BÁO CÁO GIAI ĐOẠN 1

*Thời gian dự kiến 09/09/2012 – 09/10/2012*

**Nội dung**

[1 ỨNG DỤNG CỦA KHAI PHÁ DỮ LIỆU VÀO Y HỌC: 3](#_Toc337582602)

[1.1 Mục tiêu: 3](#_Toc337582603)

[1.1.1 Các nghiên cứu: 3](#_Toc337582604)

[1.2 Khó khăn và thử thách: 7](#_Toc337582605)

[2 BỆNH TIỂU ĐƯỜNG: 8](#_Toc337582606)

[2.1 Bệnh tiểu đường là gì? 8](#_Toc337582607)

[2.2 Các biến chứng có liên quan đến bệnh tiểu đường: 8](#_Toc337582608)

[2.3 Các loại bệnh tiểu đường: 9](#_Toc337582609)

[2.3.1 Bệnh tiểu đường dạng 1: 9](#_Toc337582610)

[2.3.2 Bệnh tiểu đường dạng 2: 9](#_Toc337582611)

[2.4 Tình hình chung trên thế giới: 10](#_Toc337582612)

[2.5 Tình hình tại Việt Nam: 11](#_Toc337582613)

[2.6 Vì sao phải áp dụng khai phá dữ liệu vào chuẩn đoán bệnh tiểu đường? 11](#_Toc337582614)

[3 ỨNG DỤNG DATAMINING VÀO HỆ HỖ TRỢ RA QUYẾT ĐỊNH Y HỌC: 12](#_Toc337582615)

[3.1 Một vài phương pháp ứng dụng data mining 13](#_Toc337582616)

[3.1.1 Học có giám sát 13](#_Toc337582617)

[3.1.2 Học không giám sát 14](#_Toc337582618)

# ỨNG DỤNG CỦA KHAI PHÁ DỮ LIỆU VÀO Y HỌC:

Với những thành công của việc kết hợp khai phá dữ liệu vào xây dựng các ứng dụng trong thương mại điện tử, marketing và bán lẻ đã cho thấy một tiềm năng lớn của việc trích xuất tri thức từ cơ sở dữ liệu ( Knowledge Discovery in Database – KDD) trong tất cả các lĩnh vực hiện nay.

Trong chương này, nhóm nghiên cứu sẽ nêu ra những kỹ thuật KDD hiện tại, các ứng dụng được tích hợp khai phá dữ liệu trong y tế và sức khỏe cộng đồng, các vấn đề quan trọng và thử thách trong việc kết hợp khai phá dữ liệu với y học. Ngoài ra còn nêu lên sự phát triển của Khai phá dữ liệu tại Việt Nam.

Phương pháp luận chính là dựa trên những bài báo đã được xuất bản từ năm 2000 trở lại đây trong nhiều lĩnh vực như y khoa, khoa học máy tính, kỹ thuật máy tính.

## Mục tiêu:

* Liệt kệ các lợi ích và tầm quan trọng của khai phá dữ liệu trong y học.
* Tìm kiếm các kỹ thuật khai phá dữ liệu đã được sử dụng trong những lĩnh vực khác nhưng có thể áp dụng vào y học.
* Nếu lên sự phát triển của một lượng lớn các ứng dụng được kết hợp với khai phá dữ liệu.
* Xác định các vấn đề và thử thách trong ứng dụng khai phá dữ liệu vào y học.
* Đưa ra những khuyến nghị cho việc thu thập trị thức từ cơ sở dữ liệu thông qua khai phá dữ liệu.

### Các nghiên cứu:

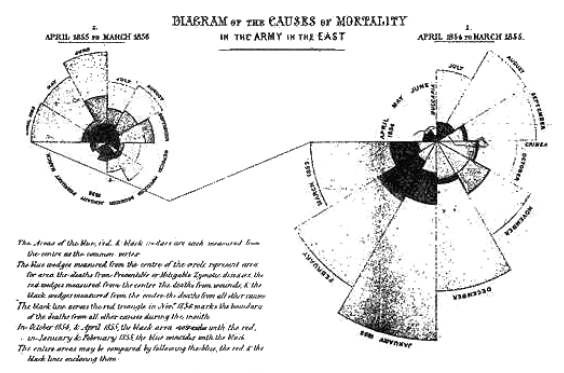
#### Khai pha dữ liệu trong y học:

* Có thể nói, việc sử dụng dữ liệu thật để hỗ trợ cho việc ra quyết định trong y học ( Evidence Bases Medicine – EBM) đã tồn tại trong nhiều thế kỷ.
* Năm 1854, John Snow, cha để của dịch tễ học hiện đại đã sử dụng các bản đồ đồ thị để phát hiện ra mầm bệnh thổ tả và đã chứng minh rằng bệnh này lây lan qua hệ thống bơm nước. Snow đã đếm số lượng bệnh nhân và vẽ sơ đồ địa chỉ của các bệnh nhân trên bản đồ bằng các thanh mà đen. Lúc đó ông đã phát hiện rằng hầu hết các ca tử vong đều nằm xung quanh một điểm bơm nước cố định tại London.



