การเปรียบเทียบวิธีการแก้ปัญหาข้อมูลไม่สมดุล สำหรับการจำแนกกลุ่มรายได้ของผู้ประกอบการร้านยาประเภท ข.ย.1

Comparison of Imbalanced Data Problem Solving for Income Classification of Type I Pharmacies Entrepreneur

นพมาศ อัครจันทโชติ (Noppamas Akarachantachote) [1] ดิเรก พนิตสุภากมล (Direk Panitsupakamol) [2]

บทคัดย่อ

ข้อมูลไม่สมดุลพบได้ในหลายสถานการณ์ ซึ่งโดยทั่วไปวิธีการจำแนกประเภทข้อมูลมีแนวโน้มที่จะทำนายข้อมูลเป็นกลุ่มส่วนมาก อันจะส่งผลถึงประสิทธิภาพที่ต่ำในการทำนายกลุ่มส่วนน้อย การสุ่มตัวอย่างเพิ่มสำหรับกลุ่มส่วนน้อยเป็นแนวทางหนึ่งในการจัดการกับปัญหาการจำแนกข้อมูลไม่สมดุล วัตถุประสงค์ของการวิจัยนี้ 1) เพื่อเปรียบเทียบประสิทธิภาพการแก้ปัญหาข้อมูลไม่สมดุลด้วยการสุ่มตัวอย่างซ้ำระหว่างวิธีการสุ่มตัวอย่างเพิ่มข้อมูลเริ่มต้นอย่างสุ่ม และการสุ่มตัวอย่างเพิ่มกลุ่มส่วนน้อยด้วยการสังเคราะห์ 2) เพื่อเปรียบเทียบประสิทธิภาพในการจำแนกข้อมูลระหว่างการถดถอยลอจิสติก และต้นไม้ตัดสินใจ สำหรับการจำแนกกลุ่มรายได้ผู้ประกอบการร้านยาประเภท ข.ย.1 โดยค่าวัดประสิทธิภาพที่ใช้ในการเปรียบเทียบได้แก่ ค่าความแม่นยำ อัตราความถูกต้องในการทำนายกลุ่มส่วนน้อย อัตราความถูกต้องในการทำนายกลุ่มส่วนมาก และค่าการวัดเอฟ ผลที่ได้ปรากฏว่า การสุ่มตัวอย่างเพิ่มกลุ่มส่วนน้อยด้วยการสังเคราะห์มีประสิทธิภาพสูงที่สุดในการจำแนกสำหรับทุกวิธีการจำแนกประเภท

คำสำคัญ การจำแนก ข้อมูลไม่สมดุล เทคนิคการสุ่มตัวอย่างเพิ่มกลุ่มส่วนน้อยด้วยการสังเคราะห์

[1] อาจารย์ประจำ สาขาวิชาคณิตศาสตร์และสถิติ มหาวิทยาลัยหัวเฉียวเฉลิมพระเกียรติ a_noppamas@yahoo.com

[2] อาจารย์ประจำ สาขาวิชาคณิตศาสตร์และสถิติ มหาวิทยาลัยหัวเฉียวเฉลิมพระเกียรติ direk7272@gmail.com

Abstract

Imbalanced data are frequently found in many situations. General classification techniques tend to biased toward the majority class. This causes low performance in predicting minority class. Oversampling for minority class is a strategy to handle class imbalance classification. This research intends 1) to compare the efficiency of resampling method between Random over-sampling (ROS) and Synthetic Minority Over-sampling TEchnique (SMOTE), and 2) to compare the efficiency of classifiers between logistic regression and decision tree in solving the imbalance data of Type I Pharmacies Entrepreneur. Performance measures for this this comparison are accuracy, true positive rate, true negative rate, and F-measure. The results show that over-sampling by SMOTE has high performance on classifying the data from minority class for all classification techniques.

