|  |
| --- |
|  |
| **TRƯỜNG ĐẠI HỌC CÔNG NGHIỆP HÀ NỘI**  **KHOA CÔNG NGHỆ THÔNG TIN**  **---------------------------------------** |
| **Logo%20HaUI%20ban%20chuan** |
| **ĐỒ ÁN TỐT NGHIỆP**  **NGÀNH KỸ THUẬT PHẦN MỀM** |
| **TÊN ĐỀ TÀI: XÂY DỰNG ỨNG DỤNG DỰ ĐOÁN MÁY HỎNG SỬ DỤNG HỌC MÁY** |
| |  |  |  | | --- | --- | --- | | **GVHD** | **:** | **ThS. Phạm Thế Anh** | | **Sinh Viên** | **:** | **Nguyễn Văn Thuỳ** | | **Mã Sinh viên** | **:** | **2019606418** | | **Lớp** | **:** | **2019DHKTPM04 – K14** | |
|  |
|  |
| *Hà Nội – Năm 2022* |

# 

# LỜI CẢM ƠN

# TÓM TẮT ĐỀ TÀI

# LỜI NÓI ĐẦU

**MỤC LỤC**

[LỜI CẢM ƠN i](#_Toc131966517)

[TÓM TẮT ĐỀ TÀI ii](#_Toc131966518)

[LỜI NÓI ĐẦU iii](#_Toc131966519)

[CHƯƠNG 1. CƠ SỞ LÝ THUYẾT VÀ XÂY DỰNG MÔ HÌNH 1](#_Toc131966520)

[1.1. Tìm hiểu bài toán về bảo trì dự đoán (predictive maintenance) 1](#_Toc131966521)

[1.2. Tìm hiểu về dữ liệu Turborfan Engine Degration Data 5](#_Toc131966522)

[1.3. Xây dựng notebook để xử lý dữ liệu và tạo mô hình học máy 7](#_Toc131966523)

[1.3.1. Chuẩn bị dữ liệu 7](#_Toc131966524)

[1.3.2. Tách dữ liệu 10](#_Toc131966525)

[1.3.3. Phân nhóm các máy có cách vận hành như nhau. 11](#_Toc131966526)

[1.3.4. Chuẩn hoá dữ liệu. 12](#_Toc131966527)

[1.3.5. Xây dựng chỉ số sức khỏe. 13](#_Toc131966528)

[1.3.6. Sử dụng hàm đa thức 15](#_Toc131966529)

[1.3.7. Đánh giá hiệu suất mô hình 16](#_Toc131966530)

[CHƯƠNG 2. PHÂN TÍCH THIẾT KẾ HỆ THỐNG 18](#_Toc131966531)

[2.1. Phân tích hệ thống 18](#_Toc131966532)

[2.1.1. Vai trò của người dùng 18](#_Toc131966533)

[2.1.2. Yêu cầu chức năng 18](#_Toc131966534)

[2.1.3. Yêu cầu phi chức năng 19](#_Toc131966535)

[2.2. Biểu đồ use case 19](#_Toc131966536)

[2.3. Mô hình thực thể liên kết 20](#_Toc131966537)

[2.4. Đặc tả use case 21](#_Toc131966538)

[2.4.1. Use case đăng nhập 21](#_Toc131966539)

[2.5. Thiết kế hệ thống 23](#_Toc131966540)

[2.5.1. Biểu đồ triển khai hệ thống 23](#_Toc131966541)

[2.5.2. Thiết kế cơ sở dữ liệu 23](#_Toc131966542)

[2.5.3. Thiết kế giao diện 23](#_Toc131966543)

[CHƯƠNG 3. KẾT QUẢ ĐẠT ĐƯỢC 24](#_Toc131966544)

[KẾT LUẬN 25](#_Toc131966545)

[TÀI LIỆU THAM KHẢO 26](#_Toc131966546)

[PHỤ LỤC 27](#_Toc131966547)

**DANH MỤC HÌNH ẢNH**

**DANH MỤC BẢNG**

# CHƯƠNG 1. CƠ SỞ LÝ THUYẾT VÀ XÂY DỰNG MÔ HÌNH

## Tìm hiểu bài toán về bảo trì dự đoán (predictive maintenance)

**Bảo trì dự đoán** hay còn gọi là **Bảo trì dự báo**hoặc **Bảo trì tiên đoán** (tiếng Anh: Predictive Maintenance) là việc sử dụng các công cụ và kỹ thuật phân tích dữ liệu để dự đoán những bất thường trong hoạt động của các thiết bị hoặc sai lỗi có thể xảy ra với trước khi dẫn đến hỏng hóc.

Mục đích chính của phương pháp này là dự đoán thời điểm sai lỗi và thiết lập kế hoạch chi tiết cho việc bảo trì máy móc. Bảo trì dự đoán giúp nhà sản xuất giảm thiểu chi phí liên quan đến bảo trì phòng ngừa, hạn chế bảo trì ngoài kế hoạch, giảm tần suất bảo trì, hạn chế những rủi ro trong sản xuất.

Theo nghiên cứu của Chương trình Quản lý Năng lượng Liên bang Mỹ (FEMP), ước tính một chương trình bảo trì dự đoán Predictive Maintenance hoạt động đúng cách có thể tiết kiệm từ 30% đến 40% so với bảo trì phản ứng (Reactive Maintenance).

**Lợi ích của bảo trì dự đoán:**

Ứng dụng bảo trì dự báo trong sản xuất hiệu quả đem lại nhiều lợi ích cho doanh nghiệp:

Tối ưu hóa thời gian dừng máy

Cơ chế hoạt động của phương pháp bảo trì dự đoán cũng giống với việc sử dụng máy tính. Việc thường xuyên vệ sinh máy, quét virus, làm sạch các dữ liệu rác sẽ giúp máy chạy ổn định, hiệu quả, từ đó, phục vụ nhu cầu của người dùng tốt hơn, lâu dài hơn. Tương tự như vậy, việc tạm dừng máy để lau chùi, bảo dưỡng và thay thế các bộ phận hỏng hóc sẽ giúp giảm thiểu rủi ro thiết bị hư hỏng bất chợt.

Bảo trì dự đoán không thực hiện bảo trì máy móc của doanh nghiệp một cách ngẫu nhiên. Tất cả thời gian máy được dừng để bảo trì, kiểm tra đều được lựa chọn kỹ càng dựa dựa trên những thu thập và phân tích dữ liệu từ quá trình vận hành. Hệ thống sẽ lên kế hoạch bảo trì thường xuyên nhưng chỉ tại các thời điểm ít ảnh hưởng đến sản xuất nhất. Điều này giúp doanh nghiệp tối ưu hóa thời gian sử dụng máy, gia tăng tuổi thọ của máy móc đắt tiền, máy móc khó thay thế.

1. **Tối thiểu việc dừng máy bất chợt**

Việc dừng máy không có kế hoạch gây ra sự trì trệ và lãng phí trong sản xuất. Với sự hỗ trợ của công cụ bảo trì dự đoán, các thiết bị sản xuất được kết nối với hệ thống này, trong quá trình vận hành, sẽ tự động thu thập hàng loạt các dữ liệu thô. Dựa vào đó, hệ thống sẽ đưa ra các dạng khuôn mẫu vận hành đặc trưng cho từng máy. Bằng việc so sánh hoạt động của máy với mô hình chuẩn, hệ thống bảo trì dự đoán sẽ xác định được thiết bị nào có khả năng xảy ra sai lỗi để có những phương án bảo trì kịp thời, hợp lý.

1. **Gia tăng tuổi thọ thiết bị**

Bảo trì dự đoán cho phép nhà sản xuất giám sát hoạt động máy móc và chất lượng sản phẩm đầu ra của họ một cách tự động và liên tục. Từ đó, hệ thống biết được thời điểm lý tưởng để bảo trì máy, cũng như khi nào thì máy móc đạt tới giới hạn tuổi thọ và cần được thay thế.

Tuổi thọ một thiết bị có mối liên hệ chặt chẽ với mức độ sử dụng của nó. Tuy nhiên, doanh nghiệp hoàn toàn có thể can thiệp vào điều này thông qua hệ thống bảo trì dự đoán vào sản xuất. Bằng việc theo dõi liên tục dữ liệu về điều kiện máy và so sánh chúng với mô hình chuẩn của máy, hệ thống có thể xác định được khi nào phát sinh nhu cầu thay thế các bộ phận cho máy. Đồng thời, việc dự báo được giới hạn tuổi thọ của máy giúp doanh nghiệp có những chuẩn bị kỹ càng và tiết kiệm đáng kể thời gian cũng như chi phí.

1. **Tối ưu năng suất nhân viên**

Thay vì để máy dừng đột ngột, bảo trì dự đoán cung cấp cho nhà sản xuất lên lịch và thời gian cần sửa vừa đủ cho máy. Điều này giúp doanh nghiệp giảm thiểu thời gian chết (downtime) khi dây chuyền bị gián đoạn, đồng thời cân đối ca làm của nhân viên để tối ưu năng suất lao động.

