



Bài giảng môn học:

**Kỹ nghệ tri thức và học máy (7080510)**

# **CHƯƠNG 3:**

# **HỌC CÓ GIÁM SÁT – Phần 4**

# **(Supervised Learning)**

**Giảng viên: Đặng Văn Nam**

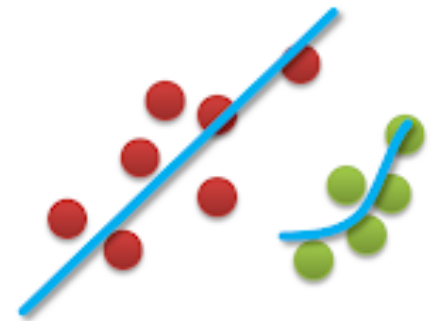
**Email: [dangvannam@hmg.edu.vn](mailto:dangvannam@hmg.edu.vn)**

## 1. Phân tích hồi quy (Regression Analysis)

1. Giới thiệu
2. Một số mô hình hồi quy cơ bản
3. Đánh giá độ chính xác của mô hình hồi quy
4. Bài tập thực hành

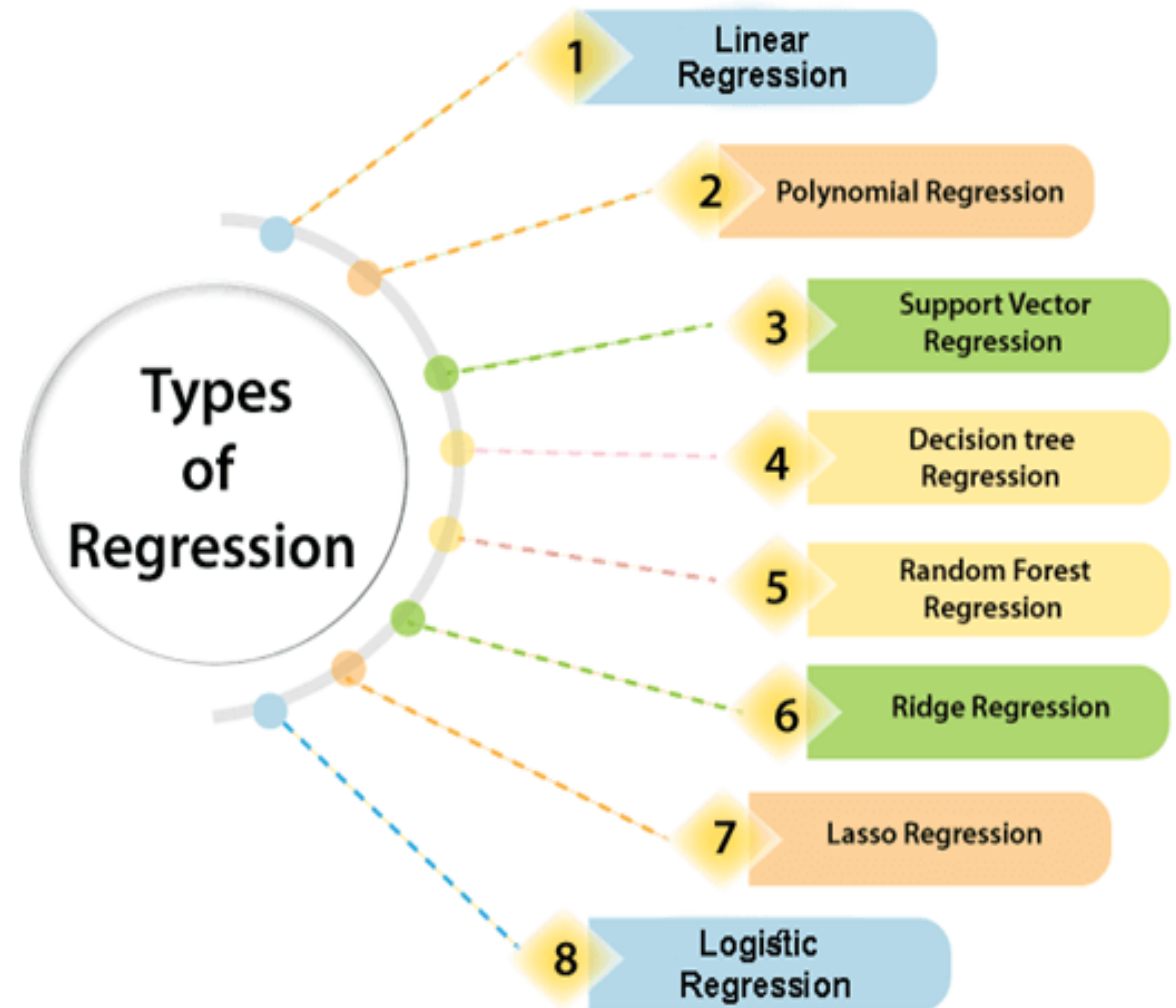
# 1. Giới thiệu bài toán hồi quy (Regression)

Regression



# Bài toán hồi quy

Advertisement	Sales
\$90	\$1000
\$120	\$1300
\$150	\$1800
\$100	\$1200
\$130	\$1380
\$200	??



## 2. Hồi quy tuyến tính (Linear Regression)

# Giới thiệu

Một căn nhà rộng  $x_1 \text{ m}^2$ , có  $x_2$  phòng ngủ và cách trung tâm thành phố  $x_3 \text{ km}$  có giá là bao nhiêu?



STT	Diện tích (m2)	Số phòng ngủ	Khoảng cách tới trung tâm thành phố (Km)	Giá bán (USD)
1	70	3	5	35,000
2	50	2	8	21,000
3	85	4	1	53,000
4	65	2	10	28,000
5	45	1	1	30,000
6	55	3	3	35,000
...	...	...	...	...

Giả sử chúng ta đã có số liệu thống kê từ 1000 căn nhà trong thành phố đó, liệu rằng khi có một căn nhà mới với các thông số về diện tích, số phòng ngủ và khoảng cách tới trung tâm, chúng ta có thể dự đoán được giá của căn nhà đó không?

Nếu có thì hàm dự đoán  $y=f(x)$  sẽ có dạng như thế nào?  
 Ở đây  $x=[x_1,x_2,x_3]$  là một vector hàng chứa thông tin *input*,  
 $y$  là một số vô hướng (scalar) biểu diễn *output* (tức giá của căn nhà trong ví dụ này).

STT	Diện tích (m2)	Số phòng ngủ	Khoảng cách tới trung tâm thành phố (Km)	Giá bán (USD)
1	70	3	5	35,000
2	50	2	8	21,000
3	85	4	1	53,000
4	65	2	10	28,000
5	45	1	1	30,000
6	55	3	3	35,000
...	...	...	...	...

