**TRƯỜNG ĐẠI HỌC CÔNG NGHIỆP HÀ NỘI**

**KHOA CÔNG NGHỆ THÔNG TIN**

**======\*\*\*======**

****

**BÁO CÁO BÀI TẬP LỚN**

**HỌC PHẦN TRÍ TUỆ NHÂN TẠO**

**Đề tài: Nghiên cứu dự đoán khả năng bị đau tim**

| GVHD: | TS Nguyễn Thị Lan Anh |
| --- | --- |
| Nhóm : | Nhóm 15 |
| Thành viên: | Nguyễn Hoàng Minh  Nguyễn Thanh Tùng  Hoàng Công Trung |
|  |  |
|  |  |
|  |  |
|  |  |

Hà Nội, năm 2021

MỤC LỤC

[Chương 1. Tổng quan về trí tuệ nhân tạo](#_30j0zll) **4**

1.1. [Khái niệm về trí tuệ nhân tạo](#_1fob9te) 4

[Hình 1.1. AI là một bộ phận của khoa học máy tính](#_yp2fstvofkp8) 5

[Hình 1.2. 1 cảnh trong bộ phim "I, Robot" nói về một AI đã tiến hóa](#_lznz03a0otca) 6

1.2. [Vai trò của trí tuệ nhân tạo](#_3dy6vkm) 7

[Hình 1.3. Áp dụng AI trong y khoa](#_x79s2vqccisg) 8

[Hình 1.4. Áp dụng AI trong lĩnh vực tài chính](#_msk0x63nbqs7) 9

[Hình 1.5. Áp dụng AI để sản xuất robot hút bụi](#_df5cpt40hfw1) 10

[Hình 1.6. Robot thay thế con người trong một số công việc](#_ratk8jv5vfh0) 13

1.3 [Kĩ thuật trong TTNT là gì và một số kĩ thuật cơ bản trong TTNT](#_1t3h5sf) 14

[1.4. Lịch sử phát triển của trí tuệ nhân tạo](#_qdqbub4ez8ji) 15

[1.5. Các thành phần trong hệ thống của trí tuệ nhân tạo](#_17dp8vu) 17

[1.6. Các lĩnh vực nghiên cứu và ứng dụng cơ bản](#_3rdcrjn) 18

[Chương 2.Cây quyết định](#_26in1rg) **20**

[2.1. Cây quyết định](#_lnxbz9) 20

[Hình 2.1 là cây quyết định dự đoán phá sản](#_4r63yevvafa5) 21

[Ưu/nhược điểm của thuật toán cây quyết định](#_1ksv4uv) 21

[2.2. Tạo cây quyết định](#_44sinio) 21

[2.3. Ví dụ minh hoạ](#_2jxsxqh) 23

[Hình 2.2. Cây quyết định sau lần phân hoạch đầu tiên](#_r6irehlyggn8) 25

[Hình 2.3. Cây quyết định bài toán dự đoán phá sản](#_izi71qjmtu95) 26

[Chương 3. Dự đoán phá sản](#_3j2qqm3) **27**

[3.1. Bài toán](#_1y810tw) 27

[3.1.1. Mô tả](#_4i7ojhp) 27

[3.1.2. Bài toán ứng dụng dự đoán phá sản](#_2xcytpi) 27

[3.2. Cơ sở dữ liệu](#_1ci93xb) 28

[3.3. Kết quả](#_qsh70q) 28

[3.3.1. Chương trình:](#_3as4poj) 28

[3.3.2. Kết quả](#_n7191pog65fj) 31

[Kết luận](#_1j4enhmx0s7h) **32**

[Tài liệu tham khảo](#_2p2csry) **33**

# **Chương 1. Tổng quan về trí tuệ nhân tạo**

## **Khái niệm về trí tuệ nhân tạo**

Theo như cha đẻ của trí tuệ nhân tạo, John McCarthy thì nó là *"*Khoa học và kỹ thuật của việc tạo ra những máy thông minh, đặc biệt là chương trình máy tính thông minh".

Trí tuệ nhân tạo là hướng đi của việc tạo ra máy tính, người máy điều khiển bằng máy tính hay là những phần mềm suy nghĩ thông minh hơn, tương tự như suy nghĩ thông minh của con người.

Trí tuệ nhân tạo được học như bộ não con người, như cách mà con người học, quyết định và làm việc khi giải quyết một vấn đề, và sau đó sử dụng kết quả của quá trình học đó như là nền tảng của việc phát triển phần mềm và hệ thống thông minh.

Ở thời điểm hiện tại, Thuật ngữ này thường dùng để nói đến các MÁY TÍNH có mục đích không nhất định và ngành khoa học nghiên cứu về các lý thuyết và ứng dụng của trí tuệ nhân tạo. Tức là mỗi loại trí tuệ nhân tạo hiện nay đang dừng lại ở mức độ những máy tính hoặc siêu máy tính dùng để xử lý một loại công việc nào đó như điều khiển một ngôi nhà, nghiên cứu nhận diện hình ảnh, xử lý dữ liệu của bệnh nhân để đưa ra phác đồ điều trị, xử lý dữ liệu để tự học hỏi, khả năng trả lời các câu hỏi về chẩn đoán bệnh, trả lời khách hàng về các sản phẩm của một công ty,...



#### Hình 1.1. AI là một bộ phận của khoa học máy tính

Nói nôm na cho dễ hiểu: đó là trí tuệ của máy móc được tạo ra bởi con người**.** Trí tuệ này có thể tư duy, suy nghĩ, học hỏi,... như trí tuệ con người. Xử lý dữ liệu ở mức rộng lớn hơn, quy mô hơn, hệ thống, khoa học và nhanh hơn so với con người.

Rất nhiều hãng công nghệ nổi tiếng có tham vọng tạo ra được những AI (trí tuệ nhân tạo) vì giá trị của chúng là vô cùng lớn, giải quyết được rất nhiều vấn đề của con người mà loài người đang chưa giải quyết được.

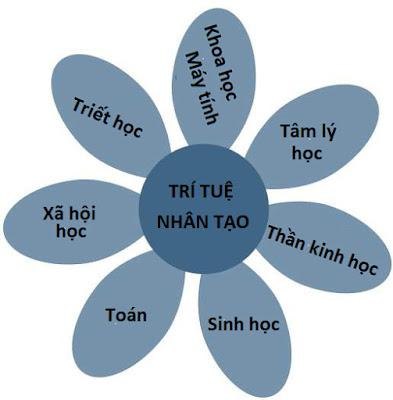
Trí tuệ nhân tạo mang lại rất nhiều giá trị cho cuộc sống loài người, nhưng cũng tiềm ẩn những nguy cơ**.** Rất nhiều chuyên gia lo lắng rằng khi trí tuệ nhân tạo đạt tới 1 ngưỡng tiến hóa nào đó thì đó cũng là thời điểm loài người bị tận diệt. Rất nhiều các bộ phim đã khai thác đề tài này với nhiều góc nhìn, nhưng qua đó đều muốn cảnh báo loài người về mối nguy đặc biệt này.



#### Hình 1.2. 1 cảnh trong bộ phim "I, Robot" nói về một AI đã tiến hóa

Trí tuệ nhân tạo là một ngành khoa học và công nghệ dựa trên nền tảng của Khoa học máy tính, Sinh học, Triết học, Ngôn ngữ học, Toán học và Kỹ thuật. Một chuyên ngành chính của Trí tuệ nhân tạo là phát triển chức năng của máy tính kết hợp với sự thông minh của con người, chẳng hạn như suy luận, học hỏi và giải quyết vấn đề.

Trong những lĩnh vực dưới đây, một hoặc nhiều lĩnh vực có thể góp thành để xây dựng hệ thống thông minh.



## **Vai trò của trí tuệ nhân tạo**

Vai trò của AI là vô tận đối với cuộc sống của chúng ta. AI có thể tiếp cận với con người thông qua nhiều lĩnh vực, ngành nghề khác nhau. Ưu điểm của trí tuệ nhân tạo AI là khả năng xử lý dữ liệu khoa học hơn, nhanh hơn, hệ thống hơn so với con người. Việc phát triển và đưa các sản phẩm AI tới tay người dùng đúng cách sẽ thúc đẩy mạnh mẽ sự phát triển của toàn nhân loại. Mở ra một thế giới hoàn toàn mới cùng các giải pháp bù đắp cho những vấn đề mà con người không thể giải quyết.

