

**BỘ GIÁO DỤC VÀ ĐÀO TẠO
TRƯỜNG ĐẠI HỌC NGUYỄN TẤT THÀNH
KHOA CÔNG NGHỆ THÔNG TIN**



TIỂU LUẬN KHAI THÁC DỮ LIỆU VÀ ỨNG DỤNG

**Ứng dụng giải thuật Naive Bayes để dự báo người
bị tiểu đường**

Giảng viên hướng dẫn : ThS. Võ Thị Hồng Thắm
Sinh viên thực hiện : Nguyễn Văn Nghĩa
MSSV : 2100009594
Khoá : 2021
Ngành/ chuyên ngành : Trí Tuệ Nhân Tạo

Tp HCM, tháng 1 năm 2024

**BỘ GIÁO DỤC VÀ ĐÀO TẠO
TRƯỜNG ĐẠI HỌC NGUYỄN TẤT THÀNH
KHOA CÔNG NGHỆ THÔNG TIN**



TIỂU LUẬN KHAI THÁC DỮ LIỆU VÀ ỨNG DỤNG

**Ứng dụng giải thuật Naive Bayes để dự báo người bị
tiểu đường**

Giảng viên hướng dẫn : ThS. Võ Thị Hồng Thắm
Sinh viên thực hiện : Nguyễn Văn Nghĩa
MSSV : 2100009594
Khoá : 2021
Ngành/ chuyên ngành : Trí Tuệ Nhân Tạo

Tp HCM, tháng 1 năm 2024

LỜI CẢM ƠN

Em xin gửi lời cảm ơn chân thành nhất tới tất cả các thầy, cô trong khoa Công nghệ thông tin đã đóng góp và hỗ trợ em trong suốt quá trình thực hiện tiểu luận môn học – Khai phá dữ liệu và ứng dụng.

Em muốn bày tỏ lòng biết ơn sâu sắc tới giảng viên hướng dẫn của em là cô Võ Thị Hồng Thắm, người đã dành thời gian và kiến thức của mình để hỗ trợ và chỉ dẫn em trong suốt khoảng thời gian thực hiện. Sự hướng dẫn và những lời khuyên quý báu của cô đã giúp em hiểu sâu hơn về trí tuệ nhân tạo và là nguồn động viên lớn để hoàn thành tiểu luận môn học này.

Em rất tự hào về thành quả và công sức đã bỏ ra để thực hiện tiểu luận môn học này.

Một lần nữa, xin chân thành cảm ơn đến cô người đã đồng hành cùng em trong hành trình này.

Em xin chân thành cảm ơn!

LỜI MỞ ĐẦU

Bệnh tiểu đường đã và đang là một “đại dịch không lây nhiễm” đáng báo động trên toàn cầu với 415 triệu người trưởng thành trên toàn thế giới mắc bệnh, chiếm khoảng 8,8% dân số thế giới.

Tính đến thời điểm hiện tại, tiểu đường là căn bệnh nguy hiểm xếp thứ 3, chỉ sau các bệnh lý tim mạch và ung thư. Tiểu đường biến chứng gây ảnh hưởng nhiều đến chất lượng cuộc sống và sức khỏe của người bệnh, làm tăng nguy cơ mắc bệnh tim mạch, thận hoặc đột quỵ...

Mặc dù chưa thể chữa khỏi hoàn toàn nhưng những người bị đái tháo đường vẫn có thể kiểm soát được bệnh và hạn chế rủi ro biến chứng bằng việc áp dụng một lối sống lành mạnh và kết hợp uống thuốc hạ đường huyết cũng như các phương pháp hỗ trợ khác. Trong tương lai vẫn cần một loại thuốc, biện pháp có thể chữa hoàn toàn, hoặc một bộ máy tính có khả năng chuẩn đoán chính xác và tiết kiệm nhất.

Trong những năm gần đây, sự phát triển mạnh mẽ của công nghệ thông tin đã làm cho khả năng thu thập và lưu trữ thông tin của các hệ thống thông tin tăng nhanh một cách chóng mặt. Sự bùng nổ này đã dẫn tới một yêu cầu cấp thiết là cần có những kỹ thuật và công cụ mới để chuyển đổi lượng dữ liệu khổng lồ kia thành các tri thức có ích. Từ đó, các kỹ thuật khai phá dữ liệu đã trở thành một lĩnh vực thời sự của nền công nghệ thông tin trên thế giới hiện nay.

Báo cáo với đề tài “Xây dựng mô hình dự đoán người bị tiểu đường dựa trên giải thuật Bayesian” khảo sát lĩnh vực khai phá dữ liệu dùng Naive Bayes. Tiểu luận tập trung vào phân lớp bằng xác suất có điều kiện theo định lý Bayes để đi sâu vào việc khai phá dữ liệu từ các thông tin thu thập được để dự đoán người có bị tiểu đường hay không.

Mặc dù đã hết sức nỗ lực, song do thời gian và kinh nghiệm nghiên cứu khoa học còn hạn chế nên không thể tránh khỏi những thiếu sót. Em rất mong nhận được sự góp ý của các thầy cô và bạn bè để hiểu biết của mình ngày một hoàn thiện hơn!

TRƯỜNG ĐẠI HỌC NGUYỄN TẤT THÀNH
TRUNG TÂM KHẢO THÍ

KỲ THI KẾT THÚC HỌC PHẦN
HỌC KỲ NĂM HỌC -

PHIẾU CHẤM THI TIỂU LUẬN/ĐỒ ÁN

Môn thi: Khai thác dữ liệu và ứng dụng Lớp học phần: 21DTH2B

Nhóm sinh viên thực hiện :

1. Huỳnh Quy Bình. Tham gia đóng góp:

2. Nguyễn Hữu Thịnh. Tham gia đóng góp:

3. Nguyễn Văn Nghĩa Tham gia đóng góp:

4. Nguyễn Hoàng Ngân Phú Tham gia đóng góp:

Ngày thi: 29/01/2024..... Phòng thi:

Đề tài tiểu luận/báo cáo của sinh viên : Ứng dụng giải thuật Native Bayes để dự đoán người bị tiểu đường

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

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

Giảng viên chấm thi
(ký, ghi rõ họ tên)

ThS. Võ Thị Hồng Thắm

MỤC LỤC

LỜI CẢM ƠN.....	i
LỜI MỞ ĐẦU	ii
MỤC LỤC	iv
DANH MỤC BẢNG	vi
DANH MỤC HÌNH	Error! Bookmark not defined.
KÍ HIỆU CÁC CỤM TỪ VIẾT TẮT.....	vii
CHƯƠNG 1 GIỚI THIỆU TỔNG QUÁT VỀ KHAI PHÁ DỮ LIỆU VÀ ỨNG DỤNG ..	1
1.1 Cơ sở lý thuyết Khai phá dữ liệu và ứng dụng.....	1
1.1.1 Các khái niệm cơ bản	1
1.1.2 Các kiến trúc của khai phá dữ liệu	2
1.1.3 Chức năng của khai phá dữ liệu	4
1.1.4 Quá trình khai phá tri thức.....	6
1.1.5 Khai phá dữ liệu và các lĩnh vực liên quan	8
1.1.6 Các yếu tố cơ bản trong khai phá dữ liệu	10
1.1.7 Quy trình và các kỹ thuật khai phá dữ liệu.....	12
CHƯƠNG 2 CƠ SỞ LÝ THUYẾT.....	15
2.1 Phát biểu định lý Bayes	15
2.2 Công thức Bayes dùng khi nào?	15
2.3 Lịch sử	16
2.4 Định lý Bayes được ứng dụng trong lĩnh vực nào?.....	16
2.5 Phân lớp Native Bayes	17
2.5.1 Định nghĩa	17
2.5.2 Giải thuật	17
2.2.3. Ưu nhược điểm của Native Bayes	18
CHƯƠNG 3	19
PHÂN TÍCH YÊU CẦU	19
3.1 Giới thiệu đề tài	19
3.1.1 Tên đề tài	19
3.1.2 Yêu cầu đề tài	19
3.1.3 Hiện trạng của bệnh tiêu đường trong đời sống ngày nay.....	19

3.1.4 Hiểu rõ về mức độ nguy hiểm của bệnh tiểu đường	21
3.2 Mô tả đề tài	26
3.3 Hướng xây dựng đề tài	27
CHƯƠNG 4	29
XÂY DỰNG MÔ HÌNH	29
4.1 Mô tả dữ liệu.....	29
4.2 Đọc và chuẩn bị dữ liệu	29
4.3 Chia dữ liệu thành tập huấn luyện và tập kiểm tra	30
4.4 Xây dựng và huấn luyện mô hình Native Bayes	31
4.5 Dự đoán và đánh giá hiệu suất	31
4.6 Trực quan hóa kết quả	32
CHƯƠNG 5	33
THỰC NGHIỆM MÔ HÌNH.....	33
5.1. Ma Trận Nhầm Lẫn (Confusion Matrix):.....	33
5.2. Độ Chính Xác (Accuracy):	33
CHƯƠNG 6	34
KẾT LUẬN VÀ HƯỚNG PHÁT TRIỂN	34
6.1 Kết quả đạt được.....	34
6.2 Hạn chế của đề tài.....	34
6.3 Hướng phát triển.....	35
TÀI LIỆU THAM KHẢO	36

DANH MỤC HÌNH

Hình 1- 1: Phân cụm cơ sở dữ liệu	1
Hình 1- 2: Khai phá tri thức đưa ra quyết định	2
Hình 1- 3: Kiến trúc khai phá dữ liệu.....	3
Hình 1- 4: Pattern evaluation.....	6
Hình 1- 5: Knowledge presentation.....	7
Hình 1- 6: Các lĩnh vực liên quan khai phá dữ liệu.....	9
Hình 1- 7: Pattern visualization and knowledge presentation	11
Hình 1- 8: Quy trình CRISP-DM	13
Hình 1- 9: Các kỹ thuật khai phá.....	14
Hình 2- 1: Định lý Bayes.....	15
Hình 3- 1: Đói và mệt	22
Hình 3- 2: Khô miệng.....	23
Hình 3- 3: Sụt cân nhiều.....	23
Hình 3- 4: Thị lực giảm	24
Hình 3- 5: Nhiễm trùng nấm men.....	25
Hình 3- 6: Vết thương lâu lành.....	25
Hình 3- 7: Thai kỳ	26
Hình 4- 1: Mô tả dữ liệu	29
Hình 4- 2: Đọc và chuẩn bị dữ liệu 1	29
Hình 4- 3: Đọc và chuẩn bị dữ liệu 2	30
Hình 4- 4: Đọc và chuẩn bị dữ liệu 3	30
Hình 4- 5: Tập huấn luyện và kiểm tra.....	30
Hình 4- 6: Xây dựng và huấn luyện mô hình	31
Hình 4- 7: Dự đoán và đánh giá hiệu suất.....	31
Hình 4- 8: Kết quả độ chính xác mô hình	32
Hình 4- 9: Trực quan hóa kết quả.....	32
Hình 5- 1: Ma trận nhầm lẫn (Confusion Matrix)	33
Hình 5- 2: Độ chính xác	33

