhằm xây dựng và triển khai một mô hình học máy để dự đoán nguy cơ động kinh dựa trên các yếu tố đầu vào cụ thể từ dữ liệu điện não đồ (EEG) và các đặc điểm lâm sàng của bệnh nhân. Mục tiêu của nghiên cứu bao gồm:

* Tìm hiểu bài toán dự đoán động kinh: Bệnh động kinh có thể được dự đoán dựa trên các đặc điểm của bệnh nhân
* Tìm hiểu và nghiên cứu các kỹ thuật học máy: Việc lựa chọn các mô hình học máy phù hợp như SVM, Iterative Dichotomiser 3, .... sẽ giúp mô hình dự đoán đạt độ chính xác cao trong việc phân loại các cơn động kinh.
* Áp dụng các kỹ thuật học máy để dự đoán động kinh: Thu thập và tiền xử lý dữ liệu lý các tín hiệu nhiễu, xây dựng và huấn luyện mô hình học máy, sau đó tối ưu hóa mô hình để đạt được hiệu quả dự đoán tốt nhất.
* Đánh giá mô hình và tính hiệu quả của thuật toán: Sử dụng các chỉ số như độ chính xác, độ nhạy, và điểm F1 để đánh giá hiệu suất của mô hình, đồng thời áp dụng các kỹ thuật đánh giá như cross-validation để kiểm tra khả năng tổng quát hóa của mô hình khi áp dụng trên dữ liệu mới.

## Đối tượng nghiên cứu và phạm vi nghiên cứu

### Đối tượng nghiên cứu

Nghiên cứu mô hình Ứng dụng các mô hình học máy trong dự đoán cơn động kinh từ những thông tin thu thập được

### Phạm vi nghiên cứu

Phạm vi nghiên cứu của đề tài " **Xây dựng ứng dụng Ứng dụng các mô hình học máy trong dự đoán cơn động kinh** " bao gồm:

* Loại dữ liệu: Dữ liệu thu thập bao gồm các yếu tố liên quan đến bệnh động kinh.
* Phương pháp học máy: SVM, Iterative Dichotomiser 3
* Đánh giá mô mình: Phạm vi nghiên cứu cũng bao gồm việc đánh giá hiệu suất của các mô hình học máy trong các bài toán phân loại thông qua các chỉ số như độ chính xác (accuracy), độ nhạy (precision), độ đặc hiệu (recall), và điểm F1 (F1-score). Các kỹ thuật đánh giá như cross-validation sẽ được sử dụng để đảm bảo mô hình có thể tổng quát hóa tốt trên dữ liệu chưa thấy, giúp đánh giá khả năng dự đoán chính xác các lớp trong các tình huống khác nhau.

# CÁC KỸ THUẬT SỬ DỤNG TRONG BÀI TOÁN

## Tổng quan về học máy

### Khái niệm học máy

Máy học là một lĩnh vực quan trọng trong trí tuệ nhân tạo, nghiên cứu và phát triển các thuật toán giúp hệ thống tự học và cải thiện khả năng xử lý dựa trên dữ liệu, mà không cần phải lập trình chi tiết cho mỗi tác vụ cụ thể.

Các hệ thống sử dụng Machine Learning được thiết kế để xử lý khối lượng dữ liệu lớn, phát hiện các mẫu ẩn trong dữ liệu và đưa ra những dự đoán hoặc phân tích hữu ích. Nhờ quá trình này, hệ thống có thể tự động học từ dữ liệu đầu vào và liên tục cải thiện độ chính xác trong các kết quả mà nó tạo ra.

Trong môi trường doanh nghiệp, Machine Learning đóng vai trò then chốt trong việc tự động hóa các quy trình thu thập và phân tích dữ liệu. Ứng dụng các thuật toán học máy giúp tối ưu hóa thời gian xử lý, nâng cao hiệu quả quản lý và tăng cường năng suất công việc. Bên cạnh đó, nó còn hỗ trợ doanh nghiệp đưa ra các quyết định chiến lược dựa trên dữ liệu thực tế và đáng tin cậy.

A cartoon robot with a backpack

Description automatically generated

Hình . Học máy

### Ứng dụng học máy

Học máy là một lĩnh vực trong trí tuệ nhân tạo đã phát triển mạnh mẽ và được ứng dụng rộng rãi trong nhiều ngành nghề khác nhau. Dưới đây là một số ví dụ về ứng dụng của học máy:

* Dự đoán bệnh: Đây là một ứng dụng quan trọng của học máy trong y tế. Các mô hình học máy có thể phân tích dữ liệu y tế như hình ảnh X-quang, MRI, hoặc các chỉ số sinh lý như huyết áp và cholesterol để phát hiện sớm các bệnh lý như ung thư, bệnh tim mạch, hay tiểu đường. Việc ứng dụng học máy trong dự đoán bệnh giúp các bác sĩ đưa ra các quyết định chính xác và kịp thời, từ đó nâng cao hiệu quả điều trị và giảm thiểu rủi ro cho bệnh nhân.
* Chatbot: Các hệ thống chatbot sử dụng học máy để giao tiếp với khách hàng, tự động trả lời các câu hỏi và hỗ trợ giải quyết vấn đề mà không cần sự can thiệp của con người. Điều này giúp rút ngắn thời gian và nâng cao hiệu quả trong việc phục vụ khách hàng.
* Hệ thống gợi ý: Học máy được sử dụng để phân tích hành vi người dùng và đưa ra các gợi ý sản phẩm hoặc dịch vụ phù hợp. Ví dụ như hệ thống gợi ý phim trên Netflix hay các sản phẩm trên Amazon.
* Định giá động: Các mô hình học máy giúp doanh nghiệp thay đổi giá sản phẩm theo thời gian thực, dựa trên nhu cầu thị trường, mức độ cạnh tranh, hay thời gian trong ngày, từ đó tối ưu hóa doanh thu.
* Tối ưu hóa bán hàng và tiếp thị: Machine Learning phân tích dữ liệu khách hàng để phát triển các chiến lược tiếp thị hiệu quả, chẳng hạn như phân tích hành vi mua sắm hoặc dự đoán khả năng khách hàng rời bỏ dịch vụ, từ đó đề xuất các giải pháp giữ chân khách hàng một cách kịp thời.
* Phát hiện gian lận: Hệ thống học máy giúp nhận diện các giao dịch tài chính bất thường bằng cách phát hiện các mẫu hành vi không phổ biến, qua đó ngăn chặn các hoạt động gian lận trước khi gây ra tổn thất.
* An ninh mạng: Machine Learning đóng vai trò quan trọng trong việc bảo vệ mạng lưới và dữ liệu tổ chức bằng cách phân tích lưu lượng mạng, phát hiện các hoạt động khả nghi, và xác định các mối đe dọa tiềm tàng
* Tối ưu hóa: Các thuật toán học máy có thể tối ưu hóa các quy trình trong sản xuất, logistics, hoặc quản lý chuỗi cung ứng, giúp nâng cao hiệu suất và giảm chi phí.
* Hỗ trợ quyết định: Các mô hình học máy phân tích dữ liệu lớn và hỗ trợ ra quyết định trong các lĩnh vực như tài chính, y tế và kinh doanh, giúp đưa ra những lựa chọn chính xác hơn.
* Bảo trì dự đoán: Học máy có khả năng dự đoán thời điểm thiết bị có thể gặp sự cố, giúp các doanh nghiệp thực hiện bảo trì chủ động, giảm thiểu thời gian dừng hoạt động và chi phí sửa chữa.
* Đảm bảo chất lượng: Trong sản xuất, học máy được sử dụng để phân tích dữ liệu và kiểm tra chất lượng sản phẩm, giúp phát hiện lỗi sớm và tối ưu hóa quy trình sản xuất.
* Phân tích cảm xúc: Học máy phân tích các phản hồi của khách hàng qua văn bản, giọng nói hoặc hình ảnh để đánh giá cảm xúc và mức độ hài lòng, từ đó hỗ trợ cải thiện dịch vụ và sản phẩm.
* Trích xuất thông tin: Học máy được ứng dụng để tự động rút ra thông tin từ các nguồn dữ liệu không cấu trúc như văn bản, hình ảnh hoặc video, giúp tổ chức và xử lý dữ liệu một cách hiệu quả.

### Cách thức hoạt động

Quy trình hoạt động của Machine Learning thường bao gồm 5 bước chính như sau:

* **Bước 1:** Thu thập dữ liệu cần thiết để làm cơ sở huấn luyện thuật toán
* **Bước 2:** Xử lý và làm sạch dữ liệu để đảm bảo chất lượng. Quá trình này bao gồm chuẩn hóa dữ liệu, gắn nhãn, mã hóa, giảm kích thước tập dữ liệu (dimensionality reduction), và trích xuất các đặc trưng quan trọng từ dữ liệu.
* **Bước 3:** Tiến hành huấn luyện mô hình bằng cách cung cấp dữ liệu đã xử lý. Mô hình học cách nhận diện các quy luật và mối quan hệ trong dữ liệu để dự đoán hoặc phân tích đầu ra.
* **Bước 4:** Đánh giá hiệu suất của mô hình trên tập dữ liệu thử nghiệm để xác định độ chính xác và khả năng áp dụng của mô hình trong thực tế.
* **Bước 5:** Điều chỉnh và tối ưu hóa mô hình để khắc phục những hạn chế. Mô hình được tiếp tục huấn luyện và cải thiện cho đến khi đạt độ chính xác tối thiểu, thường là trên 80%, tùy thuộc vào yêu cầu cụ thể của bài toá

### Phân loại nhóm học máy

Các thuật toán học máy được phân loại dựa trên mục tiêu và kết quả mong đợi. Một số loại thuật toán thường gặp bao gồm:

* Học máy có giám sát (Supervised Learning)
* Học máy không giám sát (Unsupervised Learning)
* Học máy bán giám sát (Semi-supervised Learning)
* Học máy tăng cường (Reinforcement Learning)

A diagram of machine learning

Description automatically generated

Hình 2.2 Phân loại học máy

#### Học máy có giám sát (Supervised Learning)

Học có giám sát là một phương pháp trong học máy, nhằm xây dựng một hàm dựa trên dữ liệu huấn luyện. Dữ liệu huấn luyện gồm các cặp đầu vào và đầu ra mong muốn. Đầu ra của hàm có thể là một giá trị liên tục hoặc một nhãn phân loại, tùy thuộc vào đặc điểm của đầu vào.

