การเปรียบเทียบประสิทธิภาพการจำแนกข้อมูลผู้ป่วยโรคไตเรื้อรังด้วยเทคนิคเหมืองข้อมูล

ชินวัฒน์ ภูไชยแสง[1*] , ไก้รุ่ง เฮงพระพรหม[1] และ สุพจน์ เฮงพระพรหม[1]

[1] สาขาวิชาวิทยาการข้อมูล คณะวิทยาศาสตร์และเทคโนโลยี มหาวิทยาลัยราชภัฏนครปฐม
*644285001@webmail.npru.ac.th

บทคัดย่อ

งานวิจัยนี้มีวัตถุประสงค์เพื่อ 1) ศึกษาวิธีการจำแนกข้อมูลสำหรับผู้ป่วยโรคไตเรื้อรังด้วยเทคนิคเหมืองข้อมูล 2) เปรียบเทียบประสิทธิภาพในการจำแนกข้อมูล 2 วิธีการ ได้แก่ เทคนิคเพื่อนบ้านที่ใกล้ที่สุด และเทคนิคต้นไม้ตัดสินใจ ผลการทดลองพบว่า เทคนิคต้นไม้ตัดสินใจ ให้ประสิทธิภาพที่ดีที่สุด โดยให้ค่าความถูกต้อง 93% ค่าความแม่นยำ 92% และค่าระลึก 92% ส่วนเทคนิคเพื่อนบ้านที่ใกล้ที่สุด ให้ค่าความถูกต้อง 65% ค่าความแม่นยำ 58% และค่าความระลึก 44% ตามลำดับ

คำสำคัญ: ด้วยเทคนิคเพื่อนบ้านที่ใกล้ที่สุด เทคนิคต้นไม้ตัดสินใจ เหมืองข้อมูล โรคไตเรื้อรัง การจำแนกข้อมูล

A Comparison of Data Classification Efficiency in Chronic Kidney Disease Patients Dataset Using Data Mining Techniques

Chinawat Puchaisang[1*], Kairung Hengpraprohm[1], and Supojn Hengpraprohm[1]

[1] Program in data science, Faculty of Science and Technology, Nakhon Pathom Rajabhat University
*644285001@webmail.npru.ac.th

Abstract

The objectives of this research are: 1) to study the data classification for patients with chronic kidney disease using data mining techniques, and 2) to compare the efficiency of two data classifications, namely the k nearest neighbor and decision tree. The experimental results show that the decision tree technique provides the best performance with 93% accuracy, 92% precision, and 92% recall while the k nearest neighbor gives 65% accuracy, 58% precision, and 44% recall respectively.

Keywords: K-Nearest Neighbor, Decision Tree, Data Mining, Chronic Kidney Disease, Data Classification

1. บทนำ

การทำนายโรคไตเรื้อรัง (CKD) มีความสำคัญอย่างยิ่งด้วยเหตุผลหลายประการ ประการแรก : CKD เป็นโรคที่แพร่หลายและมักเป็นภัยเงียบที่ส่งผลกระทบต่อผู้คนนับล้านทั่วโลก การตรวจหาและการจัดการ CKD ในระยะเริ่มต้นสามารถป้องกันหรือชะลอการลุกลามของโรค ปรับปรุงคุณภาพชีวิตของผู้ป่วย และลดภาระในระบบการรักษาพยาบาล ประการที่สอง : โรคไตวายเรื้อรังเป็นปัจจัยเสี่ยงที่สำคัญของโรคหัวใจและหลอดเลือด ซึ่งเป็นสาเหตุการเจ็บป่วยและการเสียชีวิตอันดับต้น ๆ ทั่วโลก การระบุบุคคลที่มีความเสี่ยงสูงต่อโรคไตตั้งแต่เนิ่นๆ สามารถนำไปสู่การรักษาในระยะแรกและป้องกันการพัฒนาของโรคหัวใจและหลอดเลือด ประการที่สาม : การทำนาย CKD สามารถช่วยในการพัฒนาแผนการรักษาเฉพาะบุคคลได้ ผู้ป่วยโรคไตวายเรื้อรังที่แตกต่างกันอาจต้องการการรักษาที่แตกต่างกัน และแผนการรักษาเฉพาะบุคคลสามารถนำไปสู่ผลลัพธ์ที่ดีขึ้นสำหรับผู้ป่วย สุดท้าย : การทำนาย CKD สามารถช่วยในการแจ้งนโยบายและกลยุทธ์ด้านสาธารณสุข ด้วยการทำความเข้าใจปัจจัยเสี่ยงของโรคไตวายเรื้อรัง นโยบายสาธารณสุขสามารถพัฒนาเพื่อกำหนดเป้าหมายประชากรที่มีความเสี่ยงสูง ซึ่งจะนำไปสู่ผลลัพธ์ที่ดีขึ้นสำหรับบุคคลและสังคมโดยรวม