KÍ HIỆU CÁC CỤM TỪ VIẾT TẮT

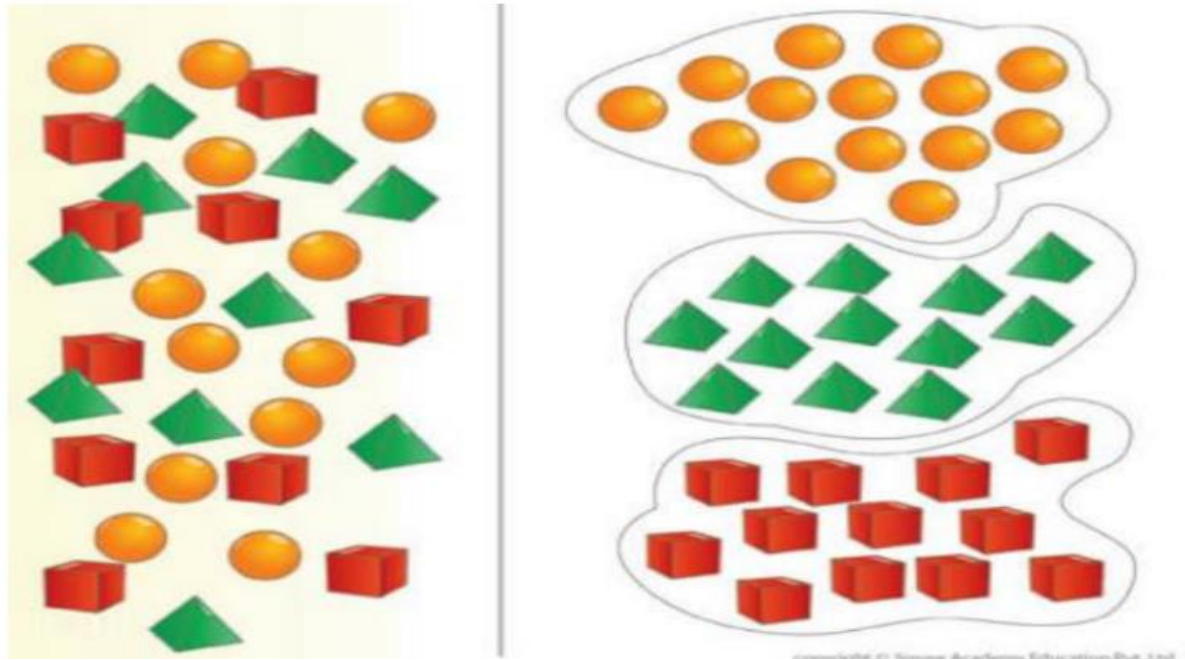
Chữ viết tắt	Ý nghĩa
KDD	Knowledge Discovery in Databases (Khám phá tri thức từ cơ sở dữ liệu)
CSDL	Cơ sở dữ liệu
CRISP-DM	Cross-Industry Standard Process for Data Mining (quy trình tiêu chuẩn cho khai thác dữ liệu trong nhiều lĩnh vực)

CHƯƠNG 1

GIỚI THIỆU TỔNG QUÁT VỀ KHAI PHÁ DỮ LIỆU VÀ ỨNG DỤNG

1.1 Cơ sở lý thuyết Khai phá dữ liệu và ứng dụng

1.1.1 Các khái niệm cơ bản

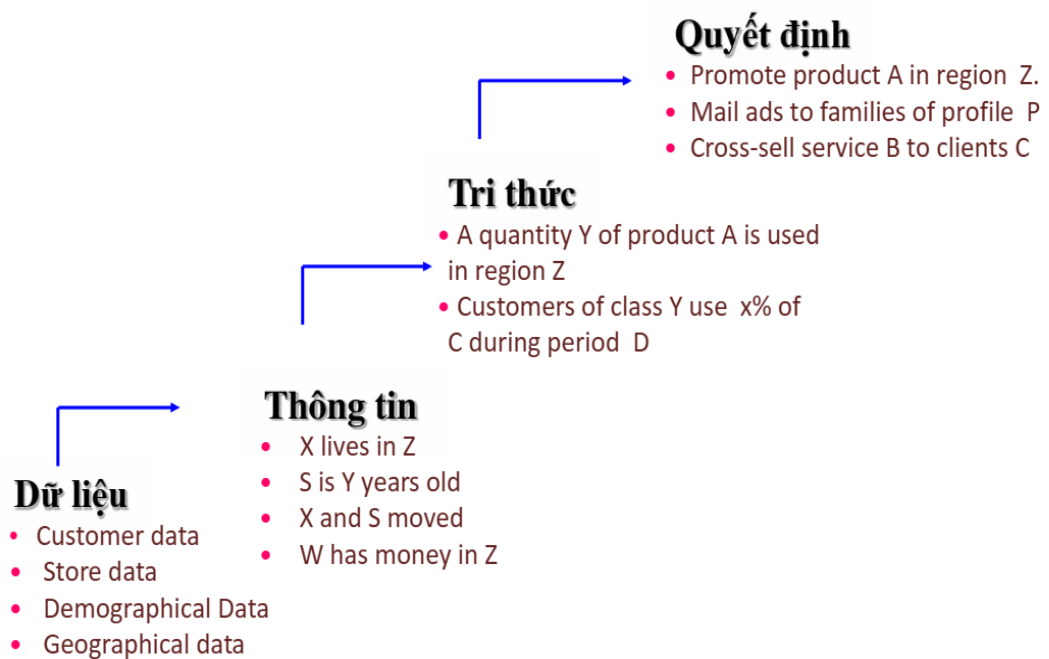


Hình 1- 1: Phân cụm cơ sở dữ liệu

- Dữ liệu (Data): có thể xem là chuỗi các bit, là số, ký tự...mà chúng ta tập hợp hàng ngày trong công việc
 - Thông tin (Information): là tập hợp của những mảnh dữ liệu đã được chất lọc dùng mô tả, giải thích đặc tính của một đối tượng nào đó
 - Tri thức (Knowledge): là tập hợp những thông tin có liên hệ với nhau, có thể xem tri thức là sự kết tinh từ dữ liệu. Tri thức thể hiện tư duy của con người về một vấn đề.
- Khai phá dữ liệu là quá trình khai thác những thông tin tiềm ẩn có tính dự đoán, những thông tin có nhiều ý nghĩa, hữu ích từ những cơ sở dữ liệu lớn,

nó được coi như là một bước trong quá trình khám phá tri thức (Knowledge Discovery in Databases – KDD). Khai phá dữ liệu là giai đoạn quan trọng nhất trong tiến trình khám phá tri thức từ cơ sở dữ liệu, các tri thức này có rất nhiều ý nghĩa, là cơ sở hỗ trợ trong việc ra quyết định trong khoa học và kinh doanh.

- Khai phá dữ liệu là giai đoạn thiết yếu, đây là bước quan trọng và tốn nhiều thời gian nhất của toàn bộ quá trình khám phá tri thức, là bước áp dụng những kỹ thuật khai phá để khai thác, trích xuất thông tin có ích, những mẫu điển hình, những mối liên hệ đặc biệt có nhiều giá trị, mang nhiều ý nghĩa từ dữ liệu.

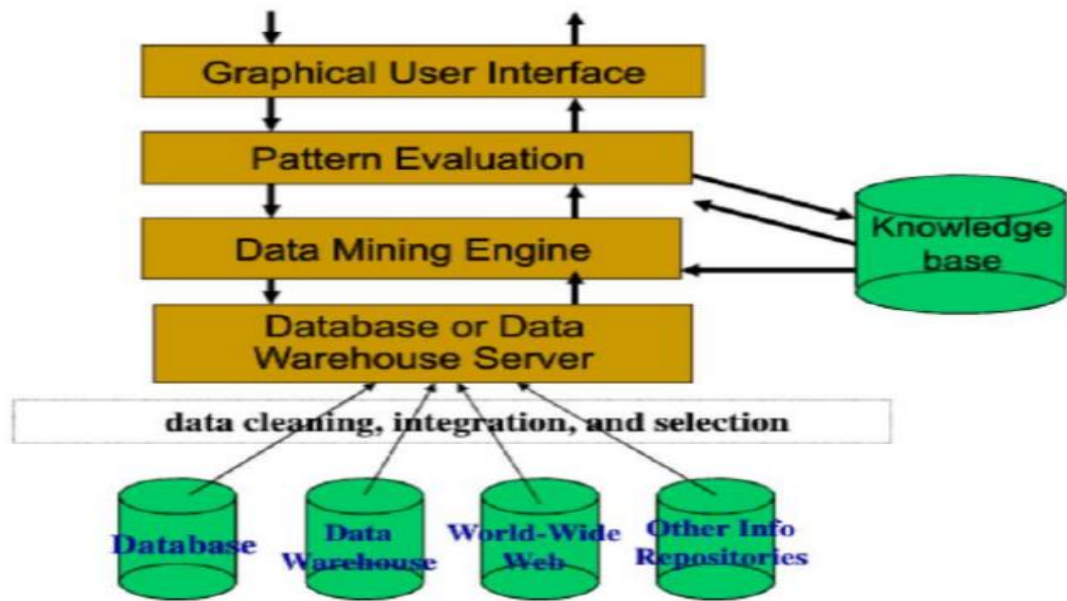


Hình 1- 2: Khai phá tri thức đưa ra quyết định

1.1.2 Các kiến trúc của khai phá dữ liệu

- Khai phá dữ liệu là quá trình rút trích thông tin bổ ích từ những kho dữ liệu lớn, là quá trình chính trong khai phá tri thức từ cơ sở dữ liệu.