Mục tiêu chính của học có giám sát là dự đoán đầu ra cho bất kỳ đầu vào hợp lệ nào sau khi học từ các ví dụ trong tập huấn luyện. Để thực hiện điều này, mô hình cần có khả năng tổng quát hóa từ dữ liệu huấn luyện để đưa ra các dự đoán chính xác cho các trường hợp chưa từng gặp, dựa trên logic "hợp lý".

Trong học có giám sát, có hai nhánh chính:

* Phân loại (Classification)
* Hồi quy (Regression)
* **Phân loại (Classification)**

Phân loại (Classification) là một quá trình trong học máy nhằm dự đoán lớp hoặc nhãn của một điểm dữ liệu dựa trên các đặc trưng mô tả của nó. Mục tiêu chính của phân loại là chia các điểm dữ liệu vào các nhóm (lớp) khác biệt, đảm bảo rằng mỗi nhóm đại diện cho một tập hợp các đặc tính riêng biệt và nhất quán.

Đặc điểm:

* Đầu vào: Dữ liệu đầu vào cho mô hình phân loại bao gồm các điểm dữ liệu với các đặc trưng mô tả (hoặc vector đặc trưng).
* Đầu ra: Đầu ra của mô hình phân loại là lớp hoặc nhãn được dự đoán cho mỗi điểm dữ liệu.

Các thuật toán phổ biến:

* Decision Trees: Xây dựng một cây quyết định dựa trên các câu hỏi về các đặc trưng để phân loại các điểm dữ liệu.
* Logistic Regression: Sử dụng hàm logistic để phân loại dữ liệu thành hai lớp.
* Support Vector Machines: Tìm ra siêu phẳng giới phân chia tối ưu giữa các lớp để phân loại dữ liệu.

Cách hoạt động:

* Mô hình phân loại học từ các tập dữ liệu huấn luyện, trong đó mỗi mẫu dữ liệu có các đặc trưng và nhãn tương ứng.
* Sau khi huấn luyện, mô hình có thể dự đoán nhãn của các mẫu dữ liệu mới dựa trên các đặc trưng của chúng.
* Cây quyết định, mô hình này xây dựng một cây quyết định dựa trên các câu hỏi về các đặc trưng của dữ liệu để đưa ra dự đoán.

Độ đo hiệu suất:

* Độ chính xác (Accuracy): Tỷ lệ các dự đoán chính xác so với tổng số dự đoán.
* Độ phủ (Recall) và Độ chính xác trung bình (Precision): Được sử dụng để đánh giá hiệu suất của mô hình đối với các lớp cụ thể.
* **Hồi quy (Regression)**

Hồi quy (Regression) là quá trình dự đoán giá trị số của một biến mục tiêu dựa trên các đặc trưng mô tả của dữ liệu. Mục tiêu của hồi quy là xác định mối quan hệ toán học giữa các đặc trưng và giá trị mục tiêu để dự đoán giá trị mục tiêu cho các điểm dữ liệu mới.

Đặc điểm:

* Đầu vào: Dữ liệu đầu vào cho mô hình hồi quy bao gồm các điểm dữ liệu với các đặc trưng liên tục.
* Đầu ra: Đầu ra của mô hình hồi quy là giá trị số liên tục

Các thuật toán phổ biến:

* Linear Regression Dự đoán giá trị mục tiêu bằng một hàm tuyến tính của các đặc trưng.
* Ridge và Lasso: Mở rộng của hồi quy tuyến tính để giảm thiểu overfitting và cải thiện độ chính xác.
* Support Vector Regression: Áp dụng SVM cho bài toán hồi quy để tìm ra mặt siêu phẳng tối ưu.

#### Học máy không giám sát (Unsupervised Learning)

Học không giám sát là một kỹ thuật trong học máy nhằm phát hiện các mô hình ẩn trong dữ liệu mà không cần biết trước nhãn đầu ra của các đầu vào. Khác với học có giám sát, phương pháp này làm việc với dữ liệu không có thông tin về kết quả đúng. Dữ liệu đầu vào được xem như một tập hợp các biến ngẫu nhiên, và từ đó, một mô hình mật độ kết hợp được xây dựng để thể hiện cấu trúc dữ liệu. Học không giám sát cũng có thể kết hợp với suy luận Bayes để ước tính xác suất có điều kiện của một biến ngẫu nhiên, dựa trên các giá trị của các biến khác.

Ngoài ra, học không giám sát còn có ứng dụng trong nén dữ liệu, bởi vì các thuật toán nén thường dựa trên việc mô hình hóa phân phối xác suất của tập dữ liệu, dù thực hiện theo cách tường minh hay ngầm định.

Một hình thức phổ biến khác của học không giám sát là phân cụm dữ liệu. Phân cụm không nhất thiết dựa trên xác suất, mà tập trung vào việc nhóm các dữ liệu có đặc điểm tương đồng thành từng cụm riêng biệt.

* **Phân cụm (Clustering)**

Phân cụm là quá trình nhóm các điểm dữ liệu có đặc điểm tương tự vào các nhóm sao cho các điểm trong cùng một nhóm có sự tương đồng cao hơn so với các điểm thuộc nhóm khác. Mục tiêu của phân cụm là xác định cách nhóm các điểm dữ liệu mà không cần phải có nhãn lớp rõ ràng.

Đặc điểm chính:

* Đầu vào: Dữ liệu đầu vào cho mô hình phân cụm là các điểm dữ liệu mô tả bằng các đặc trưng.
* Đầu ra: Đầu ra của mô hình phân cụm là các nhóm hoặc cụm, mỗi cụm có các điểm dữ liệu có tính chất tương tự nhau.

Các thuật toán phổ biến:

* K-means: Phân cụm các điểm dữ liệu bằng cách tìm các trung tâm cụm sao cho tổng bình phương khoảng cách từ các điểm đến trung tâm cụm là nhỏ nhất.
* Phân cụm phân cấp (Hierarchical Clustering): Xây dựng một cây phân cụm của các điểm dữ liệu để tìm các cụm con và cụm cha dựa trên độ tương đồng.
* Phân cụm dựa trên mô hình (Model-based Clustering): Sử dụng các mô hình thống kê để phân tích phân phối của các cụm trong không gian dữ liệu.

Ứng dụng: Phân cụm được sử dụng rộng rãi trong nhiều lĩnh vực như marketing để nhóm khách hàng, xử lý dữ liệu để nhóm tin nhắn, và phân tích hành vi người dùng trên mạng xã hội.

* **Phát hiện ngoại lệ (Anomaly Detection)**

Phát hiện ngoại lệ là quá trình nhận diện các điểm dữ liệu không giống như phần còn lại của dữ liệu, thường được gọi là ngoại lệ hoặc sự cố. Mục tiêu của phát hiện ngoại lệ là xác định những điểm dữ liệu có tính chất khác biệt so với các điểm dữ liệu bình thường.

Đặc điểm:

* Đầu vào: Dữ liệu đầu vào cho mô hình phát hiện ngoại lệ là các điểm dữ liệu mô tả bằng các đặc trưng.
* Đầu ra: Đầu ra của mô hình phát hiện ngoại lệ là các điểm dữ liệu được xác định là ngoại lệ hoặc không phải ngoại lệ.

Các thuật toán phổ biến:

* Ngưỡng (Threshold): Thiết lập một ngưỡng để xác định các điểm dữ liệu ngoại lệ dựa trên các giới hạn định sẵn.
* Phân phối (Statistical Methods): Sử dụng các phân phối thống kê như phân phối Gaussian để xác định các điểm dữ liệu có xác suất thấp.

Ứng dụng: Phát hiện ngoại lệ được áp dụng trong các lĩnh vực như an ninh mạng để phát hiện các hành vi độc hại, trong công nghiệp để phát hiện lỗi sản xuất, và trong tài chính để phát hiện gian lận.

#### Học máy bán giám sát (Semi-supervised Learning)

Trong khoa học máy tính, học bán giám sát là một phương pháp học máy sử dụng cả dữ liệu có nhãn và không có nhãn để huấn luyện mô hình.. Cụ thể, phương pháp này thường sử dụng một lượng nhỏ dữ liệu có nhãn kết hợp với một lượng lớn dữ liệu chưa nhãn. Học bán giám sát có thể được xem là sự kết hợp giữa học không giám sát (không có dữ liệu nhãn) và học có giám sát (toàn bộ dữ liệu đều có nhãn). Nhiều nghiên cứu cho thấy rằng việc kết hợp dữ liệu không nhãn với một ít dữ liệu có nhãn có thể giúp cải thiện đáng kể độ chính xác của mô hình. Việc gán nhãn dữ liệu cho các bài toán học máy thường yêu cầu chuyên gia có kỹ năng cao để phân loại thủ công các ví dụ huấn luyện. Quy trình này tốn kém và khiến việc sử dụng toàn bộ dữ liệu có nhãn trở nên không khả thi, trong khi dữ liệu không nhãn lại rẻ hơn nhiều. Do đó, học bán giám sát trở nên rất hữu ích trong các tình huống này.

Một ví dụ của học bán giám sát là kỹ thuật đồng huấn luyện (co-training), trong đó nhiều bộ học cùng được huấn luyện trên cùng một tập ví dụ, nhưng mỗi bộ sử dụng một tập đặc trưng khác nhau và lý tưởng là độc lập với nhau.

Một phương pháp khác là mô hình hóa phân phối xác suất đồng thời giữa các đặc trưng và nhãn. Khi dữ liệu không có nhãn, các nhãn có thể được xem như là "dữ liệu bị thiếu". Các kỹ thuật xử lý dữ liệu thiếu, như lấy mẫu Gibbs và tối ưu hóa kỳ vọng, có thể được sử dụng để ước tính các tham số.

#### Học máy tăng cường (Reinforcement Learning)

Trong lĩnh vực khoa học máy tính, học tăng cường là một nhánh của học máy, tập trung nghiên cứu cách một agent trong một môi trường chọn các hành động nhằm tối đa hóa phần thưởng lâu dài. Các thuật toán học tăng cường nhằm tìm kiếm một chiến lược giúp ánh xạ các trạng thái của môi trường thành các hành động mà agent nên thực hiện trong những trạng thái đó.

