|  |  |
| --- | --- |
| **bỘ GIÁO DỤC VÀ ĐÀO TẠO** | **BỘ NÔNG NGHIỆP VÀ PTNT** |

TRƯỜNG ĐẠI HỌC THỦY LỢI



NGUYỄN ĐỨC HẢI

TÊN ĐỀ TÀI: PHÂN LOẠI NGUY CƠ UNG THƯ CỔ TỬ CUng BẰNG PHƯƠNG PHÁP HỌC MÁY.

ĐỒ ÁN TỐT NGHIỆP

HÀ NỘI, NĂM 2023

|  |  |
| --- | --- |
| **bỘ GIÁO DỤC VÀ ĐÀO TẠO** | **BỘ NÔNG NGHIỆP VÀ PTNT** |

TRƯỜNG ĐẠI HỌC THỦY LỢI

NGUYỄN ĐỨC HẢI

TÊN ĐỀ TÀI: PHÂN LOẠI NGUY CƠ UNG THƯ CỔ TỬ CUng BẰNG PHƯƠNG PHÁP HỌC MÁY.

|  |  |
| --- | --- |
| Ngành : Công nghệ thông tin |  |
| Mã số:7480201 |  |

|  |  |
| --- | --- |
| NGƯỜI HƯỚNG DẪN | 1. THS. Nguyễn Đắc Hiếu |
|  |  |

HÀ NỘI, NĂM 2023

**GÁY BÌA ĐỒ ÁN TỐT NGHIỆP, KHÓA LUẬN TỐT NGHIỆP**

**NGUYỄN ĐỨC HẢI ĐỒ ÁN TỐT NGHIỆP HÀ NỘI, NĂM 2023**

LỜI CAM ĐOAN

Tác giả xin cam đoan đây là Đồ án tốt nghiệp của bản thân tác giả. Các kết quả trong Đồ án tốt nghiệp này là trung thực, và không sao chép từ bất kỳ một nguồn nào và dưới bất kỳ hình thức nào.Việc tham khảo các nguồn tài liệu (nếu có) đã được thực hiện trích dẫn và ghi nguồn tài liệu tham khảo đúng quy định.

|  |  |
| --- | --- |
|  | **Tác giả ĐATN**  *Chữ ký*  Hải  **Nguyễn Đức Hải** |

LỜI CÁM ƠN

Tác giả xin trân trọng cám ơn THS. Nguyễn Đắc Hiếu và các thầy cô tại phòng Đào tạo đóng góp ý kiến cho việc soạn thảo tài liệu Hướng dẫn Trình bày ĐATN này.

MỤC LỤC

[DANH MỤC CÁC HÌNH ẢNH iv](#_Toc149598135)

[DANH MỤC BẢNG BIỂU v](#_Toc149598136)

[DANH MỤC CÁC TỪ VIẾT TẮT VÀ GIẢI THÍCH CÁC THUẬT NGỮ vi](#_Toc149598137)

[CHƯƠNG 1 GIỚI THIỆU BÀI TOÁN, GIỚI THIỆU TỔNG QUAN 1](#_Toc149598138)

[1.1 Lý do chọn đề tài 1](#_Toc149598139)

[1.2 Mục tiêu, phạm vi 1](#_Toc149598140)

[1.3 Đối tượng, phạm vi nghiên cứu 2](#_Toc149598141)

[1.4 Khu vực nghiên cứu 3](#_Toc149598142)

[CHƯƠNG 2 CÁC KĨ THUẬT DÙNG TRONG BÀI TOÁN 3](#_Toc149598143)

[2.1 Tiếp cận cơ sở lý thuyết 3](#_Toc149598144)

[2.2 Các kĩ thuật dùng cho bài toán 3](#_Toc149598145)

[2.3 Các phương pháp đánh giá mô hình 3](#_Toc149598146)

[CHƯƠNG 3 ỨNG DỤNG PHƯƠNG PHÁP VÀ XÂY DỰNG MÔ HÌNH 3](#_Toc149598147)

[3.1 Phân tích chi tiết bài toán 3](#_Toc149598148)

[3.2 Các công cụ, thư viện dùng để xây dựng 3](#_Toc149598149)

[3.3 Xây dựng mô hình 3](#_Toc149598150)

[CHƯƠNG 4 KẾT QUẢ VÀ ĐÁNH GIÁ MÔ HÌNH 3](#_Toc149598151)

