BỘ GIÁO DỤC VÀ ĐÀO TẠO

TRƯỜNG ĐẠI HỌC NGOẠI NGỮ - TIN HỌC TP. HỒ CHÍ MINH

KHOA CÔNG NGHỆ THÔNG TIN

ĐỒ ÁN MÔN HỌC

KHAI KHOÁNG DỮ LIỆU

ĐỀ TÀI

DỰ ĐOÁN BỆNH TIỂU ĐƯỜNG DỰA TRÊN THUẬT TOÁN FREESPAN

GIẢNG VIÊN HƯỚNG DẪN: ThS. Võ Thị Hồng Tuyết

SINH VIÊN THỰC HIỆN:

Chu Đặng Bình An – 21DH113175

Bùi Tuấn Đạt – 21DH113218

Nguyễn Đức Huân – 21DH113649

Thành phố Hồ Chí Minh, Tháng 07/2024

BỘ GIÁO DỤC VÀ ĐÀO TẠO

TRƯỜNG ĐẠI HỌC NGOẠI NGỮ - TIN HỌC TP. HỒ CHÍ MINH

KHOA CÔNG NGHỆ THÔNG TIN

ĐỒ ÁN MÔN HỌC

KHAI KHOÁNG DỮ LIỆU

ĐỀ TÀI

DỰ ĐOÁN BỆNH TIỂU ĐƯỜNG DỰA TRÊN THUẬT TOÁN FREESPAN

GIẢNG VIÊN HƯỚNG DẪN: ThS. Võ Thị Hồng Tuyết

SINH VIÊN THỰC HIỆN:

Chu Đặng Bình An – 21DH113175

Bùi Tuấn Đạt – 21DH113218

Nguyễn Đức Huân – 21DH113649

Thành phố Hồ Chí Minh, Tháng 07/2024

LỜI CẢM ƠN

Lời đầu tiên, chúng em xin bày tỏ sự cảm ơn chân thành đối với giảng viên hướng dẫn báo cáo môn học cho chúng em, ThS Võ Thị Hồng Tuyết. Cô là người đã hết sức tận tình chỉ dạy và hướng dẫn trong suốt quá trình tìm hiểu, nghiên cứu và thực hiện bài báo cáo môn học khai khoáng dữ liệu này.

Chúng em xin chân thành cảm ơn các thầy cô, giảng viên Khoa Công nghệ Thông tin, Trường Đại học Ngoại ngữ - Tin học TP. Hồ Chí Minh, đã nhiệt tình giảng dạy, tạo điều kiện học tập và nghiên cứu tốt nhất, để trau dồi thêm kiến thức trong suốt thời gian học tập.

Cuối cùng, chúng em xin chân thành cảm ơn gia đình, những người thân và bạn bè đã quan tâm, động viên trong suốt thời gian học tập vừa qua.

Xin trân trọng cảm ơn!

BẢNG ĐÁNH GIÁ MỨC ĐỘ HOÀN THÀNH

CÔNG VIỆC ĐƯỢC GIAO	Chu Đặng Bình An- 21DH113175	Bùi Tuấn Đạt - 21DH113218	Nguyễn Đức Huân - 21DH113649
Viết báo cáo	60%	20%	20%
Tìm hiểu các công trình nghiên cứu	100%		
Tiền xử lý dữ liệu	100%		
Thuật toán Apriori 1. Làm thủ công 2. Code API 3. Code lại thuật toán			 100% 100%
Thuật toán DecisionTree 1. Làm thủ công 2. Code API 3. Code lại thuật toán		 100% 100% 100%	
Thuật toán FreeSpan 1. Làm thủ công 2. Code lại thuật toán	 100% 100%		

MỤC LỤC

DANH MỤC HÌNH ẢNH

DANH MỤC BẢNG BIỂU

BẢNG TỪ VIẾT TẮT

Từ viết tắt	Nội dung
FreeSpan	Frequent Pattern-projected Sequential Pattern Mining

CHƯƠNG 1. GIỚI THIỆU

1.1 Giới thiệu đề tài

Bệnh tiểu đường là một trong những căn bệnh mãn tính phổ biến và nghiêm trọng nhất hiện nay, căn bệnh này ảnh hưởng đến hàng triệu người trên toàn thế giới. Sự gia tăng các ca bệnh tiểu đường đặt ra rất nhiều thách thức cho hệ thống y tế trên toàn thế giới. Mặt khác, cùng với sự phát triển của xã hội, việc sinh hoạt không lành mạnh như tiêu thụ thức ăn nhanh, bia rượu hay thiếu vận động,… của đa số người dân hiện này, đặc biệt là giới trẻ là một trong những tác nhân làm gia tăng căn bệnh tiểu đường này. Với những biến chứng nguy hiểm, nặng nề ảnh hưởng đến nhiều cơ quan quan trọng điển hình như các bệnh về tim (tăng huyết áp, xơ vữa động mạch,…), đột quỵ, thần kinh, thận (suy thận,…),…. Đó là những lý do khiến cho căn bệnh tiểu đường trở thành mối đe dọa nghiêm trọng đến sức khỏe của cộng đồng. Trong bối cảnh đó, việc áp dụng các phương pháp khoa học, kỹ thuật hiện đại để chẩn đoán hay dự đoán khả năng mắc bệnh tiểu đường của bệnh nhân giúp cho việc đưa ra những biện pháp phòng tránh, ngăn ngừa căn bệnh với những biến chứng nghiêm trọng, khôn lường.

Với đề tài: **Dự đoán bệnh tiểu đường dựa trên thuật toán FreeSpan,** chúng em mong muốn tìm hiểu, nghiên cứu về các phương pháp, kỹ thuật khai phá luật kết hợp, khai thác mẫu trên tập dữ liệu của bệnh tiểu đường để tiến hành xây dựng các mô hình học máy đưa ra dự đoán về căn bệnh tiểu đường. Thuật toán FreeSpan (Frequent Pattern-projected Sequential Pattern Mining) là phương pháp khai phá dữ liệu dùng để tìm kiếm các mẫu tuần tự phổ biến trong cơ sở dữ liệu. Với khả năng xử lý hiệu quả và tìm ra các mẫu tuần tự có ý nghĩa, FreeSpan có thể được ứng dụng để phân tích và dự đoán nguy cơ mắc bệnh tiểu đường dựa trên dữ liệu y tế về căn bệnh tiểu đường này.

1.2 Nội dung thực hiện

Để đạt được mục tiểu của đề tài "Dự đoán bệnh tiểu đường dựa trên thuật toán FreeSpan", nội dung thực hiện sẽ bao gồm:

- Nghiên cứu tổng quan về bệnh tiểu đường và các công trình nghiên cứu hay phương pháp dự đoán hiện có.
- Nghiên cứu về thuật toán FreeSpan: Tìm hiểu chi tiết, cách hoạt động và ứng dụng của thuật toán FreeSpan trong khai thác mẫu.
- Thu thập và xử lý dữ liệu: Thu thập dữ liệu y tế về căn bệnh tiểu đường từ các nguồn dữ liệu tin cậy. Xử lý và làm sạch dữ liệu để chuẩn bị cho quá trình phân tích và dự đoán.
- Áp dụng thuật toán FreeSpan trên tập dữ liệu đã thu thập. Tìm kiếm các mẫu tuần tự phổ biến trong tập dữ liệu, từ đó xác định các yếu tố nguy cơ và các mẫu liên quan đến nguy cơ mắc bệnh tiểu đường.
- Phân lớp hoặc gom cụm trên tập dữ liệu y tế thu thập được về căn bệnh tiểu đường.
- Đánh giá và đưa ra kết quả.

1.3 Giới hạn đề tài

Đề tài sẽ tập trung vào các giới hạn sau:

- Phạm vi dữ liệu: Dữ liệu sử dụng trong báo cáo sẽ giới hạn ở các yếu tố, đặc điểm của tập dữ liệu như tuổi, giới tính, đường huyết, huyết áp, chỉ số khối cơ thể, lịch sử hút thuốc. Các yếu tố khác có thể ảnh hưởng đến nguy cơ mắc bệnh tiểu đường sẽ không được đề cập đến trong báo cáo này.
- Các thuật toán khai phá luật kết hợp, khai thác mẫu: Bài báo cáo sẽ tập trung vào thuật toán khai phá luật kết hợp Apriori và thuật toán khai thác mẫu FreeSpan.
- Thuật toán phân lớp: Bài báo cáo sẽ sử dụng thuật toán DecisionTree để thực hiện phân lớp cho dữ liệu.

1.4 Bố cục báo cáo

Bố cục của bài báo cáo với đề tài "Dự đoán bệnh tiểu đường dựa trên thuật toán FreeSpan" được xây dựng với cấu trúc các chương như sau:

CHƯƠNG 1: GIỚI THIỆU: Giới thiệu tổng quan về đề tài các nội dung nghiên cứu và hiện thực trên tập dữ liệu bệnh tiểu đường.

CHƯƠNG 2: CƠ SỞ LÝ THUYẾT: Đưa ra các khái niệm liên quan, các công trình nghiên cứu liên quan đến đề tài.

CHƯƠNG 3: DỰ ĐOÁN BỆNH TIỂU ĐƯỜNG DỰA TRÊN THUẬT TOÁN FREESPAN: Trình bày các bước thực hiện đề tài, thực hiện bằng phương pháp thủ công, thư viện và xây dựng lại các thuật toán khai thác luật kết hợp Apriori, khai thác mẫu FreeSpan và kỹ thuật phân lớp Decision Tree

CHƯƠNG 4: KẾT QUẢ - KẾT LUẬN: Nhận xét về kết quả thực hiện được, hướng phát triển tương lai của đề tài.

TÀI LIỆU THAM KHẢO: Các tài liệu tham khảo theo chuẩn IEEE

CHƯƠNG 2: CƠ SỞ LÝ THUYẾT

2.1 Các khái niệm cơ bản

2.1.1 Bệnh tiểu đường

Theo Bộ Y tế Việt Nam năm 2017, bệnh tiểu đường hay còn gọi là đái tháo đường được định nghĩa là bệnh rối loạn chuyển hóa không đồng nhất, có đặc điểm tăng glucose huyết do khiếm huyết về tiết insulin, về tác động insulin hoặc cả hai. Tăng glucose mạn tính trong thời gian dài gây nên những rối loạn chuyển hóa carbohydrate, protide, lipide, gây tổn thương ở nhiều cơ quan khác nhau, đặc biệt ở tim và mạch máu, thận, mắt, thần kinh [1].

2.1.2 Thuật toán FreeSpan

FreeSpan (Frequent Pattern – Projected Sequential Pattern Mining) [2] là một phương pháp khai thác mẫu tuần tự với ý tưởng là sử dụng các mục phổ biến để chiếu đệ quy trên cơ sở dữ liệu tuần tự vào một tập hợp các cơ sở dữ liệu được chiếu nhỏ hơn và phát triển các đoạn chuỗi con trong mỗi cơ sở dữ liệu được chiếu. Quá trình này phân vùng cả dữ liệu và tập hợp các mẫu thường xuyên được kiểm tra và giới hạn mỗi thử nghiệm được tiến hành trên cơ sở dữ liệu dự kiến nhỏ hơn tương ứng.

Nguyên lý hoạt động của thuật toán như sau:

1. Quét cơ sở dữ liệu ban để tìm các mục phổ biến (frequent item) và sắp xếp chúng giảm dần theo tần suất để tạo danh sách mục phổ biến (f-list).
2. Khai thác các mẫu tuần tự phổ biến bằng cách chiếu xen kẽ (alternative-level projection) với các bước sau: (1) Xây dựng ma trận mục phổ biến bằng cách quét cơ sở dữ liệu 1 lần, (2) tạo các mẫu tuần tự có độ dài 2 và các chú thích trên các mẫu lặp lại (item repeating), (3) quét cơ sở dữ liệu để tạo ra các mẫu lặp lại và cơ sở dữ liệu được chiếu và (4) thực hiện phép chiếu ma trận trên cơ sở dữ liệu được chiếu bằng cách đệ quy, nếu vẫn còn một mẫu dài hơn đang được khai thác.

Các bước thực hiện cụ thể như sau:

- Bước 1: Tìm mục phổ biến (Frequent Items), tạo danh sách f_list và sắp xếp lại theo tần suất theo chiều giảm dần.
- Bước 2: Xây dựng ma trận chiếu (Projection Matrix) để tìm chuỗi có độ dài 2 được hình thành từ các mục phổ biến trong f_list. Xây dựng ma trận với số lần xuất hiện của các mẫu tuần tự khác nhau dựa trên các mục phổ biến.
    - Một ma trận tam giác F[j, k] với 1 <= j <= m và 1 <= k <= j, với m là số mục phổ biến, j là mục phổ biến thứ nhất, k là mục phổ biến thứ 2. Nếu F[j, j] thì chỉ có 1 bộ đếm mà j < m. Còn lại thì F[j, k] sẽ có 3 bộ đếm (A, B, C).
    - 3 bộ đếm (A, B, C) thể hiện như sau: A thể hiện cho số lần xuất hiện của k đứng sau j. B thể hiện cho số lần xuất hiện của k đứng trước j. C thể hiện cho số lần xuất hiện đồng thời của j, k tức là j, k là 1 chuỗi con trong chuỗi mẹ.
- Bước 3: Tạo chuỗi các mẫu có độ dài 2 (Generate 2-item Sequence).
- Bước 4: Chú thích các tập item – repeating.
- Bước 5: Xây dựng cơ sở dữ liệu chiếu.
- Bước 6: Khai thác các mẫu từ các cơ sở dữ liệu chiếu

Ví dụ: *Cho 1 cơ sở dữ liệu như sau:*

id	Sequence
S1	a,b,c
S2	a,c,d
S3	a,b,c,d
S4	b,c

Bảng 1. Bảng dữ liệu ví dụ cho thuật toán FreeSpan

Khai thác mẫu với minSup = 3.