โรคไตเรื้อรัง เป็นภาวะที่ส่งผลต่อไตและอาจทำให้เกิดความเสียหายในระยะยาวหากปล่อยไว้โดยไม่รักษา ปัญหาของการทำนาย CKD มีดังนี้ ให้ชุดข้อมูลที่ประกอบด้วยข้อมูลของผู้ป่วย เช่น อายุ เพศ ความดันโลหิต ซีรั่มครีเอตินิน อัลบูมิน และระดับกลูโคส สร้างโมเดลแมชชีนเลิร์นนิงที่สามารถทำนายได้อย่างแม่นยำว่าผู้ป่วยเป็นโรคไตวายเรื้อรังหรือไม่ แบบจำลองควรสามารถที่จะจำแนกผู้ป่วยว่าเป็นโรค CKD หรือไม่ ด้วย ค่าความแม่นยำ, Precision, และ recall. ปัญหานี้เป็นเรื่องที่ซับซ้อนเนื่องจากระยะแรกของโรคไตวายเรื้อรังอาจไม่แสดงอาการที่ชัดเจน ดังนั้นจึงจำเป็นอย่างยิ่งที่จะต้องพัฒนาแบบจำลองการพยากรณ์ที่แม่นยำและเชื่อถือได้ ซึ่งสามารถระบุผู้ป่วยที่มีความเสี่ยงในการเกิดโรคไตวายเรื้อรังได้ นอกจากนี้ CKD ยังเกิดได้จากหลายปัจจัย เช่น เบาหวาน ความดันโลหิตสูง ความบกพร่องทางพันธุกรรม เป็นต้น ดังนั้น แบบจำลอง ต้องสามารถจะรองรับข้อมูลประชากรผู้ป่วยที่หลากหลายและปรับให้เป็นแบบทั่วไปในสถานพยาบาลต่างๆ

แนวทางการแก้ไขปัญหา สามารถใช้อัลกอริธึมการเรียนรู้ของเครื่องต่างๆ เช่น k – nearest neighbor, decision trees ชุดข้อมูลควรได้รับการประมวลผลล่วงหน้าเพื่อจัดการกับค่าที่ขาดหายไป ค่าผิดปกติ และปัญหาด้านคุณภาพข้อมูลอื่นๆ สามารถใช้เทคนิคการเลือกคุณลักษณะเพื่อระบุคุณลักษณะที่เกี่ยวข้องมากที่สุดสำหรับการทำนาย CKD ประสิทธิภาพของแบบจำลองสามารถประเมินได้โดยใช้เมตริกต่างๆ เช่น ค่าความแม่นยำ, precision, recall, and F1-score โมเดลสามารถปรับแต่งเพิ่มเติมได้โดยการปรับไฮเปอร์พารามิเตอร์โดยใช้เทคนิคต่างๆ เช่น การค้นหาแบบกริดหรือการค้นหาแบบสุ่ม เมื่อโมเดลได้รับการปรับให้เหมาะสมแล้ว ก็สามารถนำไปใช้เพื่อทำนาย CKD ในผู้ป่วยรายใหม่ได้ โดยรวมแล้ว การพัฒนาโมเดลแมชชีนเลิร์นนิงที่แม่นยำและเชื่อถือได้ สำหรับการทำนายโรคไตวายเรื้อรังมีศักยภาพช่วยให้แพทย์นำข้อมูลไปปรับในการรักษาให้เหมาะสมกับผู้ป่วยและลดค่าใช้จ่ายด้านการรักษาพยาบาลที่เกี่ยวข้องกับการรักษาโรคไตวายเรื้อรัง

