Supervised Learning là gì ?

- **Học có giám sát (Supervised Learning)** là một phương pháp trong Machine Learning, trong đó mô hình được huấn luyện bằng cách sử dụng các tập dữ liệu đã được gán nhãn.
- Thuật toán sẽ học cách nhận diện các mẫu và mối quan hệ giữa dữ liệu đầu vào và đầu ra, từ đó có thể dự đoán chính xác kết quả khi gặp các dữ liệu mới trong thực tế.

Quy trình của Supervised Learning:

1. Chuẩn bị dữ liệu huấn luyện có gán nhãn: Tạo ra một tập dữ liệu mẫu, trong đó mỗi mục đều được dán nhãn một cách rõ ràng, mục tiêu là để mô hình học cách nhận biết đặc điểm riêng của từng nhãn.
2. Tiền xử lý (Data preprocessing): Đây là bước quan trọng trong việc quyết định chất lượng của mô hình, vì trước khi tính toán, dữ liệu cần phải được xử lý để loại bỏ các dữ liệu không cần thiết hoặc sai lệch.
3. Chia dữ liệu thành các tập khác nhau (Train/Validation/Test): Mỗi tập dữ liệu sẽ có nhiệm vụ khác nhau, tập Train dùng để huấn luyện, tập Validation để tinh chỉnh các tham số và chọn lựa mô hình tốt nhất, còn tập Test là để đánh giá hiệu năng cuối cùng của mô hình.
4. Huấn luyện mô hình để tìm ra mối quan hệ: Mô hình sẽ xử lý khối lượng dữ liệu lớn và tìm ra quy tắc chung để phân biệt các loại dữ liệu khác nhau. Mục tiêu là giúp máy tính hiểu được các mối liên hệ giữa dữ liệu input và output.
5. Đánh giá mô hình với dữ liệu kiểm tra: Sau khi đã huấn luyện, mô hình sẽ được đánh giá bằng một tập dữ liệu chưa từng thấy. Việc này giúp đánh giá mô hình có chính xác hay không. Một kỹ thuật phổ biến để kiểm tra là xác thực chéo (cross-validation), tức là chia dữ liệu thành nhiều phần để

kiểm tra và đảm bảo mô hình không chỉ giỏi trên tệp dữ liệu cũ mà còn làm tốt trên các dữ liệu mới.

6. Tối ưu hoá mô hình để giảm sai số: Trong suốt quá trình, mô hình sẽ được tối ưu hoá để việc dự đoán càng ngày càng chính xác hơn.
7. Triển khai và giám sát: Triển khai là việc mô hình bắt đầu được sử dụng để trả về kết quả dự đoán cho người dùng.

Phân biệt giữa Classification và Regression

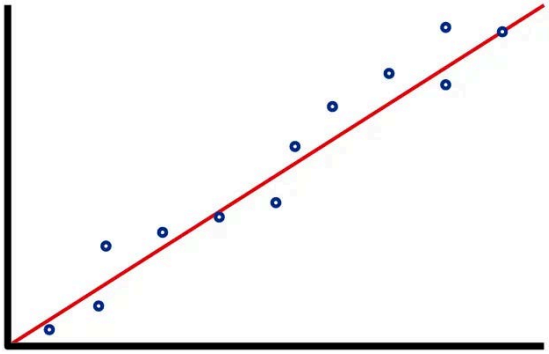
Supervised Learning thường được phân thành 2 loại khác nhau là: Phân loại (Classification) và Hồi quy (Regression). Mỗi loại có những ứng dụng và thuật toán riêng biệt, giúp mô hình có thể đưa ra các dự đoán chính xác.

- **Phân loại (Classification)**: Phân loại là phương pháp mà mô hình học máy dự đoán một nhãn hoặc danh mục cho dữ liệu đầu vào. Đây là cách giúp hệ thống sắp xếp dữ liệu vào các nhóm có sẵn. Các thuật toán phổ biến được sử dụng trong phân loại bao gồm cây quyết định, hồi quy logistic, rừng ngẫu nhiên, máy vector hỗ trợ (SVM) và Naive Bayes.
- **Hồi quy (Regression):** Hồi quy khác với phân loại ở việc nó không nhóm các dữ liệu vào các danh mục, mà dự đoán một giá trị thực liên tục dựa trên các dữ liệu đầu vào. Một số thuật toán phổ biến trong hồi quy bao gồm hồi quy tuyến tính, hồi quy phi tuyến, cây hồi quy và hồi quy đa thức.