Bước 1: Tìm tần suất xuất hiện và xây dựng f_list với minSup = 3

Xây dựng mục phổ biến {a:3, b:3, c:4, d:2} → loại {d:2} vì không thỏa điều kiện.

id	Sequence
S1	a,b,c
S2	a,c
S3	a,b,c
S4	b,c

Bảng 2. Cơ sở dữ liệu sau khi loại phần tử không đạt điều kiện

Vậy f_list được xây dựng và sắp xếp lại như sau: {c:4, a:3, b:3}.

Bước 2: Tính toán và xây dựng ma trận chiếu trên cơ sở dữ liệu:

Bắt đầu với itemset có tần suất cao nhất là c: c → a, c → b, c → c

+ S1: c đứng sau a, b

+ S2: c đứng sau a

+ S3: c đứng sau a, b

+ S4: c đứng sau b

+ Kết quả thu được a → c xuất hiện 3 lần, b → c xuất hiện 3 lần còn c → a, c → b, c → c không có lần nào

Tiếp theo với itemset a: a → a, a → b, a → c

+ S1: a đứng trước b,c

+ S2: a đứng trước c

+ S3: a đứng trước b,c

+ S4: không có a

+ Kết quả thu được là a → c xuất hiện 3 lần, a → b xuất hiện 2 lần còn a → a không có lần nào.

Tiếp theo với itemset b: b → a, b → b, b → c

+ S1: b đứng sau a, đứng trước c

+ S2: không có b

+ S3: b đứng sau a, đứng trước c

+ S4: b đứng trước c

+ Kết quả thu được là b → c xuất hiện 3 lần, b → a xuất hiện 0 lần còn b → b không có lần nào.

Vậy ta thu được 1 ma trận như sau:

	c	**a**	**b**
c	0	(0, 3, 0)	(0, 3, 0)
a	(3, 0, 0)	0	(2, 0, 0)
b	(3, 0, 0)	(0, 2, 0)	0

Bảng 3. Ví dụ ma trận chiếu

Bước 3: Tạo các mẫu có độ dài bằng 2 từ ma trận chiếu:

Từ ma trận chiếu ta có các mẫu có độ dài bằng 2 với minSup = 3 như sau:

{a, c} và {b, c}

Bước 4: Chú thích các mục lặp lại (item-repeating):

Từ các mẫu có độ dài bằng 2, tiến hành chú thích đối với những mục lặp lại. Tuy nhiên các chuỗi trong cơ sở dữ liệu trên không có mục nào lặp lại nên chúng ta bỏ qua.

Bước 5: Xây dựng cơ sở dữ liệu chiếu bằng cách chiếu các mục phổ biến lên chuỗi:

Khai thác mẫu phổ biến có độ dài bằng 2:

Bắt đầu với việc chiếu mục phổ biến c lên từng chuỗi:

+ S1: không có + S2: không có

+ S3: không có + S4: không có

Tiếp tục chiếu mục phổ biến a lên từng chuỗi:

+ S1: {b}, {c} + S2: {c}

+ S3: {b}, {c} + S4: không có

⇨ Vậy mẫu phổ biến khi chiếu a là {a, c}: 3; {a, b}: 2 (loại vì không đạt minSup)

Tiếp tục chiếu mục phổ biến b lên từng chuỗi:

+ S1: {c} + S2: không có

+ S3: {c} + S4: {c}

⇨ Vậy mẫu phổ biến khi chiếu b là {b, c} : 3

Khai thác mẫu phổ biến có độ dài bằng 3:

Tiếp tục, chúng ta sẽ chiếu các mẫu phổ biến có độ dài bằng 2 lên các chuỗi.

Bắt đầu với chiếu {a, c}:

+ S1: không có + S2: không có

+ S3: không có + S4: không có

Tương tự chiếu các mẫu còn lại thì ta không thu được mẫu phổ biến mới nào.

Bước 6: Tổng kết kết quả:

Vậy mẫu phổ biến của cở sở dữ liệu ban đầu như sau:

{a}, {b}, {c}, {a, c}, {b, c}

2.2 Các công trình nghiên cứu liên quan

2.2.1 Analysis of diabetes mellitus for early prediction using optimal features selection (2019)

Công trình nghiên cứu được thực hiện bởi N. Sneha và Tarun Gangil đến từ Học viện Công nghệ và Quản lý REVA của Ấn Độ vào năm 2019. [3]

Nội dung:

Bài nghiên cứu được thực hiện trên tập dữ liệu được thu thập từ nguồn lưu trữ UCI (Home - UCI Machine Learning Repository - Diabetes) với 15 thuộc tính và 2500 dòng dữ liệu.

Sl. no	Attribute	Description
1	Age	Age of a person
2	Gender	Male or female
3	Plasma glucose fasting	–
4	Plasma glucose post prandial	–
5	Pregnancy	Pregnancy count of women
6	Blood glucose level	Plasma glucose concentration a 2 h in an oral glucose tolerance test
7	Blood pressure	Diastolic blood pressure (mm Hg)
8	Skin thickness	Triceps skin fold thickness (mm)
9	Insulin	2-h serum insulin (mu U/ml)
10	BMI (body mass index)	Body mass index (weight in kg/(height in m)^2)
11	DPF	Diabetes pedigree function
12	Serum creatinine	Test measures the level of creatinine in the blood
13	Serum sodium	sodium content is in your blood
14	Serum potassium	Potassium content in blood
15	HBAIC	Hemoglobin A1c, a blood pigment that carries oxygen

Hình 1. Mô tả của tập dữ liệu bệnh tiểu đường UCI

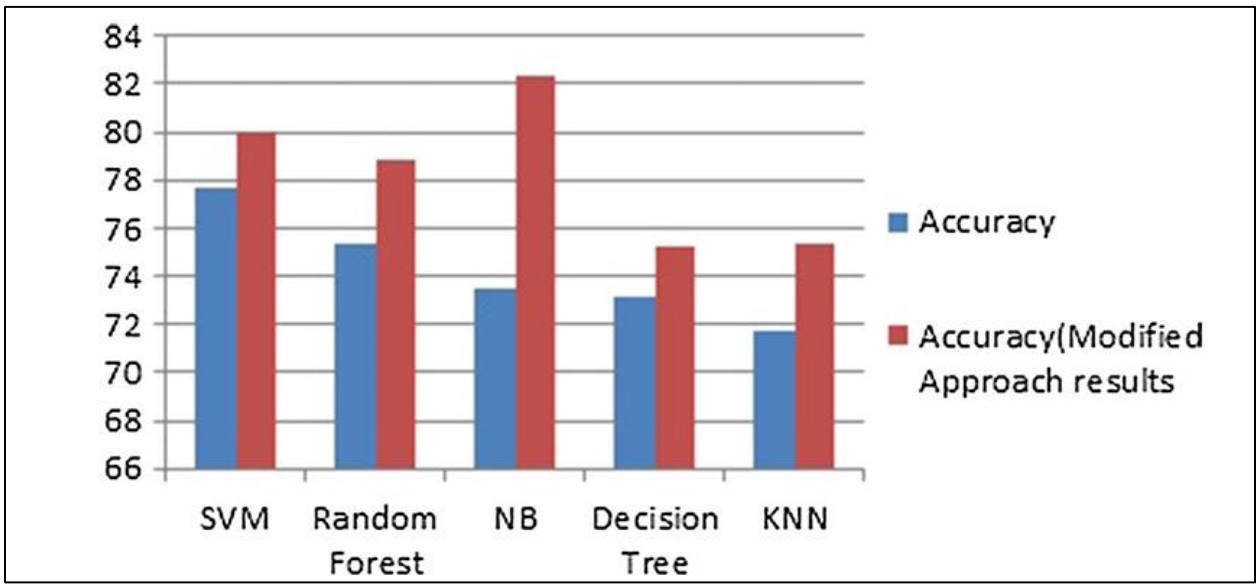
Công trình đề cập đến những nghiên cứu xoay quanh về bệnh tiểu đường như tỉ lệ mắc bệnh giữa nam và nữ, phân loại tiểu đường và các biến chứng của bệnh. Ngoài ra, công trình nghiên cứu sâu vào các phương pháp, kỹ thuật của lĩnh vực khoa học dữ liệu như khai phá dữ liệu, học máy,… để ứng dụng vào việc dự đoán. Các giải thuật, mô hình

như Decision Tree, Naïve Bayesian, Support Vectort Machine (SVM), Random Forest, K- Nearest Neighbour (KNN) để đưa ra dự đoán trên tập dữ liệu mà họ đã thu thập.

Sl. no	Classification technique	Accuracy	Correctly classified	Incorrectly classified
1	SVM	77.73	597	171
2	Random forest	75.39	579	189
3	NB	73.48	129	61
4	Decision tree	73.18	562	206
5	KNN	63.04	145	85

Hình 2. Kết quả của các kỹ thuật phân lớp của bài nghiên cứu năm 2019

Sau đó họ tiến hành cải tiến, sửa đổi các kỹ thuật, mô hình để tăng độ chính xác bằng cách chọn ra 11 thuộc tính có độ tương quan cao và loại đi 4 thuộc tính có độ tương quan thấp. Kết quả bài nghiên cứu thu được với độ chính xác cao của các mô hình.



Hình 3. So sánh kết quả trước và sau khi cải tiến của bài nghiên cứu năm 2019

2.2.2 An Improved Artificial Neural Network Model for Effective Diabetes Prediction (2021)

Công trình nghiên cứu được đăng trên tạp chí khoa học Willey Online Library thực hiện bởi các chuyên gia đến từ các trường đại học thuộc các quốc gia Pakistan, Saudi Arabia vào năm 2021. [4]

Nội dung:

Bài nghiên cứu được thực hiện với nguồn dữ liệu thu thập từ Viện Quốc gia về Bệnh Tiểu đường, Tiêu hóa và Bệnh Thận của Hoa Kỳ. Với đầu vào của tập dữ liệu gồm 9 thuộc tính liên quan đến việc chẩn đoán bệnh tiểu đường.

S/N	Input attributes	Description	Range values
1	Pregnancies	Number of times pregnant	0–17
2	Glucose	Plasma glucose concentration 2 hours in an oral glucose tolerance test	0–199
3	Blood pressure	Diastolic blood pressure (mm Hg)	0–122
4	Skin thickness	Triceps skinfold thickness (ram)	0–99
5	Insulin	2-hour serum insulin (mu U/m1)	0–846
6	BMI	Body mass index (weight in kg/(height in m)^2)	0–67.1
7	Diabetes pedigree function	Diabetes pedigree function	0.078–2.42
8	Age	Age (years)	21–81
		Output/responder variable	
Sr.	Input attributes	Description	Range values
1	Outcome	Diabetes, yes or no	01

Hình 4. Mô tả dữ liệu bệnh tiểu đường của Viện Quốc gia về Bệnh Tiểu đường, Tiêu hóa và Thận

Bài nghiên cứu tập trung vào xây dựng mạng nơ-ron nhân tạo (ANN) và cải tiến mạng nơ-ron bằng cách kết hợp lan truyền ngược với phương pháp gradient liên hợp, gọi tắt là ABP-SCGNN để tối ưu hóa quá trình huấn luyện, cải thiện hiệu suất của mô hình. Các nhà nghiên cứu đã thực hiện điều chỉnh số lượng nơ-ron trong lớp ẩn từ 5 đến 50 và thu được mô hình với độ chính xác cao nhất là 93% với 20 nơ-ron trong lớp ẩn. Tác giả sử dụng các thang đo độ chính xác, sai số toàn phương trung bình (MSE) để đánh giá mô hình.

Algorithm	Accuracy (%)
BFGS [1]	88.8
Genetic algorithm [8]	87
GRNN [9]	80.21
ABP-SCGNN (proposed)	93

Hình 5. So sánh kết quả của mô hình cải tiến ABP-SCGNN với các mô hình khác

2.2.3 AI – based smart prediction of clinical disease using random forest classifier and Naïve Bayes (2020)

Công trình nghiên cứu được đăng tải vào ngày 4 tháng 11 năm 2020, thực hiện bởi V. Jackins, S. Vimal, M. Kaliappan, Mi Young Lee đến từ các trường đại học thuộc hai quốc gia Ấn Độ và Hàn Quốc. [5]

Nội dung:

Bài nghiên cứu được thực hiện trên nguồn dữ liệu được thu thập từ Viện Quốc gia về Bệnh Tiểu đường, Tiêu hóa và Thận của Hoa Kỳ. Tập dữ liệu này được thu thập từ các bệnh nhân nữ từ 21 tuổi trở lên của người Pima Indian. Tập dữ liệu gồm 9 thuộc tính được mô tả như sau:

Pregnancies: baby deliveries happened in number of times
Glucose: the concentration test in glucose using the tolerance test for every 2
BP: diastolic BP (mm Hg)
Skin thickness: the thickness of skin in triceps fold (mm)
Insulin: insulin serum for 2-h (mu U/ml)
BMI: height/weight
Diabetes: prediction in mm
Age: in years
Outcome: true class variable either (0 or 1)

Hình 6. Mô tả về tập dữ liệu bệnh tiểu đường của phụ nữ người Pima Indian

Các nhà nghiên cứu đã để cập và giới thiệu sơ qua các mô hình như mạng nơ-ron nhân tạo (ANN), SVM, K-means,… trong việc xây dựng ứng dụng vào việc dự đoán bệnh tiểu đường. Mặt khác, các nhà nghiên cứu xoáy sâu vào hai giải thuật phân lớp đó là Random Forest (RF) và Naïve Bayes (NB) trong bài nghiên cứu này. Thông qua việc xử lý dữ liệu, chia dữ liệu để huấn luyện mô hình và các phương pháp đánh giá mô hình. Bài nghiên cứu đã cho ra kết quả với thuật toán Random Forest có độ chính xác cao hơn thuật toán Naïve Bayes. Điều này làm cho thuật toán có thể trở thành lựa chọn ưu tiến để dự đoán căn bệnh tiểu đường.

