TRƯỜNG ĐẠI HỌC NGUYỄN TẤT THÀNH
KHOA CÔNG NGHỆ THÔNG TIN

ĐỒ ÁN MÔN HỌC: HỌC MÁY VÀ ỨNG DỤNG

ỨNG DỤNG THUẬT TOÁN MACHINE LEARNING TRONG GIẢI QUYẾT BÀI TOÁN CHUẨN ĐOÁN NGUY CƠ ĐỘT QUỴ

Giảng viên hướng dẫn : TS. VÕ THỊ HỒNG THẮM

Sinh viên thực hiện : NGUYỄN BÉ LAM

MSSV : 2000001437

Lớp : 20DTH1A

Chuyên ngành : TRÍ TUỆ NHÂN TẠO

Khoá : 2020

Tp HCM, tháng 08 năm 2023

TRƯỜNG ĐẠI HỌC NGUYỄN TẤT THÀNH

KHOA CÔNG NGHỆ THÔNG TIN

ĐỒ ÁN MÔN HỌC: HỌC MÁY VÀ ỨNG DỤNG

ỨNG DỤNG THUẬT TOÁN MACHINE LEARNING TRONG GIẢI QUYẾT BÀI TOÁN CHUẨN ĐOÁN NGUY CƠ ĐỘT QUỴ

Giảng viên hướng dẫn	**:**	**TS. VÕ THỊ HỒNG THẮM**
Sinh viên thực hiện	**:**	**NGUYỄN BÉ LAM**
MSSV	**:**	**2000001437**
Lớp	**:**	**20DTH1A**
Chuyên ngành	**:**	**TRÍ TUỆ NHÂN TẠO**
Khoá	**:**	**2020**

TPHCM, tháng 08 năm 2023

LỜI MỞ ĐẦU

Trong thời buổi công nghệ trở thành một điều không thể thiếu đối với đời sống của con người, ngành Công Nghệ Thông Tin là một ngành có tầm ảnh hưởng vô cùng lớn đối với tất cả mọi lĩnh vực trong xã hội, mang đến một lượng lớn tiện ích đến cho mọi người truy cập. Về mặt nguồn nhân lực, nó đòi hỏi một nguồn lớn nhân lực với chất lượng và kỹ năng cao, không những thế còn đòi hỏi nguồn nhân lực phải vững kiến thức, dày kinh nghiệm, khả năng làm việc nhóm, sự sáng tạo và cập nhật thông tin mới thường xuyên trong lúc làm việc. Với số lượng tiện ích khổng lồ mà ngành Công Nghệ Thông Tin đem lại, các máy móc trở nên thông minh và trở thành một phần không thể thiếu của các ngành nghề, đặc biệt là y tế.

Ở trường đại học Nguyễn Tất Thành, trong ngành Công nghệ thông tin, mọi môn học đều không thể thiếu. Nhưng với niềm mong muốn có thể ứng dụng sự thông minh hữu ích cho mọi người nhất là đối với lĩnh vực hỗ trợ y tế, các bác sĩ. Vì thế, đồ án của tôi có nội dung là "Ứng dụng thuật toán Machine Learning trong giải quyết bài toán chuẩn đoán nguy cơ đột quỵ" do tôi thực hiện.

Thông qua bài báo cáo này, tôi hi vọng sẽ tổng hợp lại được kiến thức đã học trong suốt quá trình học tập. Đồng thời luyện tập để thiết kế, xây dựng trang web tiện lợi thông minh, đa dạng, đáp ứng nhu cầu của xã hội. Do khả năng, thời gian cũng như kiến thức còn nhiều hạn chế, những sai sót là không thể trách khỏi. Cho nên em rất mong được sự chỉ dạy và góp ý từ thầy cô để có thể hoàn thiện hơn trong những đồ án tiếp theo.

LỜI CẢM ƠN

Lời nói đầu xin gửi lời cảm ơn chân thành đến với gia đình tôi vì một nguồn động lực lớn đến từ họ, cảm ơn các bạn bè đã giúp đỡ trong quá trình học tập, đồng thời xin chân thành cảm ơn các Giảng Viên - Khoa Công Nghệ Thông Tin, trường Đại Học Nguyễn Tất Thành vì đã quan tâm, giúp đỡ, tạo nhiều điều kiện cho tôi trong suốt quá trình học tập.

Em xin gửi lời cảm ơn chân thành của mình tới cô Võ Thị Hồng Thắm đã rất nhiệt tình trong việc hướng dẫn trong quá trình học tập và hoàn thành đồ án để có những định hướng tốt nhất trong việc thực hiện đồ án.

Tuy nhiên do còn chưa trao dồi nhiều về mặt kinh nghiệm và cũng như kiến thức cho nên dẫn đến một vài thiếu sót cũng như những vấn đề không như ý muốn, mong nhận được sự thông cảm cũng như nhận được thêm ý kiến cũng như kiến thức đến từ cô.

Em xin trân trọng cảm ơn!

Sinh viên thực hiện

Nguyễn Bé Lam

BM-ChT-

TRƯỜNG ĐẠI HỌC NGUYỄN TẤT THÀNH
TRUNG TÂM KHẢO THÍ

KỲ THI KẾT THÚC HỌC PHẦN
HỌC KỲ 3 NĂM HỌC 2022 - 2023

PHIẾU CHẤM THI TIỂU LUẬN/ĐỒ ÁN

Môn thi: Học máy và ứng dụng Lớp học phần: 20DTH1A

Sinh viên thực hiện : Nguyễn Bé Lam

Ngày thi: 29/08/2023 Phòng thi: L.611

Đề tài tiểu luận/báo cáo của sinh viên : Ứng dụng thuật toán Machine Learning trong giải quyết bài toán chuẩn đoán nguy cơ đột quỵ

Phần đánh giá của giảng viên (căn cứ trên thang rubrics của môn học):

Tiêu chí (theo CĐR HP)	Đánh giá của GV	Điểm tối đa	Điểm đạt được
Cấu trúc của báo cáo			
Nội dung			
Các nội dung thành phần			
Lập luận			
Kết luận			
Trình bày			
TỔNG ĐIỂM			

Giảng viên chấm thi

(ký, ghi rõ họ tên)

VÕ THỊ HỒNG THẮM

NHẬN XÉT CỦA GIẢNG VIÊN HƯỚNG DẪN

.....

