การพัฒนาแบบจำลองการคัดกรองผู้ป่วยโรคไตเรื้อรังโดยใช้เทคนิคเหมืองข้อมูล

THE DEVELOPMENT OF A SCREENING MODEL FOR CHRONIC KIDNEY DISEASE PATIENTS USING DATA MINING TECHNIQUES

จีระนันท์ โคตรสา[1] ศราวุฒิ บุญญะรัง[2] และ มงคล ทะกอง[3,*]
[1]สาขาวิชาวิทยาการข้อมูลและเทคโนโลยีสารสนเทศ มหาวิทยาลัยราชภัฏอุดรธานี
[2]โรงพยาบาลส่งเสริมสุขภาพตำบลวังดารา อำเภอบ้านดุง จังหวัดอุดรธานี
[3]ศูนย์วิเคราะห์ข้อมูลและปัญญาประดิษฐ์ มหาวิทยาลัยราชภัฏอุดรธานี
Jeeranan Kotsa[1] Sarawut Bunyarang[2] Mongkol Takong[3,*]
[1]Program in Data Science and Information Technology
[2]Wangdara Subdistrict Health Promotion Hospital
[3]Data Analytics and Artificial Intelligence Center, Udon thani Rajabhat University

(Received: January 9, 2023; Revised: April 15, 2023; Accepted: April 20, 2023)

*ผู้ประสานงาน : มงคล ทะกอง อีเมล : mongkhon@udru.ac.th

บทคัดย่อ

งานวิจัยนี้มีวัตถุประสงค์เพื่อพัฒนาแบบจำลองในการคัดกรองผู้ป่วยโรคไตเรื้อรังและเปรียบเทียบประสิทธิภาพของแบบจำลองของเทคนิคเหมืองข้อมูล 4 วิธี คือ วิธีแรนดอมฟอเรสต์ วิธีนาอีฟเบย์ วิธีโครงข่ายประสาทเทียม และวิธีต้นไม้ตัดสินใจ ในการทดลองได้นำข้อมูลผู้ป่วยโรคไตจากโรงพยาบาลส่งเสริมสุขภาพตำบลวังดารา อำเภอบ้านดุง จังหวัดอุดรธานี จำนวน 813 ชุดข้อมูล 5 คุณลักษณะ และวัดประสิทธิภาพของแบบจำลองทั้ง 4 วิธีด้วย 10-Fold Cross Validation จากผลการทดลองพบว่าวิธีแรนดอมฟอเรสต์มีประสิทธิภาพสูงสุดเมื่อเทียบกับวิธีอื่นๆ ที่ใช้เปรียบเทียบร่วมกัน ดังนั้นแบบจำลองวิธีแรนดอมฟอเรสต์สามารถช่วยทำนายและสนับสนุนการตัดสินใจทางการแพทย์เกี่ยวกับการวินิจฉัยระยะการเป็นโรคไตเรื้อรังได้อย่างมีประสิทธิภาพ

คำสำคัญ: โรคไตเรื้อรัง, แรนดอมฟอเรสต์, เหมืองข้อมูล

ABSTRACT

This research subjects to develop screening models for chronic kidney disease patients and compares the effectiveness of four data mining models which are. Random Forest, Naïve Bayes, Neural Networks, and Decision Tree. The data set of nephropathies from the Wangdara District Health Promoting Hospital, Bandung Amphoe, in Udon Thani Province was utilized in this experiment. It consisted of 813 rows and 5 features. The effectiveness of the four models was measured with 10-fold cross-validation. From the experimental results, it revealed that the Random Forest method can achieve the best performance which compared with the others. Therefore, the Random Forest method can effectively predict and diagnose chronic kidney disease to support medical decisions.

Keywords: Chronic Kidney Disease, Random Forest, Data Mining.

1. บทนำ

โรคไตเรื้อรัง (Chromic kidney disease) เป็นปัญหาทางด้านสาธารณสุขทั่วโลกและยังส่งผลกระทบทางด้านเศรษฐกิจที่สำคัญ พบผู้เสียชีวิตด้วยโรคไตประมาณร้อยละ 10-16 ของผู้เสียชีวิตทั้งหมดทั่วโลก สำหรับประเทศไทยจากการสุ่มตัวอย่างประชากรทั่วไปพบผู้ป่วยโรคไตเรื้อรังระยะที่ 1 ถึงระยะที่ 5 เท่ากับร้อยละ 17.5 ของผู้ป่วยทั้งหมดของประเทศไทย [1] โดยสาเหตุปัจจัยหลักสำคัญของโรคไตเรื้อรังมาจากโรคเบาหวาน ความดันโลหิตสูง และอายุที่มากขึ้น ที่ส่งผลต่อหลอดเลือดที่มาเลี้ยงไตเกิดการตีบแข็ง ทำให้การทำงานของไตลดลง [2] โดยพบว่าผู้ป่วยเบาหวานเป็นโรคไตเรื้อรัง 37.5 % จากปัจจัยดังกล่าวจึงส่งผลให้มีผู้เสียชีวิตจากโรคไตเรื้อรังก่อนวัย [3]

ผู้ป่วยโรคไตเรื้อรังส่วนมากเกิดจากการที่ป่วยเป็นโรคเบาหวาน ที่มีระดับน้ำตาลในเลือดมากทำให้กระทบต่อเส้นเลือดฝอยในไต ต่อแรงดันและความเร็วเลือดในไตเพิ่มสูงขึ้น พบการรั่วของโปรตีนชนิดอัลบูมินออกมาพร้อมกับปัสสาวะ จากการศึกษาปัจจัยที่มีผลกระทบและส่งผลต่อการเป็นโรคไตเรื้อรัง และการเฝ้าระมัดระวัง รวมถึงการสังเกตอาการตนเองจะช่วยป้องกันได้ ยิ่งไปกว่านั้นการตรวจสุขภาพและการวิเคราะห์ผลจากข้อมูลหรือคุณลักษณะที่จำเป็นด้วยตนเองยิ่งทำให้ทราบถึงโอกาสที่จะเป็นโรคได้ [4]