Bài nghiên cứu này tổng quan cho ta thấy tầm quan trọng của việc ứng dụng học máy vào y tế, cải thiện việc chẩn đoán bệnh.

2.2.4 A Scoping Review of Artificial Intelligence-Based Methods for Diabetes Risk Prediction (2023)

Công trình nghiên cứu được thực hiện bởi Farida Mohsen, Hamada R. H. Al–Absi, Noha A. Yousri, Nady El Haij, Zubair Shah đến từ các trường đại học, các quỹ nghiên cứu của hai quốc qua Qatar và Ai Cập. [6]

Nội dung:

Công trình này trình bày tổng quan về sự gia tăng của T2DM – đái tháo đường type 2 và những biến chứng của nó. Tổng hợp những công trình nghiên cứu hiện có trên các nguồn khác nhau như IEEE-Xplore, Google Scholar, Scopus,… với khoảng 40 nghiên cứu được đưa vào bài đánh giá. Đưa ra tổng quan các nghiên cứu liên quan đến các giải thuật phân lớp, mô hình học máy, mô hình học sâu, mô hình đơn kênh và mô hình đa kênh ứng dụng vào việc dự đoán bệnh tiểu đường. Bài nghiên cứu còn chỉ ra những thách thức, khó khăn trong việc tích hợp và phát triển để đưa các mô hình AI vào ứng dụng dự đoán bệnh bao gồm các vấn đề như hiệu suất, độ chính xác,…

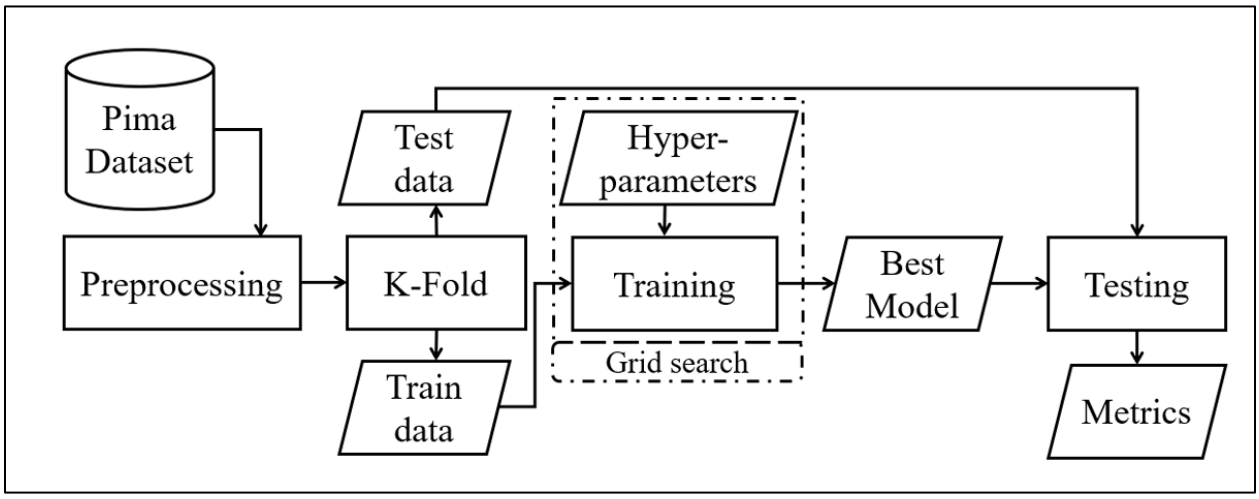
2.2.5 Diabetes Prediction Using Ensembling of Different Machine Learning Classifiers (2020)

Công trình nghiên cứu được đăng tải vào tháng 5 năm 2020, thực hiện bởi MD. Kamrul Hasan, MD. Ashraful Alam, Dola Das, Eklas Hossain, Mahmudul Hasan và thành viên của IEEE. [7]

Nội dung:

Được thực hiện trên nguồn dữ liệu Pima Indian Diabetes. Với mục tiêu tìm ra những thuật toán, mô hình học máy, học sâu tốt nhất để ứng dụng vào việc dự đoán bệnh tiểu đường. Bài nghiên cứu được thực hiện từng bước như sau:

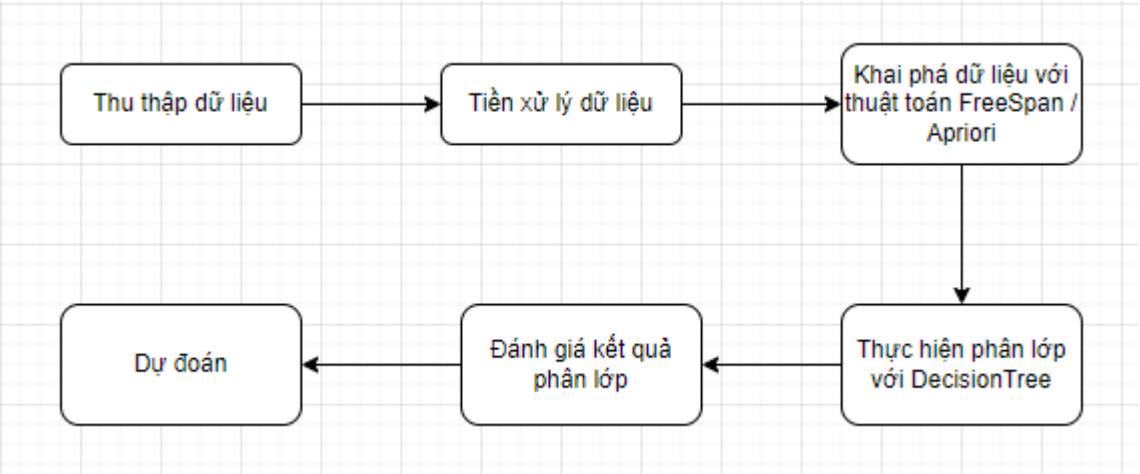
- Tiền xử lý dữ liệu, loại bỏ các giá trị thiếu.
- Chuẩn hóa dữ liệu, trích chọn đặc trưng.
- Áp dụng kỹ thuật K-fold cross-validation để đánh giá mô hình.
- Sử dụng nhiều mô hình như: Naïve Bayes, XGBoost, KNN, Decision Tree, RF, AdaBoost, Multilayer Perceptron.
- Đề xuất phương pháp tổ hợp trọng số với trọng số ước tính từ diện tích đường cong ROC(AUC) của từng mô hình.
- Tối ưu hóa tham số bằng kỹ thuật grid search.



Hình 7. Sơ đồ các bước thực hiện của bài nghiên cứu

CHƯƠNG 3: DỰ ĐOÁN BỆNH TIỂU ĐƯỜNG DỰA TRÊN THUẬT TOÁN FREESPAN

3.1 Tổng quan phương pháp hiện thực



Hình 8. Sơ đồ tổng quan các bước thực hiện

3.1.1 Mô tả dữ liệu

- Đọc dữ liệu gồm những thông tin: thuộc tính, kiểu dữ liệu, số lượng.
- Hiển thị những thông tin liên quan.
- Kiểm tra dữ liệu có xuất hiện nhiễu hay không.

3.1.2 Tiền xử lý dữ liệu

- Xử lý nhiễu bằng cách loại bỏ nhiễu hoặc thay thế.
- Chuyển dữ liệu về dạng nominal: Chia bin rồi gắn nhãn đối với những dữ liệu dạng numeric.

3.1.3 Chuẩn hóa và thực hiện khai phá dữ liệu

- Chuẩn hóa dữ liệu để phù hợp với các mô hình, thuật toán.
- Thực hiện thuật toán khai phá luật kết hợp Apriori trên tập dữ liệu.
- Thực hiện thuật toán khai thác mẫu FreeSpan trên tập dữ liệu.
- Triển khai chia tập dữ liệu train và test.
- Tiến hành thực hiện thuật toán Decision Tree để thực hiện phân lớp cho tập dữ liệu.

- Trực quan hóa bằng biểu đồ các lớp.
- Đánh giá mô hình.

3.2 Công cụ hiện thực

Bài báo cáo được thực hiện trên môi trường Python 3 trên nền Jupyter Notebook của Google Colaboratory. Đồng thời sử dụng những thư viện để hỗ trợ như: Scikit-learn, Mathplotlib, Seaborn, NumPy, Pandas, mlxtend.

3.3 Tập dữ liệu

Tập dữ liệu của bài cáo lần này được thu thập trên Kaggle với nguồn dữ liệu: Diabetes prediction dataset – Mohammed Mustafa (2023). [8]

Tập dữ liệu **Diabetes prediction dataset** là tập hợp dữ liệu gồm 9 thuộc tính và 100,000 dòng dữ liệu về lĩnh vực sức khỏe – y tế và nhân khẩu học từ các bệnh nhân, cùng với trạng thái về căn bệnh tiểu đường của họ.

Dữ liệu gồm những thông tin cơ bản sau:

STT	Đặc trưng	Kiểu dữ liệu	Ý nghĩa	Ví dụ
1	gender	Nominal	Thể hiện giới tính của bệnh nhân	Female, Male, Other
2	age	Numeric	Tuổi của bệnh nhân	80.0, 36.0, …
3	hypertension	Numeric	Bệnh nhân có bị huyết áp cao hay không?	0, 1
4	heart_disease	Numeric	Bệnh nhân có bị bệnh về tim hay không?	0, 1
5	smoking_history	Nominal	Lịch sử hút thuốc lá của bệnh nhân	Former, never, current,

				not current
6	bmi	Numeric	Chỉ số khối cơ thể được tính bằng: $\frac{Khối\ lượng\ cơ\ thể\ (kg)}{Bình\ phương\ chiều\ cao\ (m)}$	25.19, 27.32, …
7	HbA1c_level	Numeric	Chỉ số hemoglobin loại A1c, để chẩn đoán bệnh tiểu đường	6.6, 5.7, …
8	blood_glucose_level	Numeric	Mức đường huyết, là lượng đường có trong máu của bệnh nhân	100, 120, 140,…
9	diabetes	Numeric	Bệnh nhân có bị tiểu đường hay không?	0, 1

Bảng 4. Bảng mô tả thông tin cơ bản về các thuộc tính của dữ liệu

3.4 Tiền xử lý dữ liệu

Bước 1: Đọc dữ liệu và hiển thị một số thông tin liên quan.

Bước 1.1 Đọc dữ liệu vào:

```
RangeIndex: 100000 entries, 0 to 99999
Data columns (total 9 columns):
 #   Column               Non-Null Count   Dtype
---  ------               --------------   -----
 0   gender               100000 non-null  object
 1   age                  100000 non-null  float64
 2   hypertension         100000 non-null  int64
 3   heart_disease        100000 non-null  int64
 4   smoking_history      100000 non-null  object
 5   bmi                  100000 non-null  float64
 6   HbA1c_level          100000 non-null  float64
 7   blood_glucose_level  100000 non-null  int64
 8   diabetes             100000 non-null  int64
dtypes: float64(3), int64(4), object(2)
```

Hình 9. Kết quả hiển thị thông tin cơ bản về tập dữ liệu

Thấy được trên dữ liệu, với 100.000 dữ liệu. Trong đó, có 2 thuộc tính *gender* và *smoking_history* là kiểu dữ liệu định danh (nominal). Đối với 7 dữ liệu còn lại là kiểu dữ liệu số (numeric): 3 thuộc tính kiểu số thực và 4 thuộc tính kiểu số nguyên.

Bước 1.2 Hiển thị 10 dòng đầu của dữ liệu:

	gender	age	hypertension	heart_disease	smoking_history	bmi	HbA1c_level	blood_glucose_level	diabetes
0	Female	80.0	0	1	never	25.19	6.6	140	0
1	Female	54.0	0	0	No Info	27.32	6.6	80	0
2	Male	28.0	0	0	never	27.32	5.7	158	0
3	Female	36.0	0	0	current	23.45	5.0	155	0
4	Male	76.0	1	1	current	20.14	4.8	155	0
5	Female	20.0	0	0	never	27.32	6.6	85	0
6	Female	44.0	0	0	never	19.31	6.5	200	1
7	Female	79.0	0	0	No Info	23.86	5.7	85	0
8	Male	42.0	0	0	never	33.64	4.8	145	0
9	Female	32.0	0	0	never	27.32	5.0	100	0

Hình 10. Hiển thị 10 dòng dữ liệu đầu tiên của dữ liệu

Hiển thị một số thông tin về dữ liệu mà các các thuộc tính lưu trữ.

Bước 2: Kiểm tra tính duy nhất của dữ liệu trên mỗi thuộc tính.

