**BỘ GIÁO DỤC VÀ ĐÀO TẠO**

**TRƯỜNG ĐẠI HỌC NGOẠI NGỮ – TIN HỌC TP. HỒ CHÍ MINH**

**KHOA CÔNG NGHỆ THÔNG TIN**



**ĐỒ ÁN MÔN HỌC:**

**BIG DATA**

**ĐỀ TÀI:** Xây dựng hệ thống dự đoán mức độ mắc bệnh ung thư

**GIÁO VIÊN HƯỚNG DẪN**:

**Nhóm: 10**

Trần Quốc Bảo

Huỳnh Đức Trọng

TP.HCM, tháng 11 năm 2024

# Giới thiệu

Trong bối cảnh ung thư đang trở thành một trong những căn bệnh nguy hiểm hàng đầu và có xu hướng gia tăng trên toàn thế giới, việc phát hiện và chẩn đoán sớm đóng vai trò then chốt trong nâng cao hiệu quả điều trị và cải thiện tỷ lệ sống sót của bệnh nhân. Đề tài nghiên cứu này tập trung vào ứng dụng công nghệ Big Data để lưu trữ, thống kê và tìm kiếm dữ liệu y khoa từ nhiều nguồn khác nhau như bệnh án, kết quả xét nghiệm và các chỉ số sinh học của bệnh nhân. Dữ liệu được quản lý và xử lý một cách hiệu quả nhằm xây dựng một hệ thống dự đoán mức độ nguy cơ mắc bệnh ung thư.

Hệ thống dựa trên sự kết hợp giữa các phương pháp học máy tiên tiến – cụ thể là Hồi quy Logistic, Cây Quyết Định và Random Forest – được phát triển để phân tích và dự đoán nguy cơ ung thư. Qua đó, mô hình không chỉ giúp phát hiện sớm các trường hợp nguy cơ cao mà còn hỗ trợ các bác sĩ trong việc đưa ra quyết định điều trị kịp thời và tối ưu. Sự hội nhập giữa Big Data và Machine Learning hứa hẹn mang lại một giải pháp toàn diện, góp phần cải thiện quy trình chẩn đoán và chăm sóc bệnh nhân trong bối cảnh y tế hiện đại.

# Lời cảm ơn

Đầu tiên, chúng em xin gửi lời cảm ơn sâu sắc tới các Thầy, Cô của ngành Công nghệ Thông tin, trường Đại học Ngoại ngữ - Tin học TP. HCM, đặc biệt là thầy Nguyễn Thanh Bình, người đã tận tình hướng dẫn, hỗ trợ và truyền đạt cho chúng em những kiến thức quý báu trong môn Thị giác máy tính. Chính nhờ sự chỉ dạy tận tình và sự khích lệ của Thầy mà chúng em đã có thể hoàn thành đề tài này một cách tốt nhất.

Do kiến thức còn nhiều hạn chế, chúng em rất mong nhận được ý kiến đóng góp từ quý Thầy, Cô để hoàn thiện hơn trong cách trình bày và hiểu biết. Một lần nữa, chúng em xin chân thành cảm ơn và tri ân quý Thầy, Cô.

# Mục lục

Contents

[Giới thiệu i](#_Toc192695021)

[Lời cảm ơn ii](#_Toc192695022)

[Mục lục vi](#_Toc192695023)

[1 XÁC ĐỊNH ĐỀ TÀI 2](#_Toc192695024)

[1.1 Giới thiệu đề tài 2](#_Toc192695025)

[1.2 Mục tiêu và nội dung đề tài 2](#_Toc192695026)

[1.3 Giới hạn đề tài 2](#_Toc192695027)

[1.4 Cấu trúc báo cáo 4](#_Toc192695028)

[2 CƠ SỞ LÝ THUYẾT 4](#_Toc192695029)

[2.1 Tổng quan về bệnh ung thư và các yếu tố nguy cơ 4](#_Toc192695030)

[2.2 Big Data và công cụ trong hệ thống dự đoán 5](#_Toc192695031)

[2.3 Phương pháp dự đoán mức độ mắc bệnh ung thư 5](#_Toc192695032)

[3 PHƯƠNG PHÁP 5](#_Toc192695033)

[3.1 Yêu cầu bài toán: 5](#_Toc192695034)

[3.2 Giải thuật thực hiện 5](#_Toc192695035)

[3.3 Phương pháp đánh giá 6](#_Toc192695036)

[4 HIỆN THỰC KẾT QUẢ 6](#_Toc192695037)

[4.1 Yêu cầu hệ thống và tập dữ liệu thực nghiệm (nếu có) 6](#_Toc192695038)

[4.2 Tập dữ liệu 6](#_Toc192695039)

[4.3 Kết quả thực nghiệm 7](#_Toc192695040)

[4.4 Kết quả đạt được 8](#_Toc192695041)

[4.5 Ưu và nhược điểm 8](#_Toc192695042)

[4.6 Hướng mở rộng 8](#_Toc192695043)

[1 BẢNG PHÂN CÔNG NHIỆM VỤ CHUNG 9](#_Toc192695044)

[TÀI LIỆU THAM KHẢo 10](#_Toc192695045)

# XÁC ĐỊNH ĐỀ TÀI

## Giới thiệu đề tài

Đề tài nghiên cứu này tập trung vào việc ứng dụng Big Data để lưu dữ liệu để thống kê và tìm kiếm, sau đó áp dụng Machine learning trên hệ thống đã xây dựng để dự đoán khả năng mắc bệnh ung thư. Ung thư hiện nay là một trong những bệnh lý nguy hiểm và đang gia tăng trên toàn thế giới. Việc phát hiện ung thư ở giai đoạn sớm có thể giúp tăng khả năng điều trị và cải thiện tỷ lệ sống sót của bệnh nhân. Bằng cách phân tích các chỉ số y khoa, các mô hình học máy có thể giúp dự đoán nguy cơ mắc bệnh ung thư, hỗ trợ bác sĩ trong việc chẩn đoán và điều trị kịp thời.

## Mục tiêu và nội dung đề tài

* Mục tiêu của đề tài là xây dựng và thử nghiệm một mô hình dự đoán bệnh nhân có khả năng mắc ung thư dựa trên các thông tin đầu vào là các chỉ số y học. Cụ thể, các mục tiêu chính bao gồm:
* Lưu trữ và xử lý dữ liệu bệnh nhân.
* Ứng dụng các thuật toán học máy để xây dựng mô hình dự đoán.
* Đánh giá hiệu quả của mô hình dựa trên các chỉ số như độ chính xác, độ nhạy, độ đặc hiệu.
* Nội dung của đề tài:
* Cơ sở lý thuyết về bệnh ung thư và các phương pháp dự đoán.
* Các phương pháp học máy được sử dụng để xây dựng mô hình.
* Thực nghiệm và đánh giá mô hình dự đoán.

## Giới hạn đề tài

Đề tài này chỉ tập trung vào việc tổ chức lưu trữ dữ liệu về các chỉ số của bệnh nhân và phát triển một mô hình dự đoán ung thư dựa trên các dữ liệu đã được lưu trước đó, mà không đi sâu vào nghiên cứu về các phương pháp điều trị ung thư.

 **Giới hạn về dữ liệu đầu vào**: Dữ liệu sử dụng trong nghiên cứu chủ yếu đến các chỉ số đo được từ bệnh nhân không liên quan đến hồ sơ bệnh án và kết quả xét nghiệm y khoa, bao gồm các thông số về đặc điểm sinh học của bệnh nhân và các chỉ số xét nghiệm cụ thể. Hơn nữa, yếu tố như di truyền, tiền sử gia đình cũng không được xem xét. Những yếu tố này có thể ảnh hưởng đến nguy cơ mắc ung thư nhưng không nằm trong phạm vi nghiên cứu của đề tài.

 **Giới hạn về phạm vi bệnh lý**: Do giới hạn về dữ liệu đầu vào nên đề tài chỉ tập trung vào việc dự đoán nguy cơ mắc bệnh ung thư nói chung, thay vì phân loại cụ thể từng loại ung thư (như ung thư vú, ung thư phổi, ung thư gan, v.v.).

 **Giới hạn về lưu trữ và xử lý dữ liệu:** Nghiên cứu tập trung vào việc lưu trữ các chỉ số dữ liệu bệnh nhân với quy mô vừa phải, không mở rộng sang xử lý dữ liệu quy mô lớn. Phạm vi nghiên cứu sẽ chỉ áp dụng các công cụ lưu trữ và xử lý phù hợp với quy mô và đặc thù của dữ liệu được sử dụng.

 **Giới hạn về nền tảng triển khai**: Trong nghiên cứu này, mô hình dự đoán chủ yếu được thực hiện trên nền tảng Python với các thư viện hỗ trợ như Scikit-learn, Pandas, NumPy và được triển khai trên môi trường Jupyter Notebook hoặc Google Colab.

 **Giới hạn về quy mô và khả năng mở rộng của dữ liệu:** Mặc dù Big Data hướng đến việc xử lý lượng dữ liệu khổng lồ, trong nghiên cứu này dữ liệu được thu thập có thể chỉ đạt mức vừa phải, không đạt đến quy mô “big data” thực sự. Điều này hạn chế khả năng đánh giá và ứng dụng các kỹ thuật phân tích dữ liệu quy mô lớn như xử lý dữ liệu phân tán, tối ưu hóa tính toán song song, v.v.

 **Giới hạn về tính đa dạng và phức tạp của dữ liệu:** Đề tài chủ yếu tập trung vào các chỉ số đo được từ bệnh nhân ở dạng dữ liệu có cấu trúc. Các loại dữ liệu phi cấu trúc (như hình ảnh, văn bản, video) thường có tính đa dạng và phức tạp cao sẽ không được đưa vào phân tích, mặc dù chúng là một phần quan trọng trong lĩnh vực Big Data.

 **Giới hạn về tốc độ và thời gian xử lý:** Các hệ thống Big Data thường yêu cầu khả năng xử lý dữ liệu theo thời gian thực (data streaming) để đưa ra quyết định kịp thời. Tuy nhiên, nghiên cứu này không tập trung vào việc xây dựng hệ thống xử lý dữ liệu thời gian thực, mà chỉ dựa vào các tập dữ liệu lịch sử đã được thu thập.

## Cấu trúc báo cáo

1. **Giới thiệu:** Giới thiệu về đề tài nghiên cứu, mục tiêu, nội dung và giới hạn của nghiên cứu.
2. **Cơ sở lý thuyết:** Trình bày các lý thuyết cơ bản liên quan đến ung thư, tổ chức dữ liệu, công cụ và các phương pháp dự đoán.
3. **Phương pháp:** Mô tả các yêu cầu bài toán, các thuật toán học máy áp dụng, và triển khai hệ thống lên giao diện người dùng.
4. **Kết quả và thảo luận:** Trình bày kết quả thực nghiệm và thảo luận về hiệu quả của mô hình dự đoán.
5. **Kết luận:** Tổng kết kết quả nghiên cứu và đề xuất hướng nghiên cứu tiếp theo.

# CƠ SỞ LÝ THUYẾT

## Tổng quan về bệnh ung thư và các yếu tố nguy cơ

* Định nghĩa và phân loại:
  + Ung thư là hiện tượng các tế bào phát triển bất thường, có khả năng xâm lấn mô lân cận và di căn. Các loại ung thư phổ biến bao gồm ung thư vú, ung thư phổi, ung thư gan, ung thư đại trực tràng,… Tuy nhiên, hệ thống dự đoán của đề tài chủ yếu nhằm đánh giá mức độ nguy cơ mắc ung thư dựa trên các chỉ số sức khỏe và môi trường chung, thay vì chẩn đoán cụ thể từng loại ung thư.
* Các yếu tố nguy cơ:
* Ngoài các yếu tố di truyền, tiền sử gia đình hay môi trường sống, đề tài chỉ tập trung vào các chỉ số định lượng đo được từ bệnh nhân như:
  + Tuổi (Age)
  + Giới tính (Gender)
  + Mức độ ô nhiễm không khí (Air Pollution)
  + Tiêu thụ rượu (Alcohol use)
  + Dị ứng bụi (Dust Allergy)
  + Nguy cơ do môi trường làm việc (OccuPational Hazards)
  + Yếu tố di truyền (Genetic Risk)
  + Tình trạng bệnh phổi mãn tính (chronic Lung Disease)
  + Chế độ ăn uống (Balanced Diet)
  + Béo phì (Obesity)
  + Thói quen hút thuốc (Smoking)
  + Phơi nhiễm khói thuốc (Passive Smoker)
  + Các triệu chứng liên quan như đau ngực, ho ra máu, mệt mỏi, giảm cân không giải thích, khó thở, thở khò khè, khó nuốt, biến đổi móng tay, thường xuyên cảm lạnh, ho khan, ngáy,…

## Big Data và công cụ trong hệ thống dự đoán

* **Ứng dụng Big Data:**
  + Tích hợp dữ liệu từ nhiều nguồn khác nhau nhằm tạo ra một kho dữ liệu phong phú, đa dạng cho quá trình đào tạo mô hình.
* **Công cụ hỗ trợ – RapidMiner:**
  + RapidMiner cung cấp giao diện trực quan cho việc tổ chức lưu trữ và xử lý dữ liệu trước khi cho mô hình dự đoán, hiệu quả cho quá trình tiền xử lý, trực quan hóa và thử nghiệm các mô hình học máy trong giai đoạn nghiên cứu và phát triển.

## Phương pháp dự đoán mức độ mắc bệnh ung thư

* Dự đoán rủi ro:
  + Hệ thống được xây dựng nhằm ước lượng khả năng mắc bệnh ung thư của bệnh nhân dựa trên một tập hợp các chỉ số sức khỏe và môi trường. Kết quả dự đoán thường được thể hiện dưới dạng xác suất hoặc thang điểm rủi ro (ví dụ: Low, Medium, High).
* **Các phương pháp và thuật toán:**
  + **Học máy (Machine Learning):** Áp dụng các thuật toán như hồi quy logistic, cây quyết định, Random Forest,… nhằm phát hiện mối quan hệ giữa các đặc trưng và mức độ nguy cơ mắc bệnh.

# PHƯƠNG PHÁP

## Yêu cầu bài toán:

* Giao diện thân thiện, dễ thao tác
* Dữ liệu bố trí với RapidMiner
* Dự đoán được khả năng mắc bệnh ung thư dựa trên các chỉ số đã cung cấp.
* Thống kê và tìm kiếm được các bệnh nhân dựa trên bộ lọc là từng chỉ số.

## Giải thuật thực hiện

* Tiền xử lý dữ liệu:
  + Làm sạch dữ liệu, xử lý các giá trị thiếu và chuẩn hóa dữ liệu thông qua RapidMiner.
* Xây dựng mô hình dự đoán:
  + Hồi quy Logistic (Logistic Regression): Sử dụng để tính toán xác suất mắc bệnh ung thư dựa trên mối quan hệ tuyến tính giữa các chỉ số.
  + Cây Quyết Định (Decision Trees): Kết hợp nhiều cây quyết định nhằm nâng cao độ chính xác và giảm thiểu hiện tượng overfitting thông qua cơ chế bỏ phiếu đa số.
  + Random Forest: Xây dựng mô hình dưới dạng cây với các nút quyết định dựa trên giá trị của từng chỉ số, giúp xác định các yếu tố có ảnh hưởng lớn.
* Triển khai và tích hợp:
  + Sau khi huấn luyện, mô hình được tích hợp vào hệ thống có giao diện người dùng để nhập liệu, dự đoán và hiển thị kết quả.
  + Hệ thống hỗ trợ chức năng thống kê, báo cáo và tìm kiếm bệnh nhân dựa trên các tiêu chí đã thiết lập.

## Phương pháp đánh giá

* **Accuracy (Độ chính xác):** Đo lường tỷ lệ dự đoán đúng trên tổng số dự đoán, cho thấy mức độ chính xác của mô hình.
* **Recall (Độ nhạy):** Đánh giá khả năng của mô hình trong việc phát hiện các trường hợp mắc bệnh thực sự (dương tính thật).
* **Precision (Độ chính xác - Precision):** Xác định tỷ lệ các dự đoán mắc bệnh chính xác so với tổng số các trường hợp được dự đoán mắc bệnh.
* **F1 Score:** Là thước đo cân bằng giữa Recall và Precision, hữu ích trong trường hợp dữ liệu không đồng đều giữa các lớp.

# HIỆN THỰC KẾT QUẢ

## Yêu cầu hệ thống và tập dữ liệu thực nghiệm (nếu có)

**Phần mềm**:

* **Ngôn ngữ lập trình**: Python (các thư viện như Scikit-learn, TensorFlow, Keras, Pandas, NumPy, Matplotlib).
* **Môi trường phát triển**: Jupyter Notebook hoặc Google Colab.
* **Cơ sở dữ liệu**: CSV, SQL hoặc RapidMiner (tùy theo cách tổ chức dữ liệu).

## Tập dữ liệu

* **Nguồn dữ liệu**: cancer patient data sets.xlsx
* **Kích thước dữ liệu**: Khoảng vài ngàn đến hàng chục ngàn bản ghi.
* **Đặc trưng (features)**:
  + Thông tin bệnh nhân: Tuổi, giới tính
  + Kết quả xét nghiệm: Kích thước khối u, mật độ tế bào, hình thái tế bào.

Kết quả đầu ra: **0** (không mắc ung thư), **1** (mắc ung thư).

## Kết quả thực nghiệm

KẾT LUẬN

## Kết quả đạt được

**-**

## Ưu và nhược điểm

## Hướng mở rộng

# BẢNG PHÂN CÔNG NHIỆM VỤ CHUNG

|  |  |  |
| --- | --- | --- |
| **Tên thành viên** | **Nhiệm vụ** | **Múc độ đóng góp chung** |
|  |  |  |
|  |  |  |
|  |  |  |
|  |  |  |

# TÀI LIỆU THAM KHẢo