TRƯỜNG ĐẠI HỌC ĐIỆN LỰC

**KHOA CÔNG NGHỆ THÔNG TIN**



**KHAI PHÁ DỮ LIỆU**

**SỬ DỤNG THUẬT TOÁN PHÂN LỚP ID3 ĐỂ**

**CHUẨN ĐOÁN BỆNH VIÊM GAN**

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| |  |  | | --- | --- | | **Sinh viên thực hiện** | **: VI TUẤN ANH**  **NGUYỄN QUỐC VIỆT**  **LÊ TIẾN LONG** | | **Giảng viên hướng dẫn** | **: VŨ VĂN ĐỊNH** | | | **Ngành** | **: CÔNG NGHỆ THÔNG TIN** | | | **Chuyên ngành** | **: CÔNG NGHỆ PHẦN MỀM** | | | **Lớp** | **: D16CNPM6** | | | **Khóa** | **: 2022 – 2026** | | |  |
|  |  |
|  |  |

***Hà Nội, tháng 5 năm 20224***

**PHIẾU CHẤM ĐIỂM**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **STT** | **Họ và tên sinh viên** | **Nội dung thực hiện** | **Điểm** | **Chữ ký** |
| 1 | Nguyễn Quốc Việt |  |  |  |
| 2 | Lê Tiến Long |  |  |  |
| 3 | Vi Tuấn Anh |  |  |  |

|  |  |  |
| --- | --- | --- |
| **Họ và tên giảng viên** | **Chữ ký** | **Ghi chú** |
| Giảng viên chấm 1: |  |  |
| Giảng viên chấm 2: |  |  |

**MỤC LỤC**

[LỜI CẢM ƠN 1](#_Toc168255732)

[CHƯƠNG 1: TỔNG QUAN KHAI PHÁ DỮ LIỆU, TIỀN XỬ LÝ DỮ LIỆU 2](#_Toc168255733)

[1.1. Đặt vấn đề 2](#_Toc168255734)

[1.2. Tổng quan về khai phá dữ liệu 2](#_Toc168255735)

[1.2.1. Khái niệm 2](#_Toc168255736)

[1.2.2 Quy trình khai phá tri thức 3](#_Toc168255737)

[1.2.3 Lợi ích của khai phá dữ liệu 5](#_Toc168255738)

[1.2.4 Ứng dụng của khai phá dữ liệu 5](#_Toc168255739)

[1.3. Tiền xử lý dữ liệu 6](#_Toc168255740)

[1.3.1 Khái niệm 6](#_Toc168255741)

[1.3.2 Làm sạch dữ liệu (data cleaning) 7](#_Toc168255742)

[1.3.3 Tích hợp dữ liệu (data integration) 7](#_Toc168255743)

[1.3.4 Biến đổi dữ liệu (data transformation) 8](#_Toc168255744)

[1.3.5 Thu giảm dữ liệu (data reduction) 8](#_Toc168255745)

[CHƯƠNG 2: KHÁI QUÁT CÁC THUẬT TOÁN ĐƯỢC SỬ DỤNG ĐỂ KHAI PHÁ DỮ LIỆU 9](#_Toc168255746)

[2.1 Phân lớp dữ liệu 9](#_Toc168255747)

[2.1.1 Khái niệm 9](#_Toc168255748)

[2.1.2 Các thuật toán phân lớp dữ liệu 10](#_Toc168255749)

[2.2 Khai phá, kết hợp 10](#_Toc168255750)

[2.2.1 Khái Niệm. 10](#_Toc168255751)

[2.2.2 Các thuật toán khai phá kết hợp 11](#_Toc168255752)

[2.3 Phân cụm 11](#_Toc168255753)

[2.3.1 Khái niệm 11](#_Toc168255754)

[2.3.2. Các thuật toán phân cụm 13](#_Toc168255755)

[2.4 Sử dụng thuật toán phân lớp ID3 để dự đoán bệnh viêm gan 13](#_Toc168255756)

[2.4.1 Khái niệm thuật toán phân lớp ID3 13](#_Toc168255757)

[2.4.2 Ưu điểm của thuật toán ID3 15](#_Toc168255758)

[2.4.3 Nhược điểm của thuật toán ID3 15](#_Toc168255759)

[2.4.4 Sử dụng thuận toán ID3 để dự đoán bệnh viêm gan 16](#_Toc168255760)

[CHƯƠNG 3 THỬ NGHIỆM, KẾT QUẢ 18](#_Toc168255761)

[KẾT LUẬN 23](#_Toc168255762)

[TÀI LIỆU THAM KHẢO 25](#_Toc168255763)

**DANH MỤC HÌNH ẢNH**

[Hình 1. 1 Hình minh họa khai phá 4](#_Toc168255764)

[Hình 1. 2 : Tiền sử lý dữ liệu 6](#_Toc168255765)

[Hình 2. 1 : Phân cụm 12](file:///D:\Downloads\Khai_Pha_Du_lieu.docx#_Toc168255766)

[Hình 3. 1 : Dữ liệu viêm gan 18](#_Toc168176096)

[Hình 3. 2 : Tiền xử lý dữ liệu 18](#_Toc168176097)

[Hình 3. 3 : Sau khi thu giảm dữ liệu 19](#_Toc168176098)

[Hình 3. 4 : Entropy 19](#_Toc168176099)

[Hình 3. 5 : Information\_ gain 20](#_Toc168176100)

[Hình 3. 6 : Chạy thuật toán ID3 20](#_Toc168176101)

[Hình 3. 7 : Kết quả thực nghiệm 21](#_Toc168176102)

[Hình 3. 8: Test chính xác 22](#_Toc168176103)

# LỜI CẢM ƠN

Lời cảm ơn sâu sắc này được gửi đến các thầy cô giáo tại Khoa Công nghệ Thông tin, Đại học Điện Lực, những người đã không ngần ngại chia sẻ kiến thức và hỗ trợ chúng em trong suốt quá trình nghiên cứu và hoàn thành đề tài này.

Chúng em xin dành một lời cảm ơn đặc biệt tới thầy VŨ VĂN ĐỊNH, người đã không chỉ là người thầy, mà còn là người hướng dẫn, người bạn đồng hành cùng chúng em trong từng bước đi của dự án, đã giúp chúng em vượt qua mọi thách thức và hoàn thành báo cáo này.

Chúng em cũng muốn bày tỏ lòng biết ơn tới tất cả bạn bè và những người đã hỗ trợ chúng em, đã cung cấp những phản hồi và góp ý thiết thực, giúp chúng em nâng cao chất lượng của báo cáo.

Chúng em nhận thức rằng mặc dù đã nỗ lực hết mình, nhưng báo cáo này không thể hoàn hảo. Chúng em rất trân trọng mọi góp ý từ phía quý thầy cô và các bạn, những người sẽ giúp chúng tôi hoàn thiện báo cáo này hơn nữa. Xin chân thành cảm ơn!

                                                                          Nhóm sinh viên thực hiện

      Nguyễn Quốc Việt

Lê Tiến Long

Vi Tuấn Anh

# CHƯƠNG 1: TỔNG QUAN KHAI PHÁ DỮ LIỆU, TIỀN XỬ LÝ DỮ LIỆU

# 1.1. Đặt vấn đề

      Trong thời đại số hóa, việc áp dụng công nghệ thông tin vào quản lý và xử lý dữ liệu đã trở nên phổ biến, tạo ra khối lượng dữ liệu khổng lồ với tốc độ tăng trưởng chóng mặt. Điều này mở ra cánh cửa cho việc sử dụng các công cụ phân tích để chuyển đổi dữ liệu thành tri thức hữu ích, thông qua việc truy vấn, biểu diễn dữ liệu và khai phá dữ liệu.

       Khai phá dữ liệu, với nền tảng từ các ngành như toán xác suất, thống kê, và học máy, là phương pháp mạnh mẽ để khám phá kiến thức ẩn giấu trong những bộ dữ liệu lớn, nơi mà các phương pháp truyền thống không thể tiếp cận. Trong ngành y khoa, với lượng dữ liệu khổng lồ, việc áp dụng khai phá dữ liệu có thể mang lại lợi ích to lớn, cung cấp thông tin quan trọng giúp chẩn đoán và điều trị bệnh tật kịp thời, cứu sống nhiều mạng người.