[4.1 Kết quả 3](#_Toc149598152)

[4.2 Đánh giá mô hình 3](#_Toc149598153)

[CHƯƠNG 5 KẾT LUẬN 3](#_Toc149598154)

[TÀI LIỆU THAM KHẢO 4](#_Toc149598155)

[PHỤ LỤC 5](#_Toc149598156)

DANH MỤC CÁC HÌNH ẢNH

[Hình 2.1 Biểu tượng (logo) của trường Đại học Thủy lợi 6](file:///F:\BOX\Box%20Sync\HD_DATN_LVThS_LATS\Huong_dan_Trinh_bay_LATS\Huong%20dan%20trinh%20bay%20DATN1.docx#_Toc414436222)

DANH MỤC BẢNG BIỂU

[Bảng 2.1 Cách định dạng lề giấy 3](file:///D:\Box\My%20Box%20Files\Quan%20Ly%20NCS\Huong%20dan%20trinh%20bay%20LATS_updated%2001-2013\Huong%20dan%20trinh%20bay%20LATS.docx#_Toc405384698)

[Bảng 2.2 Tóm tắt các kiểu định dạng (style) cho các đề mục 7](file:///D:\Box\My%20Box%20Files\Quan%20Ly%20NCS\Huong%20dan%20trinh%20bay%20LATS_updated%2001-2013\Huong%20dan%20trinh%20bay%20LATS.docx#_Toc405384699)

DANH MỤC CÁC TỪ VIẾT TẮT VÀ GIẢI THÍCH CÁC THUẬT NGỮ

**ĐATN** Đồ án tốt nghiệp

# GIỚI THIỆU BÀI TOÁN, GIỚI THIỆU TỔNG QUAN

## Lý do chọn đề tài

Ung thư cổ tử cung là một vấn đề quan trọng trong lĩnh vực y tế. Đặc biệt là nó có liên quan đến tỷ lệ tử vong ở phụ nữ trên toàn thế giới. Ung thư cổ tử cung là một khối u ác tính xảy ra khi các tế bào mô bao phủ cổ tử cung (còn gọi là mô niêm mạc) bắt đầu phát triển và sinh sản không kiểm soát được mà không tuân theo cơ chế phân chia tế bào thích hợp. Sự phát triển không kiểm soát của tế bào này dẫn đến hình thành khối u ác tính trên cổ tử cung, và nếu không được phát hiện kịp thời, nó có thể lan ra các cơ quan và mô xung quanh. Điều này có thể gây ra nhiều biến chứng và nguy cơ nghiêm trọng cho sức khỏe của phụ nữ.

Những năm gần đây, việc ứng dụng trí tuệ nhân tạo vào các bài toán dự báo, phân loại đã được nghiên cứu và mở rộng. Đặc biệt, các mô hình học máy, học sâu đang phát triển mạnh và được ứng dụng trong nhiều lĩnh vực như: y tế, tài chính, người máy, …, bao gồm cả phân loại nguy cơ ung thư cổ tử cung. Trong lĩnh vực y tế đã có một số chuyên nghành ứng dụng AI chuẩn đoán hình ảnh nội soi… và đạt được những thành quả bước đầu rất khả quan. Với khả năng tự cập nhật trong quá trình huấn luyện, các mô hình học máy, học sâu có thể khắc phục được những hạn chế trong mô hình phân loại. Bằng cách xử lý dữ liệu lớn và phức tạp về thông tin tiền sử bệnh nhân, các kết quả xét nghiệm, dấu hiệu lâm sàng và các yếu tố nguy cơ khác, mô hình học máy có thể phân loại nguy cơ ung thư cổ tử cung với độ chính xác cao. Từ đó sẽ giúp cải thiện khả năng phát hiện sớm và đánh giá rủi ro, tiết kiệm được nhiều thời gian và chi phí. Mang lại nhiều lợi ích quan trọng trong lĩnh vực y tế và chăm sóc sức khỏe cho phụ nữ trên toàn thế giới.

Từ những vấn đề thực tế, cùng với sự phát triển của công nghệ hiện đại và ứng dụng các phương pháp học máy. Vì vậy, em lựa chọn đề tài: “**Phân loại nguy cơ ung thư cổ tử cung bằng phương pháp học máy**”.

## Mục tiêu, phạm vi

* Mục tiêu

+Chuẩn đoán và dự đoán: phát triển mô hình học máy để chuẩn đoán ung thư cổ tử cung và dự đoán nguy cơ mắc bệnh.

