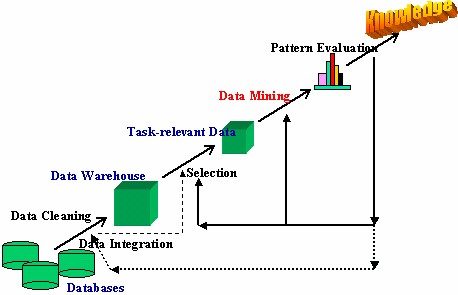
1. **Vẽ quy trình khai thác dữ liệu, mỗi bước giải thích ?**



Quy trình gồm 7 bước :

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.
8. **Phân loại là gì ?**

Là phương pháp dự báo, cho phép phân loại 1 đối tượng vào 1 hoặc 1 số lớp cho trước.

1. **Quy trình phân lớp ?**

Là quá trình phân lớp 1 đối tượng dữ liệu vào 1 hay nhiều lớp đã cho trước nhờ 1 mô hình phân lớp (model). Mô hình này được xây dựng dựa trên một tập dữ liệu được xây dựng trước đó có gán nhãn (hay còn gọi là tập huấn luyện). Quá trình phân lớp là quá trình gán nhãn cho đối tượng dữ liệu.

1. **Nêu 1 số ứng dụng liên quan đến 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
* Tin-sinh (bio-informatics)
* Tài chính và thị trường chứng khoán (finance & stock market)
* Bảo hiểm (insurance) • Nhận dạng (pattern recognition)
* .v.v.

1. **Làm rõ các khái niệm thuật ngữ**

* Classification ( phân loại ) : Là phương pháp dự báo, cho phép phân loại 1 đối tượng vào 1 hoặc 1 số lớp cho trước.
* Feature (đặc tính ) :
* Instance / simple :
* Training set : Là 1 tập dữ liệu có kích thước lớn, được dùng để training trong quá trình huấn luyện máy học.Nói dễ hiểu thì đây chính là tập dữ liệu máy dùng để học và rút trích được những đặc điểm quan trọng để ghi nhớ lại.Tập training set gồm 2 phần là Input và Output.
* Testing set : Là tập dữ liệu dùng để test sau khi máy đã học xong.Một mô hình máy học sau khi đã huấn luyện, sẽ cần phải được kiểm chứng xem nó có hiệu quả không.Khác với training set thì testing set chỉ gồm các giá trị Input mà k có các giá trị Output.
* Dimention reduction : Là quá trình giảm số lượng biến ngẫu nhiên đang được xem xét bằng cách lấy một tập hợp các biến chính.Phương pháp tiếp cận có thể được chia thành lựa chọn tính năng và trích xuất tính năng.
* Feature selection (Lựa chọn tính năng) : còn được gọi là lựa chọn biến, lựa chọn thuộc tính, là quá trình chọn 1 tập hợp con các tính năng có liên quan để sử dụng trong xây dựng mô hình.