การเปรียบเทียบประสิทธิภาพวิธีการจำแนกกลุ่มการเป็นโรคไตเรื้อรัง : กรณีศึกษาโรงพยาบาลแห่งหนึ่งในประเทศอินเดีย

Efficiency Comparison of Data Mining Classification Methods for Chronic Kidney Disease: A Case Study of a Hospital in India

สุรวัชร ศรีเปารยะ* และสายชล สินสมบูรณ์ทอง

ภาควิชาสถิติ คณะวิทยาศาสตร์ สถาบันเทคโนโลยีพระจอมเกล้าเจ้าคุณทหารลาดกระบัง

ถนนฉลองกรุง เขตลาดกระบัง กรุงเทพมหานคร 10520

Surawat Sripaoraya* and Saichon Sinsomboonthong

Department of Statistics, Faculty of Science, King Mongkut's Institute of Technology Ladkrabang,

Chalongkrung Road, Ladkrabang, Bangkok 10520

บทคัดย่อ

งานวิจัยนี้มีวัตถุประสงค์เพื่อเปรียบเทียบประสิทธิภาพของวิธีการจำแนกกลุ่ม โดยเลือกใช้วิธีความใกล้เคียงกันมากที่สุด วิธีต้นไม้ตัดสินใจ วิธีโครงข่ายประสาทเทียม วิธีซัพพอร์ตเวกเตอร์แมชชีน วิธีฐานกฎ วิธีการถดถอยลอจิสติกและวิธีนาอีฟเบย์เพื่อวัดประสิทธิภาพการจำแนกกลุ่มโดยใช้ข้อมูลผู้ป่วยโรคไตเรื้อรังของโรงพยาบาลอพอลโล ประเทศอินเดีย โดยแบ่งข้อมูลเป็นชุดสร้างตัวแบบ และชุดทดสอบตัวแบบ ในอัตราส่วน 70 และ 30 ตามลำดับ จากการเปรียบเทียบประสิทธิภาพวิธีการจำแนกกลุ่มผู้ป่วยโรคไตเรื้อรัง โดยเปรียบเทียบจากค่าความถูกต้องและค่าความคลาดเคลื่อนกำลังสองเฉลี่ย วิธีการจำแนกกลุ่มที่มีประสิทธิภาพการจำแนกดีที่สุดคือ วิธีต้นไม้ตัดสินใจ ซึ่งให้ค่าความถูกต้อง คือ 100 % และค่าความคลาดเคลื่อนกำลังสองเฉลี่ยคือ 0.0059

คำสำคัญ : ความใกล้เคียงกันมากที่สุด; ต้นไม้ตัดสินใจ; โครงข่ายประสาทเทียม; ซัพพอร์ตเวกเตอร์แมชชีน; ฐานกฎ การถดถอยลอจิสติก; นาอีฟเบย์

Abstract

The objective of this research was to compare the efficiency of several data mining classification methods–K-nearest neighbor, decision tree, artificial neural network, support vector machine, rule-based, logistic regression and Naïve Bayes–for chronic kidney disease from data obtained from Apollo Hospital, India, that are archived in a UCI machine learning database

***ผู้รับผิดชอบบทความ :** surawat.income@gmail.com

repository. This dataset were divided into a training dataset and a testing dataset at 70 : 30 respectively. The efficiency measures used were accuracy and mean square error. Based on these measures and the Apollo Hospital dataset, the best classification method for chronic kidney disease was the decision tree method that achieved an accuracy of 100 % and a mean square error of 0.0059.

Keywords: K-nearest neighbor; decision tree; artificial neural network; support vector machine; rule-based, logistic regression; Naïve Bayes

1. บทนำ

ปัจจุบันในยุคของข้อมูลข่าวสาร องค์กรส่วนใหญ่มีข้อมูลที่ต้องจัดเก็บอยู่เป็นจำนวนมากมาย เช่น ระบบร้านค้าปลีก จะเก็บข้อมูลพนักงานในองค์กร ข้อมูลการซื้อขาย ข้อมูลสินค้าและข้อมูลลูกค้า เป็นต้น จะเห็นได้ว่ายิ่งองค์กรหรือรูปแบบธุรกิจมีขนาดใหญ่เท่าไร ย่อมทำให้การเก็บสะสมข้อมูลสำหรับองค์กรต่าง ๆ มีจำนวนมากขึ้น การเก็บข้อมูลจำนวนมากเหล่านี้ ลงในฐานข้อมูลเป็นวิธีที่นิยมใช้ในหลายองค์กร แต่ระบบการจัดการฐานข้อมูลทั่วไปไม่สามารถจัดการกับข้อมูลเหล่านี้ได้อย่างมีประสิทธิภาพ เนื่องจากใช้เวลานานในการดึงข้อมูลที่มีความสำคัญออกมาวิเคราะห์ ดังนั้นจึงได้เกิดเทคโนโลยีในการวิเคราะห์ข้อมูลที่มีความสำคัญออกมาจากแหล่งเก็บข้อมูลขนาดใหญ่ เรียกเทคโนโลยีนี้ว่าการทำเหมืองข้อมูล หรือการขุดค้นข้อมูล (data mining)

ข้อมูลสารสนเทศถือว่าเป็นข้อมูลที่สำคัญในการนำมาประกอบการตัดสินใจ จัดทำงานวิจัย จัดทำแผนนโยบายและแผนกลยุทธ หรือแผนยุทธศาสตร์ของหน่วยงานภาครัฐและภาคเอกชน สำหรับข้อมูลสารสนเทศที่ดีจะต้องมีความน่าเชื่อถือมีความเป็นปัจจุบัน ทันสมัย ทันเวลา ทันต่อเหตุการณ์ และสามารถนำมาใช้งานได้อย่างมีประสิทธิภาพ เพื่อให้เกิดประโยชน์สูงสุดต่อหน่วยงาน แต่การนำข้อมูลมาใช้ในบางครั้งอาจไม่สามารถใช้งานได้อย่างเต็มประสิทธิภาพและไม่ตรงตามความต้องการของหน่วยงานหรือองค์กร เนื่องจากอาจมีการนำข้อมูลไปวิเคราะห์โดยการจำแนกผิดวิธี ทำให้ความถูกต้องและผลลัพธ์ที่ได้จากการประมวลผลมีค่าลดลงหรืออาจได้ค่าผลลัพธ์ที่คลาดเคลื่อนไม่ตรงกับข้อเท็จจริง สำหรับการจำแนกเหล่านี้สามารถทำได้หลายวิธีซึ่งแต่ละวิธีจะมีความเหมาะสมกับข้อมูลแตกต่างกันไป ดังนั้นผู้ใช้งานหรือผู้วิจัยต้องเลือกวิธีจำแนกกลุ่มให้เหมาะสมกับข้อมูลโรคไตเรื้อรัง

ในช่วงที่ผ่านมามีผู้วิจัยได้ศึกษาวิธีการจำแนกกลุ่มด้วยวิธีการต่าง ๆ เพื่อหารูปแบบที่มีความเหมาะสมและมีประสิทธิภาพ ให้ค่าผลลัพธ์ใกล้เคียงค่าจริงมากที่สุด เช่น การเปรียบเทียบประสิทธิภาพการจำแนกรูปแบบการเรียนรู้ VARK ด้วยวิธีนาอีฟเบย์ (Naïve Bayes) วิธีต้นไม้ตัดสินใจ (decision tree) และวิธีฐานกฎ (rules based) ผลการศึกษาพบว่าวิธีต้นไม้ตัดสินใจให้ประสิทธิภาพสูงสุดคืออัลกอลิทึม J48 จากการศึกษาของ อรนุช และมนชัย [1] หรือการจำแนกบุคลากรในองค์กรสำหรับการสร้างแผนที่ความรู้ โดยข้อมูลที่ใช้ในการทดลองได้จากหน่วยงานการไฟฟ้าฝ่ายผลิตแห่งประเทศไทย สายงานรองผู้ว่าการบัญชีและการเงิน สายงานรองผู้ว่าการผลิตไฟฟ้า และสายงานรองผู้ว่าการเชื้อเพลิงด้วยอัลกอริทึม 3 แบบ ได้แก่ J48, OneR และนาอีฟเบย์ จากการทดลองพบว่าอัลกอริทึมที่ให้ค่าความถูกต้องในการทำนาย คือ J48 จากการศึกษาของ บุษกร [2] หรือการเปรียบเทียบ