ดังนั้นคณะผู้วิจัยจึงได้เล็งเห็นถึงปัญหาและความสำคัญในการศึกษาข้อมูลสุขภาพของผู้ป่วยโรคไตเรื้อรังซึ่งจากผลการวัดคุณลักษณะที่สำคัญจากผลการตรวจของโรงพยาบาล เพื่อพัฒนา

แบบจำลองการคัดกรองผู้ป่วยโรคไตเรื้อรังโดยใช้เทคนิคเหมืองข้อมูล ในการจำแนกระยะการเป็นโรคไตเรื้อรัง เพื่อการวางแผนการรักษาอย่างมีประสิทธิภาพและลดอัตราการเสียชีวิตจากโรคไตเรื้อรังและโรคแทรกซ้อนอื่นๆ ได้

2. ทฤษฎีที่เกี่ยวข้อง

2.1 การทำเหมืองข้อมูล (Data Mining)

เป็นการวิเคราะห์ข้อมูลเพื่อจำแนกรูปแบบและความสัมพันธ์ของข้อมูลจากฐานข้อมูลที่มีขนาดใหญ่หรือคลังข้อมูล [5] โดยมีวิธีต่าง ๆ หลายวิธีซึ่งรูปแบบการทำเหมืองข้อมูลนั้นได้รวบรวมความรู้จากหลายแขนงเข้าไว้ด้วยกันซึ่งประกอบด้วยระบบการเรียนรู้ของเครื่องจักร (Machine Learning) ร่วมกับระบบฐานข้อมูล (Database System) วิทยาศาสตร์สารสนเทศ (Information Science) และสถิติ (Statistic) เทคนิคเหมืองข้อมูลในปัจจุบันมีหลายรูปแบบทั้งเทคนิคแบบผู้สอน (Supervised Learning) แบบไม่มีผู้สอน (Unsupervised Learning) แบบการเรียนรู้เกิดมาจากการปฏิสัมพันธ์ (Reinforcement Learning) ซึ่งผู้วิจัยได้ทำการศึกษาเทคนิคแบบมีผู้สอน เพื่อการจำแนกกลุ่มของข้อมูล ได้แก่ วิธีนาอีฟเบย์ (Naïve Bayes) วิธีโครงข่ายประสาทเทียม (Artificial Neural Networks) วิธีต้นไม้ตัดสินใจ (Decision Tree) วิธีแรนดอมฟอเรสต์ (Random Forest) เนื่องจากวิธีดังกล่าว เป็นเทคนิคที่นิยมและเหมาะสมกับการจำแนกข้อมูลที่เป็นหมวดหมู่ข้อมูล

2.1.1 วิธีนาอีฟเบย์ (Naïve Bayes) เป็นวิธีการจำแนกข้อมูลที่เหมาะกับกรณีของข้อมูลตัวอย่างที่มีจำนวนและคุณสมบัติ (Attribute) ไม่ขึ้นต่อกันมีจำนวนตัวจำแนกประเภทเบย์ไปประยุกต์ใช้งานด้านการจำแนกประเภทข้อความและพบว่าใช้งานได้ดีไม่ต่างจากการจำแนกประเภทวิธีอื่นๆ [6] ดังสมการที่ (1) และ (2)

$$P(c|x) = \frac{P(x|c)P(c)}{P(x)} \quad (1)$$

$$P(c|x) = P(x_1|c)\ x\ \dots x P(x_n|c)\ x\ P(c) \quad (2)$$

c คือ คลาสของข้อมูล

x คือ คุณสมบัติของข้อมูล

P คือ ความน่าจะเป็นของข้อมูล

$P(c|x)$ คือ ความน่าจะเป็นที่ข้อมูลที่มีแอตทริบิวต์เป็น x จะมีคลาส c

$P(x|c)$ คือ ความน่าจะเป็นที่ข้อมูลที่มีแอตทริบิวต์เป็น c จะมีคลาส x

$P(c)$ คือ จำนวนคลาสที่อาจจะเกิดขึ้นหารด้วยจำนวนคลาสทั้งหมดของคลาส c

$P(x)$ คือ จำนวนคุณสมบัติทั้งหมด

2.1.2 วิธีโครงข่ายประสาทเทียม (Artificial Neural Networks) ส่วนใหญ่จะประกอบด้วย 3 ชั้น คือ ชั้นข้อมูลเข้า ชั้นซ่อน และชั้นข้อมูลออกหรือชั้นผลลัพธ์ โครงข่ายแบบเชื่อมต่อกันอย่างสมบูรณ์เป็นโครงข่ายประสาทเทียมที่โหนดทุกโหนดในชั้นที่กำหนดเชื่อมต่อกันทุกโหนดกับชั้นถัดไป แม้จะไม่ได้เชื่อมต่อกับโหนดอื่นใดในชั้นเดียวกัน การเชื่อมต่อกันระหว่างโหนดมีการถ่วงน้ำหนักที่สัมพันธ์กันในขณะเริ่มต้น การจัดน้ำหนักถูกจัดอย่างสุ่มโดยมีค่าอยู่ระหว่าง 0 - 1 ซึ่งกำหนดให้จำนวนโหนดของข้อมูลเข้าเท่ากับจำนวนคุณลักษณะชุดข้อมูล จำนวนของชั้นซ่อนและจำนวนของโหนดในชั้นซ่อนขึ้นอยู่กับการประยุกต์ใช้งาน [7] หาได้จากสมการที่ (3) คือ

$$W_{ij}x_{ij} = W_{0j}x_{0j} + W_{1j}x_{1j} + \ \ldots + W_{lj}x_{lj} \tag{3}$$

โดยที่ x_{ij} คือ ข้อมูลเข้าที่ *i* ไปยังโหนดที่ *j*

W_{ij} คือ น้ำหนักถ่วงที่สัมพันธ์กับข้อมูลเข้าที่ *i* ไปยังโหนด *j*

2.1.3 วิธีต้นไม้ตัดสินใจ (Decision Tree) เป็นการจำแนกค่าคุณลักษะของข้อมูล โดยจะประกอบด้วยโหนด และกิ่ง ที่ต่อกับโหนด โหนดที่ปลายสุดจะเรียก ว่าใบต้นไม้ตัดสินใจจะทำโดยสร้างโหนดทีละโหนดเพื่อตรวจสอบคุณสมบัติของตัวอย่าง แล้วแยกตัวอย่างลงตามค่าของกิ่ง [8] ดังสมการที่ (4) , (5) และ (6)