- Kiến trúc của một hệ thống khai phá dữ liệu có các thành phần như sau:



Hình 1- 3: Kiến trúc khai phá dữ liệu

- **Cơ sở dữ liệu, kho dữ liệu hoặc lưu trữ thông tin khác:** Đây là một hay các tập cơ sở dữ liệu, các kho dữ liệu, các trang tính hay các dạng khác của thông tin được lưu trữ. Các kỹ thuật làm sạch dữ liệu và tích hợp dữ liệu có thể được thực hiện.
- **Máy chủ cơ sở dữ liệu (Database or Warehouse Server):** Máy chủ có trách nhiệm lấy những dữ liệu thích hợp dựa trên những yêu cầu khám phá của người dùng.
- **Cơ sở tri thức (Knowledge-base):** Đây là miền tri thức dùng để tìm kiếm hay đánh giá độ quan trọng của các mẫu kết quả thu được. Tri thức này có thể bao gồm một sự phân cấp khái niệm dùng để tổ chức các thuộc tính hay các giá trị thuộc tính ở các mức trừu tượng khác nhau.
- **Máy khai phá dữ liệu (Data mining engine):** Là một hệ thống khai phá dữ liệu cần phải có một tập các module chức năng để thực hiện công việc, chẳng hạn như kết hợp, phân lớp, phân cụm.
- **Module đánh giá mẫu (Pattern evaluation):** Bộ phận tương tác với các modul khai phá dữ liệu để tập trung vào việc duyệt tìm các mẫu đáng được quan tâm. Nó có thể dùng các ngưỡng về độ quan tâm để lọc mẫu đã

khám phá được. Cũng có thể modul đánh giá mẫu được tích hợp vào module khai phá dữ liệu, tùy theo cách cài đặt của phương pháp khai phá dữ liệu được dùng.

- **Giao diện đồ họa cho người dùng (Graphical user interface):** Bộ phận này cho phép người dùng giao tiếp với hệ thống khai phá dữ liệu. Thông qua giao diện này người dùng tương tác với hệ thống bằng cách đặc tả một yêu cầu khai phá hay một nhiệm vụ, cung cấp thông tin trợ giúp cho việc tìm kiếm và thực hiện khai phá thăm dò trên các kết quả khai phá trung gian. Ngoài ra bộ phận này còn cho phép người dùng xem các lược đồ cơ sở dữ liệu, lược đồ kho dữ liệu, các đánh giá mẫu và hiển thị các mẫu trong các khuôn dạng khác nhau.

1.1.3 *Chức năng của khai phá dữ liệu*

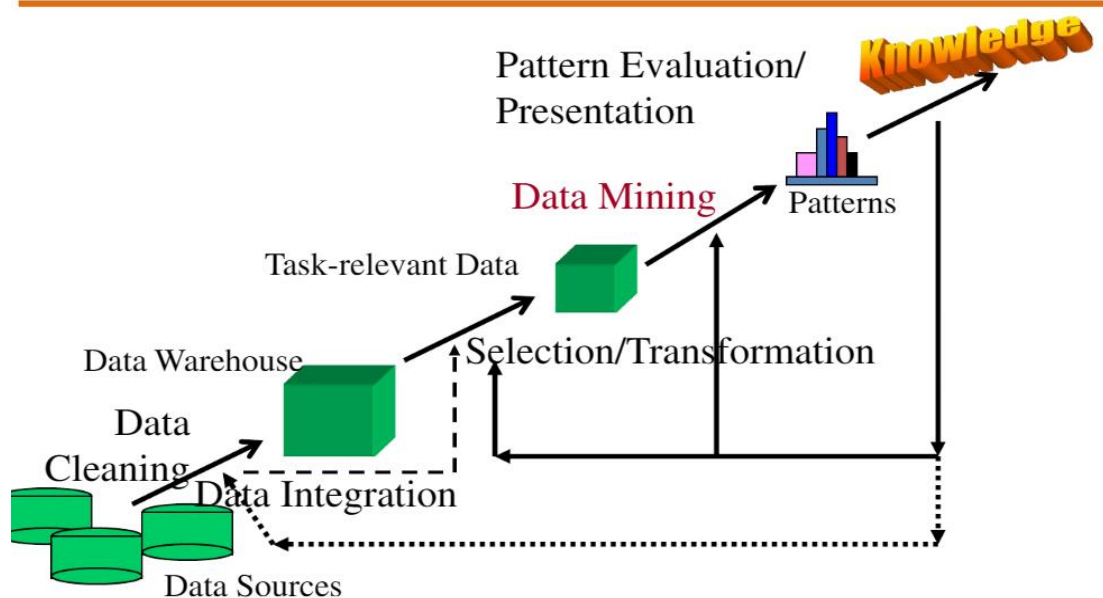
- Các kỹ thuật khai phá dữ liệu thực hiện 2 chức năng chính:
 - Chức năng mô tả: mô tả về các tính chất hoặc các đặc tính chung của dữ liệu trong cơ sở dữ liệu, các kỹ thuật này gồm có: phân cụm (Clustering), tổng hợp (Summarization), trực quan hóa (Visualization), phân tích sự phát triển và độ lệch (Evolution and deviation analysis), phân tích luật kết hợp (Association rules),...
 - Chức năng dự đoán: đưa ra các dự đoán dựa vào các suy diễn trên dữ liệu hiện thời, các kỹ thuật này gồm có: phân lớp (Classification), hồi quy (Regression), cây quyết định (Decision tree),...
- Phân lớp (classification):
 - Là việc xác định một hàm ánh xạ từ một mẫu dữ liệu vào một trong số các lớp đã được biết trước đó. Mục tiêu của thuật toán phân lớp là tìm ra mối quan hệ nào đó giữa thuộc tính dự báo và thuộc tính phân lớp. Như thế quá trình phân lớp có thể sử dụng mối quan hệ này để dự báo cho các mục mới. Các kiến thức được phát hiện biểu diễn dưới dạng các luật theo cách sau: “Nếu các thuộc tính dự báo của một mục thỏa mãn điều kiện của các tiền đề thì mục nằm trong lớp chỉ ra trong kết luận.”

- Hồi quy (regression):
 - Là việc học một hàm ánh xạ từ một mẫu dữ liệu thành một biến dự đoán có giá trị thực. Nhiệm vụ của hồi quy tương tự như phân lớp, điểm khác nhau chính là ở chỗ thuộc tính để dự báo là liên tục chứ không phải là rời rạc. Việc dự báo các giá trị số thường được làm bởi các phương pháp thống kê cổ điển, chẳng hạn như hồi quy tuyến tính. Tuy nhiên, phương pháp mô hình hóa cũng được sử dụng, ví dụ: cây quyết định.
- Phân cụm (clustering):
 - Là việc mô tả chung để tìm ra các tập hay các nhóm, loại mô tả dữ liệu. Các nhóm có thể tách nhau hoặc phân cấp hay gộp lên nhau. Có nghĩa là dữ liệu có thể vừa thuộc nhóm này vừa thuộc nhóm khác. Các ứng dụng khai phá dữ liệu có nhiệm vụ phân nhóm như phát hiện tập các khách hàng có phản ứng giống nhau trong CSDL tiếp thị; xác định các quang phổ từ các phương pháp đo tia hồng ngoại, ... liên quan chặt chẽ đến việc phân nhóm là nhiệm vụ đánh giá dữ liệu, hàm mật độ xác suất đa biến/ các trường hợp trong CSDL.
- Phân tích sự phát triển và độ lệch (evolution and deviation analysis):
 - Nhiệm vụ này tập trung vào khám phá hầu hết sự thay đổi có nghĩa dưới dạng độ đo đã biết trước hoặc giá trị chuẩn, phát hiện độ lệch đáng kể giữa nội dung của tập con dữ liệu thực và nội dung mong đợi. Hai mô hình độ lệch hay dùng là lệch theo thời gian hay lệch theo nhóm. Độ lệch theo thời gian là sự thay đổi có ý nghĩa của dữ liệu theo thời gian. Độ lệch theo nhóm là sự khác nhau giữa dữ liệu trong hai tập con dữ liệu, ở đây tính cả trường hợp tập con dữ liệu này thuộc tập con kia, nghĩa là xác định dữ liệu trong một nhóm con của đối tượng có khác đáng kể so với toàn bộ đối tượng không? Theo cách này, sai sót dữ liệu hay sai lệch so với giá trị thông thường được phát hiện.
 - Vì những nhiệm vụ này yêu cầu số lượng và các dạng thông tin rất khác nhau nên chúng thường ảnh hưởng đến việc thiết kế và chọn phương pháp khai phá dữ liệu khác nhau. Ví dụ như phương pháp cây quyết định tạo ra

được một mô tả phân biệt được các mẫu giữa các lớp nhưng không có tính chất và đặc điểm của lớp.

1.1.4 *Quá trình khai phá tri thức*

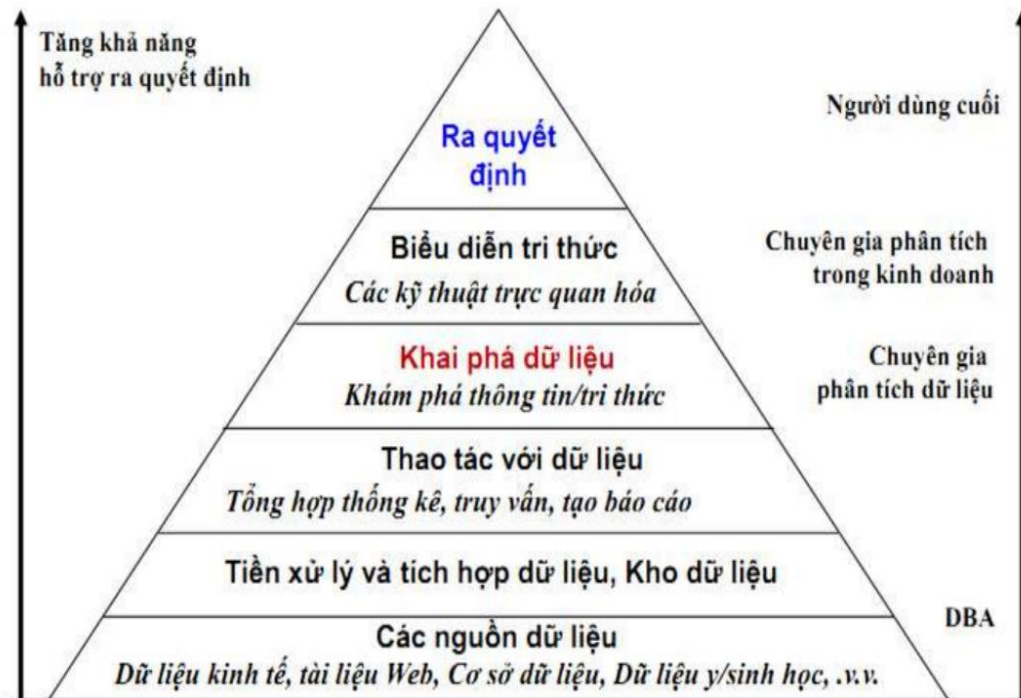
- Khám phá tri thức từ cơ sở dữ liệu (Knowledge Discovery in Databases – KDD):
 - “KDD là quá trình tự động trích rút các tri thức tiềm ẩn trong một lượng dữ liệu lớn” -Fayyad, Platetsky-Shapiro, Smyth (1996).
 - Khám phá tri thức từ cơ sở dữ liệu là quy trình bao gồm nhiều công đoạn như: xác định vấn đề, tập hợp và chọn lọc dữ liệu, khai thác dữ liệu, đánh giá kết quả, giải thích dữ liệu, áp dụng tri thức vào thực tế.
- Quá trình khám phá tri thức là một chuỗi lặp gồm các bước:
 - Data cleaning (làm sạch dữ liệu)
 - Data integration (tích hợp dữ liệu)
 - Data selection (chọn lựa dữ liệu)
 - Data transformation (biến đổi dữ liệu)
 - Data mining (khai phá dữ liệu)
 - Pattern evaluation (đánh giá mẫu)



Hình 1- 4: *Pattern evaluation*

- Knowledge presentation (biểu diễn tri thức)

Quá trình khám phá tri thức trong thông minh doanh nghiệp



Hình 1- 5: Knowledge presentation

- Tri thức đạt được từ quá trình khai phá:
 - Tri thức đạt được có thể có tính mô tả hay dự đoán tùy thuộc vào quá trình khai phá cụ thể.
 - Mô tả (Descriptive): có khả năng đặc trưng hóa các thuộc tính chung của dữ liệu được khai phá (Tình huống 5, 6).
 - Dự đoán (Predictive): có khả năng suy luận từ dữ liệu hiện có để dự đoán (Tình huống 1, 2, 3 và 4).
 - Tri thức đạt được có thể có cấu trúc, bán cấu trúc, hoặc phi cấu trúc.
 - Tri thức đạt được có thể được dùng trong việc hỗ trợ ra quyết định, điều khiển quy trình, quản lý thông tin, xử lý truy vấn ...