2. วัตถุประสงค์ของการวิจัย

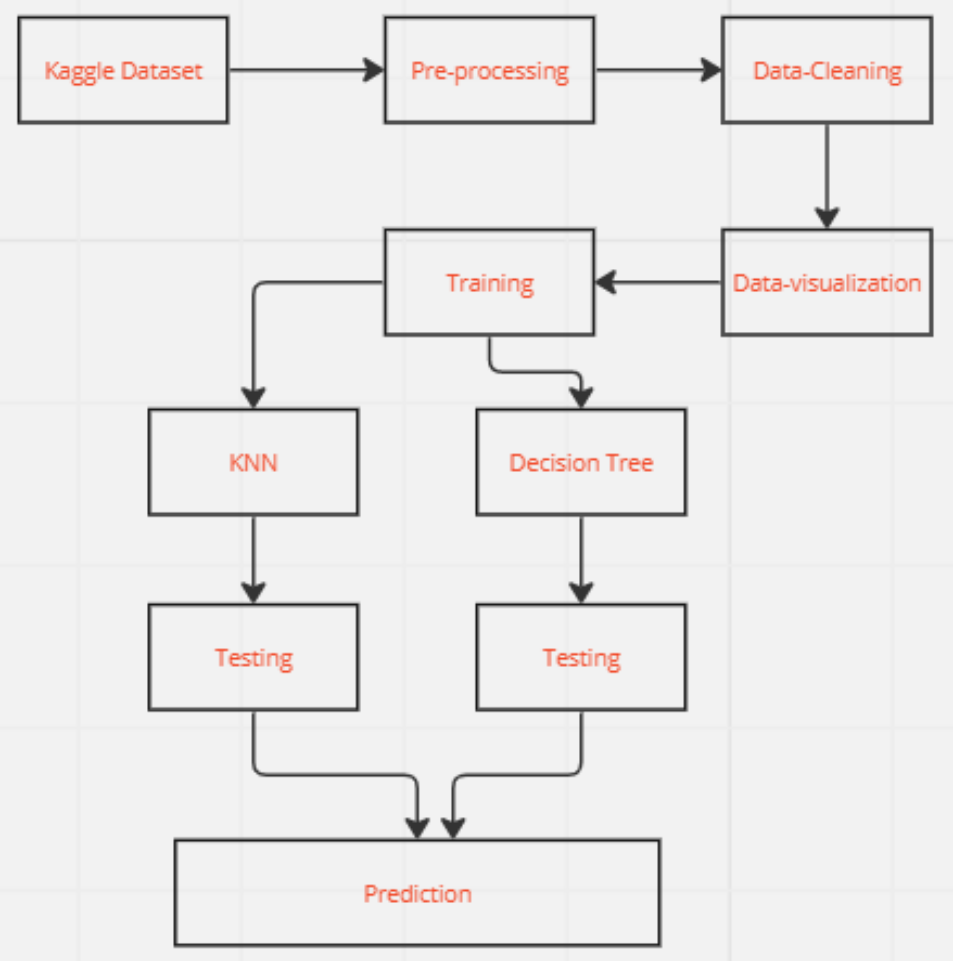
2.1 เพื่อศึกษาวิธีการจำแนกข้อมูลด้วยเทคนิคเหมืองข้อมูล

2.2 เพื่อเปรียบเทียบประสิทธิภาพในการจำแนกข้อมูล ได้แก่ เทคนิคเพื่อนบ้านที่ใกล้ที่สุด และเทคนิคต้นไม้ตัดสินใจ

3. วิธีการดำเนินการวิจัย

3.1 กรอบแนวคิดในการวิจัย

การดำเนินการวิจัย ประกอบด้วยกรอบแนวคิดในการวิจัยดังนี้



ภาพที่ 1 กรอบแนวคิดในการวิจัย

3.1.1 ทำการศึกษาลักษณะชุดข้อมูล Chronic Kidney Disease

3.1.2 ทำข้อมูลให้สมบูรณ์ (Data Cleaning) ตรวจสอบแก้ไขข้อมูลที่ไม่เกี่ยวข้องออกไป

3.1.3 จำแนกประเภทกลุ่มโรค(Classify) จากนั้นนำข้อมูลที่ผ่านการแปลงข้อมูลนำเข้าสู่วิธีการจำแนกข้อมูล ด้วยการใช้ Data Mining Techniques คือ k – nearest neighbor, Decision Tree เพื่อเปรียบเทียบประสิทธิภาพการจำแนกที่มีค่าความแม่นยำ ด้วยการใช้ซอฟต์แวร์ Google Colaboratory ในการจำแนกดังกล่าว