Keywords: Classification, Imbalanced data, SMOTE

บทนำ

ข้อมูลไม่สมดุล (Imbalanced data) หมายถึง ข้อมูลที่มีจำนวนข้อมูลในกลุ่มหนึ่งมากกว่าจำนวนข้อมูลของอีกกลุ่มหนึ่งเป็นจำนวนมาก อาจเกิดขึ้นเนื่องจากลักษณะทางธรรมชาติของข้อมูลที่มีความแตกต่างของจำนวนในแต่ละกลุ่ม หรืออาจเกิดจากข้อจำกัดในการเก็บข้อมูลซึ่งอาจเนื่องมาจากค่าใช้จ่ายในการเก็บข้อมูลที่สูง (กีระชาติ สุขสุทธิ์, 2559) ซึ่งเกิดขึ้นได้ทั่วไปในหลายๆ สถานการณ์ เช่น จำนวนผู้ที่ติดเชื้อเอชไอวีมีจำนวน 480,000 คน (hfocus, 2562) จากประชากรไทยทั้งประเทศกว่า 66 ล้านคน (ระบบสถิติทางการทะเบียน, 2562) ร้อยละของประชากรที่มีอายุ 15 ปีขึ้นไปที่เป็นผู้สูบบุหรี่ เท่ากับ 19.1 (ศูนย์วิจัยและจัดการความรู้เพื่อการควบคุมยาสูบ, 2561) ลูกค้าบัตรเครดิตที่เป็นหนี้มีจำนวนน้อยกว่าอย่างมากเมื่อเทียบกับผู้ที่ไม่เป็นหนี้ จำนวนครั้งของธุรกรรมที่มีการทุจริตมีจำนวนน้อยกว่าอย่างมากเมื่อเทียบกับธุรกรรมปกติ ผลิตภัณฑ์ที่มีข้อบกพร่องมีอัตราการผิดพลาดในการผลิตน้อยเมื่อเทียบกับผลิตภัณฑ์ที่ไม่บกพร่อง เป็นต้น

การที่ข้อมูลไม่สมดุลจะส่งผลต่อการจำแนกประเภทกลุ่มส่วนน้อย (Minority Class) ที่เป็นกลุ่มหายาก เนื่องจากวิธีการจำแนกประเภททั่วไปจะมีประสิทธิภาพในการจำแนกประเภทก็ต่อเมื่อข้อมูลในแต่ละกลุ่มมีจำนวนใกล้เคียงกัน ซึ่งหากข้อมูลไม่สมดุล วิธีการจำแนกประเภททั่วไปมีความโน้มเอียงที่จะทำนายเป็นกลุ่มส่วนมาก (Majority Class) เนื่องจากการจำแนกประเภททั่วไปมีเป้าหมายในการทำให้ภาพรวมมีความแม่นยำสูงสุด เช่น ถ้าข้อมูลชุดหนึ่ง มีกลุ่ม A อยู่ร้อยละ 5 และกลุ่ม B อยู่ร้อยละ 95 วิธีการจำแนกประเภททั่วไปมีแนวโน้มที่จะทำนายสมาชิกใหม่ว่าอยู่กลุ่ม B เนื่องจากจะทำให้มีโอกาสทำนายถูกร้อยละ 95 ซึ่งทำให้มีค่าความแม่นยำ (Accuracy) สูง แต่จะไม่สามารถทำนายกลุ่ม A ได้อย่างถูกต้อง ในหลายๆ สถานการณ์ กลุ่มส่วนน้อยมักเป็นกลุ่มที่มีความสำคัญที่ต้องการค้นหา การทำนายกลุ่มส่วนน้อยผิดพลาดว่าเป็นกลุ่มส่วนมาก (False Negative) ส่งผลเสียร้ายแรงกว่าการทำนายกลุ่มส่วนมากผิดพลาดว่าเป็นกลุ่มส่วนน้อย (False Positive) เช่น กลุ่มผู้ป่วยที่ติดเชื้อเอชไอวีมีจำนวนน้อยกว่ากลุ่มผู้ที่ไม่ติดเชื้อ และถ้ามีการทำนายผู้ป่วยที่ติดเชื้อเอชไอวีผิดพลาดว่าไม่ได้ติดเชื้อจะทำให้ผู้ป่วยไม่ได้รับการรักษา ซึ่งเป็นผลเสียร้ายแรงกว่าการทำนายผู้ที่ไม่ติดเชื้อเอชไอวีผิดพลาดว่าติดเชื้อ โดยแนวทางการแก้ปัญหาข้อมูลไม่สมดุลนี้สามารถทำได้ใน 2 แนวทาง ได้แก่ แนวทางการแก้ไขในระดับของขั้นตอนวิธี (Algorithm Approach) และ แนวทางการแก้ไขในระดับข้อมูล (Data Approach)

