**Ủ Y BAN NHÂN DÂN THÀNH PHỐ HỒ CHÍ MINH**

A blue button with white text

AI-generated content may be incorrect.**TRƯỜNG ĐẠI HỌC SÀI GÒN**

**BÁO CÁO BỆNH ĐÁI THÁO ĐƯỜNG**

**Học phần**

NHẬP MÔN MÁY HỌC

**Người hướng dẫn: TS. Đỗ Như Tài**

**Họ tên sinh viên:**

Nguyễn Tuấn Đạt – 3123410070

Châu Hải Đăng – 3123410075

Trần Đại Thắng – 3123410346

**Thành phố Hồ Chí Minh, tháng 10 năm 2025**

**LỜI MỞ ĐẦU**

Đái tháo đường (Tiểu đường) đang trở thành một trong những thách thức sức khỏe toàn cầu lớn nhất của thế kỷ 21, với số lượng người mắc bệnh ngày càng gia tăng và gây ra nhiều biến chứng nguy hiểm. Việc phát hiện sớm và chẩn đoán chính xác bệnh đái tháo đường đóng vai trò then chốt trong việc lập kế hoạch điều trị hiệu quả, qua đó giảm thiểu gánh nặng về y tế và kinh tế - xã hội.

Trong bối cảnh cuộc Cách mạng Công nghiệp 4.0, Học máy (Machine Learning) đã nổi lên như một công cụ mạnh mẽ, mang lại những hướng tiếp cận đột phá trong lĩnh vực y tế dự phòng. Khả năng phân tích các bộ dữ liệu y tế phức tạp và tìm ra những mẫu tiềm ẩn của các thuật toán Học máy mở ra tiềm năng to lớn trong việc xây dựng các mô hình dự đoán bệnh với độ chính xác cao.

Báo cáo này tập trung vào việc ứng dụng các kỹ thuật Học máy để giải quyết bài toán dự đoán nguy cơ mắc bệnh đái tháo đường. Cụ thể, nghiên cứu sử dụng bộ dữ liệu công khai "Pima Indians Diabetes" – một bộ dữ liệu kinh điển đã được sử dụng rộng rãi trong cộng đồng nghiên cứu. Mục tiêu chính của báo cáo là xây dựng, huấn luyện và đánh giá hiệu năng của một số mô hình phân loại phổ biến nhằm tìm ra mô hình tối ưu nhất, có khả năng hỗ trợ hiệu quả cho công tác sàng lọc và chẩn đoán sớm bệnh đái tháo đường.

**LỜI CẢM ƠN**

Chúng em xin chân thành cảm ơn thầy **TS. Đỗ Như Tài** đã cung cấp những kiến thức từ cơ bản đến chuyên sâu, tận tình góp ý để nhóm có thể hoàn thiện bản báo cáo một cách hợp lí và khoa học. Sự tâm huyết truyền thụ kiến thức của một người giảng viên lâu năm kinh nghiệm, và một người giảng viên trẻ năng động đầy tận tâm trong việc giảng dạy là chất xúc tác giúp chúng em hoàn thành đồ án lần này.

Trong thời gian hoàn thành bản báo cáo, nhóm đã gặp những khó khăn nhất định do các yếu tố chủ quan và khách quan. Nhóm đã cố gắng hoàn thiện đồ án từ những kinh nghiệm, kiến thức cá nhân được trao dồi trong quá trình học, cũng như từ nhiều nguồn tham khảo khác nhau tuy nhiên vẫn không tránh khỏi những sai xót. Nhóm rất trân trọng những ý kiến của các thầy để làm hành trang kiến thức cho quá trình học tập và làm việc sau này.

Một lần nữa nhóm xin gửi lời cảm ơn trân trọng nhất đến thầy cô.

*TP.Hồ Chí Minh tháng 4 năm 2025*

Nhóm thực hiện đồ án.

Nguyễn Tuấn Đạt

Châu Hải Đăng

Trần Đại Thắng

# **NHẬN XÉT CỦA GIẢNG VIÊN**

**...........................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................**

**Chương 1: Tổng quan về đề tài.**

* 1. Đặt vấn đề

Trong kỷ nguyên y học hiện đại, bệnh đái tháo đường (tiểu đường) đã và đang trở thành một trong những thách thức sức khỏe cộng đồng nghiêm trọng nhất trên toàn cầu. Theo Tổ chức Y tế Thế giới (WHO), số lượng người mắc bệnh đái tháo đường đã tăng từ 108 triệu người năm 1980 lên 422 triệu người năm 2014 và con số này vẫn tiếp tục gia tăng, đặc biệt ở các quốc gia có thu nhập thấp và trung bình. Đái tháo đường không chỉ là một căn bệnh mãn tính gây suy giảm chất lượng cuộc sống mà còn là nguyên nhân hàng đầu dẫn đến nhiều biến chứng nguy hiểm như bệnh tim mạch, mù lòa, suy thận và hoại tử chi.