ปัจจัยโรคประจำตัวผู้สูงอายุโดยใช้อัลกอริทึมการจำแนกกลุ่ม J48 และนาอีฟเบย์ ผลการทดสอบพบว่าวิธีต้นไม้ตัดสินใจโดยใช้อัลกอลิทึม J48 ให้ค่าความถูกต้องสูงกว่านาอีฟเบย์ โดยเบญจภัค [3] หรือการทดสอบประสิทธิภาพการพยากรณ์สถานะการชำระหนี้ลูกหนี้ธนาคารด้วยวิธีต้นไม้ตัดสินใจ วิธีถดถอยลอจิสติก (logistic regression) และวิธีนาอีฟเบย์ ผลการทดสอบพบว่าวิธีต้นไม้ตัดสินใจให้ค่าความแม่นยำสูงที่สุด รองลงมา คือ วิธีถดถอยลอจิสติกและนาอีฟเบย์ตามลำดับ โดย ชณัฐดาภรณ์ [4] หรือการเปรียบเทียบประสิทธิภาพการตรวจจับสิ่งผิดปกติทางเครือข่าย โดยการสแกนหาจุดอ่อน ซึ่งเป็นการบุกรุกทางเครือข่ายที่มีความสำคัญ งานวิจัยนี้ได้ศึกษาวิธีการจำแนกกลุ่ม (classification) มาตรวจจับการสแกน ได้แก่ วิธีซัพพอร์ตเวกเตอร์แมชชีน (support vector machine) วิธีต้นไม้ตัดสินใจ วิธีนาอีฟเบย์ และวิธีโครงข่ายประสาทเทียม (artificial neural network) ผลการทดสอบพบว่าวิธีต้นไม้ตัดสินใจ วิเคราะห์ได้แม่นยำที่สุดและมีค่า error ต่ำที่สุด โดยพลอยพรรณ [5] หรือการจำแนกข้อมูลโดยทดสอบประสิทธิภาพด้วยข้อมูล Austrian credit และ Bankruptey data โดยเลือกคุณลักษณะที่เหมาะสมโดยใช้ขั้นตอนวิธีเชิงพันธุกรรม เปรียบเทียบผลการวิจัยกับวิธีต้นไม้ตัดสินใจ วิธีโครงข่ายประสาทเทียม วิธีซัพพอร์ตเวกเตอร์แมชชีนกับวิธีเชิงพันธุกรรม พบว่าวิธีซัพพอร์ตเวกเตอร์แมชชีนที่ใช้ขั้นตอนวิธีเชิงพันธุกรรมจะให้ค่าความแม่นยำสูงที่สุด จากการศึกษาของ เดช และพยุง [6] และการทำนายผลของข้อมูลโรคไตเรื้อรัง ซึ่งเก็บรวบรวมข้อมูลจากกรมสาธารณสุขด้วยเทคนิคทำเหมืองข้อมูล โดยใช้วิธีซัพพอร์ทเวกเตอร์แมชชีนและวิธีโครงข่ายประสาทเทียม จากการศึกษาพบว่าวิธีโครงข่ายประสาทเทียมให้ค่าความถูกต้องคือ 87.70 % และวิธีซัพพอร์ตเวกเตอร์แมชชีนให้ค่าความถูกต้อง คือ 76.32 % ผลการทดสอบพบว่าวิธีโครงข่ายประสาทเทียมมีประสิทธิภาพในการทำนายผลได้ดีกว่าวิธีซัพพอร์ตเวกเตอร์แมชชีน จากการศึกษาของ Vijayarani and Dhayanand [7]

จากการศึกษาพบว่าผู้วิจัยได้ให้ความสำคัญกับปัญหาที่ต้องการวิธีการจำแนกกลุ่มข้อมูลที่ถูกต้องแม่นยำและรวดเร็ว โดยผู้วิจัยมีความสนใจที่จะวิจัยเกี่ยวกับข้อมูลทางด้านการแพทย์ ซึ่งจะนำข้อมูลผู้ป่วยโรคไตเรื้อรังมาวิเคราะห์ในงานวิจัยครั้งนี้ เพื่อเปรียบเทียบประสิทธิภาพวิธีการจำแนกกลุ่มด้วยวิธีทางด้านสถิติ 7 วิธี ได้แก่ วิธีความใกล้เคียงกันมากที่สุด (K-nearest neighbor) วิธีต้นไม้ตัดสินใจ วิธีโครงข่ายประสาทเทียม วิธีซัพพอร์ตเวกเตอร์แมชชีน วิธีฐานกฎ วิธีการถดถอยลอจิสติก และวิธีนาอีฟเบย์

2. ทฤษฎีที่เกี่ยวข้อง

2.1 การทำเหมืองข้อมูล (data mining)

คือการวิเคราะห์ข้อมูลเพื่อแยกประเภท จำแนกรูปแบบและความสัมพันธ์ของข้อมูลจากฐานข้อมูลที่มีขนาดใหญ่หรือคลังข้อมูล โดยมีวิธีต่าง ๆ หลายวิธี ซึ่งรูปแบบการทำเหมืองข้อมูลนั้นได้รวบรวมความรู้จากหลายแขนงเข้าไว้ด้วยกันซึ่งประกอบด้วยระบบการเรียนรู้ของเครื่องจักร (machine learning) ร่วมกับวิทยา-ศาสตร์สารสนเทศ (information science) สถิติ (statistic) และระบบฐานข้อมูล (database system) โดยทั่วไปแล้ววิธีที่นำมาใช้ส่วนใหญ่มี 5 ประเภท

2.1.1 วิธีการจำแนกกลุ่ม (classification) เป็นวิธีในการจำแนกกลุ่มข้อมูลด้วยคุณลักษณะต่าง ๆ ที่ได้มีการกำหนดไว้แล้ว วิธีนี้เหมาะกับการสร้างตัวแบบเพื่อการพยากรณ์ค่าข้อมูล (predictive modeling) ในอนาคตจากการที่ได้จำแนกกลุ่มข้อมูลตัวอย่างไว้แล้ว ซึ่งในลักษณะดังกล่าวเรียกว่าการเรียนรู้แบบมีผู้สอน (supervised learning) วิธีการจำแนกกลุ่มเป็น

กระบวนการสร้างตัวแบบเพื่อจัดข้อมูลให้อยู่ในกลุ่มที่กำหนด ตัวอย่าง เช่น การแบ่งประเภทลูกค้าว่าเชื่อถือได้หรือไม่ ซึ่งเป็นการสร้างตัวแบบโดยการเรียนรู้จากข้อมูลที่ได้กำหนดไว้เรียบร้อยแล้ว

2.1.2 วิธีการค้นหากฎความสัมพันธ์ (association rule discovery) เป็นวิธีที่ใช้ในการค้นหาความสัมพันธ์ของฐานข้อมูลที่มีขนาดใหญ่ เพื่อที่จะวิเคราะห์ข้อมูลและหาสิ่งที่ซ่อนอยู่ในข้อมูลนั้น เช่น การวิเคราะห์ข้อมูลการซื้อขายในซุปเปอร์มาร์เก็ต เพื่อวางแผนการส่งเสริมการขาย (promotion) และเตรียมการวางแผนการเรียงชั้นวางสินค้า (shelf) เช่น การวางน้ำอัดลมกับข้าวโพดคั่วไว้ใกล้กัน

2.1.3 วิธีการจัดกลุ่ม (clustering) เป็นวิธีการลดขนาดของข้อมูลด้วยการรวมกลุ่มตัวแปรที่มีลักษณะเดียวกันไว้ด้วยกัน ทำให้สามารถค้นหาข้อมูลที่ถูกละเลยไปได้ วิธีนี้มักถูกใช้เป็นขั้นตอนเบื้องต้นในการทำเหมืองข้อมูล และเหมาะกับข้อมูลที่ยังไม่มีกลุ่มอย่างชัดเจน จึงรวมกลุ่มเพื่อหากลุ่มต่าง ๆ ของข้อมูล โดยจำนวนกลุ่มของข้อมูลแทนด้วย k ซึ่งผู้ที่ใช้วิธีนี้จะเป็นผู้กำหนดจำนวนกลุ่ม วิธีนี้อาจเรียกว่าการจัดกลุ่มแบบเฉลี่ย k กลุ่ม (K-mean clustering)

2.1.4 วิธีการหาค่าที่แตกต่างจากค่ามาตรฐาน (deviation detection) เป็นวิธีในการหาค่าที่แตกต่างไปจากค่ามาตรฐาน หรือค่าที่คาดคิดไว้ว่าต่างไปมากน้อยเพียงใด โดยทั่วไปมักใช้วิธีทางสถิติหรือการแสดงให้เห็นภาพ สำหรับวิธีนี้ใช้ในการตรวจสอบลายเซ็น หรือปลอมบัตรเครดิต เป็นต้น

2.1.5 วิธีการวิเคราะห์ลำดับ (sequential analysis) เป็นวิธีในการวิเคราะห์ลำดับเพื่อค้นหารูปแบบของการปรากฏของข้อมูล ซึ่งปรากฏในรายการที่แยกออกมา เช่น ถ้าผู้ซื้อซื้อสินค้า A แล้วเขาจะซื้อสินค้า B ในภายหลัง วิธีนี้จะแตกต่างจากวิธีการค้นหาความสัมพันธ์ เพราะคำนึงถึงลำดับการซื้อด้วย

2.2 วิธีการแบ่งประเภทข้อมูลของวิธีการจำแนกกลุ่ม (classification)

