BỘ GIÁO DỤC VÀ ĐÀO TẠO

TRƯỜNG ĐẠI HỌC MỞ THÀNH PHỐ HỒ CHÍ MINH

NGUYỄN TUẤN HƯNG

KHAI PHÁ VÀ PHÂN TÍCH DỮ LIỆU VỀ COVID 19

ĐỒ ÁN NGÀNH

NGÀNH KHOA HỌC MÁY TÍNH

TP. HỒ CHÍ MINH, 2023

BỘ GIÁO DỤC VÀ ĐÀO TẠO

TRƯỜNG ĐẠI HỌC MỞ THÀNH PHỐ HỒ CHÍ MINH

NGUYỄN TUẤN HƯNG

KHAI PHÁ VÀ PHÂN TÍCH DỮ LIỆU VỀ COVID 19

Mã số sinh viên:1951012042

ĐỒ ÁN NGÀNH

NGÀNH KHOA HỌC MÁY TÍNH

Giảng viên hướng dẫn: Tiến sĩ NGUYỄN TIẾN ĐẠT

TP. HỒ CHÍ MINH, 2023

LỜI CẢM ƠN

Em là Nguyễn Tuấn Hưng qua đề tài về khai phá và phân tích dữ liệu về Covid 19 em xin gửi lời cảm ơn sâu sắc đến thầy Nguyễn Tiến Đạt vì đây là một đề tài khá mới trong đồ án ngành nên có những thắc mắc về cách trình bày bài báo cáo,cách xây dựng các nội dung của từng chương,nhờ thầy đưa ra các gợi ý,câu hỏi ,lời khuyên về đề tài này của em giúp em có cách nhìn sâu sắc hơn và cố hoàn thành tốt đề tài theo ý của em với sự tư vấn từ thầy.

NHẬN XÉT CỦA GIÁO VIÊN HƯỚNG DẪN

.....

.....

.....

.....

.....

.....

.....

.....

.....

.....

.....

.....

.....

.....

TÓM TẮT ĐỒ ÁN NGÀNH

Đồ án thể hiện quá trình khai phá và phân tích dữ liệu Covid 19 giúp hiểu rõ hơn về các thuật toán trong việc vận dụng từng thuật toán,chỉnh sửa các thông số sao cho phù hợp với mô hình huấn luyện.Đưa ra các nội dung khái quát về khai phá dữ liệu,kết quả so sánh giữa các thuật toán.Nghiên cứu kĩ hơn về thuật toán chính áp dụng trong đề tài là Neural Network.

MỤC LỤC

DANH MỤC HÌNH VẼ

DANH MỤC BẢNG

MỞ ĐẦU

Ngày nay,thời buổi công nghệ ngày càng phát triển và việc vận dụng khoa học công nghệ thông tin vào việc xử lý,phân tích dữ liệu để giải quyết các vấn đề liên quan đến ngành y tế,kinh tế,giáo dục là vô cùng cần thiết.Đặc biệt là ngành giáo dục và y tế vì đó là vấn đề sống còn của bất kỳ quốc gia nào.

Trong đề tài này em sử dụng chủ đề về Data mining để có thể xác định,đánh giá tỉ lệ tử vong do dịch Covid 19 gây ra trên toàn thế giới nhằm cho thấy sự nguy hiểm của chúng luôn tiềm tàng đối với sức khỏe con người.

Chương 1. **TỔNG QUAN ĐỀ TÀI**

1.1. Đặt vấn đề

Việc sử dụng công nghệ thông tin cho việc lưu trữ,xử lý dữ liệu hiện nay được áp dụng đối với hầu hết các lĩnh vực,việc này tạo ra một lượng lớn dữ liệu được lưu trữ với kích thước vô cùng lớn và không ngừng tăng lên.Đó cũng vừa là thử thách vừa là cơ hội cho việc quản lý,khai thác và khai phá tri thức mới từ tệp dữ liệu khổng lồ đó bằng cách sử dụng các công cụ truy vấn đó lập biểu đồ thống kê các dữ liệu cảm thấy cần thiết.

Data mining là một quá trình sắp xếp,phân loại những tệp dữ liệu lớn được áp dụng bởi rất nhiều kỹ thuật phức tạp nhằm làm sạch,chuyển đổi,phân tích mẫu dữ liệu và tích hợp dữ liệu.Bên cạnh đó do áp dụng nhiều kỹ thuật đó mà có thể tìm kiếm được các tri thức tiềm ẩn bên trong một tệp dữ liệu lớn mà con người khó có thể nhận thấy được bằng những kỹ thuật,phép tính thông thường mà phải nhờ vào máy học.

Nguồn dữ liệu về y học vô cùng lớn,việc áp dụng khai phá dữ liệu trong lĩnh vực này mang lại nhiều ý nghĩa cho ngành y tế.Việc sử dụng các công cụ và các thuật toán phù hợp sẽ cung cấp được những thông tin quý giá nhằm hỗ trợ trong việc chuẩn đoán và điều trị bệnh hợp lý hơn.

Để minh chứng cho việc khai phá dữ liệu mang ý nghĩa như thế nào đối với ngành y tế,em sử dụng bộ dữ liệu về Covid 19 để thử nghiệm và đánh giá nhằm chuẩn đoán được các nguy cơ tiềm tàng của các bệnh nhân có mang bệnh nền sẵn khi mắc phải virus corona tỉ lệ nguy hiểm đến tính mạng có cao hay không.

1.2. Lý do chọn đề tài

Theo thống kê của bộ y tế kể từ đầu dịch đến nay,Việt Nam có 11.623.858 ca nhiễm,đứng thứ 13 trên 231 quốc gia.Số ca tử vong do Covid 19 tại Viêt Nam tính đến nay là 43.206 ca,chiếm tỉ lệ 0,4% so với tổng số ca nhiễm.[1]

Trong số các ca nhiễm Covid 19 tại Việt Nam tỉ lệ bệnh nhân nặng là 6%,mức trung bình là 8,3%,nhẹ và không có triệu chứng là 85,7%.Qua phân tích của các chuyên gia số ca bệnh bị tử vong là những bệnh nhân trên 65 tuổi chiếm 47% là người có bệnh nền,36% từ 50-56 tuổi,18-49 tuổi là 15%,nhóm từ 0-17 tuổi là 0,42% [2]

Qua những thông tin trên việc áp dụng khai phá dữ liệu nhằm đánh giá tỉ lệ ,nhóm bệnh nhân có nguy cơ cao,dự đoán được mức độ nguy hiểm của từng bệnh nhân để đưa ra các biện pháp y tế kịp thời chuẩn đoán đúng bệnh giúp bệnh nhân có thể hồi phục trở lại là lý do em chọn đề tài về "khai phá và phân tích dữ liệu Covid 19".

1.3. Mục tiêu đề tài

Đề tài tập trung nghiên cứu các thuật toán trong khai phá dữ liệu nhằm đưa ra thuật toán phù hợp với bộ dữ liệu Covid 19 làm tiền đề cho việc nghiên cứu ,chuẩn đoán tỉ lệ những bệnh nhân có nguy cơ cao ảnh hưởng đến tính mạng sẽ được chăm sóc y tế theo các phác đồ của các y bác sĩ cho phù hợp nhất. Phân tích các đặc điểm,thuộc tính của dữ liệu,xây dựng và đánh giá chất lượng ,độ hiệu quả của từng thuật toán là mục tiêu của đề tài.

1.4. Ý nghĩa đề tài

1.4.1. Ý nghĩa thực tiễn

Việc đánh giá tỉ lệ tử vong do Covid 19 gây ra đối với các bệnh nhân là vô cùng quan trọng vì khi nắm rõ được tình trạng sức khỏe của bệnh nhân có mang bệnh nền hay không,độ tuổi thấp hay cao nhằm đưa ra các dự đoán về việc bệnh nhân cần được điều trị theo phác đồ của các y bác sĩ một cách hợp lý và nhanh chóng nhằm chữa trị kịp thời tránh những mất mát to lớn thiệt hại đến tính mạng.Nếu việc chuẩn đoán sai tình trạng bệnh làm việc đưa ra các phác đồ không phù hợp mang đến những hậu quả rất lớn.