Việc chẩn đoán bệnh sớm đóng vai trò then chốt trong việc kiểm soát và điều trị hiệu quả, giúp ngăn ngừa hoặc làm chậm sự tiến triển của các biến chứng. Tuy nhiên, các phương pháp chẩn đoán truyền thống thường đòi hỏi xét nghiệm y tế, tốn kém thời gian và chi phí, đồng thời không phải lúc nào cũng dễ dàng tiếp cận với mọi đối tượng.

Cùng với sự phát triển vượt bậc của khoa học máy tính, lĩnh vực Học máy (Machine Learning) đã mở ra một hướng tiếp cận mới đầy tiềm năng. Bằng cách phân tích các bộ dữ liệu y tế lớn, các mô hình Học máy có thể học và nhận diện các mẫu phức tạp ẩn sau các chỉ số sức khỏe, từ đó xây dựng nên các công cụ dự đoán bệnh với độ chính xác cao, chi phí thấp và không xâm lấn.

Xuất phát từ những lý do trên, đề tài “Ứng dụng Machine Learning để xây dựng mô hình dự đoán bệnh đái tháo đường dựa trên bộ dữ liệu Pima Indians Diabetes” được thực hiện với mong muốn ứng dụng các thuật toán tiên tiến để tạo ra một mô hình có khả năng hỗ trợ sàng lọc và phát hiện sớm nguy cơ mắc bệnh, góp phần vào công cuộc chăm sóc sức khỏe cộng đồng.

* 1. Mục tiêu nghiên cứu
     1. Mục tiêu tổng quát

Xây dựng và đánh giá các mô hình Học máy nhằm dự đoán khả năng mắc bệnh đái tháo đường của một cá nhân dựa trên các chỉ số y tế cơ bản, sử dụng bộ dữ liệu Pima Indians Diabetes.

* + 1. Mục tiêu cụ thể

Để đạt được mục tiêu tổng quát, đề tài cần thực hiện các nhiệm vụ cụ thể sau:

* Nghiên cứu tổng quan về bệnh đái tháo đường và các yếu tố nguy cơ liên quan.
* Tìm hiểu cơ sở lý thuyết về Học máy, đặc biệt là các thuật toán phân loại (Classification) như Logistic Regression, K-Nearest Neighbors (KNN), Support Vector Machine (SVM), Decision Tree và Random Forest.
* Thực hiện phân tích, khám phá và tiền xử lý bộ dữ liệu Pima Indians Diabetes để chuẩn bị cho việc xây dựng mô hình.
* Xây dựng, huấn luyện và tinh chỉnh các mô hình dự đoán dựa trên những thuật toán đã nghiên cứu.
* Đánh giá hiệu năng của các mô hình thông qua các độ đo phổ biến (Accuracy, Precision, Recall, F1-Score) và lựa chọn ra mô hình có kết quả tốt nhất.
  1. Đối tượng và phạm vi nghiên cứu

Đối tượng nghiên cứu: Các thuật toán phân loại trong Học máy và quy trình xây dựng mô hình dự đoán.

Đối tượng khảo sát: Bộ dữ liệu Pima Indians Diabetes được cung cấp trên các nền tảng như Kaggle hoặc UCI Machine Learning Repository.

Phạm vi nghiên cứu:

* Về thuật toán: Tập trung vào một số thuật toán phân loại cơ bản và phổ biến.
* Về công cụ: Sử dụng ngôn ngữ lập trình Python cùng các thư viện mã nguồn mở hỗ trợ mạnh mẽ cho khoa học dữ liệu như Scikit-learn, Pandas, NumPy, Matplotlib và Seaborn.
  1. Phương pháp nghiên cứu

Đề tài chủ yếu sử dụng phương pháp sau:

* Phương pháp nghiên cứu lý thuyết: Thu thập, tổng hợp và phân tích các tài liệu, giáo trình, bài báo khoa học liên quan đến bệnh đái tháo đường và các thuật toán Học máy.
* Phương pháp thực nghiệm: Tiến hành các bước thực nghiệm khoa học trên máy tính, bao gồm tiền xử lý dữ liệu dữ liệu, huấn luyện mô hình, kiểm thử và đánh giá kết quả một cách khách quan.
* Phương pháp so sánh, phân tích: So sánh kết quả giữa các mô hình khác nhau để đưa ra kết luận về hiệu quả của từng thuật toán đối với bài toán cụ thể này.
  1. Ý nghĩa khoa học và thực tiễn của đề tài
* Ý nghĩa khoa học: Hệ thống hóa kiến thức về việc áp dụng Học máy trong lĩnh vực y tế, cung cấp một ví dụ thực tiễn về quy trình xây dựng mô hình dự đoán và so sánh hiệu năng các thuật toán trên một bộ dữ liệu kinh điển
* Ý nghĩa thực tiễn: Kết quả của đề tài là một mô hình dự đoán, có thể được xem như một công cụ tham khảo hỗ trợ các chuyên gia y tế trong việc sàng lọc sớm, giúp xác định các đối tượng có nguy cơ cao cần được ưu tiên xét nghiệm chuyên sâu, từ đó tiết kiệm chi phí và nâng cao hiệu quả chẩn đoán.