*Bản đồ mà John Snow dùng để phát hiện mầm bệnh*

* Năm 1855, Florence Nightingale đã phát minh ra biển đồ phân cực để tìm ra nhiều ca tử vong của binh lính do mất vệ sinh trong khám lâm sàn và có thể ngăn ngừa được. Bà đã sử dụng biểu đồ này để thuyết phục chính phủ cải cách các chính sách và việc này đã làm giảm đi đáng kể số lượng các ca tử vong.



*Biểu đồ phân cực do Nightinale phát minh*

* Snow và Nightingale đã tự mình thực hiện đầy đủ các bước: thu thập dữ liệu, sàng lọc và phân tích thông qua các dữ liệu về tỉ lệ tử vong trong suốt thời gian nghiên cứu vì số lượng dữ liệu có thể quản lý được. Nhưng hiện nay, việc bùng nổ dân số, sự toàn cầu hóa và tốc độ phát bệnh của bệnh dịch làm cho việc thao tác dữ liệu gần như không thể thực hiện.
* Đó cũng là nguyên nhân tại sao khai phá dữ liệu trở nên hữu ích trong y học. Tuy còn phát triển khá châm những vẫn được áp dụng để giải quyết các vấn đề khác nhau trong việc trích xuất tri thức. Có thể nói việc ứng dụng khai phá dữ liệu trong y học là một lĩnh vực còn rất non trẻ.
* Năm 2003, Wilson đã đưa ra một kết luận hết sức bất ngờ sau khi đã thực hiện một nghiên cứu trong các trường hợp mà KDD và các kỹ thuật khai phá dữ liệu đã được áp dụng trong y học. Cho đến thời điểm này, vẫn có nhiều tác giả cho rằng khai phá dữ liệu chỉ là một quá trình thu thập tri trức, còn một số khác thì cho rằng khai phá dữ liệu sử dụng kỹ thuật thống kê quá trình thu thập tri thức.
* Những khái niệm sai lầm này vẫn tồn tại trong cộng động y khoa. Trong đó có một định nghĩa tạm chấp nhận về khai phá dữ liệu ngày nay đó là một quy trình và kỹ thuật trong khai phá mô hình và hướng của dữ liệu (Witten and Frank 2005).