Bên cạnh đó, dưới góc nhìn của nhân viên, công cụ này còn đóng vai trò quan trọng trong việc xây dựng môi trường làm việc an toàn, ổn định cho người lao động. Bằng việc bảo trì đúng thời điểm và dự đoán được trước thời gian máy đạt giới hạn tuổi thọ giúp giảm nguy hiểm trong lao động, hạn chế tối đa các sự cố nguy hiểm phát sinh bất ngờ. Người lao động từ đó cũng cảm thấy an tâm và tập trung hơn vào hoàn thành công việc được giao.

1. **Tăng doanh thu**

Một lợi ích không thể không đề cập tới của bảo trì dự báo đối với doanh nghiệp áp dụng hệ thống này trong sản xuất là việc nâng cao doanh thu, hay cụ thể hơn, là nâng cao hiệu quả cho quá trình sản xuất.

Tiết kiệm chi phí bảo trì: Mô hình bảo trì dự đoán sử dụng dữ liệu thu thập được từ hoạt động thực tế của máy để giám sát, phân tích và lập kế hoạch bảo trì vào thời điểm cần thiết. Điều này giúp nhà sản xuất giảm thiểu đáng kể tính thiếu hiệu quả trong bảo trì máy móc dư thừa.

Giảm thiểu lãng phí: Bảo trì máy móc chỉ có lợi khi nhà sản xuất thực hiện hoạt động này đúng thời gian, đúng phương pháp. Việc bảo trì không dựa trên tình trạng máy sẽ dẫn đến những lãng phí trong sản xuất: Chi phí lưu kho bãi của nguyên liệu thô, chi phí lao động, chi phí trả cho hoạt động bảo trì,… Bảo trì dự đoán giúp phát hiện những vấn đề đó trước khi nó xảy ra.

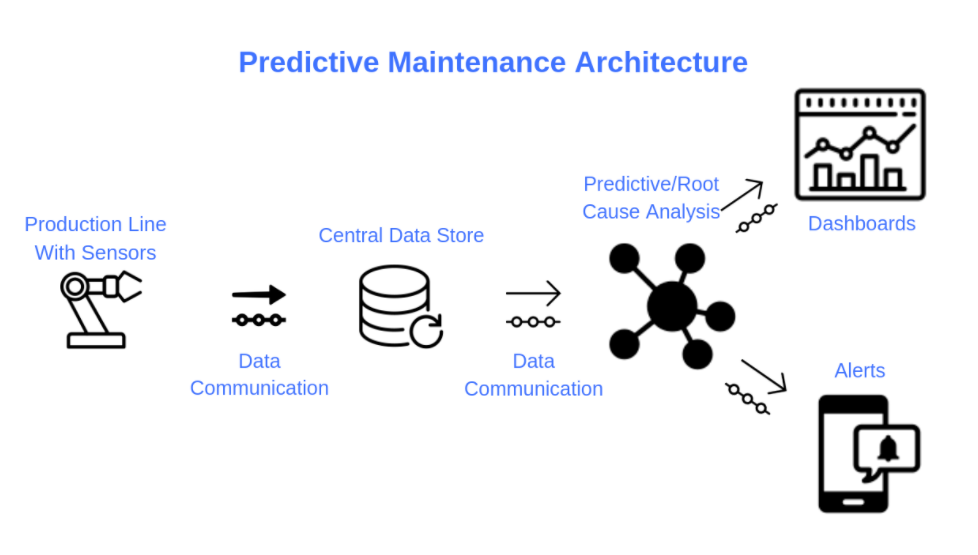
Tối ưu hóa hoạt động dây chuyền sản xuất: Các dữ liệu của máy được hệ thống bảo trì dự đoán thu thập tự động. Dựa trên những phân tích chuyên nghiệp, hệ thống sẽ xử lý dữ liệu và đưa ra kế hoạch bảo trì đảm bảo tối ưu quy trình hoạt động cho dây chuyền sản xuất của doanh nghiệp.

**Nguyên lý hoạt động của bảo trì dự đoán**

Bảo trì dự đoán giúp đánh giá tình trạng của thiết bị và dự báo thời điểm lỗi có khả năng xảy ra, với mục tiêu cuối cùng là lập kế hoạch bảo trì giúp doanh nghiệp tối ưu chi phí và tăng độ tin cậy của hệ thống.

Để theo dõi được tình trạng thiết bị và cảnh báo người giám sát về những hỏng hóc có thể xảy ra của thiết bị, một hệ thống bảo trì dự đoán phải đảm bảo có đầy đủ ba thành phần chính:

* Các cảm biến được lắp đặt để thu thập dữ liệu và phân tích tình trạng máy móc theo thời gian thực. Có nhiều loại dữ liệu khác nhau được thu thập, mỗi loại theo dõi các thiết bị khác nhau trong dây chuyền sản xuất.
* Công nghệ Internet of Things (IoT) cho phép hệ thống thu thập, phân tích, lưu trữ và xử lý được một lượng dữ liệu khổng lồ liên quan đến máy.
* Các mô hình dữ liệu dự đoán được cung cấp với tất cả dữ liệu đã xử lý giúp dự đoán được hỏng hóc và đưa ra cảnh báo để bảo trì.



Hình Kiến trúc của dự đoán bảo trì

**Các phương pháp dự đoán bảo trì**

Hiện nay, có năm phương pháp bảo trì dự báo phổ biến, đó là:

1. **Phân tích độ rung**

Nhà sản xuất thường sử dụng cảm biến rung để phát hiện các dấu hiệu về sai lỗi, xuống cấp đối với máy móc hoạt động nặng họ đang sở hữu. Nguyên lý hoạt động của phương pháp này là theo dõi liên tục tốc độ rung của máy dựa trên một độ rung tiêu chuẩn để phát hiện các điểm lệch một cách chính xác. Đây được coi là một trong những phương pháp đạt độ chính xác cao nhất trong dự báo sai lỗi xảy ra ở máy.

1. **Công nghệ hồng ngoại**

Còn có tên gọi khác là tạo ảnh nhiệt, kỹ thuật ứng dụng công nghệ hồng ngoại giúp phát hiện các bộ phận bị ma sát quá nhiều bằng việc theo dõi nhiệt lượng của các thiết bị hoạt động và xác định các điểm có nhiệt lượng cao đột biến, để đưa ra cảnh báo về bảo trì nhanh chóng.

1. **Phân tích sóng âm và siêu âm**

Phân tích sóng âm và siêu âm sử dụng các tín hiệu âm thanh để phát hiện các vết nứt, mối hàn bị hỏng từ khi chúng còn rất nhỏ, khó có thể quan sát được bằng mắt thường. Phương pháp này đặc biệt có ích trong việc phát hiện rò rỉ khí hay chất lỏng trong dây chuyền.

1. **Phân tích chất lượng dầu**

Kỹ thuật đánh giá mức độ hao mòn của các thiết bị sử dụng dầu bằng việc kiểm tra số lượng các mảnh vụn trong máy. Ngoài ra, đây cũng là kỹ thuật giúp phát hiện rò rỉ và đánh giá được độ sạch của dầu trong máy.

1. **Phân tích mạch động cơ**

Phân tích mạch động cơ được ứng dụng trong ngành công nghiệp hàng hải và sản xuất ô tô. Cơ chế hoạt động của phương pháp này dựa trên số liệu đó được từ stator và roto của động cơ để phát hiện các lỗi nối đất. Nó cũng cho phép người dùng có thể kiểm tra động cơ trước khi lắp đặt thiết bị.

**Ưu & Nhược điểm của Bảo trì dự đoán**

Ưu điểm lớn nhất phải kể đến mà bảo trì dự báo đem lại cho nhà sản xuất là tính tiết kiệm trong chi phí dành cho các hoạt động bảo dưỡng, sửa chữa máy móc. Bên cạnh đó, doanh nghiệp còn có được sự chủ động và kịp thời trong việc đưa ra các phương pháp bảo trì và kéo dài tuổi thọ của máy. Điều này là tiền đề giúp hoạt động sản xuất trong doanh nghiệp được vận hành ổn định.

Tuy vậy, bảo trì dự báo cũng có những hạn chế nhất định mà doanh nghiệp cần lưu ý. Phương pháp này đòi hỏi chi phí đầu tư ban đầu khá tốn kém về cả mặt công nghệ và lắp đặt. Ngoài ra, nó cũng đặt ra những yêu cầu khắt khe về trình độ chuyên môn và kinh nghiệm ở phía nhân sự để có thể quản lý hệ thống hiệu quả.

**Những lĩnh vực cần ứng dụng bảo trì dự đoán**

Bảo trì dự đoán được ứng dụng ở nhiều lĩnh vực trong cuộc sống với sự góp mặt của máy móc, thiết bị, có thể kể đến như:

1. **Trong Sản xuất**

Khả năng hoạt động ổn định, lâu bền của các máy móc trong toàn dây chuyền sản xuất là điều kiện sống còn để một doanh nghiệp sản xuất tồn tại và phát triển. Bảo trì dự đoán vì thế được coi là một yếu tố không thể thiếu để giám sát và phát hiện sự cố máy móc tại các đơn vị có công cụ, thiết bị với quy mô lớn và đắt tiền.

