การเปรียบเทียบประสิทธิภาพในการทำนายผลการปรับความไม่สมดุลของข้อมูลในการจำแนกด้วยเทคนิคการทำเหมืองข้อมูล

Performance Comparison in Prediction of Imbalanced Data in Data Mining Classification

พัชรียา ทองพูล*, พิมพ์ชนก จำเรือง,
รมย์นลิน บุญฤทธิ์ และสายชล สินสมบูรณ์ทอง
ภาควิชาสถิติ คณะวิทยาศาสตร์ สถาบันเทคโนโลยีพระจอมเกล้าเจ้าคุณทหารลาดกระบัง ถนนฉลองกรุง เขตลาดกระบัง กรุงเทพมหานคร 10520

Pachareeya Thongpool*, Pimchanok Jamrueng,
Romnalin Boonrit and Saichon Sinsomboonthong
Department of Statistics, Faculty of Science, King Mongkut's Institute of Technology Ladkrabang, Chalongkrung Road, Ladkrabang, Bangkok 10520

Received: May 22, 2019; Accepted: June 18, 2019

บทคัดย่อ

งานวิจัยนี้มีนี้มีวัตถุประสงค์เพื่อเปรียบเทียบวิธีการปรับข้อมูลที่ไม่สมดุล 4 วิธี คือ วิธีการสุ่มเกิน วิธีการสุ่มเกินโดยเทคนิค SMOTE วิธีการสุ่มลด และวิธีการสุ่มผสมผสาน โดยวิธีการจำแนก 4 วิธี คือ วิธีเพื่อนบ้านใกล้สุด k ตัว วิธีต้นไม้ตัดสินใจ วิธีโครงข่ายประสาทเทียม และวิธีซัพพอร์ตเวกเตอร์แมชชีน ว่าวิธีใดมีประสิทธิภาพในการจำแนกดีที่สุด โดยพิจารณาจากค่าความถูกต้อง ค่าความไว ค่าความจำเพาะ และค่าคลาดเคลื่อนกำสองเฉลี่ย โดยแบ่งข้อมูลในอัตราส่วน 70, 20 และ 10 ตามลำดับ ข้อมูลส่วนที่ 1 ข้อมูลเรียนรู้ นำไปสร้างตัวแบบร้อยละ 70 ข้อมูลส่วนที่ 2 ข้อมูลตรวจสอบความถูกต้อง นำข้อมูลไปประเมินความผิดพลาดของตัวแบบร้อยละ 20 และข้อมูลส่วนที่ 3 ข้อมูลทดสอบ นำไปทดสอบตัวแบบร้อยละ 10 โดยการกำหนดตัวสร้างเลขสุ่มเทียมเป็น 10, 20, 30, 40 และ 50 มีข้อมูลที่ไม่สมดุลในการศึกษา 3 ชุด คือ ชุดข้อมูลการรับรู้ทางหูของเด็กที่มีภาวะน้ำคั่งในหูชั้นกลางหรือหูน้ำหนวก ชุดข้อมูลยอดคงเหลือในบัตรเครดิตของลูกค้า และชุดข้อมูลคุณภาพไวน์แดง โดยใช้โปรแกรม WEKA เมื่อเปรียบเทียบข้อมูลการรับรู้ทางหูของเด็กที่มีภาวะน้ำคั่งในหูชั้นกลางหรือหูน้ำหนวก วิธีที่มีประสิทธิภาพสูงสุดคือวิธีซัพพอร์ตเวกเตอร์แมชชีนโดยการปรับความไม่สมดุลด้วยวิธีการสุ่มเกินเทคนิค SMOTE ชุดข้อมูลยอดคงเหลือในบัตรเครดิตของลูกค้า วิธีที่มีประสิทธิภาพสูงสุด คือ วิธีเพื่อนบ้านใกล้สุด k ตัว โดยการปรับความไม่สมดุลด้วยวิธีการสุ่มเกินเทคนิค SMOTE ส่วนชุดข้อมูลคุณภาพไวน์แดง วิธีที่มีประสิทธิภาพสูงสุดคือวิธีโครงข่ายประสาทเทียมโดยการปรับความไม่สมดุลด้วยวิธีการสุ่มเกิน

คำสำคัญ : ความไม่สมดุลของข้อมูล; วิธีเพื่อนบ้านใกล้สุด k ตัว; วิธีต้นไม้ตัดสินใจ; วิธีโครงข่ายประสาทเทียม; วิธีซัพพอร์ตเวกเตอร์แมชชีน

**Corresponding author: youretrorian@gmail.com*
DOI: 10.14456/tjst.2019.61

Abstract

We compared the imbalanced data of four methods: over sampling, synthetic minority over sampling technique, under sampling and hybrid method using four classification methods: k-nearest neighbor, decision tree, artificial neural network and support vector machine. Metrics were accuracy, sensitivity, specificity and mean squared error. The data sets were auditory perception in children with otitis media with effusion (OME), credit card balance and red wine quality. Each of these data sets was divided into three proportions in the ratio of 70:20:10 using the data part 1, training data are used to create a model 70 percentages; the data part 2, validaion data are used to evaluate an error a model 20 percentages and the data part 3, testing data are used to testing a model 10 percentages using the random seed 10, 20, 30, 40 and 50 by WEKA program. When we compared the OME data set, the best classification method was the support vector machine in imbalanced data, adapting the synthetic minority over sampling technique. For the credit card data sets, the best classification method was the k-nearest neighbor in imbalanced data, adapting the synthetic minority over sampling technique. For the wine data sets, the best method was the artificial neural network in imbalanced data adapting over sampling.

Keywords: imbalanced data; k-nearest neighbor; decision tree; artificial neural network; support vector machine

1. คำนำ

ปัจจุบันมีปัญหาการแบ่งข้อมูลที่กำลังได้รับความสนใจ คือ ปัญหาการแบ่งกลุ่มข้อมูลที่ไม่สมดุล ซึ่งเกิดจากการที่มีข้อมูล 2 กลุ่ม หรือมากกว่า 2 กลุ่ม โดยข้อมูลที่เป็นกลุ่มส่วนมากจะมีข้อมูลจำนวนมากกว่า ขณะเดียวกันข้อมูลกลุ่มส่วนน้อยจะมีข้อมูลจำนวนน้อยกว่า ทั้งนี้เนื่องจากโดยธรรมชาติของความเป็นจริง การที่จะกำหนดให้ขนาดของข้อมูลในกลุ่มส่วนมาก (majority) และกลุ่มส่วนน้อย (minority) มีขนาดที่เท่าเทียมกันเพื่อการสอนหรือการจัดกลุ่มข้อมูลนั้นเป็นเรื่องยากหรืออาจเป็นไปไม่ได้ ดังนั้นจึงเป็นปัญหาที่ท้าทายและมีความยากมากสำหรับการหาขั้นตอนวิธีที่เหมาะสมสำหรับการแบ่งกลุ่มข้อมูลที่ไม่สมดุล ทั้งนี้เนื่องจากถ้านำข้อมูลทั้ง 2 ชุด เข้าสู่ขั้นตอนวิธีพร้อมกันทั้งหมด จะทำให้ผลการแบ่งกลุ่มข้อมูลเกิดความผิดพลาด กล่าวคือ ข้อมูลที่อยู่ในกลุ่มส่วนน้อยถูกครอบงำหรือจะถูกจัดให้ไปอยู่ในกลุ่มส่วนมากทั้งหมด ซึ่งจะนำไปสู่ปัญหาที่เรียกว่าปัญหาการจำแนกข้อมูลผิดกลุ่ม (misclassification) (เบญจภรณ์ และคณะ, 2559)

วีระยุทธ และคณะ (2557) ศึกษางานวิจัยเรื่อง การพัฒนาแบบจำลองเพื่อการพยากรณ์การรักษาซ้ำของผู้ป่วยโรคจิตเภทโดยเทคนิคเหมืองข้อมูล พบว่าข้อมูลการรักษาของผู้ป่วยโรคจิตเภททางการแพทย์มีข้อมูลที่มีความผิดปกติและมีความไม่สมดุลของข้อมูล จึงปรับความสมดุล หลังจากนั้นนำข้อมูลที่สมดุลแล้วมาจำแนก พบว่าวิธีการกรองแล้วใช้วิธีการสุ่มเกินโดยเทคนิค SMOTE สามารถเพิ่มประสิทธิภาพให้แบบจำลองเพิ่มขึ้น โดยมีค่าความถูกต้อง (accuracy) เพิ่มขึ้นเฉลี่ยร้อยละ 46.36 ค่าความไว (sensitivity) เพิ่มขึ้นเฉลี่ยร้อยละ 20.05

และค่าความจำเพาะ (specificity) เพิ่มขึ้นร้อยละ 32.69 ซึ่งมากกว่าวิธีการกรองแล้วใช้วิธีการสุ่มลด เชาวนันท์ และคณะ (2556) ศึกษางานวิจัย เรื่อง แบบจำลองการทำนายผลการรักษาผู้ป่วยมะเร็งปากมดลูกด้วยวิธีโครงข่ายประสาทเทียม เมื่อเปรียบเทียบประสิทธิภาพการทำนาย พบว่าวิธีโครงข่ายประสาทเทียมที่มีการแก้ปัญหาความไม่สมดุลของข้อมูลด้วยวิธีการสุ่มเกินโดยเทคนิค SMOTE มีประสิทธิภาพการทำนายด้วยมีค่าความถูกต้อง 81.71 % ค่าความไว 94.47 % และค่าความจำเพาะ 55.47 % สูงกว่าวิธีโครงข่ายประสาทเทียมที่มีการแก้ปัญหาความไม่สมดุลของข้อมูลด้วยวิธีการเรียนรู้แบบมีค่าใช้จ่าย ส่วนเบญจภรณ์ และคณะ (2559) ศึกษางานวิจัยเกี่ยวกับเรื่องวิธีการที่เหมาะสมสำหรับการแบ่งกลุ่มข้อมูลที่ไม่สมดุลสูง โดยส่วนมากของการจำแนกชุดข้อมูลของวิธีการสุ่มผสมผสานดีกว่าการใช้ข้อมูลเดิมที่ไม่ได้ปรับปรุง นอกจากนี้ McCarthy และคณะ (2005) ศึกษา เรื่อง การทดสอบความสามารถในการจำแนกข้อมูลที่ค้นพบได้ยากด้วยขั้นตอนวิธี (algorithm) วิธีการเรียนรู้แบบมีค่าใช้จ่าย และวิธีการสุ่ม 2 วิธี คือ วิธีการสุ่มเกินและวิธีการสุ่มลด พบว่าวิธีการเรียนรู้แบบมีค่าใช้จ่ายมีประสิทธิภาพดีกว่าวิธีการสุ่มเกินและวิธีการสุ่มลด