        Nhóm chúng em đã chọn ứng dụng thuật toán ID3 là một thuật toán phân loại theo cách tiếp cận tham lam bằng cách chọn thuộc tính tốt nhất nhằm mang lại Information Gain (IG - lợi ích của thông tin) tối đa hoặc Entropy tối thiểu (entropy dùng để chỉ trạng thái ngẫu nhiên hoặc không có trật tự).

# 1.2. Tổng quan về khai phá dữ liệu

# 1.2.1. Khái niệm

**Khai phá dữ liệu (Data Mining)** là quá trình tính toán để tìm ra các mẫu trong các bộ dữ liệu lớn liên quan đến các phương pháp tại giao điểm của máy học, thống kê và các hệ thống cơ sở dữ liệu. Đây là một lĩnh vực liên ngành của khoa học máy tính. Mục tiêu tổng thể của quá trình khai thác dữ liệu là trích xuất thông tin từ một bộ dữ liệu và chuyển nó thành một cấu trúc dễ hiểu để sử dụng tiếp. Ngoài bước phân tích thô, nó còn liên quan tới cơ sở dữ liệu và các khía cạnh quản lý dữ liệu, xử lý dữ liệu trước, suy xét mô hình và suy luận thống kê, các thước đo thú vị, các cân nhắc phức tạp, xuất kết quả về các cấu trúc được phát hiện, hiện hình hóa và cập nhật trực tuyến. Khai thác dữ liệu là bước phân tích của quá trình "khám phá kiến thức trong cơ sở dữ liệu" hoặc KDD. Khai phá dữ liệu là một bước của quá trình khai thác tri thức (Knowledge Discovery Process), bao gồm:

 • Xác định vấn đề và không gian dữ liệu để giải quyết vấn đề (Problem understanding and data understanding).

 • Chuẩn bị dữ liệu (Data preparation), bao gồm các quá trình làm sạch dữ liệu (data cleaning), tích hợp dữ liệu (data integration), chọn dữ liệu (data selection), biến đổi dữ liệu (data transformation).

 • Khai thác dữ liệu (Data mining): xác định nhiệm vụ khai thác dữ liệu và lựa chọn kỹ thuật khai thác dữ liệu. Kết quả cho ta một nguồn tri thức thô.

 • Đánh giá (Evaluation): dựa trên một số tiêu chí tiến hành kiểm tra và lọc nguồn tri thức thu được.

 • Triển khai (Deployment).

# 1.2.2 Quy trình khai phá tri thức

- Tìm hiểu lĩnh vực của bài toán (ứng dụng): Các mục đích của bài toán, các tri thức cụ thể của lĩnh vực.

- Tạo nên (thu thập) một tập dữ liệu phù hợp.

- Làm sạch và tiền xử lý dữ liệu.

- Giảm kích thước của dữ liệu, chuyển đổi dữ liệu: Xác định thuộc tính quan trọng, giảm số chiều (số thuộc tính), biểu diễn bất biến.

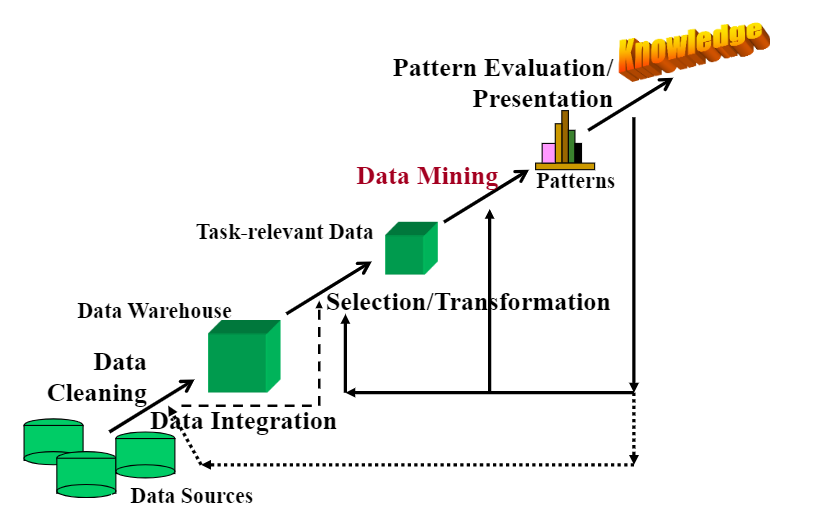
- Lựa chọn chức năng khai phá dữ liệu: Phân loại, gom cụm, dự báo, sinh ra các luật kết hợp.

- Lựa chọn/ Phát triển (các) giải thuật khai phá dữ liệu phù hợp.

- Tiến hành khai phá dữ liệu.

- Đánh giá mẫu thu được và biểu diễn tri thức: Hiển thị hóa, chuyển đổi, bỏ đi các mẫu dư thừa …

- Sử dụng tri thức được khai phá.



Hình 1. 1 Hình minh họa khai phá

**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)

- Knowledge presentation (biểu diễn tri thức)

# 1.2.3 Lợi ích của khai phá dữ liệu

**Lợi ích trong quá trình phân tích dữ liệu**

- Chọn lọc, loại bỏ tất cả các dữ liệu không liên quan và dữ liệu bị trùng

- Xác định các mẫu dữ liệu, dữ liệu có liên quan và dữ liệu trùng lặp trong tập dữ liệu

- Với data mining, chúng ta có thể phân tích khối lượng lớn dữ liệu trong thời gian ngắn và sau đó chuyển đổi dữ liệu đó thành thông tin, kiến thức

**Lợi ích sau cùng của Data mining**

- Hỗ trợ ra quyết định tự động

- Hỗ trợ đưa ra dự báo chính xác

- Hỗ trợ giảm thiểu chi phí

- Hỗ trợ khả năng thấu hiểu khách hàng

# 1.2.4 Ứng dụng của khai phá dữ liệu

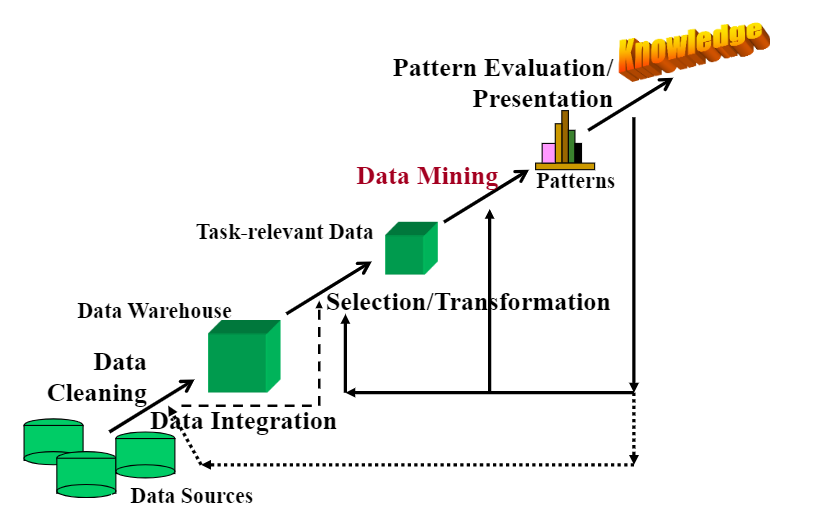
Kinh tế -ứng dụng trong kinh doanh, tài chính, tiếp thị bán hàng, bảo hiểm, thương mại, ngân hàng, … Đưa ra các bản báo cáo giàu thông tin; phân tích rủi ro trước khi đưa ra các chiến lược kinh doanh, sản xuất; phân loại khách hàng từ đó phân định thị trường, thị phần; …

Khoa học: Thiên văn học –  dự đoán đường đi các thiên thể, hành tinh, …

Công nghệ sinh học –  tìm ra các gen mới, cây con giống mới…

Web: các công cụ tìm kiếm

# 1.3. Tiền xử lý dữ liệu



Hình 1. 2 : Tiền sử lý dữ liệu

# 1.3.1 Khái niệm

Quá trình xử lý dữ liệu thô/gốc (raw/original data) nhằm cải thiện chất lượng dữ liệu (quality of the data) và từ đó cải thiện chất lượng của kết quả khai phá.