**Chương 2: Xây dựng mô hình và phân tích dữ liệu**

2.1 Xây dựng mô hình

Từ dữ liệu có sẵn sẽ được lưu thành file CSV, sau đó sẽ được chia thành 2 tập dữ liệu có tỉ lệ 80% cho tập huấn luyện và 20% cho tập kiểm tra , kế tiếp chương trình sẽ áp dụng các thuật toán như AdaBoost, Random Forest, SVC và Knearest Neighbors tiến hành phân lớp thông qua tập huấn luyện. Cuối cùng sẽ dự đoán và đánh giá mô hình thông qua tập kiểm tra.

2.2 Phân tích dữ liệu

AdaBoost

Random Forest

SVC

Knearest Neighbors

TRAIN

DATASET

PREDICTION

TEST

CLASSIFICATION

2.2.1 Giới thiệu về dữ liệu

Dữ liệu cho đề tài "Dự đoán bệnh tiểu đường" là một bộ dữ liệu chứa thông tin về các biến độc lập (các yếu tố) và biến mục tiêu (kết quả) để dự đoán nguy cơ mắc bệnh tiểu đường. Bộ dữ liệu này được lấy từ kaggle sử dụng phát triển các mô hình học máy để dự đoán xác suất mắc bệnh tiểu đường dựa trên các yếu tố khác nhau. Dữ liệu này thường được thu thập thông qua các phương pháp điều tra lâm sàng hoặc dựa trên các bệnh nhân có sẵn trong các cơ sở y tế. Dữ liệu có thể được nhập tay từ hồ sơ bệnh nhân hoặc thu thập tự động thông qua các hệ thống hồ sơ điện tử.

Các trường thông tin của dữ liệu: Dữ liệu bao gồm các thông tin sau:

* Number of times pregnant : số lần có thai.
* Plasma glucose concentration a 2 hours in an oral glucose tolerance test : bài kiểm tra độ tích tụ lượng đường trong cơ thể.
* Diastolic blood pressure (mm Hg): chỉ số đo huyết áp tâm trương (áp lực động mạch khi tim nghỉ ngơi giữa 2 lần co bóp).
* Triceps skin fold thickness (mm): độ dày da gắp nép của bắp tay sau.
* 2-Hour serum insulin (mu U/ml): chỉ số đo insulin trong cơ thể sau 2 giờ dung nạp gluco
* Body mass index (weight in kg/(height in m)^2): chỉ số đo tỉ lệ cân nặng chiều cao theo chuẩn BMI.
* Diabetes pedigree function: chỉ số đo bệnh về đái tháo đường theo phả hệ.
* Age (years): tuổi.

Để đọc dữ liệu từ tệp CSV chúng ta sẽ sử dụng thư viện pandas và sử dụng thư viện matplotlib để trực quan hoá dữ liệu bằng cách vẽ biểu đồ phân phối cho các biến. Điều này giúp chúng ta có cái nhìn tổng quan về dữ liệu trước khi bắt đầu xây dựng mô hình dự đoán bệnh tiểu đường

A screen shot of a computer

AI-generated content may be incorrect.

A screen shot of a computer

AI-generated content may be incorrect.

A screenshot of a graph

AI-generated content may be incorrect.

Để hiểu cấu trúc dữ liệu, kiểm tra giá trị thiếu, đánh giá tình trạng dữ liệu, ta sẽ sử dụng phương thức info(). Xem số lượng dòng, cột và kiểu dữ liệu, phát hiện giá trị thiếu trong các cột sẽ giúp chuẩn bị cho việc xử lý các giá trị trống một cách hiệu quả trước khi phân tích. Có cái nhìn tổng quan về tình trạng của dữ liệu, giúp xác định cần thực hiện bước tiền xử lý nào để làm sạch dữ liệu trước khi phân tích.

A screenshot of a computer

AI-generated content may be incorrect.

2.2.2 Trích chọn đặc trưng

Dữ liệu về bệnh tiểu đường là một tập hợp thông tin đa dạng về các yếu tố liên quan đến bệnh này, bao gồm thông tin Glucose, BloodPressure, SkinThickness, Insulin, BMI, DiabetesPedigreeFuntion, Age. Mỗi đặc trưng này cung cấp thông tin quan trọng về sức khỏe và yếu tố nguy cơ mắc bệnh tiểu đường.