1.4.2. Ý nghĩa khoa học

Với sự phát triển của khoa học kỹ thuật lĩnh vực y tế,với sự giúp đỡ của máy tính đề tài mang lại ý nghĩa to lớn cho việc thiết kế và thử nghiệm các thuật toán khác nhau đưa ra kết quả so sánh về thời gian xử lý từng thuật toán,kết quả dự đoán giữa các thuật toán đó nhằm hỗ trợ cho các y bác sĩ có thể nhận biết sớm tình trạng của bệnh nhân để có thể kịp thời cứu chữa.

Chương 2. Cơ sở lý thuyết

2.1. Tổng quan về khai phá dữ liệu (data mining)

2.1.1. Khái niệm

Khai phá dữ liệu bao gồm các thuật toán cốt lõi cho phép người sử dụng đạt được các kiến thức ,hiểu biết sâu sắc từ dữ liệu lớn.Đó là một lĩnh vực liên ngành kết hợp giữa các lĩnh vực liên quan như hệ thống cơ sở dữ liệu,thống kê,máy học,nhận dạng mẫu.Trên thực tế, việc khai phá dữ liệu là một phần trong quá trình khai phá tri thức mới.[3]

2.1.2. Các bước trong khai phá dữ liệu

2.1.2.1. Xác định vấn đề và hiểu về dữ liệu

- Xác định rõ mục tiêu và yêu cầu của dữ liệu cần phân tích.
- Tìm hiểu và thu thập dữ liệu từ các nguồn khác nhau để hiểu rõ và đảm bảo có đủ dữ liệu để thực hiện trong quá trình khai phá tìm tri thức mới.
- Hiểu về đặc tính của dữ liệu,các đặc trưng quan trọng và tiềm năng từ dữ liệu mang lại.

2.1.2.2. Chuẩn bị dữ liệu và tiền xử lý dữ liệu

- Làm sạch dữ liệu (data cleaning).
- Tích hợp dữ liệu (data integration).
- Biến đổi dữ liệu (data transfomation).

2.1.2.3. Lựa chọn thuật toán

- Có thể sữ dụng nhiều thuật toán để khai phá trong một tệp dữ liệu.
- Tiến hành huấn luyện và kiểm thử để dưa ra thuật toán cảm thấy phù hợp nhất với tệp dữ liệu đó nhằm tạo ra mô hình tốt nhất để sử dụng.

2.1.2.4. Triển khai và khai phá tri thức

- Sử dụng mô hình tốt nhất ở bước để đưa ra dự đoán,phân loại tạo ra báo cáo giúp đưa ra quyết định dễ dàng hơn.

2.1.2.5. Đánh giá và cải thiện

- Sau khi triển khai mô hình,đánh giá hiệu suất của mô hình và cải thiện mô hình đó nếu cảm thấy cần thiết.Bao gồm việc xem xét lại dữ liệu,phát hiện và xử lý các vấn đề phát sinh nhằm đưa ra mô hình mới cho phù hợp hơn.

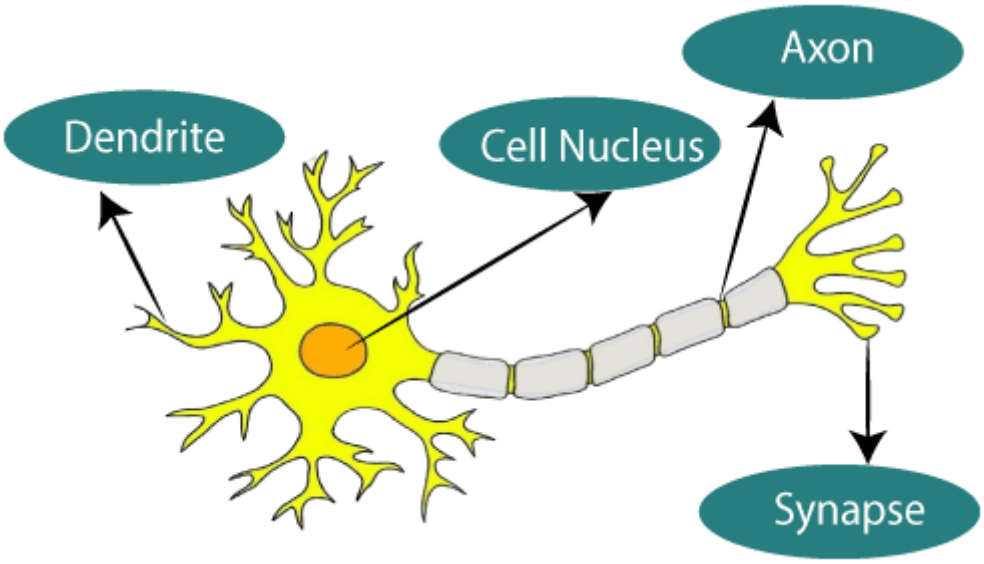
2.2. Tổng quan về thuật toán Mạng nơ ron nhân tạo(Artificial neural network)

2.2.1. Khái niệm

Là một mạng lưới thần kinh nhân tạo có nguồn gốc từ mạng lưới thần kinh sinh học được phát triển theo cấu trúc của bộ não con người.

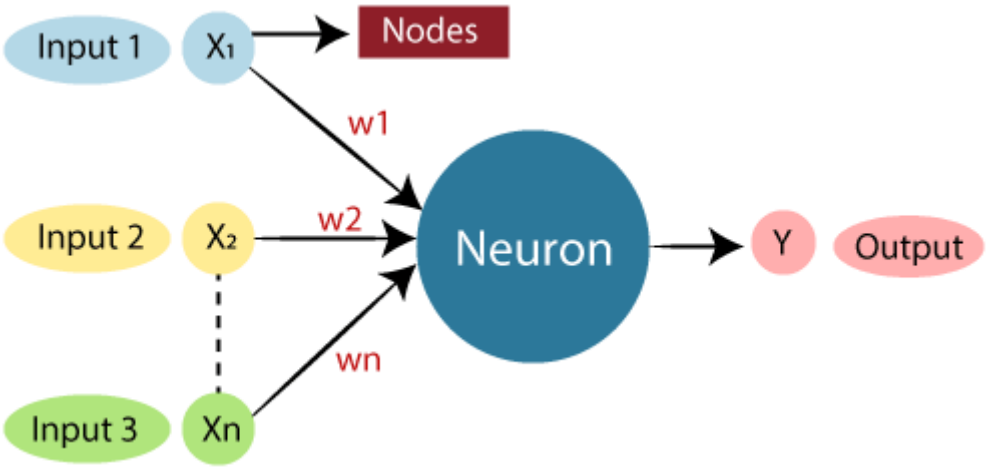
Tương tự như bộ não con người có các nơ ron được kết nối với nhau,mạng nơ ron nhân tạo cũng có các nơ ron được kết nối ở nhiều lớp khác nhau của mạng.Những nơ ron trong nhân tạo được gọi là nút.

Một mạng nơ ron nhân tạo được áp dụng đối với một bài toán cụ thể như nhận dạng mẫu,chuẩn đoán,phân loại dữ liệu phải thông qua quá trình học từ tập các mẫu huấn luyện.Về bản chất việc máy học chính là điều chỉnh các trọng số ở các lớp,các nút của mạng nơ ron.



Hình 2.2.1-1 : Mạng lưới thần kinh sinh học

.



