BỘ GIÁO DỤC VÀ ĐÀO TẠO
TRƯỜNG ĐẠI HỌC CÔNG NGHỆ TP. HCM

TRƯƠNG ANH VŨ

PHÁT TRIỂN MỘT HỆ THỐNG HỖ TRỢ CHẨN ĐOÁN BỆNH VÀ ĐỀ XUẤT CÁC HƯỚNG ĐIỀU TRỊ

LUẬN VĂN THẠC SĨ

Chuyên ngành : Công nghệ thông tin
Mã số ngành: 60480201

TP. HỒ CHÍ MINH, tháng 01 năm 2016

BỘ GIÁO DỤC VÀ ĐÀO TẠO

TRƯỜNG ĐẠI HỌC CÔNG NGHỆ TP. HCM

TRƯƠNG ANH VŨ

PHÁT TRIỂN MỘT HỆ THỐNG HỖ TRỢ CHẨN ĐOÁN BỆNH VÀ ĐỀ XUẤT CÁC HƯỚNG ĐIỀU TRỊ

LUẬN VĂN THẠC SĨ

Chuyên ngành : Công nghệ thông tin

Mã số ngành: 60480201

CÁN BỘ HƯỚNG DẪN KHOA HỌC: TS.NGUYỄN THỊ THANH SANG

TP. HỒ CHÍ MINH, tháng 01 năm 2016

CÔNG TRÌNH ĐƯỢC HOÀN THÀNH TẠI
TRƯỜNG ĐẠI HỌC CÔNG NGHỆ TP. HCM

Cán bộ hướng dẫn khoa học : **TS.NGUYỄN THỊ THANH SANG**
(Ghi rõ họ, tên, học hàm, học vị và chữ ký)

Nguyễn Thị Thanh Sang

Luận văn Thạc sĩ được bảo vệ tại Trường Đại học Công nghệ TP. HCM ngày 20 tháng 03 năm 2016

Thành phần Hội đồng đánh giá Luận văn Thạc sĩ gồm:
(Ghi rõ họ, tên, học hàm, học vị của Hội đồng chấm bảo vệ Luận văn Thạc sĩ)

TT	Họ và tên	Chức danh Hội đồng
1	PGS.TS. Võ Đình Bảy	Chủ tịch
2	GS.TSKH. Hoàng Văn Kiếm	Phản biện 1
3	TS. Lê Văn Quốc Anh	Phản biện 2
4	TS. Lê Tuấn Anh	Ủy viên
5	TS. Nguyễn Thị Thúy Loan	Ủy viên, Thư ký

Xác nhận của Chủ tịch Hội đồng đánh giá Luận sau khi Luận văn đã được sửa chữa (nếu có).

Chủ tịch Hội đồng đánh giá LV

TRƯỜNG ĐH CÔNG NGHỆ TP. HCM
PHÒNG QLKH – ĐTSĐH

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

TP. HCM, ngày 10 *tháng 01 năm 20*16

NHIỆM VỤ LUẬN VĂN THẠC SĨ

Họ tên học viên: TRƯƠNG ANH VŨ Giới tính: Nam

Ngày, tháng, năm sinh: 27/09/1982 Nơi sinh: Cần Thơ

Chuyên ngành: Công nghệ thông tin MSHV: 1441860032

I- Tên đề tài:

Phát triển một hệ thống hỗ trợ chẩn đoán bệnh và đề xuất các hướng điều trị

II- Nhiệm vụ và nội dung:

- Nghiên cứu các phương pháp học máy.
- Tìm hiểu khả năng áp dụng của phương pháp cây quyết định để ứng dụng trong lĩnh vực y tế.
- Phân tích dữ liệu học về cận lâm sàng của bệnh nhân.
- Xây dựng bộ luật nhằm hỗ trợ chẩn đoán bệnh dự vào các kết quả cận lâm sàng.
- Lập trình xây dựng một phân hệ (module) tích hợp vào hệ thống quản lý bệnh viện để hỗ trợ quá trình khám chữa bệnh.

III- Ngày giao nhiệm vụ: *20/08/2015*

IV- Ngày hoàn thành nhiệm vụ: 15/01/2016

V- Cán bộ hướng dẫn: TS.NGUYỄN THỊ THANH SANG

CÁN BỘ HƯỚNG DẪN **KHOA QUẢN LÝ CHUYÊN NGÀNH**

(Họ tên và chữ ký) (Họ tên và chữ ký)

TS.NGUYỄN THỊ THANH SANG

LỜI CAM ĐOAN

Tôi xin cam đoan đây là công trình nghiên cứu của riêng tôi và sự hướng dẫn khoa học của **TS.Nguyễn Thị Thanh Sang**. Các số liệu, kết quả nêu trong Luận văn là trung thực và chưa từng được ai công bố trong bất kỳ công trình nào khác.

Tôi xin cam đoan rằng mọi sự giúp đỡ cho việc thực hiện Luận văn này đã được cảm ơn và các thông tin trích dẫn trong Luận văn đã được chỉ rõ nguồn gốc.

Học viên thực hiện Luận văn

(Ký và ghi rõ họ tên)

Trương Anh Vũ

LỜI CÁM ƠN

Lời đầu tiên, với lòng biết ơn sâu sắc, tôi xin chân thành cảm ơn đến các thầy, cô giảng viên của trường đã tận tình truyền đạt cho học viên những kiến thức quý báu trong suốt quá trình học tập, nghiên cứu và rèn luyện tại trường.

Tôi xin chân thành cảm ơn TS.Nguyễn Thị Thanh Sang đã dành nhiều thời gian giảng dạy và tận tâm hướng dẫn tôi trong quá trình nghiên cứu chuyên môn để hoàn thành đề tài này. Một lần nữa, tôi xin gửi lời cảm ơn chân thành đến TS.Nguyễn Thị Thanh Sang.

Tôi xin chân thành cảm ơn BS.CK2.Nguyễn Quang Khả, Trưởng khoa Thận – Thận nhân tạo, bệnh viện đa khoa trung ương Cần Thơ đã tận tình hướng dẫn các qui trình và kiến thức chuyên ngành y để hoàn thành kết quả nghiên cứu này.

Cuối cùng, tôi xin gửi đến bạn bè, người thân, đồng nghiệp và lãnh đạo bệnh viện đa khoa trung ương Cần Thơ đã tạo điều kiện thuận lợi về mọi mặt trong quá trình học tập và nghiên cứu của mình.

TP. Hồ Chí Minh, ngày 15 tháng 01 năm 2016

Tác giả

Trương Anh Vũ

TÓM TẮT

Hiện nay số lượng bệnh nhân đến khám và điều trị tại các các cơ sở y tế ngày càng cao, gây quá tải cho các bệnh viện, từ đó việc khám chữa bệnh cũng như tầm soát bệnh quả các kết quả xét nghiệm chưa được quan tâm đúng mức, các bệnh lý có thể vô tình bị bỏ qua hoặc không phát hiện kịp thời dẫn đến tình trạng khi phát bệnh thì cần tốn nhiều chi phí điều trị và tốn nhiều thời gian của bệnh nhân.

Với sự hỗ trợ của hệ thống công nghệ thông tin, ngày càng có nhiều ứng dụng hữu ích để phục vụ công tác khám và điều trị. Trên cơ sở nghiên cứu các kết quả cận lâm sàng của bệnh nhân và ứng dụng một số thuật toán "học máy", đề tài này tiến tới xây dựng một phân hệ hỗ trợ chẩn đoán và gợi ý chỉ dẫn điều trị cho các bác sĩ nhằm rút ngắn khoảng cách giữa thực lý thuyết và kinh nghiệm thực tế của các bác sĩ, đồng thời có cơ sở để phát hiện các bệnh lý tiềm ẩn sớm hơn, rút ngắn thời gian điều trị và tiết kiệm chi phí.

Trên cơ sở nghiên cứu các bệnh nhân có bệnh lý thận nội khoa và dựa trên kết phân tích mẫu kết quả xét nghiệm, đề tài này hướng tới xây dựng phần mềm nhúng (module tích hợp dạng .dll) sử dụng ngôn ngữ Prolog và C# vào các phân hệ quản lý bệnh viện để hỗ trợ cảnh báo (nếu có) cho các bác sĩ trong quá trình khám và điều trị.

Bước đầu, trên cơ sở nghiên cứu của luận văn, việc áp dụng cây quyết định đã mang lại hiệu quả trong việc chẩn đoán và hỗ trợ gợi ý điều trị cho bác sĩ trong quá trình khám chữa bệnh. Tuy nhiên để kết quả ứng dụng cây quyết định trong hỗ trợ chẩn đoán và điều trị được tốt hơn cần có thời gian nghiên cứu mở rộng và đi sâu vào nghiên cứu các bệnh lý khác đặc biệt là các bệnh lý kết hợp để đưa ra phương pháp chẩn đoán tốt hơn, nhanh chóng hơn

Một số hiệu quả đạt được sau khi ứng dụng nghiên cứu này vào thực tế :

- Thời gian xác định bệnh lý nhanh hơn khi phân tích bệnh lý bằng phương pháp truyền thống.
- Tự động kết hợp các thuộc tính của người bệnh để phân tích tránh tình trạng thiếu xót chẩn đoán do không đủ điều kiện khai thác thông tin khi khám bệnh.

- Hỗ trợ bác sĩ ra quyết định điều trị nhanh chóng và có khoa học (dựa trên phát đồ điều trị)

Ngoài phần mở đầu và tổng quan, nội dung chính của luận văn được trình bày qua các nội dung sau:

- Phân tích một số thuật toán cây quyết định để đánh giá hiệu quả từng thuật toán khi áp dụng vào bài toán y tế.
- Thử nghiệm và phân tích các mẫu dữ liệu về xét nghiệm của bệnh nhân.
- Đánh giá, bàn luận, đúc kết hiệu quả của từng phương pháp và xây dựng ứng dụng tích hợp vào hệ thống quản lý bệnh viện

ABSTRACT

Currently, the patient is more and more, who go to examination and treatment at hospital, is the leading cause of overcrowding in hospitals, so the health care and medical screening have not been proper care, the disease may inadvertently overlooked or not detected in the early, this is the cause of increased severity of illness and increased costs of treatment.

With the support of information technology systems, more and more useful applications to support the examination and treatment. Based on study results of paraclinical patients with machine learning applications. This topic desire building a module to support diagnosis and suggested treatment guidelines for the doctor, shorten the gap between practice and theory of the doctor, at the same time early detection of disease, cost savings.

Based on studies of patients with kidney disease and analyzing the paraclinical test results, the topic towards will build new component software (.dll module) using Prolog and C # language integrated in the hospital management information system for warning to physician (if any).

Initially, this topic used the decision tree effective to support the examination and treatment. However, for this applications with support decision tree more effective, need further research the combined pathology to support diagnosis better and faster.

Some results when using this application into practice:

- This application is faster diagnosis against traditional methods.
- Automatic using the attributes of the patient to analyze, not lack of diagnosis because do not have time to exploited patient information.
- Support the doctor treatment decisions quickly and scientific (based on treatment protocols).

Eliminate content of introduction and overview, the main content of this topic presented the following:

- Analyse the decision tree algorithm to evaluate the effectiveness when applied to medical problems.
- Testing and analysis the paraclinical samples data of patients.
- Review, discussion, summarizing the effectiveness of each method and building component software integrated in the hospital management information system.

MỤC LỤC

DANH MỤC CÁC BẢNG

DANH MỤC CÁC BIỂU ĐỒ, ĐỒ THỊ, SƠ ĐỒ, HÌNH ẢNH

Chương 1: MỞ ĐẦU

1.1 Lý do chọn đề tài

Phát sinh từ thực tế trong quá trình khám và điều trị cho bệnh nhân.

Hằng ngày số lượng bệnh nhân đến khám và điều trị tại các cơ sở y tế ngày càng cao, gây quá tải cho các bệnh viện, từ đó việc khám chữa bệnh cũng như tầm soát bệnh quả các kết quả xét nghiệm chưa được quan tâm đúng mức, các bác sĩ chỉ quan tâm đến các chỉ số xét nghiệm ảnh hưởng nghiêm trọng đến sức khỏe người bệnh.

Việc bị bỏ qua các kết quả xét nghiệm ở tiêu chí bình thường nhưng ở mức độ cao với sự kết hợp các chỉ số xét nghiệm khác là nguyên nhân dẫn đến một số bác sĩ và bệnh nhân không phát hiện kịp thời những rối loạn trong sinh lý dẫn đến tình trạng khi phát bệnh thì cần tốn nhiều chi phí điều trị và tốn nhiều thời gian của bệnh nhân.

Trong quá trình khám và điều trị, các bác sĩ chưa có sự phận tích có tính hệ thống khi có nhiều chỉ số xét nghiệm kết hợp để sớm phát hiện những căn bệnh chưa bộc phát.

Trên cơ sở nghiên cứu các kết quả cận lâm sàng của bệnh nhân và ứng dụng một số thuật toán "học máy", đề tài này tiến tới xây dựng một phân hệ hỗ trợ chẩn đoán và gợi ý chỉ dẫn điều trị cho các bác sĩ nhằm rút ngắn khoảng cách giữa thực lý thuyết và kinh nghiệm thực tế của các bác sĩ, đồng thời có cơ sở để phát hiện các bệnh lý tiềm ẩn sớm hơn, rút ngắn thời gian điều trị và tiết kiệm chi phí.

1.2 Nội dung chính

Nghiên cứu này sẽ hướng tới hỗ trợ cho tất cả các đơn vị khám chữa bệnh có trang bị hệ thống xét nghiệm tự động và đã triển khai phần mềm quản lý thông tin bệnh viện.

Các tập dữ liệu xét nghiệm được thu thập và chọn lọc cho việc xử lý phân loại.

Các phương pháp xử lý dữ liệu dạng số, và các phương pháp máy học được tìm hiểu và chọn lựa để áp dụng vào hệ thống đề xuất trong đề tài này.

Các phương pháp thực nghiệm và đánh giá các giải thuật phân loại được áp dụng trong đề tài này.

1.3 Mục tiêu của đề tài

1.3.1 Mục tiêu tổng quát

Dựa vào kết quả các chỉ số xét nghiệm, hệ thống phân tích đánh giá và đưa ra phương án điều trị gợi ý, nhằm hỗ trợ các bác sĩ trong quá trình khám và điều trị bệnh cho bệnh nhân.

Trong khuôn khổ giới hạn, đề tài này chỉ tập trung phân tích dữ liệu liên quan đến bệnh lý thận nội khoa.

1.3.2 Mục tiêu cụ thể

Phân tích mẫu kết quả xét nghiệm của các bệnh nhân có bệnh lý về thận.

Xây dựng hệ thống phân tích các kết quả xét nghiệm thu thập được và kết quả chẩn đoán, điều trị, sử dụng phương pháp máy học phân loại các mẫu xét nghiệm.

Xây dựng phần mềm nhúng vào các phân hệ quản lý bệnh viện để phân tích kết quả xét nghiệm của bệnh nhân khi nhận được kết quả xét nghiệm từ các hệ thống xét nghiệm tự động. Cảnh báo (nếu có) sau khi phân tích kết quả dựa trên mẫu dữ liệu đã được huấn luyện trước đó.

1.4 Đối tượng nghiên cứu

Nghiên cứu kết quả khám và điều trị của các bệnh nhân tại bệnh viện đa khoa trung ương Cần Thơ trong thời gian từ năm 2014 đến 2015 (khoảng 140.000 mẫu dữ liệu).

1.5 Phạm vi nghiên cứu

Đánh giá kết quả chẩn đoán và chỉ định điều trị trên thực tế và bộ chuẩn dùng trong chẩn đoán và điều trị (phác đồ điều trị, guidelines). So sánh các tiêu chí đánh giá bệnh lý thận trên phát đồ chẩn đoán điều trị với kết quả chẩn đoán của các bác sĩ trong thực tế để xem xét việc chẩn đoán của các bác sĩ có phù hợp hay không.

Dữ liệu kết quả cận lâm sàng của bệnh nhân dùng cho "máy học" được tập hợp từ bộ lưu trữ dữ liệu của các máy xét nghiệm tự động theo chỉ định cận lâm sàng của

các bác sĩ theo mẫu (Hình 1.1) và thông tin điều trị của bệnh nhân trên hệ thống quản lý thông tin bệnh viện

BỘ Y TẾ
BVĐK TW Cần Thơ

PHIẾU XÉT NGHIỆM HÓA SINH MÁU
☐Thường ☐Cấp Cứu

MS:33/BV-01

Họ Tên Người Bệnh..... Năm Sinh ..... **Nam** **Nữ**
Địa Chỉ..... Thẻ BHYT .....
Khoa..... Buồng..... Giường .....
Chẩn Đoán.....

Tên Xét Nghiệm	Trị số bình thường	Kết quả
☐Urea	2,5 - 7,5 mmol/L	
☐Glucose	3,9 - 6,4 mmol/L	
☐Creatinine	Nam: 62 - 120 µmol/L Nữ: 53 - 100 µmol/L	
☐Acid Uric	Nam: 180 - 420 µmol/L Nữ: 150 - 360 µmol/L	
☐Bilirubin toàn phần	≤ 17 µmol/L	
☐Bilirubin trực tiếp	≤ 4,3 µmol/L	
☐Bilirubin gián tiếp	≤ 12,7 µmol/L	
☐Protid (Protein TP)	65- 82 g/L	
☐Albumin	35 -50 g/L	
☐Globulin	24 - 38 g/L	
☐Tỷ Lệ A/G	1,3 - 1,8	
☐Fibrinogen	2 - 4 g/L	
☐Cholesterol	3,9 - 5,2 mmol/L	
☐Triglyceride	0,46 - 1,88 mmol/L	
☐HDL-cholesterol	≥ 0,9 mmol/L	
☐LDL-cholesterol	≤ 3,4 mmol/L	
☐Na+	135 - 145 mmol/L	
☐K+	3,5 - 5 mmol/L	
☐Cl-	98 - 106 mmol/L	
☐Ca++	2,15 - 2,6 mmol/L	
☐Calci Ion hóa	1,17 - 1,29 mmol/L	
☐Phostpho	NL:0,9 - 1,5mmol/L TE:1,3 - 2,2mmol/L	

Tên Xét Nghiệm	Trị số bình thường	Kết quả
☐Sắt	Nam: 11 - 27 µmol/L Nữ: 7 - 26 µmol/L	
☐Magiê	0,8 - 1,00 mmol/L	
☐GOT (AST)	≤ 37 U/L - 37°C	
☐GPT (ALT)	≤ 40 U/L - 37°C	
☐Alpha-Amylase		
☐CK (CPK)	Nam: 24 - 167 U/L - 37°C Nữ: 24 - 190 U/L - 37°C	
☐CKMB	≤ 24 U/L - 37°C	
☐LDH	230 - 460 U/L - 37°C	
☐GGT (Gama GT)	Nam: ≤ 11 - 50 U/L - 37°C Nữ: ≤ 7 - 32 U/L - 37°C	
☐Cholinesterase	5300 - 12900 U/L - 37°C	
☐Phosphatase kiềm		
Các xét nghiệm khí máu		
☐pH động mạch	7,37 - 7,45	
☐pCO2	Nam : 35 - 46mmHg Nữ : 32 - 43mmHg	
☐pO2 động mạch	71 - 104mmHg	
☐HCO3 chuẩn	21 - 26mmol/L	
☐Kiềm dư	-2 đến +3 mmol/L	
Các xét nghiệm khác		
☐Alcool (Cồn)	(ĐVT: mg/100ml)	
☐HbA1c		
☐CRP hs		

..... Giờ, Ngày ..... Tháng ..... Năm .....
BÁC SĨ ĐIỀU TRỊ

..... Giờ, Ngày ..... Tháng ..... Năm .....
TRƯỞNG KHOA XÉT NGHIỆM

Hình 1.1. Mẫu phiếu xét nghiệm

Trong phạm vi giới hạn, đề tài này chỉ nghiên cứu các bệnh nhân có chẩn đoán bệnh lý thận nội khoa để xem xét đưa ra gợi ý điều trị nhằm hỗ trợ các bác sĩ trong quá trình khám bệnh và điều trị cho bệnh nhân.

1.6 Ý nghĩa thực tiễn và ý nghĩa khoa học của luận văn

1.6.1 Ý nghĩa thực tiễn

Nghiên cứu nhằm tìm ra qui luật chung để chẩn đoán bệnh lý dựa vào các kết quả cận lâm sàng của người bệnh và hỗ trợ các bác sĩ trong quá trình khám chữa bệnh, rút ngắn thời gian khám chữa bệnh cho bệnh nhân và giảm thời gian chờ đợi của bệnh nhân khác trong quá trình khám bệnh.

Nghiên cứu này hướng tới xây dựng một phân hệ (module) tích hợp vào hệ thống quản lý bệnh viện để tiến hành phân tích các kết quả cận lâm sàng và đề xuất cho bác sĩ hướng chẩn đoán và điều trị bệnh nhân dựa vào các tập luật rút trích từ việc phân tích số liệu các bệnh nhân trước đó.

1.6.2 Ý nghĩa khoa học

Nghiên cứu này nhằm tìm ra phương pháp phân tích các kết quả cận lâm sàng của bệnh nhân một cách khoa học và có hiệu quả nhất đồng thời cung cấp giải pháp xử lý thông tin cận lâm sàng của bệnh nhân và đề xuất các hướng xử lý thông tin theo hướng hợp lý (logic).

Chương 2: TỔNG QUAN

2.1 Tổng quan máy học

2.1.1 Khái niệm máy học (machine learning)

Định nghĩa của chúng ta về học tập là đủ rộng để bao gồm hầu hết các công việc mà chúng ta sẽ quy ước gọi là nhiệm vụ "học tập", như chúng ta sử dụng hàng ngày từ trong ngôn ngữ. Nó cũng là đủ rộng để bao gồm các chương trình máy tính cải thiện từ kinh nghiệm trong những cách khá đơn giản. [1]

Học là : Để có được kiến thức bằng cách nghiên cứu, đúc kinh nghiệm, hoặc được giảng dạy; Để có được nhận thức của thông tin nhờ vào quan sát; Để ghi vào bộ nhớ, trí não.