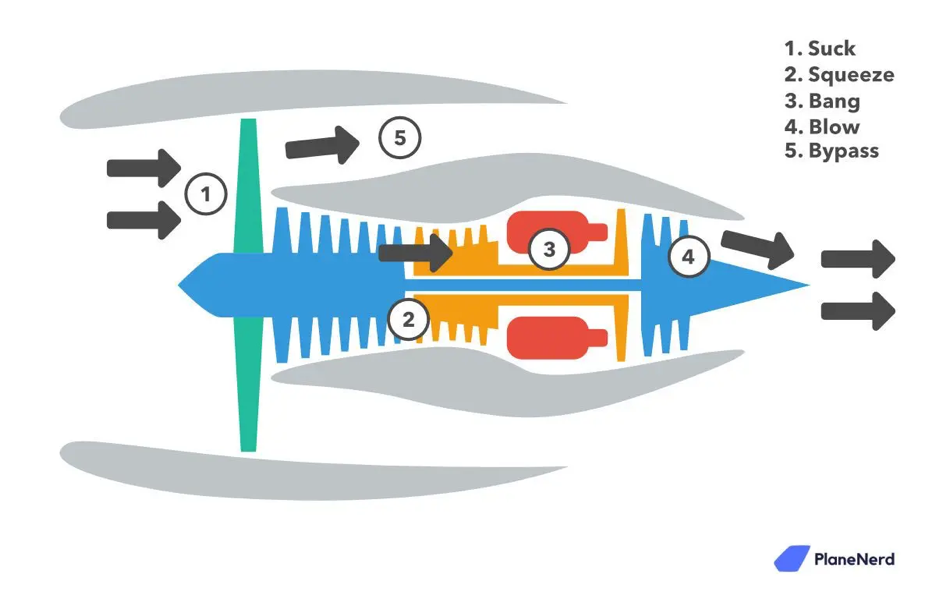
1. **Trong lĩnh vực Năng lượng**

Công nghệ bảo trì dự đoán được ứng dụng rộng rãi trong ngành sản xuất và khai thác năng lượng bởi khả năng dự đoán và cảnh báo rủi ro kịp thời của nó. Một ví dụ cụ thể ở ngành điện lực, để đảm bảo lượng điện năng được truyền tải ổn định, liên tục và an toàn đến người dùng, hệ thống máy móc trong nhà máy phải hoạt động trơn tru và hoạt động sửa chữa, bảo trì phải được lên kế hoạch trước khi sự cố diễn ra.

1. **Trong lĩnh vực Quản lý chất thải**

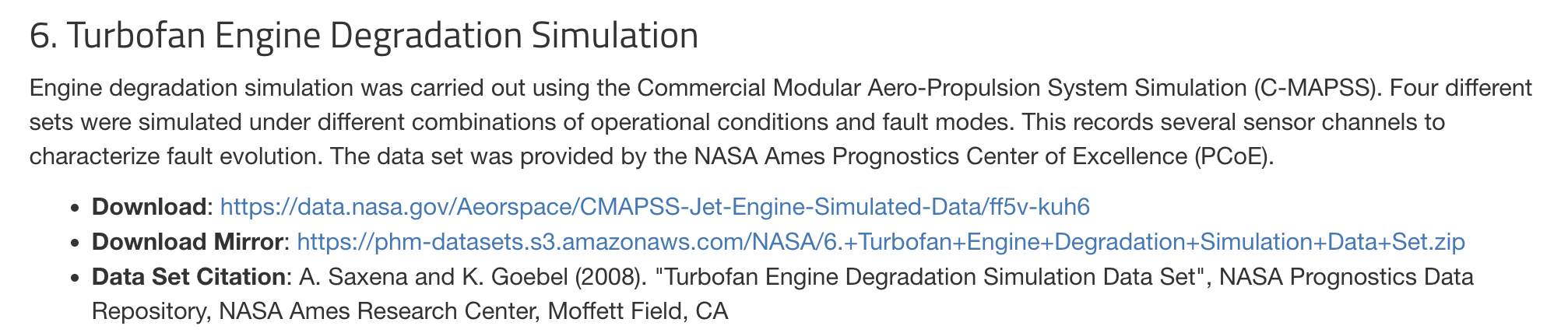
Bảo trì dự đoán đem lại sự ổn định và tính hiệu quả của máy móc và nhân công trong ngành công nghiệp tái chế và quản lý chất thải. Phương pháp này giúp hạn chế lãng phí có thể xảy ra khi bảo trì dư thừa, đồng thời tiết kiệm chi phí khi máy móc cần được thay thế.

## Tìm hiểu về dữ liệu Turborfan Engine Degration Data



Hình Động cơ turborfan engine của máy bay

Dữ liệu mô phỏng động cơ phản lực CMAPSS



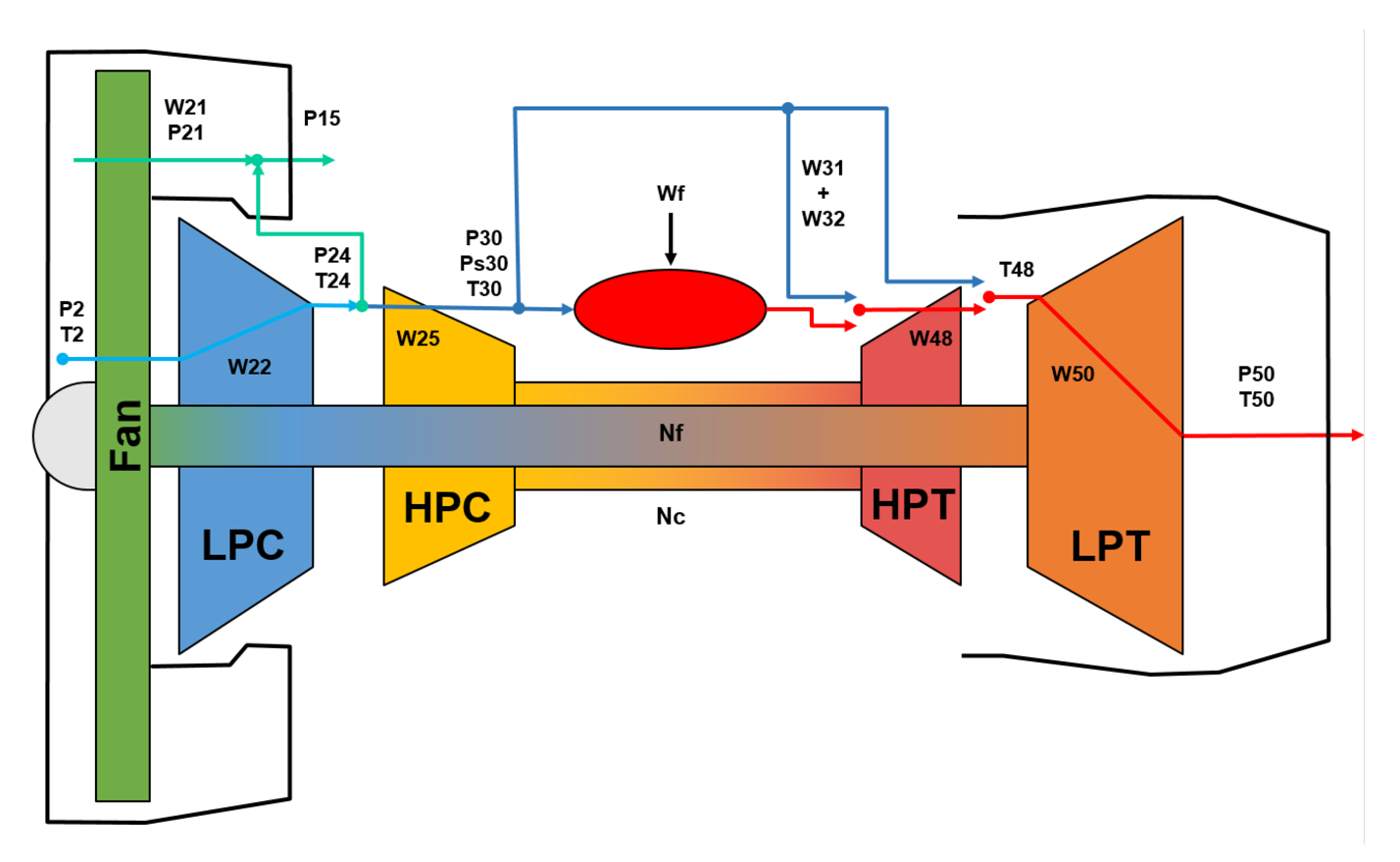
Hình Nguồn tải dữ liệu Turbofan Engine Degradation

Tập dữ liệu bao gồm nhiều chuỗi thời gian đa biến. Mỗi tập dữ liệu được chia thành các tập con huấn luyện và kiểm tra. Mỗi chuỗi thời gian là từ một động cơ khác nhau, tức là dữ liệu có thể được coi là từ một nhóm động cơ cùng loại. Mỗi động cơ khởi động với các mức độ hao mòn ban đầu và sự thay đổi trong sản xuất khác nhau mà người dùng không biết. Sự hao mòn và thay đổi này được coi là bình thường, tức là không được coi là tình trạng lỗi. Có ba cài đặt vận hành có ảnh hưởng đáng kể đến hiệu suất của động cơ. Các cài đặt này cũng được bao gồm trong dữ liệu. Dữ liệu bị nhiễm nhiễu cảm biến.