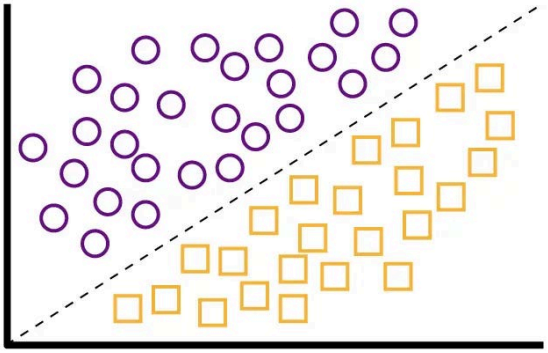
-> Với **Classification**, giá trị cốt lõi của nó là việc dự đoán và chia tách các dữ liệu thành các nhóm riêng biệt khác nhau, ví dụ cho việc này là việc phân loại Email này là spam hay không, hoặc “bức ảnh trên là chó hay mèo”. Còn với **Regression**, bài toán ấy sẽ dự đoán giá trị tốt nhất dựa theo các dữ liệu đầu vào, như việc tìm hiểu giá nhà dựa theo địa hình, kinh nghiệm,...

Bảng so sánh giữa Classification với Regression:

Đặc điểm	Classification (Phân loại)	Regression (Hồi quy)
Câu hỏi cốt lõi	"Cái này thuộc nhóm nào?" (Which one?)	"Giá trị là bao nhiêu?" (How much?)
Dạng đầu ra (Output)	Rời rạc (Discrete). Là các nhãn (labels) hoặc danh mục (categories).	Liên tục (Continuous). Là các con số thực (numbers).
Mục tiêu hình học	Tìm một đường ranh giới (Decision Boundary) để chia tách các điểm dữ liệu thành các nhóm riêng biệt.	Tìm một đường xu hướng (Best Fit Line) đi qua gần các điểm dữ liệu nhất có thể.
Ví dụ điển hình	Email là Spam hay Không Spam? Ảnh này là Chó hay Mèo?	Giá nhà là bao nhiêu? Nhiệt độ ngày mai là bao nhiêu độ?



Regression



Classification

- Theo mô hình trên, **Classification** sẽ chia tách các dữ liệu thành các nhóm riêng biệt qua một **đường ranh giới. Regression** sẽ tìm câu trả lời bằng những gì bám sát dữ liệu nhất có thể, có thể thấy điều ấy qua **đường xu hướng.**

Unsupervised Learning là gì?

- **Học không giám sát (Unsupervised Learning)** là phương pháp dùng thuật toán Machine Learning để phân tích và phân cụm dữ liệu chưa gán nhãn, phát hiện mẫu ẩn hoặc nhóm dữ liệu mà không cần con người can thiệp.
- **Unsupervised Learning** tự suy luận và sắp xếp các dữ liệu theo quy luật. Ví dụ, nếu cung cấp một tập dữ liệu lớn về thời tiết mà không có thông tin cụ thể, thuật toán có thể tự phát hiện các nhóm dữ liệu có nhiệt độ tương đồng hoặc kiểu thời tiết giống nhau. Khi quan sát kết quả, bạn có thể nhận ra rằng các nhóm nhiệt độ tương ứng với bốn mùa, hoặc các mẫu thời tiết giống nhau có thể đại diện cho các điều kiện như mưa, tuyết hay sương mù.

Mục tiêu và ứng dụng của Unsupervised Learning: Mục tiêu của việc Học không giám sát chính làm tìm ra các mẫu ẩn, nhóm dữ liệu tương tự. Học không giám sát mang nhiều lợi ích quan trọng trong bối cảnh dữ liệu lớn (Big Data) đang ngày càng phổ biến. Khả năng tự học từ những dữ liệu "thô" mà không cần sự can thiệp từ con người làm cho nó trở thành một công cụ mạnh mẽ trong nhiều lĩnh vực.

- Ứng dụng của Unsupervised Learning hiện nay:
    - Phân khúc khách hàng (Customer Segmentation): Các doanh nghiệp sử dụng các thuật toán phân cụm để chia khách hàng thành các nhóm dựa trên hành vi mua sắm hoặc sở thích.
    - Gợi ý sản phẩm (Recommendation Systems): Các hệ thống gợi ý sản phẩm, như trên các trang thương mại điện tử (Amazon,...) hoặc dịch vụ xem phim trực tuyến, sử dụng Unsupervised Learning để đề xuất các sản phẩm hoặc nội dung mà người dùng có thể quan tâm.
    - Phát hiện gian lận (Fraud Detection): Unsupervised Learning được sử dụng để xác định các giao dịch hoặc hoạt động đáng ngờ có khả năng là gian lận
    - …

Phân biệt Clustering và Dimensionality Reduction:

- **Phân cụm (Clustering):** Là một bài toán nhóm toàn bộ dữ liệu thành các nhóm nhỏ dựa trên sự liên quan giữa các dữ liệu trong mỗi nhóm. Một thuật toán phổ biến là K-Means, chia dữ liệu thành K cụm, trong đó mỗi điểm dữ liệu thuộc về cụm có trung tâm gần nhất.
- **Giảm chiều dữ liệu (Dimensionality Reduction):** Là kĩ thuật giảm bớt số lượng biến đầu vào của tập dữ liệu mà vẫn giữ lại được những thông tin quan trọng nhất.