- Một cách đơn giản nhất, chúng ta có thể thấy rằng:
  - i) diện tích nhà càng lớn thì giá nhà càng cao;
  - ii) số lượng phòng ngủ càng lớn thì giá nhà càng cao;
  - iii) càng xa trung tâm thì giá nhà càng giảm.

Một hàm số đơn giản nhất có thể mô tả mối quan hệ giữa giá nhà và 3 đại lượng đầu vào là:

$$y \approx f(x)$$

$$f(x) = w_1 x_1 + w_2 x_2 + w_3 x_3 + w_0 \quad (1)$$

**Trong đó:**  $w_1, w_2, w_3, w_0$  là các hằng số,  $w_0$  còn được gọi là bias.

Mối quan hệ  $y \approx f(x)$  bên trên là một mối quan hệ tuyến tính (linear).

Bài toán chúng ta đang làm là một bài toán thuộc loại hồi quy (regression). Bài toán đi tìm các hệ số tối ưu  $\{w_1, w_2, w_3, w_0\}$  chính vì vậy được gọi là bài toán Linear Regression.

# Hồi quy tuyến tính (Linear Regression).

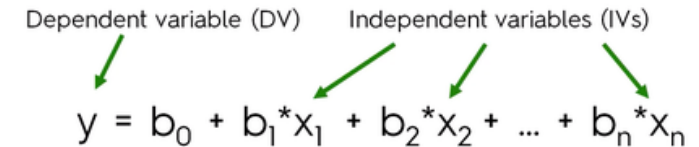
- Hồi quy tuyến tính với 1 biến độc lập  $X$  là biến đầu vào (input) để xác định 1 biến đầu ra  $y$  (target) – **Simple Linear Regression**.
- Hồi quy tuyến tính với  $n$  biến độc lập  $X_1, \dots, X_n$  để xác định 1 biến đầu ra  $y$  (target) – **Multiple Linear Regression**.

Simple  
Linear  
Regression

$$y = b_0 + b_1 * x_1$$

Multiple  
Linear  
Regression

Dependent variable (DV)      Independent variables (IVs)


$$y = b_0 + b_1 * x_1 + b_2 * x_2 + \dots + b_n * x_n$$

*Linear* hay *tuyến tính* hiểu một cách đơn giản là *thẳng, phẳng*.

- Không gian hai chiều, một hàm số được gọi là *tuyến tính* nếu đồ thị của nó có dạng một *đường thẳng*.
- Không gian ba chiều: một hàm số được gọi là *tuyến tính* nếu đồ thị của nó có dạng một *mặt phẳng*.
- Không gian nhiều hơn 3 chiều, khái niệm *mặt phẳng* không còn phù hợp nữa, thay vào đó, một khái niệm khác ra đời được gọi là *siêu mặt phẳng (hyperplane)*.



# Simple Linear Regression.

Hồi quy tuyến tính với 1 biến độc lập  $X$  là biến đầu vào (input) để xác định 1 biến đầu ra (target)

→ Xác định phương trình:

$$y = f(x)$$

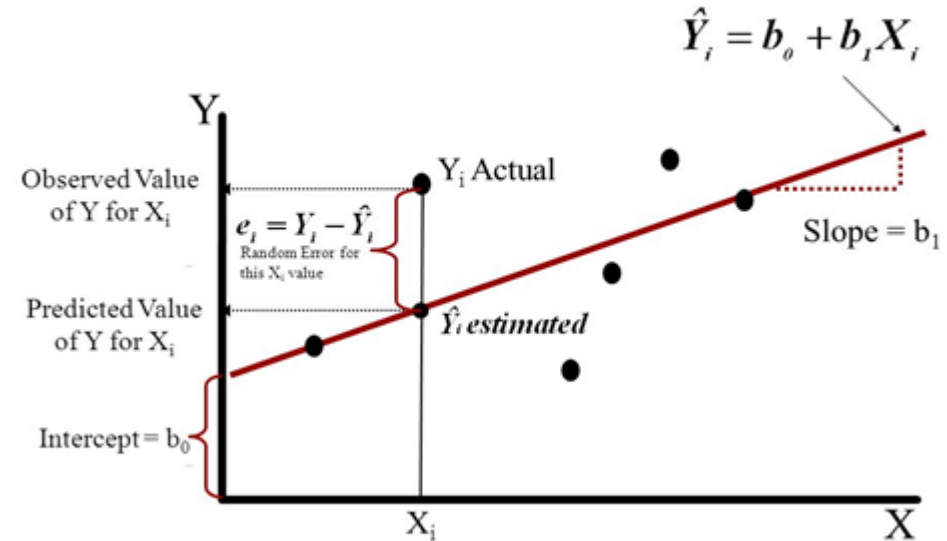
$$\hat{y} = \beta_0 + \beta_1 X$$

target
coefficients
input

Mục tiêu ước lượng các tham số  $b_i$  sao cho sai số nhỏ nhất.

$$RSS = \sum_i^n (y_i - \hat{y}_i)^2$$

Simple Linear Regression Model



- **y**: Giá trị thật trong tập train (Outcome).
- **$\hat{y}$** : Giá trị mà mô hình linear regression dự đoán được.