Động cơ đang hoạt động bình thường khi bắt đầu mỗi chuỗi thời gian và phát sinh lỗi tại một số thời điểm trong chuỗi. Trong tập huấn luyện, lỗi tăng dần về độ lớn cho đến khi lỗi hệ thống. Trong bộ thử nghiệm, chuỗi thời gian kết thúc một thời gian trước khi hệ thống bị lỗi. Mục tiêu của cuộc thi là dự đoán số chu kỳ hoạt động còn lại trước khi hỏng hóc trong tập thử nghiệm, tức là số chu kỳ hoạt động sau chu kỳ cuối cùng mà động cơ sẽ tiếp tục hoạt động. Cũng cung cấp một véc-tơ giá trị Thời gian sử dụng hữu ích còn lại (RUL) thực sự cho dữ liệu thử nghiệm

Dữ liệu được cung cấp dưới dạng tệp văn bản nén zip với 26 cột số, được phân tách bằng dấu cách. Mỗi hàng là một ảnh chụp nhanh dữ liệu được lấy trong một chu kỳ hoạt động, mỗi cột là một biến khác nhau.

Các cột tương ứng với:



Hình Chi tiết vị trí các sensor

5 bộ phận quay của động cơ: Fan, LPC (Low-Pressure Compressor), HPC (High-Pressure Compressor), HPT (High-Pressure Turbine), LPT (Low-Pressure Turbine)

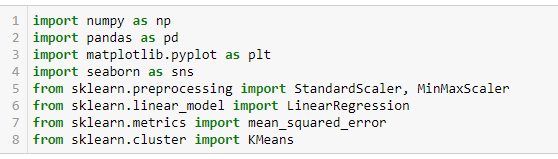
**Bảng mô tả các cột của cảm biến**

|  |  |  |
| --- | --- | --- |
| **Các cột** | **Cảm biến** | **Mô tả chi tiết** |
| Unit | Unit | id của từng máy |
| Time | Time | Thời gian chạy của từng máy |
| Op\_setting\_1 | Op\_setting\_1 | Các cài đặt hệ thống của từng máy lúc vận hành |
| Op\_setting\_2 | Op\_setting\_2 |
| Op\_setting\_3 | Op\_setting\_3 |
| sensor\_1 | T2 | Tổng nhiệt độ đầu vào của quạt |
| sensor\_2 | T24 | Tổng nhiệt độ đầu ra ở LPC |
| sensor\_3 | T30 | Tổng nhiệt độ đầu ra ở HPC |
| sensor\_4 | T50 | Tổng nhiệt độ đầu ra ở LPT |
| sensor\_5 | P2 | Áp suất đầu vào của quạt |
| sensor\_6 | P15 | Tổng áp xuất trong ông dẫn nhánh |
| sensor\_7 | P30 | Tổng áp xuất tại đầu ra HPC |
| sensor\_8 | Nf | Tốc độ quạt vật lý |
| sensor\_9 | Nc | Tốc độ tại lõi vật lý |
| sensor\_10 | epr | Tỷ lệ áp suất động cơ (P50/P2) |
| sensor\_11 | Ps30 | Áp suất tĩnh tại đầu ra HPC |
| sensor\_12 | Phi | Tỷ lệ lưu lượng nhiên liệu với Ps30 |
| sensor\_13 | Nrf | Tốc độ quạt thực sự |
| sensor\_14 | Nrc | Tốc độ ở lõi thực sự |
| sensor\_15 | Bpr | Tỷ lệ Bypass |
| sensor\_16 | farB | Tỷ lệ nhiên liệu-không khí dầu đốt |
| sensor\_17 | htBleed | Thông số Enthalpy |
| sensor\_18 | Nf\_dmd | Yêu cầu tốc độ quạt |
| sensor\_19 | PCNfR\_dmd | Yêu cầu điều chỉnh tốc độ quạt |
| sensor\_20 | W31 | Dung dịch làm mát HPT |
| sensor\_21 | W32 | Dung dịch làm mát LPT |

## Xây dựng notebook để xử lý dữ liệu và tạo mô hình học máy

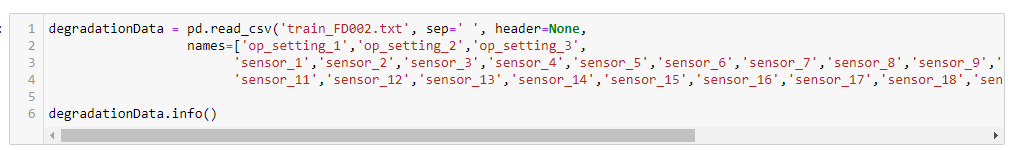
Đầu tiên để thực hiện nghiên cứu, chúng ta import các thư viên cần thiết để xử lý data.

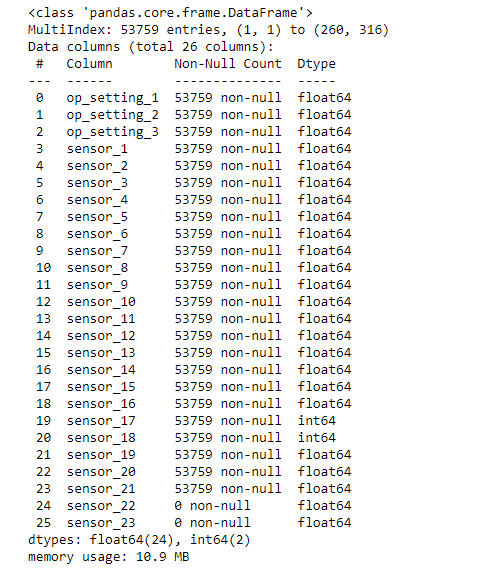
### Chuẩn bị dữ liệu



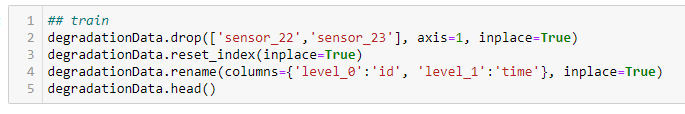
Các thư viên numpy, pandas, matplotlib.pyplot sklearn là các thư viện để đọc và phân tích dữ liệu.

Sử dụng pd.read\_csv để đọc dữ liệu, chèn thêm tên các cột để xử lý



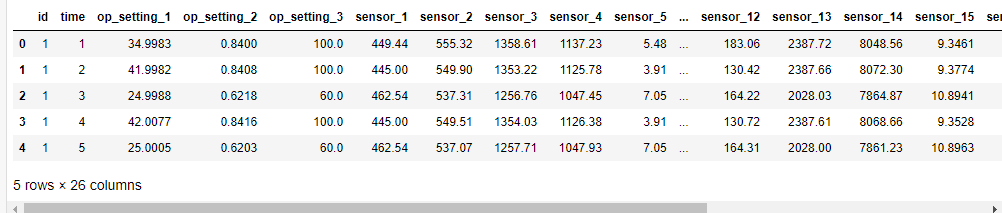


Sử dụng .info() của pandas để nắm các thông tin tổng quát của dữ liệu. Ta có thể dễ dàng thấy các cột từ 0 đến 23 đều không có giá trị null, các cột 24-25 bị thiếu dữ liệu, ta tiến hành bỏ 2 cột này, và thêm các cột level\_0, level\_1. Là các id và thời gian của mỗi máy vào cột mới.

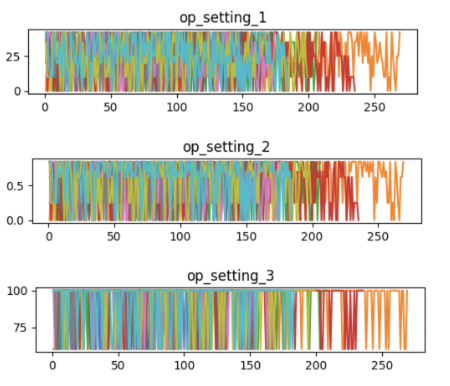


In thử 5 dòng đầu tiên của dữ liệu ta có thể thấy rõ các cột và giá trị của các cột đó.