.....

.....

.....

.....

.....

.....

.....

.....

.....

.....

.....

.....

.....

.....

.....

Tp.HCM, ngày 29 tháng 08 năm 2023

GVHD

VÕ THỊ HỒNG THẮM

MỤC LỤC

CHƯƠNG I. TỔNG QUAN ĐỀ TÀI

1. Giới thiệu

Bệnh đột quỵ, còn được gọi là tai biến mạch máu não, là một vấn đề y tế nghiêm trọng trên toàn thế giới. Đột quỵ có thể gây ra những hậu quả nghiêm trọng như tàn phế hoặc thậm chí tử vong. Do đó, việc phát hiện và dự đoán nguy cơ đột quỵ trở nên quan trọng để có thể thực hiện các biện pháp phòng ngừa và điều trị kịp thời.

Trong lĩnh vực y học, máy học đã trở thành một công cụ hữu ích để phân loại và dự đoán nguy cơ bệnh. Các thuật toán máy học như Naive Bayes, K-Nearest Neighbors (KNN), Logistic Regression, Support Vector Machine (SVM) và Random Forest đã được áp dụng rộng rãi để dự đoán nguy cơ đột quỵ.

2. Lý do chọn đề tài

Đột quỵ và các vấn đề liên quan đến tim mạch ảnh hưởng mạnh mẽ đến sức khỏe và chất lượng cuộc sống của hàng triệu người trên khắp thế giới. Việc phát triển các phương pháp dự đoán tiềm năng có thể hỗ trợ trong việc chẩn đoán sớm, đưa ra quyết định về điều trị và giảm thiểu nguy cơ mắc bệnh.

Lĩnh vực học máy đang ngày càng thể hiện khả năng ứng dụng rộng rãi trong y học. Sự kết hợp giữa kiến thức y học và công nghệ học máy có thể dẫn đến các giải pháp đột phá trong việc dự đoán, phân loại và đánh giá các vấn đề tim mạch.

Việc áp dụng phương pháp dự đoán có thể giúp tối ưu hóa quy trình y tế, giảm thiểu thời gian và nguồn lực cần thiết cho việc chẩn đoán và điều trị. Điều này có thể dẫn đến việc nâng cao hiệu quả và tiết kiệm chi phí trong lĩnh vực y tế.

3. Mục tiêu nghiên cứu

Mục tiêu của đề tài này là so sánh hiệu suất của các thuật toán máy học Naive Bayes, KNN, Logistic Regression, SVM và Random Forest trong việc dự đoán nguy cơ đột quỵ. Chúng ta sẽ xem xét cách mỗi thuật toán hoạt động, cách chúng tương tác với dữ liệu y tế và khả năng dự đoán đúng bệnh nhân có nguy cơ đột quỵ. Kết quả từ cuộc so sánh này sẽ giúp xác định thuật toán nào có hiệu suất tốt nhất để áp dụng trong việc dự đoán nguy cơ đột quỵ.

4. Đối tượng nghiên cứu

Đối tượng nghiên cứu:

- Tuổi: mọi độ tuổi; giới tính: Nam, nữ
- Trong mọi ngành nghề
- Nông thôn và thành thị

CHƯƠNG II. CƠ SỞ LÝ THUYẾT

1. CƠ SỞ LÝ THUYẾT

THUẬT TOÁN SỬ DỤNG TRONG MÔ HÌNH

1.1 Guassian Naïve Bayes

Gaussian Naive Bayes được thiết kế đặc biệt cho các biến liên tục và dữ liệu theo phân phối Gaussian (phân phối chuẩn). Trong dữ liệu y học liên quan đến đột quỵ, chúng ta thường gặp các biến như huyết áp,.. mà thường có phân phối gần với phân phối chuẩn. Vì vậy, Gaussian Naive Bayes có thể phù hợp để mô hình hóa sự tương quan giữa các biến này.

Mô hình này có thể phù hợp bằng cách đơn giản là tìm giá trị trung bình và độ lệch chuẩn của các điểm trong mỗi nhãn.

Guassian Naïve Bayes là dựa trên:

Phân phối Gaussian (Phân phối chuẩn): là một phân phối xác suất quan trọng trong thống kê. Nó được mô tả bởi hai tham số chính: trung bình (mean) và độ lệch chuẩn (standard deviation). Hàm mật độ xác suất của phân phối Gaussian có dạng:

Công thức:

$$p(x_i|c) = p(x_i|\mu_{ci}, {\sigma^2}_{ci}) = \frac{1}{\sqrt{2\pi {\sigma^2}_{ci}}} \exp\left(-\frac{(x_i - \mu_{ci})^2}{2{\sigma^2}_{ci}}\right)$$

P(x) là xác suất xảy ra giá trị x.

μ là trung bình của phân phối.

σ là độ lệch chuẩn của phân phối.

Để sử dụng Gaussian Naive Bayes, chúng ta cần ước tính các tham số như xác suất tiền nghiệm của lớp và các tham số phân phối Gaussian (trung bình và độ lệch chuẩn) cho từng biến đặc trưng dưới mỗi lớp. Các tham số này có thể được ước tính từ dữ liệu huấn luyện.

Ưu điểm và hạn chế:

Ưu điểm của Gaussian Naive Bayes bao gồm tính đơn giản, khả năng xử lý dữ liệu liên quan đến phân phối Gaussian và khả năng làm việc với số lượng biến đặc trưng lớn. Tuy nhiên, giả định "naive" có thể không phù hợp trong mọi tình huống, và mô hình có thể bị ảnh hưởng nếu các biến đặc trưng thực sự có mối tương quan đáng kể.