**Chất lượng dữ liệu (data quality)**

- Tính chính xác (accuracy): giá trị được ghi nhận đúng với giá trị thực.

- Tính hiện hành (currency/timeliness): giá trị được ghi nhận không bị lỗi thời.

- Tính toàn vẹn (completeness): tất cả các giá trị dành cho một biến/thuộc tính đều được ghi nhận.

- Tính nhất quán (consistency): tất cả giá trị dữ liệu đều được biểu diễn như nhau trong tất cả các trường hợp.

**Các kỹ thuật tiền xử lý dữ liệu**

- Làm sạch dữ liệu (data cleaning/cleansing)

- Tích hợp dữ liệu (data integration)

- Biến đổi dữ liệu (data transformation)

- Thu giảm dữ liệu (data reduction)

# 1.3.2 Làm sạch dữ liệu (data cleaning)

Data cleaning hay làm sạch dữ liệu là quy trình chuẩn bị dữ liệu trước khi phân tích thông qua xử lý hay loại bỏ những dữ liệu không chính xác, không đầy đủ, không phù hợp về định dạng, bị trùng lặp, không có giá trị, không đủ thông tin, không liên quan, …những dữ liệu có thể ảnh hưởng đến kết quả phân tích sau cùng.

# 1.3.3 Tích hợp dữ liệu (data integration)

Là quá trình trộn dữ liệu từ các nguồn khác nhau vào một kho dữ liệu sẵn sàng cho quá trình khai phá dữ liệu

# 1.3.4 Biến đổi dữ liệu (data transformation)

Là quá trình biến đổi hay kết hợp dữ liệu vào những dạng thích hợp cho quá trình khai phá dữ liệu

Làm trơn dữ liệu (smoothing)

Kết hợp dữ liệu (aggregation)

Tổng quát hoá (generalization)

Xây dựng thuộc tính/đặc tính (attribute/feature construction)

Chuẩn hoá (normalization)

# 1.3.5 Thu giảm dữ liệu (data reduction)

Tập dữ liệu được biến đổi đảm bảo các toàn vẹn, nhưng nhỏ/ít hơn nhiều về số lượng so với ban đầu.

**Các chiến lược thu giảm:**

Kết hợp khối dữ liệu (data cube aggregation)

Chọn một số thuộc tính (attribute subset selection)

Thu giảm chiều (dimensionality reduction)

Thu giảm lượng (numerosity reduction)

Rời rạc hóa (discretization)

Tạo phân cấp ý niệm (concept hierarchy generation)

# CHƯƠNG 2: KHÁI QUÁT CÁC THUẬT TOÁN ĐƯỢC SỬ DỤNG ĐỂ KHAI PHÁ DỮ LIỆU

# 2.1 Phân lớp dữ liệu

# 2.1.1 Khái niệm

Phân lớp dữ liệu là quá trình tổ chức và phân nhóm các mẫu dữ liệu vào các nhóm hoặc lớp dựa trên các đặc điểm chung của chúng. Mục tiêu của việc phân lớp dữ liệu là tạo ra các nhóm có tính chất tương đồng bên trong và khác biệt với các nhóm khác. Quá trình này thường được sử dụng trong các ứng dụng như học máy, thị trường, và Khai phá dữ liệu

Trong phân lớp dữ liệu, một mô hình hoặc thuật toán được sử dụng để xác định lớp hoặc nhóm mà một mẫu dữ liệu thuộc về. Các phương pháp phân loại dữ liệu phổ biến bao gồm cây quyết định, máy vector hỗ trợ (SVM), mạng nơ-ron nhân tạo, ID3 và các phương pháp học sâu khác.

Mục tiêu của phân lớp dữ liệu là tạo ra một mô hình hoặc hệ thống có khả năng phân loại các mẫu dữ liệu mới một cách chính xác và hiệu quả. Điều này có thể hỗ trợ trong việc ra quyết định, dự đoán kết quả, hoặc hiểu được cấu trúc của dữ liệu.

Quá trình gồm hai bước:

Bước học (giai đoạn huấn luyện): xây dựng bộ phân lớp (classifier) bằng việc phân tích/học tập huấn luyện

Bước phân lớp (classification): phân lớp dữ liệu/đối tượng mới nếu độ chính xác của bộ phân lớp được đánh giá là có thể chấp nhận được (acceptable)

# 2.1.2 Các thuật toán phân lớp dữ liệu

a) Phân lớp với cây quyết định (decision tree): Dựa trên việc phân chia không gian dữ liệu thành các phần con dựa trên các thuộc tính, cây quyết định là một phương pháp dễ hiểu và dễ giải thích. ID3, C4.5, CART là các thuật toán phổ biến trong loại này.

b) Phân lớp với Naive Bayesian: Dựa trên định lý Bayes, Naive Bayes giả định rằng các thuộc tính độc lập với nhau, và được sử dụng nhiều trong các bài toán phân loại văn bản và email spam.

c) Phân lớp với k phần tử gần nhất (k-nearest neighbor): Dựa trên việc gán nhãn cho một điểm dữ liệu mới dựa trên các điểm dữ liệu gần nhất trong tập huấn luyện, KNN là một phương pháp đơn giản nhưng hiệu quả.

d) Phân loại dữ liệu với Random Forest: Là một phương pháp kết hợp nhiều cây quyết định để tạo ra một mô hình mạnh mẽ hơn, Random Forest giảm thiểu nguy cơ overfitting và cung cấp một dự đoán ổn định hơn.

e) Thuật toán ID3 (Iterative Dichotomiser 3): ID3là một thuật toán phân lớp dữ liệu cơ bản được sử dụng để xây dựng cây quyết định. Là một công cụ hữu ích trong việc giải quyết các bài toán phân loại dữ liệu đơn giản và dễ hiểu.

# 2.2 Khai phá, kết hợp

# 2.2.1 Khái Niệm.

Khai phá kết hợp là một phương pháp trong lĩnh vực khai phá dữ liệu nhằm tìm ra các mối quan hệ tương quan giữa các mục dữ liệu trong một tập dữ liệu lớn. Mục tiêu của khai phá kết hợp là tìm ra các luật kết hợp (association rules) giữa các mục dữ liệu, từ đó giúp hiểu và dự đoán các mẫu xu hướng hoặc quy luật tiềm ẩn trong dữ liệu.

# 2.2.2 Các thuật toán khai phá kết hợp

a) Apriori Algorithm: Là một thuật toán cơ bản trong khai phá kết hợp, Apriori sử dụng kỹ thuật quét dữ liệu theo chiều rộng (breadth-first) để tìm các tập phổ biến của mục. Thuật toán này dựa trên nguyên tắc apriori để giảm bớt không gian tìm kiếm và tạo ra các luật kết hợp.

b) FP-Growth Algorithm: FP-Growth là một thuật toán hiệu quả để tìm các mẫu phổ biến trong dữ liệu. Thay vì tạo ra các tập phổ biến như Apriori, FP-Growth sử dụng cấu trúc dữ liệu cây tần suất (FP-tree) để đại diện cho dữ liệu và tìm kiếm các mẫu phổ biến dựa trên cấu trúc cây.

