**BỘ GIÁO DỤC VÀ ĐÀO TẠO**

**TRƯỜNG ĐẠI HỌC CẦN THƠ**

**TRƯỜNG CÔNG NGHỆ THÔNG TIN & TRUYỀN THÔNG**





**ĐỒ ÁN**

**HỌC PHẦN MÁY HỌC ỨNG DỤNG**

**Đề tài**

**XÂY DỰNG MÔ HÌNH HỌC MÁY DỰ ĐOÁN**

**UNG THƯ CỔ TỬ CUNG**

**Nhóm sinh viên thực hiện:**

**1. Nguyễn Duy Hưng – B2207525**

**2. Đặng Duy Khoa – B2207532**

**3. Trương Công Minh – B2207543**

#### Cần Thơ, Tháng 3/2025

**BỘ GIÁO DỤC VÀ ĐÀO TẠO**

**TRƯỜNG ĐẠI HỌC CẦN THƠ**

**TRƯỜNG CÔNG NGHỆ THÔNG TIN & TRUYỀN THÔNG**





**ĐỒ ÁN**

**HỌC PHẦN MÁY HỌC ỨNG DỤNG**

**Đề tài**

**XÂY DỰNG MÔ HÌNH HỌC MÁY DỰ ĐOÁN**

**UNG THƯ CỔ TỬ CUNG**

**Giảng viên hướng dẫn: Nhóm sinh viên thực hiện:**

**TS. Lưu Tiến Đạo** 

#### Cần Thơ, Tháng 3/2025

# NHẬN XÉT CỦA GIẢNG VIÊN

---------------------------------------------------------------------------------

---------------------------------------------------------------------------------

-------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------

-----------------------------------------------------------------------------------------------------------------------------------------------------------------

---------------------------------------------------------------------------------

-----------------------------------------------------------------------------------------------------------------------------------------------------------------

---------------------------------------------------------------------------------

*Cần Thơ, ngày tháng năm*

(Ký và ghi rõ họ tên)

# MỤC LỤC

[**NHẬN XÉT CỦA GIẢNG VIÊN 3**](#_heading=h.x3xn56tbbtwp)

[**MỤC LỤC 4**](#_heading=h.xyv7nu7m79ri)

[**PHÂN CÔNG CÔNG VIỆC 5**](#_heading=h.g41mgtnfm44j)

[**PHẦN NỘI DUNG 6**](#_heading=h.wiknm07mejtg)

[1.](#_heading=h.g7x6yad20qg8) Mô tả bài toán 6

[2.](#_heading=h.q8wk6l3gce0b) Mô tả dữ liệu, ý nghĩa của dữ liệu 6

[2.1.](#_heading=h.i4ji1bkaxead) Thông tin nhân khẩu học: 7

[2.2.](#_heading=h.ch8hbsjdwujd) Thói quen sinh hoạt: 7

[2.3.](#_heading=h.as5cw7ejlglm) Tiền sử bệnh tật: 7

[2.4.](#_heading=h.ope5amtuni4r) Kết quả xét nghiệm: 7

[3.](#_heading=h.p6aqvrdjg7tk) Phân tích dữ liệu và lựa chọn mô hình 8

[3.1.](#_heading=h.28d99r23kjjk) Tiền xử lý dữ liệu: 8

[3.2.](#_heading=h.xqq5npb1t2rp) Phân tích tương quan và phân phối: 8

[3.3.](#_heading=h.t86h3a4valzu) Lựa chọn mô hình: 8

[3.4.](#_heading=h.eqy1j6zdxl8m) Đánh giá mô hình: 9

[4.](#_heading=h.1yznalvve3tj) Cấu hình máy tính 9

[5.](#_heading=h.r6bjt1ugmwk5) Huấn luyện mô hình 9

[5.1.](#_heading=h.7affngg1ot88) Huấn luyện mô hình mạng nơ-ron nhân tạo (ANN): 9

[5.2.](#_heading=h.diu95m9t69p8) Huấn luyện mô hình Random Forest: 10

[5.3.](#_heading=h.aega634tenqa) Đánh giá hiệu suất huấn luyện: 10

[6.](#_heading=h.uq3dg4ljx96) Đánh giá mô hình 10

[6.1.](#_heading=h.ea3uctp2g2nk) Đánh giá mô hình phân lớp 10

[6.2.](#_heading=h.9k93wagtcrcr) Nhận xét kết quả thực nghiệm 11

[**PHẦN KẾT LUẬN 13**](#_heading=h.4ox4cik4sosq)

[1.](#_heading=h.xn8ygmk9728w) Kết quả đạt được 13

[2.](#_heading=h.g7ebn59boa5w) Hướng phát triển 13

[**TÀI LIỆU THAM KHẢO 14**](#_heading=h.gfgdkvyh8rxi)

# PHÂN CÔNG CÔNG VIỆC

| Công việc | Người thực hiện | Mức độ hoàn thành |
| --- | --- | --- |
| Tiền xử lí dữ liệu | Đặng Duy Khoa | 100% |
| Huấn luyện mô hình RandomForest | Nguyễn Duy Hưng | 100% |
| Huấn luyện mô hình ANN | Trương Công Minh | 100% |
| Huấn luyện mô hình KNN | Nguyễn Duy Hưng | 100% |
| Đánh giá mô hình | Trương Công Minh | 100% |
| Soạn thảo file báo cáo | Đặng Duy Khoa | 100% |
| Soạn thảo file thuyết trình | Nguyễn Duy Hưng | 100% |
| Xây dựng web demo ứng dụng | Trương Công Minh | 100% |

# PHẦN NỘI DUNG

## Mô tả bài toán

Ung thư cổ tử cung (UTCTC) là một trong những nguyên nhân gây tử vong hàng đầu ở phụ nữ trên toàn thế giới, đặc biệt là ở các nước đang phát triển . Việc phát hiện sớm và can thiệp kịp thời có vai trò then chốt trong việc giảm thiểu tỷ lệ tử vong do UTCTC. Tuy nhiên, các phương pháp sàng lọc truyền thống như xét nghiệm Pap và xét nghiệm HPV đòi hỏi nguồn lực lớn và có thể không khả thi ở những khu vực có nguồn lực hạn chế . Do đó, việc phát triển các phương pháp sàng lọc hiệu quả và tiết kiệm chi phí là vô cùng cần thiết.

Trong bối cảnh đó, việc ứng dụng học máy (machine learning) để dự đoán nguy cơ UTCTC dựa trên các yếu tố nguy cơ đã thu hút sự quan tâm của nhiều nhà nghiên cứu. Các nghiên cứu trước đây đã chỉ ra rằng các yếu tố như tuổi, số lần mang thai, tiền sử mắc các bệnh lây truyền qua đường tình dục (STDs), và kết quả xét nghiệm tế bào học có thể được sử dụng để xây dựng mô hình dự đoán nguy cơ UTCTC .

Tuy nhiên, vẫn còn nhiều thách thức trong việc xây dựng mô hình dự đoán UTCTC hiệu quả. Thứ nhất, dữ liệu y tế thường không cân bằng, với số lượng bệnh nhân mắc UTCTC ít hơn nhiều so với số lượng bệnh nhân khỏe mạnh. Điều này có thể dẫn đến việc mô hình bị thiên vị và không thể dự đoán chính xác nguy cơ UTCTC ở nhóm bệnh nhân thiểu số. Thứ hai, các yếu tố nguy cơ của UTCTC có thể tương tác phức tạp với nhau, và việc mô hình hóa các tương tác này đòi hỏi các kỹ thuật học máy tiên tiến.

Trong báo cáo này, chúng tôi đề xuất xây dựng một mô hình học máy để dự đoán nguy cơ UTCTC dựa trên một tập dữ liệu toàn diện về các yếu tố nguy cơ. Chúng tôi sẽ sử dụng các kỹ thuật học máy tiên tiến như mạng nơ-ron nhân tạo (artificial neural network) và các phương pháp xử lý dữ liệu không cân bằng để giải quyết các thách thức đã nêu. Mục tiêu của chúng tôi là xây dựng một mô hình có độ chính xác cao, có thể được sử dụng để sàng lọc UTCTC một cách hiệu quả và tiết kiệm chi phí.