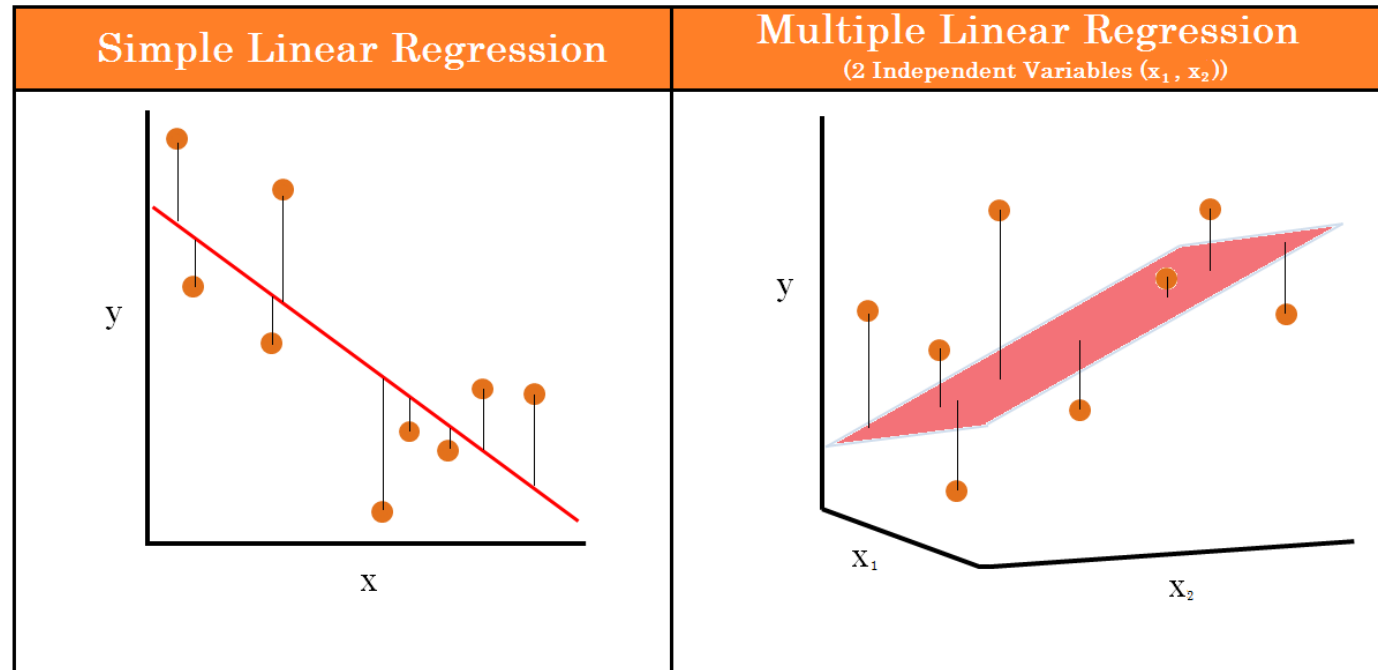
# Multiple Linear Regression.

Hồi quy tuyến tính với n biến độc lập ( $X_1, X_2 \dots X_n$ )

$$\hat{y} = \beta_0 + \beta_1 X_1 + \dots + \beta_n X_n$$

Diagram illustrating the components of the Multiple Linear Regression equation:

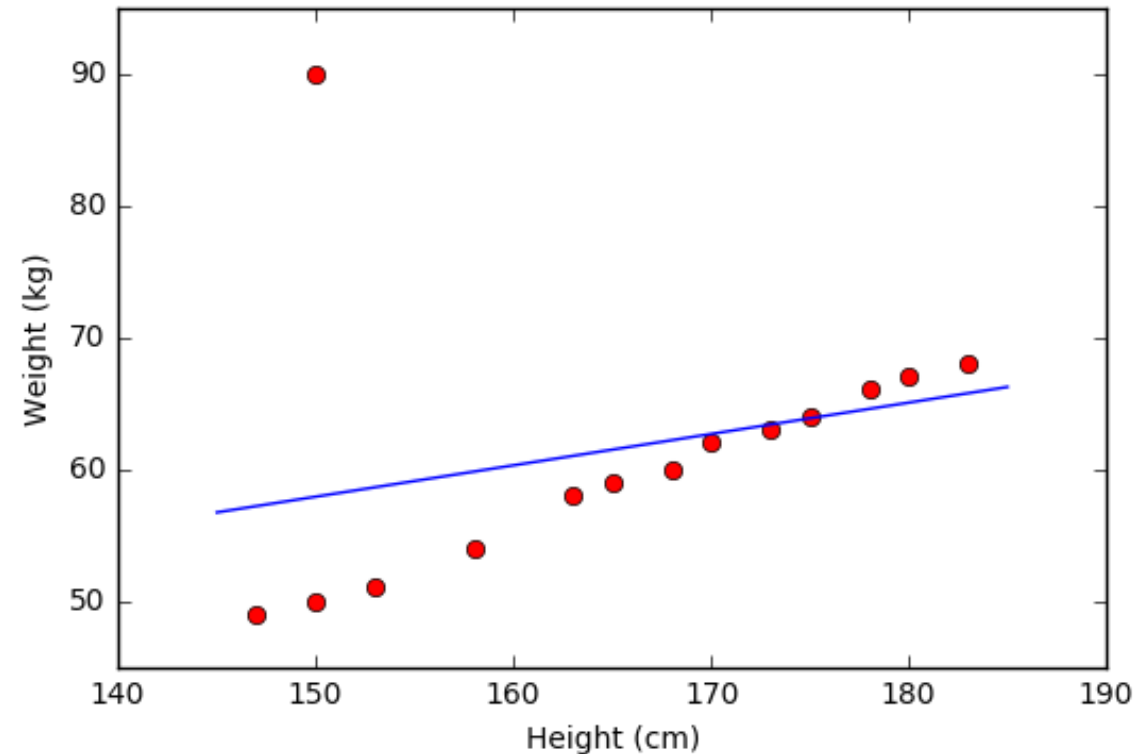
- $\hat{y}$  is labeled as the **target** (indicated by a pink arrow).
- $\beta_0, \beta_1, \dots, \beta_n$  are labeled as **coefficients** (indicated by grey arrows).
- $X_1, \dots, X_n$  are labeled as **inputs** (indicated by blue arrows).



# Nhược điểm của hồi quy tuyến tính



- Linear Regression **rất nhạy cảm với nhiễu** (sensitive to noise). Vì vậy trước khi thực hiện Linear Regression, các giá trị ngoại lai (outlier) cần phải được loại bỏ.
- Linear Regression **không biểu diễn được các mô hình phức tạp**.



