# Phần 1: Tổng quan

## 1.1 Giới thiệu về học máy

Những năm gần đây, AI - Artificial Intelligence (Trí Tuệ Nhân Tạo), và cụ thể hơn là Machine Learning (Học Máy hoặc Máy Học) nổi lên như một bằng chứng của cuộc cách mạng công nghiệp lần thứ tư (1 - động cơ hơi nước, 2 - năng lượng điện, 3 - công nghệ thông tin). Trí Tuệ Nhân Tạo đang len lỏi vào mọi lĩnh vực trong đời sống mà có thể chúng ta không nhận ra. Xe tự hành của Google và Tesla, hệ thống tự tag khuôn mặt trong ảnh của Facebook, trợ lý ảo Siri của Apple, hệ thống gợi ý sản phẩm của Amazon, hệ thống gợi ý phim của Netflix, máy chơi cờ vây AlphaGo của Google DeepMind, …

**Machine Learning** là một tập con của AI có khả năng tự học hỏi dựa trên dữ liệu đưa vào mà không cần phải được lập trình cụ thể với 2 kỹ thuật chủ yếu là **Học có giám sát** (Supervised Learning) và **Học không giám sát** (Unsupervised Learning) Trong đó:

**Học tập có giám sát** là nơi bạn có các biến đầu vào (X) và biến đầu ra (Y) và bạn sử dụng thuật toán để tìm hiểu hàm ánh xạ từ đầu vào đến đầu ra. Mục đích là để xây dựng hàm ánh xạ một cách tốt nhất có thể để khi bạn có dữ liệu đầu vào mới (X) và bạn có thể dự đoán các biến đầu ra (Y) cho dữ liệu đó. Mô hình được huấn luyện bằng cách giảm thiểu sai số lỗi **(loss)** của các dự đoán tại các vòng lặp huấn luyện. Sau quá trình huấn luyện. mô hình sẽ có khả năng đưa ra dự đoán về đầu ra với một đầu vào mới gặp (không có trong dữ liệu học). Nếu không gian đầu ra được biểu diễn dứới dạng rời rạc, ta gọi đó là bài toán phân loại **(classification)**. Nếu không gian đầu ra được biểu diễn dưới dạng liên tục, ta gọi đó là bài toán hồi quy **(regression)**

**Học không giám sát:** Là cách huấn luyện một mô hình trong đó dữ liệu học chỉ bao gồm đầu vào mà không có đầu ra. Mô hình sẽ được huấn luyện cách để tìm cấu trúc hoặc mối quan hệ giữa các đầu vào(ví dụ như phân cụm,…)

Trong kỷ nguyên số hóa, công nghệ là yếu tố thiết yếu giúp doanh nghiệp thực hiện những cải tiến, nâng cao vị thế cạnh tranh trên thị trường. Nhu cầu nguồn nhân lực chất lượng cao thông hiểu công nghệ  nổi bật liên quan đến AI như Big Data, Data Science, Machine Learning ngày càng nâng cao. Thị trường nhân lực ngành Công nghệ đã phát triển mạnh trong nửa đầu năm nay với số lượng đăng tuyển và ứng tuyển đều tăng cao. Các công ty sẽ đối mặt với thách thức giữ chân nhân tài trong lĩnh vực này bằng cách đề xuất mức lương, thưởng chế độ thăng tiến tương xứng.

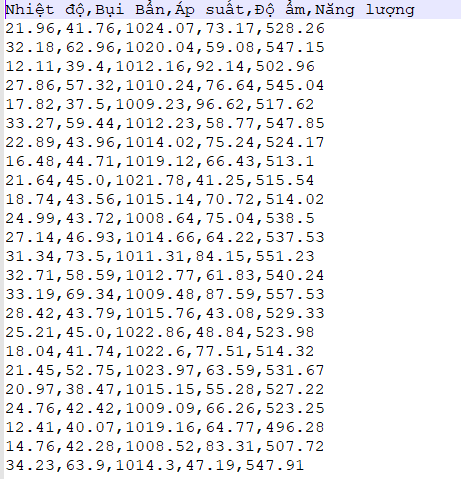
## 1.2 Giới thiệu về thuật toán Linear Regression

