วารสารมหาวิทยาลัยราชภัฏร้อยเอ็ด: วิทยาศาสตร์และเทคโนโลยี
Journal of Roi Et Rajabhat University: Science and Technology

การเปรียบเทียบประสิทธิภาพการจำแนกประเภทข้อมูลความเสี่ยงการเป็นโรคไต
Efficiency Comparison of Classification Methods for Kidney Disease

นิพาดา ธงสันเทียะ, ก้องเกียรติ สุวรรณการ, ธีระศักดิ์ ยอดดี, *ณัฐธิดา บุตรพรม
Niphada Thongsuntia, Kongkiat Suwannakan, Teerasak Yoddee, *Nattida Budprom
สาขาวิชาคอมพิวเตอร์และเทคโนโลยีสารสนเทศ คณะเทคโนโลยีสารเทศ มหาวิทยาลัยราชภัฏร้อยเอ็ด
Major of Computer and Information Technology, Faculty of Information Technology,
Roi Et Rajabhat University
*ผู้นิพนธ์หลัก: nattida.cs19@gmail.com
*Corresponding author: nattida.cs19@gmail.com

Received: 16 March., 2024; Revised: 19 May, 2024.; Accepted: 25 May, 2024

Abstract

The objective of this research is to apply data classification mining techniques to predict the risk of kidney disease and to compare their performance to identify the most efficient algorithm. Three techniques are employed: Multilayer Perceptron (Neural Networks), Decision Tree, and Bayesian Learning. The dataset used in this research is the early-stage chronic kidney disease dataset from the UCI repository, containing 400 records from the year 2015. Twenty-five factors are considered for analysis, and the model performance is compared using the 10-Fold Cross Validation method. The Weka software is used as the research tool. The results show that the neural network model is the most suitable, with an accuracy of 98.75%, a precision of 98.80%, a recall of 98.80%, and an F-measure of 99.90%.

Keywords: Kidney disease data analysis; Multilayer perceptron; Naïve Bayes; Decision Tree

บทคัดย่อ

งานวิจัยนี้มีวัตถุประสงค์เพื่อนำเทคนิคการทำเหมืองข้อมูลแบบการจำแนกประเภทข้อมูลมาประยุกต์ใช้ในการทำนายความเสี่ยงการเป็นโรคไต พร้อมทั้งเปรียบเทียบประสิทธิภาพ เพื่อให้ได้อัลกอริทึมที่มีประสิทธิภาพเหมาะสมที่สุด โดยใช้อัลกอริทึม 3 เทคนิค คือ โครงข่ายประสาทเทียม (Multilayer Perceptron) ต้นไม้ตัดสินใจ (Decision Tree) และการเรียนรู้แบบเบย์ (Bayesian

Learning) ชุดข้อมูลที่ใช้ในการวิจัย คือชุดข้อมูลระยะเริ่มต้นของโรคไตเรื้อรัง จากฐานข้อมูล UCI dataset ปี ค.ศ. 2015 มีจำนวนข้อมูล 400 ชุดข้อมูล ปัจจัยสำหรับการวิเคราะห์ 25 ปัจจัย และเปรียบเทียบประสิทธิภาพของแบบจำลองด้วยวิธี 10-Fold Cross Validation โดยเครื่องมือในการวิจัยครั้งนี้ใช้ โปรแกรม Weka ผลวิจัยพบว่าชุดข้อมูลแบบโครงข่ายประสาทเทียมเป็นชุดข้อมูลที่เหมาะสมที่สุด โดยมีค่าความถูกต้อง 98.75% ค่าความแม่นยำ 98.80% ค่าความระลึก 98.80% และค่าความถ่วงดุล 99.90%

คำสำคัญ: การวิเคราะห์ข้อมูลโรคไต; โครงข่ายประสาทเทียม; การเรียนรู้แบบเบย์; ต้นไม้ตัดสินใจ

บทนำ

โรคไตเรื้อรัง (Chronic Kidney Disease) เป็นปัญหาสาธารณสุขที่ทุกประเทศเผชิญและกำลังมีแนวโน้มเพิ่มมากขึ้นของปัจจัยเสี่ยงต่างๆ เช่น การเพิ่มความชุกโรคอ้วน โรคเบาหวาน โรคความดันโลหิตสูง และพฤติกรรมเสี่ยงที่ทำให้ไตเสื่อม เช่น การใช้ยา NSAIDs การได้รับสารเคมีโลหะหนัก หรือมีพฤติกรรมรับประทานอาหารเสี่ยง จากการคาดประมาณภายในปี ค.ศ. 2030 จะมีผู้ป่วยไตระยะสุดท้ายที่ต้องการบำบัดทดแทนไตร้อยละ 70 อยู่ในประเทศกำลังพัฒนา สำหรับประเทศไทยพบอุบัติการณ์ผู้ป่วยโรคไตเรื้อรังต้องการการบำบัดทดแทนไตเฉลี่ยปีละ 20,000 ราย และจากรายงานผู้ป่วยจากโรงพยาบาลในสังกัดกระทรวงสาธารณสุข พ.ศ. 2564 (Kamolthip Vijitsoonthornkul, 2022) มีผู้ป่วยโรคไตเรื้อรัง ระยะที่ 1-5 ทั้งหมด 1,007,251 ราย ซึ่งเป็นภาระในการจัดบริการสุขภาพและมีค่าใช้จ่ายในการรักษาจำนวนมาก

จากปัญหาดังกล่าว ผู้วิจัยได้เห็นถึงความสำคัญของปัญหาและได้ดำเนินการทำการวิจัยทดสอบข้อมูลความเสี่ยงการเป็นโรคไตด้วยเทคนิคเหมืองข้อมูล (Data Mining) (Nongyao Nai-arun, 2021) ซึ่งเป็นการวิเคราะห์ข้อมูลจากข้อมูลจำนวนมากเพื่อหาความสัมพันธ์ของข้อมูล โดยทำการจำแนกประเภท รูปแบบ และความสัมพันธ์ที่ซ่อนอยู่ในชุดข้อมูลนั้น ด้วยการใช้หลักคณิตศาสตร์และสถิติ การรู้จำรูปแบบ (Pattern Recognition) การเรียนรู้ของเครื่อง (Machine Learning) และฐานข้อมูล (Database) เพื่อให้ได้องค์ความรู้ใหม่ โดยการนำข้อมูลที่มีอยู่มาวิเคราะห์แล้วดึงความรู้ในส่วนที่สำคัญออกมาวิเคราะห์หรือทำนายสิ่งต่าง ๆ ที่เกิดขึ้น และสามารถนำไปใช้ประกอบการตัดสินใจในด้านต่าง ๆ ได้ เทคนิคการทำเหมืองข้อมูลมีหลายวิธีการ คือ การเตรียมข้อมูล (Data Preprocessing), เทคนิคการจำแนก (Classification), การวิเคราะห์การจัดกลุ่ม (Cluster Analysis), การวิเคราะห์ความสัมพันธ์ (Association Analysis), การพยากรณ์ (Prediction) ซึ่งผู้วิจัยได้เลือก 2 วิธีการในการทำวิจัยในครั้งนี้ คือ 1) การเตรียมข้อมูล (Data Preprocessing) ซึ่งเป็นขั้นตอนแรกก่อนการทำเหมืองข้อมูลที่จะสามารถปรับปรุงคุณภาพโดยรวมของรูปแบบข้อมูล เนื่องจากฐานข้อมูล UCI dataset ปี ค.ศ. 2015 (Rubini, et al., 2015) ที่ได้มานั้นมีข้อมูลขาดหายไปบางส่วน หากไม่ดำเนินการเตรียมข้อมูลก่อน เมื่อเข้าสู่กระบวนการวิเคราะห์ผลลัพธ์อาจส่งผลให้ผลการวิเคราะห์มีความคลาดเคลื่อนได้ และ 2) เทคนิคการจำแนก (Classification) เพื่อใช้การเปรียบเทียบการแบ่งกลุ่มข้อมูล เพื่อหากลุ่มที่เหมาะสมที่สุดจากอัลกอริทึม 3 คือ โครงข่ายประสาทเทียม (Multilayer Perceptron) การเรียนรู้แบบเบย์

(Bayesian Learning) (Friedman, et al., 1997) และต้นไม้ตัดสินใจ (Decision Tree) (Zorman, et al., 2002) โดยใช้ชุดข้อมูลระยะเริ่มต้นของโรคไตเรื้อรัง จากฐานข้อมูล UCI dataset ปี ค.ศ. 2015 (Rubini, et al., 2015) ซึ่งมีจำนวนข้อมูล 400 ชุดข้อมูล และปัจจัยสำหรับการวิเคราะห์ 25 ปัจจัย

วิธีดำเนินการวิจัย

งานวิจัยนี้มีวัตถุประสงค์เพื่อศึกษากระบวนการเตรียมข้อมูลก่อนการประมวลผล (Data Preprocessing) ก่อนการนำข้อมูลไปประมวลผลเกี่ยวกับการทำนายความเสี่ยงการเป็นโรคไต โดยมีกระบวนการเริ่มต้นที่การรวบรวมข้อมูล (Data collection) จนไปถึงการคัดเลือกปัจจัย (Feature Selection) มีขั้นตอนในการดำเนินการดังปรากฏใน Figure 1

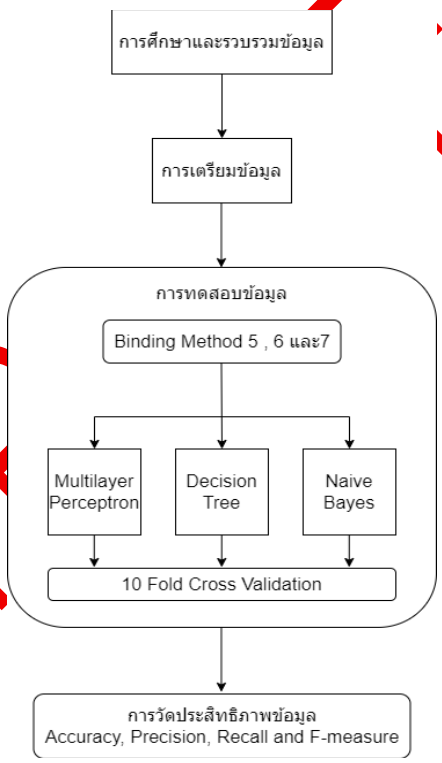

Figure 1 Research methodology

1. การศึกษาและรวบรวมข้อมูล การทดสอบข้อมูลเพื่อทำนายความเสี่ยงการเป็นโรคไตนี้ผู้วิจัยได้ใช้ชุดข้อมูลจากฐานข้อมูล UCI dataset ชุดข้อมูล Chronic Kidney Disease (Rubini, et al., 2015) โดยเป็นชุดข้อมูลที่นำมาทำนายนั้นมีความน่าเชื่อถือและได้ผ่านการวิเคราะห์อย่างถี่ถ้วนแล้ว ซึ่งชุดข้อมูลประกอบไปด้วย 25 ปัจจัย มีข้อมูล 400 รายการ โดยมีรายละเอียดดัง Table 1

Table 1. Data set details

Number	Name	Description	Example
1	age	Age	48, 62, 51, ...
2	bp	Blood pressure	80, 70, 90, ...
3	sg	Specific gravity	1.005,1.010,1.015,1.020,1.025
4	al	Albumin protein value	0,1,2,3,4,5
5	su	Sugar level	0,1,2,3,4,5
6	rbc	Red blood cells	normal, abnormal
7	pc	Pus cells	normal, abnormal
8	pcc	Lump of pus cells	present, notpresent
9	ba	Bacteria	present, notpresent
10	bgr	Random blood sugar values	117, 106, 423, ...
11	bu	Blood urea value	18, 26, 53, ...
12	sc	Serum creatinine value	1.2, 1.8, 10.8, ...
13	sod	Sodium value	131, 138, 141, ...
14	pot	Potassium value	2.5, 4.2, 5.8, ...
15	hemo	The color of red blood cells	15.4, 10.8, 5.6, ...
16	pcv	The percentage of red blood cells in the total blood volume.	44, 29, 16, ...
17	wbcc	Number of white blood cells	7800, 6700, 4500, ...
18	rbcc	Number of red blood cells	5, 3.9, 2.6, ...
19	htn	High blood pressure	yes, no
20	dm	Diabetes	yes, no
21	cad	Coronary heart disease	yes, no
22	appet	Appetite	yes, no
23	pe	Swelling of the feet	yes, no
24	ane	Anemia	yes, no
25	class	Have kidney disease, do not have kidney disease	ckd, notckd

2. การเตรียมข้อมูล (Data Preprocessing) เป็นการเตรียมข้อมูลให้พร้อมก่อนนำไปใช้วิเคราะห์ข้อมูล ในงานวิจัยนี้ใช้ Weka (Weka 3: Machine Learning Software, 2023) ในการเตรียมข้อมูล เนื่องจากชุดข้อมูลที่ได้จากฐานข้อมูล UCI dataset (Rubini, et al., 2015) จาก Table 1 ทั้งหมด 25

ปัจจัย มีค่าข้อมูลสูญหาย (Missing Value) จำนวน 21 ปัจจัย ดังนั้นผู้วิจัยจึงได้ทำการเตรียมข้อมูลด้วย 2 วิธี ดังนี้

2.1 การทำความสะอาดข้อมูล (Data Cleaning) ด้วยเทคนิคการ Replace Missing Values ซึ่งเป็นวิธีเติมข้อมูลแทนค่า Null, NA, NaN ก่อนป้อนโมเดลเทรน Machine Learning โดยการแทนที่ด้วยค่าเฉลี่ย (Mean) ของข้อมูลทั้ง Dataset ของค่าปัจจัยที่เป็นตัวเลข เช่น age bp เป็นต้น ส่วนค่าข้อมูลที่เป็นตัวอักษร เช่น rbc ba เป็นต้น ใช้เทคนิคเติมด้วยค่าจากอัลกอริทึม K-Nearest Neighbour (K-NN) ซึ่งเป็นการใช้หลักการเปรียบเทียบข้อมูลที่สนใจกับข้อมูลอื่นว่ามีความคล้ายคลึงมากน้อยเพียงใด หากข้อมูลที่กำลังสนใจนั้นอยู่ใกล้ข้อมูลใดมากที่สุดก็จะเลือกให้ข้อมูลนั้นอยู่กลุ่มใกล้เคียงกัน ดัง Table 2 และ วิธีการใน Fig 2

Table 2. Data set: Data Cleaning use Replace Missing Values

Number	Name	Missing Values	Replace Data
1	age	2%	51.483376
2	bp	3%	76.469072
3	sg	12%	1.02
4	al	12%	0
5	su	12%	0
6	rbc	38%	normal
7	pc	16%	normal
8	pcc	1%	notpresent
9	ba	1%	notpresent
10	bgr	11%	148.036517
11	bu	5%	57.425722
12	sc	4%	3.072454
13	sod	22%	137.528754
14	pot	22%	4.627244
15	hemo	13%	12.526437
16	pcv	18%	38.884498
17	wbcc	27%	8406.122449
18	rbcc	33%	4.707435
19	htn	1%	no
20	dm	1%	no
21	cad	1%	no

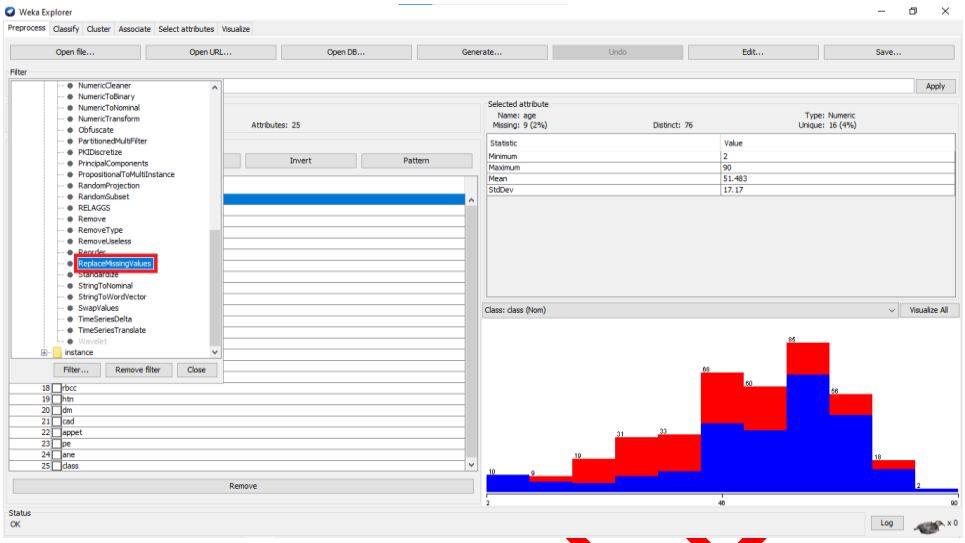

Figure 2 Data Preprocess using techniques: Replace Missing Values

2.2. การแบ่งข้อมูลออกเป็นกลุ่มๆ ซึ่งทำ Binning Method หรือที่เรียกว่าการแยกข้อมูล (Data Discretization) (Surajit Nuddamongkol, 1999) หรือ Bucketing คือการแปลงค่าของข้อมูลให้มีรายละเอียดต่ำลง โดยแบ่งค่า ของข้อมูลออกเป็นช่วงย่อย ๆ ซึ่งเป็นการช่วยลดการประมวลผลในการทำ Data Mining ได้อย่างมาก ผู้วิจัยได้ทำ Binning Method 3 กลุ่ม คือ 5, 6 และ 7 ตามลำดับ ดัง Figure 3

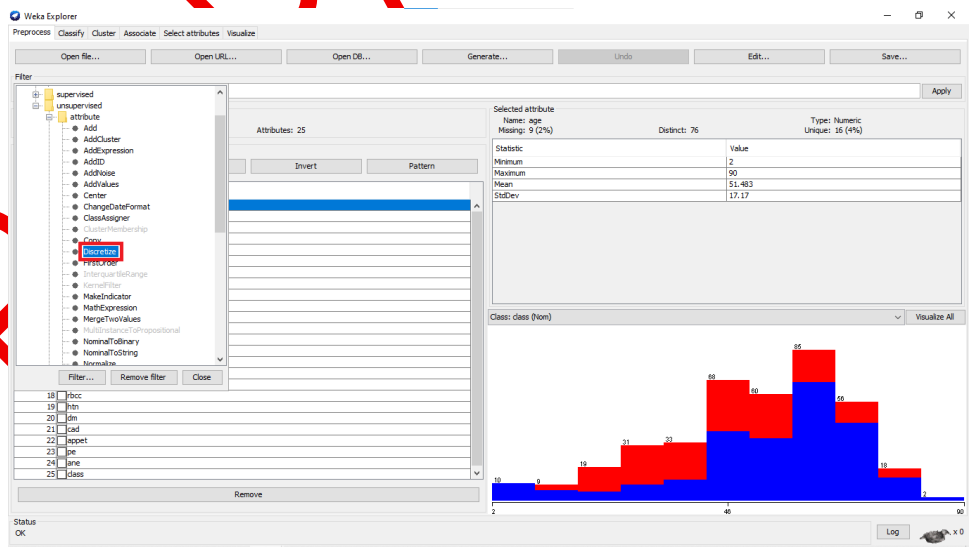

Figure 3 Data Preprocess using techniques: Discretize

3. การทดสอบข้อมูล ในการสร้างชุดข้อมูล ผู้วิจัยได้ใช้ Weka ในการทำเหมืองข้อมูล โดยใช้ประเภทการจำแนกประเภทข้อมูล (Classification) ในการทดสอบเพื่อจำแนกประเภทประเภทข้อมูล

ความเสี่ยงการเป็นโรคไต โดยนำข้อมูลที่ได้มาหลังจากการเตรียมข้อมูล นำไปวิเคราะห์ด้วยวิธีโครงข่ายประสาทเทียม ต้นไม้ตัดสินใจ และการเรียนรู้แบบเบย์ ตามลำดับ และทำการทดสอบประสิทธิภาพชุดข้อมูลด้วย 10 Fold Cross Validation ในทุกๆ กลุ่มชุดข้อมูล โดยฟังก์ชัน Cross Validation

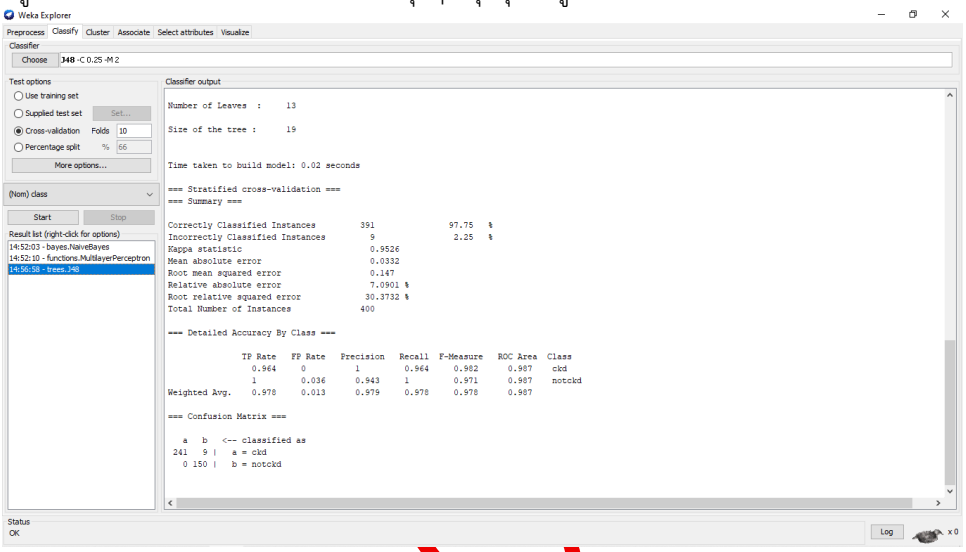

Figure 4 Data test by function: Cross Validation

4. การวัดประสิทธิภาพชุดข้อมูล ได้ใช้ Confusion Matrix คือการประเมินผลลัพธ์การทำนาย (หรือผลลัพธ์จากโปรแกรม) เปรียบเทียบกับผลลัพธ์จริง ประกอบด้วยค่า True Positive (TP) คือ ข้อมูลเป็นจริง และผลการทำนายบอกว่าจริง, ค่า True Negative (TN) คือ ข้อมูลไม่จริง และผลการทำนายบอกว่าไม่จริง, ค่า False Positive (FP) คือ ข้อมูลจริง แต่ผลการทำนายบอกว่าไม่จริง และค่า False Negative (FN) คือ ข้อมูลไม่จริง แต่ผลการทำนายบอกว่าจริง การวัดประสิทธิภาพชุดข้อมูลสามารถประเมินได้จากสมการ

1) ค่าความถูกต้อง (Accuracy) ในการจำแนกข้อมูล ซึ่งเป็นการสรุปผลลัพธ์การจำแนกข้อมูลของข้อมูลได้ถูกต้องและไม่ถูกต้อง สามารถนำมาคำนวณเพื่อวัดประสิทธิภาพ

$$Accuracy = \frac{TP+TN}{TP+TN+FP+FN}$$

2) ค่าความแม่นยำ (Precision) คำนวณจากค่าของข้อมูลที่มีการจำแนกถูกต้องโดยพิจารณาจากผลรวมของการจำแนกข้อมูลทั้งหมด

$$precision = \frac{TP}{TP+FP}$$

3) ค่าความระลึก (Sensitivity or Recall) คำนวณจากค่าของข้อมูลที่ผลลัพธ์ถูกต้องโดยพิจารณาจากข้อมูลของผลลัพธ์เดียวกัน

$$sensitivity = \frac{TP}{TP+FN}$$

4) ค่าความถ่วงดุล (F-measure) คำนวณได้จากค่าเฉลี่ยระหว่างค่าความแม่นยำและค่าความระลึก

$$F - Measure = \frac{Precision\ x\ Recall}{Precision + Recall}$$

ผลการวิจัยและอภิปรายผล

ผลการทดลองของงานวิจัยประกอบด้วย 4 ส่วน ได้แก่ 1) ผลการทดสอบชุดข้อมูลด้วยโครงข่ายประสาทเทียม 2) ผลการทดสอบชุดข้อมูลด้วยต้นไม้ตัดสินใจ 3) ผลการทดสอบชุดข้อมูลด้วยการเรียนรู้แบบเบย์ และ 4) ผลการเปรียบเทียบประสิทธิภาพชุดข้อมูล ดังนี้

1. ผลการทดสอบชุดข้อมูลด้วยโครงข่ายประสาทเทียม การทดสอบโดยใช้ชุดข้อมูลจากเทคนิคโครงข่ายประสาทเทียมแบบ Multilayer Perceptron จาก Weka โดยทำ Binning Method เป็น 5, 6 และ7 พบว่าชุดข้อมูลที่ดีที่สุด คือการ Binning Method 5 ได้ค่าความถูกต้อง 98.75% ค่าความแม่นยำ 98.80% ค่าความระลึก 98.80% ค่าความถ่วงดุล 99.90% ดัง Table 3

Table 3. Results from Artificial Neural Network techniques (Multilayer Perceptron)

N Binning Method	Accuracy (%)	Precision (%)	Recall (%)	F-measure (%)
5	98.75	98.80	98.80	99.90
6	97.75	97.80	97.80	97.80
7	97.75	97.80	97.80	97.80

2. ผลการทดสอบชุดข้อมูลด้วยต้นไม้ตัดสินใจ การทดสอบโดยใช้ชุดข้อมูลจากเทคนิคต้นไม้ตัดสินใจ จาก Weka โดยทำ Binning Method เป็น 5, 6 และ7 พบว่าชุดข้อมูลที่ดีที่สุด คือการ Binding Method 5 ได้ค่าความถูกต้อง 97.5% ค่าความแม่นยำ 97.90% ค่าความระลึก 97.80% ค่าความถ่วงดุล 97.80% ดัง Table 4

Table 4. Results from the Decision Tree technique

N Binding Method	Accuracy (%)	Precision (%)	Recall (%)	F-measure (%)
5	97.75	97.90	97.80	97.80
6	95.75	95.80	95.80	98.50
7	97.00	97.10	97.00	97.00

3. ผลการทดสอบชุดข้อมูลด้วยการเรียนรู้แบบเบย์ การทดสอบโดยใช้ชุดข้อมูลจากการเรียนรู้แบบเบย์ จาก Weka โดยทำ Binning Method เป็น 5, 6 และ7 พบว่าชุดข้อมูลที่ดีที่สุด คือการ Binning Method 7 ได้ค่าความถูกต้อง 98.25% ค่าความแม่นยำ 98.30% ค่าความระลึก 98.30% ค่าความถ่วงดุล 98.30% ดัง Table 5

Table 5. Results from Bayesian Learning techniques (Naïve Bayes)

N Binning Method	Accuracy (%)	Precision (%)	Recall (%)	F-measure (%)
5	96.50	96.80	96.50	96.50
6	97.50	97.50	97.50	97.50
7	98.25	98.30	98.30	98.30

4. ผลการเปรียบเทียบประสิทธิภาพชุดข้อมูล จากการเปรียบเทียบประสิทธิภาพการทำงานของทั้งสามเทคนิคได้ผลการทดลองดัง Table 6 และ Figure 5 พบว่าชุดข้อมูลแบบโครงข่ายประสาทเทียมมีค่าความถูกต้อง 98.75% ค่าความแม่นยำ 98.80% ค่าความระลึก 98.80% และค่าความถ่วงดุล 99.90% ชุดข้อมูลจากเทคนิคต้นไม้ตัดสินใจ มีค่าความถูกต้อง 97.50% ค่าความแม่นยำ 97.90% ค่าความระลึก 97.80% และค่าความถ่วงดุล 97.80% และชุดข้อมูลจากเทคนิคการเรียนรู้แบบเบย์ มีค่าความถูกต้อง 98.25% ค่าความแม่นยำ 98.30% ค่าความระลึก 98.30% และค่าความถ่วงดุล 98.30% ซึ่งจะเห็นได้ว่าชุดข้อมูลแบบโครงข่ายประสาทเทียมมีค่าประสิทธิภาพดีกว่าชุดข้อมูลจากเทคนิคต้นไม้ตัดสินใจ และเทคนิคการเรียนรู้แบบเบย์ โดยมีค่าความถูกต้อง 98.75% ค่าความแม่นยำ 98.80% ค่าความระลึก 98.80% และค่าความถ่วงดุล 99.90% ดังนั้นสรุปได้ว่าชุดข้อมูลแบบโครงข่ายประสาทเทียมเป็นชุดข้อมูลที่เหมาะสมที่สุดในการจำแนกประเภทข้อมูล

Table 6. Data performance comparison

Models	Accuracy (%)	Precision (%)	Recall (%)
Multilayer Perceptron	98.75	98.80	98.80
Decision Tree	97.50	97.90	97.80
Naïve Bayes	98.25	98.30	98.30

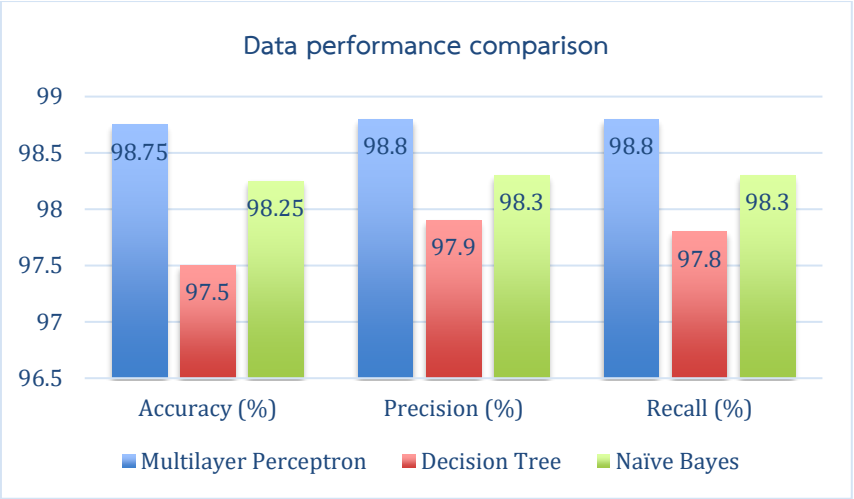

Figure 5 Data performance comparison

สรุปผลการวิจัย

งานวิจัยนี้เป็นการเปรียบเทียบประสิทธิภาพการจำแนกประเภทข้อมูลความเสี่ยงการเป็นโรคไตด้วยเทคนิคเหมืองข้อมูล ซึ่งมีวัตถุประสงค์เพื่อนำเทคนิคเหมืองข้อมูลแบบการจำแนกประเภทมาประยุกต์ใช้ในการทำนายความเสี่ยงการเป็นโรคไต พร้อมทั้งเปรียบเทียบประสิทธิภาพ เพื่อให้ได้อัลกอลิทึมที่มีประสิทธิภาพเหมาะสมด้วยอัลกอลิทึมเหมืองข้อมูล ทั้งสามแบบ คือโครงข่ายประสาทเทียม การเรียนรู้แบบเบย์ และต้นไม้ตัดสินใจ ซึ่งชุดข้อมูลที่ใช้ในครั้งนี้ คือ ชุดข้อมูลระยะเริ่มต้นของโรคไตเรื้อรังจากฐานข้อมูล UCI dataset ปี ค.ศ. 2015 (Rubini, et al., 2015) ซึ่งมีจำนวนข้อมูล 400 ชุดข้อมูล และปัจจัยสำหรับการวิเคราะห์ 25 ปัจจัย คือ อายุ ความดันโลหิต ค่าความถ่วงจำเพาะ ค่าโปรตีนอัลบูมิน ระดับน้ำตาล เซลล์เม็ดเลือดแดง เซลล์หนอง ก้อนเซลล์หนอง แบคทีเรีย ค่าน้ำตาลในเลือดแบบสุ่ม ค่ายูเรียในเลือด ค่าซีรั่มครีเอตินีน ค่าโซเดียม ค่าโพแทสเซียม ค่าสารสีของเม็ดเลือดแดง ร้อยละของเม็ดเลือดแดงต่อปริมาณเลือดทั้งหมด จำนวนเซลล์เม็ดเลือดขาว จำนวนเซลล์เม็ดเลือดแดง โรคความดันโลหิตสูง โรคเบาหวาน โรคหลอดเลือดหัวใจ ความอยากอาหาร อาการบวมเท้า และโรคโลหิตจาง การแบ่งชุดข้อมูลสำหรับใช้ในการวิเคราะห์ข้อมูล โดยการแยกข้อมูล (Data Discretization) (Surajit Ruddamongkol, 1999) เป็น 5 และทดสอบประสิทธิภาพด้วยวิธีการ 10-Fold Cross Validation ผลการวิจัยพบว่า ตัวชุดข้อมูลการจำแนกประเภทข้อมูลด้วย โครงข่ายประสาทเทียมมีประสิทธิภาพดีที่สุด โดยมีค่าความถูกต้อง 98.75% ค่าความแม่นยำ 98.80% ค่าความระลึก 98.80% และค่าความถ่วงดุล 99.90% ดังนั้นจึงสรุปได้ว่าเทคนิค Multilayer Perceptron มีความเหมาะสมในการนำมาสร้างชุดข้อมูลการจำแนกประเภทข้อมูลความเสี่ยงการเป็นโรคไต ซึ่งแตกต่างกับการวิจัยของ (อัครพล พิกุลศรี และ นิภาพร ชนะมาร, 2556) ที่ทำวิจัยการเปรียบเทียบประสิทธิภาพการจำแนกประเภทข้อมูลความเสี่ยงการเป็นโรคไตด้วยเทคนิคการทำเหมืองข้อมูล จากผลการทดลอง ผลการวิจัยพบว่าตัวแบบจำลองการจำแนกประเภทข้อมูลด้วยต้นไม้ตัดสินใจมีประสิทธิภาพดีที่สุดโดยมีค่าความถูกต้อง 98.47% ค่าความแม่นยำ 98.33% ค่าความระลึก 99.16% และค่าความถ่วงดุล 98.73% และได้กฎใน

การประกอบการตัดสินใจทั้งหมด 5 กฎ ดังนั้นจึงสรุปได้ว่าเทคนิค Decision Tree มีความเหมาะสมในการนำมาสร้างแบบจำลองการจำแนกประเภทข้อมูลความเสี่ยงการเป็นโรคไต ถึงแม้จะมีการใช้ UCI dataset แบบเดียวกัน แต่มีงานวิจัยดังกล่าวได้ใช้ปัจจัยสำหรับการวิเคราะห์ 24 ปัจจัย แต่ของผู้วิจัยได้ใช้ทั้ง 25 ปัจจัยและได้ดำเนินการเตรียมข้อมูล (Data Preprocessing) รวมไปถึงการทำ Binning Method หรือที่เรียกว่าการแยกข้อมูล (Data Discretization) (Surajit Nuddamongkol, 1999) ซึ่งอาจจะสรุปได้ว่าการปรับเปลี่ยนเทคนิคและจำนวนปัจจัยนำเข้าโครงข่ายประสาทเทียมมีผลต่อประสิทธิภาพของโครงข่ายประสาทเทียม ซึ่งมีการกระจายตัวของข้อมูลที่ดีขึ้นจะส่งเสริมให้สามารถพัฒนาโครงข่ายประสาทเทียมได้มีความแม่นยำในการทำนายได้มากขึ้น

ข้อเสนอแนะ

การศึกษาครั้งนี้เป็นการเปรียบเทียบประสิทธิภาพการจำแนกประเภทข้อมูลความเสี่ยงการเป็นโรคไต เพื่อค้นหาเทคนิคการจำแนกประเภทข้อมูลความเสี่ยงการเป็นโรคไตที่เหมาะสมกับข้อมูลปัจจัยที่จัดเก็บแบบ UCI dataset ทั้ง 25 ปัจจัย ดังนั้นในการศึกษาครั้งต่อไปสามารถนำข้อมูลผู้ป่วยจากโรงพยาบาลที่มีปัจจัยทั้ง 25 ตามแบบ UCI dataset เข้ามาทดสอบด้วยอัลกอริทึมโครงข่ายประสาทเทียม ได้พร้อมทั้งสามารถเพิ่มจำนวน และทางสถานพยาบาลสามารถนำเทคนิคโครงข่ายประสาทเทียมไปใช้เพื่อทำนายความเสี่ยงการเป็นโรคไตได้ เพื่อเพิ่มประสิทธิภาพสูงสุดในการดูแลรักษาผู้ป่วย

กิตติกรรมประกาศ

ขอบคุณแหล่งข้อมูลเปิด Machine Learning Repository จาก Center for Machine Learning and Intelligent Systems, Bren School of Information and Computer Science University of California, Irvine (UCI) สำหรับข้อมูลโรคไตที่นำมาใช้ในการวิจัยครั้งนี้

เอกสารอ้างอิง

Akkarapon, P., & Nipaporn, C. (2023). Efficiency Comparison of Classification Methods for Kidney Disease with Data Mining Techniques. *Journal of Science Engineering and Technology LOEI Rajabhat University*, 3(1), 1-17. https://ph02.tci-thaijo.org/index.php/JSET/article/view/247493/168624

Friedman, N., Geiger, D., & Goldszmidt, M. (1997). Bayesian network classifiers. *Machine learning*, 29, 131-163. https://link.springer.com/content/pdf/10.1023/A:1007465528199.pdf

Kamolthip, V. (2022). Epidemiology and review of measures to prevent chronic kidney disease. Division of Non Communicable Diseases. https://ddc.moph.go.th/uploads/publish/1308820220905025852.pdf

Nongyao, N. (2021). Performance Comparison of Cardiovascular Risk Prediction Models using Data Mining Algorithms. *Journal of Science and Technology*, 40(2), 137-

147.
https://scjmsu.msu.ac.th/pdfsplit.php?p=MTYyMjE3NzA5My5wZGZ8MTctMjc=

Rubini, L., Soundarapandian, P., & Eswaran, P. (2015). Chronic Kidney Disease. *UCI Machine Learning Repository*. https://doi.org/10.24432/C5G020

Surajit, N. (1999). Knowledge discovery in database on kohonen self-organizing map algorithm [Unpublished master dissertation, Mahidol University]. http://www.thaithesis.org/detail.php?id=44074#

Weka 3: Machine Learning Software in Java. (2023). Retrieved from https://www.cs.waikato.ac.nz/ml/weka/

Zorman, M., Masuda, G., Kokol, P., Yamanoto, R., & Stiglic, B. (2002). Mining Diabetes Database with Decision Trees and Association Rules. In Computer Based Medical Systems, 2002. (CBMS 2002). *Proceedings of the 15th IEEE Symposium*, 134-139. https://sci-hub.se/https://doi.org/10.1109/CBMS.2002.1011367