Môi trường trong học tăng cường thường được mô hình hóa như một quá trình quyết định Markov với các trạng thái hữu hạn. Các thuật toán học tăng cường trong ngữ cảnh này thường liên quan đến các phương pháp quy hoạch động. Các xác suất chuyển trạng thái và xác suất nhận phần thưởng trong quá trình quyết định Markov thường là ngẫu nhiên, nhưng lại không thay đổi trong suốt quá trình của bài toán.

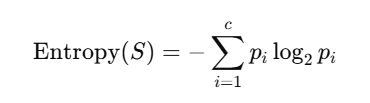
Khác với học có giám sát, học tăng cường không yêu cầu các cặp dữ liệu đầu vào và kết quả đúng. Các hành động gần tối ưu không được đánh giá rõ ràng là đúng hay sai. Hơn nữa, trong học tăng cường, việc hoạt động trực tuyến và tìm kiếm sự cân bằng giữa khám phá và khai thác là rất quan trọng. Vấn đề này thường được nghiên cứu qua bài toán multi-armed bandit, nơi việc tối ưu hóa giữa khám phá những hành động mới và khai thác những hành động đã biết được phân tích kỹ lưỡng.

## Các mô hình phân loại

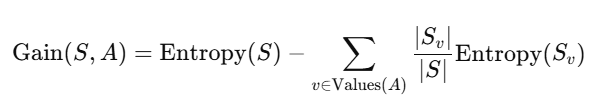
### Thuật toán Iterative Dichotomiser 3

ID3 là một thuật toán xây dựng cây quyết định. Thuật toán này sử dụng khái niệm entropy và gain thông tin để xác định thuộc tính tốt nhất tại mỗi nút trong cây quyết định. Quá trình hoạt động của ID3 dựa trên việc phân loại dữ liệu một cách đệ quy, tiếp tục chia nhỏ các dữ liệu cho đến khi đạt được phân loại hoàn chỉnh hoặc không thể chia tiếp được nữa.

* **Nguyên lý hoạt động**
* Tính Entropy



* pi​: Tỷ lệ phần tử thuộc lớp i trong tập dữ liệu S.
* **Tính Gain thông tin**

****

* A: Thuộc tính đang xét.
* Sv​: Tập con dữ liệu khi A=v
* Chọn thuộc tính tốt nhất: Chọn thuộc tính có Gain thông tin lớn nhất để làm nút phân chia tại mỗi bước.
* Lặp lại: Áp dụng đệ quy trên các tập con, tiếp tục phân chia cho đến khi tất cả các điểm dữ liệu trong tập con thuộc cùng một lớp hoặc không còn thuộc tính nào để chia.
* **Đầu vào và đầu ra**
* Đầu vào: Dữ liệu huấn luyện gồm các đặc trưng (features) và nhãn (labels).
* Đầu ra: Một cây quyết định có thể sử dụng để phân loại hoặc dự đoán dữ liệu mới.
* **Đặc điểm**
* **Phân loại dựa trên thuộc tính**: Sử dụng các thuộc tính để chia dữ liệu thành các nhóm.
* **Cây quyết định**: Kết quả của thuật toán là một cây quyết định với các nút lá biểu thị kết quả phân loại.
* **Dựa vào Entropy và Gain thông tin**: ID3 chọn thuộc tính chia nhỏ dữ liệu dựa trên giá trị Gain thông tin cao nhất.
* **Ưu điểm**
* **Dễ hiểu và dễ triển khai**: Cây quyết định có cấu trúc rõ ràng, dễ dàng giải thích và biểu diễn trực quan.
* **Tốc độ nhanh**: Xây dựng cây với tốc độ nhanh đối với tập dữ liệu nhỏ hoặc trung bình.
* **Hiệu quả với dữ liệu phân loại**: Phù hợp khi các thuộc tính có dạng phân loại.
* **Ứng dụng thực tế**
* Hệ thống chẩn đoán y khoa: Xây dựng cây quyết định để hỗ trợ bác sĩ chẩn đoán bệnh dựa trên triệu chứng.
* Quản lý rủi ro tài chính: Phân loại các khoản vay có rủi ro cao và thấp dựa trên thông tin tài chính.
* Thương mại điện tử: Phân loại khách hàng theo hành vi mua sắm để tối ưu hóa quảng cáo.
* Phân tích giáo dục: Phân loại học sinh theo thành tích dựa trên các yếu tố như thời gian học, điểm số.

### Thuật toán Support Vector Machine

Support Vector Machine là một thuật toán học máy dựa trên lý thuyết không gian hình học. Trong các bài toán phân loại, mục tiêu của Support Vector Machine là tìm ra một siêu phẳng (hyperplane) trong không gian nhiều chiều sao cho có thể phân chia các lớp dữ liệu khác nhau mà không làm sai lệch dữ liệu của lớp nào. Siêu phẳng này không chỉ phân tách các điểm dữ liệu mà còn tối đa hóa khoảng cách giữa siêu phẳng và các điểm dữ liệu gần nhất của mỗi lớp. Những điểm này được gọi là vector hỗ trợ, vì chúng là các điểm quan trọng nhất trong việc xác định siêu phẳng tối ưu.

* **Nguyên lý hoạt động**
* SVM cố gắng tìm một siêu phẳng trong không gian đặc trưng để phân chia dữ liệu thành các lớp khác nhau.
* Siêu phẳng tối ưu là siêu phẳng tối đa hóa khoảng cách giữa các điểm dữ liệu gần nhất thuộc các lớp khác nhau, được gọi là vector hỗ trợ.
* SVM có thể sử dụng hạt nhân để chuyển đổi dữ liệu từ không gian ban đầu sang không gian đặc trưng cao hơn, giúp giải quyết các bài toán không tuyến tính.
* **Đầu vào và đầu ra**
* Đầu vào:
* Tập dữ liệu huấn luyện với các đặc trưng và nhãn.
* Lựa chọn hạt nhân (kernel) phù hợp (linear, polynomial, RBF).
* Đầu ra:
* Với phân loại: Dự đoán nhãn cho dữ liệu mới.
* Với hồi quy: Dự đoán giá trị liên tục.
* **Ưu điểm**

SVM, là một kỹ thuật phân lớp phổ biến, có nhiều ưu điểm nổi bật, đặc biệt là trong việc tính toán hiệu quả trên các tập dữ liệu lớn.

* **Xử lý trong không gian chiều cao**: SVM là một công cụ mạnh mẽ khi làm việc với các không gian có số chiều cao, điều này đặc biệt hữu ích trong các bài toán phân loại văn bản và phân tích cảm xúc, nơi mà không gian đặc trưng có thể rất lớn.
* **Tiết kiệm bộ nhớ:** SVM chỉ lưu trữ một tập con các điểm dữ liệu trong quá trình huấn luyện và khi đưa ra quyết định cho các điểm mới, điều này giúp giảm thiểu việc sử dụng bộ nhớ. Chỉ những điểm dữ liệu quan trọng nhất, gọi là vector hỗ trợ, được lưu trữ và sử dụng trong quá trình dự đoán.
* **Tính linh hoạt cao**: SVM cho phép áp dụng các hàm hạt nhân (kernel), giúp chuyển đổi dữ liệu vào không gian chiều cao hơn để phân tách các lớp dữ liệu phi tuyến tính. Khả năng linh động trong việc sử dụng các kernel khác nhau giúp SVM có thể chuyển đổi giữa phương pháp phân lớp tuyến tính và phi tuyến tính, cải thiện hiệu suất phân loại.
* **Ứng dụng thực tế**
* **Nhận dạng chữ viết tay:** SVM hỗ trợ chuyển đổi chữ viết tay từ tài liệu giấy thành văn bản kỹ thuật số, giúp tự động hóa quá trình số hóa tài liệu và xử lý các mẫu đơn, tài liệu hành chính.
* **Hệ thống gợi ý:** Được sử dụng trong các hệ thống khuyến nghị, chẳng hạn như trên các nền tảng thương mại điện tử, để gợi ý sản phẩm hoặc dịch vụ cho người dùng dựa trên thói quen mua sắm và sở thích của họ.
* **Dự báo khoa học:** Sử dụng trong dự báo thời tiết, giúp phân loại các hiện tượng khí tượng như bão, mưa lớn hoặc sóng nhiệt.
* **Mạng xã hội:** SVM giúp phân loại người dùng trong các nền tảng mạng xã hội như Facebook hoặc Twitter dựa trên các yếu tố như hành vi, sở thích hoặc độ tuổi, từ đó cá nhân hóa trải nghiệm người dùng

### Thuật toán Artificial Neural Network

Mạng Nơ-ron Nhân tạo là một thuật toán học máy được thiết kế dựa trên cách thức hoạt động của hệ thần kinh con người. ANN giúp xử lý và phân tích các dữ liệu phức tạp thông qua các mô hình học sâu. Mạng này bao gồm nhiều lớp nơ-ron, với mỗi lớp có vai trò xử lý thông tin theo cách tương tự cách bộ não con người xử lý tín hiệu.

