Bài giảng môn học:

Kỹ nghệ tri thức và học máy (7080510)

CHƯƠNG 3:
HỌC CÓ GIÁM SÁT – PHẦN 1
(Supervised Learning)

Giảng viên: Đặng Văn Nam

Email: dangvannam@humg.edu.vn

Nội dung chương 3 – phần 1

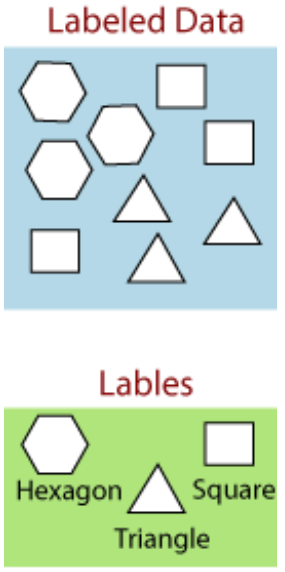
1. Giới thiệu học có giám sát (Supervised Learning)
2. Phân loại học có giám sát (Classification – Regression)
3. Một số Thuật toán phân phân lớp cơ bản
4. Đánh giá độ chính xác của các thuật toán phân lớp

1. Giới thiệu học có giám sát (supervised learning)

Học có giám sát là gì?

Một thuật toán học máy được gọi là học có giám sát (supervised learning) nếu việc xây dựng mô hình dự đoán **mối quan hệ giữa đầu vào và đầu ra** được thực hiện dựa trên các cặp (đầu vào - input, đầu ra – label) đã biết trong tập huấn luyện. Đây là nhóm thuật toán **phổ biến nhất trong các thuật toán machine learning**.

Tập dữ liệu học (Training data) bao gồm các quan sát (Examples, Observations), mà mỗi quan sát được gắn kèm với một giá trị đầu ra mong muốn (Label)



Học có giám sát là gì?

Dataset sẽ bao gồm:

- **Các thuộc tính đầu vào (Biến độc lập) – Features (input)**
- **Thuộc tính mục tiêu (Biến phụ thuộc) – Target (label)**

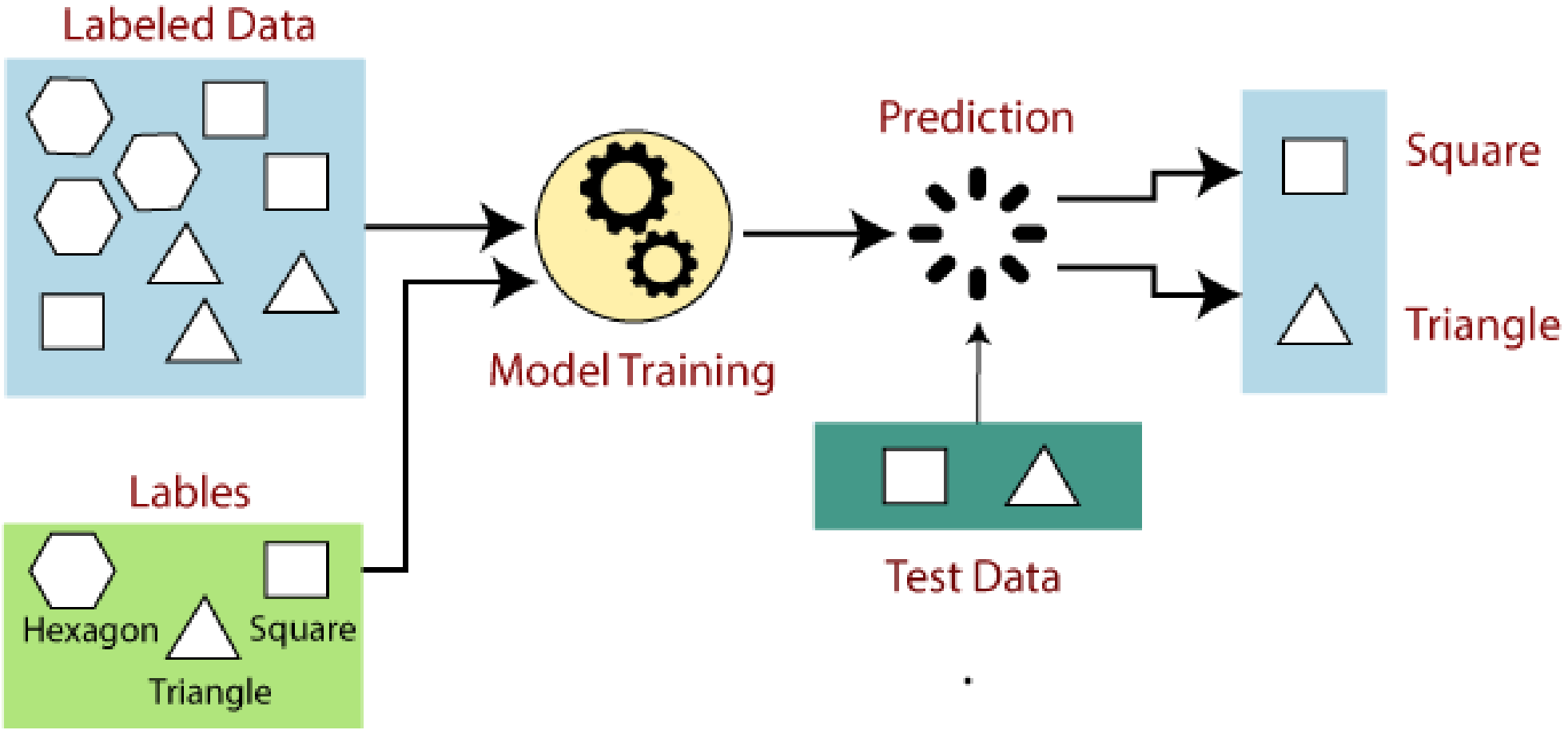
Biến độc lập (Features)

Biến phụ thuộc (Target) - Label

id	Age	Gender	Type	Blood_pressure	Cholesterol	Heartbeat	Thalassemia	Result
Patient_01	63	Male	Typical angina	145	233	150	6	0
Patient_02	67	Male	Asymptomatic	160	286	108	3	1
Patient_03	67	Male	Asymptomatic	120	229	129	7	1
Patient_04	37	Male	Non-anginal pain	130	250	187	3	0
Patient_05	41	Female	Atypical angina	130	204	172		0
Patient_16	56	Male	Atypical angina	120	236	178	3	0
Patient_07	62	Female	Asymptomatic	140	268	160	3	1
Patient_08	57	Female	Asymptomatic	120	354	163	3	0
Patient_19	63	Male	Asymptomatic	130	254	147	7	1
Patient_10	53	Male	Asymptomatic	140	203	155	7	1
Patient_11	57	Male	Asymptomatic	140	192	148	6	0
Patient_12	56	Female	Atypical angina	140	294	153	3	0
Patient_13	56	Male	Non-anginal pain	130	256	142	6	1
Patient_14	44	Male	Atypical angina	120	263	173	7	0
Patient_15	52	Male	Non-anginal pain	172	199	162	7	0
Patient_16	57	Male	Non-anginal pain	150	168	174	3	0
Patient_17	48	Male	Atypical angina	110	229	168	7	1
Patient_18	54	Male	Asymptomatic	140	239	160	3	0
Patient_19	48	Female	Non-anginal pain	130	275	139	3	0
Patient_20	49	Male	Atypical angina	130	266	171	3	0
Patient_21	64	Male	Typical angina	110	211	144		0