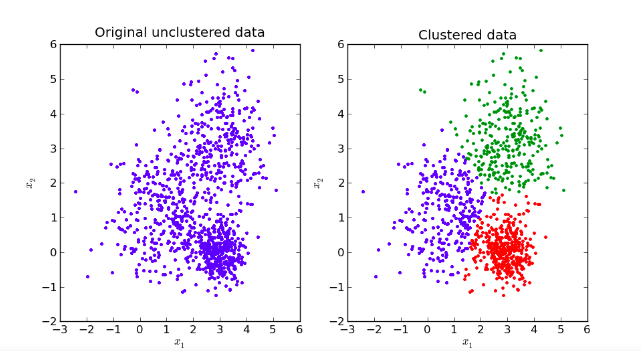
c) Eclat Algorithm: Eclat là một thuật toán khai phá kết hợp khác dựa trên kỹ thuật tối ưu hóa dựa trên hậu duệ (post-pruning). Thuật toán này tập trung vào việc tìm kiếm các tập phổ biến thông qua việc sử dụng các cấu trúc dữ liệu hiệu quả như bảng duyệt (tidset).

d) Apriori-Hybrid Algorithms: Các biến thể của thuật toán Apriori kết hợp với các kỹ thuật khác như lọc đánh giá (pruning), gom nhóm (clustering), hoặc sắp xếp lại dữ liệu (reordering) để tối ưu hóa hiệu suất và giảm thời gian tính toán.

# 2.3 Phân cụm

# 2.3.1 Khái niệm

Phân cụm là quá trình tổ chức dữ liệu thành các nhóm (cụm) dựa trên sự tương đồng giữa các mẫu dữ liệu. Mục tiêu chính của phân cụm là tạo ra các nhóm mà trong đó các mẫu dữ liệu có tính chất tương đồng cao, đồng thời các nhóm này phải khác biệt rõ rệt với nhau. Quá trình này giúp ta hiểu rõ hơn về cấu trúc và tổ chức của dữ liệu, khám phá những thông tin tiềm ẩn, nhận diện các mẫu và xu hướng, hoặc tạo ra các phân đoạn dữ liệu để dễ dàng phân tích và xử lý.



Hình 2. 1 : Phân cụm

**Các bước chính trong quá trình phân cụm bao gồm:**

* **Chuẩn bị dữ liệu:** Tiền xử lý dữ liệu để chuẩn bị và làm sạch dữ liệu trước khi phân cụm.
* **Chọn đặc trưng và độ đo tương đồng**: Xác định các đặc trưng quan trọng để đo lường sự tương đồng giữa các mẫu dữ liệu, và chọn các độ đo tương đồng phù hợp để so sánh các mẫu.
* **Chọn thuật toán phân cụm:** Chọn và áp dụng một thuật toán phân cụm phù hợp với bản chất của dữ liệu và mục tiêu của bài toán.
* **Phân cụm dữ liệu:** Áp dụng thuật toán đã chọn để phân cụm dữ liệu thành các nhóm dựa trên sự tương đồng giữa chúng.
* **Đánh giá kết quả:** Đánh giá tính hợp lý và hiệu quả của các nhóm được tạo ra bằng cách sử dụng các chỉ số đánh giá và kiểm tra thực tế.
* **Hiểu và sử dụng kết quả:** Hiểu ý nghĩa của các nhóm được tạo ra và sử dụng chúng để đưa ra quyết định hoặc thực hiện các chiến lược phù hợp.

# 2.3.2. Các thuật toán phân cụm

1. Phân hoạch(partitioning): các phân hoạch được tạo ra và đánh giá theo một tiêu chí nào đó.
2. Phân cấp(hierarchical): phân rã tập dữ liệu/đối tượng có thứ tự phân cấp theo một tiêu chí nào đó.
3. Dựa trên cấp độ (density-based): dựa trên connectivity and density functions
4. Dựa trên lưới (grid-based): dựa trên multiple-level granularity structure
5. Dựa trên mô hình (model-based): một mô hình giả thuyết được tạo ra cho mỗi cụm; sau đó hiệu chỉnh các thông số để mô hình phù hợp với cụm dữ liệu/đối tượng nhất……

# 2.4 Sử dụng thuật toán phân lớp ID3 để dự đoán bệnh viêm gan

# 2.4.1 Khái niệm thuật toán phân lớp ID3

Thuật toán ID3 (Iterative Dichotomiser 3) là một thuật toán trong lĩnh vực học máy và khai phá dữ liệu, được sử dụng để xây dựng cây quyết định từ tập dữ liệu huấn luyện. Thuật toán này được phát triển bởi Ross Quinlan vào năm 1986 và đã trở thành một trong những thuật toán phổ biến nhất để xây dựng cây quyết định trong lĩnh vực học máy.

Ý tưởng cơ bản của thuật toán ID3 là chọn ra thuộc tính tốt nhất để phân chia tập dữ liệu ở mỗi nút của cây quyết định. Thuật toán sử dụng độ đo thông tin entropy để đánh giá mức độ không chắc chắn (uncertainty) trong tập dữ liệu. Mục tiêu là chọn ra thuộc tính sao cho việc phân chia dữ liệu dựa trên thuộc tính này sẽ giảm thiểu độ không chắc chắn nhất có thể.

**Các bước chính của thuật toán ID3 bao gồm:**

**Bước 1: Tính toán Entropy và Information Gain**

* Entropy (Độ đo Entropy): Entropy đo lường độ không chắc chắn trong tập dữ liệu. Nó được tính bằng công thức:

Entropy(*S*)=−∑*i*=1*c*​*pi*​log2​(*pi*​)

Trong đó:

* + *S* là tập dữ liệu.
  + *c* là số lớp trong tập dữ liệu.
  + *pi*​ là tỷ lệ mẫu thuộc lớp *i* trong tập dữ liệu.
* Information Gain (Lợi nhuận thông tin): Là sự giảm entropy khi tập dữ liệu được phân chia dựa trên một thuộc tính cụ thể. Information Gain được tính bằng cách lấy entropy trước khi phân chia trừ đi entropy sau khi phân chia.

**Bước 2: Chọn thuộc tính tốt nhất:**

* ID3 chọn thuộc tính có Information Gain cao nhất làm thuộc tính phân chia tốt nhất. Điều này có nghĩa là thuộc tính này tạo ra phân chia dữ liệu mà giảm entropy nhiều nhất.

**Bước 3: Tạo cây quyết định**

* Sau khi chọn thuộc tính phân chia, ID3 tạo một nút cây quyết định dựa trên thuộc tính này.
* Tập dữ liệu được phân chia thành các nhánh dựa trên giá trị của thuộc tính đã chọn.

Bước 4: Lặp lại quá trình

* Quá trình trên được lặp lại cho mỗi nhánh mới trong cây, cho đến khi mỗi nhánh chỉ chứa các mẫu thuộc vào cùng một lớp (độ thuần khiết đạt đến mức cao nhất có thể) hoặc khi đạt đến một điều kiện dừng nhất định.

**Bước 5: Điều chỉnh và cắt tỉa cây**

* Để giảm thiểu overfitting, các kỹ thuật cắt tỉa (pruning) có thể được áp dụng để loại bỏ các nhánh không cần thiết của cây.

**Bước 6: Sử dụng cây quyết định**

* Khi cây quyết định được xây dựng, nó có thể được sử dụng để dự đoán lớp của các mẫu dữ liệu mới bằng cách đi xuống từ nút gốc đến lá tương ứng với giá trị của thuộc tính trong mẫu dữ liệu đó.

# 2.4.2 Ưu điểm của thuật toán ID3

* **Dễ hiểu và thực hiện**: Thuật toán ID3 dễ hiểu và thực hiện, không yêu cầu kiến thức toán học phức tạp hay kỹ năng lập trình cao. Điều này làm cho nó trở thành một lựa chọn phổ biến cho người mới bắt đầu trong lĩnh vực học máy và khai phá dữ liệu.
* **Hiệu suất tốt cho dữ liệu nhỏ hoặc có cấu trúc tốt**: ID3 thường cho kết quả tốt cho các tập dữ liệu có kích thước nhỏ hoặc dữ liệu có cấu trúc tốt, trong đó các mẫu dữ liệu tương tự được nhóm lại vào các nhóm nhỏ.
* **Hiệu suất tốt cho dữ liệu nhỏ hoặc có cấu trúc tốt**: ID3 thường cho kết quả tốt cho các tập dữ liệu có kích thước nhỏ hoặc dữ liệu có cấu trúc tốt, trong đó các mẫu dữ liệu tương tự được nhóm lại vào các nhóm nhỏ.