-> Bài toán **Phân cụm** có mục tiêu chính là tìm ra các nhóm dữ liệu có đặc điểm tương đồng nhau, còn **Giảm chiều** là tìm ra đặc trưng quan trọng nhất, giữ lại rồi giảm biến để dữ liệu gọn nhẹ hơn

Bảng so sánh giữa Clustering và Dimensionality Reduction:

Đặc điểm	Clustering (Phân cụm)	Dimensionality Reduction (Giảm chiều)
Mục tiêu chính	Tìm ra các nhóm (groups) dữ liệu có đặc điểm tương đồng nhau.	Tìm ra các đặc trưng quan trọng nhất để biểu diễn dữ liệu gọn nhẹ hơn.
Câu hỏi giải quyết	"Những điểm dữ liệu nào giống nhau?" (Who is like whom?)	"Những thông tin nào là thừa thãi?" (What is redundant?)
Tác động lên dữ liệu	Giữ nguyên số chiều, nhưng gán thêm nhãn nhóm cho từng điểm dữ liệu.	Giữ nguyên số điểm dữ liệu, nhưng giảm số lượng biến (chiều) của mỗi điểm.
Kết quả đầu ra	Một nhãn nhóm (Cluster ID) cho mỗi mẫu (VD: Khách hàng A thuộc nhóm VIP).	Một tập hợp các đặc trưng mới ít hơn (VD: Từ 100 cột giảm còn 3 cột).
Thuật toán tiêu biểu	K-Means, DBSCAN, Hierarchical Clustering.	PCA, t-SNE, Autoencoders.

Các thuật toán tiêu biểu của Supervised Learning và Unsupervised Learning:

- Linear Regression: Tìm một đường thẳng phù hợp nhất (Best Fit Line) để dự đoán giá trị đầu ra dựa trên đầu vào.

- Logistic Regression: Sử dụng hàm Sigmoid để ánh xạ đầu ra của một phép tính tuyến tính về xác suất, từ đó phân loại.

- Decision Tree: Xây dựng một cấu trúc cây bằng cách chia dữ liệu thành các nhánh dựa trên thuộc tính tốt nhất, cho đến khi mỗi nhánh chứa các mẫu dữ liệu đồng nhất.

- Random Forest: Là một tập hợp của nhiều Decision Tree. Random Forest huấn luyện từng cây trên các tập con dữ liệu ngẫu nhiên và lấy kết quả dự đoán đa số hoặc trung bình.

- Support Vector Machine (SVM): Mục tiêu của thuật toán SVM là tìm ra một siêu phẳng (hyperplane) sao cho có thể phân tách tối ưu các điểm dữ liệu thuộc hai lớp khác nhau. “Tối ưu” ở đây nghĩa là tìm siêu phẳng tạo ra khoảng cách lớn nhất (margin) giữa hai lớp dữ liệu.

- K-Nearest Neighbors (KNN): Khi cần phân loại một điểm dữ liệu mới, thuật toán sẽ tìm K điểm gần nhất (nearest neighbors) trong tập dữ liệu huấn luyện và xác định nhãn dựa trên nhãn xuất hiện nhiều nhất trong số các điểm này.

- K-Means: Chia dữ liệu thành K cụm. Thuật toán cố gắng giảm thiểu tổng bình phương khoảng cách từ mỗi điểm tới trọng tâm của cụm gần nhất.

- Hierarchical Clustering: Xây dựng một cây phân cấp để nhóm các điểm dữ liệu lại với nhau theo khoảng cách. Có hai hướng là từ nhỏ đến lớn hoặc từ lớn đến nhỏ.

- DBSCAN: Là một thuật toán được sử dụng rộng rãi trong Machine Learning. Là một cụm bao gồm một vùng điểm dày đặc, được phân tách với các cụm khác bằng các vùng có mật độ thấp hơn.

- Principal Component Analysis (PCA): Tìm ra các Thành phần chính là các trục (vector) mà dữ liệu biến thiên mạnh nhất, chiếu dữ liệu lên các trục mới để nén thông tin

- t-distributed Stochastic Neighbor Embedding): Tập trung vào việc giữ lại mối quan hệ lân cận cục bộ của dữ liệu khi chiếu dữ liệu từ không gian cao chiều xuống 2D hoặc 3D.