Học máy, có tài liệu gọi là Máy học, (tiếng Anh: machine learning) là một lĩnh vực của trí tuệ nhân tạo liên quan đến việc nghiên cứu và xây dựng các kĩ thuật cho phép các hệ thống "học" tự động từ dữ liệu để giải quyết những vấn đề cụ thể. Ví dụ như các máy có thể "học" cách phân loại thư điện tử xem có phải thư rác (spam) hay không và tự động xếp thư vào thư mục tương ứng. Học máy rất gần với suy diễn thống kê (statistical inference) tuy có khác nhau về thuật ngữ.

Một chương trình máy tính được cho là học hỏi từ kinh nghiệm của E đối với một số loại nhiệm vụ T và đo lường hiệu suất P, nếu hiệu quả của nó là những công việc ở T, được đo bằng P, cải thiện với kinh nghiệm E. [1]

Học máy hiện nay được áp dụng rộng rãi bao gồm máy truy tìm dữ liệu, chẩn đoán y khoa, phát hiện thẻ tín dụng giả, phân tích thị trường chứng khoán, phân loại các chuỗi DNA, nhận dạng tiếng nói và chữ viết, dịch tự động, chơi trò chơi và cử động rô-bốt (robot locomotion). [1]

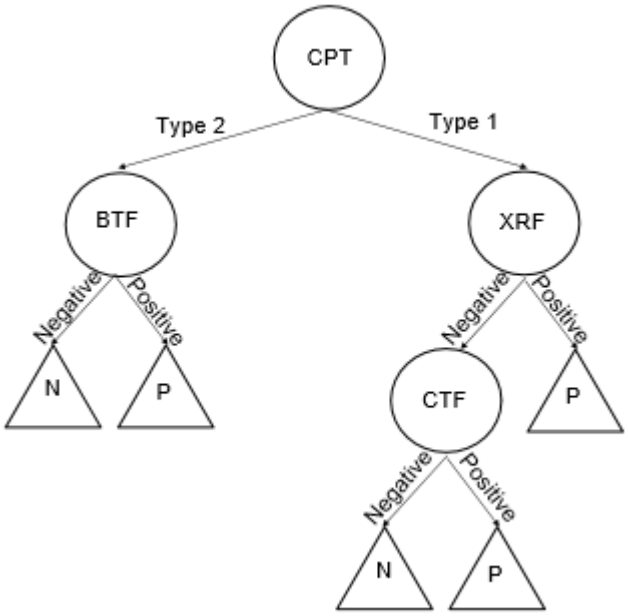
2.1.2 Cây quyết định

2.1.2.1 Tổng quan

Cây quyết định (decision tree) là một phương pháp rất mạnh và phổ biến cho cả hai nhiệm vụ của khai phá dữ liệu là phân loại và dự báo. Mặt khác, cây quyết định còn có thể chuyển sang dạng biểu diễn tương đương dưới dạng tri thức là các luật If-Then.

Cây quyết định là cấu trúc biểu diễn dưới dạng cây. Trong đó, mỗi nút trong (internal node) biểu diễn một thuộc tính, nhánh (branch) biểu diễn giá trị có thể có của thuộc tính, mỗi lá (leaf node) biểu diễn các lớp quyết định và đỉnh trên cùng của cây gọi là gốc (root). Cây quyết định có thể được dùng để phân lớp bằng cách xuất phát từ gốc của cây và di chuyển theo các nhánh cho đến khi gặp nút lá. Trên cơ sở phân lớp này chúng ta có thể chuyển đổi về các luật quyết định. [1]

VD: Minh họa quá trình chẩn đoán, sử dụng cây quyết định, bệnh nhân bị một vấn đề hô hấp nhất định. Các cây quyết định sử dụng các thuộc tính sau: CT finding (CTF); X-ray finding (XRF); loại đau ngực (CPT); và xét nghiệm máu finding (BTF). Các bác sĩ sẽ cho X-ray, nếu loại đau ngực là "1". Tuy nhiên, nếu loại đau ngực là "2", thì bác sĩ sẽ không chỉ định X-ray mà sẽ chỉ định xét nghiệm máu. Nhờ đó mà tổng chi phí cận lâm sàng sử dụng sẽ giảm (Hình 2.1). [3]



Hình 2.1. Ứng dụng cây quyết định trong y tế [3]

Một số thuật toán được sử dụng để xây dựng cây quyết định:

- ID3: xây dựng cây quyết định từ trên- xuống (top -down), tại mỗi nút chọn các thuộc tính tốt nhất phân loại các ví dụ huấn luyện. Quá trình này tiếp tục cho đến khi cây hoàn toàn phân loại các ví dụ huấn luyện, hoặc cho đến khi tất cả các thuộc tính đã được sử dụng. [1]
- C4.5: được phát triển và công bố bởi Quinlan vào năm 1996. Thuật toán C4.5 là một thuật toán được cải tiến từ thuật toán ID3 với việc cho phép xử lý trên tập dữ liệu có các thuộc tính số (numeric atributes) và làm việc được với tập dữ liệu bị thiếu và bị nhiễu. Nó thực hiện phân lớp tập mẫu dữ liệu theo chiến lược ưu tiên theo chiều sâu (Depth - First). Thuật toán xét tất cả các phép thử có thể để phân chia tập dữ liệu đã cho và chọn ra một phép thử có giá trị GainRatio tốt nhất. GainRatio là một đại lượng để đánh giá độ hiệu quả của thuộc tính dùng để thực hiện phép tách trong thuật toán để phát triển cây quyết định. [1]
- CART: được phát triển bởi Breiman et al (1984). Đặc trưng của CART là nó xây dựng cây nhị phân, mỗi nút trong cây có đúng hai cạnh đi ra. Việc chia tách được lựa chọn bằng cách sử dụng các tiêu chí Twoing Criteria và cây thu được cắt tỉa bởi Cost-Complexity. Khi sử dụng nó cũng cho phép người dùng phân phối xác suất trước. Một tính năng quan trọng của CART là khả năng tạo ra cây truy hồi. Trong cây hồi quy, các nút lá dự đoán một số thực và không phải là một lớp. Trong trường hợp hồi quy, CART tìm cách chia tách nhỏ nhất để giảm thiểu các dự đoán lỗi. Dự đoán trong mỗi nút lá được dựa trên trung bình trọng cho nút. [3]

Cây quyết định ngày nay được ứng dụng ở nhiều lĩnh vực trong đời sống xã hội. Một số lĩnh vực tiêu biểu như:

- Y tế: sử dụng cây quyết định để phân tích và đưa ra quyết định là một phần quan trọng của việc trợ giúp cho người ra quyết định chăm sóc sức khỏe phải cân đối để đạt tỉ lệ cao nhất giữa chất lượng chăm sóc với chi phí điều trị. [4]

- Kinh doanh: Việc sử dụng cây quyết định là phương pháp cho phép các dự án đầu tư được đánh giá khả năng hiệu quả (khả thi). Trong nhiều trường hợp, kết quả tương lai của các quyết định bị ảnh hưởng bởi những hành động hiện tại. Thông thường các quyết định được thực hiện mà không tính đến tác động lâu dài. Kết quả là, các quyết định ban đầu như thể hợp lý có thể gây ra những rủi ro bất lợi trong tương lai. Đối với các quyết định mà các khả năng xảy ra trong tương lai chưa được biết đến, sử dụng phương pháp cây quyết định hay còn gọi là "biểu đồ dòng chảy" là rất hiệu quả, tránh rủi ro dễ dàng và hỗ trợ ra quyết định trong quá trình đầu tư. [6]

2.1.2.2 Đánh giá việc ứng dụng cây quyết định trong y tế.

Cây quyết định là một kỹ thuật ra quyết định đáng tin cậy và hiệu quả mà cung cấp thông tin với độ chính xác cao dựa vào kiến thức thu thập được một cách đơn giản nhất. Khi sử dụng cây quyết định, quá trình ra quyết định của chính nó có thể được dễ dàng xác nhận bởi một chuyên gia. Vì những lý do đó, cây quyết định là đặc biệt thích hợp để hỗ trợ quá trình ra quyết định trong y học [4].

Từ năm 1999 đến năm 2008, Cục quản lý thực phẩm và dược phẩm Hoa Kỳ (FDA) đã phê chuẩn 259 loại thuốc mới để sử dụng tại Hoa Kỳ góp phần thay đổi các phương pháp chẩn đoán cho người bệnh, tỉ lệ của sự đổi mới trong lĩnh vực dược và liệu pháp chẩn đoán cung cấp hy vọng mới cho người bệnh và với mỗi phương pháp điều trị mới đi kèm với chi phí. Chăm sóc bệnh nhân toàn diện đòi hỏi lợi ích của phương pháp điều trị mới và cân đối với chi phí của họ. Mô hình phân tích quyết định là một phần quan trọng của việc trợ giúp cho người ra quyết định chăm sóc sức khỏe phải cân đối để đạt tỉ lệ cao nhất giữa chất lượng chăm sóc với chi phí điều trị. [5]

Từ những lý do trên, đề tài áp dụng phương pháp phân loại dùng cây quyết định vào việc chẩn đoán bệnh thận dựa vào các mẫu xét nghiệm của mỗi bệnh nhân. Hơn thế nữa, dựa vào các thuộc tính dữ liệu, giải thuật phân loại sẽ được cải tiến cho phù hợp nhằm nâng cao độ chính xác chẩn đoán.

2.2 Tổng quan về bệnh lý thận [8]

2.2.1 Khái niệm

Suy thận mạn là hậu quả các bệnh mãn tính của thận gây giảm sút từ số lượng Nephron chức năng làm giảm dần mức lọc cầu thận. Khi mức lọc cầu thận giảm xuống dưới 50% (60 ml/phút) thì được gọi là suy thận mạn.

Suy thận mạn là một hội chứng lâm sàng và sinh hóa tiến triển mạn tính qua nhiều tháng, năm, hậu quả của sự xơ hóa các Nephron chức năng gây giảm sút từ mức lọc cầu thận dẫn đến tình trạng tăng nitơ phi protein máu.

Theo PGS. TS Nguyễn Quốc Anh cho biết: "Theo thống kê của Hội Thận học Thế giới, trên thế giới ước tính khoảng 500 triệu người đang có vấn đề về bệnh lý mãn tính ở thận. Khoảng 3 triệu người bệnh trên thế giới đang sống nhờ các biện pháp thay thế. Tại Việt Nam chưa có số liệu thống kê chính thức song ước tính có khoảng 5 triệu người bị suy thận và hàng năm có khoảng 8.000 ca bệnh mới". [9]

2.2.2 Đặc trưng

Bệnh nhân có các biểu hiện:

- Có tiền sử bệnh thận tiết niệu kéo dài.
- Mức lọc cầu thận giảm.
- Nitơ phi protein máu tăng cao dần.
- Kết thúc trong hội chứng urê máu cao.

2.2.3 Đặc điểm dịch tễ học

Suy thận mạn là một bệnh tương đối phổ biến và hay gặp trong các bệnh thận tiết niệu. Theo thống kê của PGS. Trần Văn Chất và Trần Thị Thịnh (1991-1995) tại Khoa Tiết niệu Bệnh viện Bạch Mai thì suy thận mạn chiếm 40,4% và không thấy có sự khác biệt giữa nam và nữ. Riêng độ tuổi 16-24 thì thấy nam nhiều hơn nữ.

2.2.4 Chẩn đoán

2.2.4.1 Chẩn đoán xác định

- Suy thận mạn do bệnh cầu thận:
- Có tiền sử phù
- Phù - cao huyết áp - thiếu máu.
- Urê máu, creatinine máu cao, mức lọc cầu thận giảm.
- Protein niệu 2-3 g/24h.
- Suy thận mạn do bệnh viêm thận bể thận mạn.
- Có tiền sử nhiễm khuẩn tiết niệu.
- Cao huyết áp - thiếu máu.
- Urê máu, creatinine máu cao, mức lọc cầu thận giảm.
- Protein niệu có nhưng ít không quá l g/24h.
- Bạch cầu niệu bao giờ cũng có, vi khuẩn niệu có thể có hoặc không.

2.2.4.2 Chẩn đoán giai đoạn

Suy thận mạn gồm 5 giai đoạn tùy thuộc vào mức thanh trừ xuất Creatinine

Bảng 2.1. Bảng phân loại các giai đoạn bệnh thận

Giai đoạn suy thận mạn	**Mức lọc cầu thận (ml/phút)**	**Creatinine máu**		**Lâm sàng**
Bình thường	120	70 - 106	0,8 - 1,2	Bình thường
I	60 - 41	< 130	< 1,5	Gần bình thường
II	40 - 21	130 - 299	1,5 - 3,4	Gần bình thường, thiếu máu nhẹ
IIIa	20 - 11	300 - 499	3,5 - 5,9	Chán ăn, thiếu máu vừa
IIIb	10 - 5	500 - 900	6,0 - 1	Chán ăn, thiếu máu nặng, bắt đầu chỉ định lọc máu
IV	< 5	> 900	> 10	Hội chứng urê máu cao, lọc máu là bắt buộc.

Hoặc suy thận mạn có thể được chẩn đoán theo kết quả công thức Cockrofl và Gault cho phép ta tính được thanh trừ xuất Creatinine (tính bằng ml/phút) dựa trên tuổi (tính bằng năm), cân nặng (tính bằng kilô) và Creatinine (tính bằng µmol):

Công thức tính cho nam:

$$thanh\ trừ\ creatinine\ (Kr) = \frac{(140 - tuổi)\ x\ cân\ nặng}{0.814\ x\ creatinine\ máu}$$

Công thức tính cho nữ:

$$thanh\ trừ\ creatinine\ (Kr) = \frac{(140 - tuổi)\ x\ cân\ nặng}{0.85\ x\ creatinine\ máu}$$

Bảng 2.2. Bảng phân loại các giai đoạn bệnh thận theo Cockrofl và Gault

Giai đoạn	**Mô tả**	**GFR ml/phút/1,73 m^2**
1	Tổn thương thận nhẹ, lọc thận bình thường hoặc tăng	> 90
2	Chức năng thận giảm nhẹ	từ 60 đến 89
3	Chức năng thận giảm vừa phải	từ 30 đến 59
4	Chức năng thận giảm nặng	từ 15 đến 29
5	Suy thận cần phải lọc thận nhân tạo và ghép thận	< 15

Dựa vào triệu chứng lâm sàng: thiếu máu và cảm giác ăn ở tuyến cơ sở có thể chẩn đoán sớm được giai đoạn của suy thận mạn để ra quyết định điều trị sớm.

2.3 Các nghiên cứu liên quan đến đề tài

2.3.1 Chẩn đoán suy thận dựa vào hệ chuyên gia theo từng bệnh nhân [10]

Sử dụng phương pháp khai thác thông tin theo bộ câu hỏi định sẵn (ESKF) gồm 84 tiêu chí. ESKF dựa vào các bệnh án tương tự với bệnh án đang xét để suy ra thông tin có thể người dùng cung cấp thiếu. Độ tương tự của bệnh án I so với bệnh

án R được tính theo công thức (Hình 2.2). ESKF sử dụng môtơ suy luận ESS có tên là Jess.

ESKF dùng để chẩn đoán và theo dõi suy thận nên hồ sơ bệnh nhân dùng trong ESKF phải chứa đầy đủ các thông tin đặc trưng có liên quan đến suy thận bao gồm bốn nhóm thông tin: thông tin cá nhân, tiền căn của bệnh nhân, tiền căn gia đình của bệnh nhân và kết quả cận lâm sàng (Profile của bệnh nhân trong hệ chẩn đoán suy thận ESKF có tổng cộng 84 đặc trưng).

$$sim(I,R) = \frac{\sum_{i=1}^{n} w_i * sim(f_i^I, f_i^R)}{\sum_{i=1}^{n} w_i}$$

Với:

- I là bệnh án đang xét
- R là bệnh án được truy vấn hay tham khảo
- w_i là trọng số của sự kiện thứ i dùng để chẩn đoán khả năng trong bệnh án I
- f_i^I là sự kiện f_i trong bệnh án I
- f_i^R là sự kiện f_i trong bệnh án R
- $sim(f_i^I, f_i^R) = \begin{cases} 1, & f_i^I = f_i^R \\ 0, & f_i^I \neq f_i^R \end{cases}$

Hình 2.2. Công thức tính Độ tương tự của bệnh án theo ESKF.

Tuy nhiên phương pháp ESKF không yêu cầu làm xét nghiệm để chẩn đoán xác định dẫn đến chẩn đoán bị thiếu sót.

2.3.2 Hệ hỗ trợ chẩn đoán một số bệnh thông thường ở trẻ em [11]

Phương pháp ứng dụng của "hệ hỗ trợ chẩn đoán một số bệnh thông thường ở trẻ em" là khai thác thông tin của người sử dụng theo phương pháp trả lời câu hỏi dạng "Có/Không" theo các phát đồ điều trị để đưa người dùng đến với quyết định kết quả là có bệnh hay không. Các luật dùng trong suy luận của "hệ hỗ trợ chẩn đoán một số bệnh thông thường ở trẻ em" theo công thức (Hình 2.3).

Luật i:	**Nếu**	**< triệu chứng1, triệu chứng2 ...thỏa>**
(i=$\overline{1..n}$)	**Thì**	**<đưa ra kết quả về bệnh lý >**

Hình 2.3. Công thức sinh luật trong hệ hỗ trợ chẩn đoán một số bệnh thường gặp của trẻ em

Phương pháp này chưa khai thác đầy đủ các yếu tố cấu thành bệnh lý (bao gồm lâm sàng và cận lâm sàng) mà phụ thuộc vào hướng trả lời của người dùng để xác định bệnh lý nên không thể khai thác hết tình hình bệnh lý của bệnh nhân, dẫn đến có thể chẩn đoán sai lầm.

Chương 3: XÂY DỰNG HỆ THỐNG HỖ TRỢ CHẨN BỆNH THẬN VÀ ĐỀ XUẤT PHƯƠNG PHÁP ĐIỀU TRỊ

3.1 Phát biểu vấn đề

3.1.1 Tại sao chọn cây quyết định?

Cây quyết định là một kỹ thuật ra quyết định đáng tin cậy và hiệu quả, độ chính xác trong quá trình phân loại cao chỉ với một thuộc tính đơn giản của dữ liệu thu thập được. Khi sử dụng cây quyết định, quá trình ra quyết định có thể được dễ dàng xác nhận bởi một chuyên gia. Vì những lý do đó cây quyết định là đặc biệt thích hợp để hỗ trợ quá trình ra quyết định trong y học [4].

Bảng 3.1. Bảng phân loại các thuật toán trong cây quyết định

Reference	**description/ method**	**Induction approach**	**discretization method**	**Space partitioning**	**Num. decision attributes**
Quinlan, 1993	ID3, C4.5, …	heuristic	equidistant/ percentile/ dynamic	orthogonal	one
Babic, 2000	fuzzy classical	heuristic	fuzzy	orthogonal	one
Zorman, 1999	hubrid (MtDecit 2.0)	heuristic/ neural nets	dynamic	oblique	one
Breiman, 1984	classical (CART)	heuristic/ perturbations	equidistant/ percentile/ dynamic	oblique	one
Heath, 1993	classical (SADT)	heuristic/ simulated annealing	equidistant/ percentile/ dynamic	oblique	one
Murthy, 1997	classical (OC1)	heuristic/ random	equidistant/ percentile/ dynamic	oblique	one
Utgoff, 1989	classical incremental (ID5R)	heuristic/ incremental	equidistant/ percentile	orthogonal	one

Podgorelec, 2001	evolutionary (genTrees)	genetic algorithms	random	orthogonal	one
Sprogar, 2000	evolutionary vector (VEDEC)	genetic algorithms	random	orthogonal	any
Podgorelec, 2001	automatic programming (APEX)	genetic algorithms/ genetic programming	random	oblique	one

Có rất nhiều thuật toán trong cây quyết định đã được trình bày ngắn gọn trong bài báo nghiên cứu (Decision trees: an overview and their use in medicine) và được tóm tắt trong Bảng 3.1. Đương nhiên, không có các thuật toán nào là vượt trội so với những thuật toán khác; mỗi thuật toán có một số ưu điểm và nhược điểm. Để lựa chọn một thuật toán thích hợp cho một vấn đề cụ thể tốt nhất là nên sử dụng một số thuật toán khác nhau thay vì sử dụng một thuật toán duy nhất [4].

Cây quyết định chỉ đơn giản là trả lời cho một vấn đề dựa trên các yếu tố (thuộc tính) đặt ra trước và là một trong số ít các phương pháp có thể được trình bày một cách nhanh chóng, đủ để một người không chuyên xử lý dữ liệu và không cần biết về các công thức toán học. Trong bài viết so sánh về cây quyết định ID3 và C4.5 (A comparative study of decision tree ID3 and C4.5) [7], tác giả đã tập trung vào các yếu tố quan trọng để xây dựng một tập hợp các dữ liệu đồng thời tác giả đã trình bày các thuật toán ID3 và C4 và đã so sánh các thuật toán ID3 / C4.5, C4.5 / C5.0 và C5.0 / CART. Cuối cùng tác giả đã xác nhận rằng phương pháp mạnh mẽ nhất và được yêu thích trong máy học chắc chắn là C4.5 [7].

3.1.2 Tại sao sử dụng thuật toán C4.5 trong luận văn?

Giữa C4.5 và ID3 mỗi phương pháp có những ưu điểm khác nhau [7]:

- Thuật toán ID3 chọn các thuộc tính tốt nhất dựa trên khái niệm entropy và thông tin thu được để phát triển cây.
- Thuật toán C4.5 hoạt động tương tự như ID3 nhưng cải tiến nhược điểm của ID3:
- Khả năng sử dụng dữ liệu liên tục (dữ liệu số).
- Sử dụng các thuộc tính dữ liệu không xác định (bị lỗi).
- Xác định và sử dụng trọng số (trọng lượng) cho các thuộc tính khác nhau.
- Tỉa cây sau khi được tạo ra.

Vì vậy, việc chọn C4.5 làm phương pháp học máy cho dữ liệu (kết quả cận lâm sàng của bệnh nhân) sử dụng trong luận văn này vì C4.5 có những ưu thế sau:

- C4.5 là thuật toán cải tiến của ID3.
- C4.5 xử lý tốt hơn các giá trị mang tính liên tục (giá trị kết quả cận lâm sàng của bệnh nhân là giá trị số liên tục) mà nội dung luận văn này nghiên cứu.
- C4.5 cho phép thao tác với các thuộc tính có dữ liệu không xác định (do bị mất mát dữ liệu, …) phù hợp với trường hợp bệnh nhân bị khuyết một vài thuộc tính trong lúc chỉ định điều trị.
- C4.5 đưa ra phương pháp "cắt tỉa" cây và giản lược các luật để phù hợp với những bộ dữ liệu lớn (dữ liệu bệnh nhân có khoảng hơn 140.000 mẫu).