1.1.5 *Khai phá dữ liệu và các lĩnh vực liên quan*

- Dữ liệu được thu thập hàng ngày là rất lớn:
 - Các CSDL khổng lồ.
 - Dữ liệu từ Internet.
 - Nhà bác học nổi tiếng Karan Sing đã từng nói rằng “Chúng ta đang ngập chìm trong biển thông tin nhưng lại đang khát tri thức”.
- Khai phá dữ liệu (Data mining) là một bước trong quy trình khám phá tri thức, nhằm:
 - Rút trích thông tin hữu ích, chưa biết, tiềm ẩn trong khối dữ liệu lớn.
 - Phân tích dữ liệu bán tự động.
 - Giải thích dữ liệu trên các tập dữ liệu lớn.

- Các lĩnh vực liên quan:



Hình 1- 6: Các lĩnh vực liên quan khai phá dữ liệu

- Khả năng đóng góp của công nghệ cơ sở dữ liệu:
 - Công nghệ cơ sở dữ liệu cho việc quản lý dữ liệu được khai phá.
 - Dữ liệu rất lớn, có thể vượt quá khả năng của bộ nhớ chính (mainmemory).
 - Dữ liệu được thu thập theo thời gian.
- Các hệ cơ sở dữ liệu có khả năng xử lý hiệu quả lượng lớn dữ liệu với các cơ chế phân trang (paging) và hoán chuyển (swapping) dữ liệu vào/ra bộ nhớ chính.
- Các hệ cơ sở dữ liệu hiện đại có khả năng xử lý nhiều loại dữ liệu phức tạp: spatial, temporal, spatiotemporal, multimedia, text, Web...
- Các chức năng khác (xử lý đồng thời, bảo mật, hiệu năng, tối ưu hóa,...) của các hệ cơ sở dữ liệu đã được phát triển tốt.
- Học máy với khai phá dữ liệu:
 - Học máy:
 - Machine Learning.

- Cách máy tính có thể học (nâng cao năng lực) dựa trên dữ liệu.
- Các chương trình máy tính tự động học được các mẫu phức tạp và ra quyết định thông minh dựa trên dữ liệu. Ví dụ “học được chữ viết tay trên thư thông qua một tập ví dụ”.
- Là một lĩnh vực nghiên cứu phát triển nhanh.
- Một số nội dung học máy với khai phá dữ liệu:
 - Học giám sát (supervised learning): dữ liệu huấn luyện đã được gán nhãn.
 - Học không giám sát (unsupervised learning): dữ liệu huấn luyện không được gán nhãn.
 - Học bán giám sát (semi-supervised learning): sử dụng cả dữ liệu huấn luyện được gán nhãn và dữ liệu huấn luyện không gán nhãn.
 - Học tăng cường (Reinforcement learning).

1.1.6 Các yếu tố cơ bản trong khai phá dữ liệu

- Dữ liệu cụ thể sẽ được khai phá (task-relevant data).
- Loại tri thức sẽ đạt được (kind of knowledge).
- Tri thức nền (background knowledge).
- Các độ đo (interestingness measures).
- Các kỹ thuật biểu diễn tri thức/trực quan hóa mẫu (pattern visualization and knowledge presentation).
- Dữ liệu cụ thể sẽ được khai phá (task-relevant data):
 - Phần dữ liệu từ các dữ liệu nguồn được quan tâm.
 - Tương ứng với các thuộc tính hay chiều dữ liệu được quan tâm.
 - Bao gồm: tên kho dữ liệu/cơ sở dữ liệu, các bảng dữ liệu hay các khối dữ liệu, các điều kiện chọn dữ liệu, các thuộc tính hay chiều dữ liệu được quan tâm, các tiêu chí gom nhóm dữ liệu.
- Loại tri thức sẽ đạt được (kind of knowledge):

- Bao gồm: đặc trưng hóa dữ liệu, phân biệt hóa dữ liệu, mô hình phân tích kết hợp hay tương quan, mô hình phân lớp, mô hình dự đoán, mô hình gom cụm, mô hình phân tích phần tử biên, mô hình phân tích tiến hóa.
- Tri thức nền (background knowledge):
- Tương ứng với lĩnh vực cụ thể sẽ được khai phá.
 - Hỗ trợ khai phá dữ liệu ở nhiều mức trừu tượng khác nhau
 - Đánh giá các mẫu được tìm thấy.
 - Bao gồm: các phân cấp ý niệm, niềm tin của người sử dụng về các mối quan hệ của dữ liệu.
- Các kỹ thuật biểu diễn tri thức/trực quan hóa mẫu (pattern visualization and knowledge presentation):
- Xác định dạng các mẫu/tri thức được tìm thấy để thể hiện đến người sử dụng.
 - Bao gồm: luật (rules), bảng (tables), báo cáo (reports), biểu đồ (charts), đồ thị (graphs), cây (trees), và khối (cubes).



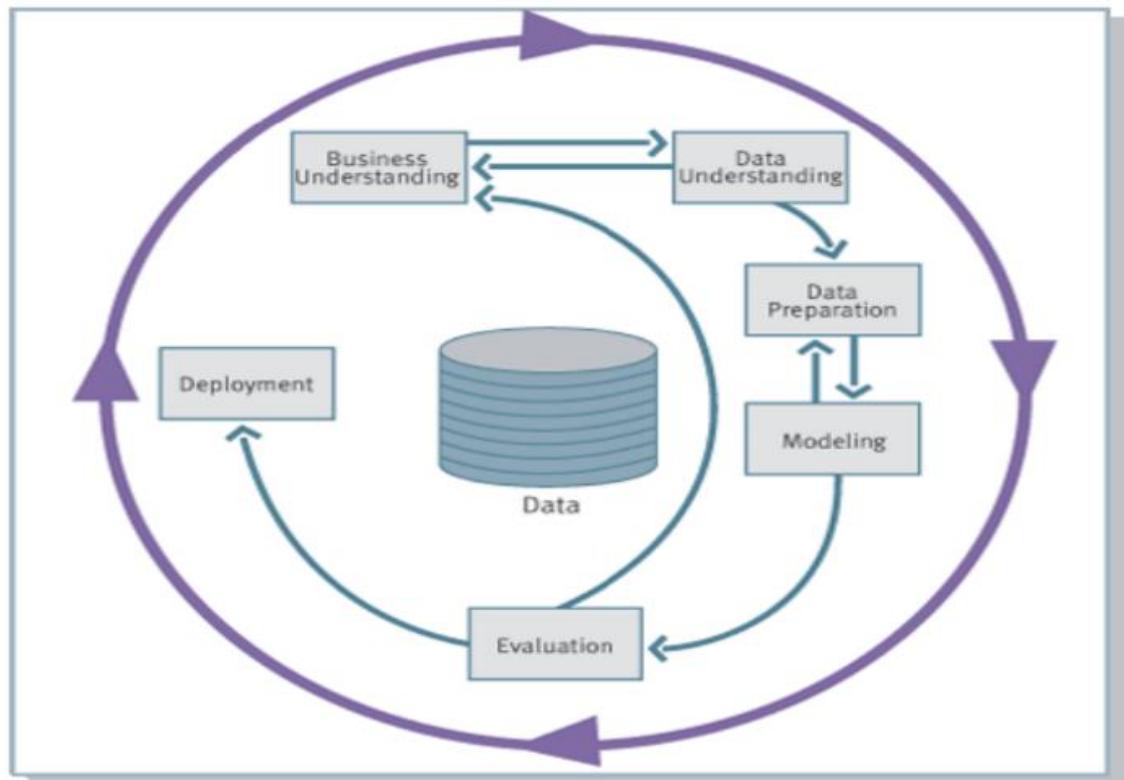
Hình 1- 7: Pattern visualization and knowledge presentation

- Gợi ý và quảng cáo trực tuyến
- Phân tích xu hướng thông tin
- Phân tích và nhận biết nội dung ảnh
- Khai phá và phân tích quan điểm
- Phân tích đánh giá người dùng
- Quản lý, cảnh báo danh tiếng
- Phân tích và dự báo thủy văn

1.1.7 *Quy trình và các kỹ thuật khai phá dữ liệu*

- Quy trình khai phá dữ liệu là một chuỗi lặp (iterative) (và tương tác (interactive)) gồm các bước (giai đoạn) bắt đầu với dữ liệu thô (raw data) và kết thúc với tri thức (knowledge of interest) đáp ứng được sự quan tâm của người sử dụng.
 - Cross Industry Standard Process for Data Mining (CRISP-DM at www.crisp-dm.org)
 - SEMMA (Sample, Explore, Modify, Model, Assess) at the SAS Institute
- Sự cần thiết của một quy trình khai phá dữ liệu:
 - Cách thức tiến hành (hoạch định và quản lý) dự án khai phá dữ liệu có hệ thống.
 - Đảm bảo nỗ lực dành cho một dự án khai phá dữ liệu được tối ưu hóa.
 - Việc đánh giá và cập nhật các mô hình trong dự án được diễn ra liên tục.