การแบ่งประเภทข้อมูลคือกระบวนการสร้างตัวแบบเพื่อจัดการข้อมูลให้อยู่ในกลุ่มที่กำหนด เป็นการสร้างตัวแบบการจัดหมวดหมู่ได้จากกลุ่มตัวอย่างของข้อมูลที่ได้กำหนดไว้ล่วงหน้า และสามารถพยากรณ์กลุ่มของข้อมูลที่ยังไม่เคยนำมาจัดหมวดหมู่ได้ ตัวแบบที่ได้อาจอยู่ในรูปแบบต้นไม้ตัดสินใจ (decision tree) หรือโครงข่ายประสาทเทียม (artificial neural network)

ในการจัดหมวดหมู่จำเป็นต้องแบ่งข้อมูลออกเป็น 2 ส่วน ส่วนแรก คือ ข้อมูลสำหรับการเรียนรู้ (training dta) เพื่อให้ข้อมูลเรียนรู้และสร้างตัวแบบ (model construction) และส่วนที่สองคือข้อมูลสำหรับการทดสอบ (testing data) เพื่อประเมินความถูกต้องของตัวแบบ (model evaluation) อีกทั้งใช้ชุดข้อมูลที่ไม่เคยเห็นมาก่อน (unseen data) เพื่อกำหนดกลุ่มให้กับข้อมูลใหม่ที่ได้มาหรือทำนายค่าออกมาตามที่ต้องการ เช่น การจัดหมวดหมู่ของผู้ยื่นบัตรเครดิต (credit) เป็นระดับต่ำ ระดับกลางและระดับสูง ของความเสี่ยงที่จะได้รับ หรือการอนุมัติบุคคลเข้ารับทำงานในลักษณะงานต่าง ๆ [8]

2.3 วิธีความใกล้เคียงกันมากที่สุด (K-nearest neighbor)

เป็นวิธีการที่ได้รับความนิยมในการใช้งานอย่างมาก สาเหตุเนื่องจากเป็นวิธีการที่ง่ายและมีประสิทธิภาพซึ่งสามารถนำไปประยุกต์ใช้กับงานได้อย่างหลากหลาย เช่น งานทางด้านการจำแนกกลุ่ม (classification) รวมถึงงานทางด้านการแทนที่ข้อมูลที่สูญหาย (missing values imputation) [9] (รูปที่ 1)

2.4 วิธีต้นไม้ตัดสินใจ (decision tree)

เป็นตัวแบบทางคณิตศาสตร์ เพื่อหาทางเลือกที่ดีที่สุด โดยการนำข้อมูลมาสร้างตัวแบบการ

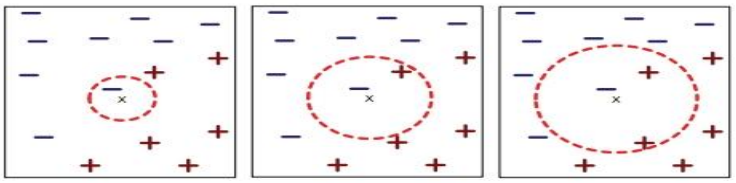
พยากรณ์ในรูปแบบของโครงสร้างต้นไม้ ซึ่งมีการเรียนรู้ข้อมูลแบบมีผู้สอน (supervised learning) สามารถสร้างตัวแบบการจัดกลุ่ม (clustering) ได้จากกลุ่มตัวอย่างของข้อมูลฝึกหัด (training data set) ได้โดยอัตโนมัติและสามารถพยากรณ์กลุ่มของรายการที่ยังไม่เคยนำมาจัดกลุ่มได้อีกด้วย

ส่วนประกอบของต้นไม้เพื่อการตัดสินใจ

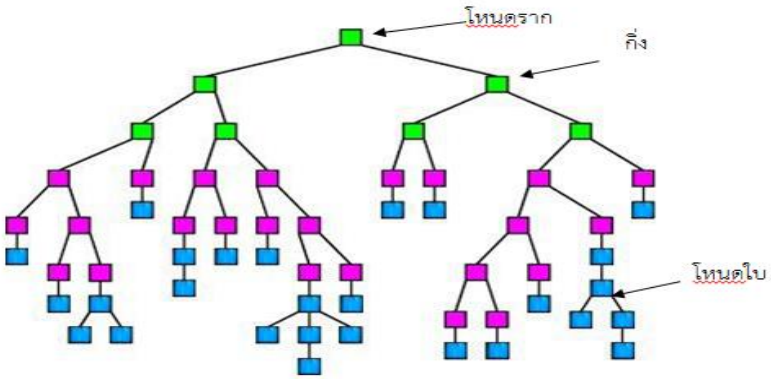
2.4.1 โหนด (node) คือ สมบัติต่าง ๆ เป็นจุดที่แยกข้อมูลว่าจะให้ไปในทิศทางใด ซึ่งโหนดที่อยู่สูงสุดเรียกว่าโหนดราก (root node)

2.4.2 กิ่ง (branch) คือ สมบัติของโหนดที่แตกออกมา โดยจำนวนของกิ่งจะเท่ากับสมบัติของโหนด

2.4.3 ใบ (leaf) คือ กลุ่มของผลลัพธ์ในการแยกแยะข้อมูล ซึ่งโหนดที่อยู่ล่างสุดเรียกว่าโหนดใบ (leaf node) โดยสามารถแสดงส่วนประกอบของต้นไม้ตัดสินใจ ดังรูปที่ 2 [8]



รูปที่ 1 ตัวอย่างของความใกล้เคียงกันมากที่สุด : (a) ความใกล้เคียงกันมากที่สุดโดยพิจารณาจากข้อมูล 1 ตัว, (b) ความใกล้เคียงกันมากที่สุดโดยพิจารณาจากข้อมูล 2 ตัว และ (c) ความใกล้เคียงกันมากที่สุดโดยพิจารณาจากข้อมูล 3 ตัว

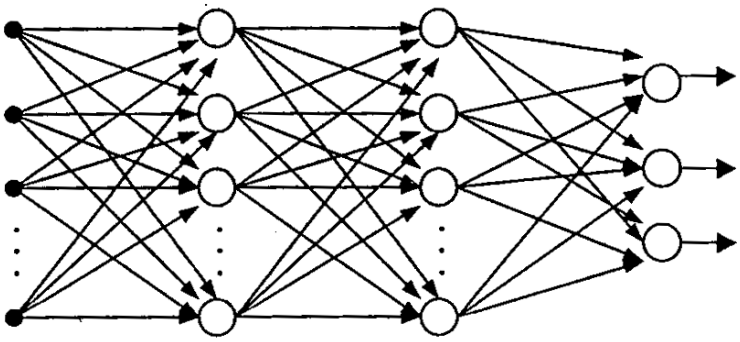


รูปที่ 2 ส่วนประกอบของต้นไม้ตัดสินใจ

2.5 วิธีโครงข่ายประสาทเทียม (artificial neural network)

มีแนวความคิดในการเรียนรู้ที่คล้ายคลึงกับระบบสมองมนุษย์ ขั้นตอนการนำโครงข่ายประสาทเทียมมาใช้ในการพยากรณ์ จะต้องอาศัยข้อมูลป้อนเข้าเพื่อสร้างแบบจำลองในการพยากรณข้อมูลในอนาคต โดยที่โครงข่ายประสาทเทียมจะพยายามละจำนวนของการทำนายที่ผิดพลาดให้ต่ำที่สุด

โครงข่ายประสาทเทียมแบบเปอรเซ็ป-ตรอนหลายชั้น เป็นการเรียนรู้แบบเปอร์เซ็ปตรอนสามารถพิสูจน์ได้ด้วยรอบการเรียนรู้ที่จำกัด อัลกอริธึมสามารถค้นหาค่าน้ำหนัก และค่าโน้มเอียงที่กำหนดเส้นขอบเขตของกลุ่ม สำหรับเซตข้อมูลที่สามารถแยกกันได้ด้วยเส้นตรง แต่สำหรับกรณีเส้นขอบเขตไม่เป็นเชิงเส้น โครงข่ายประสาทเทียมแบบเปอรเซ็ปตรอนจะไม่สามารถจำแนกกลุ่มได้ถูกต้องทั้งหมด จะมีข้อมูลบางค่าถูกจำแนกผิดกลุ่ม ซึ่งหมายถึงการเรียนรู้แบบเปอรเซ็ปตรอนจะไม่สามารถให้ค่าผิดพลาดเป็นศูนย์ได และจะวนเรียนรู้ไม่มีที่สิ้นสุด ซึ่งในขั้นตอนการเรียนรู้ต้องกำหนดจำนวนรอบในการเรียนรู้หรือกำหนดค่าผิดพลาดที่ยอมรับได้เพื่อหยุดการเรียนรู้ของอัลกอริทึม ซึ่งโครงข่ายประสาทเทียมแบบเปอรเซ็ปตรอนหลายชั้นจะใช้วิธีการเรียนรู้แบบแพร่กระจายย้อนกลับ (back-propagation learning) [10] (รูปที่ 3)



รูปที่ 3 โครงข่ายประสาทเทียมแบบเพอร์เซปตรอนหลายชั้น

2.6 วิธีซัพพอร์ตเวกเตอร์แมชชีน (support vector machine)