# Ví dụ 1: Dự báo giá nhà

# Bài toán dự đoán giá nhà.



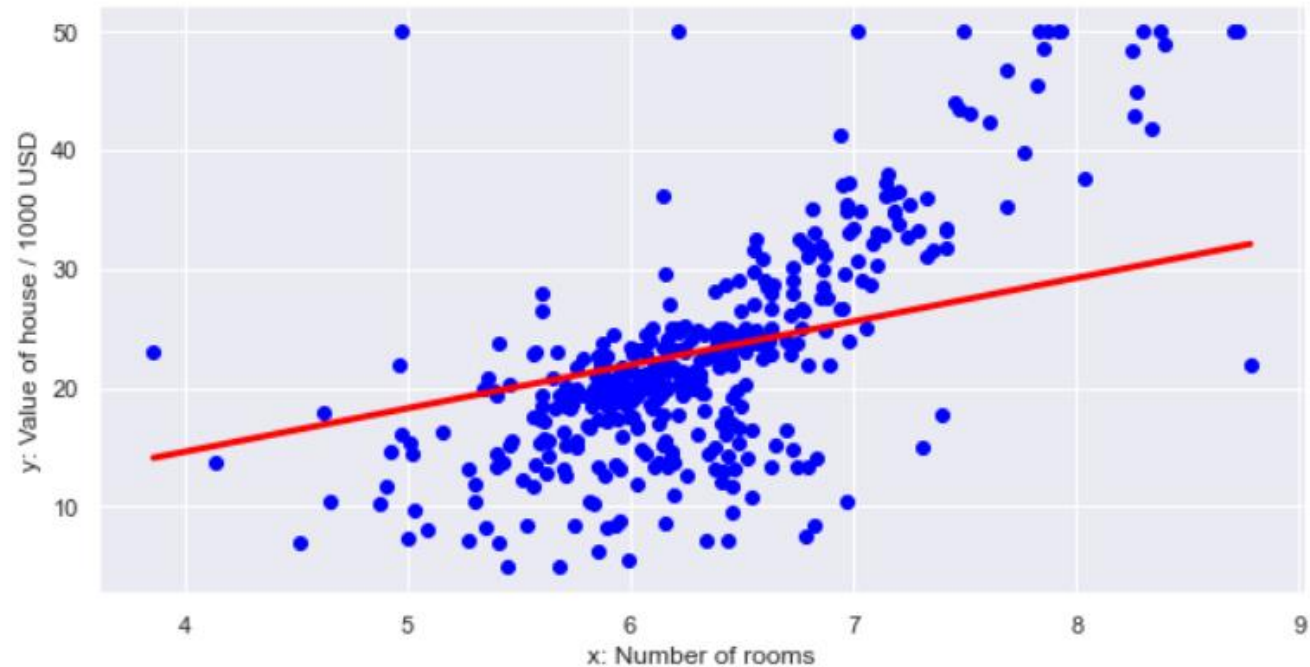
- Tập dữ liệu bao gồm 506 mẫu:
  - 13 thuộc tính đầu vào (features)
  - Thuộc tính target (MEDV)

**Tham khảo tiến trình thực hiện trong file code trên Jupyter Notebook**

# Simple Linear Regression.

Dự đoán giá nhà với 1 biến độc lập – RM (số phòng trung bình của căn nhà)

RM	MEDV
6.575	24.0
6.421	21.6
7.185	34.7
6.998	33.4
7.147	36.2
6.430	28.7
6.012	22.9
6.172	27.1
5.631	16.5
6.004	18.9
6.377	15.0



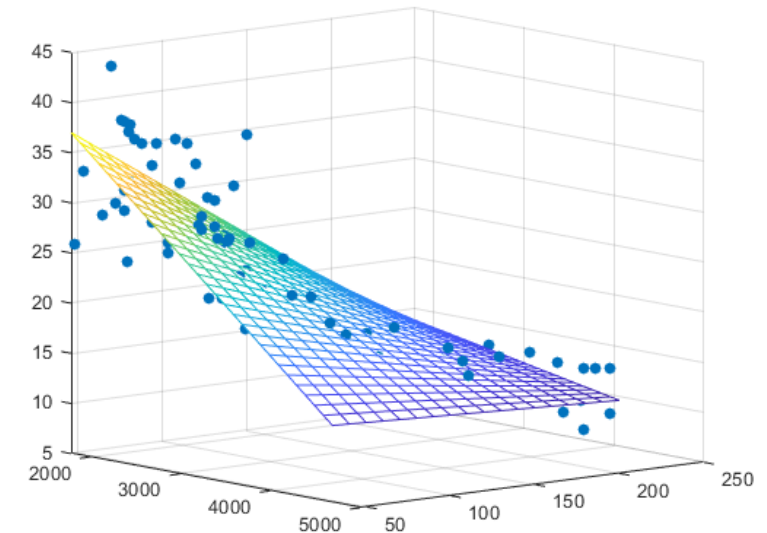
$$\hat{y}_{\text{MEDV}} = f(x) = b_0 + b_1 * X_{\text{RM}} = 0 + 3.65279843 * X_{\text{RM}}$$

Sai số RMSE: 7.67452585343132

# Multiple Linear Regression.

CRIM	ZN	INDUS	CHAS	NOX	RM	AGE	DIS	RAD	TAX	PTRATIO	B	LSTAT	MEDV
0.00632	18.0	2.31	0.0	0.538	6.575	65.2	4.0900	1.0	296.0	15.3	396.90	4.98	24.0
0.02731	0.0	7.07	0.0	0.469	6.421	78.9	4.9671	2.0	242.0	17.8	396.90	9.14	21.6
0.02729	0.0	7.07	0.0	0.469	7.185	61.1	4.9671	2.0	242.0	17.8	392.83	4.03	34.7
0.03237	0.0	2.18	0.0	0.458	6.998	45.8	6.0622	3.0	222.0	18.7	394.63	2.94	33.4
0.06905	0.0	2.18	0.0	0.458	7.147	54.2	6.0622	3.0	222.0	18.7	396.90	5.33	36.2
0.02985	0.0	2.18	0.0	0.458	6.430	58.7	6.0622	3.0	222.0	18.7	394.12	5.21	28.7
0.08829	12.5	7.87	0.0	0.524	6.012	66.6	5.5605	5.0	311.0	15.2	395.60	12.43	22.9
0.14455	12.5	7.87	0.0	0.524	6.172	96.1	5.9505	5.0	311.0	15.2	396.90	19.15	27.1
0.21124	12.5	7.87	0.0	0.524	5.631	100.0	6.0821	5.0	311.0	15.2	386.63	29.93	16.5
0.17004	12.5	7.87	0.0	0.524	6.004	85.9	6.5921	5.0	311.0	15.2	386.71	17.10	

$$\hat{y}_{MEDV} = f(x) = b_0 + b_1 * X_{CRIM} + b_2 * X_{ZN} + b_3 * X_{INDUS} + b_4 * X_{CHAS} + b_5 * X_{NOX} + b_6 * X_{RM} + b_7 * X_{AGE} + b_8 * X_{DIS} + b_9 * X_{RAD} + b_{10} * X_{TAX} + b_{11} * X_{PTRATIO} + b_{12} * X_B + b_{13} * X_{LSTAT}$$



### 3. Đánh giá độ chính xác của mô hình hồi quy



# Đánh giá mô hình hồi quy



- Giả thiết: Có một mô hình học máy thực hiện việc dự đoán giá nhà tại một khu vực?
- Mô hình sau khi được huấn luyện với dữ liệu Training, thực hiện kiểm thử mô hình trên tập dữ liệu Test với số lượng 100 mẫu.