#### Tầm quan trọng và lợi ích của khai phá dữ liệu:

* Tuy có nhiều sự khác biệt trong hướng tiếp cận nhưng lĩnh vực y học vẫn rất cần đề khai phá dữ liệu. Không chỉ trong y tế cộng đồng mà còn có y tế tư nhân.
* *Tràn dữ liệu*: Tin học hóa dữ liệu y tế đã thu được nhiều tri có giá trị nhưng hiện nay có một lượng lớn đỡ liệu làm cho việc khai phá trở nên khó khăn và gần như không thể nếu phải sàng lọc và khai phá tri thức bằng tay. Trong thực tế, một số chuyên gia cho rằng những đột phá trong lĩnh vực y học đang có dấu hiệu chậm lại do độ phức tạp của dữ liệu ngày càng tăng. Nhưng cũng chính điều này đã làm cho khai phá dữ liệu trở thành một ứng cử viên sáng giá.
* *Ngăn chặn các sai sót trong bệnh viện*: Khi các tổ chức y tế ứng dụng khai pha dữ liệu, họ có thể tìm được những tri thức và có khả năng cứu sống nhiều bệnh nhân. Một nghiên cứu đang được thực hiện trong các bệnh viện tại Mỹ cho thấy rằng 87% các ca tử vong có thể được phát hiện và ngăn chặn nếu tất cả các nhân viên ( kể cả bác sĩ) cẩn thận hơn. Thông qua việc khi phá dữ liệu, các nhà quản lý bệnh viện và chính phủ có thể tìm và giải quyết các vấn đề an toàn một cách dễ dàng hơn.
* *Hoạch định chính sách trong y tế cộng đồng*: Lavrac đã kết hợp GIS và khai dữ liệu, Weka để phân tích điểm giống nhau giữa các trung tâm y tế tại Slovenia. Sử dụng khai phá dữ liệu để tìm ra các mô hình trong trung tâm y tế nhằm đưa ra các chính sách khuyến nghị đến Viện Y Tế Cộng Đồng. Tháng 10 năm 2006 đã xảy ra một sự cố đáng tiếc tại trung tâm y tế Rizal thuộc thành phố Pasig của Indonexia. Bệnh viện đã thất bại trong việc thực hiện vệ sinh nghiêm ngặt và các biện pháp khử trùng đã dẫn đến các ca tử vong của trẻ sơ sinh do bị nhiểm khuẩn. Các bác sĩ chỉ biết được chính xác nguyên nhân cho đến khi các ca tử vong trở nên thường xuyên. Sau khi kiểm tra dữ liệu của bệnh viện, Sở Y Tế đã tìm thấy 12 trên 28 trẻ được sinh vào ngày 4 tháng 10 đã tử vong vì nhiểm khuẩn. Với khai phá dữ liệu, Sở Y Tế đã có thể phát hiện các biến cố không bình thường và hạn chế các diễn biến trở nên tệ hơn
* *Tiết kiệm tiền và chi phí*: Các tổ chức đã dùng khai phá dữ liệu nhằm rút trích tri thức từ cơ sở dữ liệu với chi phí thấp nhất. Thậm chí, có nhiều tổ chức đã áp dụng KDD và khai phá dữ liệu để tìm ra các gian lận trong thẻ tín dụng và bảo hiểm. Một hệ thống bảo hiểm thuộc quốc gia Philippin đã áp dụng khai phá dữ liệu vào việc tìm ra các bệnh nhân đặc biệt trong bảo hiểm y tế.
* *Phát hiện và ngăn chặn các bệnh nguy hiểm*: Cheng đã áp dụng thuật toán phân lớp dữ liệu để phát hiện ra các trường hợp bệnh nhân có khả năng mắc bệnh tim. Năm 2008, Cao đã sử dụng khai phá dữ liệu như một công cụ để kiểm soát các thí nghiệm vacxin lâm sàn. Sử dụng khai phá dữ liệu đã giúp cho y học có thể phát hiện ra nhiều bệnh nhân có biểu hiện khác thường hơn là nhìn vào từng tập dữ liệu riêng biệt.
* *Phát hiện, quản lý dịch bệnh và đưa ra các chính sách trong y tế cộng đồng*: Trong những năm gần đây, các chuyên gia y tế đã bắt đầu tìm cách ứng dụng khai phá dữ liệu vào việc phát hiện và quản lý dịch bệnh. Năm 2006, Kellogg đã đưa ra một kỹ thuật kế hợp mô hình không gian, mô phỏng và khai phá dữ liệu không gian để tìm ra các điểm bùng phát dịch bệnh. Kết quả phân tích được dùng để đưa ra nhằm quản lý dịch bệnh bùng phát. Năm 2005, Wong đã giới thiệu một thuật toán phát hiện dịch biện khi vừa ở giai đoạn đầu là WSARE (What’s Strange About Recent Events). Thuật toán được tạo nên dựa trên các luật kết hợp và mạng Bayes.
* *Chuẩn đoán không làm tổn thương bệnh nhân và hệ thống ra quyết định*: Một số chuẩn đoán và thí nghiệm có thể làm tổn thương bệnh nhân như sinh thiết ở phụ nữ nhằm phát hiện bệnh ung thư cổ tử cung. Năm 2006, Thangavel đã sử dụng thuật toán gom cụm K-means để phân tích các bệnh nhân ung thư cổ tử cung và đã cho thấy rằng việc gom cụm dữ liệu có thể đưa ra những kết quả chuẩn đoán tốt hơn là các phương pháp hiện có. Ông cũng đã tìm thấy một số dữ liệu có thể giúp các bác sĩ trong việc ra quyết định đưa bệnh nhân đi thiến hành thiết sinh hay không.
* *Tác hại tiềm ẩn của các loại thuốc*: Một số loại thuốc và hóa chất đã được đánh giá rằng sẽ không gây hại cho con người nhưng sau một thời gian dài sử dụng thì câu trả lời lại là có. Năm 2003, Wilson đã tiết lộ rằng tổ chức US Food và Drug Administration đã sử dụng khai phá dữ liệu để tìm ra các loại thuốc có hại khi sử dụng lâu dài trong cơ sở dữ liệu của họ. Thuật toán được áp dụng có tên là MGPS (Multi-item Gamma Poisson Shrinker) đã tìm 67% thuốc có hại và nếu dùng cách bình thường thì phải mất thêm 5 năm nữa.
* Ngoài những ứng dụng ở trên, khai phá dữ liệu vẫn còn nhiều cách khác nhau để áp dụng trong lĩnh vực y tế. Những tri thức có ích vẫn còn nằm đâu đó trong cơ sở dữ liệu và đang chờ được khai phá.

## Khó khăn và thử thách:

* Việc áp dụng khai phá dữ liệu vào lĩnh vực y học hiện đang gặp rất nhiều khó khăn do chính đặc trưng riêng của lĩnh vực này tạo nên. Shillabeer và Roddick (2007) đã đưa ra một số mâu thuẫn giữa các phương pháp khai phá dữ liệu truyền thống và y học.
* Trong khi các phương pháp khai phá dữ liệu hiện tại bắt đầu với một giả thiết và sau đó đưa ra một kết quả đã được chỉnh sửa phù hợp với giả thiết thì phương pháp khai phá dữ liệu trước đây bắt đầu với những bộ dữ liệu. Ngoài ra, khai phá dữ liệu truyền thống quan tâm đến các mô hình và hướng của các bộ dữ liệu thì khai phá dữ liệu trong y học lại quan tâm những gì nhỏ nhất không phù hợp với các mô hình. Khác biệt lớn nhất đó là cách tiếp cận của khai phá dữ liệu chuẩn chỉ quan tâm đến việc mô tả chứ không phải giải thích mô hình và hướng của dữ liệu. Ngược lại, y học lại cần những giải thích này vì một chi tiết dù là nhỏ nhất vẫn có thể cứu được nhiều bệnh nhân. Ví dụ, bệnh than và bệnh cúm có cùng những triệu chứng về đường hô hấp. Nếu bỏ đi một số đặc điểm nhỏ trong khai phá dữ liệu thì một dịch bệnh than sẽ hoàn toàn giống một dịch bệnh cúm và sẽ gây ra nhiều hậu quả đáng tiếc.
* Cho dù kết quả thu được từ khai phá dữ liệu có đáng tin cậy nhưng việc thay đổi thói quen của các bác sĩ, y tá không phải là một điều dễ dàng. Ayres (2008) đã cho biết đã có một số trường hợp khi các bác sĩ từ chối thay đổi các chính sách của bệnh viện. Một trường hợp đã tiếc đã xảy ra khi một bác sĩ quên rửa tay sau khi khám nghiệm tử thi và đã gây ra nhiều ca tử vong cho nhiều bệnh nhân được khám ngay sau đó. Chỉ khi có những điều đáng tiếc xảy ra thì họ mới chịu thay đổi những thói quen của mình. Shilabeer (2009) cũng cho biết hầu hết các bác sĩ (ở Úc) thường tiếp thu các lời khuyên từ các chuyên gia y tế hơn là xem các kết quả thu được từ khai phá dữ liệu.
* Ngoài ra, dữ liệu riêng tư của bệnh nhân cũng là một vật cản lớn trong việc ứng dụng khai phá dữ liệu vào y học. Vì để đưa ra một kết quả chính xác nhất thì cần đến một lượng lớn những dữ liệu cần thiết và cũng chính nó mới có thể giúp con người ta tránh được những bệnh chết người.