Hình 2.2.1-2 : Mạng lưới thần kinh nhân tạo

Nơ ron sinh học	Nơ ron nhân tạo
Dendrites (sợi nhánh)	Inputs
Cell nucleus (thân tế bào)	Nodes
Axon (sợi trục)	Weights
Synapse (khớp thần kinh)	Output

Bảng 2.2.1-1: Mối quan hệ giữa mạng nơ ron sinh học và nhân tạo

Đặc trưng	Nơ ron nhân tạo	Nơ ron sinh học
Tốc độ	Nhanh hơn trong việc xử lý thông tin.Thời gian phản hồi tính bằng nano/s	Chậm hơn trong việc xử lý thông tin.Thời gian phản hồi tính bằng mili/s
Xử lý	Xử lý nối tiếp	Xử lý song song
Kích thước và độ phức tạp	Kích thước nhỏ,không thực hiện các nhiệm vụ có nhiều mẫu phức tạp	Kích thước lớn,dày đặt các nơ ron liên kết với nhau .Rất phức tạp
Lưu trữ	Lưu trữ thông tin có thể thay thế được.	Lưu trữ thông tin chỉ có thể thêm vào chứ không thay thế
Kiểm soát tính toán	Có bộ phận để điều khiển hoạt động tính toán	Không có cơ chế kiểm soát tính toán cụ thể.

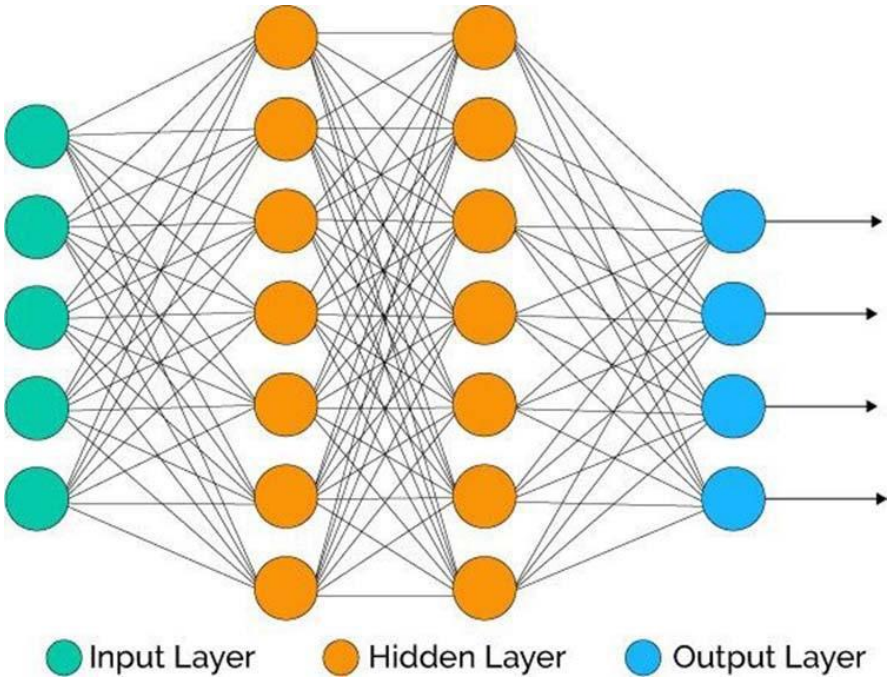
Bảng 2.2.1-2 :Sự khác biệt giữa mạng nơ ron sinh học và nhân tạo

2.2.2. Cấu trúc mạng nơ ron

Mạng nơ-ron nhân tạo bao gồm 3 lớp cơ bản:

- Lớp đầu vào (Input layer).
- Lớp ẩn (Hidden layer).
- Lớp đầu ra (output layer).

Thông tin sẽ được truyền thẳng từ đầu vào cho tới đầu ra.



Hình 2.2.2-1: Cấu trúc mạng nơ ron nhân tạo.

Lớp đầu vào sẽ có số nơ ron tương ứng với số biến phụ thuộc vào tệp dữ liệu.Lớp đầu ra sẽ có số nơ ron tương ứng với số nhóm trong bài toán phân loại của dữ liệu..Lớp ẩn sẽ có số lượng nơ ron trong mỗi lớp ẩn là tùy ý không có một quy định nào về việc chia số lượng nơ ron cho lớp ẩn.Đối với việc điều chỉnh số lượng lớp ẩn thì phải cân nhắc vì khi có nhiều lớp ẩn sẽ khiến cho mô hình huấn luyện quá vừa vặn theo tập huấn luyện. [5]

2.2.3. Huấn luyện mạng nơ ron

2.2.3.1. Học có giám sát

Là quá trình huấn luyện bằng cách cung cấp dữ liệu đầu vào và đầu ra phù hợp ,với số lớp phân loại đã biết trước.Nhiệm vụ là xác định được dữ liệu đầu vào sẽ được phân loại chính xác vào lớp của nó.

2.2.3.2. Học không giám sát

Là quá trình huấn luyện trong đó đầu ra được đào tạo để đáp ứng với một cụm mẫu trong đầu vào.Học không giám sát sử dụng thuật toán máy học để phân tích và phân cum các tập dữ liệu không được gắn nhãn thuộc tính.

2.2.3.3. Học tăng cường

Còn được gọi là học thưởng phạt,là sự kết hợp giữa việc học có giám sát và không giám sát.Phương pháp này với dữ liệu đầu vào cung cấp,quan sát dữ liệu đầu ra do máy tính được.Nếu kết quả mình đánh giá là tốt thì sẽ tiếp tục tăng các trọng số kết nối lên,nếu kết quả đánh giá được cho là không tốt các trọng số không phù hợp thì sẽ giảm xuống.

2.2.4. Ứng dụng của mạng nơ ron

Ngày nay,phương pháp mạng nơ ron nhân tạo ngày càng được sử dụng nhiều trong thực tế bởi vì tính đa dụng của nó đặc biệt đối với các bài toán nhận dạng mẫu,xử lý,lọc dữ liệu.

Với các ứng dụng quan trọng:

- Nhận dạng khuôn mặt.
- Dự đoán thị trường
- Tối ưu quãng đường di chuyển.
- Nhận dạng chữ viết tay.
- Dự đoán tỉ lệ mắc bệnh,nhiễm bệnh.

2.2.5. Ưu và nhược điểm của mạng nơ ron

2.2.5.1. Ưu điểm

- Có thể thực hiện các nhiệm vụ mà chương trình tuyến tính không thể thực hiện.
- Khi một phần tử của mạng lưới bị lỗi,tính chất song song của nó có thể tiếp tục mà không gặp vấn đề gì.
- Có thể thực hiện đối với bất kỳ bài toán nào.

2.2.5.2. Nhược điểm

- Cần trải qua quá trình huấn luyện và đánh giá trước để hoạt động.
- Kiến trúc mạng nơ ron khác với kiến trúc của vi xử lý do đó việc xử lý có thể chậm.
- Thời gian xử lý cao đối với tệp dữ liệu lớn.

Chương 3. **Phân tích và thiết kế dữ liệu**

3.1. Phân tích dữ liệu Covid 19

USMER	MEDICAL_UNIT	SEX	PATIENT_TYPE	DATE_DIED	INTUBED	PNEUMONIA	AGE	PREGNANT	DIABETES	...	ASTHMA	INMSUPR	HIPERTENSION	OTHER_DISEASE	CARDIOVASCULAR	OBESITY	RENAL_CHRONIC	TOBACCO	CLASIFFICATION_FINAL	ICU
2	1	1	1	03/05/2020	97	1	65	2	2	...	2	2	1	2	2	2	2.0	2.0	3.0	97.0
2	1	2	1	03/06/2020	97	1	72	97	2	...	2	2	1	2	2	1	1.0	2.0	5.0	97.0
2	1	2	2	09/06/2020	1	2	55	97	1	...	2	2	2	2	2	2	2.0	2.0	3.0	2.0
2	1	1	1	12/06/2020	97	2	53	2	2	...	2	2	2	2	2	2	2.0	2.0	7.0	97.0
2	1	2	1	21/06/2020	97	2	68	97	1	...	2	2	1	2	2	2	2.0	2.0	3.0	97.0

Hình 3.1-1:Dữ liệu thô ban đầu

Mô tả dữ liệu:dữ liệu chứa một số lượng lớn thông tin của các bệnh nhân bị nhiễm Covid 19. Trong đó giá trị ‘1’ có nghĩa là có và ’2’có nghĩa là không.Các giá trị 97 và 99 là dữ liệu bị thiếu.