Project của nhóm em lần này sẽ sử dụng thuật toán Linear Regression (Hồi Quy Tuyến Tính) thuộc nhóm Supervised learning ( Học có giám sát ). Hồi quy tuyến tính (Linear Regression) là một phương pháp thống kê để thiết lập mối quan hệ giữa một biến phụ thuộc và một nhóm tập hợp các biến độc lập rất đơn giản nhưng đã được chứng minh được tính hữu ích cho một số lượng lớn các tình huống:

Như ta đã biết, chịu tác động của hiệu ứng nhà kính toàn cầu nên nhiệt độ đều có xu hướng biến động và tăng sau mỗi năm. Vì vậy nhóm em sẽ ứng dụng thuật toán Hồi quy tuyến tính nhằm biểu diễn nhiệt độ trung bình sau mỗi năm, ngoài biểu diễn dữ liệu, mô hình còn đưa ra dự đoán nhiệt độ cho thời gian sắp tới nhờ vào việc liên tục cập nhật dữ liệu nhiệt độ, thời tiết qua các năm. Việc này làm cho mô hình của ta càng đáng tin cậy hơn.

## 1.3 Trình bày bài toán

**Input**: Bảng dữ liệu trình bày thông tin về các giá trị tiêu thụ điện của sinh viên dựa vào các thông tin môi trường hiện tại như nhiệt độ, độ ẩm, tỷ lệ khói bụi hay áp suất nhiệt độ.



Ví dụ bảng dữ liệu đầu vào có dạng trên.

**Ouput**: Năng lượng điện tiêu thụ điện của sinh viên. Ở trên ví dụ là giá trị tại các cột năng lượng

## 1.4 Trình bày phương pháp học máy được sư dụng trong bài tập lớn

Bài toán của chúng em sử dụng phương pháp hồi quy tuyến tính

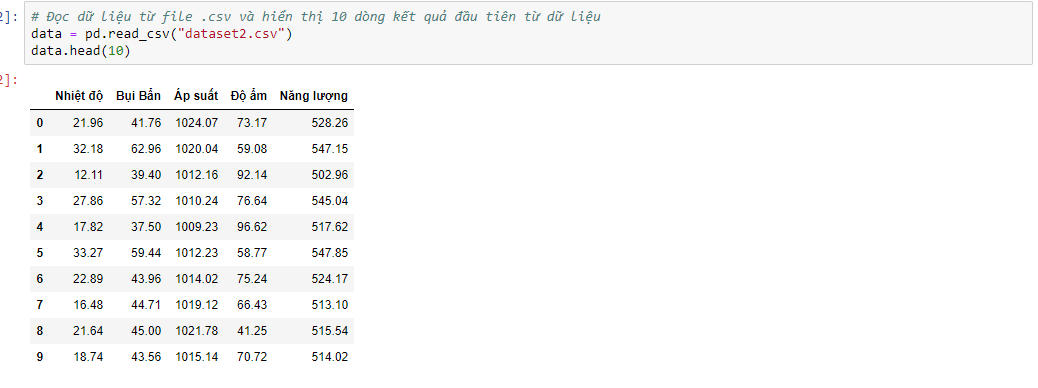
Trong việc phân tích dữ liệu, thuật ngữ Regression( Hồi quy) để nói lên mối quan hệ giữa 1 biến phụ thuộc và một nhóm tập hợp các biến độc lập

Trong khi sử dụng hồi quy tuyến tính, mục tiêu của chúng ta là để làm sao một đường thẳng có thể tạo được sự phân bố gần nhất với hầu hết các điểm. Vì vậy nó rất nạy cảm với dữ liệu đột biến/ dữ liệu nhiễu.

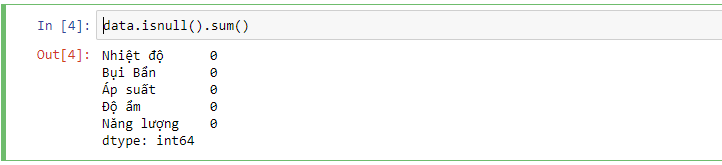
# Phần 2: Thực nghiệm

## 2.1 Mô tả tập ví dụ huấn luyện và tập test

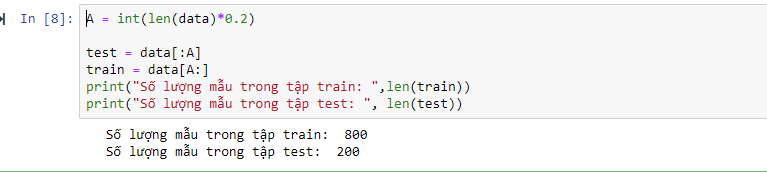
Bộ dữ liệu em thu thập bao gồm 1000 mẫu thông tin. Em sử dụng 80% trong toàn bộ dữ liệu để làm tập huấn luyện và 20% còn lại được sử dụng để làm tập kiểm tra. Do đó tập huấn luyện của em có 800 mẫu dữ liệu và tập kiểm tra là 200 dữ liệu.