# 2.4.3 Nhược điểm của thuật toán ID3

* **Dễ bị overfitting**: ID3 có xu hướng dễ bị overfitting, đặc biệt là khi cây quyết định quá phức tạp hoặc khi dữ liệu có nhiều nhiễu. Overfitting có thể dẫn đến hiệu suất kém trên dữ liệu mới mà thuật toán chưa từng thấy.
* **Không xử lý được dữ liệu liên tục:** ID3 không xử lý trực tiếp các thuộc tính liên tục. Điều này có thể dẫn đến việc phân chia không hiệu quả và làm giảm hiệu suất của cây quyết định trên dữ liệu liên tục.
* **Khả năng xử lý thuộc tính có giá trị thiếu:** ID3 không xử lý được các giá trị thiếu trong dữ liệu một cách tự động. Nếu dữ liệu có giá trị thiếu, cần phải thực hiện các biện pháp xử lý trước khi áp dụng thuật toán.
* **Khả năng chịu nhiễu thấp:** ID3 không hiệu quả trên các tập dữ liệu chứa nhiều nhiễu hoặc không đồng nhất, do độ không chắc chắn cao có thể làm cho quá trình phân loại trở nên không chính xác.

Mặc dù có nhược điểm nhất định, ID3 vẫn là một công cụ hữu ích trong việc xây dựng cây quyết định đơn giản và hiểu được cấu trúc của dữ liệu. Đối với các vấn đề cụ thể, việc kết hợp ID3 với các kỹ thuật điều chỉnh hoặc sử dụng các biến thể có thể cải thiện hiệu suất của thuật toán.

# 2.4.4 Sử dụng thuận toán ID3 để dự đoán bệnh viêm gan

- Để sử dụng thuật toán phân lớp ID3 vào dự đoán bệnh viêm gan ta cần thức hiện qua các bước sau:

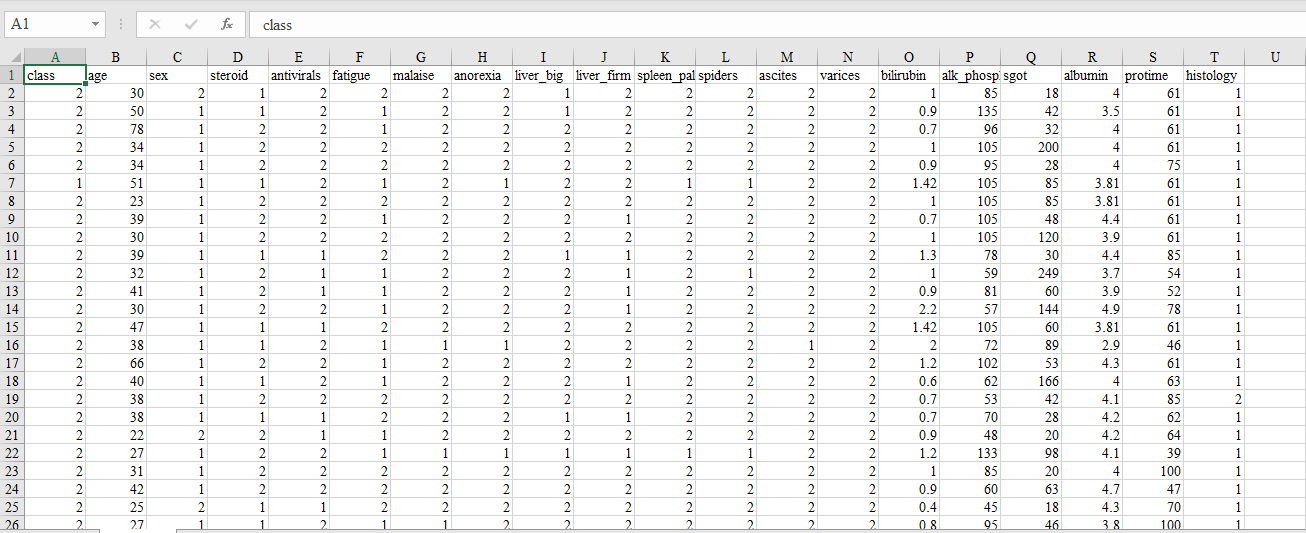
* **Chuẩn bị dữ liệu:** Thu thập dữ liệu về các yếu tố liên quan đến bệnh viêm gan như tuổi, giới tính, dấu hiệu và triệu chứng, kết quả xét nghiệm máu, tiểu cầu, và các chỉ số chức năng gan.
* **Tiền xử lý dữ liệu:** Xử lý dữ liệu để chuẩn bị cho việc sử dụng thuật toán ID3. Điều này có thể bao gồm loại bỏ các giá trị thiếu, chuyển đổi các biến phân loại thành dạng số, và chuẩn hóa dữ liệu nếu cần.
* **Xây dựng cây quyết định với ID3:** Sử dụng thuật toán ID3 để xây dựng cây quyết định từ tập dữ liệu huấn luyện. Chọn các thuộc tính quan trọng nhất để phân chia dữ liệu và dự đoán kết quả bệnh viêm gan.
* **Đánh giá mô hình:** Sử dụng tập dữ liệu kiểm tra để đánh giá hiệu suất của mô hình cây quyết định. Đo lường các chỉ số như độ chính xác, độ nhạy, độ chính xác dự đoán âm tính, và độ chính xác dự đoán dương tính.
* **Tối ưu hóa và điều chỉnh**: Kiểm tra và điều chỉnh các tham số của cây quyết định để cải thiện hiệu suất của mô hình. Các kỹ thuật như cắt tỉa cây, giảm overfitting, và tăng cường chất lượng của cây có thể được áp dụng.
* **Tối ưu hóa và điều chỉnh:** Kiểm tra và điều chỉnh các tham số của cây quyết định để cải thiện hiệu suất của mô hình. Các kỹ thuật như cắt tỉa cây, giảm overfitting, và tăng cường chất lượng của cây có thể được áp dụng.

Quan trọng nhất, việc sử dụng thuật toán ID3 trong dự đoán bệnh viêm gan cần được thực hiện trong một bối cảnh lâm sàng có sự hỗ trợ của các chuyên gia y tế và được kiểm tra kỹ lưỡng để đảm bảo tính chính xác và độ tin cậy của kết quả