### ***Vai trò của trí tuệ nhân tạo trong y học***

**

#### Hình 1.3. Áp dụng AI trong y khoa

Công nghệ AI đã mở ra một trang mới cho nền y học thế giới, đặc biệt là nền y học nước nhà. Nó mang đến cho con người những giá trị đáng kinh ngạc trong việc bảo vệ sức khỏe và điều trị bệnh tật. Tại lĩnh vực này, trí tuệ nhân tạo có vai trò quan trọng trong việc hỗ trợ điều trị y tế như định lượng thuốc, các phương pháp điều trị khác nhau cho bệnh nhân và quy trình phẫu thuật trong phòng mổ. Chúng sử dụng những thuật toán phân tích để hỗ trợ bệnh nhân theo dõi kết quả điều trị 24/7.

### ***Vai trò của AI trong tài chính***

**

#### Hình 1.4. Áp dụng AI trong lĩnh vực tài chính

Ngoài việc hỗ trợ con người chăm sóc sức khỏe, AI còn có vai trò quan trọng trong ngành tài chính ngân hàng. AI là công cụ giúp con người xử lý các hoạt động trong ngân hàng như xử lý giao dịch, theo dõi số dư, quản lý tài sản và các tài khoản tiền gửi lớn một cách nhanh chóng và chính xác nhất. Trí tuệ nhân tạo không những giúp các ngân hàng hợp lý hóa giao dịch mà còn có thể ước tính cung, cầu và định giá chứng khoán một cách dễ dàng hơn.

### ***Vai trò của AI trong trò chơi và công nghệ***

Hiện nay, những tập đoàn lớn đang ngày càng thúc đẩy việc sử dụng máy móc thông minh vào dây chuyền sản xuất. AI được sử dụng như các robot có thể thay thế một phần công việc của con người. Khối lượng công việc và thời gian hoàn thành sẽ nhanh chóng và nhẹ nhàng hơn dưới sự hoạt động của máy móc tích hợp trí tuệ nhân tạo. Tiêu biểu là với các sản phẩm như ô tô tự lái và trò chơi điện tử. Trong trò chơi điện tử, trí tuệ nhân tạo AI sẽ tự phân tích các hành vi và đưa ra những đáp án không kém cạnh với trí tuệ con người. Với ô tô tự lái, hệ thống AI tính toán tất cả các dữ liệu bên trong động cơ, tìm hiểu cách đi và ngăn chặn va chạm bởi chướng ngại vật

## ***Sự kết hợp hoàn hảo của AI và robot hút bụi***

****

#### Hình 1.5. Áp dụng AI để sản xuất robot hút bụi

Khi mọi người nghe đến trí tuệ nhân tạo, điều đầu tiên họ thường nghĩ đến là robot. Đối với lĩnh vực dọn dẹp tự động hóa gia đình, AI là điều không thể thiếu. Kết hợp các công nghệ tiên tiến cùng công nghệ AI siêu thông minh, các dòng máy robot hút bụi tự động liên tục được ra mắt trên thị trường. Tiêu biểu là dòng robot hút bụi Roomba của iRobot. Các sản phẩm tích hợp AI thường là những công cụ cao cấp nhất, đem lại hiệu quả cực lớn trong việc làm sạch sàn nhà của các hộ gia đình.

Với thời đại công nghệ 4.0 hiện nay, việc ứng dụng AI không còn xa lạ gì với cuộc sống của chúng ta. Trí tuệ nhân tạo có mặt trong mọi lĩnh vực đời sống từ giải trí cho đến y tế, xã hội. Đây chính là chìa khóa để mở ra một thế hệ mới đầy văn minh, thúc đẩy sự phát triển to lớn của loài người.

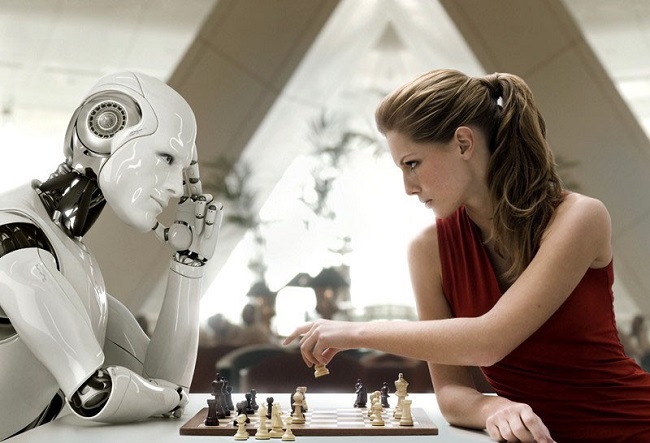
**So sánh giữa lập trình không có TTNT và lập trình có TTNT**

| Lập trình không có TTNT | Lập trình có TTNT |
| --- | --- |
| Chương trình máy tính mà không có Trí tuệ nhân tạo thì chỉ có thể trả lời những câu hỏi xác định được quy định sẵn để giải quyết vấn đề. | Chương trình máy tính mà có Trí tuệ nhân tạo thì có thể trả lời những câu hỏi chung, cùng loại để giải quyết vấn đề. |
| Chỉnh sửa chương trình dẫn đến thay đổi trong cấu trúc của nó. | Chương trình TTNT có thể tiếp thu sự cập nhật cái mới bằng cách đề cao tính độc lập của những thông tin với nhau. Vì vậy bạn có thể sửa đổi một phần thông tin trong chương trình mà không làm ảnh hưởng đến cấu trúc của nó. |
| Việc chỉnh sửa thường không nhanh và không dễ dàng. Nó có thể dẫn đến việc ảnh hưởng chương trình của bạn. | Chỉnh sửa chương trình nhanh và dễ dàng. |

**Những tác động của TTNT đến sản xuất trong nền công nghiệp 4.0 như sau:**

* Chất lượng – Năng suất dự đoán : Vai trò của trí tuệ nhân tạo đầu tiên là giảm thiểu các hao tổn trong sản xuất và ngăn ngừa các quy trình sản xuất kém hiệu quả. Khi nhu cầu ngày càng tăng để đáp ứng sự cạnh tranh thì trí tuệ nhân tạo là điều vô cùng cần thiết.
* Bảo trì sự dự đoán : Một trong những lợi ích của trí tuệ nhân tạo nữa là bảo trì dự đoán. Thay vì việc bảo trì theo lịch trình định trước thì bảo trì dự đoán sẽ sử dụng thuật toán để dự đoán lỗi tiếp theo của một bộ phận/máy móc/hệ thống. Nhờ đó có thể cảnh báo nhân viên thực hiện các quy trình bảo trì tập trung để ngăn chặn sự cố. Bảo trì dự đoán có ưu điểm là giảm đáng kể chi phí trong khi loại bỏ nhu cầu về thời gian ngừng hoạt động theo kế hoạch trong nhiều trường hợp. Ngoài ra, nhờ nó mà Tuổi thọ hữu dụng còn lại của máy móc và thiết bị lâu hơn.

### Kết hợp giữa robot và con người

****

#### Hình 1.6. Robot thay thế con người trong một số công việc

Tính đến năm 2020, ước tính có khoảng 1,64 triệu robot công nghiệp đang hoạt động trên toàn thế giới. Robot sản xuất được chấp thuận làm việc cùng với con người để tăng năng suất công việc.

Khi áp dụng robot ngày càng nhiều thì AI sẽ đóng một vai trò quan trọng trong việc đảm bảo an toàn cho con người. Đồng thời trao cho robot nhiều trách nhiệm hơn trong việc đưa ra các quyết định có thể tối ưu hóa các quy trình dựa trên dữ liệu thời gian thực được thu thập từ sàn sản xuất.

* Thiết kế sáng tạo : Nhà sản xuất có thể tận dụng trí tuệ nhân tạo vào giai đoạn thiết kế. Khi có bản tóm tắt thiết kế được xác định rõ rành làm đầu vào thì các nhà kỹ sư, thiết kế có thể sử dụng thuật toán al. Mục đích để khám phá tất cả các cấu hình có thể có của một giải pháp.
* Nhu cầu cung ứng thị trường : Vai trò của trí tuệ nhân tạo cuối cùng mà chúng tôi muốn nhắc đến là cung ứng thị trường. Hiện nay trí tuệ nhân tạo đang hiện hữu ở mọi nơi trong hệ sinh thái công nghiệp 4.0. Nhà sản xuất có thể sử dụng các thuật toán AI để tối ưu hóa chuỗi cung ứng của các hoạt động sản xuất. Đồng thời giúp họ phản ứng và dự đoán tốt hơn những thay đổi trên thị trường.