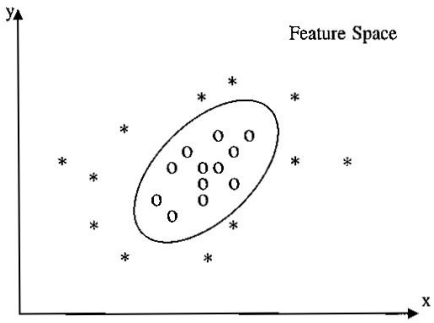
เป็นสมการที่ใช้จำแนกค่าคุณลักษณะของ 2 กลุ่ม ที่วางตัวอยู่ในพื้นที่คุณลักษณะ (feature space) ออกจากกันโดยจะสร้างเส้นแบ่ง (plane) ที่เป็นเส้นตรงขึ้นมา และเพื่อให้ทราบว่าเส้นตรงที่แบ่ง 2 กลุ่ม ออกจากกันนั้น เส้นตรงใดที่เป็นเส้นที่ดีที่สุด โดยเส้นตรงนั้นจะเพิ่มเส้นขอบ (margin) ออกไปทั้งสองข้าง โดยเส้นขอบที่เพิ่มนั้นจะขนานกับเส้นเดิมเสมอ เส้นขอบที่เพิ่มขึ้นมานี้จะขยายออกไปจนกว่าจะสัมผัสกับค่าของกลุ่มตัวอย่างที่ใกล้ที่สุด

เคอร์เนล (Kernel) ในโลกความเป็นจริงนั้น ข้อมูล 2 กลุ่ม ไม่ได้วางตัวในพื้นที่คุณลักษณะ และไม่สามารถแบ่งได้โดยเส้นตรง แต่ข้อมูลอาจจะจับกลุ่มกันในตำแหน่งต่าง ๆ ดังนั้นจึงเป็นปัญหาทำให้ไม่สามารถที่จะใช้สมการซัพพอร์ตเวกเตอร์แมชชีนแบบเชิงเส้นได้ ดังนั้นจะต้องมีเครื่องมือมาช่วยให้ข้อมูลเหล่านั้นเรียงตัวใหม่ในพื้นที่ เรียกว่าพื้นที่หลายมิติ (higher dimensional space) ดังรูปที่ 4 [11]

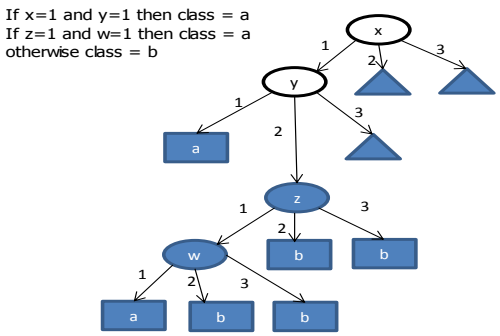
2.7 วิธีฐานกฎ (rule-based)

เป็นวิธีหนึ่งที่นิยมใช้เช่นเดียวกับวิธีต้นไม้ตัดสินใจ ข้อกำหนดหรือเงื่อนไข (antecedent or precondition) ของวิธีฐานกฎเป็นการทดสอบคล้ายกับการทดสอบของวิธีต้นไม้ตัดสินใจ แต่ผลของการทดสอบหรือผลลัพธ์ (consequent or conclusion) ที่ได้นั้น จะให้คำตอบ (class) ที่ใช้กับตัวอย่างภายใต้

กฎนั้น หรือบางครั้งก็อาจให้ค่าการแจกแจงความน่าจะเป็นของคำตอบต่าง ๆ กฎบางสูตรมีข้อกำหนดหรือเงื่อนไขที่เป็นการแสดงทางตรรกะทั่วไปมากกว่าที่จะเป็นการเชื่อมอย่างง่าย (simple conjunction) ถ้ากฎหนึ่งถูกนำไปใช้คำตอบ (หรือการแจกแจงความน่าจะเป็น) ที่กำหนดในข้อสรุปจะถูกนำไปใช้กับตัวอย่างเช่นกัน อย่างไรก็ตาม จะเกิดข้อขัดแย้งขึ้นเมื่อกฎหลายกฎมีข้อสรุปแตกต่างกัน [12]



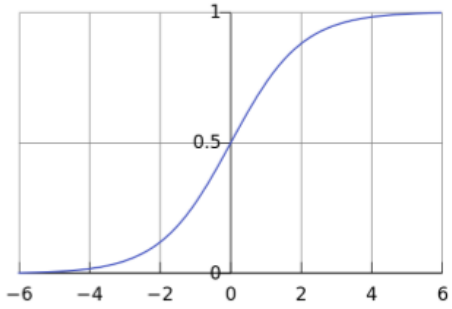
รูปที่ 4 รูปแบบการวางตัวที่ไม่สามารถแบ่งด้วยเส้นตรงได้



รูปที่ 5 อิทธิพลของกฎ

2.8 วิธีการถดถอยลอจิสติก (logistic regression)

ใช้เพื่อหาความสัมพันธ์ระหว่างตัวแปรอิสระและตัวแปรตาม โดยตัวแปรตามมีเพียงสองค่า คือ 0 และ 1 หากตัวแปรอิสระมีค่าน้อย ค่าของตัวแปรตามจะมีค่าเท่ากับ 0 และหากค่าตัวแปรอิสระมีค่ามาก ค่าของตัวแปรตามจะมีค่าเท่ากับ 1 [4]



รูปที่ 6 ฟังก์ชันลอจิสติก

2.9 วิธีนาอีฟเบย์ (Naïve Bayes)

เป็นเครื่องจักรเรียนรู้ที่อาศัยหลักการความน่าจะเป็น (probability) ตามทฤษฎีของเบย์ (Bayes theorem) ซึ่งมีอัลกอริทึมที่ไม่ซับซ้อน เป็นขั้นตอนวิธีในการจำแนกข้อมูล โดยการเรียนรู้ปัญหาที่เกิดขึ้นเพื่อนำมาสร้างเงื่อนไขการจำแนกข้อมูลใหม่ เป็นการจำแนกข้อมูลโดยใช้ความน่าจะเป็นและคำนวณการแจกแจงความน่าจะเป็นตามสมมติฐานที่ตั้งให้กับข้อมูล จากการคำนวณตัวอย่างใหม่ที่ได้จะถูกนำมาปรับเปลี่ยนการแจกแจง ซึ่งมีผลต่อการเพิ่มหรือลดความน่าจะเป็นของข้อมูล ข้อมูลใหม่ที่เกิดขึ้นและตัวแบบที่ตั้งไว้ให้กับข้อมูลจะถูกปรับเปลี่ยนไปตามข้อมูลใหม่โดยผนวกกับข้อมูลเดิมที่มี หลักการของนาอีฟเบย์ ใช้การคำนวณหาความน่าจะเป็นซึ่งถูกใช้ในการทำนายผลเป็นวิธีในการแก้ปัญหาแบบการจำแนกที่สามารถคาดการณ์ผลลัพธ์ได้ จะวิเคราะห์ความสัมพันธ์ระหว่างตัวแปรเพื่อใช้ในการสร้างเงื่อนไขความน่าจะเป็นสำหรับแต่ละความสัมพันธ์ นาอีฟเบย์เป็นวิธีจำแนกกลุ่มข้อมูลที่มีประสิทธิภาพ มีอัลกอริทึมในการทำงานที่ไม่ซับซ้อน เหมาะกับกรณีของเซตตัวอย่างที่มีจำนวนมากและสมบัติ (attribute) ของตัวอย่างไม่ขึ้นต่อกัน

[13]

2.10 การเปรียบเทียบประสิทธิภาพของวิธีการจำแนกกลุ่ม

การวัดประเมินผลมีความสำคัญเนื่องจากนำมาใช้ในการวิเคราะห์ประสิทธิภาพการทำงานของอัลกอริทึม ตัวชี้วัดการประเมินผลสามารถที่จะพัฒนาจากเมทริกซ์ความสับสน (confusion matrix) แสดงให้เห็นในตารางที่ 1

ตารางที่ 1 เมทริกซ์ความสับสน (confusion matrix)

		ผลการจำแนก	
		คำตอบเป็นบวก	คำตอบเป็นลบ
ค่าที่แท้จริง	คำตอบเป็นบวก	TP (true positive)	FN (false negative)
	คำตอบเป็นลบ	FP (false positive)	TN (true negative)

true positive (TP) คือ จำนวนข้อมูลที่จำแนกถูกว่าเป็นบวก; true negative (TN) คือ จำนวนข้อมูลที่จำแนกถูกว่าเป็นลบ; false positive (FP) คือ จำนวนข้อมูลที่จำแนกผิดว่าเป็นบวก ซึ่งค่าที่แท้จริงเป็นลบ; false negative (FN) คือ จำนวนข้อมูลที่จำแนกผิดว่าเป็นลบ ซึ่งค่าที่แท้จริงเป็นบวก

ค่าความถูกต้อง (accuracy) คือ การแสดงผลการวัดที่ได้มีความถูกต้องในรูปอัตราส่วน [1]

$$\text{Accuracy} = \frac{TP+TN}{TP+TN+FP+FN}\times 100$$

2.11 ค่าความคลาดเคลื่อนกำลังสองเฉลี่ย (mean square error, MSE)