USMER: Cho biết bệnh nhân đã được điều trị tại các đơn vị y tế được chia thành 2 cấp là 1,2.

SEX:nữ (số 1) hoặc nam (số 2).

PATIENT_TYPE: bệnh nhân đã nhập viện hoặc không.

INTUBED: bệnh nhân có được đặt ông thở hay không.

PNEUMONIA: bệnh nhân đã bị viêm phổi hay chưa.

AGE : tuổi của bệnh nhân..

PREGNANT: bệnh nhân có mang thai hay không.

DIABETES: bệnh nhân có bị bệnh tiểu đường hay không.

COPD: bệnh nhân có bị bệnh phổi tắc nghẽn mãn tính hay không.

ASTHMA:bệnh nhân có bị hen suyễn hay không.

INMSUPR:bệnh nhân có bị ức chế hệ miễn dịch hay không.

HIPERTENSION:bệnh nhân có bị cao huyết áp hay không.

CARDIOVASCULAR:bệnh nhân có bệnh liên quan đến tim mạch hay không.

OBESITY: bệnh nhân có bị béo phì không.

RENAL_CHRONIC: bệnh nhân có bị bệnh thận mãn tính không.

TOBACCO:bệnh nhân có hút thuốc hay không.

ICU:bệnh nhân đã được đưa vào đơn vị chăm sóc đặc biệt hay chưa.

DEATH:cho biết bệnh nhân đã chết hay đã hồi phục.

3.1.1. Làm sạch,chuẩn hóa dữ liệu:

Loại bỏ các hàng có những giá trị dữ liệu bị thiếu.

Bỏ các cột CLASIFFICATION_FINAL, MEDICAL_UNIT, OTHER_DISEASE, ICU, INTUBED, RENAL_CHRONIC vì các cột này không thực sự cần thiết trong quá trình phân tích và có nhiều dữ liệu bị sai số ,thiếu sót.

Ở cột DATE_DIED hiển thị giá trị '9999-99-99'có nghĩa là bệnh nhân đang hồi phục còn các giá trị khác là bệnh nhân đã mất.Do đó sem sẽ tạo ra một cột mới là DEATH1 gồm 2 giá trị là 1 và 2.

Cột PREGNANT: chuyển đổi giá trị 97 thành 2 vì hầu hết các giá trị 97 đều nằm ở cột nam.

Cột AGE :để thuận tiện cho việc rời rạc hóa dữ liệu nên thiết lập điều kiện để chỉnh sửa lại giá trị cột là '0','0.5','1' ở giá trị 0 có độ tuổi từ 1-17 tuổi,0.5 có độ tuổi từ 18-54 tuổi và cuối cùng là 1 từ 55 tuổi trở lên.

Để thuận tiện cho việc áp dụng thuật toán phân lớp chuyển đổi giá trị 1 và 2 của cột DEATH thành 'Yes' và 'No'

	USMER	SEX	PATIENT_TYPE	PNEUMONIA	AGE	PREGNANT	DIABETES	COPD	ASTHMA	INMSUPR	HIPERTENSION	CARDIOVASCULAR	OBESITY	TOBACCO	DEATH
0	1	1	1	2	0.5	2	2	2	2	2	2	2	2	2	No
1	1	2	2	1	0.5	2	2	2	2	2	1	2	1	2	Yes
2	2	2	2	1	1.0	2	1	2	2	2	1	2	2	2	Yes
3	1	2	1	1	0.5	2	2	2	2	2	2	2	2	2	No
4	1	1	1	2	0.5	2	2	2	2	2	2	2	2	2	No

Hình 3.1.1-1 : Dữ liệu sau khi được làm sạch và chuẩn hóa.

3.2. Thiết kế dữ liệu

3.2.1. Tiền xử lý dữ liệu

Sau khi đã làm sạch và chuẩn hóa dữ liệu.Chia tệp dữ liệu đó thành 2 tệp để thuận tiện cho việc áp dụng các thuật toán để huấn luyện mô hình 1 tệp train và 1 tệp test.

Ở tệp train chiếm 80% tổng số dữ liệu và tệp test chiếm 20% còn lại.

3.2.1.1. Trainning set

Tập huấn luyện dùng để huấn luyện mô hình.Ở tệp này sẽ áp dụng các thuật toán máy học để áp dụng cho quá trình học của máy.Tùy vào thuật toán và mô hình mà việc học có thể sẽ không giống nhau.

3.2.1.2. Testing set

Mục đích của việc máy học là tạo ra các mô hình có thể dự đoán tốt trên các tệp dữ,vì vậy muốn biết thuật toán có thực sự tốt hay không cần phải dùng tệp test cho vào mô hình đã huấn luyện ở tệp trainning nếu tốt thì nhận kết quả.Còn không chúng ta quay lại bước cài đặt,thiết lập các thông số của thuật toán áp dụng vào tệp trainning cho đến khi kết quả của tệp trainning và tệp test có thể nó là đồng bộ với nhau hoặc có thể chênh lệch một ít.

Chương 4. **Thực nghiệm,đánh giá kết quả**

4.1. Giới thiệu bài toán

4.1.1. Cơ sở dữ liệu

Trong đề tài sử dụng bộ dữ liệu Covid 19 để làm cơ sở đánh giá hệ thống.Bộ dữ liệu này được chia thành 2 phần là dữ liệu cho quá trình huấn luyện mô hình và dữ liệu dành cho kiểm tra.Dữ liệu dùng để huấn luyện gồm 10959 dòng chứa các thông số về tuổi,giới tính,các bệnh nền,tình trạng còn sống hay đã mất.Tất cả đã được làm sạch và chuẩn hóa để có thể phù hợp cho quá trình chạy thuật toán.V

4.2. Thực nghiệm huấn luyện bộ dữ liệu trên nhiều thuật toán

Thuật toán	Lớp đã được định nghĩa trong Weka	Độ chính xác phân lớp	Thời gian
Tree	J48	90.9298%	0.37s
	DecisionStump	89.8257%	0.04s
	RandomForest	91.8058%	2.96s
	RandomTree	91.7876%	0.04s
	REPTree	90.9572%	0.28s
	LMT	90.7108%	5.95s
Bayes	BayesNet	88.9497%	0.27s
	NaiveBayesMultinomial	81.3395%	0.03s
	NaiveBayesMultinomialText	71.8314%	0.02s
	NaiveBayesUpdateable	89.1961%	0.01s

Neural network	Mạng nơ ron với 1 lớp ẩn	90.5831%	12.27s
	Mạng nơ ron với 2 lớp ẩn	90.5192%	20.85s
	Mạng nơ ron với 3 lớp ẩn	90.4188%	28.68s
	Learning rate 0.1	90.7747%	12.58s
	Learning rate 0.3	90.5831%	12.27s
	Learning rate 0.9	90.282%	11.98s