ขั้นตอนวิธี หรือวิธีที่ใช้ในการจำแนกประเภทโดยทั่วไปจะมีประสิทธิภาพสูงเมื่อจำนวนข้อมูลในแต่ละกลุ่มมีจำนวนที่ใกล้เคียงกันหรือมีความสมดุลของข้อมูล การแก้ไขในระดับของขั้นตอนวิธีจะปรับขั้นตอนวิธีให้เพิ่มประสิทธิภาพในการทำนายกลุ่มส่วนน้อย ส่วนแนวทางการแก้ไขในระดับข้อมูลจะเป็นการปรับในขั้นของการเตรียมข้อมูล (Pre-processing Data) โดยการปรับให้ข้อมูลทั้งสองกลุ่มมีจำนวนที่ใกล้เคียงกัน ซึ่งแนวทางนี้เป็นที่นิยมในการใช้แก้ปัญหาเนื่องจากเป็นวิธีง่าย และมีความยืดหยุ่น (Minh, 2018) การทำให้ข้อมูลทั้งสองกลุ่มที่จำนวนใกล้เคียงกันทำได้โดยการสุ่มตัวอย่างซ้ำ (Re-sampling) ซึ่งทำได้ใน 2 ลักษณะ ได้แก่ การสุ่มตัวอย่างลด (Under-sampling) กลุ่มส่วนมาก และการสุ่มตัวอย่างเพิ่ม (Over-sampling) กลุ่มส่วนน้อย โดยตัวอย่างที่เพิ่มหรือลดสามารถทำด้วยการการสุ่มตัวอย่างเพิ่มจากข้อมูลเริ่มต้นอย่างสุ่ม (Random Over-sampling: ROS) หรือการสุ่มตัวอย่างลดอย่างสุ่ม (Random Under-sampling: RUS) การสุ่มตัวอย่างลดจะเหมาะกับข้อมูลขนาดใหญ่ ซึ่งมีข้อเสียคือทำให้สูญเสียสารสนเทศของข้อมูล ส่วนการสุ่มตัวอย่างเพิ่มเหมาะกับข้อมูลที่ขนาดไม่ใหญ่ แต่มีข้อเสียคือทำให้ใช้เวลาในการประมวลผลเพิ่มขึ้น

Chawla et al. (2002) ได้เสนอแนวทางการสุ่มตัวอย่างเพิ่มโดยการสร้างตัวอย่างสังเคราะห์แทนที่จะเป็นการเพิ่มจากข้อมูลเดิมของกลุ่มส่วนน้อยอย่างสุ่ม และเรียกเทคนิคนี้ว่า การสุ่มตัวอย่างเพิ่มกลุ่มส่วนน้อยด้วยการสังเคราะห์ (Synthetic