$$I(s_1, s_2 \ldots, s_n) = -\sum_{i=1}^{n} \frac{s_i}{s} \log_2 \frac{s_i}{s} \tag{4}$$

$$E(A) = \sum_{j=1}^{n} \frac{s_{1j+\cdots+s_{nj}}}{s} \ I\left(s_{1j}, s_{2j}, \ldots s_{nj}\right) \tag{5}$$

$$Gain(A) = I\left(s_{1j}, s_{2j}, \ldots s_{nj}\right) - E(A) \tag{6}$$

S เป็นเซตของข้อมูลซึ่งประกอบด้วยข้อมูล s ระเบียน

n เป็นจำนวนกลุ่มทั้งหมดที่ต่างกันของข้อมูลชุดนั้น

c_i แทนกลุ่มในลำดับที่ i โดยที่ i มีค่าระหว่าง 1 ถึง n

s_i แทนจำนวนข้อมูลสมาชิกของ s และอยู่ในกลุ่ม c_i

2.1.4 วิธีแรนดอมฟอเรสต์ (Random Forest) เป็นโมเดลอีกประเภทหนึ่งของเครื่องจักรเรียนรู้ (Machine Learning) ถูกพัฒนามาจากต้นไม้ตัดสินใจ ต่างกันที่วิธีแรนดอมฟอเรสต์เป็นการเพิ่มจำนวนต้นไม้ตัดสินใจหลายๆ ต้น ทำให้ประสิทธิภาพในการทำงานสูงขึ้นแม่นยำมากขึ้นกำหนดให้ $RFfi$ คือคุณลักษณะที่คำนวณได้จากต้นไม้ทั้งหมดในแบบจำลอง $normf_{i_{i_j}}$ คือ คุณสมบัติที่ทำให้เป็นมาตรฐานสำหรับ i ใน $tree_j$ และ T คือจำนวนต้นไม้ทั้งหมด [8] หาได้จากสมการที่ (7) ดังนี้

$$RFfij = \frac{\sum j \in all\ tree\ normf_{i_{i_j}}}{T} \quad (7)$$

2.1.5 การวัดประสิทธิภาพแบบจำลองโดยใช้วิธี K-Fold Cross Validation คือทำการแบ่งข้อมูลออกเป็นส่วนๆ โดยที่แต่ละส่วนมีจำนวนข้อมูลเท่ากันและนำแต่ละส่วนไปทดสอบ จากนั้นเก็บค่าความถูกต้องของโมเดลแต่ละรอบไว้แล้วนำมาพิจารณาค่าเฉลี่ยทุกกลุ่มข้อมูล [9 ,10] ดังสมการที่ (8) , (9) , (10) และ (11)

- ค่าความถูกต้อง (Accuracy) คือ ค่าที่ทำนายข้อมูลถูกต้องทุกคลาส

$$Accuracy = \frac{TP+TN}{TP+TN+FP+FN} \quad (8)$$

- ค่าความแม่นยำ (Precision) คือ ค่าที่ทำนายแล้วเป็นการทายถูกต้อง

$$Precision = \frac{TP}{TP+TN} \quad (9)$$

- ค่าความระลึก (Recall) คือ ค่าที่ทำนายความถูกต้อง

$$Recall = \frac{TP}{TP+FN} \quad (10)$$

- ค่าความถ่วงดุล (F-measure) คือ ค่าเฉลี่ยของค่าความแม่นยำและค่าความระลึก

$$F - Measure = \frac{Precision\ x\ Recall}{Precision + Recall} \quad (11)$$

TP = อัตราความถูกต้องเชิงบวก

TN = อัตราความถูกต้องเชิงลบ

FP = อัตราความผิดพลาดเชิงบวก

FN = อัตราความผิดพลาดเชิงลบ

2.1.6 G-mean ใช้ในการทดสอบประสิทธิภาพของโมเดลและข้อมูลที่ไม่สมดุลของคลาส ย่อมาจาก Geometric mean การหาค่า G-mean จะใช้การคำนวณจากค่า Sensitivity และ Specificity ในการคำนวณ ซึ่งสามารถคำนวณได้จากสมการ (12), (13) และ (14) ดังนี้

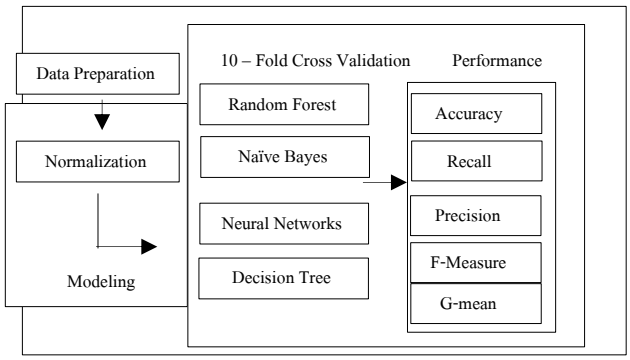
$$Sensitivity = \frac{TP}{TP+FN} \quad (12)$$

$$Specificity = \frac{TN}{TN+FP} \quad (13)$$

$$G - Mean = \sqrt{sensitivity * specificity} \quad (14)$$

3. วิธีการดำเนินงานวิจัย

การวิจัยครั้งนี้ได้มีการนำข้อมูลของผู้ป่วยโรคไตเรื้อรังมาประยุกต์ใช้งานกับเทคนิคการทำเหมืองข้อมูลซึ่งจะพัฒนาแบบจำลองการคัดกรองผู้ป่วยโรคไตเรื้อรังโดยใช้เทคนิคเหมืองข้อมูล โดยมีขั้นตอนการ ดำเนินงาน ขั้นตอนที่ 1 การเตรียมข้อมูล (Data Preparation) ผู้วิจัยศึกษาทำเข้าใจเกี่ยวกับข้อมูลผู้ป่วยโรคไตเรื้อรัง และการ Normalization ด้วยวิธีการ Min-Max Normalization เพื่อเป็นการแปลงช่วงทำให้ค่าของคุณลักษณะ (Attribute) ทั้งหมดเป็นช่วงค่าตามที่ระบุ ขั้นตอนที่ 2 เข้าสู่กระบวนการสร้างแบบจำลองโดยมี 4 วิธีประกอบด้วย Random Forest, Naïve Bayes, Neural Networks และ Decision Tree ขั้นตอนที่ 3 การทดสอบประสิทธิภาพของโมเดลโดยใช้เทคนิค 10-Fold Cross Validation เพื่อหาค่าความถูกต้อง ค่าความแม่นยำ ค่าความระลึก ค่าความถ่วงดุล และค่าความไม่สมดุล ดังรูปที่ 1



