การเปรียบเทียบประสิทธิภาพในการทำนายผลค่านอกเกณฑ์ด้วยการจำแนก 6 วิธี

An Efficiency Comparison in Prediction of Outliers 6 Classifications

สายชล สินสมบูรณ์ทอง*

ภาควิชาสถิติ คณะวิทยาศาสตร์ สถาบันเทคโนโลยีพระจอมเกล้าเจ้าคุณทหารลาดกระบัง ถนนฉลองกรุง เขตลาดกระบัง กรุงเทพมหานคร 10520

Saichon Sinsomboonthong*

Department of Statistics, Faculty of Science, King Mongkut's Institute of Technology Ladkrabang, Chalongkrung Road, Ladkrabang, Bangkok 10520

Received: October 9, 2019; Accepted: November 1, 2019

บทคัดย่อ

การศึกษานี้เป็นการเปรียบเทียบประสิทธิภาพในการทำนายผลค่านอกเกณฑ์ด้วยการจำแนก 6 วิธี วิธีการจำแนกที่นำมาเปรียบเทียบ คือ วิธีเพื่อนบ้านใกล้สุด k ตัว วิธีโครงข่ายประสาทเทียม วิธีฐานกฎ วิธีการถดถอยลอจิสติกทวิภาค วิธีเพอร์เซปตรอนให้คะแนน และวิธีลาดลงสโตแคสติก โดยมีวัตถุประสงค์เพื่อเปรียบเทียบประสิทธิภาพวิธีการจำแนก 6 วิธี และเปรียบเทียบประสิทธิภาพของโปรแกรม SPSS, MINITAB และ WEKA การเปรียบเทียบประสิทธิภาพจะใช้ค่าความถูกต้อง ค่าคลาดเคลื่อนกำลังสองเฉลี่ย และค่าคลาดเคลื่อนสัมบูรณ์เฉลี่ย การเปรียบเทียบชุดข้อมูลที่มีค่านอกเกณฑ์อยู่ในระดับต่ำ (ร้อยละ 0-3) คือ การตรวจสอบธนบัตร วิธีที่มีประสิทธิภาพสูงสุด คือ วิธีลาดลงสโตแคสติกโดยการสุ่มตัวอย่างด้วยโปรแกรม WEKA ชุดข้อมูลที่มีค่านอกเกณฑ์อยู่ในระดับปานกลาง (ร้อยละ 3-6) คือ การชี้ตัวในเฟสบุ๊ค วิธีที่มีประสิทธิภาพสูงสุด คือ วิธีเพื่อนบ้านใกล้สุด k ตัว โดยการสุ่มตัวอย่างด้วยโปรแกรม WEKA ส่วนชุดข้อมูลที่มีค่านอกเกณฑ์อยู่ในระดับสูง (ร้อยละ 6-10) คือ การเลือกวิธีการคุมกำเนิด วิธีที่มีประสิทธิภาพสูงสุด คือ วิธีโครงข่ายประสาทเทียมโดยการสุ่มตัวอย่างด้วยโปรแกรม WEKA

คำสำคัญ : วิธีเพื่อนบ้านใกล้สุด k ตัว; วิธีโครงข่ายประสาทเทียม; วิธีฐานกฎ; วิธีการถดถอยลอจิสติกทวิภาค; วิธีเพอร์เซปตรอนให้คะแนน; วิธีลาดลงสโตแคสติก

Abstract

In this study, an efficiency comparison in prediction of outliers 6 classifications were determined. The classification methods were compared the followings: (1) k-nearest neighbor method, (2) artificial

**Corresponding author: saichon.ss49@gmail.com* DOI: 10.14456/tjst.2020.16

neural network method, (3) rule-based method, (4) binary logistic regression method, (5) voted perceptron method, and (6) stochastic gradient descent method. The purposes were to compare the efficiency of 6 classifications, and to compare SPSS, MINITAB and WEKA programs. The following efficiency comparison values were employed, i.e. accuracy, mean square error (MSE), and mean absolute error (MAE). For the low outliers data set (0-3 percentage), banknote authentication, the best classification method was the stochastic gradient descent method in combination with the WEKA sampling method. The middle outliers data set (3-6 percentage), Facebook metrics, the best classification method was the k nearest neighbor method in combination with the WEKA sampling method. For the high outliers data set (6-10 percentage), contraceptive method choice, the best classification method was the artificial neural network method in combination with the WEKA sampling method.

Keywords: k-nearest neighbor; artificial neural network; rule-based; binary logistic regression; voted perceptron; stochastic gradient descent

1. คำนำ

ปัจจุบันข้อมูลที่มีคุณภาพและน่าเชื่อถือมีความสำคัญเป็นอย่างมากในการวิเคราะห์ข้อมูลเพื่อการนำข้อมูลไปใช้ประโยชน์ได้สูงสุด บางครั้งจากข้อมูลจะพบว่าข้อมูลมีค่าที่มากเกินไปหรือน้อยเกินไปแฝงอยู่ ซึ่งเรียกว่าค่านอกเกณฑ์ (outlier) เป็นค่าที่อยู่ปลายสุดซึ่งตกอยู่ใกล้กับขีดจำกัดของพิสัยข้อมูล การหาค่านอกเกณฑ์มีความสำคัญเนื่องจากแสดงค่าความคลาดเคลื่อนในข้อมูล หากนำข้อมูลที่มีค่านอกเกณฑ์ไปวิเคราะห์ จะส่งผลให้เกิดความคลาดเคลื่อนของผลลัพธ์ข้อมูลที่ได้ การกระจายของข้อมูลและค่าเฉลี่ยของข้อมูลไม่ดี ส่งผลให้ไม่เป็นไปตามข้อกำหนดเบื้องต้น (assumption) และทำให้ไม่สามารถนำข้อมูลไปใช้ประโยชน์ได้อย่างสูงสุด สำหรับสาเหตุที่ทำให้เกิดค่านอกเกณฑ์ความคลาดเคลื่อนจากการแปรผันข้อมูลที่เก็บรวบรวมมา ซึ่งเป็นความคลาดเคลื่อนที่ไม่สามารถควบคุมได้ ความคลาดเคลื่อนที่เกิดจากเครื่องมือที่ใช้วัดมีคุณภาพต่ำ ทำให้เกิดค่านอกเกณฑ์ ความคลาดเคลื่อนที่เกิดจากการบันทึกข้อมูลจากการปฏิบัติโดยไม่ตรวจสอบให้ถี่ถ้วน เพื่อให้ได้ผลการวิเคราะห์ที่เชื่อถือได้ การตรวจสอบหาค่านอกเกณฑ์จึงเป็นสิ่งที่สำคัญก่อนการนำข้อมูลไปวิเคราะห์ เพื่อให้ได้เห็นข้อมูลที่มีความผิดปกติและสามารถแก้ไขได้ บางครั้งข้อมูลอาจไม่สามารถใช้งานอย่างเต็มประสิทธิภาพหรือตรงตามความต้องการ จึงมีการจำแนกข้อมูลเพื่อให้ได้วิธีที่มีความเหมาะสมกับข้อมูลที่มีความแตกต่างกันไป ดังนั้นผู้วิจัยจึงต้องเลือกวิธีการจำแนกให้เหมาะสมกับข้อมูล (วรพรรณ, 2556)

การศึกษางานวิจัยที่เกี่ยวข้องเกี่ยวกับโรคมะเร็งเต้านมเรื่องการค้นหาวิธีการทำเหมืองข้อมูลเพื่อสร้างตัวแบบการวิเคราะห์โรคอัตโนมัติ โดยงานวิจัยนี้มุ่งเน้นการค้นหาวิธีการทำเหมืองข้อมูลเพื่อสร้างตัวแบบการวิเคราะห์โรคอัตโนมัติเพื่อค้นหาขั้นตอนวิธีที่เหมาะสมที่สุดสำหรับฐานข้อมูลทางการแพทย์ โดยวัดประสิทธิภาพจากค่าความถูกต้อง พบว่าวิธีต้นไม้ตัดสินใจให้ประสิทธิภาพสูงสุด คือ ร้อยละ 75.52 เมื่อเปรียบเทียบกับวิธีซัพพอร์ตเวกเตอร์แมชชีนและวิธีเพื่อนบ้านใกล้สุด k ตัว (นิเวศ, 2553) ซึ่งให้ผลสอดคล้องกับ ณัฐวุฒิ (2559) ที่ศึกษางานวิจัยเกี่ยวกับโรคมะเร็งเรื่อง

การเปรียบเทียบประสิทธิภาพขั้นตอนวิธีการทำเหมืองข้อมูลเพื่อวิเคราะห์ปัจจัยที่ส่งผลต่อการเกิดโรคมะเร็ง โดยงานวิจัยนี้นำวิธีการทำเหมืองข้อมูลมาประยุกต์ใช้กับการตรวจวิเคราะห์การเกิดโรคมะเร็ง เพื่อนำกฎการจำแนกข้อมูลที่ได้ไปพัฒนาเป็นระบบตรวจวิเคราะห์ปัจจัยที่ส่งผลต่อการเกิดโรคมะเร็งโดยวัดประสิทธิภาพจากค่าสัมบูรณ์ของความคลาดเคลื่อนเฉลี่ย พบว่าวิธีต้นไม้ตัดสินใจมีประสิทธิภาพสูงสุด คือ ร้อยละ 98.63 เมื่อเปรียบเทียบกับวิธีเพื่อนบ้านใกล้สุด k ตัว และวิธีนาอีฟเบส์

