**Câu 1** Xử lý dữ liệu

**\* Làm sạch dữ liệu(cleaning)**

- Tính đầy đủ, trung thực.

+ Mẫu dữ liệu cần đc thu thập từ các nguồn đáng tin cậy. Phản ánh vẫn đề cần giải quyết

+ Loại bỏ nhiễu (ngoại lai): bỏ vài mẫu dữ liệu mà có khác biệt lớn với mẫu khác

+ Một mẫu dữ liệu có thể bị trống (thiếu, chưa đầy đủ), cần có chiến lược phù hợp:

. Bỏ qua, không đưa vào phân tích

. Bổ sung các trường còn thiếu cho mẫu

- Điền các giá trị thiếu

+ Điền lại giá trị bằng tay

+ Gán giá trị nhãn đặc biệt hay ngoài khoảng biểu diễn

+ Gán giá trị trung bình cho nó

+ Gán giá trị trung bình mẫu khác thuộc cùng lớp đó

+Tìm giá trị có xác suất lớn nhất điền vào chỗ bị mất

- Tính đồng nhất: các mẫu dữ liệu cần có tính đồng nhất về cách biểu diễn, ký hiệu

**\* Tích hợp(integrating)**: là quá trình trộn dữ liệu từ các nguồn khác nhau vào một kho dữ liệu có sẵn.

- Cần tránh

+ Dư thừa dữ liệu

+ Mẫu thuẫn dữ liệu

+ Trùng lặp dữ liệu

**\* Chuyển đổi(transforming)**

- Mục tiêu: trích xuất đặc trưng và chuẩn hóa

+ Rời rạc hóa đặc trưng: một số thuộc tính tỏ ra hiệu quả hơn khi được gom các giá trị

+ Chuẩn óa đặc trưng: chuẩn hóa thuộc tính, vrrf cùng moojtmieenf giá trị, dễ dàng trong tính toán

- Giảm kích cỡ

+ Giúp giảm kích thước của dữ liệu và đồng thời giữ đc ngữ nghĩa cốt lõi của dữ liệu

+ Giúp tăng tốc quá trình học hoặc khai phá tri thức

- Một số chiến lược

+ Lựa chọn đặc trưng: các thuộc tính không liên quan, dư thừa hoặc các chiều cũng có thể xóa hay loại bỏ

+ Giảm chiều: dùng một số thuật toán (ví dụ PCA, ICA, LDA, …) để biến đổi dữ liệu ban đâu về không gian có ít chiều hơn