3.2 Các thông số, qui ước, thuộc tính sử dụng

Bảng 3.2. Bảng các qui ước, thuộc tính sử dụng

TT	Thuộc tính	Diễn giải
1	gioi_tinh	Thể hiện giới tính của bệnh nhân và được qui ước giá trị : 1 (nam), 2 (nữ)
2	muc_creatinine	Là mức phân loại dựa theo bảng giá trị phân loại các giai đoạn bệnh của bệnh nhân có bệnh lý thận nội khoa (Bảng 2.1)
3	muc_tuoi	Là độ tuổi của bệnh nhân được nhóm theo từng nhóm bệnh nhân trong khoảng từ x0 đến x9 (với x là mức tuổi)
4	muc_urea	Là phân đoạn chỉ số urea theo công thức : muc_urea = urea / 7.5 (chỉ số ngưỡng bình thường cao nhất) Ví dụ: muc_urea = 0 thì chỉ số urea của bệnh nhân từ 0 đến 7.5.
5	muc_thanh_thai	Là chỉ số phân loại bệnh thận theo từng giai đoạn dựa trên công thức của Cockrofl và Gault (Bảng 2.2)
6	ket_qua	Là mã chẩn đoán bệnh lý dựa vào bảng phân loại bệnh lý quốc tế (ICD 10)

Bảng 3.3. Bảng các qui ước các phương pháp đánh giá thuật toán

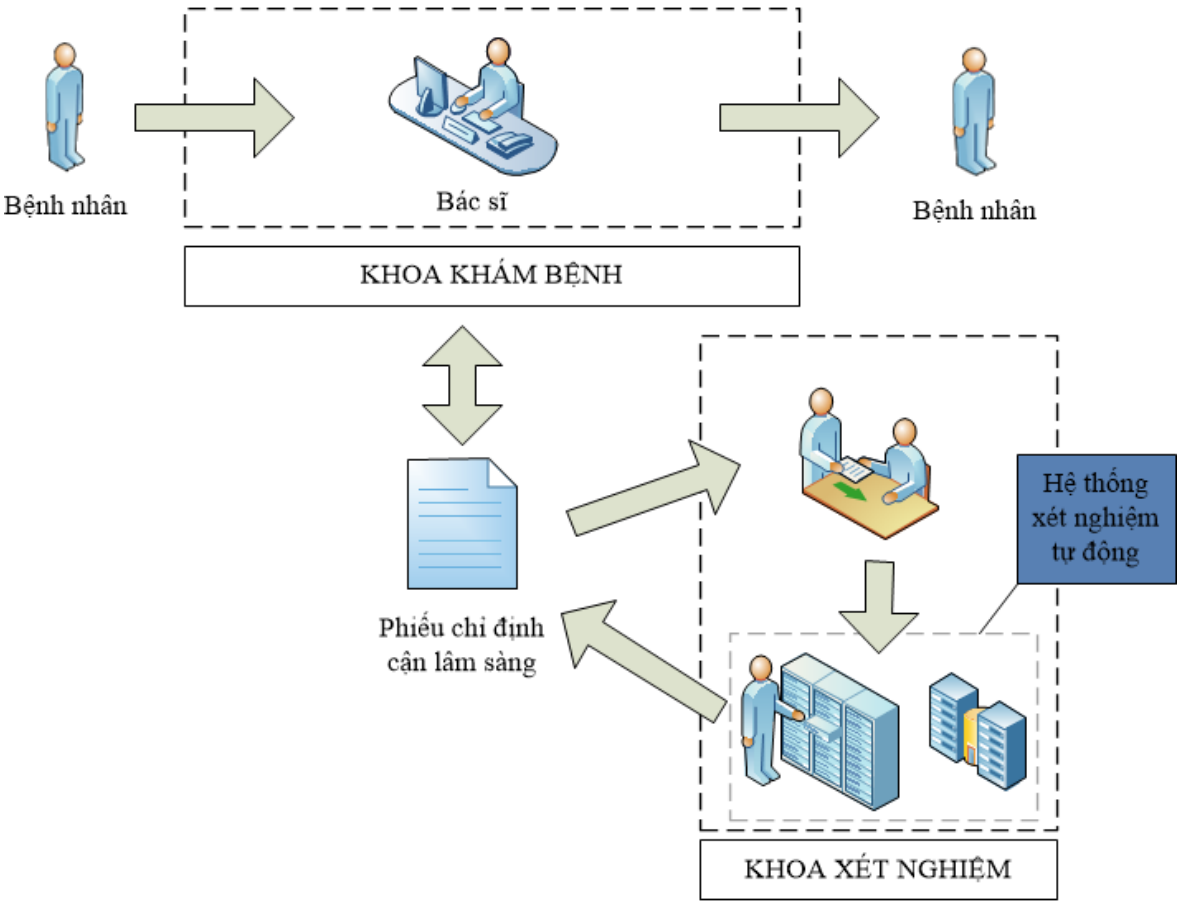
1	k-fold cross validation (k=10)	Là phương pháp đánh gia mô hình bằng cách kiểm chứng chéo. Với k = 10, nghĩa là chia tập dữ liệu thành 10 phần, 1 phần dùng làm tập kiểm tra (test set), 9 phần dùng để huấn luyện (train set).
2	Hold-out (split 66.0% train, remainder test)	Là phương pháp đánh giá mô hình theo cách chia dữ liệu thành 2 phần: 66% để xây dựng mô hình phân lớp (tập train), 34% để kiểm tra (tập test).

3.3 Qui trình chẩn đoán bệnh

Qui trình khám và chẩn đoán bệnh bao gồm :

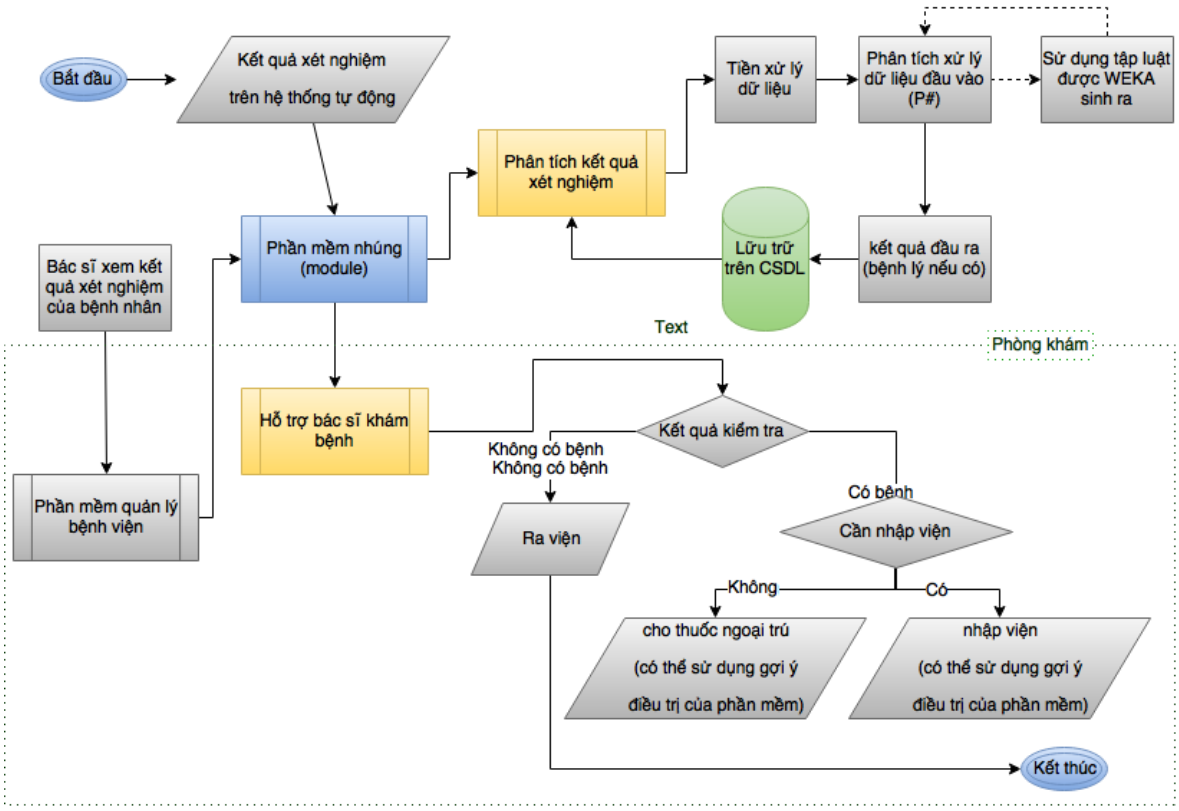
- Bước 1: Bác sĩ khám, hỏi thăm thông tin người bệnh.
- Bước 2: Chỉ định cận lâm sàng để hỗ trợ chẩn đoán và làm chứng cứ y khoa.
- Bước 3: Cận lâm sàng (xét nghiệm) được đưa vào hệ thống xét nghiệm tự động phân tích và trả kết quả trên hệ thống quản lý bệnh viện
- Bước 4: Bác sĩ dựa vào các kết quả cận lâm sàng để chẩn đoán bệnh.
- Bước 5: Ra quyết định điều trị.

Mô hình bệnh nhân đến khám và chữa bệnh (Hình 3.1):



Hình 3.1. Mô hình khám và chẩn đoán bệnh

Dựa vào qui trình chẩn đoán bệnh trên, luận văn này đề xuất hệ thống hỗ trợ chẩn đoán bệnh và đề xuất phương pháp điều trị. Hệ thống cho phép bác sĩ dễ dàng nắm bắt thông tin và nhanh chóng xác định bệnh dựa vào kết quả phân tích dữ liệu của hệ thống, từ đó có thể đề xuất hướng điều trị phù hợp. Sau đây là khung làm việc của hệ thống (Hình 3.2).



Hình 3.2. Mô hình khung làm việc của hệ thống

Dựa vào mô hình trên, phần mềm nhúng (module tích hợp) là trung gian thực hiện các thao tác xử lý thông tin của bệnh nhân và trả kết quả hỗ trợ bác sĩ trong quá trình khám chữa bệnh.

3.4 Phương pháp hỗ trợ điều trị

Dựa vào phác đồ điều trị [8] để hiển thị đề xuất các hướng điều trị cho bệnh nhân dựa vào kết quả phân tích kết quả xét nghiệm của bệnh nhân. Bảng đề xuất các hướng điều trị dựa trên phát đồ như :

Bảng 3.4. Bảng hướng dẫn điều trị theo phát đồ

Chẩn đoán	**Mức creatinine**	**Hướng dẫn điều trị**
N18	1 – 2	+ Ăn ít đạm hơn bình thường. + Điều chỉnh huyết áp: Aldomet 250mg x 2-4 viên/24h, có thể dùng Propranolol, Nifedipin... + Ăn nhạt nếu có phù và cao huyết áp. + Lợi tiểu nếu có phù và tăng huyết áp. Thuốc đề nghị sử dụng: Furosemid/ Nifedipin, …
	3	+ Chế độ ăn là biện pháp chủ đạo để hạn chế mê máu tăng, protid = 0,5kg/24h, đảm bảo vitamin, tăng cầm bằng bột ít đạm. Đảm bảo các acid amin bằng trứng, sữa trong thức ăn. Ở cuối giai đoạn III chỉ nên cho với một người sống: 20g protid đảm bảo 1800 - 2000 calo/24h. + Muối: ăn nhạt khi có phù, cao huyết áp. + Nước: chỉ uống bằng lượng nước tiểu 24h. + Kali: giai đoạn đầu thường không tăng kali máu, ở cuối giai đoạn III có thể tăng kali máu nên hạn chế các rau quả và thức ăn có nhiều kali. + Calci: cho vitamin D và calci khi có calci máu giảm. + Kiềm: cho khi có toan máu. + Trợ tim: không dùng kéo dài, giảm liều lượng khi có suy thận nặng.

		+ Chống thiếu máu: có thể truyền máu, khối hồng cầu, cho viên sắt, Erythropoietin. Thuốc đề nghị sử dụng: Furosemid/ Nifedipin/ Calci gluconat/ Erythropoietin/ Vitamin D, …
	4-5	Lọc máu nhân tạo. Thuốc đề nghị sử dụng: Erythropoietin
N17		Thuốc đề nghị sử dụng: Amlodipin/ Bromhexin/ Calci gluconat/ Captopril/ Cefepim* hoặc Ceftazidim hoặc Ciprofloxacin / Drotaverin clohydrat/ Enoxaparin (natri)/ Furosemid/ Glyceryl trinitrat

3.5 Phương pháp xử lý dữ liệu đầu vào

Dữ liệu thô về bệnh nhân được khai thác có dạng.

Bảng 3.5. Dữ liệu cận lâm sàng của bệnh nhân

Bệnh án	ID Bệnh Nhân	ID Xét Nghiệm	Họ Tên	Giới tính	Tuổi	creatinine	urea	Cân nặng	Chẩn Đoán
13174024	13105755	2413652	Lê Thị H...	Nữ	55	75	5.8	64	K62.1
13174747	13106132	2413653	Nguyễn Như V...	Nam	53	213	5.9	53	N18

Dữ liệu trên được thu thập từ 144.761 mẫu kết quả xét nghiệm của 93.997 bệnh nhân với 111.424 lượt khám và điều trị.

Dữ liệu sau khi tiền xử lý (Bảng 3.6) theo bảng qui ước (Bảng 3.2).

Bảng 3.6. Dữ liệu sau tiền xử lý thông tin

ID Xét Nghiệm	gioi_tinh	muc_tuoi	muc_creatinine	muc_urea	muc_thanh_thai	ket_qua
2413652	2	5	0	5.8	0	0
2413653	1	5	2	5.9	3	N18

Chuyển dữ liệu bệnh nhân sau khi tiền xử lý thành tập tin dữ liệu phù hợp sử dụng cho phần mềm khai thác dữ liệu WEKA (tập tin .csv).

Dữ liệu sau khi được tiền xử lý sẽ được đưa vào thành các tập thuộc tính, với mỗi tập thuộc tính là một mẫu thử nghiệm để đánh giá xem tập thuộc tính nào sẽ đem lại kết quả phân tích đạt hiệu quả nhất. Các tập thuộc tính gồm:

- Tập thứ 1: muc_creatinine, ket_qua.
- Tập thứ 2: muc_creatinine, muc_urea, ket_qua.
- Tập thứ 3: muc_creatinine, muc_urea, gioi_tinh, ket_qua.
- Tập thứ 4: muc_creatinine, muc_urea, muc_tuoi, ket_qua.
- Tập thứ 5: muc_creatinine, muc_urea, muc_tuoi, gioi_tinh, ket_qua.
- Tập thứ 6: muc_creatinine, tinh.
- Tập thứ 7: muc_creatinine (với muc_creatinine = 1), muc_urea, muc_tuoi, gioi_tinh, ket_qua.
- Tập thứ 8: muc_creatinine (với muc_creatinine = 2), muc_urea, muc_tuoi, gioi_tinh, ket_qua.
- Tập thứ 9: muc_creatinine, muc_thanh_thai, ket_qua.
- Tập thứ 10: muc_creatinine, muc_urea, muc_thanh_thai, ket_qua.
- Tập thứ 11: gioi_tinh, muc_creatinine, muc_tuoi, muc_urea, muc_thanh_thai, ket_qua.

3.6 Phương pháp khai thác dữ liệu

Ứng dụng phần mềm WEKA và thuật toán cây quyết định C4.5 (J48 trong WEKA) và CART để tìm hiểu các luật kết hợp giữa các thuộc tính trong của bệnh nhân.

Trong quá trình thử nghiệm các mẫu dữ liệu, với việc so sánh kết quả thử nghiệm của các phương pháp đánh giá khác nhau của hai thuật toán C4.5 và CART, thực nghiệm đã chỉ ra rằng sử dụng Phương pháp đánh giá: k fold cross-validation để đánh giá hiệu quả của thuật toán là tối ưu hơn. Kết quả đánh giá được thể hiện trong Bảng 4.1.

3.7 Các mẫu dữ liệu thử nghiệm

3.7.1 Mẫu thử nghiệm thứ 1a

Số lượng 144.761 mẫu với các thuộc tính : muc_creatinine, ket_qua.

Bảng 3.7. Kết quả phân tích dữ liệu (Mẫu thử nghiệm thứ 1a)

=== Run information ===

Instances: 144761

Attributes: 2

muc_creatinine

ket_qua

Test mode:split 66.0% train, remainder test

=== Classifier model (full training set) ===

J48 pruned tree

---

muc_creatinine <= 1: 0.0 (116800.0/1541.0)

muc_creatinine > 1

| *muc_creatinine <= 2: 0.0 (17390.0/5666.0)*

| *muc_creatinine > 2: N18 (10571.0/535.0)*

Number of Leaves : 3

Size of the tree : 5

Time taken to build model: 0.47 seconds

=== Stratified cross-validation ===

=== Summary ===

Correctly Classified Instances	*46547*	*94.5712 %*
Incorrectly Classified Instances	*2672*	*5.4288 %*

=== Detailed Accuracy By Class ===

	TP Rate	FP Rate	Precision	Recall	F-Measure	Class
	0.629	0.004	0.952	0.629	0.758	N18
	1	0.406	0.945	1	0.972	0
	0	0	0	0	0	N17
	0	0	0	0	0	N19
Weighted Avg.	0.946	0.356	0.934	0.946	0.935	

=== Confusion Matrix ===

a	b	c	d	<-- classified as
3475	2047	0	0	| a = N18
0	43072	0	0	| b = 0
166	419	0	0	| c = N17
9	31	0	0	| d = N19

Dựa vào kết quả phân tích, sau khi lượt bỏ các luật trùng lặp, ta rút ra được bộ luật:

- Luật 1: Nếu muc_creatinine từ 1 đến 2 thì ⇨ (kết luận) không có bệnh.
- Luật 2: Nếu muc_creatinine >= 3 thì ⇨ (kết luận) N18.

3.7.2 Mẫu thử nghiệm thứ 1b

Sử dụng lại tập dữ liệu trên (Mẫu thử nghiệm thứ 1a) nhưng sử dụng thuật toán CART với cùng phương pháp đánh giá như Mẫu thử nghiệm thứ 1a (split 66.0% train, remainder test) thay cho thuật toán C4.5 để đánh giá lại độ hiệu quả của các thuật toán.

Bảng 3.8. Kết quả phân tích dữ liệu (Mẫu thử nghiệm thứ 1b)

=== Run information ===

Instances: 144761

Attributes: 2

muc_creatinine

ket_qua

Test mode:split 66.0% train, remainder test

=== Classifier model (full training set) ===

CART Decision Tree

---

muc_creatinine < 2.5: 0.0(126983.0/7207.0)

muc_creatinine >= 2.5: N18(10036.0/535.0)

Number of Leaf Nodes: 2

Size of the Tree: 3

Time taken to build model: 8.91 seconds

=== Stratified cross-validation ===

=== Summary ===

Correctly Classified Instances	*46547*	***94.5712 %***
Incorrectly Classified Instances	*2672*	*5.4288 %*

=== Detailed Accuracy By Class ===

	TP Rate	FP Rate	Precision	Recall	F-Measure	Class
	0.629	0.004	0.952	0.629	0.758	N18
	1	0.406	0.945	1	0.972	0
	0	0	0	0	0	N17
	0	0	0	0	0	N19
Weighted Avg.	0.946	0.356	0.934	0.946	0.935	

=== Confusion Matrix ===

a	b	c	<-- classified as
3475	2047	0	| a = N18
0	43072	0	| b = 0
166	419	0	| c = N17
9	31	0	| c = N19

Dựa vào kết quả phân tích, sau khi lượt bỏ các luật trùng lặp, ta rút ra được bộ luật:

- Luật 1: Nếu muc_creatinine < 2.5 thì ⇨ (kết luận) không có bệnh.
- Luật 2: Nếu muc_creatinine >= 2.5 thì ⇨ (kết luận) N18.

Với hai thuật toán học máy C4.5 và CART với cùng phương pháp đánh giá (split 66.0% train, remainder test) thì độ chính xác được ghi nhận là như nhau (94.5712%)

3.7.3 Mẫu thử nghiệm thứ 1c

Sử dụng lại tập dữ liệu trên (Mẫu thử nghiệm thứ 1a) nhưng với cùng thuật toán C4.5 nhưng thay đổi phương pháp đánh giá từ phân sử liệu làm 3 phần (split 66.0% train, remainder test) 2/3 để học và 1/3 để kiểm tra thành chia dữ liệu làm 10 phần để luân phiên học và kiểm tra (k-fold cross-validation, với k = 10) để đánh giá lại độ hiệu quả của phương pháp đánh giá.

Bảng 3.9. Kết quả phân tích dữ liệu(Mẫu thử nghiệm thứ 1c)

=== Run information ===

Instances: 144761

Attributes: 2

muc_creatinine

ket_qua

Test mode:10-fold cross-validation

=== Classifier model (full training set) ===

J48 pruned tree

---

muc_creatinine <= 1: 0.0 (116800.0/1541.0)

muc_creatinine > 1

| muc_creatinine <= 2: 0.0 (17390.0/5666.0)

| muc_creatinine > 2: N18 (10571.0/535.0)

Number of Leaves : 3

Size of the tree : 5

Time taken to build model: 0.28 seconds

=== Stratified cross-validation ===

=== Summary ===

Correctly Classified Instances	*137019*	***94.6519 %***
Incorrectly Classified Instances	*7742*	*5.3481 %*

=== Detailed Accuracy By Class ===

	TP Rate	FP Rate	Precision	Recall	F-Measure	Class
	0.629	0.004	0.949	0.629	0.756	N18
	1	0.405	0.946	1	0.972	0
	0	0	0	0	0	N17
	0	0	0	0	0	N19
Weighted Avg.	0.947	0.356	0.935	0.947	0.936	

=== Confusion Matrix ===

a	b	c	<-- classified as
10036	5928	0	| a = N18
0	126983	0	| b = 0
499	1185	0	| c = N17
36	94	0	| c = N19

Dựa vào kết quả phân tích, sau khi lượt bỏ các luật trùng lặp, ta rút ra được bộ luật:

- Luật 1: Nếu muc_creatinine <= 2 thì ⇨ (kết luận) không có bệnh.
- Luật 2: Nếu muc_creatinine > 2 thì ⇨ (kết luận) N18.

