**TRƯỜNG ĐẠI HỌC THỦY LỢI**

**KHOA CÔNG NGHỆ THÔNG TIN**

****

**BÁO CÁO BÀI TẬP LỚN**

**MÔN: KHAI PHÁ DỮ LIỆU**

**đề tài: khai phá dữ liệu dự báo nguy cơ tiểu đường giai đoạn đầu bằng thuật toán luật kết hợp và phân lớp**

|  |  |  |
| --- | --- | --- |
| GIẢNG VIÊN : | TS. TRẦN MẠNH TUẤN | |
| Nhóm sinh viên thực hiện: |  | |
| Họ và tên | Mã sinh viên | Lớp |
| Nguyễn Văn Linh | 1651060738 | 58TH3 |
| Tho Văn Hiệu | 1851061603 |  |
| Đỗ Hoàng Long | 1851061623 |  |

Hà Nội, tháng 11 năm 2021

**Mục lục**

Contents

[CHƯƠNG 1 Tổng quan về Khai phá dữ liệu 6](#_Toc89525752)

[1.1 *GIỚI THIỆU VỀ KHAI PHÁ DỮ LIỆU (DATAMING):* 6](#_Toc89525753)

[*1.1.1* *Tại sao lại cần khai phá dữ liệu (datamining)* 6](#_Toc89525754)

[1.1.2 Khai phá dữ liệu là gì? 6](#_Toc89525755)

[1.1.3 Các chức năng của khai phá dữ liệu 8](#_Toc89525756)

[1.1.4 ứng dụng của khai phá dữ liệu 8](#_Toc89525757)

[CHƯƠNG 2 Dữ liệu đầu vào của tập dữ liệu dự báo nguy cơ bệnh tiểu đường giai đoạn đầu (nguồn UCI) 9](#_Toc89525758)

[CHƯƠNG 3 Khai phá luật kết hợp cho tập dữ liệu dự báo nguy cơ bệnh tiểu đường giai đoạn đầu 10](#_Toc89525759)

[3.1 Khái niệm 10](#_Toc89525760)

[3.2 Phương pháp khai phá tập mục phổ biến(Apriori) 10](#_Toc89525761)

[3.3 Tiền xử lý dữ liệu 11](#_Toc89525762)

[3.4 Khai phá dữ liệu 13](#_Toc89525763)

[3.5 Biểu diễn tri thức 17](#_Toc89525764)

[3.6 Nhận xét và đánh giá 20](#_Toc89525765)

[CHƯƠNG 4 Phân lớp cho tập dữ liệu dự báo nguy cơ bệnh tiểu đường giai đoạn đầu 21](#_Toc89525766)

[4.1 Khái niệm 21](#_Toc89525767)

[4.2 Phân lớp bằng cây quyết định 21](#_Toc89525768)

[4.2.1 Thuật toán xây dựng cây quyết định 22](#_Toc89525769)

[4.3 Thuật toán NaïveBayes 23](#_Toc89525770)

[4.3.1 Định lý Bayes 23](#_Toc89525771)

[4.3.2 Phân lớp Naive Bayes 24](#_Toc89525772)

[4.4 Tiền xử lý dữ liệu cho thuật toán phân lớp 25](#_Toc89525773)

[4.5 Khai phá dữ liệu 27](#_Toc89525774)

[4.5.1 Naive Bayes Classifier 27](#_Toc89525775)

[4.5.2 Thuật toán J48 - cây quyết định 30](#_Toc89525776)

[4.6 Biểu diễn tri thức 33](#_Toc89525777)

[4.6.1 Phát hiện tri thức sau khi dùng thuật toán Naïve Bayes 33](#_Toc89525778)

[4.6.2 Phát hiện tri thức sau khi dùng thuật toán cây quyết định J48 33](#_Toc89525779)

[4.7 Đánh giá mô hình bằng phương pháp Hold-out: 34](#_Toc89525780)

[*Chương 5 Python* 36](#_Toc89525781)

[5.1 Apriori bằng python 36](#_Toc89525782)

[***5.1.1*** ***Code*** 36](#_Toc89525783)

[***5.1.2*** ***Kết quả*** 36](#_Toc89525784)

[*5.2* *Cây quyết định* 40](#_Toc89525785)

[*5.2.1* *code và kết quả* 40](#_Toc89525786)

[*5.2.2* *Kết luận* 42](#_Toc89525787)

[Chương 6 TÀI LIỆU THAM KHẢO 44](#_Toc89525788)

**Phân công công việc**

|  |  |
| --- | --- |
| **Họ và tên** | **Công việc thực hiện** |
| **Nguyễn Văn Linh** |  |
| **Tho Văn Hiệu** |  |
| **Đỗ Hoàng Long** |  |

**Lời nói đầu**

Trong thời buổi hiện đại ngày nay, công nghệ thông tin cũng như những ứng dụng của nó không ngừng phát triển, lượng thông tin và cơ sở dữ liệu được thu thập và lưu trữ cũng tích lũy ngày một nhiều lên. Con người cũng vì thế mà cần có thông tin với tốc độ nhanh nhất để đưa ra quyết định dựa trên lượng dữ liệu khổng lồ đã có. Các phương pháp quản trị và khai thác cơ sở dữ liệu truyền thống ngày càng không đáp ứng được thực tế, vì thế, một khuynh hướng kỹ thuật mới là Kỹ thuật phát hiện tri thức và khai phá dữ liệu nhanh chóng được phát triển. Khai phá dữ liệu đã và đang được nghiên cứu, ứng dụng trong nhiều lĩnh vực khác nhau ở các nước trên thế giới. Ở Việt Nam, kỹ thuật này đang được nghiên cứu và dần đưa vào ứng dụng. Khai phá dữ liệu là một bước trong quy trình phất hiện tri thức. Hiện nay, mọi người không ngừng tìm tòi các kỹ thuật để thực hiện khai phá dữ liệu một cách nhanh nhất và có được kết quả tốt nhất. Trong bài tập lớn này, chúng em tìm hiểu và trình bày về một kỹ thuật trong khai phá dữ liệu để phân lớp dữ liệu cũng như tổng quan về khai phá dữ liệu, với đề tài “**dự báo nguy cơ tiểu đường giai đoạn đầu bằng thuật toán phân lớp và luật kết hợp**”. Chúng em rất mong nhận được những góp ý từ thầy. Chúng em xin chân thành cảm ơn!

# Tổng quan về Khai phá dữ liệu

## *GIỚI THIỆU VỀ KHAI PHÁ DỮ LIỆU (DATAMING):*

* + 1. *Tại sao lại cần khai phá dữ liệu (datamining)*

Khoảng hơn một thập kỷ trở lại đây, lượng thông tin được lưu trữ trên các thiết bị điện tử (đĩa cứng, CD-ROM, băng từ, .v.v.) không ngừng tăng lên. Sự tích lũy dữ liệu này xảy ra với một tốc độ bùng nổ. Người ta ước đoán rằng lượng thông tin trên toàn cầu tăng gấp đôi sau khoảng hai năm và theo đó số lượng cũng như kích cỡ của các cơ sở dữ liệu (CSDL) cũng tăng lên một cách nhanh chóng. Nói một cách hình ảnh là chúng ta đang “ngập” trong dữ liệu nhưng lại “đói” tri thức. Câu hỏi đặt ra là liệu chúng ta có thể khai thác được gì từ những “núi” dữ liệu tưởng chừng như “bỏ đi” ấy không ?

“*Necessity is the mother of invention*” - Data Mining ra đời như một hướng giải quyết hữu hiệu cho câu hỏi vừa đặt ra ở trên []. Khá nhiều định nghĩa về Data Mining và sẽ được đề cập ở phần sau, tuy nhiên có thể tạm hiểu rằng Data Mining như là một công nghệ tri thức giúp khai thác những thông tin hữu ích từ những kho dữ liệu được tích trữ trong suốt quá trình hoạt động của một công ty, tổ chức nào đó. Một số kỹ thuật Khai phá dữ liệu Một số kỹ thuật Khai phá dữ liệu

### Khai phá dữ liệu là gì?

Khai phá dữ liệu (datamining) được định nghĩa như là một quá trình chắt lọc hay khai phá tri thức từ một lượng lớn dữ liệu. Một ví dụ hay được sử dụng là là việc khai thác vàng từ đá và cát, Dataming được ví như công việc "Đãi cát tìm vàng" trong một tập hợp lớn các dữ liệu cho trước. Thuật ngữ Dataming ám chỉ việc tìm kiếm một tập hợp nhỏ có giá trị từ một số lượng lớn các dữ liệu thô. Có nhiều thuật ngữ hiện được dùng cũng có nghĩa tương tự với từ Datamining như Knowledge Mining (khai phá tri thức), knowledge extraction(chắt lọc tri thức), data/patern analysis(phân tích dữ liệu/mẫu), data archaeoloogy (khảo cổ dữ liệu), datadredging(nạo vét dữ liệu),...