## **Kĩ thuật trong TTNT là gì và một số kĩ thuật cơ bản trong TTNT**

Trong thế giới thực, Tri thức có một vài thuộc tính như sau:

* Dung lượng đồ sộ, phi thường.
* Tổ chức tốt, định dạng tốt.
* Luôn luôn cập nhật sự thay đổi.

**Kỹ thuật Trí tuệ nhân tạo** là một cách để tổ chức và sử dụng tri thức có hiệu quả trong những cách sau đây:

* Có thể nhận thức được người đã cung cấp cho nó.
* Có thể sửa đổi dễ dàng để sửa lỗi.
* Nó có thể hữu ích trong một số tình huống dù nó chưa hoàn thiện hoặc chưa chính xác lắm.

Kỹ thuật Trí tuệ nhân tạo nâng cao tốc độ thực thi của những chương trình phức tạp.

**Một số kỹ thuật Trí tuệ nhân tạo cơ bản :**

- Lý thuyết giải bài toán và suy diễn thông minh

**-** Lý thuyết tìm kiếm may rủi

**-** Các ngôn ngữ về TTNT

**-** Lý thuyết thể hiện tri thức và hệ chuyên gia

**-** Lý thuyết nhận dạng và xử lý tiếng nói

- Người máy

- Tâm lý học xử lý thông tin

- Xử lý danh sách, kỹ thuật đệ quy, kỹ thuật quay lui và xử lý cú pháp hình thức

## **1.4. Lịch sử phát triển của trí tuệ nhân tạo**

Đây là lịch sử của Trí tuệ nhân tạo trong suốt thế kỷ XX.

| Năm | Cột mốc / Phát minh |
| --- | --- |
| 1923 | Vở kịch khoa học viễn tưởng của Karel Capek tên là "Rossum's Universal Robots" (RUR) diễn ra tại Luân Đôn (nước Anh). Lần đầu tiên sử dụng từ "robot" trong tiếng Anh. |
| 1943 | Nền tảng của mạng thần kinh được đặt nền móng. |
| 1945 | Isaac Asimov, một cựu sinh viên trường Đại học Columbia, đưa ra thuật ngữ "Robotics" |
| 1950 | Alan Turing giới thiệu Bài kiểm tra Turing để đánh giá sự thông minh và công bố Máy thông minh và Sự thông minh. Claude Shannon công bố "Phân tích chi tiết của việc chơi cờ". |
| 1956 | John McCarthy đưa ra thuật ngữ Trí tuệ nhân tạo. Biểu diễn chạy chương trình trí tuệ nhân tạo đầu tiên tại trường Đại học Carnegie Mellon. |
| 1958 | John McCarthy sáng tạo ra LISP, ngôn ngữ lập trình cho trí tuệ nhân tạo. |
| 1964 | Bài luận văn của Danny Bobrow tại MIT cho thấy máy tính có thể hiểu được ngôn ngữ tự nhiên của con người. |
| 1965 | Joseph Weizenbaum tại MIT đã xây dựng ELIZA, một vấn đề tương tác được mang trong đoạn đối thoại Tiếng Anh. |
| 1969 | Cá nhà khoa học tại Viện nghiên cứu Stanford đã phát triển Shakey, một robot, được trang bị sự vận động, nhận thức, và giải quyết vấn đề. |
| 1973 | Các nhóm hội về người máy tại Đại học Edinburgh đã xây dựng Freddy. Một người máy Scotland nổi tiếng, có khả năng sử dụng thị giác để định vị và lắp ráp mô hình. |
| 1979 | Xe tự quản được điều khiển bằng máy tính đầu tiên được xây dựng. Đó là Stanford Cart. |
| 1985 | Harold Cohen tạo và trình diễn chương trình đồ họa mang tên Aaron. |
| 1990 | Những chuyên đề nâng cao trong tất cả các lĩnh vực của Trí tuệ nhân tạo là:   * Có tính chất quan trọng trong "học máy". * Suy luận theo tình huống * Lên lịch trình * Khai thác dữ liệu, thu thập web * Hiểu và dịch ngôn ngữ tự nhiên của con người * Thị giác và thực tế ảo * Ứng dụng trong trò chơi |
| 1997 | Chương trình "Deep Blue Chess" đánh bại nhà vô địch cờ thế giới, Garry Kasparov. |
| 2000 | Những robot thú cưng có sự tương tác đã được thương mại hóa. MIT đã trình diễn *Kismet* - một robot có khuôn mặt có thể biểu lộ cảm xúc. robot *Nomad* khám phá những vùng xa xôi hẻo lánh của Nam Cực và xác định thiên thạch. |

## **1.5. Các thành phần trong hệ thống của trí tuệ nhân tạo**

Hệ thống trí tuệ nhân tạo bao gồm hai thành phần cơ bản đó là biểu diễn tri thức và tìm kiếm tri thức trong miền biểu diễn:

**TTNT = Tri thức + Suy diễn**

Tri thức của bài toán có thể được phân ra làm ba loại cơ bản đó là tri thức mô tả, tri thức thủ tục và tri thức điều khiển.

Để biểu diễn tri thức người ta sử dụng các phương pháp sau đây:

* Phương pháp biểu diễn nhờ luât
* Phương pháp biểu diễn nhờ mạng ngữ nghĩa
* Phương pháp biểu diễn nhờ bộ ba liên hợp OAV
* Phương pháp biểu diễn nhờ Frame
* Phương pháp biểu diễn nhờ logic vị tư

Sau khi tri thức của bài toán đã được biểu diễn, kỹ thuật trong lĩnh vực trí tuệ nhân tạo là các phương pháp tìm kiếm trong miền đặc trưng tri thức về bài toán đó. Với mỗi cách biểu diễn sẽ có các giải pháp tương ứng. Các vấn đề này sẽ được đề cập trong chương 3.

## **1.6. Các lĩnh vực nghiên cứu và ứng dụng cơ bản**

Trí tuệ nhân tạo có những ảnh hưởng vượt trội trong nhiều lĩnh vực như:

* Game - Trí tuệ nhân tạo đóng vai trò cốt yếu trong những game chiến lược như cờ, đánh bài, tic-tac-toe (như cờ caro), ... nơi mà máy móc có thể suy nghĩ số lớn những trường hợp có khả năng xảy ra dựa trên tri thức.
* Xử lý ngôn ngữ tự nhiên - Nó có khả năng tương tác với máy tính, hiểu ngôn ngữ tự nhiên mà con người nói.
* Hệ thống chuyên môn hóa - Có một vài ứng dụng mà các máy móc thông minh, phần mềm và những thông tin đặc biệt để suy luận. Nó giải thích và đưa ra lời khuyên cho người dùng hệ thống đó.
* Hệ thống thị giác - Hệ thống có thể hiểu, phân tích và tiếp thu dữ liệu vào thuộc về thị giác ngay trên máy tính. Ví dụ như:
  + Những máy bay do thám chụp lại hình ảnh, sau đó sử dụng kỹ thuật này để mô hình hóa những thông tin không gian hay bản đồ của khu vực.
  + Bác sỹ sử dụng hệ thống buồng bệnh chuyên môn để chẩn đoán cho bệnh nhân.
  + Cảnh sát có thể sử dụng phần mềm máy tính để nhận diện khuôn mặt của tội phạm từ những hình chân dung được vẽ lại bởi những họa sĩ pháp y.
* Nhận diện lời nói - Một vài hệ thống thông minh có khả năng nghe và tiếp thu ngôn ngữ trong cấu trúc và nghĩa của câu trong khi con người nói. Nó có thể nắm bắt được độ nhấn mạnh khác nhau, từ lóng, tiếng ồn phía sau, sự thay đổi trong âm thanh của con người do trời lạnh, ...
* Nhận diện chữ viết tay - Phần mềm nhận diện chữ viết tay đọc văn bản được viết trên giấy bằng bút hoặc viết trên màn hình bằng bút cảm ứng. Nó nhận dạng được hình dạng của chữ và chuyển nó thành văn bản có thể chỉnh sửa được.
* Người máy thông minh - Người máy có khả năng thực hiện nhiệm vụ mà con người giao cho. Nó có các cảm biến để nhận dạng các dữ liệu vật lý trong thế giới thực như ánh sáng, hơi nóng, nhiệt độ, sự di chuyển, âm thanh, sự va chạm và áp lực. Nó được trang bị bộ xử lý hiệu quả, đa cảm biến và bộ nhớ lớn để thể hiện sự thông minh. Hơn thế nữa, nó có khả năng học từ lỗi sai của nó và thích nghi với môi trường mới.

