ปีที่ 3 ฉบับที่ 1 (2023) : มกราคม – มิถุนายน 2566
JSET LRU
วารสารวิทยาศาสตร์ วิศวกรรมศาสตร์ และเทคโนโลยี
มหาวิทยาลัยราชภัฏเลย
Journal of Science, Engineering and Technology
LOEI RAJABHAT UNIVERSITY

การเปรียบเทียบประสิทธิภาพการจำแนกประเภทข้อมูลความเสี่ยงการเป็นโรคไตด้วยเทคนิคการทำเหมืองข้อมูล

Efficiency Comparison of Classification Methods for Kidney Disease with Data Mining Techniques

อัครพล พิกุลศรี[1] และ นิภาพร ชนะมาร[2*]

Akkarapon Phikulsri[1] and Nipaporn Chanamarn[2*]

[1,2]สาขาวิชาคอมพิวเตอร์ คณะวิทยาศาสตร์และเทคโนโลยี มหาวิทยาลัยราชภัฏสกลนคร 47000

[1,2]Department of Computer, Faculty of Science and technology, Sakon Nakhon Rajabhat University, Sakon Nakhon Province, 47000

[*]Corresponding author E-mail: nipaporn@snru.ac.th

Received 1 October 2022, Accepted 27 December 2022, Published 1 March 2023

บทคัดย่อ

งานวิจัยนี้มีวัตถุประสงค์เพื่อนำเทคนิคการเหมืองข้อมูลแบบการจำแนกประเภทข้อมูลมาประยุกต์ใช้ในการทำนายความเสี่ยงการเป็นโรคไต พร้อมทั้งเปรียบเทียบประสิทธิภาพ เพื่อให้ได้อัลกอริทึมที่มีประสิทธิภาพเหมาะสมที่สุด โดยใช้อัลกอริทึม 3 เทคนิค คือ โครงข่ายประสาทเทียม ต้นไม้ตัดสินใจ และการเรียนรู้แบบเบย์ ชุดข้อมูลที่ใช้ในการวิจัยคือ ชุดข้อมูลระยะเริ่มต้นของโรคไตเรื้อรัง จากฐานข้อมูล UCI Machine Learning มีจำนวนข้อมูล 400 ชุดข้อมูล ปัจจัยสำหรับการวิเคราะห์ 24 ปัจจัย และเปรียบเทียบประสิทธิภาพของแบบจำลองด้วยวิธี 10-Fold Cross Validation โดยเครื่องมือในการวิจัยครั้งนี้ใช้ Scikit-Learn Library ของภาษา Python ผลการวิจัยพบว่า การจำแนกประเภทข้อมูลโดยใช้แบบจำลองจากเทคนิคต้นไม้ตัดสินใจ มีประสิทธิภาพเหมาะสมที่สุด ที่ค่าความถูกต้อง 98.47% ค่าความแม่นยำ 98.33% ค่าความระลึก 99.16% และค่าความถ่วงดุล 98.73% และให้กฎการตัดสินใจ จำนวน 5 กฎ ที่ควรนำไปพัฒนาเว็บแอปพลิเคชันเพื่อการวินิจฉัยโรคเบื้องต้นเกี่ยวกับความเสี่ยงการเกิดโรคไตได้

คำสำคัญ: โรคไต โครงข่ายประสาทเทียม การเรียนรู้แบบเบย์ ต้นไม้ตัดสินใจ การเรียนรู้ของเครื่องแบบมีผู้สอน

ABSTRACT

The objective of this research is to apply classification techniques in data mining to predict kidney disease risk and compare efficacy. To obtain the most effective algorithm, three algorithms are used: Neural Networks; Decision trees, and Naive Bayes. The dataset used in the research is the early stages of chronic kidney disease in UCI machine learning database contains 400 data sets and 24 factors. The Efficiency comparison of

the model using the 10-Fold Cross Validation method. The tools in this research use the scikit-learn library of python languages. The results showed that the best prediction performance was the decision tree model. It has an accuracy of 98.47%, precision of 98.33%, recall of 99.16%, and F-measure. of 98.73%. This Decision trees model show has 5 rules that should be used to develop web applications for early diagnosis of kidney disease risk.

Keywords: Kidney Disease, Artificial Neural Networks, Naïve Bayes, Decision Tree, Supervised Machine Learning Techniques

บทนำ

โรคไตเรื้อรัง คือภาวะที่ไตทำงานได้น้อยลงผิดปกติ อย่างต่อเนื่องในระยะเวลา หลายเดือนหรือหลายปี (สุพัฒน์, 2554) ซึ่งไตจะมีตัวกรองเล็กๆ ประมาณหนึ่งล้านตัว เรียกว่า เนฟรอน (Nephron) หากเนฟรอนหยุดทำงานลงมากขึ้นเรื่อยๆ หลังจากผ่านไประยะเวลาหนึ่ง ไตก็จะไม่สามารถกรองเลือดได้ดีพอ ส่งผลให้มีของเสียตกค้างในร่างกาย ซึ่งเป็นผลเสียต่อสุขภาพ และอาจทำให้เกิดภาวะแทรกซ้อน เช่น ภาวะไตรกลีเซอไรด์ในเลือดสูง ภาวะกระดูกพรุน และภาวะต่อพาราไทรอยด์ทำงานมากเกินไป เป็นต้น (World Kidney Day, 2022) ตามข้อมูลจาก WKD หรือ World Kidney Day 8% ถึง 10% ของประชากรบนโลก ที่อยู่ในช่วงวัยผู้ใหญ่มีภาวะที่ไตทำงานได้น้อยลงผิดปกติ และทุกๆ ปีมีคนนับล้านเสียชีวิตก่อนเวลาอันควรจากโรคแทรกซ้อนที่เกี่ยวข้องกับโรคไตเรื้อรัง (Chronic Kidney Disease) เนื่องจากโรคไตเป็นโรคเรื้อรังชนิดหนึ่งที่ต้องการการดูแลรักษาที่ยาวนานหลายรายต้องรักษาตลอดชีวิต อีกทั้งค่าใช้จ่ายก็สูง โดยเฉพาะอย่างยิ่งกรณีที่เข้าสู่ภาวะไตวายเรื้อรังระยะสุดท้ายที่ต้องรักษาด้วยวิธีที่เรียกว่าการบำบัดทดแทนไต เช่น การฟอกเลือดด้วยเครื่องไตเทียม และการผ่าตัดเปลี่ยนไตเป็นต้น

จากปัญหาดังกล่าว ผู้วิจัยได้เห็นถึงความสำคัญของปัญหา และได้ดำเนินการทำการวิจัยสร้างแบบจำลอง การจำแนกประเภทข้อมูลความเสี่ยงการเป็นโรคไตด้วยเทคนิคเหมืองข้อมูล โดยมีการเปรียบประสิทธิภาพแบบจำลองเพื่อหาแบบจำลอง ที่เหมาะสมที่สุด จากอัลกอริทึม 3 เทคนิค คือ โครงข่ายประสาทเทียม (Artificial Neural Network) การเรียนรู้แบบเบย์ (Bayesian Learning) และต้นไม้ตัดสินใจ (Decision Tree) โดยใช้ชุดข้อมูลระยะเริ่มต้นของโรคไตเรื้อรังของคนอินเดีย จากฐานข้อมูล UCI ปี 2015 ซึ่งมีจำนวนข้อมูล 400 ชุดข้อมูล และปัจจัยสำหรับการวิเคราะห์ 24 ปัจจัย

เอกสารและงานวิจัยที่เกี่ยวข้อง

1. โรคไต

โรคไตเป็นโรคเกี่ยวกับการทำงานของไตที่ลดน้อยลงผิดปกติ (โรงพยาบาลพระรามเก้า, 2563) เป็นโรคเรื้อรังที่รักษาไม่หายขาด ลักษณะของโรคไต ภาะไตวาย ไตหยุดทำงาน ของเสียจะคั่งค้างในเลือดและร่างกาย ผู้ป่วยจะมีอาการเบื่ออาหาร คลื่นไส้อาเจียน ซีด อ่อนเพลีย โลหิตจาง น้ำท่วมปอด ปวดกระดูก ถ้าของเสียค้างในสมองมากๆ อาจทำให้มีอาการชักและสมองหยุดทำงาน ถ้าเกิดโรคไตวายในเด็ก เด็กจะแคระแกรนหยุดการเจริญเติบโต (สุพัฒน์, 2554) ปัจจัยเสี่ยงในการเป็นโรคไต มีดังนี้ กรรมพันธุ์ โรคไตบางชนิดเป็นกรรมพันธุ์ เช่น โรคไตเป็นถุงน้ำ ความดันโลหิตสูง จะมีผลกระทบต่ออวัยวะที่สำคัญๆ คือหัวใจ หลอดเลือด ไต และสมอง คนที่มีความดันโลหิตสูงนานๆ จะมีผลทำให้ไตเสื่อมลง โรคเบาหวาน ผู้ป่วยโรคเบาหวานเมื่อเวลาผ่านไป 10-15 ปี จะมีการเปลี่ยนแปลงที่ไตโดยเฉพาะหลอดเลือดของไต มีไข่ขาวออกมาในปัสสาวะ ส่งผลทำให้เกิดไตเสื่อม ไตวายเรื้อรัง