ดังนั้นงานวิจัยนี้จึงศึกษาการปรับความไม่สมดุลของข้อมูลด้วยวิธีต่าง ๆ 4 วิธี คือ วิธีการสุ่มเกิน วิธีการสุ่มเกินโดยใช้เทคนิค SMOTE วิธีการสุ่มลด และวิธีการสุ่มผสมผสาน โดยจำแนกด้วยวิธีต่าง ๆ 4 วิธี คือ วิธีเพื่อนบ้านใกล้สุด k ตัว วิธีต้นไม้ตัดสินใจ วิธีโครงข่ายประสาทเทียม และวิธีซัพพอร์ตเวกเตอร์แมชชีน เพื่อต้องการเปรียบเทียบประสิทธิภาพในการจำแนกข้อมูลทั้ง 4 วิธี ว่าวิธีใดมีประสิทธิภาพและเหมาะสมกับรูปแบบของชุดข้อมูล โดยใช้การทดสอบเปรียบเทียบประสิทธิภาพด้วยค่าความถูกต้อง ค่าความไว ค่าความจำเพาะ และค่าคลาดเคลื่อนกำลังสองเฉลี่ย

2. วิธีการวิจัย

2.1 เครื่องมือที่ใช้ในงานวิจัย

เครื่องมือที่ใช้ในการวิจัยครั้งนี้ ใช้โปรแกรม WEKA 3.9.2 และ Microsoft Excel เวอร์ชั่น 2016

2.2 การเก็บรวบรวมข้อมูล

ค้นหาและการศึกษาข้อมูลที่ไม่สมดุลจากเว็บไซต์ Vincentarelbundock จำนวน 2 ชุด และจากเว็บไซต์ UCI จำนวน 1 ชุด ดังนี้

2.2.1 การรับรู้ทางหูของเด็กที่มีภาวะน้ำคั่งในหูชั้นกลางหรือหูน้ำหนวก (auditory perception in children with otitis media with effusion, OME) จำนวนข้อมูลทั้งหมด 890 ค่า พบค่าในกลุ่มส่วนมากได้ยินเสียงไม่ปกติจำนวน 850 ค่า คิดเป็นร้อยละ 95.51 และค่าในกลุ่มส่วนน้อยได้ยินเสียงปกติจำนวน 40 ค่า คิดเป็นร้อยละ 4.49 (Hogan, 2013)

2.2.2 ยอดคงเหลือในบัตรเครดิตของลูกค้า (credit card balance, credit) จำนวนข้อมูลทั้งหมด 400 ค่า พบค่าในกลุ่มส่วนมากไม่ชำระเงินจำนวน 360 ค่า คิดเป็นร้อยละ 90 และค่าในกลุ่มส่วนน้อยชำระเงินจำนวน 40 ค่า คิดเป็นร้อยละ 10 (James *et al*., 2013)

2.2.3 คุณภาพไวน์แดง (red wine quality, red wine) จำนวนข้อมูลทั้งหมด 999 ค่า พบค่าในกลุ่มส่วนมากคุณภาพดีจำนวน 804 ค่า คิดเป็นร้อยละ 80.48 และค่าในกลุ่มส่วนน้อยคุณภาพไม่ดีจำนวน 195 ค่า คิดเป็นร้อยละ 19.52 (Cortez, 2009)

2.3 การปรับข้อมูลให้มีความสมดุล

2.3.1 วิธีการสุ่มเกิน (over sampling) คือ การสุ่มข้อมูลในกลุ่มส่วนน้อยเพื่อสร้างข้อมูลใหม่ของกลุ่ม ส่วนน้อยให้มีจำนวนเพิ่มมากขึ้นให้

ใกล้เคียงหรือเท่ากับจำนวนในกลุ่มส่วนมาก โดยจะใช้วิธีการสุ่มกลุ่มตัวอย่างอย่างง่าย (simple random sampling) การศึกษาของกีระชาติ (2559) เกี่ยวกับชุดข้อมูลผู้ป่วยเป็นเนื้อร้ายได้ใช้การปรับความไม่สุมดุล 4 วิธี คือ วิธีการสุ่มเกิน วิธีการสุ่มเกินโดยเทคนิค SMOTE วิธีการสุ่มลด วิธีการสุ่มผสมผสาน วิธีที่ให้ผลการปรับดีที่สุด คือ วิธีการสุ่มเกินโดยเทคนิค SMOTE

2.3.2 วิธีการสุ่มเกินโดยเทคนิค SMOTE (synthetic minority over- sampling technique) คือ การสุ่มสร้างข้อมูลจากกลุ่มส่วนน้อยตามจำนวนที่กำหนด โดยการวัดระยะห่างจากจุดข้อมูลตัวอย่างไปยังจุดข้อมูลใกล้เคียง แล้วสุ่มสร้างข้อมูลสังเคราะห์ขึ้นให้ใกล้เคียงกับกลุ่มส่วนมาก การศึกษาของภรันยา (2559) เกี่ยวกับสัดส่วนการใช้อินเทอร์เน็ตสูงที่สุดในเยาชนอายุ 15-24 ปี โดยปรับความสมดุลของข้อมูลด้วยวิธีการสุ่มเกินเทคนิค SMOTE แล้วใช้การจำแนกด้วยวิธีต้นไม้ตัดสินใจ โดยใช้ขั้นตอนวิธี J48, ID3, LMT, CART และ random forest พบว่า ขั้นตอนวิธี random forest ดีกว่าขั้นตอนวิธี J48, ID3, LMT และ CART

2.3.3 วิธีการสุ่มลด (under sampling) คือ การปรับข้อมูลให้มีความสมดุลด้วยวิธีการสุ่มลดจำนวนข้อมูลจากกลุ่มส่วนมากลง เพื่อทำให้จำนวนข้อมูลระหว่างกลุ่มส่วนมาก และกลุ่มส่วนน้อยมีจำนวนใกล้เคียงกันมากขึ้น (กีระชาติ, 2559)

2.3.4 วิธีการสุ่มผสมผสาน (hybrid method) คือ การนำวิธีสุ่มเกินและวิธีสุ่มลดมาทำงานร่วมกัน โดยการใช้วิธีนี้จะเป็นการสุ่มลดจำนวนข้อมูลจากกลุ่มส่วนมาก และสุ่มเพิ่มข้อมูลในกลุ่มส่วนน้อย ให้จำนวนข้อมูลจากทั้ง 2 กลุ่ม มีจำนวนใกล้เคียงกันหรือเท่ากัน (กีระชาติ, 2559)

2.4 วิธีการจำแนกข้อมูล

แบ่งชุดข้อมูลโดยโปรแกรม WEKA เวอร์ชัน 3.9.2 ในอัตราส่วน 70 : 20 : 10 ส่วนที่ 1 ข้อมูลเรียนรู้ (training data) นำไปสร้างตัวแบบ (model) ร้อยละ 70 ข้อมูลส่วนที่ 2 ข้อมูลตรวจสอบความถูกต้อง (validation data) นำไปประเมินความผิดพลาดของตัวแบบร้อยละ 20 และข้อมูลส่วนที่ 3 ข้อมูลทดสอบ (testing data) นำไปทดสอบตัวแบบร้อยละ 10 (พนิดา และคณะ, 2560) โดยกำหนดตัวสร้างเลขสุ่มเทียม (random seed) 10, 20, 30, 40 และ 50

2.4.1 ผลการแบ่งข้อมูลหลังการปรับความไม่สมดุล

จำนวนข้อมูลก่อนปรับ หลังปรับ และร้อยละหลังปรับความสมดุล ดังตารางที่ 1

2.5 การศึกษาขั้นตอนวิธี (algorithm)

2.5.1 วิธีเพื่อนบ้านใกล้สุด k ตัว (k-nearest neighbor) เป็นวิธีไม่มีการสร้างตัวแบบจากข้อมูลเรียนรู้เก็บไว้ ทำนายข้อมูลใหม่โดยอาศัยการเปรียบเทียบกับข้อมูลเรียนรู้จำนวน k ตัว (ในที่นี้ k = 1) ที่อยู่ใกล้เคียงกันมากที่สุด ใช้คำตอบของข้อมูลฝึกหัดที่อยู่ใกล้เคียงกันมากที่สุด k ตัว ที่พบมากที่สุดเป็นคำตอบ โดยงานวิจัยนี้ใช้ขั้นตอนวิธี IBk (สายชล, 2560)

2.5.2 วิธีต้นไม้ตัดสินใจ (decision tree) เป็นตัวแบบทางคณิตศาสตร์เพื่อหาทางเลือกที่ดีที่สุด โดยการนำข้อมูลมาสร้างตัวแบบการพยากรณ์ในรูปแบบของโครงสร้างต้นไม้ซึ่งมีการเรียนรู้ข้อมูลแบบมีผู้สอน โดยงานวิจัยนี้ใช้ขั้นตอนวิธี J48 (C4.5) (สายชล, 2560)

2.5.3 วิธีโครงข่ายประสาทเทียม (artificial neural network) มีหลักการเลียนแบบการทำงานของสมองมนุษย์ เส้นเชื่อมแต่ละเส้นจะมีค่าถ่วงน้ำหนัก (weight) เพื่อใช้กำหนดค่าถ่วงน้ำหนักหรือความสำคัญของข้อมูลเข้า กำหนดค่าเริ่มต้นโดยการสุ่ม ในแต่ละโหนดจะคำนวณค่า

ผลรวมเชิงเส้นแบบถ่วงน้ำหนักและผ่านฟังก์ชันกระตุ้น ใช้ขั้นตอนวิธีชนิดเพอร์เซปตรอนหลายชั้น (multilayer perceptron) โดยกำหนดค่าอัตราการเรียนรู้ (learning rate) เป็น 0.1 ค่าโมเมนตัม (momentum) เป็น 0.9 จำนวนรอบการสอน (training time) 20,000 รอบ การวิจัยครั้งนี้ใช้ขั้นตอนวิธีของวิธีโครงข่ายประสาทเทียมชนิดเพอร์เซปตรอนหลายชั้นที่มีชั้นซ่อน (hidden layer) 1 ชั้น (สายชล, 2560)