# BỆNH TIỂU ĐƯỜNG:

## Bệnh tiểu đường là gì?

* Bệnh tiểu thường (còn đường gọi là bệnh đái tháo đường) là một nhóm bệnh rối loạn chuyển hóa carbohydrates khi hooc môn insulin của tụy bị thiếu hoặc giảm tác động trong cơ thể. Sự sản sinh insulin sẽ được điều chỉnh bởi lượng glucose trong máu. Nó có chức năng chuyển hóa glucose (từ carbohydrates) trong tế bào để cung cấp cho quá trình trao đổi chất và tạo thành năng lượng cho cơ thể. Sự thiếu hụt insulin hoặc không sử dụng được insulin sẽ làm giảm khả năng hấp thụ glucose và vì thế glucose sẽ tích tụ trong gan và các tế bào chất béo dẫn đến việc tăng mức đường huyết và trong nước tiểu.
* Bệnh tiểu đường thường gây nguy hiểm cho người già và những người béo phì nhiều nhất. Đồng thời những nhân tố bẩm sinh gen di truyền, chế độ dinh dưỡng không tốt, bị stress, ít vận động và thừa cân là những nguy cơ quan trọng nhất dẫn điện việc mắc bệnh tiểu đường.

## Các biến chứng có liên quan đến bệnh tiểu đường:

* Bệnh tiểu đường là một trong những nguyên nhân chính dẫn đến các bệnh hiểm nghèo như bệnh tim, tai biến, suy thận, mù mắt, hoại thư…
* Những người mắc bệnh tiểu đường có thể phát triển các tổn thương thần kinh trên cơ thể như cảm giác đau, nhói hoặc là cảm giác tê bì (mất cảm giác) ở bàn tay, cánh tay và chân nhưng một số khác lại không. Các vấn đề thần kinh có thể xuất hiện ở mọi cơ quan, bao gồm hệ tiêu hóa, tim và cơ quan sinh dục. Khoảng 60 đến 70 phần trăm những người mắc bệnh tiểu đường đều có một vài biến chứng thần kinh. Những bệnh nhân này có thể gặp các vấn đề về thần kinh bất cứ lúc nào nhưng nguy cơ tăng cao cùng với lứa tuổi và thời gian mắc bệnh tiểu đường. Tỷ lệ có những biến chứng thần kinh cao nhất ở những người đã mắc bệnh tiểu đường ít nhất là 25 năm. Các biến chứng thần kinh cũng có thể xuất hiện ngày càng phổ biến hơn những người gặp vấn đề trong việc kiểm soát mức đường huyết cũng như những người có lượng chất béo trong máu cao, những người bị cao huyết áp và những người bị béo phì.
* Bệnh nhân tiểu đường với lượng đường huyết cao trong máu gây tổn thương tế bào vi mạch thận, làm giảm chức năng lọc, bài tiết nước tiểu của thận. Bệnh nặng dẫn đến suy thận và hủy hoại chức năng của thận, dẫn đến việc đi tiểu với lượng đường cao trong nước tiểu thường thấy ở bệnh nhân tiểu đường.
* Với lượng đường huyết cao trong mạch máu, khiến cho những mạch máu nhỏ tại võng mạc bị nghẽn, có thể bị vỡ gây tấy đỏ, sưng ứ gây ra tổn thương mắt và các bệnh võng mạc. Ngoài ra, các biến chứng của bệnh tiểu đường còn gây đục thủy tinh thể, tăng nhãn áp, gây mù lòa.
* Bệnh tiểu đường còn gây ra các biến chứng nguy hiểm về các mạch máu và tim. Khi các dấu hiệu tổn thương mạch máu, tim ngày càng nặng thì bệnh nhân rất dễ bị cao huyết áp, xơ cứng động mạch, nhồi máu cơ tim, tai biến mạch máu não gây bại liệt hoặc tử vong.

## Các loại bệnh tiểu đường:

### Bệnh tiểu đường dạng 1:

* Trong trường hợp này, tuyết tụy của bệnh nhân hầu như hoặc không có khả năng sinh ra insulin. Nguyên nhân là do hệ miễn dịch tử hủy hoại các tế bào beta trong tuyến tụy có nhiệm vụ sản sinh ra insulin. Chí có khoảng 5-10% tổng số bệnh nhân mắc bệnh tiểu đường dạng 1, phần lớn xảy ra ở trẻ em và ngươi trẻ tuổi (dưới 20 tuổi). Các triệu chứng thường khởi phát đột ngột và tiến triển nhanh nếu không điều trị. Giai đoạn toàn phát thì tình trạng thiếu insulin tuyệt đối gây tăng đường huyết và nhiễm Ceton.