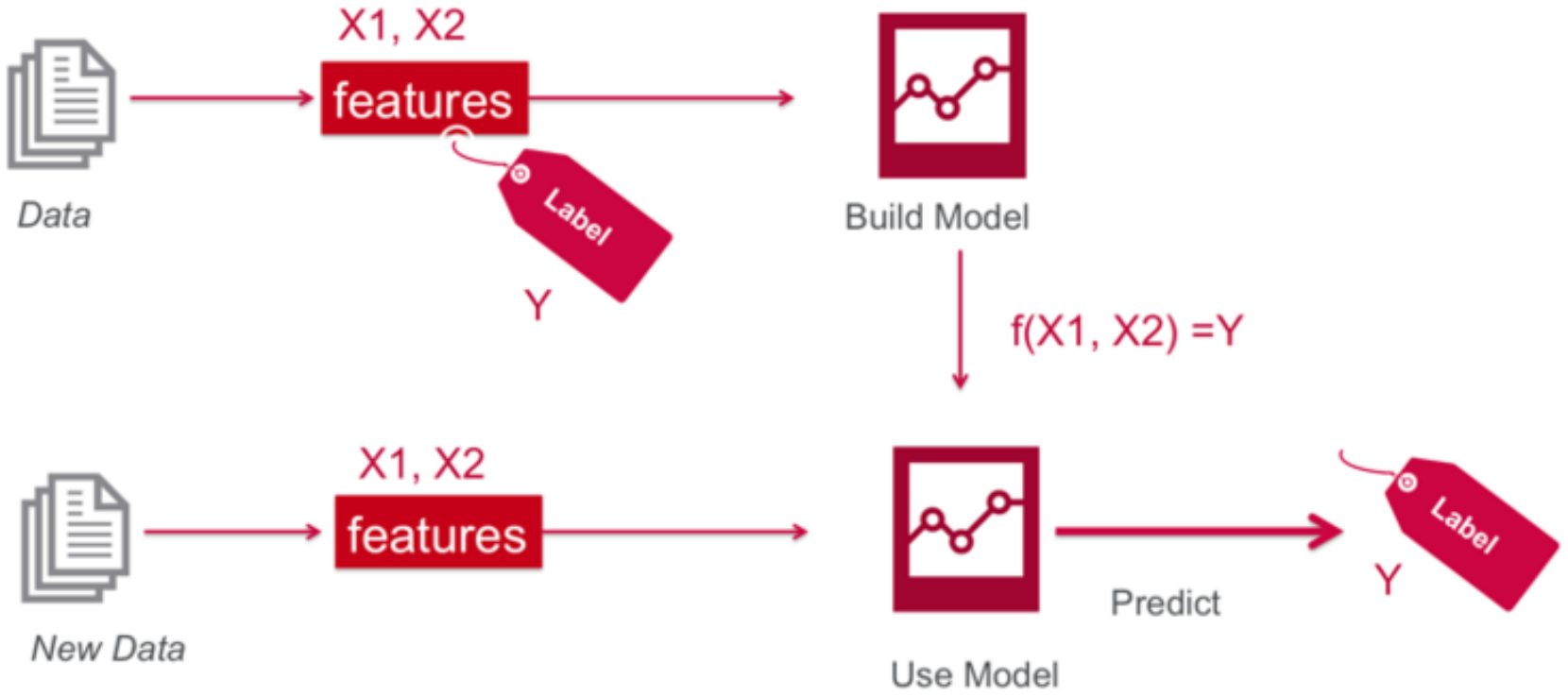
Học có giám sát là gì?

Quá trình huấn luyện và kiểm thử với Mô hình học máy có giám sát



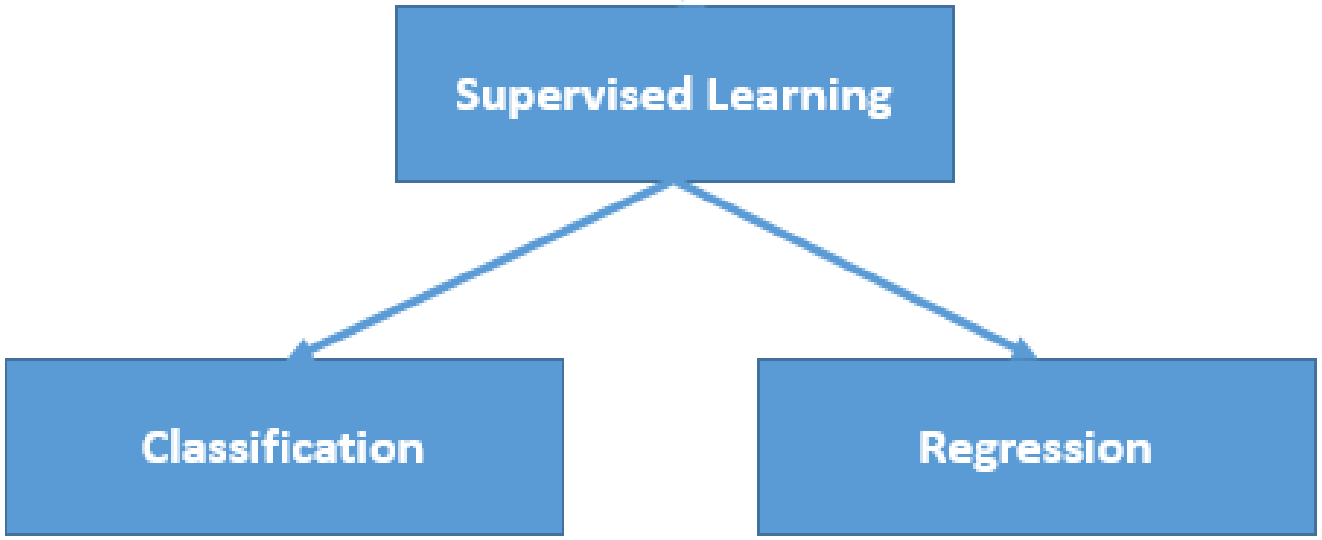
Học có giám sát là gì?

- Bản chất của Supervised learning là học một hàm f phù hợp với tập dữ liệu hiện có và có khả năng tổng quát hóa cao.
- Hàm học được sau đó sẽ dùng để dự đoán cho các quan sát mới.