# CHƯƠNG 3 THỬ NGHIỆM, KẾT QUẢ

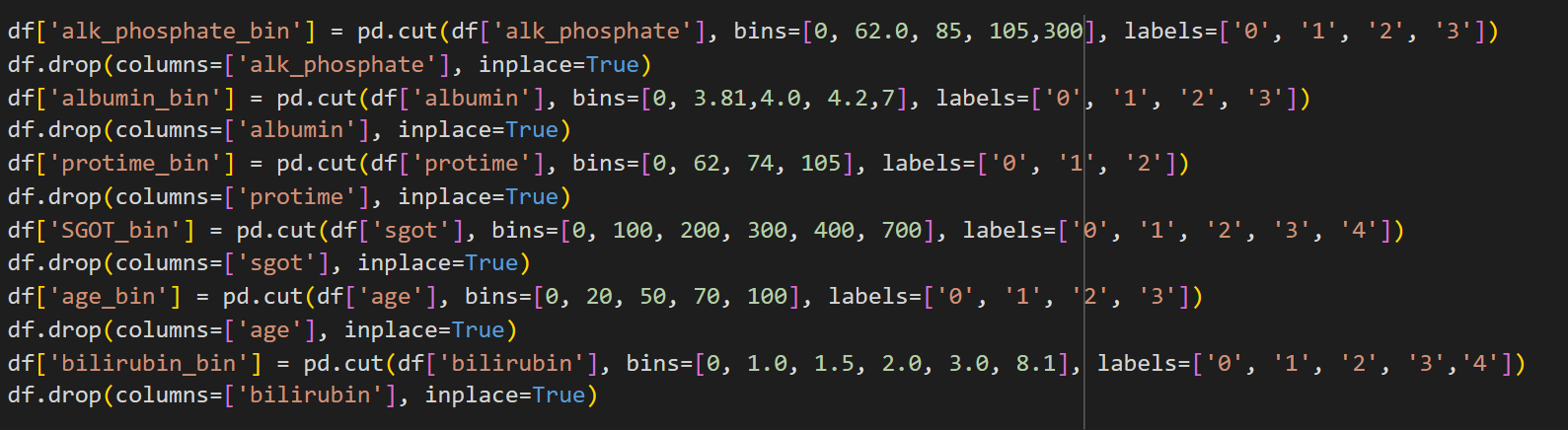
**Bước 1. Tải file dữ liệu về**

**Dữ liệu bệnh viêm gan**



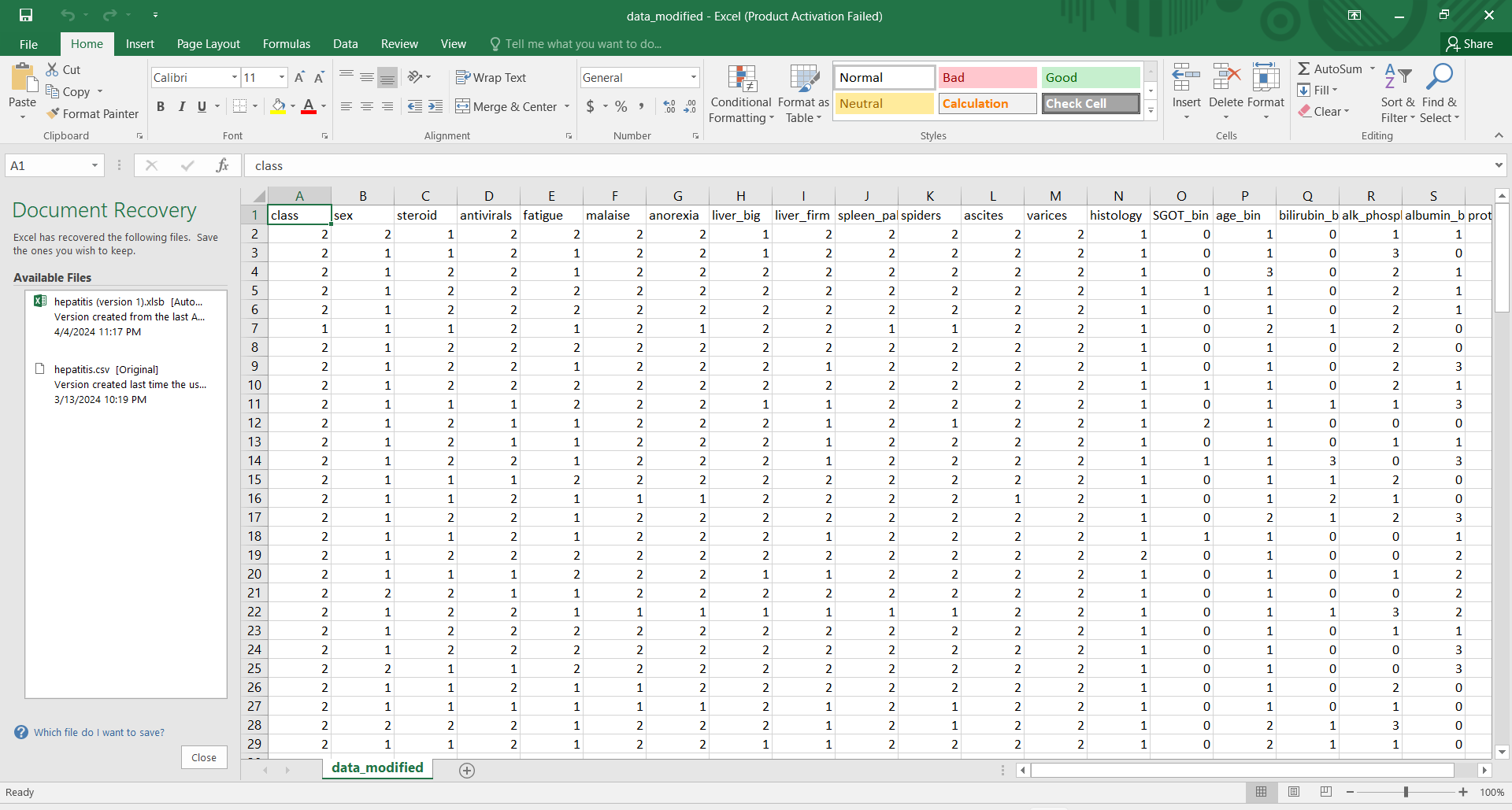
Hình 3. 1 : Dữ liệu viêm gan

**Bước 2. Thực hiện bước tiền xử lý dữ liệu (data reduction**) **khi gộp nhóm các trường dữ liệu có quá nhiều giá trị khác nhau như age, bilirubin, …**

****

Hình 3. 2 : Tiền xử lý dữ liệu

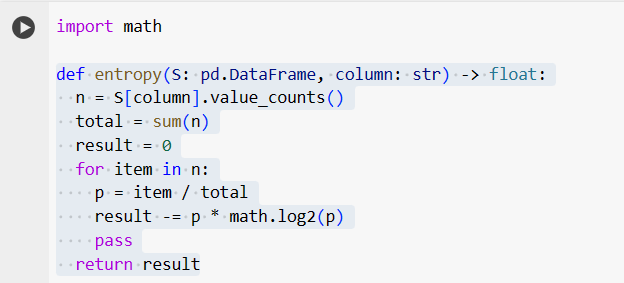
**Bước 3. Sau khi đã thu giảm dữ liệu ta sẽ có bảng dữ liệu mới như sau**

****

Hình 3. 3 : Sau khi thu giảm dữ liệu

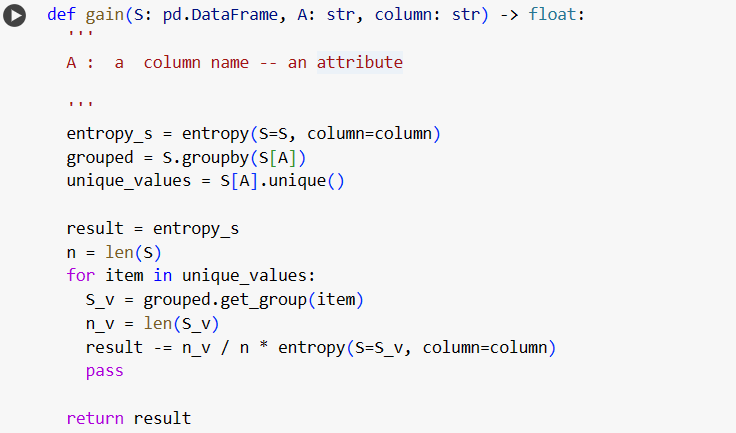
**Bước 4. Tính Information\_gain dựa vào entropy của các thuộc tính trong bảng**

