**Tiền Xử Lý Dữ Liệu**

Tiền xử lý dữ liệu là một bước không thể thiếu trong Machine Learning. Dữ liệu là một phần rất quan trọng, ảnh hưởng trực tiếp tới việc Training Model.Do đó, tiền xử lý dữ liệu trước khi đưa vào model là rất quan trọng, giúp loại bỏ hoặc bù đắp các dữ liệu chưa chuẩn.

Một số hoạt động hay gặp trong tiền xử lý dữ liệu.

**1. Xử lý Missing Data.**

khi dữ liệu bị thiếu thì sẽ ảnh hưởng đến các mô hình regression hay classification hoặc bất kỳ mô hình khác. Thường dữ liệu bị thiếu sẽ được hiển thị là Null, NAN...

Có 2 cách để xử lý là Xóa bản ghi bị missing hoặc điền vào một giá trị mới.

Để điền vào một giá trị mới thì có nhiều cách:

1: Thay bằng giá trị bất kỳ (0, 10...)

Nếu bạn có thể đoán được về giá trị còn thiếu thì bạn có thể thay thế nó bằng một số giá trị tùy ý.

2: Thay bằng giá trị trung bình (mean)

Đây là phương pháp phổ biến nhất để tính các giá trị còn thiếu của cột số. Nếu có các giá trị ngoại lệ ( độ chênh lệch quá cao, ví dụ có giá trị = 1 và giá trị = 100 thì sẽ ảnh hưởng) thì giá trị trung bình sẽ không phù hợp. Trong những trường hợp như vậy, các ngoại lệ cần được xử lý trước tiên

3: Thay bằng giá trị trung vị(median)

Giá trị trung vị là giá trị ở giữa. Tốt hơn là nên sử dụng giá trị trung bình để quy định trong trường hợp có giá trị ngoại lệ.

4: thay bằng mode:

Chế độ là giá trị xảy ra thường xuyên nhất. Nó được sử dụng trong trường hợp các tính năng phân loại. Bạn có thể sử dụng phương pháp 'fillna' để gán các cột phân loại 'Giới tính', 'Đã kết hôn'...

4: Thay bằng giá trị trước đó hoặc giá trị sau đó.

Trong một số trường hợp, việc gán các giá trị bằng giá trị trước đó thay vì giá trị trung bình, mode hoặc trung vị sẽ phù hợp hơn. Nó chủ yếu được sử dụng trong dữ liệu chuỗi thời gian. Bạn có thể sử dụng hàm ‘fillna’ với tham số ‘method = ffill’

5: Dùng phương pháp nội suy ( Interpolation).

Các giá trị bị thiếu cũng có thể được xác định bằng cách sử dụng phép nội suy. Phương pháp nội suy của Pandas có thể được sử dụng để thay thế các giá trị bị thiếu bằng các phương pháp nội suy khác nhau như polynomial,linear, và quadratic. Phương thức mặc định là linear.

6: Xử lý Categorical data.

Vì đây không phải là dữ liệu số nên không thể dùng mean, median... thay vào đó sẽ thay bằng giá trị thường xuyên nhất hoặc một hằng số.ví dụ:

collumn['a','a','b','c','null']

thì thay null bằng giá trị thường xuyên nhất là 'a', hoặc có thể điền vào 1 giá trị mới biểu diễn cho các giá trị null, ở đây thay bằng 'd', biểu thị null giờ đây là 'b'.

**2. Xử lý các collumn là giá trị Categorical data.**

Vì trong học máy chỉ xử lý các giá trị số nên những lable phải được encode ra một giá trị số.

Một số phương pháp xử lý:

One-hot Encoding

Label Encoding

Ordinal Encoding

Helmert Encoding

Binary Encoding

Frequency Encoding

Mean Encoding

Weight of Evidence Encoding

Probability Ratio Encoding

Hashing Encoding

Backward Difference Encoding

Leave One Out Encoding

James-Stein Encoding

M-estimator Encoding

Thermometer Encoder

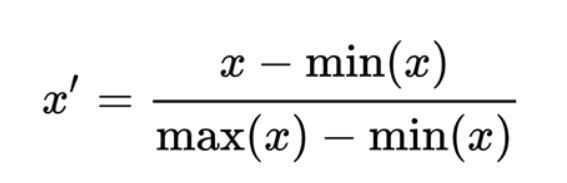
**3. Chuẩn hóa dữ liệu.**

Trong Machine Learning thì chuẩn hóa dữ liệu là một phần rất quan trọng, ảnh hưởng đáng kể đến quá trình học.

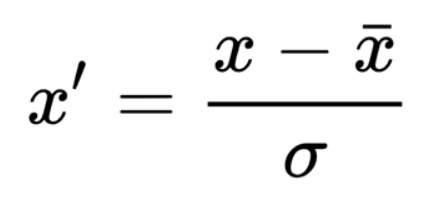
Khi mà chúng ta khai phá dữ liệu thì có thể có một số feature có độ lớn hơn hẳn các feature khác do vậy features nhỏ hơn chắc chắn sẽ bị bỏ qua khi chúng ta thực hiện Machine Learning Model.

Một số phương pháp cơ bản:

**Normalization**: còn được gọi là min-max scalling hoặc min-max normalization,đây là phương pháp đơn giản nhất và bao gồm việc thay đổi tỷ lệ phạm vi của các features để chia tỷ lệ phạm vi trong [0, 1]



Standardization: làm cho giá trị của từng features trong dữ liệu có giá trị trung bình bằng 0 và phương sai bằng 1. Phương pháp tính toán chung là xác định giá trị trung bình phân bố và độ lệch chuẩn cho từng đặc điểm và tính điểm dữ liệu mới theo công thức sau:



4. Đồ thị Histogram.

Biểu đồ là một công cụ vẽ đồ thị phổ biến. Nó được sử dụng để tóm tắt dữ liệu rời rạc hoặc liên tục được đo theo thang đo khoảng. Nó thường được sử dụng để minh họa các tính năng chính của việc phân phối dữ liệu ở dạng thuận tiện. Nó cũng hữu ích khi xử lý các tập dữ liệu lớn. Nó có thể giúp phát hiện bất kỳ quan sát bất thường nào (ngoại lệ) hoặc bất kỳ lỗ hổng nào trong dữ liệu.