Ứng dụng:

Gaussian Naive Bayes thường được sử dụng trong các bài toán phân loại, dự đoán và phân tích dữ liệu trong nhiều lĩnh vực như y học, tài chính, xử lý ngôn ngữ tự nhiên và nhiều lĩnh vực khác.

1.2 K-nearest neighbors

Thuật toán K-Nearest Neighbors (KNN) là một phương pháp học máy sử dụng cơ chế so sánh các điểm dữ liệu gần nhất để dự đoán lớp hoặc giá trị của một điểm dữ liệu mới. Ý tưởng cơ bản của KNN là dựa vào việc rằng các điểm dữ liệu cùng lớp thường có sự tương đồng về mặt không gian.

Cách hoạt động của KNN:

Lựa chọn số hàng xóm (K): Trước tiên, bạn cần xác định số lượng hàng xóm (K) mà bạn muốn sử dụng để dự đoán. K là một số nguyên dương.

Tính khoảng cách: Để tìm các điểm hàng xóm gần nhất, KNN tính khoảng cách giữa điểm dữ liệu mới và tất cả các điểm dữ liệu trong tập huấn luyện. Các phương pháp tính khoảng cách có thể là Euclidean distance, Manhattan distance, hoặc các phương pháp khác tùy thuộc vào bài toán cụ thể.

Lựa chọn hàng xóm: Chọn K điểm dữ liệu gần nhất (có khoảng cách nhỏ nhất) từ tập huấn luyện.

Phân loại hoặc dự đoán: Để phân loại một điểm dữ liệu mới, KNN đếm số lượng điểm hàng xóm thuộc mỗi lớp và chọn lớp có số lượng điểm nhiều nhất trong K điểm gần nhất. Trong trường hợp dự đoán giá trị (regression), KNN có thể trung bình hoặc là giá trị trung vị của các giá trị hàng xóm.

Ưu điểm:

Dễ hiểu và triển khai, độ phức tạp thời gian huấn luyện thấp.

Hạn chế:

Nhạy cảm với nhiễu và dữ liệu nhiều chiều, yêu cầu lưu trữ toàn bộ dữ liệu huấn luyện, hiệu suất giảm khi dữ liệu lớn.

Thuật toán KNN cho rằng những dữ liệu tương tự nhau sẽ tồn tại gần nhau trong một không gian, từ đó công việc của chúng ta là sẽ tìm k điểm gần với dữ liệu cần kiểm tra nhất.

Đối với các bài toán phân loại, nhãn lớp được chỉ định trên cơ sở đa số phiếu bầu — tức là nhãn được đại diện thường xuyên nhất xung quanh một điểm dữ liệu nhất định được sử dụng. Mặc dù điều này về mặt kỹ thuật được coi là "biểu quyết đa số".

1.3 SVM (Support Vector Machine)

Support Vector Machine - SVM là một thuật toán học máy được sử dụng cho các tác vụ phân loại và hồi quy. Nó là một trong những thuật toán phân loại mạnh mẽ và được ứng dụng rộng rãi trong nhiều lĩnh vực.

SVM là tìm một siêu phẳng (hyperplane) trong không gian nhiều chiều để phân chia các điểm dữ liệu của hai lớp khác nhau sao cho khoảng cách giữa các điểm gần nhất với siêu phẳng là lớn nhất. Các điểm dữ liệu gần nhất với siêu phẳng được gọi là "vector hỗ trợ". SVM cố gắng tối đa hóa khoảng cách này (được gọi là biên) để đảm bảo tính phân loại tốt và khả năng tổng quát hóa cao hơn.

Ứng dụng: Phân loại văn bản, Nhận dạng hình ảnh, Phân loại bệnh, dự đoán kết quả điều trị. Dự đoán thị trường chứng khoán, phân loại khách hang.

Ưu điểm:

Hiệu quả trong không gian chiều cao: SVM hoạt động tốt trong các không gian chiều cao, ngay cả khi số lượng chiều dữ liệu lớn.

Tính tổng quát hóa tốt: SVM cố gắng tối đa hóa khoảng cách giữa các lớp, giúp tránh tình trạng overfitting (quá khớp) dữ liệu huấn luyện.

Hỗ trợ vector: SVM chỉ quan tâm đến một số điểm dữ liệu quan trọng gần siêu phẳng, giúp tối ưu hóa hiệu năng và sử dụng bộ nhớ hiệu quả.

Nhược điểm:

Khó khăn trong việc xử lý dữ liệu lớn: Trong trường hợp dữ liệu lớn, SVM có thể tốn nhiều thời gian và tài nguyên tính toán.

Không trực tiếp áp dụng cho đa lớp: SVM thường được xây dựng cho bài toán phân loại hai lớp, và việc mở rộng cho nhiều lớp có thể đòi hỏi các kỹ thuật phức tạp hơn.

1.4 Random Forest

Random Forest là một thuật toán học máy dựa trên việc xây dựng một tập hợp các cây quyết định độc lập và kết hợp kết quả từ các cây này để đưa ra dự đoán cuối cùng.

Chọn mẫu ngẫu nhiên lặp lại: Đầu tiên, một tập hợp con dữ liệu được tạo ra từ tập dữ liệu gốc bằng cách chọn ngẫu nhiên các mẫu với việc lặp lại (chọn mẫu có thể lặp lại) hoặc không lặp lại (chọn mẫu không thể lặp lại).

Xây dựng cây quyết định: Với mỗi tập con dữ liệu, một cây quyết định được xây dựng bằng cách chia dữ liệu thành các nút con dựa trên các thuộc tính. Quá trình này tiếp tục cho đến khi cây đạt độ sâu tối đa hoặc không thể tiếp tục chia.