Dựa vào kết quả phân tích cho thấy khi sử dụng phương pháp đánh giá k-fold cross-validation, với k = 10 thì độ chính xác khi phân nhóm bằng thuật toán C4.5 cao hơn (đạt 94.6519 %) khi sử dụng phương pháp đánh giá bằng cách phân sử liệu làm 3 phần (split 66.0% train, remainder test) 2/3 để học và 1/3 để kiểm tra (chỉ đạt 94.5712%)

3.7.4 Mẫu thử nghiệm thứ 2

Số lượng 144.761 mẫu; bổ sung thuộc tính muc_urea (so với Mẫu thử nghiệm thứ 1c) để xem xét sự tương quan giữa việc tăng urea máu và mức tăng creatinine. Các thuộc tính gồm : muc_creatinine, muc_urea, ket_qua.

Bảng 3.10. Kết quả phân tích dữ liệu(Mẫu thử nghiệm thứ 2).

```
=== Run information ===
Instances:    144761
Attributes:   3
              muc_creatinine
              muc_urea
              ket_qua
Test mode:10-fold cross-validation
=== Classifier model (full training set) ===
J48 pruned tree
------------------
muc_creatinine <= 1: 0.0 (116800.0/1541.0)
muc_creatinine > 1
|   muc_creatinine <= 2
```

| | *muc_urea <= 1: 0.0 (12723.0/3759.0)*

| | *muc_urea > 1*

| | | *muc_urea <= 3: 0.0 (4297.0/1671.0)*

| | | *muc_urea > 3: N18 (370.0/212.0)*

| *muc_creatinine > 2: N18 (10571.0/535.0)*

Number of Leaves : 5

Size of the tree : 9

Time taken to build model: 0.89 seconds

=== Stratified cross-validation ===

=== Summary ===

Correctly Classified Instances	*137016*	***94.6498 %***
Incorrectly Classified Instances	*7745*	*5.3502 %*

=== Detailed Accuracy By Class ===

	TP Rate	FP Rate	Precision	Recall	F-Measure	Class
	0.635	0.005	0.936	0.635	0.757	N18
	0.999	0.397	0.947	0.999	0.973	0
	0	0	0	0	0	N17
	0	0	0	0	0	N19
Weighted Avg.	0.946	0.349	0.934	0.946	0.937	

=== Confusion Matrix ===

a	b	c	d	<-- classified as
10135	5829	0	0	| a = N18
102	126881	0	0	| b = 0
551	1133	0	0	| c = N17
38	92	0	0	| d = N19

Dựa vào kết quả phân tích, sau khi lượt bỏ các luật trùng lặp, ta rút ra được bộ luật:

- Luật 1: Nếu muc_creatinine từ = 1 thì ⇨ (kết luận) không có bệnh.
- Luật 2: Nếu muc_creatinine = 2 và muc_urea <= 3 thì ⇨ (kết luận) không có bệnh.
- Luật 3: Nếu muc_creatinine = 2 và muc_urea > 3 thì ⇨ (kết luận) N18.
- Luật 4: Nếu muc_creatinine >= 3 thì ⇨ (kết luận) N18.

3.7.5 Mẫu thử nghiệm thứ 3

Số lượng 144.761 mẫu; bổ sung thuộc tính gioi_tinh (so với Mẫu thử nghiệm thứ 2) để xem xét thuộc tính gioi_tinh có ảnh hưởng kết quả của sự tương quan giữa việc tăng urea máu và mức tăng creatinine. Các thuộc tính bao gồm: muc_creatinine, muc_urea, gioi_tinh, ket_qua.

Bảng 3.11. Kết quả phân tích dữ liệu (Mẫu thử nghiệm thứ 3)

```
=== Run information ===
Instances:    144761
Attributes:   4
              gioi_tinh
              muc_creatinine
              muc_urea
              ket_qua
Test mode:10-fold cross-validation
=== Classifier model (full training set) ===
J48 pruned tree
------------------

muc_creatinine <= 1: 0.0 (116800.0/1541.0)
muc_creatinine > 1
|   muc_creatinine <= 2: 0.0 (17390.0/5666.0)
|   muc_creatinine > 2: N18 (10571.0/535.0)
Number of Leaves  :   3
Size of the tree :    5
Time taken to build model: 2.17 seconds
=== Stratified cross-validation ===
=== Summary ===
Correctly Classified Instances         137019               94.6519 %
Incorrectly Classified Instances       7742                 5.3481 %
=== Detailed Accuracy By Class ===
```

	TP Rate	FP Rate	Precision	Recall	F-Measure	Class
	0.629	0.004	0.949	0.629	0.756	N18
	1	0.405	0.946	1	0.972	0
	0	0	0	0	0	N17
	0	0	0	0	0	N19
Weighted Avg.	0.947	0.356	0.935	0.947	0.936	

=== Confusion Matrix ===

a	b	c	d	<-- classified as
10036	5928	0	0	| a = N18
0	126983	0	0	| b = 0
499	1185	0	0	| c = N17
36	94	0	0	| d = N19

Dựa vào kết quả phân tích, sau khi lượt bỏ các luật trùng lặp, ta rút ra được bộ luật:

- Luật 1: Nếu muc_creatinine từ 1 đến 2 thì ⇨ (kết luận) không có bệnh.
- Luật 2: Nếu muc_creatinine >= 3 thì ⇨ (kết luận) N18.

Với bộ dữ liệu gồm các thuộc tính : muc_creatinine, muc_urea, gioi_tinh **thì việc kết luận xác định bệnh lý không phụ thuộc vào thuộc tính : muc_urea và gioi_tinh**.

3.7.6 Mẫu thử nghiệm thứ 4

Số lượng 144.761 mẫu; bổ sung thuộc tính muc_tuoi (so với Mẫu thử nghiệm thứ 2) để xem xét thuộc tính muc_tuoi có ảnh hưởng kết quả của sự tương quan giữa việc tăng urea máu và mức tăng creatinine. Các thuộc tính gồm: muc_creatinine, muc_urea, muc_tuoi, ket_qua.

Bảng 3.12. Kết quả phân tích dữ liệu (Mẫu thử nghiệm thứ 4)

```
=== Run information ===
Instances:    144761
Attributes:   4
              muc_creatinine
              muc_tuoi
              muc_urea
              ket_qua
Test mode:10-fold cross-validation
=== Classifier model (full training set) ===
J48 pruned tree
------------------
muc_urea <= 1
|   muc_creatinine <= 1: 0.0 (115346.0/1316.0)
|   muc_creatinine > 1
|   |   muc_creatinine <= 2: 0.0 (12723.0/3759.0)
|   |   muc_creatinine > 2: N18 (1748.0/130.0)
muc_urea > 1
|   muc_creatinine <= 2
|   |   muc_creatinine <= 1: 0.0 (1454.0/225.0)
|   |   muc_creatinine > 1
|   |   |   muc_urea <= 3: 0.0 (4297.0/1671.0)
|   |   |   muc_urea > 3
|   |   |   |   muc_tuoi <= 7
|   |   |   |   |   muc_tuoi <= 6: 0.0 (148.0/62.0)
|   |   |   |   |   muc_tuoi > 6: N18 (126.0/58.0)
|   |   |   |   muc_tuoi > 7
|   |   |   |   |   muc_tuoi <= 8: N17 (75.0/33.0)
|   |   |   |   |   muc_tuoi > 8: N18 (21.0/5.0)
|   muc_creatinine > 2: N18 (8823.0/405.0)
```

Number of Leaves : 10

Size of the tree : 19

Time taken to build model: 1.41 seconds

=== Stratified cross-validation ===

=== Summary ===

Correctly Classified Instances	*137093*	***94.703 %***
Incorrectly Classified Instances	*7668*	*5.297 %*

=== Detailed Accuracy By Class ===

	TP Rate	FP Rate	Precision	Recall	F-Measure	Class
	0.634	0.005	0.943	0.634	0.759	N18
	1	0.396	0.948	1	0.973	0
	0.019	0	0.542	0.019	0.037	N17
	0	0	0	0	0	N19
Weighted Avg.	0.947	0.348	0.941	0.947	0.937	

=== Confusion Matrix ===

a	b	c	d	<-- classified as
10126	5812	26	0	| a = N18
48	126935	0	0	| b = 0
523	1129	32	0	| c = N17
37	92	1	0	| d = N19

Dựa vào kết quả phân tích, sau khi lượt bỏ các luật trùng lặp, ta rút ra được bộ luật:

- Luật 1: Nếu muc_creatinine từ < 2 thì ⇨ (kết luận) không có bệnh.
- Luật 2: Nếu muc_creatinine = 2 và muc_urea <= 3 thì ⇨ (kết luận) không có bệnh.
- Luật 3: Nếu muc_creatinine = 2 và muc_urea > 3 và muc_tuoi < 7 thì ⇨ (kết luận) không có bệnh.
- Luật 4: Nếu muc_creatinine = 2 và muc_urea > 3 và muc_tuoi = 7 thì ⇨ (kết luận) N18.

- Luật 5: Nếu muc_creatinine = 2 và muc_urea > 3 và muc_tuoi = 8 thì ⇨ (kết luận) N17.
- Luật 6: Nếu muc_creatinine = 2 và muc_urea > 3 và muc_tuoi > 8 thì ⇨ (kết luận) N18.
- Luật 7: Nếu muc_creatinine >= 3 thì ⇨ (kết luận) N18.

3.7.7 Mẫu thử nghiệm thứ 5

Số lượng 144.761 mẫu; bổ sung thuộc tính gioi_tinh (so với Mẫu thử nghiệm thứ 4) để xem xét thuộc tính gioi_tinh có ảnh hưởng kết quả của sự tương quan giữa việc tăng urea máu, mức tăng creatinine, giới tính và tuổi của bệnh nhân. Các thuộc tính gồm: muc_creatinine, muc_urea, muc_tuoi, gioi_tinh, ket_qua.

Bảng 3.13. Kết quả phân tích dữ liệu (Mẫu thử nghiệm thứ 5)

```
=== Run information ===
Instances:    144761
Attributes:   5
              gioi_tinh
              muc_creatinine
              muc_tuoi
              muc_urea
              ket_qua
Test mode:10-fold cross-validation
=== Classifier model (full training set) ===
J48 pruned tree
------------------
muc_urea <= 1
|   muc_creatinine <= 1: 0.0 (115346.0/1316.0)
|   muc_creatinine > 1
|   |   muc_creatinine <= 2: 0.0 (12723.0/3759.0)
|   |   muc_creatinine > 2: N18 (1748.0/130.0)
muc_urea > 1
```

| *muc_creatinine <= 2*
| | *muc_creatinine <= 1: 0.0 (1454.0/225.0)*
| | *muc_creatinine > 1*
| | | *muc_urea <= 3: 0.0 (4297.0/1671.0)*
| | | *muc_urea > 3*
| | | | *muc_tuoi <= 7*
| | | | | *gioi_tinh <= 1: 0.0 (159.0/73.0)*
| | | | | *gioi_tinh > 1*
| | | | | | *muc_tuoi <= 4: 0.0 (14.0)*
| | | | | | *muc_tuoi > 4: N18 (101.0/40.0)*
| | | | *muc_tuoi > 7*
| | | | | *gioi_tinh <= 1: N18 (53.0/19.0)*
| | | | | *gioi_tinh > 1*
| | | | | | *muc_tuoi <= 8: N17 (35.0/9.0)*
| | | | | | *muc_tuoi > 8: N18 (8.0/3.0)*
| *muc_creatinine > 2: N18 (8823.0/405.0)*

Number of Leaves : 12

Size of the tree : 23

Time taken to build model: 2.33 seconds

=== Evaluation on test split ===

=== Summary ===

Correctly Classified Instances	*137100*	*94.7078 %*
Incorrectly Classified Instances	*7661*	*5.2922 %*

=== Detailed Accuracy By Class ===

	TP Rate	FP Rate	Precision	Recall	F-Measure	Class
	0.634	0.005	0.945	0.634	0.759	N18
	1	0.396	0.947	1	0.973	0
	0.015	0	0.578	0.015	0.03	N17
	0	0	0	0	0	N19
Weighted Avg.	0.947	0.348	0.942	0.947	0.937	

=== Confusion Matrix ===

a	b	c	d	<-- classified as
10123	5823	18	0	| a = N18
31	126951	1	0	| b = 0
525	1133	26	0	| c = N17
38	92	0	0	| d = N19

Dựa vào kết quả phân tích, sau khi lượt bỏ các luật trùng lặp, ta rút ra được bộ luật:

- Luật 1: Nếu muc_creatinine từ 1 đến 2 thì ⇨ (kết luận) không có bệnh.
- Luật 2: Nếu muc_creatinine = 2 và muc_urea <= 3 thì ⇨ (kết luận) không có bệnh.
- Luật 3: Nếu muc_creatinine = 2 và muc_urea > 3 và muc_tuoi <= 7 và gioi_tinh = 1 thì ⇨ (kết luận) không có bệnh
- Luật 4: Nếu muc_creatinine = 2 và muc_urea > 3 và muc_tuoi từ 5 đến 7 và gioi_tinh = 2 thì ⇨ (kết luận) N18.
- Luật 5: Nếu muc_creatinine = 2 và muc_urea > 3 và muc_tuoi = 8 và gioi_tinh = 1 thì ⇨ (kết luận) N18.
- Luật 6: Nếu muc_creatinine = 2 và muc_urea > 3 và muc_tuoi = 8 và gioi_tinh = 2 thì ⇨ (kết luận) N17.
- Luật 7: Nếu muc_creatinine = 2 và muc_urea > 3 và muc_tuoi > 8 và gioi_tinh = 2 thì ⇨ (kết luận) N18.
- Luật 8: Nếu muc_creatinine >= 3 thì ⇨ (kết luận) N18.

3.7.8 Mẫu thử nghiệm thứ 6

Số lượng 144.761 mẫu; bổ sung thuộc tính tinh (so với Mẫu thử nghiệm thứ 1a) để xem xét có sự thay đổi chẩn đoán đối với các bệnh nhân ở từng khu vực dân cư khác nhau. Các thuộc tính gồm : muc_creatinine, tinh.

Bảng 3.14. Kết quả phân tích dữ liệu (Mẫu thử nghiệm thứ 6)

=== Run information ===

Instances: 144761

Attributes: 3

muc_creatinine

ket_qua

tinh

Test mode:10-fold cross-validation

=== Classifier model (full training set) ===

J48 pruned tree

----

muc_creatinine <= 1: 0.0 (116800.0/1541.0)

muc_creatinine > 1

| muc_creatinine <= 2: 0.0 (17390.0/5666.0)

| muc_creatinine > 2: N18 (10571.0/535.0)

Number of Leaves : 3

Size of the tree : 5

Time taken to build model: 0.28 seconds

=== Stratified cross-validation ===

=== Summary ===

Correctly Classified Instances	*137019*	*94.6519 %*
Incorrectly Classified Instances	*7742*	*5.3481 %*

=== Detailed Accuracy By Class ===

	TP Rate	FP Rate	Precision	Recall	F-Measure	Class
	0.629	0.004	0.949	0.629	0.756	N18
	1	0.405	0.946	1	0.972	0
	0	0	0	0	0	N17
	0	0	0	0	0	N19
Weighted Avg.	0.947	0.356	0.935	0.947	0.936	

=== Confusion Matrix ===

a	b	c	d	<-- classified as
10036	5928	0	0	| a = N18
0	126983	0	0	| b = 0
499	1185	0	0	| c = N17
36	94	0	0	| d = N19

Dựa vào kết quả phân tích, sau khi lượt bỏ các luật trùng lặp, ta rút ra được bộ luật:

- Luật 1: Nếu muc_creatinine từ 1 đến 2 thì ⇨ (kết luận) không có bệnh.
- Luật 2: Nếu muc_creatinine >= 3 thì ⇨ (kết luận) N18.

Với bộ dữ liệu gồm các thuộc tính : muc_creatinine, tinh, **thì việc kết luận xác định bệnh lý không có thay đổi và không có sự ảnh hưởng của vị trí phân bố dân cư**.

3.7.9 Mẫu thử nghiệm thứ 7

Số lượng bộ dữ liệu gồm 14.076 mẫu dữ liệu có thuộc tính muc_creatinine = 1 với các thuộc tính : gioi_tinh, muc_creatinine, muc_tuoi, muc_urea, ket_qua. Ta cần xem xét đối bộ dữ liệu chỉ gồm các tập dữ liệu có liên quan đến thuộc tính muc_creatinine = 1 thì kết quả phân tích dữ liệu có sự khác biệt so với các mẫu thử nghiệm trước hay không?

Bảng 3.15. Kết quả phân tích dữ liệu (Mẫu thử nghiệm thứ 7)

=== Run information ===

Instances: 14076

Attributes: 5

gioi_tinh

muc_creatinine

muc_tuoi

muc_urea

ket_qua

Test mode:10-fold cross-validation

=== Classifier model (full training set) ===

J48 pruned tree

: 0.0 (14076.0/846.0)

Number of Leaves : 1

Size of the tree : 1

Time taken to build model: 0.05 seconds

=== Stratified cross-validation ===

=== Summary ===

Correctly Classified Instances	*13230*	*93.9898 %*
Incorrectly Classified Instances	*846*	*6.0102 %*

=== Detailed Accuracy By Class ===

	TP Rate	FP Rate	Precision	Recall	F-Measure	Class
	1	1	0.94	1	0.969	0
	0	0	0	0	0	N18
	0	0	0	0	0	N17
	0	0	0	0	0	N19
Weighted Avg.	0.94	0.94	0.883	0.94	0.911	

=== Confusion Matrix ===

a	b	c	d	<-- classified as
13230	0	0	0	| a = N18
647	0	0	0	| b = 0
191	0	0	0	| c = N17
8	0	0	0	| d = N19

Dựa vào kết quả phân tích, với chỉ số muc_creatinine = 1 ⇨ (kết luận) không có bệnh (phù hợp với kết luận theo phân tích Mẫu thử nghiệm thứ 5).

3.7.10 Mẫu thử nghiệm thứ 8

Tương tự ***Mẫu thử nghiệm thứ 7*** với số lượng 17.390 mẫu dữ liệu chỉ gồm các trường có liên quan đến thuộc tính muc_creatinine = 2 với các thuộc tính : gioi_tinh, muc_creatinine, muc_tuoi, muc_urea, ket_qua. Ta cũng đánh giá xem có sự thay đổi so với các mẫu thử nghiệm trước hay không?

Bảng 3.16. Kết quả phân tích dữ liệu (Mẫu thử nghiệm thứ 8)

```
=== Run information ===
Instances:    17390
Attributes:   5
              gioi_tinh
              muc_creatinine
              muc_tuoi
              muc_urea
              ket_qua
Test mode:10-fold cross-validation
=== Classifier model (full training set) ===
J48 pruned tree
------------------
muc_urea <= 1: 0.0 (12723.0/3759.0)
muc_urea > 1
|   muc_urea <= 3: 0.0 (4297.0/1671.0)
|   muc_urea > 3
|   |   muc_tuoi <= 7
|   |   |   gioi_tinh <= 1: 0.0 (159.0/73.0)
|   |   |   gioi_tinh > 1
|   |   |   |   muc_tuoi <= 4: 0.0 (14.0)
|   |   |   |   muc_tuoi > 4: N18 (101.0/40.0)
|   |   muc_tuoi > 7
|   |   |   gioi_tinh <= 1: N18 (53.0/19.0)
|   |   |   gioi_tinh > 1
|   |   |   |   muc_tuoi <= 8: N17 (35.0/9.0)
|   |   |   |   muc_tuoi > 8: N18 (8.0/3.0)
Number of Leaves  :    8
Size of the tree :          15
Time taken to build model: 0.09 seconds
=== Stratified cross-validation ===
```

=== Summary ===

Correctly Classified Instances	*11802*	*67.8666 %*
Incorrectly Classified Instances	*5588*	*32.1334 %*

=== Detailed Accuracy By Class ===

	TP Rate	FP Rate	Precision	Recall	F-Measure	Class
	0.997	0.972	0.68	0.997	0.808	0
	0.018	0.005	0.58	0.018	0.035	N18
	0.034	0.001	0.605	0.034	0.065	N17
	0	0	0	0	0	N19
Weighted Avg.	0.679	0.657	0.646	0.679	0.558	

=== Confusion Matrix ===

a	b	c	d	<-- classified as
11689	35	0	0	| a = N18
4737	87	17	0	| b = 0
707	26	26	0	| c = N17
64	2	0	0	| d = N19

Dựa vào kết quả phân tích, sau khi lượt bỏ các luật trùng lặp, ta rút ra được bộ luật:

- Luật 1: Nếu muc_creatinine = 2 và muc_urea <= 3 thì ⇨ (kết luận) không có bệnh.
- Luật 2: Nếu muc_creatinine = 2 và muc_urea > 3 và muc_tuoi <= 4 ⇨ (kết luận) không có bệnh.
- Luật 3: Nếu muc_creatinine = 2 và muc_urea > 3 và muc_tuoi từ 5 đến 7 và gioi_tinh = 2 thì ⇨ (kết luận) N18.
- Luật 4: Nếu muc_creatinine = 2 và muc_urea > 3 và muc_tuoi > 7 và gioi_tinh = 1 thì ⇨ (kết luận) N18.
- Luật 5: Nếu muc_creatinine = 2 và muc_urea > 3 và muc_tuoi = 8 và gioi_tinh = 2 thì ⇨ (kết luận) N17.
- Luật 6: Nếu muc_creatinine = 2 và muc_urea > 3 và muc_tuoi > 8 và gioi_tinh = 2 thì ⇨ (kết luận) N18.

So sánh với các tập luật sinh ra từ Mẫu thử nghiệm thứ 5, ta thấy khi xem xét tập dữ liệu chỉ gồm các trường có liên quan đến thuộc tính muc_creatinine = 2 thì có sự phân tích cụ thể hơn và chỉ nhỏ mức ảnh hưởng của thuộc tính muc_tuoi ra trong kết quả phân tích, tuy nhiên với kết quả phân tích này tỉ lệ phân lớp chính xác lại giảm chỉ còn ***67.8666 %.***

3.7.11 Mẫu thử nghiệm thứ 9a

Số lượng 1.027 mẫu dữ liệu được khai thác thêm thuộc tính muc_thanh_thai (mức độ thanh thải creatinine trong cơ thể) so với mẫu dữ liệu ban đầu để xem xét tính chính xác trong chẩn đoán bệnh khi khai thác thêm chỉ số cơ thể học, các thuộc tính : muc_creatinine, muc_thanh_thai, ket_qua.

