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

## **1. Executive Summary**

* **Mục tiêu dự án**: Phân tích dữ liệu sức khỏe bệnh nhân để nhận diện các yếu tố rủi ro liên quan đến ung thư phổi và xây dựng một nền tảng AI hỗ trợ dự đoán nguy cơ.
* **Kết quả chính**:
  + Đã xử lý và phân tích dữ liệu sức khỏe gồm 18 đặc trưng.
  + Huấn luyện mô hình dự đoán đạt độ chính xác XX%.
  + Triển khai chatbot Botpress hỗ trợ tư vấn người dùng.
  + Xây dựng ứng dụng Flask tích hợp đầy đủ pipeline AI.

## **2. Business Understanding**

* **Vấn đề cần giải quyết**:
  + Làm thế nào để sớm phát hiện nguy cơ ung thư phổi từ dữ liệu hành vi và sức khỏe phổ thông?
* **Giá trị mang lại**:
  + Hỗ trợ người dân đánh giá nguy cơ cá nhân.
  + Gợi ý các yếu tố nguy cơ quan trọng nhất (feature importance).
  + Giảm tải cho bác sĩ với chatbot trả lời tự động.

## **3. Data Understanding**

* **Nguồn dữ liệu**: Bộ dữ liệu tổng hợp về bệnh nhân (tên: Lung Cancer Dataset)
* **Số dòng**: XX
* **Số cột**: 18 đặc trưng gồm:
  + Nhân khẩu học: AGE, GENDER
  + Lối sống: SMOKING, ALCOHOL\_CONSUMPTION
  + Triệu chứng: BREATHING\_ISSUE, CHEST\_TIGHTNESS,...
* **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