### Bệnh tiểu đường dạng 2:

* Với những người mắc bệnh tiểu đường dạng 2, lượng insulin đặc sản sinh ra ban đầu hoàn toàn bình thường, nhưng các tế bào đã không hoặc kém nhậy cảm với sự có mặt của insulin. Đó là hiện tượng nhờn insulin (kháng insulin). Lượng đường trong máu do không được chuyển hóa thành năng lượng nên giữ ở mức cao, cơ thể bệnh nhân phản ứng bằng cách tăng sản xuất insulin lên, gây nên quá tải cho tuyến tụy và lượng insulin được tiết ra giảm dần dần.
* Bệnh tiểu đường dạng hai có nguyên nhân tiềm ẩn trong cấu tạo gen, nó làm cho bệnh phát triển nhanh. Nếu những người mang trong mình gen tạo mầm mống cho bệnh tiểu đường sớm biết được điều dó và có biện pháp phòng ngừa bằng cách sống và ăn uống tốt thì bệnh không nhất thiết xuất hiện và phát triển. Bệnh tiểu đường trong trường hợp này sẽ giữ ở dạng tiềm ẩn. Nhưng ngược lại, với cách sống không khoa học, căn bệnh sẽ phát triển rất nhanh.
* Số bệnh nhân mắc bệnh tiểu đường dạng 2 chiếm khoảng 90-95% trong tổng số bệnh nhân, thường gặp ở lứa tuổi trên 40 nhưng gần đây xuất hiện ngày càng nhiều ở lứa tuổi 30, thậm chí cả lứa tuổi thanh thiếu niên. Bệnh nhân thường ít có triệu chứng và chỉ thường được phát hiện bởi các triệu chứng của biến chứng hoặc chỉ được phát hiện tình cờ khi đi xét nghiệm máu trước khi mổ hoặc khi có biến chứng như nhồi máu cơ tim, tai biến…

## Tình hình chung trên thế giới:

* Bệnh tiểu đường trong thời đại hiện nay là một trong những căn bệnh phổ biến nhất – căn bệnh của thế kỉ 21. Phần lớn bệnh nhân mắc bệnh tiểu đường dạng 2 và tỉ lệ người bệnh tăng cao liên quan trực tiếp với cách sống của cuộc sống hiện đại ngày nay.
* Vào khoảng giữa những năm 80 của thế kỷ trước, tổng số người mắc bệnh tiểu đường trên thế giới vào khoảng 30 triệu. Ngày nay con số này đã lên tới 246 triệu và theo dự đoán tới năm 2025 số người mắc bệnh sẽ lên tới 380 triệu. Mỗi năm, thế giới có khoảng 3.2 triệu người chết vì bệnh tiểu đường, tương đương với số người chết hàng năm vì bệnh HIV/AIDS. Theo thống kê của WHO, cữ mỗi 30 giây lại có 1 người mắc bệnh tiểu đường bị cắt cụt chi, mỗi ngày có khoảng 5000 người mất khả năng nhìn do biến chứng về mắt của bệnh tiểu đường.
* Căn bệnh này làm ảnh hướng lớn tới nên kinh tế thới giới. Ước tính, mỗi năm trên thế giới người ta bỏ ra khoảng 215 đến 375 tỉ đô la để chưa căn bệnh này.
* Trước đây, người ta thường ước tính chi phí cho bệnh nhân tiểu đường tại Mỹ chiếm khoảng 174 tỷ đô la mỗi năm cho những phí tổn về y tế và mất sức lao động. Nhưng nhiều bệnh nhân tiểu đường có biến chứng bệnh trước khi được chuẩn đoán nhiều năm và những khoản chí phí này thường bị bỏ qua. Để làm rõ điều này, một nhóm nghiên cứu đã tiến hành nghiên cứu mẫu trên gần 30000 người trong khoảng thời gian 2 năm trước khi được chuẩn đoán xác định bệnh tiểu đường và họ đã so sánh chi phí cho người này so với những người không mắc bệnh tiểu đường. Kết quả là chi phí y tế cho mỗi người bệnh tiểu đường chưa được chuẩn đoán mất 2864 đô la vào năm 2007, và như vậy với 6.3 triệu người chưa biết mình mắc bệnh tiểu đường, con số chi phí y tế thực sự cho người mắc bệnh tiểu đường Mỹ tăng thêm khoảng 18 tỷ đô la mỗi năm. (*Theo* [*www.diabetes.org*](http://www.diabetes.org)). Ước tính hiện tại Mỹ có khoảng trên 18 triệu người được chuẩn đoán mắc bệnh tiểu đường và khoảng 6.3 triệu người chưa được chuẩn đoán ( số người chưa được chuẩn đoán chiếm 25% số bệnh nhân tiểu đường).
* Theo con số thống kê công bố trong cuốn Diabetes Atlas của Liên đoán Bệnh tiểu đường Quốc tế thì có 40 triệu người Ấn Độ bị mắc bệnh tiểu đường vào năm 2008 và con số này dữ đoán sẽ tiếp tục tăng tới khoảng 70 triệu vào năm 2025. Các quốc gia với dân số bị mắc bệnh tiểu đường nhiều vào năm 2025 bao gồm Ấn Độ, Trung Quốc và Mỹ. Ước tính cứ năm bệnh nhân tiểu đường thì có một bệ nhân Ấn Độ. Gánh nặng kinh tế cho bệnh nhân tiểu đường ở Ấn Độ hiện đang ở mức cao nhất thế giới.