ตารางที่ 1 ผลการแบ่งข้อมูลหลังการปรับความไม่สมดุลเป็น 3 ส่วน

วิธีการปรับ		ชุดข้อมูล		
		OME	Credit	Red wine
วิธีการสุ่มเกิน	ก่อนปรับ	890	400	999
	หลังปรับ	1650	710	1594
	ร้อยละ 70	1155	497	1115
	ร้อยละ 20	330	142	319
	ร้อยละ 10	165	71	160
วิธีการสุ่มเกินโดยเทคนิค SMOTE	ก่อนปรับ	890	400	999
	หลังปรับ	1690	720	1584
	ร้อยละ 70	1183	504	1108
	ร้อยละ 20	338	144	317
	ร้อยละ 10	169	72	159
วิธีการสุ่มลด	ก่อนปรับ	890	400	999
	หลังปรับ	100	100	390
	ร้อยละ 70	70	70	273
	ร้อยละ 20	20	20	78
	ร้อยละ 10	10	10	39
วิธีการสุ่มผสมผสาน	ก่อนปรับ	890	400	999
	หลังปรับ	890	400	998
	ร้อยละ 70	623	280	698
	ร้อยละ 20	178	80	200
	ร้อยละ 10	89	40	100

2.5.4 วิธีซัพพอร์ตเวกเตอร์แมชชีน (support vector machine) เป็นกระบวนการสอนเครื่องแบบมีผู้สอนเพื่อให้สามารถสร้างตัวจำแนกข้อมูลที่มีลักษณะทั่วไปสูง นั่นคือ สามารถทำงานได้ดีกับตัวอย่างที่ไม่รู้จัก ด้วยกระบวนการปรับรูปแบบจากข้อมูลที่มีมิติต่ำ ให้อยู่ในรูปแบบข้อมูลที่มีมิติสูง โดยงานวิจัยนี้ได้ใช้ขั้นตอนวิธี SMO ชนิดโพลิโนเมียลเคอร์เนล (polynomial Kernel) (สายชล, 2560)

2.6 การเปรียบเทียบประสิทธิภาพของวิธีการจำแนก

เมทริกซ์ความสับสน (confusion matrix) คือ ตารางสรุปจำนวนข้อมูลที่ตัวแบบมีการจำแนกได้ถูกต้องและไม่ถูกต้อง ดังแสดงในตารางที่ 2

ตารางที่ 2 เมทริกซ์ความสับสน

ค่าจริง	ค่าทำนาย	
	คำตอบเป็นบวก	คำตอบที่เป็นลบ
คำตอบเป็นบวก	TP	FN
คำตอบที่เป็นลบ	FP	TN

โดย บวกจริง (true positive, TP) คือ ค่าความถูกต้องในการจำแนกข้อมูล ซึ่งมีค่าที่แท้จริงอยู่ในกลุ่มบวก และผลการทำนายว่าอยู่ในกลุ่มบวก

ลบจริง (true negative, TN) คือ ค่าความถูกต้องในการจำแนกข้อมูล ซึ่งมีค่าที่แท้จริงอยู่ในกลุ่มลบ และผลการทำนายว่าอยู่ในกลุ่มลบ

บวกเท็จ (false positive, FP) คือ ค่าความผิดพลาดในการจำแนกข้อมูล ซึ่งมีค่าที่แท้จริงอยู่ในกลุ่มลบ แต่ผลการทำนายว่าอยู่ในกลุ่มบวก

ลบเท็จ (false negative, FN) คือ ค่าความผิดพลาดในการจำแนกข้อมูล ซึ่งมีค่าที่แท้จริงอยู่ในกลุ่มบวก แต่ผลการทำนายว่าอยู่ในกลุ่มลบ

ตารางที่ 2 สามารถนำข้อมูลในตารางมาใช้ในการคำนวณการวัดประสิทธิภาพของตัวแบบการทำนายดังนี้

2.6.1 ค่าความถูกต้อง (accuracy) คือ การแสดงการวัดที่ได้มีความถูกต้องในรูปอัตราส่วน (สุรวัชร และสายชล, 2560)

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

2.6.2 ค่าความไว คือ สัดส่วนของผลบวกที่เป็นจริงสำหรับภาวะนั้น ๆ (กีระชาติ, 2559)

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

2.6.3 ค่าความจำเพาะ คือ สัดส่วนของผลลบที่เป็นจริงสำหรับภาวะนั้น ๆ (กีระชาติ, 2559)

$$TNR = \frac{TN}{TN + FP}$$

2.6.4 ค่าคลาดเคลื่อนกำลังสองเฉลี่ยเป็นมาตรวัดการประเมินค่าได้ดี เนื่องจากค่าคลาดเคลื่อนกำลังสองเฉลี่ยประกอบด้วยทั้งความเอนเอียงและความแปรปรวน (พนิดา และคณะ, 2560)

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

โดย y_i คือ ค่าจริง และ $\hat{y}_i$ คือ ค่าทำนาย

3. ผลการวิจัย

3.1 ชุดข้อมูลการรับรู้ทางหูของเด็กที่มีภาวะน้ำคั่งในหูชั้นกลางหรือหูน้ำหนวก

ตารางที่ 3 แสดงผลการวิเคราะห์ชุดข้อมูลการรับรู้ทางหูของเด็กที่มีภาวะน้ำคั่งในหูชั้นกลางหรือหูน้ำหนวก โดยการจำแนกข้อมูลและการปรับความไม่สมดุลของข้อมูล เมื่อ random seed 10, 20, 30, 40 และ 50 พบว่าวิธีเพื่อนบ้านใกล้สุด k ตัวโดยการปรับความไม่สมดุลของข้อมูลด้วยวิธีการสุ่มเกินเทคนิค SMOTE ให้ค่าความถูกต้องเฉลี่ยสูงสุด คือ 0.8379 ให้ค่าความไวเฉลี่ยสูงสุด คือ 0.8362 ให้ค่าความจำเพาะเฉลี่ยสูงสุด คือ 0.8396 และให้ค่าเฉลี่ยของค่าคลาดเคลื่อนกำลังสองเฉลี่ยต่ำสุด คือ 0.1543

วิธีต้นไม้ตัดสินใจโดยการปรับความไม่สมดุลของข้อมูลด้วยวิธีการสุ่มเกินเทคนิค SMOTE ให้ค่าความถูกต้องเฉลี่ยสูงสุด คือ 0.8331 ให้ค่าความไวเฉลี่ยสูงสุด คือ 0.8323 ให้ค่าความจำเพาะเฉลี่ยสูงสุด คือ 0.8329 และให้ค่าเฉลี่ยของค่าคลาดเคลื่อนกำลังสองเฉลี่ยต่ำสุด คือ 0.1228

วิธีโครงข่ายประสาทเทียมโดยการปรับความไม่สมดุลของข้อมูลด้วยด้วยวิธีการสุ่มเกินเทคนิค SMOTE ให้ค่าความถูกต้องเฉลี่ยสูงสุด คือ 0.8355 ให้ค่าความไวเฉลี่ยสูงสุด คือ 0.8157 ให้ค่าความจำเพาะเฉลี่ยสูงสุด คือ 0.8571 และให้ค่าเฉลี่ยของค่าคลาดเคลื่อนกำลังสองเฉลี่ยต่ำสุด คือ 0.1303

วิธีซัพพอร์ตเวกเตอร์แมชชีนโดยการปรับความไม่สมดุลของข้อมูลด้วยวิธีการสุ่มเกินด้วยเทคนิค SMOTE ให้ค่าความถูกต้องเฉลี่ยสูงสุด คือ 0.8509 ให้ค่าความไวเฉลี่ยสูงสุด คือ 0.8210 ให้ค่าความจำเพาะเฉลี่ยสูงสุด คือ 0.8838 และให้ค่าเฉลี่ยของค่าคลาดเคลื่อนกำลังสองเฉลี่ยต่ำสุด คือ 0.1490

3.2 ชุดข้อมูลยอดคงเหลือในบัตรเครดิตของลูกค้า

ตารางที่ 4 แสดงผลการวิเคราะห์ชุดข้อมูลยอดคงเหลือในบัตรเครดิตของลูกค้า โดยการจำแนกข้อมูลและการปรับความไม่สมดุลของข้อมูล เมื่อ random seed 10, 20, 30, 40 และ 50 พบว่าวิธีเพื่อนบ้านใกล้สุด k ตัว โดยการปรับความไม่สมดุลของข้อมูลด้วยวิธีการสุ่มเกินเทคนิค SMOTE ให้ค่าความถูกต้องเฉลี่ยสูงสุด คือ 0.7778 ให้ค่าความไวเฉลี่ยสูงสุด คือ 0.7834 ให้ค่าความจำเพาะเฉลี่ยสูงสุด คือ 0.7680 และให้ค่าเฉลี่ยของค่าคลาดเคลื่อนกำลังสองเฉลี่ยต่ำสุด คือ 0.2157

วิธีต้นไม้ตัดสินใจโดยการปรับความไม่สมดุลของข้อมูลด้วยวิธีการสุ่มเกินเทคนิค SMOTE ให้ค่าความถูกต้องเฉลี่ยสูงสุด คือ 0.7556 ให้ค่าความไวเฉลี่ยสูงสุด คือ 0.7551 ให้ค่าความจำเพาะเฉลี่ยสูงสุด คือ 0.7497 และให้ค่าเฉลี่ยของค่าคลาดเคลื่อนกำลังสองเฉลี่ยต่ำสุด คือ 0.2114

วิธีโครงข่ายประสาทเทียมโดยการปรับความไม่สมดุลของข้อมูลด้วยวิธีการสุ่มเกินเทคนิค SMOTE ให้ค่าความถูกต้องเฉลี่ยสูงสุด คือ 0.7278 ให้ค่าความไวเฉลี่ยสูงสุด คือ 0.7298 ให้ค่าความจำเพาะเฉลี่ยสูงสุด คือ 0.7262 และให้ค่าเฉลี่ยของค่าคลาดเคลื่อนกำลังสองเฉลี่ยต่ำสุด คือ 0.2291