**Định nghĩa**: Khai phá dữ liệu là một tập hợp các kỹ thuật được sử dụng để tự động khai thác và tìm ra các mối quan hệ lẫn nhau của dữ liệu trong một tập hợp dữ liệu khổng lồ và phức tạp, đồng thời cũng tìm ra các mẫu tiềm ẩn trong tập dữ liệu đó. Khai phá dữ liệu là một bước trong bảy bước của quá trình KDD (Knowleadge Discovery in Database) và KDD được xem như 7 quá trình khác nhau theo thứ tự sau:

1. Làm sạch dữ liệu (data cleaning & preprocessing)s: Loại bỏ nhiễu và các dữ liệu không cần thiết.
2. Tích hợp dữ liệu: (data integration): quá trình hợp nhất dữ liệu thành những kho dữ liệu (data warehouses & data marts) sau khi đã làm sạch và tiền xử lý (data cleaning & preprocessing).
3. Trích chọn dữ liệu (data selection): trích chọn dữ liệu từ những kho dữ liệu và sau đó chuyển đổi về dạng thích hợp cho quá trình khai thác tri thức. Quá trình này bao gồm cả việc xử lý với dữ liệu nhiễu (noisy data), dữ liệu không đầy đủ (incomplete data), .v.v.
4. Chuyển đổi dữ liệu: Các dữ liệu được chuyển đổi sang các dạng phù hợp cho quá trình xử lý
5. Khai phá dữ liệu(data mining): Là một trong các bước quan trọng nhất, trong đó sử dụng những phương pháp thông minh để chắt lọc ra những mẫu dữ liệu.
6. Ước lượng mẫu (knowledge evaluation): Quá trình đánh giá các kết quả tìm được thông qua các độ đo nào đó.
7. Biểu diễn tri thức (knowledge presentation): Quá trình này sử dụng các kỹ thuật để biểu diễn và thể hiện trực quan cho người dùng.

Diagram, schematic

Description automatically generated

### Các chức năng của khai phá dữ liệu

Data Mining được chia nhỏ thành một số hướng chính như sau:

* Mô tả khái niệm (concept description): thiên về mô tả, tổng hợp và tóm tắt khái niệm. Ví dụ: tóm tắt văn bản.
* Luật kết hợp (association rules): là dạng luật biểu diễn tri thứ ở dạng khá đơn giản. Ví dụ: “60 % nam giới vào siêu thị nếu mua bia thì có tới 80% trong số họ sẽ mua thêm thịt bò khô”. Luật kết hợp được ứng dụng nhiều trong lĩnh vực kính doanh, y học, tin-sinh, tài chính & thị trường chứng khoán, .v.v.
* Phân lớp và dự đoán (classification & prediction): xếp một đối tượng vào một trong những lớp đã biết trước. Ví dụ: phân lớp vùng địa lý theo dữ liệu thời tiết. Hướng tiếp cận này thường sử dụng một số kỹ thuật của machine learning như cây quyết định (decision tree), mạng nơ ron nhân tạo (neural network), .v.v. Người ta còn gọi phân lớp là học có giám sát (học có thầy).
* Phân cụm (clustering): xếp các đối tượng theo từng cụm (số lượng cũng như tên của cụm chưa được biết trước. Người ta còn gọi phân cụm là học không giám sát (học không thầy).
* Khai phá chuỗi (sequential/temporal patterns): tương tự như khai phá luật kết hợp nhưng có thêm tính thứ tự và tính thời gian. Hướng tiếp cận này được ứng dụng nhiều trong lĩnh vực tài chính và thị trường chứng khoán vì nó có tính dự báo cao.

### ứng dụng của khai phá dữ liệu

Data Mining tuy là một hướng tiếp cận mới nhưng thu hút được rất nhiều sự quan tâm của các nhà nghiên cứu và phát triển nhờ vào những ứng dụng thực tiễn của nó. Chúng ta có thể liệt kê ra đây một số ứng dụng điển hình:

* Phân tích dữ liệu và hỗ trợ ra quyết định (data analysis & decision support)
* Điều trị y học (medical treatment)
* Text mining & Web mining
* Nhận dạng (pattern recognition)
* .v.v.

# Dữ liệu đầu vào của tập dữ liệu dự báo nguy cơ bệnh tiểu đường giai đoạn đầu (nguồn UCI)

Bộ dữ liệu được sử dụng để khai phá dữ liệu là bộ dữ liệu dự báo nguy cơ tiểu đường *diabetes\_data\_upload.csv* được lấy từ trang:*https://archive.ics.uci.edu/ml/machine-learning-databases/00529/*

|  |  |  |  |
| --- | --- | --- | --- |
| Thuộc tính | Giá trị | | Kiểu |
| Age (tuổi) | Từ 16 tuổi đến 90 tuổi | | Numeric |
| Gender (giới tính) | Male (nam giới) | Female (nữ giới) | Nominal |
| Polyuria (đa niệu/đi tiểu nhiều) | Yes (có) | No (không) | Nominal |
| Polydipsia (Thường xuyên khát nước) | Yes (có) | No (không) | Nominal |
| Sudden weight loss (giảm cân đột ngột) | Yes (có) | No (không) | Nominal |
| Weakness (Thể trạng yếu) | Yes (có) | No (không) | Nominal |
| Polyphagia (Ăn quá nhiều) | Yes (có) | No (không) | Nominal |
| Genital thrush (Tưa miệng) | Yes (có) | No (không) | Nominal |
| Visual blurring (Mờ thị giác) | Yes (có) | No (không) | Nominal |
| Itching (Ngứa) | Yes (có) | No (không) | Nominal |
| Irritability (Dễ cáu gắt) | Yes (có) | No (không) | Nominal |
| Delayed healing (Lâu lành vết thương) | Yes (có) | No (không) | Nominal |
| Partial paresis (Liệt một phần) | Yes (có) | No (không) | Nominal |
| Muscle stiness (Cứng cơ) | Yes (có) | No (không) | Nominal |
| Alopecia (Rụng lông tóc) | Yes (có) | No (không) | Nominal |
| Obesity (Béo phì) | Yes (có) | No (không) | Nominal |
| Class (Phân loại) | Positive (dương tính) | Negative (âm tính) | Nominal |

# Khai phá luật kết hợp cho tập dữ liệu dự báo nguy cơ bệnh tiểu đường giai đoạn đầu

### Khái niệm

Khai phá luật kết hợp là tìm ra các mẫu có tần suất cao, các mẫu kết hợp, liên quan hoặc các cấu trúc tồn tại giữa các tập hợp đối tượng trong cơ sở dữ liệu các giao dịch, cơ sở dữ liệu quan hệ hoặc các kho chứa thông tin khác. Nói cách khác là chúng ta đi tìm tất cả các tập phổ biến từ trong dữ liệu.

### Phương pháp khai phá tập mục phổ biến(Apriori)

Apriori là một giải thuật được R. Agrawal, R. Srikant đề xuất lần đầu vào năm 1994 nhằm khai phá tập mục phổ biến nhị phân. Thuật toán này thực hiện lặp lại việc tìm kiếm theo mức, sử dụng thông tin ở mức *k* để duyệt mức *k+1*. Đầu tiên, tập các mục thường xuyên có độ dài là 1 được xây dựng bằng việc duyệt qua toàn bộ dữ liệu để đếm sự xuất hiện của từng phần tử và giá trị này phải lớn hơn hoặc bằng độ hỗ trợ nhỏ nhất(minsup). Kết quả của việc đếm này được ký hiệu là . Tiếp theo này được sử dụng dụng để tìm là tập mục thường xuyên có độ dài 2. Tác vụ này được thực hiện lặp lại đến khi không tìm được tập mục thường xuyên có độ dài k thỏa mãn điều kiện minsup. Lưu ý rằng mỗi lần thực hiện việc tìm tập các mục thường xuyên yêu cầu duyệt toàn bộ dữ liệu. Từ tập mục thường xuyên này ta sinh ra luật kết hợp mạnh bằng cách tìm các luật trong tập mục thường xuyên thỏa mãn ngưỡng minconf.

Độ hỗ trợ của một luât X Y là tỉ % các giao dịch trong T chứa cả X và Y cả X và Y.

Support(XY) = = P(XY)

Độ tin cậy của luật X Y là tỉ lệ % các giao dịch trong T chứa cả X và Y trên tổng số các giao dịch trong T chỉ chứa X.