ส่วนงานวิจัยเกี่ยวกับโรคเบาหวานเรื่องวิธีซัพพอร์ตเวกเตอร์แมชชีนและวิธีโครงข่ายประสาทเทียม การวินิจฉัยโรคเบาหวานในจอประสาทตาซึ่งเป็นโรคที่มีสาเหตุเกิดจากโรคแทรกซ้อนของโรคเบาหวาน โดยจำแนกเป็น 2 กลุ่ม คือ ผู้ที่เป็นโรคเบาหวานในจอประสาทตาและไม่เป็นโรคเบาหวานในจอประสาทตา พบว่าวิธีซัพพอร์ตเวกเตอร์แมชชีนให้ค่าความถูกต้องร้อยละ 97.61 มากกว่าวิธีโครงข่ายประสาทเทียม (Priya and Aruna, 2012) แต่ให้ผลไม่สอดคล้องกับ Sa-nga soongsong และ Chongwatpol (2012) ซึ่งศึกษาเรื่องปัจจัยเสี่ยงการเป็นโรคเบาหวานด้วยวิธีการทำเหมืองข้อมูล ซึ่งแบ่งกลุ่มคนไข้เป็น 3 กลุ่ม โดยพิจารณาจากค่าใช้จ่าย การวิเคราะห์เบื้องต้นพบว่าความดันเลือดสูง อายุ คลอเรสเตอรอล ดัชนีมวลกาย รายได้ทั้งหมด เพศ การเป็นโรคหัวใจ สถานภาพสมรส การตรวจฟัน และการวินิจฉัยโรคหอบหืด เป็นปัจจัยเสี่ยงที่สำคัญ ถ้าพิจารณาตัวแบบทั้งหมดพบว่าวิธีการถดถอยลอจิสติกให้ผลอัตราการจำแนกผิดทั้งหมดดีที่สุดร้อยละ 22.89 แม้ว่าวิธีโครงข่ายประสาทเทียมมีอัตราความผิดพลาดเชิงลบต่ำสุด คือ ร้อยละ 20.55 ก็ยังคงเลือกตัวแบบการถดถอยลอจิสติกเป็นตัวแบบสุดท้ายเพื่อทำนายคนไข้ที่เป็นโรคเบาหวาน และให้ผลไม่สอดคล้องกับกิตติพล และคณะ (2552) ที่ศึกษาเกี่ยวกับโรคเบาหวานเรื่องการวิเคราะห์ปัจจัยเสี่ยงของโรคเบาหวานเพื่อนำมาเป็นเครื่องมือประเมินการเกิดโรคเบาหวานโดยไม่ต้องอาศัยการตรวจเลือด พบว่าวิธีโครงข่ายประสาทเทียมแบบแพร่กระจายย้อนกลับให้ค่าความถูกต้องสูงสุด เมื่อเปรียบเทียบกับวิธีโครงข่ายประสาทเทียมแบบธรรมดาและวิธีนาอีฟเบส์

การศึกษางานวิจัยเกี่ยวกับการให้คะแนนสินเชื่อโดยวิธีการทำเหมืองข้อมูลด้วยวิธีซัพพอร์ตเวกเตอร์แมชชีน รวมทั้งการเลือกใช้ลักษณะที่เหมาะสมร่วมกับการหาค่าพารามิเตอร์ที่เหมาะสมด้วยวิธีค้นหาแบบกริช เพื่อลดความเสี่ยงสำหรับการให้เครดิตสินเชื่อแก่ลูกค้าที่มีความเสี่ยงในการผิดสัญญาหรือขาดการชำระเงินในการหาตัวแบบที่เหมาะสม พบว่าวิธีต้นไม้ตัดสินใจมีค่าความถูกต้องสูงสุด คือ ร้อยละ 79.48 เมื่อเปรียบเทียบกับวิธีโครงข่ายประสาทเทียมแบบแพร่กระจายย้อนกลับและวิธีซัพพอร์ตเวกเตอร์แมชชีน (เดช และคณะ, 2552) ซึ่งให้ผลสอดคล้องกับ ทิพย์ธิดา (2555) ที่ศึกษางานวิจัยเกี่ยวกับสินเชื่อเรื่องการใช้เหมืองข้อมูลช่วยในการตัดสินใจการให้สินเชื่อ เพื่อเป็นแนวทางการสนับสนุนการตัดสินใจการอนุมัติสินเชื่อของบริษัทได้อย่างมีประสิทธิภาพมากขึ้น ซึ่งมีส่วนช่วยลดปริมาณหนี้สินสูญได้ พบว่าวิธีต้นไม้ตัดสินใจให้ค่าความถูกต้องสูงสุด คือ ร้อยละ 90.47 มากกว่าวิธีนาอีฟเบส์ แต่ให้ผลตรงข้ามกับ เดช และพยุง (2553) ซึ่งนำเสนอการจำแนกประเภทโดยทำการทดสอบประสิทธิภาพด้วยข้อมูล Austrian credit และ Bankrupt data โดยนำเอาวิธีซัพพอร์ตเวกเตอร์แมชชีนและหาค่าพารามิเตอร์ที่เหมาะสมร่วมด้วยการเลือกคุณลักษณะที่เหมาะสมโดยใช้ขั้นตอนวิธีเชิงพันธุกรรมเปรียบเทียบผลการวิจัยกับวิธีต้นไม้ตัดสินใจ วิธีโครงข่ายประสาทเทียม วิธีซัพพอร์ตเวกเตอร์แมชชีนกับวิธีเชิงพันธุกรรม พบว่าวิธีซัพ

พอร์ตเวกเตอร์แมชชีนที่ใช้ขั้นตอนวิธีเชิงพันธุกรรมจะให้ค่าความแม่นยำสูงสุด

นอกจากนี้การศึกษาการเปรียบเทียบประสิทธิภาพในการจำแนกเมื่อข้อมูลมีค่านอกเกณฑ์ การทำเหมืองข้อมูลพบว่าชุดข้อมูลโรคมะเร็งเต้านมของรัฐวิสคอนซินที่มีค่านอกเกณฑ์อยู่ในระดับต่ำ ที่ random seed 10, 20 วิธีโครงข่ายประสาทเทียมโดยใช้โปรแกรม SPSS ให้ค่าความถูกต้องสูงสุด คือ ร้อยละ 100 ค่าความคลาดเคลื่อนกำลังสองเฉลี่ยและค่าส่วนเบี่ยงเบนสัมบูรณ์เฉลี่ยต่ำสุด คือ 0 และ 0.0013 ตามลำดับ และ random seed 30 วิธีต้นไม้ตัดสินใจโดยใช้โปรแกรม WEKA และวิธีซัพพอร์ตเวกเตอร์แมชชีนโดยใช้โปรแกรม SPSS ให้ค่าความถูกต้องสูงสุดเท่ากันคือร้อยละ 100 ค่าคลาดเคลื่อนกำลังสองเฉลี่ยและค่าส่วนเบี่ยงเบนสัมบูรณ์เฉลี่ยต่ำสุดเท่ากัน คือ 0 และ 0 ตามลำดับ ส่วนชุดข้อมูลโรคเบาหวานของชาวพิม่าประเทศอินเดีย ที่มีค่านอกเกณฑ์อยู่ในระดับปานกลาง และชุดข้อมูลการชำระเงินด้วยบัตรเครดิตของลูกค้าที่มีค่านอกเกณฑ์อยู่ในระดับสูงให้ผลสอดคล้องกัน คือ วิธีเพื่อนบ้านใกล้สุด k ตัว โดยใช้โปรแกรม SPSS และ WEKA ให้ค่าความถูกต้องสูงสุดเท่ากัน คือ ร้อยละ 100 ส่วนค่าความคลาดเคลื่อนกำลังสองเฉลี่ยต่ำสุด คือ 0.00016129 และ 0.00022201 และค่าส่วนเบี่ยงเบนสัมบูรณ์เฉลี่ยต่ำสุด คือ 0.0127 และ 0.0149 ตามลำดับ (พนิดา และคณะ, 2560)

ดังนั้นผู้วิจัยจึงให้ความสนใจในการเปรียบเทียบประสิทธิภาพในการทำนายผลค่านอกเกณฑ์ด้วยการจำแนก 6 วิธี คือ วิธีเพื่อนบ้านใกล้สุด k ตัว (k-nearest neighbor) วิธีโครงข่ายประสาทเทียม (artificial neural network) วิธีฐานกฎ (rule based) วิธีการถดถอยลอจิสติกทวิภาค (binary logistic regression) วิธีเพอร์เซปตรอนให้คะแนน (voted perceptron) และวิธีลาดลงสโตแคสติก (stochastic gradient descent) เพื่อต้องการเปรียบเทียบประสิทธิภาพวิธีการจำแนกทั้ง 6 วิธี ว่าวิธีใดมีประสิทธิภาพและเหมาะสมกับรูปแบบของชุดข้อมูลระหว่างโปรแกรม SPSS, MINITAB และ WEKA เพื่อเปรียบเทียบว่าโปรแกรมใดให้ค่าความถูกต้อง (accuracy) ในการทำนายสูงสุด ส่วนค่าคลาดเคลื่อนกำลังสองเฉลี่ย (mean square error, MSE) และค่าคลาดเคลื่อนสัมบูรณ์เฉลี่ย (mean absolute error, MAE) ต่ำสุด

2. วิธีการวิจัย

2.1 เครื่องมือที่ใช้ในการวิจัย

โปรแกรมที่ใช้ในการวิจัยครั้งนี้ คือ WEKA (Waikato Environment for Knowledge Analysis) เวอร์ชั่น 3.9.2 โปรแกรม SPSS เวอร์ชั่น 25 และโปรแกรม MINITAB เวอร์ชั่น 18

2.2 การเก็บรวบรวมข้อมูล การแบ่งข้อมูล การศึกษาขั้นตอนวิธี และการเปรียบเทียบประสิทธิภาพของวิธีการจำแนก

2.2.1 การเก็บรวบรวมข้อมูล ศึกษาข้อมูลที่มีค่านอกเกณฑ์จากเว็บไซต์ UCI โดยแบ่งข้อมูลค่านอกเกณฑ์เป็น 3 ระดับ คือ ค่านอกเกณฑ์ระดับต่ำ (ร้อยละ 0-3) ค่านอกเกณฑ์ระดับ ปานกลาง (ร้อยละ 3-6) และค่านอกเกณฑ์ระดับสูง (ร้อยละ 6-10) โดยได้ข้อมูล 3 ชุด คือ