Minority Over-ampling Technique: SMOTE) SMOTE สร้างกลุ่มส่วนน้อยตัวอย่างใหม่อย่างสุ่มบนแนวเส้นที่เชื่อมต่อระหว่างตัวอย่างกลุ่มส่วนน้อยใดๆ และหนึ่งในกลุ่มเพื่อนบ้านที่ใกล้ที่สุดจำนวน k โดย Chawla และคณะ ได้เปรียบเทียบวิธี ROS และ SMOTE โดยใช้ต้นไม้การตัดสินใจในการจำแนกประเภท และทำการทดลองสุ่มเพิ่มที่ระดับร้อยละ 100 200 300 400 และ 500 ของขนาดเริ่มต้น ซึ่งได้เปรียบเทียบโดยใช้ชุดข้อมูลผลการตรวจเต้านม ซึ่งกลุ่มที่มีผลปกติมีจำนวน 10,923 และกลุ่มที่มีผลผิดปกติมีจำนวน 260 ผลการทดลองพบว่า ความถูกต้องในการทำนายกลุ่มส่วนน้อยโดยวิธี ROS จะน้อยกว่าวิธี SMOTE เมื่อร้อยละของการสุ่มเพิ่มมีค่าสูง SMOTE ได้ถูกใช้อย่างแพร่หลายในการปรับปรุงประสิทธิภาพของการทำนาย (Blagus, & Lusa, 2015) อีกทั้งมีความแกร่งกว่าการสุ่มตัวอย่างลดอย่างสุ่มในสถานการณ์ที่มีข้อมูลรบกวนจากการทดลองเปรียบเทียบวิธีการสุ่มซ้ำระหว่าง SMOTE และ RUS ของข้อมูลที่ไม่สมดุลภายใต้สถานการณ์ที่มีข้อมูลรบกวน (Kaur, & Gosain2018)

Mishra (2017) ได้เปรียบเทียบการแก้ปัญหาข้อมูลไม่สมดุล โดยการสุ่มตัวอย่างซ้ำ ซึ่งเปรียบเทียบวิธี SMOTE ที่ใช้ Cross Validation (SMOTE with CV) วิธี SMOTE ที่ไม่ใช้ Cross Validation (SMOTE without CV) และ วิธี RUS และใช้วิธีการจำแนกประเภท ได้แก่ Random Forest และ XGBoost ผลการทดลองพบว่า วิธี RUS สามารถทำนายกลุ่มส่วนน้อยได้ดีกว่าวิธีอื่น เมื่อใช้วิธีการจำแนกประเภท Random Forest นอกจากนี้ Jagelid and Movin (2017) ได้ทำนายการใช้บริการสตรีมเพลงดิจิทัล พอดแคสต์ และวิดีโอ ซึ่งมีบริการสองประเภท ได้แก่ ผลิตภัณฑ์ที่ใช้บริการฟรีแบบมีข้อจำกัดและผลิตภัณฑ์ที่ใช้บริการแบบเสียเงิน ซึ่งผู้วิจัยต้องการทำนายการเปลี่ยนประเภทของผลิตภัณฑ์ที่ใช้บริการจากฟรีเป็นแบบเสียเงิน โดยแบ่งลูกค้าเป็นสองกลุ่มได้แก่ กลุ่มที่เปลี่ยน และกลุ่มที่ไม่เปลี่ยน ซึ่งลูกค้าที่เป็นกลุ่มส่วนน้อยได้แก่กลุ่มที่เปลี่ยนโดยมีร้อยละ 2 ของลูกค้าที่ใช้บริการฟรี วิธีการจำแนกประเภทที่ใช้ ได้แก่ การถดถอยลอจิสติก (Logistic Regression) ต้นไม้ตัดสินใจ (Decision Tree) และ Gradient Boosting Trees และเนื่องจากข้อมูลที่ได้เป็นข้อมูลไม่สมดุลจึงมีการสุ่มตัวอย่างซ้ำ โดยในงานวิจัยได้เปรียบเทียบระหว่างวิธี ROS, RUS และ SMOTE ซึ่งผลการวิจัยพบว่า การสุ่มตัวอย่างซ้ำช่วยเพิ่มประสิทธิภาพของวิธีการจำแนกประเภท โดยวิธีที่มีประสิทธิภาพสูงสุดคือการจำแนกประเภทโดย การถดถอยลอจิสติก และ Gradient Boosting Trees ร่วมกับการสุ่มตัวอย่างเพิ่มให้จำนวนทั้งสองกลุ่มเท่ากันด้วยวิธี ROS