Đoạn code đọc thông tin dữ liệu từ file dataset2.csv. Kết quả hiển thị trên 10 dòng dữ liệu. Chúng ta có thể thấy rằng 4 cột thông tin là nhiệt độ, bụi bẩn, áp suất và độ ẩm. Mục tiêu của chúng ta là dựa vào các giá trị thông tin này để dự đoán năng lượng điện tiêu thụ của sinh viên.

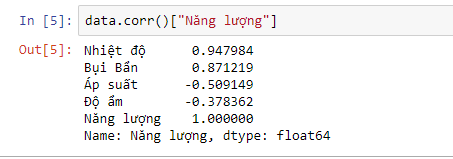


Đoạn code kiểm tra trong giá trị có chứa giá trị Null hay không. Nếu có chúng ta cần phải xử lý chúng trước khi đưa vào mô hình hồi quy tuyến tính.



### Phân tích dữ liệu

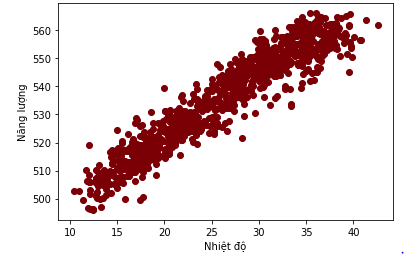
Trong bước này, em sẽ tiến hành phân tích dữ liệu để tìm hiểu mối liên quan giữa các thuộc tính và giá trị điện tiêu thụ. Đầu tiên em sẽ tìm hiểu mối quan hệ tuyến tính giữa các thuộc tính với giá trị đầu ra.



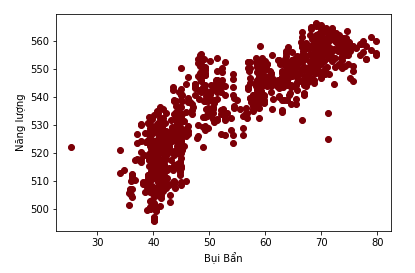
Chúng ta có thể thấy rằng nhiệt độ có và bụi bẩn có giá trị tương đồng với năng lượng tiêu độ điện. Điều này cũng dễ hiểu bởi vì trời nóng thì nhu cầu sử dụng năng lượng điện sẽ nhiều hơn so với bình thường . Tiếp theo em sẽ trực quan hóa mối quan hệ giữa các thuộc tính với giá trị đầu ra.

## 2.2 Ý tưởng giải quyết

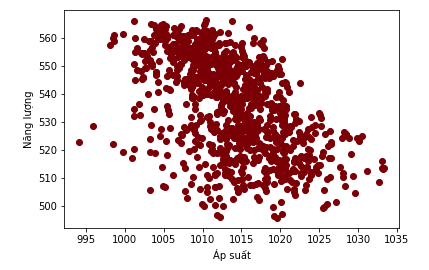
Bới nhiệt độ trung bình của các tháng được ghi nhận không có sự biến động quá lớn( dữ liệu gây nhiễu) nên lẽ dĩ nhiên ta sẽ tìm ra được mối quan hệ giữa nhiệt độ và các tháng trong năm thông qua 1 đường hồi quy tuyến tính với mức tin tưởng ở một ngưỡng nhất định nào đó.



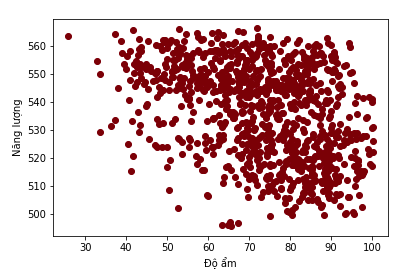
Hình 2: Gía trị tương quan giữa nhiệt độ và năng lượng điện tiêu thụ