* Phạm vi

+Có thể mở rộng nghiên cứu đối với nhiều quốc gia, tùy thuộc vào khả năng thu thập dữ liệu và mức độ phổ biến của bệnh.

+Dữ liệu: thu thập dữ liệu về lối sống chi tiết, về các yếu tố nguy cơ, lịch sử bệnh, và kết quả kiểm tra xét nghiệm của những người phụ nữ.

+Thuật toán: Áp dụng và đánh giá các thuật toán học máy như SVM, Decision Trees để xây dựng mô hình.

+Đánh giá: Đánh giá hiệu suất của các mô hình trên qua các độ đo như: Precision, Recall, F1 Score.

## Đối tượng, phạm vi nghiên cứu

* Đối tượng nghiên cứu

+Tiêu chuẩn lựa chọn:

+Giới tính: Chỉ chọn phụ nữ vì ung thư cổ tử cung liên quan chặt chẽ đến hệ sinh sản ở

nữ giới.

+Độ tuổi: Chọn phụ nữ ở các độ tuổi có nguy cơ cao về ung thư cổ tử cung. Điều này

có thể là phụ nữ trưởng thành có thể bắt đầu quan hệ tình dục và có nguy cơ nhiễm HPV,

một trong những yếu tố gây ra ung thư cổ tử cung.

+Số lượng đối tác tình dục: Những người có số lượng đối tác tình dục nhiều có thể có

nguy cơ cao hơn.

+Có thông tin y tế đầy đủ.

+Tiêu chuẩn loại trừ:

+Không phải là phụ nữ: Nếu là nam giới thì loại trừ hết vì vấn đề nghiên cứu này chỉ

liên quan đến phụ nữ.

+Dữ liệu thiếu: Nếu thiếu các dữ liệu thông tin quan trọng như tuổi, số lượng đối tác

tình dục, hoặc các chỉ số liên quan đến ý tế, xét nghiệm, có thể loại trừ.

* Phạm vi nghiên cứu

+Sử dụng dữ liệu y tế và lối sống chi tiết để xây dựng mô hình.

+Dữ liệu có thể thu thập được từ các cơ sở y tế, cuộc khảo sát và các nguồn thông tin y

tế khác.

## Khu vực nghiên cứu

Khía cạnh địa lý: Nghiên cứu sẽ chọn một số khu vực địa lý quan trọng có tình trạng

ung thư cổ tử cung đáng chú ý. Các khu vực này có thể được xác định dựa trên mức độ

phổ biến của bệnh, độ phức tạp của hệ thống ý tế, và các yếu tố xã hội địa phương. Điều

này giúp tập trung nghiên cứu vào những vùng có nhu cầu cấp bách về việc sớm phát

hiện và quản lý ung thư cổ tử cung.

# CÁC KĨ THUẬT DÙNG TRONG BÀI TOÁN

## Tiếp cận cơ sở lý thuyết

* Cây quyết định (Decision tree): là một mô hình có giám sát (supervised learning), có

thể được áp dụng cho cả bài toán phân lớp (classification) và hồi quy (regression).

Các thuộc tính của cây quyết định là: Thuộc tính rời rạc và không có thứ tự và thuộc tính liên tục.

* Cây quyết định:

+ Có cấu trúc cây giống như biểu đồ luồng.

+ Mỗi nút trong thể hiện một sự kiểm tra trên một thuộc tính.

+ Mỗi nhánh đại diện cho một kết quả của sự kiểm tra.

+ Các nút lá đại diện cho các nhãn lớp hoặc phân khối lớp

* Thuật toán cơ bản (thuật toán tham lam)

+ Cây quyết định được xây dựng theo cách chia để trị từ trên xuống (top-down).

+ Tại các vị trí khởi tạo, tất cả các mẫu thuộc nút gốc.

+ Các thuộc tính được phân loại (nếu giá trị của thuộc tính là liên tục, thì phải được rời rạc hóa trước).

+ Các mẫu được phân chia đệ quy dựa vào các thuộc tính được chọn.

+ Thuộc tính kiểm tra được lựa chọn vào kinh nghiệm hoặc độ đo thống kê.

* Điều kiện dừng phân chia

+ Tất cả các mẫu của nút xem xét thuộc tính cùng một lớp.