(1) การตรวจสอบธนบัตร (banknote authentication) มีจำนวนข้อมูลทั้งหมด 603 ค่า พบค่านอกเกณฑ์ 7 ค่า คิดเป็นร้อยละ 1.16 จัดเป็นชุดข้อมูลที่มีค่านอกเกณฑ์ระดับต่ำ โดยตัวแปรอิสระประกอบด้วยอายุของธนบัตร ความสมบูรณ์ของธนบัตร การปนเปื้อนสารเสพติด และธนบัตรเสีย ส่วนตัวแปรตามคือการอนุมัติให้แลกเปลี่ยนธนบัตร (ไม่อนุมัติ/อนุมัติ)

(2) การชี้ตัวในเฟสบุ๊ค (Facebook metric) มีจำนวนข้อมูลทั้งหมด 500 ค่า พบค่านอก

เกณฑ์ 17 ค่า คิดเป็นร้อยละ 3.40 จัดเป็นชุดข้อมูลที่มีค่านอกเกณฑ์ระดับปานกลาง โดยตัวแปรอิสระประกอบด้วยเพศ อายุ ชั่วโมงใช้งานเฉลี่ยต่อวัน จำนวนโพสต์สเตตัส จำนวนการโพสต์วีดีโอ จำนวนการโพสต์รูป และจำนวนเพื่อนในเฟสบุ๊ค ส่วนตัวแปรตามคือผู้ใช้งานที่สามารถยืนยันตัวตนได้ (ไม่สามารถยืนยันตัวตนได้/สามารถยืนยันตัวตนได้)

(3) การเลือกวิธีการคุมกำเนิด (contraceptive method choice) มีจำนวนข้อมูลทั้งหมด 651 ค่า พบค่านอกเกณฑ์ 41 ค่า คิดเป็นร้อยละ 6.30 จัดเป็นชุดข้อมูลที่มีค่านอกเกณฑ์ระดับสูง โดยตัวแปรอิสระประกอบด้วยอายุของภรรยา อายุของสามี ระดับการศึกษาของภรรยา (ปริญญาตรี/ไม่ใช่ปริญญาตรี) ระดับการศึกษาของสามี (ปริญญาตรี/ไม่ใช่ปริญญาตรี) ศาสนาของภรรยา (อิสลาม/ไม่ใช่อิสลาม) สถานะการทำงานของภรรยา (ทำงาน/ไม่ทำงาน) และสถานะการทำงานของสามี (ทำงาน/ไม่ทำงาน) ส่วนตัวแปรตามคือการคุมกำเนิด (ไม่คุมกำเนิด/คุมกำเนิด)

2.2.2 การแบ่งข้อมูล

แบ่งชุดข้อมูลโดยโปรแกรม SPSS เวอร์ชั่น 25 โปรแกรม MINITAB เวอร์ชั่น 18 และโปรแกรม WEKA เวอร์ชั่น 3.9.2 ด้วยวิธีการสุ่มตัวอย่างอย่างง่าย (simple random sampling, SRS) สุ่มจำนวน 5 รอบ โดยการกำหนดตัวเลขสุ่มเทียม (random seed) เป็น 10, 20, 30, 40 และ 50 ตามลำดับ อัตราส่วน 70 : 20 : 10 (พยูน, 2548) ส่วนที่ 1 ข้อมูลเรียนรู้ (training data) นำไปสร้างตัวแบบ (model) ร้อยละ 70 ข้อมูลส่วนที่ 2 ข้อมูลตรวจสอบความถูกต้อง (validation data) นำไปประเมินความผิดพลาดของตัวแบบร้อยละ 20 และข้อมูลส่วนที่ 3 ข้อมูลทดสอบ (testing data) นำไปทดสอบตัวแบบร้อยละ 10

2.2.3 การศึกษาขั้นตอนวิธี (algorithm)

(1) วิธีเพื่อนบ้านใกล้สุด k ตัว เป็นวิธีการที่ได้รับความนิยมอย่างมาก เนื่องจากเป็นวิธีการที่ง่ายและมีประสิทธิภาพ ซึ่งสามารถนำไปประยุกต์ใช้กับงานได้อย่างหลากหลาย เช่น งานด้านการจำแนก รวมถึงงานทางด้านการแทนที่ข้อมูลที่สูญหาย ใช้ขั้นตอนวิธี IBK ซึ่งมีวิธีการดำเนินการดังนี้ (Troyanskaya, *et al*., 2001)

(1.1) กำหนดค่า k เพื่อใช้พิจารณาสมาชิกที่อยู่ใกล้กันมากที่สุด เช่น k=3 จะพิจารณาเฉพาะข้อมูล 3 ตัวแรก ที่อยู่ใกล้กับจุดที่ต้องการทำนาย

(1.2) คำนวณหาระยะห่างระหว่างข้อมูลตัวอย่างที่สนใจกับข้อมูลอื่น ๆ ทุกตัวด้วยระยะห่างยูคลิเดียน (Euclidian distance) จากสมการที่ 1 ดังนี้

$$D_{\text{Euclidian}}(x_i, y_i) = \sqrt{\sum_{k=1}^{n}(x_i, y_i)^2} \quad (1)$$

โดยที่ $D_{\text{Euclidian}}(x_i, y_i)$ คือ ระยะห่างระหว่างตัวอย่าง x_i กับตัวอย่าง y_i ; k คือ คุณลักษณะทั้งหมดของตัวอย่าง

(1.3) เลือกค่าข้อมูลที่มีค่าระยะห่างน้อยที่สุด k ตัว เพื่อนำมาพิจารณาหาคำตอบ

(2) วิธีโครงข่ายประสาทเทียม ใช้ขั้นตอนวิธีชนิดเพอร์เซปตรอนหลายชั้น โดยกำหนดค่าอัตราการเรียนรู้เป็น 0.1 ค่าโมเมนตัมเป็น 0.9 จำนวนรอบการสอน 20,000 รอบ การวิจัยครั้งนี้ใช้ขั้นตอนวิธีของวิธีโครงข่ายประสาทเทียมชนิดเพอร์เซปตรอนหลายชั้นที่มีชั้นซ่อน 1 ชั้น การเชื่อมโยงกันระหว่างเซลล์ประสาทโดยทั่วไปนิยมใช้การเชื่อมโยงแบบแพร่ย้อนกลับ(back-propagation) ซึ่งเป็นขั้นตอนที่ใช้สอนโครงข่ายประสาทเทียมแบบเพอร์เซปตรอนหลายชั้น โดยแบบจำลองโครงข่ายประสาทเทียมมีการเชื่อมโยงกันเป็นโครงข่ายแบบเป็นชั้น ๆ โครงข่ายชนิดนี้มีการเชื่อมโยงกัน 3 ชั้น ประกอบด้วยชั้นข้อมูลเข้า (input layer) ถัดมาเป็น

ชั้นซ่อน (hidden layer) และชั้นสุดท้าย คือ ชั้นข้อมูลออก (output layer) (Berson and Smith, 1997) โดยส่วนประกอบที่ถูกบรรจุอยู่ในเซลล์ประสาทแต่ละตัวประกอบด้วย 2 ฟังก์ชันย่อย คือ ฟังก์ชันผลรวม (summation function) และฟังก์ชันกระตุ้น (activation function)

(2.1) ฟังก์ชันผลรวม ทำหน้าที่ในการคำนวณผลรวมของข้อมูลที่ได้จากชั้นข้อมูลเข้า ซึ่งคำนวณได้ดังสมการที่ 2 (Hagan *et al*., 1996) กำหนดให้ตัวแปร x คือ ค่าข้อมูลเข้าตัวที่ i; ตัวแปร w คือ ค่าถ่วงน้ำหนักของข้อมูลเข้าตัวที่ i; ตัวแปร g คือ ข้อมูลออกจากฟังก์ชันผลรวม; ตัวแปร z คือ จำนวนเซลล์ประสาทของข้อมูลเข้า; ตัวแปร β คือ ค่าความเอนเอียง (bias)

$$g = \sum_{i=1}^{z} x_i w_i + \beta \qquad (2)$$

(2.2) ฟังก์ชันกระตุ้น ทำหน้าที่ปรับเปลี่ยนค่าของข้อมูลที่ได้จากฟังก์ชันผลรวมให้อยู่ในช่วงที่ต้องการ ฟังก์ชันกระตุ้นที่นิยม ได้แก่ ฟังก์ชันเชิงเส้น (linear function) ฟังก์ชันซิกมอยด์ (sigmoid function) และฟังก์ชันไฮเปอร์โบลิกแทนเจนต์ (hyperbolic tangent function) (ธนาวุฒิ, 2552)

(3) วิธีฐานกฎ (rule-based) ใช้ชุดลำดับของกฎมาสร้างรูปแบบการแยกประเภทข้อมูล โดยส่วนใหญ่แล้วจะใช้กฎที่เป็น If ... then ซึ่งเป็นกฎอย่างง่าย ใช้ขั้นตอนวิธี decision table เป็นเครื่องมือที่ใช้แสดงเงื่อนไขการตัดสินใจและเลือกการทำงานหรือกระทำกิจกรรมภายใต้เหตุการณ์ของเงื่อนไขที่ระบุ วิธีการตัดสินใจแบบ decision table จะเป็นตาราง 2 มิติ วิธีฐานกฎเป็นวิธีหนึ่งที่นิยมใช้เช่นเดียวกับวิธีต้นไม้ตัดสินใจ ข้อกำหนดหรือเงื่อนไข (antecedent or precondition) ของวิธีฐานกฎเป็นการทดสอบคล้ายกับการทดสอบที่โหนดของวิธีต้นไม้ตัดสินใจ แต่ผลของการทดสอบหรือผลลัพธ์ (consequent or conclusion) ที่ได้นั้นจะให้คำตอบ (class) ที่ใช้กับตัวอย่างที่อยู่ภายใต้กฎนั้น หรือบางครั้งก็จะให้ค่าการแจกแจงความน่าจะเป็นของคำตอบต่าง ๆ (Murti and Mahantappa, 2012)