**Bước 4.1. Entropy**

****

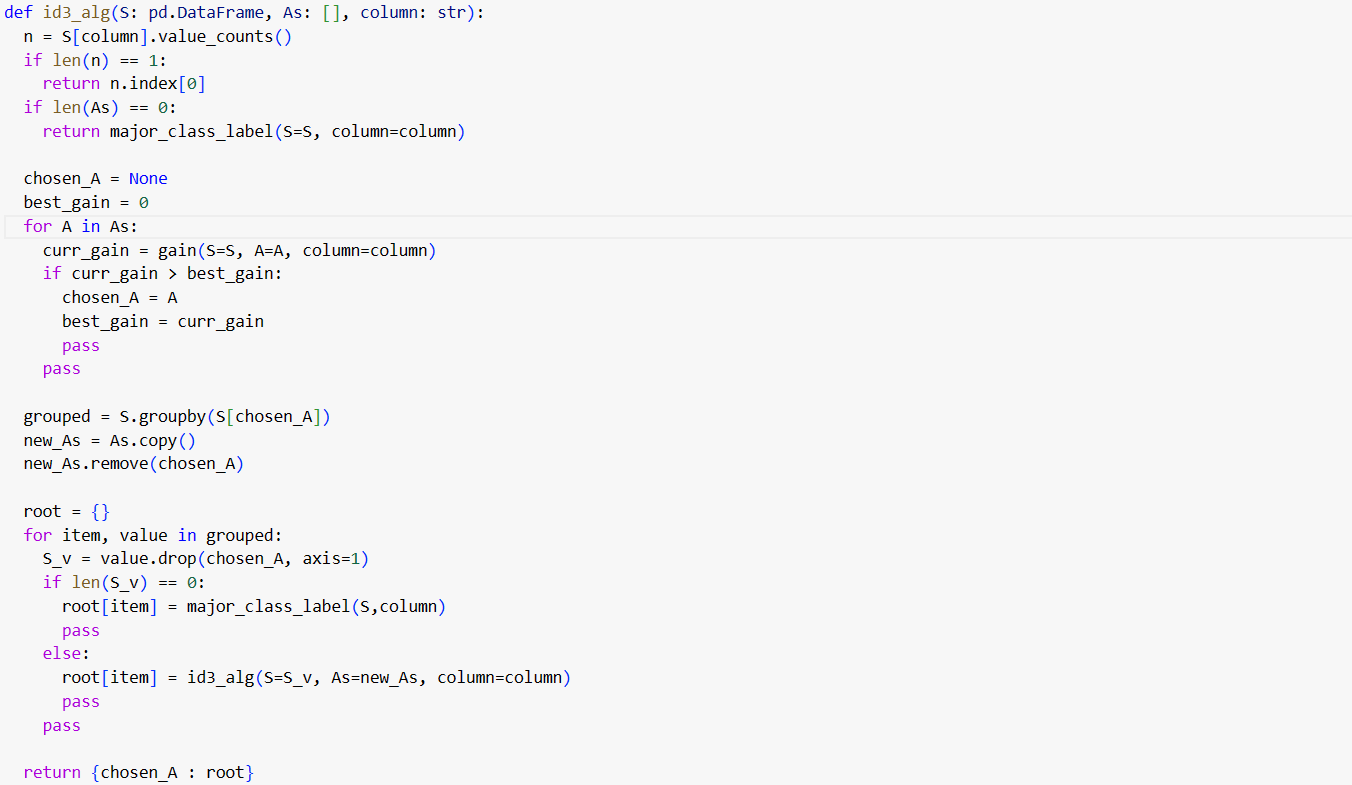
Hình 3. 4 : Entropy

**Bước 4.2. Information\_gain**

****

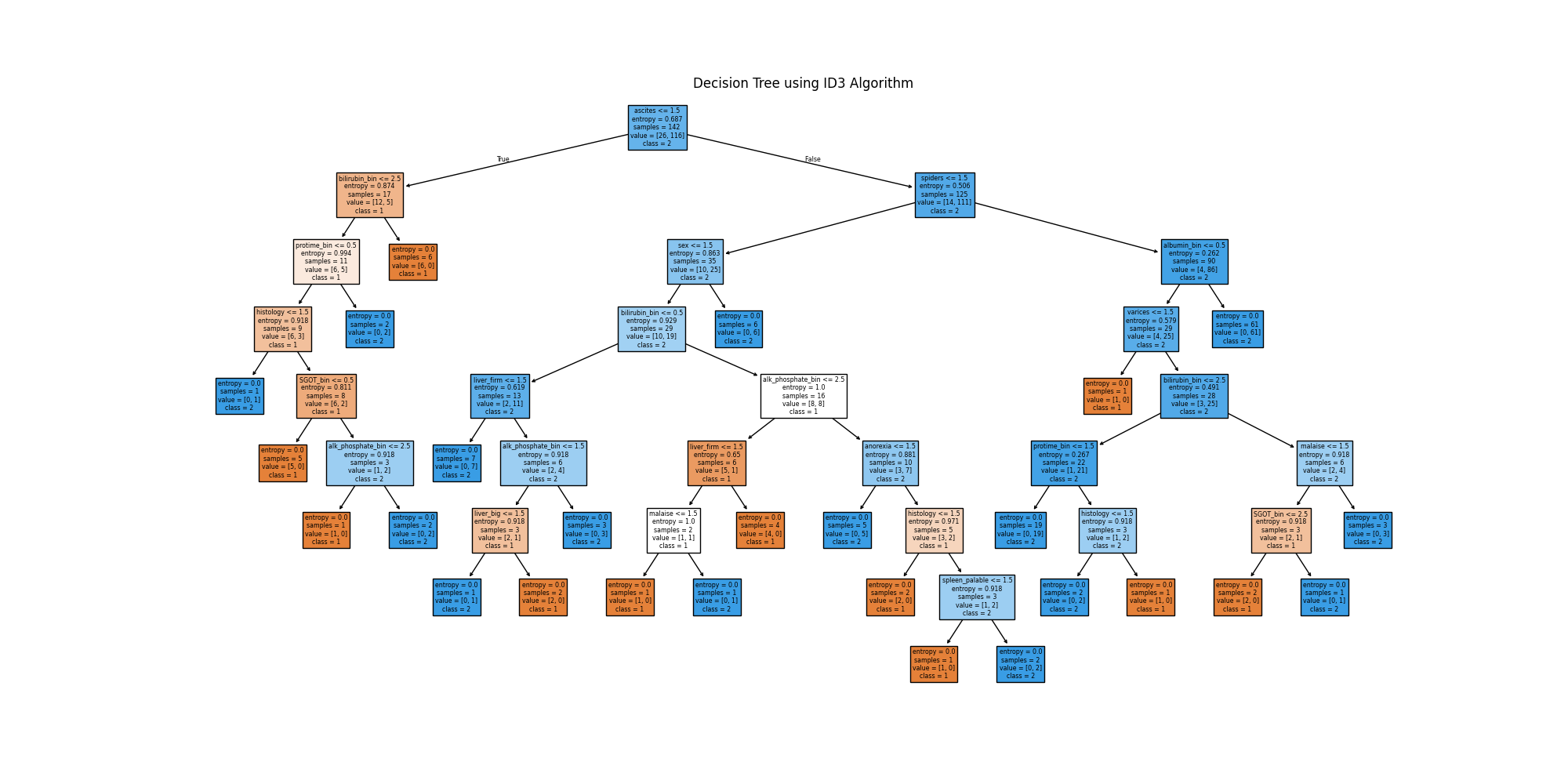
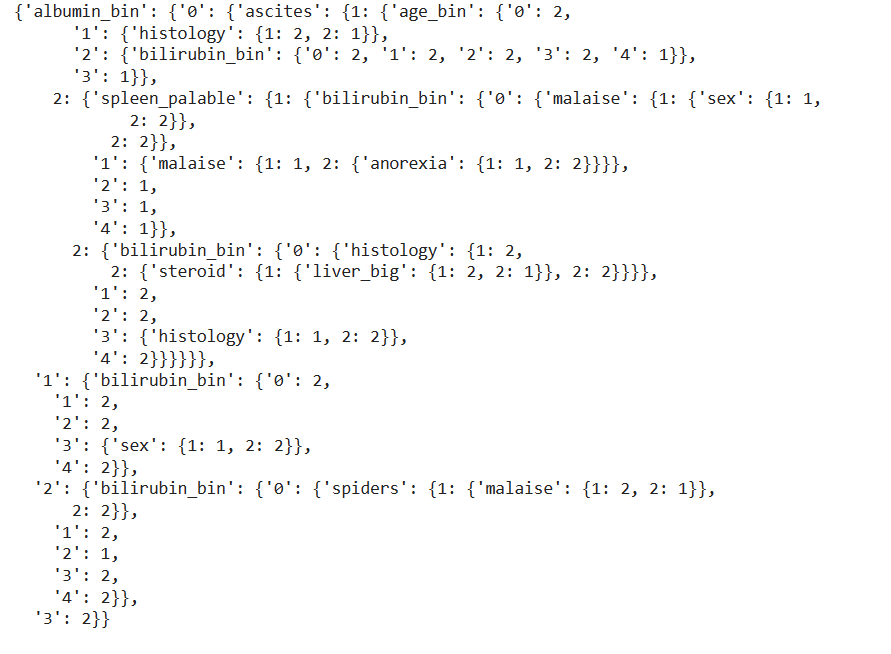
Hình 3. 5 : Information\_ gain