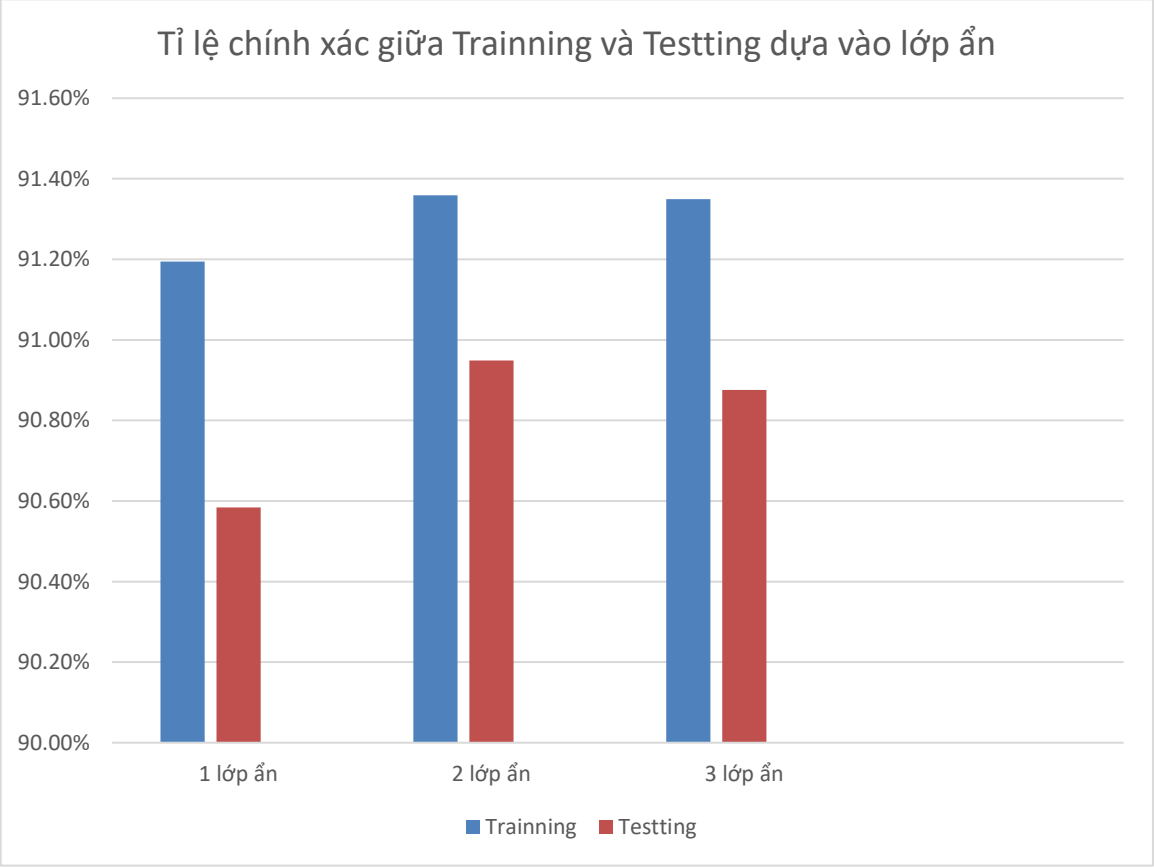
Bảng 4.2-1: Bảng so sánh giữa các thuật toán

4.3. Áp dụng mô hình huấn luyện của thuật toán Neural network vào tệp test

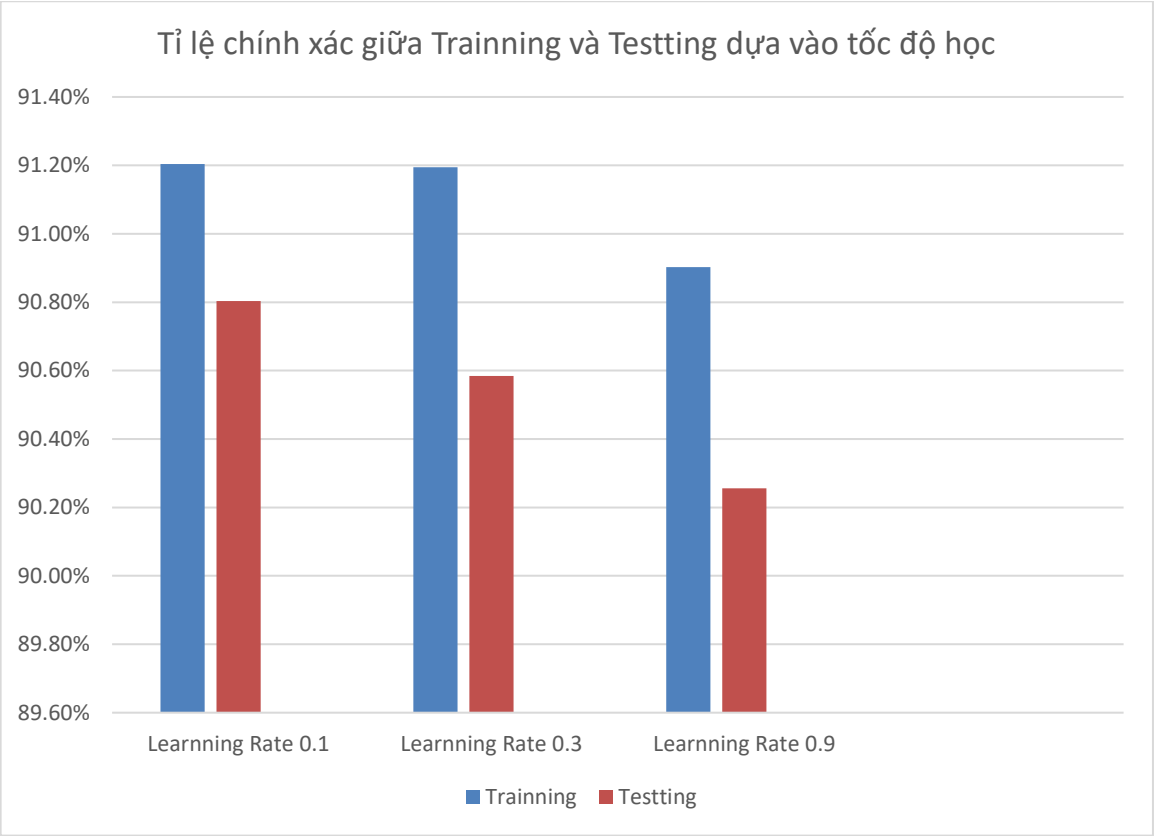
Thuật toán	Thông số	Trainning	Testting
	Mạng nơ ron với 1 lớp ẩn	91.1945%	90.5839%
	Mạng nơ ron với 2 lớp ẩn	91.3587%	90.9489%
	Mạng nơ ron với 3 lớp ẩn	91.3496%	90.8759%

Neural network	Learning rate 0.1	91.2036%	90.8029%
	Learning rate 0.3	91.1945%	90.5839%
	Learning rate 0.9	90.9025%	90.2555%

Bảng 4.3-1: Tỉ lệ chính xác của tập train và test khi thay đổi thông số



Hình 4.3-1: Tỉ lệ chính xác dựa vào lớp ẩn



Hình 4.3-2: Tỉ lệ chính xác dựa vào Learning Rate

Dựa vào các tỉ lệ vừa phân tích và thể hiện trực quan qua bảng 4.3,4.2,và biểu đồ 4.3.1,4.3.2 .Có thể thấy việc lựa chọn thông số của lớp ẩn và tốc độ học của mô hình ảnh hưởng nhiều đối với thời gian huấn luyện và kết quả từ độ chính xác mà từng thông số mang lại.Tùy vào yêu cầu của người sử dụng mô hình để làm gì mà có thể quyết định thông số nếu như đề tài không cần độ chính xác chênh lệch quá nhiều nhưng quan trọng thời gian nhanh hơn thì có thể sử dụng thông số Learning rate =0.9 và lớp ẩn =1 và ngược lại số lớp ẩn càng ít và thông số learning rate =0.1 sẽ cho ra độ chính xác cao nhất nhưng tốc độ sẽ chậm nhất.

4.4. Nhận xét và đánh giá về các thuật toán

Qua quá trình thực nghiệm và thử nghiệm các thuật toán.