รูปที่ 1 ขั้นตอนการดำเนินงานวิจัย

3.1 ขั้นตอนการเตรียมข้อมูล ผู้วิจัยได้ศึกษาข้อมูลที่ได้จากการลงพื้นที่เก็บข้อมูลผู้ป่วยโรคไตเรื้อรัง โดยการทบทวนเวชระเบียนผู้ป่วยโรคไตเรื้อรังจากโรงพยาบาลส่งเสริมตำบลวังดารา ตั้งแต่ปี พ.ศ. 2563-2565 ทั้งหมด ซึ่งอยู่ในรูปแบบตารางไฟล์ Excel จำนวน 1,145 รายการ จากนั้นได้ทำความสะอาดข้อมูล (Data Cleaning) ทำการลบข้อมูลซ้ำซ้อน และข้อมูลที่ไม่สมบูรณ์ (Incomplete Data) ให้เหลือเฉพาะข้อมูลที่สำคัญและมีคุณลักษณะที่เพียงพอต่อการวิเคราะห์สภาวะความเสี่ยงของการเกิดโรคไตเรื้อรังเพื่อให้ได้ข้อมูลที่สมบูรณ์ในการพัฒนาแบบจำลองคัดกรองผู้ป่วยโรคไตเรื้อรัง ประกอบด้วยข้อมูลผู้ป่วยโรคไตเรื้อรังระยะที่ 1 (Class 1) จำนวน 424 คน , ระยะที่ 2 (Class 2) จำนวน 286 คน , ระยะที่ 3 (Class 3) จำนวน 94 คน และระยะที่ 4 (Class 4) จำนวน 9 คน และผ่านกระบวนการ Normalization ซึ่งข้อมูลที่เหลือทั้งหมดเท่ากับ 813 รายการ 5 คุณลักษณะ ดังตัวอย่างข้อมูลแสดงในตารางที่ 1 โดยการกำหนดของแพทย์ผู้ร่วมวิจัยซึ่งได้พิจารณาเลือกคุณลักษณะในการดำเนินงานวิจัยและได้ให้ข้อสังเกตว่าคุณลักษณะที่เลือกสามารถนำมาใช้ในการวิจัยและเพียงพอที่จะสามารถวิเคราะห์แสดงถึงการเกิดโรคไตเรื้อรังในแต่ละระยะได้ ประกอบด้วย Sex , Age , Microalbumin Result , Creatinine Result และ eGFR Result โดยรายละเอียดของคุณลักษณะดังตารางที่ 2 และตารางที่ 3 ทั้งนี้ผู้วิจัยได้มีการขอจริยธรรมการวิจัยในมนุษย์และได้ปฏิบัติตามกฎระเบียบ ข้อบังคับ แนวทางมาตรฐานวิชาชีพและวางแผนควบคุมคุณภาพของการเก็บข้อมูลอย่างเป็นระบบ เพื่อให้ได้ข้อมูลที่ถูกต้องตามความเป็นจริงมากที่สุด โดยมีขั้นตอนการดำเนินการคือ ภายหลังการจัดเก็บข้อมูลจากประวัติผู้ป่วยในโรงพยาบาล ซึ่งเป็นคุณหมอประจำโรงพยาบาลส่งเสริมตำบลวังดารา เป็นผู้วิจัยร่วมดำเนินการตรวจสอบข้อมูลว่าเป็นไปตามกระบวนการวิจัยได้ระบุไว้อย่างละเอียด และให้ตรงตามความเป็นจริงที่ปรากฏในเวชระเบียนของผู้ป่วยโรคไตเรื้อรัง

ตารางที่ 1 ตัวอย่างข้อมูลผู้ป่วยโรคไตเรื้อรัง

Sex	Age	Microalbumin Result	Creatinine Result	eGFR Result	Class
2	21	0	0.520	138	1
1	10	0	1.120	66	2
2	78	1	1.210	43	3
2	92	0	1.630	27	4

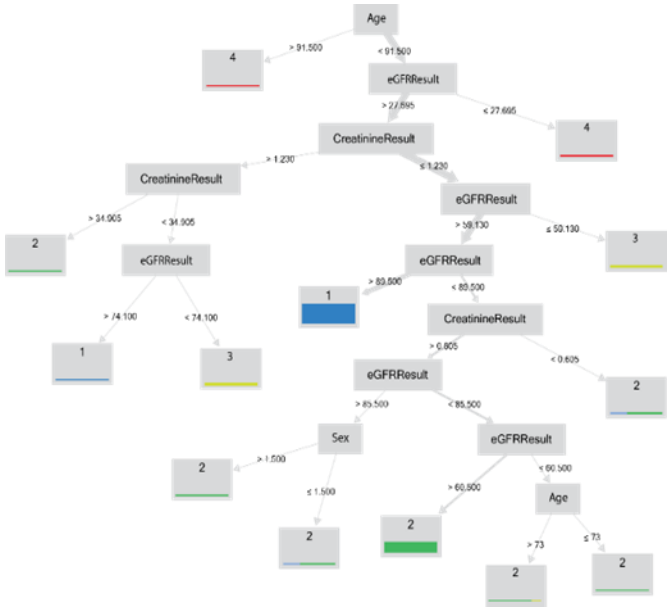
ตารางที่ 2 ระยะโรคไตเรื้อรังพิจารณาจากค่า GFR