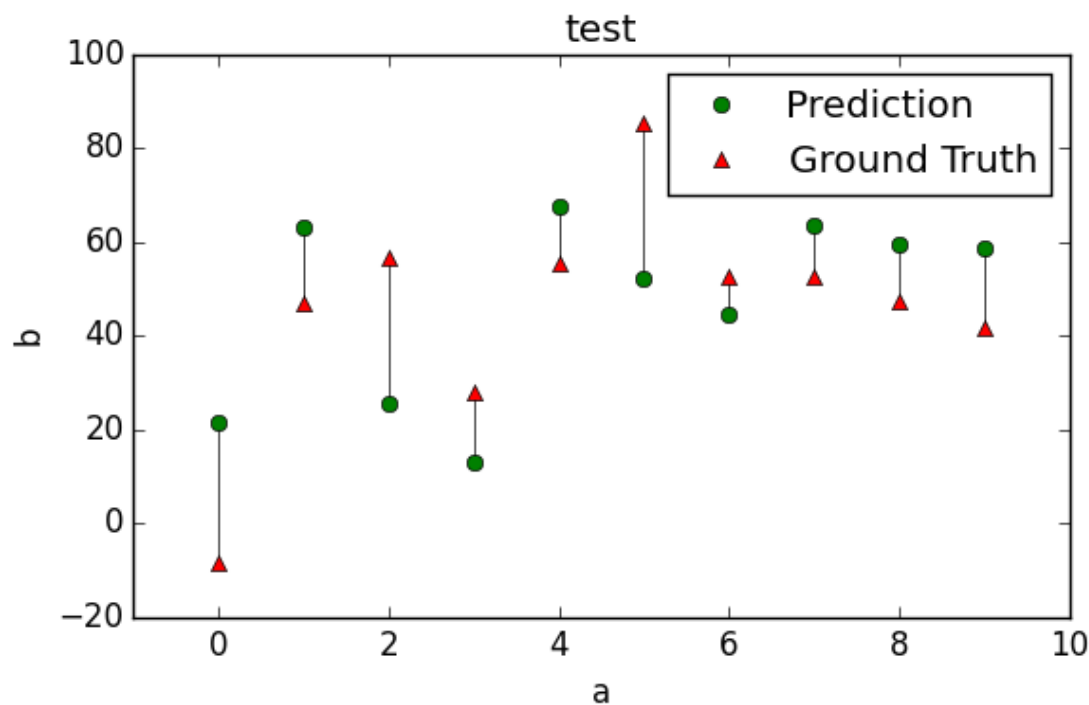
Evaluating Machine Learning Models

y_predict	y_groundtruth
22 890	23 432
19 120	18 850
9 590	10 500
20 231	22 567
7 498	5 235
13 675	11 563
22 453	25 005
24 645	19 214
30 654	27 087
5 643	8 675
14 087	13 675
8 000	7 465
25 986	29 875

# Đánh giá mô hình hồi quy



- Các chỉ số cơ bản để đánh giá độ chính xác của mô hình hồi quy:



## Regression

- $MAE$   
(mean abs. error)
- $MSE$   
(mean sq. error)
- $RMSE$   
(Root mean sq. error)
- $RMSLE$   
(Root mean sq. error  
log error)
- $R^2$  and Adjusted  
 $R^2$

# 1. Sai số MAE



- Sai số tuyệt đối trung bình (MAE – Mean Absolute Error) nằm trong khoảng  $(0, +\infty)$ . MAE biểu thị biên độ trung bình của sai số mô hình nhưng không nói lên xu hướng lệch của giá trị dự đoán (predicted) và giá trị thực (Actual). Khi  $MAE = 0$ , các giá trị dự đoán hoàn toàn trùng khớp với các giá trị thực, khi đó mô hình được xem là “lý tưởng”

$$MAE = \frac{1}{n} \sum \left| y - \hat{y} \right|$$

Diagram illustrating the MAE formula with annotations:

- Divide by the total number of data points**: Points to the  $\frac{1}{n}$  term.
- Sum of**: Points to the summation symbol  $\sum$ .
- Actual output value**: Points to the  $y$  term inside the absolute value.
- Predicted output value**: Points to the  $\hat{y}$  term inside the absolute value.
- The absolute value of the residual**: Points to the entire absolute value expression  $|y - \hat{y}|$ .

## 2. Sai số MSE



- **Sai số bình phương trung bình (MSE)** nằm trong khoảng  $(0, +\infty)$ , MSE phản ánh mức độ dao động giữa giá trị dự đoán với giá trị thực.

$$MSE = \frac{1}{n} \sum \underbrace{\left( y - \hat{y} \right)^2}_{\substack{\text{The square of the difference} \\ \text{between actual and} \\ \text{predicted}}}$$

### 3. Sai số RMSE



- **Sai số bình phương trung bình quân phương (RMSE)** là một trong những đại lượng cơ bản và thường được sử dụng phổ biến trong đánh giá độ tin cậy của mô hình hồi quy. Người ta thường hay sử dụng RMSE biểu thị độ lớn trung bình của sai số. Đặc biệt RMSE rất nhạy với những giá trị sai số lớn. Giống như MAE, RMSE không chỉ ra độ lệch giữa giá trị dự báo và giá trị thực. Giá trị của RMSE nằm trong khoảng  $(0, +\infty)$

$$\text{RMSE} = \sqrt{\frac{1}{n} \sum_{i=1}^n (y_i - \hat{y}_i)^2}$$

# 4. Hệ số $R^2$



**R<sup>2</sup>: Đánh giá tỷ lệ giải thích của mô hình ước lượng, hệ số này nằm giữa 0 và 1, càng gần 1 tỷ lệ giải thích được của mô hình càng tốt.**

- Giá trị R bình phương dao động từ 0 đến 1. R bình phương càng gần 1 thì mô hình đã xây dựng càng phù hợp với bộ dữ liệu dùng chạy hồi quy. R bình phương càng gần 0 thì mô hình đã xây dựng càng kém phù hợp với bộ dữ liệu dùng chạy hồi quy. Trường hợp đặc biệt, phương trình hồi quy đơn biến (chỉ có 1 biến độc lập) thì R<sup>2</sup> chính là bình phương của hệ số tương quan r giữa hai biến đó.

$$\text{Coefficient of Determination} \rightarrow R^2 = \frac{SSR}{SST} = 1 - \frac{SSE}{SST}$$