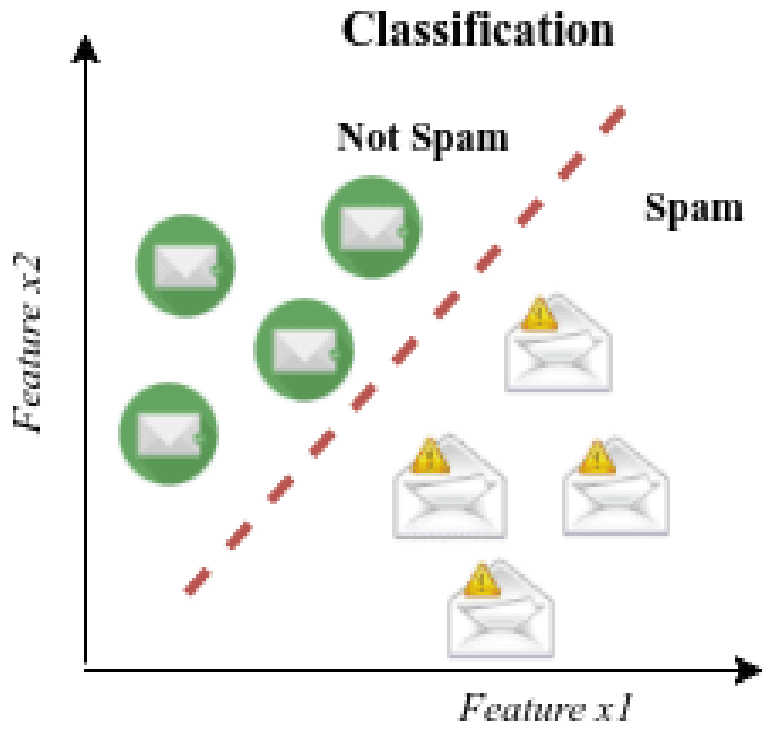
Các loại học có giám sát

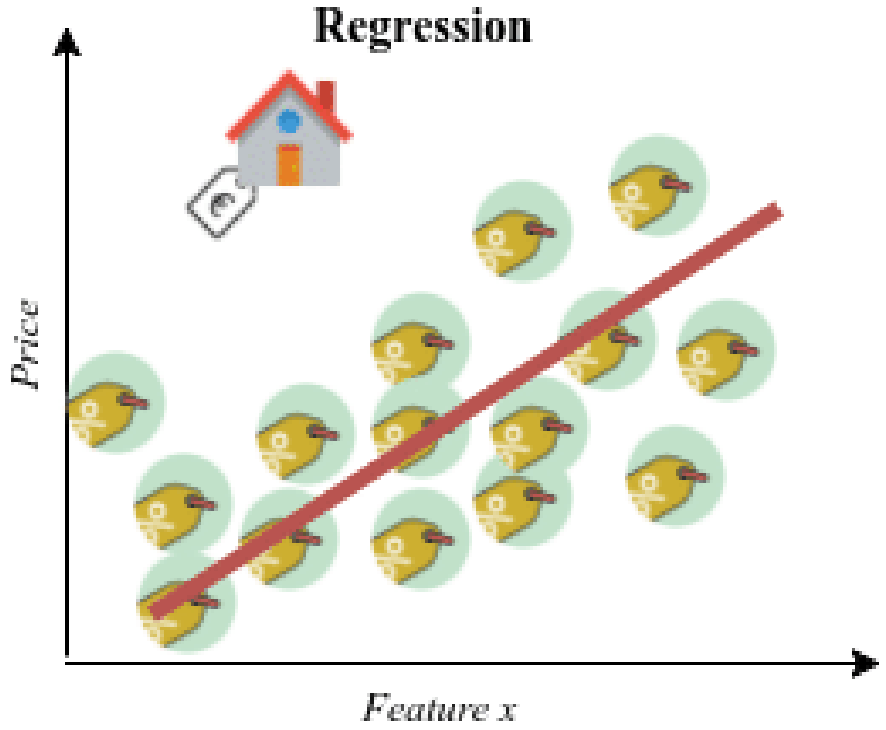
- Học có giám sát bao gồm 2 loại:
  - **Phân loại (Classification)**
  - **Hồi quy (Regression)**



Các loại học có giám sát

Phân loại (Classification) & Hồi quy (Regression)



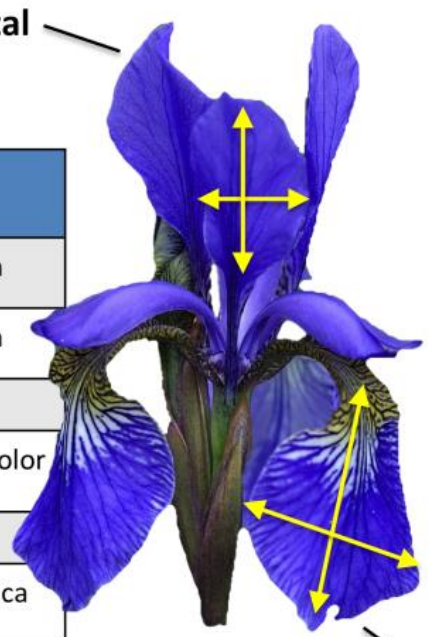


Phân loại (Classification)

Phân loại (Classification): Nếu nhãn (y – Target) thuộc tập rời rạc và hữu hạn

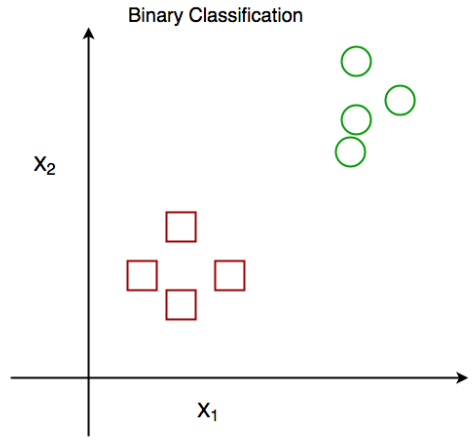
Samples
(instances, observations)

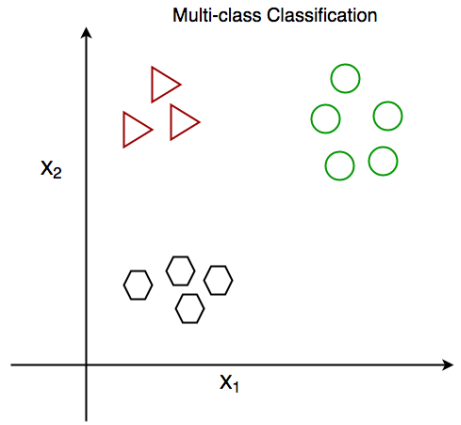
	Sepal length	Sepal width	Petal length	Petal width	Class label
1	5.1	3.5	1.4	0.2	Setosa
2	4.9	3.0	1.4	0.2	Setosa
...					
50	6.4	3.5	4.5	1.2	Versicolor
...					
150	5.9	3.0	5.0	1.8	Virginica



Features
(attributes, measurements, dimensions)

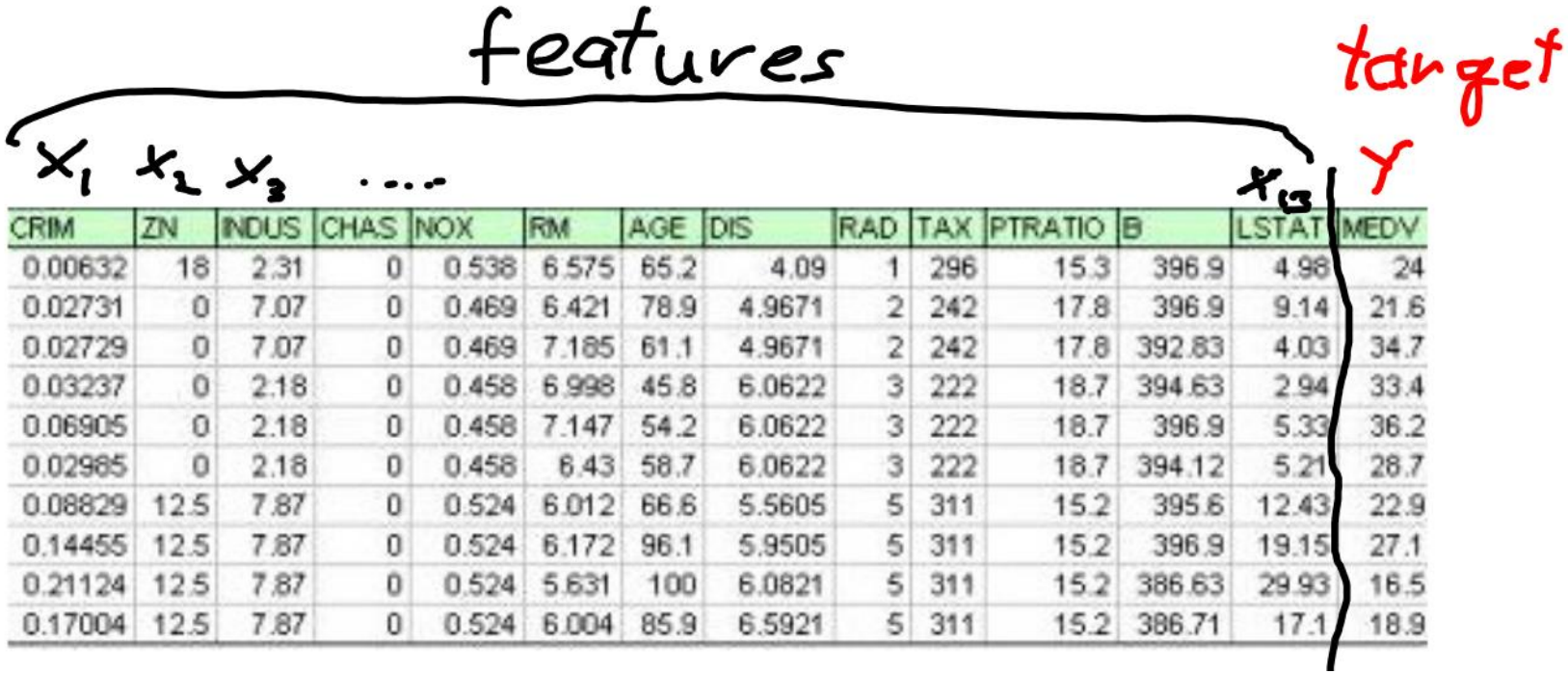
Class labels
(targets)