- Tác động của số lớp ẩn đối với thuật toán Neural Network
  Việc lựa chọn số lớp ẩn trong thuật toán mạng nơ ron nhân tạo hay số lượng nơ ron trong lớp ẩn hiện tại vẫn chưa có một quy tắc cụ thể.Thực nghiệm đã sử dụng cùng số nơ ron và chia thành 1 lớp ,2 lớp,3 lớp với các tỉ lệ chênh lệch về độ chính xác là không nhiều tuy nhiên về thời gian thực hiện với một mô hình có nhiều lớp mạng nơ ron hơn thì thời gian để học có sự chênh lệch không nhỏ.Có thể nhận xét việc tăng số lớp ẩn làm cho mô hình mạng nơ ron nhân tạo trở nên phức tạp kéo dài thời gian hoành thành quá trình huấn luyện hơn.

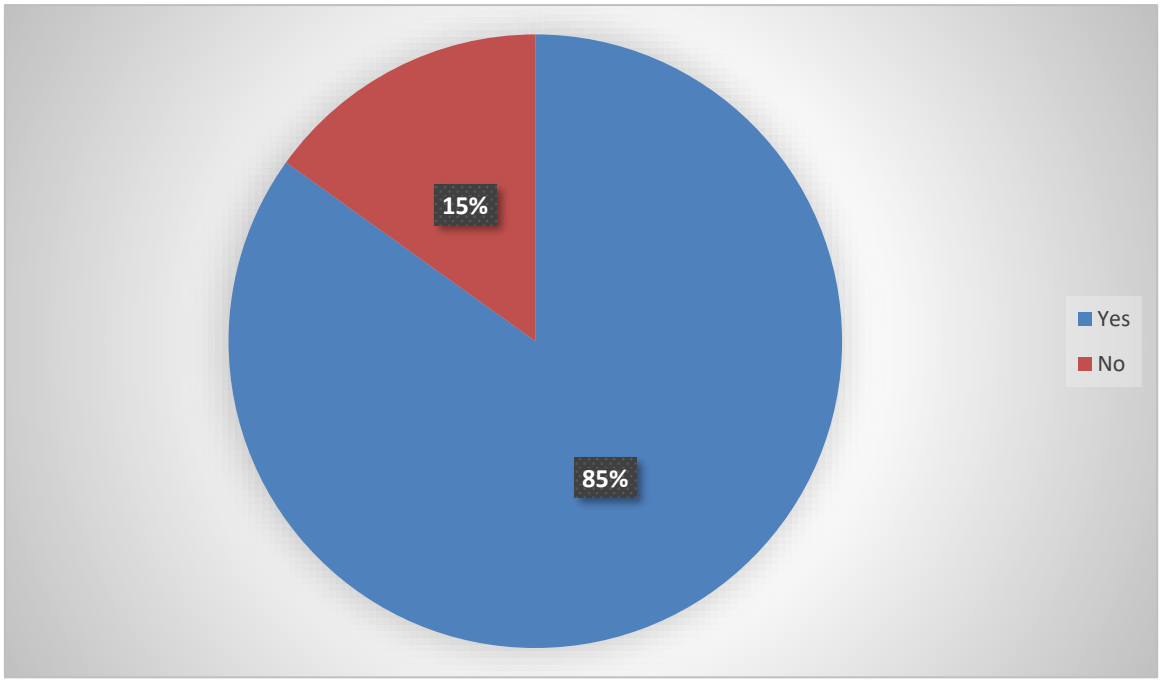
- Tác động của tỉ lệ học- Learning Rate (LR) đối với thuật toán Neural Network
  Tốc độ học của máy được sử dụng trong việc huấn luyện các mạng nơ ron.Là một số dương nằm ở khoảng từ 0-1.Đóng vai trò quan trọng đối với tỉ lệ chính xác của thuật toán.Nếu tốc độ học càng lớn thì mô hình huấn luyện sẽ nhanh hơn tuy nhiên độ chính xác sẽ giảm đi.

- Qua bảng so sánh 4.2 có thể thấy đưa một tập dữ liệu chạy trên các thuật toán khác nhau như Tree,Bayes,Neural network thì thuật toán Neural network đưa ra tỉ lệ chính xác cao nhất nhưng lại có thời gian thực hiện khá lâu so với 2 thuật toán còn lại.
- Đối với thuật toán Bayes tốc độ nhanh nhất nhưng tỉ lệ chính xác lại thấp nhất
- Đối với thuật toán Tree đánh giá về tốc độ học và tỉ lệ chính xác có thể thấy là ổn định nhất vì thời gian không tốn lâu để học và tỉ lệ chính xác của thuật toán mang lại cũng khá ổn.
- Tuy nhiên tùy vào tệp dữ liệu lớn hay nhỏ và các đặc trưng về thuộc tính của tệp dữ liệu đó ra sao muốn nhận thấy thuật toán nào ổn định nhất cần phải có sự phân tích kĩ càng để lựa chọn ra thuật toán phù hợp với bộ dữ liệu đó.

4.5. Nhận xét và đánh giá về dữ liệu Covid 19

- Qua quá trình thực nghiệm và nghiên cứu thuật toán
  Thuật toán Neural Network giúp em có thể hiểu rõ hơn về các yếu tố ảnh hưởng đến tính mạng khi mắc phải virus Corona.

4.5.1. Mối liên quan giữa người hút thuốc và bệnh phổi mãn tính

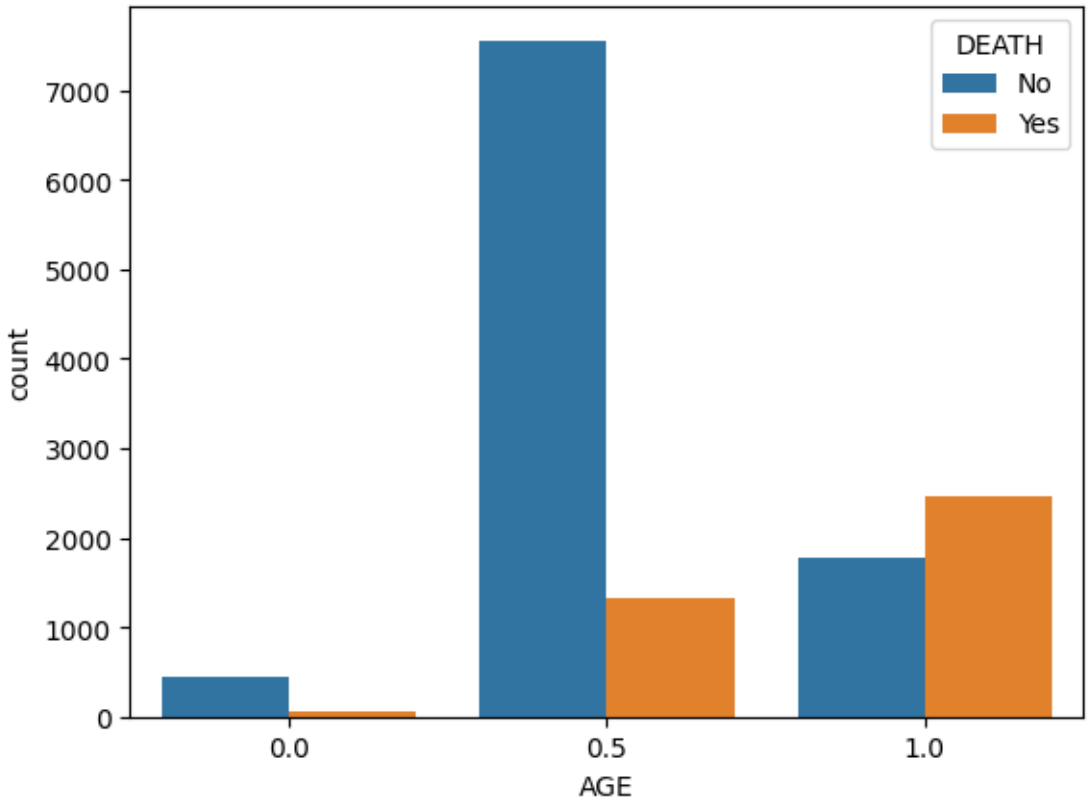


Hình 4.5.1-1 : Số lượng người hút thuốc và mắc bệnh tắc nghẽn phổi mãn tính bị mất do Covid

Trong dữ liệu gồm 13699 người có 106 người có hút thuốc và mắc bệnh tắc nghẽn phổi mãn tính

Trong đó ở hình 4.5.1 hiển thị tỉ lệ phần trăm người mất do có hút thuốc và mắc phải bệnh là 85% tỉ lệ tử vong và 15% là còn sống. Có thể thấy việc hút thuốc và mắc phải bệnh này có mối liên quan đến nhau và chiếm tỉ lệ rất cao dẫn đến tử vong.

4.5.2. Tuổi



Hình 4.5.2-1:Số lượng người tử vong

Đối với tuổi gồm 3 giá trị là 0,0.5,1 tương ứng với 3 khoảng độ tuổi

Giá trị 0 : từ 1 đến 17 tuổi

Giá trị 0.5 : từ 18-54 tuổi

Giá trị 1 : từ 55 trở đi

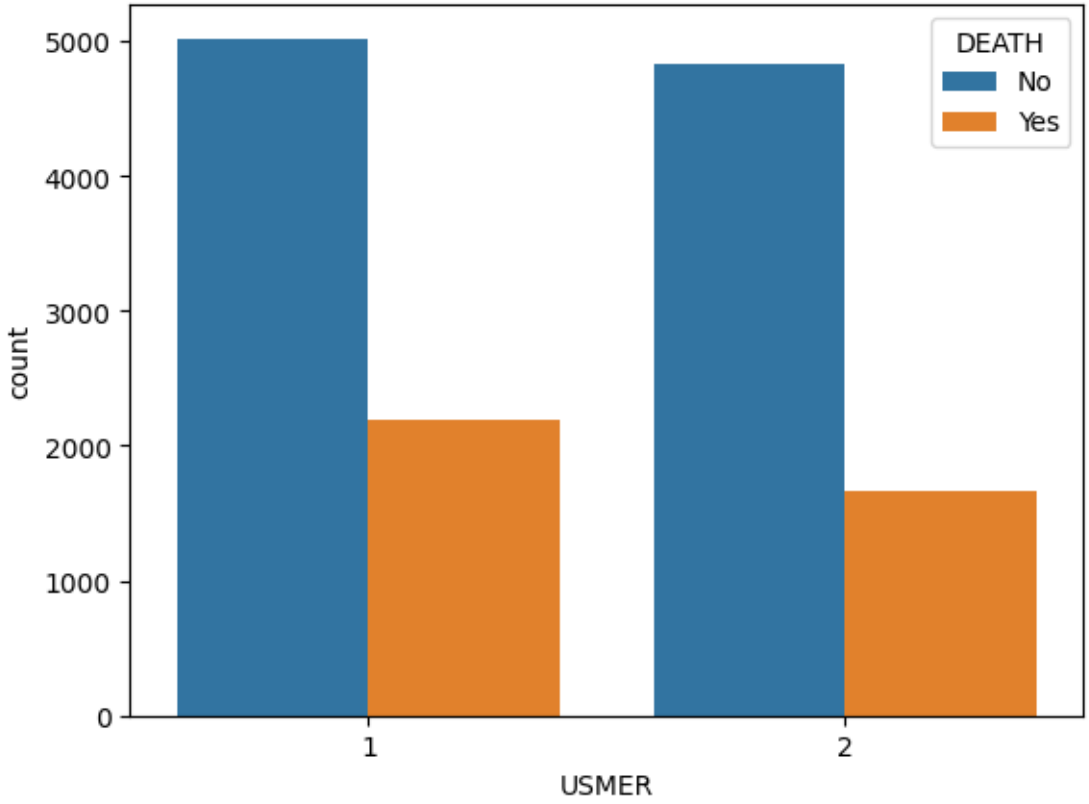
Dựa vào hình 4.5.2 có thể thấy số lượng người mất cao nhất chủ yếu là người lớn tuổi và chiếm 58% tỉ lệ tử vong đối với người lớn tuổi.

Đối với người từ 1-17 tuổi tỉ lệ tử vong rất thấp và người mắc phải covid cũng rất ít.

Người từ 18-54 chiếm số lượng lớn người mắc phải covid nhưng tỉ lệ mất do covid là 14,8% trên tổng số 8872 người nhiễm.

4.5.3. Đơn vị y tế (USMER)

Ở thuộc tính này được chia thành 2 giá trị là 1,2 tương trưng cho 2 cấp độ của đơn vị y tế đó cấp 1 các đơn vị y tế địa phương,cấp 2 là các bệnh viện.



Hình 4.5.3-1 : Số lượng người tử vong được điều trị y tế theo cấp 1,2

Qua hình 4.5-3 số lượng người mất khi điều trị ở các đơn vị y tế địa phương nhiều hơn là số người mất khi điều trị tại các bệnh viện.

Ở địa phương gồm 7204 số ca mắc được điều trị và số ca tử vong là 2185 chiếm tỉ lệ 30,3% tỉ lệ mất.

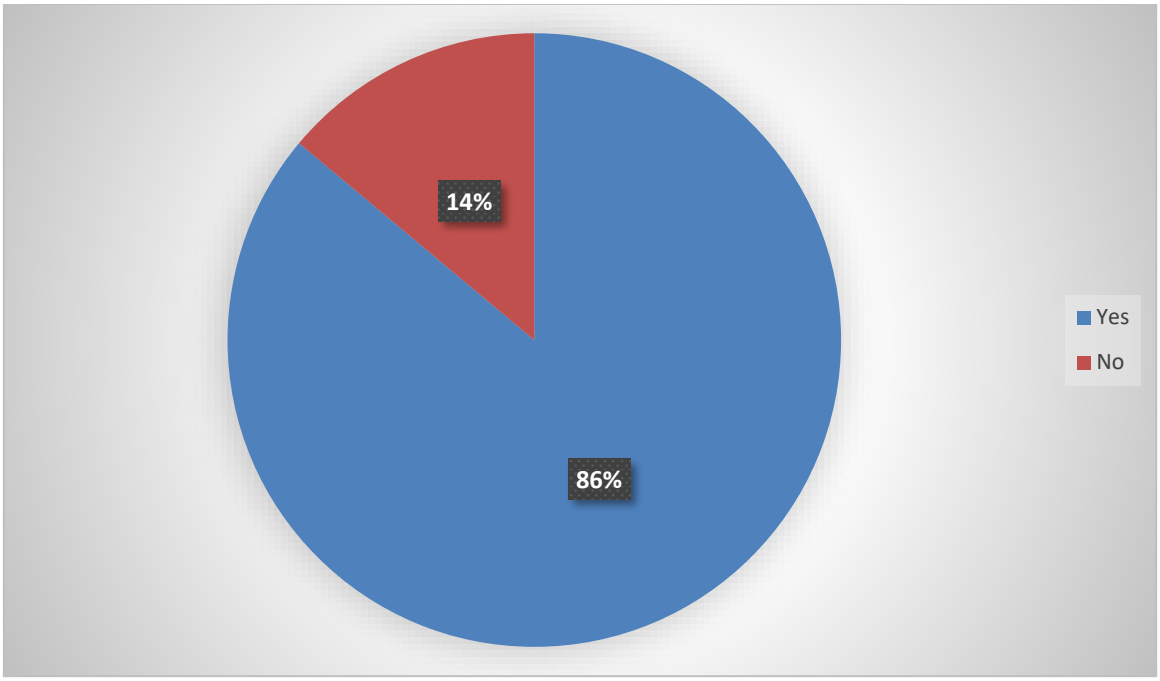
Ở bệnh viện gồm 6495 số ca mắc được điều trị và số ca tử vong là 1667 chiếm tỉ lệ 25,66%.