Bước 2.1 Kiểm tra tổng quan số lượng tính duy nhất dữ liệu trên mỗi mỗi tính:

```
gender                3
age                 102
hypertension          2
heart_disease         2
smoking_history       6
bmi                4247
HbA1c_level          18
blood_glucose_level  18
diabetes              2
```

Hình 11. Số lượng dữ liệu mang tính duy nhất của từng thuộc tính

Thực hiện kiểm tra tính độc nhất của toàn bộ dữ liệu để xác định tổng quan về dữ liệu có bao nhiêu thể hiện của từng thuộc tính trong tập dữ liệu này. Xác định được khoảng giá trị thể hiện của từng thuộc tính đó. Theo kết quả, chúng ta sẽ tiến hành tiếp tục thực hiện kiểm tra những dữ liệu độc nhất gồm những thể hiện nào đối với những

thuộc tính có số lượng dữ liệu độc nhất thấp gồm: *gender, hypertension, heart_disease, smoking_history, diabetes.*

Bước 2.2 Kiểm tra dữ liệu độc nhất của các thuộc tính:

```
['Female' 'Male' 'Other']
[0 1]
[1 0]
['never' 'No Info' 'current' 'former' 'ever' 'not current']
[0 1]
```

Hình 12. Kết quả kiểm tra tính duy nhất của 6 thuộc tính

Kết quả kiểm tra cho thấy tính duy nhất của 6 thuộc tính theo thứ tự: *gender, hypertension, heart_disease, smoking_history, diabetes.*

- Với 3 thuộc tính *hypertension, heart_disease, diabetes* đều có kết quả là giá trị [0, 1], tức là điều này thể hiện trạng thái tương đương đó là [No, Yes].
- Với thuộc tính *gender* trả về kết quả là [‘Female’, ‘Male’, ‘Other’] tương đương với giới tính [‘Nữ’, ‘Nam’, ‘Khác’]. Tuy nhiên, đây là dữ liệu liên quan đến y học hay nói cách khác về mặt sinh học thì chúng ta chỉ xét ở hai giới tính nam hoặc nữ. Cho nên đối với dữ liệu [‘Other’] nằm ở thuộc tính *gender* chúng ta sẽ thực hiện loại bỏ những dòng dữ liệu có dữ liệu đấy.
- Với thuộc tính *smoking_history* trả về kết quả là ['never', 'No Info', 'current', 'former', 'ever', 'not current']. Ở đây, chúng ta có dữ liệu ‘No Info’, tức là không có thông tin gì liên quan đến việc bệnh nhân có hút thuốc hay không. Vì tần suất hút thuốc của bệnh nhân cũng là một trong những tác nhân ảnh hưởng đến chẩn đoán bệnh tiểu đường. Cho nên, tương tự ở *gender,* chúng ta sẽ loại bỏ những dòng dữ liệu chứa thông tin ‘No Info’ của thuộc tính *smoking_history.*

Bước 3: Loại bỏ những dòng dữ liệu không hợp lệ

```
Index: 64172 entries, 0 to 99999
Data columns (total 9 columns):
 #   Column               Non-Null Count  Dtype
---  ------               --------------  -----
 0   gender               64172 non-null  object
 1   age                  64172 non-null  float64
 2   hypertension         64172 non-null  int64
 3   heart_disease        64172 non-null  int64
 4   smoking_history      64172 non-null  object
 5   bmi                  64172 non-null  float64
 6   HbA1c_level          64172 non-null  float64
 7   blood_glucose_level  64172 non-null  int64
 8   diabetes             64172 non-null  int64
dtypes: float64(3), int64(4), object(2)
```

Hình 13. Tập dữ liệu sau khi tiến hành xóa bỏ những dữ liệu không hợp lệ

Như đã trình bày ở bước 2, chúng ta tiến hành loại bỏ những dòng dữ liệu của 2 thuộc tính *gender, smoking_history*. Sau khi thực hiện loại bỏ chúng ta còn 64.172 dòng dữ liệu.

Bước 4: Thực hiện gắn nhãn dữ liệu phương pháp phân hoạch theo khoảng cách (Binning) với thuộc tính có kiểu dữ liệu là dạng số với miền giá trị trải dài.

Bước 4.1 Hiển thị thống kê mô tả của tập dữ liệu

	age	hypertension	heart_disease	bmi	HbA1c_level	blood_glucose_level	diabetes
count	64172.000000	64172.000000	64172.000000	64172.000000	64172.000000	64172.000000	64172.000000
mean	46.547268	0.097909	0.047045	28.424262	5.564279	139.629792	0.109799
std	19.539695	0.297194	0.211738	6.515975	1.095535	42.166693	0.312641
min	0.160000	0.000000	0.000000	10.080000	3.500000	80.000000	0.000000
25%	31.000000	0.000000	0.000000	24.600000	4.800000	100.000000	0.000000
50%	47.000000	0.000000	0.000000	27.320000	5.800000	140.000000	0.000000
75%	61.000000	0.000000	0.000000	31.100000	6.200000	159.000000	0.000000
max	80.000000	1.000000	1.000000	91.820000	9.000000	300.000000	1.000000

Hình 14. Thống kê mô tả của tập dữ liệu

Chúng ta thực hiện thống kê mô tả để xác định min, max của dữ liệu. Sau đó chúng ta tiến hành phân hoạch theo khoảng cách.

Bước 4.2 Chia bin cho thuộc tính *age*

- Đầu tiên, nhìn vào thống kê mô tả, ta thấy giá trị lớn nhất của dữ liệu ở thuộc tính *age* là 80 và nhỏ nhất là 0.16. Tiến hành chia dữ liệu thành 5 bin khác nhau. Thực hiện tính toán cho miền giá trị cho mỗi bin theo công thức như sau:

$$W = \frac{max - min}{N} \qquad (1)$$

Trong đó:

+ *W* là miền giá trị của mỗi bin.

+ *max, min* lần lượt là các giá trị lớn nhất, nhỏ nhất của tập dữ liệu.

+ *N* là số lượng bin muốn chia.

Áp dụng công thức vào để tiến hành chia bin cho thuộc tính *age:*

$$W = \frac{80 - 0.16}{5} \approx 16$$

- Tiến hành chia 5 bin với khoảng cách miền giá trị là 16 như sau:

1. Bin thứ nhất: từ 0 đến 16, gắn nhãn là [‘0 – 16’]
2. Bin thứ hai: từ 16 đến 32, gắn nhãn là [‘16 – 32’]
3. Bin thứ ba: từ 32 đến 48, gắn nhãn là [’32 – 48’]
4. Bin thứ tư: từ 48 đến 64, gắn nhãn là [’48 – 64’]
5. Bin thứ năm: từ 64 trở lên, gắn nhãn là [‘64+’]

Bước 4.3 Chia bin cho thuộc tính *bmi*

- Đối với thuộc tính *bmi* ta sẽ chia khoảng theo tiêu chuẩn được quốc tế công nhận thành 3 bin như sau:

1. Bin thứ nhất (Dưới chuẩn): BMI thấp hơn 18.5, gắn nhãn [‘Underweight’]
2. Bin thứ hai (Đạt chuẩn): BMI từ 18.5 – 25, gắn nhãn [‘Balance’]
3. Bin thứ ba (Thừa cân): BMI từ 25 trở lên, gắn nhãn [‘Overweight’]

**Chú thích*: Trong quy chuẩn quốc tế, trên mức thừa cân còn 2 mức đó là béo và rất béo. Tuy nhiên, ta sẽ gộp chung vào mức thừa cân.

Bước 4.4 Chia bin cho thuộc tính *HbA1c_level*

- Tương tự với *bmi,* chúng ta sẽ chia thành 3 bin với miền giá trị của từng bin được quy định trong quyết định năm 2017 của Bộ Y Tế [1]**:**
    1. Bin thứ nhất: từ 0 đến dưới 5.7, gắn nhãn là [‘HbA1c Normal’]
    2. Bin thứ hai: từ 5.7 đến 6.4, gắn nhãn là [‘Borderline’]
    3. Bin thứ ba: từ 6.5 trở lên, gắn nhãn là [‘High’]

Bước 4.5 Chia bin cho thuộc tính *blood_glucose_level*

- Tương tự với *HbA1c_level,* chúng ta sẽ chia thành 3 bin với miền giá trị của từng bin được quy định trong quyết định năm 2017 của Bộ Y Tế [1]**:**
    1. Bin thứ nhất: từ 0 đến dưới 90, gắn nhãn là [‘Blood Glucose Normal’]
    2. Bin thứ hai: từ 90 đến 125, gắn nhãn là [‘Pre_diabetic’]
    3. Bin thứ ba: từ 125 trở lên, gắn nhãn là [‘Diabetic’]

Bước 5: Gắn nhãn dữ liệu đối với các thuộc tính *hypertension, heart_disease, diabetes* bằng phương pháp mapping.

- Với thuộc tính *hypertension,* chúng ta thực hiện mapping dữ liệu phù hợp sao cho {0: 'No Hypertension', 1: 'Hypertension'}.
- Với thuộc tính *heart_disease,* chúng ta thực hiện mapping dữ liệu phù hợp sao cho {0: 'No Heart Disease', 1: 'Heart Disease'}.
- Với thuộc tính *diabetes,* chúng ta thực hiện mapping dữ liệu phù hợp sao cho {0: 'No', 1: 'Yes'}

	gender	age	hypertension	heart_disease	smoking_history	bmi	HbA1c_level	blood_glucose_level	diabetes
0	Female	64+	No Hypertension	Heart Disease	never	Overweight	High	Diabetic	No
2	Male	16-32	No Hypertension	No Heart Disease	never	Overweight	HbA1c Normal	Diabetic	No
3	Female	32-48	No Hypertension	No Heart Disease	current	Balance	HbA1c Normal	Diabetic	No
4	Male	64+	Hypertension	Heart Disease	current	Balance	HbA1c Normal	Diabetic	No
5	Female	16-32	No Hypertension	No Heart Disease	never	Overweight	High	Blood Glucose Normal	No
6	Female	32-48	No Hypertension	No Heart Disease	never	Balance	Borderline	Diabetic	Yes
8	Male	32-48	No Hypertension	No Heart Disease	never	Overweight	HbA1c Normal	Diabetic	No
9	Female	16-32	No Hypertension	No Heart Disease	never	Overweight	HbA1c Normal	Pre_diabetic	No
10	Female	48-64	No Hypertension	No Heart Disease	never	Overweight	Borderline	Blood Glucose Normal	No
11	Female	48-64	No Hypertension	No Heart Disease	former	Overweight	Borderline	Pre_diabetic	No
12	Female	64+	No Hypertension	No Heart Disease	former	Overweight	HbA1c Normal	Diabetic	No
13	Female	64+	No Hypertension	No Heart Disease	never	Overweight	Borderline	Diabetic	No

Hình 15. Dữ liệu sau khi được gắn nhãn bằng phương pháp Binning và Mapping

3.5 Triển khai khai thác luật kết hợp và khai thác mẫu

3.5.1 Khai thác luật kết hợp với Apriori

3.5.1.1 Triển khai bằng phương pháp thủ công

Triển khai bằng phương pháp thủ công trên 13 dòng dữ liệu đầu tiên.

ID	**Transaction (Itemset)**
0	{Male, 0-16, No hypertension, No heart disease, never, Overweight, Borderline, Diabetic, No}
1	{Male, 32-48, Hypertension, No heart disease, never, Overweight, HbA1c Normal, Diabetic, No}
2	{Female, 16-32, No hypertension, No heart disease, never, Balance, HbA1c Normal, Pre_diabetic, No}
3	{Male, 64+, No hypertension, No heart disease, former, Overweight, High, Diabetic, Yes}
4	{Female, 48-64, No hypertension, No heart disease, never, Underweight, HbA1c Normal, Diabetic, No}

5	{Female, 48-64, No hypertension, No heart disease, former, Overweight, HbA1c Normal, Pre_diabetic, No}
6	{Female, 48-64, No hypertension, No heart disease, current, Balance, High, Blood Glucose Normal, No}
7	{Female, 64+, Hypertension, Heart disease, never, Overweight, High, Diabetic, Yes}
8	{Male, 32-48, No hypertension, No heart disease, not current, Overweight, Borderline, Diabetic, Yes}
9	{Female, 32-48, No hypertension, No heart disease, not current, Balance, HbA1c Normal, Blood Glucose Normal, No}
10	{Female, 0-16, No hypertension, No heart disease, never, Underweight, Borderline, Diabetic, Yes}
11	{Male, 0-16, No hypertension, Heart disease, never, Underweight, Borderline, Diabetic, Yes}
12	{Female, 64+, Hypertension, Heart disease, never, Overweight, Borderline, Pre_diabetic, No}

Bảng 5. Transaction của 13 dòng dữ liệu đầu tiên

1. Tính tần suất của các 1-itemset, cho minSup = 3:

Item	**Count**
Male	5
Female	9
0-16	3
16-32	1
32-48	3
48-64	3
64+	4
No hypertension	10
Hypertension	4
No heart disease	10
Heart disease	3
never	6

former	2
current	2
not current	2
Overweight	6
Balance	3
Underweight	4
Borderline	4
HbA1c Normal	5
High	3
Diabetic	9
Pre_diabetic	3
Blood Glucose Normal	2
No	7
Yes	6

Bảng 6. Bảng tính tần suất 1-itemset theo phương pháp Apriori

- Loại bỏ những 1-itemset không đạt minSup =3:

