**Xử lý missing data**

1. **Những kiểu missing data:**
2. *Missing at Random - Dữ liệu khuyết ngẫu nhiên:*

Sự mất mát dữ liệu ở đây là ngẫu nhiên, tuy nhiên vẫn có mối quan hệ hệ thống giữa dữ liệu bị mất và dữ liệu được quan sát.

1. *Missing Completely at Random - Dữ liệu thiếu hoàn toàn ngẫu nhiên:*

Như tên gọi của nó đã nói lên tất cả. Sự mất mát dữ liệu ở đây là hoàn toàn ngẫu nhiên, và không có bất kỳ một mối quan hệ hay sự liên quan nào giữa dữ liệu và bất kì dữ liệu nào, missing hoặc dữ liệu quan sát.

1. *Missing Not at Random - Dữ liệu khuyết không ngẫu nhiên:*

Sự mất mát dữ liệu không phải là ngẫu nhiên mà có một mối quan hệ xu hướng giữa giá trị bị missing và giá trị không bị missing trong một biến.

1. **Giải quyết vấn đề missing:**

* Để giải quyết missing data, thì ta cần phải hiểu rõ được data mình có.

1. Removing data:

* Trong trường hợp có số lượng missing values không nhiều và chúng ta có thể phán đoán rằng biến đó thực sự không quan trọng. Chúng ta có thể xóa cột đó luôn.
* Nhưng chúng ta có thể sẽ gặp trường hợp, dữ liệu bị xóa chứa thông tin quan trọng sẽ dẫn đến mất mát dữ liệu. Không phù hợp với trường dữ liệu lớn.

1. Thay thế bằng giá trị trung bình của column tương ứng:

* Ta thay thế các missing values bằng giá trị trung bình của cột tương ứng. Việc này đơn giản và không làm giảm số lượng dữ liệu.
* Không phù hợp với dữ liệu phân bố đều, có nhiễu, đặc biệt không phù hợp với dữ liệu có độ biến động lớn.

1. Thay thế bằng mode hoặc median:

* Dễ thực hiện: Việc thay thế missing data bằng mode hoặc median là phương pháp đơn giản và dễ thực hiện.
* Không làm giảm số lượng dữ liệu: Phương pháp này không làm giảm số lượng dữ liệu và giữ lại nhiều thông tin quan trọng nhất có thể.
* Không làm sai lệch dữ liệu: Việc thay thế missing data bằng mode hoặc median không làm sai lệch dữ liệu và giữ nguyên tính chính xác của dữ liệu ban đầu.
* Phù hợp với dữ liệu phân phối không đồng đều: Thay vì sử dụng giá trị trung bình, việc sử dụng mode hoặc median phù hợp hơn với dữ liệu phân phối không đồng đều.
* Nhược điểm:

+ Không phù hợp với dữ liệu có biến động lớn: Nếu dữ liệu có biến động lớn, việc thay thế bằng mode hoặc median có thể làm mất đi nhiều thông tin quan trọng và giảm độ chính xác của phân tích.

+ Không phù hợp với dữ liệu có nhiễu: Nếu dữ liệu có nhiễu, việc thay thế bằng mode hoặc median có thể làm giảm độ chính xác của phân tích.

1. Sử dụng mô hình hồi quy hoặc phân tích đa biến:

* Giữ lại nhiều thông tin hơn: Phương pháp này có thể giữ lại nhiều thông tin hơn về mối quan hệ giữa các biến trong dữ liệu và có thể cải thiện độ chính xác của phân tích.
* Phù hợp với dữ liệu có tính tương quan cao: Nếu dữ liệu có tính tương quan cao, việc sử dụng mô hình hồi quy hoặc phân tích đa biến có thể cải thiện độ chính xác của phân tích so với các phương pháp đơn giản hơn.
* Giảm thiểu ảnh hưởng của outliers: Sử dụng mô hình hồi quy hoặc phân tích đa biến có thể giảm thiểu ảnh hưởng của outliers và giảm thiểu sai số trong dự đoán giá trị của missing data.
* Nhược điểm:

+ Phức tạp: Phương pháp này phức tạp hơn so với các phương pháp đơn giản như thay thế bằng giá trị trung bình, mode hoặc median. Việc xác định mô hình phù hợp và huấn luyện mô hình cần tốn nhiều thời gian và công sức.

+ Cần phải có đủ dữ liệu: Việc sử dụng mô hình hồi quy hoặc phân tích đa biến đòi hỏi phải có đủ dữ liệu để huấn luyện mô hình. Nếu dữ liệu không đủ, mô hình có thể bị overfitting và không đưa ra kết quả chính xác.

+ Không phù hợp với dữ liệu không tuyến tính: Nếu mối quan hệ giữa các biến trong dữ liệu là không tuyến tính, sử dụng mô hình hồi quy hoặc phân tích đa biến có thể không phù hợp và không đưa ra kết quả chính xác.