- Quy trình CRISP-DM:



Hình 1- 8: Quy trình CRISP-DM

- Được khởi xướng từ 09/1996 và được hỗ trợ bởi hơn 200 thành viên.
- Chuẩn mở.
- Hỗ trợ công nghiệp/ứng dụng và công cụ khai phá dữ liệu hiện có.
- Tập trung vào các vấn đề nghiệp vụ cũng như phân tích kỹ thuật.
- Tạo ra một khung thức hướng dẫn quy trình khai phá dữ liệu.
- Có nền tảng kinh nghiệm từ các lĩnh vực ứng dụng.
- Quy trình CRISP-DM là một quy trình lặp, có khả năng quay lui (backtracking) gồm 6 giai đoạn:
 - Tìm hiểu nghiệp vụ (Business understanding)
 - Tìm hiểu dữ liệu (Data understanding)
 - Chuẩn bị dữ liệu (Data preparation)
 - Mô hình hoá (Modeling)
 - Đánh giá (Evaluation)

- Triển khai (Deployment)
- Các kỹ thuật khai phá:

◆ Các kỹ thuật khai phá



Hình 1- 9: Các kỹ thuật khai phá

CHƯƠNG 2

CƠ SỞ LÝ THUYẾT

2.1 Phát biểu định lý Bayes

Định lý Bayes là một nguyên tắc quan trọng trong xác suất và thống kê, được đặt tên theo nhà toán học người Anh Thomas Bayes. Định lý này thường được sử dụng để cập nhật xác suất của một giả thuyết dựa trên bằng chứng mới. Phát biểu của định lý Bayes có thể được diễn đạt như sau:

- Cho A và B là hai sự kiện khác nhau với $P(B) > 0$, thì xác suất có điều kiện của A khi đã biết B đã xảy ra được tính bằng công thức:

$$P(A|B) = (P(B|A) \cdot P(A)) / P(B)$$

- Trong đó:

$P(A|B)$ là xác suất có điều kiện của sự kiện A khi đã biết B đã xảy ra.

$P(B|A)$ là xác suất có điều kiện của sự kiện B khi đã biết A đã xảy ra.

$P(A)$ là xác suất ban đầu của sự kiện A.

$P(B)$ là xác suất ban đầu của sự kiện B.

$$\boxed{P(A|B)} = \boxed{P(A)} \times \frac{\boxed{P(B|A)}}{\boxed{P(B)}}$$

The diagram shows the formula for Bayes' Theorem with four terms highlighted in colored boxes and labeled below: $P(A|B)$ is in a pink box labeled 'posterior'; $P(A)$ is in a cyan box labeled 'prior'; $P(B|A)$ is in a green box labeled 'likelihood'; and $P(B)$ is in a blue box labeled 'marginal'.

Hình 2- 1: Định lý Bayes

2.2 Công thức Bayes dùng khi nào?

Công thức Bayes được sử dụng trong các bài toán xác suất và thống kê để tính toán xác suất điều kiện dựa trên thông tin có sẵn. Một số trường hợp khi cần sử dụng công thức Bayes bao gồm:

1. Khi có thông tin xác suất ban đầu và muốn tính toán xác suất sau khi có thêm thông tin mới.
2. Khi có thông tin xác suất rời rạc (discrete) và muốn tính toán xác suất xảy ra của một sự kiện khác.
3. Khi muốn tính toán xác suất hậu nghiệm (posterior probability) của một giả thuyết dựa trên dữ liệu thực tế.
4. Khi muốn tính toán xác suất của một giả thuyết dựa trên xác suất liên tục (uniform prior probability).
5. Khi muốn phân loại dữ liệu dựa trên các đặc trưng và thông tin xác suất đã biết. Dùng công thức Bayes giúp tối ưu hóa việc tính toán xác suất và đưa ra các quyết định dựa trên thông tin có sẵn.

2.3 Lịch sử

Định lý Bayes được đặt tên cho Bộ trưởng và nhà thống kê người Anh, Mục sư Thomas Bayes, người đã xây dựng một phương trình cho công trình của mình "Một bài luận hướng tới giải quyết vấn đề trong học thuyết cơ hội". Sau cái chết của Bayes, bản thảo được Richard Price biên tập và sửa chữa trước khi xuất bản vào năm 1763. Sẽ chính xác hơn nếu coi định lý này là quy tắc Bayes-Price, vì đóng góp của Price là rất đáng kể. Công thức hiện đại của phương trình do nhà toán học người Pháp Pierre-Simon Laplace nghĩ ra vào năm 1774, người không biết gì về công trình của Bayes. Laplace được công nhận là nhà toán học chịu trách nhiệm về sự phát triển của xác suất Bayes .

2.4 Định lý Bayes được ứng dụng trong lĩnh vực nào?

Định lý Bayes có ứng dụng rất rộng trong nhiều lĩnh vực, bao gồm xác suất thống kê, trí tuệ nhân tạo, học máy, khoa học dữ liệu, và thị giác máy tính. Một số ví dụ cụ thể về ứng dụng của định lý Bayes bao gồm:

1. **Phân loại email:** Định lý Bayes được sử dụng để xác định xem một email có phải là rác hay không dựa trên các từ khóa và thuộc tính khác.
2. **Nhận dạng giọng nói:** Định lý Bayes có thể được áp dụng để nhận dạng giọng nói của một người dựa trên các đặc điểm âm thanh và thông tin khác.

3. **Xác định tội phạm:** Định lý Bayes có thể được sử dụng để xác định khả năng nghi phạm là tội phạm dựa trên các chứng cứ và thông tin khác.
4. **Dự đoán thị trường tài chính:** Định lý Bayes có thể được áp dụng để dự đoán biến động của thị trường tài chính dựa trên các chỉ số kinh tế và thông tin khác.
5. **Xác định tình trạng bệnh:** Định lý Bayes có thể được sử dụng để xác định xác suất mắc bệnh dựa trên các triệu chứng và thông tin y tế khác.

2.5 Phân lớp Native Bayes

2.5.1 Định nghĩa

Mô hình phân loại Native Bayes là một mô hình học máy dựa trên định lý Bayes, được sử dụng để thực hiện tác vụ phân loại. Nó là một trong những phương pháp đơn giản và hiệu quả cho các vấn đề phân loại với dữ liệu có nhiều đặc trưng. Mô hình này dựa trên giả định "ngây thơ" (naïve), giả sử rằng các đặc trưng đầu vào độc lập với nhau khi đã biết lớp của dữ liệu.

2.5.2 Giải thuật

- Giải thuật của mô hình phân loại Native Bayes bao gồm các bước cơ bản sau:
 - a) **Thu thập dữ liệu:** Xây dựng tập dữ liệu huấn luyện, trong đó mỗi mẫu dữ liệu được gán một nhãn (lớp).
 - b) **Tính toán xác suất prior:** Tính toán xác suất của mỗi lớp dựa trên tần suất xuất hiện của nó trong tập dữ liệu huấn luyện.
 - c) **Tính toán xác suất điều kiện:** Tính toán xác suất của từng đặc trưng dựa trên lớp của dữ liệu. Đối với mỗi đặc trưng, tính xác suất của nó khi biết lớp.
 - d) **Tính toán xác suất posterior:** Sử dụng định lý Bayes để tính toán xác suất của mỗi lớp khi đã biết các đặc trưng.
 - e) **Phân loại:** Dựa trên xác suất posterior, mô hình chọn lớp có xác suất cao nhất làm dự đoán cho mẫu dữ liệu mới.
 - f) **Đánh giá mô hình:** Đánh giá hiệu suất của mô hình bằng cách sử dụng tập dữ liệu kiểm thử.

2.2.3. Ưu nhược điểm của Native Bayes

Ưu điểm của mô hình Native Bayes:

- Dễ triển khai và nhanh chóng huấn luyện: Mô hình Native Bayes thường dễ triển khai và nhanh chóng huấn luyện, đặc biệt là với dữ liệu lớn.
- Hiệu suất tốt cho các tập dữ liệu lớn: Với các tập dữ liệu lớn, mô hình Native Bayes có thể cung cấp hiệu suất tốt và độ chính xác cao. Điều chỉnh dễ dàng: Mô hình không có nhiều siêu tham số cần điều chỉnh, giúp đơn giản hóa quá trình huấn luyện và tối ưu hóa mô hình.
- Phù hợp với dữ liệu có số chiều cao: Thường hoạt động tốt trên dữ liệu có số chiều (đặc trưng) cao.
- Xử lý tốt với các biến không liên tục: Native Bayes có thể xử lý tốt với dữ liệu không liên tục, chẳng hạn như dữ liệu văn bản.

Nhược điểm của mô hình Native Bayes:

- Giả định ngây thơ (Native Assumption): Giả định rằng các đặc trưng là độc lập có thể không đúng trong mọi trường hợp, đặc biệt là khi có sự tương quan giữa các đặc trưng.
- Khả năng xử lý tốt khi các đặc trưng tương quan mạnh: Nếu có sự tương quan mạnh giữa các đặc trưng, Native Bayes có thể không hoạt động tốt.
- Yếu đối với dữ liệu không cân bằng: Nếu một lớp có số lượng mẫu lớn hơn rất nhiều so với lớp khác, mô hình có thể bị chệch và không làm tốt trên lớp thiểu số.
- Không thể ước lượng xác suất 0: Nếu một từ hoặc đặc trưng nào đó không xuất hiện trong tập huấn luyện, xác suất của nó sẽ bằng 0, và mô hình sẽ không thể đưa ra dự đoán.
- Khả năng xử lý tốt với dữ liệu liên tục hạn chế: Mô hình này thường không xử lý tốt với dữ liệu liên tục và phân phối không chuẩn.

CHƯƠNG 3

PHÂN TÍCH YÊU CẦU

3.1 Giới thiệu đề tài

3.1.1 *Tên đề tài*

- Ứng dụng giải thuật Native Bayes để dự đoán người bị tiểu đường.

3.1.2 *Yêu cầu đề tài*

- Phải thu thập dữ liệu phù hợp về các yếu tố ảnh hưởng đến bệnh tiểu đường, bao gồm dữ liệu về đường huyết, trọng lượng, tuổi, v.v.
- Phải xử lý và làm sạch dữ liệu để chuẩn bị cho việc huấn luyện mô hình Naive Bayes. Huấn luyện mô hình trên dữ liệu đã chuẩn bị.
- Đánh giá hiệu suất của mô hình sử dụng các phương pháp đánh giá hợp lý như cross-validation hoặc holdout validation.
- Triển khai mô hình để có thể dự đoán bệnh tiểu đường cho dữ liệu mới.