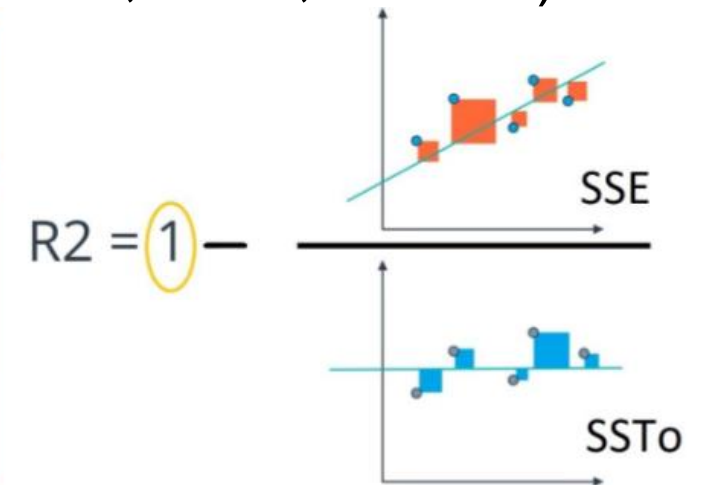
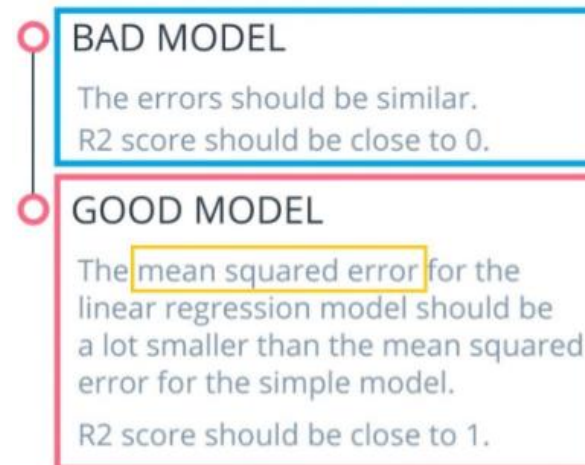
$$\text{Sum of Squares Total} \rightarrow SST = \sum (y - \bar{y})^2$$

$$\text{Sum of Squares Regression} \rightarrow SSR = \sum (y' - \bar{y}')^2$$

$$\text{Sum of Squares Error} \rightarrow SSE = \sum (y - y')^2$$

# 4. Hệ số $R^2$

**Ý nghĩa  $R$  bình phương:** Giả sử  $R$  bình phương là 0.60, thì mô hình hồi quy tuyến tính này phù hợp với tập dữ liệu ở mức 60%. Nói cách khác, 60% biến thiên của biến phụ thuộc được giải thích bởi các biến độc lập. (còn 40% còn lại ở đâu, dĩ nhiên là do sai số đo lường, do cách thu thập dữ liệu, do có thể có biến độc lập khác giải thích cho biến phụ thuộc mà chưa được đưa vào mô hình nghiên cứu...vv). Thông thường, ngưỡng của  $R^2$  phải trên 50%, vì như thế mô hình mới phù hợp. Tuy nhiên tùy vào dạng nghiên cứu, như các mô hình về tài chính, không phải tất cả các hệ số  $R^2$  đều bắt buộc phải thỏa mãn lớn hơn 50%. (do rất khó để dự đoán giá vàng, giá cổ phiếu mà chỉ đơn thuần dựa vào các biến độc lập ví dụ GDP, ROA, ROE....)



# THỰC HÀNH 9



- Sinh viên tìm hiểu về tập dữ liệu mẫu Diabetes Dataset của Sklearn (xác định các features đầu vào (input) và label đầu ra (target))

## Data Set Characteristics:

<b>Number of Instances:</b>	442
<b>Number of Attributes:</b>	First 10 columns are numeric predictive values
<b>Target:</b>	Column 11 is a quantitative measure of disease progression one year after baseline
<b>Attribute Information:</b>	<ul style="list-style-type: none"> <li>• age age in years</li> <li>• sex</li> <li>• bmi body mass index</li> <li>• bp average blood pressure</li> <li>• s1 tc, T-Cells (a type of white blood cells)</li> <li>• s2 ldl, low-density lipoproteins</li> <li>• s3 hdl, high-density lipoproteins</li> <li>• s4 tch, thyroid stimulating hormone</li> <li>• s5 ltg, lamotrigine</li> <li>• s6 glu, blood sugar level</li> </ul>

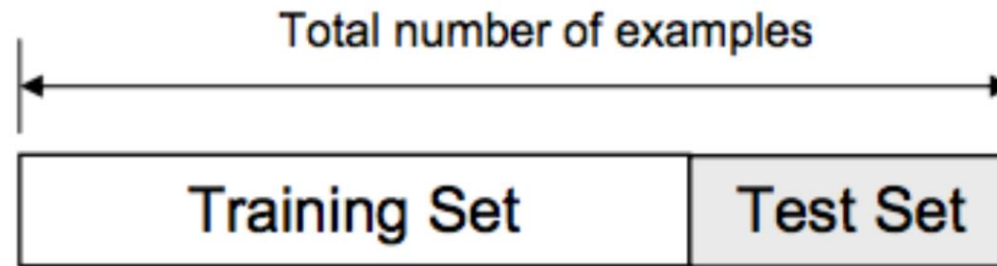
Samples total	442
Dimensionality	10
Features	real, $-0.2 < x < 0.2$
Targets	integer 25 - 346

Patient	AGE x1	SEX x2	BMI x3	BP x4	... x5	Serum Measurements x6	x7	x8	x9	x10	Response y
1	59	2	32.1	101	157	93.2	38	4	4.9	87	151
2	48	1	21.6	87	183	103.2	70	3	3.9	69	75
3	72	2	30.5	93	156	93.6	41	4	4.7	85	141
4	24	1	25.3	84	198	131.4	40	5	4.9	89	206
5	50	1	23.0	101	192	125.4	52	4	4.3	80	135
6	23	1	22.6	89	139	64.8	61	2	4.2	68	97
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
441	36	1	30.0	95	201	125.2	42	5	5.1	85	220
442	36	1	19.6	71	250	133.2	97	3	4.6	92	57

## Yêu cầu 9.2



- Trong tập dữ liệu Diabetes xác định thuộc tính có ảnh hưởng lớn nhất (hệ số tương quan cao nhất) tới thuộc tính target.
- Tách tập dữ liệu thành 2 phần Train – Test với tỷ lệ 75%-25%