จากผลการวิจัยที่ผ่านมา สรุปได้ว่าถ้าข้อมูลไม่มีความสมดุล ควรมีการปรับข้อมูลให้มีความสมดุลก่อน ซึ่งจะช่วยให้การจำแนกประเภทสามารถทำงานได้อย่างมีประสิทธิภาพยิ่งขึ้น โดยในงานวิจัยนี้ได้ทำการศึกษากรณีตัวอย่าง ในการจำแนกกลุ่มรายได้ผู้ประกอบการร้านยา ซึ่งร้านยาเป็นทางเลือกแรก ๆ ของประชาชนในการดูแลรักษาสุขภาพ เนื่องจากมีอยู่ทั่วไปและใกล้ชิดกับประชาชน การดำเนินธุรกิจร้านยาในอดีตมักเป็นการดำเนินการภายในครอบครัวโดยสถานที่ดำเนินการมักอยู่ตามห้องแถว แต่ในปัจจุบันมีการเปลี่ยนแปลงไป โดยร้านยาเริ่มมีการขยายไปสู่ห้างสรรพสินค้า มีการดำเนินธุรกิจในรูปแบบแฟรนไชส์ ซึ่งมีการแข่งขันเพิ่มขึ้น การใช้กลยุทธ์ต่าง ๆ ทางการตลาดจึงมีความจำเป็นเพื่อให้รายได้เพิ่มขึ้น ในงานวิจัยนี้มีแนวคิดในการสร้างตัวแบบเพื่อการทำนายกลุ่มรายได้ผู้ประกอบการร้านยาจากตัวแปรต่าง ๆ ให้มีความแม่นยำ ซึ่งต้องเลือกวิธีการจำแนกประเภทที่มีความเหมาะสม แต่เนื่องจากข้อมูลผู้ประกอบการร้านยา แบ่งเป็นกลุ่มผู้ประกอบการที่มีรายได้น้อย และกลุ่มผู้ประกอบการที่มีรายได้สูง โดยมีจำนวน 72 และ 39 ราย ตามลำดับ ซึ่งมีจำนวนในแต่ละกลุ่มที่แตกต่างกันเกือบเท่าตัว ดังนั้นวิธีการจำแนกประเภทจะให้ผลที่เอนเอียงเนื่องจากจำนวนของทั้งสองกลุ่มแตกต่างกันมาก จึงควรมีการปรับข้อมูลให้มีความสมดุล และเนื่องจากจำนวนข้อมูลทั้งหมดนั้นมีน้อย ดังนั้นการสุ่มตัวอย่างเพิ่มเป็นทางเลือกที่เหมาะสม การสุ่มตัวอย่างเพิ่มที่ใช้ในงานวิจัยนี้ได้แก่ การสุ่มตัวอย่างเพิ่มข้อมูลเริ่มต้นอย่างสุ่ม และการสุ่มตัวอย่างเพิ่มกลุ่มส่วนน้อยด้วย

การสังเคราะห์โดยวิธี SMOTE (Chawla et al., 2002) โดยใช้วิธีการจำแนกประเภทได้แก่ การถดถอยลอจิสติก และต้นไม้ตัดสินใจ