## Mô tả dữ liệu, ý nghĩa của dữ liệu

Báo cáo này sử dụng bộ dữ liệu "Cervical Cancer Risk Classification" từ UC Irvine, một tập hợp dữ liệu toàn diện về các yếu tố nguy cơ liên quan đến ung thư cổ tử cung (UTCTC). Bộ dữ liệu này bao gồm thông tin từ 858 bệnh nhân, với 36 đặc trưng đầu vào và một biến mục tiêu nhị phân (Biopsy) biểu thị sự hiện diện của UTCTC.

Các đặc trưng đầu vào bao gồm:

### Thông tin nhân khẩu học:

* Tuổi (Age): Tuổi của bệnh nhân.
* Số lượng bạn tình (Number of sexual partners): Số lượng bạn tình của bệnh nhân.
* Tuổi lần quan hệ tình dục đầu tiên (First sexual intercourse): Tuổi của bệnh nhân khi lần đầu quan hệ tình dục.
* Số lần mang thai (Num of pregnancies): Số lần mang thai của bệnh nhân.

*A table with numbers and a number of people

AI-generated content may be incorrect.Table 1:Bảng thông tin về nhân khẩu học*

### Thói quen sinh hoạt:

* + Hút thuốc (Smokes): Biến nhị phân biểu thị bệnh nhân có hút thuốc hay không.
  + Số năm hút thuốc (Smokes (years)): Số năm bệnh nhân hút thuốc.
  + Số lượng thuốc lá hút mỗi năm (Smokes (packs/year)): Số lượng thuốc lá bệnh nhân hút mỗi năm.
  + Sử dụng thuốc tránh thai (Hormonal Contraceptives): Biến nhị phân biểu thị bệnh nhân có sử dụng thuốc tránh thai hay không.
  + Số năm sử dụng thuốc tránh thai (Hormonal Contraceptives (years)): Số năm bệnh nhân sử dụng thuốc tránh thai.
  + Sử dụng vòng tránh thai (IUD): Biến nhị phân biểu thị bệnh nhân có sử dụng vòng tránh thai hay không.
  + Số năm sử dụng vòng tránh thai (IUD (years)): Số năm bệnh nhân sử dụng vòng tránh thai.

*A table with numbers and numbers

AI-generated content may be incorrect.Table 2:Thói quen sinh hoạt*

### Tiền sử bệnh tật:

* Các bệnh lây truyền qua đường tình dục (STDs): Biến nhị phân biểu thị bệnh nhân có mắc các bệnh lây truyền qua đường tình dục hay không.
* Các bệnh lây truyền qua đường tình dục cụ thể(STDs:cervical condylomatosis, STDs:AIDS...)
* Tiền sử ung thư(Dx:Cancer), CIN(Dx:CIN), HPV(Dx:HPV), Dx.

A screenshot of a computer

AI-generated content may be incorrect.

*Table 3:Tiền sử bệnh tật*

### Kết quả xét nghiệm:

* Hinselmann: Kết quả xét nghiệm Hinselmann.
* Schiller: Kết quả xét nghiệm Schiller.
* Citology: Kết quả xét nghiệm tế bào học.
* Biopsy: Kết quả xét nghiệm sinh thiết.

A table with numbers and letters

AI-generated content may be incorrect.

*Table 4:Kết quả xét nghiệm*

Ý nghĩa của các đặc trưng này là cung cấp thông tin về các yếu tố nguy cơ có thể ảnh hưởng đến sự phát triển của UTCTC. Các nghiên cứu trước đây đã chỉ ra rằng các yếu tố như nhiễm HPV, hút thuốc, và tiền sử mắc các bệnh lây truyền qua đường tình dục có liên quan đến tăng nguy cơ UTCTC .

Trong báo cáo này, chúng tôi sẽ sử dụng bộ dữ liệu này để xây dựng và đánh giá các mô hình học máy có khả năng dự đoán nguy cơ UTCTC. Việc phân tích và khai thác thông tin từ dữ liệu này sẽ giúp hiểu rõ hơn về các yếu tố nguy cơ và phát triển các công cụ sàng lọc hiệu quả.

## Phân tích dữ liệu và lựa chọn mô hình

Trước khi tiến hành xây dựng mô hình, chúng tôi đã thực hiện một loạt các phân tích dữ liệu để hiểu rõ hơn về cấu trúc và đặc điểm của bộ dữ liệu. Phân tích này bao gồm các bước sau:

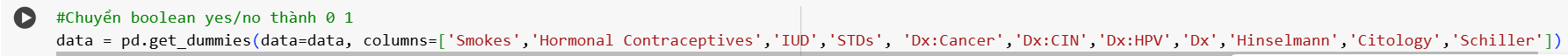
### Tiền xử lý dữ liệu:

* Xử lý giá trị thiếu: Dữ liệu y tế thường chứa nhiều giá trị thiếu, điều này có thể ảnh hưởng đến hiệu suất của mô hình. Chúng tôi đã xử lý giá trị thiếu bằng cách thay thế chúng bằng giá trị trung vị cho các biến liên tục và 0/1 cho các biến rời rạc, tùy thuộc vào phân phối và bản chất của từng biến.

A screenshot of a computer code

AI-generated content may be incorrect.

* Mã hóa biến phân loại: Các biến phân loại như "Smokes", "Hormonal Contraceptives", và "IUD" đã được mã hóa bằng phương pháp one-hot encoding để chuyển đổi chúng thành dạng số, phù hợp cho việc huấn luyện mô hình.



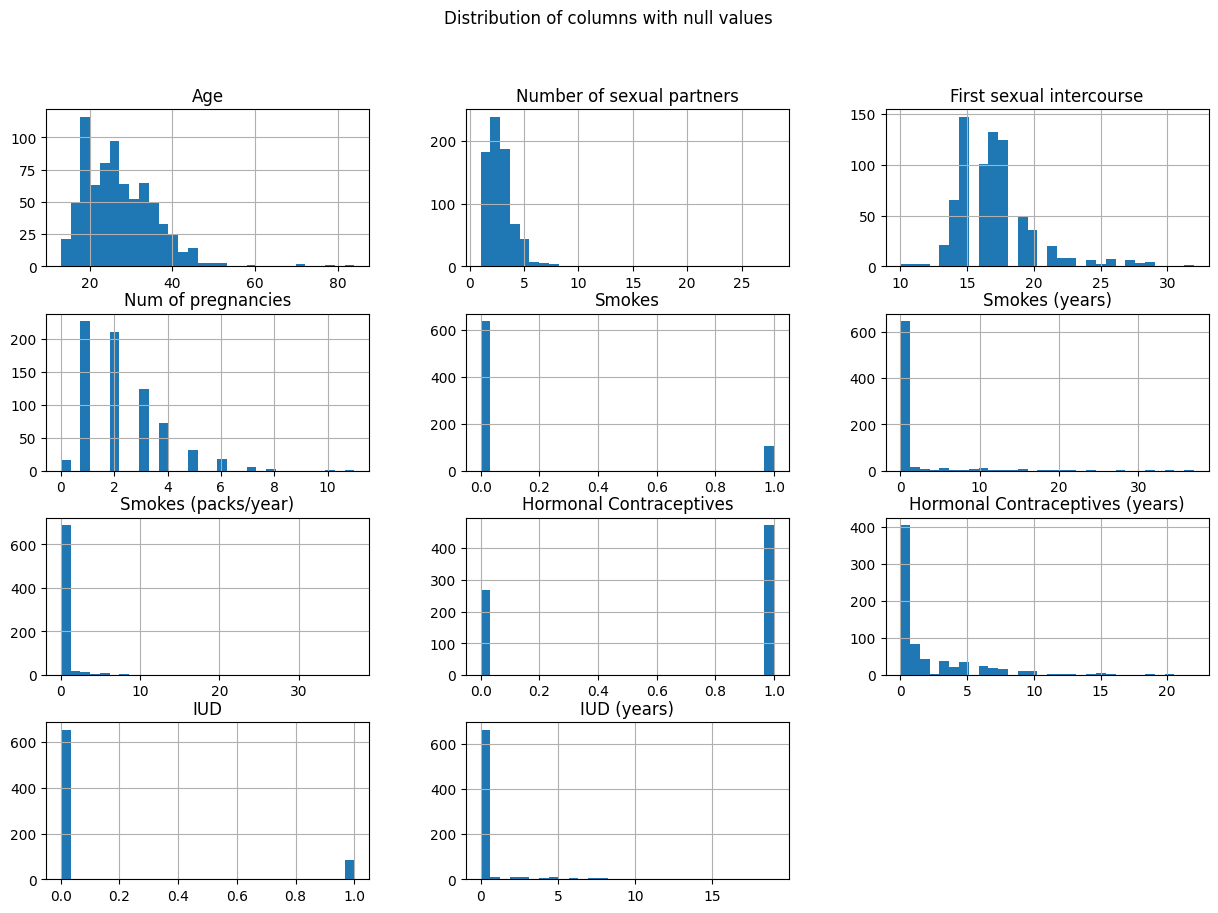
* Chuẩn hóa dữ liệu: Để giảm ảnh hưởng của ngoại lai và đảm bảo sự ổn định của mô hình, chúng tôi đã sử dụng StandardScaler để chuẩn hóa dữ liệu.