(4) วิธีการถดถอยลอจิสติกทวิภาค (binary logistic regression method) เป็นการวิเคราะห์การถดถอยแบบหนึ่งโดยที่ตัวแปรตามเป็นตัวแปรเชิงคุณภาพมีค่าได้เพียง 2 ค่า ส่วนตัวแปรอิสระอาจเป็นตัวแปรเชิงปริมาณหรือเชิงคุณภาพหรืออาจมีทั้งตัวแปรเชิงปริมาณและตัวแปรเชิงคุณภาพก็ได้ ซึ่งจะพบว่าลักษณะของตัวแปรอิสระและตัวแปรตามข้างต้นเหมือนกับวิธีการจำแนกประเภท (discriminant method) กรณีที่ตัวแปรตามมีเพียง 2 กลุ่ม โดยวิธีการถดถอยลอจิสติกทวิภาคไม่มีเงื่อนไขเกี่ยวกับการแจกแจงของตัวแปรอิสระและไม่มีเงื่อนไขเกี่ยวกับเมทริกซ์ความแปรปรวนและความแปรปรวนร่วมของแต่ละกลุ่ม และวิธีการถดถอยลอจิสติกทวิภาคเป็นการพยากรณ์โอกาสที่แต่ละหน่วยจะอยู่ในกลุ่มใดกลุ่มหนึ่ง (กัลยา, 2552)

(5) วิธีเพอร์เซปตรอนให้คะแนน (voted perceptron method) ขั้นตอนวิธีนี้มีข้อดี คือ ข้อมูลสามารถแบ่งแยกเชิงเส้นด้วยข้อมูลที่อยู่ริมที่มีค่ามาก วิธีนี้ง่ายในการดำเนินการและมีประสิทธิภาพมากในด้านระยะเวลาในการคำนวณเมื่อเปรียบเทียบกับวิธีซัพพอร์ทเวกเตอร์แมชชีน ขั้นตอนวิธีนี้สามารถใช้ในสเปซมิติสูงโดยใช้ฟังก์ชันเคอร์เนล ขั้นตอนวิธีเพอร์เซปตรอนให้คะแนนสามารถใช้แทนที่ข้อมูลสูญหายและแปลงคุณลักษณะเชิงกลุ่มให้เป็นทวิภาค และช่วยทำนายผลลัพธ์ทวิภาค (Freund and Schapire, 1998) ในขั้นตอนวิธีเพอร์เซปตรอนให้คะแนน ข้อมูลจำนวนมากสามารถเก็บไว้ในระหว่างฝึกหัด แล้วใช้ข้อมูลนี้เพื่อสร้างการทำนายที่ดีขึ้นกับชุดข้อมูลทดสอบ (Singh and Bansal, 2013)

(6) วิธีลาดลงสโตแคสติก (stochastic

gradient descent method) เป็นวิธีที่มีประสิทธิภาพในการจำแนกการเรียนรู้ของตัวจำแนกเชิงเส้นภายใต้ฟังก์ชันการสูญเสียโค้งนูน ได้แก่ ซัพพอร์ทเวกเตอร์แมชชีนและการถดถอยลอจิสติก (LeCun *et al*., 1998) ใช้การเรียนรู้ตัวแบบเชิงเส้นต่าง ๆ ได้แก่ ซัพพอร์ทเวกเตอร์แมชชีนที่มีคำตอบเป็นทวิภาค การถดถอยลอจิสติกที่มีคำตอบเป็นทวิภาค และการถดถอยเชิงเส้น แทนที่ข้อมูลสูญหายและแปลงคุณลักษณะเชิงกลุ่มให้เป็นทวิภาค แปลงคุณลักษณะต่าง ๆ ให้อยู่ในรูปปรกติ (normalization) ดังนั้นค่าสัมประสิทธิ์ของผลลัพธ์เป็นข้อมูลที่อยู่ในรูปปรกติ (Nektarios, 2013) สำหรับคุณลักษณะที่มีคำตอบเป็นนามบัญญัติจะใช้ฟังก์ชันสูญเสียไฮนด์ (hinge loss function) หรือฟังก์ชันการสูญเสียล็อก (log loss function) ส่วนคุณลักษณะที่มีคำตอบเป็นเชิงตัวเลขจะใช้ฟังก์ชันการสูญเสียกำลังสอง (squared loss function) ฟังก์ชันการสูญเสียเอฟซิลอน (epsilon-insensitive loss function) หรือฟังก์ชันการสูญเสียฮูเบอร์ (Huber loss function)

หลังจากนั้นนำข้อมูลที่แบ่งเป็น 3 ส่วน มาวิเคราะห์โดยใช้โปรแกรม WEKA ซึ่งวิเคราะห์จากวิธีการจำแนกทั้ง 6 วิธี ข้างต้น

2.2.4 การเปรียบเทียบประสิทธิภาพของวิธีการจำแนก (classification)

นำผลการวิเคราะห์ของแต่ละวิธีทั้ง 6 วิธี มาเปรียบเทียบประสิทธิภาพโดยพิจารณาจากเมทริกซ์ความสับสน

เมทริกซ์ความสับสน (confusion matrix) เป็นรูปแบบตารางที่เฉพาะเจาะจงที่นำผลลัพธ์จากการทำนายมาใส่ในรูปตารางเมทริกซ์ ซึ่งจะช่วยให้ง่ายต่อการมองเห็นค่าทำนายของขั้นตอนวิธีดังตารางที่ 1

(1) ค่าความถูกต้อง (accuracy) ในการทำนาย คือ การแสดงการวัดที่ได้มีความถูกต้องในรูปอัตราส่วนโดยคิดเป็นร้อยละ (สุรวัชร และสายชล, 2560)

Accuracy = (จำนวนข้อมูลที่จำแนกถูกว่าเป็นชั้น A และ B ÷ จำนวนข้อมูลทั้งหมด) x 100 %

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

(2) ค่าคลาดเคลื่อนกำลังสองเฉลี่ย (mean square error, MSE) เป็นมาตรวัดการประเมินค่าได้ดี เนื่องจากค่าคลาดเคลื่อนกำลังสองเฉลี่ยประกอบด้วยความเอนเอียงและความแปรปรวน (พนิดา และคณะ, 2560)

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

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

(3) ค่าคลาดเคลื่อนสัมบูรณ์เฉลี่ย (mean absolute error, MAE) คือ ค่าวัดความถูกต้องของการพยากรณ์ที่วัดจากค่าความคลาดเคลื่อนโดยไม่คำนึงถึงทิศทางของความคลาดเคลื่อน MAE มีหน่วยวัดหน่วยเดียวกับค่าสังเกต (สายชล, 2560)

$$MAE = \frac{\sum_{i=1}^{n}\left|y_i - \hat{y}_i\right|}{n}$$

3. ผลการวิจัย

งานวิจัยครั้งนี้ผู้วิจัยใช้การจำแนกโดยใช้วิธีการทำเหมืองข้อมูล โดยนำชุดข้อมูลที่ค้นคว้าจำนวน 3 ชุด มาวิเคราะห์ข้อมูล ซึ่งสุ่มแบ่งข้อมูลเป็น 3 ส่วน คือ ส่วนที่ 1 ข้อมูลเรียนรู้ นำไปสร้างตัวแบบ ร้อยละ 70 ข้อมูลส่วนที่ 2 ข้อมูลตรวจสอบความถูกต้อง นำไปประเมินความผิดพลาดของตัวแบบ ร้อยละ 20 และข้อมูลส่วนที่ 3 ข้อมูลทดสอบนำไปทดสอบตัวแบบ ร้อยละ 10 และผู้วิจัยได้นำมา

ตารางที่ 1 เมทริกซ์ความสับสน

		ผลลัพธ์จากสมการหรือการทดสอบ	
		คำตอบเป็นบวก	คำตอบที่เป็นลบ
ผลลัพธ์ที่เกิดขึ้นจริง	คำตอบเป็นบวก	TP (true positive)	FN (false negative)
	คำตอบที่เป็นลบ	FP (false positive)	TN (true negative)

โดยที่ บวกจริง (true positive, TP) คือ จำนวนข้อมูลที่จำแนกถูกว่าเป็นบวก ซึ่งค่าที่แท้จริงเป็นบวก; ลบจริง (true negative, TN) คือ จำนวนข้อมูลที่จำแนกถูกว่าเป็นลบ ซึ่งค่าที่แท้จริงเป็นลบ; บวกเท็จ (false positive, FP) คือ จำนวนข้อมูลที่จำแนกผิดว่าเป็นบวก ซึ่งค่าที่แท้จริงเป็นลบ; ลบเท็จ (false negative, FN) คือ จำนวนข้อมูลที่จำแนกผิดว่าเป็นลบ ซึ่งค่าที่แท้จริงเป็นบวก

เปรียบเทียบประสิทธิภาพในการทำนายผลของวิธีการจำแนกโดยพิจารณาจากค่าความถูกต้อง ค่าคลาดเคลื่อนกำลังสองเฉลี่ย และค่าส่วนเบี่ยงเบนสัมบูรณ์เฉลี่ย ซึ่งวิธีที่ใช้ในการทดสอบครั้งนี้มี 6 วิธี คือ วิธีเพื่อนบ้านใกล้สุด k ตัว วิธีโครงข่ายประสาทเทียม วิธีฐานกฎ วิธีการถดถอยลอจิสติกทวิภาค วิธีเพอร์เซปตรอนให้คะแนน และวิธีลาดลงสโตแคสติก ผลการวิเคราะห์ข้อมูลจากตารางที่ 2 ซึ่งเป็นกรณีที่ค่านอกเกณฑ์อยู่ในระดับต่ำ วิธีลาดลงสโตแคสติกโดยการสุ่มตัวอย่างด้วยโปรแกรม WEKA ให้ค่าความถูกต้องสูงสุด คือ ร้อยละ 56.3934 และให้ค่าคลาดเคลื่อนสัมบูรณ์เฉลี่ยต่ำสุด คือ 0.4360 ส่วนวิธีฐานกฎโดยการสุ่มตัวอย่างด้วยโปรแกรม MINITAB ให้ค่าคลาดเคลื่อนกำลังสองเฉลี่ยต่ำสุด คือ 0.2464 ส่วนตารางที่ 3 ซึ่งเป็นกรณีที่ค่านอกเกณฑ์อยู่ในระดับปานกลาง วิธีเพื่อนบ้านใกล้สุด k ตัว โดยการสุ่มตัวอย่างด้วยโปรแกรม WEKA ให้ค่าความถูกต้องสูงสุด คือ ร้อยละ 58.0000 และให้ค่าคลาดเคลื่อนสัมบูรณ์เฉลี่ยต่ำสุด คือ 0.4216 ส่วนวิธีฐานกฎโดยการสุ่มตัวอย่างด้วยโปรแกรม SPSS ให้ค่าคลาดเคลื่อนกำลังสองเฉลี่ยต่ำสุด คือ 0.2446 และจากตารางที่ 4 ซึ่งเป็นกรณีที่ค่านอกเกณฑ์อยู่ในระดับสูง วิธีโครงข่ายประสาทเทียมโดยการสุ่มตัวอย่างด้วยโปรแกรม WEKA ให้ค่าความถูกต้องสูงสุด คือ ร้อยละ 58.4849 และให้ค่าคลาดเคลื่อนสัมบูรณ์เฉลี่ยต่ำสุด คือ 0.4277 ส่วนวิธีฐานกฎโดยการสุ่มตัวอย่างด้วยโปรแกรม MINITAB ให้ค่าคลาดเคลื่อนกำลังสองเฉลี่ยต่ำสุด คือ 0.2509