และไตวายเรื้อรังระยะสุดท้าย และโรคเบาหวานยังมีผลกระทบทำให้เกิดความดันโลหิตสูงซึ่งมีผลกระทบต่อไตอีกด้วย อายุ ไตคนปกติจะเจริญเต็มที่เมื่ออายุประมาณ 2 ขวบ และจะเริ่มเสื่อมเมื่ออายุ 35 ปี ยาและอาหาร ยาหลายชนิด อาหารบางประเภทมีพิษต่อไต เช่น ยาแก้ข้อกระดูกอักเสบ ประเภท NSAID อาหารรสจัดเป็นต้น (ทวี, สัฆวี และวิริณธิ์, 2563) โรคไตวายเรื้อรังเกิดจากความผิดปกติของระบบต่างๆ ในร่างกาย เช่น ความผิดปกติของระบบเผาผลาญ ระบบหลอดเลือด ระบบทางอิมมูน โรคในระบบเหล่านี้ได้แก่ โรคเบาหวาน โรคความดันโลหิตสูง โรคเก๊าท์ โรคแพ้ภูมิตนเอง โรคนิ่วในไต และระบบท่อทางเดินปัสสาวะ กรวยไตอักเสบเรื้อรัง เป็นต้น ปัจจัยที่ส่งผลต่อการทำหน้าที่ของไตลดลง และส่งผลให้เกิดโรคไตเรื้อรัง ประกอบไปด้วย 3 ปัจจัยหลัก ได้แก่ ปัจจัยส่วนบุคคล เช่น เพศ อายุ ดัชนีมวลกาย เป็นต้น ปัจจัยด้านการเจ็บป่วย ได้แก่ โรคประจำตัวร่วมเช่น โรคเก๊าท์ โรคหัวใจและหลอดเลือด โรคความดันโลหิตสูง เป็นต้น โรคถ่ายทอดทางพันธุกรรม เช่น โรคถุงน้ำในไต กลุ่มอาการอัลพอร์ต เป็นต้น ประวัติโรคไตเรื้อรังในครอบครัว ระยะเวลาในการป่วยด้วยโรคเบาหวาน การควบคุมระดับความดันโลหิต การควบคุมระดับน้ำตาลในเลือด เป็นต้น และในส่วนของปัจจัยด้านพฤติกรรมสุขภาพ เช่น การกลั้นปัสสาวะ การดื่มน้ำน้อย การใช้ยา NSAIDs การบริโภคอาหารหวาน มันและเค็ม การสูบบุหรี่ การขาดกิจกรรมทางกาย เป็นต้น

2. การทำเหมืองข้อมูล

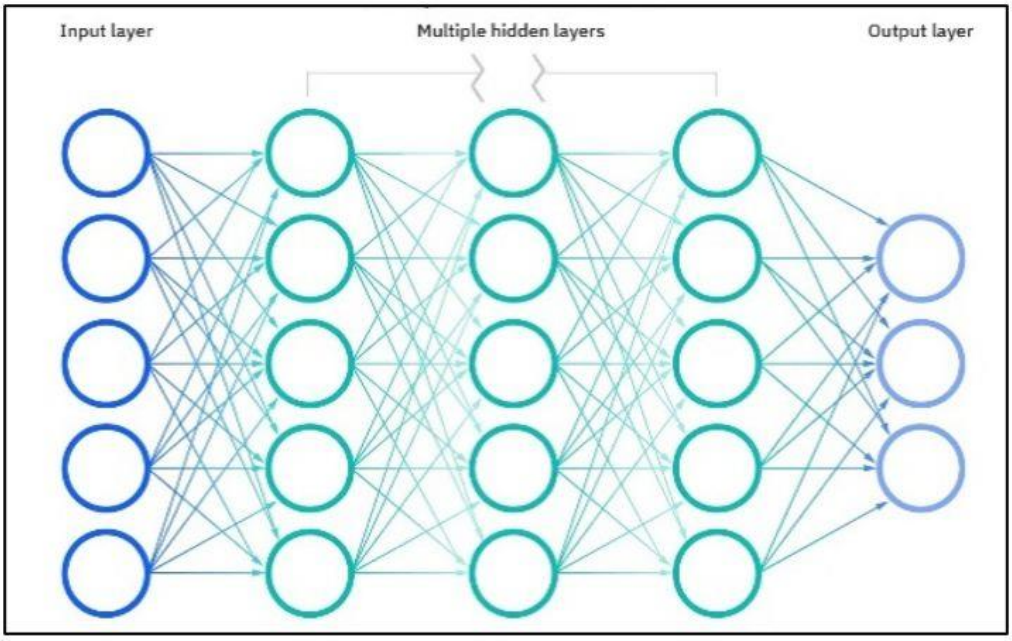
การทำเหมืองข้อมูล (Data Mining) (ณัฐวดี และประภาสิริ, 2562) เป็นการค้นหาองค์ความรู้จากฐานข้อมูลขนาดใหญ่ ซึ่งเป็นกระบวนการที่จะกระทำกับข้อมูลที่มีจำนวนมาก เป็นการค้นหาความเชื่อมโยงที่ซ่อนอยู่และคาดการณ์แนวโน้มที่จะเกิดขึ้น โดยอาศัยหลักสถิติ การเรียนรู้ของเครื่อง (Machine Learning) และการจดจำรูปแบบ (Pattern Recognition) สารทนเทศที่ได้จากกระบวนการทางเหมืองข้อมูลอาจนำมาสร้างแบบจำลองเพื่อพยากรณ์ หรือจำแนกหน่วย หรือแสดงความสัมพันธ์ระหว่างหน่วยต่างๆ อัลกอริทึมการทำเหมืองข้อมูลมีหลายประเภทการทำงาน ในการวิจัยนี้ใช้อัลกอริทึมการจำแนกประเภทข้อมูล (Classification) เป็นเทคนิคการวิเคราะห์จากการเรียนรู้จดจำหรือจากรูปแบบของข้อมูลแบบมีป้ายบอกทาง ซึ่งเป็นการเรียนรู้ของเครื่องแบบมีผู้สอน (Supervised Machine Learning Techniques)

3. การเรียนรู้ของเครื่องแบบมีผู้สอน

การเรียนรู้ของเครื่องแบบมีผู้สอน (Noyunsan, Katanyukul and Saikaew, 2018) เป็นวิธีการเรียนรู้ของเครื่องเพื่อสร้างแบบจำลองการพยากรณ์ โดยมีการกำหนดชุดข้อมูล X หรือเป็นอินพุตให้กับแบบจำลอง พร้อมคลาสคำตอบคือ Y เพื่อให้แบบจำลองใช้ในการเรียนรู้ โดยผลลัพธ์การพยากรณ์ อาจจะเป็นค่าต่อเนื่อง (Continuous Value) กรณีเช่นนี้ เรียกว่าปัญหาการถดถอย (Regression Problem) แต่เมื่อผลลัพธ์เป็นค่าไม่ต่อเนื่อง (Discrete Value) กรณีเช่นนี้ เรียกว่า ปัญหาการจำแนกประเภท (Classification Problem) ซึ่งผลลัพธ์จากการจำแนกประเภทมักเรียกว่าคลาส (Class) หรือ ป้ายกำกับ (Label) ซึ่งค่าพารามิเตอร์ของแบบจำลองแบบจำแนกประเภทจะประกอบไปด้วยข้อมูล X เป็นชุดข้อมูล Feature หรือ Attributes และ y คือ คลาสคำตอบ สำหรับข้อมูลของ Feature นั้นๆ กล่าวคือ มีชุดข้อมูล X และ ชุดข้อมูล y ประกอบกัน มักเรียกว่า ข้อมูลการฝึกฝนแบบมีป้ายกำกับ อัลกอริทึมแบบจำแนกประเภท เช่น ต้นไม้ตัดสินใจ (Decision Tree) ซัพพอร์ตเวกเตอร์แมชชีน (Support Vector Machine) การเรียนรู้แบบเบย์ (Naïve Bayes) และ โครงข่ายประสาทเทียม (Artificial Neural Network) เช่น Multilayer Perceptron ในงานวิจัยนี้เป็นการใช้แบบจำลองการจำแนกประเภท (Classification) 3 อัลกอริทึม ได้แก่ โครงข่ายประสาทเทียม (Artificial Neural Network) ต้นไม้ตัดสินใจ (Decision Tree) และการเรียนรู้แบบเบย์ (Naïve Bayes)