Kết hợp kết quả: Khi cần thực hiện dự đoán cho một dữ liệu mới, các cây trong tập hợp sẽ được sử dụng để đưa ra dự đoán riêng lẻ. Kết quả sau đó có thể được kết hợp (để đưa ra dự đoán cuối cùng.

Ứng dụng:

Phân loại, Hồi quy: Dự đoán giá nhà, dự đoán doanh số bán hàng

Nhận dạng đối tượng: Nhận dạng đối tượng trong hình ảnh hoặc video.

Khai phá dữ liệu: Phân tích dữ liệu để khám phá mô hình hoặc mối quan hệ.

Ưu điểm:

Tính đa dạng: Random Forest giúp tránh tình trạng quá khớp bằng cách kết hợp nhiều cây quyết định, mỗi cây là một mô hình khác nhau.

Khả năng xử lý dữ liệu lớn: Random Forest có thể xử lý tập dữ liệu lớn mà không cần nhiều tiền xử lý dữ liệu trước.

Tự động xử lý thuộc tính thiếu hoặc nhiễu: Nó có khả năng làm việc tốt với các tập dữ liệu có nhiễu hoặc thiếu dữ liệu.

Nhược điểm: Random Forest tạo ra nhiều cây, vì vậy có thể tạo ra một lượng lớn các mô hình phụ thuộc vào cấu hình.

1.5 Logistic Regression

Hồi quy Logistic (Logistic Regression) là một thuật toán học máy được sử dụng chủ yếu cho các tác vụ phân loại. Hồi quy logistic thực hiện phân loại dựa trên xác suất.

Công thức của hàm logistic là:

$$\sigma(z)= \frac{1}{1 + e^{-z}}$$

Trong đó:

z là biểu thức tuyến tính của đặc trưng đầu vào, có thể được tính bằng cộng dồn của trọng số của từng đặc trưng nhân với giá trị tương ứng của đặc trưng.

Khi có xác suất, ngưỡng có thể được áp dụng để quyết định phân loại thành hai lớp: một lớp với xác suất lớn hơn ngưỡng và một lớp với xác suất nhỏ hơn hoặc bằng ngưỡng.

Ứng dụng:

Dự đoán xác suất bệnh tật, phân loại email, phân loại hình ảnh: Phân loại hình ảnh của các đối tượng, như phân loại loại động vật dựa trên hình ảnh.

Ưu điểm:

Dễ hiểu và triển khai

Tính tương thích: Nó hoạt động tốt với các bài toán phân loại nhị phân (2 lớp).

Nhược điểm:

Giới hạn trong phân loại đa lớp: Hồi quy logistic ban đầu được thiết kế cho phân loại nhị phân, và mở rộng lên phân loại đa lớp có thể đòi hỏi kỹ thuật phức tạp hơn.

Dễ bị ảnh hưởng bởi nhiễu: Hồi quy logistic có thể bị ảnh hưởng bởi dữ liệu nhiễu hoặc outliers.

2. Bài báo nghiên cứu

Data Science and Management

Bài báo này trình bày một cái nhìn tổng quan về lĩnh vực học máy (Machine Learning), một lĩnh vực đang phát triển mạnh mẽ. Tiến bộ gần đây trong ML đã được thúc đẩy cả bởi việc phát triển lý thuyết về các thuật toán học mới, cũng như bởi sự bùng nổ liên tục về lượng lớn dữ liệu.

Bài báo cũng nói về hướng nghiên cứu và tiềm năng trong lĩnh vực này, một số lĩnh vực phổ biến như y tế, tài chính, hệ thống bảo mật, nông nghiệp, quản lí dữ liệu.

Giới thiệu:

Bài này nhấn mạnh vai trò của dữ liệu xung quanh chúng ta, là một phần thiết yếu trong cuộc sống hằng ngày. Dữ liệu tạo ra ứng dụng thông minh trong nhiều lĩnh vực và hỗ trợ rất nhiều cho đời sống con người.

Sự phát triển của trí tuệ nhân tạo và học máy chính là công cụ để phân tích thông tin dữ liệu, phát triển các úng dụng thực tế. Máy học có thể giúp giải quyết xử lý và ứng dụng kho dữ liệu lớn hiện nay. Ứng dụng của ML có ảnh hưởng lớn trong các lĩnh vực khác nhau, như hỗ trợ các nhà khoa học phát hiện và phân loại ung thư dựa trên phân tích mảng ADN, giải quyết vấn đề xác định cấu trúc ba chiều của protein từ dãy axit amin, và thậm chí trong việc dự đoán chẩn đoán SARS-CoV-2 dựa trên triệu chứng của bệnh và tìm kiếm thuốc và vắc-xin tiềm năng dựa trên mô phỏng máy tính.

Bài viết nhấn mạnh về sự cần thiết của các thuật toán học máy hiệu quả và phải dáng tin cậy, đề cập đến những thách thức trong việc bảo vệ quyền riêng tư khi xử lý dữ liệu. Đồng thời, sự cần thiết của tính minh bạch và khả năng của các thuật toán học máy, đặc biệt khi chúng xử lý thông tin cá nhân.

Phân loại học máy: Học có giám sát, học không giám sát, học bán giám sát, học tăng cường.

Học có giám sát: Là các thuật toán học máy mà dữ liệu có đầu vào và nhãn của đầu ra. Quá trình học này dựa vào việc so sánh đầu ra tính toán và đầu ra dự đoán, việc học liên quan đến tính toán lỗi và điều chỉnh lỗi để đạt được đầu ra mong đợi. ví dụ về các thuật

toán như Naïve Bayes classification, hồi quy tuyến tính và hồi quy logistic, SVM. Học có giám sát được ứng dụng như nhận dạng khuôn mặt, gắn nhãn, phân loại.

Học không giám sát: Là phân tích và học các thông tin dữ liệu nhưng không được gán nhãn đầu ra. Học không giám sát sẽ phân tích các tập dữ liệu và tách các mẫu thành các cụm (các lớp) khác nhau dựa trên đặc điểm của của dữ liệu. Một số thuật toán như là K-mean,..