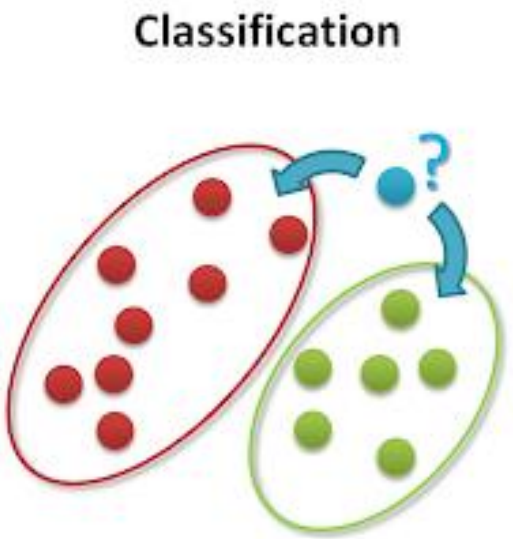
Hồi quy (Regression)

Hồi quy (Regression): Nếu nhãn (y – Target) là biến liên tục (các số thực) ví dụ như dự báo nhiệt độ, giá nhà, mức tiêu thụ điện năng…

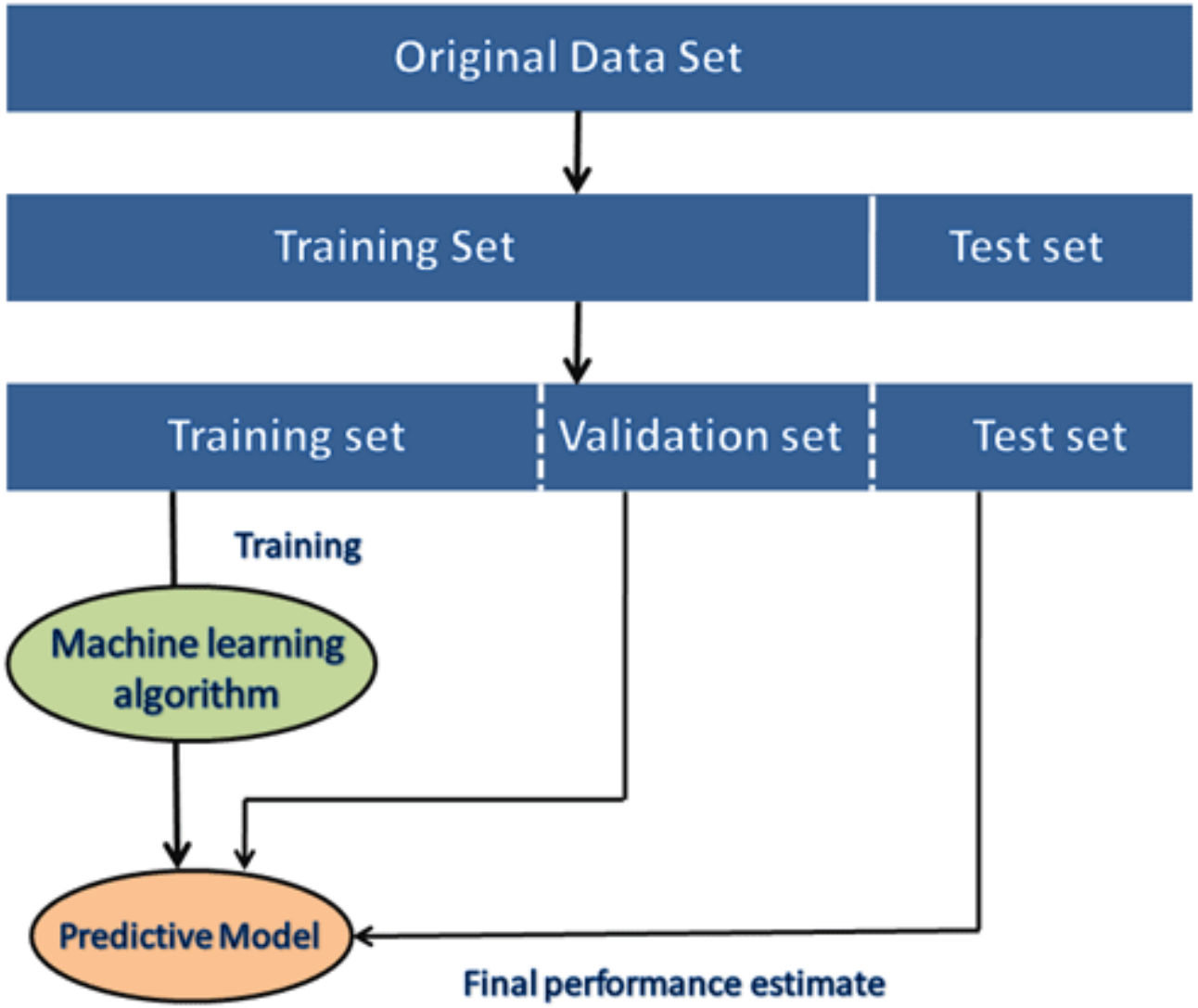


CRIM	ZN	INDUS	CHAS	NOX	RM	AGE	DIS	RAD	TAX	PTRATIO	B	LSTAT	MEDV
0.00632	18	2.31	0	0.538	6.575	65.2	4.09	1	296	15.3	396.9	4.98	24
0.02731	0	7.07	0	0.469	6.421	78.9	4.9671	2	242	17.8	396.9	9.14	21.6
0.02729	0	7.07	0	0.469	7.185	61.1	4.9671	2	242	17.8	392.83	4.03	34.7
0.03237	0	2.18	0	0.458	6.998	45.8	6.0622	3	222	18.7	394.63	2.94	33.4
0.06905	0	2.18	0	0.458	7.147	54.2	6.0622	3	222	18.7	396.9	5.33	36.2
0.02985	0	2.18	0	0.458	6.43	58.7	6.0622	3	222	18.7	394.12	5.21	28.7
0.08829	12.5	7.87	0	0.524	6.012	66.6	5.5605	5	311	15.2	395.6	12.43	22.9
0.14455	12.5	7.87	0	0.524	6.172	96.1	5.9505	5	311	15.2	396.9	19.15	27.1
0.21124	12.5	7.87	0	0.524	5.631	100	6.0821	5	311	15.2	386.63	29.93	16.5
0.17004	12.5	7.87	0	0.524	6.004	85.9	6.5921	5	311	15.2	386.71	17.1	18.9