A graph of different colored bars

AI-generated content may be incorrect.

Dựa theo đồ thị thì ta thấy 2 biến Isulin và Glucose là 2 thuộc tính có sự khác biệt dẫn rõ ràng dẫn đến 2 thuộc tính này là không thể thiếu trong việc dự đoán.

A computer code on a black background

AI-generated content may be incorrect.

A diagram of a heatmap

AI-generated content may be incorrect.

- Cặp chỉ số BMI và da gấp nép có độ tương quang cao (.66), có nghĩa là BMI càng lớn thì da gấp nép càng dày. Sự tương quan này có thể là dữ liệu tốt trong việc dự đoán bệnh đái tháo đường vì nó cho biết sự mất cân đối của 1 cá thể.

- Tuổi và số lần có thai (.68) có độ tương quang cao nhất và ngoài ra thì theo các bài báo khoa học ta biết được phụ nữ có thai dễ mắc phải bệnh hơn khi càng lớn tuổi và có thai nhiều lần.

- Ngoài ra thì còn chỉ số thực nghiệm Insulin và Gluco (.58) là một trong những công cụ mạnh nhất hay thông tin tốt nhất và quan hệ mật thiết nhất với bệnh đái tháo đường trong y học.

A black screen with white text

AI-generated content may be incorrect.

A chart of a diagram

AI-generated content may be incorrect.

Nhận thấy 2 thuộc tính Isulin và Glucose là 2 thuộc tính có phân cụm dễ nhìn và rõ ràng nhất.

Chúng ta không cần trích chọn đặc trưng trong trường hợp này là vì mỗi thông tin đều có thể đóng góp vào việc dự đoán bệnh tiểu đường một cách quan trọng. Tuổi có thể là một yếu tố quan trọng trong việc đánh giá nguy cơ mắc bệnh, trong khi chỉ số BMI có thể liên quan mật thiết đến việc đo lường tình trạng cơ thể. Mức độ đường trong máu và Isulin cũng cung cấp thông tin chi tiết về kiểm soát đường huyết, một yếu tố quan trọng trong quản lý bệnh lý. Điều này sẽ mang lại những lợi ích:

* Đa dạng thông tin quan trọng: Các thông tin về sức khỏe, lối sống và yếu tố di truyền đều có thể đóng góp vào việc dự đoán nguy cơ mắc bệnh tiểu đường. Không trích chọn đặc trưng giúp giữ lại đầy đủ thông tin để mô hình có thể tận dụng mối quan hệ phức tạp giữa các yếu tố này.
* Tiềm năng ẩn của thông tin: Đôi khi, các thông tin mà ban đầu có vẻ không quan trọng có thể liên quan mật thiết đến bệnh lý. Việc loại bỏ các đặc trưng này có thể làm mất đi thông tin quan trọng trong việc dự đoán bệnh tiểu đường.
* Tránh mất mát thông tin: Trích chọn đặc trưng có thể dẫn đến mất mát thông tin quan trọng. Trong trường hợp bệnh tiểu đường, việc giữ lại tất cả thông tin có thể giúp mô hình học máy có cái nhìn tổng quan và toàn diện hơn về bệnh lý.
* Khả năng mô hình hóa tốt hơn: Một số thuật toán học máy như mạng nơ-ron sâu (deep neural networks) có thể học và tận dụng thông tin từ nhiều đặc trưng khác nhau một cách hiệu quả, do đó không cần phải trích chọn đặc trưng một cách cẩn thận.

Vì vậy, việc giữ lại tất cả các đặc trưng này giúp mô hình học máy hiểu rõ hơn về mối quan hệ phức tạp giữa chúng và bệnh tiểu đường. Điều này cung cấp cho mô hình cái nhìn toàn diện hơn về các yếu tố có thể ảnh hưởng đến bệnh lý và giúp trong việc dự đoán chính xác nguy cơ mắc bệnh, cũng như đưa ra các khuyến nghị quản lý sức khỏe hiệu quả.