**Bước 5. Chạy thuật toán ID3**

****

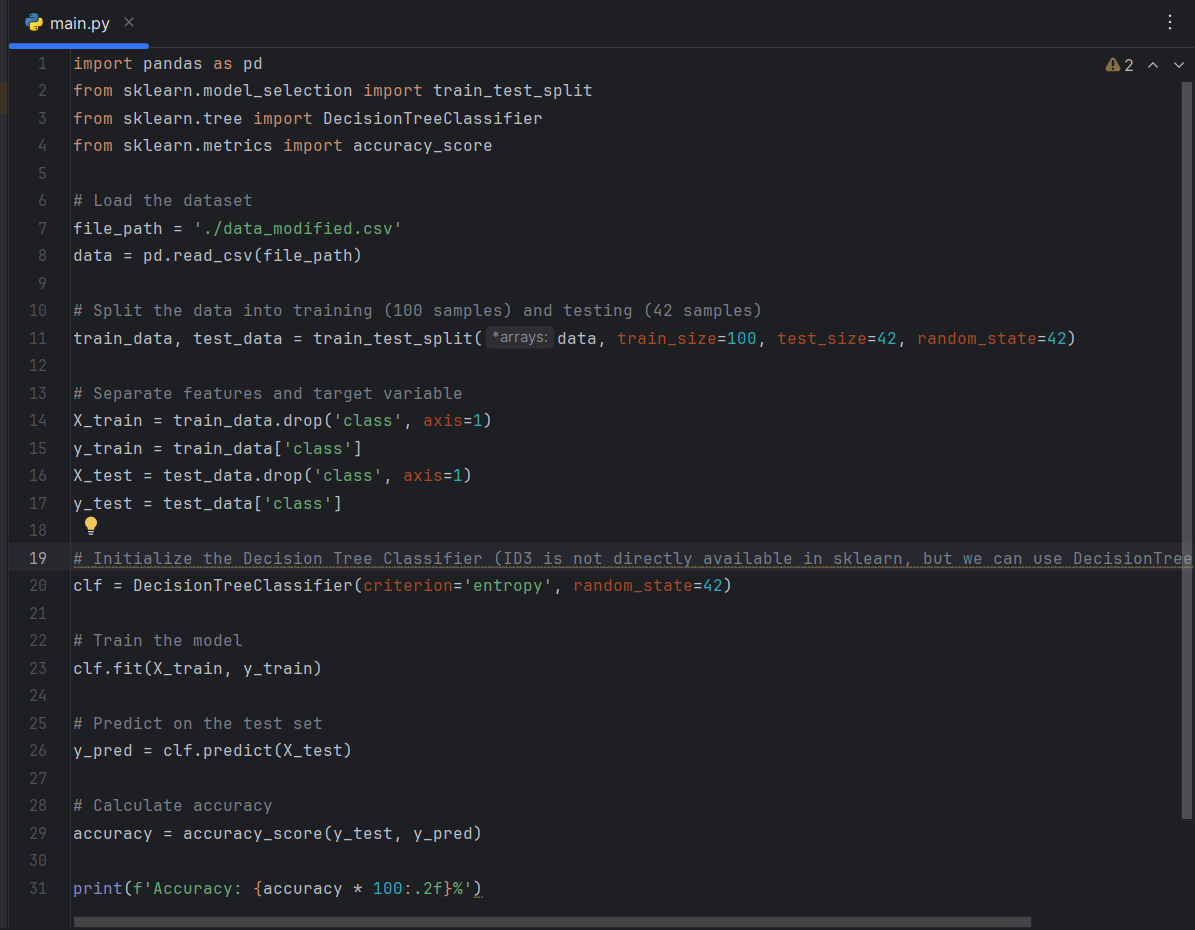
Hình 3. 6 : Chạy thuật toán ID3

**Bước 6. Kết quả thực nghiệm**

****

Hình 3. 7 : Kết quả thực nghiệm

**Bước 7. Thực hiện test độ chính xác của thuật toán với 100 bản train và 42 bản test**

****

Hình 3. 8: Test chính xác

**Kết quả ta ra được độ chính xác của thuật toán là 80,5%**

# KẾT LUẬN

Trong nghiên cứu này, chúng em đã ứng dụng thuật toán phân lớp ID3 để chuẩn đoán bệnh viêm gan, nhằm mục đích kiểm tra khả năng của thuật toán trong việc phân loại các trường hợp mắc và không mắc bệnh dựa trên dữ liệu y học. Kết quả thu được đã chứng minh tính hiệu quả của thuật toán ID3 trong lĩnh vực này, đồng thời cũng làm nổi bật một số hạn chế cần được khắc phục để nâng cao độ chính xác và độ tin cậy của mô hình.

* Kết quả đạt được:
* Độ chính xác cao: Thuật toán ID3 đã đạt được độ chính xác cao trong việc phân loại các trường hợp mắc và không mắc bệnh viêm gan
* Hiệu quả trong xử lý dữ liệu phức tạp: ID3 đã chứng tỏ khả năng xử lý và phân tích một lượng lớn dữ liệu y học phức tạp, từ các kết quả xét nghiệm đến các triệu chứng lâm sàng.
* Hỗ trợ quyết định cho bác sĩ: Cây quyết định được tạo ra từ thuật toán ID3 cung cấp một công cụ trực quan và dễ hiểu cho các bác sĩ, hỗ trợ họ trong việc đưa ra các quyết định chẩn đoán dựa trên dữ liệu thực tế.
* Hạn chế
* Overfitting: Một trong những hạn chế chính của thuật toán ID3 là xu hướng bị overfitting, đặc biệt khi dữ liệu huấn luyện không đủ đa dạng hoặc bị nhiễu. Overfitting khiến mô hình hoạt động kém hiệu quả trên dữ liệu mới, chưa từng gặp.
* Độ phức tạp của cây quyết định: Cây quyết định được tạo ra từ ID3 có thể trở nên rất phức tạp và khó hiểu nếu có nhiều thuộc tính và giá trị thuộc tính. Điều này có thể làm giảm tính thực tiễn của mô hình trong việc sử dụng hàng ngày.
* Yêu cầu về dữ liệu: ID3 yêu cầu dữ liệu đầy đủ và chính xác. Nếu dữ liệu bị thiếu hoặc không chính xác, hiệu suất của mô hình sẽ bị ảnh hưởng đáng kể.
* Tổng kết lại, việc sử dụng thuật toán phân lớp ID3 trong chẩn đoán bệnh viêm gan đã mang lại những kết quả rất tích cực, với khả năng phân loại chính xác và hỗ trợ hiệu quả cho các bác sĩ trong quá trình chẩn đoán. Tuy nhiên, một số hạn chế như overfitting và độ phức tạp của cây quyết định cần được khắc phục thông qua các phương pháp cải tiến như cắt tỉa cây quyết định (pruning) hoặc kết hợp với các thuật toán khác. Với những cải tiến này, ID3 có thể trở thành một công cụ mạnh mẽ và hữu ích hơn nữa trong lĩnh vực y tế, đặc biệt là trong việc chẩn đoán và điều trị bệnh viêm gan.

# TÀI LIỆU THAM KHẢO

[1] Han, J., Pei, J., & Kamber, M. (2011). Data Mining: Concepts and Techniques. Morgan Kaufmann.

[2] Witten, I. H., Frank, E., Hall, M. A., & Pal, C. J. (2016). Data Mining: Practical Machine Learning Tools and Techniques. Morgan Kaufmann.

[3] Tan, P.-N., Steinbach, M., & Kumar, V. (2005). Introduction to Data Mining. Addison-Wesley.

[4] Aggarwal, C. C. (2015). Data Mining: The Textbook. Springer.