ตารางที่ 3 ผลการวิเคราะห์ข้อมูลการรับรู้ทางหูของเด็กที่มีภาวะน้ำคั่งในหูชั้นกลางหรือหูน้ำหนวก เมื่อ random seed 10, 20, 30, 40 และ 50 โดยใช้ข้อมูลส่วนที่ 3 ข้อมูลทดสอบร้อยละ 10

วิธีการจำแนก	วิธีการปรับความไม่สมดุลของข้อมูล	ค่าความถูกต้อง	ค่าความไว	ค่าความจำเพาะ	ค่าคลาดเคลื่อนกำลังสองเฉลี่ย
Random seed 10					
เพื่อนบ้านใกล้สุด k ตัว	วิธีการสุ่มเกิน	0.7697	0.8919	0.6703	0.2229
	วิธีการสุ่มเกินเทคนิค SMOTE	0.8639	0.8651	0.8625	0.1316
	วิธีการสุ่มลด	0.5000	0.6000	0.4000	0.4173
	วิธีการสุ่มผสมผสาน	0.7079	0.6522	0.7674	0.2803
ต้นไม้ตัดสินใจ	วิธีการสุ่มเกิน	0.6667	0.6757	0.6593	0.2395
	วิธีการสุ่มเกินเทคนิค SMOTE	**0.8817**	**0.8989**	0.8625	**0.1046**
	วิธีการสุ่มลด	0.5000	0.6000	0.4000	0.4028
	วิธีการสุ่มผสมผสาน	0.6567	0.6522	0.6511	0.2606
โครงข่ายประสาทเทียม	วิธีการสุ่มเกิน	0.6465	0.8649	0.4725	0.2228
	วิธีการสุ่มเกินเทคนิค SMOTE	0.8462	0.7303	0.9750	0.1223
	วิธีการสุ่มลด	0.8000	0.6000	**1.0000**	0.1944
	วิธีการสุ่มผสมผสาน	0.6630	0.5217	0.8140	0.2272
ซัพพอร์ตเวกเตอร์แมชชีน	วิธีการสุ่มเกิน	0.6485	0.7838	0.5385	0.3515
	วิธีการสุ่มเกินเทคนิค SMOTE	0.8343	0.6966	0.9875	0.1656
	วิธีการสุ่มลด	0.5000	0.6000	0.4000	0.4999
	วิธีการสุ่มผสมผสาน	0.5955	0.7609	0.4186	0.4045
Random seed 20					
เพื่อนบ้านใกล้สุด k ตัว	วิธีการสุ่มเกิน	0.7454	0.7303	0.7632	0.1887
	วิธีการสุ่มเกินเทคนิค SMOTE	0.8639	**0.8795**	0.8488	0.1322
	วิธีการสุ่มลด	0.8000	0.8571	0.6667	0.1719
	วิธีการสุ่มผสมผสาน	0.6854	0.7872	0.5714	0.2811

ตารางที่ 3 (ต่อ)

วิธีการจำแนก	วิธีการปรับความไม่สมดุลของข้อมูล	ค่าความถูกต้อง	ค่าความไว	ค่าความจำเพาะ	ค่าคลาดเคลื่อนกำลังสองเฉลี่ย
ต้นไม้ตัดสินใจ	วิธีการสุ่มเกิน	0.6788	0.6629	0.6974	0.2410
	วิธีการสุ่มเกินเทคนิค SMOTE	0.7929	0.7470	0.8372	0.1427
	วิธีการสุ่มลด	0.5000	0.7143	0.0000	0.4605
	วิธีการสุ่มผสมผสาน	0.7303	0.7021	0.7619	0.2124
โครงข่ายประสาทเทียม	วิธีการสุ่มเกิน	0.6182	0.5843	0.6579	0.2430
	วิธีการสุ่มเกินเทคนิค SMOTE	0.8521	0.7229	0.9767	0.1176
	วิธีการสุ่มลด	**0.9000**	0.8571	**1.0000**	**0.0991**
	วิธีการสุ่มผสมผสาน	0.6630	0.5217	0.8140	0.2272
ซัพพอร์ตเวกเตอร์แมชชีน	วิธีการสุ่มเกิน	0.6727	0.8202	0.5000	0.3273
	วิธีการสุ่มเกินเทคนิค SMOTE	0.8343	0.6867	0.9767	0.1656
	วิธีการสุ่มลด	**0.9000**	0.8570	**1.0000**	0.1000
	วิธีการสุ่มผสมผสาน	0.6067	0.6429	0.6271	0.3933
Random seed 30					
เพื่อนบ้านใกล้สุด k ตัว	วิธีการสุ่มเกิน	0.8060	0.7582	0.8649	0.1676
	วิธีการสุ่มเกินเทคนิค SMOTE	0.8047	0.8023	0.8072	0.1930
	วิธีการสุ่มลด	0.7000	0.2500	**1.0000**	0.2537
	วิธีการสุ่มผสมผสาน	0.6854	0.7143	0.6500	0.2855
ต้นไม้ตัดสินใจ	วิธีการสุ่มเกิน	0.7394	0.7143	0.7703	0.2008
	วิธีการสุ่มเกินเทคนิค SMOTE	0.8284	0.7674	0.8916	**0.1220**
	วิธีการสุ่มลด	0.3000	0.0000	0.5000	0.5236
	วิธีการสุ่มผสมผสาน	0.7865	**0.8163**	0.7500	0.1997
โครงข่ายประสาทเทียม	วิธีการสุ่มเกิน	0.7212	0.5604	0.9189	0.1934
	วิธีการสุ่มเกินเทคนิค SMOTE	0.8166	0.7442	0.8916	0.1422
	วิธีการสุ่มลด	0.8000	0.5000	**1.0000**	0.1987
	วิธีการสุ่มผสมผสาน	0.5955	0.7551	0.4000	0.2698
ซัพพอร์ตเวกเตอร์แมชชีน	วิธีการสุ่มเกิน	0.5879	0.7253	0.4189	0.4122
	วิธีการสุ่มเกินเทคนิค SMOTE	**0.8580**	0.7326	0.9880	0.1420
	วิธีการสุ่มลด	0.7000	0.5000	0.8333	0.3000
	วิธีการสุ่มผสมผสาน	0.6517	0.7755	0.5000	0.3483

ตารางที่ 3 (ต่อ)

วิธีการจำแนก	วิธีการปรับความไม่สมดุลของข้อมูล	ค่าความถูกต้อง	ค่าความไว	ค่าความจำเพาะ	ค่าคลาดเคลื่อนกำลังสองเฉลี่ย
Random seed 40					
เพื่อนบ้านใกล้สุด k ตัว	วิธีการสุ่มเกิน	0.7152	0.7380	0.6670	0.2636
	วิธีการสุ่มเกินเทคนิค SMOTE	0.8580	0.8430	0.8720	0.1162
	วิธีการสุ่มลด	0.5000	0.6000	0.4000	0.4167
	วิธีการสุ่มผสมผสาน	0.6629	0.7670	0.5650	0.3085
ต้นไม้ตัดสินใจ	วิธีการสุ่มเกิน	0.6606	0.5220	0.7600	0.2288
	วิธีการสุ่มเกินเทคนิค SMOTE	0.8106	0.7950	0.8260	0.1253
	วิธีการสุ่มลด	0.5000	0.6000	0.4000	0.5001
	วิธีการสุ่มผสมผสาน	0.7415	0.7440	0.7390	0.1966
โครงข่ายประสาทเทียม	วิธีการสุ่มเกิน	0.6424	0.7830	0.5420	0.2233
	วิธีการสุ่มเกินเทคนิค SMOTE	0.8461	0.9160	0.7790	0.1285
	วิธีการสุ่มลด	0.4000	0.6000	0.2000	0.5826
	วิธีการสุ่มผสมผสาน	0.6741	0.7210	0.6300	0.2406
ซัพพอร์ตเวกเตอร์แมชชีน	วิธีการสุ่มเกิน	0.6970	0.4490	0.8750	0.3031
	วิธีการสุ่มเกินเทคนิค SMOTE	0.8757	0.9880	0.7670	0.1242
	วิธีการสุ่มลด	0.5000	0.6000	0.4000	0.4999
	วิธีการสุ่มผสมผสาน	0.6292	0.4880	0.7610	0.3708
Random seed 50					
เพื่อนบ้านใกล้สุด k ตัว	วิธีการสุ่มเกิน	0.6848	0.6560	0.7220	0.2371
	วิธีการสุ่มเกินเทคนิค SMOTE	0.7988	0.7910	0.8070	0.1988
	วิธีการสุ่มลด	0.4000	0.0000	0.5710	0.4991
	วิธีการสุ่มผสมผสาน	0.7079	0.7070	0.7080	0.2479
ต้นไม้ตัดสินใจ	วิธีการสุ่มเกิน	0.6303	0.6130	0.6530	0.2481
	วิธีการสุ่มเกินเทคนิค SMOTE	0.8520	0.9530	0.7470	0.1192
	วิธีการสุ่มลด	0.6000	0.0000	0.8570	0.2731
	วิธีการสุ่มผสมผสาน	0.0065	0.5120	0.7710	0.2361
โครงข่ายประสาทเทียม	วิธีการสุ่มเกิน	0.5879	0.4840	0.7220	0.2404
	วิธีการสุ่มเกินเทคนิค SMOTE	0.8165	0.9650	0.6630	0.1410
	วิธีการสุ่มลด	0.1000	0.1430	0.1000	0.7588
	วิธีการสุ่มผสมผสาน	0.6067	0.3410	0.8330	0.2556

ตารางที่ 3 (ต่อ)