* **Nguyên lý hoạt động**
* Dữ liệu đầu vào được đưa vào các nơ-ron của lớp đầu tiên.
* Các nơ-ron trong mỗi lớp sẽ tính toán giá trị đầu ra bằng cách sử dụng trọng số và hàm kích hoạt.
* Kết quả đầu ra từ lớp cuối cùng được so sánh với giá trị mục tiêu để tính toán lỗi.
* Thuật toán sẽ sử dụng các phương pháp tối ưu để điều chỉnh trọng số nhằm giảm thiểu lỗi.
* **Đầu vào và đầu ra**
* Đầu vào: Tập dữ liệu huấn luyện với các đặc trưng và nhãn.
* Đầu ra: Dự đoán nhãn cho dữ liệu mới.
* **Đặc điểm**
* **Cấu trúc mạng:** Bao gồm các lớp nơ-ron, bao gồm lớp đầu vào, một hoặc nhiều lớp ẩn, và lớp đầu ra.
* **Học giám sát và không giám sát:** Có thể huấn luyện với cả dữ liệu có nhãn hoặc không có nhãn.
* **Hàm kích hoạt:** Các hàm kích hoạt như ReLU, Sigmoid, hoặc Tanh được sử dụng để điều chỉnh tín hiệu trong các nơ-ron.
* **Khả năng tự học:** ANN có khả năng tự điều chỉnh và học từ dữ liệu mà không cần phải lập trình chi tiết.
* **Ưu điểm**
* **Khả năng học phi tuyến tính**: Mạng nơ-ron có thể mô hình hóa các mối quan hệ phi tuyến tính rất tốt, điều mà các phương pháp học máy truyền thống khó thực hiện.
* **Khả năng tự học**: Không cần phải biết rõ quy luật chi tiết của vấn đề, mạng có thể tự học từ dữ liệu.
* **Xử lý dữ liệu phức tạp**: ANN có thể xử lý các vấn đề phức tạp như nhận diện hình ảnh, phân loại văn bản, dự đoán thời gian, ...
* **Khả năng mở rộng**: Có thể mở rộng kích thước mạng, thêm các lớp và nơ-ron để tăng độ chính xác và hiệu suất.
* **Ứng dụng thực tế**
* **Nhận dạng hình ảnh**: Các ứng dụng như nhận diện khuôn mặt, nhận diện chữ viết tay, nhận diện đối tượng trong video.
* **Xử lý ngôn ngữ tự nhiên**: Dùng trong các hệ thống dịch máy, chatbot, phân tích cảm xúc, nhận dạng giọng nói.
* **Dự báo và phân tích dữ liệu**: Dự đoán giá cổ phiếu, phân tích tín dụng, dự báo nhu cầu tiêu dùng.
* **Xe tự lái**: Các mạng nơ-ron giúp xe tự lái nhận diện môi trường xung quanh và đưa ra quyết định điều khiển.
* **Y tế**: Chẩn đoán bệnh qua hình ảnh y tế, phân tích gen, dự đoán hiệu quả điều trị.

## Quy trình tổng quan

### Xác định vấn đề

Việc xác định bài toán học máy là một bước hết sức quan trọng trong quy trình phát triển mô hình học máy, vì nó không chỉ định hướng toàn bộ quá trình mà còn quyết định các phương pháp và công cụ cần sử dụng, từ việc thu thập dữ liệu cho đến việc lựa chọn mô hình và các chỉ số đánh giá. Để có thể xây dựng một mô hình học máy hiệu quả, bạn phải hiểu rõ vấn đề mà mình đang cố gắng giải quyết, bởi vì mỗi bài toán có những đặc điểm và yêu cầu riêng biệt

A group of people standing on a question mark

Description automatically generated

Hình . Xác định vấn đề

### Thu thập dữ liệu

Thu thập dữ liệu là quá trình thu nhận và tập hợp thông tin từ nhiều nguồn khác nhau nhằm phục vụ các mục đích cụ thể như nghiên cứu, phân tích, hoặc phát triển mô hình học máy. Đây là một bước quan trọng và cần thiết trong quy trình xây dựng mô hình, vì dữ liệu chất lượng cao sẽ là nền tảng để phát triển mô hình học máy hiệu quả. Việc thu thập dữ liệu chính xác và đầy đủ là yếu tố quyết định trong việc đảm bảo rằng mô hình sẽ hoạt động chính xác và đạt được kết quả mong muốn.

A group of people working on a cellphone

Description automatically generated

Hình . Thu thập dữ liệu

### Tiền xử lý dữ liệu

Tiền xử lý dữ liệu là một bước thiết yếu trong quy trình xây dựng mô hình học máy, nhằm chuyển đổi và chuẩn bị dữ liệu thô thành một dạng mà các thuật toán học máy có thể xử lý và học từ đó. Mục tiêu chính của tiền xử lý dữ liệu là nâng cao chất lượng dữ liệu, loại bỏ những yếu tố gây nhiễu hoặc không chính xác, đồng thời cải thiện khả năng áp dụng của mô hình học máy. Khi dữ liệu được chuẩn hóa, xử lý missing values, và loại bỏ các nhiễu không cần thiết, mô hình sẽ học được các mẫu và mối quan hệ trong dữ liệu một cách hiệu quả hơn, từ đó nâng cao độ chính xác và hiệu suất tổng thể.

A person and person standing next to a computer

Description automatically generated

Hình . Tiền xử lý dữ liệu

### Lựa chọn mô hình

Lựa chọn mô hình là một bước quan trọng trong quy trình xây dựng mô hình học máy, nhằm xác định thuật toán phù hợp để giải quyết vấn đề cụ thể. Quyết định này ảnh hưởng trực tiếp đến hiệu suất dự đoán và tính khả thi của giải pháp, vì một mô hình không phù hợp có thể dẫn đến kết quả không chính xác hoặc khó triển khai trong thực tế.

Lựa chọn mô hình học máy phụ thuộc vào các yếu tố sau:

* **Loại bài toán**: Phân loại, hồi quy, phát hiện bất thường hay clustering, mỗi loại yêu cầu mô hình khác nhau (ví dụ: Logistic Regression cho phân loại, Linear Regression cho hồi quy).
* **Bản chất của dữ liệu**: Dữ liệu có thể là có cấu trúc hoặc phi cấu trúc, và mô hình phải phù hợp với đặc điểm dữ liệu.
* **Yêu cầu về độ chính xác**: Mô hình đơn giản có thể đủ cho yêu cầu thấp, trong khi mô hình phức tạp cần cho yêu cầu độ chính xác cao.
* **Hạn chế về tài nguyên tính toán**: Mô hình phức tạp như deep learning đòi hỏi tài nguyên mạnh mẽ, trong khi các mô hình như Iterative Dichotomiser 3 hay SVM ít tốn kém hơn và phù hợp với nguồn lực hạn chế.

A person with his hand on his chin

Description automatically generated

Hình . Lựa chọn mô hình

### Huấn luyện mô hình

Huấn luyện là bước quan trọng trong xây dựng mô hình học máy, nơi dữ liệu đã được chuẩn bị được cung cấp cho mô hình để học từ các mẫu và mối quan hệ trong dữ liệu. Trong quá trình này, mô hình điều chỉnh các tham số của mình thông qua các thuật toán tối ưu hóa, như Gradient Descent, nhằm giảm sai số dự đoán. Mục tiêu là giúp mô hình học được mối quan hệ giữa đặc trưng và giá trị mục tiêu, từ đó dự đoán chính xác hơn. Quá trình huấn luyện cần được thực hiện cẩn thận để tránh overfitting hoặc underfitting.

A cartoon of a robot reading a book

Description automatically generated

Hình . Huấn luyện mô hình

### Đánh giá mô hình

Đánh giá mô hình là quá trình kiểm tra và xác định hiệu suất của mô hình học máy thông qua các chỉ số và kỹ thuật đánh giá khác nhau. Mục đích chính của việc đánh giá là kiểm tra xem mô hình có khả năng hoạt động tốt trên dữ liệu mới, chưa từng thấy, hay không. Điều này rất quan trọng để đảm bảo rằng mô hình không chỉ phù hợp với dữ liệu huấn luyện mà còn có thể tổng quát và dự đoán chính xác trên dữ liệu thực tế.

Quá trình đánh giá mô hình bao gồm việc sử dụng các chỉ số như độ chính xác (accuracy), độ chính xác của từng lớp (precision, recall), F1-score, hoặc các chỉ số khác tùy thuộc vào loại bài toán (phân loại, hồi quy, v.v.). Các kỹ thuật như cross-validation cũng có thể được sử dụng để đảm bảo rằng kết quả đánh giá là đáng tin cậy và không bị ảnh hưởng bởi việc chia dữ liệu ngẫu nhiên.

Kết quả đánh giá sẽ giúp bạn quyết định xem mô hình có đủ hiệu quả để triển khai trong thực tế hay không, đồng thời đưa ra những cải tiến nếu cần thiết để tối ưu hóa mô hình trước khi áp dụng vào các tình huống thực tế.

A person sitting at a desk with a robot on it

Description automatically generated

Hình . Đánh giá mô hình

### Tinh chỉnh mô hình

Tinh chỉnh mô hình (Hyperparameter Tuning) là quá trình điều chỉnh các siêu tham số của mô hình học máy nhằm tối ưu hóa hiệu suất dự đoán. Siêu tham số (hyperparameters) là các tham số không được học từ dữ liệu trong quá trình huấn luyện mà phải được thiết lập trước khi bắt đầu. Ví dụ về siêu tham số bao gồm tốc độ học (learning rate), số lượng lớp trong mạng nơ-ron, số lượng cây trong Iterative Dichotomiser 3, hoặc giá trị của tham số C trong SVM.

Quá trình tinh chỉnh này thường đòi hỏi thử nghiệm với nhiều giá trị khác nhau của các siêu tham số để tìm ra bộ giá trị tối ưu. Các phương pháp như grid search (tìm kiếm qua lưới) hoặc random search (tìm kiếm ngẫu nhiên) có thể được sử dụng để đánh giá hiệu suất mô hình với các giá trị khác nhau của siêu tham số. Khi các siêu tham số được tinh chỉnh đúng, mô hình có thể đạt được hiệu suất cao hơn và dự đoán chính xác hơn trên dữ liệu chưa thấy.

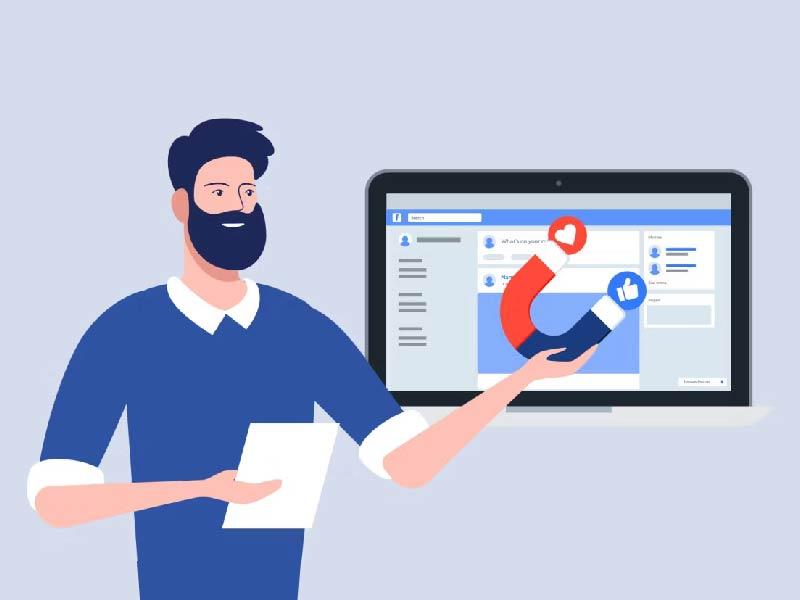
A hand holding a finger up

Description automatically generated

Hình . Tinh chỉnh mô hình

### Triển khai mô hình

Triển khai mô hình là quá trình chuyển mô hình học máy từ môi trường phát triển sang môi trường sản xuất, nơi mô hình có thể thực hiện các dự đoán hoặc phân tích trong các ứng dụng thực tế. Quá trình này bao gồm các bước quan trọng như chuẩn bị mô hình, thiết lập cơ sở hạ tầng phù hợp, và triển khai mô hình vào hệ thống sản xuất. Sau khi triển khai, mô hình cần được giám sát và bảo trì thường xuyên để đảm bảo hiệu suất và tính ổn định trong suốt quá trình vận hành.