# **Chương 2.Cây quyết định**

## **2.1. Cây quyết định**

Cây quyết định được dùng để đưa ra tập luật if – then nhằm mục đích dự báo, giúp con người nhận biết về tập dữ liệu. Cây quyết định cho phép phân loại đối tượng tùy thuộc vào các điều kiện tại các nút trong cây, bắt đầu từ gốc cây tới các nút sát lá-Nút xác định phân loại đối tượng. Mỗi nút trong của cây xác định điều kiện đối với thuộc tính mô tả của đối tượng. Mỗi nhánh tương ứng với điều kiện: Nút (thuộc tính) bằng giá trị nào đó. Đối tượng được phân loại nhờ tích hợp các điều kiện bắt đầu từ nút gốc của cây và các thuộc tính mô tả với giá trị của thuộc tính đối tượng.

#### 

### **Ưu/nhược điểm của thuật toán cây quyết định**

**Ưu điểm**

Cây quyết định là một thuật toán đơn giản và phổ biến. Thuật toán này được sử dụng rộng rãi bởi những lợi ích của nó:

* Mô hình sinh ra các quy tắc dễ hiểu cho người đọc, tạo ra bộ luật với mỗi nhánh lá là một luật của cây.
* Dữ liệu đầu vào có thể là là dữ liệu missing, không cần chuẩn hóa hoặc tạo biến giả
* Có thể làm việc với cả dữ liệu số và dữ liệu phân loại
* Có thể xác thực mô hình bằng cách sử dụng các kiểm tra thống kê
* Có khả năng là việc với dữ liệu lớn

**Nhược điểm**

Kèm với đó, cây quyết định cũng có những nhược điểm cụ thể:

* Mô hình cây quyết định phụ thuộc rất lớn vào dữ liệu của bạn. Thậm chí, với một sự thay đổi nhỏ trong bộ dữ liệu, cấu trúc mô hình cây quyết định có thể thay đổi hoàn toàn.
* Cây quyết định hay gặp vấn đề [overfitting](https://trituenhantao.io/tu-dien-thuat-ngu/overfitting/).

## **2.2. Tạo cây quyết định**

Xét bảng dữ liệu T = (A, D) trong đó A = {A1, A2,..., An } là tập thuộc tính dẫn xuất, D = {r1, r2, ..., rn} là thuộc tính mục tiêu. Vấn đề đặt ra là trong tập thuộc tính A ta phải chọn thuộc tính nào để phân hoạch? Một trong các phương pháp đó là dựa vào độ lợi thông tin. Hay còn gọi là thuật giải ID3.

Lựa chọn chủ yếu trong giải thuật ID3 là chọn thuộc tính nào để đưa vào mỗi nút trong cây. Ta sẽ chọn thuộc tính phân rã tập mẫu tốt nhất. Thước đo độ tốt của việc chọn lựa thuộc tính là gì? Ta cần xác định một độ đo thống kê, gọi là thông tin thu được, đánh giá từng thuộc tính được chọn tốt như thế nào còn phụ thuộc vào việc phân loại mục tiêu của tập mẫu. ID3 sử dụng thông tin thu được đánh giá để chọn ra thuộc tính cho mỗi bước giữa những thuộc tính ứng viên, trong quá trình phát triển cây.

****

Để đánh giá chính xác thông tin thu được, dùng Entropy(S): Độ bất định (độ pha trộn/độ hỗn tạp) của S liên quan đến sự phân loại đang xét

Trong đó là xác suất xuất hiện trạng thái i của hệ thống. Theo lý thuyết thông tin: mã có độ dài tối ưu là mã gán bits cho thông điệp có xác suất là p. S là một tập huấn luyện.

Nếu gọi là xác suất xuất hiện các ví dụ dương trong tập S, là xác suất xuất hiện các ví dụ âm trong tập S. Entropy đo độ bất định của tập S sẽ là:

Entropy (S) = − −

Quy định 0.log 0 = 0

Chẳng hạn với tập S gồm 14 mẫu có chung một vài giá trị logic gồm 9 mẫu dương và 5 mẫu âm. Khi đó đại lượng Entropy của tập S liên quan đến sự phân loại logic này là:

Entropy([9+, 5-]) = - (9/14) (9/14) - (5/14) (5/14) = 0,940

**Chú ý :**

Đại lượng Entropy = 0 nếu tất cả thành viên của tập S cùng thuộc một lớp (vì nếu tất cả là dương (P+ = 1), do đó P- = 0, Entropy(S) = −11− 00 = 0 ).

Đại lượng Entropy(S) = 1 khi tập S chứa tỉ lệ tập mẫu âm và mẫu dương là như nhau. Nếu tập S chứa tập mẫu âm và tập mẫu dương có tỉ lệ P+ khác P- thì Entropy(S) ∈ (0,1).

Dựa trên sự xác định entropy, ta tính Gain(S, A) = Lượng giảm entropy mong đợi qua việc chia các ví dụ theo thuộc tính A

Gain(S, A) = Entropy(S) - Entropy(

## **2.3. Ví dụ minh hoạ**

Xem xét nhiệm vụ học được đưa ra bởi tập mẫu dưới đây , thuộc tính mục tiêu ở đây là: class có giá trị là N hoặc NB, giá trị thuộc tính này dự đoán dựa vào các thuộc tính mô tả.

Kí hiệu: P=Positive,A-Average,N-negative,B-Bankruptcy,NB-Non-Bankruptcy

Financial\_Flexibility : khả năng linh hoạt tài chính

Credibility : Sự uy tín

Competitiveness : khả năng cạnh tranh

Operating\_Risk : Rủi ro hoạt động

Bảng tập mẫu:

| Object | Financial\_Flexibility | Credibility | Competitiveness | Operating\_Risk | Class |
| --- | --- | --- | --- | --- | --- |
| D1 | N | P | P | P | NB |
| D2 | P | P | P | N | B |
| D3 | P | P | P | N | B |
| D4 | A | P | P | P | NB |
| D5 | P | A | N | P | B |
| D6 | A | N | P | N | NB |
| D7 | A | P | N | P | NB |
| D8 | N | N | A | A | B |
| D9 | N | N | A | N | B |
| D10 | N | A | N | N | NB |
| D11 | A | A | P | N | B |
| D12 | N | N | N | N | NB |
| D13 | P | A | N | A | B |
| D14 | N | N | A | P | NB |

Bước 1: Tạo nút đỉnh cho cây quyết định:

Giá trị thông tin thu được cho mỗi thuộc tính:

Entropy(S)=1

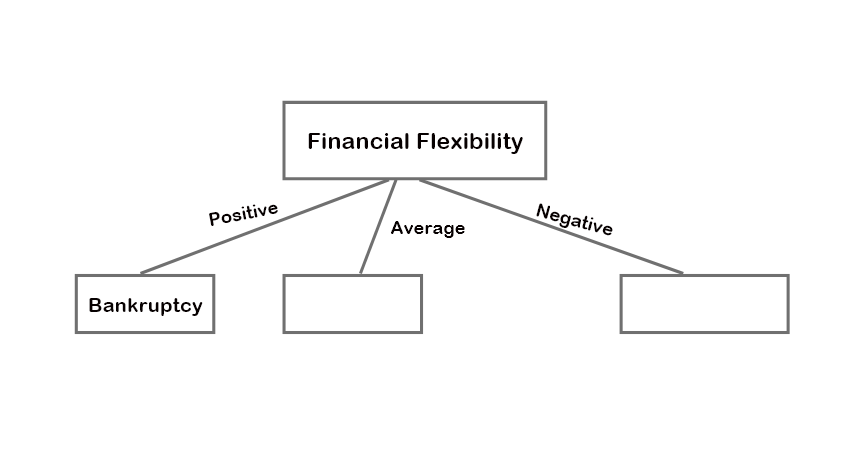
Gain(S, Financial Flexibility)=0.401

Gain(S, Credibility)=0.044

Gain(S, Competitiveness)=0.028

Gain(S, Operating Risk)=0.25

Theo đánh giá thông tin thu được, thuộc tính Financial Flexibility cung cấp dự đoán tốt nhất về thuộc tính mục tiêu “Class” trên tập mẫu. Do đó, thuộc tính “Financial Flexibility” được chọn là thuộc tính quyết định cho nút gốc, nhánh được tạo ra dưới nút gốc tương ứng với mỗi giá trị của thuộc tính như Credibility, Competitiveness, Operating Risk cùng với tập mẫu sẽ thêm vào mỗi nút con mới.



