1. **Chuẩn bị dữ liệu**

Bộ dữ liệu tham khảo trên Kaggle gồm 10000 mẫu và 14 thuộc tính thể hiện các thông số có sẵn và đo được từ máy điện quay. Bộ dữ liệu này chỉ mang tính chất mô phỏng vì các bộ dữ liệu thật về quá trình thí nghiệm trên các động cơ thường có tính bảo mật và không được công khai, vì vậy trong đồ án này ta chỉ dùng bộ số liệu mô phỏng này để dự báo khả năng xảy ra lỗi của một máy điện quay.

A screenshot of a computer

Description automatically generated

Với:

* UID: là số thứ tự của mỗi mẫu trong tập số liệu
* ProductID: biểu thị chất lượng của loại động cơ tiến hành đo, ở đây có ba loại L (thấp), M (trung bình), H (cao).
* Air temperature [K]: nhiệt độ không khí ( nhiệt độ môi trường khi động cơ làm việc)
* Process temperature [K]: nhiệt độ thực khi làm việc (thường có độ chênh lệch nhỏ so với nhiệt độ không khí.
* Rotational speed [rpm]: Tốc độ quay của Rotor
* Torque [Nm]: Lực moment trung bình đầu trục sinh ra
* Tool wear [min]: biểu thị thời gian dụng cụ có thể hoạt động hiệu quả trước khi cần thay thế hoặc bảo trì
* Failure type: chỉ thị loại lỗi động cơ đang gặp phải hoặc hoạt động bình thường
* Tool wear failure (TWF): lỗi động cơ liên quan đến sự mài mòn thiết bị
* Heat dissipation failure (HDF): lỗi quá nhiệt động cơ khi hoạt động
* Power failure (PWF): Lỗi công suất đầu ra của động cơ không đáp ứng đủ nhu cầu cho tải (<3500W) và quá tải (>9000W)
* Overstrain failure (OSF): hỏng về mặt cơ học do động cơ hoạt động quá tải
* Random failures (RNF): các lỗi ngẫu nhiên khác xuất hiện

1. **Xử lý dữ liệu**
   1. **Loại bỏ dữ liệu**

Để xây dựng một mô hình dự đoán hiệu quả, việc loại bỏ các dữ liệu không cần thiết là một bước quan trọng trong quá trình tiền xử lý dữ liệu. Những dữ liệu này có thể bao gồm các thông tin không liên quan, không có ý nghĩa hoặc không có tác động đáng kể đến việc xác định kết quả đầu ra của mô hình, cụ thể là chuẩn đoán xem động cơ có lỗi hay không. Ví dụ, nếu chúng ta đang nghiên cứu về tình trạng kỹ thuật của động cơ, các thuộc tính như màu sơn của vỏ động cơ hoặc các thông tin ít liên quan khác không đóng góp vào việc dự đoán tình trạng lỗi có thể bị loại bỏ. Việc giữ lại những dữ liệu không liên quan có thể làm tăng độ phức tạp của mô hình, giảm độ chính xác và kéo dài thời gian xử lý.

A screenshot of a computer

Description automatically generated

*Hình x. Dữ liệu sau khi loại bỏ thuộc tính UDI và Product ID*

Khi xử lý dữ liệu, việc kiểm tra xem các thuộc tính trong dữ liệu đầu vào có chứa giá trị **None** hoặc bị thiếu (Missing Value) là một bước vô cùng quan trọng. Missing Value có thể xuất hiện do nhiều nguyên nhân như lỗi thu thập dữ liệu, lỗi trong quá trình nhập liệu, hoặc sự không đồng bộ giữa các nguồn dữ liệu. Những giá trị này có thể gây ảnh hưởng tiêu cực đến kết quả phân tích hoặc mô hình dự đoán nếu không được xử lý đúng cách.

Khi xử lý Missing Value, có nhiều cách thức tùy thuộc vào bản chất dữ liệu và mục tiêu của bài toán:

1. **Loại bỏ Missing Value**: Phương pháp này thích hợp khi số lượng giá trị bị thiếu nhỏ và không ảnh hưởng lớn đến dữ liệu tổng thể.
2. **Thay thế Missing Value**: Phương pháp này phổ biến hơn, trong đó giá trị bị thiếu được thay thế bằng:
   * **Giá trị trung bình, trung vị hoặc mode**: Thường dùng cho dữ liệu số.
   * **Giá trị suy đoán**: Áp dụng các mô hình dự đoán hoặc thuật toán như **k-NN imputation** để dự đoán giá trị bị thiếu dựa trên các thuộc tính khác.
   * **Giá trị mặc định**: Với dữ liệu dạng danh mục (categorical), có thể thay thế bằng một giá trị cụ thể như "Unknown".
3. **Gán giá trị Missing Value với dấu hiệu đặc biệt**: Thêm một giá trị đặc biệt (như -1 hoặc "N/A") để đánh dấu và phân biệt các giá trị bị thiếu trong phân tích tiếp theo.

A screenshot of a computer

Description automatically generated

*Hình x. Tổng quan thống kê chung về các chỉ số của dữ liệu*

Nhìn vào hình trên ta thấy dữ liệu không có Missing Value vì các thuộc tính đều đạt 10000 có giá trị /10000 mẫu

* 1. **Chuyển đổi đơn vị nhiệt độ**

Nhằm tăng độ tương quan cho người dùng về số liệu ta chuyển từ nhiệt độ K sang độ C.

A screenshot of a computer

Description automatically generated

A blue rectangular object with white text

Description automatically generated

* 1. **Biểu diễn mối quan hệ các thuộc tính**
  2. Chuẩn hóa số liệu
  3. Mã hóa các loại lỗi

1. **Xây dựng mô hình**
   1. Phân chia dữ liệu
   2. Mô hình Decision tree
   3. Mô hình KNN
   4. Mô hình Naïve bayes
   5. Mô hình CNN
2. **Đánh giá hiệu suất mô hình**
   1. Đánh giá hiệu suất từng mô hình
   2. Kết luận