Bảng 3.17. Kết quả phân tích dữ liệu (Mẫu thử nghiệm thứ 9a)

```
=== Run information ===
Instances:    1027
Attributes:   3
              muc_creatinine
              muc_thanh_thai
              ket_qua
Test mode:split 66.0% train, remainder test
=== Classifier model (full training set) ===
J48 pruned tree
------------------
muc_creatinine <= 1: 0.0 (558.0/23.0)
muc_creatinine > 1
|   muc_creatinine <= 2
|   |   muc_thanh_thai <= 3: 0.0 (71.0/21.0)
|   |   muc_thanh_thai > 3: N18 (141.0/59.0)
|   muc_creatinine > 2: N18 (257.0/24.0)
Number of Leaves  :   4
Size of the tree :    7
Time taken to build model: 0.02 seconds
```

=== Stratified cross-validation ===

=== Summary ===

Correctly Classified Instances	*298*	***85.3868 %***
Incorrectly Classified Instances	*51*	*14.6132 %*

=== Detailed Accuracy By Class ===

	TP Rate	FP Rate	Precision	Recall	F-Measure	Class
	0.937	0.134	0.911	0.937	0.924	0
	0	0	0	0	0	N17
	0.889	0.138	0.765	0.889	0.822	N18
Weighted Avg.	0.854	0.126	0.797	0.854	0.824	

=== Confusion Matrix ===

a	b	c	<-- classified as
194	0	13	| a = 0
6	0	19	| b = N17
13	0	104	| c = N18

Dựa vào kết quả phân tích, sau khi lượt bỏ các luật trùng lặp, ta rút ra được bộ luật:

- Luật 1: Nếu muc_creatinine <= 1 thì ⇨ (kết luận) không có bệnh.
- Luật 2: Nếu muc_creatinine = 2 và muc_thanh_thai <= 3 thì ⇨ (kết luận) không có bệnh.
- Luật 3: Nếu muc_creatinine = 2 và muc_thanh_thai > 3 thì ⇨ (kết luận) N18.
- Luật 4: Nếu muc_creatinine > 2 thì ⇨ (kết luận) N18.

Với việc thêm thuộc tính muc_thanh_thai kết hợp với muc_creatinine thì độ chính xác giảm xuống chỉ còn 85.3868 %

3.7.12 Mẫu thử nghiệm thứ 9b

Sử dụng lại tập dữ liệu trên (***Mẫu thử nghiệm thứ 9a***) nhưng sử dụng thuật toán CART với cùng phương pháp đánh giá như Mẫu thử nghiệm thứ 1a (split 66.0%

train, remainder test) thay cho thuật toán C4.5 để đánh giá lại độ hiệu quả của các thuật toán

Bảng 3.18. Kết quả phân tích dữ liệu (Mẫu thử nghiệm thứ 9b)

=== Run information ===

Instances: 1027

Attributes: 3

muc_creatinine

muc_thanh_thai

ket_qua

Test mode:split 66.0% train, remainder test

=== Classifier model (full training set) ===

CART Decision Tree

---

muc_creatinine < 1.5: 0.0(535.0/23.0)

muc_creatinine >= 1.5

| muc_thanh_thai < 3.5: 0.0(50.0/21.0)

| muc_thanh_thai >= 3.5

| | muc_creatinine < 2.5

| | | muc_thanh_thai < 4.5: N18(67.0/49.0)

| | | muc_thanh_thai >= 4.5: N18(15.0/10.0)

| | muc_creatinine >= 2.5: N18(233.0/24.0)

Number of Leaf Nodes: 5

Size of the Tree: 9

Time taken to build model: 0.05 seconds

=== Stratified cross-validation ===

=== Summary ===

*Correctly Classified Instances 298 **85.3868 %***

Incorrectly Classified Instances 51 14.6132 %

=== Detailed Accuracy By Class ===

	TP Rate	FP Rate	Precision	Recall	F-Measure	Class
	0.937	0.134	0.911	0.937	0.924	0

	0	0	0	0	0	N17
	0.889	0.138	0.765	0.889	0.822	N18
Weighted Avg.	0.854	0.126	0.797	0.854	0.824	

=== Confusion Matrix ===

a	b	c	<-- classified as
194	0	13	| a = 0
6	0	19	| b = N17
13	0	104	| c = N18

Dựa vào kết quả phân tích, sau khi lượt bỏ các luật trùng lặp, ta rút ra được bộ luật:

- Luật 1: Nếu muc_creatinine < 1.5 thì ⇨ (kết luận) không có bệnh.
- Luật 2: Nếu muc_creatinine > 1.5 và muc_thanh_thai < 3.5 thì ⇨ (kết luận) không có bệnh.
- Luật 3: Nếu muc_creatinine >= 2 thì ⇨ (kết luận) N18.

Với hai thuật toán học máy C4.5 và CART với cùng phương pháp đánh giá (split 66.0% train, remainder test) thì độ chính xác được ghi nhận là như nhau (85.3868 %) tương tự kết quả Mẫu thử nghiệm thứ 1b.

3.7.13 Mẫu thử nghiệm thứ 9c

Sử dụng lại tập dữ liệu trên (***Mẫu thử nghiệm thứ 9a***) nhưng với cùng thuật toán C4.5 nhưng thay đổi phương pháp đánh giá từ phân sử liệu làm 3 phần (split 66.0% train, remainder test) 2/3 để học và 1/3 để kiểm tra thành chia dữ liệu làm 10 phần để luân phiên học và kiểm tra (k-fold cross-validation, với k = 10) để đánh giá lại độ hiệu quả của phương pháp đánh giá.

Bảng 3.19. Kết quả phân tích dữ liệu(Mẫu thử nghiệm thứ 9c)

=== Run information ===

Instances: 1027

Attributes: 3

muc_creatinine

muc_thanh_thai

ket_qua

Test mode:10-fold cross-validation

=== Classifier model (full training set) ===

J48 pruned tree

---

muc_creatinine <= 1: 0.0 (558.0/23.0)

muc_creatinine > 1

| muc_creatinine <= 2

| | muc_thanh_thai <= 3: 0.0 (71.0/21.0)

| | muc_thanh_thai > 3: N18 (141.0/59.0)

| muc_creatinine > 2: N18 (257.0/24.0)

Number of Leaves : 4

Size of the tree : 7

Time taken to build model: 0.09 seconds

=== Stratified cross-validation ===

=== Summary ===

Correctly Classified Instances	*900*	*87.6339 %*
Incorrectly Classified Instances	*127*	*12.3661 %*

=== Detailed Accuracy By Class ===

	TP Rate	FP Rate	Precision	Recall	F-Measure	Class
	0.941	0.109	0.93	0.941	0.935	0
	0	0	0	0	0	N17
	0.918	0.121	0.791	0.918	0.85	N18
Weighted Avg.	0.876	0.106	0.828	0.876	0.85	

=== Confusion Matrix ===

a	b	c	<-- classified as
585	0	37	| a = 0
16	0	46	| b = N17
28	0	315	| c = N18

Dựa vào kết quả phân tích, sau khi lượt bỏ các luật trùng lặp, ta rút ra được bộ luật:

- Luật 1: Nếu muc_creatinine < 2 thì ⇨ (kết luận) không có bệnh.

- Luật 2: Nếu muc_creatinine = 2 và muc_thanh_thai <= 3 thì ⇨ (kết luận) không có bệnh.
- Luật 3: Nếu muc_creatinine = 2 và muc_thanh_thai > 3 thì ⇨ (kết luận) N18.
- Luật 4: Nếu muc_creatinine > 2 thì ⇨ (kết luận) N18.

Dựa vào kết quả phân tích cho thấy khi sử dụng phương pháp đánh giá k-fold cross-validation, với k = 10 thì độ chính xác khi phân nhóm bằng thuật toán C4.5 cao hơn (đạt 87.6339 %) khi sử dụng phương pháp đánh giá bằng cách phân sử liệu làm 3 phần (split 66.0% train, remainder test) 2/3 để học và 1/3 để kiểm tra (chỉ đạt 85.3868 %). Điều này tương tự kết quả Mẫu thử nghiệm thứ 1c.

3.7.14 Mẫu thử nghiệm thứ 10

Số lượng 1.027 mẫu dữ liệu gồm các thuộc tính : muc_creatinine, muc_urea, muc_thanh_thai, ket_qua.

Bảng 3.20. Kết quả phân tích dữ liệu(Mẫu thử nghiệm thứ 10)

```
=== Run information ===
Instances:    1027
Attributes:   4
              muc_creatinine
              muc_urea
              muc_thanh_thai
              ket_qua
Test mode:10-fold cross-validation
=== Classifier model (full training set) ===
J48 pruned tree
------------------
muc_creatinine <= 1: 0.0 (558.0/23.0)
muc_creatinine > 1
|   muc_creatinine <= 2
|   |   muc_thanh_thai <= 3: 0.0 (71.0/21.0)
```

| | *muc_thanh_thai > 3*

| | | *muc_urea <= 3: N18 (134.0/52.0)*

| | | *muc_urea > 3: N17 (7.0)*

| *muc_creatinine > 2: N18 (257.0/24.0)*

Number of Leaves : 5

Size of the tree : 9

Time taken to build model: 0 seconds

=== Stratified cross-validation ===

=== Summary ===

Correctly Classified Instances	*907*	*88.3155 %*
Incorrectly Classified Instances	*120*	*11.6845 %*

=== Detailed Accuracy By Class ===

	TP Rate	FP Rate	Precision	Recall	F-Measure	Class
	0.941	0.109	0.93	0.941	0.935	0
	0.113	0	1	0.113	0.203	N17
	0.918	0.111	0.806	0.918	0.858	N18
Weighted Avg.	0.883	0.103	0.893	0.883	0.865	

=== Confusion Matrix ===

a	b	c	<-- classified as
585	0	37	| a = 0
16	7	39	| b = N17
28	0	315	| c = N18

Dựa vào kết quả phân tích, sau khi lượt bỏ các luật trùng lặp, ta rút ra được bộ luật:

- Luật 1: Nếu muc_creatinine <= 1 thì ⇨ (kết luận) không có bệnh.
- Luật 2: Nếu muc_creatinine = 2 và muc_thanh_thai <= 3 thì ⇨ (kết luận) không có bệnh.
- Luật 3: Nếu muc_creatinine = 2 và muc_thanh_thai > 3 và muc_urea <= 3 thì ⇨ (kết luận) N18.
- Luật 4: Nếu muc_creatinine = 2 và muc_thanh_thai > 3 và muc_urea > 3 thì ⇨ (kết luận) N17.

- Luật 5: Nếu muc_creatinine > 2 thì ⇨ (kết luận) N18.

Dựa vào kết quả phân tích khi thêm thuộc tính muc_urea so với Mẫu thử nghiệm thứ 9c thì độ chính xác tăng lên 88.3155 % so với 87.6339 %

3.7.15 Mẫu thử nghiệm thứ 11

Số lượng 1.027 mẫu dữ liệu gồm các thuộc tính : gioi_tinh, muc_creatinine, muc_tuoi, muc_urea, muc_thanh_thai, ket_qua.

Bảng 3.21. Kết quả phân tích dữ liệu (Mẫu thử nghiệm thứ 11)

```
=== Run information ===
Instances:    1027
Attributes:   6
              gioi_tinh
              muc_creatinine
              muc_tuoi
              muc_urea
              muc_thanh_thai
              ket_qua
Test mode:10-fold cross-validation
=== Classifier model (full training set) ===
J48 pruned tree
------------------
muc_creatinine <= 1: 0.0 (558.0/23.0)
muc_creatinine > 1
|   muc_creatinine <= 2
|   |   muc_thanh_thai <= 3: 0.0 (71.0/21.0)
|   |   muc_thanh_thai > 3
|   |   |   muc_urea <= 3: N18 (134.0/52.0)
|   |   |   muc_urea > 3: N17 (7.0)
|   muc_creatinine > 2
|   |   gioi_tinh <= 1
|   |   |   muc_tuoi <= 4: N18 (37.0)
```

| | | *muc_tuoi > 4*

| | | | *muc_creatinine <= 3*

| | | | | *muc_tuoi <= 7: N17 (10.0/1.0)*

| | | | | *muc_tuoi > 7: N18 (17.0/3.0)*

| | | | *muc_creatinine > 3: N18 (59.0/8.0)*

| | *gioi_tinh > 1: N18 (134.0/4.0)*

Number of Leaves : 9

Size of the tree : 17

Time taken to build model: 0.02 seconds

=== Stratified cross-validation ===

=== Summary ===

Correctly Classified Instances	*912*	*88.8023 %*
Incorrectly Classified Instances	*115*	*11.1977 %*

=== Detailed Accuracy By Class ===

	TP Rate	FP Rate	Precision	Recall	F-Measure	Class
	0.937	0.109	0.93	0.937	0.934	0
	0.274	0.003	0.85	0.274	0.415	N17
	0.91	0.099	0.821	0.91	0.863	N18
Weighted Avg.	0.888	0.099	0.889	0.888	0.879	

=== Confusion Matrix ===

a	b	c	<-- classified as
583	0	39	| a = 0
16	17	29	| b = N17
28	3	312	| c = N18

Dựa vào kết quả phân tích, sau khi lượt bỏ các luật trùng lặp, ta rút ra được bộ luật:

- Luật 1: Nếu muc_creatinine <= 1 thì ⇨ (kết luận) không có bệnh.
- Luật 2: Nếu muc_creatinine = 2 và muc_thanh_thai <= 3 thì ⇨ (kết luận) không có bệnh.
- Luật 3: Nếu muc_creatinine = 2 và muc_thanh_thai > 3 và muc_urea <= 3 thì ⇨ (kết luận) N18.
- Luật 4: Nếu muc_creatinine = 2 và muc_thanh_thai > 3 và muc_urea > 3 thì ⇨ (kết luận) N17.
- Luật 5: Nếu muc_creatinine > 2 và gioi_tinh = 2 thì ⇨ (kết luận) N18.
- Luật 6: Nếu muc_creatinine = 3 và gioi_tinh = 1 và muc_tuoi từ 5 đến 7 thì ⇨ (kết luận) N17.
- Luật 7: Nếu muc_creatinine = 3 và gioi_tinh = 1 và muc_tuoi > 7 thì ⇨ (kết luận) N18.
- Luật 8: Nếu muc_creatinine > 3 thì ⇨ (kết luận) N18.

Dựa vào kết quả phân tích khi thêm thuộc tính gioi_tinh và muc_tuoi so với Mẫu thử nghiệm thứ 10 thì độ chính xác tăng lên 88.8023 % so với 88.3155 %

Chương 4: TRÌNH BÀY, ĐÁNH GIÁ, BÀN LUẬN CÁC KẾT QUẢ

4.1 Đánh giá hiệu quả của thuật toán

Dựa vào kết quả của các mẫu thử nghiệm: Mẫu thử nghiệm thứ 1a, Mẫu thử nghiệm thứ 1b, Mẫu thử nghiệm thứ 1c; ta rút ra được kết luận:

Bảng 4.1. So sánh các phương pháp học máy 1

Thuật toán	J48	CART	J48
Tỉ lệ dữ liệu học	Split 66.0% train, remainder test	split 66.0% train, remainder test	10-fold, cross-validation
Độ chính xác (Precision)	0.934	0.934	0.935
Tỉ lệ gọi lại (recall)	0.946	0.946	0.947
Thời gian học (training)	0.47	8.91	0.28
Số luật sinh ra	2	2	2
Tỉ lệ chính xác (Correctly Classified Instances)	94.5712%	94.5712%	94.6519%
Tỉ lệ sai (Incorrectly Classified Instances)	5.4288%	5.4288%	5.3481%

Dựa vào kết quả của các mẫu thử nghiệm :Mẫu thử nghiệm thứ 9a, Mẫu thử nghiệm thứ 9b, Mẫu thử nghiệm thứ 9c; ta rút ra được kết luận:

Bảng 4.2. So sánh các phương pháp học máy 2

Thuật toán	J48	CART	J48
Phương pháp học	split 66.0% train, remainder test	split 66.0% train, remainder test	10-fold, cross-validation
Độ chính xác (Precision)	0.797	0.797	0.828
Tỉ lệ gọi lại (recall)	0.854	0.854	0.876
Thời gian học (training)	*0.02*	*0.05*	*0.09*
Số luật sinh ra	4	3	4
Tỉ lệ chính xác (Correctly Classified Instances)	*85.3868*	*85.3868*	*87.6339*
Tỉ lệ sai (Incorrectly Classified Instances)	*14.6132*	*14.6132*	*12.3661*

Dựa vào 2 bảng phân tích thuật toán áp dụng đối với mẫu dữ liệu cận lâm sàng bệnh nhân, ta thấy phương pháp cây quyết định sử dụng thuật toán CART tuy có các chỉ số tương tự khi phân tích dữ liệu theo bằng thuật toán C4.5 (J48) nhưng có thời gian thực hiện thuật toán CART lớn hơn thời gian thực hiện bằng thuật toán C4.5

Đối với phương pháp sử dụng thuật toán C4.5: nếu ta sử dụng phương pháp đánh giá Hold-out (split 66.0% train, remainder test) so với phương pháp k-fold cross validation (với k = 10) cho thấy hiệu quả hơn về mặt thời gian và tỉ lệ mẫu chính xác cao hơn.

Từ các phân tích trên cho thấy với bộ mẫu dữ liệu kết quả chỉ số cận lâm sàng của bệnh nhân, thì việc chọn thuật toán C4.5 và mô hình phân lớp bằng phương

pháp k-fold cross validation (với k = 10) là sự lựa chọn tốt hơn phương pháp Hold-out (split 66.0% train, remainder test).

4.2 Đánh giá kết quả thử nghiệm

Bảng 4.3. Bảng phân tích kết quả thực nghiệm

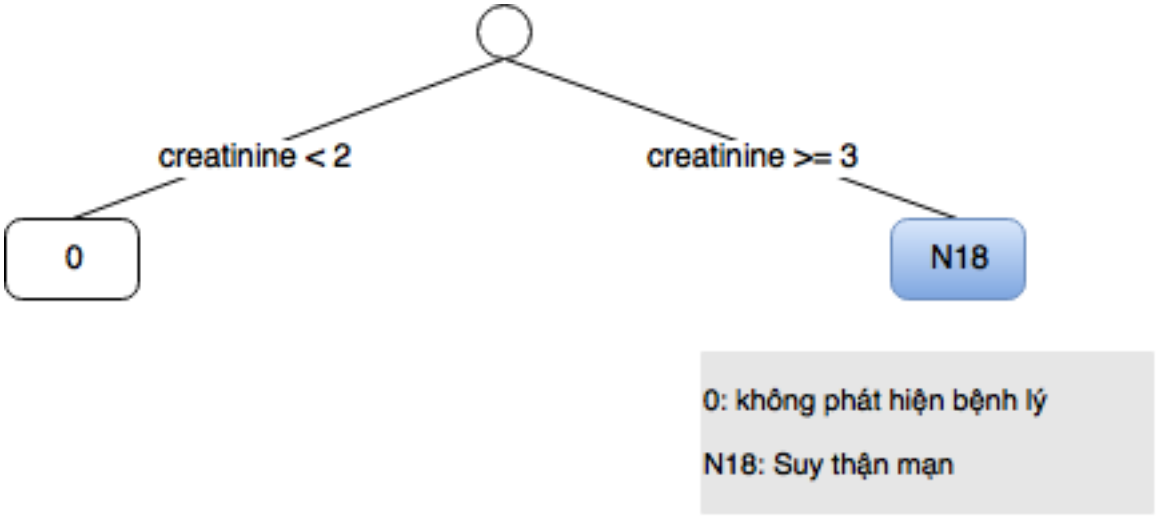
TT	Mẫu	Tập thuộc tính	Thời gian thực thi (Time taken to build model) (giây)	Tỉ lệ phân lớp đúng (Correctly Classified Instances) (%)	Độ chính xác (Precision)	Tỉ lệ gọi lại (recall)	Luật sinh ra
1	Mẫu thử nghiệm thứ 1c	muc_creatinine	0.28	94.6519 %	0.935	0.947	2
2	Mẫu thử nghiệm thứ 2	muc_creatinine, muc_urea	0.89	94.6498 %	0.934	0.946	4
3	Mẫu thử nghiệm thứ 3	gioi_tinh, muc_creatinine, muc_urea	2.17	94.6519 %	0.935	0.947	2
4	Mẫu thử nghiệm thứ 4	muc_creatinine muc_tuoi, muc_urea	1.41	94.703 %	0.941	0.947	7
5	Mẫu thử nghiệm thứ 5	gioi_tinh, muc_creatinine, muc_tuoi, muc_urea	2.33	94.7078 %	0.942	0.947	8
6	Mẫu thử nghiệm thứ 6	muc_creatinine, ket_qua, tinh	0.28	94.6519 %	0.935	0.947	2

7	Mẫu thử nghiệm thứ 7	gioi_tinh, muc_creatinine, muc_tuoi, muc_urea	0.05	93.9898 %	0.883	0.94	0
8	Mẫu thử nghiệm thứ 8	gioi_tinh, muc_creatinine, muc_tuoi, muc_urea	0.09	67.8666 %	0.646	0.679	6

4.3 Bàn luận kết quả

Dựa vào kết quả Mẫu thử nghiệm thứ 1c, ta rút ra được tập luật (tập thứ 1) khi sử dụng mô hình cây quyết định với thuật toán C4.5 (J48 trong WEKA) cho bảng mẫu dữ liệu phân tích thông tin cận lâm sàng bệnh nhân gồm:

- Luật 1: Nếu muc_creatinine <= 2 thì ⇨ (kết luận) không có bệnh.
- Luật 2: Nếu muc_creatinine > 2 thì ⇨ (kết luận) N18.