## Tình hình tại Việt Nam:

* Việt Nam vào năm 1990 tỷ lệ mắc bệnh tiểu đường chỉ ở mức 0.9% (Huế) cho đến 2.52% (Tp Hồ Chí Minh), nhưng chỉ sau 10 năm, năm 2001 tỷ lệ này ở các thành phố lớn đã 4.1%, năm 2002 tăng lên 4.4% - với mức tính ở cả cộng đồng là 2.7% dân số; nếu tính ở nhóm người có yếu tố mắc bệnh cao thì tỷ lệ này đã tăng lên 10%. Theo thống kê năm 2008, tỷ lệ mắc bệnh tiểu đường trong cả nước là trên 5% (khoảng 4.5 triệu người), tại các thành phố và khu công nghiệp có tỷ lệ từ 7 – 10%.
* Việt Nam hiện đang là một nước đang phát triển, trong khi các bệnh nhiễm trùng, bệnh lây lan còn đang là phổ biến thì nay các bệnh của một xã hội công nghiệp – bệnh không lây lan lại bùng phát với tốc độ đáng lo ngại. Do những thay đổi đột ngột về kinh tế, xã hội kéo theo những thay đổi về lối song đã làm cho tỷ lệ bệnh không lây lan tăng cao, tăng nhanh, trong khi chúng ta không có kinh nghiệm trong lĩnh vực này. Thậm chí hầu hết ở các địa phương trong cả nước không có bác sỹ chuyên khoa chuyên ngành Nội tiết và Rối loạn chuyển khóa. Theo thống kê, báo cáo của Vụ Điều Trị - Bộ Y tế, năm 2005, một trăm phần trăm người mắc các bệnh nội tiết và rối loạn chuyển hóa phải chuyển lên tuyến trên. Về mặt dự phòng, nước ta cũng chưa có hệ thống để phát hiện sớm, ngăn ngừa khả năng tiến tới bệnh tiểu đường ở nhóm người có yếu tố nguy cơ cao. Đó cũng là nguyên nhân để giải thích tại sao tỷ lệ người mắc bệnh đái tháo đường không được chuẩn đoán ở Việt Nam còn cao (trên 64%), có tới 78-80% số người tham gia phỏng vấn không hiểu biết về bệnh và cách phòng bệnh.
* Theo thống kê của WHO, chi phí để điều trị cho những người mắc bệnh tiểu đường gấp từ 2-3 lần người không có bệnh. Trên thực tế, chi phí cho việc phòng bệnh tiểu đường thấp hơn nhiều lần so với chi phí điều trị các biến chứng của bệnh. Trong khi bệnh tiểu đường ở nước ta là phổ biến và có tốc độ phát triển thuộc nhóm nhanh nhất thì nhận thức của cộng đồng về bệnh tiểu đường, nhất là kiến thức về phòng bệnh lại rất thấp.

## Vì sao phải áp dụng khai phá dữ liệu vào chuẩn đoán bệnh tiểu đường?

* Theo những thống kê trên, ta có thể thấy được rằng hiện nay Việt Nam vẫn chưa có hệ thống phát hiện sớm, ngăn ngừa khả năng tiến tới bệnh tiểu đường. Điều đó đồng nghĩa với việc vẫn chưa có một hệ thống chuẩn đoán bệnh tiểu đường nào được xây dựng tại Việt Nam.
* Trên thực tế cho thấy, việc chi phí để xây dựng một hệ thống phát hiện sớm khả năng mắc bệnh tiểu đường hoàn toàn thấp hơn so với chi phí dùng để điều trị bệnh tiểu đường.
* Nhưng cũng trong chính những thuận lợi nhất lại có những khó khăn như vấn đề thu thập dữ liệu. Khi ý thức phòng bệnh của người dân còn kém thì việc khám tổng quát, khám định thì lại không được thực hiện thường xuyên. Nhất là ở những vùng nông thôn của nước ta khi mà điều kiện vật chất ở bệnh viện vẫn chưa đủ để thực hiện những xét nghiện cần thiết.
* Vì vậy, việc áp dụng khai phá dữ liệu vào chuẩn đoán bệnh tiểu đường chỉ có thực hiện ở một số thành phố lớn của đất nước nơi có tỷ lệ mắc bệnh tiểu đường cao nhất. Và nhóm nghiên cứu dự định sẽ thực hiện tại Tp Hồ Chí Minh.