Confidence(XY) = = P(XY | X)

1. **Thuật toán Apriori**
2. Duyệt toàn bộ CSDL giao dịch để tính giá trị hỗ trợ là phẩn tử của tập phổ biến tiềm năng của 1-itemset, so sánh với minsup, để có được 1-itemset()
3. nối(phép join) để sinh ra 2-itemset là tập phổ biến tiềm năng. Loại bỏ các tập mục không phải là tập phổ biến thu được 2-itemset
4. Duyệt toàn bộ CSDL giao dịch để tính giá trị hỗ trợ của mỗi ứng viên 2-itemset, so sánh từng phần tử với minsup để thư được tập mục thường xuyên 2-itemset()
5. Lặp lại từ bước 2 cho đến khi tập ứng cử tiềm năng C =
6. Với mỗi mục phổ biến *I*, sinh tất cả các tập con s không rỗng của *I*
7. Với mỗi tập con s không rỗng của *I*, sinh ra các luật s => (I-s) nếu độ tin cậy (Confidence) của nó minconf

### Tiền xử lý dữ liệu

Trong bộ dữ liệu dự báo nguy cơ tiểu đường thuộc tính Age có kiểu dữ liệu là số mà thuật toán Apriori không làm việc với dữ liệu là số. Chính vì vậy để áp dụng thuật toán Apriori với bộ dữ liệu trên ta cần áp dụng phương pháp rời rạc hóa dữ liệu cho thuộc tính Age và chuyển kiểu dữ liệu cho thuộc tính này từ Numeric trở thành Nominal. Ta thấy thuộc tính Age có giá trị nhỏ nhất là 16 và giá trị lớn nhất là 90.

A screenshot of a computer

Description automatically generated

*Hình 3.1 Tập dữ liệu về bệnh tiểu đường trước khi được tiền xử lý*

A screenshot of a computer

Description automatically generated

*Hình 3.2 Tiền xử lý dữ liệu trên phần mềm Weka*

Sử dụng Weka với bộ lọc weka ->filters.unsupervised ->Discretize để rời rạc hóa dữ liệu cho thuộc tính Age.

Sử dụng phương pháp chia giá trị của thuộc tính Age theo khoảng rộng và chia thành 3 bins.

Chúng ta nhận được kết quả giá trị của thuộc tính Age đã được chuyển thành 3 giá trị và từ kiểu dữ liệu Numeric đã được chuyển thành kiểu Nominal.

* Giá trị thứ nhất là (-inf-40.7]: thể hiển những người có độ tuổi nhỏ hơn hoặc bằng 40.7.
* Giá trị thứ hai là (-40.7-65.3]: thể hiện những người có độ tuổi trên 40.7 đến 65.3.
* Giá trị thứ 3 là (65.3-inf): thể hiện những người có độ tuổi lớn hơn 65.3.

### Khai phá dữ liệu

A screenshot of a computer

Description automatically generated

*Hình 3.5 Sử dụng thuật toán Apriori*

Chúng ta chọn Associate và sau đó lựa chọn thuật toán weka ->associations ->Apriori để khai phá dữ liệu luật kết hợp bằng thuật toán Apriori.

Ở đây nhóm chúng em cài đặt thông số cho thuật toán Apriori là:

lowerBoundMinSupport là 0.3: Có nghĩa thuật toán chỉ lấy những tập phổ biến có độ hỗ trợ lớn hơn hoặc bằng 0.3.

metricType là Confidence và minMetric là 0.9: Có nghĩa tập luật sẽ chỉ lấy những luật có độ tin cậy lớn hơn hoặc bằng 0.9.

numRules: Thuật toán sẽ chỉ lấy ra tối đa 30 luật có độ hỗ trợ lớn nhất và lớn hơn hoặc bằng lowerBoundMinSupport.

Sau khi chạy thuật toán Apriori với thông số đã cài đặt trên chúng ta nhận được kết quả với những thông số sau:

Graphical user interface, text

Description automatically generated A computer screen capture

Description automatically generated with medium confidence

*Hình 3.7 Kết quả sau khi chạy thuật toán Apriori – luật kết hợp*

Minimum support: 0.3 (156 instances)

Minimum metric <confidence>: 0.9

Number of cycles performed: 14

Generated sets of large itemsets:

* Size of set of large itemsets L(1): 31
* Size of set of large itemsets L(2): 180
* Size of set of large itemsets L(3): 232
* Size of set of large itemsets L(4): 66
* Size of set of large itemsets L(5): 2

Best rules found:

1. Polyuria=Yes Polydipsia=Yes 193 ==> class=Positive 193 <conf:(1)> lift:(1.63) lev:(0.14) [74] conv:(74.23)

2. Polyuria=Yes Alopecia=No 187 ==> class=Positive 187 <conf:(1)> lift:(1.63) lev:(0.14) [71] conv:(71.92)

3. Polyuria=Yes Polydipsia=Yes Alopecia=No 166 ==> class=Positive 166 <conf:(1)> lift:(1.63) lev:(0.12) [63] conv:(63.85)

4. Polydipsia=Yes weakness=Yes 179 ==> class=Positive 175 <conf:(0.98)> lift:(1.59) lev:(0.12) [64] conv:(13.77)

5. partial paresis=No class=Negative 168 ==> Polydipsia=No 164 <conf:(0.98)> lift:(1.77) lev:(0.14) [71] conv:(15.06)

6. Polyuria=Yes sudden weight loss=Yes 165 ==> class=Positive 161 <conf:(0.98)> lift:(1.59) lev:(0.11) [59] conv:(12.69)

Polyuria=No partial paresis=No class=Negative 160 ==> Polydipsia=No 156 <conf:(0.97)> lift:(1.77) lev:(0.13) [67] conv:(14.34)

8. sudden weight loss=No class=Negative 171 ==> Polydipsia=No 166 <conf:(0.97)> lift:(1.76) lev:(0.14) [71] conv:(12.77)

9. Polydipsia=Yes 233 ==> class=Positive 225 <conf:(0.97)> lift:(1.57) lev:(0.16) [81] conv:(9.96)

10. Polyuria=Yes Obesity=No 202 ==> class=Positive 195 <conf:(0.97)> lift:(1.57) lev:(0.14) [70] conv:(9.71)

11. Gender=Female Alopecia=No 165 ==> class=Positive 159 <conf:(0.96)> lift:(1.57) lev:(0.11) [57] conv:(9.07)

12. Polydipsia=Yes Alopecia=No 191 ==> class=Positive 184 <conf:(0.96)> lift:(1.57) lev:(0.13) [66] conv:(9.18)

13. Polyuria=Yes weakness=Yes 185 ==> class=Positive 178 <conf:(0.96)> lift:(1.56) lev:(0.12) [64] conv:(8.89)

14. Polydipsia=Yes Obesity=No 184 ==> class=Positive 177 <conf:(0.96)> lift:(1.56) lev:(0.12) [63] conv:(8.85)

15. sudden weight loss=No Irritability=No muscle stiffness=No 176 ==> Obesity=No 169 <conf:(0.96)> lift:(1.16) lev:(0.04) [22] conv:(3.72)

16. class=Negative 200 ==> Polydipsia=No 192 <conf:(0.96)> lift:(1.74) lev:(0.16) [81] conv:(9.96)

17. Obesity=No class=Negative 173 ==> Polyuria=No 166 <conf:(0.96)> lift:(1.9) lev:(0.15) [78] conv:(10.73)

18. Obesity=No class=Negative 173 ==> Polydipsia=No 166 <conf:(0.96)> lift:(1.74) lev:(0.14) [70] conv:(9.69)

19. Polyuria=Yes partial paresis=Yes 168 ==> class=Positive 161 <conf:(0.96)> lift:(1.56) lev:(0.11) [57] conv:(8.08)

20. Polydipsia=No Obesity=No class=Negative 166 ==> Polyuria=No 159 <conf:(0.96)> lift:(1.9) lev:(0.14) [75] conv:(10.3)

21. Polyuria=No Obesity=No class=Negative 166 ==> Polydipsia=No 159 <conf:(0.96)> lift:(1.74) lev:(0.13) [67] conv:(9.3)

22. Irritability=No Obesity=No class=Negative 165 ==> Polyuria=No 158 <conf:(0.96)> lift:(1.9) lev:(0.14) [74] conv:(10.23)

23. Irritability=No Obesity=No class=Negative 165 ==> Polydipsia=No 158 <conf:(0.96)> lift:(1.73) lev:(0.13) [66] conv:(9.24)

24. Polyuria=No class=Negative 185 ==> Polydipsia=No 177 <conf:(0.96)> lift:(1.73) lev:(0.14) [74] conv:(9.21)

25. Irritability=No class=Negative 184 ==> Polydipsia=No 176 <conf:(0.96)> lift:(1.73) lev:(0.14) [74] conv:(9.16)

26. Gender=Male class=Negative 181 ==> Polydipsia=No 173 <conf:(0.96)> lift:(1.73) lev:(0.14) [73] conv:(9.01)

27. Polydipsia=Yes Genital thrush=No 178 ==> class=Positive 170 <conf:(0.96)> lift:(1.55) lev:(0.12) [60] conv:(7.61)

28. Obesity=No class=Negative 173 ==> Irritability=No 165 <conf:(0.95)> lift:(1.26) lev:(0.07) [33] conv:(4.66)

29. Polyuria=No Irritability=No class=Negative 173 ==> Polydipsia=No 165 <conf:(0.95)> lift:(1.73) lev:(0.13) [69] conv:(8.61)

30. Polyuria=No sudden weight loss=No Irritability=No Obesity=No 170 ==> Polydipsia=No 162 <conf:(0.95)> lift:(1.73) lev:(0.13) [68] conv:(8.46)

### Biểu diễn tri thức

1. Polyuria=Yes Polydipsia=Yes 193 ==> class=Positive 193 <conf:(1)> lift:(1.63) lev:(0.14) [74] conv:(74.23)

* Người có triệu chứng đi tiểu nhiều và thường xuyên thấy khát nước sẽ có nguy cơ mắc bệnh tiểu đường.