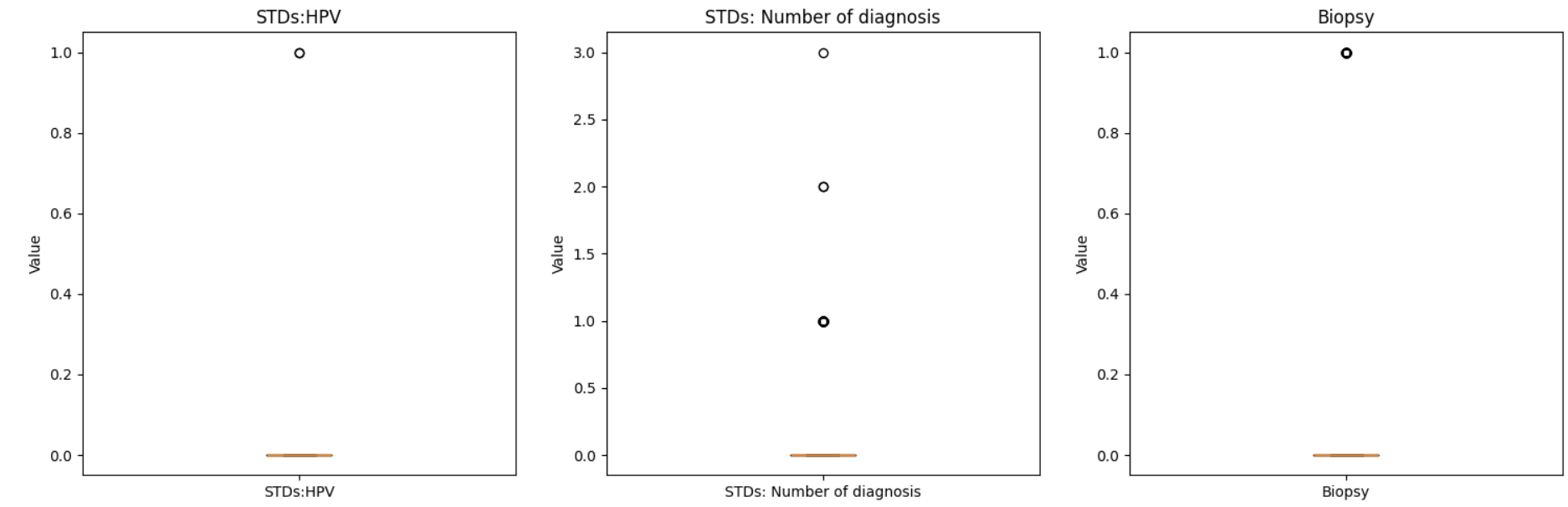
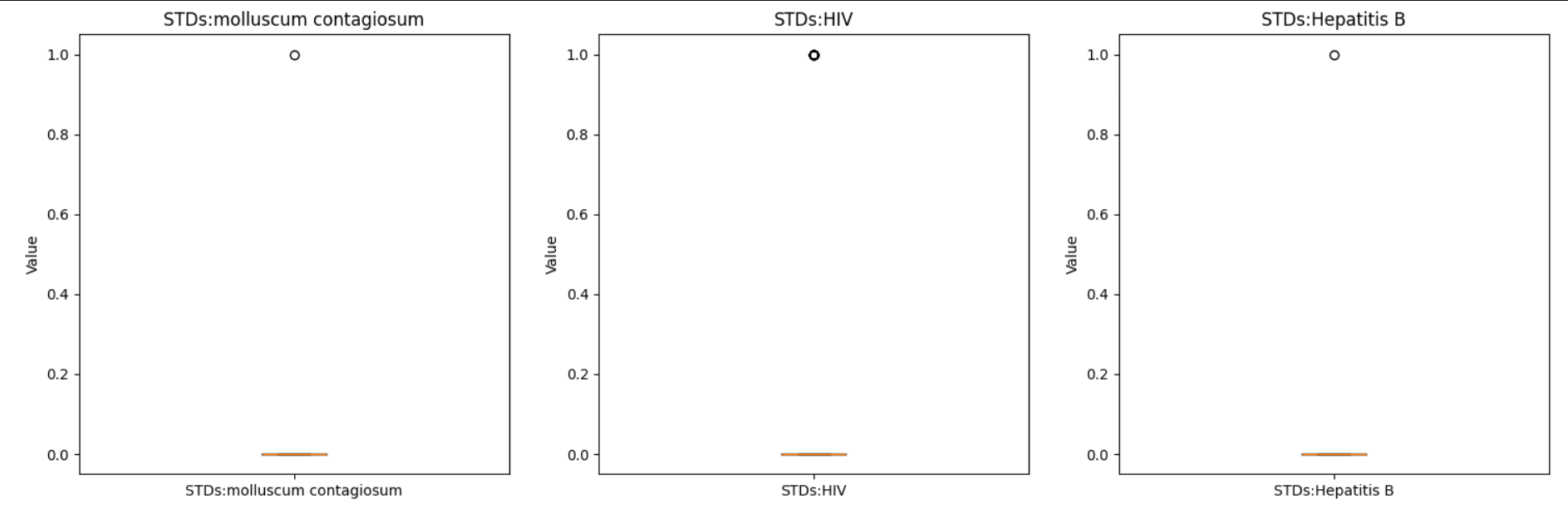
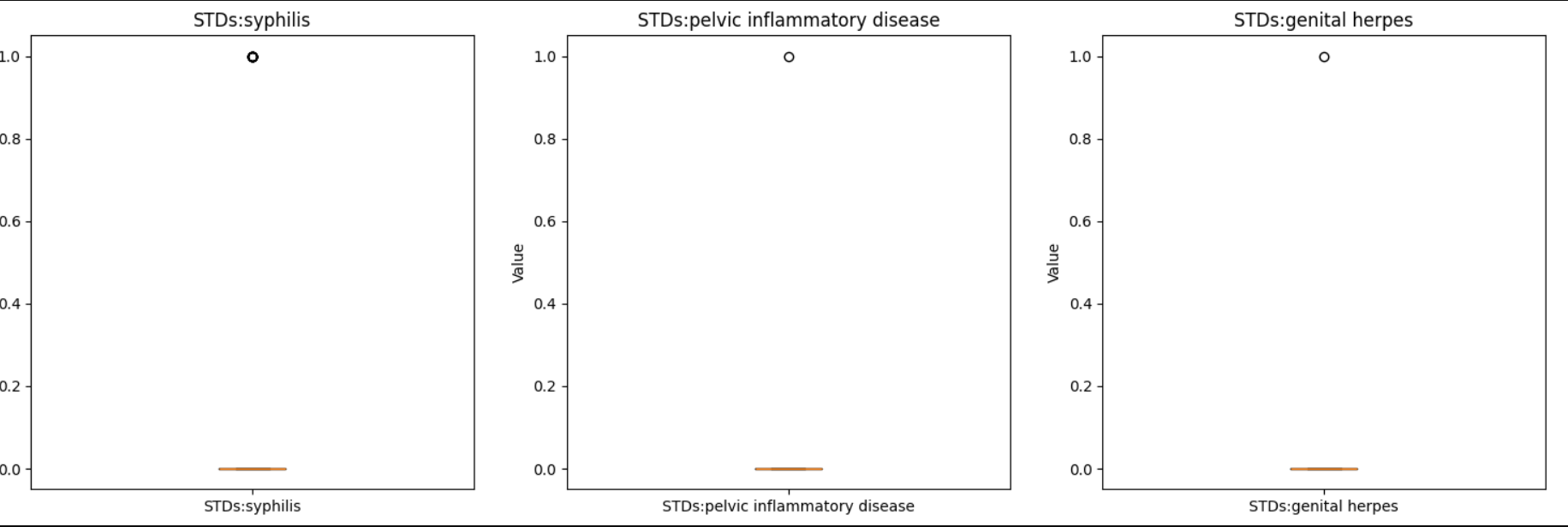
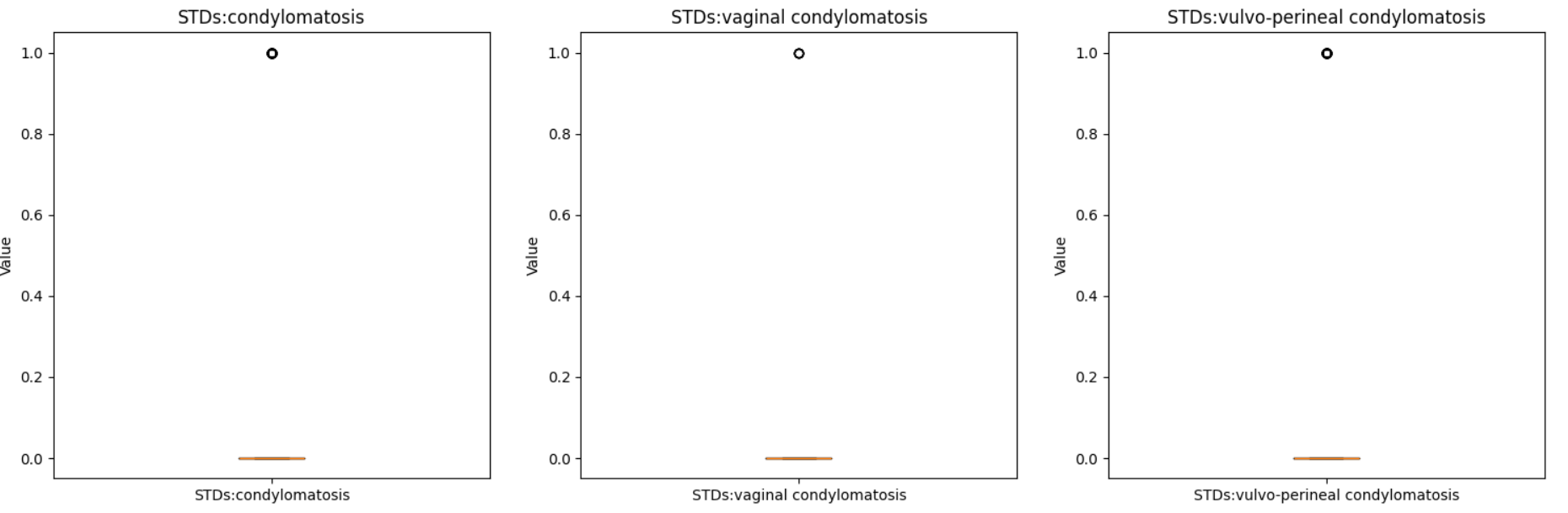