Học bán giám sát: có thể xem là sự kết hợp giữa học có giám sát và học không giám sát, vì nó sử dụng được dữ liệu gán nhãn và không gán nhãn. Mục tiêu của nó là mang lại kết quả tốt hơn. Nó được sử dụng rộng rãi trong dịch máy, ghi nhãn dữ liệu và phân loại văn bản.

Học tăng cường: là thuật toán hoạt động tuần tự để tự động đánh giá hành vi tối ưu trong môi trường cụ thể nhằm cải thiện hiệu quả của nó.

Học liên kết: Google đã đề xuất khái niệm này vào năm 2016. Ý tưởng của nó là xây dựng mô hình machine learning dựa trên các tập dữ liệu trên nhiều thiết bị và tránh rò rỉ dữ liệu. Công việc của học liên kết là học dữ liệu và cho phép huấn luyện các mô hình cục bộ không đồng nhất để giải quyết các vấn đề. Trong tương lai gần, nó sẽ phá vỡ các rào cản giữa các ngành và thiết lập một cộng đồng nơi dữ liệu và kiến thức được chia sẽ an toàn.

- ⇨ Sự thành công của Machine Learning có thể là do sự cần thiết của rất nhiều lĩnh vực như y học, sản xuất, giá dục, tài chính, nông nghiệp,.. cũng như sự bùng nổ của dữ liệu lớn hiện nay.

Ứng dụng trong các lĩnh vực:

Y học: Sử dụng để hỗ trợ bác sĩ dự đoán, quản lí bệnh, can thiệp phẫu thuật, dự đoán nguy cơ dịch bệnh hay phát hiện ra thuốc mới.

An ninh: AI và ML đang được ứng dụng trong việc nâng cao hiệu suất hoạt động của các cơ quan chính phủ, tạo ra sự dự đoán về các sự kiện quan trọng, và cải thiện an ninh mạng.

Tài chính: giúp các tổ chức làm việc hiệu quả, giúp đưa ra quyết định tài chính, quản lí chi phí.

Công nghệ nano: ML được sử dụng để thiết kế và dự đoán tính chất của các vật liệu nano, giúp tạo ra những sản phẩm mới và tiên tiến hơn.

Nông nghiệp: ML giúp cải thiện quản lý nông trại, dự đoán năng suất cây trồng, phát hiện bệnh và cỏ dại, quản lý đàn gia súc và chăm sóc sức khỏe, quản lý nguồn nước và đất đai.

Thách thức pháp lý:

Sự tăng trưởng nhanh chóng của ML đặt ra các thách thức pháp lý, bao gồm vấn đề trách nhiệm và bảo mật thông tin.

Sự không rõ ràng và không thể đoán trước trong hành vi của AI và ML làm cho việc xác định người chịu trách nhiệm trở nên khó khăn.

Vấn đề trách nhiệm pháp lý cũng cần được giải quyết cẩn thận vì thực tế là các công cụ ML và AI có thể được áp dụng liên quan đến các hoạt động có rủi ro cao, ví dụ: xe tự lái.

Các quốc gia đang cố gắng thiết lập quy định pháp lý để giải quyết vấn đề trách nhiệm và đảm bảo an toàn cho người dùng cuối.

Kết luận:

Sau khi đọc bài báo này , em có cái nhìn tổng quan hơn về học máy và phân tích dữ liệu. Các loại học máy và cách sử dụng các trong nhiều trường hợp khác nhau để giải quyết vấn đề. Kết quả và hiệu suất của Machine learning phụ thuộc vào dữ liệu mà nó được học tập. Thấy được sự quan tâm của các quốc gia về nhiều rủi ro và trách nhiệm khi ứng dụng và sử dụng máy học thông minh.

CHƯƠNG III. XÂY DỰNG MÔ HÌNH

1. Tóm tắt đề tài

Đề tài này tập trung vào việc xử lý dữ liệu đầu vào và so sánh các mô hình thuật toán và lựa chọn mô hình có kết quả tốt nhất để dự đoán nguy cơ đột quỵ. Bệnh đột quỵ được coi là một vấn đề y tế nghiêm trọng, vì vậy khả năng dự đoán nguy cơ sẽ giúp phát hiện sớm và thực hiện biện pháp phòng ngừa cũng như điều trị kịp thời. Các thuật toán sẽ được so sánh để xem xét hiệu suất của chúng trong việc dự đoán bệnh đột quỵ.

Giới thiệu về Naive Bayes, K-Nearest Neighbors, Logistic Regression, Support Vector Machine và Random Forest, cung cấp cái nhìn tổng quan về cách chúng hoạt động trong việc dự đoán.

Tiền xử lý dữ liệu: Mô tả về việc thu thập và tiền xử lý dữ liệu y tế để sử dụng cho huấn luyện và kiểm tra mô hình.