2. Polyuria=Yes Alopecia=No 187 ==> class=Positive 187 <conf:(1)> lift:(1.63) lev:(0.14) [71] conv:(71.92)

* Người có triệu chứng đi tiểu nhiều và **không** có triệu chứng rụng lông tóc nhiều sẽ có nguy cơ mắc bệnh tiểu đường.

3. Polyuria=Yes Polydipsia=Yes Alopecia=No 166 ==> class=Positive 166 <conf:(1)> lift:(1.63) lev:(0.12) [63] conv:(63.85)

* Người có triệu chứng đi tiểu nhiều, thường xuyên thấy khát nước và **không** có triệu chứng rụng lông tóc nhiều sẽ có nguy cơ mắc bệnh tiểu đường

4. Polydipsia=Yes weakness=Yes 179 ==> class=Positive 175 <conf:(0.98)> lift:(1.59) lev:(0.12) [64] conv:(13.77)

* Người có triệu chứng đi tiểu nhiều, thể trạng yếu sẽ có nguy cơ mắc bệnh tiểu đường

5. partial paresis=No class=Negative 168 ==> Polydipsia=No 164 <conf:(0.98)> lift:(1.77) lev:(0.14) [71] conv:(15.06)

* Người **không** bị liệt một phần sẽ sẽ **không** có triệu chứng thường xuyên khát nước

6. Polyuria=Yes sudden weight loss=Yes 165 ==> class=Positive 161 <conf:(0.98)> lift:(1.59) lev:(0.11) [59] conv:(12.69)

* Người có triệu chứng đi tiểu nhiều, bị giảm cân đột ngột sẽ có nguy cơ mắc bệnh tiểu đường

1. Polyuria=No partial paresis=No class=Negative 160 ==> Polydipsia=No 156 <conf:(0.97)> lift:(1.77) lev:(0.13) [67] conv:(14.34)

* Người **không** có triệu chứng đi tiểu nhiều,**không** bị liệt một phần, **không** bị tiểu đường sẽ **không** có triệu chứng thường xuyên khát nước

8. sudden weight loss=No class=Negative 171 ==> Polydipsia=No 166 <conf:(0.97)> lift:(1.76) lev:(0.14) [71] conv:(12.77)

* Người **không** bị giảm cân đột ngột, **không** mắc bệnh tiểu đường sẽ **không** có triệu chứng thường xuyên khát nước

9. Polydipsia=Yes 233 ==> class=Positive 225 <conf:(0.97)> lift:(1.57) lev:(0.16) [81] conv:(9.96)

* Người có triệu chứng thường xuyên khát nước sẽ có nguy cơ mắc bệnh tiểu đường

10. Polyuria=Yes Obesity=No 202 ==> class=Positive 195 <conf:(0.97)> lift:(1.57) lev:(0.14) [70] conv:(9.71)

* Người có triệu chứng đi tiểu nhiều**, không** bị béo phì có nguy cơ mắc bệnh tiểu đường

11. Gender=Female Alopecia=No 165 ==> class=Positive 159 <conf:(0.96)> lift:(1.57) lev:(0.11) [57] conv:(9.07)

* Nữ giới, **không** có triệu chứng rụng lông tóc sẽ có nguy cơ mắc bệnh tiểu đường

12. Polydipsia=Yes Alopecia=No 191 ==> class=Positive 184 <conf:(0.96)> lift:(1.57) lev:(0.13) [66] conv:(9.18)

* Người có triệu chứng thường xuyên khát nước,**không** rụng lông tóc sẽ có nguy cơ mắc bệnh tiểu đường

13. Polyuria=Yes weakness=Yes 185 ==> class=Positive 178 <conf:(0.96)> lift:(1.56) lev:(0.12) [64] conv:(8.89)

* Người có triệu chứng đi tiểu nhiều, thể trạng yếu sẽ có nguy cơ mắc bệnh tiểu đường

14. Polydipsia=Yes Obesity=No 184 ==> class=Positive 177 <conf:(0.96)> lift:(1.56) lev:(0.12) [63] conv:(8.85)

* Người có triệu chứng thường xuyên khát nước, bị béo phì sẽ có nguy cơ mắc bệnh tiểu đường

15. sudden weight loss=No Irritability=No muscle stiffness=No 176 ==> Obesity=No 169 <conf:(0.96)> lift:(1.16) lev:(0.04) [22] conv:(3.72)

* Người **không** có triệu chứng giảm cân đột ngột, **không** dễ cáu gắt, **không** bị cứng cơ sẽ **không** có nguy cơ mắc bện tiểu đường

16. class=Negative 200 ==> Polydipsia=No 192 <conf:(0.96)> lift:(1.74) lev:(0.16) [81] conv:(9.96)

* Người **không** mắc bệnh tiểu đường sẽ **không** có triệu chứng thường xuyên khát nước

17. Obesity=No class=Negative 173 ==> Polyuria=No 166 <conf:(0.96)> lift:(1.9) lev:(0.15) [78] conv:(10.73)

* Người không bị béo phì, không mắc bệnh tiểu đường sẽ **không** có triệu chứng đi tiểu nhiều

18. Obesity=No class=Negative 173 ==> Polydipsia=No 166 <conf:(0.96)> lift:(1.74) lev:(0.14) [70] conv:(9.69)

* Người **không** bị béo phì, **không** mắc bệnh tiểu đường sẽ **không** có triệu chứng thường xuyên khát nước

19. Polyuria=Yes partial paresis=Yes 168 ==> class=Positive 161 <conf:(0.96)> lift:(1.56) lev:(0.11) [57] conv:(8.08)

* Người có triệu chứng đi tiểu nhiều,bị liệt một phần sẽ có nguy cơ mắc bệnh tiểu đường

20. Polydipsia=No Obesity=No class=Negative 166 ==> Polyuria=No 159 <conf:(0.96)> lift:(1.9) lev:(0.14) [75] conv:(10.3)

* Người **không** có triệu chứng thường xuyên khát nước, **không** bị béo phì, **không** mắc bệnh tiểu đường sẽ **không** có triệu chứng đi tiểu nhiều

21. Polyuria=No Obesity=No class=Negative 166 ==> Polydipsia=No 159 <conf:(0.96)> lift:(1.74) lev:(0.13) [67] conv:(9.3)

* Người **không** có triệu chứng đi tiểu nhiều, **không** bị béo phì, **không** mắc bệnh tiểu đường sẽ **không** có triệu chứng thường xuyên khát nước

22. Irritability=No Obesity=No class=Negative 165 ==> Polyuria=No 158 <conf:(0.96)> lift:(1.9) lev:(0.14) [74] conv:(10.23)

* Người **không** có triệu chứng dễ cáu gắt**, không** bị béo phì, **không** mắc bệnh tiểu đường sẽ **không** có triệu chứng đi tiểu nhiều

23. Irritability=No Obesity=No class=Negative 165 ==> Polydipsia=No 158 <conf:(0.96)> lift:(1.73) lev:(0.13) [66] conv:(9.24)

* Người **không** có triệu chứng dễ cáu gắt, **không** bị béo phì, **không** mắc bệnh tiểu đường sẽ **không** có triệu chứng thường xuyên khát nước

24. Polyuria=No class=Negative 185 ==> Polydipsia=No 177 <conf:(0.96)> lift:(1.73) lev:(0.14) [74] conv:(9.21)

* Người không có triệu chứng đi tiểu nhiều, không mắc bệnh tiểu đường sẽ **không** có triệu chứng thường xuyên khát nước

25. Irritability=No class=Negative 184 ==> Polydipsia=No 176 <conf:(0.96)> lift:(1.73) lev:(0.14) [74] conv:(9.16)

* Người **không** có triệu chứng dễ cáu gắt, **không** mắc bệnh tiểu đường, sẽ **không** có triệu chứng thường xuyên khát nước

26. Gender=Male class=Negative 181 ==> Polydipsia=No 173 <conf:(0.96)> lift:(1.73) lev:(0.14) [73] conv:(9.01)

* Nam giới **không** mắc bệnh tiểu đường sẽ **không** có triệu chứng thường xuyên khát nước

27. Polydipsia=Yes Genital thrush=No 178 ==> class=Positive 170 <conf:(0.96)> lift:(1.55) lev:(0.12) [60] conv:(7.61)

* Người có triệu chứng thường xuyên khát nước, **không** có triệu chứng tưa miệng sẽ có nguy cơ mắc bệnh tiểu đường