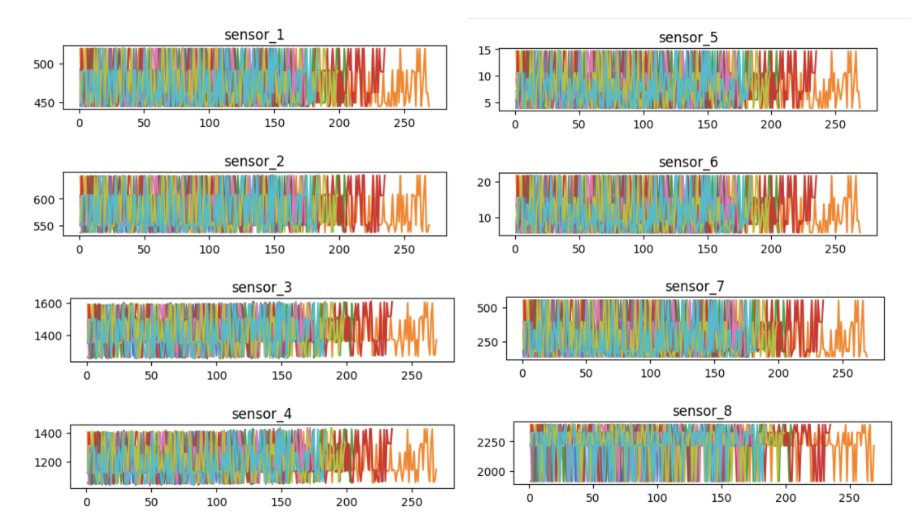
* Cột id đại diện cho các máy, mỗi id là một máy
* Cột time đại diện mỗi lần tính cho đến lúc hỏng sẽ bắt đầu chạy từ 1.
* Cột op\_setting\_ là các cột cài đặt vận hành
* Cột sensor\_. Là các cột cảm biến thông tin được đặt tại các vị trí trong turbor fan



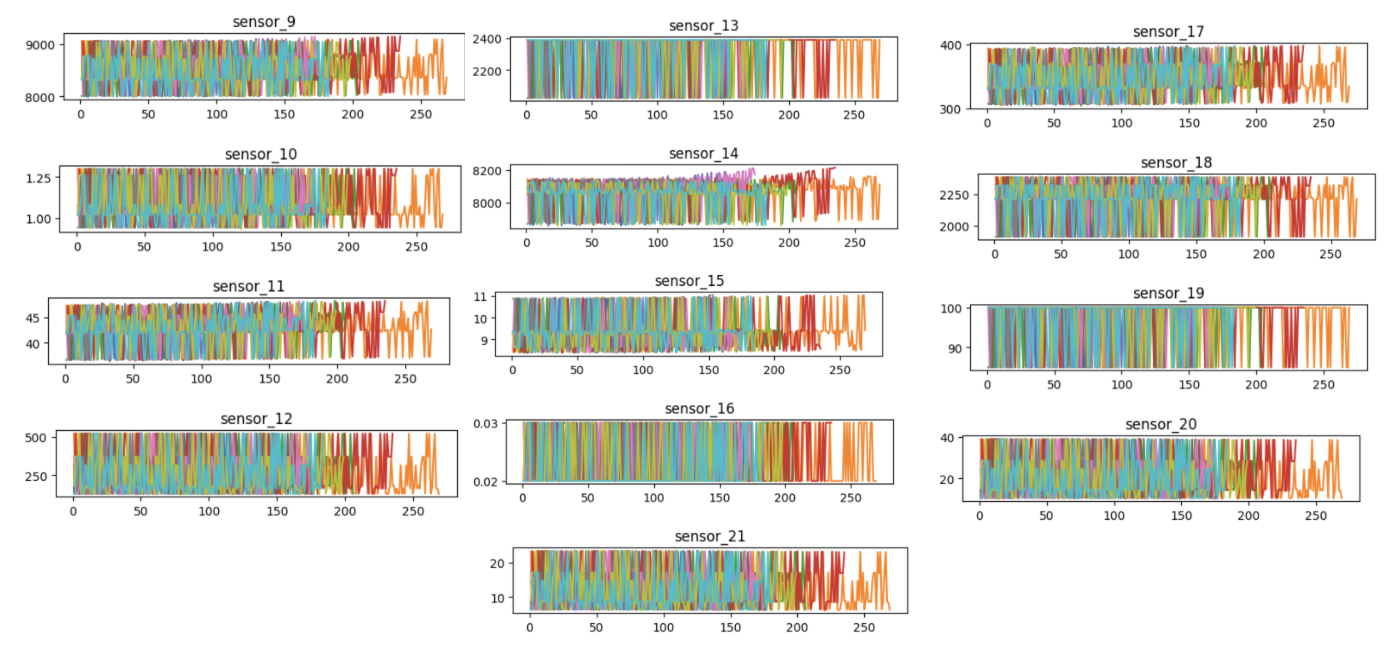
Hiện thử 10 mẫu máy có các cột op\_setting, và sensor\_ để hình dung dữ liệu



Hình 2.12. 10 mẫu op\_setting 1,2 và 3



Hình 2.13. Hiển thị 10 mẫu cảm biến 1-8

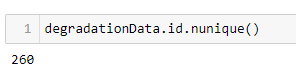


Hình Hiển thị 10 mẫu có các cảm biến từ 9 - 21

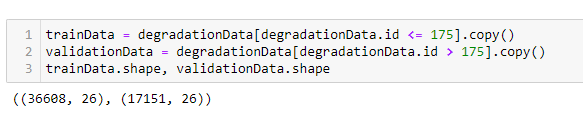
Nhìn vào dữ liệu chúng ta sẽ không biết dữ liệu nào tốt, chúng chưa thực sự mang lại nhiều thông tin cho bài toán.

### Tách dữ liệu

Để xem có bao nhiêu máy trong tập dữ liệu, sử dụng hàm nuique() sẽ biết số các giá trị trong cột id. Từ đó ta thấy có 260 máy trong tập dữ liệu.

****

Ở đây sẽ tách dữ liệu 175 xe cho tập training, và 85 xe cho tập đánh giá

****

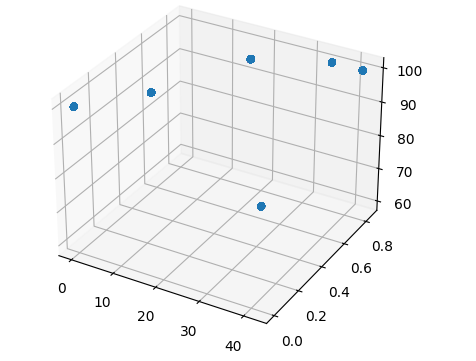
Việc tách dữ liệu là một bước rất quan trọng để đánh giá hiệu suất của mô hình. Tách dữ liệu bao gồm chia dữ liệu thành tập huấn luyện (training set) và tập kiểm tra (test set) để đánh giá khả năng tổng quát hóa của mô hình trên dữ liệu mới.

Tách dữ liệu có các lợi ích sau:

* Đánh giá hiệu suất mô hình: Tách dữ liệu giúp đánh giá khả năng tổng quát hóa của mô hình trên dữ liệu mới mà mô hình chưa từng nhìn thấy. Nếu mô hình chỉ học trên tập huấn luyện, và không được đánh giá trên tập kiểm tra, thì mô hình có thể dẫn đến overfitting (quá khớp) - mô hình hoạt động tốt trên tập huấn luyện nhưng không hoạt động tốt trên dữ liệu mới.
* Cải thiện khả năng dự đoán của mô hình: Tách dữ liệu giúp cải thiện khả năng dự đoán của mô hình bằng cách đánh giá các thay đổi trong hiệu suất của mô hình khi thay đổi các tham số hoặc phương pháp đào tạo.
* Hiểu rõ hơn về dữ liệu: Tách dữ liệu giúp ta hiểu rõ hơn về dữ liệu bằng cách phân tích các giá trị trong tập huấn luyện và tập kiểm tra, đánh giá các biến quan trọng và tìm kiếm các tương tác giữa các biến.

### Phân nhóm các máy có cách vận hành như nhau.

Đầu tiên hiển thị 3 biến cài đặt khởi động máy chúng ta dễ dàng thấy phân loại rõ ràng 6 cụm. Từ đó tiến hành sử dụng thuật k-means để phân cụm các máy có cách khởi độn giống nhau.

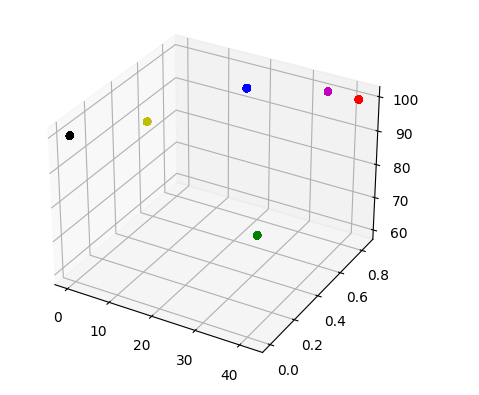


Hình Ảnh các cảm biến vận vận hành

Thuật toán k-means là một thuật toán phân cụm (clustering) dữ liệu trong khai phá dữ liệu (data mining) và học máy (machine learning). Nó được sử dụng để tìm cụm (cluster) dữ liệu trong một tập hợp dữ liệu cho trước.

Thuật toán k-means bắt đầu bằng cách chọn k cụm ban đầu ngẫu nhiên và gán từng điểm dữ liệu vào cụm gần nó nhất. Sau đó, trung tâm của mỗi cụm được tính bằng trung bình cộng của tất cả các điểm dữ liệu trong cụm. Tiếp theo, các điểm dữ liệu được gán lại vào cụm có trung tâm gần nhất. Quá trình này lặp lại cho đến khi không còn thay đổi trong việc phân cụm.