ระยะ	ค่า GFR	คำอธิบาย
1	>=90	ปกติ
2	60-89	ไตผิดปกติ และอัตรากรองไตลดลง
3A	45-59	อัตรากรองไตลดลงปานกลางและเสี่ยงต่อไตล้มเหลวต่ำ
3B	30-44	อัตรากรองไตลดลงปานกลางและเสี่ยงต่อไตล้มเหลวต่ำ
4	15-29	อัตรากรองไตลดลงมากและมีความเสี่ยงต่อไตล้มเหลวสูง
5	<15	ไตล้มเหลว (โรคไตเรื้อรังระยะสุดท้าย)

ตารางที่ 3 คุณลักษณะของผู้ป่วยโรคไตเรื้อรัง

คุณลักษณะ	ความหมาย
Sex	เพศ
Age	อายุ
Microalbumin Result	ค่าระดับโปรตีนในปัสสาวะ
Creatinine Result	ค่าระดับการทำงานของไต
eGFR Result	อัตราการกรองของเสียของไต
Class	ผลการเป็นโรคไตเรื้อรัง 5 ระยะ 1: ระยะที่ 1 ปกติ 2: ระยะที่ 2 ไตผิดปกติ และอัตรากรองไตลดลง 3: ระยะที่ 3 อัตรากรองไตลดลงปานกลางและเสี่ยงต่อไตล้มเหลวต่ำ 4: ระยะที่ 4 อัตรากรองไตลดลงมากและมีความเสี่ยงต่อไตล้มเหลวสูง 5: ระยะที่ 5 ไตล้มเหลว (โรคไตเรื้อรังระยะสุดท้าย)

3.2 การสร้างแบบจำลองงานวิจัยนี้ผู้วิจัยได้ใช้เครื่องมือในการวิเคราะห์ข้อมูลคือ โปรแกรม RapidMiner Version.9.8 และเทคนิคการจำแนกข้อมูลประกอบด้วย วิธีแรนดอมฟอเรสต์, วิธีนาอีฟเบย์, วิธีโครงข่ายประสาทเทียม และวิธีต้นไม้ตัดสินใจ ทั้งนี้ในการสร้างแบบจำลองทั้ง 4 วิธี แบบจำลองการจำแนกผู้ป่วยโรคไตเรื้อรังด้วยแรน ดอมฟอเรสต์ สามารถจำแนกการเกิดโรคไตเรื้อรังแต่ละระยะได้ดังรูปที่ 2 สีน้ำเงินระยะที่ 1 ปกติ, สีเขียวระยะที่ 2 ไตผิดปกติและอัตรากรองไตลดลง, สีเหลืองระยะที่ 3 อัตรากรองไตลดลงปานกลางและเสี่ยงต่อไตล้มเหลวต่ำ และสีแดงระยะที่ 4 อัตรากรองไตลดลงมากและมีความเสี่ยงต่อไตล้มเหลวสูง ตามลำดับ และกฎการจำแนกด้วยแรนดอมฟอเรสต์ 13 กฎ ดังรูปที่ 3



รูปที่ 2 แบบจำลองการจำแนกผู้ป่วยโรคไตเรื้อรังด้วยแรนดอมฟอเรสต์

```
Tree

Age > 91.500: 4  (1=0, 2=0, 3=0, 4=7)
Age ≤ 91.500: 4
|   eGFRResult > 27.695
|   |   CreatinineResult > 1.230
|   |   |   CreatinineResult > 34.905: 2 (1=0, 2=3, 3=0, 4=0)
|   |   |   CreatinineResult ≤ 34.905
|   |   |   |   eGFRResult > 74.100: 1 (1=5, 2=0, 3=0, 4=0)
|   |   |   |   eGFRResult ≤ 74.100: 3 (1=0, 2=0, 3=48, 4=0)
|   |   CreatinineResult ≤ 1.230
|   |   |   eGFRResult > 59.130
|   |   |   |   eGFRResult > 89.500: 1 (1=422, 2=0, 3=0, 4=0)
|   |   |   |   eGFRResult ≤ 89.500
|   |   |   |   |   CreatinineResult > 0.605
|   |   |   |   |   |   eGFRResult > 85.500
|   |   |   |   |   |   |   Sex > 1.500: 2 (1=0, 2=22, 3=0, 4=0)
|   |   |   |   |   |   |   Sex ≤ 1.500: 2 (1=3, 2=6, 3=0, 4=0)
|   |   |   |   |   |   eGFRResult ≤ 85.500
|   |   |   |   |   |   |   eGFRResult > 60.500: 2 (1=0, 2=231, 3=0, 4=0)
|   |   |   |   |   |   |   eGFRResult ≤ 60.500
|   |   |   |   |   |   |   |   Age > 73: 2  (1=0, 2=5, 3=1, 4=0)
|   |   |   |   |   |   |   |   Age ≤ 73: 2  (1=0, 2=8, 3=0, 4=0)
|   |   |   |   |   CreatinineResult ≤ 0.605: 2 (1=1, 2=3, 3=0, 4=0)
|   |   |   eGFRResult ≤ 59.130: 3 (1=0, 2=0, 3=41, 4=0)
|   eGFRResult ≤ 27.695: 4 (1=0, 2=0, 3=0, 4=7)
```

รูปที่ 3 กฎการจำแนกด้วยแรนดอมฟอเรสต์

ยกตัวอย่างกฎการจำแนกด้วยแรนดอมฟอเรสต์ 6 กฎ เช่น

กฎข้อที่ 1 ถ้า Age (อายุ) มากกว่า 91.500 ผลลัพธ์คือระยะที่ 4 อัตรากรองไตลดลงมาก และมีความเสี่ยงต่อไตล้มเหลวสูง

กฎข้อที่ 2 ถ้า Age (อายุ) น้อยกว่าหรือเท่ากับ 91.500 ให้ไปดู eGFRResult มากกว่าเท่ากับ 27.695 ถ้ามากกว่า ให้ไปดู CreatinineResult มากกว่า 34.905 ผลลัพธ์คือระยะที่ 2 ไตผิดปกติ และอัตรากรองไตลดลง