วิธีการจำแนก	วิธีการปรับความไม่สมดุลของข้อมูล	ค่าความถูกต้อง	ค่าความไว	ค่าความจำเพาะ	ค่าคลาดเคลื่อนกำลังสองเฉลี่ย
ซัพพอร์ตเวกเตอร์แมชชีน	วิธีการสุ่มเกิน	0.5758	0.7310	0.3750	0.4242
	วิธีการสุ่มเกินเทคนิค SMOTE	0.8521	1.0000	0.6990	0.1479
	วิธีการสุ่มลด	0.7000	0.0000	1.0000	0.2999
	วิธีการสุ่มผสมผสาน	0.5056	0.5610	0.4580	0.4943
ค่าเฉลี่ย					
เพื่อนบ้านใกล้สุด k ตัว	วิธีการสุ่มเกิน	0.7378	0.7619	0.7177	0.2211
	วิธีการสุ่มเกินเทคนิค SMOTE	**0.8379**	**0.8362***	**0.8396**	**0.1543**
	วิธีการสุ่มลด	0.5800	0.4614	0.6075	0.3517
	วิธีการสุ่มผสมผสาน	0.5892	0.611	0.557	0.3872
ต้นไม้ตัดสินใจ	วิธีการสุ่มเกิน	0.6752	0.6375	0.708	0.2316
	วิธีการสุ่มเกินเทคนิค SMOTE	**0.8331**	**0.8323**	**0.8329**	**0.1228***
	วิธีการสุ่มลด	0.4800	0.3829	0.4314	0.4320
	วิธีการสุ่มผสมผสาน	0.7123	0.6853	0.7346	0.2211
โครงข่ายประสาทเทียม	วิธีการสุ่มเกิน	0.6436	0.6553	0.6627	0.2246
	วิธีการสุ่มเกินเทคนิค SMOTE	**0.8355**	**0.8157**	**0.8571**	**0.1303**
	วิธีการสุ่มลด	0.7148	0.6556	0.7660	0.2631
	วิธีการสุ่มผสมผสาน	0.5142	0.4756	0.5484	0.3551
ซัพพอร์ตเวกเตอร์แมชชีน	วิธีการสุ่มเกิน	0.6364	0.7019	0.5415	0.3636
	วิธีการสุ่มเกินเทคนิค SMOTE	**0.8509***	**0.8210**	**0.8838***	**0.1490**
	วิธีการสุ่มลด	0.6600	0.5114	0.7267	0.3399
	วิธีการสุ่มผสมผสาน	0.5720	0.468	0.6220	0.4280

ตัวทึบ หมายถึง ค่าสูงสุดของค่าความถูกต้อง ค่าความไว ค่าความจำเพาะ และค่าต่ำสุดของค่าคลาดเคลื่อนกำลังสองเฉลี่ย; * หมายถึง ค่าความถูกต้อง ค่าความไว ค่าความจำเพาะเฉลี่ยสูงสุด และค่าเฉลี่ยของค่าคลาดเคลื่อนกำสองเฉลี่ยต่ำสุดของทั้ง 4 การจำแนก

วิธีซัพพอร์ตเวกเตอร์แมชชีนโดยการปรับความไม่สมดุลของข้อมูลด้วยวิธีการสุ่มเกินเทคนิค SMOTE ให้ค่าความถูกต้องเฉลี่ยสูงสุด คือ 0.7722 ให้ค่าความไวเฉลี่ยสูงสุด คือ 0.7706 ให้ค่าความจำเพาะเฉลี่ยสูงสุด คือ 0.7742 และให้ค่าเฉลี่ยของค่าคลาดเคลื่อนกำลังสองเฉลี่ยต่ำสุด คือ 0.2278

3.3 ชุดข้อมูลคุณภาพไวน์แดง

ตารางที่ 5 แสดงผลการวิเคราะห์ชุดข้อมูลคุณภาพไวน์แดงโดยการจำแนกข้อมูลและการปรับความไม่สมดุลของข้อมูล เมื่อ random seed 10, 20, 30, 40 และ 50 พบว่าวิธีเพื่อนบ้านใกล้สุด k ตัว โดยการปรับความไม่สมดุลของข้อมูลด้วยวิธีการสุ่มเกินให้ค่าความถูกต้องเฉลี่ยสูงสุด คือ 0.6200 ให้ค่าความไวเฉลี่ยสูงสุด คือ 0.6409 และให้ค่าเฉลี่ยของค่าคลาดเคลื่อนกำลังสองเฉลี่ยต่ำสุด คือ 0.3747 ส่วนการปรับความไม่สมดุลของข้อมูลด้วยวิธีการสุ่ม

ผสมผสานให้ค่าความจำเพาะเฉลี่ยสูงสุด คือ 0.6595

วิธีต้นไม้ตัดสินใจโดยการปรับความไม่สมดุลของข้อมูลด้วยวิธีการสุ่มเกินเทคนิค SMOTE ให้ค่าความถูกต้องเฉลี่ยสูงสุด คือ 0.5572 ให้ค่าความไวเฉลี่ยสูงสุด คือ 0.6403 และให้ค่าเฉลี่ยของค่าคลาดเคลื่อนกำลังสองเฉลี่ยต่ำสุด คือ 0.3105 ส่วนการปรับความไม่สมดุลของข้อมูลด้วยวิธีการสุ่มผสมผสานให้ค่าความจำเพาะเฉลี่ยสูงสุด คือ 0.6167

วิธีโครงข่ายประสาทเทียมโดยการปรับความไม่สมดุลของข้อมูลด้วยวิธีการสุ่มเกินให้ค่าความถูกต้องเฉลี่ยสูงสุด คือ 0.5888 ให้ค่าความจำเพาะเฉลี่ยสูงสุด คือ 0.6703 และให้ค่าเฉลี่ยของค่าคลาดเคลื่อนกำลังสองเฉลี่ยต่ำสุด คือ 0.2861 ส่วนการปรับความไม่สมดุลของข้อมูลด้วยวิธีการสุ่มลดให้ค่าความไวเฉลี่ยสูงสุด คือ 0.6772

วิธีซัพพอร์ตเวกเตอร์แมชชีนโดยการปรับความไม่สมดุลของข้อมูลด้วยด้วยวิธีการสุ่มลดให้ค่าความถูกต้องเฉลี่ยสูงสุด คือ 0.6513 ให้ค่าความไวเฉลี่ยสูงสุด คือ 0.6500 และให้ค่าเฉลี่ยของค่าคลาดเคลื่อนกำลังสองเฉลี่ยต่ำสุด คือ 0.3487 ส่วนการปรับความไม่สมดุลของข้อมูลด้วยวิธีการสุ่มผสมผสานให้ค่าความจำเพาะเฉลี่ยสูงสุด คือ 0.5805

ตารางที่ 4 ผลการวิเคราะห์ชุดข้อมูลยอดคงเหลือในบัตรเครดิตของลูกค้า เมื่อ random seed 10, 20, 30, 40 และ 50 โดยใช้ข้อมูลส่วนที่ 3 ข้อมูลทดสอบร้อยละ 10

วิธีการจำแนก	วิธีการปรับความไม่สมดุลของข้อมูล	ค่าความถูกต้อง	ค่าความไว	ค่าความจำเพาะ	ค่าคลาดเคลื่อนกำลังสองเฉลี่ย
Random seed 10					
เพื่อนบ้านใกล้สุด k ตัว	วิธีการสุ่มเกิน	0.7324	0.7778	0.6856	0.2604
	วิธีการสุ่มเกินเทคนิค SMOTE	0.7630	0.6774	0.8293	0.2292
	วิธีการสุ่มลด	0.5000	0.6000	0.4000	0.4173
	วิธีการสุ่มผสมผสาน	0.4000	0.3158	0.4762	0.5691
ต้นไม้ตัดสินใจ	วิธีการสุ่มเกิน	**0.7887**	**0.8056**	0.7714	**0.1860**
	วิธีการสุ่มเกินเทคนิค SMOTE	0.6944	0.5484	0.8049	0.2735
	วิธีการสุ่มลด	0.6000	0.8000	0.4000	0.3549
	วิธีการสุ่มผสมผสาน	0.4250	0.4210	0.4286	0.4970
โครงข่ายประสาทเทียม	วิธีการสุ่มเกิน	0.6619	0.6111	0.7143	0.2735
	วิธีการสุ่มเกินเทคนิค SMOTE	0.7222	0.5806	0.8293	0.2408
	วิธีการสุ่มลด	0.3000	0.2000	0.4000	0.6374
	วิธีการสุ่มผสมผสาน	0.5000	0.5789	0.4286	0.3750
ซัพพอร์ตเวกเตอร์แมชชีน	วิธีการสุ่มเกิน	0.6338	0.5000	0.7714	0.3661
	วิธีการสุ่มเกินเทคนิค SMOTE	0.7083	0.5161	**0.8537**	0.2917
	วิธีการสุ่มลด	0.5000	0.6000	0.4000	0.4999
	วิธีการสุ่มผสมผสาน	0.5250	0.4211	0.6190	0.4750

ตารางที่ 4 (ต่อ)

วิธีการจำแนก	วิธีการปรับความไม่สมดุลของข้อมูล	ค่าความถูกต้อง	ค่าความไว	ค่าความจำเพาะ	ค่าคลาดเคลื่อนกำลังสองเฉลี่ย
Random seed 20					
เพื่อนบ้านใกล้สุด k ตัว	วิธีการสุ่มเกิน	0.7183	0.8333	0.6342	0.2757
	วิธีการสุ่มเกินเทคนิค SMOTE	0.7083	0.7838	0.6286	0.2831
	วิธีการสุ่มลด	0.6000	0.7142	0.3333	0.3356
	วิธีการสุ่มผสมผสาน	0.6000	0.5714	0.6154	0.3808
ต้นไม้ตัดสินใจ	วิธีการสุ่มเกิน	0.6760	0.7333	0.6341	0.2910
	วิธีการสุ่มเกินเทคนิค SMOTE	0.7222	0.7838	0.6571	0.2375
	วิธีการสุ่มลด	0.5000	0.7142	0.0000	0.3298
	วิธีการสุ่มผสมผสาน	0.7500	0.8462	0.5714	**0.2183**
โครงข่ายประสาทเทียม	วิธีการสุ่มเกิน	0.6619	0.7000	0.6341	0.2536
	วิธีการสุ่มเกินเทคนิค SMOTE	0.6805	0.7568	0.6000	0.2748
	วิธีการสุ่มลด	0.4000	0.4286	0.3333	0.5196
	วิธีการสุ่มผสมผสาน	0.5750	0.4231	0.8571	0.2517
ซัพพอร์ตเวกเตอร์แมชชีน	วิธีการสุ่มเกิน	0.6056	0.4000	0.7561	0.3944
	วิธีการสุ่มเกินเทคนิค SMOTE	**0.7638**	**0.8649**	0.6571	0.2361
	วิธีการสุ่มลด	0.6000	0.8571	0.0000	0.4001
	วิธีการสุ่มผสมผสาน	0.6250	0.1429	**0.8846**	0.3750
Random seed 30					
เพื่อนบ้านใกล้สุด k ตัว	วิธีการสุ่มเกิน	0.7042	0.8372	0.5000	0.2875
	วิธีการสุ่มเกินเทคนิค SMOTE	**0.8333**	0.9231	0.7273	**0.1627**
	วิธีการสุ่มลด	0.5000	0.5000	0.5000	0.4173
	วิธีการสุ่มผสมผสาน	0.4500	0.5909	0.2778	0.5236
ต้นไม้ตัดสินใจ	วิธีการสุ่มเกิน	0.7183	0.8140	0.5714	0.2572
	วิธีการสุ่มเกินเทคนิค SMOTE	0.7639	0.7865	**0.7353**	0.2175
	วิธีการสุ่มลด	0.5000	0.2500	0.6667	0.4445
	วิธีการสุ่มผสมผสาน	0.4250	0.5454	0.2778	0.5010
โครงข่ายประสาทเทียม	วิธีการสุ่มเกิน	0.6901	0.7674	0.5714	0.2428
	วิธีการสุ่มเกินเทคนิค SMOTE	0.7777	0.8684	0.6765	0.1920
	วิธีการสุ่มลด	0.5000	0.5000	0.5000	0.5054
	วิธีการสุ่มผสมผสาน	0.5250	0.6818	0.3333	0.3448