Hình . Triển khai mô hình

# ỨNG DỤNG VÀO DƯ ĐOÁN CƠN ĐỘNG KINH

## Công nghệ sử dụng

### Ngôn ngữ

#### Python

Python là một ngôn ngữ lập trình mạnh mẽ và phổ biến, được ứng dụng rộng rãi trong các lĩnh vực như phát triển web, phần mềm, khoa học dữ liệu và học máy (Machine Learning). Nguyên nhân chính khiến Python được yêu thích là nhờ vào tính hiệu quả, dễ học và dễ sử dụng. Cú pháp của ngôn ngữ này đơn giản và dễ hiểu, giúp các lập trình viên nhanh chóng làm quen và phát triển ứng dụng mà không cần phải học các cú pháp phức tạp. Chính vì vậy, Python là lựa chọn ưu tiên cho cả người mới bắt đầu và các chuyên gia trong ngành.

Một trong những ưu điểm của Python là tính tương thích cao với nhiều hệ điều hành, bao gồm Windows, macOS, Linux và các nền tảng khác. Điều này cho phép các nhà phát triển triển khai ứng dụng Python trên nhiều hệ thống mà không gặp phải khó khăn do sự khác biệt giữa các môi trường hoạt động.

#### HTML CSS

HTML (HyperText Markup Language) là ngôn ngữ đánh dấu được sử dụng để cấu trúc và tổ chức nội dung của một trang web. Nó định nghĩa các phần tử của trang web như tiêu đề, đoạn văn, danh sách, hình ảnh, và các liên kết. HTML cung cấp cấu trúc cơ bản giúp trình duyệt hiển thị nội dung.

CSS (Cascading Style Sheets) là ngôn ngữ dùng để định dạng và thiết kế giao diện của một trang web. CSS cho phép điều chỉnh màu sắc, font chữ, khoảng cách, vị trí các phần tử, và nhiều thuộc tính khác, giúp tạo ra một giao diện đẹp mắt và dễ sử dụng. CSS giúp tách biệt phần cấu trúc (HTML) và phần giao diện (CSS) của trang web, làm cho việc duy trì và thay đổi giao diện trở nên dễ dàng hơn.

### Các thư viện sử dụng

#### Numpy

NumPy là một thư viện quan trọng trong Python, được phát triển để xử lý các mảng (arrays) và thực hiện các phép toán số học. Đây là một công cụ thiết yếu trong khoa học dữ liệu và học máy, vì hầu hết các thư viện khác đều xây dựng trên nền tảng của NumPy.

Nổi bật của NumPy bao gồm:

* **Mảng n chiều (ndarray):** NumPy cung cấp đối tượng mảng, cho phép lưu trữ và xử lý dữ liệu hiệu quả, hỗ trợ các phép toán số học.
* **Tính toán nhanh chóng:** Các phép toán trong NumPy được tối ưu hóa với C và Fortran, giúp thực hiện nhanh chóng.
* **Hỗ trợ các phép toán toán học, đại số và thống kê:** Thư viện này cung cấp các hàm cho phép thực hiện các phép toán ma trận, thống kê, biến đổi Fourier và nhiều công cụ hữu ích khác trong khoa học dữ liệu.

#### Pandas

Pandas là một thư viện mạnh mẽ trong Python, chuyên dụng cho việc xử lý và phân tích dữ liệu, đặc biệt là dữ liệu dạng bảng. Thư viện này giúp bạn thực hiện các công việc như làm sạch, phân tích và trích xuất thông tin từ dữ liệu một cách dễ dàng.

Các tính năng nổi bật của Pandas bao gồm:

* **DataFrame và Series:** Pandas cung cấp hai cấu trúc dữ liệu chính là DataFrame và Series, giúp việc xử lý và thao tác với dữ liệu trở nên dễ dàng và linh hoạt
* **Xử lý dữ liệu thiếu (missing data):** Pandas cung cấp các công cụ đơn giản để thay thế, loại bỏ hoặc xử lý dữ liệu thiếu một cách hiệu quả.
* **Hỗ trợ dữ liệu theo chuỗi thời gian:** Pandas có khả năng mạnh mẽ trong việc xử lý và phân tích dữ liệu thời gian, bao gồm các tính năng lọc, nhóm và chuyển đổi dữ liệu theo thời gian.
* **Tương thích với nhiều nguồn dữ liệu:** Pandas hỗ trợ nhập và xuất dữ liệu từ các định dạng phổ biến như CSV, Excel, SQL, JSON, và nhiều loại khác.

#### Scikit-learn

Scikit-learn là một thư viện Python mạnh mẽ dành cho học máy, cung cấp nhiều thuật toán học máy hiệu quả cho các tác vụ như phân loại, hồi quy, phân cụm, giảm chiều dữ liệu và các nhiệm vụ khác.

Các tính năng nổi bật của Scikit-learn bao gồm:

* **Đa dạng thuật toán học máy:** Thư viện hỗ trợ nhiều thuật toán học máy như SVM, hồi quy tuyến tính, cây quyết định, K-means, Naive Bayes và nhiều thuật toán khác.
* **Giao diện dễ sử dụng:** Scikit-learn cung cấp một API thống nhất, dễ sử dụng, giúp người dùng dễ dàng triển khai các mô hình học máy mà không phải lo lắng về các chi tiết kỹ thuật phức tạp.
* **Tiền xử lý dữ liệu:** Thư viện này cung cấp các công cụ mạnh mẽ để xử lý dữ liệu, như chuẩn hóa, mã hóa phân loại, giảm chiều dữ liệu (PCA), và nhiều kỹ thuật khác.
* **Đánh giá và chọn mô hình:** Scikit-learn cung cấp các công cụ như cross-validation, grid search và các phương pháp khác để đánh giá hiệu quả của các mô hình học máy

#### Matplotlib

Matplotlib là một thư viện vẽ đồ thị trong Python, được sử dụng phổ biến trong khoa học dữ liệu và phân tích số liệu. Thư viện này cho phép người dùng tạo ra các loại đồ thị và biểu đồ khác nhau, từ biểu đồ đường (line plot), biểu đồ cột (bar plot) đến biểu đồ phân tán (scatter plot) và biểu đồ 3D. Matplotlib hỗ trợ các tùy chọn trực quan hóa dữ liệu mạnh mẽ, giúp các nhà phân tích dữ liệu và nhà nghiên cứu có thể trình bày kết quả một cách rõ ràng và dễ hiểu.

Các tính năng nổi bật của Matplotlib:

* Tạo các biểu đồ đơn giản hoặc phức tạp, tùy chỉnh các tham số như màu sắc, kiểu đường, và hình dạng điểm.
* Hỗ trợ trực quan hóa dữ liệu 3D và các loại đồ thị phức tạp khác.
* Tích hợp dễ dàng với các thư viện khác như Pandas để tạo ra biểu đồ từ DataFrame.

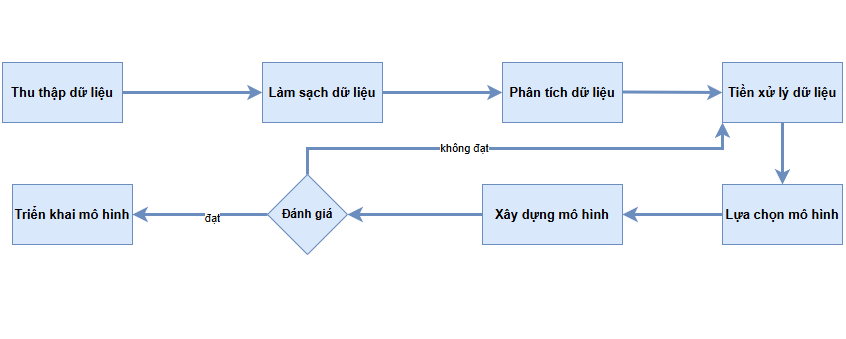
## Phát biểu bài toán

Bài toán dự đoán cơn động kinh thông qua các phương pháp học máy nhằm mục tiêu phát hiện và dự đoán các cơn co giật bằng cách phân tích dữ liệu điện não đồ. Đặc trưng của bài toán này là xử lý khối lượng lớn dữ liệu phức tạp, trong đó chứa nhiều tín hiệu nhiễu và biến đổi liên tục, đòi hỏi các kỹ thuật phân tích tiên tiến.

Mục tiêu chính là phát triển một mô hình học máy có khả năng nhận diện các mẫu tín hiệu đặc trưng của cơn động kinh, từ đó dự đoán thời điểm xảy ra cơn co giật trước khi nó xuất hiện. Việc dự đoán chính xác không chỉ nâng cao độ nhạy trong chẩn đoán mà còn giúp bệnh nhân chuẩn bị tâm lý và vật lý, giảm thiểu nguy cơ chấn thương và các biến chứng nghiêm trọng.

Các phương pháp như ANN, SVM, Iterative Dichotomiser 3 sẽ được áp dụng để tối ưu hóa khả năng phát hiện và dự đoán, nhằm cải thiện hiệu quả chẩn đoán. Thông qua việc khai thác các kỹ thuật học sâu và xử lý tín hiệu tiên tiến, nghiên cứu này không chỉ khắc phục những hạn chế của phương pháp truyền thống mà còn mở ra hướng đi mới trong việc nâng cao chất lượng cuộc sống cho bệnh nhân động kinh.