{Male}: 5 - {Female}: 9

{64+}: 4 - {No hypertension}: 10

{Hypertension}: 4 - {No heart disease}: 10

{never}: 6 - {Overweight}: 6

{Underweight}: 4 - {Borderline}: 4

{HbA1c Normal}: 5 - {Diabetic}: 9

{No}: 7 - {Yes}: 6

2.Tạo các tập 2-itemset từ các 1-itemset phổ biến:

Itemset	**Count**
{Male, No hypertension}	4
{Male, Diabetic}	3
{Female, Diabetic}	6

{Female, No hypertension}	6
{Female, never}	4
{Female, Overweight}	4
{Female, Borderline}	3
{Female, HbA1c Normal}	3
{Female, No}	4
{64+, Diabetic}	3
{64+, Hypertension}	3
{No hypertension, Diabetic}	6
{Hypertension, Diabetic}	3
{No heart disease, Diabetic}	6
{never, Diabetic}	6
{Overweight, Diabetic}	6
{Borderline, Diabetic}	3
{HbA1c Normal, Diabetic}	3
{No, Diabetic}	6
{Yes, Diabetic}	3

Bảng 7. Các cặp 2-itemset và tần suất của chúng theo phương pháp Apriori

- Loại bỏ những 2-itemset không đạt minSup =3:

  {Male, No hypertension}: 4,

  {Male, Diabetic}: 3,

  {Female, Diabetic}: 6,

  {Female, No hypertension}: 6,

  {Female, never}: 4,

  {Female, Overweight}: 4,

  {Female, Borderline}: 3,

  {Female, HbA1c Normal}: 3,

  {Female, No}: 4,

  {64+, Diabetic}: 3,

  {64+, Hypertension}: 3,

  {No hypertension, Diabetic}: 6,

  {Hypertension, Diabetic}: 3,

  {No heart disease, Diabetic}: 6,

  {Heart disease, Diabetic}: 3,

{never, Diabetic}: 6,

{Overweight, Diabetic}: 6,

{Borderline, Diabetic}: 3,

{HbA1c Normal, Diabetic}: 3,

{No, Diabetic}: 6,

{Yes, Diabetic}: 3

3.Tạo các tập 3-itemset từ các 2-itemset phổ biến:

Itemset	**Count**
{Female, No hypertension, Diabetic}	4
{Female, never, Diabetic}	4
{Female, Overweight, Diabetic}	4
{No hypertension, never, Diabetic}	4
{No hypertension, Overweight, Diabetic}	4
{never, Overweight, Diabetic}	4
{No heart disease, never, Diabetic}	4
{No heart disease, Overweight, Diabetic}	4
{Male, Diabetic, No hypertension}	3
{Female, Borderline, Diabetic}	3
{Female, HbA1c Normal, Diabetic}	3
{Female, No, Diabetic}	3
{64+, Diabetic, Hypertension}	3
{Heart disease, Diabetic, Hypertension}	3
{Hypertension, Diabetic, Yes}	3

Bảng 8. Các cặp 3-itemset và tần suất của chúng theo phương pháp Apriori

- Loại bỏ những 3-itemset không đạt minSup = 4:

  {Female, No hypertension, Diabetic}: 4,

  {Female, never, Diabetic}: 4,

  {Female, Overweight, Diabetic}: 4,

  {No hypertension, never, Diabetic}: 4,

  {No hypertension, Overweight, Diabetic}: 4,

{never, Overweight, Diabetic}: 4,
{No heart disease, never, Diabetic}: 4,
{No heart disease, Overweight, Diabetic}: 4,
{Male, Diabetic, No hypertension}: 3,
{Female, Borderline, Diabetic}: 3,
{Female, HbA1c Normal, Diabetic}: 3,
{Female, No, Diabetic}: 3,
{64+, Diabetic, Hypertension}: 3,
{Heart disease, Diabetic, Hypertension}: 3,
{Hypertension, Diabetic, Yes}: 3

4.Kết quả sinh ra các luật kết hợp:

- **1-itemset phổ biến**: {Male}, {Female}, {0-16}, {32-48}, {48-64}, {64+}, {No hypertension}, {Hypertension}, {No heart disease}, {Heart disease}, {never}, {Overweight}, {Balance}, {Underweight}, {Borderline}, {HbA1c Normal}, {High}, {Diabetic}, {Pre_diabetic}, {No}, {Yes}
- **2-itemset phổ biến**: {Male, No hypertension}, {Male, Diabetic}, {Female, Diabetic}, {Female, No hypertension}, {Female, never}, {Female, Overweight}, {Female, Borderline}, {Female, HbA1c Normal}, {Female, No}, {64+, Diabetic}, {64+, Hypertension}, {No hypertension, Diabetic}, {Hypertension, Diabetic}, {No heart disease, Diabetic}, {Heart disease, Diabetic}, {never, Diabetic}, {Overweight, Diabetic}, {Borderline, Diabetic}, {HbA1c Normal, Diabetic}, {No, Diabetic}, {Yes, Diabetic}
- **3-itemset phổ biến**: {Female, No hypertension, Diabetic}, {Female, never, Diabetic}, {Female, Overweight, Diabetic}, {No hypertension, never, Diabetic}, {No hypertension, Overweight, Diabetic}, {never, Overweight, Diabetic}, {No heart disease, never, Diabetic}, {No heart disease, Overweight, Diabetic}, {Male, Diabetic, No hypertension}, {Female, Borderline, Diabetic}, {Female, HbA1c Normal, Diabetic}, {Female, No,

Diabetic}, {64+, Diabetic, Hypertension}, {Heart disease, Diabetic, Hypertension}, {Hypertension, Diabetic, Yes}

3.5.1.2 Triển khai bằng thư viện có sẵn:

1. Cài đặt thư viện và tiến hành chuẩn hóa dữ liệu 13 dòng bằng TransactionEnconder:

	0-16	16-32	32-48	48-64	64+	Balance	Blood Glucose Normal	Borderline	Diabetic	Female	...	No Hypertension	Overweight	Pre_diabetic	Underweight	Yes	current	ever	former	never	not current
0	False	False	True	False	False	True	False	False	True	True	...	True	False	False	False	False	False	False	False	True	False
1	False	False	True	False	False	False	False	False	True	False	...	True	True	False	False	False	False	True	False	False	False
2	False	False	True	False	False	False	True	True	False	False	...	True	True	False	False	False	True	False	False	False	False
3	False	False	False	False	True	True	True	False	False	True	...	True	False	False	False	False	False	False	False	True	False
4	False	False	False	False	True	False	False	True	True	True	...	True	True	False	False	False	False	False	True	False	False
5	False	True	False	False	False	False	False	True	True	False	...	True	True	False	False	False	False	False	False	True	False
6	False	False	False	False	True	False	False	True	True	False	...	True	True	False	False	True	False	False	False	False	True
7	False	False	True	False	False	False	True	True	False	False	...	True	True	False	False	False	False	False	True	False	False
8	False	False	True	False	False	False	False	False	True	False	...	False	True	False	False	False	False	False	False	True	False
9	False	False	True	False	False	False	False	True	True	False	...	True	True	False	False	False	False	False	False	True	False
10	False	False	True	False	False	True	False	True	True	True	...	True	False	False	False	False	True	False	False	False	False
11	False	True	False	False	False	True	False	False	False	True	...	True	False	True	False	False	False	False	False	True	False
12	False	False	False	False	True	True	True	False	False	True	...	True	False	False	False	False	False	False	False	True	False

Hình 16.Kết quả sau khi chuẩn hóa 13 dòng dữ liệu

4. Triển khai sinh luật với thư viện Apriori

support	itemsets
0.38461538461538464	frozenset({'Borderline'})
0.6153846153846154	frozenset({'Diabetic'})
0.6153846153846154	frozenset({'Female'})
0.38461538461538464	frozenset({'HbA1c Normal'})
0.38461538461538464	frozenset({'Male'})
0.6153846153846154	frozenset({'No'})
0.7692307692307693	frozenset({'No heart disease'})
0.7692307692307693	frozenset({'No hypertension'})
0.5384615384615384	frozenset({'Overweight'})
0.38461538461538464	frozenset({'Yes'})
0.6153846153846154	frozenset({'never'})
0.38461538461538464	frozenset({'Diabetic', 'Male'})
0.46153846153846156	frozenset({'Diabetic', 'No heart disease'})
0.46153846153846156	frozenset({'Diabetic', 'No hypertension'})
0.38461538461538464	frozenset({'Diabetic', 'Overweight'})
0.38461538461538464	frozenset({'Diabetic', 'Yes'})
0.46153846153846156	frozenset({'Diabetic', 'never'})
0.46153846153846156	frozenset({'Female', 'No'})
0.46153846153846156	frozenset({'Female', 'No heart disease'})
0.46153846153846156	frozenset({'No hypertension', 'Female'})
0.38461538461538464	frozenset({'Female', 'never'})
0.38461538461538464	frozenset({'HbA1c Normal', 'No'})
0.38461538461538464	frozenset({'HbA1c Normal', 'No heart disease'})
0.5384615384615384	frozenset({'No', 'No heart disease'})
0.46153846153846156	frozenset({'No hypertension', 'No'})

support	
0.38461538461538464	frozenset({'never', 'No'})
0.6923076923076923	frozenset({'No hypertension', 'No heart disease'})
0.38461538461538464	frozenset({'Overweight', 'No heart disease'})
0.38461538461538464	frozenset({'never', 'No heart disease'})
0.38461538461538464	frozenset({'No hypertension', 'never'})
0.38461538461538464	frozenset({'Diabetic', 'No hypertension', 'No heart disease'})
0.38461538461538464	frozenset({'Female', 'No', 'No heart disease'})
0.38461538461538464	frozenset({'No hypertension', 'Female', 'No'})
0.46153846153846156	frozenset({'No hypertension', 'Female', 'No heart disease'})
0.38461538461538464	frozenset({'HbA1c Normal', 'No', 'No heart disease'})
0.46153846153846156	frozenset({'No hypertension', 'No', 'No heart disease'})
0.38461538461538464	frozenset({'No hypertension', 'Female', 'No', 'No heart disease'})

Hình 17. Kết quả các mẫu phổ biến của 13 dòng bằng thư viện Apriori

3.5.1.3 Xây dựng lại thuật toán Apriori

Bước 1: Xây dựng hàm tính tần suất các itemset.

Bước 2: Xây dựng hàm loại bỏ những itemset không đạt minSup.

Bước 3: Xây dựng các hàm tạo ra các mẫu phổ biến có độ dài từ 2-itemset trở lên.

Bước 4: Xây dựng hàm tổng hợp, tạo các luật kết hợp từ các mẫu phổ biến

Bước 5: Kết quả trên 13 dòng dữ liệu.

Các luật kết hợp:

	Antecedent	Consequent	Confidence
0	(Diabetic, No hypertension)	(Borderline)	2.00
1	(Diabetic, Borderline)	(No hypertension)	3.00
2	(No hypertension, Borderline)	(Diabetic)	3.00
3	(No hypertension, never)	(Borderline)	1.80
4	(never, Borderline)	(No hypertension)	2.25
..	...	...	...
107	(HbA1c Normal, Female)	(No hypertension)	3.00
108	(No hypertension, HbA1c Normal)	(Female)	3.00
109	(Diabetic, No hypertension)	(Yes)	2.00
110	(Diabetic, Yes)	(No hypertension)	2.40
111	(Yes, No hypertension)	(Diabetic)	3.00

Hình 18. Kết quả 13 dòng dữ liệu trên thuật toán được xây dựng

3.5.2 Khai thác mẫu với FreeSpan

3.5.2.1 Triển khai với phương pháp thủ công

gender	age	hypertension	heart_disease	smoking_history	bmi	HbA1c_level	blood_glucose_level	diabetes
Male	15	0	0	never	30	6,1	200	0
Male	45	1	0	never	26	4	158	0
Female	20	0	0	never	22	3,5	100	0
Male	73	0	0	former	26	9	160	1
Female	60	0	0	never	18	4	159	0
Female	54	0	0	former	55	6	100	0
Female	57	0	0	current	22	6,6	90	0
Female	65	1	1	never	34	8,2	140	1
Male	36	0	0	not current	46	6,2	130	1
Female	38	0	0	not current	24	4,8	85	0
Female	9	0	0	never	16	6,1	200	1
Male	6	0	1	never	17	6,5	240	1
Female	80	1	1	never	30	6,1	100	0
Female	73	1	0	current	18	5,8	159	1

Bảng 9. Lấy 14 dòng dữ liệu

Chia tập dữ liệu theo thuộc tính các cột					
Cột age	0-16	16-32	32-48	48-64	64+

Cột bmi	0-18.5	18.5-25	25+		
	Underweight	Balance	Overweight		
Cột HbA1c_level	0-6	6-6.5	6.5+		
	HbA1c Normal	Borderline	High		
Cột blood_glucose_level	0-90	90-125	125+		
	Blood Glucose Normal	Pre_diabetic	Diabetic		

Bảng 19. Gắn nhãn lại cho 14 dòng dữ liệu

gender	age	hypertension	heart_disease	smoking_history	bmi	HbA1c_level	blood_glucose_level	diabetes
Male	0-16	No hypertension	No heart disease	never	Overweight	Borderline	Diabetic	No
Male	32-48	Hypertension	No heart disease	never	Overweight	HbA1c Normal	Diabetic	No
Female	16-32	No hypertension	No heart disease	never	Balance	HbA1c Normal	Pre_diabetic	No
Male	64+	No hypertension	No heart disease	former	Overweight	High	Diabetic	Yes
Female	48-64	No hypertension	No heart disease	never	Underweight	HbA1c Normal	Diabetic	No
Female	48-64	No hypertension	No heart disease	former	Overweight	HbA1c Normal	Pre_diabetic	No
Female	48-64	No hypertension	No heart disease	current	Balance	High	Blood Glucose Normal	No
Female	64+	Hypertension	Heart disease	never	Overweight	High	Diabetic	Yes
Male	32-48	No hypertension	No heart disease	not current	Overweight	Borderline	Diabetic	Yes
Female	32-48	No hypertension	No heart disease	not current	Balance	HbA1c Normal	Blood Glucose Normal	No
Female	0-16	No hypertension	No heart disease	never	Underweight	Borderline	Diabetic	Yes
Male	0-16	No hypertension	Heart disease	never	Underweight	Borderline	Diabetic	Yes
Female	64+	Hypertension	Heart disease	never	Overweight	Borderline	Pre_diabetic	No
Female	64+	Hypertension	No heart disease	current	Underweight	HbA1c Normal	Diabetic	Yes

Bảng 11. Dữ liệu 14 dòng sau khi gắn nhãn

Đầu tiên, nhìn vào cột *diabetes* có tất cả là 8 dữ liệu thể hiện là "No". Với mong muốn khai thác mẫu theo mẫu "No" cho nên chọn minSup = 8.