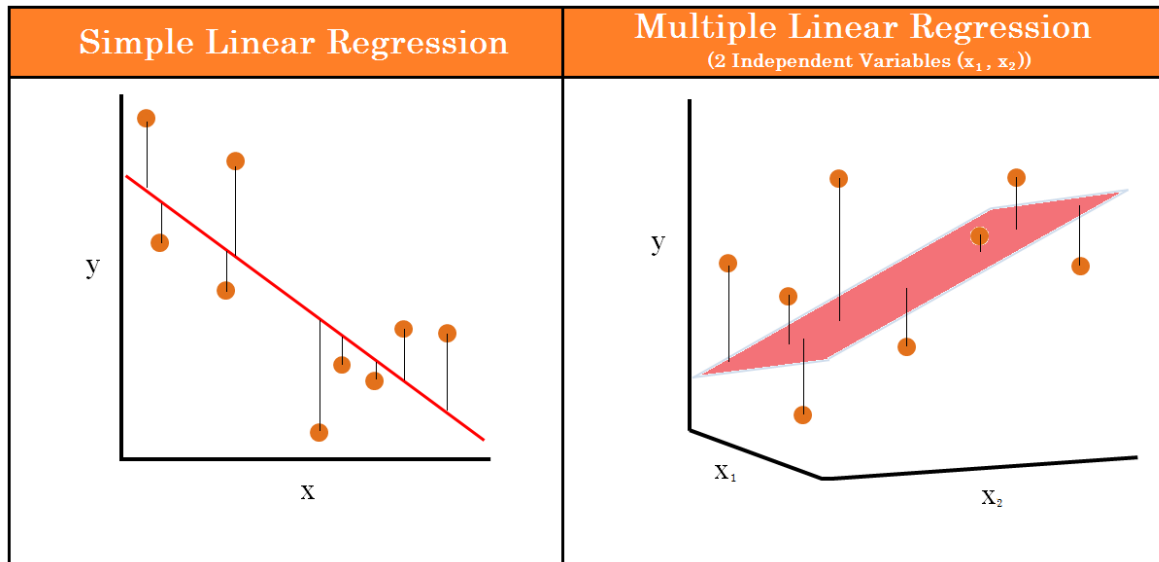


## Yêu cầu 9.3:

1. Xây dựng mô hình hồi quy tuyến tính đơn giản (Simple Linear Regression) với thuộc tính có ảnh hưởng cao nhất tới thuộc tính Target. Xác định sai số RMSE và  $R^2$  trên tập Train và Test.

2. Xây dựng mô hình hồi quy tuyến tính với tất các thuộc tính đầu vào (input). Xác định sai số RMSE và  $R^2$  trên tập Train và Test.

3. Xây dựng mô hình hồi quy tuyến tính với các thuộc tính đầu vào (input) có hệ số tương quan  $>|0.5|$ . Xác định sai số RMSE và  $R^2$  trên tập Train và Test.

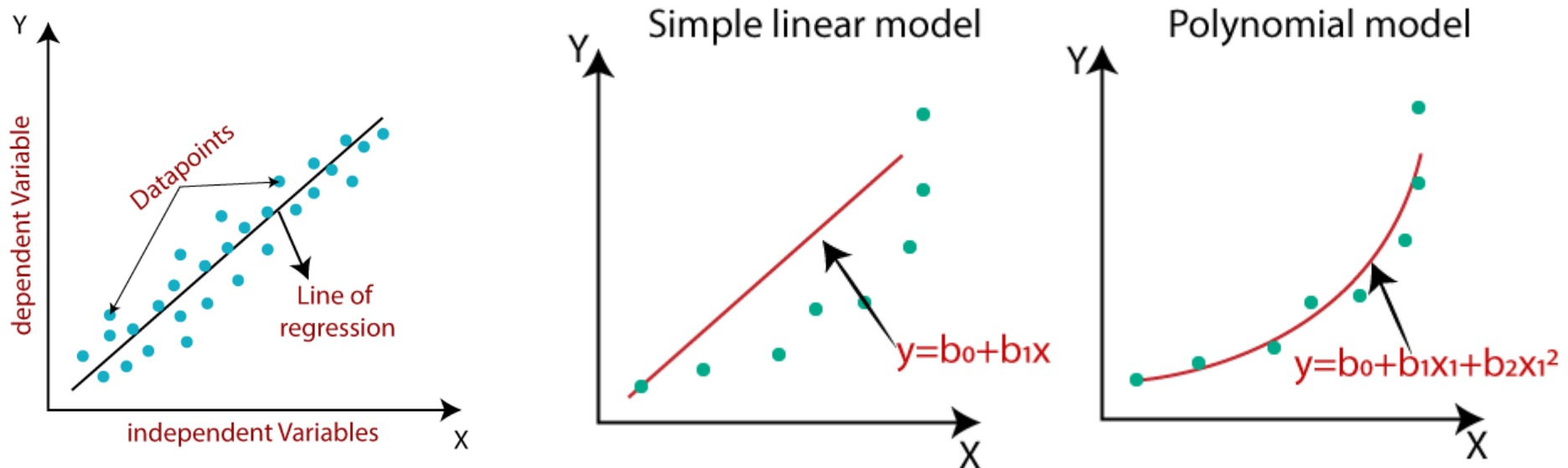


## 4. Một số mô hình hồi quy khác (Sinh viên tìm hiểu thêm)

## 4.1 Hồi quy đa thức (Polynomial Regression)

# Hồi quy đa thức

Trong trường hợp dữ liệu không tuyến tính việc áp dụng mô hình tuyến tính sẽ không hiệu quả tỷ lệ lỗi cao, độ chính xác giảm.



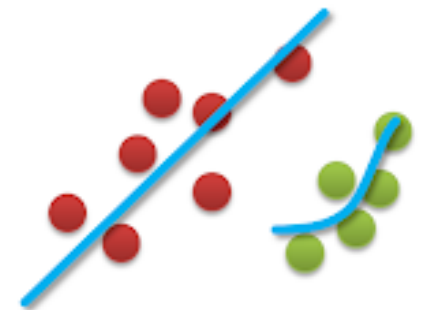
Hồi quy đa thức bậc n của biến độc lập  $x_1$