4. เทคนิคโครงข่ายประสาทเทียม

โครงข่ายประสาทเทียม (Artificial Neural Network) (ปริญญา, 2562) นั้นเป็นแบบจำลองทางคณิตศาสตร์ เพื่ออธิบายการทำงานอันซับซ้อนของสมองมนุษย์ จากทฤษฎีการเรียนรู้ของเฮบบ์ (Hebb's Rule) เซลล์ที่ทำงานร่วมกันจะเชื่อมต่อกัน เป็นหลักในการอธิบายการเรียนรู้ด้วยรูปแบบของการประกอบเซลล์ประสาทเข้าด้วยกันเป็นโครงข่าย สามารถนำมาแก้ปัญหาพื้นฐานการเรียนรู้แบบไม่มีผู้สอน (Unsupervised Learning) แต่สำหรับการเรียนรู้แบบมีผู้สอน (Supervised Learning) นั้นจะเป็นขั้นตอนที่เรียกว่า "เพอร์เซปตรอน" (Perceptron) ซึ่งสามารถทำการเรียนรู้จากตัวอย่างที่นำเข้าไปได้ แต่ก็มีข้อจำกัดบางประการ เช่น ปัญหา XOR (Exclusive OR) แต่ปัจจุบันพื้นฐานสำคัญของโครงข่ายประสาทเทียม เรียกว่า โครงข่ายประสาทเทียมแบบแพร่กลับ (Backpropagation) ซึ่งขั้นตอนวิธีนี้สามารถแก้ปัญหา XOR โดยอาศัยโครงข่ายแบบหลายชั้น (Multi-Layer) ซึ่งในงานวิจัยนี้ผู้วิจัยได้ใช้โครงข่ายประสาทเทียมแบบหลายชั้น (Multilayer Perceptron) จาก Python Library Scikit-Learn ในการทดสอบ (วิทยา, 2555) ซึ่งโครงข่ายประสาทเทียมเป็นเทคนิคที่ได้มาจากการศึกษาโครงข่ายไฟฟ้าชีวภาพ (Bioelectric Network) ในสมอง ประกอบด้วย เซลล์ประสาท (Neurons) จุดประสานประสาท (Synapses) แต่ละเซลล์ประสาทประกอบด้วยปลายในการรับกระแสประสาท (Dendrite) ซึ่งเป็น Input และปลายประสาทในการส่งกระแสประสาท (Axon) ซึ่งเป็นเหมือน Output ของเซลล์ ลักษณะของโครงข่ายประสาทเทียม มีโครงสร้างเป็นกลุ่มของโหนด (Node) ที่เชื่อมโยงถึงกันในแต่ละชั้น (Layer) ได้แก่ Input Layer, Hidden Layer, Output Layer ดังภาพที่ 1



ภาพที่ 1 โครงสร้างโครงข่ายประสาทเทียม
ที่มา: (IBM, 2020)

5. เทคนิคการเรียนรู้แบบเบย์

การเรียนรู้แบบเบย์ (Bayesian Learning) (ปริญญา, 2562) เป็นการจำแนกประเภทรูปแบบหนึ่งที่อาศัยหลักการของความน่าจะเป็น (Probability) มาช่วยในการหาคำตอบ ตามกฎของเบย์ (Bayes' Law) เพื่อหาว่าสมมติฐานใดน่าจะถูกต้องที่สุด เช่น

คลาสหรือประเภทของตัวอย่างสมมติฐานว่า ปริมาณที่สนใจจะอยู่ภายใต้การแจกแจงความน่าจะเป็น (Probability Distribution) ดังนั้น ความน่าจะเป็นของตัวอย่างจึงสามารถใช้ในการประกอบการตัดสินใจอย่างมีเหตุผลได้ในงานจำแนกประเภท (Classification) ซึ่งทฤษฎีบทของเบย์ (Bayes' Theorem) หรือกฎของเบย์ (Bayes' Law) ตั้งชื่อตาม โทมัส เบย์ (Thomas Bayes) นักสถิติและนักปราชญ์ชาวอังกฤษ กล่าวถึงเหตุการณ์ในปัจจุบันและสิ่งที่เกิดขึ้นก่อนหน้า โดยมีความน่าจะเป็นแบบมีเงื่อนไข แสดงดังสมการที่ 1 และการใช้กฎของเบย์ในการจำแนกประเภท (Classification) แสดงดังสมการที่ 2

$$P(A|B) = \frac{\mathrm{P(B|A)P(A)}}{\mathrm{P(B)}} \quad (1)$$

$$P(h|D) = \frac{\mathrm{P(D \mid h)P(h)}}{\mathrm{P(D)}} \quad (2)$$

$P(D)$ หมายถึง ความน่าจะเป็นก่อน (Prior Probability) ของเซตตัวอย่างฝึกฝน D

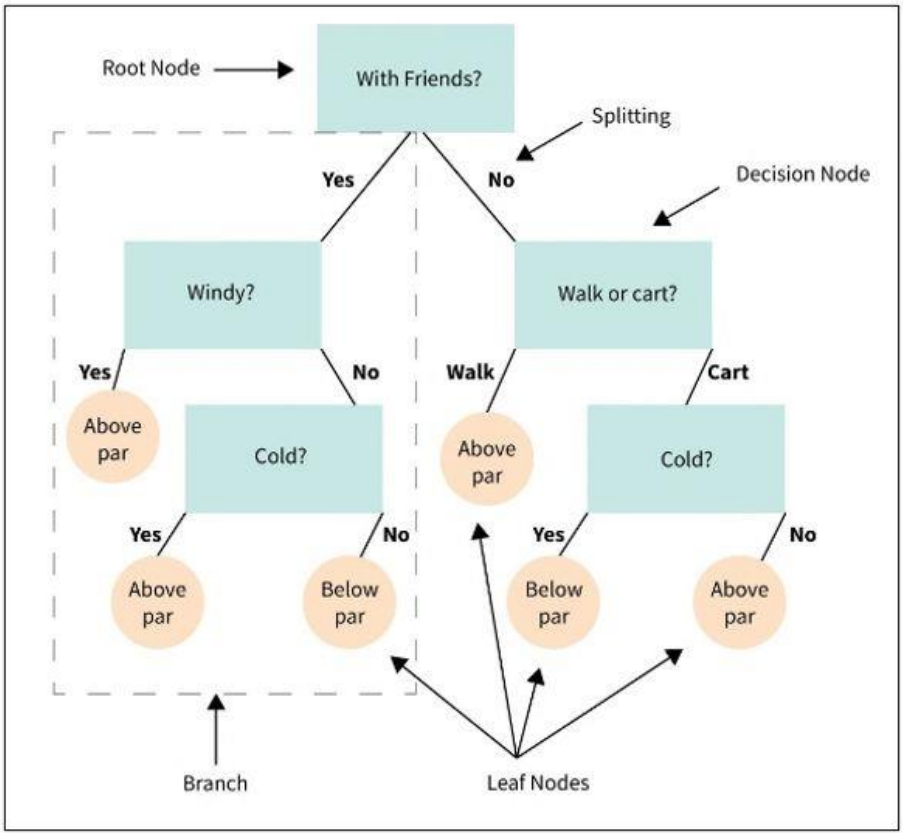
$P(h)$ หมายถึง ความน่าจะเป็นก่อน (Prior Probability) ของสมมติฐาน $\mathrm{h} \in \mathrm{H}$

$P(h|D)$ หมายถึง ความน่าจะเป็นภายหลัง (Posterior Probability) ของสมมติฐาน h เมื่อกำหนดเซตตัวอย่างฝึกฝน D