Sau khi phân cụm ta được kết quả như sau:



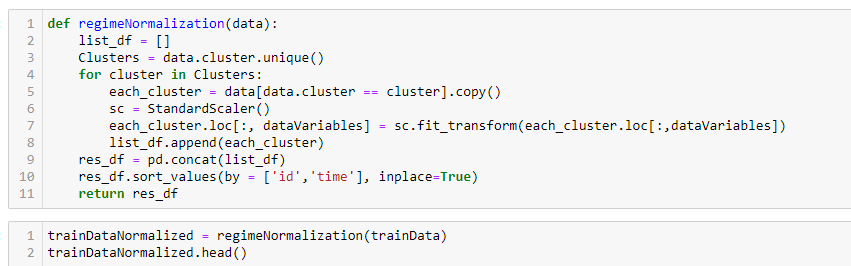
Hình Ảnh các cảm biến vận hành sau khi đã phân cụm

### Chuẩn hoá dữ liệu.

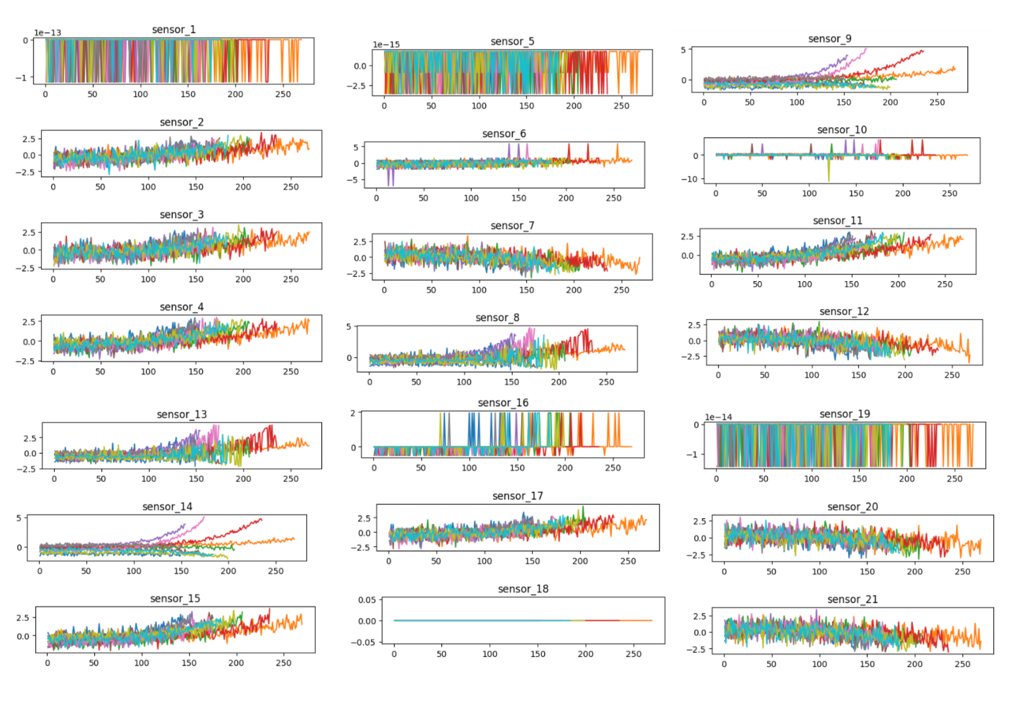
Sau khi phân cụm các loại hoạt động xong, thì tiến hành chuẩn hoá dữ liệu. Với mỗi cụng sẽ chuẩn hoá riêng, ở đây chúng ta dùng StandardScaler để chuẩn hoá dữ liệu.

StandardScaler là một trong những kỹ thuật chuẩn hóa dữ liệu phổ biến trong machine learning. Khi áp dụng StandardScaler, các giá trị của các thuộc tính dữ liệu sẽ được biến đổi sao cho có trung bình bằng 0 và độ lệch chuẩn bằng 1. Việc chuẩn hóa dữ liệu này có rất nhiều lợi ích trong quá trình huấn luyện mô hình, bao gồm:

* Loại bỏ sự phụ thuộc vào tỷ lệ của các thuộc tính.
* Giảm thiểu ảnh hưởng của các ngoại lệ (outliers).
* Cải thiện tốc độ hội tụ của các thuật toán tối ưu hóa.
* Cải thiện hiệu suất của mô hình.



Sau khi chuẩn hóa dữ liệu theo từng cụm có thể dễ dàng thấy thay đổi của các cảm biến sau các lần chạy đến gần hỏng thì sẽ có thay đổi.

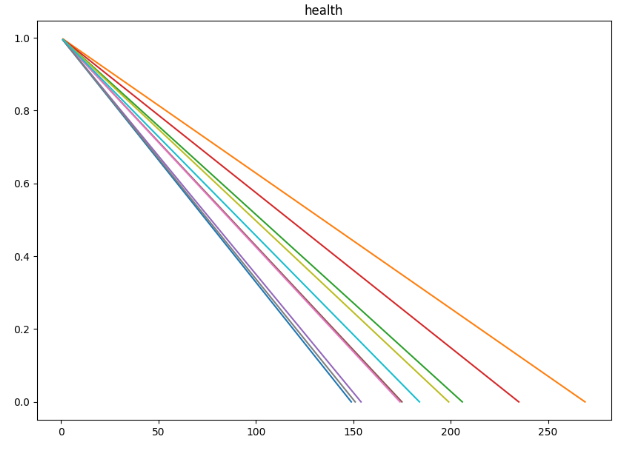


*Hình ảnh 21 sensor sau khi chuẩn hoá dữ liệu*

### Xây dựng chỉ số sức khỏe.

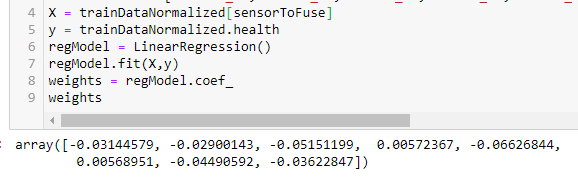
Phần này tập trung vào việc hợp nhất các phép đo của cảm biến thành một chỉ số sức khỏe duy nhất, trong đó một mô hình dựa trên sự tương đồng được đào tạo.

Tất cả dữ liệu chạy đến hỏng hóc được cho là bắt đầu với tình trạng khỏe mạnh. Tình trạng sức khỏe lúc đầu được gán giá trị là 1 và tình trạng sức khỏe khi thất bại được gán giá trị 0. Tình trạng sức khỏe được giả định là suy giảm tuyến tính từ 1 đến 0 theo thời gian. Sự xuống cấp tuyến tính này được sử dụng để giúp kết hợp các giá trị cảm biến.

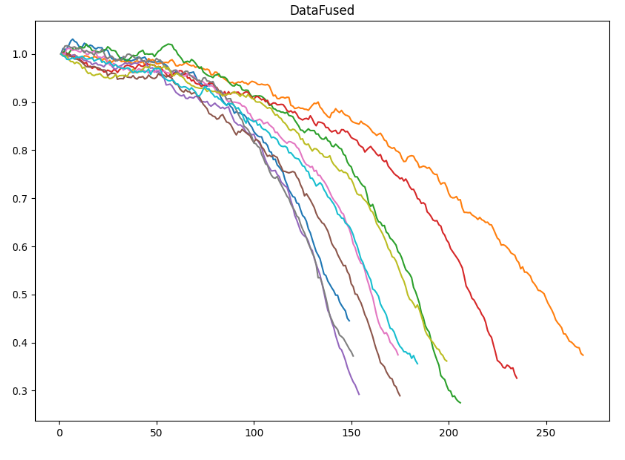
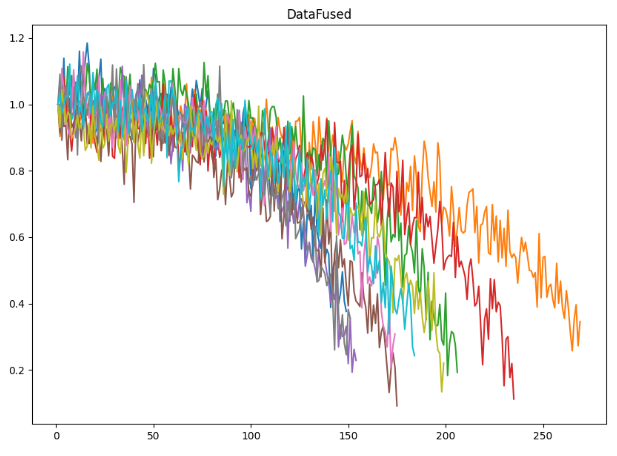


Hình Ảnh chỉ số sức khoẻ của 10 xe

Tình trạng sức khỏe của tất cả các thành viên trong nhóm thay đổi từ 1 thành 0 với tốc độ suy giảm khác nhau.