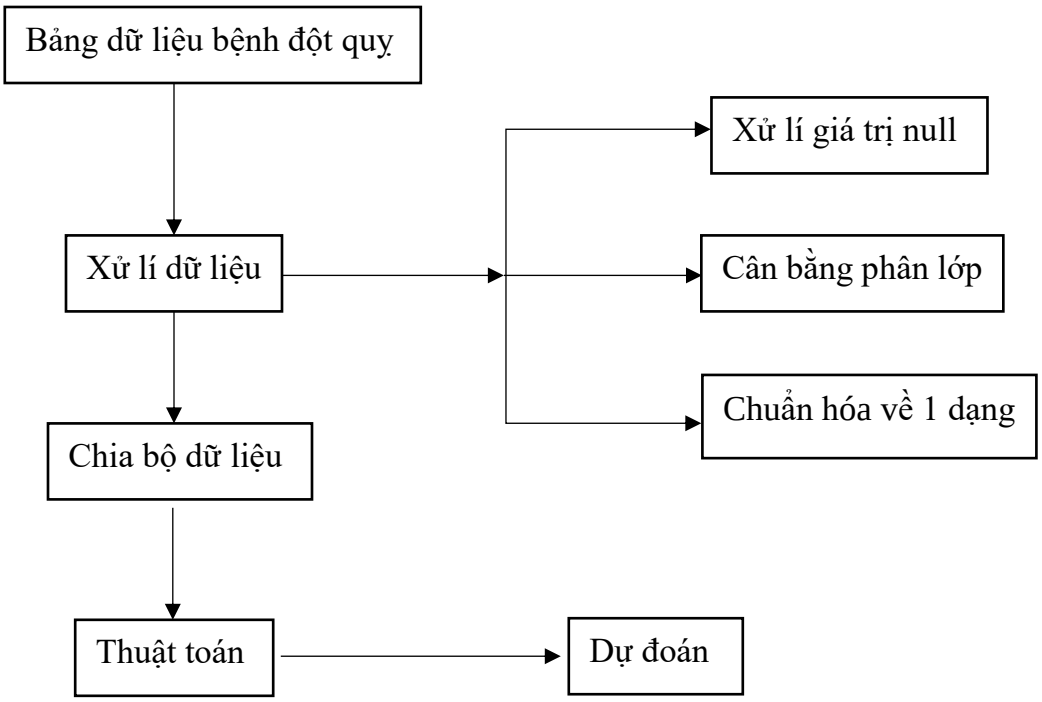
Huấn luyện và đánh giá: Thực hiện quá trình huấn luyện các mô hình trên tập dữ liệu, sau đó đánh giá chất lượng của chúng bằng cách sử dụng các chỉ số như accuracy

So sánh hiệu suất: Đưa ra so sánh về hiệu suất của các thuật toán, xác định thuật toán nào cho kết quả tốt nhất trong việc dự đoán nguy cơ đột quỵ.

Kết luận và triển vọng: Tổng kết kết quả so sánh và đề xuất hướng phát triển tiếp theo, để cải thiện khả năng dự đoán và ứng dụng thực tế trong lĩnh vực y tế.

Đề tài này hướng đến việc cung cấp cái nhìn rõ ràng về khả năng của các thuật toán máy học trong việc dự đoán nguy cơ đột quỵ, với mong muốn cung cấp thông tin hữu ích cho việc cải thiện quản lý sức khỏe và chăm sóc y tế.

2. Mô hình



3. Các bước xây dựng mô hình

Bước 1: Thu thập dữ liệu, xử lí giá trị thiếu, chuẩn hóa về 1 dạng, cân bằng nhãn.

Bước 2: Chuẩn bị dữ liệu bằng phương pháp K-Fold (K-Fold Cross-Validation)

Bước 3: Chuẩn hóa dữ liệu bằng phương pháp StandardScaler và đưa vào mô hình huấn luyện

Bước 4: Xây dựng các mô hình Naive Bayes, KNN, Logistic Regression, SVM, Random Forest.

Bước 5: Huấn luyện và đánh giá Accuracy.

Bước 6: Lựa chọn mô hình dự đoán

CHƯƠNG 4. ĐÁNH GIÁ VÀ THỰC NGHIỆM

1. Chuẩn bị dữ liệu

Sau khi có bảng dữ liệu về nguy cơ đột quỵ. Em xác định các giá trị đầu vào và biến đầu ra mục tiêu.

	id	gender	age	hypertension	heart_disease	ever_married	work_type	Residence_type	avg_glucose_level	bmi	smoking_status	stroke
0	9046	Male	67.0	0	1	Yes	Private	Urban	228.69	36.6	formerly smoked	1
1	51676	Female	61.0	0	0	Yes	Self-employed	Rural	202.21	NaN	never smoked	1
2	31112	Male	80.0	0	1	Yes	Private	Rural	105.92	32.5	never smoked	1
3	60182	Female	49.0	0	0	Yes	Private	Urban	171.23	34.4	smokes	1
4	1665	Female	79.0	1	0	Yes	Self-employed	Rural	174.12	24.0	never smoked	1