2.3 Chuẩn hóa dữ liệu

Bước tiếp theo cần thực hiện để chuẩn bị dữ liệu cho máy học là làm sạch dữ liệu đó. Làm sạch dữ liệu liên quan đến việc tìm và sửa lỗi, sự không nhất quán và các giá trị bị thiếu. Có một số cách tiếp cận để làm điều đó:

* Xử lý dữ liệu bị thiếu: Thiếu giá trị là một vấn đề phổ biến trong học máy. Nó có thể được xử lý bằng cách quy nạp (nghĩ: điền các giá trị còn thiếu bằng dữ liệu dự đoán hoặc ước tính), nội suy (lấy các giá trị còn thiếu từ các điểm dữ liệu xung quanh) hoặc xóa (chỉ cần xóa các hàng hoặc cột có giá trị bị thiếu khỏi tập dữ liệu.)
* Xử lý ngoại lệ: Ngoại lệ là các điểm dữ liệu khác biệt đáng kể so với phần còn lại của tập dữ liệu. Các ngoại lệ có thể xảy ra do lỗi đo lường, lỗi nhập dữ liệu hoặc đơn giản vì chúng đại diện cho các quan sát bất thường hoặc cực đoan. Ví dụ, trong một bộ dữ liệu về lương của nhân viên, một ngoại lệ có thể là một nhân viên kiếm được nhiều hơn hoặcít hơn đáng kể so với những người khác. Các ngoại lệ có thể được xử lý bằng cách loại bỏ, biến đổi chúng để giảm tác động của chúng, chiến thắng (nghĩ: thay thế các giá trị cực đoan bằng các giá trị gần nhất nằm trong phạm vi phân phối thông thường) hoặc coi chúng là một loại dữ liệu riêng biệt.
* Loại bỏ trùng lặp: Một bước khác trong quá trình chuẩn bị dữ liệu cho máy học là loại bỏ các dữ liệu trùng lặp. Các bản sao không chỉ làm sai lệch các dự đoán ML mà còn lãng phí dung lượng lưu trữ và tăng thời gian xử lý, đặc biệt là trong các bộ dữ liệu lớn. Để loại bỏ các bản sao, các nhà khoa học dữ liệu sử dụng nhiều kỹ thuật nhận dạng trùng lặp (như khớp chính xác, khớp mờ, băm hoặc liên kết bản ghi). Sau khi được xác định, chúng có thể được loại bỏ hoặc hợp nhất. Tuy nhiên, trong các bộ dữ liệu không cân bằng, trên thực tế, các bản sao có thể được hoan nghênh để đạt được phân phối bình thường.
* Xử lý dữ liệu không liên quan: Dữ liệu không liên quan đề cập đến dữ liệu không hữu ích hoặc không thể áp dụng để giải quyết vấn đề. Xử lý dữ liệu không liên quan có thể giúp giảm nhiễu và cải thiện độ chính xác của dự đoán. Để xác định dữ liệu không liên quan, các nhóm dữ liệu sử dụng các kỹ thuật như phân tích thành phần chính, phân tích tương quan hoặcđơn giản là dựa vào kiến thức miền của họ. Sau khi được xác định, các điểm dữ liệu đó sẽ bị xóa khỏi tập dữ liệu.
* Xử lý dữ liệu không chính xác: Chuẩn bị dữ liệu cho học máy cũng phải bao gồm việc xử lý dữ liệu không chính xác và sai sót. Các kỹ thuật phổ biến để xử lý dữ liệu đó bao gồm chuyển đổi dữ liệu (thay đổi dữ liệu để dữ liệu đáp ứng các tiêu chí đã đặt) hoặc xóa hoàn toàn các điểmdữ liệu không chính xác.
* Xử lý dữ liệu mất cân bằng: Tập dữ liệu không cân bằng là tập dữ liệu trong đó số điểm dữ liệu trong một lớp thấp hơn đáng kể so với số điểm dữ liệu trong lớp khác. Điều này có thể dẫn đến một mô hình thiên lệch ưu tiên cho tầng lớp đa số, trong khi bỏ qua tầng lớp thiểu số. Để giải quyết vấn đề, các nhóm dữ liệu có thể sử dụng các kỹ thuật như lấy mẫu lại (lấy mẫu quá mức của lớp thiểu số hoặc lấy mẫu dưới lớp đa số để cân bằng việc phân phối dữ liệu), tạo dữ liệu tổng hợp (tạo điểm dữ liệu bổ sung cho lớp thiểu số một cách tổng hợp), chi phí -học tập nhạy cảm (gán trọng số cao hơn cho lớp thiểu số trong quá trình đào tạo), học tập đồng bộ (kết hợp nhiều mô hình được đào tạo trên các tập hợp con dữ liệu khác nhau bằng các thuật toán khác nhau) và các mô hình khác.

Các hoạt động này giúp đảm bảo rằng dữ liệu đào tạo là chính xác, đầy đủ và nhất quán. Mặc dù là một thành tích lớn nhưng vẫn chưa đủ để tạo ra một mô hình ML đáng tin cậy. Vì vậy, bước tiếp theo trong hành trình chuẩn bị dữ liệu cho máy học liên quan đến việc đảm bảo các điểm dữ liệu trong tập dữ liệu huấn luyện tuân thủ các quy tắc và tiêu chuẩn cụ thể. Và giai đoạn đó trong quy trình quản lý dữ liệu được gọi là chuyển đổi dữ liệu

A screen shot of a computer code

AI-generated content may be incorrect.