ค่าความคลาดเคลื่อนกำลังสองเฉลี่ยใช้หลักการเดียวกันกับการหาค่าความแปรปรวนในทางสถิติ การวัดค่าความคลาดเคลื่อนด้วยวิธีนี้จะได้ค่าความคลาดเคลื่อนที่สูง เนื่องจากเป็นการนำความคลาดเคลื่อน ณ เวลาใด ๆ มายกกำลังสองก่อนที่จะหาผลรวม แล้วจึงนำมาหาค่าเฉลี่ยอีกครั้งหนึ่ง นั่นคือ ค่า MSE ยิ่งน้อย หมายถึง การพยากรณ์ยิ่งแม่นยำ มีสูตรในการคำนวณดังนี้ [14]

$$\text{MSE} = \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{n}$$

โดยที่ y_i แทน ค่าจริง; $\hat{y}_i$ แทน ค่าพยากรณ์

3. วิธีการดำเนินงานวิจัย

โรคไตเรื้อรัง (chronic kidney disease) เปนปญหาทางดานสาธารณสุขที่สำคัญของคนทั่วไป มีผลกระทบตอประชาชนทุกอายุ เชื้อชาติและทุกสถานะทางเศรษฐกิจ ความชุกและอุบัติการณของโรคที่เพิ่มขึ้นเนื่องมาจากโรคเบาหวาน ความดันโลหิตสูงและโรคอวน นอกจากนี้ผู้ป่วยโรคไตเรื้อรังมีความเสี่ยงเพิ่มขึ้นที่จะเกิดภาวะหลอดเลือดตีบแข็ง ทำให้เกิดโรคหัวใจและโรคหลอดเลือดตามมา และเป็นสาเหตุสำคัญที่ทำให้ผู้ป่วยเสียชีวิต นอกจากนี้โรคไตเรื้อรังเรื้อรังถือเป็นปัญหาสาธารณสุขที่สำคัญของทั่วโลกรวมถึงประเทศอินเดีย ผู้วิจัยจึงได้มีความสนใจที่จะนำข้อมูลเกี่ยวกับโรคไตเรื้อรังมาวิเคราะห์ในการทำวิจัยครั้งนี้ ซึ่งได้มีการค้นคว้าข้อมูลจากฐานข้อมูล UCI Machine Learning Repository จำนวน 1 ชุด คือ ข้อมูลผู้ป่วยโรคไตเรื้อรังของโรงพยาบาลอพอลโล ประเทศอินเดีย (Chronic Kidney Disease of Apollo Hospitals, India) รวบรวมข้อมูลผู้ป่วยโรคไตเรื้อรังโดย Rubini นักวิชาการจากมหาวิทยาลัยอลากัปปา (L. Jerlin Rubini Research Scholar Alagappa University) บริจาคข้อมูล ณ วันที่ 3 กรกฎาคม 2558 ซึ่งได้มีการเก็บรวบรวมข้อมูลเป็นระยะเวลา 2 เดือนจำนวนข้อมูลที่เก็บรวบรวมทั้งหมด 400 ระเบียน ประกอบด้วยตัวแปรอิสระ 24 ตัวแปร และตัวแปรตาม

1 ตัวแปร โดยมีรายละเอียดดังนี้ [15]

ตารางที่ 2 คุณลักษณะและรายละเอียดของข้อมูลผู้ป่วยโรคไตเรื้อรัง

ตัวแปร	ความหมาย	มาตรวัด	หน่วย
X_1	อายุ (age)	numerical	ปี (year)
X_2	ความดันโลหิต (blood pressure)	numerical	มิลลิเมตรต่อปรอท (mm/Hg)
X_3	ความถ่วงจำเพาะของปัสสาวะ (urine specific gravity)	nominal	ไม่มีหน่วย
X_4	ระดับโปรตีนอัลบูมิน (albumin)	nominal	มิลลิกรัมต่อเดซิลิตร (mgs/dl)
X_5	ระดับน้ำตาล (sugar)	nominal	ความเข้มข้นโมลาร์ (molar concentration)
X_6	เซลล์เม็ดเลือดแดง (red blood cell)	nominal	ปกติ, ไม่ปกติ (normal, abnormal)
X_7	เซลล์เม็ดเลือดขาวปนปัสสาวะขุ่น (pus cell)	nominal	ปกติ, ไม่ปกติ (normal, abnormal)
X_8	เซลล์เม็ดเลือดขาวปนปัสสาวะขุ่นเป็นกลุ่มก้อน (pus cell clumps)	nominal	ปรากฏ, ไม่ปรากฏ (present, not present)
X_9	แบคทีเรีย (bacteria)	nominal	ปรากฏ, ไม่ปรากฏ (present, not present)
X_{10}	ปริมาณน้ำตาลในเลือดแบบสุ่ม (blood glucose random)	numerical	มิลลิกรัมต่อเดซิลิตร (mgs/dl)
X_{11}	ปริมาณไนโตรเจนในกระแสเลือด (blood urea nitrogen)	numerical	มิลลิกรัมต่อเดซิลิตร (mgs/dl)
X_{12}	ปริมาณครีเอตินินในเลือด (serum creatinine)	numerical	มิลลิกรัมต่อเดซิลิตร (mgs/dl)
X_{13}	ปริมาณโซเดียมในเลือด (sodium)	numerical	มิลลิอิควิวาเลนซ์ต่อลิตร (mEq/L)
X_{14}	ปริมาณโปแตสเซียมในเลือด (potassium)	numerical	มิลลิอิควิวาเลนซ์ต่อลิตร (mEq/L)
X_{15}	ปริมาณฮีโมโกลบิน (hemoglobin)	numerical	กรัมต่อเลือด 100 มิลลิลิตร (gms)
X_{16}	ปริมาณเม็ดเลือดแดงต่อปริมาณเลือดทั้งหมด (packed cell volume)	numerical	-
X_{17}	จำนวนเซลล์เม็ดเลือดขาว (white blood cell count)	numerical	เซลล์ต่อลูกบาศก์มิลลิเมตร (cells/cumm)
X_{18}	จำนวนเซลล์เม็ดเลือดแดง (red blood cell count)	numerical	ล้านเซลล์ต่อลูกบาศก์มิลลิเมตร (millions/cumm)
X_{19}	ภาวะความดันโลหิตสูง (hypertension)	nominal	เป็น, ไม่เป็น (yes, no)
X_{20}	เบาหวาน (diabetes mellitus)	nominal	เป็น, ไม่เป็น (yes, no)
X_{21}	โรคหลอดเลือดหัวใจ (coronary artery disease)	nominal	เป็น,ไม่เป็น (yes, no)
X_{22}	ความอยากอาหาร (appetite)	nominal	ต้องการมาก, ต้องการน้อย (good, poor)
X_{23}	อาการเท้าบวม (pedal edema)	nominal	เป็น, ไม่เป็น (yes, no)
X_{24}	โรคโลหิตจาง (anemia)	nominal	เป็น, ไม่เป็น (yes, no)
Y	การเป็นโรคไตเรื้อรัง (chronic kidney)	nominal	0 = ผู้ป่วยไม่เป็นโรคไตเรื้อรัง 1 = ผู้ป่วยเป็นโรคไตเรื้อรัง

3.1 ศึกษาและวิเคราะห์ปัญหาที่เกี่ยวข้องกับงานวิจัย

3.2 ศึกษาและหาข้อมูลเพื่อนำมาเปรียบเทียบประสิทธิภาพในการทำนายผลการจำแนก ด้วยวิธีความใกล้เคียงกันมากที่สุด (K-nearest neighbor) วิธีต้นไม้ตัดสินใจ (decision tree) วิธีโครงข่ายประสาทเทียม (artificial neural network) วิธีซัพพอร์ทเวกเตอร์-แมชชีน (support vector machine) วิธีฐานกฎ (rule-based) วิธีการถดถอยลอจิสติก (logistic regression) และวิธีนาอีฟเบย์ (Naïve Bayes)

3.3 ศึกษาอัลกอริทึมของแต่ละวิธี ดังนี้

3.3.1 วิธีความใกล้เคียงกันมากที่สุด (K-nearest neighbor) ใช้อัลกอริทึมชนิด IBk เนื่องจากเป็นฟังก์ชันหลักที่สนใจ ซึ่งเป็นพื้นฐานของอัลกอริทึม อย่างไรก็ตาม อัลกอริทึม IBk ยังสามารถกำหนดน้ำหนัก ระยะห่างและทางเลือก (option) เพื่อกำหนดค่า k โดยใช้ cross-validation [16] ซึ่งในการศึกษาครั้งนี้ ผู้วิจัยได้กำหนดค่า k = 1 เนื่องจากให้ค่าความถูกต้องมากที่สุดและค่าความคลาดเคลื่อนกำลังสองเฉลี่ยน้อยที่สุด