3.1.3 *Hiện trạng của bệnh tiểu đường trong đời sống ngày nay*

- Thực trạng bệnh tiểu đường hiện nay trên thế giới được Hiệp hội đái tháo đường thế giới (IDF) thống kê với con số hơn 425 triệu người, nghĩa là cứ 11 người thì có 1 người mắc bệnh tiểu đường. Trong đó, cứ 2 người mắc bệnh tiểu đường thì 1 người không biết mình bị bệnh (không đi kiểm tra chuẩn đoán bệnh tiểu đường). Việc điều trị muộn sẽ dẫn đến nhiều biến chứng nguy hiểm và ảnh hưởng trực tiếp tới chất lượng cuộc sống và sức khỏe của bệnh nhân.
- Theo dự đoán, số người mắc bệnh tiểu đường trên thế giới sẽ tăng lên 522 triệu người vào năm 2030, và con số này sẽ còn tăng nhanh hơn nữa nếu mọi người chủ quan đối với căn bệnh này.
- Một số con số ấn tượng khác về thực trạng bệnh tiểu đường hiện nay cũng đã được thống kê dưới đây:

- Trên thế giới, mỗi năm có khoảng 132.600 trẻ em được chẩn đoán mắc bệnh tiểu đường tuýp 1. Con số bệnh nhân tiểu đường tuýp 1 ở trẻ em từ (0-19 tuổi) là hơn 1 triệu người.
- Có hơn 21 triệu phụ nữ bị tăng đường huyết và có khả năng dung nạp đường kém trong thai kỳ, tương đương 1/6 đối tượng phụ nữ mang thai.
- 2/3 đối tượng mắc bệnh tiểu đường là người cao tuổi, tuy nhiên, số người trẻ tuổi bị tiểu đường ngày càng tăng.
- Cứ 6 giây có 1 người tử vong do các biến chứng tiểu đường
- Năm 2017 con số người chết do bệnh tiểu đường khoảng 4 triệu người.
- Thống kê chi phí điều trị bệnh tiểu đường năm 2017 trên toàn thế giới là 727 tỷ đô la. Nhiều hơn chi phí quốc phòng của cả Hoa Kỳ và Trung Quốc. Có thể thấy, bệnh tiểu đường đang trở thành gánh nặng trên toàn thế giới.
- Trong những năm gần đây, bệnh tiểu đường (hay còn gọi là bệnh đái tháo đường) đang là một trong những bệnh mạn tính phổ biến nhất với tần suất gia tăng nhanh chóng đã trở thành một vấn đề nghiêm trọng đối với thế giới nói chung và Việt Nam nói riêng. Bệnh tiểu đường là một trong các nguyên nhân gây tử vong hàng đầu, dẫn đến gánh nặng rất lớn cho nền y tế cũng như ảnh hưởng nghiêm trọng đối với sức khỏe cộng đồng, ấy vậy mà sự hiểu biết của mọi người về “kẻ giết người thầm lặng” vẫn còn rất mơ hồ.
- Theo các số liệu thống kê của hiệp hội phòng chống tiểu đường thế giới, cho biết ở năm 2021, tỉ lệ người bị tiểu đường là 20% ở các quốc gia và ngày tăng, nghĩa là cứ 10 người trưởng thành thì có hơn một người bị mắc bệnh tiểu đường. Theo thống kê từ năm 2000 cho tới nay cho thấy rằng tỉ lệ người từ 20 đến 79 tuổi bị bệnh tiểu đường đã tăng gấp 3 lần dẫn đến việc các chi phí y tế cho căn bệnh này cũng tăng lên gấp 3 lần trong vòng 15 năm trở lại đây.
- Tại Việt Nam hiện nay cũng đang gặp tình trạng tỉ lệ người bị bệnh tiểu đường đang tăng nhanh. Không những các vùng đô thị mà cả ở những khu vực miền núi, trung du cho đến đồng bằng đều có sự xuất hiện của căn bệnh này. Hiện tại, ở Việt Nam có khoảng 7 triệu người bị bệnh tiểu đường, Nghiêm trọng là

trong đó có hơn 55% bệnh nhân đã có biến chứng, trong đó 34% biến chứng về tim mạch, 39.5% bị về mắt và thần kinh còn lại thì bị biến chứng về thận. Điều này không những gây ra các gánh nặng gia tăng chi phí y tế mà nó còn làm giảm chất lượng cuộc sống.

3.1.4 *Hiểu rõ về mức độ nguy hiểm của bệnh tiểu đường*

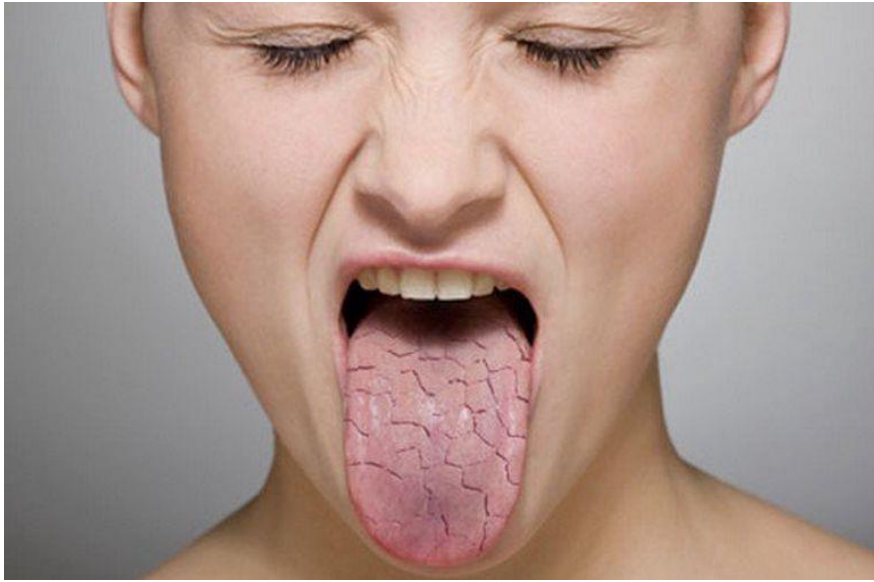
Như đã nói ở trên, bệnh tiểu đường (hay còn gọi là bệnh đái tháo đường) là một căn bệnh mạn tính gây ra tình trạng bệnh lý rối loạn chuyển hóa không đồng nhất, đặc điểm của bệnh này là làm tăng lượng đường có trong máu của cơ thể. Nguyên nhân thông thường là do Insulin không ổn định (có thể thiếu hoặc thừa). Dựa vào đặc điểm và các thay đổi của bệnh, có thể chia ra 4 loại bệnh đái tháo đường: Đái tháo đường typ1, đái tháo đường typ2, đái tháo đường tứ phát và đái tháo đường thai kỳ.

- **Đái tháo đường typ1:** chỉ trong vài ngày hoặc vài tuần là các triệu chứng của bệnh sẽ xuất hiện rất nhanh, và thường có 4 dấu hiệu điển hình:
 - **Đói và mệt:** Thông thường cơ thể sẽ chuyển đổi thức ăn thành glucose được các tế bào sử dụng để nạp năng lượng, nhưng chúng cần có insulin để hấp thụ glucose. Nếu cơ thể không tạo ra đủ insulin hoặc kháng lại insulin mà cơ thể tạo ra thì glucose không thể thâm nhập vào tế bào và cơ thể sẽ không có năng lượng dẫn đến cơ thể bị đói và mệt thường xuyên.



Hình 3- 1: Đói và mệt

- **Đi tiểu thường xuyên và khát hơn:** Khi bệnh tiểu đường đẩy lượng đường có trong máu lên cao, điều này dẫn đến việc thận sẽ bị quá tải và không thể đưa tất cả trở lại, làm cho cơ thể sản sinh nhiều nước tiểu và phải mất nước. Kết quả: bệnh nhân sẽ đi nhiều hơn và cũng vì đi nhiều nên sẽ khiến họ khát hơn. Cứ như vậy, họ càng uống nhiều thì họ cũng sẽ đi tiểu nhiều hơn.
- **Khô miệng, khát nước nhiều và ngứa da:** Vì cơ thể cần đã bài tiết hết chất lỏng làm cho độ ẩm của những thứ khác sẽ ít hơn, dẫn đến việc cơ thể bị mất nước và miệng thì luôn cảm thấy khô, da khô có thể gây ngứa.



Hình 3- 2: Khô miệng

- **Sụt cân nhiều:** Mặc dù ăn nhiều nhưng bệnh nhân vẫn bị sụt cân rất nhiều.



Hình 3- 3: Sụt cân nhiều

- **Thị lực giảm:** Việc mật độ chất lỏng trong cơ thể bị thay đổi liên tục làm cho tròng kính trong mắt sưng lên khiến mắt mờ và thị giác giảm.



Hình 3- 4: Thị lực giảm

- **Đái tháo đường typ2:** Khác với typ1, các triệu chứng ở typ2 diễn ra một cách thầm lặng hoặc thậm chí là không có triệu chứng gì. Bạn chỉ có thể được chuẩn đoán bệnh tiểu đường chỉ vì bạn đi khám bác sĩ về một căn bệnh khác mà cần phải làm xét nghiệm glucose máu hoặc phát hiện bệnh vì các triệu chứng khác như vết nhiễm trùng khó lành. Tóm lại, người bệnh có thể không nhận thức được căn bệnh một cách rõ ràng, nó có thể phát triển trong nhiều năm và các dấu hiệu thì rất khó chuẩn đoán. Một số dấu hiệu như:
 - **Nhiễm trùng nấm men:** Nấm men thường ăn glucose, vì vậy nếu có nhiều glucose ở xung quanh rất dễ làm cho nấm men phát triển nhanh. Các vết nhiễm trùng có thể phát triển ở bất kỳ nếp gấp ẩm và ẩm của da, bao gồm: giữa ngón tay và ngón chân, dưới ngực, trong hoặc xung quanh cơ quan sinh dục.



Hình 3- 5: Nhiễm trùng nấm men

- **Vết loét hoặc vết cắt chậm lành:** Lượng đường trong máu cao có thể ảnh hưởng đến lưu lượng máu, gây ra tổn thương thần kinh khiến cho cơ thể khó lành các vết thương. Có thể khiến bệnh nhân đau hoặc tê ở chân, đây cũng là một kết quả khác của tổn thương thần kinh.



Hình 3- 6: Vết thương lâu lành

- **Đái tháo đường thai kỳ:** Lượng đường trong máu cao khi mang thai thường không có triệu chứng. Bạn có thể cảm thấy hơi khát hơn bình thường hoặc phải đi tiểu thường xuyên hơn. Thường phát hiện chủ yếu khi làm nghiệm pháp 3 mẫu glucose lúc thai 28 tuần.



