# **CHƯƠNG 1. INTRODUCTION**

## **Lý do chọn đề tài**

Hiện nay nhu cầu sống của người dân ngày càng tăng cao theo thu nhập bình quân đầu người. Xuất phát từ ý nghĩ “An cư lạc nghiệp” mà có ý định đầu tư một căn căn hộ chất lượng, cao cấp của người dân đã dần trở nên phổ biến. Nhưng không phải bất kỳ ai cũng có thể định giá chính xác cho ngôi nhà hoặc căn hộ của mình. Đôi khi những sự nhầm lẫn hoặc sự thiếu kinh nghiệm cũng có thể dẫn đến dự đoán sai và dẫn đến thua lỗ là điều không thể tránh khỏi. Ngay cả những doanh nghiệp về bất động sản thì đây cũng là một vấn đế khá nan giải. Vậy làm sao, dựa vào đâu để đưa ra một báo giá chính xác nhất, gần với thực tế nhất, cũng chính là lo nhóm em chọn đề tài “Dự báo giá nhà đất”.

## **Ý nghĩa**

Xuất phát từ ý định mang đến không chỉ người mua mà còn người bán một công cụ có thể báo giá chính xác nhất có thể, từ đó góp phần giúp thị trường bất động sản chuyên nghiệp hơn, giảm tối thiểu tổn thất thua lỗ nhất có thể.

## **Mục tiêu**

* Mang đến cho người dùng một công cụ báo giá gần thực tế nhất, đáng tin cậy.
* Dữ liệu cần cho việc dự đoán phải tinh gọn, không quá phức tạp. Nghĩa là bất kỳ ngôi nhà cũng đáp ứng đầy đủ những yêu cầu cho việc dự đoán.
* Mang ứng dụng phổ biến đến người quan tâm đến mua bán nhà nói riêng và bất động sản nói chung.
* Đáp ứng được nhu cầu phát triền đối với dữ liệu toàn cầu.

## **Triển khai**

### **Khảo sát, thu thập dữ liệu**

### **Phân tích, lọc dữ liệu phù hợp**

### **Định hướng giải quyết vấn đề, chọn giải thuật phù hợp**

### **Triển khai dự án theo giải thuật đã chọn**

### **Giới thiệu ứng dụng đến rộng rãi người dùng**

# **CHƯƠNG 2. LÝ THUYẾT – LINEAR REGRESSION**

## **Định nghĩa**

### **Linear là gì?**

Vì ta cần tìm ra 1 đường thẳng, đường thẳng trong tiếng Anh là line, linear là “có thể biểu diễn bằng 1 đường thẳng” — trong toán học Việt Nam còn gọi là “tuyến tính” (tuyến trong tiếng Hán Việt nghĩa là đường thẳng — tuyến tính là có tính chất của đường thẳng).

### **Regression là gì?**

Là một cái tên lấy từ bộ môn [thống kê (statistics)](https://medium.com/pymi/d%C3%B9ng-python-%C4%91%E1%BB%83-h%E1%BB%8Dc-th%E1%BB%91ng-k%C3%AA-8e41dfdaaf97), một thuật toán/khái niệm mang từ bên statistics về dùng.  
Trong ngành thống kê: regression là mối quan hệ giữa giá trị đầu ra (y) với các giá trị (biến) đầu vào (x, t …).

### **Linear regression là gì?**

Linear regression là một phương pháp (thuật toán) thuộc loại đơn giản nhất trong Machine Learning, trong tiếng Việt gọi là *hồi quy tuyến tính*, với mục đích TÌM (vẽ) ra một đường thẳng, sao cho nó đi qua hoặc đi gần nhất với các điểm cho trước. Từ một tập dữ liệu cho trước, khi ta vẽ được một đường thẳng như vậy, ta có thể đoán xem các điểm khác sẽ nằm ở đâu.

Có thể nói mục tiêu của giải thuật hồi quy tuyến tính là dự đoán giá trị của một hoặc nhiều biến mục tiêu liên tục (continuous target variable) **y** dựa trên một véc-to đầu vào **x**.