4. สรุป อภิปรายผล และข้อเสนอแนะ

4.1 สรุปผลการวิจัย

งานวิจัยนี้ได้ศึกษาประสิทธิภาพในการทำนายผลค่านอกเกณฑ์ด้วยวิธีการจำแนก 6 วิธี ได้แก่ วิธีเพื่อนบ้านใกล้สุด k ตัว วิธีโครงข่ายประสาทเทียม วิธีฐานกฎ วิธีการถดถอยลอจิสติกทวิภาค วิธีเพอร์เซปตรอนให้คะแนน และวิธีลาดลงสโตแคสติก รวมทั้งเปรียบเทียบค่าความถูกต้องในการทำนาย ค่าคลาดเคลื่อนกำลังสองเฉลี่ย และค่าคลาดเคลื่อนสัมบูรณ์เฉลี่ยของวิธีการสุ่มตัวอย่างโดยใช้โปรแกรม SPSS, MINITAB และ WEKA เมื่อข้อมูลมีค่านอกเกณฑ์ 3 ระดับ คือ ระดับต่ำ (ร้อยละ 0-3) ระดับปานกลาง (ร้อยละ 3-6) และระดับสูง (ร้อยละ 6-10) โดยข้อมูลที่นำมาวิเคราะห์ คือ การตรวจสอบธนบัตร การชี้ตัวในเฟสบุ๊ค และการเลือกวิธีการคุมกำเนิด เกณฑ์ที่ใช้ในการเปรียบเทียบประสิทธิภาพ คือ ค่าความถูกต้องในการทำนาย (ร้อยละ) พิจาณาจากค่าที่สูงกว่า ส่วนค่าคลาดเคลื่อนกำลังสองเฉลี่ยและค่าคลาดเคลื่อนสัมบูรณ์เฉลี่ยพิจารณาจากค่าที่ต่ำกว่า

ตารางที่ 2 ผลการเปรียบเทียบประสิทธิภาพของวิธีการจำแนกโดยใช้โปรแกรม SPSS, MINITAB และ WEKA สำหรับข้อมูลการตรวจสอบธนบัตร (ข้อมูลชุดที่ 1 ค่านอกเกณฑ์อยู่ในระดับต่ำ)

วิธีการจำแนก	ชนิดโปรแกรม	ตัวสร้างเลขสุ่มเทียม	ค่าความถูกต้องในการทำนาย (ร้อยละ)	ค่าคลาดเคลื่อนกำลังสองเฉลี่ย	ค่าคลาดเคลื่อนสัมบูรณ์เฉลี่ย
วิธีเพื่อนบ้านใกล้สุด k ตัว	SPSS	10, 20, 30, 40, 50	50.8197, 52.4590, 54.0984, 49.1803, 52.4590	0.4749, 0.4513, 0.4432, 0.4907, 0.4590	0.4921, 0.4763, 0.4605, 0.5079, 0.4763
		ค่าเฉลี่ย	51.8033	0.4638	0.4826
	MINITAB	10, 20, 30, 40, 50	54.0984, 59.0164, 47.5410, 57.3770, 36.0656	0.4432, 0.3958, 0.5065, 0.4115, 0.6172	0.4604, 0.4130, 0.5237, 0.4288, 0.6345
		ค่าเฉลี่ย	50.8197	0.4747	0.4921
	WEKA	10, 20, 30, 40, 50	54.0984, 44.2633, 49.1803, 44.2623, 52.4590	0.4432, 0.5380, 0.4906, 0.5380, 0.4590	0.4604, 0.5554, 0.5079, 0.5554, 0.4763
		ค่าเฉลี่ย	48.8527	0.4938	0.5111
วิธีโครงข่ายประสาทเทียม	SPSS	10, 20, 30, 40, 50	57.3770, 50.8197, 59.0164, 52.4590, 49.1803	0.3031, 0.3035, 0.3037, 0.3538, 0.3287	0.4513, 0.5054, 0.4421, 0.5157, 0.4763
		ค่าเฉลี่ย	53.7705	0.3186	0.4782
	MINITAB	10, 20, 30, 40, 50	60.6557, 47.5410, 47.5410, 36.6560, 50.8197	0.3136, 0.3595, 0.3739, 0.3520, 0.3364	0.4299, 0.5302, 0.5272, 0.5442, 0.5224
		ค่าเฉลี่ย	48.9181	0.3471	0.5108
	WEKA	10, 20, 30, 40, 50	60.6557, 49.1803, 42.6230, 37.7049, 36.0565	0.2785, 0.3517, 0.3785, 0.3573, 0.3965	0.4306, 0.5140, 0.5517, 0.5194, 0.5549
		ค่าเฉลี่ย	45.2441	0.3525	0.5141
วิธีฐานกฎ	SPSS	10, 20, 30, 40, 50	52.4590, 54.0984, 45.9016, 50.8197, 55.7377	0.2499, 0.2497, 0.2815, 0.2500, 0.2486	0.4993, 0.4981, 0.5230, 0.4999, 0.4954
		ค่าเฉลี่ย	51.8033	0.2559	0.5031
	MINITAB	10, 20, 30, 40, 50	57.3770, 54.0984, 57.3770, 52.4590, 55.7377	0.2459, 0.2450, 0.2432, 0.2499, 0.2481	0.4910, 0.4893, 0.4873, 0.4973, 0.4923
		ค่าเฉลี่ย	55.4098	***0.2464***	0.4914
	WEKA	10, 20, 30, 40, 50	60.6557, 54.0984, 55.7377, 49.1803, 60.6557	0.2397, 0.2501, 0.2635, 0.2503, 0.2404	0.4793, 0.4753, 0.4653, 0.5001, 0.4796
		ค่าเฉลี่ย	56.0656	0.2488	0.4799
วิธีการถดถอยลอจิสติกทวิภาค	SPSS	10, 20, 30, 40, 50	55.7377, 34.4262, 57.3770, 55.7377, 50.8197	0.2517, 0.2797, 0.2626, 0.2596, 0.2630	0.4711, 0.5212, 0.4811, 0.4827, 0.4900
		ค่าเฉลี่ย	50.8197	0.2633	0.4892
	MINITAB	10, 20, 30, 40, 50	60.6557, 54.0984, 44.2623, 52.4590, 47.5410	0.2472, 0.2750, 0.2675 0.2821, 0.2678	0.4672, 0.5023, 0.4962, 0.5188, 0.4995
		ค่าเฉลี่ย	51.8033	0.2679	0.4968
	WEKA	10, 20, 30, 40, 50	60.6557, 44.2623, 44.2623, 57.3770, 52.4590	0.2774, 0.2871, 0.2864, 0.2586, 0.2683	0.5072, 0.5286, 0.5195, 0.4872, 0.4971
		ค่าเฉลี่ย	51.8033	0.2756	0.5079
วิธีเพอร์เซปตรอนให้คะแนน	SPSS	10, 20, 30, 40, 50	52.4590, 44.2623, 65.5738, 47.5410, 44.2623	0.4275, 0.5473, 0.3344, 0.4609, 0.5300	0.4743, 0.5691, 0.3633, 0.5188, 0.5651
		ค่าเฉลี่ย	50.8197	0.4600	0.4981
	MINITAB	10, 20, 30, 40, 50	60.6557, 54.0984, 55.7377, 49.1803, 47.5410	0.3848, 0.4445, 0.4292, 0.4731, 0.4987	0.3934, 0.4637, 0.4544, 0.5063, 0.5151
		ค่าเฉลี่ย	53.4426	0.4460	0.4666
	WEKA	10, 20, 30, 40, 50	59.0164, 40.9836, 59.0164, 60.6557, 59.0164	0.3911, 0.5703, 0.3851, 0.3647, 0.3973	0.4034, 0.5899, 0.4036, 0.4070, 0.4215
		ค่าเฉลี่ย	55.7377	0.4217	0.4451
วิธีลาดลงสโตแคสติก	SPSS	10, 20, 30, 40, 50	54.0984, 42.6230, 49.1803, 49.1803, 55.7377	0.4590, 0.5738, 0.5082, 0.5082, 0.4426	0.4590, 0.5738, 0.5082, 0.5082, 0.4426
		ค่าเฉลี่ย	50.1639	0.4984	0.4984
	MINITAB	10, 20, 30, 40, 50	60.6557, 55.7377, 49.1803, 52.4590, 44.2623	0.3935, 0.4426, 0.5082, 0.4754, 0.5574	0.3934, 0.4426, 0.5082, 0.4754, 0.5574
		ค่าเฉลี่ย	52.4590	0.4754	0.4754
	WEKA	10, 20, 30, 40, 50	62.2951, 45.9016, 50.8197, 62.2951, 60.6557	0.3770, 0.5410, 0.4918, 0.3778, 0.3935	0.3770, 0.5410, 0.4918, 0.3770, 0.3934
		ค่าเฉลี่ย	***56.3934***	0.4374	***0.4360***