3.1.4 ประเมินผล(Evaluation) เพื่อนำผลที่ได้จากการนำเข้าWeka มาพิจารณาว่าเทคนิคเหมืองข้อมูลชนิดไหนมีประสิทธิภาพที่เหมาะสมที่สุด

3.1.5 นำแบบจำลองไปใช้งาน (Deployment) นำผลการวิเคราะห์ของแบบจำลองที่ทำการเปรียบเทียบประสิทธิภาพเพื่อได้แบบจำลองเทคนิคที่มีความน่าเชื่อถือ และสามารถนำผลที่ได้มาใช้ให้เกิดประโยชน์เกี่ยวกับทางการแพทย์ได้ต่อไป

3.2 ข้อมูลสำหรับการวิจัย

ชุดข้อมูล Chronic Kidney Disease ที่นำมาใช้ในงานวิจัยได้นำมาจากแหล่งข้อมูลจากเว็บไซต์ของ Kaggle เป็นชุดข้อมูลเกี่ยวกับผู้ป่วยโรคไตเรื้อรัง ภายในจะมี 26 คอลัมน์ และ 401แถว โดยข้อมูลมีคุณลักษณะประกอบด้วยรายละเอียดคือ รหัส (ID), อายุ (Age), ค่าความดันโลหิต (bp), ค่าถ่วงจำเพราะ (sg), ค่าการทำงานของตับ (al), ระดับน้ำตาลในเลือด (su), เซลล์เม็ดเลือดแดง (rbc), pus cell (pc), pus cell clumps (pcc), bacteria (bc), การตรวจน้ำตาลแบบสุ่ม (bgr), ปริมาณ

ไนโตรเจนในเลือด (bu), serum creatinine (sc), sodium (sod), Potassium (pot), hemoglobin (hemo), เปอร์เซ็นต์ของเม็ดเลือดแดง (pcv), จำนวนเม็ดเลือดขาว (wc), จำนวนเม็ดเลือดแดง (rc), ค่าความดันโลหิตสูง (htn), โรคเบาหวาน (dm), โรคหลอดเลือดหัวใจ (cad), appetite, อาการบวมน้ำ (pe), โลหิตจาง (ane) และการจำแนกกลุ่มข้อมูล (Classification)

ตารางที่ 1 แสดงโครงสร้างและข้อมูลโรคไตวายเรื้อรัง (Chronic Kidney Disease)

A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T	U	V	W	X	Y	Z
id	age	bp	sg	al	su	rbc	pc	pcc	ba	bgr	bu	sc	sod	pot	hemo	pcv	wc	rc	htn	dm	cad	appet	pe	ane	classification
90	63	100	1.01	2	2	normal	normal	notpresent	present	280	35	3.2	143	3.5	13	40	9800	4.2	yes	no	yes	good	no	no	ckd
91	56	70	1.015	4	1	abnormal	normal	notpresent	notpresent	210	26	1.7	136	3.8	16.1	52	12500	5.6	no	no	no	good	no	no	ckd
92	71	70	1.01	3	0	normal	abnormal	present	present	219	82	3.6	133	4.4	10.4	33	5600	3.6	yes	yes	yes	good	no	no	ckd
93	73	100	1.01	3	2	abnormal	abnormal	present	notpresent	295	90	5.6	140	2.9	9.2	30	7000	3.2	yes	yes	yes	poor	no	no	ckd

3.4 การประเมินผลการวิจัย

การวิจัยนี้ใช้วิธีการประเมินประสิทธิภาพด้วยค่าความแม่นยำในการจำแนก (Accuracy) คือ มาตรวัดค่าความแม่นตรง คือ ค่าที่บอกถึงความแม่นยำในการจำแนกข้อมูล จากสมการ

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

โดยที่ TP คือค่า True Positive, TN คือค่า True Negative,
FP คือค่า False Positive, FN คือค่า False Negative