Linear regression đóng một vai trò quan trọng trong lĩnh vực [trí thông minh nhân tạo](https://en.wikipedia.org/wiki/Artificial_intelligence) như Machine Learning . Thuật toán hồi quy tuyến tính là một trong những thuật toán [học máy được giám sát](https://en.wikipedia.org/wiki/Supervised_learning) cơ bản do tính đơn giản tương đối và các thuộc tính nổi tiếng của nó.

## **Các giả định trong phân tích hồi qui tuyến tính**

Phân tích hồi qui tuyến tính không chỉ là việc mô tả các dữ liệu quan sát được trong mẫu (sample) nghiên cứu mà cần phải suy rộng cho mối liên hệ trong dân số (population). Vì vậy, trước khi trình bày và diễn dịch mô hình hồi qui tuyến tính cần phải dò tìm vi phạm các giả định. Nếu các giả định bị vi phạm thì các kết quả ước lượng không đáng tin cậy được.

Các giả định cần thiết trong hồi qui tuyến tính:

* là biến số cố định, không có sai sót ngẫu nhiên trong đo lường.
* Phần dư (trị số quan sát trừ cho trị số ước đoán) phân phối theo luật phân phối chuẩn
* Phần dư có trị trung bình bằng 0 và phương sai không thay đổi cho mọi trị
* Không có tương quan giữa các phần dư

## **Một vài tính chất của hồi quy tuyến tính**

* Đường hồi quy luôn luôn đi qua trung bình của biến độc lập (x) cũng như trung bình của biến phụ thuộc (y)
* Đường hồi qui tối thiểu hóa tổng của "Diện tích các sai số". Đó là lý do tại sao phương pháp hồi quy tuyến tính được gọi là "Ordinary Least Square (OLS)"
* Β1 giải thích sự thay đổi trong Y với sự thay đổi X bằng một đơn vị. Nói cách khác, nếu chúng ta tăng giá trị của X bởi một đơn vị thì nó sẽ là sự thay đổi giá trị của Y

## **Giải thuật**

1. **Bài toán tìm giá nhà**

* Nếu ta có số liệu về diện tích của 3 ngôi nhà, và giá của chúng đã bán được, liệu với ngôi nhà rộng 3000 mét vuông, ta nên bán với giá bao nhiêu?
* Vì thế dựa vào lợi thế chúng ta có được là Collection Giá nhà bán theo từng yếu tố thu thập được, chúng ta sẽ tiến hành phân tích theo nhiều hướng: Số lượng phòng ngủ được người mua ưa chuộng khi chọn mua nhà, tọa lạc , khu vực nhà được giá, diện tích căn nhà , Từ đó có được những định giá đúng nhất về giá trị căn nhà bằng việc kết hợp với các thuật toán trong Machine Learning

1. **Phân tích bài toán và chọn mô hình**

* Ban đầu ta chưa biết được phải giải bài này thế nào, nên ta vẽ số liệu này lên đồ thị 2 chiều, trục x sẽ chứa diện tích ngôi nhà, trục y sẽ chứa giá của ngôi nhà. Nhìn vào hình vẽ thì thấy có vẻ như, ta có thể vẽ 1 đường thẳng đi qua tất cả các điểm này. Vậy ta sẽ thử đi tìm đường thẳng đó.

1. **Đường thẳng trong toán học**

* Cụ thể trong bài toán này ta sẽ sử dụng các thuật toán : Linear regression, Gradient descend

Đường thẳng trong toán học được biểu diễn bằng biểu thức sau:

y = ax + b

mối quan hệ giữa diện tích ngôi nhà và giá nhà là một đường thẳng, ta có thể tính giá nhà bằng công thức:

giá nhà = a \* (diện tích ngôi nhà) + b

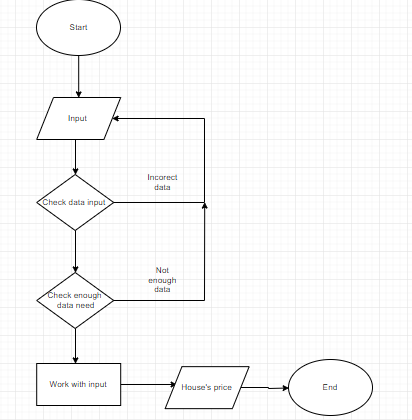
Ta đã có một tập giá nhà (y) và diện tích ngôi nhà (x), giờ chỉ việc đi tìm a và b. Sau khi tìm được a và b, ta có thể đoán (tính) được y khi có x.