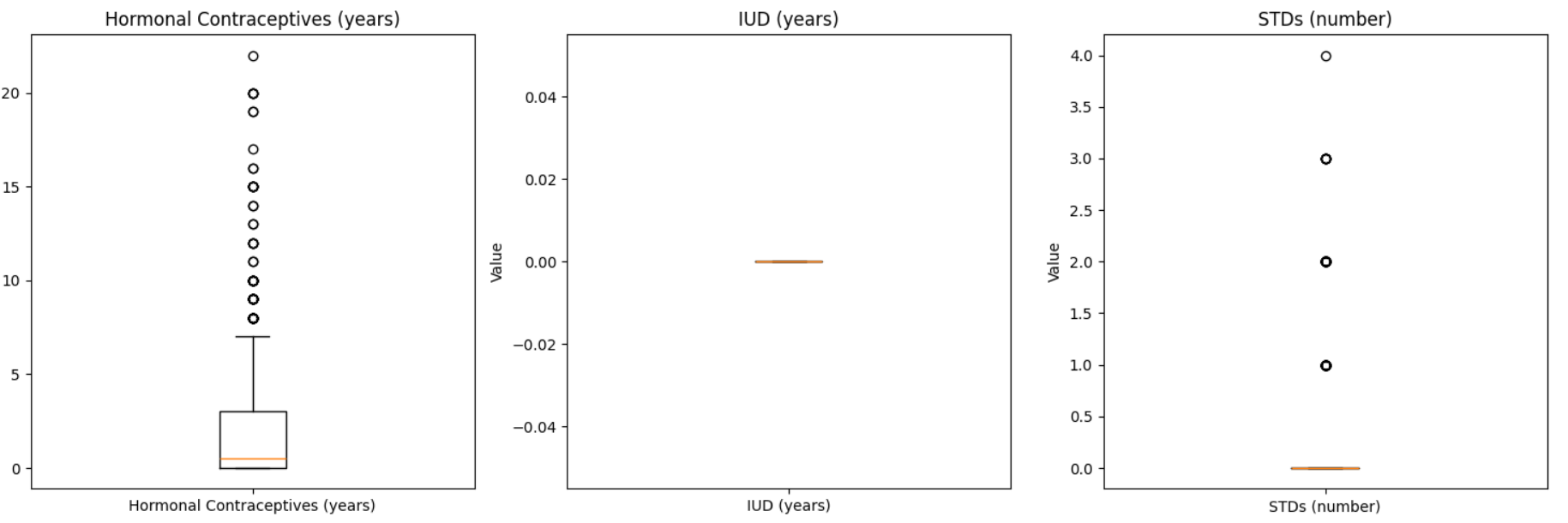
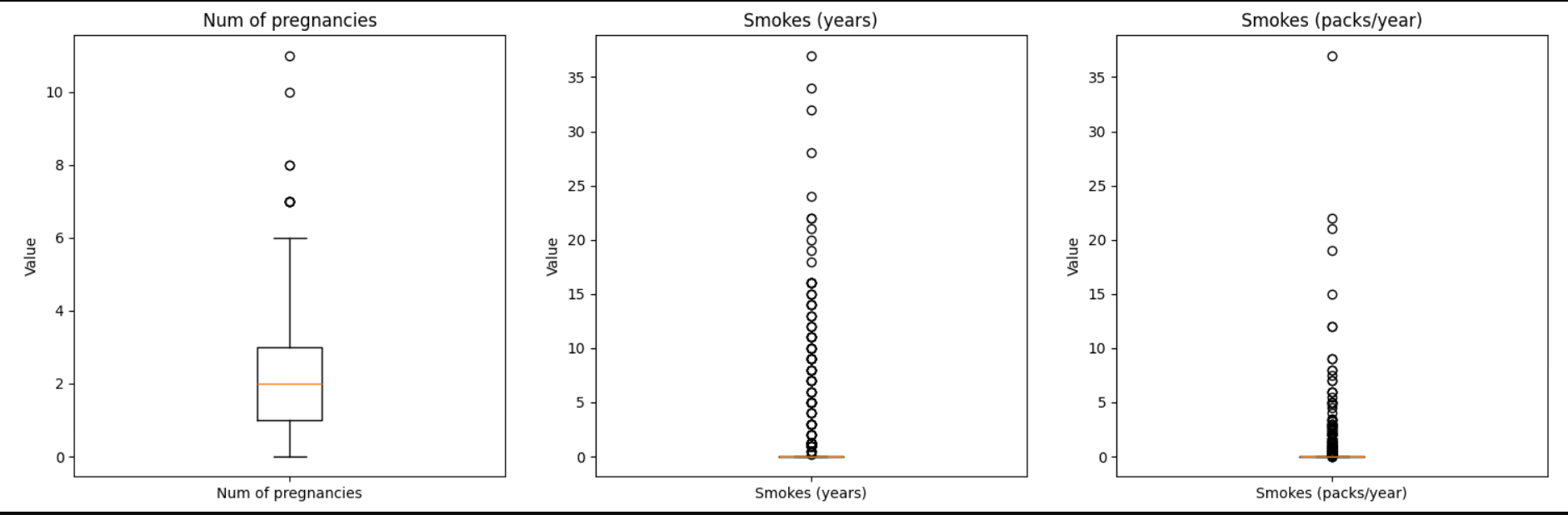
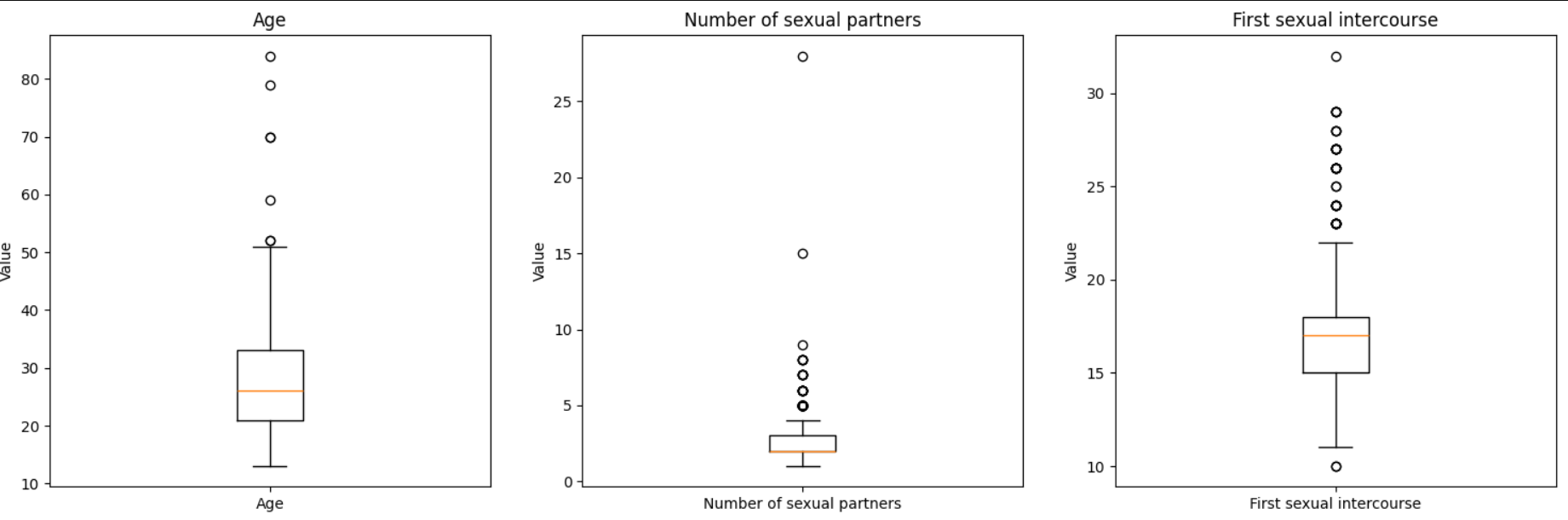
A screenshot of a computer code