3.3.2 วิธีต้นไม้ตัดสินใจ (decision tree) ใช้อัลกอริทึมชนิด J48 ซึ่งพัฒนาจาก ID3 สามารถใช้ได้กับข้อมูลแบบไม่ต่อเนื่องและแบบต่อเนื่อง ต่างจาก ID3 ที่ใช้ได้เพียงข้อมูลแบบไม่ต่อเนื่องเท่านั้น [8]

3.3.3 วิธีโครงข่ายประสาทเทียม (artificial neural network) ใช้อัลกอริทึมชนิดเพอร์เซปตรอนหลายชั้น (multilayer perceptron) โดยกำหนดค่าอัตราการเรียนรู้ (learning rate) เป็น 0.1 ค่าโมเมนตัม (momentum) เป็น 0.9 จำนวนรอบการสอน (training time) 20,000 รอบ การวิจัยครั้งนี้ใช้อัลกอริทึมของวิธีโครงข่ายประสาทเทียมแบบเพอร์-เซปตรอนหลายชั้น ที่มีชั้นซ่อน (hidden layer) 1 ชั้น แม้ว่าโครงสร้างโครงข่ายประสาทเทียมที่ซับซ้อนสามารถมีชั้นซ่อนมากกว่า 1 ชั้น แต่ในทางปฏิบัติการกำหนดชั้นซ่อน 1 ชั้น ก็เพียงพอต่อการวิเคราะห์ข้อมูล [17]

3.3.4 วิธีซัพพอร์ตเวกเตอร์แมชชีน (support vector machine) ใช้อัลกอริทึมชนิดโพลิ-โนเมียลเคอร์เนล (polynomial Kernel) เนื่องจากงานวิจัยที่อ้างอิง [18] ได้ผลลัพธ์ว่าวิธีซัพพอร์ตเวกเตอร์แมชชีนที่ใช้อัลกอริทึมชนิด polynomial Kernel ดีที่สุด

3.3.5 วิธีฐานกฎ (rule-based) ใช้อัลกอริทึม decision table เป็นเครื่องมือที่ใช้แสดงเงื่อนไขการตัดสินใจและเลือกการทำงานหรือกระทำกิจกรรมภายใต้เหตุการณ์ของเงื่อนไขที่ระบุ วิธีการตัดสินใจแบบ decision table จะเป็นตาราง 2 มิติอย่างง่าย [1]

3.3.6 วิธีถดถอยลอจิสติก (logistic regression) เนื่องจากงานวิจัยที่อ้างอิงจาก [4] กรณีศึกษาวิธีการถดถอยลอจิสติกเปรียบเทียบค่าความแม่นยำ ค่าเฉลี่ยความคลาดเคลื่อนสัมพัทธ์ และรากที่สองของค่าเฉลี่ยคลาดเคลื่อนกำลังสอง

3.3.7 วิธีนาอีฟเบย์ (Naïve Bayes) เป็นวิธีจำแนกกลุ่มข้อมูลที่มีประสิทธิภาพ มีอัลกอริทึมในการทำงานที่ไม่ซับซ้อน เหมาะกับกรณีของเซตตัวอย่างที่มีจำนวนมากและสมบัติ (attribute) [13]

3.4 แบ่งข้อมูลออกเป็น 2 ส่วน แบ่งข้อมูลส่วนที่ 1 สำหรับการสร้างตัวแบบ 70 เปอร์เซ็นต์ หรือจำนวน 280 ระเบียน และแบ่งข้อมูลส่วนที่ 2 สำหรับการทดสอบตัวแบบ 30 เปอร์เซ็นต์ หรือจำนวน 120 ระเบียน [8] โดยใช้โปรแกรม SPSS ในการสุ่มข้อมูลครั้งนี้

3.5 นำข้อมูลหลังจากแบ่งออกเป็น 2 ส่วน แล้วมาวิเคราะห์ด้วยโปรแกรม WEKA

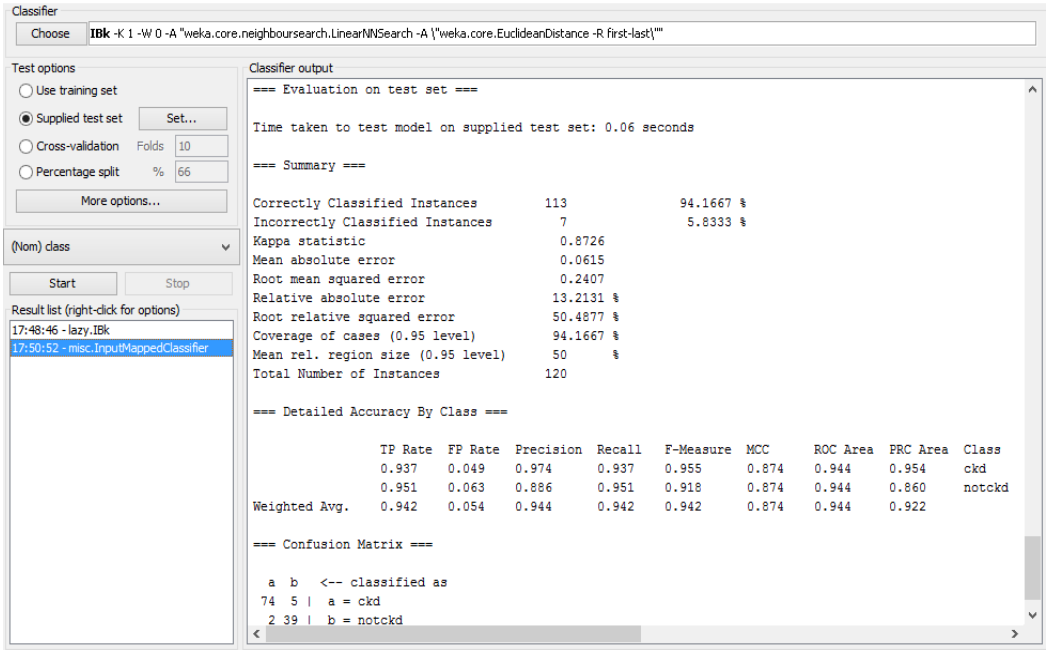
3.6 นำผลการวิเคราะห์ของแต่ละวิธีทั้ง 7 วิธี

มาเปรียบเทียบประสิทธิภาพในการทำนายผลการจำแนก ซึ่งในงานวิจัยอาศัยค่าความถูกต้อง (accuracy) เนื่องจากเป็นค่าที่บ่งบอกถึงความสามารถของเครื่องมือวัด (instrument) ในการอ่านค่าหรือแสดงค่าที่วัดได้เข้าใกล้ค่าจริง นอกจากนี้ งานวิจัยที่ผู้วิจัยได้ไปศึกษาค้นคว้ามา ส่วนใหญ่แล้วได้ใช้ค่าความถูกต้อง (accuracy) ทุกงานวิจัย และค่าความคลาดเคลื่อนกำลังสองเฉลี่ย (mean square error, MSE) เป็นมาตราวัดการประเมินค่าได้ดี เนื่องจากความคลาดเคลื่อนกำลังสองเฉลี่ยประกอบด้วยทั้งความแปรปรวนและความเอนเอียง

3.7 สรุปผลงานวิจัย อภิปรายผล และข้อเสนอแนะ

4. ผลการวิเคราะห์

ตัวอย่างการวิเคราะห์ข้อมูลวิธีใกล้เคียงกันมากที่สุด (รูปที่ 7)



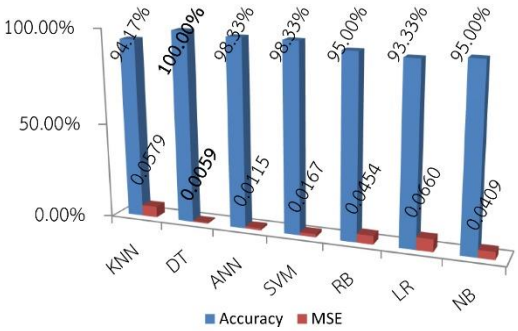
รูปที่ 7 ผลการวิเคราะห์ข้อมูลสำหรับการทดสอบตัวแบบด้วยวิธีความใกล้เคียงกันมากที่สุด

จากรูปที่ 7 พบว่าข้อมูล 120 ระเบียน ตัวแบบสามารถทำนายข้อมูลได้ถูกต้อง 113 ระเบียน คิดเป็น 94.17 % โดยมีจำนวนข้อมูลที่จำแนกถูกว่าเป็นโรคไตเรื้อรังมี 74 ระเบียน และจำนวนข้อมูลที่จำแนกถูกว่าไม่เป็นโรคไตเรื้อรังมี 39 ระเบียน ตัวแบบทำนายข้อมูลไม่ถูกต้องมี 7 ระเบียน คิดเป็น 5.83 % โดยมีจำนวนข้อมูลที่จำแนกผิดว่าเป็นโรคไตเรื้อรัง ซึ่งแท้จริงแล้วไม่ได้เป็นโรคไตเรื้อรังมี 2 ระเบียน และจำนวนข้อมูลที่จำแนกผิดว่าไม่เป็นโรคไตเรื้อรัง ซึ่งแท้จริงแล้วเป็นโรคไตเรื้อรังมี 5 ระเบียน โดยความคลาดเคลื่อนกำลังสองเฉลี่ยมีค่าเท่ากับ $(0.2407)^2 = 0.0579$