# ỨNG DỤNG DATAMINING VÀO HỆ HỖ TRỢ RA QUYẾT ĐỊNH Y HỌC:

Với sự phát triển của sức mạnh điện toán và công nghệ y học , các bộ dữ liệu lớn cũng như các phương pháp phân loại dữ liệu đa dạng và phức tạp đã được phát triển và nghiên cứu. Kết quả là, khai phá dữ liệu đã thu hút được sự chú ý đáng kể trong vài thập kỷ vừa qua, và đã có một số lượng lớn các ứng dụng bao gồm cả khai phá dữ liệu và các hệ  hỗ trợ quyết định lâm sàng . Hệ hỗ trợ quyết định được đề cập như là một lớp học của hệ dựa trên máy tính nhằm hỗ trợ quá trình ra quyết định. Bảng 3.1 liệt kê một số ví dụ về các hệ hỗ trợ quyết định sử dụng các công cụ khai phá dữ liệu trong cơ sở y tế.

|  |  |
| --- | --- |
| **Hệ thống (tham khảo)** | **Giải thích** |
| **Hệ thống nhận diện và diễn giải hình ảnh y khoa** | |
| Máy tính hỗ trợ chẩn đoán ung thư vú | Sự khác biệt giữa các nốt vú lành tính và ác tính, dựa trên nhiều tính năng siêu âm |
| Chẩn đoán của chứng rối loạn thần kinh cơ | Phân loại tín hiệu điện đồ (EMG), dựa trên các hình dạng và tỷ lệ phát của đơn vị động cơ hành động tiềm năng(MUAPs) |
| **Hệ thống giáo dục** | |
| Khai thác tài liệu sinh y học | Hệ thống tự động để khai phá MEDLINE cho các tham khảo về gen và protein và đánh giá sự phù hợp của mỗi tham khảo được phân công |

Bảng . - Ví dụ về hệ hỗ trỡ quyết định lâm sàn và công cụ khai phá dữ liệu dùng nhận dạng mô hình thống kê.

Một hệ hỗ trợ quyết định điển hình bao gồm năm thành phần: quản lý dữ liệu, quản lý mô hình, động cơ tri ​​thức, giao diện người dùn, và người dùng. Một trong những sự khác biệt lớn giữa các hệ hỗ trợ quyết định sử dụng các công cụ khai phá dữ liệu và những hệ thống chuyên gia dựa trên luật là các động cơ tri thức.Trong các hệ thống hỗ trợ quyết định sử dụng hệ thống chuyên gia dựa trên nguyên tắc, các công cụ suy luận phải được cung cấp các sự kiện và các quy tắc liên quan đến chúng, thường được thể hiện trong luật "nếu-thì".Trong ý nghĩa này, hệ thống hỗ trợ quyết định đòi hỏi một lượng lớn các kiến thức tiên nghiệm của người ra quyết định để cung cấp câu trả lời đúng cho các câu hỏi .Ngược lại, hệ thống hỗ trợ quyết định sử dụng công cụ khai thác dữ liệu không đòi hỏi phải có kiến thức tiên nghiệm trên người ra quyết định.Thay vào đó, hệ thống được thiết kế để tìm mô hình mới và các mối quan hệ trong một tập hợp các dữ liệu, hệ thống sau đó áp dụng tri thức mới được phát hiện với một tập dữ liệu mới. Điều này là hữu ích khi một kiến ​​thức tiên nghiệm là hạn chế hoặc không tồn tại.

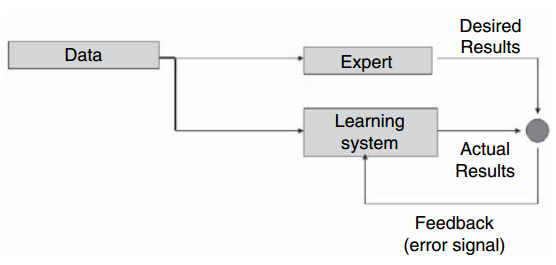
Nhiều hệ thống hỗ trợ quyết định lâm sàng đã thành công bằng cách sử dụng các hệ thống chuyên gia dựa trên nguyên tắc đã được phát triển cho các khu vực chuyên ngành trong lĩnh vực chăm sóc sức khoẻ.Một thí dụ ban đầu của một hệ thống chuyên gia dựa trên nguyên tắc là MYCIN, sử dụng các quy tắc của nó để xác định các vi sinh vật gây ra nhiễm khuẩn huyết và viêm màng não.Tuy nhiên, các hệ thống như vậy có thể được thử thách khả năng duy trì do thực tế các hệ thống thường chứa hàng ngàn các quy tắc hay. Ngoài ra, các hệ thống quy tắc "nếu-thì" có khó khăn trong việc đối phó với sự không chắc chắn. Hệ thống Bayesian là một cách để giải quyết vấn đề không chắc chắn.

## Một vài phương pháp ứng dụng data mining

Khai phá dữ liệu và mô hình dự báo có thể được xem như việc học từ dữ liệu. Trong bối cảnh này, khai phá dữ liệu có hai hướng: học có giám sát và học không giám sát.

### Học có giám sát

Học có giám sát, còn gọi là phá dữ liệu định hướng, giả định rằng người sử dụng biết trước các lớp và có mẫu của các lớp hiện có.(Hình 4.1) Tri thức này được chuyển tới hệ thống thông qua một quá trình gọi là huấn luyện. Các tập dữ liệu được sử dụng trong quá trình này được gọi là mẫu huấn luyện. Các mẫu huấn luyện bao gồm các biến phụ thuộc hoặc biến mục tiêu, và các biến độc lập hoặc đầu vào. Hệ thống này được điều chỉnh dựa trên các mẫu huấn luyện và tín hiệu lỗi (sự khác biệt giữa các phản hồi mong đợi và phản hồi thực tế của hệ thống).Nói cách khác, một hệ thống học có giám sát có thể được xem như là một hoạt động làm giảm sự khác biệt giữa các giá trị mong đợi và giá trị quan sát đoực như là sự tiến bộ quá trình đào tạo.Với đủ các mẫu trong dữ liệu huấn luyện, sự khác biệt này sẽ được giảm thiểu và sự nhận dạng mô hình sẽ được chính xác hơn.