กฎข้อที่ 3 ถ้า Age (อายุ) น้อยกว่าหรือเท่ากับ 91.500 ให้ไปดู eGFRResult มากกว่าเท่ากับ 27.695 ถ้ามากกว่า ให้ไปดู CreatinineResult น้อยกว่าเท่ากับ 34.905 ถ้าน้อยกว่าเท่ากับ ให้ไปดู eGFRResult มากกว่า 74.100 ผลลัพธ์คือระยะที่ 1 ปกติ

กฎข้อที่ 4 ถ้า Age (อายุ) น้อยกว่าหรือเท่ากับ 91.500 ให้ไปดู eGFRResult มากกว่าเท่ากับ 27.695 ถ้ามากกว่า ให้ไปดู CreatinineResult น้อยกว่าเท่ากับ 34.905 ถ้าน้อยกว่าเท่ากับ ให้ไปดู eGFRResult น้อยกว่าเท่ากับ 74.100 ผลลัพธ์คือระยะที่ 3 อัตรากรองไตลดลงปานกลางและเสี่ยงต่อไตล้มเหลวต่ำ

กฎข้อที่ 5 Age (อายุ) น้อยกว่าหรือเท่ากับ 91.500ให้ไปดู eGFRResult มากกว่าเท่ากับ 27.695 ถ้ามากกว่า ให้ไปดู CreatinineResult น้อยกว่าเท่ากับ 1.230 ถ้าน้อยกว่ากับเท่ากับ ให้ไปดู eGFRResult มากกว่า 89.500 ถ้ามากกว่า ผลลัพธ์คือระยะที่ 1 ปกติ

กฎข้อที่ 6 Age (อายุ) น้อยกว่าหรือเท่ากับ 91.500ให้ไปดู eGFRResult มากกว่าเท่ากับ 27.695 ถ้ามากกว่า ให้ไปดู CreatinineResultน้อยกว่าเท่ากับ 1.230 ถ้าน้อยกว่ากับเท่ากับ ให้ไปดู eGFRResult น้อยกว่าเท่ากับ 89.500 ถ้าน้อยกว่ากับเท่ากับ ให้ไปดู CreatinineResult มากกว่า 0.605 ถ้ามากกว่า ให้ไปดู eGFRResult มากกว่า 85.500 ถ้ามากกว่า ให้ไปดู Sex มากกว่า 1.500 ผลลัพธ์คือระยะที่ 2 ไตผิดปกติ และอัตรากรองไตลดลง

3.3 การวัดประสิทธิภาพด้วยวิธี10-Fold Cross Validation การวัดประสิทธิภาพด้วยวิธีนี้จะทำการแบ่งข้อมูลออกเป็น 10 ส่วน โดยที่แต่ละส่วนมีจำนวนข้อมูลเท่ากันแล้วนำข้อมูลเข้าทดสอบในแบบจำลอง โดยนำคุณลักษณะทั้ง 5 คุณลักษณะ 813 รายการเข้าสู่กระบวนการสร้างแบบจำลองทั้ง 4 วิธี เพื่อวัดค่าประสิทธิภาพและเปรียบเทียบการทำนายของแบบจำลองแต่และวิธี ผู้วิจัยได้มีการเลือกใช้เกณฑ์ที่เหมาะสมกับแต่ละวิธี วิธีแรนดอมฟอเรสต์ มีการเลือกใช้เกณฑ์แบบ Gain Ratio คือ รับข้อมูลสำหรับแต่ละคุณลักษณะ (Attribute) เพื่อให้ค่าคุณลักษณะมีความกว้างและสม่ำเสมอ โดยมีจำนวนต้นไม้เท่ากับ 100 ต้น และความลึกของต้นไม้เท่ากับ 10 ซึ่งเป็นชุดพารามิเตอร์ที่ให้ผลการทดสอบที่มีค่าความถูกต้องสูงสุดเมื่อเทียบกับอีก 3 วิธี วิธีนาอีฟเบย์ เป็นเทคนิคการจำแจกตามการใช้ทฤษฎีเบย์ เพื่อหาคำตอบในแต่ละคลาส (Class) และมีการกำหนดคุณลักษณะ โดยแทนค่าตามสมการที่ (1) และ (2) วิธีโครงข่ายประสาทเทียมมีการกำหนดชั้นซ่อน (Hidden Layer) เท่ากับ 2 ซึ่งส่งผลต่อชุดพารามิเตอร์ที่ให้ผลทดสอบที่ดีที่สุด วิธีต้นไม้ตัดสินใจเลือกใช้เกณฑ์แบบ Gini Index การวัดความไม่เท่าเทียมกันระหว่างการกระจายของคุณลักษณะการแยก ที่เลือกส่งผลต่อชุดพารามิเตอร์และกำหนดความลึกของต้นไม้เท่ากับ 10 ซึ่งเป็นชุดพารามิเตอร์ที่ให้ผลการทดสอบที่ดีที่สุด ดังตารางที่ 4 และตารางที่ 5

4. ผลการวิจัย

ผลการทดลองวิเคราะห์ค่าความถูกต้องของแบบจำลองการคัดกรองผู้ป่วยโรคไตเรื้อรังโดยใช้เทคนิคเหมืองข้อมูลแต่ละวิธี ดังตารางที่ 4 และตารางที่ 5

ตารางที่ 4 ผลการวัดประสิทธิภาพของโมเดลด้วย 10-Fold Cross Validation ของข้อมูลทั้ง 5 คุณลักษณะ

Method	Performance				
	Accuracy	Recall	Precision	F-Measure	G-mean
Random Forest	**98.89%**	**98.56%**	98.67%	**98.61%**	**99.76%**
Naïve Bayes	75.53%	81.88%	73.58%	77.51%	81.64%
Neural Networks	94.71%	69.23%	69.94%	69.58%	97.69%
Decision Tree	98.77%	98.47%	**99.03%**	98.47%	99.38%