$P(D|h)$ หมายถึง ความน่าจะเป็นภายหลัง (Posterior Probability) ของเซตตัวอย่างฝึกฝน D เมื่อกำหนดสมมติฐาน h

6. เทคนิคต้นไม้ตัดสินใจ

ต้นไม้ตัดสินใจ (Decision Tree) (MastersInDataScience, 2022) เป็นการนำข้อมูลมาสร้างแบบจำลองการพยากรณ์ในรูปแบบโครงสร้างต้นไม้ โดยต้นไม้ตัดสินใจถือเป็นเครื่องมือหนึ่งที่ช่วยวิเคราะห์เหตุการณ์ หรือสถานการณ์เพื่อตัดสินใจได้อย่างเป็นระบบและรวดเร็ว มีลักษณะเป็นกราฟรูปต้นไม้ แสดงรากแขนงต่างๆ แตกออกมาจากต้นไม้ไปในทิศทางเดียว จนกระทั่งนำไปสู่ข้อสรุปสำหรับการตัดสินใจ ประกอบไปด้วย โหนดราก (Root Node) โหนดลูก (Child Node) หรือโหนดตัดสินใจ (Decision Node) และโหนดใบ (Leaf Node) แสดงดังภาพที่ 2 (ปริญญา, 2562) ต้นไม้ตัดสินใจสามารถนำไปใช้ได้กับปัญหาและข้อมูลหลายประเภท เช่น ค่าแบบไม่ต่อเนื่องสำหรับการจำแนกประเภท และค่าแบบต่อเนื่องสำหรับการวิเคราะห์การถดถอย ทั้งนี้ต้นไม้ตัดสินใจยังสามารถแทนค่าข้อมูลจำนวนมาก หรือแทนกฎแบบถ้า...แล้ว (If-then Rule) เพื่อให้มนุษย์สามารถแปรผลได้ง่ายขึ้น ดังภาพที่ 2



ภาพที่ 2 โครงสร้างต้นไม้ตัดสินใจ

ที่มา: (MastersInDataScience, 2022)

7. งานวิจัยที่เกี่ยวข้อง

ผู้วิจัยได้ทำการค้นคว้างานวิจัยด้านเหมืองข้อมูลที่เกี่ยวข้องวิธีการนำมาใช้ในการวิจัย เพื่อนำมาเป็นข้อมูลในการศึกษาและอ้างอิง ผู้วิจัยได้ศึกษางานวิจัยที่มีความเกี่ยวข้อง ดังต่อไปนี้

ณัฐวดี หงษ์บุญมี และประภาสิริ ตรีพาณิชกุล (ณัฐวดี และประภาสิริ, 2562) ได้ศึกษาการเปรียบเทียบประสิทธิภาพการจำแนกข้อมูลเพื่อวิเคราะห์ปัจจัยความเสี่ยงที่ส่งผลต่อการเกิดโรคไฮเปอร์ไทรอยด์ด้วยเทคนิคเหมืองข้อมูล ผลการเปรียบเทียบพบว่าการจำแนกข้อมูลโดยใช้โครงข่ายประสาทเทียมให้ค่าประสิทธิภาพสูงสุดมีค่าความถูกต้อง 82.97% ซึ่งมากกว่าต้นไม้ตัดสินใจและการเรียนรู้แบบเบย์ที่ให้ค่าความถูกต้อง 79.87% และ 68.11% ตามลำดับ

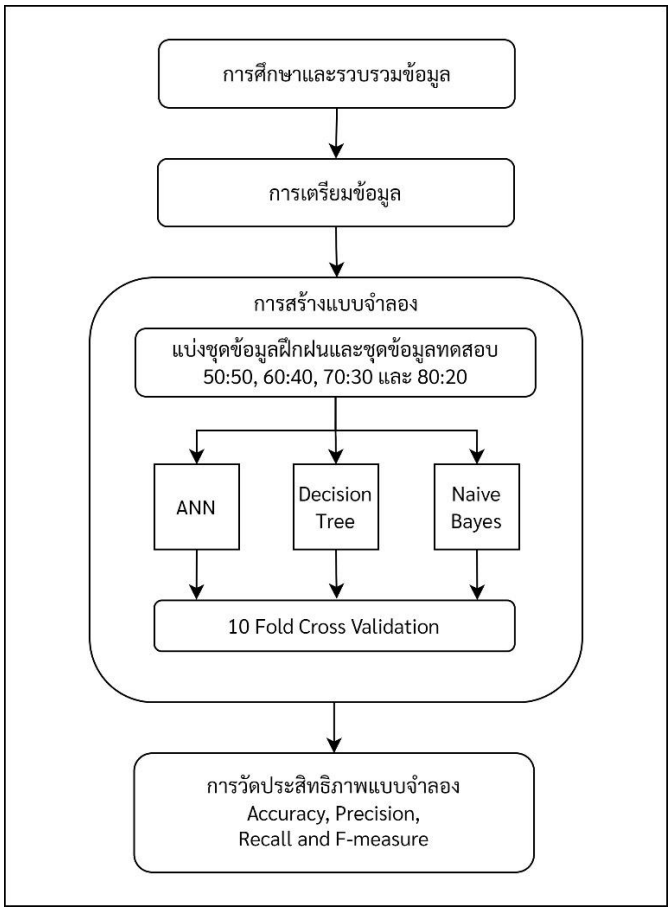
นงเยาว์ ในอรุณ (นงเยาว์, 2564) ได้ทำการเปรียบเทียบประสิทธิภาพของแบบจำลองการทำนายความเสี่ยงโรคหัวใจและหลอดเลือดโดยใช้อัลกอริทึมเหมืองข้อมูลพบว่าแบบจำลองโครงข่ายประสาทเทียมพร้อมการเลือกคุณสมบัติ มีค่าความถูกต้อง 99.29% และต่ำสุดคือ แบบจำลองต้นไม้ตัดสินใจ มีค่าความถูกต้อง 70.39%

อุกฤษฏ์ ศรีสุข (อุกฤษฏ์, 2564) ได้ทำการเปรียบเทียบประสิทธิภาพของเทคนิคเหมืองข้อมูลสำหรับอุบัติการณ์ของผู้ป่วย โดยข้อมูลทั้งหมดถูกรวบรวมมาจากฐานข้อมูล UCI จำนวนทั้งหมด 3 ชุดข้อมูล ในงานวิจัยนี้ได้นำเอาเทคนิคในเหมืองข้อมูล 5 เทคนิค ได้แก่ Decision Tree C4.5, Naive Bayes, Neural Networks, Random Forest และ Deep Learning มาทำการ

สร้างแบบจำลองเพื่อการพยากรณ์การเกิดโรค พบว่าเทคนิค Decision Tree เป็นเทคนิคที่ดีที่สุดในการสร้างแบบจำลองในการพยากรณ์โรคไฮโปไทรอยด์ โดยให้ค่าความถูกต้อง 99.86% ค่าความไว 99.85% และค่าความจำเพาะ 100%

วิธีการวิจัย

การวิจัยครั้งนี้มีขั้นตอนการดำเนินงาน 5 ขั้นตอน ได้แก่ 1) การศึกษาและรวบรวมข้อมูล 2) การเตรียมข้อมูล 3) การสร้างแบบจำลอง 4) การวัดประสิทธิภาพแบบจำลอง ดังภาพที่ 3 และมีรายละเอียดแต่ละขั้นตอนดังต่อไปนี้



ภาพที่ 3 ขั้นตอนการดำเนินการวิจัย

1. การศึกษาและรวบรวมข้อมูล

โรคไตเรื้อรังเป็นภาวะที่ไตทำงานได้น้อยลงผิดปกติ เป็นระยะเวลาหลายเดือนหรือหลายปี เป็นเวลานานเมื่อเซลล์ที่ทำหน้าที่เป็นตัวกรองของเสีย หรือเนฟรอน (Nephron) ที่มีอยู่เพียง 1 ล้านเซลล์หยุดการทำงาน ส่งผลทำให้ไตไม่สามารถกรองของเสียในเลือดได้ ดังนั้นการสร้างแบบจำลองเพื่อจำแนกความเสียงการเป็นโรคไตจะช่วยให้วินิจฉัยโรค ได้ถูกต้องและเป็นประโยชน์ต่อการวางแผนการรักษาผู้ป่วยได้ โดยการศึกษาวิจัยนี้ ผู้วิจัยได้ใช้ชุดข้อมูลจากฐานข้อมูล UCI ชุดข้อมูล Early Stage of Indians Chronic Kidney Disease (CKD) ประกอบไปด้วย 25 แอตทริบิวต์ มีข้อมูล 400 รายการ โดยมีรายละเอียดดังตารางที่ 1