## Quy trình tổng quan bài toán



Hình . Quy trình tổng quan bài toán

## Thu thập dữ liệu

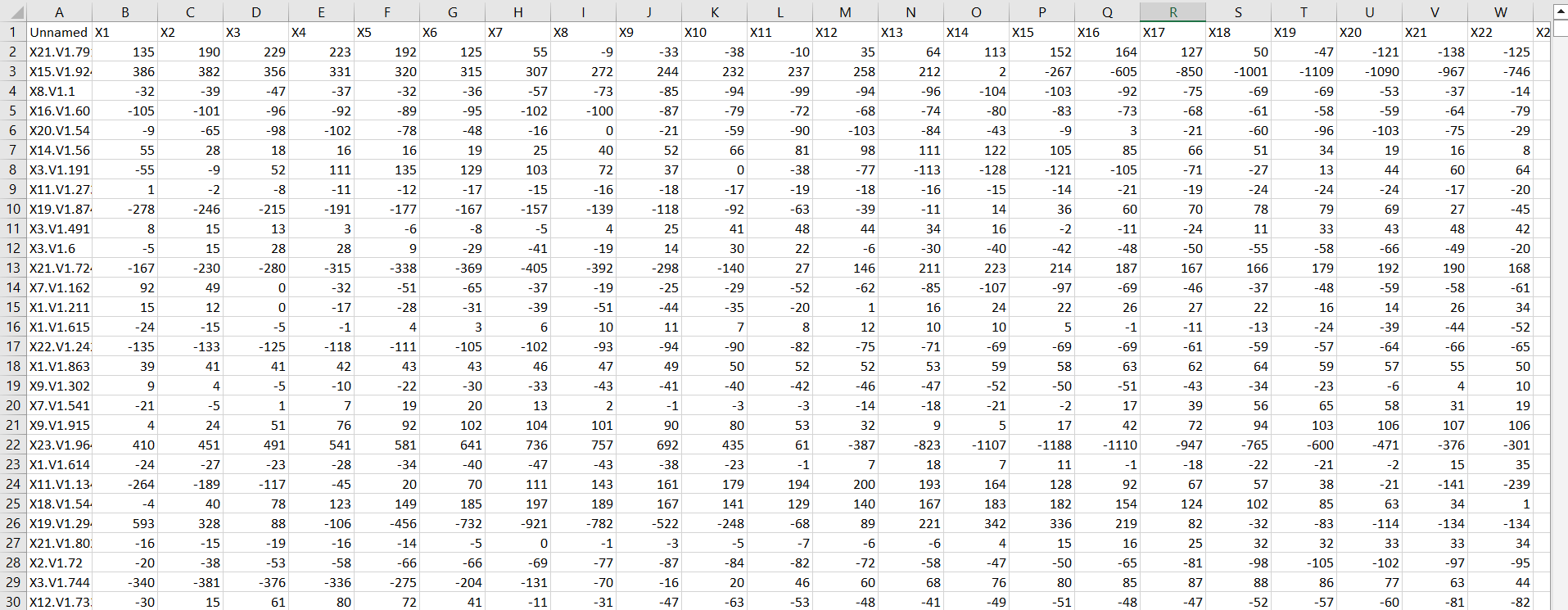
Tập dữ liệu này là một phiên bản được tinh chỉnh và tổ chức lại từ một nguồn dữ liệu phổ biến chuyên dùng để phát hiện cơn động kinh. Nó bao gồm thông tin từ 500 cá nhân, mỗi người có một đoạn ghi lại hoạt động điện não đồ (EEG) kéo dài 23,6 giây. Dữ liệu ban đầu được lấy mẫu với 4097 điểm dữ liệu, sau đó chia thành các khối nhỏ hơn, mỗi khối chứa 178 điểm dữ liệu (tương ứng 1 giây). Tổng cộng, tập dữ liệu tạo ra 11.500 hàng, với 178 cột biểu thị các đặc trưng và một cột phản hồi chứa nhãn kết quả, đại diện cho các trạng thái.

Trong bài toán phân loại cơn động kinh từ dữ liệu EEG, có một tập dữ liệu bao gồm 178 đặc trưng EEG và 5 lớp nhãn khác nhau. Mục tiêu chính của tập dữ liệu này là xác định chính xác các cơn động kinh từ tín hiệu EEG, thực hiện phân loại nhị phân giữa lớp nhãn 1 (đại diện cho cơn động kinh) và các lớp còn lại 2, 3, 4, 5(không phải cơn động kinh).

Tập dữ liệu này mang lại giá trị lớn trong nghiên cứu phát triển thuật toán học máy nhằm tối ưu hóa việc phát hiện và phân loại cơn động kinh dựa trên tín hiệu EEG. Những ứng dụng này có tiềm năng hỗ trợ hiệu quả các bác sĩ trong việc chẩn đoán và điều trị bệnh nhân, góp phần nâng cao chất lượng chăm sóc y tế.

Nguồn dữ liệu:

<https://www.kaggle.com/code/harunshimanto/machine-learning-algorithms-for-epileptic-seizures/notebook>

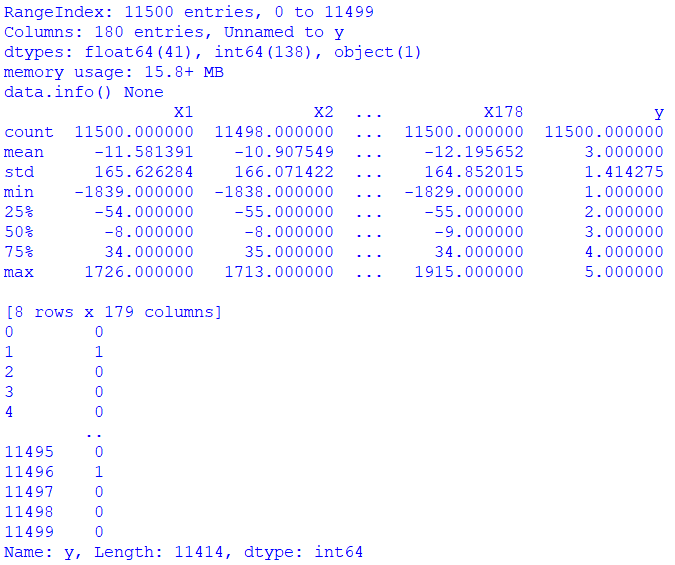


Hình . Thông tin dữ liệu thu thập

## Làm sạch dữ liệu

### Xóa các thông tin bị thiếu

Trong quá trình thu thập dữ liệu thiếu có thể xuất hiện do nhiều nguyên nhân, chẳng hạn như lỗi trong quá trình thu thập dữ liệu hoặc sự cố trong thiết bị đo tín hiêu não bộ. Việc xóa các dòng bị thiếu là một bước quan trọng nhằm đảm bảo tính chính xác và độ tin cậy của các phân tích sau này.



Hình . Hình ảnh thông tin dữ liệu

|  |
| --- |
| data.dropna(inplace=True) |

### Xóa dữ liệu trùng lặp

Thông tin trùng lặp không mang lại giá trị bổ sung cho mô hình mà có thể làm giảm hiệu quả và độ chính xác của nó. Do đó, việc loại bỏ các dữ liệu dư thừa không cần thiết là một bước quan trọng, giúp tối ưu hóa mô hình và cải thiện khả năng học tập, đồng thời giảm thiểu sự phức tạp không cần thiết.

|  |
| --- |
| data.drop\_duplicates(inplace=True) |

## Phân tích dữ liệu

Dữ liệu ban đầu được thu thập với 4097 điểm dữ liệu từ tín hiệu EEG. Sau đó, dữ liệu này được chia thành các khối nhỏ hơn, mỗi khối chứa 178 điểm dữ liệu (tương đương với 1 giây). Mỗi mẫu dữ liệu trong tập chứa 178 đặc trưng EEG, phản ánh hoạt động của não bộ trong khoảng thời gian 1 giây, và một nhãn kết quả thể hiện trạng thái của tín hiệu EEG trong khoảng thời gian đó.

Chính vì thế nên ta sẽ sử dụng tất cả các đặc trưng để huấn luyện mô hình.

## Tiền xử lý dữ liệu

### Chuẩn hóa dữ liệu

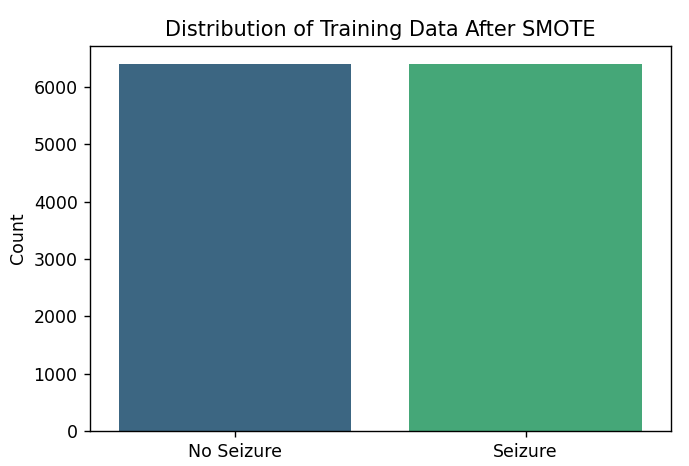
Chuẩn hóa dữ liệu là bước quan trọng trong tiền xử lý dữ liệu, mang lại nhiều lợi ích cho việc phát hiện cơn động kinh từ tín hiệu EEG. Việc này giúp cải thiện tính chính xác của mô hình bằng cách đưa các biến về cùng một thang đo, từ đó giảm thiểu sự thiên lệch và tăng tốc độ hội tụ của các thuật toán tối ưu hóa. Chuẩn hóa cũng làm giảm độ phức tạp của dữ liệu, tăng cường khả năng so sánh giữa các biến và giảm thiểu ảnh hưởng của các biến ngoại lai. Nhờ đó, nó tạo ra một nền tảng vững chắc cho phân tích và diễn giải kết quả, đặc biệt quan trọng trong các ứng dụng nhạy cảm như chẩn đoán cơn động kinh.