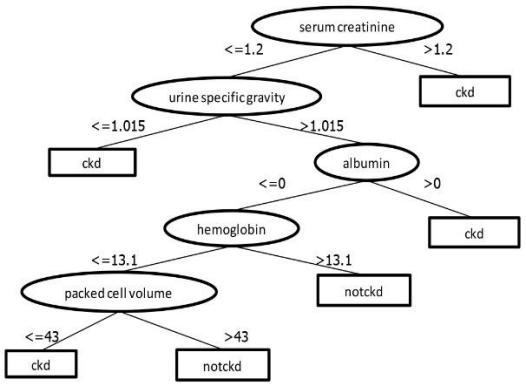
ผลลัพธ์การวิเคราะห์ข้อมูลค่าความถูกต้องและค่าความคลาดเคลื่อนกำลังสองเฉลี่ยสำหรับการทดสอบตัวแบบแต่ละวิธี (ตารางที่ 3)

ตารางที่ 3 การเปรียบเทียบค่าความถูกต้องและค่าความคลาดเคลื่อนกำลังสองเฉลี่ยของข้อมูลผู้ป่วยโรคไตเรื้อรัง

ข้อมูลโรคไตเรื้อรัง	ค่าความถูกต้อง (%)	ค่าความคลาดเคลื่อนกำลังสองเฉลี่ย
วิธีความใกล้เคียงกันมากที่สุด	94.17	0.0579
วิธีต้นไม้ตัดสินใจ	**100.00**	**0.0059**
วิธีโครงข่ายประสาทเทียม	98.33	0.0115
วิธีซัพพอร์ตเวกเตอร์แมชชีน	98.33	0.0167
วิธีฐานกฎ	95.00	0.0454
วิธีถดถอยลอจิสติก	93.33	0.0660
วิธีนาอีฟ เบย์	95.00	0.0409



รูปที่ 8 การเปรียบเทียบค่าความถูกต้องและค่าความคลาดเคลื่อนกำลังสองเฉลี่ยของข้อมูลผู้ป่วยโรคไตเรื้อรัง



รูปที่ 9 แบบจำลองของกฎต้นไม้ตัดสินใจเพื่อใช้ในการทำนาย

จากตารางที่ 3 และรูปที่ 8 เมื่อพิจารณาจากค่าความถูกต้อง และค่าความคลาดเคลื่อนกำลังสองเฉลี่ยประกอบกัน พบว่าวิธีต้นไม้ตัดสินใจมีค่าความถูกต้องมากที่สุด คือ 100.00 % และยังมีค่าความคลาดเคลื่อนกำลังสองเฉลี่ยน้อยที่สุด คือ 0.0059 จึงสรุปว่าวิธีการจำแนกกลุ่มที่ดีที่สุดที่ใช้สำหรับจำแนกข้อมูลผู้ป่วยโรคไตเรื้อรังของโรงพยาบาลอพอลโล ประเทศอินเดีย คือ วิธีต้นไม้ตัดสินใจ

การนำไปใช้สำหรับต้นไม้ตัดสินใจ สามารถแปลงเป็นกฎได้ดังนี้

กฎข้อที่ 1 : ถ้า (ปริมาณครีเอตินินในเลือด (serum creatinine) <= 1.2 และ ความถ่วงจำเพาะของปัสสาวะ (urine specific gravity) <= 1.015)

สถานะ = เป็นโรคไตเรื้อรัง

กฎข้อที่ 2 : ถ้า (ปริมาณครีเอตินินในเลือด (serum creatinine) <= 1.2 และ ความถ่วงจำเพาะของปัสสาวะ (urine specific gravity) > 1.015 และ ระดับโปรตีนอัลบูมิน (albumin) <= 0 และ ปริมาณฮีโมโกลบิน (hemoglobin) <= 13.1 และ ปริมาณเม็ดเลือดแดงต่อปริมาณเลือดทั้งหมด (packed cell

volume) <= 43)

สถานะ = เป็นโรคไตเรื้อรัง

กฎข้อที่ 3 : ถ้า (ปริมาณครีเอตินินในเลือด (serum creatinine) <= 1.2 และ ความถ่วงจำเพาะของปัสสาวะ (urine specific gravity) > 1.015 และ ระดับโปรตีนอัลบูมิน (albumin) <= 0 และ ปริมาณฮีโมโกลบิน (hemoglobin) <= 13.1 และ ปริมาณเม็ดเลือดแดงต่อปริมาณเลือดทั้งหมด (packed cell volume) > 43)

สถานะ = ไม่เป็นโรคไตเรื้อรัง

กฎข้อที่ 4 : ถ้า (ปริมาณครีเอตินินในเลือด (serum creatinine) <= 1.2 และ ความถ่วงจำเพาะของปัสสาวะ (urine specific gravity) > 1.015 และ ระดับโปรตีนอัลบูมิน (albumin) <= 0 และ ปริมาณฮีโมโกลบิน (hemoglobin) > 13.1)

สถานะ = ไม่เป็นโรคไตเรื้อรัง

กฎข้อที่ 5 : ถ้า (ปริมาณครีเอตินินในเลือด (serum creatinine) <= 1.2 และ ความถ่วงจำเพาะของปัสสาวะ (urine specific gravity) > 1.015 และ ระดับโปรตีนอัลบูมิน (albumin) > 0)

สถานะ = เป็นโรคไตเรื้อรัง

กฎข้อที่ 6 : ถ้า (ปริมาณครีเอตินินในเลือด (serum creatinine) > 1.2)

สถานะ = เป็นโรคไตเรื้อรัง

5. สรุป

การทำวิจัยครั้งนี้มีวัตถุประสงค์เพื่อเปรียบเทียบประสิทธิภาพวิธีการจำแนกกลุ่มของผู้ป่วยโรคไตเรื้อรังด้วยวิธีความใกล้เคียงกันมากที่สุด วิธีต้นไม้ตัดสินใจ วิธีโครงข่ายประสาทเทียม วิธีซัพพอร์ตเวกเตอร์แมชชีน วิธีฐานกฎ วิธีการถดถอยลอจิสติก และวิธีนาอีฟ เบย์ เนื่องจากเป็นอัลกอริทึมที่นิยมนำมาใช้ในการจัดประเภทข้อมูลที่ทราบผลลัพธ์แน่นอน โดยใช้หลักการของการทำเหมืองข้อมูลมาใช้ในการจำแนกข้อมูล ซึ่งสรุปผลได้ดังนี้ วิธีความใกล้เคียงกันมากที่สุดมีค่าความถูกต้อง คือ 94.17 % และมีค่าความคลาดเคลื่อนกำลังสองเฉลี่ย คือ 0.0579 วิธีต้นไม้ตัดสินใจมีค่าความถูกต้อง คือ 100 % และมีค่าความคลาดเคลื่อนกำลังสองเฉลี่ย คือ 0.0059 วิธีโครงข่ายประสาทเทียมมีค่าความถูกต้อง คือ 98.33 % และมีค่าความคลาดเคลื่อนกำลังสองเฉลี่ย คือ 0.0115 วิธีซัพพอร์ตเวกเตอร์แมชชีนมีค่าความถูกต้อง คือ 98.33 % และมีค่าความคลาดเคลื่อนกำลังสองเฉลี่ย คือ 0.0167 วิธีฐานกฎมีค่าความถูกต้อง คือ 95 % และมีค่าความคลาดเคลื่อนกำลังสองเฉลี่ย คือ 0.0454 วิธีการถดถอยลอจิสติกมีค่าความถูกต้อง คือ 93.33 % และมีค่าความคลาดเคลื่อนกำลังสองเฉลี่ย คือ 0.0660 และวิธีนาอีฟเบย์มีค่าความถูกต้อง คือ 95 % และมีค่าความคลาดเคลื่อนกำลังสองเฉลี่ย คือ 0.0409 ซึ่งวิธีการจำแนกกลุ่มที่มีประสิทธิภาพในการจำแนกดีที่สุดสำหรับข้อมูลโรคไตเรื้อรัง คือ วิธีต้นไม้ตัดสินใจ โดยใช้อัลกอริทึมชนิด J48

การสรุปผลการวิจัยครั้งนี้ วิธีต้นไม้ตัดสินใจมีประสิทธิภาพในการจำแนกของโรคไตเรื้อรังมากที่สุด ดังนั้นผู้วิจัยจึงเลือกวิธีต้นไม้ตัดสินใจสำหรับการจำแนกกลุ่มผู้ป่วยโรคไตเรื้อรังโรงพยาบาลอพอลโล ประเทศอินเดีย เพราะว่ามีค่าความถูกต้องอยู่ในระดับสูงที่สุดและความคลาดเคลื่อนกำลังสองเฉลี่ยต่ำที่สุด ผลการสรุปดังกล่าวใกล้เคียงกับในช่วงที่ผ่านมาที่มีผู้วิจัยได้ศึกษาวิธีการจำแนกกลุ่มด้วยข้อมูลต่าง ๆ ซึ่งผู้วิจัยส่วนใหญ่ได้สรุปว่าวิธีต้นไม้ตัดสินใจเป็นวิธีที่มีประสิทธิภาพมากที่สุดในการจำแนกกลุ่ม และจากผลการวิจัย ผู้วิจัยจึงสรุปได้ว่าวิธีต้นไม้ตัดสินใจเป็นวิธีที่ดีที่สุด ซึ่งสามารถนำผลลัพธ์ที่ได้จากแบบจำลองของกฎต้นไม้ตัดสินใจเพื่อใช้ในการทำนาย เป็นแนวทางในการสนับสนุนการตัดสินใจทางการแพทย์เกี่ยวกับการ