ตารางที่ 3 ผลการเปรียบเทียบประสิทธิภาพของวิธีการจำแนกโดยใช้โปรแกรม SPSS, MINITAB และ WEKA สำหรับข้อมูลการชี้ตัวในเฟสบุ๊ค (ข้อมูลชุดที่ 2 ค่านอกเกณฑ์อยู่ในระดับปานกลาง)

วิธีการจำแนก	ชนิดโปรแกรม	ตัวสร้างเลขสุ่มเทียม	ค่าความถูกต้องในการทำนาย (ร้อยละ)	ค่าคลาดเคลื่อนกำลังสองเฉลี่ย	ค่าคลาดเคลื่อนสัมบูรณ์เฉลี่ย
วิธีเพื่อนบ้านใกล้สุด k ตัว	SPSS	10, 20, 30, 40, 50	50.0000, 41.1765, 52.0000, 44.0000, 50.0000	0.4791, 0.5642, 0.4601, 0.5366, 0.4791	0.5000, 0.5845, 0.4809, 0.5574, 0.5000
		ค่าเฉลี่ย	47.4353	0.5038	0.5246
	MINITAB	10, 20, 30, 40, 50	58.0000, 44.0000, 36.0000, 50.0000, 46.0000	0.4026, 0.5366, 0.6133, 0.4791, 0.5175	0.4234, 0.5574, 0.6340, 0.5000, 0.5383
		ค่าเฉลี่ย	46.8000	0.5098	0.5306
	WEKA	10, 20, 30, 40, 50	48.0000, 52.0000, 58.0000, 64.0000, 68.0000	0.3507, 0.4601, 0.4026, 0.3452, 0.3068	0.5098, 0.4809, 0.4234, 0.3660, 0.3277
		ค่าเฉลี่ย	***58.0000***	0.3731	***0.4216***
วิธีโครงข่ายประสาทเทียม	SPSS	10, 20, 30, 40, 50	68.0000, 47.0588, 34.0000, 38.0000, 58.0000	0.3172, 0.4456, 0.4976, 0.5061, 0.3334	0.4119, 0.5271, 0.5988, 0.5980, 0.4268
		ค่าเฉลี่ย	49.0118	0.4200	0.5125
	MINITAB	10, 20, 30, 40, 50	68.0000, 34.0000, 48.0000, 60.0000, 52.0000	0.2733, 0.5310, 0.4193, 0.3407, 0.4118	0.3628, 0.6125, 0.5202, 0.4254, 0.4962
		ค่าเฉลี่ย	52.4000	0.3952	0.4834
	WEKA	10, 20, 30, 40, 50	52.0000, 42.0000, 60.0000, 66.0000, 56.0000	0.4123, 0.4571, 0.3546, 0.2979, 0.3612	0.5070, 0.5474, 0.4198, 0.3621, 0.4449
		ค่าเฉลี่ย	55.2000	0.3766	0.4562
วิธีฐานกฎ	SPSS	10, 20, 30, 40, 50	52.0000, 47.0588, 46.0000, 42.0000, 70.0000	0.2531, 0.2517, 0.2531, 0.2529, 0.2124	0.5022, 0.5015, 0.5022, 0.5026, 0.4246
		ค่าเฉลี่ย	51.4118	***0.2446***	0.4866
	MINITAB	10, 20, 30, 40, 50	52.0000, 42.0000, 42.0000, 46.0000, 54.0000	0.2863, 0.2529, 0.2527, 0.2830, 0.2513	0.5046, 0.5026, 0.5022, 0.5264, 0.4991
		ค่าเฉลี่ย	47.2000	**0.2652**	0.5070
	WEKA	10, 20, 30, 40, 50	54.0000, 56.0000, 58.0000, 52.0000, 46.0000	0.2497, 0.2480, 0.2441, 0.2713, 0.2529	0.4895, 0.4943, 0.4883, 0.5108, 0.5026
		ค่าเฉลี่ย	53.2000	0.2532	0.4971
วิธีการถดถอยลอจิสติกทวิภาค	SPSS	10, 20, 30, 40, 50	48.0000, 41.1765, 50.0000, 40.0000, 62.0000	0.2977, 0.2723, 0.3028, 0.3218, 0.2574	0.4950, 0.4970, 0.5213, 0.5400, 0.4420
		ค่าเฉลี่ย	48.2353	0.2904	0.4991
	MINITAB	10, 20, 30, 40, 50	50.0000, 36.0000, 50.0000, 56.0000, 46.0000	0.3037, 0.3089, 0.3037, 0.2883, 0.3130	0.5045, 0.5305, 0.5045, 0.4778, 0.5452
		ค่าเฉลี่ย	47.6000	0.3035	0.5125
	WEKA	10, 20, 30, 40, 50	50.0000, 48.0000, 60.0000, 60.0000, 42.0000	0.2920, 0.3101, 0.2616, 0.2405, 0.3069	0.5235, 0.5152, 0.4488, 0.4156, 0.5397
		ค่าเฉลี่ย	52.0000	0.2822	0.4886
วิธีเพอร์เซปตรอนให้คะแนน	SPSS	10, 20, 30, 40, 50	52.0000, 47.0588, 46.0000, 46.0000, 70.0000	0.4138, 0.4889, 0.4674, 0.3693, 0.2981	0.4979, 0.5436, 0.5380, 0.5168, 0.2991
		ค่าเฉลี่ย	52.2118	0.4075	0.4791
	MINITAB	10, 20, 30, 40, 50	58.0000, 48.0000, 44.0000, 50.0000, 50.0000	0.4207, 0.3888, 0.4798, 0.4072, 0.4334	0.4254, 0.5070, 0.5544, 0.5073, 0.4848
		ค่าเฉลี่ย	50.0000	0.4260	0.4958
	WEKA	10, 20, 30, 40, 50	54.0000, 58.0000, 50.0000, 56.0000, 42.0000	0.4133, 0.3447, 0.4498, 0.4039, 0.4591	0.4666, 0.4454, 0.4839, 0.4466, 0.5425
		ค่าเฉลี่ย	52.0000	0.4142	0.4770
วิธีลาดลงสโตแคสติก	SPSS	10, 20, 30, 40, 50	58.0000, 47.0588, 42.0000, 38.0000, 70.0000	0.4200, 0.5294, 0.5800, 0.6200, 0.3000	0.4200, 0.5294, 0.5800, 0.6200, 0.3000
		ค่าเฉลี่ย	51.0118	0.4899	0.4899
	MINITAB	10, 20, 30, 40, 50	72.0000, 36.0000, 48.0000, 60.0000, 42.0000	0.2801, 0.6400, 0.5200, 0.4001, 0.5800	0.2800, 0.6400, 0.5200, 0.4000, 0.5800
		ค่าเฉลี่ย	51.6000	0.4840	0.4840
	WEKA	10, 20, 30, 40, 50	56.0000, 46.0000, 58.0000, 64.0000, 44.0000,	0.4400, 0.5399, 0.4200, 0.3600, 0.5600	0.4400, 0.5400, 0.4200, 0.3600, 0.5600
		ค่าเฉลี่ย	53.6000	0.4640	0.4640

ตารางที่ 4 ผลการเปรียบเทียบประสิทธิภาพของวิธีการจำแนกโดยใช้โปรแกรม SPSS, MINITAB และ WEKA สำหรับข้อมูลการเลือกวิธีคุมกำเนิด (ข้อมูลชุดที่ 3 ค่านอกเกณฑ์อยู่ในระดับสูง)