Bước 2: Bắt đầu xử lí dữ liệu.

```
1 # Xem thông tin dữ liệu
2 df.info()
```

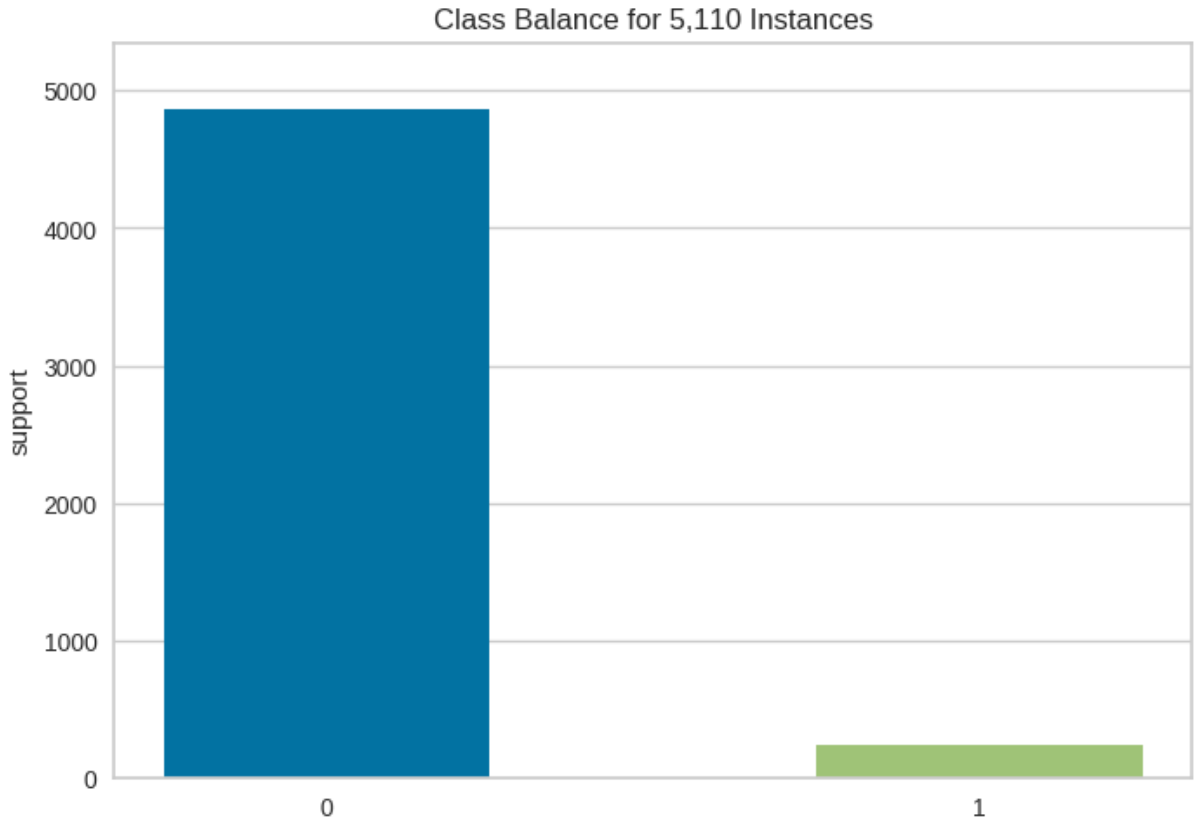
```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 5110 entries, 0 to 5109
Data columns (total 12 columns):
 #   Column             Non-Null Count  Dtype
---  ------             --------------  -----
 0   id                 5110 non-null   int64
 1   gender             5110 non-null   object
 2   age                5110 non-null   float64
 3   hypertension       5110 non-null   int64
 4   heart_disease      5110 non-null   int64
 5   ever_married       5110 non-null   object
 6   work_type          5110 non-null   object
 7   Residence_type     5110 non-null   object
 8   avg_glucose_level  5110 non-null   float64
 9   bmi                4909 non-null   float64
 10  smoking_status     5110 non-null   object
 11  stroke             5110 non-null   int64
dtypes: float64(3), int64(4), object(5)
memory usage: 479.2+ KB
```

Đây là thông tin bảng dữ liệu, ta thấy bảng data này không đồng nhất một dạng dữ liệu, nó tồn tại dữ liệu object. Nên em sẽ chuẩn hóa nó về dạng số.

Ngoài ra còn có các giá trị null ở cột ‘bmi’, em sẽ thực hiện thay thế giá trị null là giá trị trung bình (mean)

```
id                   0
gender               0
age                  0
hypertension         0
heart_disease        0
ever_married         0
work_type            0
Residence_type       0
avg_glucose_level    0
bmi                201
smoking_status       0
stroke               0
dtype: int64
```

Biểu diễn sự cân bằng lớp để xử lí nếu sự chênh lệch quá lớn:



Từ biểu đồ trên ta thấy được sự mất cân bằng giữa hai phân lớp của kết quả đầu ra mục tiêu quá lớn. Nếu hai lớp trong tập dữ liệu mất cân bằng quá lớn, có nghĩa là một

lớp có số lượng điểm dữ liệu lớn hơn nhiều so với lớp còn lại, thì việc dự đoán và huấn luyện mô hình có thể gặp một số thách thức.

Vấn đề mất cân bằng lớp có thể ảnh hưởng đến hiệu suất của mô hình, vì mô hình có thể dễ dàng dự đoán hầu hết các điểm dữ liệu vào lớp đa số, trong khi lớp thiểu số thường bị dự đoán sai. Điều này là do mô hình có xu hướng học từ lớp có số lượng lớn hơn nên hiếm khi phát hiện được các trường hợp hiếm hơn.

Vì vậy em chọn phương pháp SMOTE (Synthetic Minority Over-sampling Technique) để xử lì trường hợp này, lấy lại cân bằng cho hai lớp nhãn đầu ra. Một phương pháp để tạo dữ liệu nhân tạo cho lớp thiểu số bằng cách tạo ra các mẫu giả tạo từ các mẫu có sẵn trong lớp này. Điều này giúp cân bằng lại tập dữ liệu và cung cấp thêm sự đa dạng cho lớp thiểu số.

Sơ đồ hai lớp nhãn đầu ra sau khi thực hiện cân bằng, dữ liệu cũng được tăng lên gấp đôi.

Chia dữ liệu để huấn luyện bằng phương pháp K-Fold (K-Fold Cross-Validation) với k=5. Dữ liệu sẽ được chia đều ra để train và test giúp đạt hiệu quả cao hơn.

Chuẩn hóa dữ liệu đầu vào bằng phương pháp StandardScaler. StandardScaler được sử dụng để chuẩn hóa dữ liệu bằng cách loại bỏ trung bình và chia tỷ lệ chuẩn của từng đặc trưng.