A screenshot of a computer screen

AI-generated content may be incorrect.

Ở đây nhóm tôi xử lý dữ liệu bị thiếu theo trung vị của lớp output.

Xử lý theo Normalization và Standardization

Normalization: chuẩn hóa min max

Xnormalized =

Standardization (Tiêu chuẩn hóa Z-score)

A screen shot of a computer program

AI-generated content may be incorrect.

A screenshot of a computer screen

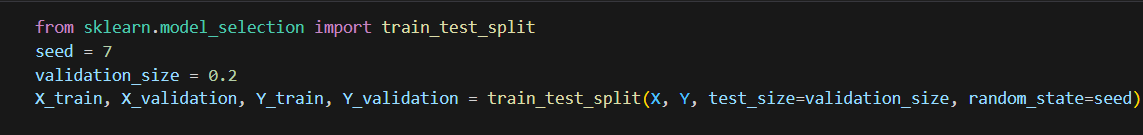
AI-generated content may be incorrect.

2.4 Phân chia dữ liệu

Bước tiếp theo trong quá trình chuẩn bị dữ liệu cho máy học bao gồm việc chia tất cả dữ liệu đã thu thập thành các tập hợp con — quá trình này được gọi là phân tách dữ liệu. Thông thường, dữ liệu được chia nhỏ thành tập dữ liệu đào tạo, xác thực và thử nghiệm.

* Tập dữ liệu huấn luyện được sử dụng để thực sự dạy một mô hình máy học nhận
* Tập dữ liệu huấn luyện được sử dụng để thực sự dạy một mô hình máy học để nhận dạng các mẫu và mối quan hệ giữa các biến đầu vào và mục tiêu. Bộ dữ liệu này thường là lớn nhất.
* Tập dữ liệu xác thực là một tập hợp con dữ liệu được sử dụng để đánh giá hiệu suất của mô hình trong quá trình đào tạo. Nó giúp tinh chỉnh mô hình bằng cách điều chỉnh các siêu tham số (nghĩ: các tham số của quy trình đào tạo được đặt thủ công trước khi đào tạo, như tốc độ học tập, cường độ chuẩn hóa hoặc số lớp ẩn). Tập dữ liệu xác thực cũng giúp ngăn chặn việc khớp quá mức với dữ liệu huấn luyện.
* Tập dữ liệu thử nghiệm là một tập hợp con dữ liệu được sử dụng để đánh giá hiệu suất của mô hình được đào tạo. Mục tiêu của nó là đánh giá độ chính xác của mô hình trên dữ liệu mới, chưa từng thấy. Tập dữ liệu thử nghiệm chỉ được sử dụng một lần — sau khi mô hình đã được huấn luyện và tinh chỉnh trên tập dữ liệu huấn luyện và xác thực.

Bằng cách chia tách dữ liệu, chúng tôi có thể đánh giá mức độ hiệu quả của một mô hình máy học trên dữ liệu mà nó chưa từng thấy trước đây. Nếu không có sự phân tách, rất có thể mô hình sẽ hoạt động kém trên dữ liệu mới. Điều này có thể xảy ra vì mô hình có thể chỉ ghi nhớ các điểm dữ liệu thay vì học các mẫu và khái quát hóa chúng thành dữ liệu mới.



Ở đây chung tôi sẽ sử dụng phương K-Fold Cross Validation để đánh giá mô hình.

Để lựa chọn mô hình học máy có độ chính xác cao nhất trong bài toán dự đoán bệnh đái tháo đường, nhóm sử dụng kỹ thuật K-Fold Cross Validation với K = 10.  
Phương pháp này chia dữ liệu huấn luyện thành 10 phần (fold) khác nhau, trong đó 9 phần được dùng để huấn luyện và 1 phần còn lại dùng để kiểm thử, quá trình được lặp lại 10 lần để đảm bảo tính khách quan và tránh hiện tượng overfitting.

Nhóm tiến hành so sánh bốn mô hình phân loại phổ biến:

* + K-Nearest Neighbors (KNN)
  + Random Forest Classifier
  + Support Vector Machine (SVC)
  + AdaBoost Classifier

A screen shot of a computer program

AI-generated content may be incorrect.

Kết quả:

A screenshot of a computer screen

AI-generated content may be incorrect.

Kết quả cho thấy AdaBoost Classifier đạt độ chính xác trung bình cao nhất với 87.62%, do đó được chọn làm mô hình tốt nhất trong giai đoạn này.

Sau khi chọn được mô hình có hiệu quả cao nhất, nhóm tiến hành huấn luyện lại toàn bộ các mô hình trên tập huấn luyện (X\_train, Y\_train) và kiểm thử trên tập dữ liệu mới (X\_validation, Y\_validation) — tức là dữ liệu chưa từng xuất hiện trong quá trình học.