วัตถุประสงค์

1. เพื่อเปรียบเทียบประสิทธิภาพในการสุ่มตัวอย่างซ้ำระหว่างวิธี ROS และ SMOTE ของข้อมูลกลุ่มรายได้ของผู้ประกอบการร้านยาประเภท ข.ย.1
2. เพื่อเปรียบเทียบประสิทธิภาพในการจำแนกข้อมูลระหว่างการถดถอยลอจิสติก และต้นไม้ตัดสินใจ ของข้อมูลกลุ่มรายได้ของผู้ประกอบการร้านยาประเภท ข.ย.1

ระเบียบวิธีวิจัย

ขอบเขตการวิจัย

ข้อมูลที่ใช้ในการเปรียบเทียบเป็นข้อมูลของผู้ประกอบการร้านยาประเภท ข.ย.1 จำนวน 111 ราย โดยจำแนกเป็นผู้ประกอบการที่มีรายได้ไม่เกิน 150,000 ต่อเดือน และผู้ประกอบการที่มีรายได้เกิน 150,000 บาทต่อเดือน ซึ่งมีจำนวน 72 ราย และ 39 ราย ตามลำดับ โดยปัจจัยที่ใช้ในการทำนาย ได้แก่ ปัจจัยด้านลักษณะทั่วไปของผู้ประกอบการ ลักษณะของธุรกิจ และส่วนประสมทางการตลาด ที่มีความสัมพันธ์กับรายได้ รวมทั้งสิ้น 14 ตัวแปร (นพมาศ และคณะ, 2561) ได้แก่

1) ระดับการศึกษา
2) รูปแบบของธุรกิจร้านยา
3) ประเภทของการขาย
4) การขายวิตามินอาหารเสริม
5) การขายเวชสำอาง
6) การขายเครื่องมือแพทย์
7) ความหลากหลายของตราสินค้า
8) การกำหนดราคาสินค้าหรือบริการ
9) ขนาดพื้นที่ร้าน
10) ปริมาณแผ่นพับให้ความรู้
11) การแจกผลิตภัณฑ์ทดลองใช้
12) การแจกของแถมหรือลดราคา
13) ความสามารถในการจ้างเภสัชกรเต็มเวลา
14) ค่าตอบแทนเภสัชกร

ขั้นตอนการดำเนินงาน

1. ศึกษางานวิจัยและทฤษฎีที่เกี่ยวข้องในการแก้ปัญหาข้อมูลไม่สมดุล

2. สร้างชุดข้อมูลสอน (Training Set) และชุดข้อมูลทดสอบ (Test Set) ด้วยอัตราส่วน 7:3 (สุรวัชร และสายชล, 2560)

3. สุ่มตัวอย่างเพิ่มกลุ่มส่วนน้อยด้วยวิธีการสุ่มตัวอย่างเพิ่มข้อมูลเดิมอย่างสุ่ม (ROS) และวิธีการสุ่มตัวอย่างเพิ่มด้วยข้อมูลสังเคราะห์ (SMOTE) บนข้อมูลสอน

4. จำแนกประเภทโดยใช้วิธีการถดถอยลอจิสติก (Logistic Regression) และวิธีการต้นไม้ตัดสินใจ (Decision Tree) เมื่อใช้

1) ข้อมูลเริ่มต้น (Original data)

2) ข้อมูลที่เพิ่มกลุ่มส่วนน้อยด้วยวิธีการเพิ่มตัวอย่างข้อมูลเริ่มต้นอย่างสุ่ม (Random Over-Sampling: ROS)

3) ข้อมูลที่เพิ่มกลุ่มส่วนน้อยด้วยวิธี SMOTE โดยข้อมูลที่สังเคราะห์ขึ้นใหม่จากวิธี SMOTE คำนวณดังนี้

$$x_s = x_i + u \cdot (\hat{x}_i - x_i)$$

โดยที่ x_s แทน ข้อมูลที่สังเคราะห์ใหม่

x_i แทน ข้อมูลเดิมที่สุ่มมา

$\hat{x}_i$ แทน ข้อมูลที่เป็นเพื่อนบ้านของ x_i

u แทน ค่าสุ่มที่อยู่ระหว่าง 0 – 1

5. วัดประสิทธิภาพของแต่ละวิธีการด้วยค่าความแม่นยำ (Accuracy) อัตราความถูกต้องในการทำนายกลุ่มส่วนน้อย (True Positive Rate) อัตราความถูกต้องในการทำนายกลุ่มส่วนมาก (True Negative Rate) และค่าการวัดเอฟ (F-measure) บนข้อมูลทดสอบ (Tan, Steinbach, & Kumar, 2006) โดยแสดงการคำนวณดังนี้