|  |
| --- |
| scaler = StandardScaler()  X = scaler.fit\_transform(X) |

### Cân bằng dữ liệu

Nhận thấy rằng dữ liệu có sự phân phối không đồng đều giữa các nhãn kết quả. Để khắc phục vấn đề này, em sẽ sử dụng phương pháp SMOTEđể cân bằng dữ liệu. SMOTE giúp tạo ra các mẫu tổng hợp cho lớp thiểu số, từ đó giảm thiểu sự thiên lệch trong dữ liệu.

|  |
| --- |
| smote = SMOTE(random\_state=42)  X\_train\_smote, y\_train\_smote = smote.fit\_resample(X\_train, y\_train) |



Hình . Giá trị sau khi cân bằng

## Lựa chọn mô hình

Sử dụng các mô hình phân loại:

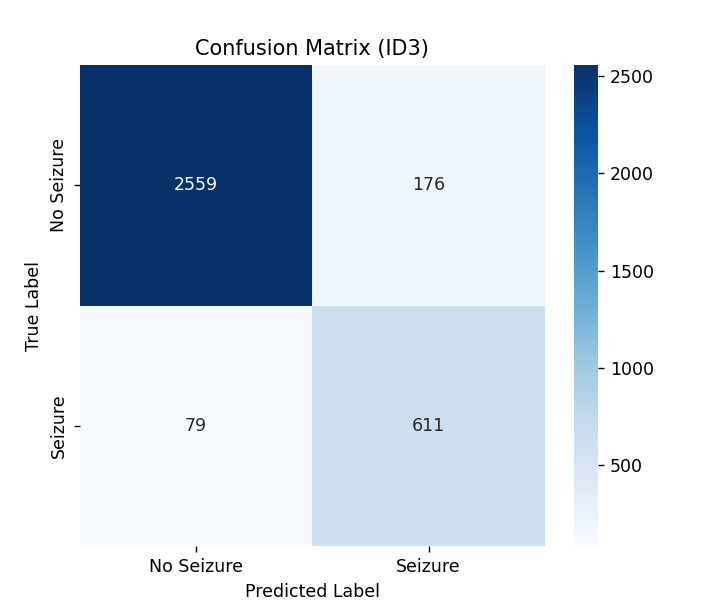
* SVM
* ID3
* ANN

## Huấn luyện mô hình

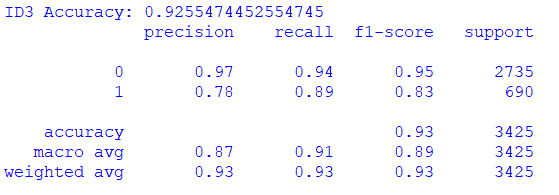
|  |
| --- |
| # 5. Huấn luyện mô hình Iterative Dichotomiser 3  rf = RandomForestClassifier(n\_estimators=100, random\_state=42)  rf.fit(X\_train\_smote, y\_train\_smote)  y\_pred\_rf = rf.predict(X\_test)  y\_pred\_prob\_rf = rf.predict\_proba(X\_test)[:, 1] # Xác suất dự đoán để vẽ ROC  print("Iterative Dichotomiser 3 Accuracy:", accuracy\_score(y\_test, y\_pred\_rf))  print(classification\_report(y\_test, y\_pred\_rf))  # 5. Huấn luyện mô hình SVM  svm = SVC(probability=True, random\_state=42)  svm.fit(X\_train\_smote, y\_train\_smote)  y\_pred\_svm = svm.predict(X\_test)  y\_pred\_prob\_svm = svm.predict\_proba(X\_test)[:, 1] # Xác suất dự đoán để vẽ ROC  print("SVM Accuracy:", accuracy\_score(y\_test, y\_pred\_svm))  print(classification\_report(y\_test, y\_pred\_svm))  # Xây dựng mô hình ANN  model = Sequential([      Dense(128, activation='relu', input\_shape=(X\_train.shape[1],)),      Dropout(0.3),      Dense(64, activation='relu'),      Dropout(0.3),      Dense(1, activation='sigmoid')  # Output layer for binary classification  ])  model.compile(optimizer=Adam(learning\_rate=0.001), loss='binary\_crossentropy', metrics=['accuracy']) |

## Đánh giá mô hình

### Iterative Dichotomiser 3



Hình . Ma trận matrix thuật toánIterative Dichotomiser 3

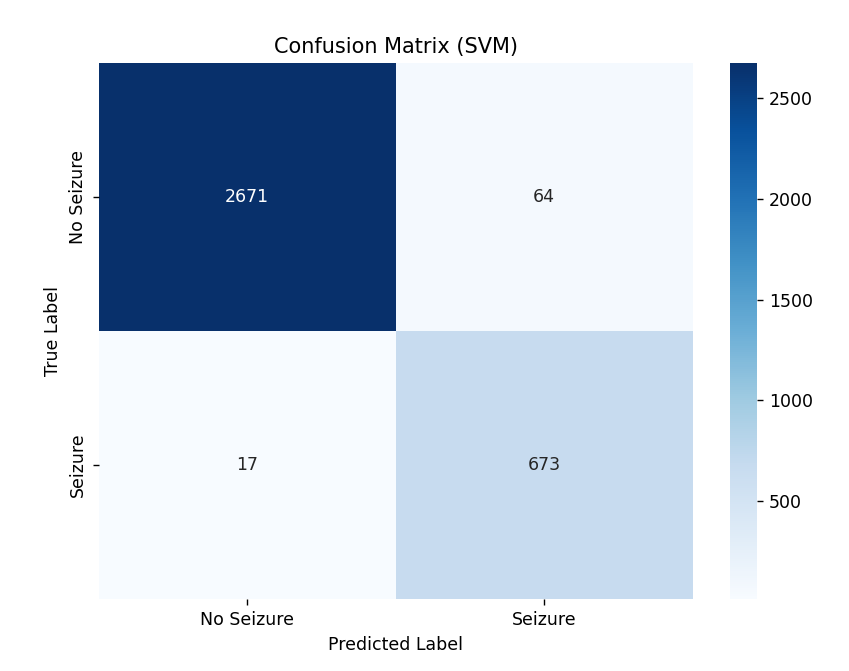


Hình . Kết quả báo cáo phân loại thuật toánIterative Dichotomiser 3

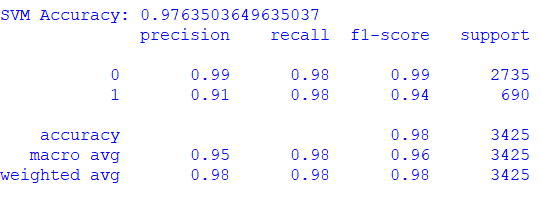
Nhận xét:

* Tỷ lệ dự đoán của mô hình: 92%
* Số lượng dự đoán "Không co giật" đúng là 2599.
* Số lượng dự đoán "Co giật" đúng là 611.
* Số lượng dự đoán "Không co giật" nhưng thực tế là "Co giật" là 176.
* Số lượng dự đoán "Co giật" nhưng thực tế là "Không co giật" là 79.

### SVM



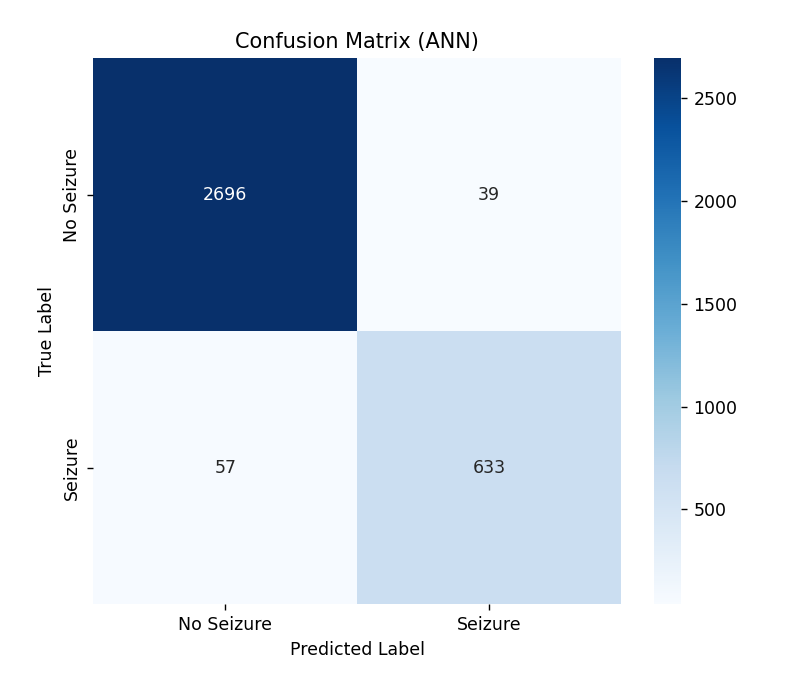
Hình . Ma trận matrix thuật toán SVM



Hình . Kết quả báo cáo phân loại thuật toán SVM

* Nhận xét:
* Tỷ lệ dự đoán của mô hình: 98%
* Số lượng dự đoán "Không co giật" đúng là 2671.
* Số lượng dự đoán "Co giật" đúng là 673.
* Số lượng dự đoán "Không co giật" nhưng thực tế là "Co giật" là 64.
* Số lượng dự đoán "Co giật" nhưng thực tế là "Không co giật" là 17.

### ANN



Hình . Ma trận matrix thuật toán ANN

* Nhận xét
* Tỷ lệ dự đoán mô hình: 97%
* Số lượng dự đoán "Không co giật" đúng là 2696.
* Số lượng dự đoán "Co giật" đúng là 633.
* Số lượng dự đoán "Không co giật" nhưng thực tế là "Co giật" là 39.
* Số lượng dự đoán "Co giật" nhưng thực tế là "Không co giật" là 57.

### Nhận xét chung

Bảng . Bảng dự đoán của 3 mô hình

|  |  |  |
| --- | --- | --- |
| **STT** | **Thuật toán** | **Tỷ lệ dự đoán** |
| **1** | **Iterative Dichotomiser 3** | 92% |
| **2** | **SVM** | 98% |
| **3** | **ANN** | 97% |

Ba mô hình Iterative Dichotomiser 3, SVM, và ANN đều thể hiện hiệu suất tốt trong bài toán nhận diện cơn động kinh từ tín hiệu EEG, với độ chính xác lần lượt là 92%, 98%, và 97%. Mỗi mô hình có ưu điểm riêng: Iterative Dichotomiser 3 xử lý hiệu quả dữ liệu phức tạp và không cần nhiều điều chỉnh, ANN có khả năng học các mối quan hệ phi tuyến tính từ dữ liệu lớn, và SVM vượt trội trong phân loại dữ liệu nhị phân nhờ tìm siêu phẳng phân tách tối ưu. Trong số đó, SVM được xem là lựa chọn tối ưu nhất với độ chính xác cao nhất (98%) và khả năng phân loại hiệu quả trên tập dữ liệu đã chuẩn hóa.