AI-generated content may be incorrect.

### Phân tích tương quan và phân phối:

* Biểu đồ phân phối: Các biểu đồ histogram và boxplot đã được sử dụng để phân tích phân phối của các biến liên tục và rời rạc. Điều này giúp chúng tôi nhận diện các ngoại lai và hiểu rõ hơn về đặc điểm của từng biến.





### Lựa chọn mô hình:

* Mạng nơ-ron nhân tạo (ANN): ANN được lựa chọn vì khả năng học các mối quan hệ phức tạp trong dữ liệu. Với kiến trúc nhiều lớp, ANN có thể nắm bắt các tương tác phi tuyến giữa các yếu tố nguy cơ và biến mục tiêu.
* Random Forest: Random Forest là một thuật toán mạnh mẽ và hiệu quả, đặc biệt là trong các bài toán phân loại với dữ liệu không cân bằng. Chúng tôi đã sử dụng Random Forest để so sánh hiệu suất và đánh giá tính ổn định của mô hình.
* **K-Nearest Neighbors (KNN):** KNN là thuật toán phi tham số dựa trên khoảng cách, được dùng để so sánh hiệu suất với ANN và Random Forest. Chúng tôi tối ưu số láng giềng (K) bằng grid search để cân bằng giữa nhiễu và tổng quát hóa. Để xử lý mất cân bằng, chúng tôi áp dụng tái mẫu hoặc điều chỉnh trọng số khoảng cách, giúp cải thiện khả năng nhận diện nhóm thiểu số.

### Đánh giá mô hình:

* Ma trận nhầm lẫn và các chỉ số đánh giá: Chúng tôi đã sử dụng ma trận nhầm lẫn, độ chính xác (accuracy), độ chính xác (precision), độ bao phủ (recall), và F1-score để đánh giá hiệu suất của mô hình.

## Cấu hình máy tính

Hệ điều hành: Windows 11

CPU: AMD Ryzen 5

RAM: 8GB

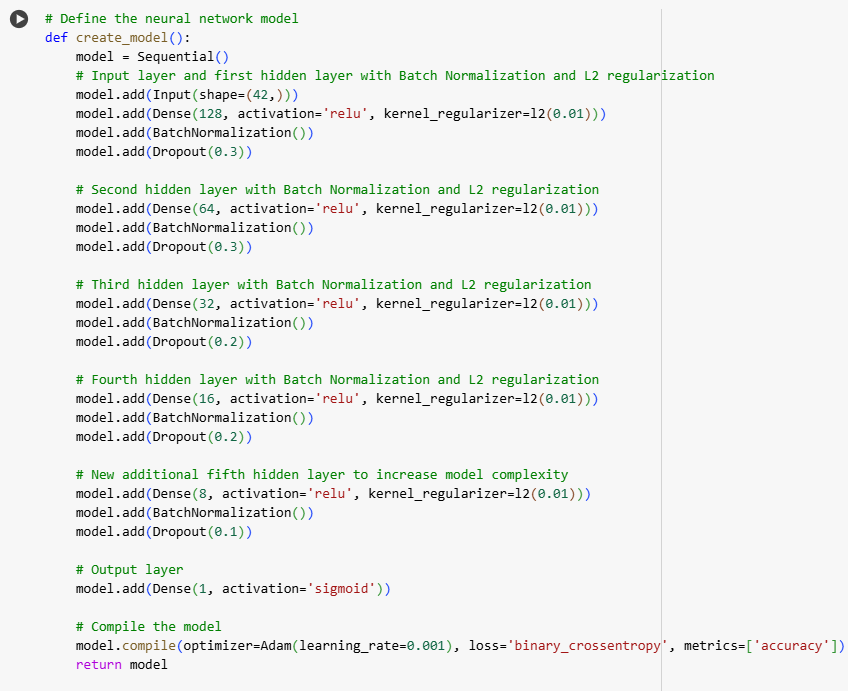
Python: 3.11

Các thư viện: NumPy, Pandas, Scikit-learn, TensorFlow/Keras, Matplotlib, Seaborn, imblearn...