A computer screen with white text

AI-generated content may be incorrect.

Kết quả cho thấy mặc dù KNN và AdaBoost có hiệu quả tốt trên tập huấn luyện, nhưng khi áp dụng lên dữ liệu mới, Random Forest và AdaBoost vẫn giữ được độ ổn định tốt hơn.  
Vì vậy, nhóm quyết định tập trung tinh chỉnh (tune) các siêu tham số cho hai mô hình này ở bước tiếp theo để cải thiện hiệu suất dự đoán.

**Chương 3: Huấn luyện mô hình**

3.1 Mô hình AdaBoost

AdaBoost không phải là một thuật toán phân loại đơn lẻ dựa trên hàm logistic, mà là một thuật toán học tập kết hợp (Ensemble Learning) có giám sát, được thiết kế để cải thiện hiệu suất của các bộ phân loại yếu (weak learners), biến chúng thành một bộ phân loại mạnh (strong learner).

Ý tưởng chính của AdaBoost là kết hợp tuần tự nhiều bộ phân loại yếu (thường là các cây quyết định rất nông, hay còn gọi là *stumps*). Mô hình này tập trung vào các điểm dữ liệu bị phân loại sai ở các bước trước bằng cách tăng trọng số của chúng.

* Đầu vào: Một tập hợp dữ liệu huấn luyện và một bộ phân loại yếu.
* Mục tiêu: Xây dựng một mô hình cuối cùng là tổng hợp có trọng số của các bộ phân loại yếu.

Quá trình huấn luyện của AdaBoost được thực hiện lặp đi lặp lại qua $T$ vòng:

1. Khởi tạo Trọng số Dữ liệu (Initialize Weights):
   * Ban đầu, tất cả các điểm dữ liệu trong tập huấn luyện được gán một trọng số bằng nhau với N là số lượng điểm dữ liệu).
2. Huấn luyện Bộ phân loại Yếu (Train Weak Learner):
   * Tại mỗi vòng lặp t, một bộ phân loại yếu ht được huấn luyện trên tập dữ liệu đã điều chỉnh trọng số. Bộ phân loại này cố gắng phân loại đúng các điểm dữ liệu có trọng số cao.
3. Tính toán Mức độ Đóng góp (Calculate Learner Weight):
   * Tính toán lỗi của bộ phân loại yếu ht dựa trên tổng trọng số của các điểm bị phân loại sai.
   * Tính toán mức độ đóng góp (hay còn gọi là *trọng số*) của bộ phân loại này vào mô hình cuối cùng. Bộ phân loại nào có lỗi càng nhỏ thì mức độ đóng góp càng lớn.
4. Cập nhật Trọng số Dữ liệu (Update Data Weights):
   * Tăng trọng số của các điểm dữ liệu bị bộ phân loại ht phân loại sai.
   * Giảm trọng số của các điểm dữ liệu được phân loại đúng.
   * Việc này đảm bảo rằng bộ phân loại yếu tiếp theo (ở vòng t + 1) sẽ tập trung hơn vào các điểm dữ liệu "khó".

Đưa ra dự đoán: Khi có các tham số tối ưu (tức là mức độ đóng góp cho mỗi bộ phân loại yếu), AdaBoost đưa ra dự đoán bằng cách:

* Tổng hợp có Trọng số: Dự đoán cuối cùng là tổng hợp có trọng số của tất cả các bộ phân loại yếu ht đã được huấn luyện.
* Quyết định Phân loại: Lớp cuối cùng được dự đoán bằng cách lấy dấu (sign) của tổng có trọng số này (đối với bài toán phân loại nhị phân).

Đánh giá mô hình với tập test

A screen shot of a computer code

AI-generated content may be incorrect.

Vậy có thể thấy AdaBoost có độ chính xác khoảng 83.77%

3.2 Mô hình Random Forest

Random Forest là một thuật toán học máy phổ biến thường được sử dụng cho cả bài toán phân loại và hồi quy. Nó kết hợp kết quả từ nhiều cây quyết định để đưa ra một kết quả duy nhất. Điều này giúp cải thiện độ chính xác và kiểm soát hiện tượng quá khớp (overfitting) mà thường gặp khi sử dụng một cây quyết định duy nhất. Dùng thuật toán Random Forest để đưa ra dự đoán bệnh tiểu đường

Hoạt Động của Random Forest:

* Khi có một dữ liệu mới cần được dự đoán, Random Forest thực hiện các bước sau để tạo dự đoán cuối cùng.
* Thực hiện dự đoán trên từng cây quyết định: Mô hình sử dụng mỗi cây quyết định để dự đoán kết quả trên dữ liệu mới.
* Kết hợp dự đoán từ các cây con: Dự đoán từ mỗi cây con được tính toán và sau đó kết hợp để tạo ra dự đoán cuối cùng. Trong bài toán phân loại, kết quả cuối cùng có thể dựa trên sự biểu quyết (voting) của các cây. Trong bài toán hồi quy, kết quả cuối cùng thường là trung bình của các dự đoán từ các cây.
* Kết quả dự đoán cuối cùng: Kết quả dự đoán cuối cùng được trả về.