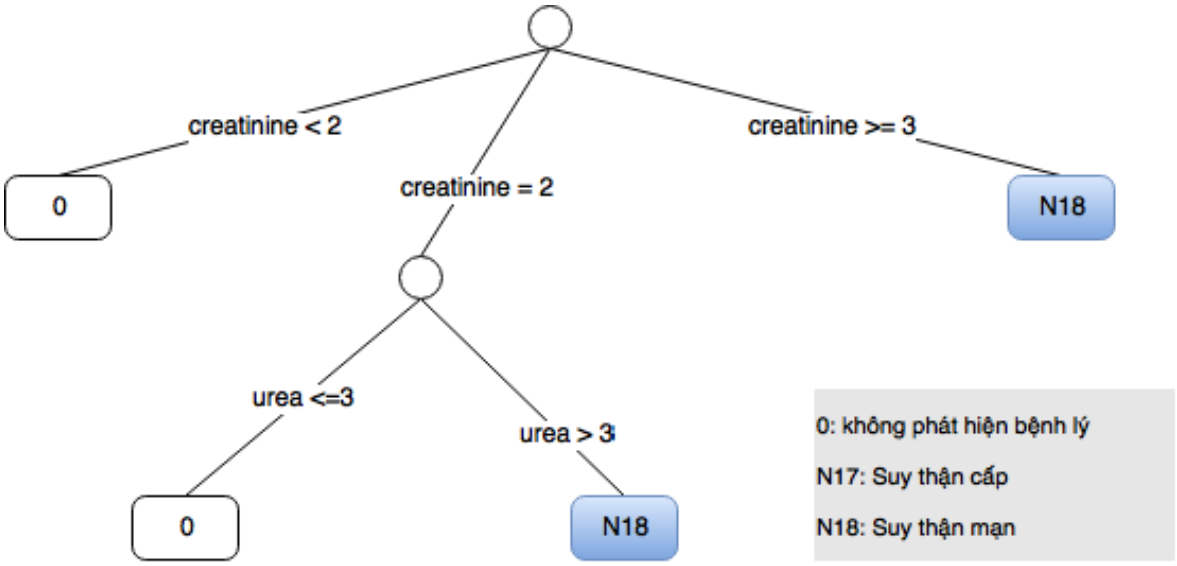
Hình 4.1. Cây quyết định theo tập luật thứ 1

Dựa vào cây quyết định trên (Hình 4.1), ta thấy các bệnh nhân có mức creatinine > 2 đều có kết luận có bệnh lý suy thận mạn (N18). Với kết luận này tuy có phù hợp hợp với các phát đồ điều trị chung nhưng thông tin còn chung chưa đánh giá hết được được các yếu tố khác ảnh hưởng đến kết quả chẩn đoán. Vì vậy ta xem xét các mẫu dữ liệu khác để tìm ra những yếu cố cấu thành trong việc chẩn đoán chính xác bệnh lý cho bệnh nhân.

Xem xét kết quả Mẫu thử nghiệm thứ 2, ta thấy khi có sự kết hợp giữa việc phân tích muc_urea trong máu, việc hình thành các tập luật có thay đổi và cụ thể hơn, chính xác hơn khi chẩn đoán. Các tập luật (tập thứ 2) của Mẫu thử nghiệm thứ 2 bao gồm:

- Luật 1: Nếu muc_creatinine từ 1 đến 2 thì ⇨ (kết luận) không có bệnh.
- Luật 2: Nếu muc_creatinine = 2 và muc_urea <= 3 thì ⇨ (kết luận) không có bệnh.
- Luật 3: Nếu muc_creatinine = 2 và muc_urea > 3 thì ⇨ (kết luận) N18.
- Luật 4: Nếu muc_creatinine >= 3 thì ⇨ (kết luận) N18.

Dựa vào tập luật thứ 2 ta có mô hình cây quyết định thứ 2 như sau (Hình 4.2):



Hình 4.2. Cây quyết định theo tập luật thứ 2

Dựa trên tập luật vừa có, ta thấy cây quyết định được tạo ra có sự phân tách cụ thể khi muc_creatinine = 2. Cụ thể khi ở tập dữ liệu theo Mẫu thử nghiệm thứ 1c

muc_creatinine = 2 thì kết luận không phát hiện bệnh lý nhưng với tập dữ liệu có sự tham gia của thuộc tính muc_urea (Mẫu thử nghiệm thứ 2) thì việc kết luận bệnh lý còn phụ thuộc vào muc_urea trong máu của bệnh nhân. Qua đó ta cũng thấy cây quyết định thứ 2 (Hình 4.2) mang tính bao trùm và tổng quát hơn cây quyết định thứ 1 (Hình 4.1), vì vậy ta chọn mẫu dữ liệu với các thuộc tính *(muc_creatinine, muc_urea)* thì kết quả chính xác hơn và hiệu quả hơn.

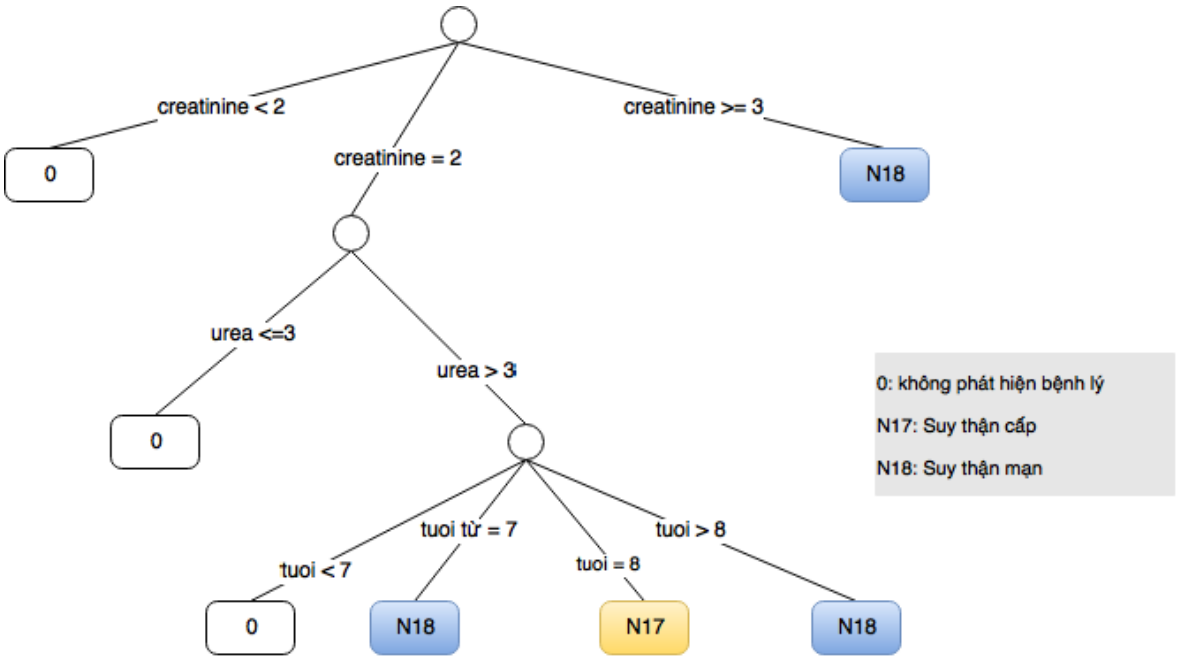
Tiếp tục xem xét tập luật (tập thứ 3) được rút ra từ Mẫu thử nghiệm thứ 3 ta thấy khi ta xem xét các thuộc tính *(muc_creatinine, muc_urea, gioi_tinh)* trên dữ liệu cận lâm sàng bệnh nhân thì tập các luật được tạo ra không có sự khác biệt so với tập luật thứ 1. Vì vậy mô hình cây quyết định không có sự thay đổi so với Mẫu thử nghiệm thứ 2. Các luật được sinh ra của Mẫu thử nghiệm thứ 3 bao gồm:

- Luật 1: Nếu muc_creatinine từ 1 đến 2 thì ⇨ (kết luận) không có bệnh.
- Luật 2: Nếu muc_creatinine >= 3 thì ⇨ (kết luận) N18.

Tiếp tục xem xét tập luật (tập thứ 4) được rút ra từ Mẫu thử nghiệm thứ 4 ta thấy khi ta xem xét các thuộc tính *(muc_creatinine, muc_urea, muc_tuoi)*, mô hình chẩn đoán bệnh lý có sự phân tách đổi với trường hợp muc_urea > 3 và ta cũng tìm thấy mối liên quan giữa tuổi của bệnh nhân với các chỉ số cận lâm sàng.

- Luật 1: Nếu muc_creatinine từ 1 đến 2 thì ⇨ (kết luận) không có bệnh.
- Luật 2: Nếu muc_creatinine = 2 và muc_urea <= 3 thì ⇨ (kết luận) không có bệnh.
- Luật 3: Nếu muc_creatinine = 2 và muc_urea > 3 và muc_tuoi = 7 thì ⇨ (kết luận) N18.
- Luật 4: Nếu muc_creatinine = 2 và muc_urea > 3 và muc_tuoi = 8 thì ⇨ (kết luận) N17.
- Luật 5: Nếu muc_creatinine = 2 và muc_urea > 3 và muc_tuoi > 8 thì ⇨ (kết luận) N18.
- Luật 6: Nếu muc_creatinine >= 3 thì ⇨ (kết luận) N18

Dựa vào tập luật (tập thứ 4) ta có mô hình cây quyết định thứ 3 như sau (Hình 4.3):



Hình 4.3. Cây quyết định theo tập luật thứ 4

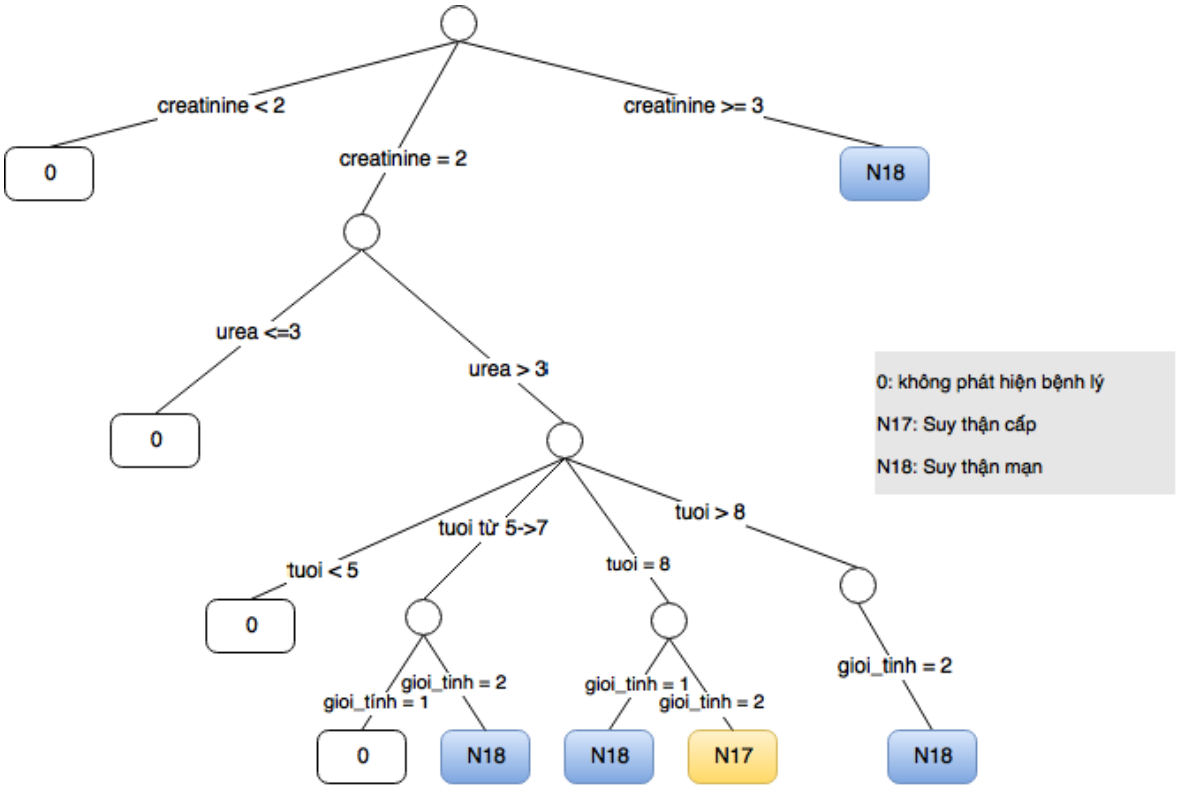
Qua mô hình cây quyết định vừa vẽ, ta cũng thấy cây quyết định thứ 3 (Hình 4.3) là sự mở rộng của cây quyết định thứ 2 (Hình 4.2), vì vậy ta chọn mẫu dữ liệu với các thuộc tính *(muc_creatinine, muc_urea, muc_tuoi)* thì kết quả chính xác hơn và hiệu quả hơn mô hình cây quyết định theo tập luật thứ 2.

Tiếp tục xem xét tập luật (tập thứ 5) được rút ra từ Mẫu thử nghiệm thứ 5 ta thấy khi ta xem xét các thuộc tính *(muc_creatinine, muc_urea, muc_tuoi, gioi_tinh)*, mô hình chẩn đoán bệnh lý theo tập luật cũng sự thay đổi và việc kết hợp nhiều thuộc tính của bệnh nhân sẽ cho ra kết quả chẩn đoán tốt hơn khi xem xét các mẫu dữ liệu trước đó.

- Luật 1: Nếu muc_creatinine từ 1 đến 2 thì ⇨ (kết luận) không có bệnh.
- Luật 2: Nếu muc_creatinine = 2 và muc_urea <= 3 thì ⇨ (kết luận) không có bệnh.
- Luật 3: Nếu muc_creatinine = 2 và muc_urea > 3 và muc_tuoi <= 7 và gioi_tinh = 1 thì ⇨ (kết luận) không có bệnh

- Luật 4: Nếu muc_creatinine = 2 và muc_urea > 3 và muc_tuoi từ 5 đến 7 và gioi_tinh = 2 thì ⇨ (kết luận) N18.
- Luật 5: Nếu muc_creatinine = 2 và muc_urea > 3 và muc_tuoi = 8 và gioi_tinh = 1 thì ⇨ (kết luận) N18.
- Luật 6: Nếu muc_creatinine = 2 và muc_urea > 3 và muc_tuoi = 8 và gioi_tinh = 2 thì ⇨ (kết luận) N17.
- Luật 7: Nếu muc_creatinine = 2 và muc_urea > 3 và muc_tuoi > 8 và gioi_tinh = 2 thì ⇨ (kết luận) N18.
- Luật 8: Nếu muc_creatinine >= 3 thì ⇨ (kết luận) N18.

Dựa vào tập luật (tập thứ 5) ta có mô hình cây quyết định thứ 4 như sau (Hình 4.4):



Hình 4.4. Cây quyết định theo tập luật thứ 5

Qua mô hình cây quyết định vừa vẽ, ta cũng thấy cây quyết định thứ 4 (Hình 4.4) là sự mở rộng của cây quyết định thứ 3 (Hình 4.3), vì vậy ta chọn mẫu dữ liệu

với các thuộc tính *(muc_creatinine, muc_urea, muc_tuoi, gioi_tinh)* thì kết quả chính xác hơn và hiệu quả hơn mô hình cây quyết định theo tập luật thứ 3.

Tiếp tục xem xét tập luật (tập thứ 6) được rút ra từ Mẫu thử nghiệm thứ 6 ta thấy khi ta xem xét các thuộc tính *(muc_creatinine, muc_thanh_thai)* thì tập các luật được tạo ra đã được bao trùm trong tập luật thứ 5. Vì vậy mô hình cây quyết định không có sự thay đổi so với Mẫu thử nghiệm thứ 5. Các luật được sinh ra của Mẫu thử nghiệm thứ 6 bao gồm:

- Luật 1: Nếu muc_creatinine từ 1 đến 2 thì ⇨ (kết luận) không có bệnh.
- Luật 2: Nếu muc_creatinine >= 3 thì ⇨ (kết luận) N18.

Do mẫu dữ liệu thứ 7 (Mẫu thử nghiệm thứ 7) không rút trích được tập luật vì với mức creatinine máu = 1 thì không phát sinh bệnh lý đã phù hợp với các tập luật trước đó và nằm trong luật 1 trong tập luật 5 do đó ta không xem xét dữ liệu trong tập thuộc này.

Tiếp tục xem xét tập luật (tập thứ 8) được rút ra từ Mẫu thử nghiệm thứ 8 ta thấy khi ta xem xét các thuộc tính *(gioi_tinh, muc_creatinine, muc_tuoi, muc_urea),* trong đó muc_creatinine = 2, thì mô hình chẩn đoán bệnh lý theo tập luật cũng không sự thay đổi vì tập luật thứ 8 được sinh ra đã được tập luật thứ 5 bao trùm vì vậy mô hình cây quyết định không có thay đổi so vây quyết định thứ 4 (Hình 4.4) .

- Luật 1: Nếu muc_creatinine = 2 và muc_urea <= 3 thì ⇨ (kết luận) không có bệnh.
- Luật 2: Nếu muc_creatinine = 2 và muc_urea > 3 và muc_tuoi <= 4 ⇨ (kết luận) không có bệnh.
- Luật 3: Nếu muc_creatinine = 2 và muc_urea > 3 và muc_tuoi từ 5 đến 7 và gioi_tinh = 2 thì ⇨ (kết luận) N18.
- Luật 4: Nếu muc_creatinine = 2 và muc_urea > 3 và muc_tuoi > 7 và gioi_tinh = 1 thì ⇨ (kết luận) N18.
- Luật 5: Nếu muc_creatinine = 2 và muc_urea > 3 và muc_tuoi = 8 và gioi_tinh = 2 thì ⇨ (kết luận) N17.

- Luật 6: Nếu muc_creatinine = 2 và muc_urea > 3 và muc_tuoi > 8 và gioi_tinh = 2 thì ⇨ (kết luận) N18.

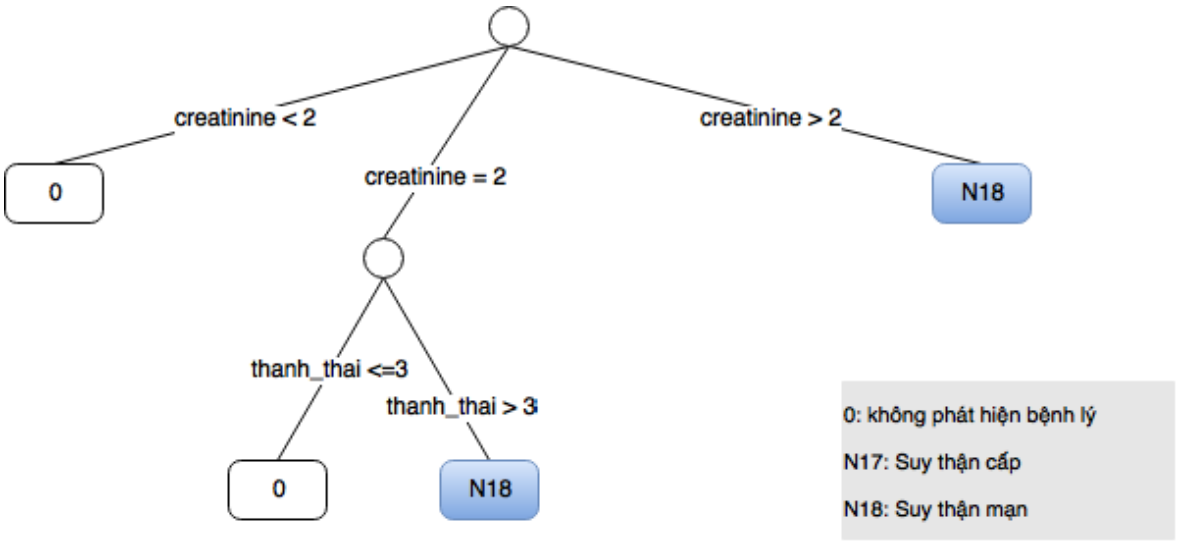
Vậy với việc phân tích dữ liệu kết quả cận lâm sàng (tập trung vào phân tích các chỉ số urea và creatinine) của bệnh nhân, ta rút ra được tập luật theo mô hình cây quyết định dùng để hỗ trợ chẩn đoán trong quá trình khám chữa bệnh cho bệnh nhân một cách khoa học, nhanh chóng và chính xác.

Tuy nhiên để tìm hiểu mối tương quan giữa mức độ thanh thải creatinine với các chỉ số sinh học của bệnh nhân dựa vào công thức Cockrofl và Gault (Bảng 2.2), trong phạm vi giới hạn của đề tài khai thác được (khoảng 1.000 mẫu dữ liệu), ta cần so sánh mô hình cây quyết định được xây dựng dựa trên tập thuộc tính *(muc_creatinine, muc_thanh_thai)* với mô hình cây quyết định chỉ sử dụng phương pháp khai thác kết quả cận lâm sàng của bệnh nhân.

Sau khi xem xét tập luật (tập thứ 9) được rút ra từ Mẫu thử nghiệm thứ 9c ta thấy khi ta xem xét các thuộc tính *(muc_creatinine, muc_thanh_thai)*, ta được tập luật gồm:

- Luật 1: Nếu muc_creatinine < 2 thì ⇨ (kết luận) không có bệnh.
- Luật 2: Nếu muc_creatinine = 2 và muc_thanh_thai <= 3 thì ⇨ (kết luận) không có bệnh.
- Luật 3: Nếu muc_creatinine = 2 và muc_thanh_thai > 3 thì ⇨ (kết luận) N18.
- Luật 4: Nếu muc_creatinine > 2 thì ⇨ (kết luận) N18.