28. Obesity=No class=Negative 173 ==> Irritability=No 165 <conf:(0.95)> lift:(1.26) lev:(0.07) [33] conv:(4.66)

* Người **không** mắc bệnh béo phì, **không** mắc bệnh tiểu đường, sẽ **không** dễ bị cáu gắt

29. Polyuria=No Irritability=No class=Negative 173 ==> Polydipsia=No 165 <conf:(0.95)> lift:(1.73) lev:(0.13) [69] conv:(8.61)

* Người **không** có triệu chứng đi tiểu nhiều, **không** dễ bị cáu gắt **không** mắc bệnh tiểu đường sẽ **không** có triệu chứng thường xuyên khát nước

30. Polyuria=No sudden weight loss=No Irritability=No Obesity=No 170 ==> Polydipsia=No 162 <conf:(0.95)> lift:(1.73) lev:(0.13) [68] conv:(8.46)

* Người **không** có triệu chứng đi tiểu nhiều, **không** có triệu chứng giảm cân đột ngột, **không** dễ bị cáu gắt, **không** mắc bệnh béo phì, sẽ **không** có triệu chứng thường xuyên khát nước

### Nhận xét và đánh giá

Sau bước biểu diễn tri thức từ kết quả nhận được sau khi sử dụng thuật toán Apriori để khai phá luật kết hợp đã cho thấy các triệu chứng đi tiểu nhiều, thường xuyên thấy khát nước, dễ cáu gắt hay căn bệnh béo phì với căn bệnh tiểu đường giai đoạn đầu chúng có một mối liên hệ mật thiết với nhau.

Nhưng cũng từ tri thức mà chúng ta nhận được, cho thấy việc sử dụng thuật toán Apriori trong khai phá dữ liệu về bệnh nhân đôi khi còn nhiều những hạn chế. Đối với ngành y học việc chẩn đoán bệnh cần phải có độ chính xác cao nên việc khai phá dữ liệu trong ngành này sẽ cần một tập dữ liệu lớn với số lượng mẫu lớn và đa dạng thì đó lại là mặt hạn chế của thuật toán Apriori khi phải làm việc với lượng dữ liệu lớn.

Không những vậy việc lựa chọn độ hỗ trợ tối thiểu và độ đo lường tối thiểu (ở đây nhóm sử dụng độ tin cậy) phù hợp cũng đóng góp cực kỳ lớn trong việc khám phá ra tri thức chính xác và không bỏ qua những tri thức quan trọng. Đây là một việc không hề dễ dàng mà thậm chí còn nó còn là khó khăn để lựa chọn được độ hỗ trợ tối thiểu và độ đo lường tối thiểu tốt.

# Phân lớp cho tập dữ liệu dự báo nguy cơ bệnh tiểu đường giai đoạn đầu

## Khái niệm

Phân lớp (Classification/Categorization) thực hiện việc xây dựng (mô tả) các mô hình (hàm) dự báo nhằm mô tả hoặc phát hiện các lớp hoặc khái niệm cho các dự báo tiếp theo. Một số phương pháp điển hình là cây quyết định, luật phân lớp, mạng neuron. Nội dụng của phân lớp chính là học một hàm ánh xạ các dữ liệu vào một trong một số lớp đã biết.

Bản chất của bài toán phân lớp là dự đoán các nhãn (hay lớp) của các phần tử dữ liệu đầu vào và các nhãn (hay lớp) này là các giá trị rời rạc. Thông thường, các giải thuật phân lớp thường hoạt động thông qua 2 bước. Bước đầu tiên nó sẽ phân tích tập dữ liệu đã gán nhãn để tìm ra mô hình phù hợp mô tả tập dữ liệu đó. Bước này được gọi là bước học (learning step) hay pha học (learning phase) và tập dữ liệu gán nhãn phục vụ quá trình học này được gọi là dữ liệ huấn luyện (training data). Dữ liệu huấn luyện là một tập các phần tử dữ liệu (data point) có gán nhãn, hay còn được gọi là bản ghi (tuple) mô tả dữ liệu và nhãn (hay lớp) tương ứng của bản ghi đó. Bước thứ hai là bước kiểm tra hiệu năng của chúng. Để tránh hiện tượng quá phù hợp (overfit), một tập dữ liệu khác gọi là tập dữ liệu kiểm thử (testing set) sẽ được sử dụng để đo độ chính xác của thuật toán.

## Phân lớp bằng cây quyết định

Giải thuật học dựa trên cây quyết định hoạt động trên tập dữ liệu được biểu diễn bằng các thuộc tính có giá trị liên tục thì cần thực hiện bước rời rạc hóa. Các giải thuật phân lớp bằng cây quyết định (ID3, CART, C4.5) đều áp dụng cách tiếp cận ăn tham (greedy) (một thuật toán không quay lui) (non-backtracking) để xây dựng cây theo hướng từ trên xuống. Tập dữ liệu huấn luyện được sẽ được chia thành các tập nhỏ hơn trong quá trình xây dựng cây theo cơ chế chia để trị (devide-and-conquer).

* Mô tả sơ lược về cây quyết định:

Để nạp dữ liệu mushroom và Weka, ta cần thêm thông tin các thuộc tính nằm trong file mô tả vào tập dữ liệu, sau đó chuyển đuôi mở rộng thành \*.csv

Dữ liệu đầu vào để dự báo nguy cơ mắc bệnh tiểu đường giai đoạn đầu lấy link sau:

*https://archive.ics.uci.edu/ml/machine-learning-databases/00529/*

* Tập dữ liệu mô tảdấu hiệu của bệnh nhân mới mắc bệnh tiểu đường hoặc sẽ là bệnh nhân tiểu đường:
  + Số lượng mẫu: 520
  + Số lượng thuộc tính: 17
  + Kiểu của mỗi thuộc tính: numeric.
  + Sự phân bố của dữ liệu vào các phân lớp khá cân bằng. Số lượng các phân lớp không áp đảo nhau (imbalanced).
  + Ta dùng filter -> unsupervised -> attribute -> **ReplaceMissingValues** để điền các giá trị thiếu.

### Thuật toán xây dựng cây quyết định

Đầu vào: Tập D chứa dữ liệu huấn luyện

Attribute\_list chứa danh sách các thuộc tính ứng cử

Đầu ra: Cây quyết định

**Generate\_decision\_tree (D, attribute\_list)**

1. Tạo một nút gốc N cho cây quyết định
2. If toàn bộ dữ liệu trong D đều thuộc lớp C, return nút N là nút lá có nhãn C
3. If attribute\_list là rỗng, return nút N với nhãn là lớp xuất hiện nhiều nhất trong D
4. Splitting\_attribute = attribute\_selection\_method ( D, attribute\_list ) tìm thuộc tính phân chia tốt nhất
5. Gán cho nút N nhãn là splitting\_attribute
6. Attribute\_list 🡸 attribute\_list \ {splitting\_attribute} (loại bỏ thuộc tính splitting\_attribute khỏi attribute\_list)
7. For each giá trị j của thuộc tính splitting\_attribute
8. Gọi Dj là tập chứa các phần tử dữ liệu mà thuộc tính splitting\_attribute có giá trị j
9. If Dj là rỗng thì thêm một nút lá Nj cho nút N có nhãn là nhãn phổ biến nhất xuất hiện trong D
10. Else gắn cây trả về bởi Generate\_decision\_tree (Dj, attribute\_list) vào nút N
11. Return N

## Thuật toán NaïveBayes

### Định lý Bayes

* Gọi A, B là hai biến cố

Với 



Suy ra:



Công thức Bayes:



* Công thức Bayes tổng quát

Với  và  là một hệ đầy đủ các biến cố:

Tổng xác suất của hệ bằng 1:



Từng đôi một xung khắc:



Khi đó ta có:



Trong đó ta gọi A là một chứng cứ (evidence) (trong bài toán phân lớp A sẽ là một phần tử dữ liệu), B là một giả thiết nào để cho A thuộc về một lớp C nào đó. Trong bài toán phân lớp chúng ta muốn xác định giá trị P(B/A) là xác suất để giả thiết B là đúng với chứng cứ A thuộc vào lớp C với điều kiện ra đã biết các thông tin mô tả A. P(B|A) là một xác suất hậu nghiệm (posterior probability hay posteriori probability) của B với điều kiện A.  
    **Posterior = Likelihood \* Prior / Evidence**

### Phân lớp Naive Bayes

    Bộ phân lớp Naive bayes hay bộ phân lớp Bayes (simple byes classifier) hoạt động như sau:

1. Gọi D là tập dữ liệu huấn luyện, trong đó mỗi phần tử dữ liệu X được biểu diễn bằng một vector chứa n giá trị thuộc tính A1, A2,...,An = {x1,x2,...,xn}
2. Giả sử có m lớp C1, C2, .. , Cm. Cho một phần tử dữ liệu X, bộ phân lớp sẽ gán nhãn cho X là lớp có xác suất hậu nghiệm lớn nhất. Cụ thể, bộ phân lớp Bayes sẽ dự đoán X thuộc vào lớp Ci  nếu và chỉ nếu:   
   Giá trị này sẽ tính dựa trên định lý Bayes.
3. Để tìm xác suất lớn nhất, ta nhận thấy các giá trị P(X) là giống nhau với mọi lớp nên không cần tính. Do đó ta chỉ cần tìm giá trị lớn nhất của P(X|Ci) \* P(Ci). Chú ý rằng P(Ci) được ước lượng bằng |Di|/|D|, trong đó Di là tập các phần tử dữ liệu thuộc lớp Ci. Nếu xác suất tiền nghiệm P(Ci) cũng không xác định được thì ta coi chúng bằng nhau P(C1) = P(C2) = ... = P(Cm), khi đó ta chỉ cần tìm giá trị P(X|Ci) lớn nhất.
4. Khi số lượng các thuộc tính mô tả dữ liệu là lớn thì chi phí tính toàn P(X|Ci) là rất lớn, dó đó có thể giảm độ phức tạp của thuật toán Naive Bayes giả thiết các thuộc tính độc lập nhau. Khi đó ta có thể tính:  
   **P(X|Ci) = P(x1|Ci)...P(xn|Ci)**

## Tiền xử lý dữ liệu cho thuật toán phân lớp

Tập dữ liệu gồm 17 thuộc tính trong đó có 16 thuộc tính rời rạc (có hoặc không, dương tính hoặc âm tính). Riêng thuộc tính Age (tuổi) là dữ liệu liên tục (từ 16 đến 90 tuổi). Ta cần rời rạc hóa dữ liệu thuộc tính tuổi bằng phương pháp bins (đóng thùng), tức là ta sẽ phân độ tuổi từ 16 đến 90 thành 3 loại trẻ (dưới 35 tuổi), trung niên (35 đến 65 tuổi) và già (trên 65 tuổi).

Chart

Description automatically generated

*Hình 4.1 Thuộc tính Age trước khi rời rạc hóa dữ liệu (độ tuổi min = 16 và max = 90)*

Chart

Description automatically generated with medium confidence

*Hình 4.2 Thuộc tính Age (tuổi) sau khi được rời rạc hóa dữ liệu thành 3 bins.*

## Khai phá dữ liệu

Sau khi đã làm xong bước tiền xử lý dữ liệu, ta lần lượt chạy các giải thuật J48 và NaiveBayes trên phần mềm Weka để khai phá dữ liệu.

### Naive Bayes Classifier

*Bảng 4.1 Kết quả phân lớp sau khi chạy thuật toán NaiveBayes*

|  |  |  |
| --- | --- | --- |
| Attribute (thuộc tính) | Class (lớp) | |
| Positive (dương tính) | Negative (âm tính) |
| 0.61 | 0.39 |
| * Age (tuổi) | | |
| Dưới 35 tuổi | 52 | 43 |
| Từ 35 đến 65 tuổi | 235 | 143 |
| Trên 65 tuổi | 36 | 17 |
| Tổng | 323 | 203 |
| * Gender (giới tính) | | |
| Male (nam giới) | 148 | 182 |
| Female (nữ giới) | 174 | 20 |
| Tổng | 322 | 202 |
| * Polyuria (đa niệu) | | |
| Yes (có) | 244 | 16 |
| No (không) | 78 | 186 |
| Tổng | 322 | 202 |
| * Polydipsia | | |
| Yes (có) | 226 | 9 |
| No (không) | 96 | 193 |
| Tổng | 322 | 202 |
| * Sudden weight loss (sút cân đột ngột) | | |
| Yes (có) | 189 | 30 |
| No (không) | 133 | 172 |
| Tổng | 322 | 202 |
| * Weakness | | |
| Yes (có) | 219 | 88 |
| No (không) | 103 | 114 |
| Tổng | 322 | 202 |
| * Polyphagia | | |
| Yes (có) | 190 | 49 |
| No (không) | 132 | 153 |
| Tổng | 322 | 202 |
| * Genital thrush (mọc mụn ở miệng) | | |
| Yes (có) | 84 | 34 |
| No (không) | 238 | 168 |
| Tổng | 322 | 202 |
| * Visual bluring (mắt mờ) | | |
| Yes (có) | 176 | 59 |
| No (không) | 146 | 143 |
| Tổng | 322 | 202 |
| * Itching (ngứa) | | |
| Yes (có) | 155 | 100 |
| No (không) | 167 | 102 |
| Tổng | 322 | 202 |
| * Irritability (cáu gắt) | | |
| Yes (có) | 111 | 17 |
| No (không) | 211 | 185 |
| Tổng | 322 | 202 |
| * Delayed healing (vết thương lâu lành) | | |
| Yes (có) | 154 | 87 |
| No (không) | 168 | 115 |
| Tổng | 322 | 202 |
| * Partial paresis (chứng liệt một phần) | | |
| Yes (có) | 193 | 33 |
| No (không) | 129 | 169 |
| Tổng | 322 | 202 |
| * Muscle stiffness (co cứng cơ bắp) | | |
| Yes (có) | 136 | 61 |
| No (không) | 186 | 141 |
| Tổng | 322 | 202 |
| * Alopecia (rụng tóc) | | |
| Yes (có) | 79 | 102 |
| No (không) | 243 | 100 |
| Tổng | 322 | 202 |
| * Obesity (Béo phì) | | |
| Yes (có) | 62 | 28 |
| No (không) | 260 | 174 |
| Tổng | 322 | 202 |

Graphical user interface, text, application

Description automatically generated

*Hình 4.3 Độ chính xác khi sử dụng thuật toán NaiveBaye*

### Thuật toán J48 - cây quyết định

Kết quả phân lớp sau khi sử dụng thuật toán J48 cây quyết định

Polyuria = No

| Polydipsia = Yes

| | Irritability = No

| | | Itching = Yes

| | | | Gender = Male

| | | | | Alopecia = Yes: Positive (5.0/1.0)

| | | | | Alopecia = No: Negative (7.0)

| | | | Gender = Female: Positive (4.0)

| | | Itching = No: Positive (6.0)

| | Irritability = Yes: Positive (18.0)

| Polydipsia = No

| | Gender = Male

| | | Irritability = No: Negative (153.0/6.0)

| | | Irritability = Yes

| | | | Genital thrush = No

| | | | | Polyphagia = No: Negative (7.0)

| | | | | Polyphagia = Yes

| | | | | | visual blurring = No: Positive (2.0)

| | | | | | visual blurring = Yes: Negative (4.0)

| | | | Genital thrush = Yes: Positive (5.0)

| | Gender = Female

| | | Alopecia = Yes: Negative (14.0/1.0)

| | | Alopecia = No: Positive (37.0/6.0)

Polyuria = Yes

| Polydipsia = Yes: Positive (193.0)

| Polydipsia = No

| | Itching = Yes

| | | delayed healing = Yes

| | | | Alopecia = Yes

| | | | | Gender = Male

| | | | | | visual blurring = No

| | | | | | | sudden weight loss = No: Positive (2.0)

| | | | | | | sudden weight loss = Yes: Negative (4.0)

| | | | | | visual blurring = Yes: Negative (11.0)

| | | | | Gender = Female: Positive (2.0)

| | | | Alopecia = No: Positive (5.0)

| | | delayed healing = No: Positive (11.0)

| | Itching = No: Positive (30.0)

Số lượng nút lá: 20

Kích thước của cây: 39

Diagram

Description automatically generated

*Hình 4.4 Cây quyết định sau khi chạy thuật toán J48*

Graphical user interface, text, application

Description automatically generated

*Hình 4.5 Độ chính xác khi sử dụng thuật toán cây quyết định J48*

## Biểu diễn tri thức

### Phát hiện tri thức sau khi dùng thuật toán Naïve Bayes

Nhìn vào bảng 4.2 kết quả phân lớp sau khi chạy thuật toán NaiveBayes chúng ta có thể phát hiện được các tri thức như sau:

* Nếu xét về độ tuổi thì độ tuổi trên 65 tuổi có tỷ lệ mắc tiểu đường cao nhất 68%
* Nếu xét về giới tính thì nữ giới có tỷ lệ mắc tiểu đường cao hơn nam giới
* Nếu người nào có Polyuria thì tỷ lệ cao là mắc tiểu đường

### Phát hiện tri thức sau khi dùng thuật toán cây quyết định J48

Sau khi nhìn vào cây quyết định được sinh ra từ thuật toán J48 ta có thể phát hiện ra được các tri thức như sau:

* Nếu người nào mà có triệu chứng đa niệu và có Polydipsia thì sẽ mắc tiểu đường
* Nếu người nào không có triệu chứng đa niệu nhưng có Polydipsia kèm theo cáu gắt thì chuẩn đoán mắc tiểu đường
* Nếu người nào mà không có đa niệu không có Polydipsia giới tính nữ không hói đầu mà có mắt mờ thì chuẩn đoán mắc tiểu đường

## Đánh giá mô hình bằng phương pháp Hold-out:

Chúng ta sẽ chia dữ liệu thành 2 phần: 67% để xây dựng mô hình phân lớp (train), 33% để kiểm tra (test).