ตารางที่ 1 รายละเอียดชุดข้อมูล

ลำดับ	ชื่อ	รายละเอียด	ลักษณะข้อมูล	ตัวอย่างข้อมูล
1	age	อายุ	numerical	48, 62, 51, ...
2	bp	ความดันโลหิต	numerical	80, 70, 90, ...
3	sg	ค่าความถ่วงจำเพาะ	nominal	1.005,1.010,1.015,1.020,1.025
4	al	ค่าโปรตีนอัลบูมิน	nominal	0,1,2,3,4,5
5	su	ระดับน้ำตาล	nominal	0,1,2,3,4,5
6	rbc	เซลล์เม็ดเลือดแดง	nominal	normal=1, abnormal=0
7	pc	เซลล์หนอง	nominal	normal=1, abnormal=0
8	pcc	ก้อนเซลล์หนอง	nominal	present=1, notpresent=0
9	ba	แบคทีเรีย	nominal	present=1, notpresent=0
10	bgr	ค่าน้ำตาลในเลือดแบบสุ่ม	numerical	117, 106, 423, ...
11	bu	ค่ายูเรียในเลือด	numerical	18, 26, 53, ...
12	sc	ค่าซีรั่มครีเอตินีน	numerical	1.2, 1.8, 10.8, ...
13	sod	ค่าโซเดียม	numerical	131, 138, 141, ...
14	pot	ค่าโพแทสเซียม	numerical	2.5, 4.2, 5.8, ...
15	hemo	ค่าสารสีของเม็ดเลือดแดง	numerical	15.4, 10.8, 5.6, ...
16	pcv	ร้อยละของเม็ดเลือดแดงต่อปริมาณเลือดทั้งหมด	numerical	44, 29, 16, ...
17	wbcc	จำนวนเซลล์เม็ดเลือดขาว	numerical	7800, 6700, 4500, ...
18	rbcc	จำนวนเซลล์เม็ดเลือดแดง	numerical	5, 3.9, 2.6, ...
19	htn	โรคความดันโลหิตสูง	nominal	yes=1, no=0
20	dm	โรคเบาหวาน	nominal	yes=1, no=0
21	cad	โรคหลอดเลือดหัวใจ	nominal	yes=1, no=0
22	appet	ความอยากอาหาร	nominal	yes=1, no=0
23	pe	อาการบวมเท้า	nominal	yes=1, no=0
24	ane	โรคโลหิตจาง	nominal	yes=1, no=0
25	class	เป็นโรคไต, ไม่เป็นโรคไต	nominal	ckd=1, notckd=0

2. การเตรียมข้อมูล

การเตรียมข้อมูล (Data Preparation) เป็นการเตรียมข้อมูลให้พร้อมก่อนนำไปใช้วิเคราะห์ข้อมูล ในงานวิจัยนี้ ใช้ภาษา Python ในการเตรียมข้อมูลด้วย Pandas Library และเครื่องมือที่ใช้คือ Google Colab ในการดำเนินการวิจัย เนื่องจากชุดข้อมูลที่ได้จากฐานข้อมูล UCI มีค่าข้อมูลสูญหาย (Missing Value) ผู้วิจัยได้ทำความสะอาดข้อมูล (Data Cleaning) ด้วยเทคนิค Nearest Value โดยใช้ฟังก์ชันการแทนค่าข้อมูล ด้วยข้อมูลจากแถวก่อนหน้า ใน Pandas Library สำหรับการแปลงข้อมูล (Data Transformation) ดำเนินการแปลงข้อมูลให้อยู่ในรูปแบบที่สามารถนำไปสร้างแบบจำลองได้ โดยใช้ฟังก์ชันจาก Scikit-Learn ดังภาพที่ 4

JSET LRU
วารสารวิทยาศาสตร์ วิศวกรรมศาสตร์ และเทคโนโลยี
มหาวิทยาลัยราชภัฏเลย
Journal of Science, Engineering and Technology
LOEI RAJABHAT UNIVERSITY

```python
def transform_data():
  # transform data
  df_kidney['htn'] = transform_custom(df_kidney, 'htn', 'y_n')
  df_kidney['cad'] = transform_custom(df_kidney, 'cad', 'y_n')
  df_kidney['pe'] = transform_custom(df_kidney, 'pe', 'y_n')
  df_kidney['ane'] = transform_custom(df_kidney, 'ane', 'y_n')
  df_kidney['dm'] = transform_custom(df_kidney, 'dm', 'y_n')
  df_kidney['pc'] = transform_custom(df_kidney, 'pc', 'n_abn')
  df_kidney['rbc'] = transform_custom(df_kidney, 'rbc', 'n_abn')
  df_kidney['pcc'] = transform_custom(df_kidney, 'pcc', 'p_np')
  df_kidney['ba'] = transform_custom(df_kidney, 'ba', 'p_np')
  df_kidney['appet'] = pd.Series(map(lambda x: dict(good=1, poor=0)[x], df_kidney['appet'].values.tolist()), df_kidney.index)
  df_kidney['class'] = pd.Series(map(lambda x: dict(ckd=1, notckd=0)[x], df_kidney['class'].values.tolist()), df_kidney.index


def transform_custom(dataframe, key, option="y_n"):
  """
  option:
    y_n: yes, no
    n_an: normal, abnormal
    p_np: present, notpresent
  """
  if(option == "y_n"):
    print(f"transform ${key} from yes no to 1 0")
    return pd.Series(map(lambda x: dict(yes=1, no=0)[x], dataframe[key].values.tolist()), dataframe.index)
  if(option == "n_abn"):
    print(f"transform ${key} from normal abnormal to 1 0")
    return pd.Series(map(lambda x: dict(normal=1, abnormal=0)[x], dataframe[key].values.tolist()), dataframe.index)
  if(option == "p_np"):
    print(f"transform ${key} from present notpresent to 1 0")
    return pd.Series(map(lambda x: dict(present=1, notpresent=0)[x], dataframe[key].values.tolist()), dataframe.index)

X = df_kidney.drop('class', axis=1)
y = df_kidney['class']

from sklearn.model_selection import train_test_split

X_train,X_test,y_train,y_test = train_test_split(X,y, test_size=0.2,random_state=200)

# transform data to the same standard scaler.
from sklearn.preprocessing import StandardScaler
scaler = StandardScaler()
scaler.fit(X_train)

X_train = scaler.transform(X_train)
X_test = scaler.transform(X_test)
```

ภาพที่ 4 การแปลงข้อมูลให้อยู่ในสเกลเดียวกันและแบ่งชุดข้อมูลเป็นชุดฝึกฝนและชุดทดสอบ

3. การสร้างแบบจำลอง

ในการสร้างแบบจำลอง ผู้วิจัยได้ใช้ Scikit-Learn ซึ่งเป็น Python Library ในการทำเหมืองข้อมูล โดยใช้ประเภทการจำแนกประเภทข้อมูล (Classification) ในการสร้างแบบจำลองเพื่อจำแนกประเภทประเภทข้อมูลความเสียงการเป็นโรคไต โดยนำข้อมูลที่ได้มาหลังจากการเตรียมข้อมูล นำไปสร้างแบบจำลองด้วย วิธีโครงข่ายประสาทเทียม ต้นไม้ตัดสินใจ และการเรียนรู้แบบเบย์ ดังภาพที่ 5-7 ตามลำดับ ซึ่งทำการทดลองด้วยการแบ่งชุดข้อมูลฝึกฝน (Training Data) ชุดข้อมูลทดสอบ

(Testing Data) โดยแบ่งแบบเปอร์เซ็นด้วยการสุ่ม โดยมีอัตราส่วนชุดฝึกฝนต่อชุดทดสอบ ที่ใช้ทดลองดังนี้ 50:50, 60:40, 70:30 และ 80:20 ตามลำดับ และทำการทดสอบประสิทธิภาพแบบจำลองด้วย 10 Fold Cross Validation ในทุกๆ อัตราส่วนชุดฝึกฝนและชุดทดสอบ โดยโค้ดฟังก์ชัน Cross Validation แสดงดังภาพที่ 8

```python
from sklearn.neural_network import MLPClassifier

mlp = MLPClassifier(
    hidden_layer_sizes=(12,),
    max_iter=250,
    learning_rate='invscaling',
    verbose=False,
    momentum=0.3
)
mlp.fit(X_train, y_train)
```