Đưa các thông tin sensor vào mô hình linearRegression để tạo ra đường sức khỏe của mỗi máy. Mỗi máy sau khi training xong sẽ có một đường chỉ số sức khỏe mới, sau đó tiến hành sử dụng kĩ thuật làm mịn (data smothing) để dễ dàng xử lý với bước tiếp theo hơn



*Hình 2.19. Ảnh trước và sau khi dùng kĩ thuật làm mịn*

Làm mịn (data smoothing) là một kỹ thuật trong xử lý tín hiệu để giảm nhiễu hoặc các biến động ngắn hạn trong dữ liệu bằng cách tạo ra một phiên bản mới của dữ liệu với các giá trị trơn tru hơn. Mục đích của data smoothing là để giúp tìm ra các xu hướng chung và giảm thiểu ảnh hưởng của các biến động ngắn hạn hoặc nhiễu trong dữ liệu.

Có nhiều phương pháp để thực hiện data smoothing, ví dụ như :

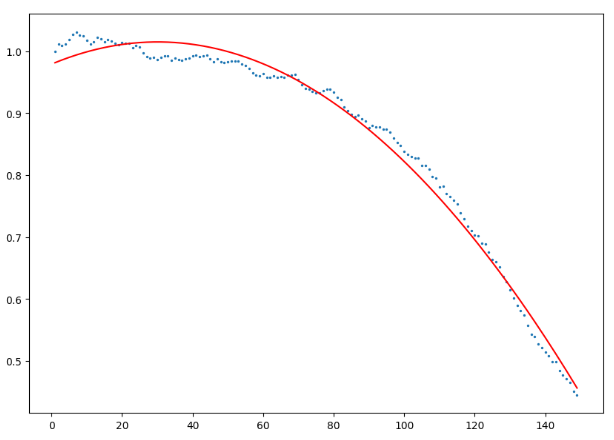
* Moving Average : phương pháp này tính trung bình của một tập hợp các giá trị liên tiếp và sử dụng kết quả này làm giá trị mới. Ví dụ, nếu chúng ta muốn áp dụng moving average với cửa sổ kích thước 3, thì giá trị mới tại mỗi thời điểm sẽ là trung bình của giá trị hiện tại và hai giá trị trước đó.
* Exponential Smoothing : phương pháp này sử dụng một hệ số mô hình để trọng số các giá trị cũ và mới, và tính toán giá trị mới dựa trên các giá trị trước đó. Hệ số mô hình thường được đặt tùy ý và phải được điều chỉnh để phù hợp với dữ liệu cụ thể.
* Median Smoothing : phương pháp này sử dụng giá trị trung vị của một tập hợp các giá trị liên tiếp để tính toán giá trị mới. Phương pháp này thường được sử dụng khi dữ liệu có nhiễu hoặc outlier.

### Sử dụng hàm đa thức

Hàm đa thức (polynomial) là một kỹ thuật trong machine learning để biến đổi các đặc trưng ban đầu của một mẫu dữ liệu thành các đặc trưng mới, bao gồm các lũy thừa và các tích của các đặc trưng ban đầu. Điều này cho phép mô hình có thể học được các quan hệ phi tuyến giữa các đặc trưng và đầu ra một cách hiệu quả hơn.

Cụ thể, trong kỹ thuật này, các đặc trưng ban đầu được nâng lên bậc, bao gồm lũy thừa bậc 2, bậc 3 và cao hơn. Ví dụ, nếu ban đầu chúng ta chỉ có một đặc trưng là chiều cao của một người, thì bằng cách sử dụng Polynomial features, chúng ta có thể tạo ra các đặc trưng mới bao gồm chiều cao bình phương, chiều cao mũ 3, v.v.

Việc sử dụng Polynomial features có thể giúp cho mô hình có khả năng học được các quan hệ phức tạp hơn giữa các đặc trưng và đầu ra. Tuy nhiên, việc sử dụng quá nhiều đặc trưng mới có thể dẫn đến hiện tượng overfitting, vì mô hình có thể học được quá nhiều chi tiết không cần thiết trong dữ liệu. Do đó, cần thận trọng khi sử dụng kỹ thuật Polynomial features và điều chỉnh tham số để đạt được kết quả tốt nhất.



*Hình 2.20. Hàm đa thức bậc 2*

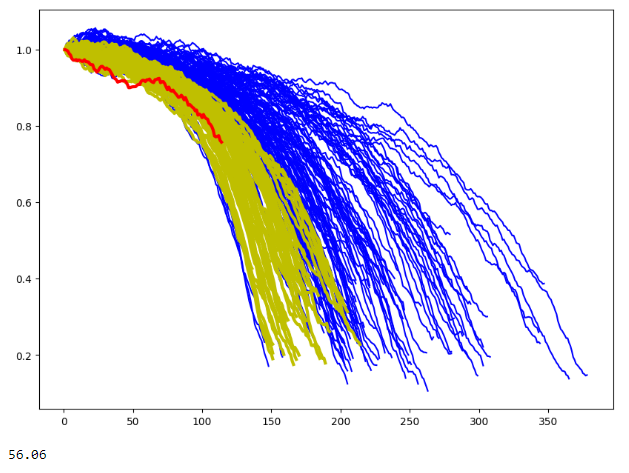
Bây giờ chúng ta sẽ xây dựng mô hình RUL sử dụng dữ liệu tương ứng. Bước này, mỗi model sẽ huấn luyện một hàm bậc hai sao cho khớp với dữ liệu.

Khoảng cách giữa dữ liệu và dữ liệu được tính bằng 1-norm:

Với là chỉ số sức khoẻ của máy , là chỉ số sức khoẻ ước tính của máy sử dụng đa thức bậc hai được xác định trong máy .

Khoảng cách tương đồng sẽ được tính theo công thứ sau:

Từ các khảng cách này, mô hình sẽ tìm 50 thành viên nhóm gần nhất trong tập dữ liệu huấn luyện, điều chỉnh phân phối xác xuất dựa trên 50 thành viên nhóm và sử dụng giá trị trung bình của phân phối làm ước tính RUL.



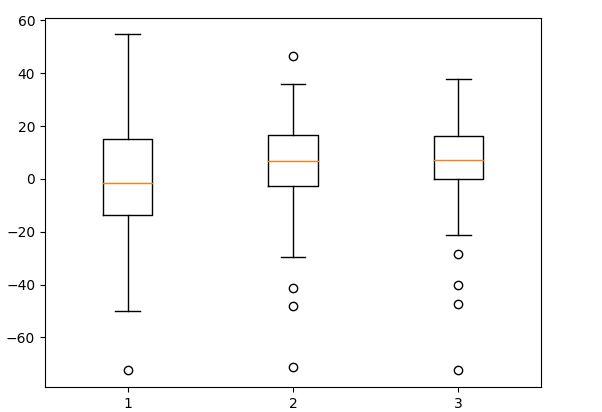
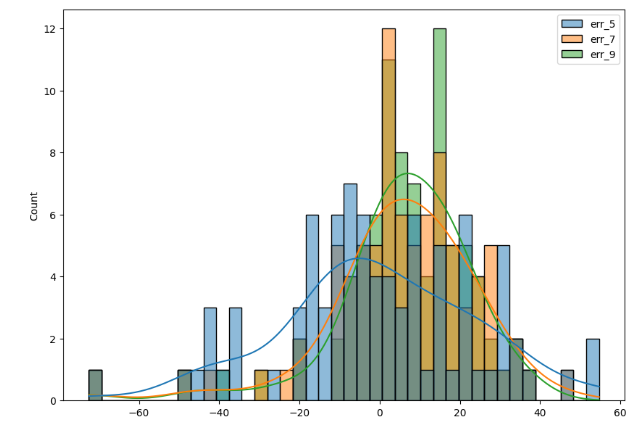
*Hình Kết quả khi thử nghiệm 1 máy*

Mô tả biểu đồ trên:

* Đường màu đỏ – Là mô tả mô mình 1 máy đang chạy là máy cần dự đoán.
* Các đường màu vàng – Là các đường chạy từ lúc bắt đầu đến lúc hỏng của 50 máy có khoảng cách gần đến màu màu đỏ nhất.
* Các đường màu xanh – Là các đường chạy của tất cả dữ liệu huấn luyện từ lúc chạy đến lúc hỏng.
* Trục x: mô tả khoảng thời gian của máy.
* Trục y: mô tả giả thiết suy hao khi chạy của động cơ, càng giảm về không động cơ càng khó khả năng hỏng cao.

### Đánh giá hiệu suất mô hình

Lặp lại quy trình đánh giá tương tự cho toàn bộ dữ liệu xác thực và tính toán sai số giữa ước tính RUL và RUL thực cho mỗi điểm dừng là: 50%, 70%, 90%.



Hình. Biểu đồ histogram và boxplot để đánh giá dữ liệu

Nhìn bảo biểu đồ ta thấy, kết quả vẫn còn sai lệch khá là nhiều. Dù chỉ khi máy chạy đến 70% hay 90% cho đến khi hỏng thì kết quả chỉ có xu hướng ổn định hơn chứ không tốt hơn đảng kể.