*Bảng 4.2 Đánh giá thuật toán Naïve Bayes và Bảng quyết định (J48)*

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Classifier | Precision | Recall | F-measure | Confusion matrix |
| Naive Bayes | 0,931 | 0,848 | 0,888 | a b 🡸 classified as  95 17 | a = Positive  7 53 | b = Negative |
| J48 decision tree | 0,960 | 0,866 | 0,911 | a b 🡸 classified as  97 15 | a = Positive  4 56 | b = Negative |

Qua kết quả phân lớp trên, ta thấy mô hình Naive Bayes và J48 đều không cho kết quả phân lớp chính xác hoàn toàn. Dựa vào cây quyết định, ta có thể biết được một người là bệnh nhân tiểu đường mới hoặc sẽ là bệnh nhân tiểu đường thông qua việc đánh giá các triệu chứng và các mẹo hữu ích để kiểm soát các yếu tố nguy cơ đã được đề xuất

Về đặc điểm đa phần những người Polyuria ( Đa niệu) là Yes thường mắc bệnh theo chiều hướng tích cực.

# *Chương 5 Python*

### Apriori bằng python

* + 1. ***Code***

Graphical user interface, text, application

Description automatically generated

Graphical user interface, text, application, email

Description automatically generated

* + 1. ***Kết quả***

***Title1 Title2 Support Confidence Lift***

***0 No 1.0 1.0 1.0 1.0***

***Title1 Title2 Support Confidence Lift***

***0 No 1.0 1.0 1.0 1.0***

***1 Female 0.36923076923076925 0.36923 1.0 1.0***

***Title1 Title2 Support Confidence Lift***

***0 No 1.0 1.0 1.0 1.0***

***1 Female 0.36923076923076925 0.36923 1.0 1.0***

***2 Positive 0.3326923076923077 0.33269 0.90104 1.46419***

***Title1 Title2 Support Confidence Lift***

***0 No 1.0 1.0 1.0 1.0***

***1 Female 0.36923076923076925 0.36923 1.0 1.0***

***2 Positive 0.3326923076923077 0.33269 0.90104 1.46419***

***3 Yes 0.35192307692307695 0.35192 0.95312 1.06129***

***Title1 Title2 Support Confidence Lift***

***0 No 1.0 1.0 1.0 1.0***

***1 Female 0.36923076923076925 0.36923 1.0 1.0***

***2 Positive 0.3326923076923077 0.33269 0.90104 1.46419***

***3 Yes 0.35192307692307695 0.35192 0.95312 1.06129***

***4 Male 0.34807692307692306 0.34807 0.90499 1.43475***

***Title1 Title2 Support Confidence Lift***

***0 No 1.0 1.0 1.0 1.0***

***1 Female 0.36923076923076925 0.36923 1.0 1.0***

***2 Positive 0.3326923076923077 0.33269 0.90104 1.46419***

***3 Yes 0.35192307692307695 0.35192 0.95312 1.06129***

***4 Male 0.34807692307692306 0.34807 0.90499 1.43475***

***5 Male 0.6307692307692307 0.63076 1.0 1.0***

***Title1 Title2 Support Confidence Lift***

***0 No 1.0 1.0 1.0 1.0***

***1 Female 0.36923076923076925 0.36923 1.0 1.0***

***2 Positive 0.3326923076923077 0.33269 0.90104 1.46419***

***3 Yes 0.35192307692307695 0.35192 0.95312 1.06129***

***4 Male 0.34807692307692306 0.34807 0.90499 1.43475***

***5 Male 0.6307692307692307 0.63076 1.0 1.0***

***6 Negative 0.38461538461538464 0.38461 1.0 1.0***

***Title1 Title2 Support Confidence Lift***

***0 No 1.0 1.0 1.0 1.0***

***1 Female 0.36923076923076925 0.36923 1.0 1.0***

***2 Positive 0.3326923076923077 0.33269 0.90104 1.46419***

***3 Yes 0.35192307692307695 0.35192 0.95312 1.06129***

***4 Male 0.34807692307692306 0.34807 0.90499 1.43475***

***5 Male 0.6307692307692307 0.63076 1.0 1.0***

***6 Negative 0.38461538461538464 0.38461 1.0 1.0***

***7 Positive 0.6153846153846154 0.61538 1.0 1.0***

***Title1 Title2 Support Confidence Lift***

***0 No 1.0 1.0 1.0 1.0***

***1 Female 0.36923076923076925 0.36923 1.0 1.0***

***2 Positive 0.3326923076923077 0.33269 0.90104 1.46419***

***3 Yes 0.35192307692307695 0.35192 0.95312 1.06129***

***4 Male 0.34807692307692306 0.34807 0.90499 1.43475***

***5 Male 0.6307692307692307 0.63076 1.0 1.0***

***6 Negative 0.38461538461538464 0.38461 1.0 1.0***

***7 Positive 0.6153846153846154 0.61538 1.0 1.0***

***8 Yes 0.8980769230769231 0.89807 1.0 1.0***

***Title1 Title2 Support Confidence Lift***

***0 No 1.0 1.0 1.0 1.0***

***1 Female 0.36923076923076925 0.36923 1.0 1.0***

***2 Positive 0.3326923076923077 0.33269 0.90104 1.46419***

***3 Yes 0.35192307692307695 0.35192 0.95312 1.06129***

***4 Male 0.34807692307692306 0.34807 0.90499 1.43475***

***5 Male 0.6307692307692307 0.63076 1.0 1.0***

***6 Negative 0.38461538461538464 0.38461 1.0 1.0***

***7 Positive 0.6153846153846154 0.61538 1.0 1.0***

***8 Yes 0.8980769230769231 0.89807 1.0 1.0***

***9 Positive 0.6038461538461538 0.60384 0.98124 1.09261***

***Title1 Title2 Support Confidence Lift***

***0 No 1.0 1.0 1.0 1.0***

***1 Female 0.36923076923076925 0.36923 1.0 1.0***

***2 Positive 0.3326923076923077 0.33269 0.90104 1.46419***

***3 Yes 0.35192307692307695 0.35192 0.95312 1.06129***

***4 Male 0.34807692307692306 0.34807 0.90499 1.43475***

***5 Male 0.6307692307692307 0.63076 1.0 1.0***

***6 Negative 0.38461538461538464 0.38461 1.0 1.0***

***7 Positive 0.6153846153846154 0.61538 1.0 1.0***

***8 Yes 0.8980769230769231 0.89807 1.0 1.0***

***9 Positive 0.6038461538461538 0.60384 0.98124 1.09261***

***10 Positive 0.3326923076923077 0.33269 0.90104 1.46419***

***Title1 Title2 Support Confidence Lift***

***0 No 1.0 1.0 1.0 1.0***

***1 Female 0.36923076923076925 0.36923 1.0 1.0***

***2 Positive 0.3326923076923077 0.33269 0.90104 1.46419***

***3 Yes 0.35192307692307695 0.35192 0.95312 1.06129***

***4 Male 0.34807692307692306 0.34807 0.90499 1.43475***

***5 Male 0.6307692307692307 0.63076 1.0 1.0***

***6 Negative 0.38461538461538464 0.38461 1.0 1.0***

***7 Positive 0.6153846153846154 0.61538 1.0 1.0***

***8 Yes 0.8980769230769231 0.89807 1.0 1.0***

***9 Positive 0.6038461538461538 0.60384 0.98124 1.09261***

***10 Positive 0.3326923076923077 0.33269 0.90104 1.46419***

***11 Yes 0.35192307692307695 0.35192 0.95312 1.06129***

***Title1 Title2 Support Confidence Lift***

***0 No 1.0 1.0 1.0 1.0***

***1 Female 0.36923076923076925 0.36923 1.0 1.0***

***2 Positive 0.3326923076923077 0.33269 0.90104 1.46419***

***3 Yes 0.35192307692307695 0.35192 0.95312 1.06129***

***4 Male 0.34807692307692306 0.34807 0.90499 1.43475***

***5 Male 0.6307692307692307 0.63076 1.0 1.0***

***6 Negative 0.38461538461538464 0.38461 1.0 1.0***

***7 Positive 0.6153846153846154 0.61538 1.0 1.0***

***8 Yes 0.8980769230769231 0.89807 1.0 1.0***

***9 Positive 0.6038461538461538 0.60384 0.98124 1.09261***

***10 Positive 0.3326923076923077 0.33269 0.90104 1.46419***

***11 Yes 0.35192307692307695 0.35192 0.95312 1.06129***

***12 Positive 0.3211538461538462 0.32115 0.96531 1.07487***

***Title1 Title2 Support Confidence Lift***

***0 No 1.0 1.0 1.0 1.0***

***1 Female 0.36923076923076925 0.36923 1.0 1.0***

***2 Positive 0.3326923076923077 0.33269 0.90104 1.46419***

***3 Yes 0.35192307692307695 0.35192 0.95312 1.06129***

***4 Male 0.34807692307692306 