Hình 3: Gía trị tương quan giữa bụi bẩn và năng lượng điện tiêu thụ



Hình 4: Gía trị tương quan giữa áp suất không khí và năng lượng điện tiêu thụ



Hình 4: Gía trị tương quan giữa độ ẩm và năng lượng điện tiêu thụ

## 2.3 Mô tả chương trình demo

Sau khi phân tích dữ liệu xong, em sẽ xây dựng chương trình minh họa thuật toán hồi quy tuyến tính để huấn luyện trên tập dữ liệu huấn luyện và kiểm tra trên tập dữ liệu kiểm tra. Để đánh giá mô hình hồi quy tuyến tính em sử dụng các độ đo như sau:

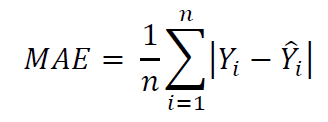
\* Root Mean Square Error (RMSE) RMSE cho biết mức độ phân tán các giá trị dự đoán từ các giá trị thực tế. Công thức tính RMSE là:



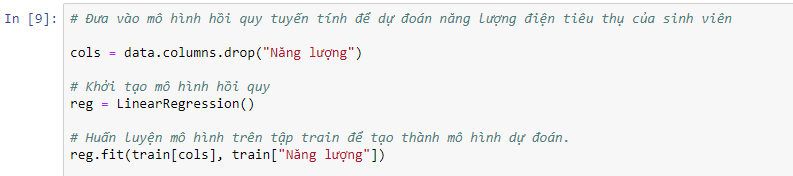
\* Mean Square Error: Sai số bình phương trung bình được tính theo công thức như sau.

# Machine learning: an introduction to mean squared error and ...

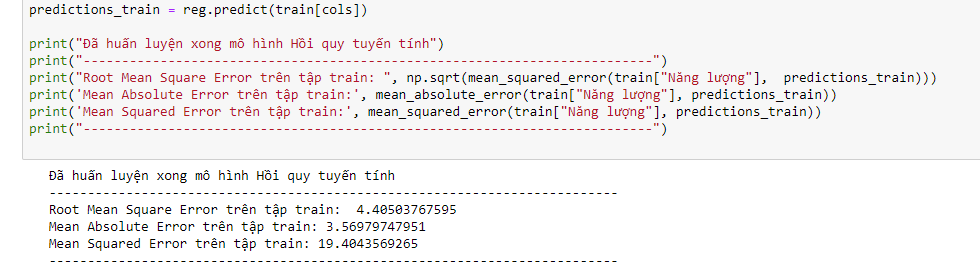
\* Mean Absolute Error được tính theo công thức như sau:



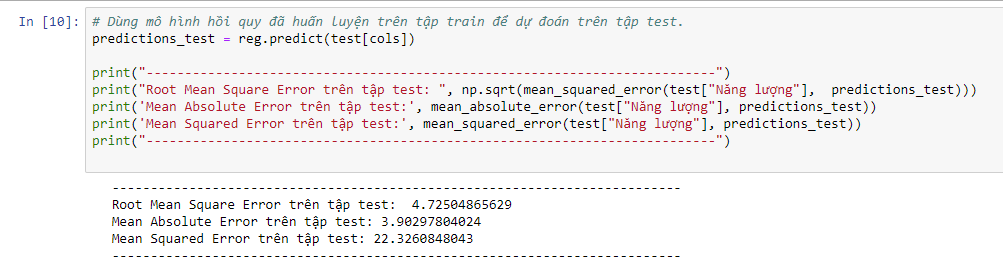
Đầu tiên chúng ta khởi tạo mô hình hồi quy tuyến tính và đưa dữ liệu huấn luyện vào mô hình học. Mô hình được khởi tạo như sau.



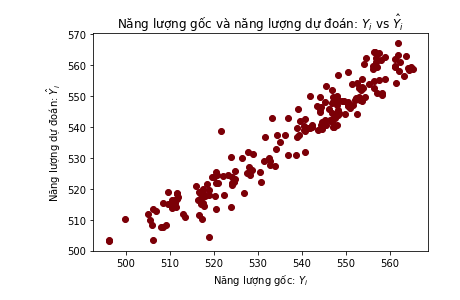
Sau khi huấn luyện mô hình xong, chúng ta sẽ đánh giá lại trên tập huấn luyện để xem hiệu quả mô hình sau khi học như thế nào.