กำหนด Confusion Matrix แสดงผลของค่าจริงและผลการทำนาย ดังนี้

ค่าจริง	ผลการทำนาย	
	Positive (กลุ่มส่วนน้อย)	Negative (กลุ่มส่วนมาก)
Positive (กลุ่มส่วนน้อย)	TP	FN
Negative (กลุ่มส่วนมาก)	FP	TN

โดยที่ TP หมายถึง True Positive แสดงถึงจำนวนข้อมูลที่อยู่ในกลุ่ม Positive และทำนายว่าอยู่ในกลุ่ม Positive

FN หมายถึง False Negative แสดงถึง จำนวนข้อมูลที่อยู่ในกลุ่ม Positive แต่ทำนายว่าอยู่ในกลุ่ม Negative

FP หมายถึง False Positive แสดงถึงจำนวนข้อมูลที่อยู่ในกลุ่ม Negative แต่ทำนายว่าอยู่ในกลุ่ม Positive

TN หมายถึง True Negative แสดงถึงจำนวนข้อมูลที่อยู่ในกลุ่ม Negative และทำนายว่าอยู่ในกลุ่ม Negative

1) ค่าความแม่นยำ (Accuracy) แสดงถึงความถูกต้องในการทำนายในภาพรวม ทั้งกลุ่ม Positive และ Negative ซึ่งคำนวณได้ ดังนี้

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

2) อัตราความถูกต้องในการทำนายกลุ่มส่วนน้อย (True Positive Rate) แสดงถึงความถูกต้องในการทำนายกลุ่มส่วนน้อยว่าอยู่กลุ่มส่วนน้อย

$$\text{True Positive Rate} = \frac{TP}{TP + FN}$$

3) อัตราความถูกต้องในการทำนายกลุ่มส่วนมาก (True Negative Rate) แสดงถึงความถูกต้องในการทำนายกลุ่มส่วนมากว่าอยู่กลุ่มส่วนมาก

$$\text{True Negative Rate} = \frac{TN}{TN + FP}$$

4) ค่าการวัดเอฟ (F-measure) เป็นการวัดความถูกต้องโดยการใช้ค่าเฉลี่ยฮาร์โมนิคระหว่าง True Positive Rate และ Precision

$$\text{F-measure} = \frac{2 \times (TP\,\text{rate} \times \text{Precision})}{TP\,\text{rate} + \text{Precision}}$$

โดยที่ Precision หมายถึง ความเที่ยง ซึ่งแสดงถึงความถูกต้องของการทำนายกลุ่ม Positive เมื่อเทียบกับผลการทำนาย Positive ทั้งหมด ซึ่งคำนวนจาก

$$\text{Precision} = \frac{TP}{TP + FP}$$

ซึ่งค่าวัดแต่ละค่า ถ้ามีค่าสูงแสดงถึงประสิทธิภาพในการจำแนกประเภท

ผลการวิจัย

จากการเปรียบเทียบวิธีการสุ่มตัวอย่างซ้ำ ได้แก่ วิธี ROS และวิธี SMOTE โดยใช้วิธีการจำแนกประเภทด้วยการถดถอยลอจิสติก และต้นไม้ตัดสินใจ และวัดประสิทธิภาพบนข้อมูลทดสอบ ได้ผลการเปรียบเทียบค่าวัดประสิทธิภาพต่างๆ ดังนี้