## Huấn luyện mô hình

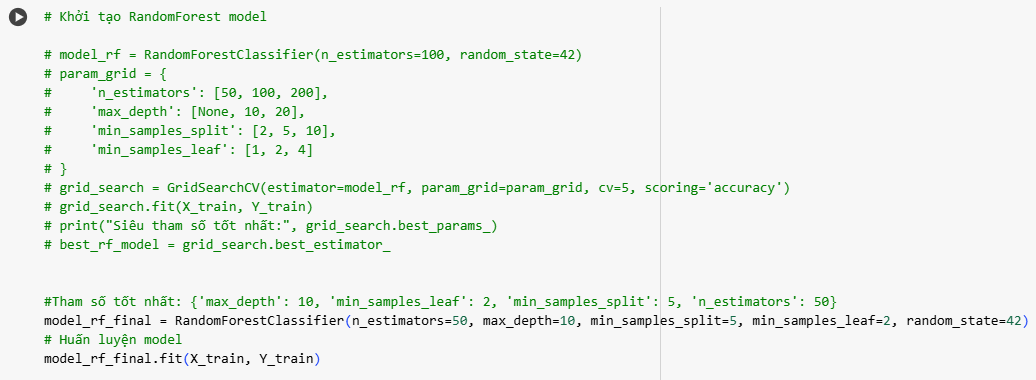
Sau khi hoàn thành các bước tiền xử lý và phân tích dữ liệu, chúng tôi tiến hành huấn luyện các mô hình học máy đã được lựa chọn. Quá trình huấn luyện này được thực hiện với mục tiêu tối ưu hóa hiệu suất dự đoán và đảm bảo tính ổn định của mô hình.

### Huấn luyện mô hình mạng nơ-ron nhân tạo (ANN):

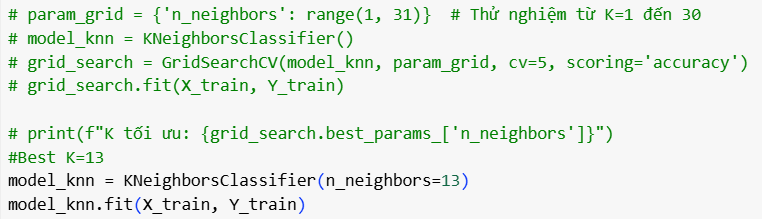


* Kiến trúc mô hình: Chúng tôi xây dựng một mô hình ANN với nhiều lớp ẩn, sử dụng hàm kích hoạt ReLU cho các lớp ẩn và hàm kích hoạt sigmoid cho lớp đầu ra.
  + Hàm kích hoạt ReLU: f(x) = max(0, x)
  + Hàm kích hoạt sigmoid: σ(x) = 1 / (1 + e^(-x))
* Hàm mất mát và tối ưu hóa: Hàm mất mát binary crossentropy được sử dụng để đo lường sai số giữa dự đoán và giá trị thực tế. Thuật toán tối ưu hóa Adam được lựa chọn để cập nhật trọng số của mạng.
  + Hàm mất mát binary crossentropy: L(y, ŷ) = -[y log(ŷ) + (1-y) log(1-ŷ)]
* Huấn luyện và kiểm tra: Mô hình được huấn luyện trên tập huấn luyện và kiểm tra trên tập kiểm tra để đánh giá hiệu suất.

### Huấn luyện mô hình Random Forest:



* Tham số mô hình: Mô hình Random Forest được huấn luyện với các tham số được tối ưu bằng Grid Search, bao gồm độ sâu tối đa (max\_depth): 10, số lượng mẫu lá tối thiểu (min\_samples\_leaf): 2, số mẫu tối thiểu để tách (min\_samples\_split): 5, số cây trong rừng (n\_estimators): 50.
* Huấn luyện và kiểm tra: Mô hình được huấn luyện trên tập huấn luyện và kiểm tra trên tập kiểm tra để đánh giá hiệu suất.
  1. **Huấn luyện mô hình KNN:**

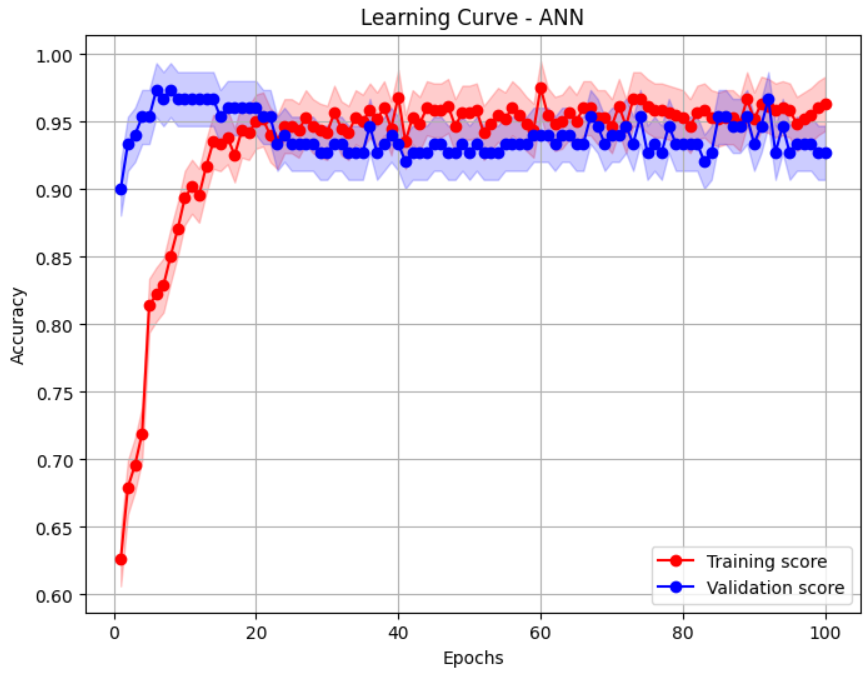
****

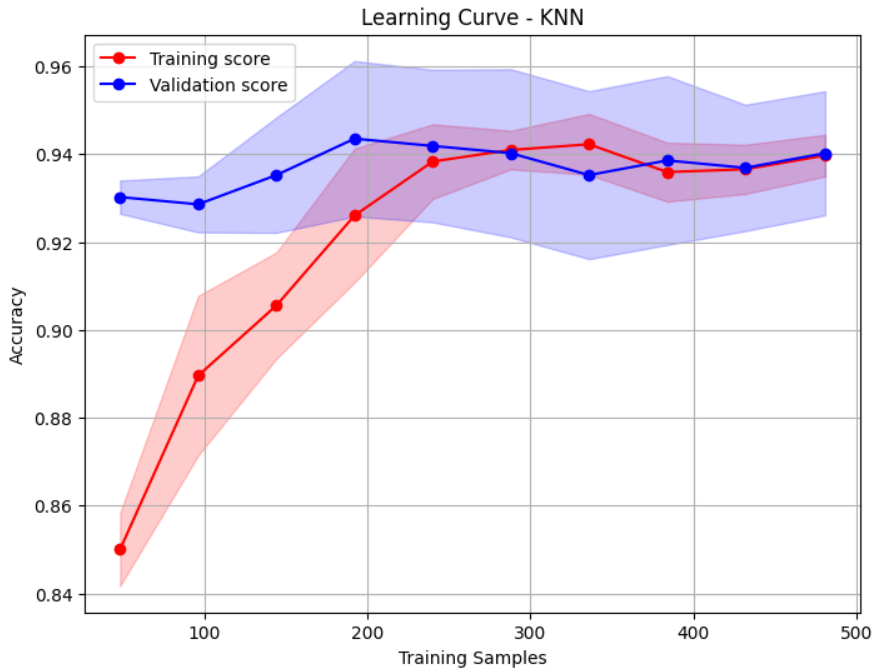
* Tham số mô hình: Mô hình KNN được huấn luyện với các tham số được tối ưu bằng Grid Search, bao gồm số láng giềng (K): 13.
* Huấn luyện và kiểm tra: Mô hình được huấn luyện trên tập huấn luyện và kiểm tra trên tập kiểm tra để đánh giá hiệu suất.