จากตารางที่ 4 ผลการทดลองพบว่าวิธีแรนดอมฟอเรสต์ มีค่าความถูกต้อง, ค่าความระลึก, ค่าถ่วงน้ำหนัก และค่า G-mean สูงที่สุดเท่ากับ 98.89% , 98.56% , 98.61% และ 99.76% ตามลำดับ ส่วนค่าความแม่นยำของวิธีแรนดอมฟอเรสต์มีค่าน้อยกว่าวิธีต้นไม้ตัดสินใจเท่ากับ 99.03% ตามลำดับ

ตารางที่ 5 ผลการวัดประสิทธิภาพของโมเดลด้วย 10-Fold Cross Validation ของข้อมูล 4 คุณลักษณะ (ยกเว้นค่า eGFR Result)

Method	Performance				
	Accuracy	Recall	Precision	F-Measure	G-mean
Random Forest	**94.96%**	**98.56%**	**98.67%**	**98.61%**	**96.10%**
Naïve Bayes	56.59%	59.75%	44.22%	50.83%	64.16%
Neural Networks	61.02%	32.75%	29.02%	30.77%	65.89%
Decision Tree	85.12%	87.52%	88.33%	87.92%	87.36%

จากตารางที่ 5 ประสิทธิภาพของวิธีแรนดอมฟอเรสต์ มีค่าความถูกต้อง, ค่าความระลึก ค่าความแม่นยำ ค่าถ่วงน้ำหนัก และค่า G-mean สูงที่สุดเท่ากับ 94.96% , 98.56% , 98.67% , 98.61% และ 96.10% ตามลำดับ เมื่อเทียบกับอีก 3 วิธี

จากผลการทดลองตารางที่ 4 และตารางที่ 5 จะเห็นได้ว่าวิธีแรนดอมฟอเรสต์ มีประสิทธิภาพในการจำแนกคัดกรองผู้ป่วยโรคไตเรื้อรังแต่ละระยะ เช่น ถ้า Age (อายุ) มากกว่า 91.500 ผลลัพธ์คือระยะที่ 4 อัตรากรองไตลดลงมากและมีความเสี่ยงต่อไตล้มเหลวสูง ซึ่งแม่นยำและถูกต้องที่สุดเมื่อเปรียบเทียบกับ วิธีนาอีฟเบย์ วิธีโครงข่ายประสาทเทียม และวิธีต้นไม้ตัดสินใจ ทั้งในกรณีที่ใช้คุณลักษณะทั้ง 5 คุณลักษณะ และกรณีทดลองด้วยการใช้คุณลักษณะเพียง 4 คุณลักษณะ (ยกเว้นค่า eGFR Result)

5. สรุปผลและอภิปราย

บทความนี้มีวัตถุประสงค์เพื่อพัฒนาแบบจำลองการคัดกรองผู้ป่วยโรคไตเรื้อรังโดยใช้เทคนิคเหมืองข้อมูลกรณีข้อมูลของโรงพยาบาลส่งเสริมตำบลวังดารา อำเภอบ้านดุง จังหวัดอุดรธานี ที่เกิดจากการทบทวนเวชทะเบียนผู้ป่วยโรคไตเรื้อรังย้อนหลังตั้งแต่ปีพ.ศ. 2563-2565 จำนวน 813 รายการ 5 คุณลักษณะ จากผลการวิจัยจากตารางที่ 4 และตารางที่ 5 สามารถสรุปผลได้ประสิทธิภาพของโมเดลทั้ง 4 วิธี ประกอบด้วยวิธีแรนดอมฟอเรสต์ วิธีนาอีฟเบย์ วิธีโครงข่ายประสาทเทียม และ

วิธีต้นไม้ตัดสินใจ ซึ่งวัดประสิทธิภาพในการทำนายโดยใช้ 10-Fold Cross Validation ซึ่งวิธีแรนดอมฟอเรสต์มีประสิทธิภาพสูงสุดประกอบด้วยค่าความถูกต้อง , ค่าความระลึก, ค่าถ่วงน้ำหนัก และค่า G-mean เท่ากับ 98.89% , 98.56% , 98.61% และ 99.76% ตามลำดับ เมื่อเปรียบเทียบกับวิธีนาอีฟเบย์ วิธีโครงข่ายประสาทเทียม และวิธีต้นไม้ตัดสินใจ ในกรณีของค่าความแม่นยำและค่าถ่วงน้ำหนักของวิธีแรน ดอมฟอเรสต์มีค่าสูงกว่าวิธีวิธีนาอีฟเบย์ และวิธีโครงข่ายประสาทเทียมแต่มีค่าน้อยกว่าวิธีต้นไม้ตัดสินใจเนื่องจากค่าผลลัพธ์ของอัตราความถูกต้องเชิงลบมีค่าสูง จึงส่งผลทำให้ค่าความแม่นยำและค่าถ่วงน้ำหนักของวิธีต้นไม้ตัดสินใจมีค่าสูงกว่าวิธีแรนดอมฟอเรสต์

แต่เนื่องจากข้อมูลผู้ป่วยโรคไตเรื้อรังที่ใช้ในการทดลองวิจัยอยู่ในรูปแบบข้อมูลที่ไม่สมดุล (Imbalanced Data) จึงได้มีการนำค่า G-mean มาใช้ร่วมในการวัดค่าประสิทธิภาพของแต่ละวิธีดังตารางที่ 4 และตารางที่ 5 ซึ่งจากผลการทดลองพบว่า วิธีแรนดอมฟอเรสต์มีค่าสูงสุดเมื่อเทียบกับอีก 3 วิธี

การสรุปผลการวิจัยในครั้งนี้เห็นได้ว่าการพัฒนาแบบจำลองด้วยวิธีแรนดอมฟอเรสต์มีประสิทธิภาพมากที่สุด เมื่อเทียบกับวิธีที่ใช้เปรียบเทียบร่วมกัน เพื่อการคัดกรองผู้ป่วยโรคไตเรื้อรัง เนื่องจากวิธีแรนดอมฟอเรสต์ เป็นวิธีการที่ใช้ต้นไม้หลายๆ ต้นในการตัดสินใจส่งผลทำให้การตัดสินใจมีประสิทธิภาพมากที่สุดในการจำแนกกลุ่มซึ่งสอดคล้องกับงานวิจัย [11, 12] ได้ให้ข้อสังเกตเกี่ยวกับการจำแนกผู้ป่วยไว้ ดังนั้นการพัฒนาแบบจำลองและใช้วิธีแรนดอมฟอเรสต์เพื่อใช้ในการทำนายและสนับสนุนการวินิจฉัยโรคไตเรื้อรังอย่างมีประสิทธิภาพต่อไป