Ở đây, ta sẽ sử dụng ký hiệu theo [lớp học ML của Stanford](https://www.coursera.org/learn/machine-learning):

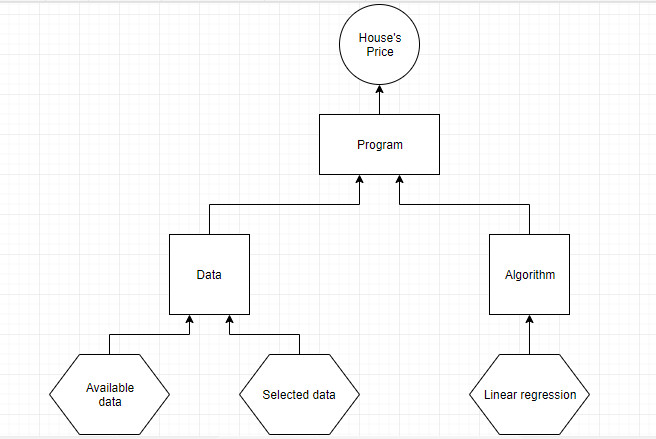
h(x) = theta\_0 + theta\_1 \* x

* Ta mong muốn vẽ được 1 đường thẳng đi qua tất cả các điểm, nhưng thực tế thì sẽ có vài điểm không nằm trên đường thẳng mà chỉ nằm gần. Vậy cách nào để đánh giá đường thẳng vừa vẽ có đủ tốt hay không? có đường nào tốt hơn không? Ta đưa ra một giá trị để đo xem đường thẳng này có tốt không, dễ thấy giá trị đó có thể là tính bằng tổng khoảng các của tất cả các điểm đã biết tới đường thẳng. Ta gọi tên giá trị này là “cost” (giá tiền), khi cost = 0, tức là điểm không cách đường thẳng — hay chúng nằm trên đường thẳng — đường thẳng ta vẽ ra là đúng tuyệt đối. Khi cost càng lớn, thì các điểm càng lệch khỏi đường thẳng, tức đường thẳng này không được tốt cho lắm. Hàm số biểu diễn các giá trị cost gọi là J(theta\_0, theta\_1). Mục tiêu của ta là đi tìm đường thẳng mà có giá trị J nhỏ nhất.
* Cost function là một hàm số, và đạo hàm là công cụ giúp ta tìm được xem tại 1 điểm, đồ thị đang đi lên hay đi xuống. Đó là điều duy nhất ta cần hiểu ở đây.
* Nếu derivative là khái niệm dành cho hàm số với một biến, thì cách nào để ta tìm được derivative đối với hàm h(x) có 2 biến theta\_1 và theta\_2?
* [Gradient](https://en.wikipedia.org/wiki/Gradient) là khái niệm tương tự như derivative, nhưng dành cho hàm nhiều biến (multi-variable)  
  Nếu derivative giúp ta biết hướng xuống dốc / lên dốc trong function 1 biến tại một điểm thì gradient giúp ta biết điều này trong function nhiều biến (2 trở lên)
* Gradient descend (là một thuật toán giúp tìm điểm cực tiểu cục bộ (local minimum) của hàm số bằng cách đi dần dần về phía giảm của gradient, và Các thư viện Python đã giúp ta tính sẵn rồi, không cần phải tính lại nữa, chỉ việc dùng function có sẵn )

# **CHƯƠNG 3. MODEL**

**FLOWCHART DIAGRAM**

**BLOCKED DIAGRAM**

****