ภาพที่ 5 โค้ดแบบจำลองโครงข่ายประสาทเทียม

```python
from sklearn.naive_bayes import GaussianNB

gnb = GaussianNB()
gnb.fit(X_train, y_train)
y_pred_kidney_NB = gnb.predict(X_test)
```

ภาพที่ 6 โค้ดแบบจำลองการเรียนรู้แบบเบย์

```python
from sklearn import tree

DT = tree.DecisionTreeClassifier(
    max_depth=7,
    splitter='best',
    criterion='gini'
)
DT.fit(X_train, y_train)
y_predict_DT = DT.predict(X_test)
```

ภาพที่ 7 โค้ดแบบจำลองต้นไม้ตัดสินใจ

JSET LRU
วารสารวิทยาศาสตร์ วิศวกรรมศาสตร์ และเทคโนโลยี
มหาวิทยาลัยราชภัฏเลย
Journal of Science, Engineering and Technology
LOEI RAJABHAT UNIVERSITY

```python
from sklearn.model_selection import cross_validate

def cross_validation(model, X_data, y_data, fcv=10):
    scoring_data = ['accuracy', 'precision', 'recall', 'f1']
    results = cross_validate(estimator=model,
                             X=X_data,
                             y=y_data,
                             cv=fcv,
                             scoring=scoring_data,
                             return_train_score=True)
    return  {"ค่าความถูกต้อง ขณะเทรน (train accuracy)": results['train_accuracy'],
             "---> ค่าเฉลี่ย ของค่าความถูกต้อง ขณะเทรน - (mean train accuracy)": results['train_accuracy'].mean(),
             "ค่าความแม่นยำ ขณะเทรน (train precision)": results['train_precision'],
             "---> ค่าเฉลี่ย ของค่าความแม่นยำ ขณะเทรน  (mean train precision)": results['train_precision'].mean(),
             "ค่าความระลึก ขณะเทรน  (train recall)": results['train_recall'],
             "---> ค่าเฉลี่ย ของค่าความระลึก ขณะเทรน (mean train recall)": results['train_recall'].mean(),
             "ค่า f1 ขณะเทรน (train f1 scores)": results['train_f1'],
             "---> ค่าเฉลี่ย ของค่า f1  ขณะเทรน (mean train f1 scores)": results['train_f1'].mean(),
             "ค่าความถูกต้อง ขณะทดสอบ (test_accuracy)": results['test_accuracy'],
             "---> ค่าเฉลี่ย ของค่าความถูกต้อง ขณะทดสอบ - (mean test accuracy)": results['test_accuracy'].mean(),
             "ค่าความแม่นยำ ขณะทดสอบ (test precision)": results['test_precision'],
             "---> ค่าเฉลี่ย ของค่าความแม่นยำ ขณะทดสอบ (test precision)": results['test_precision'].mean(),
             "ค่าความระลึก ขณะทดสอบ  (test recall)": results['test_recall'],
             "---> ค่าเฉลี่ย ของค่าความระลึก ขณะทดสอบ  (mean test recall)": results['test_recall'].mean(),
             "ค่า f1 ขณะทดสอบ (test f1 scores)": results['test_f1'],
             "---> ค่าเฉลี่ย ของค่า f1  ขณะทดสอบ (mean test f1 scores)": results['test_f1'].mean()
             }
```

ภาพที่ 8 ฟังก์ชันการ Cross Validation

4. การวัดประสิทธิภาพแบบจำลอง

การวัดประสิทธิภาพแบบจำลอง (ณัฐวดี และประภาสิริ, 2562) สามารถประเมินได้จากสมการ 1) ค่าความถูกต้อง (Accuracy) ในการจำแนกข้อมูล ในตาราง Confusion Matrix ซึ่งเป็นตารางสรุปการจำแนกข้อมูลของแบบจำลองได้ถูกต้องและไม่ถูกต้อง สามารถนำมาคำนวณเพื่อวัดประสิทธิภาพโดยการหา 2) ค่าความแม่นยำ (Precision) คำนวณจากค่าของข้อมูลที่มีการจำแนกถูกต้องโดยพิจารณาจากผลรวมของการจำแนกข้อมูลทั้งหมด 3) ค่าความระลึก (Recall) คำนวณจากค่าของข้อมูลที่ผลลัพธ์ถูกต้องโดยพิจารณาจากข้อมูลของผลลัพธ์เดียวกัน 4) ค่าความถ่วงดุล (F-measure) คำนวณได้จากค่าเฉลี่ยระหว่างค่าความแม่นยำและค่าความระลึก รายละเอียดดังสมการที่ 3-6 ต่อไปนี้

$$\text{Accuracy} = \frac{\text{TP+TN}}{\text{TP+FP+FN+TN}} \quad (3)$$

$$\text{Precision} = \frac{\text{TP}}{\text{TP+FP}} \quad (4)$$

$$\text{Recall} = \frac{\text{TP}}{\text{TP+FN}} \quad (5)$$

$$\text{F} - \text{measure} = \frac{(\text{Precision x Recall})}{(\text{Precision+Recall})} \text{ x } 2 \quad (6)$$

TP = ข้อมูลเป็นจริงและผลการทำนายบอกว่าจริง

TN = ข้อมูลเป็นข้อมูลไม่จริงและผลการทำนายบอกว่าไม่จริง

FP = ข้อมูลเป็นจริงแต่ผลการทำนายบอกว่าไม่จริง

FN = ข้อมูลเป็นข้อมูลไม่จริงแต่ผลการทำนายบอกว่าจริง

ผลการวิจัยและวิจารณ์ผลการวิจัย

ผลการทดลองของงานวิจัยประกอบด้วย 4 ส่วน ได้แก่ 1) ผลการสร้างแบบจำลองด้วยโครงข่ายประสาทเทียม 2) ผลการสร้างแบบจำลองด้วยต้นไม้ตัดสินใจ 3) ผลการสร้างแบบจำลองด้วยการเรียนรู้แบบเบย์ และ 4) ผลการเปรียบเทียบประสิทธิภาพแบบจำลอง ดังนี้

1. ผลการสร้างแบบจำลองด้วยโครงข่ายประสาทเทียม

การทดสอบโดยใช้แบบจำลองจากเทคนิคโครงข่ายประสาทเทียมแบบ Multilayer Perceptron จาก Scikit-Learn ทำการกำหนดโมเมนตัม (Momentum) ที่ 0.3 อัตราการเรียนรู้ (Learning Rate) แบบ Invscaling ทำการเทรนทั้งหมด 250 รอบ และกำหนด 1 Hidden Layer 12 Unit แบ่งชุดข้อมูลด้วยการสุ่มออกเป็น 2 ส่วน ด้วยอัตราส่วน 50:50 60:40 70:30 และ 80:20 พบว่าแบบจำลองที่ดีที่สุด คือการแบ่งชุดข้อมูลด้วยอัตราส่วน 60:40 ได้ค่าความถูกต้อง 97.50% ค่าความแม่นยำ 100% ค่าความระลึก 95.88% ค่าความถ่วงดุล 97.83% ดังตารางที่ 2

ตารางที่ 2 ค่าประสิทธิภาพจากเทคนิคโครงข่ายประสาทเทียม

Ratio Train: Test	Accuracy (%)	Precision (%)	Recall (%)	F-measure (%)
50:50	93.49	100	89.09	94.10
60:40	97.50	100	95.88	97.83
70:30	94.16	100	90.00	94.35
80:20	93.75	100	88.50	93.01

2. ผลการสร้างแบบจำลองด้วยต้นไม้ตัดสินใจ

การทดสอบโดยใช้แบบจำลองจากเทคนิคต้นไม้ตัดสินใจ จาก Scikit-Learn ทำการกำหนดค่าความลึกสูงสุดของต้นไม้ (Max depth) ที่ 7 Splitter แบบ Best และ Criterion แบบ Gini แบ่งชุดข้อมูลด้วยการสุ่มออกเป็น 2 ส่วน ด้วยอัตราส่วน 50:50 60:40 70:30 และ 80:20 พบว่าแบบจำลองที่ดีที่สุด คือการแบ่งชุดข้อมูลด้วยอัตราส่วน 50:50 ได้ค่าความถูกต้อง 98.47% ค่าความแม่นยำ 98.33% ค่าความระลึก 99.16 % ค่าความถ่วงดุล 98.73% ดังตารางที่ 3