ตารางที่ 4 (ต่อ)

วิธีการจำแนก	วิธีการปรับความไม่สมดุลของข้อมูล	ค่าความถูกต้อง	ค่าความไว	ค่าความจำเพาะ	ค่าคลาดเคลื่อนกำลังสองเฉลี่ย
ซัพพอร์ตเวกเตอร์แมชชีน	วิธีการสุ่มเกิน	0.6619	0.8140	0.4286	0.3380
	วิธีการสุ่มเกินเทคนิค SMOTE	0.8194	**0.9474**	0.6765	0.1805
	วิธีการสุ่มลด	0.3000	0.0000	0.5000	0.7000
	วิธีการสุ่มผสมผสาน	0.5750	0.7273	0.3889	0.4250
Random seed 40					
เพื่อนบ้านใกล้สุด k ตัว	วิธีการสุ่มเกิน	0.5915	0.7030	0.4710	0.3972
	วิธีการสุ่มเกินเทคนิค SMOTE	0.6806	0.6760	0.6840	0.3101
	วิธีการสุ่มลด	0.3000	0.4000	0.2000	0.5809
	วิธีการสุ่มผสมผสาน	0.6000	0.5600	0.6670	0.3814
ต้นไม้ตัดสินใจ	วิธีการสุ่มเกิน	0.6338	0.6760	0.5880	0.2911
	วิธีการสุ่มเกินเทคนิค SMOTE	0.7500	0.8240	0.6840	0.2081
	วิธีการสุ่มลด	0.3000	0.4000	0.2000	0.6389
	วิธีการสุ่มผสมผสาน	0.6250	0.7200	0.4670	0.3031
โครงข่ายประสาทเทียม	วิธีการสุ่มเกิน	0.5211	0.5680	0.4710	0.3395
	วิธีการสุ่มเกินเทคนิค SMOTE	0.7361	0.8240	0.6580	0.2224
	วิธีการสุ่มลด	0.7000	0.800	0.6000	0.2955
	วิธีการสุ่มผสมผสาน	0.7415	0.7440	0.7390	0.1966
ซัพพอร์ตเวกเตอร์แมชชีน	วิธีการสุ่มเกิน	0.5634	0.5950	0.5290	0.4367
	วิธีการสุ่มเกินเทคนิค SMOTE	0.7778	0.8820	0.6840	0.2222
	วิธีการสุ่มลด	0.4000	0.4000	0.4000	0.6000
	วิธีการสุ่มผสมผสาน	0.6500	0.8400	0.3330	0.3500
Random seed 50					
เพื่อนบ้านใกล้สุด k ตัว	วิธีการสุ่มเกิน	0.7042	0.8160	0.5760	0.2883
	วิธีการสุ่มเกินเทคนิค SMOTE	0.8750	0.8330	0.9330	0.1214
	วิธีการสุ่มลด	0.6000	0.3330	0.7140	0.3355
	วิธีการสุ่มผสมผสาน	0.4500	0.5790	0.3330	0.5245
ต้นไม้ตัดสินใจ	วิธีการสุ่มเกิน	0.5634	0.6840	0.4240	0.3864
	วิธีการสุ่มเกินเทคนิค SMOTE	0.8472	0.8330	0.8670	0.1202
	วิธีการสุ่มลด	0.5000	0.0000	0.7140	0.3335
	วิธีการสุ่มผสมผสาน	0.7000	0.8420	0.5710	0.2496

ตารางที่ 4 (ต่อ)

วิธีการจำแนก	วิธีการปรับความไม่สมดุลของข้อมูล	ค่าความถูกต้อง	ค่าความไว	ค่าความจำเพาะ	ค่าคลาดเคลื่อนกำลังสองเฉลี่ย
โครงข่ายประสาทเทียม	วิธีการสุ่มเกิน	0.4789	0.6050	0.3330	0.3428
	วิธีการสุ่มเกินเทคนิค SMOTE	0.7222	0.6190	0.8670	0.2156
	วิธีการสุ่มลด	0.5000	0.0000	0.7140	0.4976
	วิธีการสุ่มผสมผสาน	0.6000	0.7370	0.4760	0.3194
ซัพพอร์ตเวกเตอร์แมชชีน	วิธีการสุ่มเกิน	0.4085	0.6050	0.1820	0.5915
	วิธีการสุ่มเกินเทคนิค SMOTE	0.7917	0.6430	1.0000	0.2083
	วิธีการสุ่มลด	0.5000	0.0000	0.7140	0.5000
	วิธีการสุ่มผสมผสาน	0.6500	0.6840	0.6190	0.3500
ค่าเฉลี่ย					
เพื่อนบ้านใกล้สุด k ตัว	วิธีการสุ่มเกิน	0.6901	0.7735	0.5734	0.3018
	วิธีการสุ่มเกินเทคนิค SMOTE	**0.7778***	**0.7834***	**0.768**	**0.2157**
	วิธีการสุ่มลด	0.5000	0.5094	0.4295	0.4173
	วิธีการสุ่มผสมผสาน	0.5000	0.5234	0.4739	0.4759
ต้นไม้ตัดสินใจ	วิธีการสุ่มเกิน	0.6761	0.7426	0.5978	0.2823
	วิธีการสุ่มเกินเทคนิค SMOTE	**0.7556**	**0.7551**	**0.7497**	**0.2114***
	วิธีการสุ่มลด	0.4800	0.4328	0.3961	0.4203
	วิธีการสุ่มผสมผสาน	0.5850	0.6749	0.4632	0.3538
โครงข่ายประสาทเทียม	วิธีการสุ่มเกิน	0.6254	0.6719	0.5682	0.2808
	วิธีการสุ่มเกินเทคนิค SMOTE	**0.7278**	**0.7298**	**0.7262**	**0.2291**
	วิธีการสุ่มลด	0.4800	0.3857	0.5095	0.4911
	วิธีการสุ่มผสมผสาน	0.5883	0.6330	0.5668	0.2975
ซัพพอร์ตเวกเตอร์แมชชีน	วิธีการสุ่มเกิน	0.5746	0.5828	0.5334	0.4253
	วิธีการสุ่มเกินเทคนิค SMOTE	**0.7722**	**0.7706**	**0.7742***	**0.2278**
	วิธีการสุ่มลด	0.4600	0.3714	0.4028	0.5400
	วิธีการสุ่มผสมผสาน	0.6050	0.5631	0.5689	0.3950

ตัวทึบ หมายถึง ค่าสูงสุดของค่าความถูกต้อง ค่าความไว ค่าความจำเพาะ และค่าต่ำสุดของค่าคลาดเคลื่อนกำลังสองเฉลี่ย; * หมายถึง ค่าความถูกต้อง ค่าความไว ค่าความจำเพาะเฉลี่ยสูงสุด และค่าเฉลี่ยของค่าคลาดเคลื่อนกำสองเฉลี่ยต่ำสุดของทั้ง 4 การจำแนก

4. สรุปผลการวิจัย

การค้นคว้าและศึกษาในการหาข้อมูลที่ไม่สมดุลได้ข้อมูล 3 ชุด คือ ชุดข้อมูลการรับรู้ทางหูของเด็กที่มีภาวะน้ำคั่งในหูชั้นกลางหรือหูน้ำหนวก มีจำนวนกลุ่มส่วนน้อยร้อยละ 4.49 ชุดข้อมูลยอดคงเหลือในบัตรเครดิตของลูกค้า มีจำนวนกลุ่มส่วน

น้อยร้อยละ 10 และชุดข้อมูลคุณภาพไวน์แดง มีจำนวนกลุ่มส่วนน้อยร้อยละ 19.52 สำหรับชุดข้อมูลการทางหูของเด็กที่มีภาวะน้ำคั่งในหูชั้นกลางหรือหูน้ำหนวก วิธีที่มีประสิทธิภาพสูงสุด คือ ซัพพอร์ตเวกเตอร์แมชชีน โดยการปรับความไม่สมดุลด้วยวิธีการสุ่มเกินเทคนิค SMOTE ชุดข้อมูลยอดคงเหลือในบัตรเครดิตของลูกค้า วิธีที่มีประสิทธิภาพสูงสุด คือ วิธีเพื่อนบ้านใกล้สุด k ตัว โดยการปรับความไม่สมดุลด้วยวิธีการสุ่มเกินเทคนิค SMOTE ส่วนชุดข้อมูลคุณภาพไวน์แดง วิธีที่มีประสิทธิภาพสูงสุด คือ วิธีโครงข่ายประสาทเทียม โดยการปรับความไม่สมดุลด้วยวิธีการสุ่มเกิน