### Đánh giá hiệu suất huấn luyện:

* Đường cong học tập: Chúng tôi theo dõi đường cong học tập (learning curves) để đánh giá quá trình huấn luyện của mô hình ANN. Đường cong học tập giúp chúng tôi xác định xem mô hình có bị overfitting hay underfitting hay không.





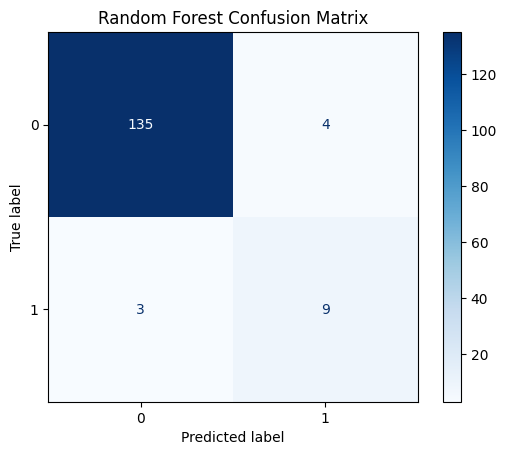


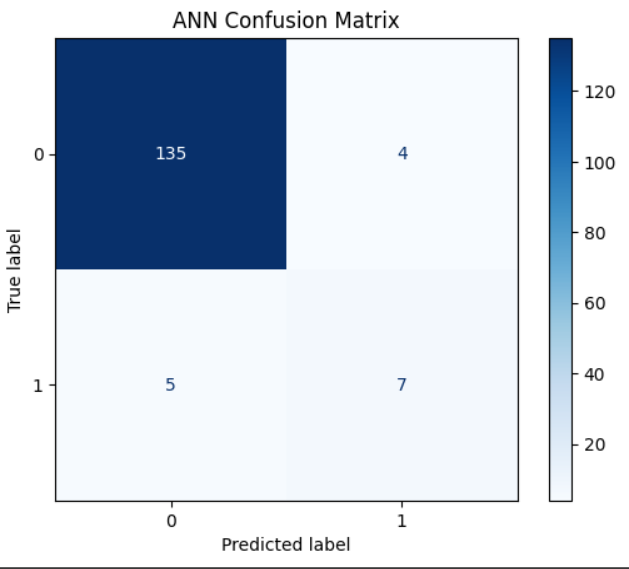
## Đánh giá mô hình

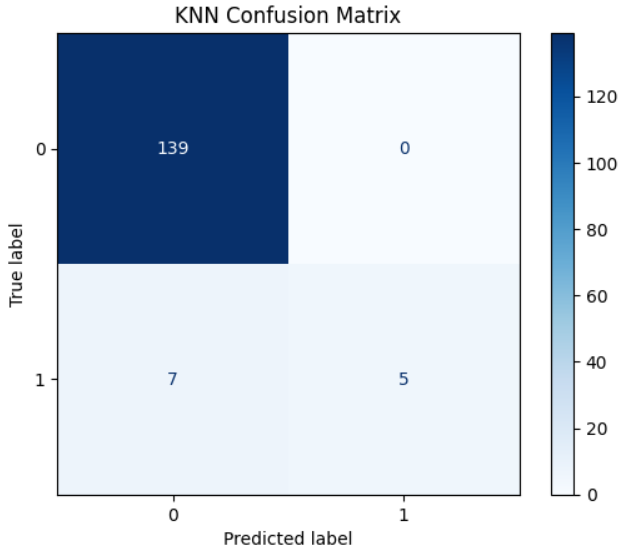
### Đánh giá mô hình phân lớp

#### Phương pháp đánh giá:

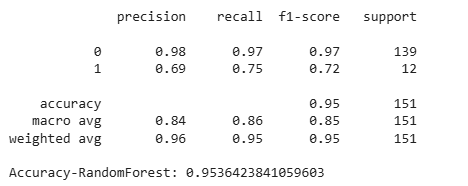
Ma trận nhầm lẫn (Confusion Matrix): Ma trận nhầm lẫn được sử dụng để đánh giá chi tiết khả năng phân loại của mô hình, bao gồm số lượng dự đoán đúng (True Positives - TP, True Negatives - TN) và sai (False Positives - FP, False Negatives - FN) cho từng lớp.

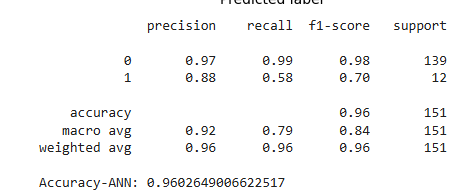


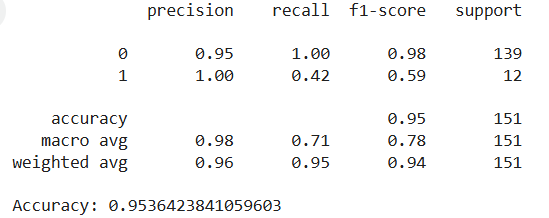




Báo cáo phân loại (Classification Report): Báo cáo phân loại được sử dụng để cung cấp một cái nhìn tổng quan về hiệu suất của mô hình cho từng lớp, bao gồm các chỉ số precision, recall, F1-score và support.





****

#### Kết quả đánh giá:

Mô hình ANN: Mô hình ANN cho thấy hiệu suất tốt trong việc dự đoán nguy cơ UTCTC. Tuy nhiên, mô hình có thể gặp khó khăn trong việc dự đoán chính xác lớp dương tính (có nguy cơ UTCTC) do sự mất cân bằng dữ liệu.

Mô hình Random Forest: Mô hình Random Forest cho thấy hiệu suất tốt, đặc biệt là trong việc dự đoán lớp dương tính. Mô hình này có khả năng chống chịu ngoại lai tốt và phù hợp với dữ liệu không cân bằng.

Mô hình KNN: Mô hình KNN cho thấy hiệu suất cạnh tranh trong việc dự đoán nguy cơ UTCTC. Tuy nhiên, do dựa trên khoảng cách, mô hình có thể bị ảnh hưởng bởi sự phân bố không đồng đều của dữ liệu. Sau khi tối ưu tham số K và áp dụng kỹ thuật xử lý mất cân bằng, KNN cải thiện khả năng dự đoán lớp dương tính nhưng vẫn nhạy cảm với nhiễu và dữ liệu ngoại lai.

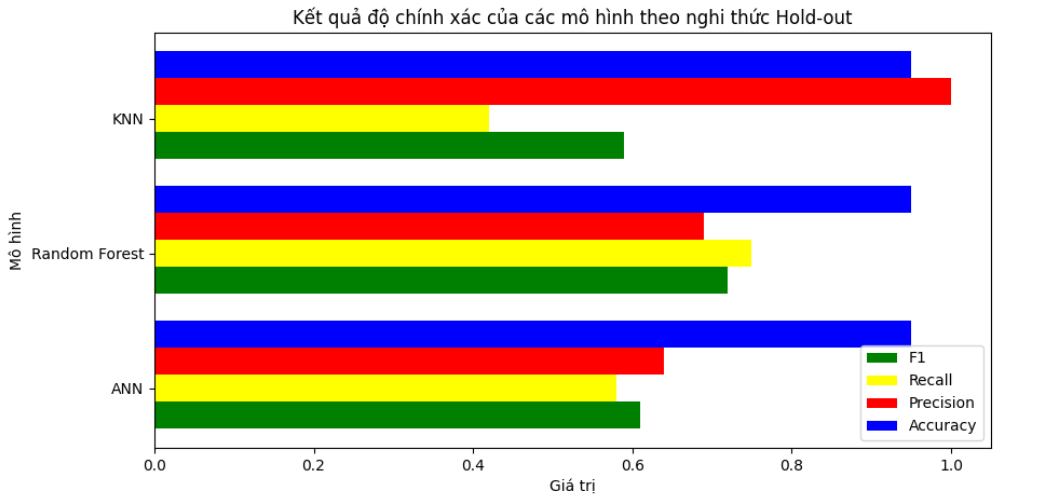
So sánh hiệu suất: Kết quả đánh giá cho thấy rằng mô hình Random Forest có hiệu suất tốt nhất trong việc dự đoán nguy cơ UTCTC, đặc biệt là đối với lớp dương tính. Mô hình ANN gặp khó khăn trong việc dự đoán chính xác lớp dương tính. Mô hình KNN có hiệu suất thấp hơn do nhạy cảm với phân bố dữ liệu và nhiễu, mặc dù đã được tối ưu tham số.

