**CLEANNING DATA**

**1.Vấn đề**

-Với quá nhiều dữ liệu được tạo ra, điều quan trọng là dữ liệu chúng tôi sử dụng cho các ứng dụng Khoa học dữ liệu như Học máy và Lập mô hình dự đoán phải sạch. Nhưng dữ liệu sạch nghĩa là gì? Và điều gì làm cho dữ liệu bị bẩn ngay từ đầu?

-Dữ liệu bẩn chỉ đơn giản là dữ liệu có sai sót. Sự trùng lặp của các bản ghi, dữ liệu không đầy đủ hoặc lỗi thời và phân tích cú pháp không đúng có thể làm cho dữ liệu bị bẩn. Dữ liệu này cần được làm sạch. Làm sạch dữ liệu (hoặc làm sạch dữ liệu) đề cập đến quá trình “làm sạch” dữ liệu bẩn này, bằng cách xác định các lỗi trong dữ liệu và sau đó sửa chúng.

🡪Cần phải làm sạch dữ liệu

**2.Các bước làm sạch dữ liệu:**

Bước 1:Đọc dữ liệu,phân chia các cột dữ liệu số với các cột dữ liệu phân loại(khác số)

Bước 2:Thống kê các loại dữ liệu xấu

-Dữ liệu xấu:

* Dữ liệu trống
* Dữ liệu sai định dạng
* Dữ liệu lỗi
* Dữ liệu trùng lặp

Bước 3:Xử lý dữ liệu xấu

-Đối với loại dữ liệu trống:

* Xoá hàng có chứa ô dữ liệu trống
* Thay thế các ô trống bằng giá trị thay thế.

-Đối với dữ liệu sai định dạng:

* Chỉnh sửa lại định dạng
* Xoá hàng sai định dạng

-Đối với dữ liệu lỗi:

* Thay thế bằng các dữ liệu khác
* Xoá hàng có chứa dữ liệu lỗi

-Đối với dữ liệu trùng lặp:

* Xoá hàng có chứa dữ liệu lặp.