## Triển khai mô hình

### Lưu mô hình

Thực hiện lưu các mô hình để tiến hành đưa vào sử dụng

|  |
| --- |
| with open('rf\_model.pkl', 'wb') as file:  pickle.dump(rf, file) |

Việc lưu mô hình học máy sau khi huấn luyện mang lại nhiều lợi ích quan trọng, bao gồm:

* **Tiết kiệm thời gian:** Quá trình huấn luyện mô hình có thể mất nhiều thời gian, đặc biệt với các mô hình phức tạp và dữ liệu lớn. Lưu mô hình đã huấn luyện giúp không phải huấn luyện lại từ đầu mỗi khi cần sử dụng.
* **Tái sử dụng mô hình:** Mô hình đã được lưu có thể được tái sử dụng trong các ứng dụng thực tế hoặc các dự án tương tự mà không cần xử lý lại dữ liệu hay huấn luyện từ đầu.
* **Triển khai mô hình:** Một mô hình đã lưu có thể được tích hợp vào các ứng dụng thực tế như API, ứng dụng web, hoặc hệ thống dự báo mà không yêu cầu truy cập mã nguồn hoặc dữ liệu huấn luyện ban đầu.
* **Chia sẻ mô hình:** Có thể chia sẻ mô hình đã lưu với mọi người để sử dụng hoặc kiểm tra.
* **Khả năng phục hồi:** Nếu có sự cố xảy ra, chẳng hạn mất dữ liệu hoặc thay đổi môi trường, mô hình đã lưu sẽ đảm bảo rằng bạn vẫn có thể tiếp tục sử dụng nó mà không mất tiến trình trước đó.

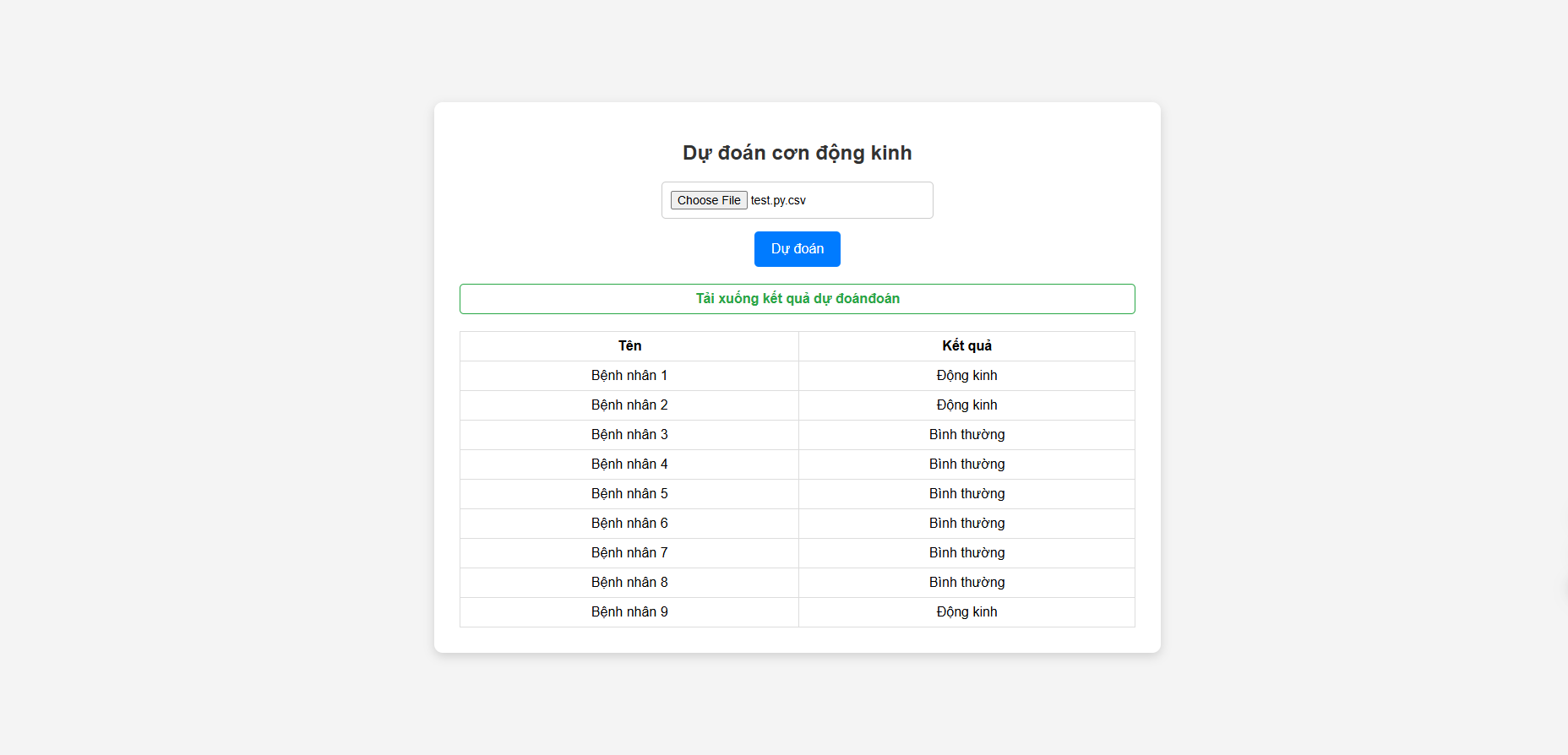
### Triển khai mô hình

Thực hiện xây dựng giao diện người dùng bằng html và flask và triển khai đến người sử dùng. Cụ thể sẽ xây dựng giao diện cho người dùng nhập file csv thông tin điện não đồ để đưa vào dự đoán(Hình 3.10)

Hệ thống sẽ cho phép người dùng tải lên file CSV chứa thông tin về các bệnh nhân. Sau khi tải lên, người dùng chỉ cần nhấn nút dự đoán và hệ thống sẽ tiến hành xử lý dữ liệu và trả về một file kết quả dự đoán. File này sẽ được cung cấp cho người dùng để tải về, giúp họ dễ dàng xem xét và sử dụng các kết quả từ hệ thống.

Code giao diện sẽ có càng hàm chính sau:

* index(): Thực hiện hiển thị trang giao diện người dùng cho ứng dụng dự đoán cơn động kinh
* predict():Thực hiện tiếp nhận dữ liệu do người dùng nhập từ giao diện, sau đó thực hiện các bước tiền xử lý để đảm bảo dữ liệu đầu vào đáp ứng đúng định dạng và yêu cầu của mô hình dự đoán. Sau khi chuẩn bị dữ liệu, chức năng sẽ đưa dữ liệu vào mô hình đã huấn luyện và lưu trữ, thực hiện quá trình dự đoán, rồi xử lý kết quả để trả về thông tin dự báo một cách rõ ràng và dễ hiểu cho người dùng.



Hình . Giao diện người dùng

# KẾT LUẬN

Trong đồ án này, em đã phát triển một hệ thống dự đoán cơn động kinh dựa trên dữ liệu EEG, sử dụng các mô hình học máy khác nhau như Iterative Dichotomiser 3, SVM và ANN để phân loại các tín hiệu EEG thành các lớp tương ứng. Các mô hình này đã được huấn luyện và đánh giá hiệu quả với độ chính xác cao, với SVM đạt hiệu suất tốt nhất. Hệ thống không chỉ mang lại kết quả dự đoán chính xác mà còn cung cấp giao diện người dùng đơn giản, dễ sử dụng, cho phép người dùng tải lên dữ liệu và nhận kết quả dự đoán trong thời gian ngắn. Việc ứng dụng các kỹ thuật học máy vào phân tích dữ liệu y tế sẽ góp phần nâng cao khả năng phát hiện và chẩn đoán bệnh một cách tự động, hỗ trợ các bác sĩ trong công tác điều trị và quản lý bệnh nhân

Hoàn thành các mục tiêu ban đầu đề ra:

* Mục tiêu 1: Tìm hiểu bài toán dự đoán cơn động kinh
* Mục tiêu 2: Tìm hiểu, nghiên cứu một số mô hình học máy
* Mục tiêu 3: Huấn luyện mô hình với bộ dữ liệu
* Mục tiêu 4: Đánh giá mô hình
* Mục tiêu 5: Kết luận

Trong tương lai, hệ thống có thể được cải tiến theo nhiều hướng để nâng cao độ chính xác và tính ứng dụng thực tế.

* Cải thiện mô hình học máy: Tinh chỉnh và thử nghiệm các thuật toán tiên tiến để tối ưu hóa độ chính xác.
* Mở rộng tập dữ liệu: Thu thập thêm dữ liệu từ nhiều nguồn khác nhau để tăng tính tổng quát.
* Tích hợp vào hệ thống giám sát y tế: Kết nối hệ thống với các thiết bị giám sát để cung cấp cảnh báo sớm và hỗ trợ bác sĩ trong việc chăm sóc bệnh nhân.

TÀI LIỆU THAM KHẢO

|  |  |
| --- | --- |
| [1] | https://www.who.int/news-room/fact-sheets/detail/epilepsy |
| [2] | https://www.elcom.com.vn/12-ung-dung-may-hoc-machine-learning-hang-dau-trong-thuc-tien-1693998666 |
| [3] | https://viettelidc.com.vn/tin-tuc/machine-learning-la-gi |
| [4] | https://www.ibm.com/topics/random-forest#Random+forest+applications |
| [5] | https://viblo.asia/p/support-vector-machine-trong-hoc-may-mot-cai-nhin-don-gian-hon-XQZkxoQmewA |
| [6] | https://www.w3schools.com/python/default.asp |
| [7] | https://www.kaggle.com/code/harunshimanto/machine-learning-algorithms-for-epileptic-seizures |
| [8] | https://vjol.info.vn/index.php/TCDHDL/article/view/46662 |

PHỤ LỤC