ถ้าพิจารณาวิธีการจำแนกเพียงอย่างเดียว จะพบว่าวิธีการจำแนกที่ดีที่สุด คือ วิธีซัพพอร์ตเวกเตอร์แมชชีน ส่วนถ้าพิจารณาวิธีการปรับความไม่สมดุลของข้อมูลเพียงอย่างเดียว จะพบว่าวิธีการปรับความไม่สมดุลของข้อมูลดีที่สุด คือ วิธีการสุ่มเกินโดยเทคนิค SMOTE

ตารางที่ 5 ผลการวิเคราะห์ชุดข้อมูลคุณภาพไวน์แดง เมื่อ random seed 10, 20, 30, 40 และ 50 โดยใช้ข้อมูลส่วนที่ 3 ข้อมูลทดสอบร้อยละ 10

วิธีการจำแนก	วิธีการปรับความไม่สมดุลของข้อมูล	ค่าความถูกต้อง	ค่าความไว	ค่าความจำเพาะ	ค่าคลาดเคลื่อนกำลังสองเฉลี่ย
Random seed 10					
เพื่อนบ้านใกล้สุด k ตัว	วิธีการสุ่มเกิน	0.5375	0.6250	0.4500	0.4596
	วิธีการสุ่มเกินเทคนิค SMOTE	0.6226	0.6625	0.5822	0.3722
	วิธีการสุ่มลด	0.5641	0.3846	0.6538	0.4132
	วิธีการสุ่มผสมผสาน	0.5600	0.5370	0.5870	0.4306
ต้นไม้ตัดสินใจ	วิธีการสุ่มเกิน	0.5375	0.5250	0.5500	0.3390
	วิธีการสุ่มเกินเทคนิค SMOTE	0.6037	0.6875	0.5190	0.3145
	วิธีการสุ่มลด	0.4615	0.3077	0.5384	0.4718
	วิธีการสุ่มผสมผสาน	0.5700	0.5926	0.5435	0.3741
โครงข่ายประสาทเทียม	วิธีการสุ่มเกิน	0.5437	0.5875	0.5000	0.2924
	วิธีการสุ่มเกินเทคนิค SMOTE	0.5974	**0.7750**	0.4177	**0.2724**
	วิธีการสุ่มลด	0.5641	0.5385	0.5769	0.3716
	วิธีการสุ่มผสมผสาน	0.5200	0.5185	0.5217	0.3191
ซัพพอร์ตเวกเตอร์แมชชีน	วิธีการสุ่มเกิน	0.4750	0.4250	0.5250	0.5250
	วิธีการสุ่มเกินเทคนิค SMOTE	0.6289	0.725	0.5316	0.3711
	วิธีการสุ่มลด	**0.6667**	0.0769	**0.9615**	0.3334
	วิธีการสุ่มผสมผสาน	0.5300	0.6481	0.3913	0.4700
Random seed 20					
เพื่อนบ้านใกล้สุด k ตัว	วิธีการสุ่มเกิน	0.6187	0.6400	0.6000	0.3759
	วิธีการสุ่มเกินเทคนิค SMOTE	0.5723	0.5443	0.6000	0.4243
	วิธีการสุ่มลด	0.5897	0.7083	0.4000	0.3895
	วิธีการสุ่มผสมผสาน	0.6200	0.5957	0.6415	0.3725

ตารางที่ 5 (ต่อ)

วิธีการจำแนก	วิธีการปรับความไม่สมดุลของข้อมูล	ค่าความถูกต้อง	ค่าความไว	ค่าความจำเพาะ	ค่าคลาดเคลื่อนกำลังสองเฉลี่ย
ต้นไม้ตัดสินใจ	วิธีการสุ่มเกิน	0.5062	0.3067	0.6824	0.3232
	วิธีการสุ่มเกินเทคนิค SMOTE	0.5660	0.7848	0.3500	0.2739
	วิธีการสุ่มลด	0.5128	0.6250	0.3333	0.4196
	วิธีการสุ่มผสมผสาน	0.5400	0.4681	0.6038	0.2916
โครงข่ายประสาทเทียม	วิธีการสุ่มเกิน	0.5625	0.2400	0.8471	0.2814
	วิธีการสุ่มเกินเทคนิค SMOTE	0.5849	0.8228	0.3500	0.2678
	วิธีการสุ่มลด	0.6153	0.7500	0.4000	0.3328
	วิธีการสุ่มผสมผสาน	**0.7000**	0.4680	**0.9057**	**0.2192**
ซัพพอร์ตเวกเตอร์แมชชีน	วิธีการสุ่มเกิน	0.5188	0.2133	0.7882	0.4812
	วิธีการสุ่มเกินเทคนิค SMOTE	0.5220	0.5950	0.4500	0.4780
	วิธีการสุ่มลด	0.6667	**0.8333**	0.4000	0.3334
	วิธีการสุ่มผสมผสาน	0.5700	0.3617	0.5747	0.4299
Random seed 30					
เพื่อนบ้านใกล้สุด k ตัว	วิธีการสุ่มเกิน	0.6562	0.7093	0.5946	0.3392
	วิธีการสุ่มเกินเทคนิค SMOTE	0.4968	0.4815	0.5128	0.4994
	วิธีการสุ่มลด	0.7179	0.7727	0.6471	0.2675
	วิธีการสุ่มผสมผสาน	0.6700	0.5349	**0.7719**	0.3260
ต้นไม้ตัดสินใจ	วิธีการสุ่มเกิน	0.6000	0.6279	0.5676	0.3390
	วิธีการสุ่มเกินเทคนิค SMOTE	0.4779	0.7284	0.2180	0.3169
	วิธีการสุ่มลด	0.6154	0.7727	0.4118	0.3446
	วิธีการสุ่มผสมผสาน	0.6500	0.5116	0.7544	0.3190
โครงข่ายประสาทเทียม	วิธีการสุ่มเกิน	0.5812	0.5465	0.6216	0.2834
	วิธีการสุ่มเกินเทคนิค SMOTE	0.4905	0.2346	0.7564	0.3010
	วิธีการสุ่มลด	0.6923	0.7727	0.5882	0.2929
	วิธีการสุ่มผสมผสาน	0.5200	0.5349	0.5087	0.3185
ซัพพอร์ตเวกเตอร์แมชชีน	วิธีการสุ่มเกิน	0.5687	**0.7791**	0.3243	0.4313
	วิธีการสุ่มเกินเทคนิค SMOTE	0.4905	0.5926	0.3846	0.5094
	วิธีการสุ่มลด	**0.7436**	0.7727	0.7059	**0.2564**
	วิธีการสุ่มผสมผสาน	0.6400	0.3953	0.8246	0.3600

ตารางที่ 5 (ต่อ)

วิธีการจำแนก	วิธีการปรับความไม่สมดุลของข้อมูล	ค่าความถูกต้อง	ค่าความไว	ค่าความจำเพาะ	ค่าคลาดเคลื่อนกำลังสองเฉลี่ย
Random seed 40					
เพื่อนบ้านใกล้สุด k ตัว	วิธีการสุ่มเกิน	0.6438	0.5070	0.7420	0.3469
	วิธีการสุ่มเกินเทคนิค SMOTE	0.5849	0.6430	0.5200	0.4094
	วิธีการสุ่มลด	0.5385	0.6670	0.3330	0.4373
	วิธีการสุ่มผสมผสาน	0.6000	0.5200	0.6800	0.3917
ต้นไม้ตัดสินใจ	วิธีการสุ่มเกิน	0.5438	0.3880	0.6560	0.3461
	วิธีการสุ่มเกินเทคนิค SMOTE	0.5408	0.4880	0.6000	0.3195
	วิธีการสุ่มลด	0.4615	0.6670	0.1330	0.3109
	วิธีการสุ่มผสมผสาน	0.4700	0.4600	0.4800	0.3778
โครงข่ายประสาทเทียม	วิธีการสุ่มเกิน	0.5438	0.3880	0.6560	0.3461
	วิธีการสุ่มเกินเทคนิค SMOTE	0.5911	0.5950	0.5870	0.2740
	วิธีการสุ่มลด	0.4103	0.6250	0.0670	0.4601
	วิธีการสุ่มผสมผสาน	0.5500	0.5000	0.6000	0.3540
ซัพพอร์ตเวกเตอร์แมชชีน	วิธีการสุ่มเกิน	0.5875	0.1190	0.9250	0.4125
	วิธีการสุ่มเกินเทคนิค SMOTE	0.5409	0.5600	0.5200	0.4591
	วิธีการสุ่มลด	0.5641	0.9170	0.0000	0.4359
	วิธีการสุ่มผสมผสาน	0.5500	0.5200	0.5800	0.4500
Random seed 50					
เพื่อนบ้านใกล้สุด k ตัว	วิธีการสุ่มเกิน	0.6438	0.7230	0.5300	0.3518
	วิธีการสุ่มเกินเทคนิค SMOTE	0.5786	0.6050	0.5540	0.4156
	วิธีการสุ่มลด	0.5128	0.6000	0.4210	0.4616
	วิธีการสุ่มผสมผสาน	0.5900	0.5660	0.6170	0.4014
ต้นไม้ตัดสินใจ	วิธีการสุ่มเกิน	0.6063	0.6600	0.5300	0.2894
	วิธีการสุ่มเกินเทคนิค SMOTE	0.5974	0.5130	0.6750	0.3279
	วิธีการสุ่มลด	0.6154	0.6000	0.6320	0.3357
	วิธีการสุ่มผสมผสาน	0.5100	0.3400	0.7020	0.3333
โครงข่ายประสาทเทียม	วิธีการสุ่มเกิน	0.6625	0.6170	0.7270	0.2274
	วิธีการสุ่มเกินเทคนิค SMOTE	0.4654	0.3160	0.6020	0.3275
	วิธีการสุ่มลด	0.5128	0.6500	0.3680	0.3870
	วิธีการสุ่มผสมผสาน	0.6400	0.6790	0.5960	0.3029

ตารางที่ 5 (ต่อ)