Dựa vào tập luật (tập thứ 9) ta có mô hình cây quyết định thứ 5 như sau (Hình 4.5):

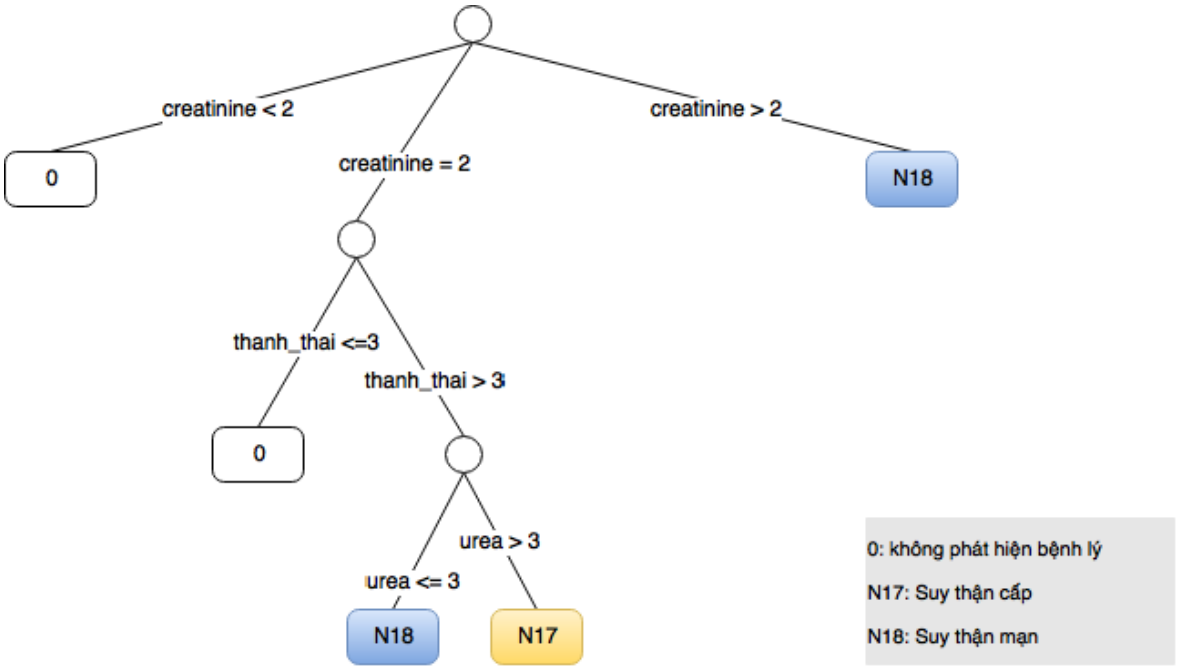


Hình 4.5. Cây quyết định theo tập luật thứ 9

Tiếp tục xem xét tập luật (tập thứ 10) được rút ra từ Mẫu thử nghiệm thứ 10 ta thấy khi ta xem xét các thuộc tính *(muc_creatinine, muc_urea, muc_thanh_thai)*, thì mô hình chẩn đoán bệnh lý theo tập luật cũng có sự thay đổi theo sự ảnh hưởng của chi số urea, ta được tập luật gồm:

- Luật 1: Nếu muc_creatinine <= 1 thì ⇨ (kết luận) không có bệnh.
- Luật 2: Nếu muc_creatinine = 2 và muc_thanh_thai <= 3 thì ⇨ (kết luận) không có bệnh.
- Luật 3: Nếu muc_creatinine = 2 và muc_thanh_thai > 3 và muc_urea <= 3 thì ⇨ (kết luận) N18.
- Luật 4: Nếu muc_creatinine = 2 và muc_thanh_thai > 3 và muc_urea > 3 thì ⇨ (kết luận) N17.
- Luật 5: Nếu muc_creatinine > 2 thì ⇨ (kết luận) N18.

Dựa vào tập luật (tập thứ 10) ta có mô hình cây quyết định thứ 6 như sau (Hình 4.6):



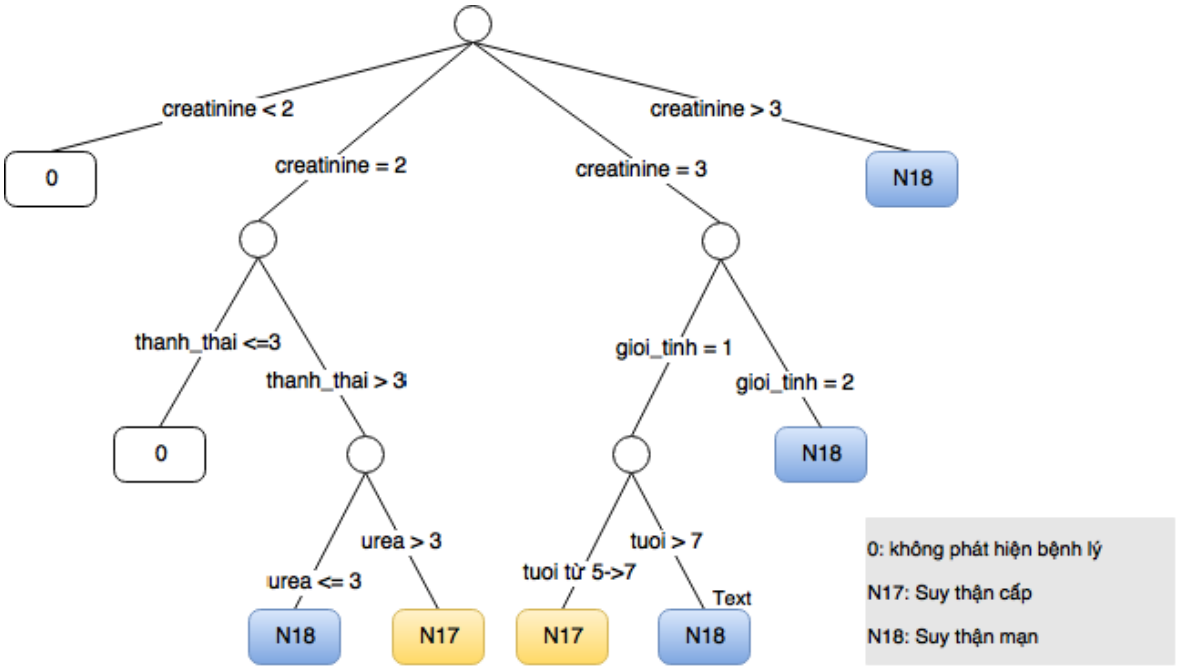
Hình 4.6. Cây quyết định theo tập luật thứ 10

Tiếp tục xem xét tập luật (tập thứ 11) được rút ra từ Mẫu thử nghiệm thứ 11 ta thấy khi ta xem xét các thuộc tính *(gioi_tinh, muc_creatinine, muc_tuoi, muc_urea, muc_thanh_thai)*, mô hình chẩn đoán bệnh lý theo tập luật cũng sự thay đổi và việc kết hợp nhiều thuộc tính của bệnh nhân sẽ cho ra kết quả chẩn đoán tốt hơn khi xem xét các mẫu dữ liệu trước đó (tập luật thứ 10) , ta được tập luật gồm:

- Luật 1: Nếu muc_creatinine <= 1 thì ⇨ (kết luận) không có bệnh.
- Luật 2: Nếu muc_creatinine = 2 và muc_thanh_thai <= 3 thì ⇨ (kết luận) không có bệnh.
- Luật 3: Nếu muc_creatinine = 2 và muc_thanh_thai > 3 và muc_urea <= 3 thì ⇨ (kết luận) N18.
- Luật 4: Nếu muc_creatinine = 2 và muc_thanh_thai > 3 và muc_urea > 3 thì ⇨ (kết luận) N17.
- Luật 5: Nếu muc_creatinine > 2 và gioi_tinh = 2 thì ⇨ (kết luận) N18.

- Luật 6: Nếu muc_creatinine = 3 và gioi_tinh = 1 và muc_tuoi từ 5 đến 7 thì ⇨ (kết luận) N17.
- Luật 7: Nếu muc_creatinine = 3 và gioi_tinh = 1 và muc_tuoi > 7 thì ⇨ (kết luận) N18.
- Luật 8: Nếu muc_creatinine > 3 thì ⇨ (kết luận) N18.

Dựa vào tập luật (tập thứ 11) ta có mô hình cây quyết định thứ 7 như sau (Hình 4.7):



Hình 4.7. Cây quyết định theo tập luật thứ 11

Qua việc phân tích thêm chỉ số mức thanh thải creatinine theo công thức Cockrofl và Gault (Bảng 2.2), ta thấy mô hình cây quyết định được xây dựng có nội dung tương đồng với mô hình cây quyết định khi chỉ khai thác kết quả của hệ thống phân tích chỉ số cận lâm sàng của bệnh nhân, mô hình cây quyết định khi khai thác thêm các chỉ số sinh học của bệnh nhân cũng cho thấy một số ưu điểm:

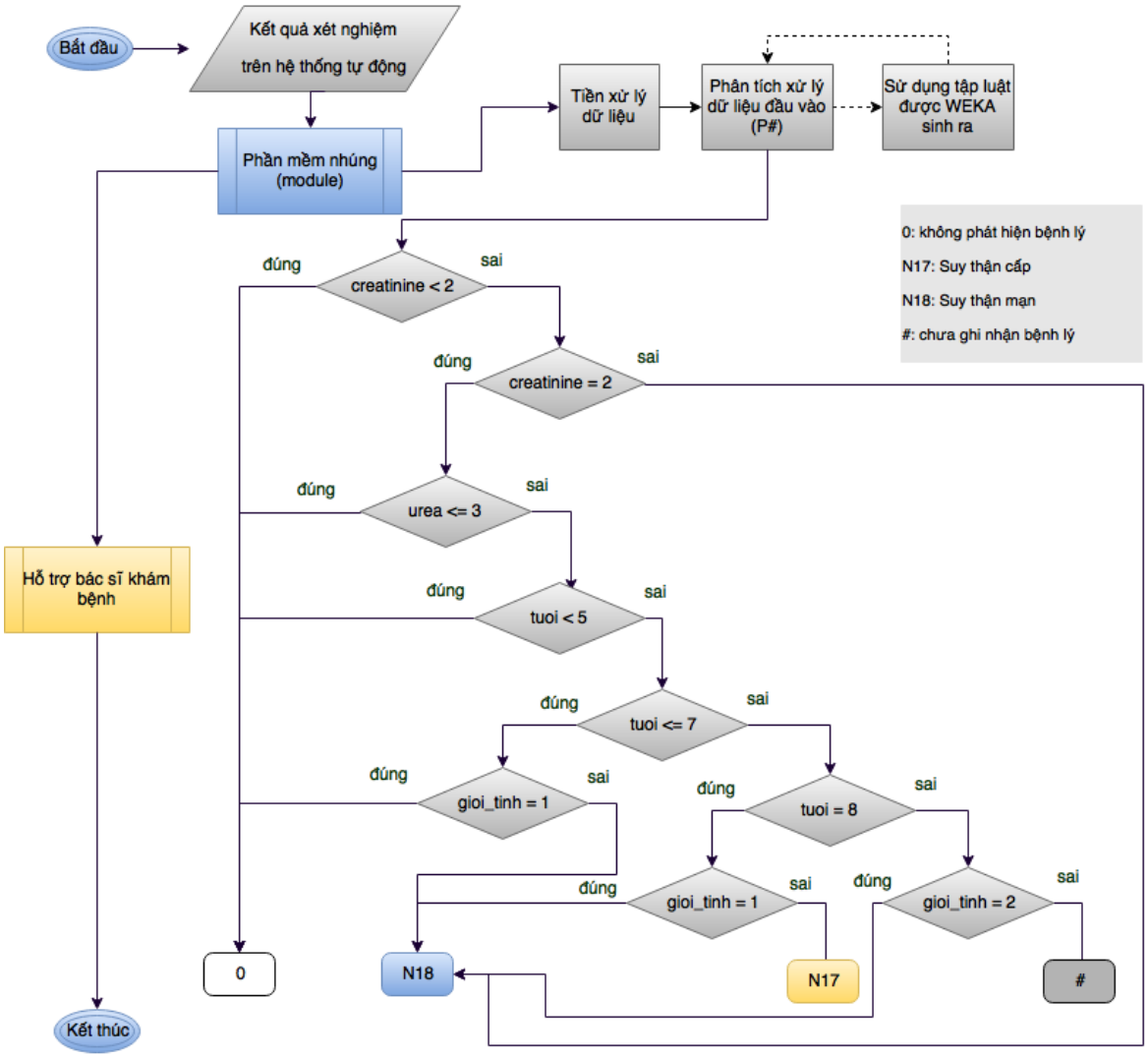
Độ sâu của cây quyết định thứ 7 (Cây quyết định theo tập luật thứ 11) thấp hơn cây quyết định thứ 4 (Cây quyết định theo tập luật thứ 5) do đó thời gian xử lý thông tin sẽ nhanh hơn.

Các giá trị trên cây quyết định thứ 7 (Cây quyết định theo tập luật thứ 11) cụ thể hơn cây quyết định thứ 4 (Cây quyết định theo tập luật thứ 5).

Thời gian để xây dựng quyết định thứ 7 (Cây quyết định theo tập luật thứ 11) ngắn hơn (chỉ có 0.02 giây) thời gian xây dựng cây quyết định thứ 4 (Cây quyết định theo tập luật thứ 5) (2.33 giây). Tuy nhiên chỉ số này chưa phản ánh đúng bản chất vì số lượng mẫu sử dụng cho cây quyết định thứ 7 chỉ bằng 1/10 số mẫu dữ liệu sử dụng đế xây dựng cây quyết định thứ 4, đồng thời tỉ lệ mẫu được phân lớp đúng của cây quyết định thứ 7 chỉ đạt 88.8023% so với tỉ lệ mẫu được phân lớp đúng của cây quyết định thứ 4 đạt 94.7078%; đây cũng là một lý do để cần có nghiên cứu sâu hơn trong việc khai thác các thông tin liên quan đến bệnh nhân để có được một kết luận chính xác.

Tuy nhiên do trong khuôn khổ đề tài chưa khai thác được số lượng mẫu lớn tương đương với số lượng mẫu sử dụng để xây dựng cây quyết định dựa trên chỉ số cận lâm sàng (Cây quyết định theo tập luật thứ 5), để có một đánh giá hoàn chỉnh về những hiệu quả khi xây dựng cây quyết định dựa vào nhiều thuộc tính sinh học và cận lâm sàng của bệnh nhân, ta cần phải có một nghiên cứu sâu rộng hơn về lợi ích khi khai thác triệt để thông tin bệnh nhân và hiệu quả của cây quyết định trong việc hỗ trợ và chẩn đoán bệnh lý cho người bệnh.

Vậy với dữ liệu kết quả xét nghiệm của bệnh nhân như đã nêu trên việc sử dụng mô hình cây quyết định với thuật toán C4.5 và phương pháp đánh giá k-fold cross validation (với k = 10) là đạt hiệu quả tốt hơn. Đồng thời dựa vào mô hình cây quyết định được xây dựng nên cùng với bảng phân tích kết quả thực nghiệm (Bảng 4.3) cho thấy việc chọn Cây quyết định theo tập luật thứ 5 (Hình 4.4) với tập các thuộc tính cần phân tích gồm: *muc_creatinine*, *muc_urea*, *muc_tuoi*, *gioi_tinh*, là phù hợp nhất trong tập các thuộc tính của bệnh nhân. Với mô hình đã chọn, ta có sơ đồ mô tả quá trình phân tích dữ liệu như Hình 4.8.



Hình 4.8. Mô hình xử lý thông tin của phần mềm nhúng (module .dll)

4.4 Ứng dụng xây dựng chương trình

Thiết kế một phần mềm nhúng (dạng .dll) trên nền tảng .Net để sử dụng kết quả phân tích.

Sử dụng ngôn ngữ P# để hiện thực các tập luật đã được xây dựng ở trên. (P# là một thể hiện Prolog được thiết kế để sử dụng với .NET Framework của Microsoft. Đặc biệt nó dịch Prolog vào mã nguồn C#, cho phép liên kết với C# để tạo thành ngôn ngữ khác trong .NET)

Nhúng P# (Psharp.dll) vào phần mềm để phân tích dữ liệu kết quả xét nghiệm của bệnh nhân tự động theo mô hình cây quyết định đã xây dựng được. Nội dung thuật toán xử lý thông tin trên prolog theo mô hình cây quyết định đã xây dựng được thể hiện như sau :

---

compXY(IN, X, Y, KQ) :- IN >= X, IN =< Y -> KQ = true; KQ = false.
compX(IN, X, KQ) :- IN = X -> KQ = true; KQ = false.

gender1(GT, KQ) :- compX(GT, 1, Y1), Y1 = true -> KQ = false; KQ = 'N18'.
gender2(GT, KQ) :- compX(GT, 1, Y1), Y1 = true -> KQ = 'N18'; KQ = 'N17'.

age8(TU, GT, KQ) :- compX(TU, 8, Y1), Y1 = true -> gender2(GT, KQ); gender1(GT, KQ).

age7(TU, GT, KQ) :- compXY(TU, 5, 7, Y1), Y1 = true -> gender1(GT, KQ); age8(TU, GT, KQ).

ure4(TU, GT, KQ) :- compXY(TU, 0, 4, Y1), Y1 = true -> KQ = false; age7(TU, GT, KQ).

ure3(UR, TU, GT, KQ) :- compXY(UR, 0, 3, Y1), Y1 = true -> KQ = false; ure4(TU, GT, KQ).

---

creatinine2(CR, UR, TU, GT, KQ) :- compX(CR, 2, Y1), Y1 = true -> ure3(UR, TU, GT, KQ); KQ = 'N18'.

creatinine(CR, UR, TU, GT, KQ) :- compXY(CR, 0, 1, Y1), Y1 = true -> KQ = false; creatinine2(CR, UR, TU, GT, KQ).

input(CR, UR, TU, GT, KQ) :- creatinine(CR, UR, TU, GT, KQ).

Chương 5: KẾT LUẬN

Nội dung luận văn với mục đích nghiên cứu, phân tích các chỉ số kết quả cận lâm sàng của bệnh án nội khoa nhằm tìm ra phương pháp chẩn đoán bệnh lý nhanh chóng và hiệu quả. Một số kết quả đã đạt được:

5.1 Về nội dung

Về cơ bản, nội dung của luận văn đã đáp ứng được nhu đánh giá, khai thác số liệu và ứng dụng học máy (data mining) đế đáp ứng nhu cầu trong hoạt động khám chữa bệnh.

Nội dung kết quả phân tích dữ liệu theo mô hình cây quyết định sử dụng thuật toán C4.5 phù hợp với phát đồ chẩn đoán và điều trị bệnh cho bệnh nhân có bệnh lý thận nội khoa.

Ứng dụng kết quả phân tích để xây dựng được phần mềm tích hợp (module nhúng dạng .dll) để đưa vào sử dụng trên hệ thống phần mềm quản lý bệnh viện.

Nghiên cứu đã góp phần xây dựng nên một phương thức chẩn đoán bệnh không chỉ dựa trên các kết quả cận lâm sàng mà còn kết hợp giữa các yếu tố (thuộc tính) khác của người bệnh để hỗ trợ chẩn đoán. Đồng thời đây cũng là một phương pháp dự trên cơ sở lý luận khoa học, đảm bảo tính đúng đắn trong khi đánh giá bệnh lý của người bệnh, giúp bác sĩ nhanh chóng đưa ra quyết định điều trị, tránh bỏ xót những thông tin liên quan đến bệnh lý của bệnh nhân nhưng chưa được khai thác triệt để trong quá trình thăm khám bệnh.

5.2 Về xây dựng chương trình

Xây dựng được phần mềm nhúng (module) tích hợp với hệ thống quản lý bệnh viện để phân tích kết quả xét nghiệm sinh hóa và tự động đưa ra gợi ý điều trị cho bác sĩ tham khảo khi thăm khám bệnh cho bệnh nhân.

Sử dụng ngôn ngữ Prolog kết hợp với C# để tăng hiệu năng xử lý thông tin trong phần mềm.

5.3 Về áp dụng thực tế

Sau khi ứng dụng thực tế mô hình cây quyết định vào quá trình khám chữa bệnh để hỗ trợ các bác sĩ trong việc phân tích kết quả cận lâm sàng của bệnh nhân (liên quan đến chỉ số creatinine và urea trong xét nghiệm sinh hóa máu) đã đạt được một số hiệu quả (kèm theo phụ lục 1):

Thời gian xác định bệnh lý nhanh hơn khi phân tích bệnh lý bằng phương pháp truyền thống.

Tự động kết hợp các thuộc tính của người bệnh để phân tích tránh tình trạng thiếu xót chẩn đoán do không đủ điều kiện khai thác thông tin khi khám bệnh.

Hỗ trợ bác sĩ ra quyết định điều trị nhanh chóng và có khoa học (dựa trên phát đồ điều trị)

5.4 Về kết quả mới thực hiện được

Bước đầu, trên cơ sở nghiên cứu của luận văn, việc áp dụng cây quyết định đã mang lại hiệu quả trong việc chẩn đoán và hỗ trợ gợi ý điều trị cho bác sĩ trong quá trình khám chữa bệnh. Tuy nhiên để kết quả ứng dụng cây quyết định trong hỗ trợ chẩn đoán và điều trị được tốt hơn cần có thời gian nghiên cứu mở rộng và đi sâu vào nghiên cứu các bệnh lý khác đặc biệt là các bệnh lý kết hợp để đưa ra phương pháp chẩn đoán tốt hơn, nhanh chóng hơn.

5.5 Một số vấn đề còn tồn tại

- Chưa phân tích được các chỉ số cận lâm sàng kết hợp để đánh giá bệnh lý của bệnh nhân toàn diện hơn.
- Các thông tin về người bệnh cần được khai thác đầy đủ hơn nhằm khai thác tối đa hệ thống phân tích dữ liệu, tăng cường khả năng quyết định của hệ thống hỗ trợ chẩn đoán bệnh.
- Cần xây dựng chương trình tự động phân tích dữ liệu bệnh nhân định kỳ và chuyển sang ngôn ngữ Prolog từ từ kết quả phân tích của WEKA để cập nhật các tập luật trước đó nhằm củng cố kết quả học máy và tăng khả năng chẩn đoán chính xác.

Chương 6: KIẾN NGHỊ NHỮNG NGHIÊN CỨU TIẾP THEO

Trên cơ sở những đạt được khi ứng dụng phần mềm vào hỗ trợ quá trình khám chữa bệnh, chương trình sẽ tiến tới phát triển các chức năng mới nhằm hỗ trợ người dùng thuận lợi và nhanh chóng hơn nữa trong việc sử dụng và đồng thời ứng dụng các công nghệ mới để hỗ trợ tăng tính hiệu quả của phần mềm như:

- Mở rộng phân tích các chỉ số xét nghiệm khác để phát hiện nhiều bệnh lý khác nhau.
- Phân tích nhiều chỉ số xét nghiệm đồng thời, các chỉ số xét nghiệm phức tạp nhằm đưa ra hướng hỗ trợ đánh giá sát với thực tế chẩn đoán.
- So sánh, phân tích, đánh giá các chỉ số của xét nghiệm của người bệnh trong quá trình theo dõi và điều trị để cảnh báo.
- Sử dụng hình ảnh siêu âm để nâng cao độ chính xác chẩn đoán.
- Tăng cường khai thác các thông tin khác liên quan đến bệnh nhân để tăng độ chính xác trong quá trình phân tích dữ liệu và đưa ra kết luận đầy đủ. .., Bổ sung phân tích nhiều chỉ số cận lâm sàng trên cùng một bệnh nhân để có cơ sở kết luận chẩn đoán đầy đủ các bệnh lý khi khám chữa bệnh cho bệnh nhân.