4. ทฤษฎีและงานวิจัยที่เกี่ยวข้อง

4.1 ทฤษฎีที่เกี่ยวข้อง

1. เหมืองข้อมูล เหมืองข้อมูล (data Mining) เป็นเทคนิคในการวิเคราะห์ข้อมูลอย่างหนึ่ง ซึ่งมาจากคำว่า เหมืองข้อมูล นั่นคือ เป็นการค้นหาสิ่งที่มีประโยชน์จากฐานข้อมูลที่มีขนาดใหญ่ เช่น ข้อมูลการซื้อขายสินค้าในซุปเปอร์มาร์เก็ตต่าง ๆ โดยข้อมูลเหล่านี้จะเก็บจากรายการสินค้าที่ลูกค้าซื้อในแต่ละครั้ง โดยเมื่อทำการวิเคราะห์ข้อมูลด้วยเทคนิค Data Mining แล้วจะได้สิ่งที่เป็นประโยชน์

2. วิธีการเพื่อนบ้านใกล้ที่สุด ขั้นตอนวิธีการเพื่อนบ้านใกล้ที่สุด (K-Nearest Neighbor Algorithm : Knn) เป็นวิธีที่ใช้ในการจัดแบ่งคลาส โดยเทคนิคนี้จะตัดสินใจว่า คลาสใดที่จะแทนเงื่อนไขหรือกรณีใหม่ๆ ได้บ้าง โดยการตรวจสอบจำนวนบางจำนวน ในขั้นตอนวิธีการเพื่อนบ้านใกล้ที่สุด ของกรณีหรือเงื่อนไขที่เหมือนกันหรือใกล้เคียงกันมากที่สุด โดยจะหาผลรวม (Count Up) ของจำนวนเงื่อนไข หรือกรณีต่างๆ สำหรับแต่ละคลาส และกำหนดเงื่อนไขใหม่ๆ ให้คลาสที่เหมือนกันกับคลาสที่ใกล้เคียงกันมากที่สุด

3. ต้นไม้การตัดสินใจ ต้นไม้การตัดสินใจ (decision tree) เป็นเครื่องมือที่ช่วยให้วิเคราะห์เหตุการณ์ หรือสถานการณ์เพื่อการตัดสินใจได้อย่างเป็นระบบและรวดเร็ว ต้นไม้การตัดสินใจมีลักษณะเป็นกราฟรูปต้นไม้ ซึ่งแสดงที่ตั้งต้นที่มีรากและแขนงต่างๆแตกออกมาจากต้นไม้ไปในทิศทางเดียว จนกระทั่งนำไปสู่ข้อสรุปสำหรับการตัดสินใจได้ ต้นไม้การตัดสินใจมีประโยชน์ในการสรุปการตัดสินใจที่มีความซับซ้อนให้ง่ายต่อความเข้าใจ ปัจจุบันต้นไม้การตัดสินใจเป็นที่นิยมใช้ในงานหลายอย่าง เช่น การแพทย์ ธุรกิจ การเขียนโปรแกรม การสร้างเครื่องที่เรียนรู้ได้เอง การสร้างระบบผู้เชี่ยวชาญ ฯลฯ

