# **Report: Lung Cancer Risk Analysis and AI Platform**

## **1. Executive Summary**

### **Mục tiêu Dự án**

Mục tiêu chính của dự án là phân tích một cách toàn diện dữ liệu sức khỏe của bệnh nhân nhằm nhận diện các yếu tố rủi ro có liên quan đến ung thư phổi – một trong những nguyên nhân gây tử vong hàng đầu trên toàn cầu. Dựa trên các đặc trưng như hành vi sinh hoạt (hút thuốc, sử dụng rượu), môi trường sống (ô nhiễm không khí, điều kiện nghề nghiệp), tiền sử bệnh lý (di truyền, bệnh phổi mãn tính) và các triệu chứng lâm sàng (ho ra máu, đau ngực, khó thở,...), dự án hướng đến việc khai thác các mối quan hệ ẩn trong dữ liệu nhằm phát hiện sớm khả năng mắc bệnh.

Trên cơ sở đó, nhóm phát triển xây dựng một nền tảng trí tuệ nhân tạo (AI) có khả năng học hỏi từ dữ liệu thực tế và đưa ra các dự đoán có ý nghĩa y học. Nền tảng này không chỉ phục vụ cho mục đích học thuật mà còn có tính ứng dụng thực tiễn cao, giúp người dân dễ dàng tiếp cận với các công cụ đánh giá nguy cơ ung thư phổi thông qua các giao diện đơn giản và dễ sử dụng.

### **Kết quả Chính**

* **Thu thập và xử lý dữ liệu:** Dự án sử dụng bộ dữ liệu Lung Cancer Dataset, bao gồm **1000 dòng dữ liệu tương ứng với 1000 bệnh nhân**, cùng **21 đặc trưng đầu vào** liên quan đến hành vi cá nhân, lối sống, môi trường sống và các biểu hiện sức khỏe. Dữ liệu đã được làm sạch, chuẩn hóa và mã hóa để sẵn sàng phục vụ cho phân tích và huấn luyện mô hình.
* **Phân tích dữ liệu:** Đã tiến hành các kỹ thuật phân tích thống kê và trực quan hóa dữ liệu để khám phá các xu hướng, phân bố và mối tương quan giữa các yếu tố đầu vào với nguy cơ ung thư phổi. Qua đó, bước đầu xác định được những đặc trưng nổi bật như hút thuốc, ô nhiễm không khí và yếu tố di truyền có mối liên hệ chặt chẽ với khả năng mắc bệnh.
* **Huấn luyện mô hình dự đoán:** Ứng dụng các thuật toán học máy như Decision Tree, Random Forest để xây dựng mô hình dự đoán nguy cơ ung thư phổi dựa trên dữ liệu đầu vào. Các mô hình được đánh giá bằng các chỉ số như độ chính xác, độ nhạy, độ đặc hiệu và AUC. Kết quả bước đầu cho thấy mô hình đạt độ chính xác 92**%**
* **Triển khai chatbot hỗ trợ:** Sử dụng nền tảng **Botpress**, một chatbot thông minh đã được tích hợp vào hệ thống để trả lời các câu hỏi phổ biến về nguy cơ ung thư phổi. Chatbot có khả năng hiểu ngôn ngữ tự nhiên, tương tác trực tiếp với người dùng và hỗ trợ tư vấn sơ bộ dựa trên dữ liệu đầu vào cung cấp.
* **Xây dựng ứng dụng web AI:** Phát triển một ứng dụng web hoàn chỉnh bằng **Flask**, tích hợp toàn bộ pipeline xử lý dữ liệu – từ tiếp nhận dữ liệu người dùng, xử lý đầu vào, đưa vào mô hình dự đoán, cho đến hiển thị kết quả và phản hồi bằng chatbot. Ứng dụng có giao diện đơn giản, thân thiện, dễ triển khai trên các nền tảng cloud hoặc nội bộ.

## **2. Business Understanding**

**Vấn đề cần giải quyết:**  
Ung thư phổi là nguyên nhân hàng đầu gây tử vong do ung thư trên toàn thế giới, một phần do bệnh thường tiến triển âm thầm và được phát hiện khi đã ở giai đoạn muộn. Việc xét nghiệm, chụp CT hoặc sinh thiết để chẩn đoán sớm không phải lúc nào cũng khả thi với mọi người dân, đặc biệt tại các vùng có điều kiện y tế hạn chế. Tuy nhiên, nhiều yếu tố nguy cơ của ung thư phổi – như hút thuốc, ô nhiễm không khí, tiền sử bệnh phổi – lại là những yếu tố có thể được theo dõi, ghi nhận thông qua dữ liệu phổ thông.

Dự án đặt ra bài toán cốt lõi:

**Làm thế nào để sớm phát hiện nguy cơ ung thư phổi từ dữ liệu hành vi và sức khỏe phổ thông, mà không cần đến các phương pháp y học tốn kém và xâm lấn?**

Câu hỏi này mở ra cơ hội ứng dụng các phương pháp khoa học dữ liệu và trí tuệ nhân tạo (AI) để xây dựng một hệ thống cảnh báo nguy cơ sớm, hỗ trợ cá nhân và y bác sĩ trong việc theo dõi, phòng ngừa bệnh hiệu quả hơn.