วิธีการจำแนก	วิธีการปรับความไม่สมดุลของข้อมูล	ค่าความถูกต้อง	ค่าความไว	ค่าความจำเพาะ	ค่าคลาดเคลื่อนกำลังสองเฉลี่ย
ซัพพอร์ตเวกเตอร์แมชชีน	วิธีการสุ่มเกิน	0.5563	0.8300	0.1670	0.4437
	วิธีการสุ่มเกินเทคนิค SMOTE	0.5220	0.3420	0.6870	0.4780
	วิธีการสุ่มลด	0.6154	0.6500	0.5790	0.3846
	วิธีการสุ่มผสมผสาน	0.6100	0.6790	0.5320	0.3900
ค่าเฉลี่ย					
เพื่อนบ้านใกล้สุด k ตัว	วิธีการสุ่มเกิน	**0.6200**	**0.6409**	0.5833	**0.3747**
	วิธีการสุ่มเกินเทคนิค SMOTE	0.5711	0.5872	0.5538	0.4240
	วิธีการสุ่มลด	0.5846	0.6265	0.4910	0.3938
	วิธีการสุ่มผสมผสาน	0.6080	0.5507	**0.6595**	0.3844
ต้นไม้ตัดสินใจ	วิธีการสุ่มเกิน	0.5558	0.5015	0.5972	0.3274
	วิธีการสุ่มเกินเทคนิค SMOTE	**0.5572**	**0.6403**	0.4724	**0.3105**
	วิธีการสุ่มลด	0.5333	0.5945	0.4097	0.3765
	วิธีการสุ่มผสมผสาน	0.5480	0.4745	**0.6167**	0.3392
โครงข่ายประสาทเทียม	วิธีการสุ่มเกิน	**0.5888**	0.4758	**0.6703***	**0.2861***
	วิธีการสุ่มเกินเทคนิค SMOTE	0.5459	0.5487	0.5426	0.2885
	วิธีการสุ่มลด	0.5590	**0.6672***	0.4000	0.3689
	วิธีการสุ่มผสมผสาน	0.5860	0.5401	0.6264	0.3027
ซัพพอร์ตเวกเตอร์แมชชีน	วิธีการสุ่มเกิน	0.5413	0.4733	0.5459	0.4588
	วิธีการสุ่มเกินเทคนิค SMOTE	0.5409	0.5630	0.5148	0.4591
	วิธีการสุ่มลด	**0.6513***	**0.6500**	0.5293	**0.3487**
	วิธีการสุ่มผสมผสาน	0.5800	0.5208	**0.5805**	0.4200

ตัวทึบ หมายถึง ค่าสูงสุดของค่าความถูกต้อง ค่าความไว ค่าความจำเพาะ และค่าต่ำสุดของค่าคลาดเคลื่อนกำลังสองเฉลี่ย; * หมายถึง ค่าความถูกต้อง ค่าความไว ค่าความจำเพาะเฉลี่ยสูงสุด และค่าเฉลี่ยของค่าคลาดเคลื่อนกำสองเฉลี่ยต่ำสุดของทั้ง 4 การจำแนก

5. อภิปรายผล

การสรุปผลงานวิจัยครั้งนี้ ชุดข้อมูลการรับรู้ทางหูของเด็กที่มีภาวะน้ำคั่งในหูชั้นกลางหรือหูน้ำหนวก วิธีซัพพอร์ตเวกเตร์แมชชีน ใช้ขั้นตอนวิธี SMO โดยการปรับความไม่สมดุลด้วยวิธีการสุ่มเกินเทคนิค SMOTE มีประสิทธิภาพในการจำแนกดีที่สุด แต่การจำแนกข้อมูลให้ผลไม่สอดคล้องกับเชาวนันท์ และคณะ (2556) เรื่อง แบบจำลองการทำนายผลการรักษาผู้ป่วยมะเร็งปากมดลูก วิธีโครงข่ายประสาทเทียมโดยการปรับด้วยวิธีการสุ่ม

เกินโดยเทคนิค SMOTE มีประสิทธิภาพในการจำแนกดีที่สุด เนื่องจากข้อมูลมีลักษณะแตกต่างกัน โดยชุดข้อมูลการรับรู้ทางหูของเด็กที่มีภาวะน้ำคั่งในหูชั้นกลางมีกลุ่มส่วนน้อยร้อยละ 4.49 ส่วนชุดข้อมูลผลการรักษาผู้ป่วยมะเร็งปากมดลูกอาจมีกลุ่มส่วนน้อยมากกว่านี้ ชุดข้อมูลยอดคงเหลือในบัตรเครดิต วิธีวิธีเพื่อนบ้านใกล้สุด k ตัว โดยใช้ขั้นวิธี IBk โดยการปรับความไม่สมดุลด้วยวิธีการสุ่มเกินโดยเทคนิค SMOTE มีประสิทธิภาพในการจำแนกดีที่สุด แต่การจำแนกข้อมูลให้ผลไม่สอดคล้องกับเชาวนันท์ และคณะ (2556) เรื่อง แบบจำลองการทำนายผลการรักษาผู้ป่วยมะเร็งปากมดลูก วิธีโครงข่ายประสาทเทียมโดยการปรับด้วยวิธีการสุ่มเกินโดยเทคนิค SMOTE มีประสิทธิภาพในการจำแนกดีที่สุด เนื่องจากข้อมูลมีลักษณะแตกต่างกัน ชุดข้อมูลยอดคงเหลือในบัตรเครดิตมีกลุ่มส่วนน้อยร้อยละ 10 ส่วนชุดข้อมูลผลการรักษาผู้ป่วยมะเร็งปากมดลูกอาจมีกลุ่มส่วนน้อยน้อยกว่าหรือมากกว่านี้ และชุดข้อมูลคุณภาพไวน์แดง วิธีโครงข่ายประสาทเทียม โดยใช้ขั้นตอนวิธีชนิดเพอร์เซปตรอนหลายชั้น โดยการปรับความไม่สมดุลด้วยวิธีการสุ่มเกิน มีประสิทธิภาพในการจำแนกดีที่สุด ให้ผลไม่สอดคล้องกับ He และ Ghodsi (2010) เรื่อง การศึกษาเกี่ยวกับการจำแนกข้อมูลที่ค้นพบได้ยาก วิธีซัพพอร์ตเวกเตอร์แมชชีนโดยการปรับด้วยวิธีการสุ่มเกินมีประสิทธิภาพในการจำแนกที่ดีที่สุด เนื่องจากข้อมูลมีลักษณะแตกต่างกัน โดยชุดข้อมูลคุณภาพไวน์แดงมีกลุ่มส่วนน้อยร้อยละ 19.52 ส่วนชุดข้อมูลการศึกษาเกี่ยวกับการจำแนกข้อมูลที่ค้นพบได้ยากมีกลุ่มส่วนน้อยน้อยกว่าร้อยละ 19.52 ทำให้ผลการจำแนกและการปรับความไม่สมดุลแตกต่างกัน

6. ข้อเสนอแนะ

6.1 ชุดข้อมูลที่นำมาปรับความไม่สมดุลอาจเพิ่มจำนวนชุดข้อมูลที่มากกว่านี้

6.2 อาจวิเคราะห์ข้อมูลด้วยวิธีการจำแนกวิธีอื่น ๆ ได้แก่ นาอีฟเบส์ เบส์เนท การถดถอยลอจิสติกทวิภาค และฐานกฎ เป็นต้น

6.3 อาจศึกษาวิธีการปรับความไม่สมดุลของข้อมูลวิธีอื่น ๆ เช่น cost sensitive learning

7. รายการอ้างอิง

กีระชาติ สุขสุทธิ์, 2559, การจำแนกข้อมูลไม่สมดุลโดยใช้การปรับปรุงข้อมูลร่วมกับการหาค่าพารามิเตอร์ที่เหมาะสมด้วยขั้นตอนวิธีทางพันธุกรรมที่มีการเริ่มต้นใหม่, วิทยานิพนธ์ปริญญาเอก, มหาวิทยาลัยเทคโนโลยีสุรนารี, นครราชสีมา.

เชาวนันท์ โสโท, พุธษดี ศิริแสงตระกูล และวรชัย ตั้งวรพงศ์ชัย, 2556, แบบจำลองการทำนายผลการรักษาผู้ป่วยมะเร็งปากมดลูกด้วยโครงข่ายประสาทเทียม, ว.วิจัยมหาวิทยาลัยขอนแก่น 13(1): 39-50.

เบญจภรณ์ จันทรกองกุล, สุวรรณา รัศมีขวัญ, สุนิสา ริมเจริญ, ภูสิต กุลเกษม, กฤษณะ ชินสาร, อัณณ์นุพันธ์ รอดทุกข์, ปิยนุช วรบุตร และจรรยา อันปันส์, วิธีการที่เหมาะสมสำหรับการแบ่งกลุ่มข้อมูลที่ไม่สมดุลสูง, แหล่งที่มา : http://digital_collect.lib.buu.ac.th/dcms/files/2559_047.pdf, 24 มิถุนายน 2561.

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

ภรันยา ปาลวิสุทธิ์, 2559, การเพิ่มประสิทธิภาพเทคนิคต้นไม้ตัดสินใจบนชุดข้อมูลที่ไม่สมดุล

โดยวิธีการการสุ่มเพิ่มตัวอย่างกลุ่มน้อยสำหรับสำหรับข้อมูลการเป็นโรคอินเตอร์เน็ต, ว.เทคโนโลยีสารสนเทศ 12(1): 54-63.

วีระยุทธ มายุศิริ, จารี ทองคำ และวาทินี สุขมาก, 2557, การพัฒนาแบบจำลองเพื่อการพยากรณ์การรักษาซ้ำของผู้ป่วยโรคจิตเภทโดยเทคนิคเหมืองข้อมูล, ว.วิทยาศาสตร์และเทคโนโลยี 10(1): 144-153.

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

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

Cortez, P. , 2009, Wine Quality Data Set, Available Source: http://archive.ics.uci.edu/m0l/datasets/Wine+Quality, June 24, 2018.

He, H. and Ghodsi, A. 2010, Rare class classification by support vector machine, pp. 548-551, In 20th International Conference on Pattern Recognition.

Hogan, S., 2013, Tests of Auditory Perception in Children with OME, Available Source: https://vincentarebundock.gitub.io/Rdatasets/doc/MASS/OME.html, June 24, 2018.

James, G., Witten, D., Hastie, T. and Tibshirani, R. , 2013, Credit Card Balance Data, Available Source: https://vincentarelbundock.github.io/Rdataset/doc/ISLR/Credit.html, June 24, 2018.

McCarthy, K. , Zabar, B. and Weiss, G. , 2005, Cost-sensitive learning vs. sampling: Which is best for handling unbalanced classes with unequal error costs, Proceedings of the 2007 International Conference on Data 7(1): 35-41.