4.2 งานวิจัยที่เกี่ยวข้อง

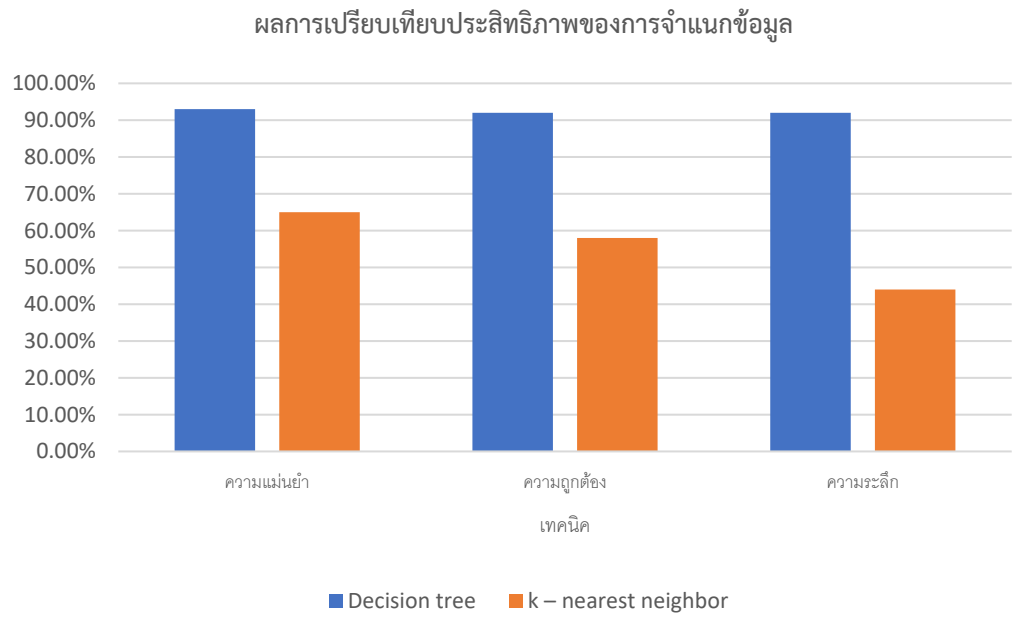
การวิจัยครั้งนี้มีเป้าหมายเพื่อช่วยให้การวินิจฉัยและการรักษาโรคไตเรื้อรังมีความแม่นยำและเหมาะสมมากขึ้น จากการศึกษางานวิจัยที่เกี่ยวข้องกับการทำนายโรคไตเรื้อรังด้วย Machine Learning ประกอบด้วยแนวทางในการศึกษาหลากหลายดังนี้ 1) งานวิจัยนี้ได้ใช้ Machine Learning [2] ในการวิเคราะห์ข้อมูลสำหรับการทำนายโรคไตเรื้อรัง โดยอาจใช้ข้อมูลเชิงพันธุกรรม ข้อมูลการตรวจทางแล็ป ข้อมูลอาการและประวัติโรคของผู้ป่วย เป็นต้น 2) การสร้างโมเดล Machine Learning [3] เพื่อทำนายความเสี่ยงในการเป็นโรคไตเรื้อรัง โดยอาจใช้ข้อมูลเชิงพันธุกรรม ข้อมูลประวัติการรักษาของผู้ป่วย และตัวชี้วัดสุขภาพ เช่น ค่าความดันโลหิต ระดับน้ำตาลในเลือด เป็นต้น 3) การสร้างโมเดล Machine Learning [4] เพื่อทำนายผลของการรักษาโรคไตเรื้อรัง โดยอาจใช้ข้อมูลการรักษาที่ผ่านมาของผู้ป่วย เช่น การใช้ยา การผ่าตัด และการบำบัดทางกายภาพ เป็นต้น 4) งานวิจัยนี้มีวัตถุประสงค์ [5] เพื่อเปรียบเทียบประสิทธิภาพของวิธีการจำแนกกลุ่ม โดยเลือกใช้วิธีความใกล้เคียงกันมากที่สุด วิธีต้นไม้ตัดสินใจ วิธีโครงข่ายประสาทเทียม วิธีซัพพอร์ตเวกเตอร์แมชชีน วิธีฐานกฎ วิธีการถดถอยลอจิสติก และวิธีนาอีฟเบย์ สำหรับวัดประสิทธิภาพการจำแนกกลุ่ม โดยใช้ข้อมูลผู้ป่วยโรคไตเรื้อรังของโรงพยาบาลอพอลโลประเทศอินเดีย โดยแบ่งข้อมูลเป็นชุดสร้างตัวแบบ และชุดทดสอบตัวแบบ ในอัตราส่วน 70 และ 30 ตามลำดับ จากการเปรียบเทียบประสิทธิภาพวิธีการจำแนกกลุ่มผู้ป่วยโรคไตเรื้อรัง วิธีการจำแนกกลุ่มที่มีประสิทธิภาพการจำแนกดีที่สุดคือวิธีต้นไม้ตัดสินใจ ซึ่งให้ค่าความถูกต้อง คือ 100 % และค่าความคลาดเคลื่อนกำลังสองเฉลี่ย คือ 0.0059

5. ผลการวิจัย

ผลการแปลงชุดข้อมูลและ จำแนกประเภทข้อมูลจากชุดข้อมูลการแปลงแต่ละเทคนิค ผ่านเทคนิคการจำแนกข้อมูลโดยแบ่งข้อมูลในการเรียนรู้ (Training Data) และข้อมูลในการทดสอบ (Testing Data) ตามซอฟต์แวร์ Google Collaboratory ได้แก่ ค่าความแม่นยำ (Accuracy) ค่าความถูกต้อง (precision) และค่าความระลึก (recall) แสดงผลในรูปแบบตารางที่ 2 และภาพที่ 2-3