Dựa trên dữ liệu của bảng 13, ta thực hiện pháp pháp FreeSpan với minSup = 8 như sau:

Bước 1: *Tìm tần suất xuất hiện của các item và tạo danh sách f_list*

- Đầu tiên chúng ta tiến hành đếm tần suất xuất hiện của các item có trong cơ sở dữ liệu.
- Tiếp đến, loại ra những item không thỏa điều kiện nhỏ hơn hoặc bằng minSup, giữ lại những item thỏa điều kiện.
- Sau đó, lấy những item thỏa điều kiện đưa vào danh sách f_list và sắp xếp chúng lại theo chiều giảm dần.

item	f	item	f	item	f	item	f
Female	9	No Hypertension	10	Overweight	7	Blood_glucose_normal	2

Male	5	Hypertension	4	Balance	3	Yes	6
64+	4	Heart Disease	3	Underweight	4	No	8
16-32	1	No Heart Disease	11	Borderline	5		
32-48	3	never	8	HbA1c Normal	6		
48-64	3	current	2	High	3		
0-16	3	former	2	Diabetic	9		
		Not_current	2	Pre_diabetic	3		

Bảng 110. Bảng tần suất các item của 14 dòng dữ liệu

- Loại bỏ những item không đạt yêu cầu và sau đó tạo danh sách như sau:

f_list = {No Heart Disease: 12, No Hypertension: 10, Female: 9, Diabetic: 9, never: 8, No: 8}

Bước 2: *Tạo ma trận chiếu từ f_list*

	No Heart Disease	**No Hypertension**	**Female**	**Diabetic**	**never**	**No**
No Heart Disease	0	(0, 9, 0)	(0, 7, 0)	(7, 0, 0)	(5, 0, 0)	(7, 0, 0)
No Hypertension	(9, 0, 0)	0	(0, 6, 0)	(6, 0, 0)	(5, 0, 0)	(6, 0, 0)
Female	(7, 0, 0)	(6, 0, 0)	0	(4, 0, 0)	(5, 0, 0)	(6, 0, 0)
Diabetic	(0, 7, 0)	(0, 6, 0)	(0, 4, 0)	0	(0, 6, 0)	(3, 0, 0)
never	(0, 5, 0)	(0, 5, 0)	(0, 5, 0)	(6, 0, 0)	0	(5, 0, 0)
No	(0, 7, 0)	(0, 6, 0)	(0, 6, 0)	(0, 3, 0)	(0, 5, 0)	0

Bảng 13. Ma trận chiếu từ f_list

Bước 3: *Tìm các mẫu có độ dài 2 dựa vào ma trận chiếu*

Các mẫu có độ dài 2 thỏa minSup dựa theo ma trận chiếu như sau:

1. {No Hypertension, No Heart disease}: 9

Bước 4: *Chú thích item-repeating*

Các chuỗi trong cơ sở dữ liệu không có item lặp lại trong chuỗi.

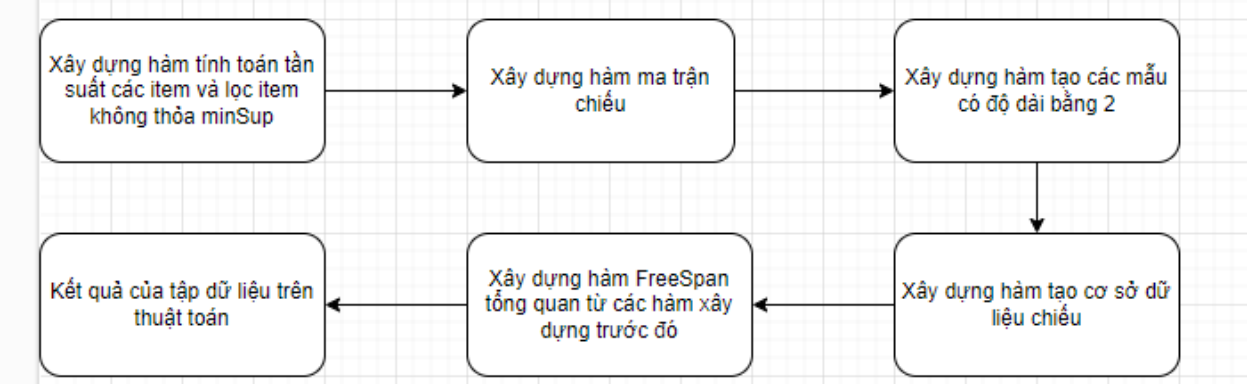
Bước 5: *Thực hiện chiếu toàn bộ chuỗi của cơ sở dữ liệu và lọc theo minSup:*

- Chiếu mẫu {No Hypertension, No Heart disease} lên toàn bộ chuỗi ta có để kiếm thêm những mẫu phổ biến mới có độ dài là 3. Tuy nhiên, các mẫu mới sinh ra nhỏ hơn hơn minSup cho nên không có mẫu phổ biến nào được sinh ra.

Bước 6: *Kết quả*

-Vậy mẫu phổ biến thu được của cơ sở dữ liệu ở bảng 13 với minSup = 8 chỉ có {No Hypertension, No Heart disease}.

3.5.2.3 Xây dựng lại thuật toán FreeSpan



Hình 19. Sơ đồ các bước xây dựng thuật toán FreeSpan

Bước 1: Xây dựng hàm đếm tần suất các item và hàm lọc item không thỏa minSup

- Ở đây sử dụng thư viện defaultdict để lưu trữ tần suất của cá item, với đặc điểm là thiết lập mặc định kiểu giá trị trả về của dict. Nếu không đúng kiểu giá trị trả về thì nó sẽ đưa ra lỗi KeyError.

Bước 2: Xây dựng hàm ma trận chiếu

- Đầu tiên khởi tạo 1 ma trận với 3 thông số A, B, C được thiết lập mặc định ở giá trị ban đầu là (0, 0, 0). Sau đó tiến hành lặp mục phổ biến trong cơ sở

dữ liệu ban đầu để lọc ra các item phổ biến. Sau đó, tiến hành lặp các item trong đấy để thu được thông số A, B, C để xây dựng ma trận. Điều này lặp đi lặp lại cho đến khi hết mục phổ biến.

Bước 3: Tạo mẫu có độ dài bằng 2 từ ma trận chiếu

- Đầu tiên khởi tạo một list rỗng. Sau đó tiến hành lặp lấy các giá trị item1 qua từng hàng trong ma trận. Tiếp đến lặp item 2 lấy giá trị đếm của 3 thông số A, B, C. Nếu thỏa điều kiện >= minSup thì đưa vào list.

Bước 4: Xây dựng cơ sở dữ liệu chiếu

- Đầu tiên khởi tạo một danh sách rỗng projected_db.
- Sau đó, tiến hành lặp qua từng bản ghi trong dữ liệu data.
- Kiểm tra xem phần tử cuối cùng của prefix có trong bản ghi hay không:

  +Nếu phần tử này có trong bản ghi, tìm vị trí của nó.

  +Lấy phần còn lại của bản ghi bắt đầu từ vị trí ngay sau phần tử cuối cùng của prefix.

  +Lọc các mục trong phần còn lại, chỉ giữ lại các mục thường xuyên (frequent_items).

  Nếu phần còn lại sau khi lọc không rỗng, thêm nó vào danh sách projected_db.

- Trả về danh sách projected_db chứa các chuỗi chiếu đã được lọc.

Bước 5: Xây dựng tổng quan hàm FreeSpan

- Chúng ta khởi tạo hàm FreeSpan với các khai báo hàm liên quan bao gồm ma trận chiếu, tìm mẫu phổ biến, cơ sở dữ liệu chiếu. Sau chúng ta tiến hành đệ quy để khai thác các mẫu tuần tự.

Bước 6: Kết quả của thuật toán trên 14 dòng dữ liệu ở bảng 13

```
MẪU PHỔ BIẾN:
No hypertension -> No heart disease: 9
```

Hình 20. Kết quả khai thác mẫu FreeSpan trên 14 dòng dữ liệu

3.5.1.3 So sánh

Kết quả của thuật toán được xây dựng trên Python của JupyterNotebook cho ra kết quả mẫu phổ biến giống với kết quả thực hiện thuật toán bằng phương pháp thủ công.

3.6 Phân lớp bằng thuật toán cây quyết định (Decision Tree)

3.6.1 Phương pháp thủ công

Triển khai thuật toán ID3 trên tập dữ liệu theo bảng 13 như sau:

1. Tính Entropy của tập dữ liệu:

Entropy(S)	**0,985228136**

2. Tính Gain của S với các thuộc tính:

gender	Male	Female			
	3Y, 2N	3Y, 6N			
	0.346768069	0.590333036			
Gain(S, gender)	**0.04812703**				
age	0-16	16-32	32-48	48-64	64+
	2Y, 1N	0Y, 1N	2Y, 1N	0Y, 3N	3Y, 1N
	0.196777679	0	0.196777679	0	0.23179375
Gain(S, age)	**0.359879029**				
hypertension	Hypertension	No hypertension			
	2Y, 2N	4Y, 6N			
	0.285714286	0.693536139			
Gain(S, hypertension)	**0.005977711**				
heart_disease	Heart disease	No heart disease			
	2Y, 1N	4Y, 7N			
	0.196777679	0.743018811			
Gain(S, heart_disease)	**0.045431647**				
smoking_history	never	former	current	not current	
	3Y, 5N	1Y, 1N	1Y, 1N	1Y, 1N	
	0.545390859	0.142857143	0.142857143	0.142857143	
Gain(S, smoking_history)	**0.011265849**				
bmi	Overweight	Balance	Underweight		
	3Y, 4N	0Y, 3N	3Y, 1N		
	0.492614068	0	0.23179375		
Gain(S, bmi)	**0.260820318**				
HbA1c_level	HbA1c Normal	Borderline	High		
	0Y, 4N	4Y, 3N	2Y, 1N		
	0	0.492614068	0.196777679		
Gain(S, HbA1c_level)	**0.295836389**				
blood_glucose_level	Pre_diabetic	Diabetic	Blood Glucose Normal		
	0Y, 3N	6Y, 3N	0Y, 2N		
	0	0.590333036	0		
Gain(S, blood_glucose_level)	**0.3948951**				

3.