2. Train, Test, Validation set



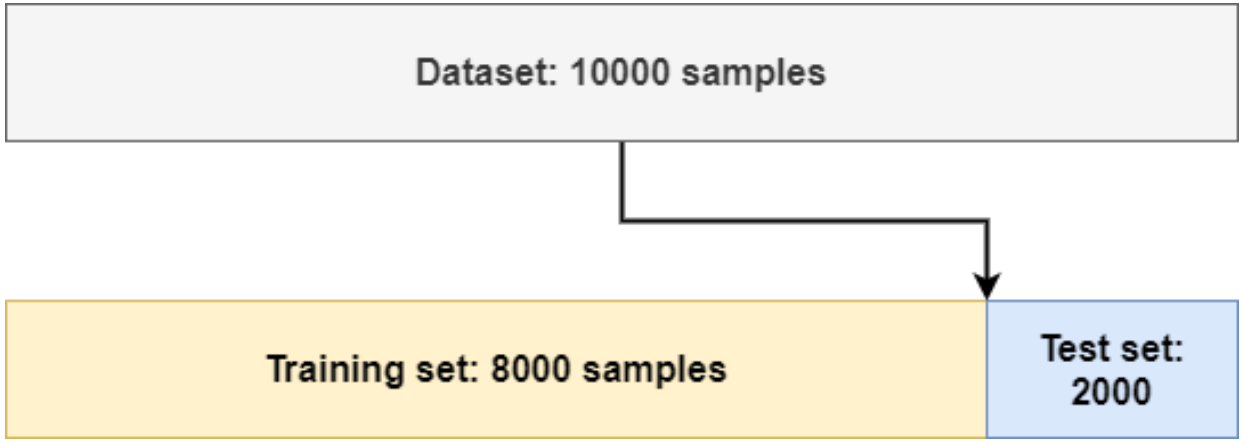
Train, Test, Validation set



1. Tập huấn luyện (Training Set)
2. Tập kiểm tra (Test Set)
3. Tập kiểm chéo (Validation Set)

Train, Test, Validation set

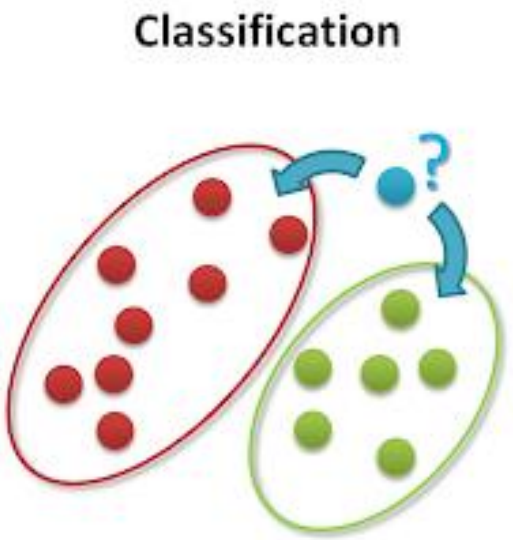
- **Tập huấn luyện (Training Set)** bao gồm các điểm dữ liệu sử dụng trực tiếp trong việc xây dựng mô hình.
- **Tập kiểm tra (Test set)** gồm các dữ liệu được dùng để đánh giá hiệu quả của mô hình. Tập kiểm tra đại diện cho dữ liệu mà mô hình chưa từng thấy, có thể xuất hiện trong quá trình vận hành mô hình trên thực tế.



➢ Để đảm bảo tính phổ quát, dữ liệu kiểm tra không được sử dụng trong quá trình xây dựng mô hình.

➢ Điều kiện cần để một mô hình hiệu quả: Kết quả đánh giá trên tập huấn luyện và tập kiểm tra đều cao.

3. Một số thuật toán phân lớp (Classification)



Top 10 Algorithms every Machine Learning Engineer should know

1. Naïve Bayes Classifier Algorithm

2. K Means Clustering Algorithm

3. Support Vector Machine Algorithm

4. Apriori Algorithm

5. Linear Regression Algorithm

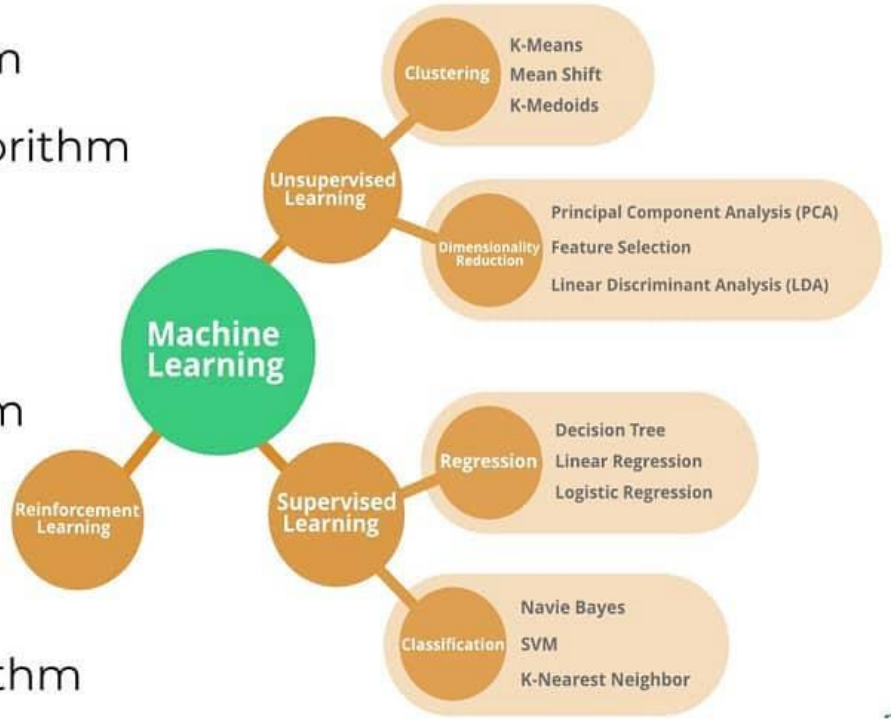
6. Logistic Regression Algorithm

7. Decision Trees Algorithm

8. Random Forests Algorithm

9. K Nearest Neighbours Algorithm

10. Artificial Neural Networks Algorithm



3.1 KNN
(K – Nearest Neighbors)

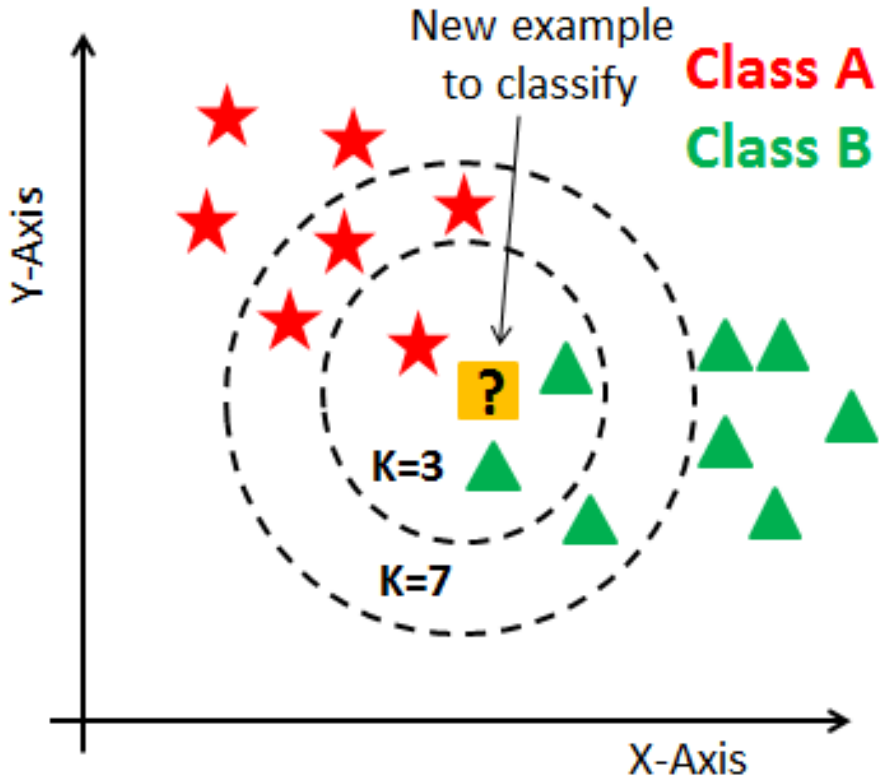
KNN

K-Nearest neighbors(k-NN) là một trong những thuật toán đơn giản nhất và phổ biến trong học máy. Một số tên gọi khác:

- **Instance-based learning**
- **Lazy learning**
- **Memory-based learning**

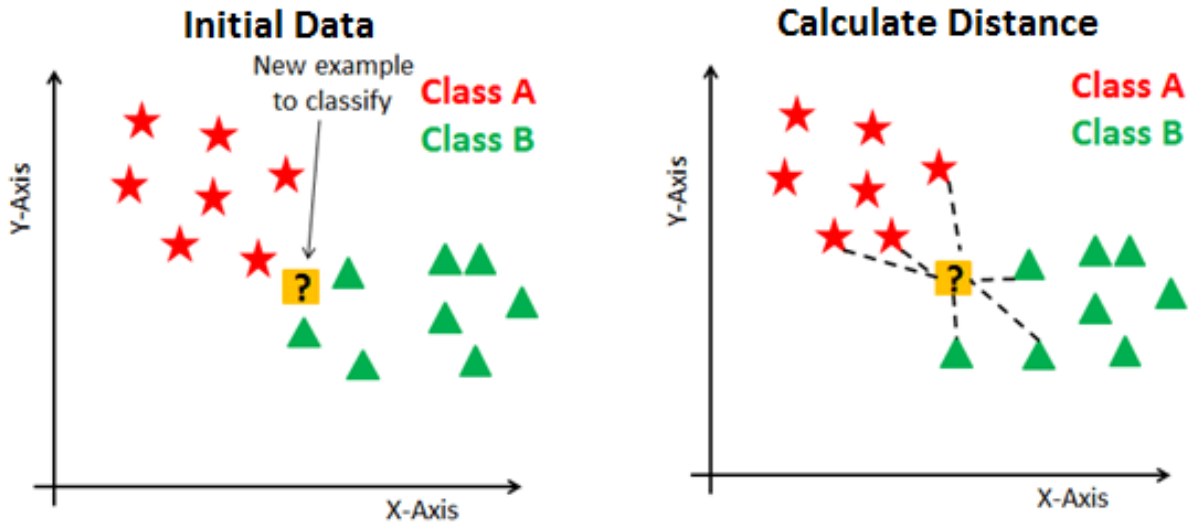
KNN

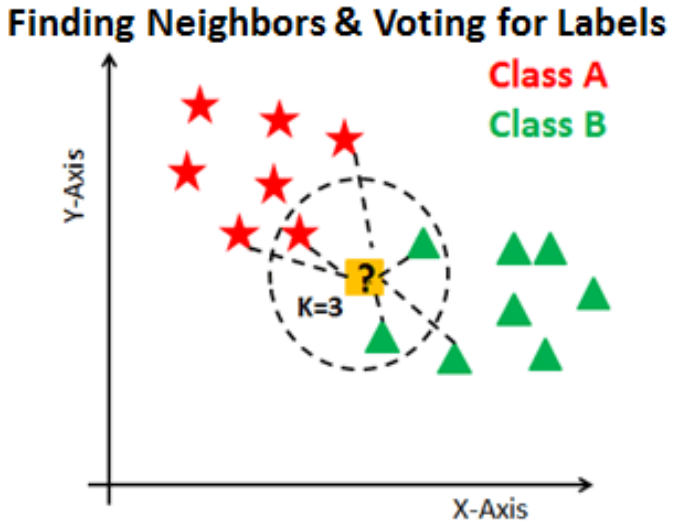
Bản chất, KNN là thuật toán đi tìm đầu ra của một điểm dữ liệu mới bằng cách chỉ dựa trên thông tin của K điểm dữ liệu trong tập huấn luyện gần nó nhất (K - lân cận)



KNN được sử dụng cho cả bài toán phân loại và hồi quy

KNN

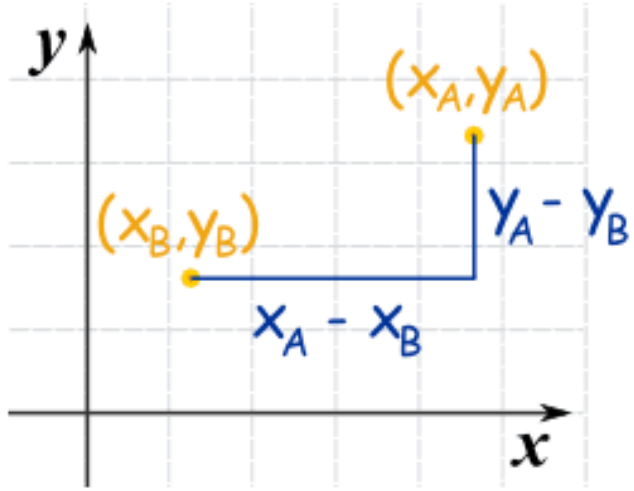




Những hàng xóm nào sẽ được sử dụng cho việc dự đoán?

KNN

Tính khoảng cách giữa hai điểm A - B



Now label the coordinates of points A and B.

x_A means the x-coordinate of point A

y_A means the y-coordinate of point A

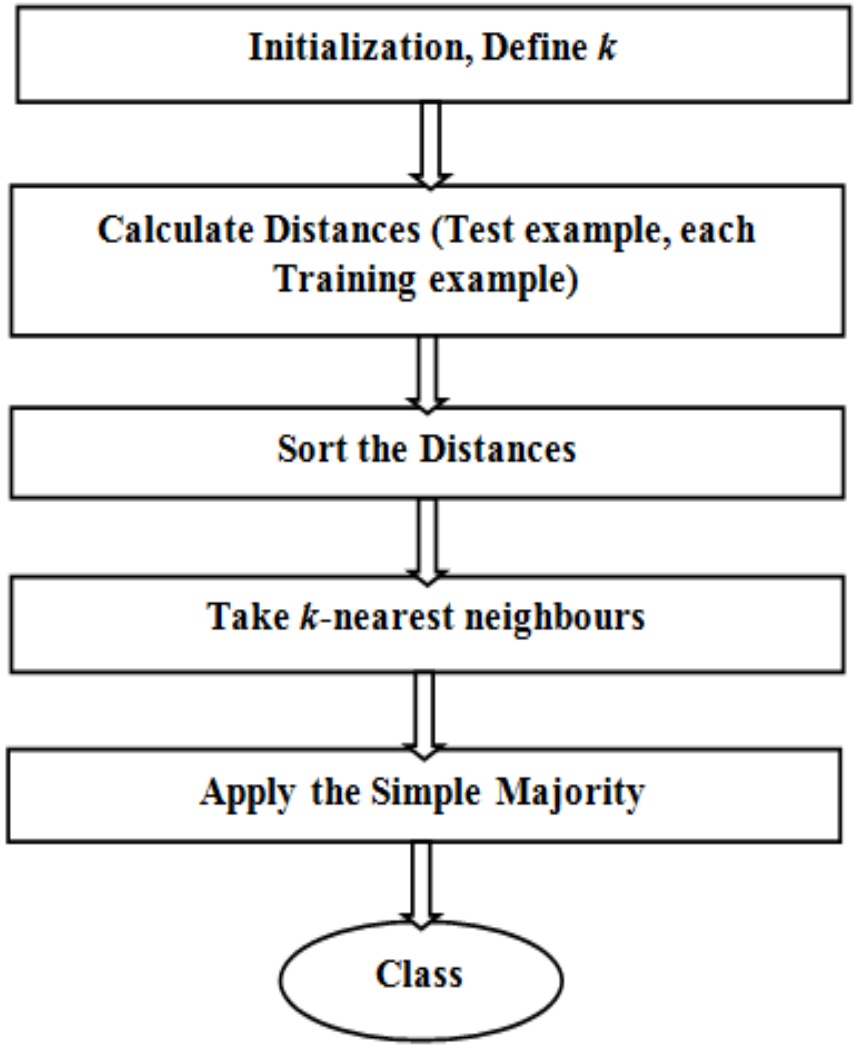
The horizontal distance **a** is $(x_A - x_B)$

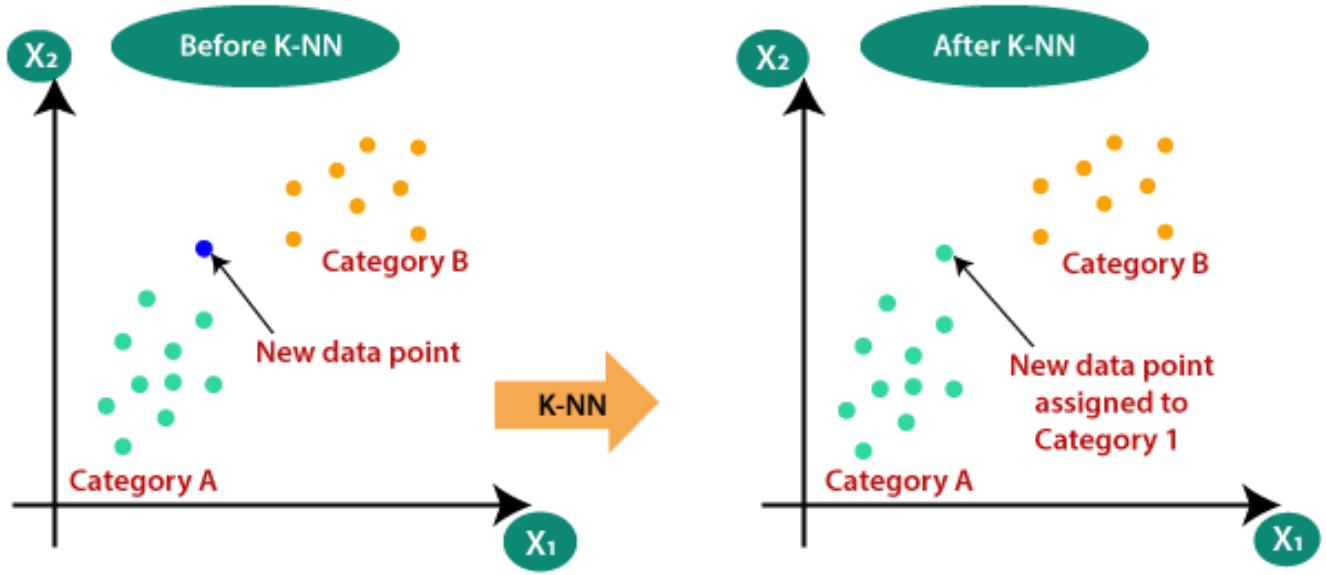
The vertical distance **b** is $(y_A - y_B)$

