1. **MỘT SỐ KHÁI NIỆM CƠ BẢN:**
2. **Feature (đặc trưng, thuộc tính):**

* **Định nghĩa:** Là các yếu tố hoặc biến độc lập mà bạn sử dụng để dự đoán kết quả. Mỗi feature đại diện cho một đặc điểm của dữ liệu mà bạn sử dụng trong mô hình máy học.
* **Ví dụ:** Trong một bài toán dự đoán giá nhà, các feature có thể là diện tích nhà, số phòng ngủ, vị trí, năm xây dựng, v.v.

1. **Label:**

* **Định nghĩa:** Là kết quả hoặc biến phụ thuộc mà mô hình dự đoán. Trong một bài toán học có giám sát (supervised learning), label là những giá trị mà mô hình sẽ học để dự đoán dựa trên các feature. Label thường là dữ liệu thực tế mà bạn đã có và sẽ dùng để huấn luyện mô hình.
* **Ví dụ:** Trong bài toán dự đoán giá nhà, label sẽ là giá nhà thực tế.

(Trong **tabular data** (dữ liệu dạng bảng), **feature** và **label** thường được biểu diễn dưới dạng các cột. Trong một bảng dữ liệu, bạn có thể có nhiều feature (nhiều cột thông tin), nhưng thường chỉ có một label (cột mục tiêu) mà mô hình cần dự đoán.)

1. **Histogram:**

* Định nghĩa: **Histogram** là một biểu đồ dạng cột được sử dụng để biểu diễn phân bố tần suất của dữ liệu số. Nó cho thấy cách các giá trị trong một tập dữ liệu được phân chia thành các khoảng giá trị (bins).

1. **PHƯƠNG PHÁP HỌC MÁY SUPERVISED LEARNING VÀ UN-SUPERVISED LEARNING:**
2. **Phương Pháp Supervised Learning (Học Máy có giám sát):**

* **Dữ liệu:** Được gắn nhãn (label) đầy đủ, bao gồm cả đầu vào (features) và đầu ra (labels).
* **Mục tiêu:** Học từ dữ liệu có nhãn để dự đoán nhãn cho dữ liệu mới chưa biết.
* **Ứng dụng:** Phân loại (classification), hồi quy (regression).
* **Ví dụ:** Dự đoán giá nhà dựa trên các đặc điểm như diện tích, vị trí.

1. **Phương Pháp Unsupervised Learning (Học không giám sát):**

* **Dữ liệu:** Không có nhãn, chỉ bao gồm đầu vào (features).
* **Mục tiêu:** Tìm ra cấu trúc ẩn, mẫu (patterns) hoặc nhóm (clusters) trong dữ liệu.
* **Ứng dụng:** Phân cụm (clustering), giảm số chiều (dimensionality reduction).
* **Ví dụ:** Phân cụm khách hàng dựa trên hành vi mua sắm.

1. **So sánh:**

* **Dữ liệu:** Supervised cần dữ liệu có nhãn; Unsupervised không cần.
* **Mục tiêu:** Supervised dự đoán kết quả dựa trên dữ liệu quá khứ; Unsupervised khám phá cấu trúc ẩn trong dữ liệu.
* **Ứng dụng:** Supervised dùng trong dự đoán chính xác; Unsupervised dùng trong khám phá dữ liệu.

1. **ỨNG DỤNG CỦA PHƯƠNG PHÁP HỌC MÁY:**
   1. **Kỹ Thuật Classification:**

* **Mục tiêu:** Dự đoán nhãn (label) của các quan sát mới, trong đó nhãn là các giá trị rời rạc (discrete).
* **Ví dụ:** Phân loại email là "spam" hay "not spam", dự đoán một bệnh nhân mắc bệnh hay không (có/không).
  1. **Kỹ thuật Regression:**
* **Mục tiêu:** Dự đoán giá trị liên tục (continuous) cho các quan sát mới.
* **Ví dụ:** Dự đoán giá nhà dựa trên diện tích, dự đoán doanh số bán hàng.
  1. **So Sánh:**
* **Đầu ra:** Classification cho kết quả là nhãn rời rạc; Regression cho kết quả là giá trị liên tục.
* **Ứng dụng:** Classification dùng trong bài toán phân loại, quyết định; Regression dùng trong bài toán dự đoán giá trị cụ thể.
* **Ví dụ:** Classification phân loại khách hàng theo nhóm; Regression dự đoán thu nhập dựa trên kinh nghiệm.