Hình 3- 7: Thai kỳ

3.2 Mô tả đề tài

- Tìm hiểu về bệnh tiểu đường: Nắm vững kiến thức cơ bản về bệnh tiểu đường, bao gồm nguyên nhân, loại bệnh, triệu chứng, biến chứng và cách điều trị. Tìm hiểu về các yếu tố nguy cơ, ví dụ: tiền sử gia đình, lối sống, chế độ ăn uống, hoạt động thể chất, v.v.
- Thu thập dữ liệu: Tìm kiếm và thu thập dữ liệu liên quan đến bệnh tiểu đường, bao gồm thông tin về bệnh nhân, kết quả xét nghiệm, thông tin về yếu tố nguy cơ và các biến số khác có thể liên quan đến bệnh.
- Tiền xử lý dữ liệu: Thực hiện các bước tiền xử lý dữ liệu như xóa bỏ dữ liệu thiếu, xử lý dữ liệu ngoại lệ, chuẩn hóa dữ liệu, v.v.
- Phân tích đặc trưng: Xác định các đặc trưng quan trọng trong dữ liệu để phân tích tác động của chúng đến bệnh tiểu đường. Có thể sử dụng các phương pháp như phân tích thành phần chính (PCA), phân tích biến thể (ANOVA), hoặc các phương pháp khác để tìm ra các đặc trưng quan trọng.

- Xây dựng mô hình dự đoán: Sử dụng các thuật toán học máy hoặc học sâu để xây dựng mô hình dự đoán bệnh tiểu đường. Các thuật toán phổ biến có thể bao gồm cây quyết định, hồi quy logistic, máy vector hỗ trợ (SVM), mạng nơ-ron, v.v.
- Đánh giá mô hình: Đánh giá hiệu suất của mô hình bằng cách sử dụng các độ đo như độ chính xác, độ nhạy, độ đặc hiệu, độ F1, v.v. Đồng thời, phân tích các yếu tố ảnh hưởng đến kết quả dự đoán của mô hình.
- Hiểu biết và kết luận: Dựa trên kết quả phân tích, đưa ra những hiểu biết và kết luận về yếu tố nguy cơ, đặc trưng quan trọng và mô hình dự đoán bệnh tiểu đường.
- Gợi ý giải pháp: Dựa trên những phân tích và kết luận, đề xuất các giải pháp để phòng ngừa, chẩn đoán sớm hoặc quản lý bệnh tiểu đường một cách hiệu quả.

3.3 Hướng xây dựng đề tài

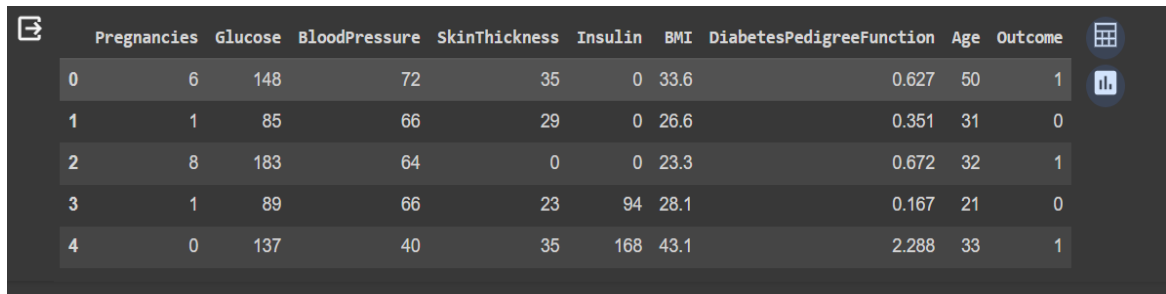
- Phân tích yếu tố nguy cơ: Nghiên cứu và phân tích các yếu tố nguy cơ gây ra bệnh tiểu đường, bao gồm tiền sử gia đình, tuổi tác, mức độ hoạt động thể chất, chế độ ăn uống, cân nặng, v.v. Điều này giúp hiểu rõ hơn về những yếu tố có thể làm tăng nguy cơ mắc bệnh tiểu đường và có thể sử dụng để xây dựng các biện pháp phòng ngừa.
- Dự đoán bệnh tiểu đường: Xây dựng một mô hình dự đoán bệnh tiểu đường dựa trên các yếu tố nguy cơ và thông tin bệnh nhân. Bạn có thể sử dụng các thuật toán học máy như hồi quy logistic, cây quyết định, máy vector hỗ trợ, hoặc mạng nơ-ron để phát triển mô hình dự đoán. Điều này có thể giúp xác định nguy cơ mắc bệnh tiểu đường cho một cá nhân dựa trên thông tin cá nhân của họ.
- Quản lý bệnh tiểu đường: Nghiên cứu các phương pháp quản lý bệnh tiểu đường, bao gồm chế độ ăn uống, hoạt động thể chất, kiểm soát đường huyết, quản lý dược phẩm, và các biện pháp chăm sóc sức khỏe khác. Tìm hiểu về những phương pháp hiệu quả để kiểm soát bệnh tiểu đường và cải thiện chất lượng cuộc sống của người bệnh.

- Tích hợp công nghệ: Nghiên cứu và phát triển các ứng dụng công nghệ thông tin để hỗ trợ người bị bệnh tiểu đường trong việc quản lý bệnh, theo dõi đường huyết, ghi chú chế độ ăn uống và hoạt động thể chất, và nhận thông tin hữu ích về bệnh tiểu đường.

CHƯƠNG 4

XÂY DỰNG MÔ HÌNH

4.1 Mô tả dữ liệu



	Pregnancies	Glucose	BloodPressure	SkinThickness	Insulin	BMI	DiabetesPedigreeFunction	Age	Outcome
0	6	148	72	35	0	33.6	0.627	50	1
1	1	85	66	29	0	26.6	0.351	31	0
2	8	183	64	0	0	23.3	0.672	32	1
3	1	89	66	23	94	28.1	0.167	21	0
4	0	137	40	35	168	43.1	2.288	33	1

Hình 4- 1: Mô tả dữ liệu

- Dữ liệu trong tệp bao gồm 8 đặc trưng chính hay còn được gọi là các features và biến nhãn được biểu diễn dưới dạng số 1 và 0 tượng trưng cho có mắc bệnh và không mắc bệnh. Dưới đây là mô tả về từng cột dữ liệu:

4.2 Đọc và chuẩn bị dữ liệu

```
[3] import pandas as pd
    from sklearn.model_selection import train_test_split
    from sklearn.naive_bayes import GaussianNB
    import matplotlib.pyplot as plt

    # Đọc dữ liệu từ tệp CSV
    data = pd.read_csv('diabetes.csv')
    data.head()
    data.shape

    (768, 9)
```

Hình 4- 2: Đọc và chuẩn bị dữ liệu 1

- Dữ liệu được đọc từ tệp 'diabetes.csv' bằng cách sử dụng thư viện Pandas và được lưu trong biến data. Đây là bước đầu tiên để nắm bắt thông tin trong tập dữ liệu.

```
# Tạo các đặc trưng
features = ['Pregnancies', 'Glucose', 'BloodPressure', 'SkinThickness', 'Insulin', 'BMI', 'Age']

x = data[features]
y = data["Outcome"]
```

Hình 4- 3: Đọc và chuẩn bị dữ liệu 2

- Tập dữ liệu được chia thành hai phần: X chứa các đặc trưng (features) và y chứa nhãn (labels). Điều này chuẩn bị dữ liệu cho việc huấn luyện và kiểm tra mô hình.

	Pregnancies	Glucose	BloodPressure	SkinThickness	Insulin	BMI	Age
0	6	148	72	35	0	33.6	50
1	1	85	66	29	0	26.6	31
2	8	183	64	0	0	23.3	32
3	1	89	66	23	94	28.1	21
4	0	137	40	35	168	43.1	33

```
y.head()
```

```
0    1
1    0
2    1
3    0
4    1
Name: Outcome, dtype: int64
```

Hình 4- 4: Đọc và chuẩn bị dữ liệu 3

4.3 Chia dữ liệu thành tập huấn luyện và tập kiểm tra

```
# Chia dữ liệu
from sklearn.model_selection import train_test_split

X_train,X_test,y_train,y_test = train_test_split(X,y,random_state=0,train_size=0.8,test_size=0.2)
```

Hình 4- 5: Tập huấn luyện và kiểm tra

- Dữ liệu thường được chia thành hai phần: tập huấn luyện (X_train, y_train) và tập kiểm tra (X_test, y_test). Việc này cho phép bạn huấn luyện mô hình trên một phần của dữ liệu và sau đó kiểm tra hiệu suất của nó trên một tập dữ liệu độc lập.
- Khi đặt train_size=0.8 có nghĩa là 80% dữ liệu sẽ được chọn làm tập huấn luyện còn 20% dữ liệu còn lại sẽ được đặt làm tập kiểm tra.

4.4 Xây dựng và huấn luyện mô hình Native Bayes

```
# Mô hình Bayes
model = GaussianNB()

model.fit(X_train, y_train)

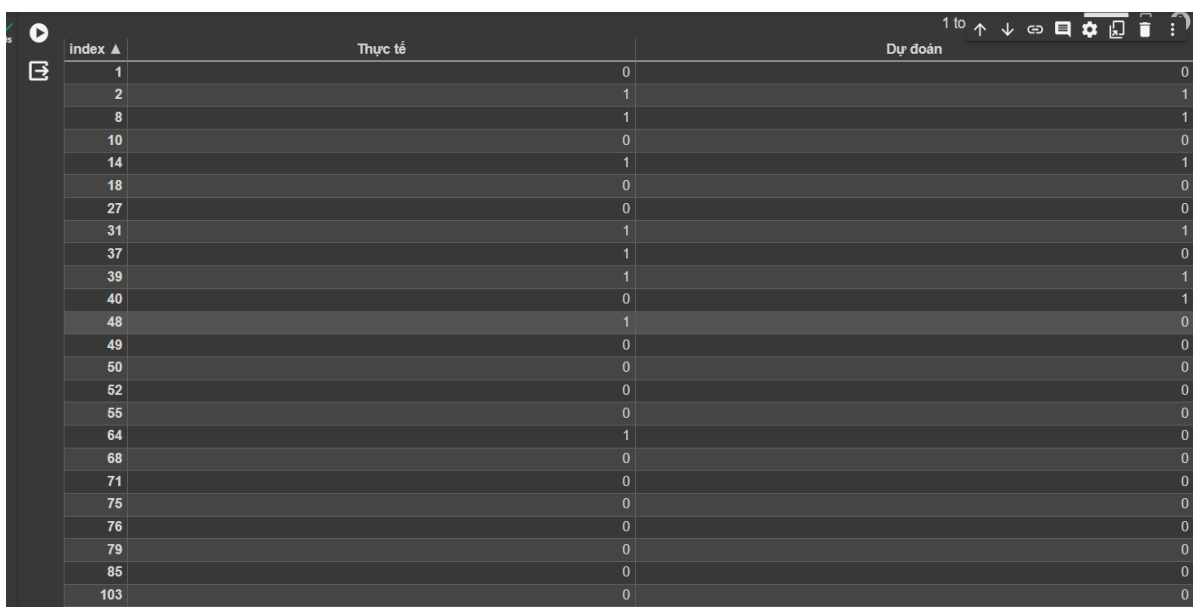
y_pred = model.predict(X_test)

pd.DataFrame({'Thực tế': y_test, 'Dự đoán': y_pred })
```

Hình 4- 6: Xây dựng và huấn luyện mô hình

- Để xây dựng một mô hình Naive Bayes với phân phối Gaussian (hay còn gọi là Gaussian Naive Bayes) trong Python sử dụng thư viện scikit-learn (sklearn), chúng ta sử dụng lớp GaussianNB từ mô-đun sklearn.naive_bayes. Mô hình này được dùng để học cách dự đoán nhãn dựa trên các đặc trưng của mẫu dữ liệu.
- Dữ liệu huấn luyện (X_train, y_train) được sử dụng để đào tạo mô hình Naive Bayes bằng cách gọi phương thức .fit(X_train, y_train) trên đối tượng mô hình. Trong quá trình này, mô hình học cách phân loại các mẫu dữ liệu vào các nhãn dự đoán dựa trên các đặc trưng trong tập dữ liệu huấn luyện (X_train) và các nhãn tương ứng (y_train). Điều này bao gồm việc học các phân phối xác suất của các đặc trưng cho từng nhãn và tính toán xác suất của mỗi nhãn dựa trên các đặc trưng của mẫu dữ liệu mới.