Bảng 111. Kết quả tính Gain của tập dữ liệu với các thuộc tính

Ta thấy được Gain (S, blood_glucose_level) cao hơn so với các thuộc tính còn lại

=> Chọn blood_glucose_level

Xét nhánh Diabetic:

gender	age	hypertension	heart_disease	moking_history	bmi	HbA1c_level	blood_glucose_level	diabetes
Male	0-16	No hypertension	No heart disease	never	Overweight	Borderline	Diabetic	No
Male	32-48	Hypertension	No heart disease	never	Overweight	HbA1c Normal	Diabetic	No
Male	64+	No hypertension	No heart disease	former	Overweight	High	Diabetic	Yes
Female	48-64	No hypertension	No heart disease	never	Underweight	HbA1c Normal	Diabetic	No
Female	64+	Hypertension	Heart disease	never	Overweight	High	Diabetic	Yes
Male	32-48	No hypertension	No heart disease	not current	Overweight	Borderline	Diabetic	Yes
Female	0-16	No hypertension	No heart disease	never	Underweight	Borderline	Diabetic	Yes
Male	0-16	No hypertension	Heart disease	never	Underweight	Borderline	Diabetic	Yes
Female	64+	Hypertension	No heart disease	current	Underweight	Borderline	Diabetic	Yes

Bảng 15. Xét theo nhánh Diabetic của blood_glucose_level

4. Tính Entropy(S_Diabetic):

Entropy(S_Diabetic)	**0,918295834**

5. Tính Gain của Diabetic với các thuộc tính còn lại:

gender	Male	Female			
	3Y, 2N	3Y, 1N			
	0.539416997	0.360568055			
Gain(S, gender)	**0.018310782**				
age	0-16	16-32	32-48	48-64	64+
	2Y, 1N	0Y, 0N	1Y, 1N	0Y, 1N	3Y, 0N
	0.306098611	0	0.222222222	0	0
Gain(S, age)	**0.389975**				
hypertension	Hypertension	No hypertension			
	2Y, 1N	4Y, 2N			
	0.306098611	0.612197223			
Gain(S, hypertension)	**0**				
heart_disease	Heart disease	No heart disease			
	2Y, 0N	4Y, 3N			
	0	0.76628855			
Gain(S, heart_disease)	**0.152007284**				
smoking_history	never	former	current	not current	
	3Y, 3N	1Y, 0N	1Y, 0N	1Y, 0N	
	0.666666667	0	0	0	
Gain(S, smoking_history)	**0.251629167**				
bmi	Overweight	Balance	Underweight		
	3Y, 2N	0Y, 0N	3Y, 1N		
	0.539416997	0	0.360568055		
Gain(S, bmi)	**0.018310782**				
HbA1c_level	HbA1c Normal	Borderline	High		
	0Y, 2N	4Y, 1N	2Y, 0N		
	0	0.401071164	0		
Gain(S, HbA1c_level)	**0.51722467**				

Bảng 112. Gain của Diabetic với các thuộc tính còn lại

Ta thấy được Gain của Diabetic với HbA1c_level cao hơn các thuộc tính còn lại

=> chọn HbA1c_level

Xét nhánh Borderline của HbA1c_level:

gender	age	hypertension	heart_disease	smoking_history	bmi	HbA1c_level	blood_glucose_level	diabetes
Male	0-16	No hypertension	No heart disease	never	Overweight	Borderline	Diabetic	No
Male	32-48	No hypertension	No heart disease	not current	Overweight	Borderline	Diabetic	Yes
Female	0-16	No hypertension	No heart disease	never	Underweight	Borderline	Diabetic	Yes
Male	0-16	No hypertension	Heart disease	never	Underweight	Borderline	Diabetic	Yes
Female	64+	Hypertension	No heart disease	current	Underweight	Borderline	Diabetic	Yes

Bảng 17. Bảng dữ liệu xét theo nhánh Borderline của HbA1c_level

6. Tính Entropy của Borderline của thuộc tính HbA1c_level:

Entropy(S_Borderline)	**0.721928095**

7. Tính Gain của Borderline với các thuộc tính còn lại:

gender	Male	Female		
	2Y, 1N	2Y, 0N		
	0.5509775	0		
Gain(S, gender)	**0.170950594**			
hypertension	Hypertension	No hypertension		
	1Y, 0N	3Y, 1N		
	0	0.6490225		
Gain(S, hypertension)	**0.072905595**			
heart_disease	Heart disease	No heart disease		
	1Y, 0N	3Y, 1N		
	0	0.6490225		
Gain(S, heart_disease)	**0.072905595**			
smoking_history	never	former	current	not current
	2Y, 1N	0Y, 0N	1Y, 0N	1Y, 0N
	0.5509775	0	0	0
Gain(S, smoking_history)	**0.170950594**			
bmi	Overweight	Balance	Underweight	
	1Y, 1N	0Y, 0N	3Y, 0N	
	0.4	0	0	
Gain(S, bmi)	**0.321928095**			
age	0-16	32-48	64+	
	2Y, 1N	1Y, 0N	1Y, 0N	
	0.5509775	0	0	
Gain(S, age)	**0.170950594**			

Bảng 18. Gain của Borderline với các thuộc tính còn lại

Ta thấy được Gain của Borderline với bmi cao hơn so với các thuộc tính còn lại

⇨ chọn thuộc tính bmi

Tiếp tục xét nhánh Overweight của thuộc tính bmi:

gender	age	hypertension	heart_disease	smoking_history	bmi	HbA1c_level	blood_glucose_level	diabetes
Male	32-48	Hypertension	No heart disease	never	Overweight	HbA1c Normal	Diabetic	No
Male	32-48	No hypertension	No heart disease	not current	Overweight	Borderline	Diabetic	Yes

Bảng 19. Dữ liệu xét theo nhánh Overweight của thuộc tính bmi

8. Tính Entrotry của Overweight thuộc tính bmi:

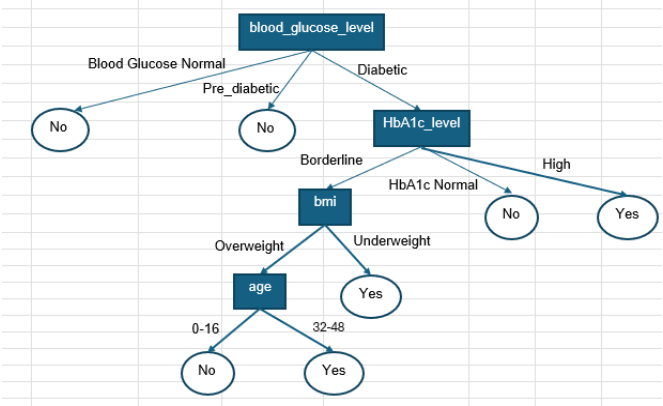
Entropy(S_32-48)	**1**

9. Tính Gain của Overweight với các thuộc tính còn lại

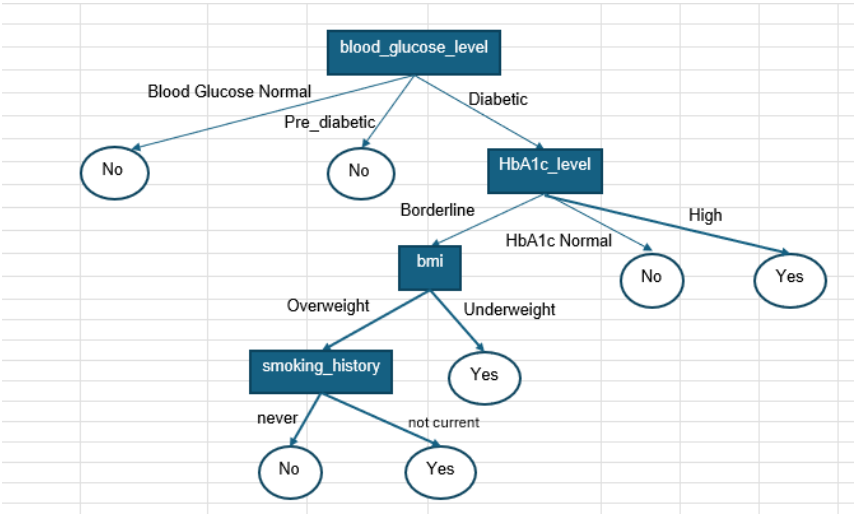
gender	Male	Female		
	1Y, 1N	0Y, 0N		
	1	0		
Gain(S, gender)	**0**			
hypertension	Hypertension	No hypertension		
	0Y, 0N	1Y, 1N		
	0	1		
Gain(S, hypertension)	**0**			
heart_disease	Heart disease	No heart disease		
	0Y, 0N	1Y, 1N		
	0	1		
Gain(S, heart_disease)	**0**			
smoking_history	never	former	current	not current
	0Y, 1N	0Y, 0N	0Y, 0N	1Y, 0N
	0	0	0	0
Gain(S, smoking_history)	**1**			
age	0-16	32-48		
	0Y, 1N	1Y, 0N		
	0	0		
Gain(S, age)	**1**			

Bảng 20. Gain của Overweight với các thuộc tính

Ta thấy có cả 2 thuộc tính đều có Gain = 1, nên ta chọn cả 2 thuộc tính và hình thành cây quyết định như sau:

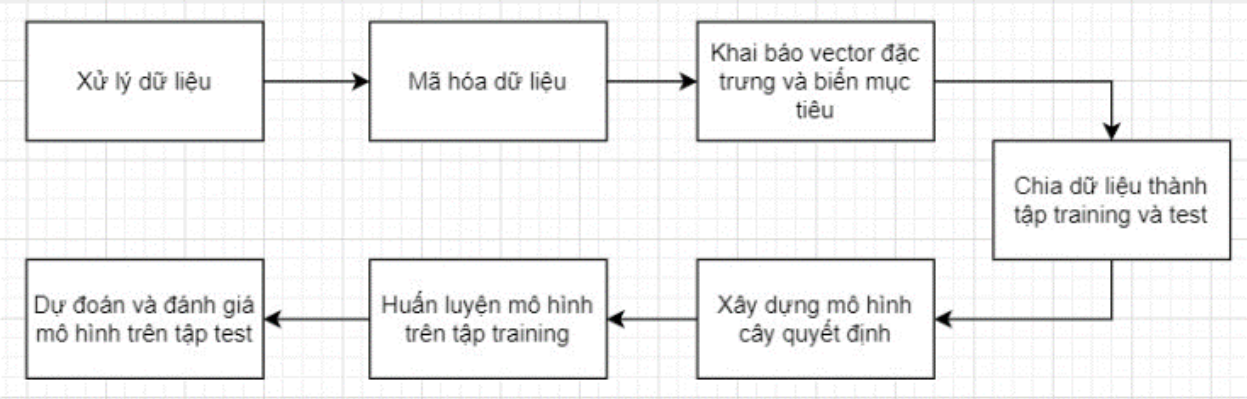


Hình 21. Cây quyết định theo age



Hình 22. Cây quyết định theo smoking_history

3.6.2 Xây dựng cây quyết định (Decision Tree)



Hình 23. Tổng quan các bước xây dựng thuật toán Decision Tree

Bước 1: Tính toán entropy của đầu ra phân lớp của tập dữ liệu

Công thức tính Entropy dựa trên tập dữ liệu hiện tại cần phân lớp với thuộc tính *diabetes* với 2 thể hiện đó là "Yes", "No":

$$Entropy = -\frac{y}{s}log_2\left(\frac{y}{s}\right) - \frac{n}{s}log_2(\frac{n}{s})$$

- Khởi tạo hàm *find_entropy(df)* với *df* là dữ liệu dạng DataFrame.
- Đầu tiên, chúng ta khai báo 1 biến *class* là biến lấy cột cuối cùng trong dữ liệu là *diabetes* là đích cuối để phân lớp.
- Tiếp theo, khai báo biến *values* lấy các giá trị unique của biến mục tiêu.
- Thực hiện vòng lặp tính toán tỉ lệ *fraction* của từng giá trị mục tiêu. Đồng thời thực hiện tính Entropy.

Bước 2: Xây dựng hàm tính toán Entropy của các thuộc tính còn lại

- Khởi tạo hàm *find_entropy_attribute(df, attribute)* với *df* là DataFrame, *attribute* là các thuộc tính cần tính.
- Khởi tạo biến *Class* là lấy cột cuối cùng của DataFrame.
- *target_variables* là biến lấy các giá trị duy nhất của biến mục tiêu.
- *variables* là các giá trị duy nhất của thuộc tính cần tính toán entropy.

- Vòng lặp tính toán entropy của từng giá trị của thuộc tính dựa trên phân bố của biến mục tiêu.

Bước 3: Xây dựng hàm tính toán Gain của các thuộc tính còn lại:

Tính Gain của các thuộc tính còn lại dựa trên Entropy của thuộc tính phân lớp.

$$Entropy(Diabetes, x) = Entropy(Diabetes) - \sum_{giá\ trị\ x/Diabestes} Entropy(x)$$

- Khởi tạo hàm *find_winner(df)* để đi tìm Gain cao nhất của các thuộc tính
- Tính toán entropy của toàn bộ tập dữ liệu (find_entropy(df)).
- Tính toán entropy của từng thuộc tính (find_entropy_attribute(df, key)).
- Tính toán Information Gain cho từng thuộc tính và chọn thuộc tính có IG cao nhất.

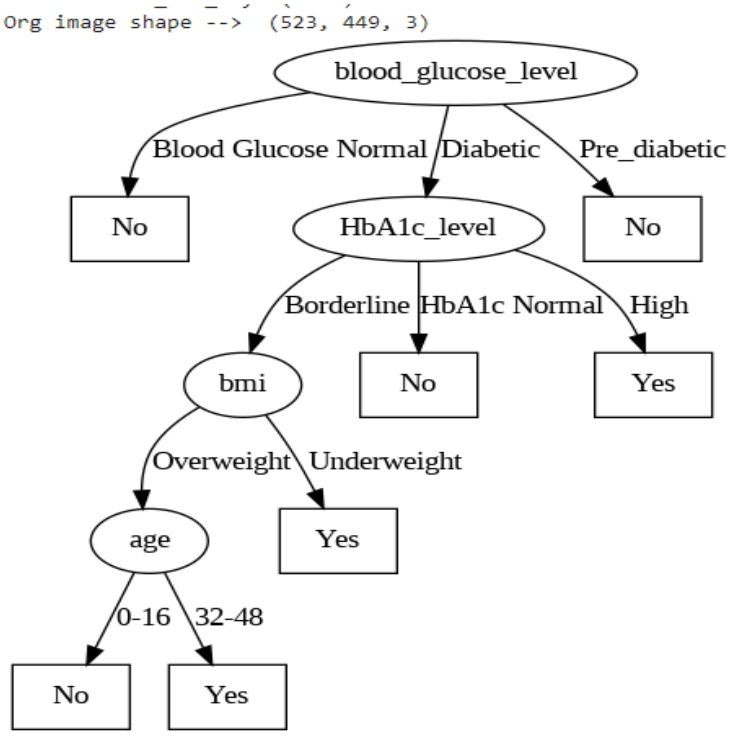
Bước 4: Tạo bảng con lưu trữ các kết quả các node

- Khởi tạo hàm *get_subtable(df, node, value)* để trả về một bảng con chứa các *node* kết quả.
- Trả về các hàng của DataFrame mà thuộc tính *node* có giá trị *value.*

Bước 5: Xây dựng hàm cây quyết định với đệ quy

- Khởi tạo hàm *buildTree(df, tree = None)* với *df* là dữ liệu dưới dạng DataFrame, *tree* là giá trị cây ở thời điểm hiện tại với ban đầu là rỗng.
- Tìm thuộc tính tốt nhất để phân chia (*find_winner(df)*).
- Phân chia DataFrame thành các bảng con dựa trên giá trị của thuộc tính tốt nhất.
- Kiểm tra nếu tất cả các giá trị của biến mục tiêu trong bảng con là giống nhau (*len(counts) == 1*), thì gán giá trị của nút lá.
- Nếu không, tiếp tục xây dựng cây quyết định đệ quy cho các bảng con.