A screenshot of a computer program

AI-generated content may be incorrect.

Đánh giá mô hình với tập test

A screen shot of a computer code

AI-generated content may be incorrect.

Vậy RandomForest sau khi tunned có độ chính xác khoảng 88,31%

**Chương 4: Kết luận**

4.1 Kết quả đã triển khai

Dựa trên kết quả có thể đánh giá các mô hình máy học như sau:

* Mô hình AdaBoost (Adaptive Boosting) là một mô hình học tập kết hợp (Ensemble Learning) được thiết kế để cải thiện hiệu suất tổng thể, đặc biệt là cân bằng giữa các chỉ số của các lớp.Mô hình AdaBoost của chúng ta đạt được độ chính xác (Accuracy) tổng thể cao hơn, khoảng 0.8377 (83,77%) trên tập kiểm tra. Đặc điểm hiệu suất quan trọng nhất của nó là sự cân bằng tốt hơn giữa Độ chuẩn xác (Precision) và Độ nhớ (Recall) cho lớp 1 (bệnh tiểu đường) so với mô hình Logistic Regression cơ bản.
* Mô hình RandomForest đạt được độ chính xác (accuracy) là 88.31% trên tập dữ liệu kiểm tra. Điều này có nghĩa là mô hình dự đoán đúng khoảng 88% tổng số mẫu trong tập kiểm tra — một kết quả thể hiện khả năng khái quát hóa khá tốt đối với dữ liệu mới. Nếu xem xét sâu hơn qua các chỉ số đánh giá khác (precision, recall, và AUC), ta có thể nhận thấy rằng mô hình Random Forest không chỉ đạt độ chính xác cao mà còn có độ chuẩn xác (precision) và độ bao phủ (recall) tốt cho cả hai lớp (bệnh và không bệnh).  
  Cụ thể, với lớp “không mắc bệnh”, mô hình duy trì precision và recall ở mức cao, trong khi với lớp “mắc bệnh”, tuy các giá trị này thấp hơn đôi chút, nhưng vẫn đủ thể hiện khả năng phát hiện bệnh đáng tin cậy. Như vậy, Random Forest sau khi tinh chỉnh là mô hình có hiệu suất tổng thể tốt nhất trong số các mô hình được thử nghiệm, cân bằng giữa độ chính xác, độ ổn định và khả năng khái quát hóa.

4.2 Hướng phát triển của đề tài

- Tối ưu hóa các mô hình hiện tại: Có thể tìm cách tối ưu hóa cho các mô hình AdaBoost và RandomForest để cải thiện độ chính xác. Điều này có thể bao gồm việc điều chỉnh các tham số, sử dụng các phương pháp chọn lọc đặc trưng, hoặc áp dụng các kỹ thuật chuẩn hóa và mở rộng dữ liệu

- Sử dụng các mô hình học sâu: Các mô hình học sâu như mạng nơ ron nhân tạo (ANN) và mạng nơ ron tích chập (CNN) có thể được sử dụng để xử lý các tác vụ phức tạp hơn và cung cấp kết quả chính xác hơn.

- Phân tích dữ liệu lớn: Với sự tăng trưởng của dữ liệu y tế, việc sử dụng các kỹ thuật phân tích dữ liệu lớn có thể giúp cải thiện độ chính xác của mô hình dự đoán.

- Sử dụng dữ liệu thời gian thực: Việc sử dụng dữ liệu thời gian thực từ các thiết bị theo dõi sức khỏe có thể giúp cải thiện khả năng dự đoán của các mô hình và cho phép phát hiện sớm các dấu hiệu của bệnh tiểu đường.

- Tích hợp với hệ thống y tế: Các mô hình dự đoán có thể được tích hợp vào hệ thống y tế để hỗ trợ các bác sĩ trong việc đưa ra quyết định lâm sàn cho bệnh tiểu đường.

**Tài liệu tham khảo**

“Diabetes prediction using machine learning and explainable AI” — trình bày việc kết hợp các thuật toán ML và kỹ thuật giải thích (LIME, SHAP) để dự đoán đái tháo đường.

[1] Về Lý thuyết Học máy (Machine Learning)- Hastie, T., Tibshirani, R., Friedman, J. (2009).

[2] Về Thuật toán Boosting (AdaBoost)- Freund, Y., Schapire, R. E. (1997)