4.5 Dự đoán và đánh giá hiệu suất



	Thực tế	Dự đoán
1	0	0
2	1	1
8	1	1
10	0	0
14	1	1
18	0	0
27	0	0
31	1	1
37	1	0
39	1	1
40	0	1
48	1	0
49	0	0
50	0	0
52	0	0
55	0	0
64	1	0
68	0	0
71	0	0
75	0	0
76	0	0
79	0	0
85	0	0
103	0	0

Hình 4- 7: Dự đoán và đánh giá hiệu suất

- Mô hình được sử dụng để dự đoán nhãn trên tập kiểm tra (X_{test}) bằng cách gọi `model.predict(X_test)`. Kết quả dự đoán được lưu trong biến `y_pred`.
- Để biểu diễn kết quả dự đoán một cách trực quan ta sử dụng thư viện Pandas và nhận được kết quả như hình trên.

```
print("Độ chính xác của mô hình: {:.2f}%".format(accuracy_score(y_test, y_pred) * 100))
```

Độ chính xác của mô hình: 78.57%

Hình 4- 8: Kết quả độ chính xác mô hình

- Độ Chính Xác: Độ chính xác (`accuracy_score`) được tính để biết tỷ lệ mẫu dự đoán đúng trên tổng số mẫu trong tập kiểm tra.
- Để đo độ chính xác của mô hình Bayes trong python ta sử dụng hàm `accuracy_score` trong thư viện `sklearn.metrics`.

4.6 Trực quan hóa kết quả

```
plt.figure(figsize=(10, 6))
plt.scatter(range(len(y_test)), y_test, label='Thực tế', marker='o', color='b', alpha=0.5)
plt.scatter(range(len(y_pred)), y_pred, label='Dự đoán', marker='x', color='r', alpha=0.5)
plt.xlabel('Số mẫu')
plt.ylabel('Nhãn')
plt.legend()
plt.title('Kết quả Dự đoán và Thực tế')
plt.grid(True, linestyle='--', alpha=0.6)
plt.xticks(range(len(y_test)))
plt.show()
```

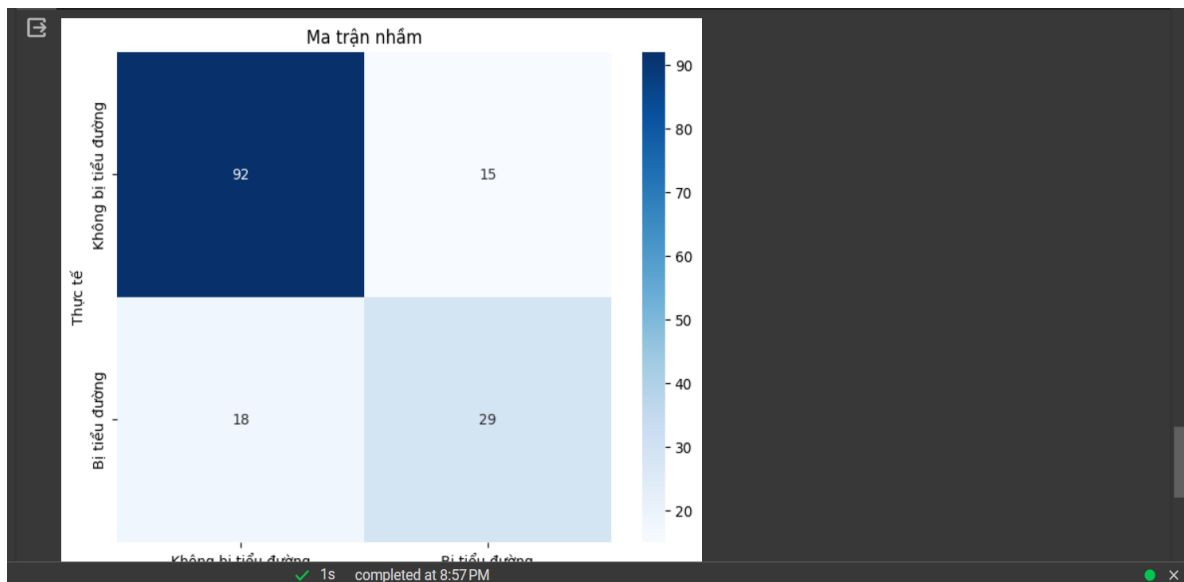
Hình 4- 9: Trực quan hóa kết quả

- Kết quả dự đoán (`y_pred`) và nhãn thực tế (`y_test`) được trực quan hóa bằng biểu đồ scatter plot. Biểu đồ này cho phép so sánh giữa các điểm dự đoán và điểm thực tế trên một đồ thị.

CHƯƠNG 5

THỰC NGHIỆM MÔ HÌNH

5.1. Ma Trận Nhầm Lẫn (Confusion Matrix):



Hình 5- 1: Ma trận nhầm lẫn (Confusion Matrix)

- Ma trận nhầm lẫn hiển thị số lượng dự đoán đúng và sai cho từng lớp nhãn (0 và 1). Trong trường hợp này, có thể thấy rằng có một số lượng nhỏ các dự đoán sai (False Positives và False Negatives), nhưng tỷ lệ này không lớn.

5.2. Độ Chính Xác (Accuracy):



Hình 5- 2: Độ chính xác

- Độ chính xác là khoảng 78.5%, có nghĩa là mô hình dự đoán đúng khoảng 78.5% tổng số mẫu trên tập kiểm tra. Điều này chỉ ra rằng mô hình có khả năng dự đoán tương đối tốt trên tập kiểm tra, nhưng nó còn có thể được cải thiện.

CHƯƠNG 6

KẾT LUẬN VÀ HƯỚNG PHÁT TRIỂN

6.1 Kết quả đạt được

- Xây dựng Mô hình Dự đoán Tiểu đường: Chúng ta đã thành công trong việc xây dựng một mô hình học máy bằng giải thuật Naive Bayes để dự đoán tiểu đường dựa trên dữ liệu đầu vào.
- Đánh Giá Hiệu Suất: Mô hình đã được đánh giá thông qua báo cáo phân loại và ma trận nhầm lẫn, cho thấy khả năng dự đoán tương đối tốt với độ chính xác khoảng 75%. Điều này có thể hữu ích trong việc hỗ trợ việc chẩn đoán tiểu đường.
- Phân Tích Kết Quả: Chúng ta đã phân tích kết quả dự đoán để hiểu hiệu suất của mô hình trên từng lớp nhãn và trực quan hóa kết quả để dễ dàng thấy sự trùng khớp giữa dự đoán và thực tế.

6.2 Hạn chế của đề tài

- Dữ liệu Hạn Chế: Một trong những hạn chế lớn của đề tài là dữ liệu. Dữ liệu tiểu đường có thể không đủ đại diện và có thể cần thêm dữ liệu mới và đa dạng để cải thiện hiệu suất mô hình.
- Tính Cân Bằng Dữ Liệu: Tập dữ liệu có sự không cân bằng giữa các lớp nhãn (tiểu đường và không tiểu đường). Việc xử lý sự không cân bằng này có thể cải thiện hiệu suất mô hình.
- Tối Ưu Hóa Mô Hình: Mô hình Naive Bayes có thể được tối ưu hóa thêm để đạt được hiệu suất cao hơn. Các biến thể khác của Naive Bayes và siêu tham số cần được kiểm tra.

6.3 Hướng phát triển

- Thu Thập Thêm Dữ Liệu: Thu thập thêm dữ liệu từ nhiều nguồn khác nhau để cải thiện độ chính xác và đa dạng của mô hình.
- Tối Ưu Hóa Mô Hình: Tiến hành tối ưu hóa mô hình bằng cách sử dụng các phương pháp tinh chỉnh siêu tham số và thử nghiệm với các biến thể của Naive Bayes.
- Xử Lý Dữ Liệu Không Cân Bằng: Áp dụng các biện pháp xử lý dữ liệu không cân bằng như oversampling hoặc undersampling để cân bằng tập dữ liệu.
- Tích Hợp Vào Ứng Dụng Thực Tế: Phát triển ứng dụng thực tế hoặc hệ thống hỗ trợ quyết định trong lĩnh vực chăm sóc sức khỏe để giúp các chuyên gia y tế trong việc dự đoán tiểu đường.
- Nghiên Cứu Thêm: Tiến hành nghiên cứu tiếp về việc sử dụng học máy trong lĩnh vực y tế, và tìm hiểu các giải thuật và phương pháp mới để cải thiện dự đoán và chẩn đoán bệnh tiểu đường.

TÀI LIỆU THAM KHẢO

1. Bạch Hưng Khang, Hoàng Kiếm Trí tuệ nhân tạo: Các phương pháp và ứng dụng. Nhà xuất bản Khoa học và Kỹ thuật, 1989.
2. Đinh Mạnh Tường Giáo trình Trí tuệ nhân tạo, Đại học Quốc gia Hà nội.
3. Nguyễn Thanh Thủy Trí tuệ nhân tạo: Các phương pháp giải quyết vấn đề và kỹ thuật xử lý tri thức. Nhà xuất bản Giáo dục, 1996.
4. N. Nilson Artificial Intelligence. Ed. McGrawhill, 1971
5. Patrick Henry Winston Artificial Intelligence. Ed. Addison Wesley, 1992.