0.34807 0.90499 1.43475***

***5 Male 0.6307692307692307 0.63076 1.0 1.0***

***6 Negative 0.38461538461538464 0.38461 1.0 1.0***

***7 Positive 0.6153846153846154 0.61538 1.0 1.0***

***8 Yes 0.8980769230769231 0.89807 1.0 1.0***

***9 Positive 0.6038461538461538 0.60384 0.98124 1.09261***

***10 Positive 0.3326923076923077 0.33269 0.90104 1.46419***

***11 Yes 0.35192307692307695 0.35192 0.95312 1.06129***

***12 Positive 0.3211538461538462 0.32115 0.96531 1.07487***

***13 Male 0.34807692307692306 0.34807 0.90499 1.43475***

***Title1 Title2 Support Confidence Lift***

***0 No 1.0 1.0 1.0 1.0***

***1 Female 0.36923076923076925 0.36923 1.0 1.0***

***2 Positive 0.3326923076923077 0.33269 0.90104 1.46419***

***3 Yes 0.35192307692307695 0.35192 0.95312 1.06129***

***4 Male 0.34807692307692306 0.34807 0.90499 1.43475***

***5 Male 0.6307692307692307 0.63076 1.0 1.0***

***6 Negative 0.38461538461538464 0.38461 1.0 1.0***

***7 Positive 0.6153846153846154 0.61538 1.0 1.0***

***8 Yes 0.8980769230769231 0.89807 1.0 1.0***

***9 Positive 0.6038461538461538 0.60384 0.98124 1.09261***

***10 Positive 0.3326923076923077 0.33269 0.90104 1.46419***

***11 Yes 0.35192307692307695 0.35192 0.95312 1.06129***

***12 Positive 0.3211538461538462 0.32115 0.96531 1.07487***

***13 Male 0.34807692307692306 0.34807 0.90499 1.43475***

***14 Male 0.5461538461538461 0.54615 1.0 1.0***

***Title1 Title2 Support Confidence Lift***

***0 No 1.0 1.0 1.0 1.0***

***1 Female 0.36923076923076925 0.36923 1.0 1.0***

***2 Positive 0.3326923076923077 0.33269 0.90104 1.46419***

***3 Yes 0.35192307692307695 0.35192 0.95312 1.06129***

***4 Male 0.34807692307692306 0.34807 0.90499 1.43475***

***5 Male 0.6307692307692307 0.63076 1.0 1.0***

***6 Negative 0.38461538461538464 0.38461 1.0 1.0***

***7 Positive 0.6153846153846154 0.61538 1.0 1.0***

***8 Yes 0.8980769230769231 0.89807 1.0 1.0***

***9 Positive 0.6038461538461538 0.60384 0.98124 1.09261***

***10 Positive 0.3326923076923077 0.33269 0.90104 1.46419***

***11 Yes 0.35192307692307695 0.35192 0.95312 1.06129***

***12 Positive 0.3211538461538462 0.32115 0.96531 1.07487***

***13 Male 0.34807692307692306 0.34807 0.90499 1.43475***

***14 Male 0.5461538461538461 0.54615 1.0 1.0***

***15 Positive 0.6038461538461538 0.60384 0.98124 1.09261***

***Title1 Title2 Support Confidence Lift***

***0 No 1.0 1.0 1.0 1.0***

***1 Female 0.36923076923076925 0.36923 1.0 1.0***

***2 Positive 0.3326923076923077 0.33269 0.90104 1.46419***

***3 Yes 0.35192307692307695 0.35192 0.95312 1.06129***

***4 Male 0.34807692307692306 0.34807 0.90499 1.43475***

***5 Male 0.6307692307692307 0.63076 1.0 1.0***

***6 Negative 0.38461538461538464 0.38461 1.0 1.0***

***7 Positive 0.6153846153846154 0.61538 1.0 1.0***

***8 Yes 0.8980769230769231 0.89807 1.0 1.0***

***9 Positive 0.6038461538461538 0.60384 0.98124 1.09261***

***10 Positive 0.3326923076923077 0.33269 0.90104 1.46419***

***11 Yes 0.35192307692307695 0.35192 0.95312 1.06129***

***12 Positive 0.3211538461538462 0.32115 0.96531 1.07487***

***13 Male 0.34807692307692306 0.34807 0.90499 1.43475***

***14 Male 0.5461538461538461 0.54615 1.0 1.0***

***15 Positive 0.6038461538461538 0.60384 0.98124 1.09261***

***16 Positive 0.3211538461538462 0.32115 0.96531 1.07487***

* 1. *Cây quyết định*
     1. *code và kết quả*

import pandas as pd

from sklearn import tree

from sklearn import metrics

import matplotlib.pyplot as pl

wine= pd.read\_csv("./diabetes\_data\_upload.csv")

Gender={'Male':0,'Female':1}

wine['Gender']=wine['Gender'].map(Gender)

Polyuria={'Yes':0,'No':1}

wine['Polyuria']=wine['Polyuria'].map(Polyuria)

Polydipsia={'Yes':0,'No':1}

wine['Polydipsia']=wine['Polydipsia'].map(Polydipsia)

SuddenWeightLoss={'Yes':0,'No':1}

wine['SuddenWeightLoss']=wine['SuddenWeightLoss'].map(SuddenWeightLoss)

weakness={'Yes':0,'No':1}

wine['weakness']=wine['weakness'].map(weakness)

Polyphagia={'Yes':0,'No':1}

wine['Polyphagia']=wine['Polyphagia'].map(Polyphagia)

GenitalThrush={'Yes':0,'No':1}

wine['GenitalThrush']=wine['GenitalThrush'].map(GenitalThrush)

visualBlurring={'Yes':0,'No':1}

wine['visualBlurring']=wine['visualBlurring'].map(visualBlurring)

Itching={'Yes':0,'No':1}

wine['Itching']=wine['Itching'].map(Itching)

Irritability={'Yes':0,'No':1}

wine['Irritability']=wine['Irritability'].map(Irritability)

delayedHealing={'Yes':0,'No':1}

wine['delayedHealing']=wine['delayedHealing'].map(delayedHealing)

partialParesis={'Yes':0,'No':1}

wine['partialParesis']=wine['partialParesis'].map(partialParesis)

muscleStiffness={'Yes':0,'No':1}

wine['muscleStiffness']=wine['muscleStiffness'].map(muscleStiffness)

Alopecia={'Yes':0,'No':1}

wine['Alopecia']=wine['Alopecia'].map(Alopecia)

Obesity={'Yes':0,'No':1}

wine['Obesity']=wine['Obesity'].map(Obesity)

result={'Positive':0,'Negative':1}

wine['result']=wine['result'].map(result)

data =['Age','Gender','Polyuria','Polydipsia','SuddenWeightLoss','weakness',

'Polyphagia','GenitalThrush','visualBlurring','Itching','Irritability','delayedHealing',

'partialParesis','muscleStiffness','Alopecia','Obesity']

x=wine[data]

y=wine['result']

model= tree.DecisionTreeClassifier()

model=model.fit(x,y)

fig=pl.figure(figsize=(250,200))

\_=tree.plot\_tree(model)

Graphical user interface

Description automatically generated

* + 1. *Kết luận*

Sự bùng nổ thông tin cùng với sự phát triển ứng dụng ngày càng rộng rãi của công nghệ thông tin trong mọi lĩnh vực đã khiến nhu cầu xử lý những khỗi dữ liệu khổng lồ để kết xuất ra những thông tin, tri thức hữu ích cho người sử dụng một cách tự động, nhanh chóng và chính xác, trở thành nhân tố quan trọng hàng đầu cho mọi lĩnh vực. Để kết xuất được những thông tin, tri thức quý giá đó thì việc sử dụng khai phá dữ liệu trong việc xử lý dữ liệu là cần thiết. Khai phá dữ liệu là một trong những kỹ thuật quan trọng, mang tính thời sự không chỉ với lĩnh vực công nghệ thông tin mà còn quan trọng đối với rất nhiều lĩnh vực khác trong đó có cả lĩnh vực y tế.

Trong lĩnh vực y tế, khai phá dữ liệu cho ta những tri thức về bệnh tật. Tuy việc áp dụng kỹ thuật này vào y tế còn một số hạn chế như việc lựa chọn thuật toán phù hợp để khai phá dữ liệu, cần một lượng lớn dữ liệu về bệnh nhân, tốc độ xử lý của máy tính. Nhưng cũng đã phần nào hỗ trợ cho việc chẩn đoán bệnh trở lên nhanh chóng và tăng thêm độ chính xác cho các bác sĩ trong việc chẩn đoán bệnh nhân.

# Chương 6 TÀI LIỆU THAM KHẢO

* Slide bài giảng thầy : TS. TRẦN MẠNH TUẤN
* <https://www.youtube.com/watch?v=SVM_pX0oTU8&t=535s&ab_channel=ProgrammingKnowledge>
* *https://archive.ics.uci.edu/ml/machine-learning-databases/00529/*