Euclidean Distance = $\sqrt{(x_A - x_B) + (y_A - y_B)}$

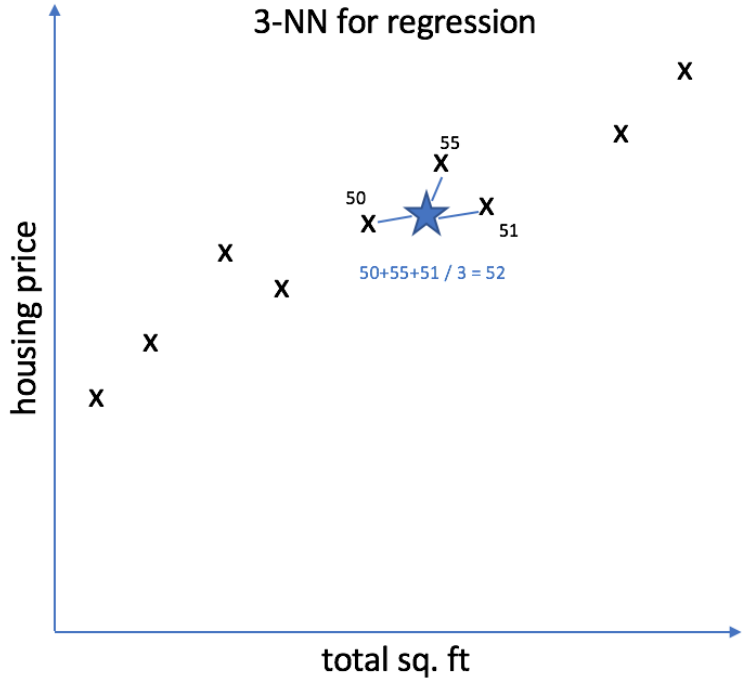
Manhattan Distance = $|(x_A - x_B)| + |(y_A - y_B)|$

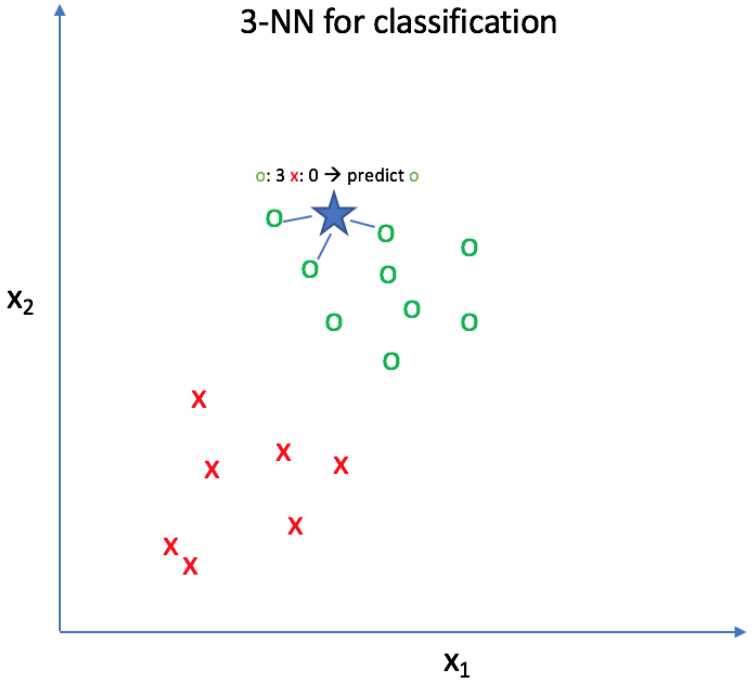
Các bước thực hiện thuật toán KNN





KNN cho bài toán phân lớp và hồi quy





Hồi quy (regression): 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.

Phân loại (classification): nhãn của điểm dữ liệu mới được suy ra trực tiếp từ K điểm dữ liệu gần nhất.

Ưu nhược điểm của KNN

Ưu điểm:

- Độ phức tạp tính toán trong quá trình huấn luyện bằng 0
- Việc dự đoán kết quả của dữ liệu mới rất đơn giản
- Không cần giả sử gì về phân phối của các class

Nhược điểm:

- KNN rất nhạy với nhiễu khi K nhỏ.
- Tính toán khoảng cách tới từng điểm dữ liệu trong trong tập huấn luyện tốn rất nhiều thời gian, đặc biệt với các CSDL có số chiều lớn và có nhiều điểm dữ liệu. K càng lớn thì độ phức tạp càng tăng.
- Lưu toàn bộ dữ liệu trong bộ nhớ ảnh hưởng tới hiệu năng của KNN

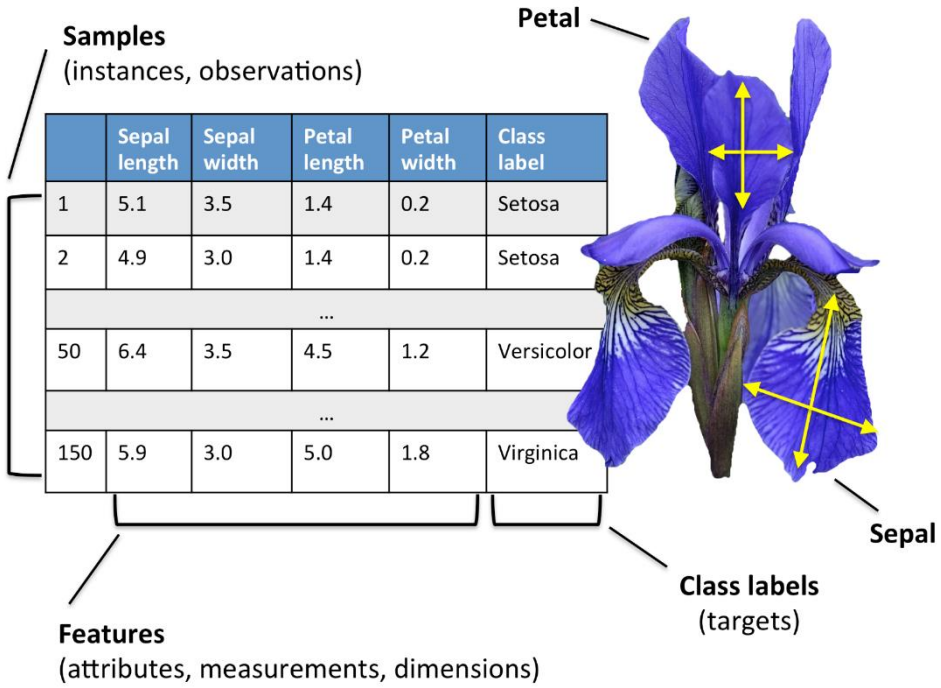
Ví dụ 1:
Phân lớp hoa lan với KNN

Ví dụ: Phân lớp hoa lan với KNN

- Tập dữ liệu bao gồm 150 mẫu về thông số chiều rộng, chiều dài của lá hóa và cánh hoa của 3 loại hoa Lan

IRIS DATASET

Classes	3
Samples per class	50
Samples total	150
Dimensionality	4
Features	real, positive



	Sepal length	Sepal width	Petal length	Petal width	Class label
1	5.1	3.5	1.4	0.2	Setosa
2	4.9	3.0	1.4	0.2	Setosa
...					
50	6.4	3.5	4.5	1.2	Versicolor
...					
150	5.9	3.0	5.0	1.8	Virginica