### Nhận xét kết quả thực nghiệm

#### Tổng quan kết quả:

Hiệu suất mô hình: Kết quả đánh giá cho thấy rằng mô hình Random Forest đạt hiệu suất tốt nhất trong việc dự đoán nguy cơ ung thư cổ tử cung (UTCTC), tiếp theo là mô hình mạng nơ-ron nhân tạo (ANN). Mô hình KNN có hiệu suất thấp hơn do nhạy cảm với phân bố dữ liệu và nhiễu.

Sự khác biệt giữa các mô hình: Mô hình Random Forest cho thấy hiệu suất tốt nhất, đặc biệt trong việc dự đoán lớp dương tính (có nguy cơ UTCTC), nhờ khả năng chống chịu ngoại lai và xử lý tốt dữ liệu không cân bằng. Mô hình ANN cũng đạt hiệu suất cao nhưng vẫn gặp khó khăn với lớp dương tính. Trong khi đó, KNN có độ chính xác thấp hơn do bị ảnh hưởng bởi sự phân bố không đồng đều của dữ liệu và tính nhạy cảm với nhiễu.



#### Hạn chế

Hạn chế dữ liệu: Dữ liệu được sử dụng trong báo cáo này có thể không đại diện cho tất cả các nhóm bệnh nhân, điều này có thể ảnh hưởng đến khả năng tổng quát hóa của mô hình.

Hạn chế mô hình: Các mô hình được sử dụng trong báo cáo này có thể chưa tối ưu hóa hoàn toàn. Cần có thêm nghiên cứu để thử nghiệm các mô hình và kỹ thuật khác nhau để cải thiện hiệu suất.

Cần thiết của nghiên cứu lâm sàng: Cần có các nghiên cứu lâm sàng để đánh giá tính ứng dụng của mô hình trong thực tế và xác định tác động của mô hình đến kết quả sức khỏe của bệnh nhân.

#### Kết luận

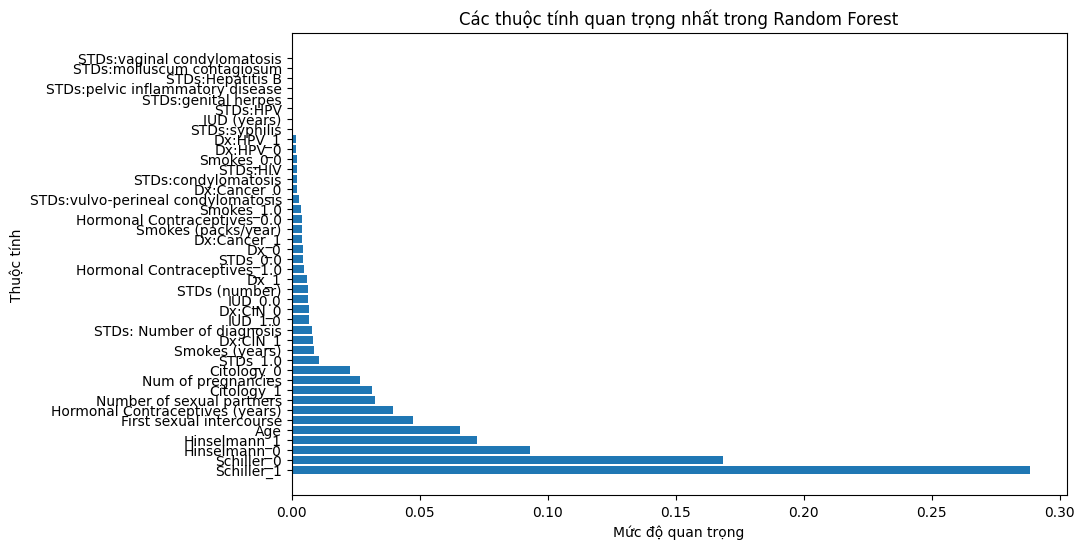
Kết quả thực nghiệm cho thấy rằng các mô hình học máy có tiềm năng trong việc dự đoán nguy cơ UTCTC. Tuy nhiên, cần có thêm nghiên cứu để giải quyết các hạn chế và đánh giá tính ứng dụng của mô hình trong thực tiễn lâm sàng.

# PHẦN KẾT LUẬN

## Kết quả đạt được

Xây dựng và đánh giá mô hình: Chúng tôi đã xây dựng và đánh giá hiệu suất của cả mô hình ANN, Random Forest và KNN trên tập dữ liệu UTCTC. Kết quả cho thấy cả ba mô hình đều đạt được hiệu suất tốt, với Random Forest cho thấy hiệu suất tốt hơn trong việc dự đoán lớp dương tính (có nguy cơ UTCTC).

Phân tích yếu tố nguy cơ: Phân tích tầm quan trọng của các yếu tố nguy cơ cho thấy rằng một số yếu tố như kết quả xét nghiệm (Hinselmann, Schiller, Citology) và tiền sử bệnh tật (Dx:Cancer, Dx:CIN, Dx:HPV) có ảnh hưởng lớn đến nguy cơ UTCTC.



## Hướng phát triển

Cải thiện hiệu suất mô hình: Cần có thêm nghiên cứu để tối ưu hóa các tham số của mô hình và thử nghiệm các kỹ thuật học máy khác nhau để cải thiện hiệu suất, đặc biệt là khả năng dự đoán lớp dương tính.

Mở rộng dữ liệu: Việc mở rộng tập dữ liệu với nhiều thông tin hơn và từ nhiều nguồn khác nhau có thể giúp cải thiện khả năng tổng quát hóa của mô hình.

Nghiên cứu lâm sàng: Cần có các nghiên cứu lâm sàng để đánh giá tính ứng dụng của mô hình trong thực tế và xác định tác động của mô hình đến kết quả sức khỏe của bệnh nhân.

Phát triển công cụ hỗ trợ quyết định: Việc phát triển các công cụ hỗ trợ quyết định dựa trên mô hình học máy có thể giúp các chuyên gia y tế sàng lọc UTCTC một cách hiệu quả và đưa ra các quyết định lâm sàng chính xác hơn.

Nghiên cứu đa ngành: Việc nghiên cứu và phát triển các mô hình dự đoán UTCTC đòi hỏi sự hợp tác giữa các chuyên gia trong nhiều lĩnh vực, bao gồm khoa học máy tính, y học và thống kê.

# TÀI LIỆU THAM KHẢO

[1]. Arbyn, M., Sankaranarayanan, R., & Muwonge, R. (2020). Early detection and screening of cervical cancer. *Journal of Clinical Pathology*,

[2]. Plummer, M., Herrero, R., Franceschi, S., Meijer, C. J., Snijders, P. J., & Bosch, F. X. (2003). Smoking and cervical cancer: pooled analysis of the IARC multi-centric case-control study. *Cancer Causes & Control*

[3]. Sokolova, M., & Lapalme, G. (2009). A systematic analysis of performance measures for classification tasks. *Information Processing & Management*