#### Hình 2.2. Cây quyết định sau lần phân hoạch đầu tiên

Bước 2: Xét nhánh Average với các thuộc tính còn lại trong bảng.

Entropy(S Average)= 0.811

Gain (S Average, Credibility)=0.811

Gain(S Average, Competitiveness)=0.123

Gain(S Average, Operating Risk)=0.311

Theo đánh giá thông tin thu được, thuộc tính Credibility cung cấp dự đoán tốt nhất.

Chọn Credibility là nút tiếp sau Average.

Bước 3: Xét nhánh Negative với các thuộc tính còn lại trong bảng.

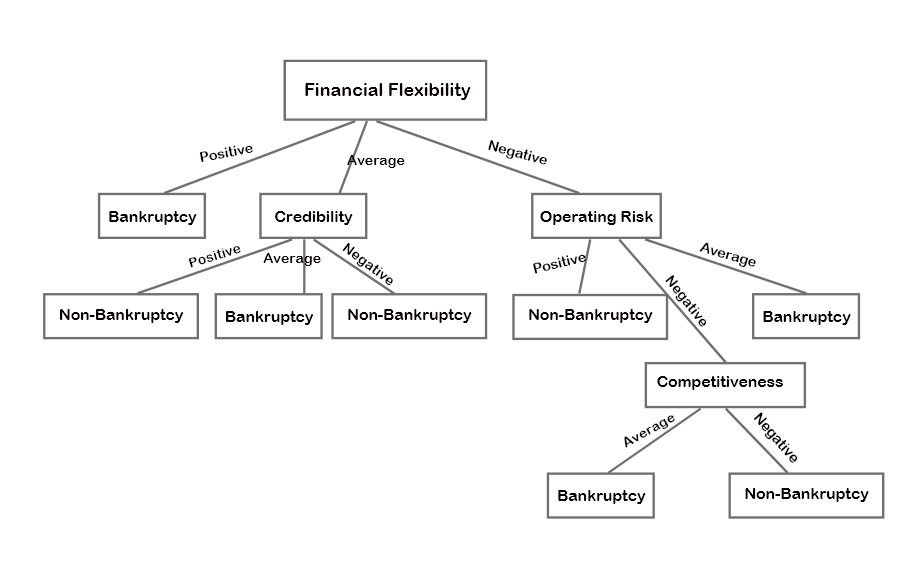
Entropy(S Negative)= 0.918

Gain (S Negative, Credibility)=0.251

Gain(S Negative, Competitiveness)=0.251

Gain(S Negative, Operating Risk)=0.459

Theo đánh giá thông tin thu được, thuộc tính Operating Risk cung cấp dự đoán tốt nhất. Chọn Operating Risk là nút tiếp sau Negative.



#### Hình 2.3. Cây quyết định bài toán dự đoán phá sản

# **Chương 3: Hồi quy Logistic ( Logistic Regression)**

# **3.1. Khái niệm**

# Logistic Regression là 1 thuật toán phân loại được dùng để gán các đối tượng cho 1 tập hợp giá trị rời rạc (như 0, 1, 2, ...). Một ví dụ điển hình là phân loại Email, gồm có email công việc, email gia đình, email spam, ... Giao dịch trực tuyến có là an toàn hay không an toàn, khối u lành tính hay ác tình.

# Thuật toán trên dùng hàm sigmoid logistic để đưa ra đánh giá theo xác suất. Ví dụ: Khối u này 80% là lành tính, giao dịch này 90% là gian lận, ...

3.2. Ưu điểm

Hồi quy Logistic là một thuật toán đơn giản và phổ biến. Thuật toán này được sử dụng rộng rãi bởi những ưu điểm của nó:

* Hồi quy logistic không yêu cầu bất kỳ dạng phân phối cụ thể nào của các biến độc lập và các vấn đề như phương sai thay đổi.
* Các mô hình hồi quy logistic có thể xử lý khối lượng lớn dữ liệu ở tốc độ cao
* Hồi quy logistic không yêu cầu mối quan hệ tuyến tính giữa các biến độc lập và các biến phụ thuộc
* Có thể giải quyết các hiệu ứng phi tuyến tính ngay cả khi các số hạng mũ và đa thức không được thêm vào một cách rõ ràng dưới dạng các biến độc lập bổ sung
* Có khả năng là việc với dữ liệu lớn
* Không yêu cầu dữ liệu phải *linearly separable*.

3.3. Nhược điểm

Bên cạnh những ưu điểm, hồi quy Logistic cũng có những nhược điểm cụ thể:

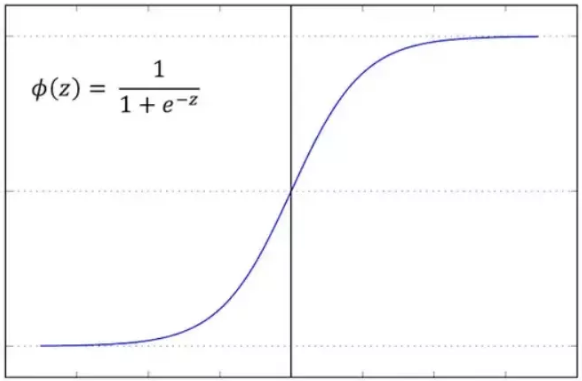
* Hồi quy logistic không dự đoán được kết quả liên tục.
* Hồi quy logistic giả định tính tuyến tính giữa biến dự đoán (phụ thuộc) và biến dự báo (độc lập).
* Hồi quy logistic có thể không chính xác nếu kích thước mẫu quá nhỏ.

3.4. Thiết lập thuật toán

3.4.1. Hàm sigmoid

Yêu cầu dự đoán của bài toán là đưa ra kết quả giá trị trong đoạn [0, 1]. Bởi vậy, chúng ta cần một hàm số liên tục nhận giá trị thực, bị chặn trong khoảng (0,1). Nếu coi điểm có tung độ là 1/2 làm điểm phân chia thì các điểm càng xa điểm này về phía bên trái có giá trị càng gần 0. Ngược lại, các điểm càng xa điểm này về phía phải có giá trị càng gần 1.Trong số các hàm số hội tụ đủ các yêu cầu nêu trên, hàm *sigmoid là hàm số được chọn nhiều nhất.*

**



3.4.2. Model

Thuật toán hồi qui Logistic nói riêng, và các thuật toán Machine Learning/ Deep Learning nói chung. Đầu tiền, chúng ta bắt đầu xây dựng model cho bài toán.



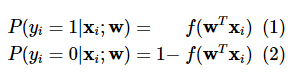
# **P(y=1|x;w) chính là xác suất có điều kiện để xảy ra sự kiện y=1 tương ứng với đầu vào x, và trọng số w.**

3.4.3 Hàm mất mát( Loss function)

Hàm số L(w) được gọi là **hàm mất mát** (loss function) của bài toán Linear Regression. Chúng ta luôn mong muốn rằng sự mất mát (sai số) là nhỏ nhất, điều đó đồng nghĩa với việc tìm vector hệ số w sao cho giá trị của hàm mất mát này càng nhỏ càng tốt. Giá trị của w làm cho hàm mất mát đạt giá trị nhỏ nhất được gọi là *điểm tối ưu* (optimal point).

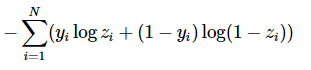
Với mỗi điểm trong tập dữ liệu, gọi hàm loss function, loss function này có tên gọi là binary\_crossentropy.