6. กิตติกรรมประกาศ

การศึกษาครั้งนี้สำเร็จสมบูรณ์ได้ด้วยความกรุณาและความช่วยเหลืออย่างดียิ่ง จากโรงพยาบาลส่งเสริมตำบลวังดารา อำเภอบ้านดุง จังหวัดอุดรธานี และคุณศราวุฒิ บุญญะรัง ตำแหน่งพยาบาลวิชาชีพชำนาญการ ที่ได้ให้ข้อมูลและคำแนะนำ ช่วยเหลือและตรวจแก้ไขข้อบกพร่องต่าง ๆ ผู้วิจัยขอกราบขอบพระคุณท่านเป็นอย่างสูงไว้ ณ โอกาสนี้

เอกสารอ้างอิง

[1] วราทิพย์ แก่นการ เกษม ดำนอก และศิริรัตน์ อนุตระกูลชัย. (2562). ประสิทธิผลของรูปแบบการดูแลผู้ป่วยโรคไตเรื้อรังโดยใช้การจัดการโรคเชิงบูรณาการและการจัดการรายกรณีในหน่วยบริการปฐมภูมิ ภาคตะวันออกเฉียงเหนือ. **วารสารการพยาบาลและการดูแลสุขภาพ, 37**(3), 173 – 182.

[2] กวิศรา สอนพูด และลัฆวี ปิยะบัณฑิตกุล. (2563). การดูแลสุขภาพเพื่อชะลอไตเสื่อมของผู้ป่วยโรคเบาหวานความดันโลหิตสูงและประชาชนในชุมชนแห่งหนึ่ง จังหวัดศรีสะเกษ. **วารสารพยาบาลสงขลานครินทร์, 40**(1), 101-114.

[3] สุปราณี สูงแข็ง และสมพร แวงแก้ว (2560). การวิเคราะห์ปัจจัยที่มีอิทธิพลต่อการเกิดภาวะไตเรื้อรังในผู้ป่วยเบาหวานในจังหวัดอุดรธานี. **สำนักงานป้องกันควบคุมโรคที่ 7 ขอนแก่น**, **24**(2) , 1-9.

[4] Thammasudjarit, R., Ingsathit, P., Saputro, S.A., Ingsthit, A. and Thakkinstian, A. (2021). Comparison of Machine Learning With Logistic Regression for Prediction of Chronic Kidney Disease in the Thai Adult Population. **Rama Med J, 44**(4), 1-12.

[5] Ngmaldin, S., Mohamed Ahmed, T. and Abdalkarim Abdalfadi Mohammed, T. (2021). Data mining classification algorithms: An overview. **International Journal of Advanced and Applied Sciences, 8**(2), pp.1 - 5.

[6] นพรัตน์ นนท์ศิริ ราตรี มนัสศิลา และ กริช สมกันธา. (2565). การจำแนกข้อมูลเพื่อวินิจฉัยความเสี่ยงการเป็นโรคเบาหวานโดยใช้เทคนิคเหมืองข้อมูล. **วารสารวิชาการพระจอมเกล้าพระนครเหนือ**, **33**(2), 1-10

[7] ปิยวรรณ นิลถนอม และ ธนพร มาลัย. (2564). การเปรียบเทียบประสิทธิภาพการทำนายผลการแปลงข้อมูลในการจำแนกด้วยเทคนิคการทำเหมืองข้อมูล**. วารสารวิชาการพระจอมเกล้าพระนครเหนือ**, **10**(1), 14- 25.

[8] นพรัตน์ นนท์ศิริ พิศณุ ชัยจิตวณิชกุล และ กริช สมกันธา. (2565). การจำแนกข้อมูลเพื่อวินิจฉัยความเสี่ยงการเป็นโรคเบาหวานโดยโช้เทคนิควิธีแบบร่วมกันตัดสินใจวิธีเลือกคุณลักษณะเด่นไปข้างหน้า. **วารสารวิทยาศาสตร์และเทคโนโลยี มหาวิทยาลัยราชภัฏอุดรธานี, 10**(2), 107-122.

[9] ภาภรณ์ เหล่าพิลัย และ จรัญ แสนราช. (2562). การวิเคราะห์การลาออกกลางคันของนักศึกษาระดับปริญญาตรีโดยใช้เทคนิควิธีการทำเหมืองข้อมูล. **วารสารวิทยาศาสตร์ แห่งมหาวิทยาลัยราชภัฏเพชรบุรี, 16**(2), 61 - 72.

[10] ปิยะนันท์ คงไพ่ จิรพันธุ์ ศรีสมพันธุ์ และ จรัญ แสนราช. การวิเคราะห์รูปแบบความสนใจเลือกอาชีพของผู้เรียนการแบบทดสอบตามทฤษฎีของจอห์น ฮอลแลนด์ ด้วยเทคนิคเหมืองข้อมูล. **วารสารวิชาการพระจอมเกล้าพระนครเหนือ, 9**(1), 80 -90.

[11] สุรวัชร ศรีเปารยะ และสายชล สินสมบูรณ์ทอง. (2560). การเปรียบเทียบประสิทธิภาพวิธีการจำแนกกลุ่มการเป็นโรคไตเรื้อรัง : กรณีศึกษาโรงพยาบาลแห่งหนึ่งในประเทศอินเดีย. **วารสารวิทยาศาสตร์และเทคโนโลยี, 25**(5), 839 – 853.

[12] อัครพล พิกุลศรี และนิภาพร ชนะมาร. (2566). การเปรียบเทียบประสิทธิภาพวิธีการจำแนกประเภทข้อมูลความเสี่ยงเป็นโรคไตด้วยเทคนิคการทำเหมืองข้อมูล. **วารสารวิทยาศาสตร์ วิศวกรรมศาสตร์ และเทคโนโลยี มหาวิทยาลัยราชภัฏเลย, 3**(1), 839 – 853.