วิธีการจำแนก	ชนิดโปรแกรม	ตัวสร้างเลขสุ่มเทียม	ค่าความถูกต้องในการทำนาย (ร้อยละ)	ค่าคลาดเคลื่อนกำลังสองเฉลี่ย	ค่าคลาดเคลื่อนสัมบูรณ์เฉลี่ย
วิธีเพื่อนบ้านใกล้สุด k ตัว	SPSS	10, 20, 30, 40, 50	57.5758, 34.8485, 43.9394, 57.5758, 60.6061	0.4106, 0.6306, 0.5426, 0.4106, 0.3813	0.4267, 0.6465, 0.5586, 0.4267, 0.3974
		ค่าเฉลี่ย	50.9091	0.4751	0.4912
	MINITAB	10, 20, 30, 40, 50	60.6061, 30.3030, 40.9091, 54.5455, 33.3333	0.3813, 0.6745, 0.5718, 0.4400, 0.6453	0.3974, 0.6905, 0.5879, 0.4560, 0.6613
		ค่าเฉลี่ย	43.9394	0.5426	0.5586
	WEKA	10, 20, 30, 40, 50	54.5455, 60.6061, 53.0303, 34.8485, 54.5455	0.4400, 0.3814, 0.4547, 0.6306, 0.4400	0.4559, 0.3974, 0.4707, 0.6466, 0.4560
		ค่าเฉลี่ย	51.5152	0.4693	0.4853
วิธีโครงข่ายประสาทเทียม	SPSS	10, 20, 30, 40, 50	50.0000, 39.3939, 42.4242, 43.9394, 60.6061	0.3999, 0.5037, 0.4735, 0.4695, 0.3486	0.4893, 0.5776, 0.5617, 0.5556, 0.4239
		ค่าเฉลี่ย	47.2727	0.4390	0.5216
	MINITAB	10, 20, 30, 40, 50	56.0606, 60.6061, 43.9394, 51.5152, 36.3636	0.3727, 0.3109, 0.4556, 0.4017, 0.4742	0.4663, 0.4217, 0.5709, 0.4858, 0.5777
		ค่าเฉลี่ย	49.6970	0.4030	0.5045
	WEKA	10, 20, 30, 40, 50	53.0303, 69.6970, 66.6667, 46.9697, 56.0606	0.3868, 0.2336, 0.2643, 0.4194, 0.3536	0.4959, 0.3187, 0.3370, 0.5367, 0.4502
		ค่าเฉลี่ย	***58.4849***	0.3315	***0.4277***
วิธีฐานกฎ	SPSS	10, 20, 30, 40, 50	53.0303, 45.4545, 53.0303, 54.5455, 48.4848	0.2541, 0.2738, 0.2500, 0.2483, 0.2657	0.4936, 0.5132, 0.4990, 0.4962, 0.5109
		ค่าเฉลี่ย	50.9091	0.2584	0.5026
	MINITAB	10, 20, 30, 40, 50	53.0303, 54.5455, 53.0303, 56.0606, 48.4848	0.2493, 0.2585, 0.2493, 0.2469, 0.2506	0.4985, 0.4858, 0.4958, 0.4935, 0.5003
		ค่าเฉลี่ย	53.0303	***0.2509***	0.4948
	WEKA	10, 20, 30, 40, 50	62.1212, 56.0606, 48.4848, 54.5455, 46.9697	0.2365, 0.2469, 0.2501, 0.2481, 0.2809	0.4724, 0.4935, 0.4997, 0.4962, 0.5122
		ค่าเฉลี่ย	53.6364	0.2525	0.4948
วิธีการถดถอยลอจิสติกทวิภาค	SPSS	10, 20, 30, 40, 50	54.5455, 54.5455, 53.0303, 63.6360, 57.5758	0.2660, 0.2782, 0.2964, 0.2507, 0.3032	0.4658, 0.4542, 0.5146, 0.4497, 0.5172
		ค่าเฉลี่ย	56.6666	0.2789	0.4803
	MINITAB	10, 20, 30, 40, 50	53.0303, 57.5758, 45.4545, 40.9091, 45.4545	0.2965, 0.2676, 0.2915, 0.2969, 0.3027	0.5187, 0.4669, 0.5255, 0.5304, 0.5260
		ค่าเฉลี่ย	48.4848	0.2910	0.5135
	WEKA	10, 20, 30, 40, 50	53.0303, 63.6364, 62.1212, 48.4848, 60.6061	0.2882, 0.2469, 0.2462, 0.2913, 0.2500	0.5190, 0.4436, 0.4235, 0.5227, 0.4468
		ค่าเฉลี่ย	57.5758	0.2645	0.4711
วิธีเพอร์เซปตรอนให้คะแนน	SPSS	10, 20, 30, 40, 50	53.0303, 46.9697, 53.0303, 54.5455, 53.0303	0.4659, 0.3965, 0.4326, 0.4516, 0.4111	0.4719, 0.5135, 0.4799, 0.4548, 0.4780
		ค่าเฉลี่ย	52.1212	0.4316	0.4796
	MINITAB	10, 20, 30, 40, 50	53.0303, 62.1212, 46.9697, 54.5455, 50.0000	0.3896, 0.3788, 0.4464, 0.4199, 0.4168	0.4888, 0.3788, 0.5043, 0.4522, 0.5006
		ค่าเฉลี่ย	53.3333	0.4103	0.4649
	WEKA	10, 20, 30, 40, 50	50.0000, 56.0606, 42.4242, 53.0303, 48.4848	0.3980, 0.4394, 0.4605, 0.4163, 0.3920	0.5033, 0.4394, 0.5654, 0.4641, 0.4981
		ค่าเฉลี่ย	50.0000	0.4213	0.4941
วิธีลาดลงสโตแคสติก	SPSS	10, 20, 30, 40, 50	51.5152, 65.1515, 50.0000, 60.6061, 48.4848	0.4848, 0.3485, 0.5000, 0.3939, 0.5151	0.4848, 0.3485, 0.5000, 0.3939, 0.5152
		ค่าเฉลี่ย	55.1515	0.4485	0.4485
	MINITAB	10, 20, 30, 40, 50	51.5152, 66.6667, 50.0000, 39.3939, 53.0303	0.4848, 0.3334, 0.5000, 0.6061, 0.4696	0.4848, 0.3333, 0.5000, 0.6061, 0.4697
		ค่าเฉลี่ย	52.1212	0.4788	0.4788
	WEKA	10, 20, 30, 40, 50	50.0000, 66.6667, 53.0303, 46.9697, 62.1212	0.5000, 0.3334, 0.4696, 0.5303, 0.3788	0.5000, 0.3333, 0.4697, 0.5303, 0.3788
		ค่าเฉลี่ย	55.7576	0.4424	0.4424

การเปรียบเทียบชุดข้อมูลที่มีค่านอกเกณฑ์อยู่ในระดับต่ำ คือ การตรวจสอบธนบัตร วิธีที่มีประสิทธิภาพสูงสุด คือ วิธีลาดลงสโตแคสติก โดยการสุ่มตัวอย่างด้วยโปรแกรม WEKA ชุดข้อมูลที่มีค่านอกเกณฑ์อยู่ในระดับปานกลาง คือ การชี้ตัวในเฟสบุ๊ค วิธีที่มีประสิทธิภาพสูงสุด คือ วิธีเพื่อนบ้านใกล้สุด k ตัว โดยการสุ่มตัวอย่างด้วยโปรแกรม WEKA ชุดข้อมูลที่มีค่านอกเกณฑ์อยู่ในระดับสูง คือ การเลือกวิธีการคุมกำเนิด วิธีที่มีประสิทธิภาพสูงสุด คือ วิธีโครงข่ายประสาทเทียม โดยการสุ่มตัวอย่างด้วยโปรแกรม WEKA ถ้าเปรียบเทียบเฉพาะวิธีการจำแนกที่มีประสิทธิภาพดีที่สุด วิธีที่มีประสิทธิภาพดีที่สุด คือ วิธีฐานกฎ รองลงมา คือ วิธีลาดลงสโตแคสติก วิธีเพื่อนบ้านใกล้สุด k ตัว และวิธีโครงข่ายประสาทเทียม ส่วนถ้าเปรียบเทียบเฉพาะโปรแกรมที่มีประสิทธิภาพดีที่สุด คือ โปรแกรม WEKA รองลงมา คือ โปรแกรม MINITAB และ SPSS ตามลำดับ

4.2 อภิปรายผล

การสรุปผลงานวิจัยครั้งนี้ ชุดข้อมูลที่มีค่านอกเกณฑ์อยู่ในระดับต่ำ ได้แก่ การตรวจสอบธนบัตร วิธีลาดลงสโตแคสติกโดยการสุ่มตัวอย่างด้วยโปรแกรม WEKA ให้ค่าความถูกต้องสูงสุดและค่าคลาดเคลื่อนสัมบูรณ์เฉลี่ยต่ำสุด วิธีฐานกฎโดยการสุ่มตัวอย่างด้วยโปรแกรม MINITAB ให้ค่าคลาดเคลื่อนกำลังสองเฉลี่ยต่ำสุด ซึ่งให้ผลไม่สอดคล้องกับ พนิดา และคณะ (2560) ที่ศึกษาในชุดข้อมูลโรคมะเร็งเต้านมของรัฐวิสคอนซิน พบว่าวิธีเพื่อนบ้านใกล้สุด k ตัว โดยใช้โปรแกรม SPSS และ WEKA วิธีต้นไม้ตัดสินใจโดยใช้โปรแกรม WEKA วิธีโครงข่ายประสาทเทียมโดยใช้โปรแกรม SPSS และ WEKA วิธีซัพพอร์ตเวกเตอร์แมชชีนโดยใช้โปรแกรม SPSS ให้ค่าความถูกต้องสูงสุด วิธีต้นไม้ตัดสินใจโดยใช้โปรแกรม WEKA วิธีซัพพอร์ตเวกเตอร์แมชชีนโดยใช้โปรแกรม SPSS ให้ค่าความคลาดเคลื่อนกำลังสองเฉลี่ยและค่าคลาดเคลื่อนสัมบูรณ์เฉลี่ยต่ำสุด เนื่องจากในงานวิจัยของ พนิดา และคณะ (2560) ไม่ได้ศึกษาวิธีลาดลงสโตแคสติกและวิธีฐานกฎ และข้อมูลที่นำมาศึกษาก็มีความแตกต่างกัน

ชุดข้อมูลที่มีค่านอกเกณฑ์อยู่ในระดับปานกลาง ได้แก่ การชี้ตัวในเฟสบุ๊ค วิธีเพื่อนบ้านใกล้สุด k ตัว โดยการสุ่มตัวอย่างด้วยโปรแกรม WEKA ให้ค่าความถูกต้องสูงสุดและค่าคลาดเคลื่อนสัมบูรณ์เฉลี่ยต่ำสุด วิธีฐานกฎโดยการสุ่มตัวอย่างด้วยโปรแกรม SPSS ให้ค่าคลาดเคลื่อนกำลังสองเฉลี่ยต่ำสุด ให้ผลสอดคล้องกับ พนิดา และคณะ (2560) ซึ่งศึกษาในชุดข้อมูลโรคเบาหวานของชาวพิม่า ประเทศอินเดีย พบว่าวิธีเพื่อนบ้านใกล้สุด k ตัว โดยการสุ่มตัวอย่างด้วยโปรแกรม SPSS และ WEKA ให้ค่าความถูกต้องสูงสุด ค่าความคลาดเคลื่อนกำลังสองเฉลี่ยและค่าคลาดเคลื่อนสัมบูรณ์เฉลี่ยต่ำสุด แต่ในงานวิจัยของ พนิดา และคณะ (2560) ไม่ได้ศึกษาวิธีฐานกฎ

ชุดข้อมูลที่มีค่านอกเกณฑ์อยู่ในระดับสูง ได้แก่ การเลือกวิธีการคุมกำเนิด วิธีโครงข่ายประสาทเทียมโดยการสุ่มตัวอย่างด้วยโปรแกรม WEKA ให้ค่าความถูกต้องสูงสุดและค่าคลาดเคลื่อนสัมบูรณ์เฉลี่ยต่ำสุด วิธีฐานกฎโดยการสุ่มตัวอย่างด้วยโปรแกรม MINITAB ให้ค่าคลาดเคลื่อนกำลังสองเฉลี่ยต่ำสุด ให้ผลไม่สอดคล้องกับ พนิดา และคณะ (2560) ซึ่งศึกษาในชุดข้อมูลการชำระเงินด้วยบัตรเครดิตของลูกค้า พบว่าวิธีเพื่อนบ้านใกล้สุด k ตัว โดยการสุ่มตัวอย่างด้วยโปรแกรม SPSS และ WEKA ให้ค่าความถูกต้องสูงสุด ค่าคลาดเคลื่อนกำลังสองเฉลี่ยและค่าคลาดเคลื่อนสัมบูรณ์เฉลี่ยต่ำสุด เนื่องจากในงานวิจัยนี้กับงานวิจัยของ พนิดา และคณะ (2560) ไม่ได้ศึกษาวิธีฐานกฎ ข้อมูลที่นำมาศึกษามีความแตกต่างกัน