# CHƯƠNG 2. PHÂN TÍCH THIẾT KẾ HỆ THỐNG

## 2.1. Phân tích hệ thống

### Vai trò của người dùng

* Hệ thống chia làm 3 nhóm người dùng chính:
  + Khách truy cập vãng lai
  + Khách hàng
  + Người quản trị
* Mô tả nhóm người dùng.
* Khách truy cập vãng lai: những người dùng truy cập trang web và chưa có tài khoản đăng nhập trên trang web. Xem giới thiệu về mô hình dự đoán, có thể gửi thông tin của mình nếu muốn hợp tác với công ty.
* Khách hàng (thành viên của trang web): những khách hàng đã có tài khoản đăng nhập trên trang web. Người dùng có thể quan sát thông tin của máy mình khi đang vận hành, nhận được cảnh báo khi lỗi, xuất báo cáo.
* Người quản trị: là người quản lý trang web và được truy cập vào các chức năng quản lý của hệ thống.

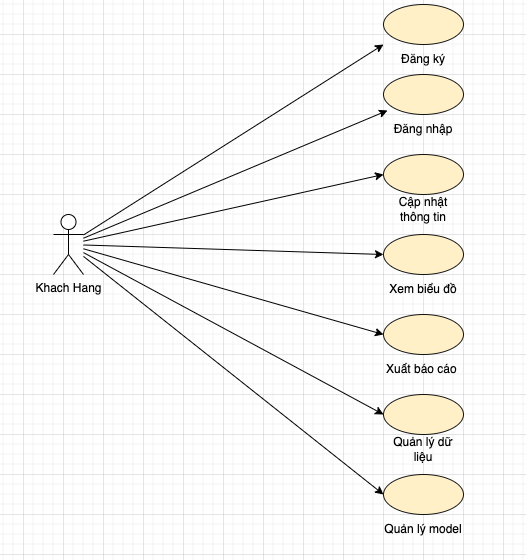
### Yêu cầu chức năng

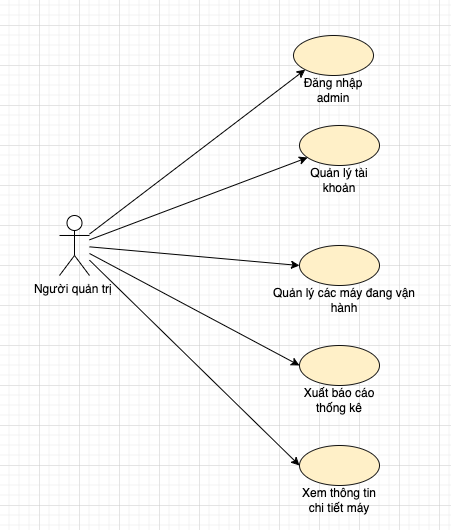
* Khách truy cập vãng lai:
  + Đăng ký tài khoản
  + Xem giới thiệu về bài toán bảo trì dự đoán
* Thành viên của trang web:
  + Đăng nhập
  + Đăng xuất
  + Xem các biểu đồ của máy khi đang hoạt động
  + Xuất báo cáo
  + Sửa thông tin cá nhân
  + Quên mật khẩu
* Quản trị viên:
  + Sửa thông tin cá nhân
  + Quản lý tài khoản
  + Quản lý các máy đang vận hành
  + Xuất báo cáo tổng thể
  + Xem thông tin chi tiết máy

### Yêu cầu phi chức năng

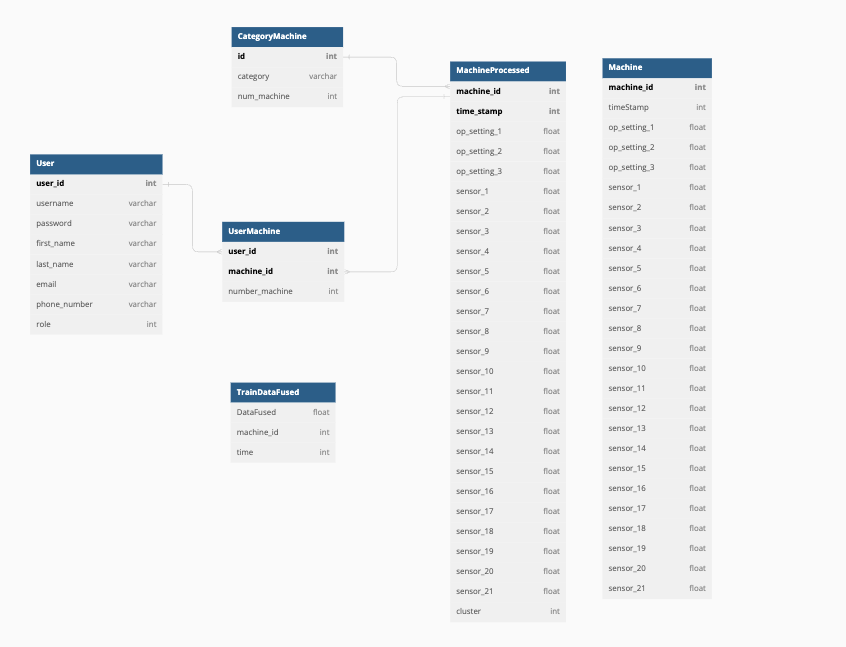
* Giao diện trang web được thiết kế dễ sử dụng và có tính thẩm mỹ cao.
* Được viết trên nền tảng web
* Hiệu năng: Tốc độ, khả năng, độ tin cây của hệ thống.
* Website bảo mật phân quyền để thực hiện các chức năng khác của hệ thống

## Biểu đồ use case





## Mô hình thực thể liên kết



## Đặc tả use case

### Use case đăng nhập

1. Biểu đồ use case:



1. Đặc tả use case:

Use case này cho phép người dùng đăng nhập để xác định quyền truy cập vào hệ thống.

* Luồng sự kiện:
  + Luồng cơ bản:

1. Use case này bắt đầu khi người dùng bấm vào biểu tượng người dùng sau đó bấm vào nút “Đăng nhập” trên góc phải màn hình.
2. Hệ thống yêu cầu người dùng nhập các thông tin bao gồm: tên đăng nhập và mật khẩu.
3. Người dùng nhập các thông tin: tên đăng nhập và mật khẩu và bấm nút “Đăng nhập”.
4. Hệ thống kiểm tra tên đăng nhập và mật khẩu đã nhập và chuyển hướng sang trang chủ.
5. Use case kết thúc.
   * Luồng rẽ nhánh:
6. Sai tên đăng nhập hoặc mật khẩu: tại bước 3 trong luồng cơ bản nếu người dùng nhập tên đăng nhập hoặc mật khẩu sai, hệ thống hiển thị thông báo lỗi. Người dùng có thể nhập lại hoặc bỏ qua thao tác, khi đó use case kết thúc.
7. Đăng ký: tại bước 3 trong luồng cơ bản, nếu người dùng bấm nút “Đăng ký” use case kết thúc.
8. Quên mật khẩu: tại bước 3 trong luồng cơ bản, nếu người dùng bấm nút “Quên mật khẩu” use case kết thúc.
9. Tại bất kỳ thời điểm nào trong quá trình thực hiện use case nếu không kết nối được với cơ sở dữ liệu thì hệ thống sẽ hiển thị thông báo lỗi và use case kết thúc.

* Các yêu cầu đặc biệt:

Không có.

* Tiền điều kiện:

Không có.

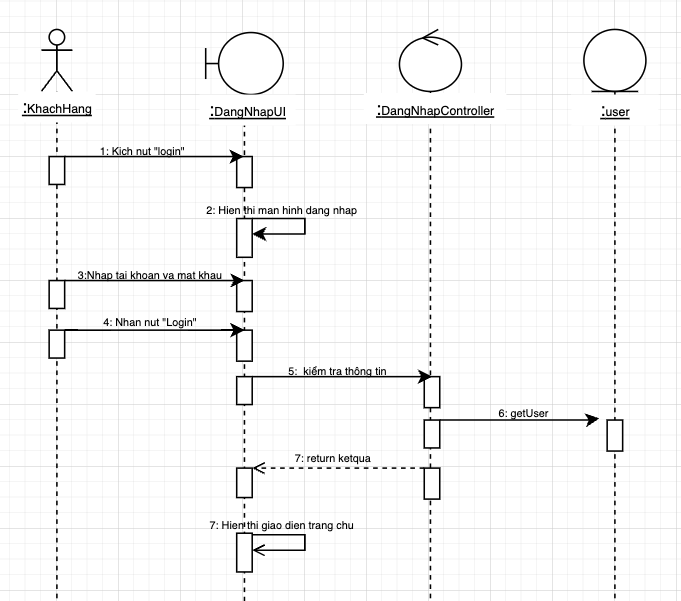
* Hậu điều kiện:

Nếu use case thành công, người dùng sẽ đăng nhập được vào hệ thống. Nếu không trạng thái của hệ thống không thay đổi.

* Điểm mở rộng:

Không có.

1. Biểu đồ trình tự



*Hình. Biểu đồ trình tự đăng nhập*

## Thiết kế hệ thống

### Biểu đồ triển khai hệ thống

### Thiết kế cơ sở dữ liệu

### Thiết kế giao diện

# CHƯƠNG 3. KẾT QUẢ ĐẠT ĐƯỢC

# KẾT LUẬN

# TÀI LIỆU THAM KHẢO

# PHỤ LỤC