2. Cấu hình máy để thực nghiệm

Cấu hình máy sử dụng để giải quyết bài toán: sử dụng google Colaboratory

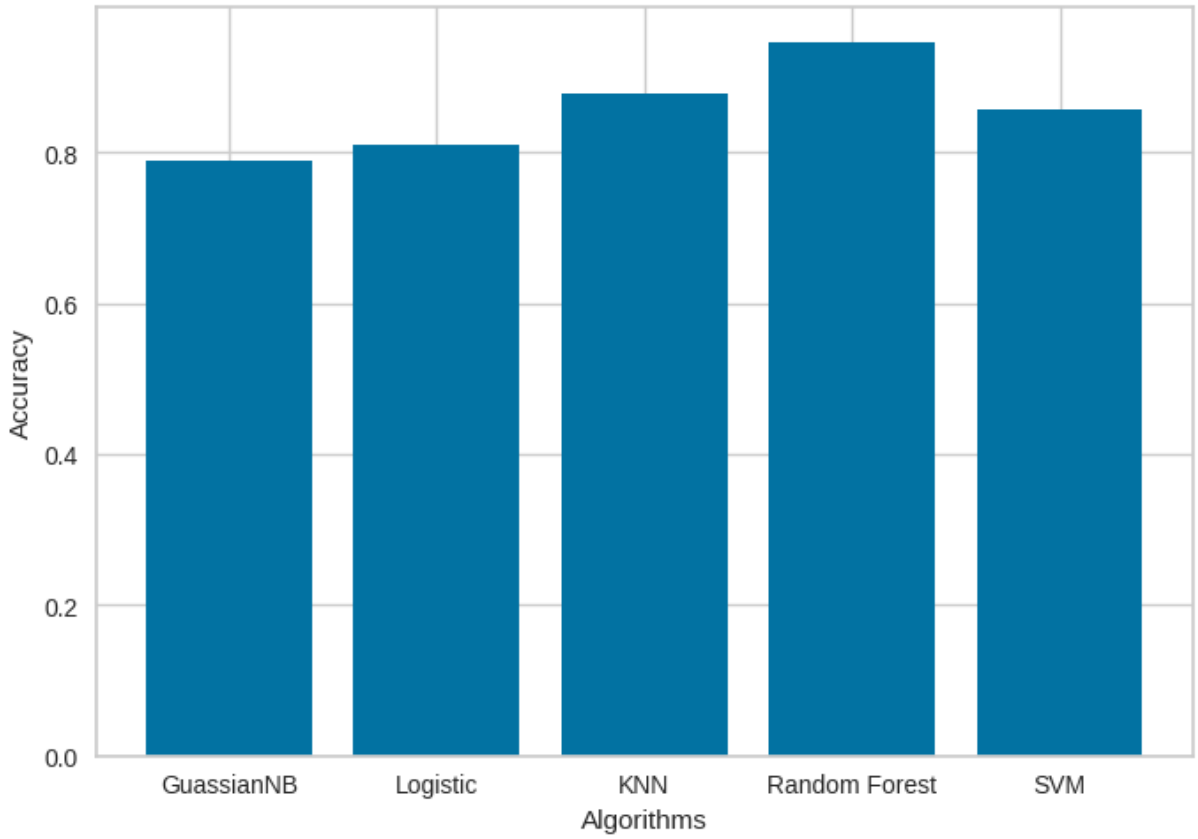
3. Mô tả mục tiêu đánh giá

Sử dụng các thuật toán Naive Bayes, K-Nearest Neighbors, Logistic Regression, Support Vector Machine và Random Forest để áp dụng lên bảng dữ liệu. Sử dụng các thư viện sẵn có trên python và bắt đầu thực nghiệm hai mô hình để đưa ra dự đoán.

4. Đánh giá và so sánh các tiêu chí

Sau khi huấn luyên mô hình qua các k-fold thì sẽ lấy trung bình

Thuật toán	Accuracy
Guassian Naïve Bayes	0.7883
K-Nearest Neighbors	0.8791
Logistic Regression	0.8094
Support Vector Machine	0.8576
Random Forest	0.9451



Từ sơ đồ trên, ta thấy được Random forest là thuật toán có độ chính xác cao nhất, vì vậy em sẽ lấy nó làm mô hình dự đoán. Test thử với mẫu đầu tiên của bảng dữ liệu

gender	age	hypertension	heart_disease	ever_married	work_type	Residence_type	avg_glucose_level	bmi	smoking_status	stroke
1	67.0	0	1	1	2	1	228.69	36.600000	1	1

```
Nhập giới tính (0: Nữ, 1: Nam): 1
Nhập tuổi: 67
Nhập tình trạng tăng huyết áp (0: Không, 1: Có): 0
Có bị tim (0: không, 1:có)1
Nhập tình trạng kết hôn (0: Không, 1: Có): 1
Nhập loại công việc (0-4): 2
Nhập loại nơi cư trú (0: Nông thôn, 1: Thành thị): 1
Nhập mức độ glucose trung bình: 228.69
Nhập chỉ số BMI: 36.6
Nhập tình trạng hút thuốc (0-3): 1
Có nguy cơ đột quỵ
```

Đầu ra mong muốn là 1 và đầu ra thực tế cũng là 1. Như vậy mô hình random forest đạt hiệu quả cao trong bộ dữ liệu này

CHƯƠNG V. KẾT LUẬN

Dựa trên quá trình so sánh 5 mô hình khác nhau trong việc dự đoán nguy cơ đột quỵ, mô hình Random Forest đã cho thấy có độ chính xác (accuracy) cao nhất trong số các mô hình được kiểm tra. Điều này cho thấy mô hình Random Forest có khả năng dự đoán chính xác cao hơn so với các mô hình khác trong tập dữ liệu cụ thể này.

Dưới góc độ tổng thể, nghiên cứu này đã cung cấp cái nhìn sâu rộng về khả năng của các thuật toán máy học trong việc dự đoán nguy cơ đột quỵ. Điều này có thể hỗ trợ quyết định lựa chọn phương pháp dự đoán trong thực tế y tế và đóng góp vào việc cải thiện quản lý sức khỏe và chăm sóc y tế.

TÀI LIỆU THAM KHẢO

Link mã nguồn:

https://drive.google.com/drive/folders/10XhhGI8PXuZZZDaGKzV5Joc2TY-LhjAw?usp=sharing

1. Võ Thị Hồng Thắm, Slide bài giảng Học máy và ứng dụng
2. Trần Minh Quang (2020), Khai phá dữ liệu và kỹ thuật phân lớp, Nhà xuất bản Đại học Quốc Gia TP. Hồ Chí Minh
3. https://www.kaggle.com/datasets/rishidamarla/heart-disease-prediction
4. scikit-learn: machine learning in Python — scikit-learn 1.2.0 documentation