Có thể giả sử rằng xác suất để một điểm dữ liệu x rơi vào class 1 là f(wTx) và rơi vào class 0 là 1−f(wTx). Với mô hình được giả sử như vậy, với các điểm dữ liệu training (đã biết đầu ra y), ta có thể viết như sau:



Mục đích của chúng ta là tìm các hệ số w sao cho f(wTxi) càng gần với 1 càng tốt với các điểm dữ liệu thuộc class 1 và càng gần với 0 càng tốt với những điểm thuộc class 0.

Một phương pháp thường được sử dụng đó là lấy logarit tự nhiên (cơ số e) của *likelihood function* biến phép nhân thành phép cộng và để tránh việc số quá nhỏ. Sau đó lấy ngược dấu để được một hàm và coi nó là hàm mất mát. Lúc này bài toán tìm giá trị lớn nhất (maximum likelihood) trở thành bài toán tìm giá trị nhỏ nhất của hàm mất mát (hàm này còn được gọi là negative log likelihood):

. 

3.4.4. Tối ưu hàm mất mát

Với mỗi điểm (,), gọi hàm mất mát

trong đólà giá trị mà model dự đoán, còn yi là giá trị thật của dữ liệu.

Để model nâng cao hiệu quả dự đoán, chúng ta cần tối ưu hàm mất mát. Mục tiêu là hàm mất mát nhỏ nhất, điều đó đồng nghĩa rằng phải tìm bộ weight (W) sao cho giá trị của hàm mất mát với bộ W càng nhỏ càng tốt.

Có nhiều thuật toán giúp tối ưu hàm mất mát, nhưng ở đây chúng ta sử dụng thuật toán Gradient Descent.

Gradient descent là thuật toán tìm giá trị nhỏ nhất của hàm số f(x) dựa trên đạo hàm. Đây là một trong những thuật toán đơn giản nhưng chỉ sử dụng được cho những hàm mất mát liên tục và xác định trên khoảng được xét.

Thuật toán:

1. Khởi tạo giá trị x = x0 ban đầu tùy ý

2. Gán x = x - learning\_rate \* f’(x) ( learning\_rate là hằng số dương ví dụ learning\_rate = 0.001)

3. Tính lại f(x):

Nếu f(x) đủ nhỏ hoặc đạt số lần lặp bước 2 theo yêu cầu thì dừng lại.

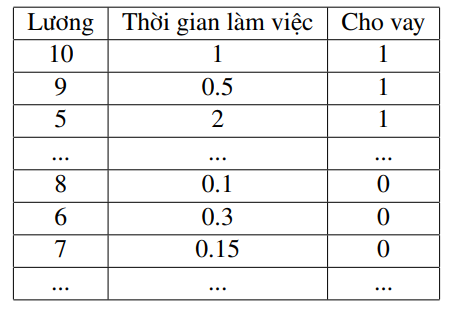
Ngược lại tiếp tục bước 2 .

Đến đây, sau khi thực hiện tối ưu hàm mất mát, chúng ta có được bộ trọng số tối ưu để thực hiện các nghiên cứu dự đoán theo yêu cầu.

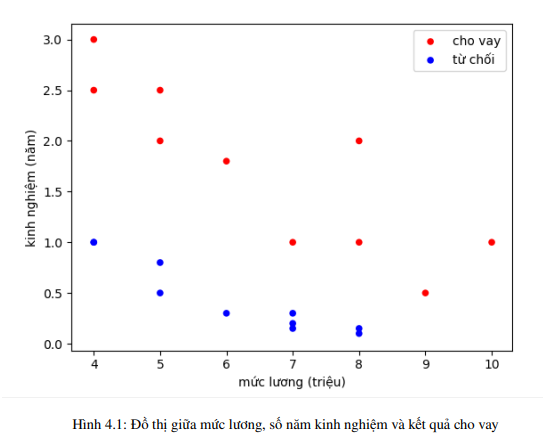
3.4.5. Bài toán minh họa

3.4.5.1. Bài toán

Ngân hàng bạn đang làm có chương trình cho vay ưu đãi cho các đối tượng mua chung cư. Tuy nhiên gần đây có một vài chung cư rất hấp dẫn (giá tốt, vị trí đẹp,...) nên lượng hồ sơ người nộp cho chương trình ưu đãi tăng đáng kể. Bình thường bạn có thể duyệt 10-20 hồ sơ một ngày để quyết định hồ sơ có được cho vay hay không, tuy nhiên gần đây bạn nhận được 1000-2000 hồ sơ mỗi ngày. Bạn không thể xử lý hết hồ sơ và bạn cần có một giải pháp để có thể dự đoán hồ sơ mới là có nên cho vay hay không. Sau khi phân tích thì bạn nhận thấy là hai yếu tố chính quyết định đến việc được vay tiền đó là mức lương và thời gian công tác. Đây là dữ liệu bạn có từ trước đến nay:



Về mặt logic, giờ ta cần tìm đường thẳng phân chia giữa các điểm cho vay và từ chối. Rồi quyết định hồ sơ mới có nên có vay hay không?



Ví dụ đường xanh là đường phân chia. Dự đoán cho hồ sơ của người có mức lương 6 triệu và 1 năm kinh nghiệm là không cho vay. Tuy nhiên, do ngân hàng đang trong thời kỳ khó khăn nên việc cho vay bị thắt lại, chỉ những hồ sơ nào chắc chắn trên 80% mới được vay. Vậy nên bây giờ bạn không những tìm là hồ sơ ấy cho vay hay không cho vay mà cần tìm xác suất nên cho hồ sơ ấy vay là bao nhiêu. Hay trong nhiều trường hợp khác trong bài toán phân loại người ta quan tâm hơn đến xác suất hay vì chỉ 1 hay 0. Ví dụ: bác sĩ sẽ thông báo ca mổ này 80% thành công cho người nhà bệnh nhân.

3.4.5.2. Model

Với dòng thứ i trong bảng dữ liệu, gọi x (i) 1 là lương và x (i) 2 là thời gian làm việc của hồ sơ thứ i . p(x (i) = 1) = yˆi là xác suất mà model dự đoán hồ sơ thứ i được cho vay.

p(x (i) = 0) = 1−yˆi là xác suất mà model dự đoán hồ sơ thứ i không được cho vay.

=> p(x (i) = 1) + p(x (i) = 0) = 1

Hàm sigmoid: 



3.4.5.3 Loss function

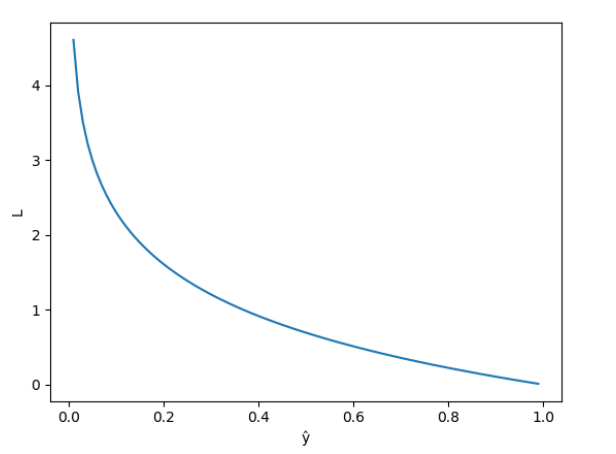
Giờ cũng cần một hàm để đánh giá độ tốt của model. Như bài trước là yˆ càng gần y càng tốt, giờ cũng vậy:

• Nếu hồ sơ thứ i là cho vay, tức yi = 1 thì ta cũng mong muốn yˆi càng gần 1 càng tốt hay model dự đoán xác suất người thứ i được vay vốn càng cao càng tốt.

• Nếu hồ sơ thứ i không được vay, tức yi = 0 thì ta cũng mong muốn yˆi càng gần 0 càng tốt hay model dự đoán xác suất người thứ i được vay vốn càng thấp càng tốt.