+ Không có thuộc tính nào để phân chia – biểu quyết đa số được sử dụng để gán nhãn phân loại cho lá.

+ Không còn mẫu nào.

* Cây quyết định bao gồm một số thuật toán để xây dựng cây và thực hiện các quyết

định:

+ ID3 (Iterative Dichotomiser 3): Thuật toán này sử dụng độ do Entropy để tìm cách chia dữ liệu hiệu quả nhất. Nó chia cây theo cách tốt nhất để giảm độ không chắc chắn trong quyết định.

+ CART (Classification and Regression Trees): Sử dụng Gini Index để đo lường độ không thuần khiết của dữ liệu. Nó giảm Gini Index bằng cách chọn cách chia tốt nhất.

* Support Vector Machine (SVM): là một thuật toán học máy thuộc lớp các thuật toán

phân loại và hồi quy. Mục tiêu của SVM là tạo ra một đường ranh giới (decision boundary) sao cho khoảng cách từ mỗi điểm dữ liệu đến đường ranh giới là lớn nhất. SVM thường được sử dụng trong bài toán phân loại nhưng cũng có thể áp dụng vào bài toán hồi quy.

* Thuật toán SVM thường hiệu quả trong việc xử lý các dữ liệu tuyến tính và phi tuyến

tính và nó đã được áp dụng trong nhiều lĩnh vực như thị giác máy tính, dự đoán chuỗi thời gian, phân loại y tế và phân loại văn bản.

## Các kĩ thuật dùng cho bài toán

* Thu thập dữ liệu: Gom dữ liệu của các bệnh viện, cơ sở y tế hoặc các nguồn khác liên

quan đến thông tin người bệnh như Kaggle.

* Tiền xử lý dữ liệu: Loại bỏ dữ liệu thiếu, chuyển đổi các biến không phải số thành

dạng số, xử lý nhiễu, và chuẩn hóa dữ liệu để tạo điều kiện cho mô hình học máy.

* Chọn đặc trưng: Lựa chọn các trường quan trọng để xây dựng mô hình. Các trường

này có thể liên quan đến tuổi, số lần mang thai, số lần hút thuốc và các thông tin về y tế.

* Chia dữ liệu: Phân chia dữ liệu thành tập huấn luyện (training) và tập kiểm thử

(testing) để đào tạo đánh giá mô hình.

* Chọn mô hình: Sử dụng các thuật toán học máy như SVM, cây quyết định để xây

dựng mô hình.

* Đào tạo mô hình: Mô hình được đào tạo trên tập huấn luyện để học cách dự đoán

nguy cơ ung thư cổ tử cung dựa trên các đặc trưng đã chọn.

* Đánh giá mô hình: Sử dụng tập kiểm thử để đánh giá hiệu suất của mô hình. Các độ

đo như độ chính xác, độ nhạy (recall), độ đặc hiệu (precision) và F1-Score.

* Điều chỉnh tham số: Sau khi đã huấn luyện và đánh giá được mô hình học máy của

mình, chúng ta có thể điều chỉnh lại tham số nếu cần để tăng độ chính xác cho hệ thống.

* Đưa ra dự đoán: Cuối cùng, chúng ta sẽ sử dụng mô hình mà chúng ta đã tạo để phân

tích thông tin và đưa ra dự đoán một cách chính xác.

## Các phương pháp đánh giá mô hình

* Precision và Recall: Precision đo lường tỷ lệ dự đoán Positive chính xác, Recall đo

lường tỷ lệ Positive thực sự dự đoán đúng.

* F1-score: Trung bình điều hòa giữa Precision và Recall.
* Độ chính xác: Đo lường tỷ lệ dự đoán chính xác trên toàn bộ tập kiểm thử.

# ỨNG DỤNG PHƯƠNG PHÁP VÀ XÂY DỰNG MÔ HÌNH

## Phân tích chi tiết bài toán

## Các công cụ, thư viện dùng để xây dựng

## Xây dựng mô hình

# KẾT QUẢ VÀ ĐÁNH GIÁ MÔ HÌNH

## Kết quả

## Đánh giá mô hình

# KẾT LUẬN

Insert Caption

Bảng được chèn vào

TÀI LIỆU THAM KHẢO

|  |  |
| --- | --- |
| [1] | [DOANVANNGOC-LA.pdf (hmu.edu.vn)](https://sdh.hmu.edu.vn/images/DOANVANNGOC-LA.pdf) |

PHỤ LỤC