TÀI LIỆU THAM KHẢO

[1] T.M. Mitchell. Machine Learning. McGraw Hill, 1997

[2] ROKACH, Lior. Data Mining with Decision Trees. Series in Machine Perception and Artificial Intelligence: Volume 69.

[3] WITTEN, Ian H.; FRANK, Eibe. Data Mining: Practical machine learning tools and techniques. Morgan Kaufmann, 2005.

[4] PODGORELEC, Vili, et al. Decision trees: an overview and their use in medicine. Journal of medical systems, 2002, 26.5: 445-463.

[5] IRINA, Burd. Creating Decision Trees to Assess Cost-Effectiveness in Clinical Research. Journal of Biometrics & Biostatistics, 2012.

[6] BRUSEVA, Mariya. FINANCIAL RISK EVALUATION BY THE "TREE OF PROBABILITY DECISIONS" METHOD. In: Knowledge as Business Opportunity: Proceedings of the Management, Knowledge and Learning International Conference 2011. International School for Social and Business Studies, Celje, Slovenia, 2011. p. 327-334.

[7] HSSINA, Badr, et al. A comparative study of decision tree ID3 and C4. 5. International Journal of Advanced Computer Science and Applications, 2014, 4.2.

[8] Hướng dẫn chẩn đoán và điều trị một số bệnh về thận - tiết niệu (Ban hành kèm theo Quyết định số 3931/QĐ-BYT ngày 21/9/2015 của Bộ trưởng Bộ Y tế)

[9] Báo cáo của bệnh viện Bạch Mai, 2015

[10] Bùi Nhật Hằng. Hệ chẩn đoán suy thận dựa vào hệ chuyên gia theo từng bệnh nhân, 2012

[11] Thái Thị Bích Thủy; Nguyễn Thị Kim Ngân; Nguyễn Thị Diễm Thúy - xây dựng hệ chuyên gia "hệ hỗ trợ chẩn đoán một số bệnh thông thường ở trẻ em"

PHỤ LỤC

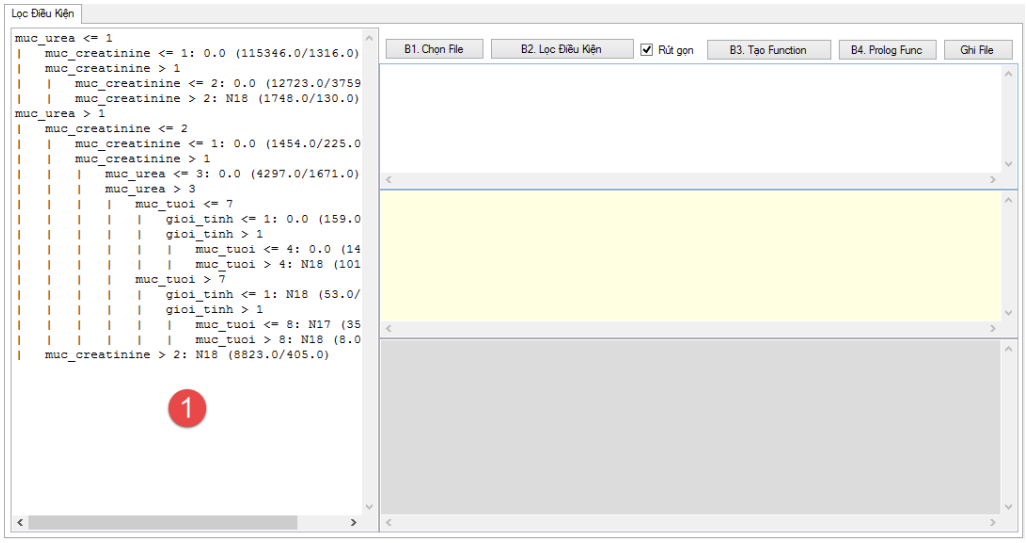
1. Chương trình demo cách hỗ trợ chẩn đoán và hướng dẫn điều trị
2. Bản nhận xét của cán bộ hỗ trợ chuyên môn.
3. Bảng xác nhận nguồn dữ liệu sử dụng trong luận văn của bệnh viện đa khoa trung ương Cần Thơ.

DEMO CHƯƠNG TRÌNH SỬ DỤNG PHẦN MỀM NHÚNG ĐỂ PHÂN TÍCH DỮ LIỆU BỆNH NHÂN

1. Phân tích kết quả từ phần mềm WEKA

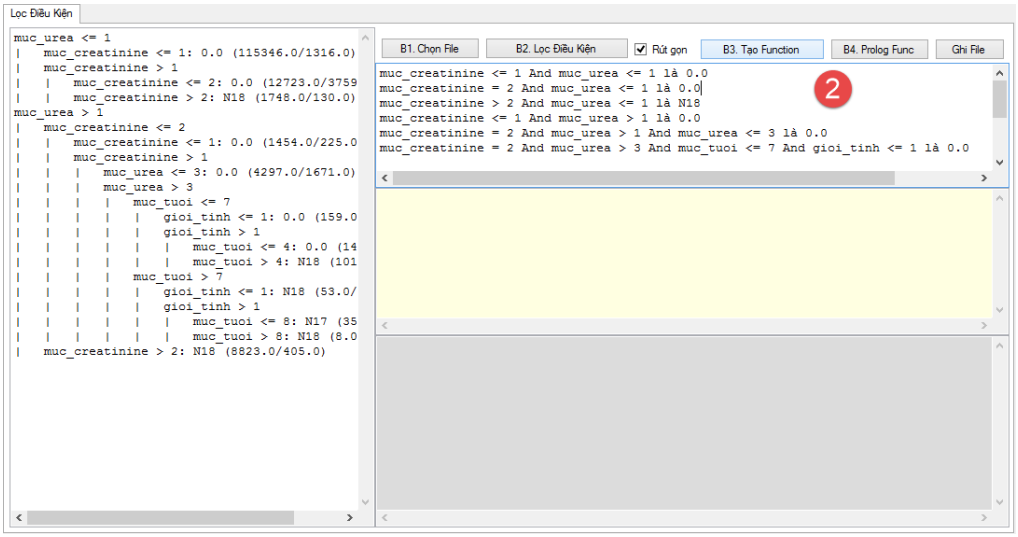
Qui trình phần mềm nhúng phân tích dữ liệu từ kết quả phân tích của WEKA thành các tập luật Prolog gồm:

- Bước 1) sử dụng kết quả phân tích từ chương trình WEKA



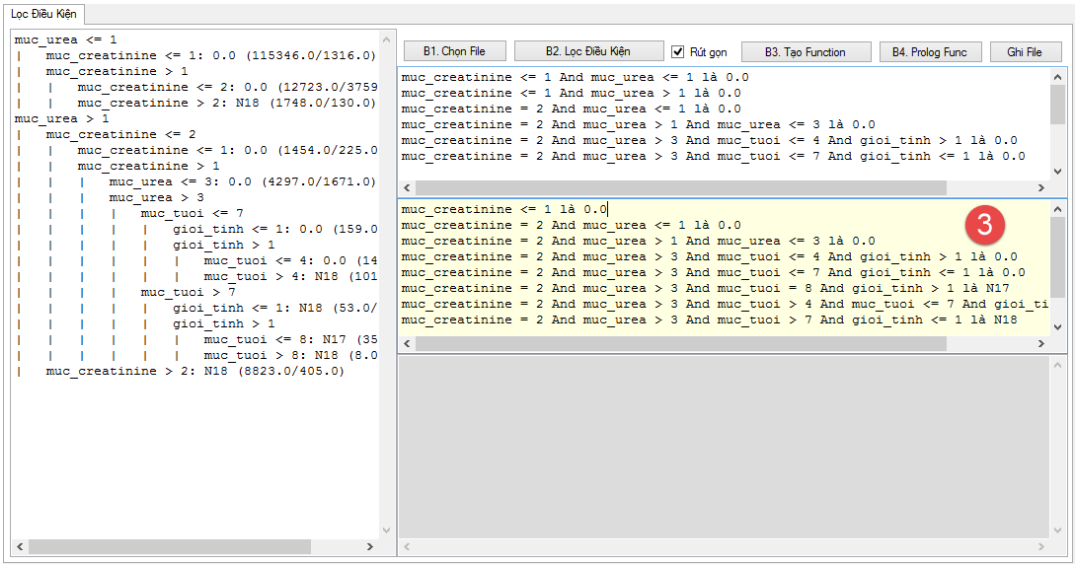
Hình Demo 1. Quá trình phân tích kết quả của WEKA

- Bước 2) chuyển đổi thành tập các điều kiện



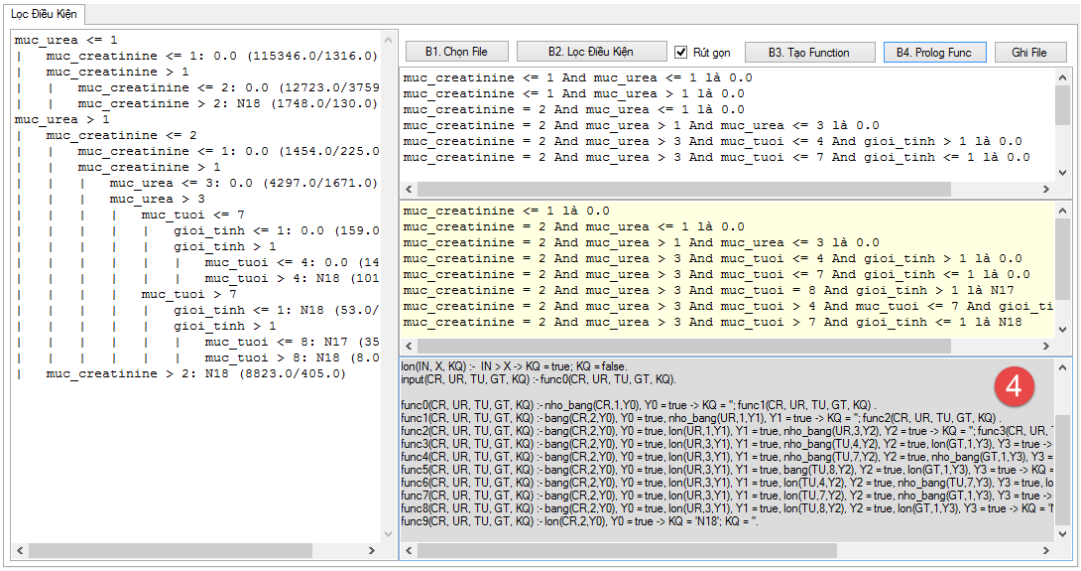
Hình Demo 2. Quá trình phân tích kết quả của WEKA

- Bước 3) Rút gọn các điều kiện



Hình Demo 3. Quá trình phân tích kết quả của WEKA

- Bước 4) chuyển đổi thành các câu lệnh theo cú pháp Prolog để thư viện P# sử dụng.

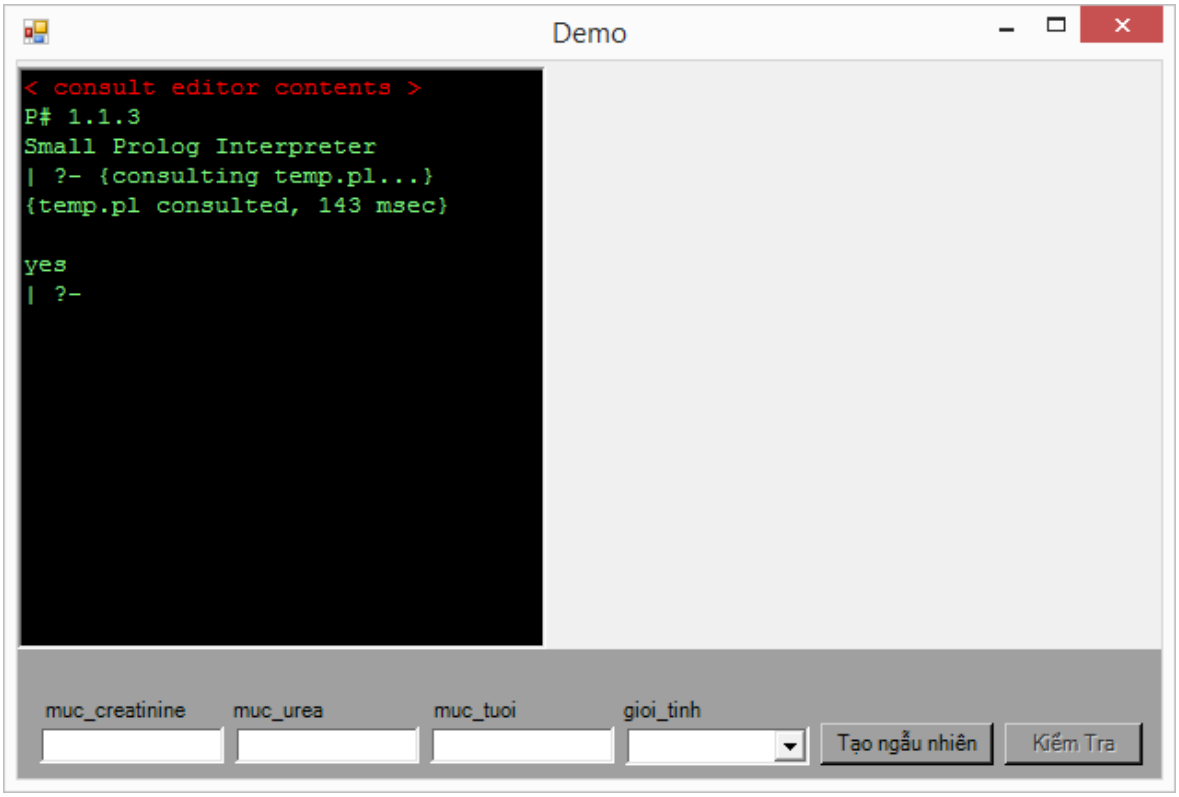


Hình Demo 4. Quá trình phân tích kết quả của WEKA

2. Các bước xây dựng chương trình nhúng (module tích hợp):

Bước 1) Tích hợp thư viện Psharp.dll.
Bước 2) Xây dựng tập tin Prolog chứa các tập luật được xây dựng từ WEKA
Bước 3) Thiết kết giao diện sử dụng (Windows Application Form) của Visual Studio.

Giao diện chương trình được thể hiện như sau:

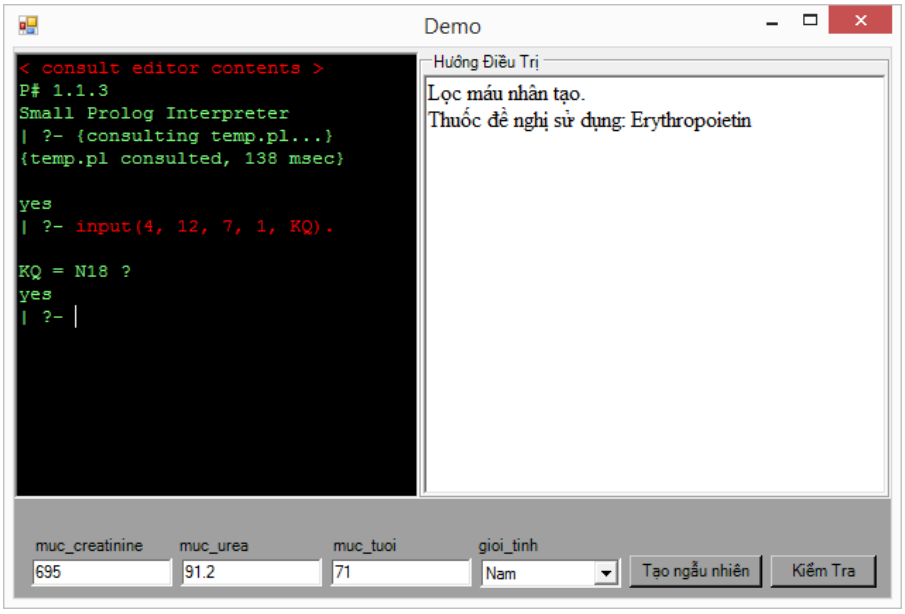


Hình Demo 5. Giao diện

3. Demo kết quả phân tích dữ liệu bệnh nhân.

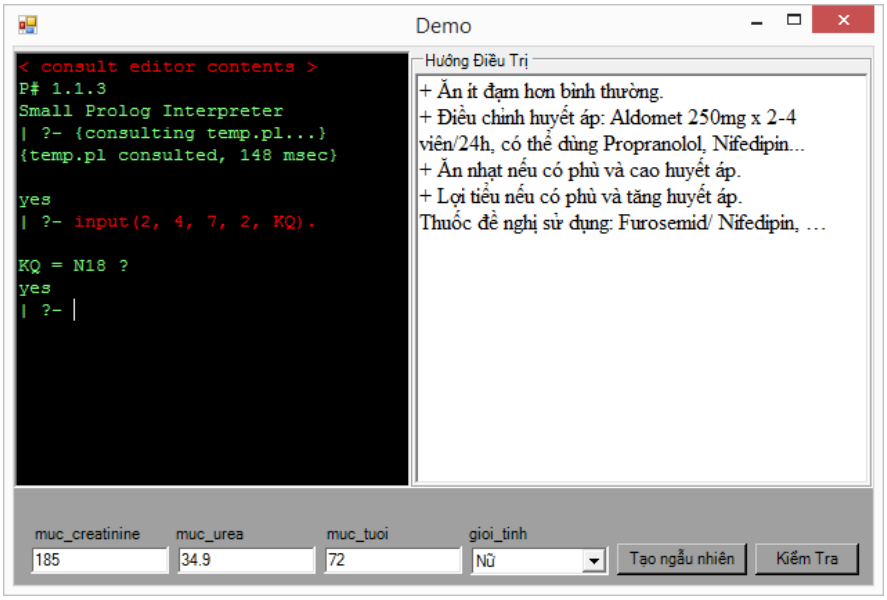
Chương trình demo cho phép tạo ngẫu nhiên các giá trị để kiểm tra kết quả phân tích dữ liệu của chương trình.

Với kết quả thử nghiệm như các hình bên dưới, các chỉ số sẽ được tiền xử lý và đưa thành dữ liệu đầu vào (input) để chương trình phân tích, trả kết quả và đề xuất các hướng điều trị.

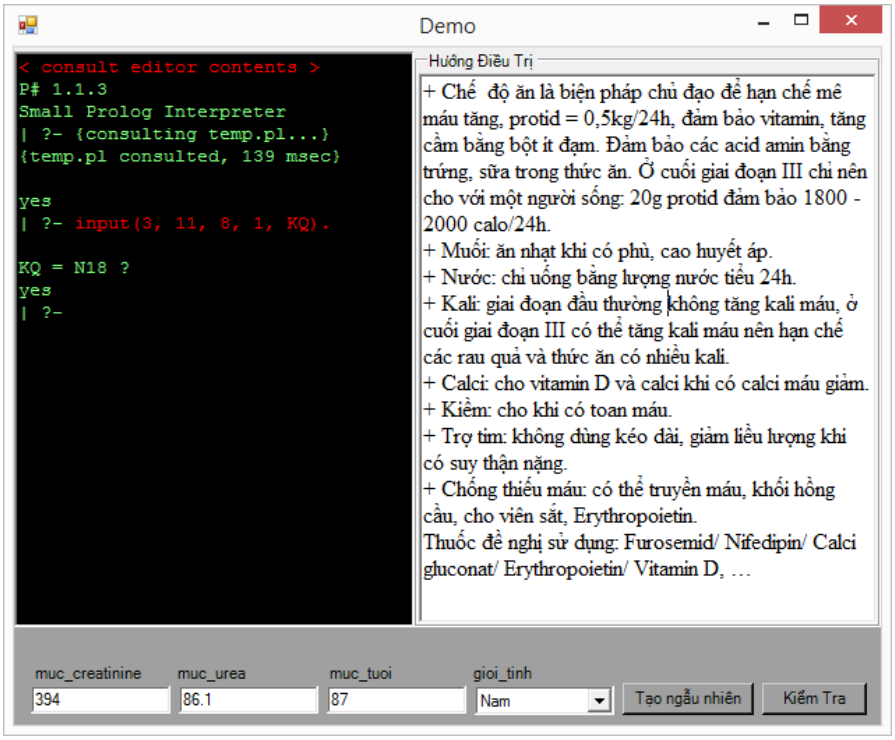


Hình Demo 6. Kết quả thử nghiệm

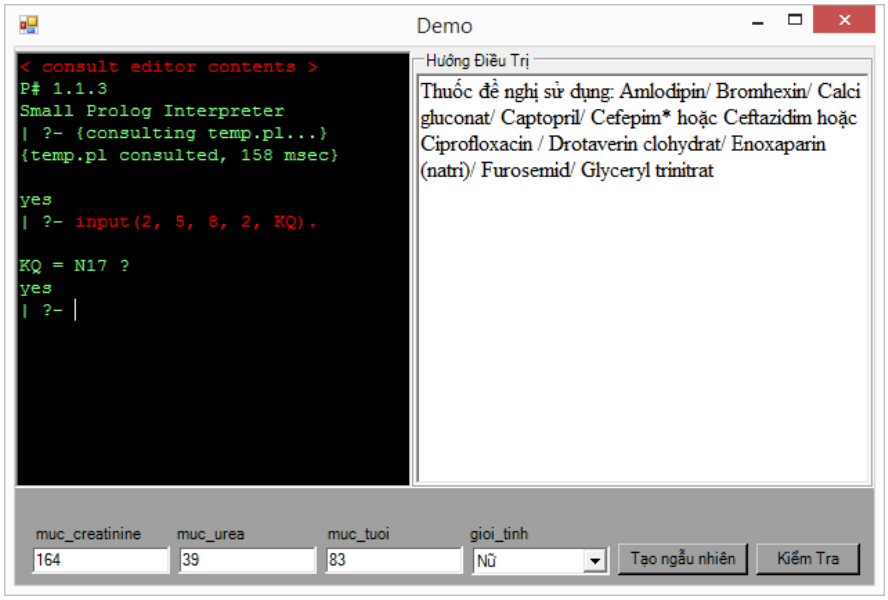
Hình Demo 7. Kết quả thử nghiệm



Hình Demo 8. Kết quả thử nghiệm



Hình Demo 9. Kết quả thử nghiệm



Hình Demo 10. Kết quả thử nghiệm