ตารางที่ 3 ค่าประสิทธิภาพจากเทคนิคต้นไม้ตัดสินใจ

Ratio Train: Test	Accuracy (%)	Precision (%)	Recall (%)	F-measure (%)
50:50	98.47	98.33	99.16	98.73
60:40	96.12	99.00	94.55	96.35
70:30	95.00	97.50	94.28	95.56
80:20	93.75	96.00	93.49	94.34

3. ผลการสร้างแบบจำลองด้วยการเรียนรู้แบบเบย์

การทดสอบโดยใช้แบบจำลองจากการเรียนรู้แบบเบย์ จาก Scikit-Learn แบ่งชุดข้อมูลด้วยการสุ่มออกเป็น 2 ส่วน ด้วยอัตราส่วน 50:50 60:40 70:30 และ 80:20 พบว่าแบบจำลองที่ดีที่สุด คือการแบ่งชุดข้อมูลด้วยอัตราส่วน 50:50 ได้ค่าความถูกต้อง 98.00% ค่าความแม่นยำ 100% ค่าความระลึก 96.66% ค่าความถ่วงดุล 98.26% ดังตารางที่ 4

ตารางที่ 4 ค่าประสิทธิภาพจากเทคนิคการเรียนรู้แบบเบย์

Ratio Train: Test	Accuracy (%)	Precision (%)	Recall (%)	F-measure (%)
50:50	98.00	100	96.66	98.26
60:40	97.45	100	95.77	97.77
70:30	97.50	100	95.89	97.79
80:20	97.50	100	95.50	97.46

4. ผลการเปรียบเทียบประสิทธิภาพแบบจำลอง

จากการเปรียบเทียบประสิทธิภาพการทำงานของทั้งสามเทคนิคได้ผลการทดลองดังตารางที่ 5 พบว่าแบบจำลองแบบโครงข่ายประสาทเทียม มีค่าความถูกต้อง 97.50% ค่าความแม่นยำ 100% ค่าความระลึก 95.88% และค่าความถ่วงดุล 98.00% แบบจำลองจากเทคนิคต้นไม้ตัดสินใจ มีค่าความถูกต้อง 98.47% ค่าความแม่นยำ 98.33% ค่าความระลึก 99.16% และค่าความถ่วงดุล 98.73% และแบบจำลองจากเทคนิคการเรียนรู้แบบเบย์ มีค่าความถูกต้อง 98.00% ค่าความแม่นยำ 100% ค่าความระลึก 96.66% และค่าความถ่วงดุล 98.26% ซึ่งจะเห็นได้ว่าแบบจำลองจากเทคนิคต้นไม้ตัดสินใจมีค่าประสิทธิภาพดีกว่าแบบจำลองจากเทคนิคโครงข่ายประสาทเทียม (ANN) และเทคนิคการเรียนรู้แบบเบย์ โดยมีค่าความถูกต้อง 98.47% ค่าความแม่นยำ 98.33% ค่าความระลึก 99.16% และค่าความถ่วงดุล 98.73% ดังนั้นสรุปได้ว่า เทคนิคต้นไม้ตัดสินใจเป็นแบบจำลองที่เหมาะสมที่สุดในการจำแนกประเภทข้อมูล ซึ่งสอดคล้องกับงานวิจัยของ (อุกฤษฏ์, 2564) ที่มีการทดลองใช้เทคนิคทั้ง 3 เทคนิคเช่นกัน พบว่า เทคนิค Decision Tree เป็นเทคนิคที่ดีที่สุดในการสร้างแบบจำลองในการพยากรณ์โรคไฮโปไทรอยด์ โดยให้ค่าความถูกต้อง 99.86% ค่าความไว 99.85% และค่าความจำเพาะ 100%

ตารางที่ 5 ตารางเปรียบเทียบประสิทธิภาพแบบจำลอง

Models	Accuracy (%)	Precision (%)	Recall (%)	F-measure (%)
Neural Network	97.50	100	95.88	97.83
Decision Tree	98.47	98.33	99.16	98.73
Naïve Bayes	98.00	100	96.66	98.26

ดังนั้นผลของการจำแนกประเภทข้อมูลของแบบจำลองต้นไม้ตัดสินใจ สามารถสร้างกฎในการตัดสินใจได้ 5 กฎ โดยการแปลความหมายจากต้นไม้ตัดสินใจที่ได้ ดังแสดงภาพที่ 10 สามารถแปลความหมายได้ดังนี้

จากต้นไม้ตัดสินใจที่ได้จากการสร้างแบบจำลอง คลาสคำตอบ คือ y[0] หมายถึง ไม่เป็นโรคไต และ y[1] หมายถึงเป็นโรคไต สำหรับฟีเจอร์นั้น จากภาพที่ 9 ที่แสดงตำแหน่งของฟีเจอร์ในชุดข้อมูลจากตัวแปร X โดย X[2] หมายถึง sg หรือ ค่าความถ่วงจำเพาะ (Specific Gravity) X[14] หมายถึง hemo หรือ ค่าสารสีของเม็ดเลือดแดง (Hemoglobin) X[15] หมายถึง pcv หรือ ร้อยละของเม็ดเลือดแดงต่อปริมาณเลือดทั้งหมด (Packed Cell Volume) และ X[19] หมายถึง dm หรือ โรคเบาหวาน (Diabetes Mellitus) ดังนั้น กฎของต้นไม้ตัดสินใจที่ได้จากภาพที่ 10 มีดังนี้

กฎข้อที่ 1 ถ้า hemo <= 0.288 และ pcv <= 1.273 คลาสคำตอบคือ y[1] หมายถึง เป็นโรคไต

กฎข้อที่ 2 ถ้า hemo <= 0.288 และ pcv > 1.273 คลาสคำตอบคือ y[0] หมายถึง ไม่เป็นโรคไต

กฎข้อที่ 3 ถ้า hemo > 0.288 และ sg <= 0.112 คลาสคำตอบคือ y[1] หมายถึง เป็นโรคไต

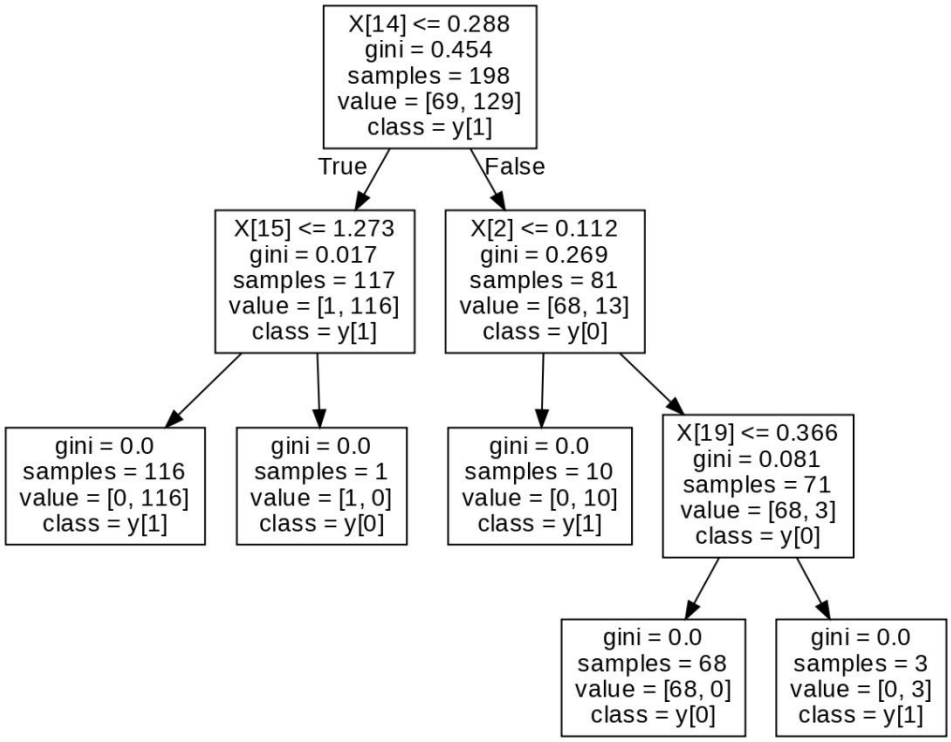
กฎข้อที่ 4 ถ้า hemo > 0.288 และ sg > 0.112 และ dm <= 0.366 คลาสคำตอบคือ y[0] หมายถึง ไม่เป็นโรคไต

