**Câu 1.** So sánh sự khác nhau giữa lập trình hệ thống và lập trình trí tuệ nhân tạo?

**Câu 2.** Làm rõ khái niệm của các thuật ngữ sau: Classification, Instance/Sample, Label, Training dataset, Testing dataset, Feature, Feature Selection, Feature Reduction. Vẽ mô hình tổng quan cho quy trình Classification. Cho ví dụ về một tập dataset áp dụng cho mô hình phân lớp.

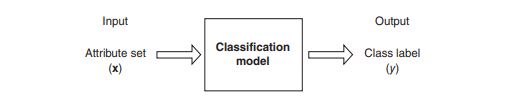
**Câu 3:** Confusion matrix là gì? Trình bày công thức tính toán cho confusion matrix, giải thích rõ từng tham số.

BÀI LÀM

**Bài 2.**

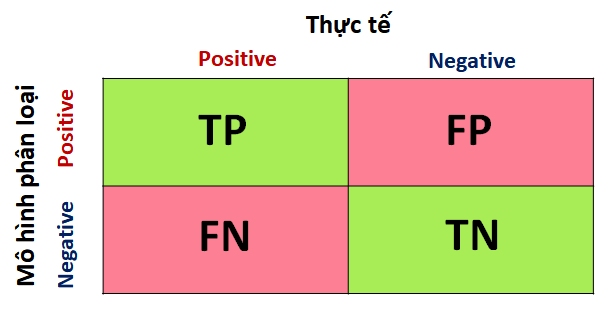
* Classification : Classification là việc sử dụng input là một tập các mẫu dữ liệu huấn luyện, với một nhãn phân lớp cho mỗi mẫu dữ liệu. Output là mô hình (bộ phân lớp) dựa trên tập huấn luyện và những nhãn phân lớp . Hệ thống Classification để dự đoán những nhãn phân lớp cho các bộ dữ liệu/mẫu mới.
* **instance**(bản ghi/mẫu dữ liệu). Mỗi instance được đặc tả bởi một tuple(x, y), trong đó x là tập giá trị của các thuộc tính mô tả instance đó, và y là **class label**(nhãn lớp) của instance.
* ***Tập dữ liệu huấn luyện* (Training dataset):** là một tập dữ liệu dùng để huấn luyện cho mô hình của thuật toán Máy học. Nói một cách khác, thuật toán Máy học sẽ học từ tập dữ liệu này.
* ***Tập dữ liệu kiểm tra* (Testing dataset):** là một tập dữ liệu dùng để kiểm chứng độ chính xác của mô hình thuật toán Máy học. Tập dữ liệu này không được dùng để huấn luyện mô hình.
* **Đặc trưng (feature):** là một thuộc tính hoặc đặc tính có thể đo lường riêng lẻ của một mẫu dữ liệu đang được quan sát.

**\*Mô hình tổng quan của quá trình phân lớp :**



**Bài 3.**

* Confusion Matrix(ma trận lỗi): thể hiện có bao nhiêu điểm dữ liệu thực sự thuộc vào một class, và được dựđoán là rơi vào một class.
* Một confusion matrix gồm 4 chỉ số sau đối với mỗi lớp phân loại:



* **TP (True Positive)**: Số lượng dự đoán chính xác. VD:Là khi mô hình dự đoán đúng một người bị ung thư.
* **TN (True Negative)**: Số lương dự đoán chính xác một cách gián tiếp. VD:Là khi mô hình dự đoán đúng một người không bị ung thư
* **FP (False Positive)**: Số lượng các dự đoán sai lệch. VD:Là khi mô hình dự đoán một người bị ung thư và người đó hoàn toàn khỏe mạnh.
* **FN (False Negative)**: Số lượng các dự đoán sai lệch một cách gián tiếp. VD:Là khi mô hình dự đoán một người không bị ung thư nhưng người đó bị ung thư

**\*Công thức tính:**

**\*Độ chính xác:** Trong tất cả các dự đoán , tỷ lệ dự đoán đúng là bao nhiêu?

*Accuracy=*

***=***

***\*Tỷ lệ lỗi(Error rate):***

***Error rate = =***

**\*Precision**: Trong tất cả các dự đoán Positive được đưa ra, bao nhiêu dự đoán là chính xác? Chỉ số này được tính theo công thức:

**Precision=**

**\*Recall**: Trong tất cả các trường hợp Positive, bao nhiêu trường hợp đã được dự đoán chính xác? Chỉ số này được tính theo công thức:

**Recall=**