วินิจฉัยโรคไตเรื้อรัง เพื่อให้ได้ผลการวินิจฉัยที่รวดเร็วและแม่นยำมากขึ้น จะทำให้กลุ่มผู้ป่วยโรคไตเรื้อรังลดน้อยลง

นอกจากนี้ งานวิจัยที่ผ่านมาได้มีการศึกษาเกี่ยวกับข้อมูลโรคไตเรื้อรัง โดยใช้วิธีซัพพอร์ทเวกเตอร์แมชชีนและโครงข่ายประสาทเทียม จากการศึกษาพบว่าวิธีโครงข่ายประสาทเทียมมีประสิทธิภาพในการทำนายผลได้ดีกว่าวิธีซัพพอร์ตเวกเตอร์แมชชีน [7] ซึ่งงานวิจัยดังกล่าวได้ให้ผลลัพธ์ที่สอดคล้องกัน เนื่องจากในงานวิจัยในครั้งนี้วิธีโครงข่ายประสาทเทียมมีประสิทธิภาพในการทำนายผลได้ดีกว่าวิธีซัพพอร์ตเวกเตอร์แมชชีนเช่นกัน แต่อาจมีวิธีการจำแนกกลุ่มอื่น ๆ ที่มีประสิทธิภาพในการวิเคราะห์ข้อมูลโรคไตเรื้อรังมากกว่าวิธีดังกล่าว ดังนั้นผู้วิจัยจึงได้ศึกษาวิธีการจำแนกกลุ่มด้วยวิธีอื่น ๆ ทั้งหมด 7 วิธี เพื่อหาวิธีที่ให้ประสิทธิภาพมากที่สุดในการวิเคราะห์ข้อมูลสำหรับโรคไตเรื้อรัง ทำให้ได้ผลการวิจัยที่แตกต่างกัน

6. รายการอ้างอิง

[1] อรนุช พันโท และมนชัย เทียนทอง, 2557, การเปรียบเทียบประสิทธิภาพการจำแนกรูปแบบการเรียนรู้ VARK ด้วยเทคนิคเหมืองข้อมูล, ว. เทคโนโยโลยีอุตสาหกรรม ม.ราชภัฏอุบลราชธานี 4(1): 1-11.

[2] บุษกร สุคนธวงศาโรจน์, 2551, อัลกอรึทึมในการจำแนกบุคลากรในองค์กรสำหรับการสร้างแผนที่ความรู้, วิทยานิพนธ์ปริญญาโท, มหาวิทยาลัยเกษตรศาสตร์, กรุงเทพฯ, 106 น.

[3] เบญจภัค จงหมื่นไวย์, 2558, การเปรียบเทียบปัจจัยโรคประจำตัวผู้สูงอายุโดยใช้อัลกอริทึมการจัดกลุ่ม J48 และ Naïve Bayes กรณีศึกษาสาธารณสุขโพธิ์กลางนครราชสีมา, ว.การประชุมวิชาการระดับชาติ การจัดการเทคโนโลยีและนวัตกรรม 1: 91-98.

[4] ชณัฐดาภรณ์ เย็นประเสริฐ, 2557, การเปรียบเทียบความแม่นยำการพยากรณ์สถานการณ์ชำระหนี้ของลูกหนี้ โดยใช้เทคนิคการถดถอยลอจิสติก นาอีฟย์เบย์ และต้นไม้การตัดสินใจ, สารนิพนธ์ปริญญาโท, มหาวิทยาลัยเทคโนโลยีพระจอมเกล้าพระนครเหนือ, กรุงเทพฯ, 79 น.

[5] พลอยพรรณ สอนสุวิทย์, 2552, การเปรียบเทียบประสิทธิภาพการตรวจจับสิ่งผิดปกติทางเครือข่ายชนิด, วิทยานิพนธ์ปริญญาโท, มหาวิทยาลัยเชียงใหม่, เชียงใหม่, 40 น.

[6] เดช ธรรมศิริ และพยุง มีสัจ, 2552, การจำแนกข้อมูลด้วยเทคนิคซัพพอร์ตเวกเตอร์แมชชีนโดยการปรับพารามิเตอร์และเลือกคุณลักษณะที่เหมาะสมด้วยขั้นตอนวิธีเชิงพันธุกรรม, ว. วิชาการพระจอมเกล้าพระนครเหนือ 21(2): 293-303.

[7] Vijayarani, S. and Dhayanand, S., 2015, Kidney disease prediction using SVM and annalgorithms, Int. J. Cybernet. Inform. 4(4): 13-25.

[8] รุจิรา ธรรมสมบัติ, 2554, ระบบสนันสนุนการตัดสินใจในการเลือกใช้แพคเกจอินเทอร์เน็ตมือถือ โดยใช้ต้นไม้ตัดสินใจ, รายงานวิจัย, สาขาคอมพิวเตอร์ธุรกิจ คณะบริหารธุรกิจ วิทยาลัยราชพฤกษ์, กรุงเทพฯ, 68 น.

[9] Troyanskaya, O., Cantor, M., Sherlock, G., Brown, P., Hastie, T., Tibshirani, R., Botstein, D. and Altman, R.B., 2001, Missing values estimation methods for DNA microarrays, Bioinformatics 17: 520-525.

[10] จารุมน หนูคง, 2552, การศึกษาเปรียบเทียบ

เทคนิควิธีการพยากรณ์ข้อมูลอนุกรมเวลาราคายางด้วยวิธีโครงข่ายประสาทเทียม สมการถดถอยแบบโพลิโนเมียล และซัพพอร์ตเวกเตอร์รีเกรสชัน, วิทยานิพนธ์ปริญญาโท, สถาบันเทคโนโลยีพระจอมเกล้าพระนครเหนือ, กรุงเทพฯ, 203 น.

[11] อานนท์ นามสนิท, 2549, การจำแนกกลุ่มเพลงไทยโดยใช้ซัพพอร์ตเวกเตอร์แมชชีน, วิทยานิพนธ์ปริญญาโท, สถาบันเทคโนโลยีพระจอมเกล้าพระนครเหนือ, กรุงเทพฯ, 53 น.

[12] Witten, I.H., Frank, E., Hall, M.A. and Pal, C.J., 2011, Data Mining: Practical Machine Learning Tools and Techniques, 3rd Ed., Morgan Kaufmann, Amsterdam, 629 p.

[13] วรรณสิริ ธุระชน, วรพจน์ สุเมธาวัฒนพงศ์ และณัฐวิภา ส่งสุข, 2557, ระบบการจำแนกพันธุ์ยางพาราโดยใช้ตัวจำแนกนาอีฟเบย์, น. 20-25, การประชุมทางวิชาการระดับชาติ ด้านคอมพิวเตอร์และเทคโนโลยีสารสนเทศ ครั้งที่ 10, มหาวิทยาลัยเทคโนโลยีพระจอมเกล้าพระนครเหนือ, กรุงเทพฯ.

[14] จุฑามาส สิทธิโชคสถาพร, 2555, พยากรณ์ราคายางแผ่นดิบโดยใช้ซัพพอร์ตเวกเตอร์แมชชีน, วิทยานิพนธ์ปริญญาโท, มหาวิทยาลัยสงขลานครินทร์, สงขลา, 78 น.

[15] Chronic Kidney Disease Data Set of India 2015, UCI Machine Learning Repository, Available Source: https://archive.ics.uci.edu/ml/datasets/Chronic_Kidney_Disease, March 10, 2016.

[16] Wu, X. and Kumar, V., 2009, The Top Ten Algorithms in Data Mining, University of Minnesota Department of Computer Science and Engineering, Minnesota, Minneapolis, 201 p.

[17] Berson, A. and Smith, S.J., 2001, Data Warehousing Data Mining and OLAP, McGraw-Hill, Boston, 612 p.

[18] วาทินี นุ้ยเพียร, ภรัณยา อำมฤครัตน์, เดช ธรรมศิริ ณรงค์ โพธิ และพยุง มีสัจ, 2553, การเปรียบเทียบประสิทธิภาพและวิเคราะห์การจำแนกข้อมูลด้วยวิธีการทางเครือข่ายประสาทเทียม, น. 131-138, การประชุมทางวิชาการระดับชาติ ด้านคอมพิวเตอร์และเทคโนโลยีสารสนเทศ ครั้งที่ 5, มหาวิทยาลัยเทคโนโลยีพระจอมเกล้าพระนครเหนือ, กรุงเทพฯ.