4.5.4. Bệnh tim mạch và bệnh béo phì (Cardiovascular,obesity)

Béo phì là một căn bệnh mãn tính, xảy ra do sự gia tăng quá mức chất béo trong cơ thể, làm thúc đẩy và gây rối loạn chức năng mô mỡ.

Béo phì tác động trên bệnh mạch vành thông qua cơ chế xơ vữa động mạch. Một nghiên cứu dịch tễ chứng minh rằng béo phì có liên quan nguy cơ mắc bệnh động mạch vành cao hơn. Ở mỗi mức BMI, chỉ số mỡ trung tâm, bao gồm vòng eo và tỷ lệ eo hông, cao hơn liên quan nguy cơ bệnh mạch vành và tử vong tim mạch cao hơn, kể cả người có cân nặng bình thường theo BMI. Mối liên quan giữa béo phì và bệnh mạch vành theo các nghiên cứu lớn do tăng huyết áp, rối loạn lipid máu, đái tháo đường và các bệnh đi kèm khác.[9]

Trong tệp dữ liệu về covid 19 để thể hiện mối liên hệ giữa bệnh béo phì và tim mạch sau quá trình lọc dữ liệu có 162 bệnh nhân nhiễm covid 19 đều mang cả 2 bệnh béo phì và tim mạch.



Hình 4.5.4-1: Tỉ lệ phần trăm người mất do dịch covid của bệnh nhân mang 2 bệnh lý tim mạch và béo phì

Có thể thấy tỉ lệ phần trăm số người mất do bị nhiễm covid và mang 2 căn bệnh béo phì và tim mạch là rất cao.Chiếm đến 86% tỉ lệ mất.

Chương 5. **Tổng kết**

5.1. Kết quả đạt được

- Đồ án môn học về khai phá và phân tích dữ liệu Covid 19 được thể hiện các nội dung cơ bản về khai phá dữ liệu,thuật toán chính được sử dụng trong đề tài,phân tích các thuộc tính của dữ liệu, sự so sánh giữa các thuật toán trong quá trình triển khai mô hình khai phá dữ liệu,đưa ra các nhận xét về các kết quả trong việc huấn luyện mô hình các thuật toán,hiển thị rõ một số thuộc tính dữ liệu quan trọng.Đưa ra các so sánh về các thông số ảnh hưởng đến thuật toán chính được sử dụng trong đề tài là Neural Network.

- Việc em đưa ra một vài thuộc tính mang tính biểu trưng về tỉ lệ tử vong ở độ tuổi ,đơn vị y tế điều trị, mối liên hệ giữa thuốc lá và bệnh tắc nghẽn phổi mãn tính.Tính riêng các tỉ lệ chỉ thể hiện các yếu tố chủ quan gây ra cái chết.Việc áp dụng thuật toán trong quá trình huấn luyện không chỉ là việc so sánh một vài yếu tố chủ quan đó mà đó là quá trình máy học và đánh giá kết quả từ các thuộc tính khác của tệp dữ liệu để đưa ra kết quả cuối cùng.Việc này giúp cho việc chuẩn đoán về mức độ nghiêm trọng khi mắc phải covid 19 được chuẩn xác hơn để các y bác sĩ có thể đưa ra các phác đồ ,phương pháp điều trị phù hợp.

Nhờ có việc áp dụng khai phá và phân tích dữ liệu giúp cho mọi chuẩn đoán về tình trạng bệnh trở nên nhanh chóng và có tỉ lệ chính xác cao hơn.

5.2. Các hạn chế,cần tìm tòi và nghiên cứu thêm.

-Chủ đề về khai phá và phân tích dữ liệu về Covid 19 liên quan đến vấn đề y học vì vậy gặp khó khăn trong việc tìm hiểu về các thuộc tính trong tệp dữ liệu,mối tương quan giữa các thuộc tính đó.

-Việc tổng quát về nội dung của khai phá dữ liệu nói chung hay thuật toán chính là Neural Network nói riêng còn có nhiều thiếu sót.

-Việc xây dựng chương trình từ các mô hình thực nghiệm có thể chuẩn đoán khi người sử dụng chỉ cần nhập các chỉ số liên quan đến các thuộc tính của tệp dữ liệu chương trình sẽ đưa ra chuẩn đoán.

5.3. Bài học và kinh nghiệm rút ra sau khi thực hiện đồ án

-Việc khai phá và phân tích dữ liệu là một bước trong quá trình khai phá tri thức mới vì vậy cần có sự đầu tư về thời gian cho việc phân tích các thuộc tính của dữ liệu, phải thực hiện lặp đi lặp lại rất nhiều lần từ việc thu thập dữ liệu,tiền xử lý dữ liệu, không phải chọn một tệp tài liệu nào cũng có thể đem vào khai phá và phân tích.Đó là cả một quy trình dài.

-Việc khai phá một tệp dữ liệu lớn cần phải có sự kiên nhẫn,tỉ mỉ trong việc lọc dữ liệu làm sao cho hợp lý,rời rạc hóa dữ liệu đó,phân tích một vài thuộc tính đặc trưng của dữ liệu.

-Thực hiện quá trình huấn luyện với nhiều mô hình thuật toán khác nhau nhằm đưa ra được những nhận thức về tính năng của từng thuật toán mang lại cho bản thân.

TÀI LIỆU THAM KHẢO

[1] D.Thu, "Dịch Covid-19 hôm nay:Ca mắc và bệnh nhân nặng lại tăng," 09/10/2023. [Trực tuyến].Địa chỉ: https://nld.com.vn/suc-khoe/dich-covid-19-hom-nay-ca-mac-va-benh-nhan-nang-lai-tang-20231009161407411.htm [Truy cập 09/10/2023]

[2]T.Giang, "Những nguyên nhân dẫn tới tỷ lệ tử vong do Covid 19 cao ở Việt Nam," 30/12/2021 .[Trực tuyến].Địa chỉ: https://www.vietnamplus.vn/nhung-nguyen-nhan-dan-toi-ty-le-tu-vong-do-covid19-cao-o-viet-nam/765740.vnp#lnk43vx2zhk1uq3t6yl [Truy cập 10/10/2023].

[3] M.J.Zaki and Wagner.M.JR, "Data Mining and Analysis",*Data Mining and Analysis Fundamental Concepts and Algorithms*. United States of America: Cambridge University Press,2014,pp. 25-26.

[4] N.S.Gill ,"Artificial Neural Networks Applications and Algorithms," 19 June 2023.[Online]. Availabel: https://www.xenonstack.com/blog/artificial-neural-network-applications [Accessed 10/11/2023].

[5] KyoHB,"Mạng thần kinh nhân tạo-Artificial Neural Network,"7/3/2021.[Trực tuyến].Địa chỉ : https://aiwithmisa.com/2021/03/07/aml-bai15/ [Truy cập 11/10/2023]

[6] Charu C.Aggarwal,"Neural Networks",*Data Mining* .New York:Springer International Publishing Switzerland ,2015,pp.326-330.

[7] Ian H. Witten, Eibe Frank, Mark A. Hall (2011), Data Mining Practical Machine Learning Tools and Techniques.

[8] Eibe Frank, Mark A. Hall, and Ian H. Witten(2016), The WEKA Workbench.

[9] Bs. Nguyễn Hoàng Minh Phương – Bệnh viện Tim mạch An Giang,"BÉO PHÌ VÀ BỆNH TIM MẠCH,"18/9/2023.[Trực tuyến].Địa chỉ : https://cdcangiang.vn/index.php/2023/09/18/beo-phi-va-benh-tim-mach/ [Truy cập 15/10/2023]