I – Introduce:

1. Machine Learning ?
2. Lí do chọn đề tài

* Thực tế:
  + Lí do chọn ngành y khoa
    - Nhu cầu hỗ trợ bác sĩ trong y khoa càng lúc càng tang theo thời gian, và số lượng bệnh nhân càng lúc càng tang và các loại bệnh cơ bản lặp đi lặp lại nhiều nên cần có công cụ hỗ trợ các bác sĩ chuẩn đoán nhanh các bệnh cơ bản đó.
  + Lí do chọn bệnh phổi
    - Vì quá trình công nghiệp hóa hiện đại hóa đang diễn ra với tốc độ nhanh chóng, từ đó kéo theo rất vấn đề như: ô nhiễm môi trường đặc biệt là không khí ,… Dẫn đến các bệnh liên quan về đường hô hấp gia tang và ngày càng phổ biến. Nên có 1 lượng dữ liệu lớn về bệnh phổi - sau khi chuẩn đoán (bằng phương pháp chụp X - ray).
    - Số liệu thực tế: … (Lấy từ bệnh viện mô ? Thời gian ?)
* Tương lai
  + Machine Learning nói riêng và AI nói chung trong tương lai sẽ phát triển rất nhanh chóng và hỗ trợ, phục vụ con người trong nhiều lĩnh vực như: kinh tế, công nghiệp, đời sống, giáo dục,… và đặt biệt là y khoa ().
  + Bệnh càng ngày càng nhiều và trở nên phổ biến kéo theo lượng dữ liệu sẽ tang lên rất nhiều lần so với hiện tại nên cần các công cụ hỗ trợ mang tính bước ngoặt trong lĩnh vực y tế

1. Ý nghĩa đề tài

* Áp dụng Machine Learning & AI vào trong y khoa sẽ giúp các bác sĩ dễ dàng và nhanh chóng hơn trong việc chuẩn đoán bệnh để từ đó kê đơn thuốc đúng và cho từng loại bệnh của các bệnh nhân
* Giảm thiểu sự sai sót trong chuẩn đoán bệnh
* Hỗ trợ các bác sĩ từ các quốc gia có nên y khoa chưa phát triển từ đó giúp giảm thiểu rủi ro tử vong của bệnh nhân vì chẩn đoán không chính xác

VD: Mỗi năm, riêng ở Mỹ, ước tính có khoảng 40.000 đến 80.000 người chết do chẩn đoán sai. (theo BCC News: Lỗi chẩn đoán gây tử vong cho bệnh nhân)

II – Goal:

* Hoàn thành dự án đúng mong đợi.
* Từ dữ liệu có sẵn, training ra được bộ model cho chẩn đoán sau này.
* Với hình ảnh test sẽ dự đoán ra được có bệnh hay không có bệnh
* Dự đoán bệnh với độ chính xác cao (trên 80%)
* Có thể tận dụng dự án vào trong thực tế

III – Objective

Các bước thực hiện dự án:

1. Đặt đúng vấn đề:
   * Tìm hiểu đúng thực trạng và nhu cầu sử dụng
   * Phân tích vấn đề thực tế
   * Hội ý và đưa ra giải pháp
2. Chuẩn bị dữ liệu:
   * Tìm nguồn cung cấp dữ liệu thực tế
   * Phân loại dữ liệu
   * Phân tích dữ liệu
   * Chuyển đổi dữ liệu dạng hình, nhãn thành dạng dữ liệu matrix theo vector
3. Chọn thuật toán
   * Nghiên cứu và tìm hiểu các thuật toán cơ bản và phổ biến trong Machine Learing
   * Xác định các thuật toán phù hợp với dự án
   * Áp dụng các thuật toán đã chọn vào dự án
4. Huấn luyện
   * Từ dữ liệu đã xử lí trước , thông qua việc áp dụng các thuật toán để tạo ra 1 bộ model để từ đó so sánh với dữ liệu test và cho ra kết quả theo mong đợi
5. Kiểm thử
   * Input hình cần chẩn đoán sau khi dự đoán sẽ cho ra kết quả với số liệu về độ chính xác (theo %)

IV – Theory:

1. Tiền xử lí:
2. Thuật toán train:
   1. Vấn đề:

Sau khi có một tập các ảnh dữ liệu training là ảnh X-ray về bệnh phổi, và đã kéo được vector ảnh số cho tập dữ liệu này, làm sau để hình thành Model cho bài toán từ tập dữ liệu vector để biết được ảnh nào (bộ vector nào) là có bệnh, và ảnh nào (bộ vector nào) là không có bệnh?

* 1. Hướng giải quyết:

Sau khi chuyển tất cả các ảnh X-ray về dạng vector ảnh số, ta phải tìm ra quy luật phân biệt của các vector ảnh, sau đó dùng thuật toán Support Vector Machine để tìm ra một siêu phẳng để phân chia các ảnh về 2 phía, dựa theo label ta sẽ biết được ảnh nào là có bệnh và ảnh nào là không có bệnh.

Cách giải quyết là phải tìm được một siêu phẳng để phân chia 2 phần có bệnh và không có bệnh sao cho công bằng nhất.

* + Giả sử ban đầu ta có bộ dữ liệu training với label có bệnh và không có bệnh có dạng sau:

()

Trong đó:

* : Là là vector ảnh số được kéo ra từ ảnh X-ray
* : Là label với 2 giá trị là có bệnh / không có bệnh. (1/-1 : Tùy theo cách đặt nhãn)
  + - : sẽ là mặt phẳng phân chia giữa 2 lớp có bệnh và không có bệnh
* Vậy với cặp dữ liệu thì khoảng cách từ điểm đó đến mặt phẳng phân chia là:

Với trong đó d là số chiều không gian trên vector ảnh số

* Nếu ta tìm được min(|(1)|) thì ta sẽ tìm được khoảng gần nhất từ một điểm trong 2 lớp đến siêu phẳng (margin).
* Để tối ưu hóa bài toán này ta phải tìm được **w** và b sao cho margin là lớn nhất

V – Model: