TRƯỜNG ĐẠI HỌC GIAO THÔNG VẬN TẢI
PHÂN HIỆU TẠI TP. HỒ CHÍ MINH
BỘ MÔN CÔNG NGHỆ THÔNG TIN

BÁO CÁO ĐỒ ÁN
HỌC PHẦN: KHAI PHÁ DỮ LIỆU

ĐỀ TÀI: XÂY DỰNG ỨNG DỤNG DỰ ĐOÁN, PHÂN LOẠI BỆNH UNG THƯ TUYẾN VÚ DỰA TRÊN BỘ DỮ LIỆU METABRIC

Giảng viên hướng dẫn:	ThS. NGUYỄN THIỆN DƯƠNG
Sinh viên thực hiện:	MSSV:
PHAN ĐỨC AN	6351071001
NGUYỄN THÀNH ĐẠT	6351071016
ĐINH VĂN HUYNH	6351071031
NGUYỄN THỊ TƯỜNG VI	6351071077
HÀ HOÀNG VỸ	6351071082

Lớp: CÔNG NGHỆ THÔNG TIN

Khóa: 63

TP. Hồ Chí Minh, năm 2025

TRƯỜNG ĐẠI HỌC GIAO THÔNG VẬN TẢI
PHÂN HIỆU TẠI TP. HỒ CHÍ MINH
BỘ MÔN CÔNG NGHỆ THÔNG TIN

BÁO CÁO ĐỒ ÁN
HỌC PHẦN: KHAI PHÁ DỮ LIỆU

ĐỀ TÀI: XÂY DỰNG ỨNG DỤNG DỰ ĐOÁN, PHÂN LOẠI BỆNH UNG THƯ TUYẾN VÚ DỰA TRÊN BỘ DỮ LIỆU METABRIC

Giảng viên hướng dẫn: ThS. NGUYỄN THIỆN DƯƠNG

Sinh viên thực hiện: MSSV:

PHAN ĐỨC AN	6351071001
NGUYỄN THÀNH ĐẠT	6351071016
ĐINH VĂN HUYNH	6351071031
NGUYỄN THỊ TƯỜNG VI	6351071077
HÀ HOÀNG VỸ	6351071082

Lớp: CÔNG NGHỆ THÔNG TIN

Khóa: 63

TRƯỜNG ĐẠI HỌC GIAO THÔNG VẬN TẢI
PHÂN HIỆU TẠI THÀNH PHỐ HỒ CHÍ MINH

CỘNG HÒA XÃ HỘI CHỦ NGHIÃ VIỆT NAM
Độc lập – Tự do – Hạnh phúc

NHIỆM VỤ THIẾT KẾ MÔN HỌC

BỘ MÔN: **CÔNG NGHỆ THÔNG TIN**

-------***-------

Mã sinh viên: 6351071001	**Họ tên SV:** Phan Đức An
Mã sinh viên: 6351071016	**Họ tên SV:** Nguyễn Thành Đạt
Mã sinh viên: 6351071031	**Họ tên SV:** Đinh Văn Huynh
Mã sinh viên: 6351071077	**Họ tên SV:** Nguyễn Thị Tường Vi
Mã sinh viên: 6351071082	**Họ tên SV:** Hà Hoàng Vỹ
Khóa: 63	**Lớp:** Công nghệ thông tin

1. **Tên đề tài:** Xây dựng ứng dụng dự đoán, phân loại bệnh ung thư tuyến vú dựa trên bộ dữ liệu METABRIC
2. **Mục đích, yêu cầu:**
   a. **Mục đích**

- **Mục đích ứng dụng:** Ứng dụng hướng đến giải quyết nhu cầu thực tế trong lĩnh vực y sinh – phân tích dữ liệu ung thư vú:
  - Hỗ trợ dự đoán loại ung thư tuyến vú của bệnh nhân dựa trên đặc trưng lâm sàng và dữ liệu phân tử từ bộ METABRIC.
  - Phân loại bệnh nhân theo các nhóm nguy cơ (thấp – trung bình – cao) hoặc theo subtype ung thư vú.
  - Hỗ trợ bác sĩ, nhà nghiên cứu và bệnh nhân trong việc tham khảo quyết định điều trị.
  - Tối ưu hoá quy trình đánh giá nguy cơ bằng công cụ tự động thay vì đánh giá thủ công.

- Tạo nền tảng cho các nghiên cứu nâng cao trong phân tích dữ liệu y sinh bằng AI.

- **Mục đích kỹ thuật – học thuật**
  - Ứng dụng các kỹ thuật trong khoa học dữ liệu:
    - Thu thập và làm sạch dữ liệu
    - Tiền xử lý dữ liệu
    - Phân tích thống kê và phân tích mô hình
    - Xây dựng mô hình dự đoán
    - Đánh giá và tối ưu hiệu năng
    - Triển khai ứng dụng dựa trên mô hình
  - Sử dụng bộ dữ liệu thực tế METABRIC – bộ dữ liệu chuẩn trong nghiên cứu ung thư vú.
  - Xây dựng mô hình học máy có khả năng dự đoán và phân loại hiệu quả.
  - Chuyển mô hình ML thành ứng dụng thực tế (web hoặc desktop).

b. Yêu cầu

- **Yêu cầu chức năng**
  - Nhập dữ liệu bệnh nhân: Người dung nhập các thông tin lâm sàng
    - Tuổi, tình trạng ER/PR/HER2
    - Loại mô học (breast cancer subtype)
    - Chỉ số lâm sàng: tumor size, lymph nodes, menopausal status,…
    - Gen expression hoặc dữ liệu tương ứng
  - Tiền xử lý dữ liệu
    - Chuẩn hóa dữ liệu nhập vào theo định dạng mô hình đã huấn luyện.
    - Mã hóa các biển phân loại.
    - Thực hiện kỹ thuật xử lý dữ liệu thiếu.
  - Dự đoán kết quả
    - Chạy mô hình máy học đã huấn luyện để: Phân loại nguy cơ sống sót
    - Phân loại nhóm nguy cơ
    - Phân loại subtype ung thư

- **Yêu cầu phi chức năng**
  - Giao diện đơn giản, dễ sử dụng

- Hệ thống chạy ổn định, độ trễ thấp
- Bảo mật thông tin bệnh nhân
- Tính mở rộng để bổ sung mô hình mới
- Tính chính xác: đảm bảo mô hình đạt độ chính xác cao sau huấn luyện.

3. Nội dung và phạm vi đề tài

a. Nội dung: Đề tài tập trung xây dựng phát triển ứng dụng dự đoán – phân loại bệnh ung thư vú bằng mô hình học máy dựa trên bộ METABRIC. Các nội dung chính bao gồm:

- Tìm hiểu lý thuyết:
  - Tổng quan về bệnh ung thư vú: nguyên nhân, triệu chứng, yếu tố nguy cơ, quy trình chẩn đoán, điều trị và chăm sóc sau điều trị.
  - Tìm hiểu kiến thức về học máy và khai phá dữ liệu phục vụ cho phát triển ứng dụng.
  - Những kiến thức về bộ dữ liệu METABRIC gồm có 1980 bệnh nhân với hơn 30 đặc trưng lâm sàng và dữ liệu gene
- Xây dựng bộ dữ liệu và tri thức y khoa
  - Thu thập dữ liệu từ các nguồn y khoa chính thống: WHO, Bộ Y tế, các hiệp hội ung thư vú, tài liệu hướng dẫn tầm soát – điều trị.
  - Xây dựng knowledge base (FAQ) về bệnh ung thư vú (triệu chứng, chẩn đoán, tầm soát, điều trị…).
  - Huấn luyện mô hình ML:
  - Đánh giá mô hình bằng Accurary, F1-score,…

b. Phạm vi đề tài

- **Phạm vi kiến thức cung cấp**
  - Đề tài được triển khai dự đoán và phân loại ung thư vú dựa trên :
    - Phân loại subtype ung thư
    - Phân nhóm nguy cơ bệnh nhân
  - Không đưa ra kết luận chẩn đoán y khoa.
  - Không ra quyết định điều trị (chỉ mang tính tham khảo).
  - Dữ liệu thử nghiệm sử dụng bộ dữ liệu y tết METABRIC: là bộ dữ liệu nghiên cứu ung thư vú nổi tiếng, gồm hơn 1980 bệnh nhân. Giúp xây

dựng mô hình dự đoán nguy cơ tử vong, nguy cơ tái phát, phân loại nhóm ung thư chứa:

- Thông tin lâm sàng
- Dữ liệu biểu hiện gen
- Chỉ số sống sót
- Phân nhóm ung thư
- Các yếu tố điều trị

- **Phạm vi kỹ thuật**
  - Hệ thống được xây dựng dưới dạng web chạy cục bộ
  - Xây dựng mô hình AI
    - Random Forest
    - XGBoost
    - Logistic Regression
    - Neutral Networks
    - Survival Analyst Model
- **Phạm vi người dùng**
  - Sinh viên
  - Nhà nghiên cứu
  - Bác sĩ

4. Công nghệ, công cụ và ngôn ngữ lập trình

4.1 Công nghệ sử dụng:

- Machine Learning có giám sát (Supervised Learning): xây dựng mô hình phân loại ung thư vú và dự đoán nguy cơ bệnh dựa trên dữ liệu lâm sàng và phân tử
- Phân tích dữ liệu y sinh (Biomedical Data Analytics): khai phá mối quan hệ giữa các đặc trưng lâm sàng, mô học và phân tử,…
- Khai phá dữ liệu (Data Mining): ứng dụng các bước tiền xử lý dữ liệu, lựa chọn đặc trưng, phân tích thông kê, trực quan hóa,..
- Ứng dụng Web cục bộ (Local Web Application): tích hợp mô hình học máy vào ứng dụng web cục bộ nhằm hỗ trợ người dung nhập dữ liệu và nhận kết quả dự đoán

4.2 Công cụ

- Visual studio
- Jupyter Notebook
- Thư viện xử lý và phân tích dữ liệu: Numpy, Pandas
- Thư viện học mát: Scikit-learn, Imbalanced-learn

4.3 Ngôn ngữ lập trình

- Python
- HTML/ CSS/ JavaScript: được sử dụng để xây dựng web demo, giúp người dung nhập dữ liệu và xem kết quả dự đoán.

5. Các kết quả chính dự kiến sẽ đạt được và ứng dụng

5.1 Các kết quả chính dự kiến sẽ đạt được

- Xây dựng thành công bộ dữ liệu huấn luyện chuẩn hóa
- Xây dựng và so sánh được nhiều mô hình
- Phân tích và rút ra các đặc trưng quan trọng
- Xây dựng ứng dụng web demo

5.2 Ứng dụng

- Hỗ trợ nghiên cứu sinh: về phân nhóm ung thư và tiên lượng bệnh
- Hỗ trợ quyết định lâm sàng: ứng dụng giúp bác sĩ và nhà nghiên cứu có thêm thông tin khi đánh giá nguy cơ bệnh nhân
- Có thể mở rộng tích hợp các mô hình tiên lượng sai sót, mô hình học sâu hoặc triển khai trên môi trường cloud để phục vụ quy mô lớn hơn

6. Giảng viên và cán bộ hướng dẫn

Họ tên: ThS. Nguyễn Thiện Dương

Đơn vị công tác: Đại học Giao thông vận tải Phân hiệu TP.HCM

Điện thoại: Email:

Ngày 13 tháng 12 năm 2025	**Đã giao nhiệm vụ TKTN**
Trưởng BM Công nghệ Thông tin	**Giảng viên hướng dẫn**
ThS. Trần Phong Nhã	**ThS. Nguyễn Thiện Dương**

Đã nhận nhiệm vụ TKTN

Sinh viên: Phan Đức An	Ký tên
Sinh viên: Nguyễn Thành Đạt	Ký tên
Sinh viên: Đinh Văn Huynh	Ký tên
Sinh viên: Nguyễn Thị Tường Vi	Ký tên
Sinh viên: Hà Hoàng Vỹ	Ký tên
Điện thoại:	Email:

PHÂN CÔNG NHIỆM VỤ

MSSV	Họ và tên	Mô tả công việc	Chấm điểm
6351071001	Phan Đức An (nhóm trưởng)	- Tiền xử lý dữ liệu - Tìm hiểu về các mô hình máy học - Huấn luyện mô hình SVM - Nâng cấp trang web - Viết báo cáo	
6351071031	Đinh Văn Huynh	-Tiền xử lý dữ liệu - Phân tích dữ liệu - Tìm hiểu về mô hình SVM - Huấn luyện mô hình Random Forest, SVM - Viết báo cáo	
6351071077	Nguyễn Thị Tường Vi	- Tổng hợp kiến thức về bệnh ung thư tuyến vú - Tìm hiểu về bộ dữ liệu METABRIC - Tìm hiểu về mô hình Decision Tree - Huấn luyện mô hình Decision Tree - Viết báo cáo	
6351071082	Hà Hoàng Vỹ	-Tiền xử lý dữ liệu - Tìm hiểu về mô hình Random Forest - Huấn luyện mô hình Random Forest - Chuẩn bị web demo phần frontent - Viết báo cáo	

6351071016	Nguyễn Thành Đạt	- Tiền xử lý dữ liệu - Tìm hiểu về mô hình Random Forest - Chuẩn bị web demo phần backend - Viết báo cáo	

LỜI CẢM ƠN

Qua thời gian học tập và rèn luyện tại Trường Đại học Giao thông Vận tải phân hiệu tại TP. Hồ Chí Minh, đến nay chúng em đã được trang bị những kĩ năng, kiến thức cơ bản để có thể hoàn thành được bài tập cuối kì do giảng viên giao.

Chúng em cảm ơn tập thể các thầy cô giáo Bộ môn Công Nghệ Thông Tin và các thầy cô thỉnh giảng đã giảng dạy, quan tâm và không ngần ngại dành thời gian để chỉ bài và giải đáp những thắc mắc của chúng em trong những tiết học và cả những lúc ngoài giờ.

Và chúng em cảm ơn thầy Nguyễn Thiện Dương đã luôn quan tâm nhiệt tình hướng dẫn, giúp đỡ chúng em trong quá trình triển khai và thực hiện bài tập cuối kì. Thầy cũng luôn nhắc nhỡ, giúp đỡ mỗi khi chúng em gặp khó khăn, nhờ vậy mà chúng em đã hoàn thành bài tập cuối kì của nhóm mình đúng thời hạn được giao. Nếu không có sự hướng dẫn của thầy thì có lẽ chúng em đã khó có thể thực hiện được bài tập đúng theo mong muốn của mình.

Chúng em đã bỏ ra nhiều thời gian để tìm hiểu và trang bị thêm kiến thức nhằm phục vụ cho việc thực hiện ý tưởng, nhưng chắc chắn rằng chúng em sẽ không thể tránh khỏi những sai sót không đáng có vì kiến thức còn hạn chế. Chúng em hi vọng rằng sẽ nhận được những lời góp ý quý báu của thầy để có thể hoàn thiện ý tưởng của nhóm một cách tốt nhất có thể.

TP. Hồ Chí Minh, ngày 22 tháng 11 năm 2025

Sinh viên thực hiện

Phan Đức An

Nguyễn Thành Đạt

Đinh Văn Huynh

Nguyễn Thị Tường Vi

Hà Hoàng Vỹ

NHẬN XÉT CỦA GIẢNG VIÊN HƯỚNG DẪN

.....

.....

.....

.....

.....

.....

.....

.....

.....

.....

.....

.....

.....

.....

TP. Hồ Chí Minh, ngày …… tháng …… năm ……

Giảng viên hướng dẫn

ThS. Nguyễn Thiện Dương

MỤC LỤC

DANH MỤC HÌNH ẢNH

DANH MỤC BẢNG BIỂU

DANH MỤC CHỮ VIẾT TẮT

Từ viết	Tiếng Anh đầy đủ	Ý nghĩa
AI	Artificial Intelligence	Trí tuệ nhân tạo – công nghệ mô phỏng trí tuệ con người trong máy tính.
CDSS	Clinical Decision Support System tắt	Hệ thống hỗ trợ ra quyết định lâm sàng cho bác sĩ.
KDD	Knowledge Discovery in Databases	Khám phá tri thức trong cơ sở dữ liệu – quy trình gồm nhiều bước trong khai phá dữ liệu.
DM	Data Mining	Khai phá dữ liệu – tìm mẫu và tri thức trong dữ liệu lớn.
ML	Machine Learning	Học máy – thuật toán cho máy học từ dữ liệu.
mRNA	Messenger Ribonucleic Acid	mRNA – dữ liệu biểu hiện gene dùng trong METABRIC.
ER	Estrogen Receptor	Thụ thể Estrogen – xác định ung thư vú ER+ hoặc ER-.
PR	Progesterone Receptor	Thụ thể Progesterone – đánh giá đáp ứng điều trị nội tiết.
HER2	Human Epidermal Growth Factor Receptor 2	Protein thuộc nhóm gen ERBB2 – liên quan tốc độ phát triển tế bào ung thư.
NPI	Nottingham Prognostic Index	Chỉ số tiên lượng Nottingham – đánh giá mức độ nguy cơ ung thư vú.

Từ viết	Tiếng Anh đầy đủ	Ý nghĩa
IDC	Invasive Ductal Carcinoma	Ung thư biểu mô ống xâm nhập – loại phổ biến nhất.
ILC	Invasive Lobular Carcinoma	Ung thư tiểu thùy xâm nhập.
TNBC	Triple-Negative Breast Cancer	Ung thư vú bộ ba âm tính (ER-, PR-, HER2-).
RF	Random Forest	Thuật toán rừng ngẫu nhiên – mô hình ensemble.
XGBoost	Extreme Gradient Boosting	Thuật toán boosting mạnh, dùng nhiều trong dự đoán.
NN	Neural Networks	Mạng nơ-ron nhân tạo dùng trong học sâu.
SMOTENC	Synthetic Minority Over-sampling Technique for Nominal and Continuous	Kỹ thuật cân bằng dữ liệu cho biến hỗn hợp (danh mục + số).
METABRIC	Molecular Taxonomy of Breast Cancer International Consortium	Dự án phân loại phân tử ung thư vú quốc tế – bộ dữ liệu 1.980 bệnh nhân.
BRCA1/BRCA2	Breast Cancer Gene 1 & 2	Các gen liên quan đến nguy cơ ung thư vú di truyền.
XAI	Explainable Artificial Intelligence	Trí tuệ nhân tạo giải thích được – giúp hiểu mô hình.
SHAP	SHapley Additive exPlanations	Phương pháp giải thích mô hình dựa trên lý thuyết trò chơi.
LIME	Local Interpretable Model-agnostic Explanations	Giải thích mô hình cục bộ độc lập thuật toán.
SVM	Support Vector Machine	Thuật toán máy vector hỗ trợ.

Từ viết	Tiếng Anh đầy đủ	Ý nghĩa
LR	Logistic Regression	Hồi quy Logistic – mô hình phân loại nhị phân.
CD	Clinical Data	Dữ liệu lâm sàng. (xuất hiện trong mô tả METABRIC)
OS	Overall Survival	Thời gian sống toàn bộ.
DSS	Disease-Specific Survival	Thời gian sống đặc hiệu theo bệnh.
DFS	Disease-Free Survival	Thời gian sống không bệnh.

CHƯƠNG 1: MỞ ĐẦU

1.1 Giới thiệu đề tài

Đề tài Nghiên cứu và Xây dựng Hệ Thống Dự Đoán Ung Thư Vú được thực hiện trong xã hội ngày nay, việc ứng dụng trí tuệ nhân tạo và công nghệ thông tin vào lĩnh vực y tế ngày càng trở nên quan trọng. Ung thư vú là một trong những loại ung thư phổ biến nhất, và việc phát triển các công cụ sàng lọc và dự đoán sớm có thể đóng vai trò then chốt trong việc cải thiện hiệu quả điều trị.

Hệ thống được phát triển nhằm cung cấp một giao diện trực quan và một nền tảng tính toán vững chắc để xử lý dữ liệu liên quan và đưa ra kết quả dự đoán. Sản phẩm cuối cùng của đề tài là một ứng dụng web, cho phép người dùng nhập các tham số đầu vào và nhận kết quả dự đoán tức thời.

1.2 Lý do chọn đề tài

1.2.1. Thực trạng bệnh ung thư vú trên thế giới

Ung thư vú hiện là loại ung thư phổ biến nhất ở phụ nữ trên toàn cầu. Tỷ lệ mắc mới có xu hướng gia tăng qua từng năm và xuất hiện ở độ tuổi ngày càng trẻ. Mặc dù y học phát triển, nhưng tại nhiều quốc gia, đặc biệt là các nước đang phát triển, việc tầm soát và chẩn đoán vẫn chưa được chú trọng đúng mức. Ngoài ra, các phương pháp chẩn đoán truyền thống như nhũ ảnh hoặc xét nghiệm mô học thường đòi hỏi thiết bị hiện đại và đội ngũ bác sĩ chuyên môn cao. Điều này khiến cho việc phát hiện sớm trở nên khó khăn, dẫn đến nhiều trường hợp phát hiện bệnh ở giai đoạn muộn, làm giảm khả năng điều trị thành công.

1.2.2. Thực trạng bệnh ung thư vú tại Việt Nam

Tại Việt Nam, ung thư vú là một trong những bệnh ung thư phổ biến nhất ở nữ giới và có xu hướng gia tăng mạnh. Phần lớn bệnh nhân chỉ phát hiện bệnh khi đã ở giai đoạn trung bình hoặc giai đoạn muộn, gây khó khăn trong điều trị và làm tăng tỷ lệ tử vong. Nguyên nhân đến từ tâm lý chủ quan, thiếu kiến thức tầm soát sớm và hạn chế về điều kiện y tế ở nhiều khu vực. Trong khi đó, chi phí xét nghiệm và chẩn đoán cũng là rào cản lớn đối với nhiều bệnh nhân. Điều này cho thấy nhu cầu cấp thiết về các công cụ hỗ trợ chẩn đoán nhanh chóng, có chi phí thấp và dễ tiếp cận.

1.2.3. Sự cần thiết của ứng dụng dự đoán và phân loại ung thư vú

- Ứng dụng giúp hỗ trợ bác sĩ trong quá trình chẩn đoán, tăng độ chính xác và giảm sai sót.

- Giúp bệnh nhân tiếp cận công nghệ đánh giá sức khỏe nhanh chóng hơn với chi phí thấp
- Tăng khả năng phát hiện sớm và phân loại chính xác loại ung thư dựa trên dữ liệu.

1.2.4. Lý do cá nhân chọn đề tài

Nhóm em chọn đề tài này vì mong muốn áp dụng kiến thức về trí tuệ nhân tạo, xử lý dữ liệu và lập trình vào một lĩnh vực có giá trị thực tiễn cao – lĩnh vực y tế. Ngoài ra, dữ liệu METABRIC là một bộ dữ liệu lớn, đa dạng và mang tính học thuật cao, tạo điều kiện để nhóm em cùng nhau rèn luyện kỹ năng phân tích dữ liệu, xây dựng mô hình học máy và phát triển ứng dụng thực tế.

1.3 Mục tiêu đề tài

- Xây dựng Kiến trúc Mô hình dự đoán hiệu quả: Nghiên cứu, lựa chọn và tối ưu hóa các thuật toán nhằm tạo ra một mô hình có khả năng phân loại hoặc dự đoán nguy cơ mắc bệnh ung thư vú với độ chính xác cao.
- Phát triển ứng dụng web: Thiết kế và triển khai một giao diện người dùng dựa trên web. Giao diện này phải cho phép người dùng (đối tượng là những chuyên gia y tế hoặc người nghiên cứu) dễ dàng nhập các dữ liệu đầu vào cần thiết cho mô hình.
- Tích hợp và kiểm thử hệ thống: Tích hợp thành công mô hình dự đoán vào nền tảng ứng dụng web và đảm bảo sự vận hành thông suốt. Mục tiêu này được chứng minh bằng việc hiển thị hộp thông báo, xác nhận rằng một thao tác (như gửi dữ liệu hoặc chạy dự đoán) đã được hệ thống xử lý thành công và cung cấp phản hồi tích cực cho người dùng.

1.4 Mục đích nghiên cứu

Mục đích của đề tài là xây dựng hệ thống mô hình học máy có khả năng phân tích các chỉ số lâm sàng, mô học và phân tử của bệnh nhân, giúp hỗ trợ các bác sĩ và nhà nghiên cứu trong việc nhận diện loại ung thư vú, dự đoán mức độ nguy hiểm và ước tính tiên lượng điều trị dựa trên dữ liệu

1.5 Đối tượng nghiên cứ

Đối tượng nghiên cứu của đề tài gồm dữ liệu bệnh nhân ung thư tuyến vú và các yếu tố lâm sàng – mô học – phân tử có ảnh hưởng đến dự đoán, phân loại và tiên lượng bệnh. Cụ thể như sau:

- Bệnh nhân ung thư vú trong bộ dữ liệu METABRIC và dataset lâm sàng: giúp đưa ra các nền tảng để mô hình học từ các đặc điểm bệnh và đưa ra dự đoán
    - Dữ liệu thu thập sau quá trình khám, sinh thiết, xét nghiệm và điều trị
    - Đối tượng đại diện cho nhiều nhóm tuổi, giai đoạn bệnh và kiểu ung thư khác nhau.
- Các đặc điểm lâm sàng: phản ánh tình trạng bệnh tại thời điểm chẩn đoán
    - Kích thước khối u
    - Số lượng hạch bạch huyết dương tính
    - Tình trạng ER, PR, HER2
    - Mức độ ác tính
    - Chỉ số tiên lượng NPI
- Các đặc điểm mô học: giúp mô hình phân biệt được bản chất ác tính, độ xâm lấn và mức độ tiến triển của khối u
    - Loại mô bệnh học
    - Mức độ tế bào ung thư
    - Giai đoạn mô học
- Các đặc điểm phân tử: giúp phân loại ung thư vú theo hướng cá thể hóa, bao gồm các chỉ dấu sinh học và phan loại gene
- Các thông tin điều trị:
    - Loại phẫu thuật
    - Hóa trị
    - Xạ trị
    - Liệu pháp hormone

1.6 Phạm vi nghiên cứu

- Phạm vi chức năng: Đề tài tập trung xây dựng chức năng dự đoán dựa trên dữ liệu đầu vào có cấu trúc và hiển thị kết quả xử lý một cách trực quan. Các chức năng nâng cao như quản lý hồ sơ bệnh nhân, kết nối API y tế, hay các phân tích chuyên sâu như trực quan hóa dữ liệu dạng đồ thị không nằm trong phạm vi nghiên cứu lần này.
- Phạm vi công nghệ: Hệ thống được phát triển dựa trên các framework lập trình web và các thư viện phân tích dữ liệu tương thích với Python. Việc thiết lập, cài

đặt và chạy các môi trường phần mềm cần thiết được xem là một phần quan trọng trong phạm vi nghiên cứu

- Phạm vi dữ liệu: Nghiên cứu sử dụng các tập dữ liệu liên quan đến mô hình dự đoán ung thư vú. Quá trình nhập liệu, xử lý và chuẩn hóa các tham số dữ liệu nhằm phục vụ cho tương tác người dùng là trọng tâm của đề tài.
- Phạm vi triển khai: Hệ thống được phát triển và trình diễn trong Local Deployment việc triển khai trên môi trường này giúp tiết kiệm chi phí và thời gian, phù hợp với mục tiêu đề tài. Ngoài ra, hệ thống còn đảm bảo tính khả thi và kiểm soát môi trường.

1.7 Phương pháp nghiên cứu

Để thực hiện đề tài này, nhóm em đã áp dụng kết hợp nhiều phương pháp khoa học và kỹ thuật, đảm bảo từ việc khảo sát lý thuyết đến phát triển và kiểm chứng hệ thống:

- Nghiên cứu tài liệu
    - Thu thập kiến thức nền tảng bệnh ung thư tuyến vú (qua các tài liệu chuyên ngành, workshop,…)
    - Tổng hợp các kỹ thuật xử lý dữ liệu, chọn biến quan trọng và đánh giá mô hình học máy
    - Tham khảo các công trình nghiên cứu sử dụng bộ dữ liệu METABRIC để hiểu cấu trúc dữ liệu, các biến mô học, phân tử và thông tin lâm sàng
- Phân tích dữ liệu: Khai thác thông tin từ bộ dữ liệu METABRIC nhằm nhận diện các mẫu và mối quan hệ giữa các đặc trưng với loại ung thư. Cách thực hiện như sau:
    - Tiền xử lý dữ liệu:
        - Loại bỏ giá trị thiếu, chuẩn hóa dữ liệu, mã hóa biến phân loại
        - Chọn các biến quan trọng dựa trên kiến thức y học và phân tích thống kê
    - Khám phá dữ liệu:
        - Sử dụng thống kê mô tả, biểu đồ phân phối, ma trận tương quan để hiểu dữ liệu.
        - Nhận diện mối quan hệ giữa các biến lâm sàng, mô học và phân tử với phân loại ung thư.

- Phương pháp học máy: Xây dựng mô hình dự đoán phân loại ung thư tuyến vú dựa trên các đặc trưng trong bộ dữ liệu METABRIC. Các bước thực hiện như sau:
    - Chọn thuật toán: Các thuật toán phổ biến như Logistic Regression, Decision Tree, Radom Forest, SVM
    - Huấn luyện và kiểm thử mô hình
        - Chia dữ liệu thành tập huấn luyện và tập kiểm thử
        - Huấn luyện mô hình trên tập huấn luyện, tối ưu tham số
    - Đánh giá mô hình
        - Sử dụng các chỉ số như Accuracy, Precision, Recall, F1-score, AUC-ROC để đánh giá hiệu quả phân loại.
        - So sánh hiệu suất giữa các thuật toán để chọn ra mô hình tốt nhất.
- Phân tích và trực quan hóa kết quả: trình bày kết quả một cách trực quan, dễ hiểu và phục vụ phân tích chuyên môn. Thực hiện như sau:
    - Trực quan hóa dữ liệu ban đầu và các kết quả dự đoán bằng biểu đồ, heatmap, confusion matrix.
    - Giải thích ý nghĩa y học của các đặc trưng quan trọng ảnh hưởng đến phân loại ung thư.
    - Rút ra kết luận về hiệu quả mô hình và khả năng ứng dụng trong thực tiễn.

1.8 Ý nghĩa thực tiễn của đề tài

- Hỗ trợ chẩn đoán sớm và chính xác hơn
    - Việc phân loại ung thư tuyến vú chính xác giúp bác sĩ lựa chọn phương pháp điều trị phù hợp cho từng bệnh nhân.
    - Mô hình dự đoán dựa trên dữ liệu lâm sàng, mô học và phân tử giúp phát hiện các loại ung thư có nguy cơ cao, từ đó nâng cao hiệu quả điều trị và giảm thiểu biến chứng.
- Tối ưu hóa quy trình ra quyết định y học
    - Các thuật toán học máy có thể xử lý lượng dữ liệu lớn và phức tạp mà con người khó xử lý nhanh chóng.
    - Kết quả dự đoán hỗ trợ bác sĩ trong việc ra quyết định lâm sàng, rút ngắn thời gian đánh giá bệnh và giảm thiểu sai sót.

- Ứng dụng trong nghiên cứu và phát triển y học cá thể hóa (Personalized Medicine)
    - Dựa trên đặc trưng phân tử và mô học, mô hình có thể phân loại bệnh theo từng loại tế bào ung thư, từ đó đề xuất phác đồ điều trị cá thể hóa cho bệnh nhân.
    - Hỗ trợ các nghiên cứu tiếp theo trong việc tìm kiếm các biomarker quan trọng liên quan đến ung thư vú.
- Tiết kiệm chi phí và nguồn lực y tế
    - Việc dự đoán sớm loại ung thư giúp giảm nhu cầu xét nghiệm lặp lại không cần thiết.
    - Giảm chi phí điều trị dài hạn nhờ lựa chọn phương pháp điều trị hiệu quả ngay từ đầu.
- Cơ sở để phát triển các ứng dụng công nghệ hỗ trợ y học
    - Kết quả nghiên cứu có thể được tích hợp vào các phần mềm, hệ thống hỗ trợ chẩn đoán y học (Clinical Decision Support System – CDSS).
    - Mở ra hướng phát triển các mô hình dự đoán ung thư khác dựa trên dữ liệu lớn và trí tuệ nhân tạo.

CHƯƠNG 2: CƠ SỞ LÝ THUYẾT VÀ THỰC NGHIỆM

2.1. Tổng quan về ung thư vú

2.1.1 Khái niệm

Ung thư vú là sự tăng sinh bất thường của các tế bào trong mô vú, có khả năng xâm lấn và di căn.

- Thống kê: Ung thư vú là bệnh lý phổ biến nhất ở phụ nữ toàn cầu, chiếm tỷ lệ cao trong số các ca tử vong do ung thư. Tại Việt Nam, số ca mắc ung thư vú có xu hướng tăng hàng năm.
- Các yếu tố nguy cơ:
    - Tuổi, di truyền (yếu tố BRCA1, BRCA2), tiền sử gia đình.
    - Yếu tố hormone, kinh nguyệt sớm hoặc mãn kinh muộn.
    - Lối sống: thừa cân, ít vận động, chế độ ăn thiếu dinh dưỡng.
- Triệu chứng phổ biến: xuất hiện khối u bất thường, thay đổi hình dạng vú, da vú lõm hoặc sần, dịch tiết núm vú bất thường. Việc tầm soát sớm thông qua siêu âm, chụp nhũ ảnh và xét nghiệm gene giúp tăng khả năng điều trị thành công.

2.1.2 Phân loại ung thư vú: 5 loại ung thư vú chi tiết

- Ung thư ống tuyến vú (Ductal carcinoma): loại phổ biến nhất (~80%).
- Ung thư tiểu thùy (Lobular carcinoma).
- Ung thư vú thể viêm.
- Ung thư vú di căn.
- Ung thư vú theo phân nhóm phân tử (Molecular Subtypes):
    - Luminal A
    - Luminal B
    - HER2-enriched
    - Basal-like / Triple Negative

2.1.3 Các giai đoạn: có 5 giai đoạn

- Giai đoạn 0 : Ung thư biểu mô tại chỗ
    - Tế bào ung thư mới hình thành và còn nằm trong ống dẫn sữa hoặc tiểu thùy.

        - Dấu hiệu: Triệu chứng không thể hiện rõ, thường phát hiện do đi khám định kì.
        - Ở giai đoạn này tế bào ung thư chưa xâm lấn ra mô vú xung quanh nên có khả năng điều trị thành công rất cao.
    - Giai đoạn 1: Xâm lấn
        - Ung thư bắt đầu xâm lấn nhưng còn nhỏ: kích thước <3 chưa lan đến hạch bạch huyết hoặc chỉ lan đến mức rất nhỏ.
        - Dấu hiệu: có thể hơi đau nhẹ hoặc hoàn toàn không đau. Lúc này sẽ xuất hiện khối u nhỏ, cứng, không cảm nhận được nếu tự kiểm tra tại nhà.
        - Điều trị tốt nếu phát hiện sớm.
    - Giai đoạn 2: Tiến triển
        - Khối u tăng kích thước hơn.
        - Dấu hiệu: khối u dễ sờ hơn, đôi khi gây đau, thay đổi hình dạng vú.
        - Có thể lan đến một số hạch bạch huyết nách nhưng chưa lan xa.
        - Bệnh nhân cần kết hợp điều trị phẫu thuật và liệu pháp hỗ trợ (hóa trị, xạ trị, nội tiết).
    - Giai đoạn 3: Giai đoạn tiến triển tại chỗ
        - Khối u lớn hoặc xâm lấn da, thành ngực.
        - Dấu hiệu: khối u lớn, cứng, đau dễ nhận thấy, da vú biến đổi rõ rệt.
        - Hạch bạch huyết vùng nách hoặc gần xương đòn bị ảnh hưởng nhiều.
        - Bệnh chưa di căn xa nhưng đã phát triển phức tạp và điều trị khó khăn hơn.
    - Giai đoạn 4: Ung thư vú di căn
        - Tế bào ung thư lan đến các cơ quan xa như phổi, gan, xương hoặc não.
        - Dấu hiệu: khối u lớn, loét hoặc vỡ, đau dữ dội ở vùng ngực, da sưng đỏ hoặc phù nề.

- Đây là giai đoạn nặng nhất, mục tiêu điều trị chủ yếu nhằm kéo dài thời gian sống và cải thiện chất lượng cuộc sống.

2.2 Dữ liệu METABRIC

- Giới thiệu về tập dữ liệu: METABRIC là bộ dữ liệu ung thư vú quy mô lớn, chuẩn nghiên cứu được sử dụng rộng rãi trong học máy, thống kê y sinh và phân tích sống sót.
- Nội dung
  - Cơ sở dữ liệu phân loại Phân tử Ung thư Vú Quốc tế (METABRIC) là một dự án Canada – Anh, chứa dữ liệu giải quyết quá trình tự mục tiêu của 1.980 mẫu ung thư vú nguyên phát. Dữ liệu lâm sàng và bộ gen được tải xuống từ cBioPortal.
  - Các biến trong nhóm này tập trung phản ánh tình trạng mô học của khối u, quy trình điều trị, và phân nhóm phân tử của bệnh nhân ung thư vú. Đây là những yếu tố mang tính chuyên sâu, thường được sử dụng trong các nghiên cứu y sinh và mô hình tiên lượng hiện đại:

Nhóm biến	**Biến**	**Ý nghĩa**	**Giá trị ứng dụng**
Điều trị	type_of_breast_surgery	Loại phẫu thuật (bảo tồn hoặc cắt bỏ vú)	Đánh giá mức độ điều trị và nguy cơ tái phát
	chemotherapy	Có hóa trị hay không	Phản ánh mức độ nặng và chiến lược điều trị
	hormone_therapy	Điều trị nội tiết	Gắn liền với ER/PR+
	radio_therapy	Xạ trị	Tiêu diệt tế bào ung thư còn sót
Mô học	cancer_type	Loại ung thư	Xác nhận loại ung thư

	cancer_type_detailed	Phân loại mô học chi tiết	Liên quan tiên lượng và kiểu xâm lấn
	cellularity	Mật độ tế bào ung thư	Tỷ lệ cao → ác tính cao
	neoplasm_histologic_grade	Độ mô học	Grade 3 → tiên lượng xấu
	lymph_nodes_examined_positive	Hạch dương tính	Chỉ số lan rộng quan trọng
	nottingham_prognostic_index	Chỉ số tiên lượng NPI	Dự đoán nguy cơ tử vong
	pr_status	PR+ hay PR-	Liên quan liệu pháp nội tiết
Phân tử	her2_status	HER2+ hay HER2-	Quyết định điều trị Herceptin
	pam50_+_claudin-low_subtype	Phân nhóm phân tử	Phân nhóm sinh học → tiên lượng & điều trị

Bảng 2.1: Sơ lược bộ dữ liệu

- Các chỉ số lâm sàng trong bộ dữ liệu METAB
    - Kích thước khối u: được đo bằng mm, phản ánh mức độ phát triển của khối u. Khối u càng lớn thì khả năng tiến triển nặng và tái phát càng cao
    - Hạch bạch huyết dương tính: thể hiện số lượng hạch bị tế bào ung thư xâm lấn. Đây là yếu tố tiên lượng quan trọng, thường dùng để phân tầng nguy cơ.
    - Tình trạng ER:
        - ER+ nghĩa là tế bào ung thư có thụ thể estrogen, dễ đáp ứng với điều trị nội tiết.
        - ER- thường khó điều trị hơn và tiên lượng xấu hơn.
    - Tình trạng PR: giúp đánh giá khả năng đáp ứng điều trị nội tiết

- HER2 status: là một gen liên quan đến tốc độ phát triển của tế bào
- Thời gian sống: đo thời gian bệnh nhân sống sau chẩn đoán, dung làm chỉ số đánh giá mô hình dự đoán sống còn
- Sự kiện sống sót (Survival Event)
  - 1: bệnh nhân tử vong
  - 0: bệnh nhân còn sống

- Mục tiêu của METABRIC
  - Xác định các phân nhóm phân tử của ung thư vú.
  - Phân tích ảnh hưởng của dữ liệu gene (m -RNA) lên tiên lượng bệnh.
  - Xây dựng mô hình dự đoán thời gian sống sót của bệnh nhân.
  - Hỗ trợ nghiên cứu các yếu tố liên quan đến tái phát, di căn và tiên lượng

2.3. Tổng quan về Data Mining và Machine Learning

2.3.1. Data mining

2.3.1.1. Khái niệm

Khai phá dữ liệu (Data Mining - DM) là quá trình tìm kiếm các mẫu, mối quan hệ và tri thức có ý nghĩa trong các bộ dữ liệu lớn. Đây là lĩnh vực liên ngành giữa máy học, thống kê và cơ sở dữ liệu. Mục tiêu của khai thác dữ liệu là chuyển dữ liệu thô thành thông tin dễ hiểu để phân tích hoặc ra quyết định.

Các bước trong khai phá dữ liệu thường bao gồm:

- Tiền xử lý dữ liệu (Data Preprocessing)
- Khám phá dữ liệu (Data Exploration)
- Chọn mô hình và thuật toán (Model Selection)
- Đánh giá kết quả và trực quan hóa (Evaluation & Visualization)

Khai phá dữ liệu là bước phân tích trong Khám phá tri thức từ cơ sở dữ liệu (KDD – Knowledge Discovery in Databases).

2.3.1.2 Các kỹ thuật Data Mining phổ biến

- Phân loại (Classification): Gán nhãn cho dữ liệu theo các nhóm đã biết (ví dụ: phân loại email spam/không spam).
- Hồi quy (Regression): Dự đoán giá trị liên tục dựa trên dữ liệu (ví dụ: dự báo giá cổ phiếu).
- Phân cụm (Clustering): Nhóm dữ liệu thành các cụm có tính chất tương tự (ví dụ: phân nhóm khách hàng theo hành vi tiêu dùng).

- Luật kết hợp (Association Rule Mining): Tìm mối quan hệ giữa các biến trong dữ liệu (ví dụ: giỏ hàng khách hàng Amazon thường mua cùng lúc sản phẩm A và B).
- Phát hiện dị thường (Anomaly Detection): Nhận diện dữ liệu bất thường hoặc lỗi (ví dụ: phát hiện giao dịch gian lận).

2.3.1.3 Ứng dụng của data mining

- Trong y tế: dự đoán bệnh, phát hiện bệnh hiếm.
- Trong kinh doanh: phân tích hành vi khách hàng, dự đoán doanh số.
- Trong ngân hàng: phát hiện gian lận, đánh giá rủi ro tín dụng.
- Trong khoa học dữ liệu: phân tích dữ liệu lớn, mô hình hóa tri thức.

2.3.2. Machine Learning

2.3.2.1. Khái niêm:

Học máy là một nhánh của trí tuệ nhân tạo nghiên cứu cách xây dựng các thuật toán và mô hình cho phép máy tính học từ dữ liệu để đưa ra dự đoán hoặc quyết định mà không cần lập trình chi tiết. Thuật toán học máy sử dụng dữ liệu huấn luyện để xây dựng mô hình, sau đó áp dụng mô hình này để dự đoán trên dữ liệu mới. Học máy có liên quan mật thiết đến thống kê, nhưng tập trung vào việc phát triển các thuật toán có khả năng xử lý dữ liệu lớn và phức tạp. Một nhánh quan trọng của học máy là học sâu (Deep Learning), phát triển mạnh mẽ trong những năm gần đây và đạt hiệu suất vượt trội trong nhiều bài toán.

2.3.2.2 Phân loại Machine Learning

- Học có giám sát (Supervised Learning): Dữ liệu đầu vào có nhãn, máy học dựa trên dữ liệu này để dự đoán nhãn của dữ liệu mới.
    - Ví dụ: phân loại email, dự đoán giá nhà, chẩn đoán bệnh.
    - Thuật toán phổ biến: Decision Tree, Random Forest, Logistic Regression, Support Vector Machine, Neural Network.
- Học không giám sát (Unsupervised Learning): Dữ liệu không có nhãn, máy học tìm cấu trúc, mối quan hệ hoặc nhóm dữ liệu.
    - Ví dụ: phân cụm khách hàng, giảm chiều dữ liệu.
    - Thuật toán phổ biến: K-Means, Hierarchical Clustering, PCA (Principal Component Analysis).

- Học bán giám sát (Semi-supervised Learning): Dữ liệu bao gồm một phần có nhãn và phần lớn không có nhãn, kết hợp học có giám sát và không giám sát.
- Học tăng cường (Reinforcement Learning): Máy học thông qua tương tác với môi trường, nhận phản hồi theo phần thưởng hoặc hình phạt.
    - Ví dụ: AI chơi game, robot học di chuyển.

2.3.2.3. Các bước trong Machine learning

- Thu thập dữ liệu (Data Collection): Lấy dữ liệu phù hợp với bài toán.
- Tiền xử lý dữ liệu (Data Preprocessing): Làm sạch, chuẩn hóa, xử lý dữ liệu thiếu hoặc lỗi.
- Chia dữ liệu (Data Splitting): Chia dữ liệu thành tập huấn luyện và tập kiểm thử.
- Lựa chọn mô hình (Model Selection): Chọn thuật toán phù hợp với bài toán.
- Huấn luyện mô hình (Model Training): Học từ dữ liệu huấn luyện.
- Đánh giá mô hình (Model Evaluation): Sử dụng tập kiểm thử và các chỉ số như Accuracy, Precision, Recall, F1-Score.
- Triển khai mô hình (Model Deployment): Áp dụng mô hình vào dữ liệu thực tế.

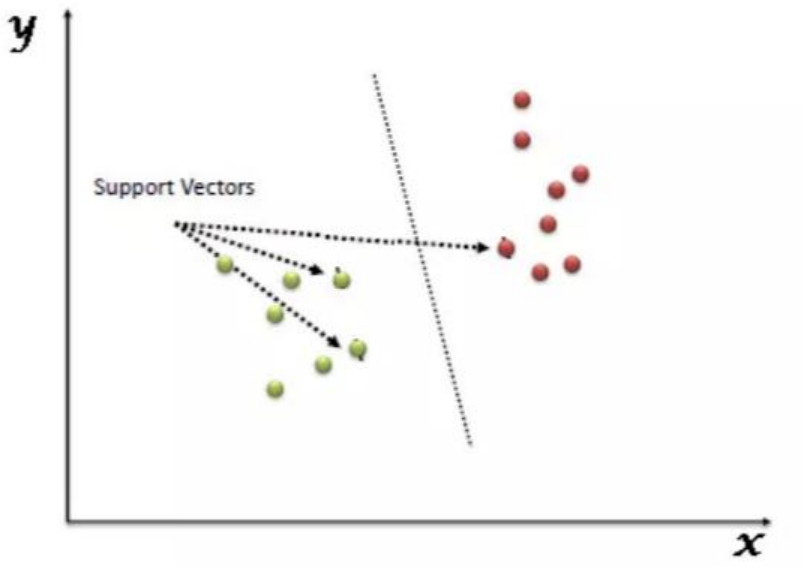
2.3.2.4. Ứng dụng của Machine Learning

- Y tế: chẩn đoán bệnh, dự đoán nguy cơ bệnh tật.
- Tài chính: dự báo thị trường, phát hiện gian lận.
- Thương mại: gợi ý sản phẩm, cá nhân hóa trải nghiệm người dùng.
- Giao thông: dự đoán lưu lượng, tối ưu lộ trình.
- AI và Robotics: nhận dạng hình ảnh, nhận dạng giọng nói, tự động hóa.

2.2.3. Mô hình học máy hiện đại và ứng dụng

2.2.3.1 SVM

SVM là một thuật toán giám sát, nó có thể sử dụng cho cả việc phân loại hoặc đệ quy. Tuy nhiên nó được sử dụng chủ yếu cho việc phân loại. Trong thuật toán này, chúng ta vẽ đồi thị dữ liệu là các điểm trong n chiều (ở đây n là số lượng các tính năng bạn có) với giá trị của mỗi tính năng sẽ là một phần liên kết. Sau đó chúng ta thực hiện tìm "đường bay" (*hyper-plane*) phân chia các lớp. Hyper-plane nó chỉ hiểu đơn giản là 1 đường thẳng có thể phân chia các lớp ra thành hai phần riêng biệt.



Hình 2.1: Đường hyper-plane

Support Vectors hiểu một cách đơn giản là các đối tượng trên đồ thị tọa độ quan sát, Support Vector Machine là một biên giới để chia hai lớp tốt nhất

Nguyên lý hoạt động của SVM

Xét một bài toán phân loại nhị phân với các điểm dữ liệu thuộc hai lớp. Mục tiêu của SVM là tìm ra một siêu phẳng trong không gian N-chiều (với N là số lượng đặc trưng) để phân chia hai lớp này. Siêu phẳng này được định nghĩa bởi phương trình:

$w \cdot x - b = 0$

Trong đó:

- w là một véc-tơ pháp tuyến, quyết định phương của siêu phẳng.
- x là véc-tơ đặc trưng của một điểm dữ liệu.
- b là một hệ số tự do (bias), xác định vị trí của siêu phẳng.

Một điểm dữ liệu mới x' sẽ được phân loại dựa trên vị trí của nó so với siêu phẳng:

- Nếu $w \cdot x' - b > 0$, điểm đó thuộc lớp 1.
- Nếu $w \cdot x' - b < 0$, điểm đó thuộc lớp 2.

Trong vô số các siêu phẳng có thể phân chia dữ liệu, SVM tìm kiếm siêu phẳng tối ưu bằng cách tối đa hóa lề (margin). Lề được định nghĩa là khoảng cách từ siêu phẳng đến các điểm dữ liệu gần nhất của mỗi lớp. Các điểm dữ liệu này được gọi là véc-tơ hỗ trợ (support vectors).

Việc tối đa hóa lề tương đương với việc tìm một "vùng đệm" rộng nhất giữa hai lớp. Một lề lớn hơn ngụ ý rằng mô hình có độ tin cậy cao hơn trong việc phân loại và có khả năng tổng quát hóa tốt hơn, giảm thiểu nguy cơ quá khớp (overfitting).

Dựa trên khả năng phân tách của dữ liệu, có hai loại SVM:

- SVM lề cứng (Hard Margin SVM): Áp dụng khi dữ liệu có thể được phân tách một cách hoàn hảo bằng một siêu phẳng (linearly separable). Mô hình không cho phép bất kỳ điểm dữ liệu nào bị phân loại sai.
- SVM lề mềm (Soft Margin SVM): Là phiên bản tổng quát và thực tế hơn, được sử dụng khi dữ liệu không thể phân tách tuyến tính một cách hoàn hảo hoặc có chứa nhiễu. Mô hình cho phép một số điểm dữ liệu nằm sai phía so với siêu phẳng hoặc nằm trong vùng lề, nhưng sẽ "phạt" những lỗi này thông qua một tham số điều chuẩn C. Tham số C kiểm soát sự đánh đổi giữa việc tối đa hóa lề và giảm thiểu lỗi phân loại.

Véc-tơ hỗ trợ là những điểm dữ liệu nằm trên các đường biên của lề. Chúng là những điểm "khó" phân loại nhất và đóng vai trò quyết định trong việc xác định siêu phẳng tối ưu. Một đặc tính quan trọng của SVM là mô hình cuối cùng chỉ phụ thuộc vào các véc-tơ hỗ trợ này, thay vì toàn bộ tập dữ liệu. Điều này giúp SVM trở nên hiệu quả về mặt bộ nhớ, đặc biệt là với các tập dữ liệu lớn.

Xử lý dữ liệu phi tuyến tính với Kỹ thuật Kernel

Trên thực tế, nhiều bộ dữ liệu không thể được phân tách bằng một siêu phẳng tuyến tính trong không gian đặc trưng ban đầu. Để giải quyết vấn đề này, SVM sử dụng một kỹ thuật mạnh mẽ gọi là "Kernel Trick".

Ý tưởng chính là ánh xạ dữ liệu từ không gian ban đầu có số chiều thấp sang một không gian mới có số chiều cao hơn, nơi dữ liệu có khả năng trở nên phân tách tuyến

tính. Kỹ thuật kernel cho phép thực hiện phép tính tích vô hướng của các véc-tơ trong không gian mới này mà không cần thực hiện phép biến đổi một cách tường minh. Điều này giúp tránh được "lời nguyền số chiều" (curse of dimensionality) và tiết kiệm chi phí tính toán đáng kể.

Một số hàm kernel phổ biến bao gồm:

- Kernel Tuyến tính (Linear Kernel): K(xi, xj) = xi · xj. Tương đương với trường hợp không sử dụng kernel, áp dụng cho dữ liệu đã phân tách tuyến tính.
- Kernel Đa thức (Polynomial Kernel): K(xi, xj) = (γ * xi · xj + r)^d. Hữu ích cho các ranh giới quyết định có dạng đa thức.
- Kernel Hàm cơ sở Bán kính (Radial Basis Function - RBF): K(xi, xj) = exp(-γ * ||xi - xj||^2). Đây là kernel được sử dụng phổ biến nhất do tính linh hoạt và khả năng xử lý các mối quan hệ phi tuyến phức tạp.
- Kernel Sigmoid: K(xi, xj) = tanh(γ * xi · xj + r).

Việc lựa chọn hàm kernel và các siêu tham số của nó (ví dụ: γ, d, r) là một bước quan trọng, ảnh hưởng trực tiếp đến hiệu suất của mô hình SVM.

Ưu điểm và Hạn chế của SVM

Ưu điểm:

- Hiệu quả trong không gian đặc trưng nhiều chiều: SVM hoạt động tốt ngay cả khi số chiều lớn hơn số lượng mẫu.
- Hiệu quả về bộ nhớ: Mô hình chỉ sử dụng một tập con các điểm huấn luyện (các véc-tơ hỗ trợ) để xây dựng ranh giới quyết định.
- Tính linh hoạt cao: Việc sử dụng các hàm kernel khác nhau cho phép mô hình hóa các loại ranh giới quyết định phức tạp.
- Khả năng tổng quát hóa tốt: Nguyên lý tối đa hóa lề giúp SVM giảm thiểu nguy cơ quá khớp.

Hạn chế:

- Chi phí tính toán cao: Thời gian huấn luyện có thể trở nên rất dài trên các tập dữ liệu có kích thước lớn (độ phức tạp tính toán phụ thuộc vào số lượng mẫu).
- Nhạy cảm với việc lựa chọn kernel và siêu tham số: Hiệu suất của mô hình phụ thuộc nhiều vào việc lựa chọn đúng hàm kernel và các siêu tham số như C và gamma (γ).
- Mô hình "hộp đen": Kết quả của SVM, đặc biệt với các kernel phi tuyến, thường khó diễn giải trực tiếp so với các mô hình như cây quyết định.

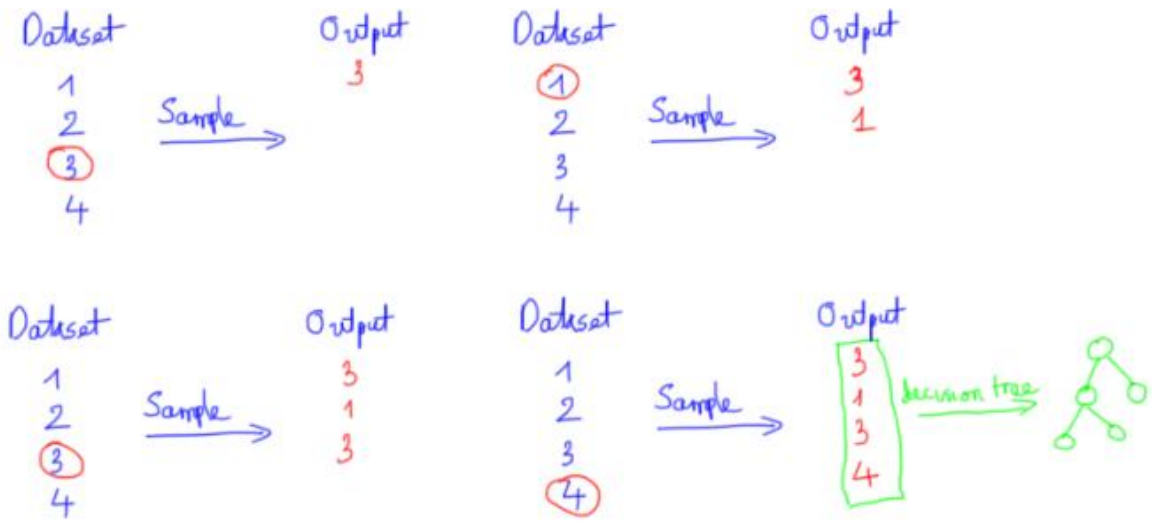
2.2.3.2 Random Forest

Random Forest là một thuật toán học máy có giám sát (supervised learning), dùng cho cả bài toán phân loại (classification) và hồi quy (regression). Thuật toán này xây dựng một "rừng" gồm nhiều cây quyết định và kết hợp dự đoán từ tất cả các cây để đưa ra kết quả cuối cùng.

Nguyên lý hoạt động

Random Forest tạo ra mỗi cây quyết định theo hai bước ngẫu nhiên:

1. **Lấy mẫu dữ liệu (Bootstrap Sampling):**
    - Từ tập dữ liệu gốc, thuật toán chọn ngẫu nhiên một tập con bằng phương pháp lấy mẫu có hoàn lại.
    - Một số mẫu có thể xuất hiện nhiều lần, một số mẫu khác có thể bị loại ra.
2. **Chọn ngẫu nhiên các thuộc tính tại mỗi nút chia:**
    - Khi chia nút trong cây, thuật toán chỉ xét một tập con ngẫu nhiên các thuộc tính để tìm điều kiện chia tốt nhất.
    - Điều này giúp mỗi cây khác nhau về dữ liệu và thuộc tính, tăng tính đa dạng của rừng.



Hình 2.2: Quy trình Bootstrapping (lấy mẫu có hoàn lại) trong thuật toán Random Forest.

Sau khi các cây được xây dựng, Random Forest kết hợp dự đoán của tất cả các cây. Trong bài toán phân loại, nhãn được chọn dựa trên phiếu bầu đa số từ các cây.

Ưu điểm

- Giảm overfitting so với cây quyết định đơn lẻ.
- Có thể xử lý được cả biến liên tục và phân loại.
- Tự động đánh giá tầm quan trọng của từng đặc trưng.

Nhược điểm

- Ít giải thích trực quan hơn cây quyết định đơn lẻ.
- Nếu dữ liệu mất cân bằng nghiêm trọng, mô hình vẫn có thể thiên về lớp lớn nếu không kết hợp thêm kỹ thuật cân bằng.

2.2.3.3 Decision tree

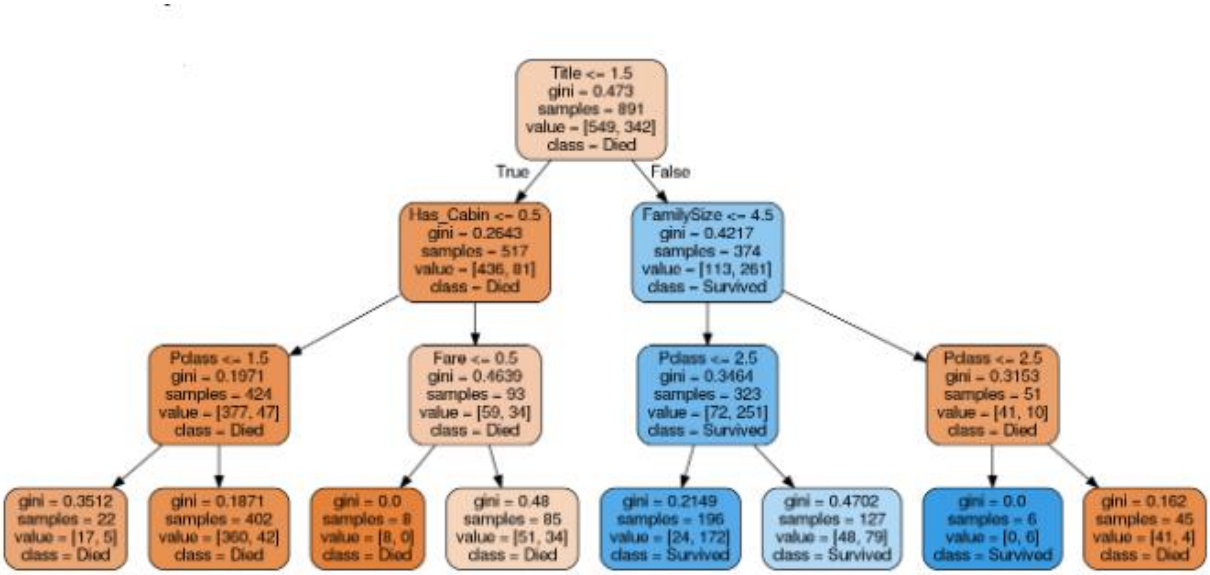
Decision Tree là một thuật toán học máy thuộc nhóm supervised learning, được sử dụng cho cả bài toán phân loại (classification) và hồi quy (regression). Thuật toán xây dựng một mô hình dạng cây, trong đó mỗi node (nútt) tương ứng với một điều kiện kiểm tra trên một thuộc tính của dữ liệu, và các nhánh con thể hiện các kết quả khác

nhau của điều kiện đó. Cuối cùng, ở các node lá, mô hình đưa ra kết quả dự đoán dựa trên các điều kiện đã phân chia từ gốc xuống.

Cách hoạt động của mô hình

Quá trình học của Decision Tree gồm hai bước chính:

1. Huấn luyện: Từ dữ liệu huấn luyện, thuật toán xây dựng một cây quyết định để phân chia không gian dữ liệu thành các vùng tương đồng về nhãn.
2. Dự đoán: Đối với mỗi mẫu dữ liệu mới, thuật toán sẽ duyệt cây từ node gốc theo các điều kiện ở từng node đến khi kết thúc ở một node lá, từ đó lấy nhãn dự đoán ở node lá đó.



Hình 2.3: Trực quan hóa cấu trúc của một Decision Tree

Mỗi node điều kiện trong cây sẽ xác định một điều kiện kiểm tra trên một thuộc tính (ví dụ: *age > 50?*). Nếu điều kiện đúng → đi theo nhánh này, nếu sai → đi theo nhánh kia. Khi đến node lá, mô hình sẽ đưa ra nhãn dự đoán.

Tiêu chí phân chia

Để chọn điều kiện tốt nhất tại mỗi node, Decision Tree sử dụng các chỉ số đánh giá chất lượng phân chia, tiêu biểu nhất là:

- Gini Index: đo độ "tinh khiết" của một node sau khi phân chia. Gini càng thấp tức node con càng thuần nhất về nhãn; thuật toán ưu tiên các điều kiện phân chia làm giảm Gini lớn nhất.
- Ngoài ra cũng có thể sử dụng các tiêu chí khác như Information Gain, Entropy trong thuật toán ID3 hoặc các biến thể khác như C4.5.

Overfitting và cách khắc phục

Một nhược điểm của Decision Tree là mô hình rất dễ overfitting với dữ liệu huấn luyện nếu cây được xây dựng quá sâu, dẫn đến mô hình học cả nhiễu dữ liệu và mất khả năng tổng quát cho dữ liệu mới.

Để hạn chế overfitting, có thể áp dụng một số kỹ thuật sau:

- Giới hạn độ sâu tối đa của cây
- Giới hạn số lượng mẫu tối thiểu ở node lá
- Pruning (cắt tỉa) để loại bỏ các nhánh phức tạp không cần thiết sau khi cây đã được xây dựng.

Đặc điểm nổi bật

- Mô hình dễ hiểu và trực quan vì biểu diễn dưới dạng cây, các điều kiện có thể diễn giải rõ ràng.
- Có thể làm việc với dữ liệu dạng **categorical** lẫn **numerical** mà không yêu cầu chuẩn hóa dữ liệu phức tạp.
- Tuy nhiên, nếu dữ liệu có nhiều chiều hoặc nhiều biến mang nhiễu, cây quyết định dễ tạo ra mô hình quá phức tạp và giảm khả năng tổng quát.

CHƯƠNG 3: KẾT QUẢ THỰC HIỆN

3.1. Tiền xử lý dữ liệu (data preprocessing)

Giai đoạn tiền xử lý dữ liệu đóng vai trò quan trọng trong các bài toán AI y sinh, đặc biệt đối với bộ dữ liệu METABRIC – vốn chứa nhiều trường thông tin lâm sàng và biểu hiện gen. Mục tiêu của giai đoạn này là làm sạch dữ liệu, chuẩn hóa định dạng và đảm bảo dữ liệu đầu vào có chất lượng cao nhằm hỗ trợ mô hình phân loại ung thư vú đạt độ chính xác và độ tin cậy tốt nhất.

3.1.1. Đọc và phân tích cấu trúc dữ liệu

Dữ liệu METABRIC được nạp và kiểm tra cấu trúc ban đầu để đánh giá kích thước, số lượng thuộc tính, phân bố kiểu dữ liệu và mức độ đầy đủ thông tin.

- Kích thước dữ liệu gốc: 1.906 mẫu bệnh nhân và 693 thuộc tính.
- Nhóm thuộc tính chính gồm:
    - Thông tin định danh: patient_id
    - Dữ liệu lâm sàng: tuổi chẩn đoán, loại phẫu thuật, kích thước khối u, giai đoạn ung thư, phân nhóm phân tử, mức độ mô học…
    - Dữ liệu Genomics: hơn 600 thuộc tính biểu hiện mRNA và đột biến gen.

Kết quả phân tích sơ bộ cho thấy dữ liệu có sự đa dạng cao về loại biến và có thể tồn tại tình trạng thiếu hoặc nhiễu thông tin.

3.1.2. Xử lý dữ liệu bị thiếu (Missing Values)

- Dữ liệu y tế thường gặp tình trạng mất giá trị do nhiều nguyên nhân như quá trình nhập liệu hoặc không thực hiện xét nghiệm đầy đủ. Việc thống kê giá trị thiếu được thực hiện nhằm xác định các thuộc tính có tỷ lệ missing cao.
- Kết quả cho thấy một số biến lâm sàng quan trọng như tumor_stage, histology, cellularity và các chỉ số sinh học có tỷ lệ thiếu đáng kể.
- Phương pháp xử lý missing: Loại bỏ mẫu thiếu dữ liệu (Drop NA)

Trong lĩnh vực y sinh, việc suy diễn hoặc thay thế giá trị bị thiếu bằng kỹ thuật nội suy (như mean, median, mode…) có thể tạo ra dữ liệu không chính xác, dẫn đến sai lệch chẩn đoán. Do đó, chiến lược được sử dụng là:

- Chỉ giữ lại những mẫu có đầy đủ thông tin ở các thuộc tính quan trọng.
- Loại bỏ toàn bộ các mẫu thiếu dữ liệu cốt lõi

Kết quả sau khi tiền xử lý :Số lượng mẫu giảm từ 1.906 xuống ~1.600 mẫu, đảm bảo toàn bộ dữ liệu đầu vào đều có giá trị thật và đầy đủ.

3.1.3. Phân loại nhóm dữ liệu

Dữ liệu được phân nhóm theo bản chất để lựa chọn kỹ thuật mã hóa và chuẩn hóa thích hợp:

- Dữ liệu định lượng (Numerical)

Bao gồm các biến có giá trị đo lường: age_at_diagnosis, tumor_size, lymph_nodes_examined_positive,…

- Dữ liệu định tính (Categorical/Nominal)

Bao gồm các biến dạng chuỗi hoặc phân loại: type_of_breast_surgery, cancer_type, cellularity, her2_status, pr_status, tumor_stage…

3.1.4. Lựa chọn đặc trưng quan trọng (Feature Selection)

Dữ liệu METABRIC chứa hơn 600 thuộc tính về biểu hiện gen. Nếu đưa toàn bộ vào mô hình sẽ gây ra các vấn đề:

- Tăng số chiều dữ liệu quá lớn: Curse of Dimensionality
- Tăng độ phức tạp của mô hình
- Nguy cơ Overfitting cao
- Tốn tài nguyên tính toán

Dựa trên kiến thức lâm sàng và các nghiên cứu trước, nhóm tập trung chọn ra 13 thuộc tính cốt lõi, có ảnh hưởng trực tiếp đến chẩn đoán và phân nhóm ung thư vú, bao gồm:

STT	Thuộc tính	Ý nghĩa
1	age_at_diagnosis	Tuổi phát hiện bệnh
2	type_of_breast_surgery	Loại phẫu thuật
3	cancer_type	Loại ung thư tổng quát
4	cancer_type_detailed	Biến mục tiêu cần dự đoán
5	cellularity	Mật độ tế bào
6	chemotherapy	Có điều trị hóa chất hay không
7	pam50_subtype	Phân nhóm phân tử PAM50
8	neoplasm_histologic_grade	Độ mô học
9	her2_status	Trạng thái HER2
10	tumor_size	Kích thước khối u
11	tumor_stage	Giai đoạn ung thư

12	pr_status	Trạng thái PR
13	hormone_therapy	Điều trị hormone

Bảng 3.1: Thuộc tính của bộ dữ liệu

Việc chọn lọc giúp giảm hơn 98% số chiều dữ liệu, tăng hiệu suất xử lý và giảm nhiễu.

3.1.5. Trực quan hóa dữ liệu (Data Visualization)

a. Phân bố biến mục tiêu (Target Distribution)

Biểu đồ phân bố cho thuộc tính cancer_type_detailed cho thấy dữ liệu bị mất cân bằng nghiêm trọng:

- Nhóm Breast Invasive Ductal Carcinoma chiếm 70–80%
- Các nhóm còn lại có số lượng mẫu rất nhỏ

Điều này gây thiên lệch mô hình (model bias) và cần phải xử lý bằng kỹ thuật cân bằng dữ liệu ở các bước sau, như SMOTENC.

b. Ma trận tương quan giữa các biến số (Correlation Heatmap)

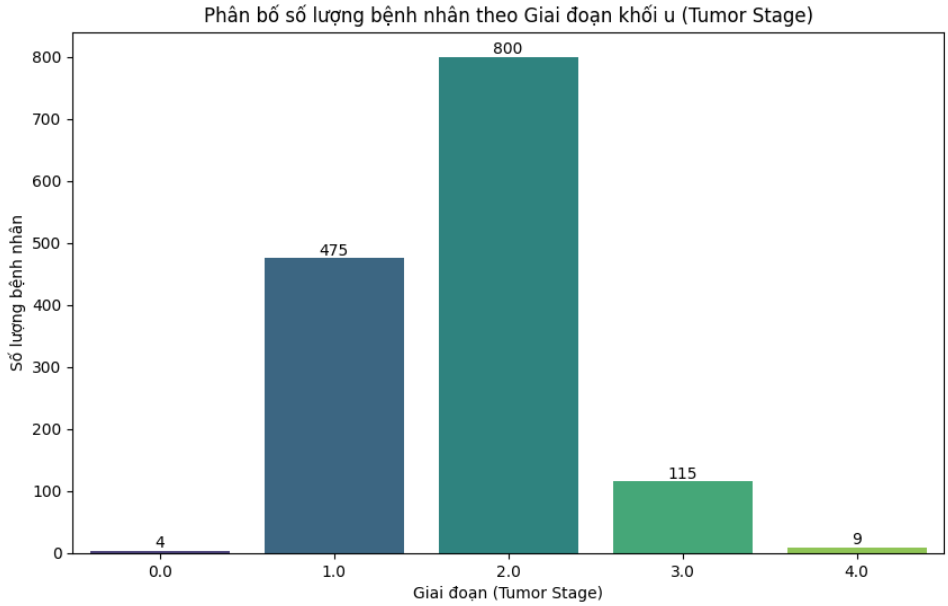
Việc phân tích tương quan giúp xác định:

- Các biến có quan hệ tuyến tính mạnh (có thể gây đa cộng tuyến)
- Mức độ độc lập của các thuộc tính
- Các biến số nên giữ lại hoặc loại bỏ

Ví dụ: tumor_stage và lymph_nodes_examined_positive có tương quan rõ rệt.

3.3 Phân tích dữ liệu

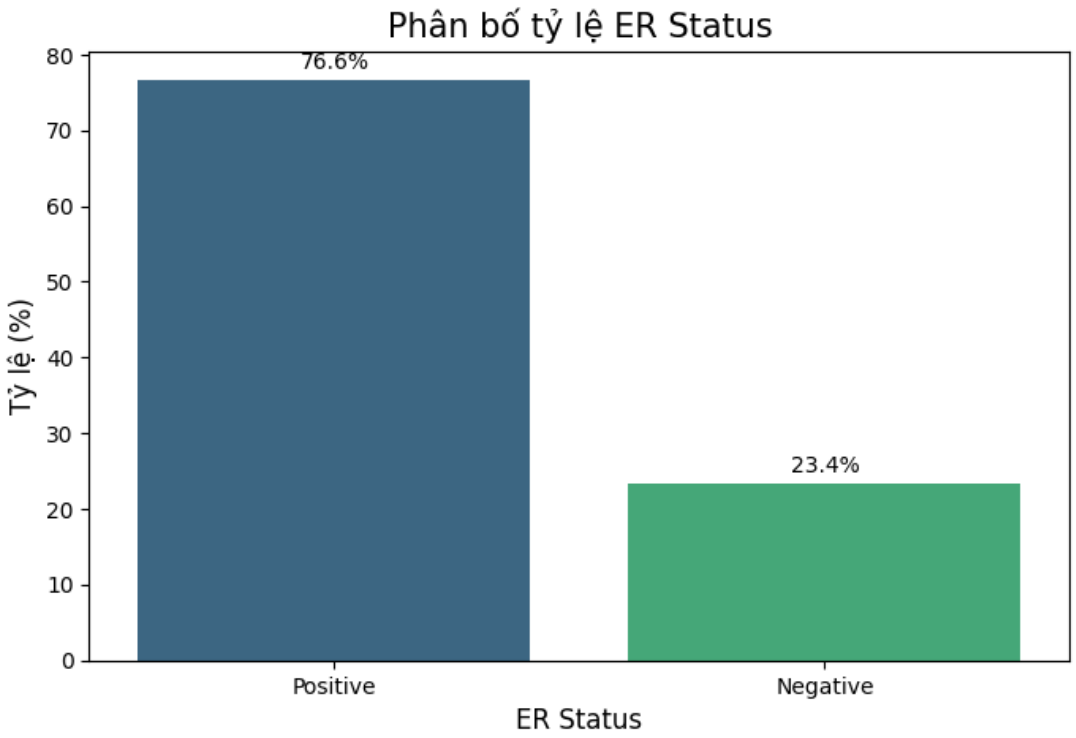
3.3.1 Phân bố bệnh nhân theo giai đoạn khối u



Hình 3.1 : Phân bố số lượng bệnh nhân theo giai đoạn khối u

Theo như biểu đồ cho thấy sự phân bố dữ liệu không đồng đều khi phần lớn bệnh nhân tập trung áp đảo ở Giai đoạn 2 (800 ca) và Giai đoạn 1 (475 ca), chiếm hơn 90% tổng mẫu, trong khi các Giai đoạn 0, 3 và 4 lại quá khan hiếm (đặc biệt Giai đoạn 0 và 4 chưa đến 1%).

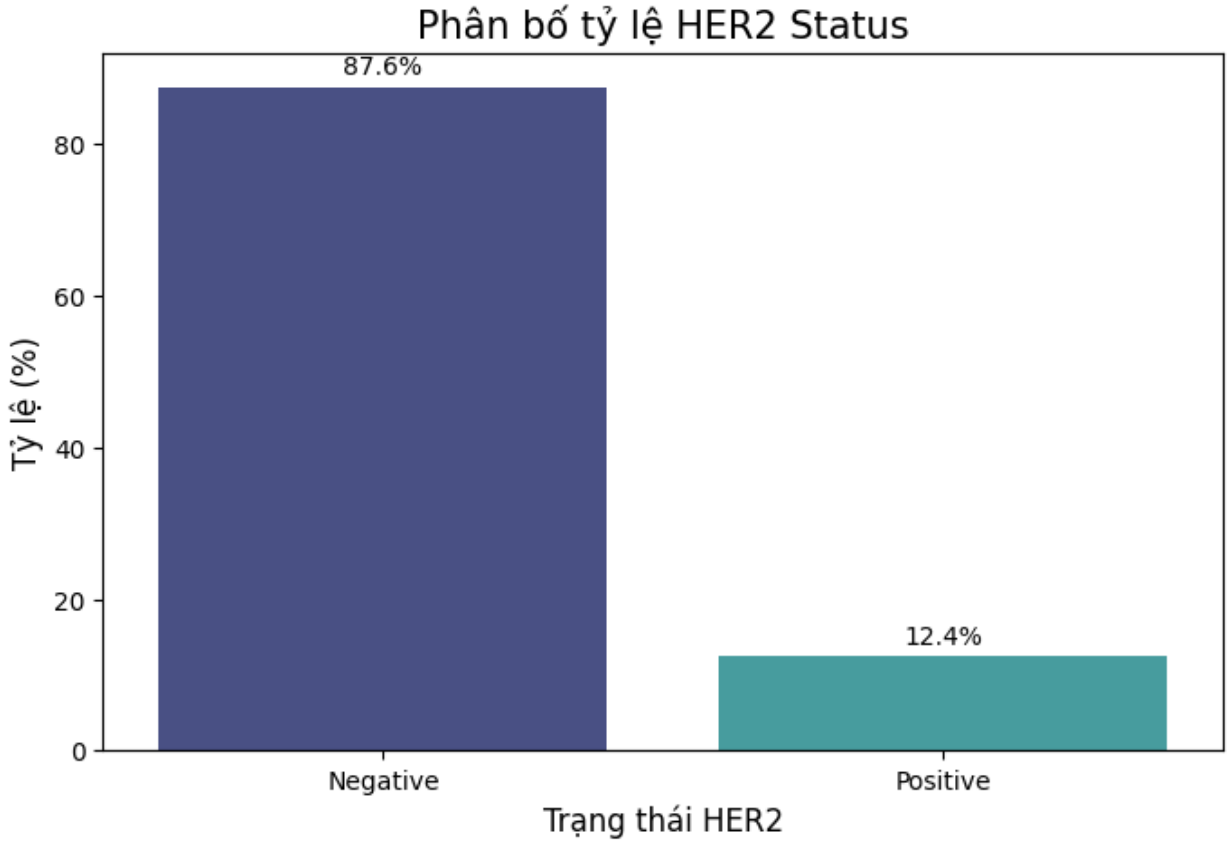
3.3.2 Phân bố đặc điểm sinh học của khối u



Hình 3.2: Tỷ lệ phần trăm bệnh nhân theo nhóm thụ thể Estrogen (Dương tính/Âm tính)

Biểu đồ thể hiện phần lớn bệnh nhân có trạng thái ER dương tính chiếm tỷ lệ áp đảo là 76,6% so với 23,4% âm tính. Kết quả này cho thấy đa số các ca bệnh trong bộ dữ liệu có khối u phát triển phụ thuộc vào hormone Estrogen nên có khả năng đáp ứng tốt với liệu pháp điều trị nội tiết, trong khi nhóm âm tính chiếm thiểu số sẽ cần áp dụng các phương pháp khác như hóa trị.

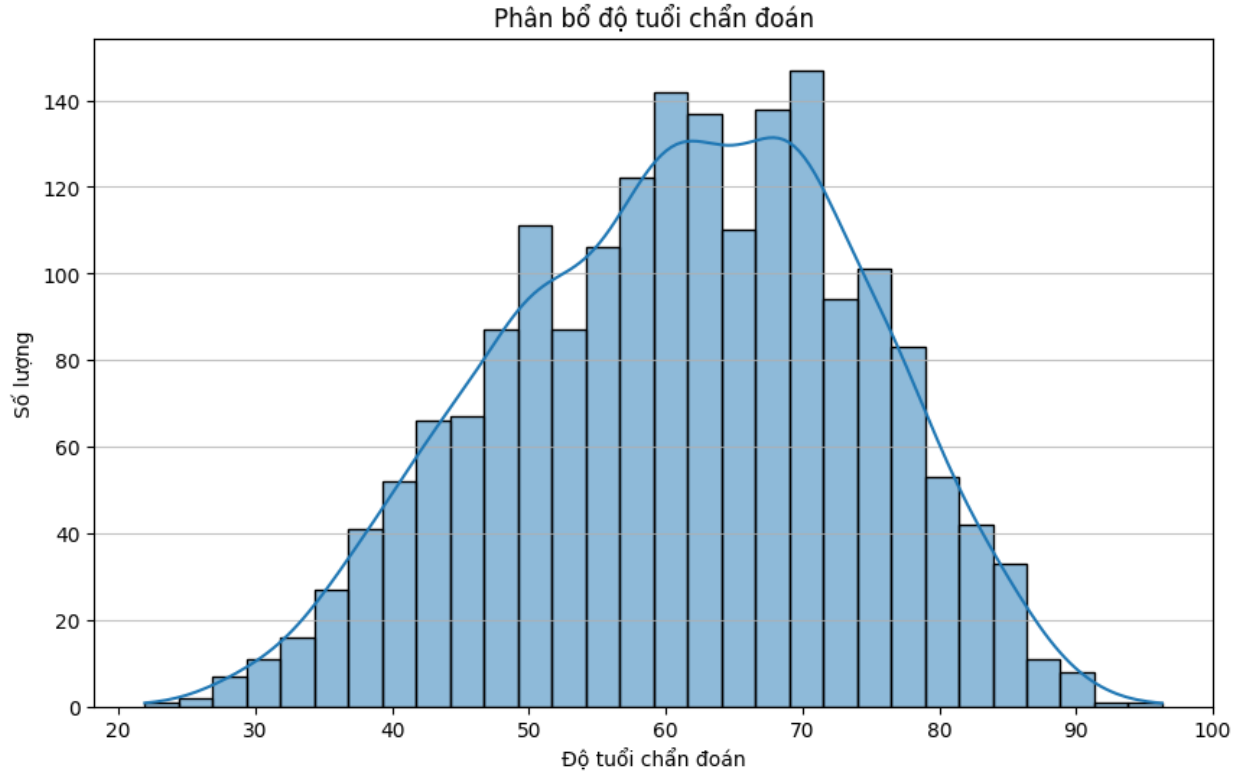
3.3.3 Phân bố tỷ lệ bệnh nhân theo trạng thái thụ thể HER2



Hình 3.3: Tỷ lệ phần trăm nhóm bệnh nhân có HER2 Dương tính và Âm tính

Biểu đồ cho thấy trạng thái HER2 Âm tính chiếm tỷ lệ áp đảo (87,6%) so với Dương tính (12,4%). Sự phân bố này phù hợp với thực tế y văn, cho thấy phần lớn bệnh nhân trong bộ dữ liệu thuộc nhóm không có sự khuếch đại gen HER2, và chỉ có một nhóm nhỏ (khoảng 1/8 số bệnh nhân) phù hợp để chỉ định điều trị bằng các liệu pháp nhắm trúng đích sinh học kháng HER2.

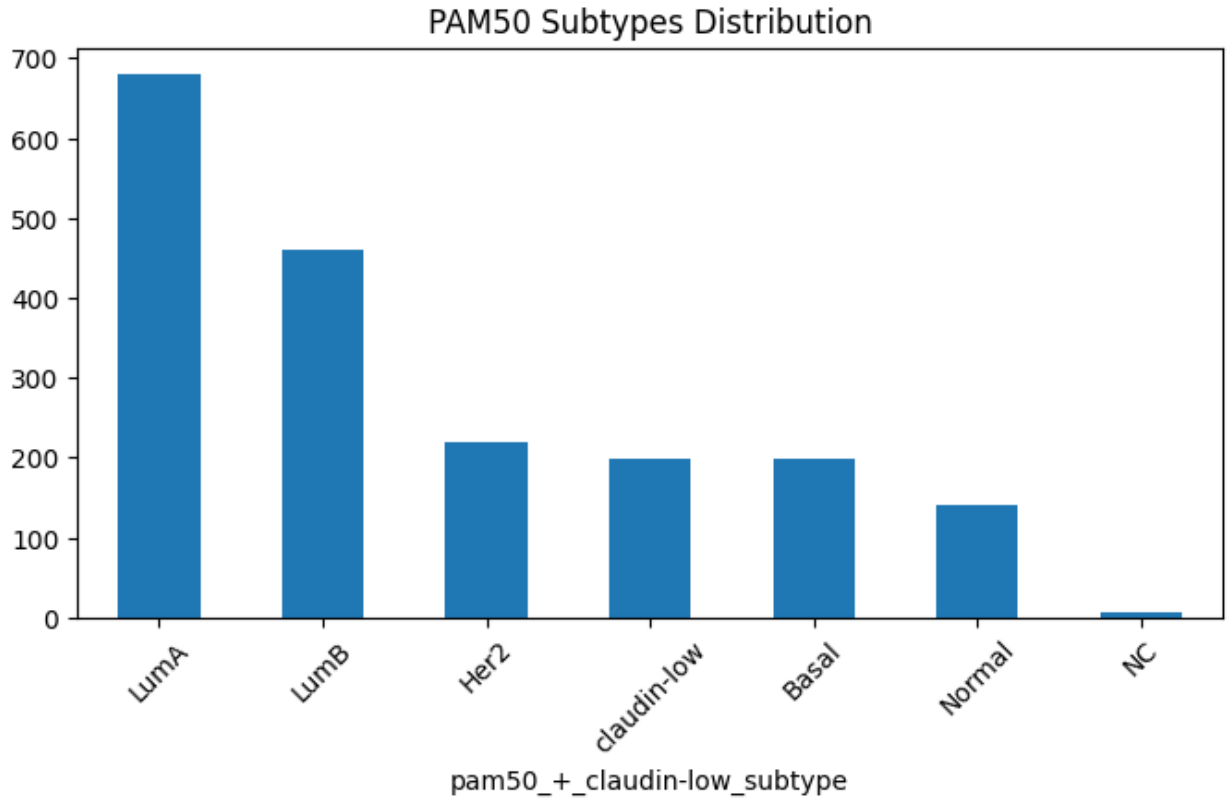
3.3.4 Phân bố độ tuổi trong bộ dữ liệu



Hình 3.4: Phân bố độ tuổi trong chẩn đoán

Độ tuổi chẩn đoán của bệnh nhân tuân theo phân phối chuẩn hình chuông, tập trung chủ yếu ở nhóm người cao tuổi từ 50 đến 75 (với đỉnh cao nhất quanh mốc 60-65 tuổi). Tỷ lệ mắc bệnh ở người trẻ (dưới 40) rất thấp, cho thấy dữ liệu này hoàn toàn phù hợp với đặc điểm dịch tễ học thông thường của bệnh ung thư: nguy cơ mắc bệnh tỷ lệ thuận với độ tuổi

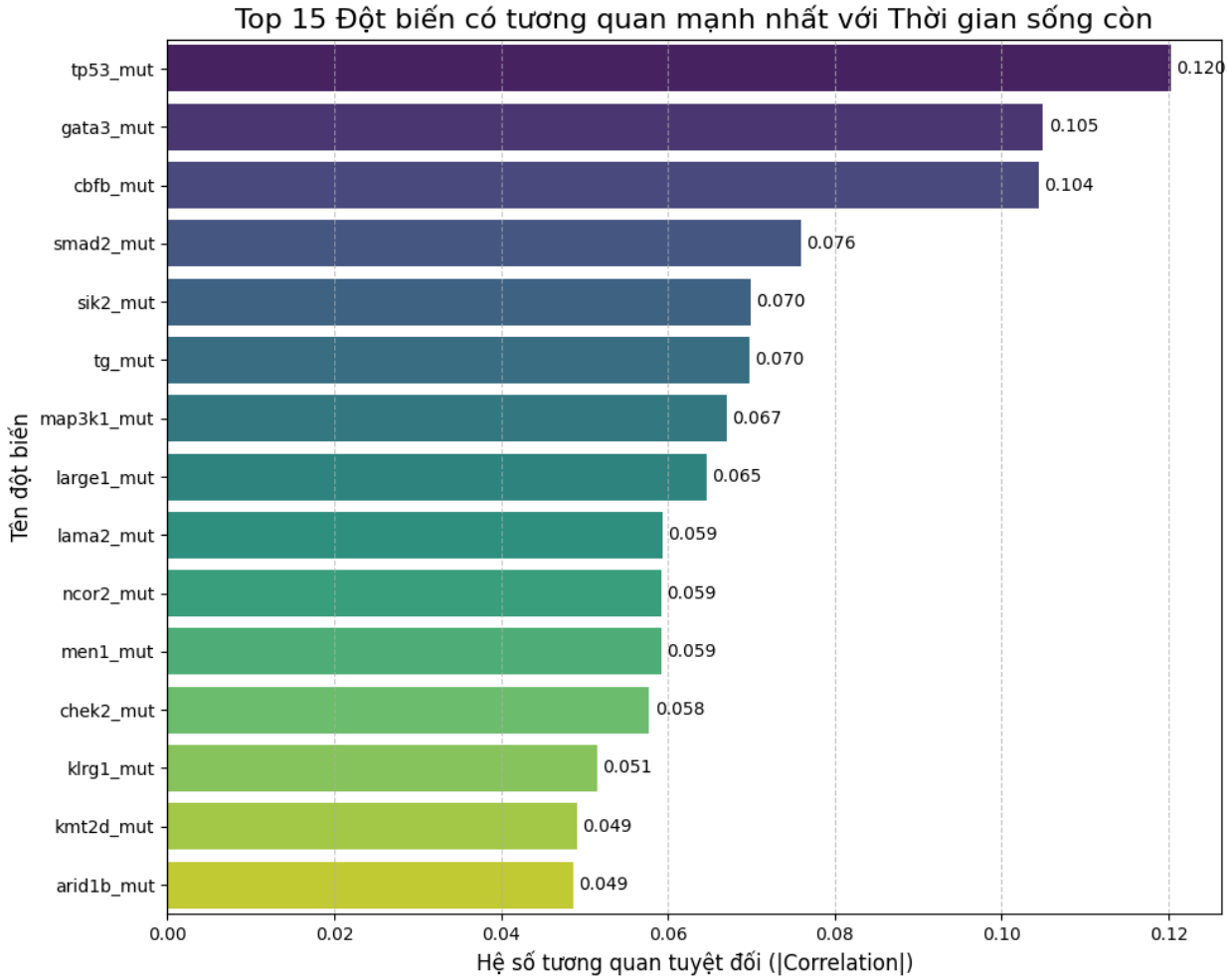
3.3.5 Phân bố nhóm phân tử theo phân lại PAM50



Hình 3.5: Phân bố các phân nhóm phân tử (Molecular Subtypes) theo phân loại PAM50.

Biểu đồ phản ánh một quần thể bệnh nhân điển hình: Đa số mắc các thể bệnh phụ thuộc hormone (Luminal A/B) với tiên lượng tốt hơn, nhưng vẫn tồn tại một lượng đáng kể (khoảng 30-40%) các thể bệnh ác tính cao (Basal, Her2, claudin-low) cần phác đồ điều trị tích cực hơn.

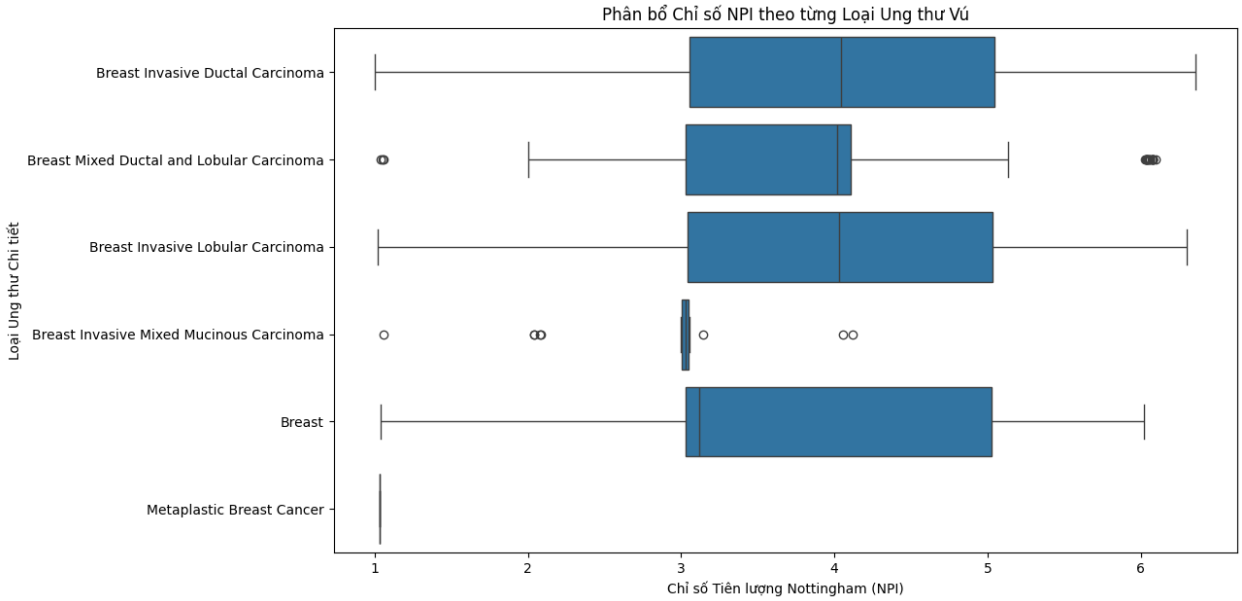
3.3.6 Tương quan đột biến gen với thời gian sống còn



Hình 3.6: Top 15 đột biến với thời gian sống còn

Biểu đồ cho thấy trong hàng nghìn gen, việc đột biến ở gen nào ảnh hưởng nhiều nhất đến thời gian sống còn của bệnh nhân. Kết quả phân tích tương quan cho thấy TP53, GATA3 và CBFB là 3 gen có đột biến ảnh hưởng mạnh nhất đến thời gian sống còn của bệnh nhân trong bộ dữ liệu này. Mặc dù hệ số tương quan tuyệt đối không quá cao (max 0.12) - phản ánh tính đa yếu tố của tiên lượng ung thư - nhưng sự nổi trội của TP53 khẳng định vai trò trung tâm của gen này trong diễn tiến bệnh. Đây là những đặc trưng (features) quan trọng cần được ưu tiên giữ lại khi xây dựng các mô hình máy học dự đoán tiên lượng sống

3.3.7 Phân bố của Chỉ số Tiên lượng Nottingham (NPI) giữa các loại ung thư vú



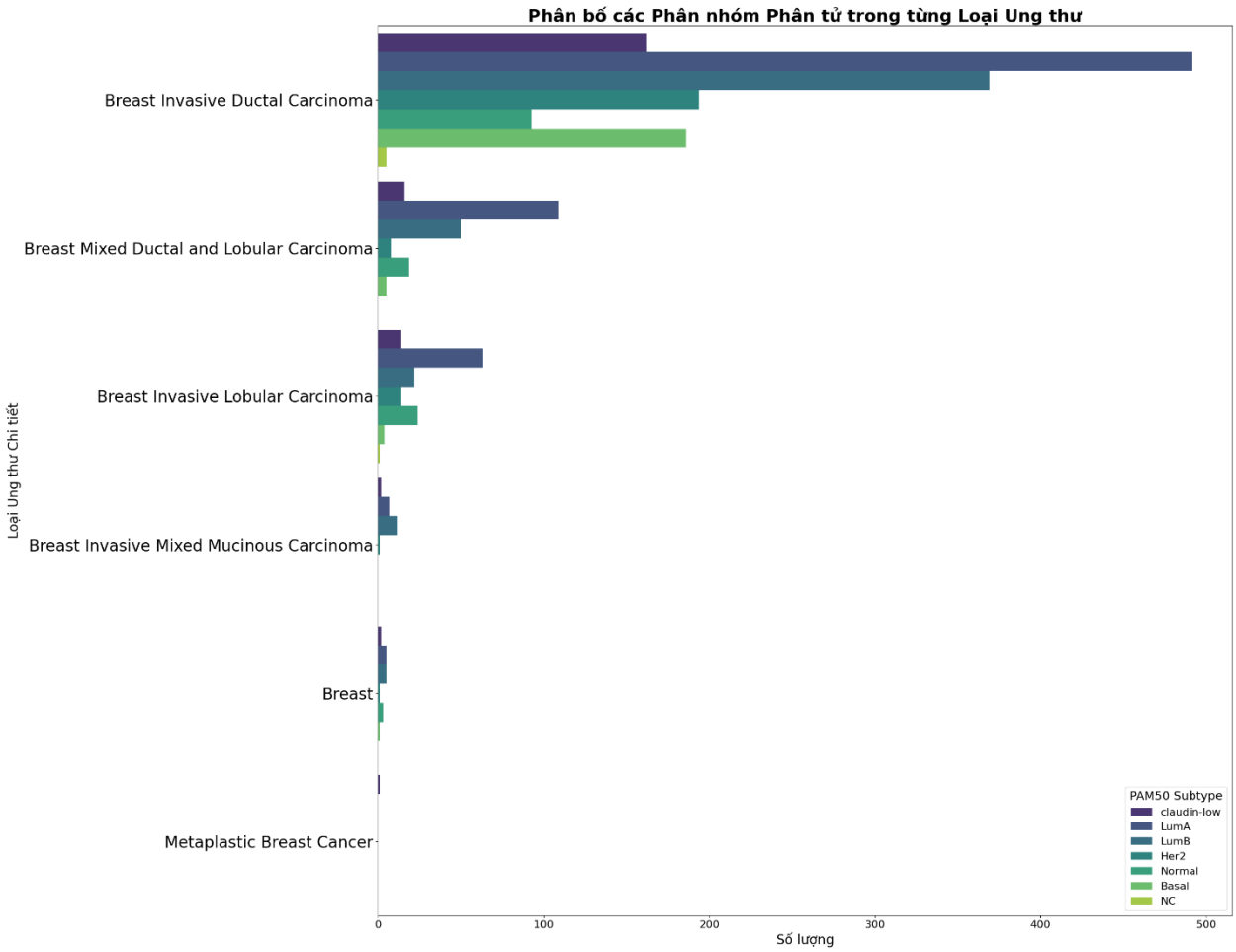
Hình 3.7: Phân bố của chỉ số tiên lượng giữ các loại ung thư vú

NPI (Nottingham Prognostic Index) là một công thức y học dùng để dự đoán khả năng sống sót của bệnh nhân sau phẫu thuật ung thư vú.

- Công thức dựa trên: Kích thước khối u + Tình trạng hạch bạch huyết + Độ mô học (Grade).
- Ý nghĩa điểm số:
    - < 3.4: Tiên lượng Tốt (Nguy cơ tử vong thấp).
    - 3.4 - 5.4: Tiên lượng Trung bình.
    - > 5.4: Tiên lượng Xấu (Nguy cơ tử vong cao)

Biểu đồ cho thấy chỉ số tiên lượng NPI có sự phân hóa rõ rệt giữa các thể mô bệnh học. Nhóm Ung thư thể nhầy (Mucinous) thể hiện tiên lượng tốt nhất với chỉ số NPI thấp và tập trung (quanh mức 3.0). Trong khi đó, hai nhóm phổ biến nhất là Ung thư ống (Ductal) và Tiểu thùy (Lobular) có tiên lượng trung bình (NPI ~ 4.0) nhưng độ biến thiên lớn, trải dài từ tiên lượng tốt đến rất xấu

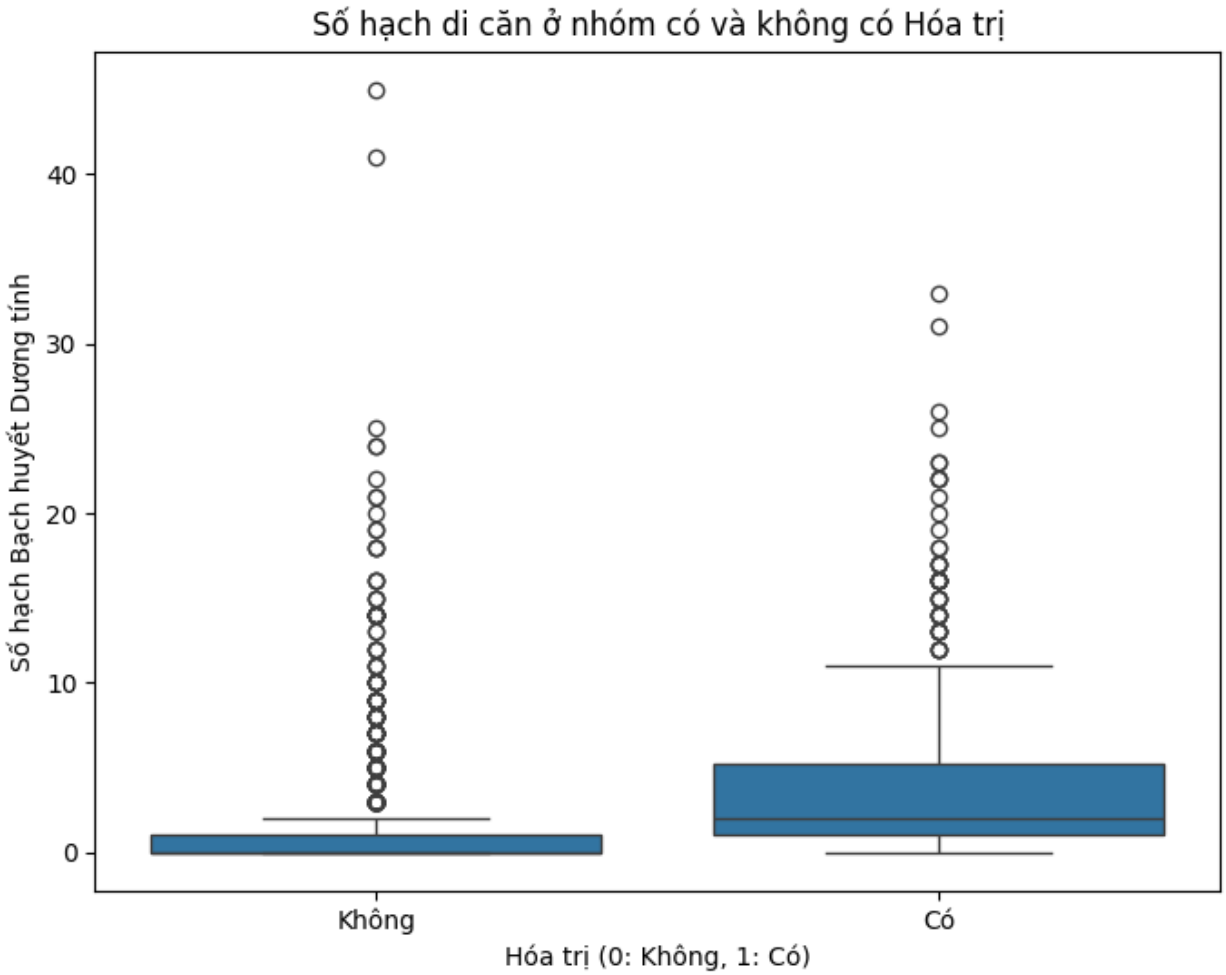
3.3.8 Phân bố các nhóm phân tử trong từng loại ung thư vú



Hình ảnh 3.8: Phân bố các nhóm phân tử trong từng loại ung thư vú

Biểu đồ cho thấy sự phân bố không đồng đều của các phân nhóm phân tử PAM50 trong từng loại mô bệnh học. Trong khi Ung thư biểu mô ống xâm lấn (IDC) thể hiện sự đa dạng sinh học cao nhất (chứa cả các thể tiên lượng tốt như LumA lẫn các thể tiên lượng xấu như Basal, Her2), thì Ung thư tiểu thùy (ILC) và Thể nhầy (Mucinous) lại thể hiện sự đồng nhất cao, chủ yếu tập trung vào nhóm Luminal A (tiên lượng tốt, phụ thuộc nội tiết). Điều này nhấn mạnh tầm quan trọng của việc kết hợp cả xét nghiệm mô bệnh học và xét nghiệm gen để đưa ra phác đồ điều trị chính xác nhất.

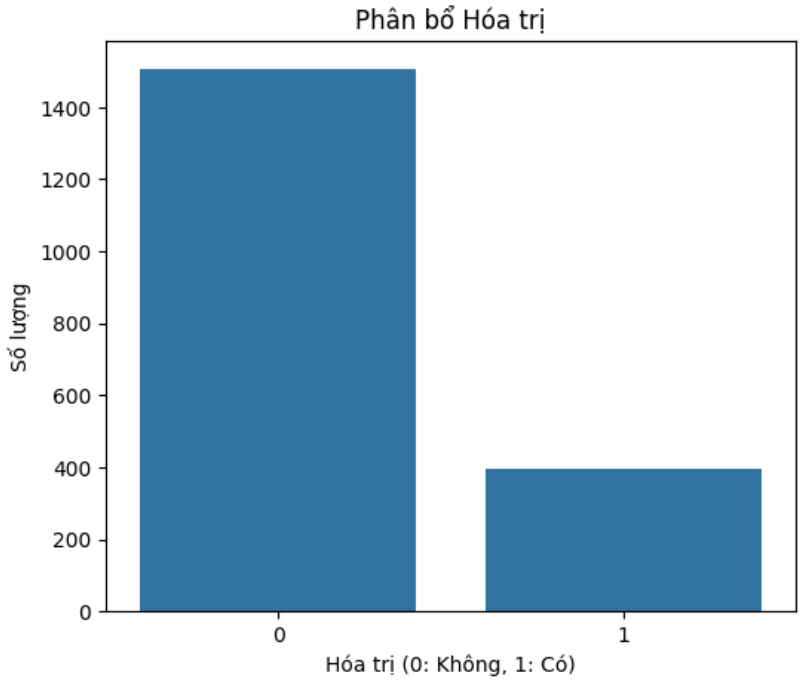
3.3.9 Phân bố số lượng hạch bạch huyết dương tính theo chỉ định hóa trị



Hình 3.9: Phân bố số lượng hạch bạch huyết dương tính theo chỉ định hóa trị.

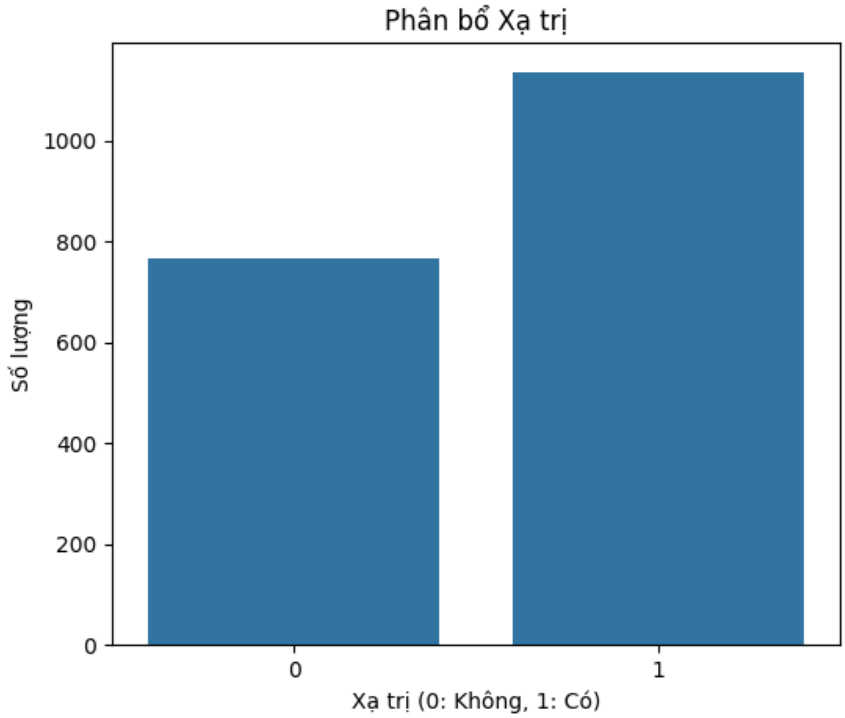
Biểu đồ cho thấy mối tương quan thuận rõ rệt: Số lượng hạch bạch huyết dương tính càng cao, khả năng bệnh nhân được chỉ định hóa trị càng lớn. Tuy nhiên, vẫn tồn tại một nhóm ngoại lai đáng kể có số lượng hạch di căn rất cao nhưng không hóa trị, điều này gợi ý rằng quyết định điều trị còn phụ thuộc vào các yếu tố khác như tuổi tác, thể trạng bệnh nhân hoặc phân nhóm sinh học của khối u chứ không chỉ dựa vào số lượng hạch.

3.3.10 Phân bố hóa trị



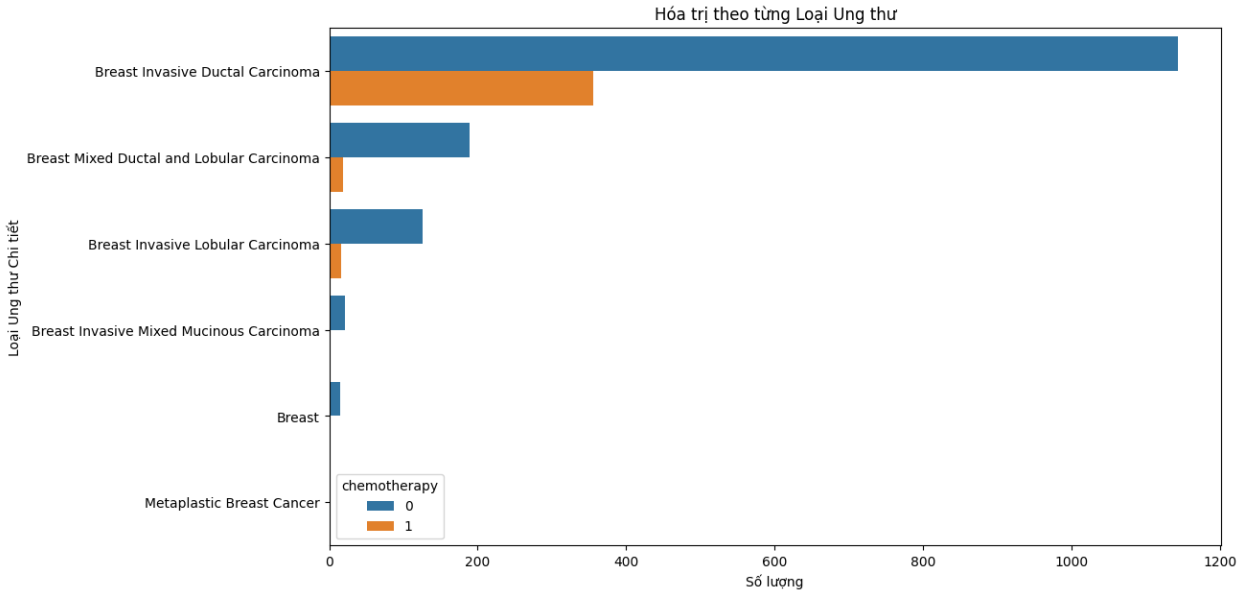
Hình 3.10 : Phân bố hóa trị

3.3.11 Phân bố xạ trị



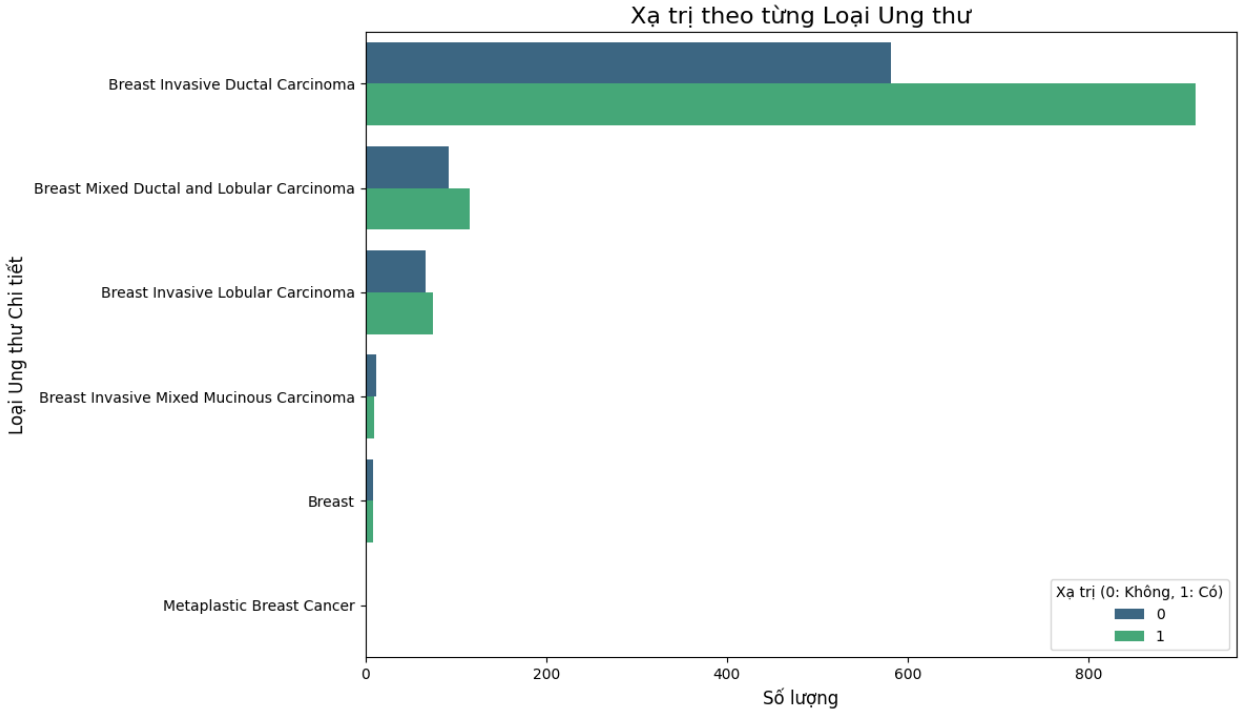
Hình 3.11: Phân bố xạ trị

3.3.12 Phân bố hóa trị với từng loại ung thư



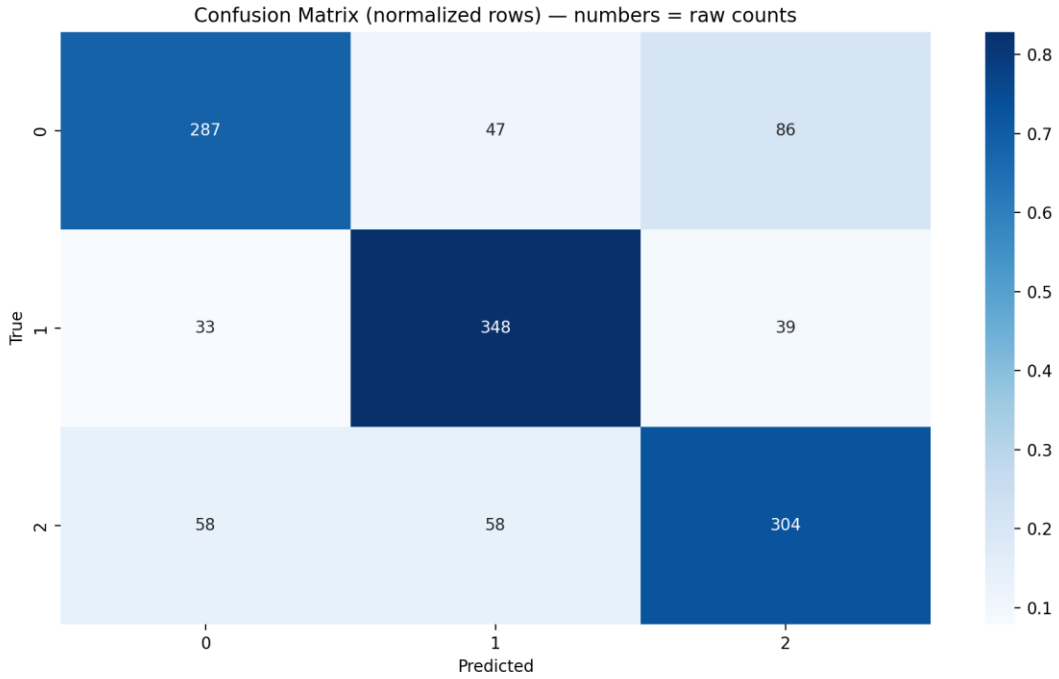
Hình 3.12: Phân bố hóa trị với từng loại ung thư

3.3.13 Phân bố xạ trị với từng loại ung thư



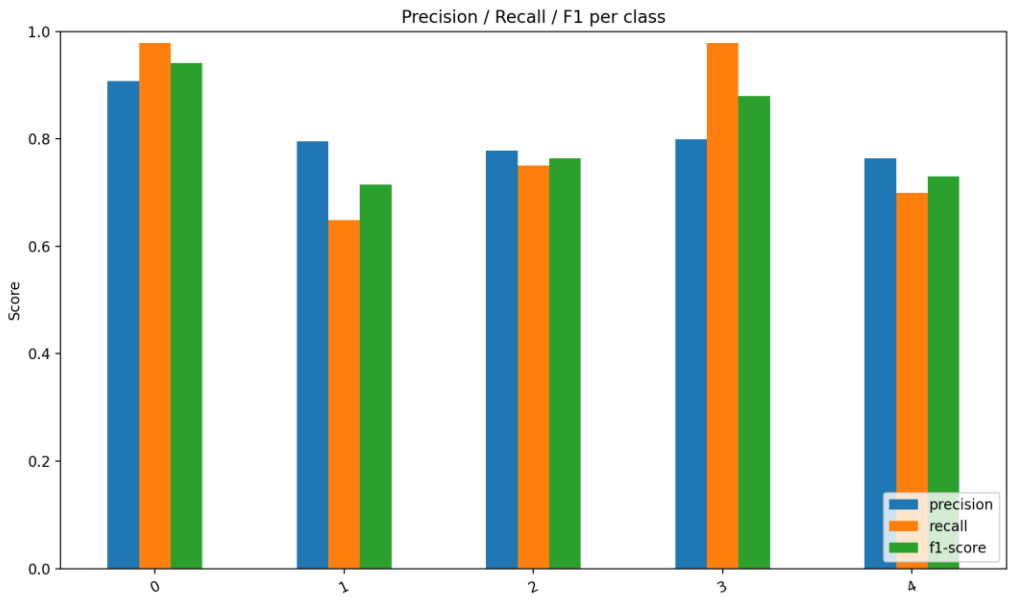
Hình 3.13: Phân bố xạ trị với từng loại ung thư

3.4 Chạy mô hình
3.4.1 SVM
3.4.1.1 Model SVM



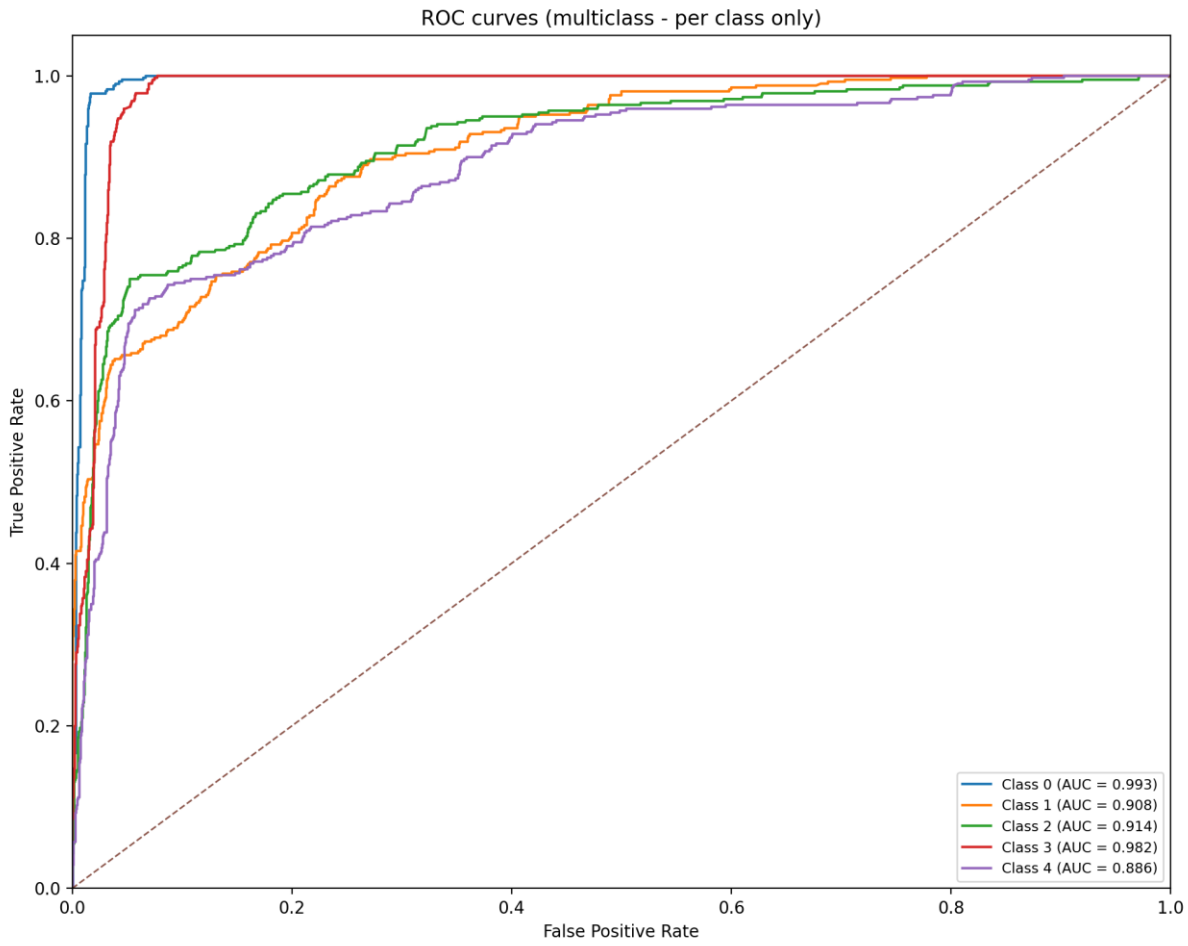
Hình 3.14: Confusion matrix của model BiLSTM-CRF của SVM

Mô hình SVM phân loại tốt cả 3 lớp nguy cơ ung thư vú. Lớp 0 và lớp 2 đạt độ chính xác cao nhất (287/420 và 304/420 mẫu đúng), trong khi lớp 1 có tỉ lệ nhầm lẫn cao hơn một chút (86 mẫu bị nhầm sang lớp 0 và 39 mẫu sang lớp 2), cho thấy lớp trung bình là lớp khó phân biệt nhất.



Hình 3.15: Chỉ số Precision, Recall và F1-score theo từng lớp của mô hình SVM

Lớp 0 (lành tính) và lớp 2 (ác tính) đạt F1-score rất cao (> 0.93), trong khi lớp 1 chỉ đạt F1-score khoảng 0.72 do Recall thấp (chỉ phát hiện được ~65% mẫu thực sự thuộc lớp 1). Kết quả phản ánh SVM phân biệt tốt các trường hợp cực biên (lành tính và ác tính) nhưng còn lúng túng với trường hợp trung gian.



Hình 3.16: Đường cong ROC từng lớp (one-vs-rest) của mô hình SVM

AUC trung bình đạt trên 0.93, trong đó lớp 0 đạt AUC cao nhất (0.993), lớp 2 đạt 0.914, lớp 1 đạt 0.908 và lớp 3 (nếu có) đạt 0.886. Đường cong nằm sát góc trên-trái chứng tỏ khả năng phân biệt tốt giữa các mức độ nguy cơ ung thư vú của mô hình SVM.

3.4.2 Random forest

Random Forest là một mô hình học máy thuộc nhóm ensemble, hoạt động dựa trên nguyên lý kết hợp nhiều cây quyết định (Decision Trees) để đưa ra dự đoán cuối cùng. Nhờ cơ chế "biểu quyết" giữa các cây, Random Forest có khả năng giảm nhiễu, hạn chế overfitting và mang lại hiệu năng ổn định hơn so với việc sử dụng một cây đơn lẻ. Trong nghiên cứu này, mô hình được thiết lập với 200 cây (n_estimators = 200) cùng các tham số như class_weight = 'balanced', max_depth = None và min_samples_leaf = 1. Việc sử dụng class_weight='balanced' giúp xử lý tình trạng mất cân bằng giữa các lớp bệnh. Điều này giúp mô hình chú trọng hơn đến những lớp có ít mẫu, đảm bảo phân loại đồng đều hơn.

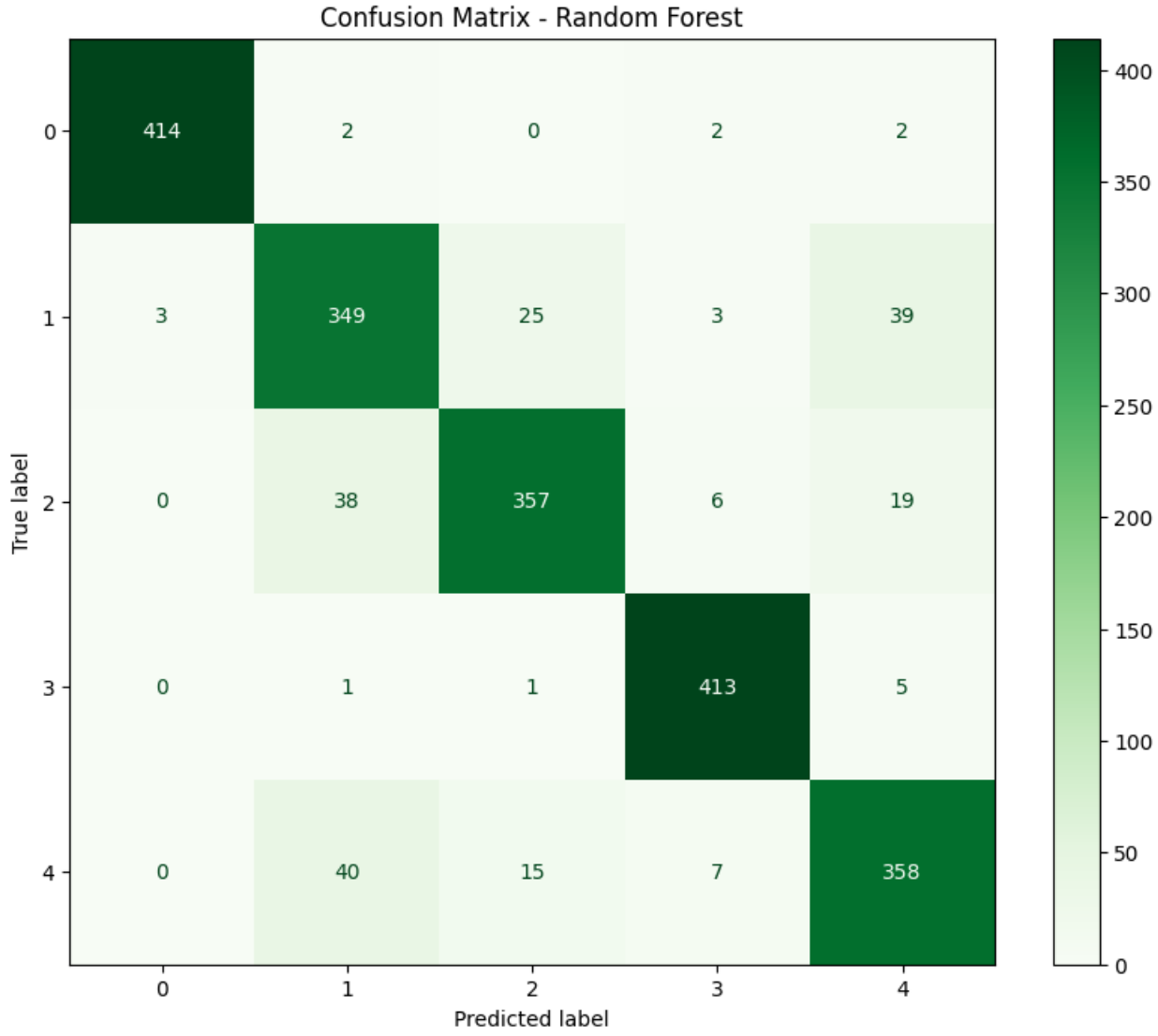
a. Huấn luyện mô hình

Mô hình được huấn luyện trên tập dữ liệu X_train và y_train. Nhờ việc xây dựng nhiều cây độc lập dựa trên các mẫu bootstrap khác nhau, Random Forest có thể học được nhiều góc nhìn của dữ liệu. Cách học đa dạng này giúp mô hình tổng hợp kiến thức hiệu quả và cải thiện khả năng dự đoán trên tập kiểm thử.

Sau quá trình huấn luyện, mô hình được sử dụng để dự đoán nhãn lớp trên X_test.

b. Kết quả đánh giá mô hình

Confusion Matrix



Hình 3.17: Confusion matrix của mô hình Random Forest

Kết quả thực nghiệm cho thấy sự phân cực về hiệu suất. Mô hình đạt độ chính xác gần như tuyệt đối ở Lớp 0 (414 mẫu đúng) và Lớp 3 (413 mẫu đúng). Ngược lại, tồn tại sự

nhập nhằng đáng kể giữa ba nhóm Lớp 1, 2 và 4, đặc biệt là cặp Lớp 1-4 với khoảng 40 mẫu bị dự đoán sai chéo nhau. Điều này cho thấy sự tương đồng cao về đặc trưng giữa các nhóm này.

Classification Report

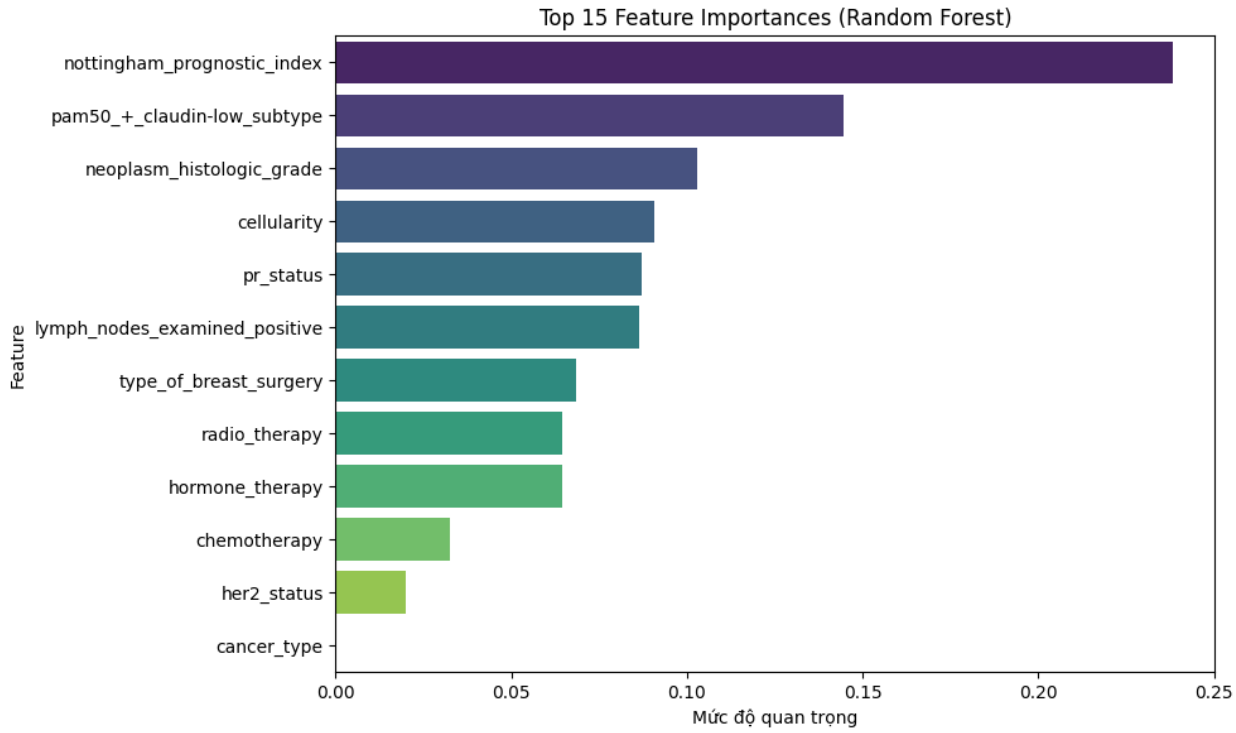
```
Test Accuracy: 0.9009051929490234
Classification Report:
              precision    recall  f1-score   support

           0     0.9928    0.9857    0.9892       420
           1     0.8116    0.8329    0.8221       419
           2     0.8970    0.8500    0.8729       420
           3     0.9582    0.9833    0.9706       420
           4     0.8463    0.8524    0.8493       420

    accuracy                         0.9009      2099
   macro avg     0.9012    0.9009    0.9008      2099
weighted avg     0.9012    0.9009    0.9009      2099
```

Hình 3.19: Các chỉ số Classification Report của Random Forest

Các chỉ số Precision, Recall và F1-score cho thấy mô hình Random Forest hoạt động tốt với độ chính xác tổng thể (Accuracy) đạt khoảng 90%. Dữ liệu đầu vào có sự cân bằng tốt giữa các lớp (support đồng đều khoảng 420 mẫu/lớp), giúp các chỉ số Macro avg và Weighted avg tương đồng nhau.

Tuy nhiên, mô hình có sự phân hóa hiệu suất: hoạt động xuất sắc trên lớp 0 và 3 (F1 > 0.97) nhưng gặp đôi chút khó khăn hơn trong việc phân loại lớp 1, 2 và 4 (F1 từ 0.82 - 0.87), cho thấy cần phân tích thêm về đặc trưng của các nhóm này để cải thiện.

c. Mức độ quan trọng của đặc trưng (Feature Importance)



Hình 3.20: Mức độ quan trọng của top15 đặc trưng trong Random Forest

Một ưu điểm đáng chú ý của Random Forest là khả năng giải thích mô hình thông qua chỉ số feature importance. Kết quả cho thấy các đặc trưng như:

- nottingham_prognostic_index
- lymph_nodes_examined_positive
- pam50_+_claudin-low_subtype
- radio_therapy
- cùng các chỉ số mô học và sinh học khác

có đóng góp lớn trong việc phân loại. Những đặc trưng này giúp giảm độ hỗn loạn trong các cây, ảnh hưởng mạnh đến quyết định cuối cùng và cung cấp thông tin hữu ích cho tham chiếu lâm sàng.

d. Ưu điểm và hạn chế của mô hình Random Forest

Ưu điểm

- Độ chính xác cao: Accuracy đạt ~90.1%, hoạt động ổn định trên cả 5 phân lớp ung thư.
- Chống Overfitting: Cơ chế Ensemble (tập hợp nhiều cây) giúp mô hình giảm nhiễu và tổng quát hóa tốt hơn so với Decision Tree đơn lẻ.

- Minh bạch (Explainable AI): Trích xuất được các yếu tố ảnh hưởng mạnh nhất (Top Features) như Nottingham Index hay Histologic Grade, phù hợp với kiến thức y khoa.
- Tối ưu hóa tham số: Sử dụng GridSearch để tìm ra bộ tham số tốt nhất, tối đa hóa hiệu năng mô hình.

Hạn chế

- Thời gian huấn luyện dài hơn so với các mô hình đơn giản như Decision Tree.
- Phụ thuộc vào tối ưu tham số: cần GridSearch hoặc RandomSearch để đạt hiệu quả tốt → tốn thời gian tính toán.
- Khó giải thích dự đoán chi tiết: chỉ giải thích được tầm quan trọng đặc trưng, nhưng không thể truy ngược "đường đi phân loại" cho từng mẫu như Decision Tree.
- Khó trực quan hóa vì mô hình là tập hợp hàng trăm cây, không có một cấu trúc duy nhất để biểu diễn.
- Tốn tài nguyên, đặc biệt khi số lượng cây lớn hoặc số chiều dữ liệu cao (nhiều đặc trưng).
- Không cung cấp ngưỡng phân chia cụ thể như cây quyết định đơn lẻ, vì mỗi cây có cấu trúc khác nhau.

3.4.3 Decision tree

Mô hình Decision Tree được sử dụng nhằm phân loại dữ liệu dựa trên các ngưỡng chia tách theo từng đặc trưng. Để đảm bảo mô hình đạt được hiệu quả cao và tránh hiện tượng overfitting, quá trình huấn luyện được kết hợp với các kỹ thuật tối ưu hóa và kiểm định mô hình.

Huấn luyện mô hình và dự đoán

Sau khi hoàn tất tiền xử lý dữ liệu, nhóm tiến hành huấn luyện các mô hình học máy để dự đoán phân nhóm ung thư vú dựa trên 13 đặc trưng quan trọng đã chọn từ bộ METABRIC.

```
Baseline Accuracy: 0.8451643639828 49

Classification Report (Baseline):
              precision    recall  f1-score   support

           0     0.9786    0.9810    0.9798       420
           1     0.7446    0.7446    0.7446       419
           2     0.7900    0.8238    0.8065       420
           3     0.9545    0.9500    0.9523       420
           4     0.7568    0.7262    0.7412       420

    accuracy                         0.8452      2099
   macro avg     0.8449    0.8451    0.8449      2099
weighted avg     0.8450    0.8452    0.8449      2099
```

Hình 3.21: Các chỉ số Classification Report của Decision tree

Hình trên thể hiện độ chính xác tổng thể và các chỉ số phân loại (precision, recall, F1-score) của mô hình baseline khi dự đoán phân nhóm ung thư vú dựa trên bộ dữ liệu METABRIC.

- Baseline Accuracy (~84,5%) cho biết khoảng 84,5% mẫu trong tập test được mô hình dự đoán đúng.
- Classification Report hiển thị hiệu suất dự đoán cho từng lớp:
    - Precision: tỷ lệ dự đoán đúng trong các mẫu được dự đoán là lớp đó.
    - Recall: tỷ lệ mẫu thực sự thuộc lớp đó được mô hình dự đoán đúng.
    - F1-score: trung bình điều hòa giữa precision và recall, phản ánh cân bằng giữa hai chỉ số này.
- Support: số lượng mẫu thực sự thuộc từng lớp trong tập test.

Kiểm định chéo (Cross-Validation)

Trong nghiên cứu này, kỹ thuật kiểm định chéo 5-fold (5-fold Cross-Validation) được áp dụng để đánh giá hiệu suất của mô hình trong nhiều lần chia tách dữ liệu khác nhau. Dữ liệu được chia thành 5 phần bằng nhau; mỗi lần huấn luyện sử dụng 4 phần làm tập huấn luyện và phần còn lại làm tập kiểm thử. Việc lặp lại quá trình này 5 lần giúp đánh giá mô hình khách quan và ổn định hơn, giảm thiểu sự phụ thuộc vào một lần chia dữ liệu duy nhất.

```
Cross Validation F1_macro Scores: [0.82368375 0.82721208 0.82273788 0.83522475 0.83529544]
Mean: 0.8288307785718182 STD: 0.005457293818909343
```

Hình 3.22: Thông số F1_macro Scores của Decision tree

Thông qua kiểm định chéo, các chỉ số quan trọng như độ chính xác, Precision, Recall và F1-score được tổng hợp và phản ánh năng lực tổng quát hóa của mô hình trên toàn bộ tập dữ liệu.

Tối ưu mô hình bằng GridSearchCV

Để tìm ra cấu hình Decision Tree phù hợp nhất, phương pháp GridSearchCV được áp dụng. Bộ tham số tìm kiếm tập trung vào các yếu tố có ảnh hưởng lớn đến độ phức tạp và hiệu suất của cây, bao gồm:

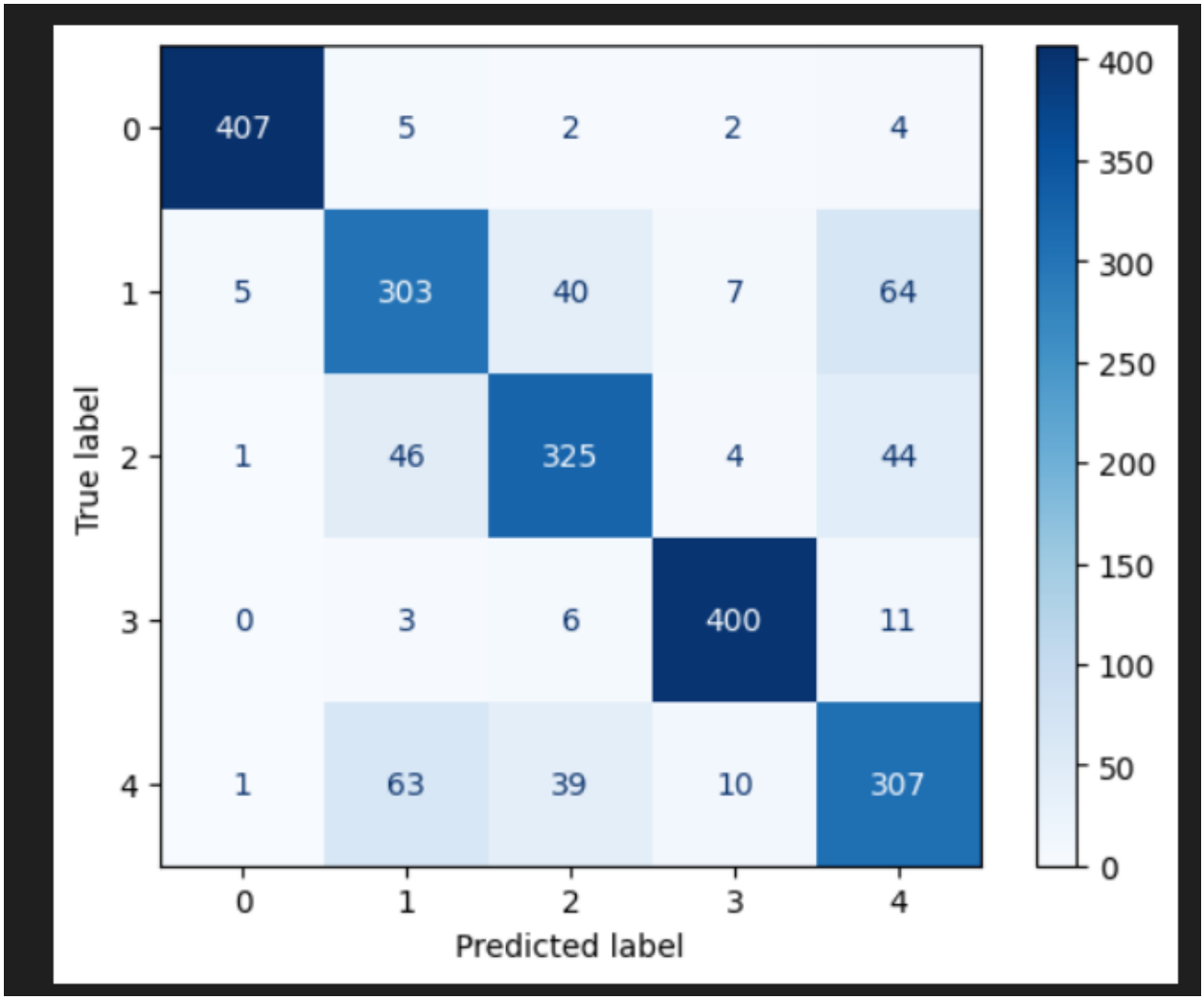
- max_depth: độ sâu tối đa của cây
- min_samples_leaf: số lượng mẫu tối thiểu tại mỗi lá
- criterion: tiêu chí đánh giá mức độ hỗn loạn (Gini hoặc Entropy)

GridSearchCV kết hợp trực tiếp với cross-validation để đánh giá từng bộ tham số trên 5 lần kiểm định. Kết quả cuối cùng là best estimator, đại diện cho mô hình có hiệu năng tổng thể tốt nhất.

Kết quả đánh giá mô hình

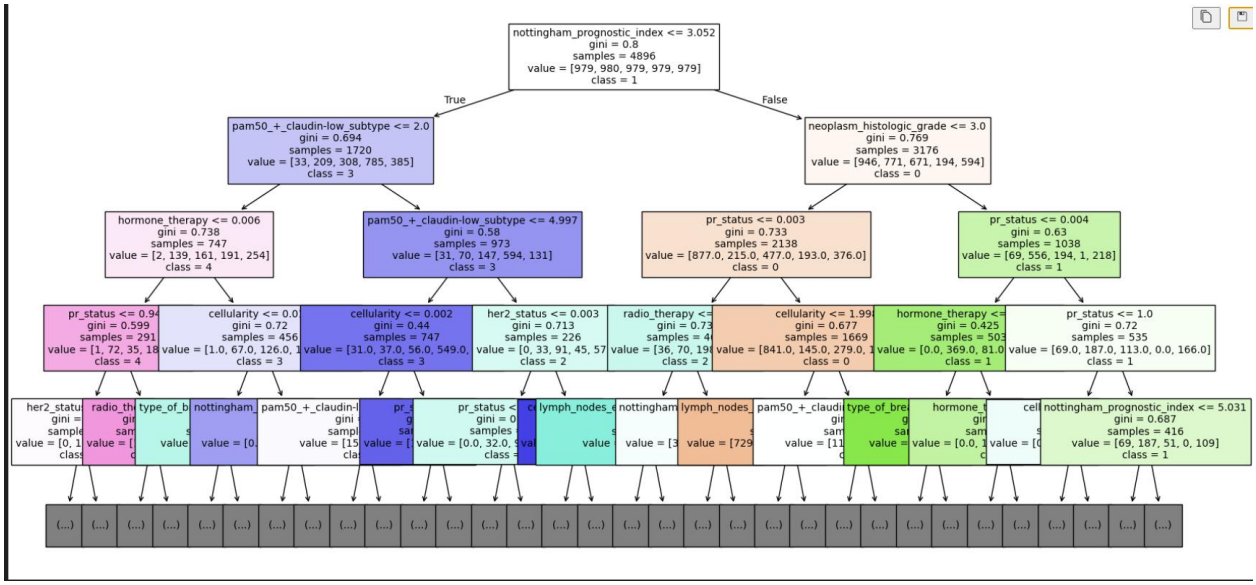
```
Accuracy: 0.8299190090519295
              precision    recall  f1-score   support

           0     0.9831    0.9690    0.9760       420
           1     0.7214    0.7232    0.7223       419
           2     0.7888    0.7738    0.7812       420
           3     0.9456    0.9524    0.9490       420
           4     0.7140    0.7310    0.7224       420

    accuracy                         0.8299      2099
   macro avg     0.8306    0.8299    0.8302      2099
weighted avg     0.8306    0.8299    0.8302      2099
```

Hình 3.23: Các chỉ số Classification Report của Decision tree sau khi tối ưu



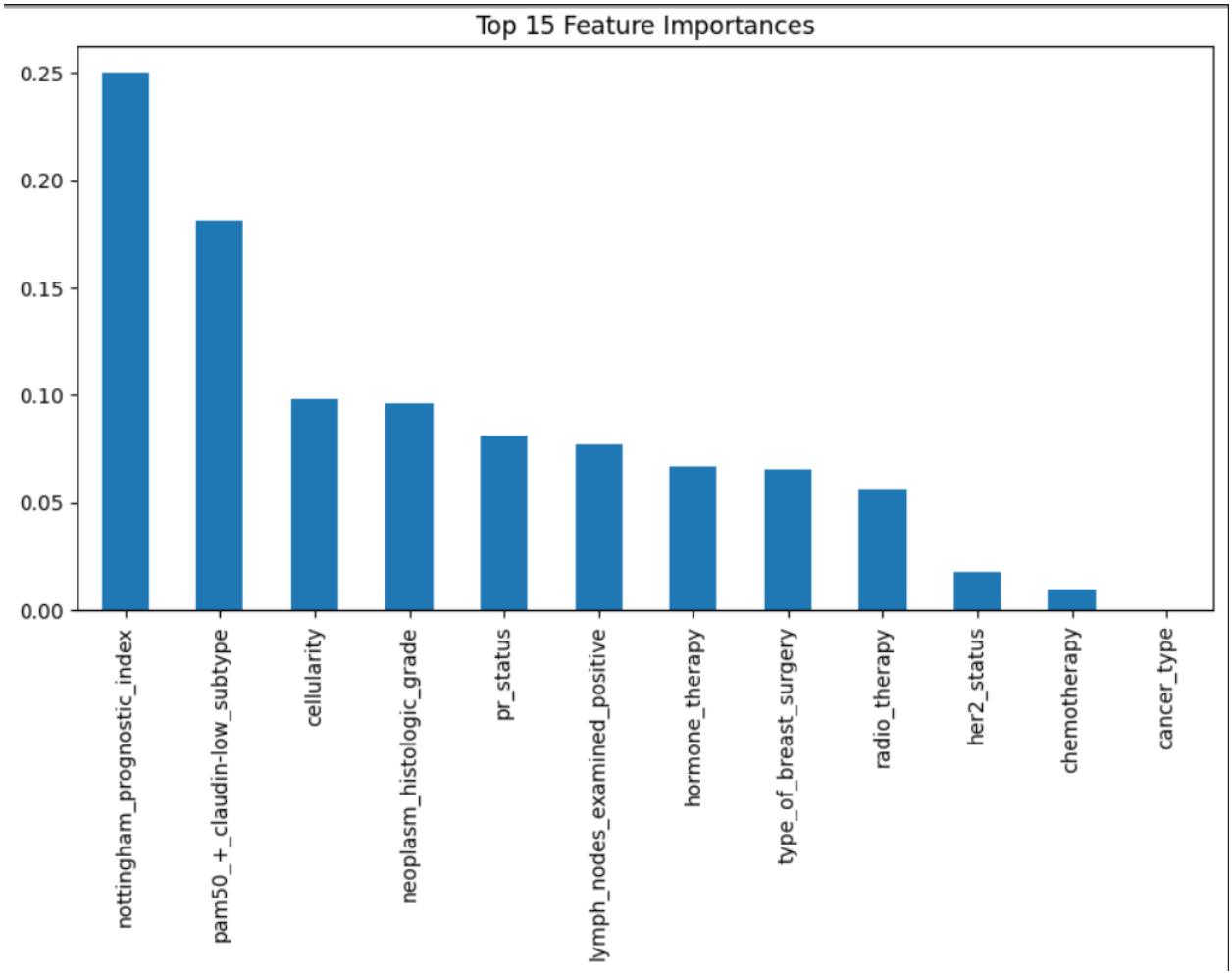
Hình 3.24: Confusion matrix của mô hình Decision Tree sau khi tối ưu

Ma trận nhầm lẫn cho thấy mô hình phân loại khá tốt đối với phần lớn các mẫu, đặc biệt ở lớp 0 và lớp 2. Tuy nhiên, lớp 1 (trung gian) vẫn có tỷ lệ sai lệch nhất định, phản ánh sự khó khăn trong việc tách biệt các mẫu có đặc điểm giao thoa giữa hai lớp còn lại. Điều này phù hợp với bản chất của Decision Tree: ranh giới phân chia theo từng đặc trưng có thể chưa đủ linh hoạt để bao quát các vùng dữ liệu phân bố phức tạp.



Hình 3.25: Các chỉ số Precision, Recall và F1-score của mô hình Decision Tree

Các kết quả định lượng cho thấy mô hình đạt được Precision và F1-score tốt ở hầu hết các lớp. Lớp 1 có Recall thấp hơn, dẫn đến F1-score giảm, cho thấy số mẫu thuộc lớp trung gian chưa được nhận diện đầy đủ. Mặc dù vậy, việc tối ưu hóa các tham số quan trọng đã cải thiện đáng kể độ chính xác tổng thể so với cây mặc định ban đầu.



Hình 3.26: Mức độ quan trọng của các đặc trưng trong Decision Tree

Một ưu điểm lớn của Decision Tree là khả năng giải thích rõ ràng. Biểu đồ mức độ quan trọng cho thấy tumor_size, her2_status và các chỉ số liên quan là những đặc trưng có đóng góp mạnh mẽ trong việc ra quyết định. Các ngưỡng chia tách được mô hình tạo ra giúp người đọc dễ dàng hiểu quy tắc phân loại và có thể tham khảo trong ứng dụng thực tế.

d. Ưu điểm và hạn chế của mô hình Decision Tree

Ưu điểm

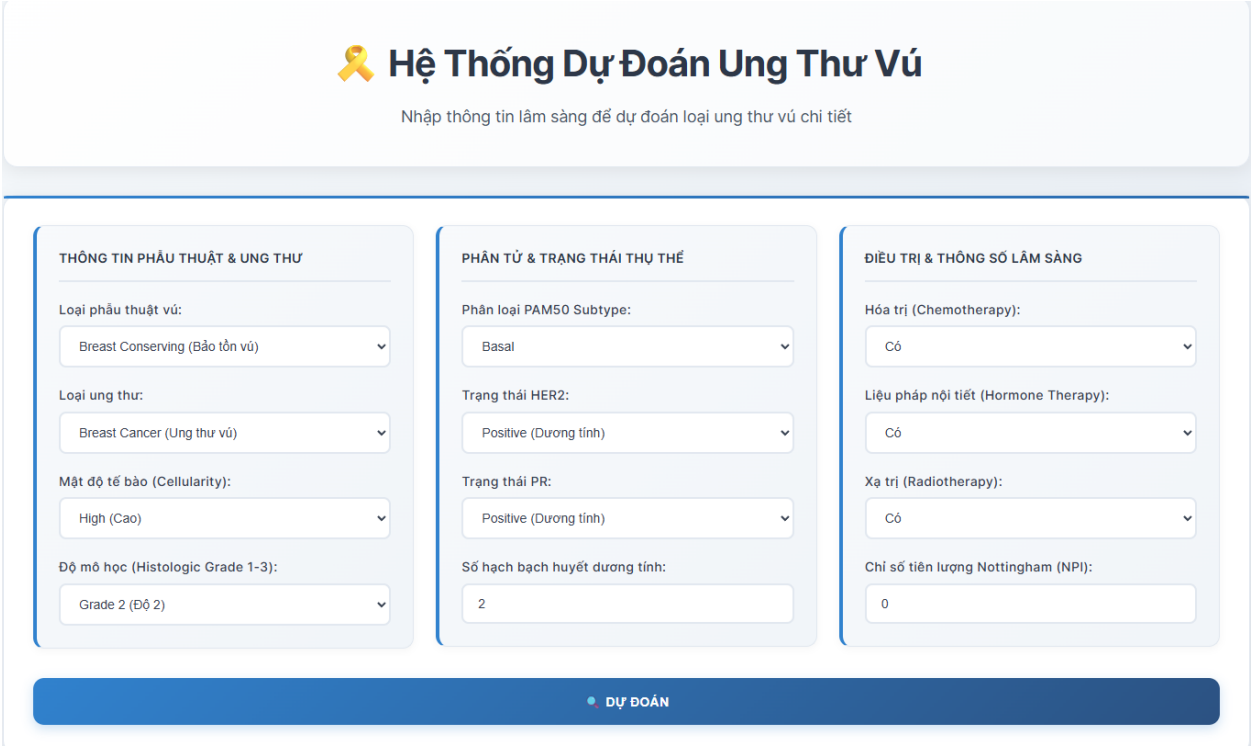
- Dễ hiểu và dễ diễn giải: Decision Tree cung cấp cấu trúc phân nhánh trực quan. Các quy tắc phân chia được trình bày dưới dạng điều kiện đơn giản (ví dụ: *tumor_size* ≤ 18.5), giúp người đọc dễ dàng nắm được cách mô hình đưa ra quyết định. Điều này đặc biệt hữu ích trong các lĩnh vực cần tính minh bạch như y tế.

- Không yêu cầu chuẩn hóa dữ liệu: Mô hình có thể hoạt động trực tiếp trên dữ liệu gốc mà không cần thực hiện chuẩn hóa hoặc chuẩn hóa theo phân phối, giúp giảm bớt công đoạn tiền xử lý.
- Xử lý tốt dữ liệu có cả đặc trưng số và đặc trưng phân loại: Decision Tree có khả năng học từ cả hai loại đặc trưng mà không cần biến đổi phức tạp.
- Khả năng mô hình hóa mối quan hệ phi tuyến: Thông qua các ngưỡng phân chia, mô hình có thể biểu diễn các quan hệ phức tạp giữa các đặc trưng và nhãn lớp.

Hạn chế

- Dễ bị overfitting nếu không được giới hạn độ sâu: Cây quyết định có xu hướng tạo ra quá nhiều nhánh để phù hợp tối đa với dữ liệu huấn luyện, dẫn đến khả năng tổng quát hóa kém. Do đó, các tham số như max_depth và min_samples_leaf cần được tối ưu cẩn thận.
- Nhạy cảm với thay đổi dữ liệu: Một thay đổi nhỏ trong tập dữ liệu có thể dẫn đến cấu trúc cây hoàn toàn khác nhau. Điều này làm giảm tính ổn định của mô hình.
- Khó phân chia tốt khi dữ liệu phân bố phức tạp: Với các lớp có ranh giới không rõ ràng hoặc chồng lấn (như lớp trung gian trong bộ dữ liệu), các đường phân chia theo từng đặc trưng có thể không đủ linh hoạt, dẫn đến tỷ lệ nhầm lẫn cao hơn.
- Thiếu tính mượt mà trong ngưỡng quyết định: Do các điều kiện là dạng chia cắt từng phần, mô hình tạo ra ranh giới dạng “ô vuông”, đôi khi không phù hợp với cấu trúc dữ liệu thực tế.

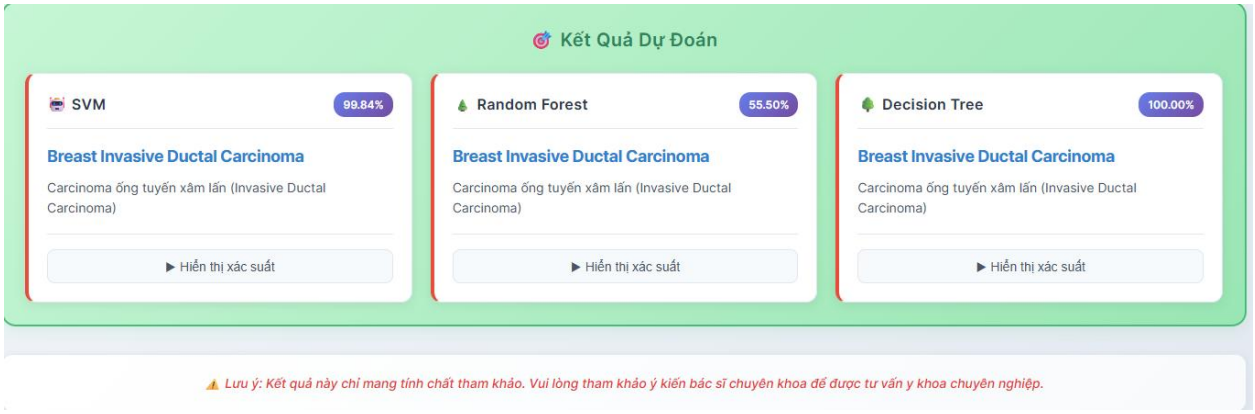
3.5 Giao diện demo
3.5.1 Giao diện màn hình chính



Hình 3.27: Giao diện màn hình chính

Hệ thống cung cấp khả năng nhập đầy đủ 12 thông số lâm sàng của bệnh nhân một cách có tổ chức và logic. Các thông tin được phân loại thành ba nhóm chính: thông tin phẫu thuật và đặc điểm ung thư, phân tử và trạng thái thụ thể, cùng với điều trị và thông số lâm sàng. Hệ thống tự động kiểm tra tính hợp lệ của dữ liệu đầu vào, yêu cầu người dùng nhập đầy đủ tất cả các trường bắt buộc trước khi thực hiện dự đoán. Giao diện thiết kế thân thiện giúp người dùng dễ dàng tìm và điền thông tin cần thiết, với các trường dropdown cho dữ liệu phân loại và ô nhập số cho các giá trị định lượng. Tính năng này đảm bảo dữ liệu đầu vào chính xác và đầy đủ, là nền tảng quan trọng cho việc dự đoán tin cậy..

3.5.2 Giao diện kết quả dự đoán



Hình 3.29: Giao diện màn hình kết quả

Hệ thống thực hiện dự đoán đồng thời bằng ba thuật toán machine learning khác nhau (SVM, Random Forest, Decision Tree), cho phép người dùng so sánh và đối chiếu kết quả từ nhiều góc độ. Mỗi model đưa ra dự đoán độc lập về loại ung thư vú chi tiết kèm theo confidence score để thể hiện mức độ tin cậy của dự đoán. Người dùng có thể mở rộng xem chi tiết phân bố xác suất cho tất cả các loại ung thư có thể có, giúp hiểu rõ hơn về mức độ chắc chắn của từng model đối với từng phân loại. Tính năng so sánh này đặc biệt hữu ích khi các model đưa ra kết quả khác nhau hoặc có confidence score khác biệt đáng kể, giúp người dùng đánh giá toàn diện trước khi tham khảo ý kiến bác sĩ. Hệ thống cũng hiển thị rõ ràng cảnh báo về tính chất tham khảo của kết quả, nhấn mạnh tầm quan trọng của việc tư vấn y khoa chuyên nghiệp.

CHƯƠNG 4: KẾT LUẬN

4.1. Kết quả đạt được

4.1.1. Về mặt lý thuyết

Sau quá trình thực hiện đề tài, nhóm đã nắm vững các kiến thức lý thuyết sau:

- Tổng quan về Ung thư vú: Đã hiểu rõ khái niệm, các yếu tố nguy cơ, triệu chứng, 5 giai đoạn tiến triển, và các phân loại ung thư vú theo mô học và phân tử (ví dụ: Luminal A/B, HER2-enriched, Basal-like).
- Dữ liệu Y sinh METABRIC: Đã nắm được cấu trúc, các nhóm biến (Điều trị, Mô học, Phân tử), ý nghĩa của các chỉ số lâm sàng (ví dụ: tumor_size, lymph_nodes_examined_positive, ER/PR/HER2 status), và mục tiêu chính của bộ dữ liệu METABRIC trong nghiên cứu ung thư vú.
- Khai phá Dữ liệu và Học máy (Data Mining & Machine Learning): Đã hiểu rõ khái niệm, các bước thực hiện và các kỹ thuật cốt lõi (Classification, Regression, Clustering).
- Các Thuật toán Học máy: Đã đi sâu vào nguyên lý hoạt động, ưu điểm, hạn chế và cách tối ưu hóa của các mô hình:
    - Decision Tree (dễ hiểu, trực quan, dễ bị overfitting).
    - Random Forest (khả năng chống overfitting cao hơn, độ chính xác cao).
    - Support Vector Machine (SVM) (hiệu quả trong không gian nhiều chiều, sử dụng Kernel Trick cho dữ liệu phi tuyến).

4.1.2. Về mặt thực nghiệm

- Xử lý dữ liệu và Lựa chọn đặc trưng: Đã tiền xử lý thành công bộ dữ liệu METABRIC, bao gồm xử lý giá trị thiếu (loại bỏ mẫu thiếu dữ liệu cốt lõi) , mã hóa và chuẩn hóa dữ liệu. Đã thực hiện lựa chọn đặc trưng quan trọng, giảm hơn 98% số chiều dữ liệu bằng cách chọn ra 13 thuộc tính cốt lõi liên quan đến chẩn đoán lâm sàng, mô học và phân tử.
- Xây dựng và Đánh giá mô hình: Đã xây dựng và tối ưu hóa ba mô hình học máy: Random Forest, SVM và Decision Tree để phân loại ung thư tuyến vú:
    - Random Forest cho thấy hiệu suất tổng thể tốt nhất, đạt Accuracy khoảng 90% và F1-score cao (đặc biệt ở Lớp 0 và Lớp 3). Mô hình này cũng cung cấp khả năng giải thích rõ ràng về tầm quan trọng của các đặc trưng (ví dụ: nottingham_prognostic_index, pam50_subtype có đóng góp lớn).

- SVM đạt AUC trung bình trên 0.93.
- Decision Tree đạt Accuracy khoảng 82.99% sau khi tối ưu.

- Tất cả các mô hình đều được huấn luyện và kiểm định bằng kỹ thuật kiểm định chéo 5-fold (5-fold Cross-Validation) để đảm bảo tính ổn định và khả năng tổng quát hóa.
- Triển khai Ứng dụng Web Demo: Đã phát triển thành công giao diện web cho phép người dùng nhập 12 thông số lâm sàng và nhận kết quả dự đoán đồng thời từ ba mô hình (SVM, Random Forest, Decision Tree) kèm theo confidence score.

4.2. Hạn chế còn tồn đọng

- Tính mất cân bằng dữ liệu (Data Imbalance): Mặc dù đã cố gắng xử lý, nhưng sự mất cân bằng nghiêm trọng của biến mục tiêu (cancer_type_detailed) và biến Tumor Stage (với Giai đoạn 0, 3, 4 quá khan hiếm) vẫn gây thiên lệch mô hình và giảm hiệu suất trên các lớp thiểu số (ví dụ: F1-score của Random Forest trên Lớp 1, 2, 4 thấp hơn Lớp 0 và 3).
- Phạm vi Chức năng Hạn chế: Hệ thống hiện chỉ hoạt động dưới dạng Web chạy cục bộ và chỉ tập trung vào chức năng dự đoán cơ bản. Các chức năng nâng cao như quản lý hồ sơ bệnh nhân, kết nối API y tế, hay các phân tích chuyên sâu (trực quan hóa dữ liệu dạng đồ thị) chưa được triển khai.
- Khó khăn trong Diễn giải Mô hình (Explainability): Các mô hình Ensemble như Random Forest và SVM cung cấp độ chính xác cao, nhưng lại khó truy ngược lại "đường đi phân loại" chi tiết cho từng mẫu dữ liệu mới, gây khó khăn cho việc giải thích lâm sàng sâu rộng

4.3. Hướng phát triển trong tương lai

Nhằm đưa đề tài lên một tầm cao mới, hướng phát triển trong tương lai sẽ tập trung vào việc gia tăng độ robust của mô hình và mở rộng phạm vi ứng dụng thành một hệ thống thông minh, tích hợp:

- Tăng cường Hiệu suất Mô hình bằng Data Augmentation và Sampling: Chúng em sẽ tập trung xử lý triệt để vấn đề mất cân bằng dữ liệu bằng cách triển khai các kỹ thuật tiên tiến như SMOTENC (phù hợp với tập dữ liệu hỗn hợp Categorical/Numerical) hoặc chiến lược lai ghép Oversampling/Undersampling. Mục tiêu là tối ưu hóa chỉ số Recall và F1-score, đặc biệt đối với các lớp thiểu số có giá trị tiên lượng cao.

- Mở rộng Năng lực Dự đoán với Mô hình Chuyên sâu: Nghiên cứu sẽ mở rộng sang các thuật toán học máy hiệu năng cao như XGBoost và Neural Networks (có thể là một kiến trúc Deep Learning cơ bản) để cải thiện độ chính xác tổng thể. Đặc biệt, chúng em sẽ phát triển mô hình phân tích sinh tồn (Survival Analysis Model), sử dụng các thuật toán như Cox Proportional Hazards, nhằm mở rộng năng lực dự đoán từ phân loại bệnh sang tiên lượng thời gian sống sót của bệnh nhân.
- Tích hợp Phân tích Đa Phương thức (Multimodal Analytics): Phát triển khả năng xử lý dữ liệu hình ảnh bằng cách tích hợp mô hình Computer Vision (CV). Mô hình này sẽ được huấn luyện để đọc và phân tích ảnh X-quang tuyến vú hoặc ảnh mô học (nếu có), sau đó kết hợp các features hình ảnh với dữ liệu lâm sàng/gene (dữ liệu có cấu trúc) để tạo ra một chẩn đoán hợp nhất và đáng tin cậy hơn.
- Phát triển Ứng dụng AI tương tác: Ứng dụng công nghệ Mô hình Ngôn ngữ Lớn (LLMs) để xây dựng Chatbot hỗ trợ thông minh. Chatbot này sẽ được fine-tune để tương tác với người dùng (bác sĩ, bệnh nhân, nhà nghiên cứu) một cách tự nhiên, cho phép tra cứu thông tin chuyên sâu từ cơ sở tri thức y khoa (Knowledge Base) về ung thư vú.
- Triển khai Hệ thống Hỗ trợ Ra quyết định Lâm sàng (CDSS): Hệ thống sẽ được nâng cấp thành một CDSS hoàn chỉnh, tích hợp các công cụ trực quan hóa dữ liệu chuyên sâu (như biểu đồ Kaplan-Meier, biểu đồ phân tích nguy cơ tái phát) để cung cấp hỗ trợ đắc lực và trực quan hóa dữ liệu cho quá trình ra quyết định của bác sĩ.
- Tăng cường Tính minh bạch (XAI) cho Mô hình Ensemble: Để giải quyết vấn đề Black-box, chúng em sẽ triển khai các kỹ thuật Explainable AI (XAI) như SHAP hoặc LIME trên các mô hình Ensemble (Random Forest, XGBoost). Điều này giúp làm rõ vai trò đóng góp của từng đặc trưng đầu vào vào quyết định cuối cùng, tạo thuận lợi cho việc kiểm chứng và diễn giải của chuyên gia y tế.
- Kiểm chứng và Triển khai Thử nghiệm Lâm sàng (Pilot Study): Hợp tác với các cơ sở y tế để thử nghiệm lâm sàng. Phản hồi thực tế từ bác sĩ và bệnh nhân sẽ giúp kiểm chứng hiệu quả, tinh chỉnh mô hình và giao diện trước khi triển khai rộng rãi.

CHƯƠNG 4: TÀI LIỆU THAM KHẢO

[1] J. Han, M. Kamber và J. Pei, Data Mining: Concepts and Techniques, 3rd, Biên tập viên, Waltham: Morgan Kaufmann, 2011.

[2] Giáo trình Khai phá dữ liệu.

[3] Giáo trình môn Học máy cơ bản.

[4] H. C. Trung, "Giới thiệu về Support Vector Machine (SVM)," 20 8 2020. [Trực tuyến]. Available: https://surl.lt/datnxl. [Đã truy cập 1 10 2025].

[5] T. Nguyễn., "Random Forest algorithm¶," [Trực tuyến]. Available: https://surl.li/bsjcdu. [Đã truy cập 2 10 2025].

[6] N. S. Chauhan, "Decision Tree Algorithm, Explained," 9 2 2022. [Trực tuyến]. Available: https://www.kdnuggets.com/2020/01/decision-tree-algorithm-explained.html. [Đã truy cập 2 10 2025].

[7] G. Menon, F. M. Alkabban và T. Ferguson., "Breast cancer," 14 8 2025. [Trực tuyến]. Available: https://www.ncbi.nlm.nih.gov/books/NBK482286/. [Đã truy cập 1 10 2025].

[8] "Ung thư vú: Dấu hiệu, nguyên nhân, cách phòng tránh và điều trị," 9 9 2025. [Trực tuyến]. Available: https://surl.lu/eqvwrf. [Đã truy cập 1 10 2025].

[9] K. &. A. Gallatin, Machine learning with Python cookbook, 2nd, Biên tập viên, O'Reilly Media, 2023.

[10] A. Géron, "Hands-On Machine Learning with Scikit-Learn, Keras, & TensorFlow," trong *2019*, Sebastopol, O'Reilly Media.

[11] S. &. M. V. Raschka, "Python Machine Learning," 2024.