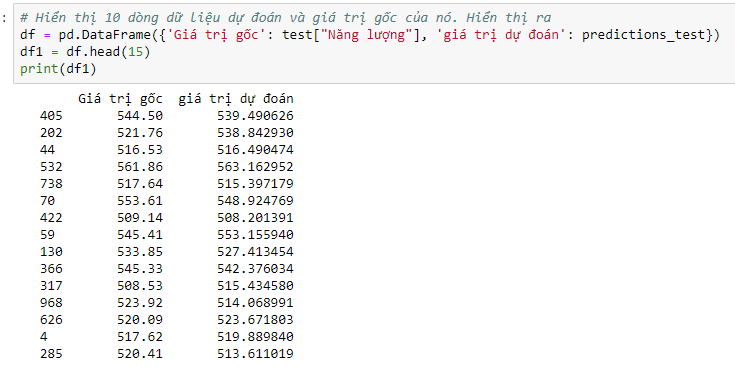
Kết quả trên chúng ta thấy rang độ đo RMSE là 4.4 còn các độ đo như MAE là 3.56 và MSE là 19.4. Tỷ lệ kết quả đạt tương đối chưa hiệu quả trên bộ dữ liệu này. Tiếp theo em sẽ đưa dữ liệu kiểm tra vào dự đoán kết quả dự đoán xem thế nào.



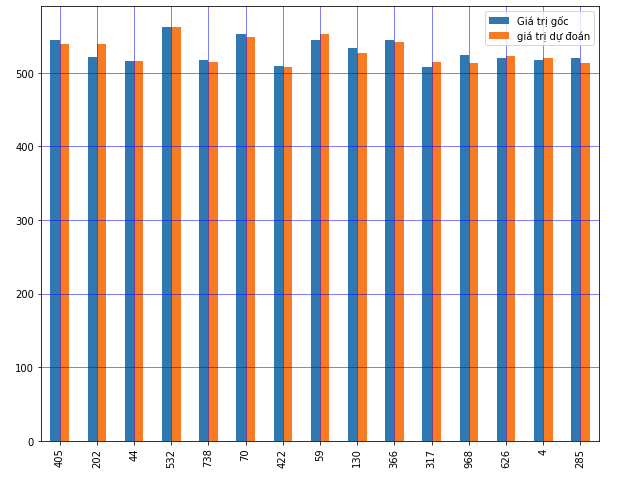
Kết quả thấy rằng mô hình dự đoán trên tập test cũng tương đối so với tập train nhưng tỷ lệ sai lệnh tương đối. Sau đó em trực quan hóa dữ liệu dự đoán giữa năng lượng tiêu thụ gốc và năng lượng dự đoán. Chúng ta thấy các điểm dữ liệu tang dần theo một cách tuyến tính.



Sau đó em tiếp tục thể hiện kết quả dự đoán trên 10 mẫu đầu tiên trong tập test để kiểm tra đánh giá giữa mô hình dự đoán và giá trị đúng thật sự của năng lượng nó là bao nhiêu. Kết quả được trình bày như sau:



Ví dụ 15 dòng dữ liệu đầu tiên trong tập kiểm tra bao gồm giá trị gốc và giá trị dự đoán của nó.



Biểu đồ thể hiện sự so sánh giữa giá trị năng lượng tiêu thụ gốc và năng lượng tiêu thụ của mô hình hồi quy dự đoán.

# 3. Kết luận:

Bài toán của nhóm em ứng dụng hoàn toán kiến thức đã học được từ các thầy cô cũng như kiến thức tìm hiểu được về Học máy nói chung và Hồi quy tuyến tính nói riêng nhằm đưa ra một cách chính xác, đáng tin cậy nhất về dự báo tiêu thụ điện .Tuy nhiên do quỹ thời gian và kiến thức hạn chế nên bài toán, ý tưởng của nhóm chưa đi đến được như ý muốn. Nhóm rất mong nhận được những gợi ý và nhận xét của thầy/cô để đi tới 1 mô hình hoàn thiện, có ý nghĩa thực tế.

# 4. Tài liệu tham khảo

**Tham khảo:**

1. Các slide bài giảng của PGS Nguyễn Hữu Quỳnh và kiến thức được giảng dạy qua các buổi thực hành của thầy Cù Mạnh Dũng
2. Lý thuyết về Linear Regression trên trang Machine Learning cơ bản
3. <https://machinelearningcoban.com/2016/12/28/linearregression/>
4. <https://www.geeksforgeeks.org/ml-rainfall-prediction-using-linear-regression/>

**Thư viện:**

1. <https://numpy.org/>
2. <https://pandas.pydata.org/>
3. <https://matplotlib.org/>
4. <https://pypi.org/project/future/>