Bước 6: Thực hiện vẽ cây



Hình 24. Cây quyết định với 14 dòng dữ liệu

3.6.3 So sánh kết quả

Phương pháp thủ công cho ra kết quả cuối cùng là 2 trường hợp xét theo nhánh *blood_glucose_level → HbA1c_level → bmi → age* và *blood_glucose_level → HbA1c_level → bmi → smoking_history.*

Còn đối với trường hợp thuật toán cây quyết định khi xây dựng lại thì trả về 1 trường hợp đầu tiên đó là *blood_glucose_level → HbA1c_level → bmi → age.*

Vậy kết quả của phương pháp thủ công và thuật toán được xây dựng lại là giống nhau.

3.7 Kết quả trên toàn tập dữ liệu

3.7.2 Kết quả của thuật toán FreeSpan

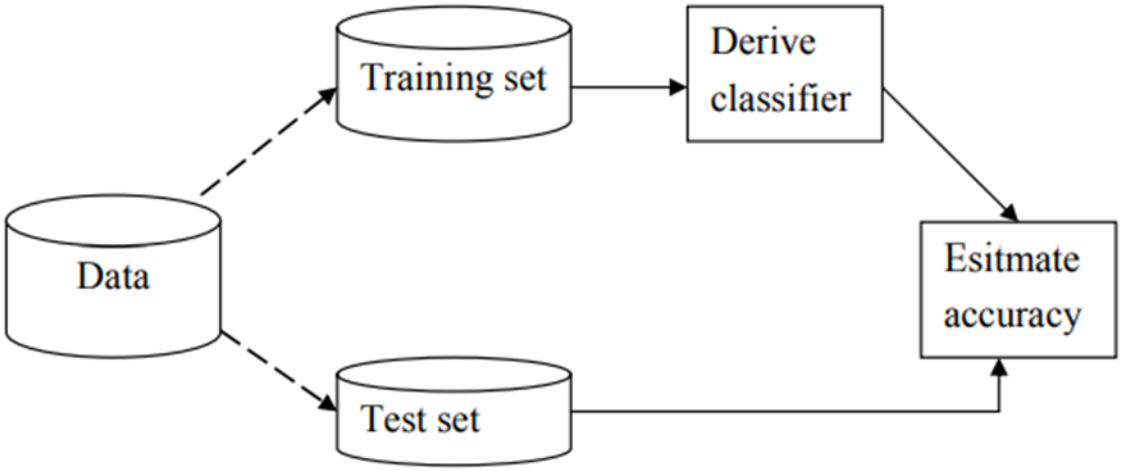
```
MẪU PHỔ BIẾN:
Female -> No Hypertension: 35310
Female -> No Hypertension -> No: 51168
Female -> No Hypertension -> No: 51168
Female -> No Hypertension -> Overweight: 39610
Female -> No Hypertension -> Diabetic: 39908
Female -> No Hypertension -> No: 51168
Female -> No: 35145
Female -> No Heart Disease: 37636
Female -> No Heart Disease -> Diabetic: 32419
Female -> No Heart Disease -> No: 38244
Female -> No Heart Disease -> No: 38244
No Hypertension -> Overweight: 41409
No Hypertension -> Diabetic: 41695
No Hypertension -> No: 52679
No Hypertension -> No Heart Disease: 55640
No Hypertension -> No Heart Disease -> Diabetic: 32419
No Hypertension -> No Heart Disease -> No: 38244
No Hypertension -> No Heart Disease -> No: 38244
Overweight -> Diabetic: 34396
Overweight -> No: 40530
Diabetic -> No: 39603
No Heart Disease -> never: 33995
No Heart Disease -> Overweight: 44337
No Heart Disease -> Diabetic: 44238
No Heart Disease -> No: 55159
```

Hình 33. Kết quả FreeSpan trên toàn tập dữ liệu với minSup = 0.5

Kết quả của thuật toán thu được 25 mẫu có độ dài 2 trở lên với độ hỗ trợ là 0.5.

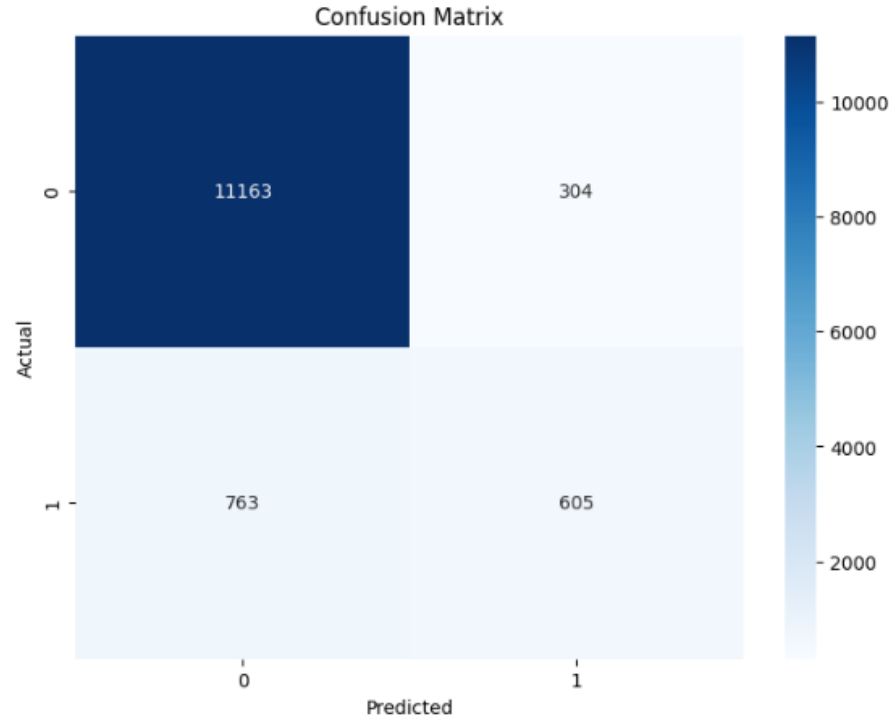
3.7.3 Kết quả của phân lớp với cây quyết định

Sử dụng phương pháp chia phần (Holdout method): Tập dữ liệu được chia ngẫu nhiên thành 2 tập dữ liệu độc lập là tập dữ liệu huấn luyện và tập kiểm định mô hình. Tỷ lệ phân chia cho training data set là 80% và test data set là 20%.



Hình 25. Phương pháp Hold-out

Ở bài toán phân lớp này, chúng em sử dụng tập dữ liệu với biến mục tiêu có 2 giá trị, tui em sử dụng ma trận đúng sai (Confusion matrix) để đánh giá hiệu xuất của mô hình dựa trên số lượng các dự đoán đúng và sai lầm của mô hình trên tập dữ liệu kiểm tra.



Hình 26. Ma trận đúng sai của tập dữ liệu

Qua ma trận ta thấy đối với lớp 0, mô hình hoạt động rất tốt với số lượng mẫu phân loại đúng là 11163, có 763 mẫu dự đoán sai nhưng so với số mẫu dự đoán đúng thì không đáng kể. Tuy nhiên đối với lớp 1, số lượng mẫu phân loại đúng là 605, nhưng có tận 304 mẫu dữ đoán sai, cho thấy lớp 1 có độ chính xác khá thấp. Cũng một phần do số lượng mẫu lớp 1 khá ít nên ảnh hưởng đến việc huấn luyện mô hình.

Qua đó tính toán các độ đo và thống kê quan trọng như độ chính xác, precision, recall, F1-score, và nhiều độ đo khác.

	precision	recall	f1-score	support
0	0.94	0.98	0.96	11424
1	0.70	0.45	0.55	1411
accuracy			0.92	12835
macro avg	0.82	0.72	0.75	12835
weighted avg	0.91	0.92	0.91	12835

Hình 27. Các thang đo cho mô hình

Báo cáo phân loại (Classification Report) cho thấy các chỉ số đánh giá như độ chuẩn xác (precision), độ bao phủ (recall), F1-score và số lượng mẫu (support) cho từng lớp. Mô hình đạt độ chính xác tổng thể là 0.92, cho thấy hiệu suất cao. Lớp '0' có độ chuẩn xác (precision) và độ bao phủ (recall) rất cao (0.94 và 0.98), cho thấy mô hình phân loại rất tốt đối với lớp này. Tuy nhiên, lớp '1' có độ chuẩn xác (precision) và độ bao phủ (recall) rất thấp, số lượng lớp 1 cũng ít hơn rất nhiều so với lớp 0 cho thấy mô hình gặp khó khăn trong việc phân loại đúng lớp này.

CHƯƠNG 4: KẾT QUẢ - KẾT LUẬN

4.1 Nhận xét kết quả đề tài

Trong quá trình thực hiện đề tài, dữ liệu bệnh tiểu đường được thu thập từ các nguồn đáng tin cậy và tiền xử lý để loại bỏ các giá trị bị thiếu hoặc bất thường, đồng thời chuẩn hóa để phù hợp cho việc khai thác mẫu hay luật kết hợp. Thuật toán Apriori và FreeSpan là 2 thuật toán được áp dụng để thực hiện việc khai phá các mẫu phổ biến để hỗ trợ việc dự đoán bệnh tiểu đường. Ngoài ra, thuật toán DecisionTree là thuật toán phân lớp để hỗ trợ việc phân lớp xây dựng mô hình dự đoán. Với điểm đánh giá trên lớp 0 theo f1-score là 96% và trên lớp 1 là 55%. Tổng quan độ chính xác (accuracy) là 92%. Chứng tỏ thuật toán này có thể được sử dụng để đưa ra dự đoán cho bệnh tiểu đường.

4.2 Ưu – nhược điểm của đề tài

***Ưu điểm:**

- Đưa ra các phương pháp tiền xử lý dữ liệu phù hợp, đưa ra các đánh giá phù hợp để đưa vào xây dựng các mô hình dự đoán.

- Đưa ra các phương pháp khai phá dữ liệu phù hợp với tập dữ liệu. Có thể ứng dụng trong lĩnh vực y tế để hỗ trợ đưa các dự đoán khả quan.

***Nhược điểm:**

- Thuật toán khai phá FreeSpan đã quá cũ so với hiện nay. Cho nên có thể không phù hợp với nhiều trường hợp khai phá, xử lý dữ liệu.

- Việc phân lớp dự đoán chỉ mang tính chất tham khảo, không đảm bảo các dự đoán chuẩn xác.

4.3 Hướng dẫn phát triển

Mở rộng đề tài nghiên cứu trên nhiều nguồn dữ liệu khác nhau. Cải tiến thuật toán để phù hợp với nhiều yêu cầu khác nhau, giảm chi phí hệ thống, tối ưu được khả năng dự đoán của các mô hình.

TÀI LIỆU THAM KHẢO

[1] Bộ Y Tế, *Quyết định số 3319/QĐ-BYT ngày 19 tháng 7 năm 2017 Về việc ban hành tài liệu chuyên môn "Hướng dẫn chẩn đoán và điều trị đái tháo đường típ 2"*, Hà Nội: Bộ Y Tế, 2017.

[2] H. Jiawei, M.-A. Behazad, P. Jian, C. Qiming, D. Umeshwar và H. Mei-Chun, "FreeSpan: Frequent pattern-projected sequential pattern mining," *ACM SIGKDD Int'l Conf. Knowledge Discovery,* 2000.

[3] Tarun Gangil, N. Sneha, "Analysis of diabetes mellitus for early prediction using optimal fearures selection," *Journal of Big Data*, 2019.

[4] M. M. Bukhari, B. F. Alkhamess, S. Hussain, A. Gumaei, A. Adel và S. S. Ullah, "An Improved Artificial Neural Network Model for Effective Diabetes Prediction," *Wiley Library*, p. 10, 2021.

[5] V.Jackins, S.Vimal, M.Kaliappan và Mi Young Lee, "AIbased smart prediction of clinical disease using random forest classifier and Naive Bayes," *The Jounal of Supercomputing*, 2020.

[6] Farida Mohsen, Hamada R. H. Al-Abs, Noha A. Yousri, Nady El Hajj và Zubair Shah, "A scoping review of artificial intelligence-based methods for diabetes risk prediction," *npj Digital Medicine*, 2023.

[7] MD. Kamrul Hasan, MD. Ashraful Alam, Dola Das, Eklas Hossain, Senior Member of IEEE và Mahmudul Hansan, " Diabetes Prediction Using Ensembling of Different Machine Learning Classifiers," *IEEE Access*, 2020.

[8] M. Mustafa, "Kaggle," Google LLC, 2023. [Online]. Available: https://www.kaggle.com/datasets/iammustafatz/diabetes-prediction-dataset. [Accessed 15 6 2024].