Hình . - Học có giám sát.

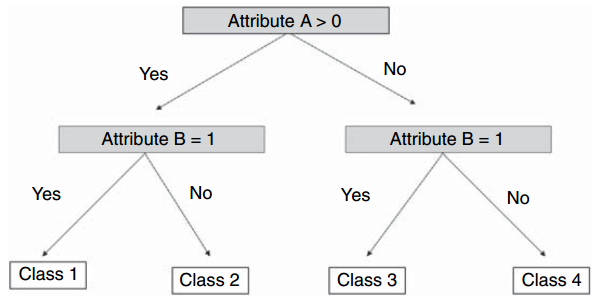
Mục tiêu của phương pháp này là để thiết lập một mối quan hệ hoặc dự đoán mô hình giữa các biến phụ thuộc và độc lập. Mô hình dự đoán thuộc về học có giám sát vì một biến được qui định là mục tiêu sẽ được giải thích như là một chức năng của biến khác.Mô hình dự báo thường được xây dựng để dự đoán giá trị trong tương lai hoặc hành vi của một đối tượng hoặc thực thể. Bản chất của biến mục tiêu/phụ thuộc xác định loại mô hình: một mô hình được gọi là một mô hình phân loại nếu biến mục tiêu là rời rạc, và một mô hình hồi quy nếu biến mục tiêu là liên tục.

Ví dụ:

Goldman et al. mô tả việc xây dựng một hệ hỗ trợ quyết định lâm sàng để dự đoán sự hiện diện của nhồi máu cơ tim trong 4.770 bệnh nhân có biểu hiện đau ngực cấp tính tại hai bệnh viện trường đại học và bốn bệnh viện cộng đồng. Dựa trên các triệu chứng và dấu hiệu của bệnh nhân, các hệ hỗ trợ quyết định lâm sàng có độ nhạy tương tự nhau (88,0% so với 87,8%) nhưng đặc biệt cao hơn đáng kể (74% so với 71%) ở dự đoán các trường hợp không có nhồi máu cơ tim khi so với quyết định của bác sĩ nếu bệnh nhân đã được yêu cầu vào khoa tim mạch. Nếu quyết định thừa nhận được chỉ dựa trên hệ hỗ trợ quyết định, việc tiếp nhận bệnh nhân không có nhồi máu cơ tim đến các đơn vị chăm sóc mạch vành đã được giảm 11,5% mà không gây ảnh hưởng xấu đến kết quả của bệnh nhân hoặc chất lượng chăm sóc.

#### Phân lớp với học có giám sát

Trong học có giám sát, phân lớp đề cập đến việc bản đồ hoá của các mục dữ liệu vào một trong những lớp đã được xác định trước. Trong sự phát triển của các công cụ khai phá dữ liệu và các hệ hỗ trợ quyết định lâm sàng sử dụng phương pháp thống kê như mô tả sau, một trong những nhiệm vụ quan trọng là tạo ra một mô hình phân lớp, mô hình sẽ dự đoán các lớp của một số thực thể hoặc các mẫu dựa trên các giá trị của các thuộc tính đầu vào. Chọn đúng phương pháp phân lớp là một bước quan trọng trong quá trình nhận dạng mô hình. Một loạt các kỹ thuật đã được sử dụng để có được sự phân lớp tốt. Một số các kỹ thuật được sử dụng rộng rãi và được nhiều người biết đến được sử dụng trong khai phá dữ liệu bao gồm các cây quyết định, hồi qui logic, mạng thần kinh, và cách tiếp cận gần nhất.



Hình . – Cây quyết định đơn giản.

### Học không giám sát

Trong học không giám sát hoặc học vô hướng, hệ thống được trình bày với một tập các dữ liệu, nhưng không có thông tin như thế nào để nhóm các dữ liệu vào các lớp học có ý nghĩa hơn (Hình 4.2). Dựa trên nhận thức sự tương đồng mà hệ thống học phát hiện trong tập dữ liệu, hệ thống phát triển các lớp hoặc cụm cho đến khi một tập các mô hình có thể định nghĩa được bắt đầu xuất hiện. Không có biến mục tiêu, tất cả các biến được đối xử theo cùng một cách mà không có sự phân biệt giữa biến phụ thuộc và biến độc lập.

C:\Users\Binh\Pictures\Capture2.PNG

Hình . - Học không giám sát

Ví dụ:

Avanzolini phân tích 13 yếu tố sinh lý được theo dõi thường xuyên trong một nhóm 200 bệnh nhân trong khoảng thời gian sáu giờ ngay sau phẫu thuật tim trong một nỗ lực để xác định bệnh nhân có nguy cơ phát triển các biến chứng sau phẫu thuật. Sử dụng một phương pháp học không giám sát (clustering), các nhà điều tra cho thấy sự tồn tại của hai loại nguy cơ được xác định rõ bệnh nhân: những người có nguy cơ phát triển các biến chứng sau mổ và những người có nguy cơ thấp.