ตารางที่ 2 เปรียบเทียบแสดงค่าความแม่นยำ ความถูกต้อง และค่าความระลึกในการจำแนกประเภทจากเทคนิคกับชุดข้อมูลการแปลงข้อมูล

เทคนิค	ความแม่นยำ	ความถูกต้อง	ความระลึก
Decision tree	93.00%	92.00%	92.00%
k – nearest neighbor	65.00%	58.00%	44.00%



ภาพที่ 2 ผลการเปรียบเทียบแสดงค่าความแม่นยำ ความถูกต้อง และค่าความระลึกในการจำแนกประเภทจากเทคนิคกับชุดข้อมูลการแปลงข้อมูล

จากตารางที่ 2 และภาพที่ 2 ผลการเปรียบเทียบแสดงค่าความแม่นยำในการจำแนกประเภทจากเทคนิคกับชุดข้อมูลการแปลงข้อมูล ผลการจำแนกข้อมูล เทคนิค Decision tree ให้ประสิทธิภาพที่ดีที่สุด โดยให้ผลความแม่นยำ (accuracy) 93% ค่าความถูกต้อง (precision) 92% และค่าความระลึก (recall) 92%

6. สรุปผล

ผลการวิจัยนี้ได้ทำการทรานฟอร์มข้อมูลด้วยวิธีการใช้ซอฟต์แวร์ Google Collaboratory เพื่อทำชุดข้อมูลให้เป็นปกติและทำการเปรียบเทียบการจำแนกกลุ่มจากชุดข้อมูลที่แปลงให้เป็นปกติทั้ง 2 วิธีการด้วยเทคนิคการจำแนกกลุ่มจำนวน 2 เทคนิควิธี คือ Decision tree และ k – nearest neighbor จากชุดข้อมูลการทรานฟอร์ม ซึ่งผลการประเมินประสิทธิภาพตัวแบบ คือ เทคนิค Decision tree ให้ประสิทธิภาพที่ดีที่สุด โดยให้ผลความแม่นยำ (accuracy) 93% ค่าความถูกต้อง (precision) 92% และค่าความระลึก (recall) 92% จึงสรุปได้ว่า Decision tree เป็นตัวแบบที่เหมาะสมที่สุดที่จะนำมาใช้จำแนกข้อมูลผู้ป่วยโรคไตวายเรื้อรัง

7. เอกสารอ้างอิง

[1] Kaggle. (2021). Chronic Kidney Disease dataset. Retrieved March 21, 2023 from https://www.kaggle.com/datasets/mansoordaku/ckdisease.

[2] Kanmanee Sungkhapan, Piyawan Kaewthon, Phacharee Kwankapo, Wiparat Yokphuang, Tum Boonrod, Wichada Simala, and Sirirat Sriraksa. (2020). The duration of chronic kidney disease development amongst type 2 diabetes patients: A systematic review and meta-analysis. **J Med Health Sci**. Vol.27, No.3, P.83-99. (In Thai).

[3] Uraiwan Pantong. (2018). Chronic kidney Disease Management with Chronic Care Model at Primary Care in Nakhon si Thammarat Province. **Maharaj Nakhon Si Thammarat Medical Journal.** Retrieved 2 January 2018 from http://www.mecnst.com/NSTMJ/file_content/202102102050444589.pdf . (In Thai).

[4] Akkarapol Pikulsri and Nipaporn Chanamarn. (2022). Efficiency Comparison of Classification Methods for Kidney Disease with Data Mining Techniques. **Journal of Science, Engineering and Technology Loei Rajabhat University.** Retrieved October 1, 2022 from https://ph02.tci-thaijo.org/index.php./JSET/article/view/247493 . (In Thai).

[5] Surawat Sripaoraya and Saichon Sinsomboonthong.(2017). Efficiency Comparison of Data Mining Classification Methods for Chronic Kidney Disease: A Case Study of a Hospital in India. Journal of Science and Technology. Vol.25, No.5. P.839-852. (In Thai)