กฎข้อที่ 5 ถ้า hemo > 0.288 และ sg > 0.112 และ dm > 0.366 คลาสคำตอบคือ y[1] หมายถึง เป็นโรคไต

```
[12]  1 for idx, key in enumerate(X):
      2   print(idx, key)

      0 age
      1 bp
      2 sg
      3 al
      4 su
      5 rbc
      6 pc
      7 pcc
      8 ba
      9 bgr
      10 bu
      11 sc
      12 sod
      13 pot
      14 hemo
      15 pcv
      16 wbcc
      17 rbcc
      18 htn
      19 dm
      20 cad
      21 appet
      22 pe
      23 ane
```

ภาพที่ 9 ตำแหน่งของฟีเจอร์ในชุดข้อมูลจากตัวแปร X



ภาพที่ 10 กฎต้นไม้ตัดสินใจที่ได้จากการสร้างแบบจำลอง

สรุปผลการวิจัย

งานวิจัยนี้เป็นการเปรียบเทียบประสิทธิภาพการจำแนกประเภทข้อมูลความเสี่ยงการเป็นโรคไตด้วยเทคนิคเหมืองข้อมูล ซึ่งมีวัตถุประสงค์เพื่อนำเทคนิคเหมืองข้อมูลแบบการจำแนกประเภทมาประยุกต์ใช้ในการทำนายความเสี่ยงการเป็นโรคไต พร้อมทั้งเปรียบเทียบประสิทธิภาพ เพื่อให้ได้อัลกอลิทึมที่มีประสิทธิภาพเหมาะสมด้วยอัลกอลิทึมเหมืองข้อมูล ทั้งสามแบบ คือโครงข่ายประสาทเทียม การเรียนรู้แบบเบย์ และต้นไม้ตัดสินใจ ซึ่งชุดข้อมูลที่ใช้ในครั้งนี้ คือ ชุดข้อมูลระยะเริ่มต้นของโรคไตเรื้อรังของคนอินเดีย จากฐานข้อมูล UCI ปี 2015 ซึ่งมีจำนวนข้อมูล 400 ชุดข้อมูล และปัจจัยสำหรับการวิเคราะห์ 24 ปัจจัย คือ อายุ ความดันโลหิต ค่าความถ่วงจำเพาะ ค่าโปรตีนอัลบูมีน ระดับน้ำตาล เซลล์เม็ดเลือดแดง เซลล์หนอง ก้อนเซลล์หนอง แบคทีเรีย ค่าน้ำตาลในเลือดแบบสุ่ม ค่ายูเรียในเลือด ค่าซีรั่มครีเอตินีน ค่าโซเดียม ค่าโพแทสเซียม ค่าสารสีของเม็ดเลือดแดง ร้อยละของเม็ดเลือดแดงต่อปริมาณเลือดทั้งหมด จำนวนเซลล์เม็ดเลือดขาว จำนวนเซลล์เม็ดเลือดแดง โรคความดันโลหิตสูง โรคเบาหวาน โรคหลอดเลือดหัวใจ ความอยากอาหาร อาการบวมเท้า และโรคโลหิตจาง การแบ่งข้อมูลสำหรับใช้ในการทดลองสร้างแบบจำลองกรจำแนกประเภทข้อมูล แบ่งออกเป็น 2 ชุดข้อมูล คือ ชุดข้อมูลฝึกฝน (Training Data) และชุดข้อมูลทดสอบ (Testing Data) โดยแบ่งแบบเปอร์เซ็นด้วยการสุ่ม โดยมีอัตราส่วนชุดข้อมูลฝึกฝนต่อชุดข้อมูลทดสอบ คือ 50:50, 60:40, 70:30 และ 80:20 และทดสอบประสิทธิภาพด้วยวิธีการ 10-Fold Cross Validation ผลการวิจัยพบว่า ตัวแบบจำลองการจำแนกประเภทข้อมูลด้วยต้นไม้ตัดสินใจมีประสิทธิภาพดีที่สุด โดยมีค่าความถูกต้อง 98.47% ค่าความแม่นยำ 98.33% ค่าความระลึก 99.16% และค่าความ

ถ่วงดุล 98.73% และได้กฎในการประกอบการตัดสินใจทั้งหมด 5 กฎ ดังนั้นจึงสรุปได้ว่าเทคนิค Decision Tree มีความเหมาะสมในการนำมาสร้างแบบจำลองการจำแนกประเภทข้อมูลความเสี่ยงการเป็นโรคไต

ข้อเสนอแนะการทำวิจัยครั้งต่อไป ควรมีวิธีการแบ่งชุดข้อมูลที่หลากหลายรูปแบบสำหรับใช้ในการฝึกฝนและการทดสอบ อีกทั้งควรมีการนำอัลกอริทึมหลายแบบมาทำการทดลอง เช่นเทคนิคการเรียนรู้ร่วมกัน (Ensemble Learning) ได้แก่ อัลกอริทึม Vote, Bagging หรือ Boosting เพื่อหาประสิทธิภาพในการจำแนกหรือการทำนายที่เหมาะสมกว่า

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

ณัฐวดี หงษ์บุญมี และประภาสิริ ตรีพาณิชกุล. (2562). การเปรียบเทียบประสิทธิภาพการจำแนกข้อมูลเพื่อวิเคราะห์ปัจจัยความเสี่ยงที่ส่งผลต่อการเกิดโรคไฮเปอร์ไทรอยด์ด้วยเทคนิคเหมืองข้อมูล. **วารสารวิทยาศาสตร์และเทคโนโลยีสารสนเทศ**, 9(1), 41-51.

ทวี ศิลารักษ์, สัฆวี ปิยะปัณฑิตกุล, และวิริณธิ์ กิตติพิชัย. (2563). ปัจจัยทำนายการเกิดโรคไตเรื้อรังในผู้ป่วยโรคเบาหวานในจังหวัดศรีสะเกษ. **วารสารพยาบาลสงขลานครินทร์**, 40(2), 109-121.

นงเยาว์ ในอรุณ. (2564). การเปรียบเทียบประสิทธิภาพของแบบจำลองการทำนายความเสี่ยงโรคหัวใจและหลอดเลือดโดยใช้อัลกอริทึมเหมืองข้อมูล. **วารสารวิทยาศาสตร์และเทคโนโลยี มหาวิทยาลัยมหาสารคาม**, 40(2), 137-147.

ปริญญา สงวนสัตย์. (2562). **Artificial Intelligence with Machine Learning, AI สร้างได้ด้วยแมชชีนเลิร์นนิ่ง**. นนทบุรี: ไอดีซี พรีเมียร์.

โรงพยาบาลพระรามเก้า. (2563). **โรคไตเรื้อรัง**. ค้นจาก https://www.praram9.com/โรคไตเรื้อรัง-2

วิทยา พรพัชรพงศ์. (2555). **โครงข่ายประสาทเทียม (Artificial Neural Networks - ANN)**. ค้นจาก https://www.gotoknow.org/posts/163433

สุพัฒน์ วาณิชย์การ. (2554). ปัจจัยเสี่ยงต่อการเป็นโรคไต. **วารสารมูลนิธิโรคไตแห่งประเทศไทย,** 25(49), 14-17.

อุกฤษฏ์ ศรีสุข. (2564). การเปรียบเทียบประสิทธิภาพของเทคนิคเหมืองข้อมูลสำหรับอุบัติการณ์ของผู้ป่วย. **วารสารวิทยาศาสตร์และเทคโนโลยี มหาวิทยาลัยมหาสารคาม,** 40(2), 157-163.

IBM. (2020). **What are neural networks**. Retrieved from https://www.ibm.com/cloud/learn/neural-networks

MastersInDataScience. (2022). **What Is a Decision Tree**. Retrieved from https://www.mastersindatascience.org/learning/machine-learning-algorithms/decision-tree

Noyunsan, C., Katanyukul, T., and Saikaew, K. (2018). Performance evaluation of supervised learning algorithms with various training data sizes and missing attributes. **Engineering and Applied Science Research**, 45(3), 221–229.

World Kidney Day. (2022). **What is chronic kidney disease**. Retrieved from https://www.worldkidneyday.org/facts/chronic-kidney-disease/