Với mỗi điểm (x (i) , yi), gọi hàm loss function L = −(yi ∗ log(yˆi) + (1−yi) ∗ log(1−yˆi)), loss function này có tên gọi là binary\_crossentropy

Thử đánh giá hàm L nhé. Nếu yi = 1 => L = −log(yˆi)

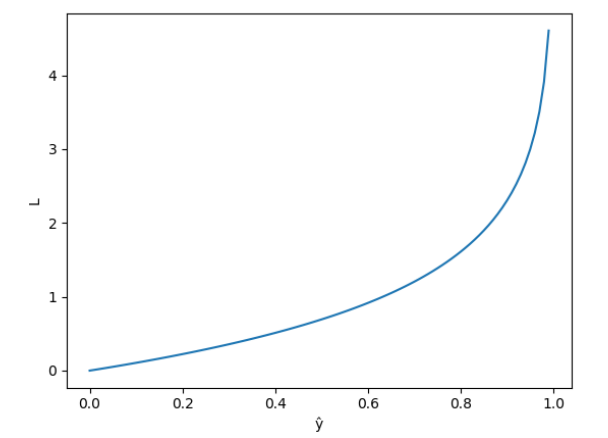


Hình 4.4: Đồ thị hàm loss function trong trường hợp yi = 1

• Hàm L giảm dần từ 0 đến 1.

• Khi model dự đoán yˆi gần 1, tức giá trị dự đoán gần với giá trị thật yi thì L nhỏ, xấp xỉ 0

• Khi model dự đoán yˆi gần 0, tức giá trị dự đoán ngược lại giá trị thật yi thì L rất lớn Ngược lại, nếu yi = 0 => L = −log(1−yˆi)



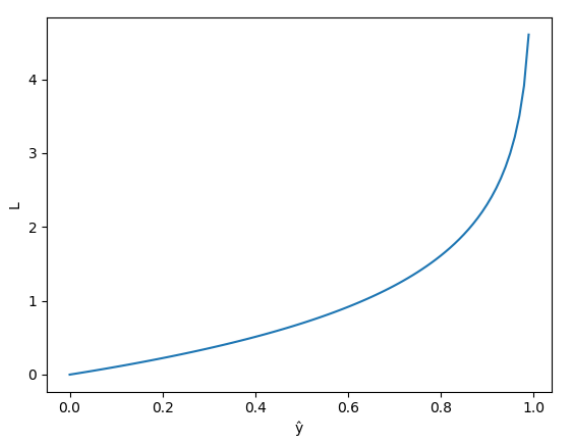
Hình 4.4: Đồ thị hàm loss function trong trường hợp yi = 1

Nhận xét:

• Hàm L giảm dần từ 0 đến 1.

• Khi model dự đoán yˆi gần 1, tức giá trị dự đoán gần với giá trị thật yi thì L nhỏ, xấp xỉ 0

• Khi model dự đoán yˆi gần 0, tức giá trị dự đoán ngược lại giá trị thật yi thì L rất lớn. Ngược lại, nếu yi = 0 => L = −log(1−yˆi)



Nhận xét:

• Hàm L tăng dần từ 0 đến 1

• Khi model dự đoán yˆi gần 0, tức giá trị dự đoán gần với giá trị thật yi thì L nhỏ, xấp xỉ 0

• Khi model dự đoán yˆi gần 1, tức giá trị dự đoán ngược lại giá trị thật yi thì L rất lớn

=> Hàm L nhỏ khi giá trị model dự đoán gần với giá trị thật và rất lớn khi model dự đoán sai, hay nói cách khác L càng nhỏ thì model dự đoán càng gần với giá trị thật. => Bài toán tìm model trở thành tìm giá trị nhỏ nhất của L

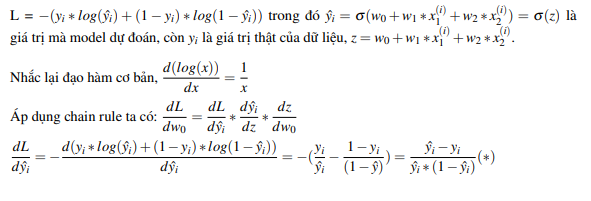
Hàm loss function trên toàn bộ dữ liệu

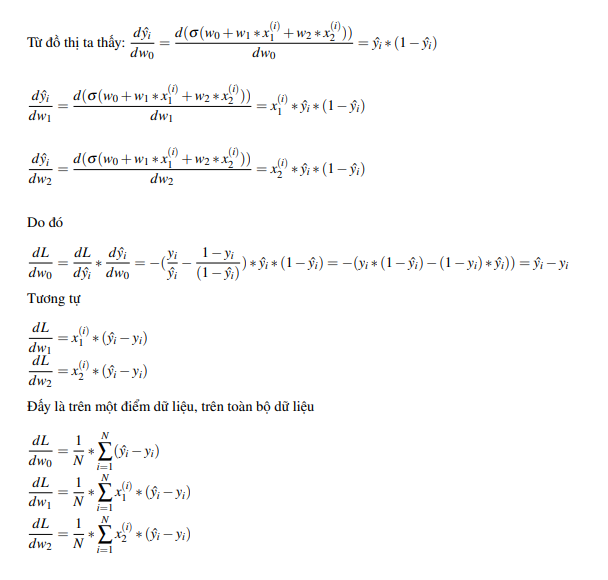
J = − 1/N ∗ N ∑ i=1 (yi ∗ log(yˆi) + (1−yi) ∗ log(1−yˆi))

3.4.5.4. Tối ưu hàm mất mát

Để áp dụng thuật toán gradient descent tìm tối ưu loss function mình cần tính đạo hàm của loss function với w. Với mỗi điểm (x (i) , yi), gọi hàm loss function

cho 2 cái ảnh này vào thành word không dán ảnh vào





# **Chương 4. Ứng dụng bài toán nghiên cứu dự đoán khả năng bị đau tim**

4.1. Đặt vấn đề

Đau tim (Các bệnh về tim mạch) là bệnh xảy ra khi dòng chảy của máu đến cơ tim đột ngột bị tắc nghẽn. Đau tim nói riêng và các bệnh tim mạch nói chung để lại di chứng, biến chứng và hậu quả nghiêm trọng cho mọi người như suy tim, đột quỵ,.. thậm chí chết người.

Bệnh tim mạch đã và đang trở thành gánh nặng cho xã hội, kéo theo nhiều hệ lụy trong đời sống. Theo WHO - Tổ chức Y tế Thế giới, mỗi năm có khoảng 17,5 triệu người tử vong do mắc các bệnh lý về tim và mạch máu, chiếm tỷ lệ cao nhất so với tất cả các nhóm bệnh lý khác. Theo thống kê từ tổ chức y tế nước ta cho biết, Việt Nam hiện nay có hơn 25% dân số đang mắc bệnh về tim mạch. Đáng quan ngại hơn, tỷ lệ này đang có xu hướng trẻ hóa. Là một bệnh đặc biệt nguy hiểm nhưng đau tim lại có thể giảm bớt nghiêm trọng của bệnh tình nếu được phát hiện sớm. Từ đây, việc phát hiện bệnh tim sớm là một việc vô cùng quan trọng và cấp thiết cần được quan tâm.

4.2. Dữ liệu bài toán

4.2.1. Dữ liệu đầu vào

Kích thước dữ liệu: 14 trường x 1025 bản ghi.

| Tên cột | Ý Nghĩa |
| --- | --- |
|  |  |
|  |  |
|  |  |
|  |  |
|  |  |
|  |  |
|  |  |
|  |  |
|  |  |
|  |  |
|  |  |
|  |  |
|  |  |
|  |  |

4.2.2. Dữ liệu đầu ra

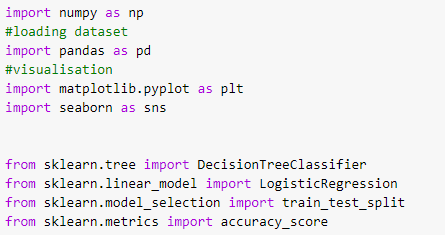
|  |  |
| --- | --- |
|  |  |
|  |  |

4.2.3. Nguồn

## **4.3. Ứng dụng thuật toán**

### **4.3.1. Chương trình:**

B1, Import thư viện cần thiết

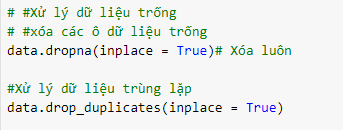


* Thư viện Numpy xử lý tính toán với mảng
* Thư viện pandas để đọc dữ liệu
* Thư viện matplotlib, seaborn để biểu diễn dữ liệu dưới dạng biểu đồ
* DecisionTreeClassifier của thư viện sklearn dùng để giải quyết bài toán cây quyết định
* Logistic Regression của Thư viện sklearn dùng để giả quyết bài toán hồi quy logistic
* train\_ test\_split của Thư viện sklearn dùng để tách bộ dữ liệu train và test trong cùng 1 file dữ liệu.
* accuracy\_score của thư viện sklearn dùng để tính độ chính xác của model