Polynomial  
Linear  
Regression

$$y = b_0 + b_1x_1 + b_2x_1^2 + \dots + b_nx_1^n$$

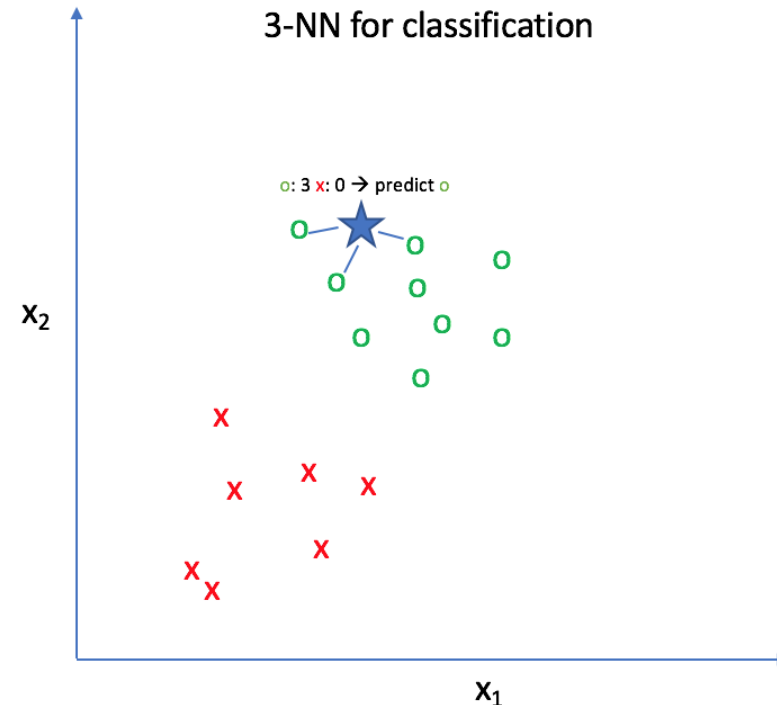
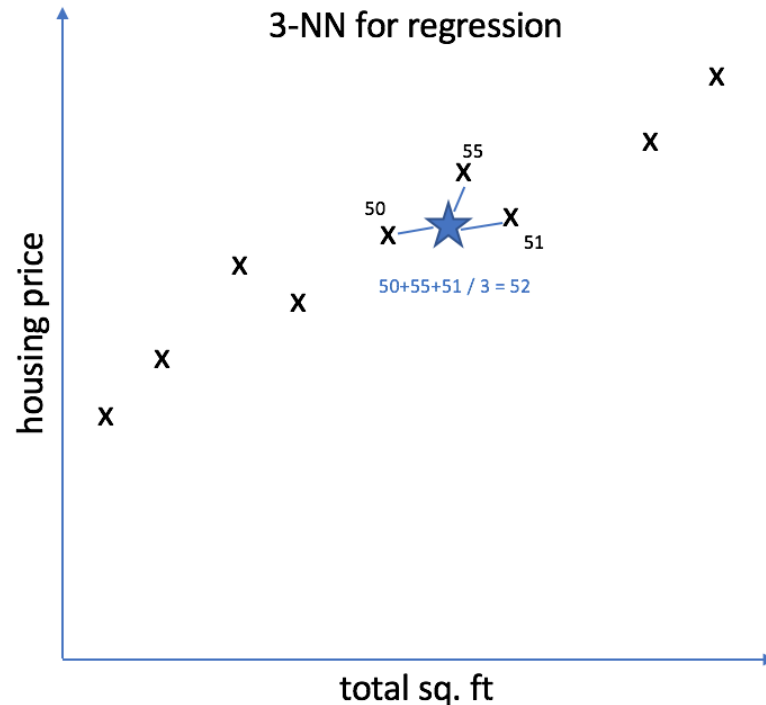
## 4.2 KNN cho bài toán Hồi quy

Regression



Tương tự như đối với bài toán phân lớp. Xác định những điểm dữ liệu gần nhất với điểm dữ liệu mới.

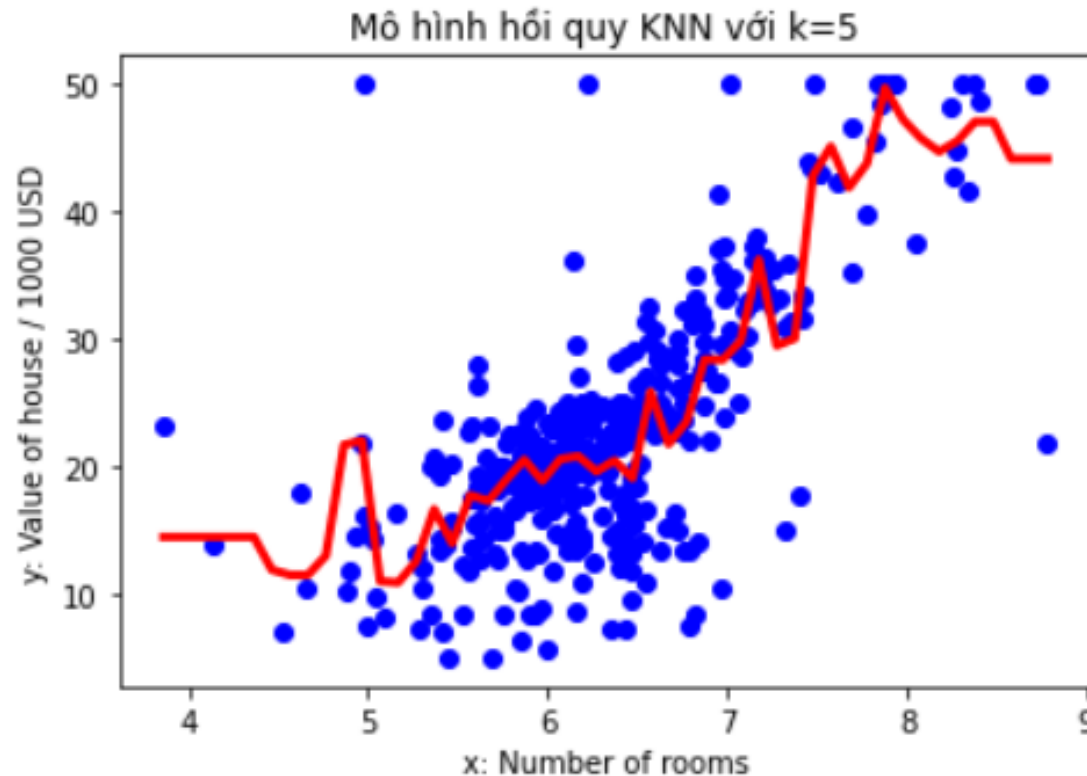
Nhãn của điểm dữ liệu mới được là nhãn của điểm dữ liệu đã biết gần nhất ( $K=1$ ) hoặc trung bình có trọng số của những điểm gần nhất.





## Dự đoán giá nhà với 1 biến độc lập – RM (số phòng trung bình của căn nhà)

RM	MEDV
6.575	24.0
6.421	21.6
7.185	34.7
6.998	33.4
7.147	36.2
6.430	28.7
6.012	22.9
6.172	27.1
5.631	16.5
6.004	18.9
6.377	15.0



MÔ HÌNH HỒI QUY KNN SỬ DỤNG 1 BIẾN ĐỘC LẬP-RM  
Độ chính xác của mô hình trên tập huấn luyện:

Sai số RMSE 5.166827358301712  
Sai số R2 0.6906745137827078

Độ chính xác của mô hình trên tập kiểm thử:

Sai số RMSE 6.498274280718765  
Sai số R2 0.4508457090315743



# Thank you!