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

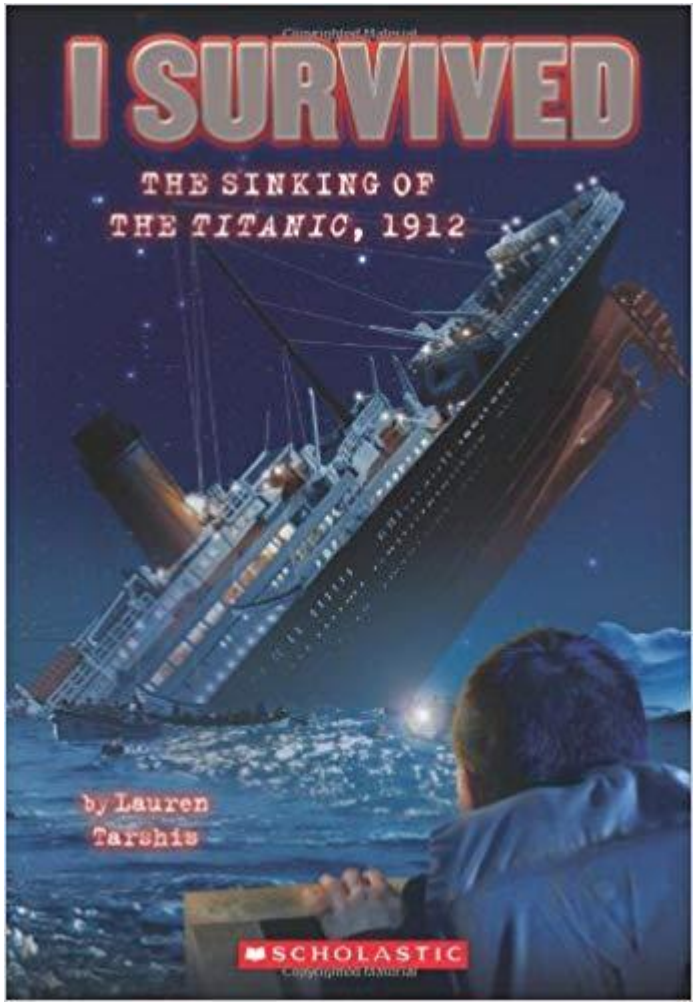
Ví dụ 2:
Bài toán Titanic

Ví dụ: Bài toán Titanic

- Xây dựng model học máy sử dụng KNN dự đoán khả năng được cứu (1), Không được cứu (0) của hành khách trên tập dữ liệu đã được chuẩn bị ở chương 2:

	A	B	C	D	E	F	G
1	Survived	Pclass	Sex	Age	SibSp	Parch	Embarked
2	0	3	0	1	1	0	0
3	1	1	1	2	1	0	1
4	1	3	1	1	0	0	0
5	1	1	1	2	1	0	0
6	0	3	0	2	0	0	0
7	0	3	0	1	0	0	2
8	0	1	0	3	0	0	0
9	0	3	0	0	3	1	0
10	1	3	1	1	0	2	0
11	1	2	1	0	1	0	1
12	1	3	1	0	1	1	0
13	1	1	1	3	0	0	0
14	0	3	0	1	0	0	0
15	0	3	0	2	1	5	0

Data_Titanic_ok



- **Sinh viên làm trên Jupyter Notebook**

THỰC HÀNH 6

Yêu cầu 1:

- Sinh viên tìm hiểu về tập dữ liệu mẫu wine trong Dataset của Sklearn (xác định các features và label) – **Đã tìm hiểu trong chương 2**

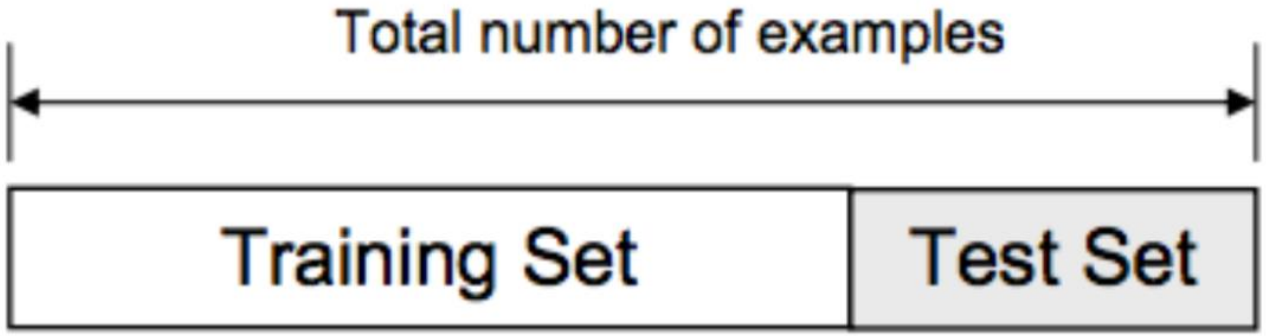
Number of Instances:	178 (50 in each of three classes)
Number of Attributes:	13 numeric, predictive attributes and the class
Attribute Information:	• Alcohol • Malic acid • Ash • Alcalinity of ash • Magnesium • Total phenols • Flavanoids • Nonflavanoid phenols • Proanthocyanins • Color intensity • Hue • OD280/OD315 of diluted wines • Proline

- **class:**
  - class_0
  - class_1
  - class_2

Classes	3
Samples per class	[59,71,48]
Samples total	178
Dimensionality	13
Features	real, positive

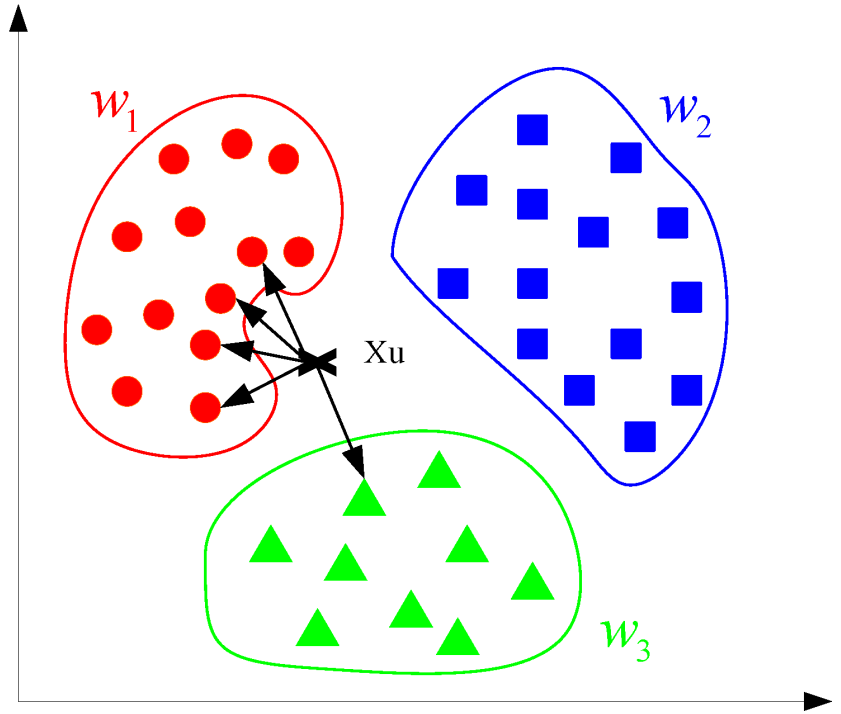
Yêu cầu 2:

- Tách tập dữ liệu data_wine thành 2 phần train – test theo tỷ lệ 75% - 25%



Yêu cầu 3:

- Sử dụng thuật toán KNN với các trường hợp: K= 5, 7, 11, 13 cho biết độ chính xác ứng với từng trường hợp của K trên tập Test.
- Áp dụng thuật toán KNN với K=9 và có đánh trọng số các điểm lân cận. cho biết độ chính xác của thuật toán trên tập Test và ma trận Confusion tương ứng




HANOI UNIVERSITY OF MINING AND GEOLOGY
KHOA
CÔNG NGHỆ THÔNG TIN
FACULTY OF INFORMATION TECHNOLOGY

Thank you!