4.3 ข้อเสนอแนะ

4.3.1 เพื่อให้ได้ข้อสรุปของผลการวิเคราะห์ข้อมูลที่มีความสมบูรณ์มากขึ้น อาจวิเคราะห์ข้อมูลด้วยวิธีการอื่น ๆ ได้แก่ วิธีนาอีฟเบส์ วิธีนาอีฟเบส์ปรับปรุง วิธีเบส์เนท วิธีเบส์เนทปรับปรุง เป็นต้น เนื่องจากเป็นวิธีที่นิยมใช้กันพอสมควรในการเปรียบเทียบประสิทธิภาพในการทำนายผลด้วยการจำแนก

4.3.2 ควรศึกษาวิธีการหาค่านอกเกณฑ์ด้วยโปรแกรมอื่น ๆ ได้แก่ R, NCSS เป็นต้น เนื่องจากเป็นโปรแกรมที่มีวิธีการจัดการข้อมูลและตัวสถิติทดสอบให้เลือกใช้เป็นจำนวนมาก นอกจากนี้ทั้ง 2 โปรแกรม ได้รับความนิยมในการใช้งาน

4.3.3 ควรเพิ่มจำนวนชุดข้อมูลที่มีค่านอกเกณฑ์ให้หลากหลายมากขึ้น เช่น ชุดข้อมูลที่มีค่านอกเกณฑ์ระดับต่ำ ระดับปานกลาง และระดับสูงอย่างละ 2 ชุด เพื่อให้ได้ข้อสรุปที่ชัดเจน

4.3.4 การวิเคราะห์ค่านอกเกณฑ์ไม่มีหลักเกณฑ์การแบ่งที่แน่นอน ผู้วิจัยได้ทำการค้นหาข้อมูลจำนวนหนึ่งมาตรวจสอบค่านอกเกณฑ์ พบว่ามีค่าร้อยละ 0-10 จึงแบ่งข้อมูลเป็น 3 ช่วง คือ ต่ำ ปานกลาง และสูง ซึ่งอาจมีข้อมูลอื่น ๆ ที่มีค่านอกเกณฑ์ไม่ได้อยู่ในช่วงดังกล่าว แล้วไม่ได้นำมาศึกษา

5. กิตติกรรมประกาศ

ขอขอบคุณคณะวิทยาศาสตร์ สถาบันเทคโนโลยีพระจอมเกล้าเจ้าคุณทหารลาดกระบัง ที่สนับสนุนการให้ทุนวิจัย ประจำปี 2562 เรื่อง การเปรียบเทียบประสิทธิภาพในการทำนายผลค่านอกเกณฑ์ด้วยการจำแนก 6 วิธี

6. รายการอ้างอิง

กิตติพล วิแสง, สิรภัทร เชี่ยวชาญวัฒนา และคำรณ สุนัติ, 2552, การวิเคราะห์ปัจจัยเสี่ยงของโรคเบาหวาน, 8 น., ใน รายงานการประชุมวิชาการระดับชาติด้านคอมพิวเตอร์และเทคโนโลยีสารสนเทศ (NCCIT) ครั้งที่ 5, มหาวิทยาลัยเทคโนโลยีพระจอมเกล้าพระนครเหนือ, กรุงเทพฯ.

กัลยา วานิชย์บัญชา, 2552, การวิเคราะห์ข้อมูลหลายตัวแปร, บริษัทธรรมสาร จำกัด, กรุงเทพฯ.

ณัฐวุฒิ ศรีวิบูลย์, การเปรียบเทียบประสิทธิภาพอัลกอริทึมเหมืองข้อมูลเพื่อวิเคราะห์ปัจจัยที่ส่งผลต่อการเกิดโรคมะเร็ง, แหล่งที่มา : http://snrujst.snru.ac.th/th/articles-in-press, 25 ตุลาคม 2560.

ทิพย์ธิดา วงศ์พิพันธ์, 2555, การใช้เหมืองข้อมูลช่วยในการตัดสินใจการให้สินเชื่อ,วิทยานิพนธ์ปริญญาโท, มหาวิทยาลัยธุรกิจบัณฑิตย์, กรุงเทพฯ.

ธนาวุฒิ ประกอบผล, 2552, โครงข่ายประสาทเทียม, ว.มฉก.วิชาการ 12(24): 73-87.

นิเวศ จิระวิชิตชัย, การค้นหาเทคนิคเหมืองข้อมูลเพื่อสร้างโมเดลการวิเคราะห์โรคอัตโนมัติ, แหล่งที่มา : http://www.ssruir.ssru.ac.th/bitstream/ssruir/377/1/080-53.pdf, 25 ตุลาคม 2560.

พนิดา สมบัติมาก, ภัสสร จันทร์หอม, ศุภกร รัศมี และโอพาร รุ่งมณีธรรมคุณ, 2560, การเปรียบเทียบประสิทธิภาพในการจำแนกกลุ่มเมื่อข้อมูลมีค่านอกเกณฑ์ในการทำเหมืองข้อมูล, ปัญหาพิเศษปริญญาตรี, สถาบันเทคโนโลยีพระจอมเกล้าคุณทหารลาดกระบัง, กรุงเทพฯ.

พยูน พาณิชย์กุล, 2548, การพัฒนาระบบดาต้าไมน์นิ่งโดยใช้ Decision Tree, วิทยานิพนธ์ปริญญาโท, สถาบันเทคโนโลยีพระจอมเกล้าคุณทหารลาดกระบัง, กรุงเทพฯ.

วรพรรณ เจริญขำ, 2556, การตรวจสอบค่านอก

เกณฑ์ในตัวอย่างสุ่มจากประชากรปรกติ, วิทยานิพนธ์ปริญญาโท, สถาบันบัณฑิตพัฒนบริหารศาสตร์, กรุงเทพฯ.

สายชล สินสมบูรณ์ทอง, 2560, การทำเหมืองข้อมูล เล่ม 1 : การค้นหาความรู้จากข้อมูล, พิมพ์ครั้งที่ 2, จามจุรีโปรดักส์ จำกัด, กรุงเทพฯ.

สุรวัชร ศรีเปารยะ และสายชล สินสมบูรณ์ทอง, 2560, การเปรียบเทียบประสิทธิภาพวิธีการจำแนกกลุ่มการเป็นโรคไตเรื้อรัง : กรณีศึกษาโรงพยาบาลแห่งหนึ่งในประเทศอินเดีย, ว. วิทยาศาสตร์และเทคโนโลยี 25(5): 839-853.

เดช ธรรมศิริ และพยุง มีสัจ, 2553, การจำแนกข้อมูลด้วยเทคนิคซัพพอร์ตเวกเตอร์แมชชีนโดยการปรับพารามิเตอร์และเลือกคุณลักษณะที่เหมาะสมด้วยขั้นตอนวิธีเชิงพันธุกรรม, , 12 น., ใน รายงานการประชุมทางวิชาการเสนอผลงานวิจัย ระดับบัณฑิตศึกษา ครั้งที่ 11, มหาวิทยาลัยขอนแก่น, ขอนแก่น.

เดช ธรรมศิริ, วาทินี นุ้ยเพียร, ภัทราวุฒิ แสงศิริ, ภรัณยา อำมฤครัตน์, ณรงค์ โพธิ และพยุง มีสัจ, 2552, การให้คะแนนสินเชื่อโดยวิธีการทำเหมืองข้อมูลด้วยเทคนิคซับพอร์ตเวกเตอร์แมชชีนรวมทั้งการเลือกใช้ลักษณะที่เหมาะสมร่วมกับการหาค่าพารามิเตอร์ที่เหมาะสมด้วยวิธีค้นหาแบบกริช, 11 น., ใน การประชุมวิชาการระดับชาติด้านคอมพิวเตอร์และเทคโนโลยีสารสนเทศ (NCCIT) ครั้งที่ 5, มหาวิทยาลัยเทคโนโลยีพระจอมเกล้าพระนครเหนือ, กรุงเทพฯ.

Berson, A. and Smith, S.J., 1997, Data Warehousing, Data Mining and OLAP, McGraw-Hll, New York.

Freund, Y. and Schapire, R.E., 1998, Large Margin Classification Using the Perceptron Algorithms, NCCLT, New York, 13 p.

Hagan, M., Demuth, H., and Beale, M., 1996, Neural Network Design, Martin T. Hagan, Oklahoma.

LeCun, Y., Bottou, L., Orr, G. and Muller, K., Efficiency BackProp, In Neural Networks, Available Source: https://scholar.google.co.th/scholar?q=LeCun,+Y.+and+Bottou,+L.&hl=th&as_sdt=o&as_vis=1&oi=scholart#d=gs_qabs&u=%23p%3DQzVcWsIB3yQJ, January 20, 2018.

Murti, S. and Mahantappa, M., Using Rule Based Classifiers for the Predictive Analysis of Breast Cancer Recurrence, Available Source: https://archive.ics.uci.edu/ml/datasets/pima+indians+diabetes, February 1, 2018.

Nektarios, T. G., Weka Classify Summary, Athens University of Economics and Business, Available Source: https://www.academia.edu/5167325/Weka_Classifiers_Summary, January 20, 2018.

Priya, R. and Aruna, P., 2012, Support vector machine and neural network based diagnosis of diabetic retinopathy, Int. J. Comput. Appl. 41: 15-27.

Sa-ngasoongsong, A. and Chongwatpol, J., 2012, An Analysis of Diabetes Risk Factors using Data Mining Approach, Oklahoma State University, Stillwater, 11 p.

Singh, S. and Bansal, M., 2013, Improvement of intrusion detection system in data mining using neural network, Int. J. Adv. Res. Comput. Sci. Software Eng. 3: 1124-1130.

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