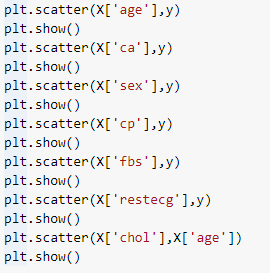
B2, Đọc data set



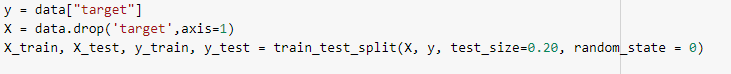
B3, Tiền xử lý dữ liệu trước khi thực hiện bài toán



B4: Mô tả dữ liệu dưới dạng biểu đồ



B5 Tách dữ liệu để train và test



B6 Train dữ liệu

Thuật toán cây QĐ - ID3:

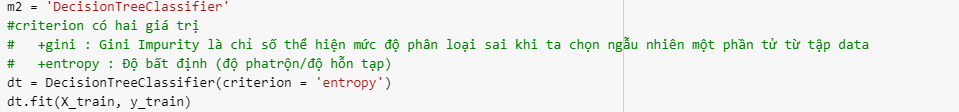
-Hàm DecisionTreeClassifier dùng để giải bài toán cây quyết định

+Tham số criterion là tiêu chí đo lường của một lần tách. Criterion có 2 giá trị gini(default) và entropy

+Entropy: là một thước đo thông tin chỉ ra sự rối loạn của các thuộc tính dẫn xuất với mục tiêu,

+Information entropy : thể hiện mức độ hỗn loạn hay độ nhiễu của data dùng để xác định thuộc tính nào mang lại nhiều thông tin hơn

-Ta dùng method fit để train dữ liệu với X\_train là dữ liệu về các thuộc tính dẫn xuất, y\_train là dữ liệu thuộc tính mục tiêu tương ứng



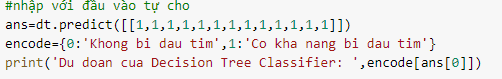
Thuật toán hồi quy Logistic:



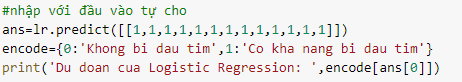
B7 Dự đoán và in ra

* Dự đoán trong bộ dữ liệu test

Thuật toán Cây QĐ:



Thuật toán Hồi Quy Logistic:



* Dự đoán theo yêu cầu ngoài bộ dữ liệu test

Thuật toán Cây QĐ:

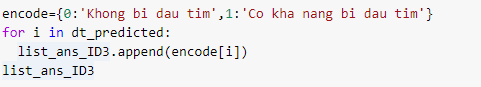


Thuật toán Hồi Quy Logistic:

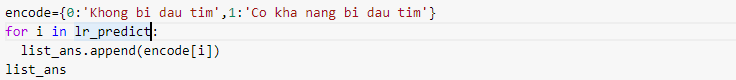


B7, Giải mã dữ liệu

### **Thuật toán cây QĐ**



Thuật toán hồi quy Logistic

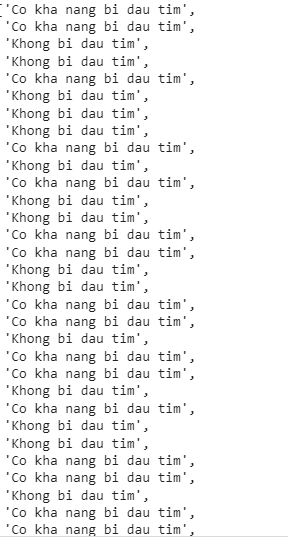


### **3.3.2. Kết quả**

1. Thuật toán cây QĐ

* Dự đoán trong bộ dữ liệu test



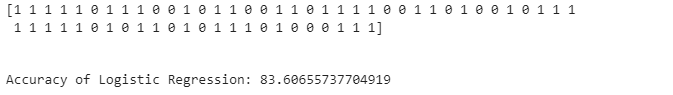


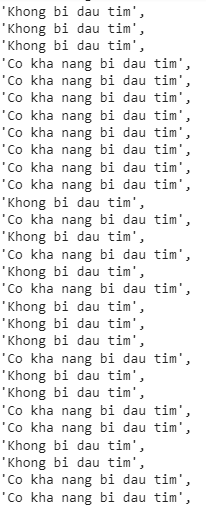
* Dự đoán theo yêu cầu ngoài bộ dữ liệu test



1. Thuật toán hồi qui Logistic

* Dự đoán trong bộ dữ liệu test





* Dự đoán theo yêu cầu ngoài bộ dữ liệu test



# 

# 

# **Tổng kết và Kiến Nghị**

1. Tổng kết

Ngày nay khi mà nền kinh tế, đời sống của con người ngày càng phát triển. Các công cụ, và giải phát ngày càng trở nên thông minh trong nền kinh tế 4.0 hiện nay. Chăm sóc sức khỏe được mọi người quan tâm và đặc biệt là đề tài dự đoán khả năng bị đau tim. Từ những thực tiễn đó nhóm em nghĩ đề tài này có thể được nâng cấp và phát triển nhiều hơn. Từ đó có thể đưa vào thực tiễn để ứng dụng trong cuộc sống.

Những kết quả đạt được:

* Sự hiểu biết về thuật toán Decision Tree, hồi quy Logistic cơ bản tương đối tốt
* Từ những gì đã làm được, từ đó biết thêm về AI, ứng dụng của AI vào cuộc sống công nghệ hiện đại
* Làm quen ngôn ngữ lập trình Python.
* Sử dụng thành tạo công cụ, làm quen với môi trường Colab.

Trong quá trình thực hiện bài tập lớn, nhóm đã cố gắng hết sức để tìm hiểu và học hỏi nhưng vì khả năng còn giới hạn không tránh khỏi những sai sót, nên có thể chưa giải quyết được tất cả những vấn đề đặt ra. Rất mong nhận được sự thông cảm của quý thầy cô và các bạn. Em xin chân thành cảm ơn.

1. Kiến Nghị

Mặc dù đạt được kết quả tưởng đối tốt khi test nhưng nhóm em cũng đã nhận thấy được rất nhiều điểm yếu của ứng dụng đề tài này.

**Điểm yếu :**

* Dữ liệu thu thập còn hạn chế chưa đủ lớn, dữ liệu còn bị nhiễu cần tiền xử lý nhiều.
* Thuật toán còn đơn giản, cần đạt độ chính xác cao hơn.

Hướng phát triển:

* Nghiên cứu thêm về các kĩ thuật, thuật toán về bài toán phân loại. · Từ đó tăng được tốc độ cũng như độ chính xác của việc dự đoán khả năng bị đau tim.
* Tiếp tục nghiên cứu các kĩ thuật xây dựng model, tối ưu bộ trọng số từ đó nâng cấp Model nhanh hơn, nhẹ hơn, chính xác hơn.

# **Tài liệu tham khảo**

**[1] Giáo trình trí tuệ nhân tạo - Trường Đại Học Công Nghiệp Hà Nội**

* **Nguyễn Phương Nga - Trần Hùng Cường**

**[2] Machine Learning Cơ bản - Vũ Hữu Tiệp**

**[3] Deep Learning Cơ Bản - Nguyễn Thanh Tuấn**

**[4] Machine Learning for Dummies – IBM**

**[5] Introduce to Machine Learning – Ethem Alpaydin**

**[6] Machine Learning, A Probabilistic Perspective**

**[7] Machine Learning Yearning - Andrew NG**

**[8] Micael DaGraca – Practical Game AI Programming-Packt Publishing**

**[9] Deep Learning- Ian Godfellow**

**[10] Tài liệu trên internet:**

<https://scikit-learn.org/stable/modules/tree.html>

<https://scikit-learn.org/stable/modules/generated/sklearn.linear_model.LogisticRegression.html>

<https://pandas.pydata.org/>

<https://www.coursera.org/learn/machine-learning/home>

<https://www.coursera.org/specializations/machine-learning-introduction>

https://www.coursera.org/specializations/deep-learning

Dataset:https://www.kaggle.com/code/nareshbhat/heart-attack-prediction-using-different-ml-models