**Giá trị mang lại:**

1. **Hỗ trợ người dân tự đánh giá nguy cơ cá nhân:**  
   Hệ thống AI có thể tiếp nhận thông tin đơn giản như tuổi, giới tính, tình trạng hút thuốc, mức độ tiếp xúc với ô nhiễm, triệu chứng hô hấp,... để đưa ra nhận định sơ bộ về mức độ rủi ro mắc ung thư phổi. Qua đó, người dùng có thể ý thức hơn về tình trạng sức khỏe của mình và chủ động đi khám khi cần thiết.
2. **Xác định các yếu tố nguy cơ quan trọng nhất (feature importance):**  
   Thông qua việc huấn luyện mô hình học máy, hệ thống có khả năng xác định các yếu tố có ảnh hưởng lớn nhất đến kết quả dự đoán. Điều này giúp người dùng hiểu rõ hơn các nguyên nhân tiềm ẩn gây bệnh và có thể điều chỉnh hành vi sinh hoạt để giảm thiểu nguy cơ. Đồng thời, thông tin này cũng giúp bác sĩ và nhà quản lý y tế định hướng các biện pháp phòng ngừa trong cộng đồng.
3. **Giảm tải cho bác sĩ với chatbot trả lời tự động:**  
   Một chatbot AI có khả năng trả lời các câu hỏi phổ biến về bệnh lý, triệu chứng, yếu tố nguy cơ và hướng dẫn ban đầu sẽ giúp giảm áp lực lên đội ngũ y tế, đặc biệt là trong giai đoạn đầu tương tác với người bệnh. Chatbot có thể hoạt động 24/7, tiếp cận dễ dàng qua nền tảng web hoặc mạng xã hội, góp phần nâng cao hiệu quả giáo dục sức khỏe cộng đồng.

## **3. Data Understanding**

* **Nguồn dữ liệu**: Bộ dữ liệu tổng hợp về bệnh nhân (tên: Lung Cancer Dataset)
* **Số dòng**: XX
* **Số cột**:
* **Mô tả ban đầu**:
  + Có missing values ở các trường: ...
  + Một số thuộc tính dạng categorical cần encoding.

## **4. Data Preparation**

* **Xử lý missing values**: Phương pháp áp dụng: mean/mode imputation.
* **Encoding**: LabelEncoder hoặc OneHotEncoder áp dụng cho GENDER, SMOKING...
* **Chuẩn hóa dữ liệu**: Sử dụng StandardScaler với dữ liệu số.
* **Tập dữ liệu sau xử lý**: lung\_cancer\_cleaned.csv, số dòng còn lại: XX

## **5. Exploratory Data Analysis (EDA)**

* **Phân tích phân phối thuộc tính**:
  + Biểu đồ histogram theo độ tuổi, giới tính
* **Tương quan giữa các yếu tố**:
  + Heatmap: CORRELATION giữa các yếu tố với nguy cơ ung thư
* **Phân tích đặc trưng nổi bật**:
  + SMOKING, OXYGEN\_SATURATION, CHEST\_TIGHTNESS có tương quan mạnh nhất với nhãn mục tiêu.

## **6. Data Visualization**

* **Các biểu đồ minh họa đã sử dụng**:
  + Bar chart: Tỷ lệ người mắc theo giới tính
  + Pie chart: Tỷ lệ hút thuốc
  + Scatter plot: SMOKING vs OXYGEN\_SATURATION
  + Word Cloud: Từ khóa mô tả triệu chứng (nếu có text)
* **Ý nghĩa rút ra từ biểu đồ**:
  + Có xu hướng rõ ràng giữa hút thuốc và tỷ lệ mắc bệnh
  + Mức bão hòa oxy thấp là yếu tố nguy cơ tiềm tàng

## **7. Modeling**

* **Thuật toán sử dụng**: (ví dụ) Random Forest Classifier
* **Chia tập train/test**: 80/20
* **Kết quả đánh giá**:
  + Accuracy: XX%
  + Precision, Recall, F1-score: trình bày theo bảng
  + Confusion Matrix
  + ROC Curve & AUC: XX
* **Feature Importance**: Xếp hạng các đặc trưng quan trọng nhất

## **8. Application Development**

* **Framework**: Flask
* **Các chức năng chính của ứng dụng**:
  + Upload dữ liệu bệnh nhân
  + Xem kết quả phân tích dữ liệu
  + Dự đoán nguy cơ mắc bệnh
  + Tương tác với chatbot hỗ trợ
* **Ảnh minh họa**: ảnh chụp màn hình giao diện từng trang

## **9. Chatbot Design (Botpress)**

* **Mục đích**: Hướng dẫn người dùng hiểu về dữ liệu, yếu tố nguy cơ và mô hình
* **Tính năng chính**:
  + Trả lời câu hỏi về đặc trưng dữ liệu (e.g., "Thế nào là CHEST\_TIGHTNESS?")
  + Giải thích mô hình dự đoán
  + Chuyển tiếp các câu hỏi khó sang chuyên viên
* **Flow thiết kế**: trình bày sơ đồ (diagram hoặc mô tả bằng text)

## **10. Deployment**

* **Flask** chạy trên: localhost:5000
* **Botpress** nhúng qua iframe: localhost:3000/...

## **11. Challenges & Limitations**

* Kích thước dataset còn hạn chế
* Một số biến chưa rõ ý nghĩa lâm sàng (giả định)
* Mô hình không thay thế chẩn đoán y tế thực tế
* Chatbot không thể trả lời các câu hỏi cá nhân nhạy cảm

## **12. Conclusion & Future Work**

* **Kết luận**:
  + Nền tảng đã đạt mục tiêu cơ bản về phân tích, trực quan hóa và dự đoán
* **Hướng phát triển tiếp**:
  + Mở rộng chatbot bằng RAG để truy xuất kiến thức y khoa
  + Triển khai web app trên cloud (Heroku, Render, GCP)
  + Fine-tune mô hình với dữ liệu thực tế từ bệnh viện (nếu có)
  + Thêm phân tích thời gian (longitudinal risk)

## **13. Appendix**

* Mã nguồn quan trọng (trích đoạn)
* Tài liệu tham khảo (WHO, CDC, Papers)
* Hướng dẫn sử dụng ứng dụng