ตารางที่ 1 การเปรียบเทียบประสิทธิภาพในวิธีการสุ่มตัวอย่างซ้ำแบบต่างๆ เมื่อจำแนกประเภทด้วยวิธีการถดถอยลอจิสติก

ค่าวัดประสิทธิภาพ	ข้อมูลเริ่มต้น	วิธีการเพิ่มข้อมูล	
		ROS	SMOTE
Accuracy	**0.7353**	0.7059	**0.7353**
True Positive Rate	0.5833	**0.7500**	**0.7500**
True Negative Rate	**0.8182**	0.6818	0.7273
F-measure	0.6087	0.6429	**0.6667**

ตัวเลขที่เป็นตัวเข้มแสดงถึงประสิทธิภาพสูงสุด

จากตารางที่ 1 ค่าความแม่นยำเมื่อไม่มีการสุ่มตัวอย่างซ้ำ และการสุ่มตัวอย่างซ้ำด้วยวิธี SMOTE จะให้ค่าสูงสุดเท่ากับ 0.7353 อัตราความถูกต้องในการทำนายกลุ่มส่วนน้อยมีค่าสูงสุดเมื่อมีการสุ่มตัวอย่างซ้ำทั้งวิธี ROS และ SMOTE มีค่าเท่ากับ 0.75 อัตราความถูกต้องในการทำนายกลุ่มส่วนมากมีค่าสูงสุดเมื่อไม่มีการสุ่มตัวอย่างซ้ำ ซึ่งมีค่าเท่ากับ 0.8182 และค่าการวัดเอฟจะสูงสุดเมื่อมีการสุ่มตัวอย่างซ้ำด้วยวิธี SMOTE ซึ่งมีค่าเท่ากับ 0.6667

ตารางที่ 2 การเปรียบเทียบประสิทธิภาพในวิธีการสุ่มตัวอย่างซ้ำแบบต่างๆ เมื่อจำแนกประเภทด้วยต้นไม้ตัดสินใจ

ค่าวัดประสิทธิภาพ	ข้อมูลเริ่มต้น	วิธีการเพิ่มข้อมูล	
		ROS	SMOTE
Accuracy	**0.7941**	0.7647	**0.7941**
True Positive Rate	0.6667	0.5833	**0.7500**
True Negative Rate	**0.8636**	**0.8636**	0.8182
F-measure	0.6957	0.6364	**0.7200**

ตัวเลขที่เป็นตัวเข้มแสดงถึงประสิทธิภาพสูงสุด

จากตารางที่ 2 ค่าความแม่นยำเมื่อไม่มีการสุ่มตัวอย่างซ้ำ และการสุ่มตัวอย่างซ้ำด้วยวิธี SMOTE จะให้ค่าสูงสุดเท่ากับ 0.7941 อัตราความถูกต้องในการทำนายกลุ่มส่วนน้อยมีค่าสูงสุดเมื่อมีการสุ่มตัวอย่างซ้ำด้วยวิธี SMOTE ซึ่งมีค่า

เท่ากับ 0.75 อัตราความถูกต้องในการทำนายกลุ่มส่วนมากมีค่าสูงสุดเมื่อไม่มีการสุ่มตัวอย่างซ้ำ และการสุ่มตัวอย่างซ้ำด้วยวิธี ROS ซึ่งมีค่าเท่ากับ 0.8636 และค่าการวัดเอฟจะสูงสุดเมื่อมีการสุ่มตัวอย่างซ้ำด้วยวิธี SMOTE ซึ่งมีค่าเท่ากับ 0.72

อภิปรายผลการวิจัย

การสุ่มตัวอย่างซ้ำโดยเพิ่มข้อมูลกลุ่มส่วนน้อยให้มีจำนวนใกล้เคียงกับกลุ่มส่วนมากด้วยวิธี SMOTE ให้ประสิทธิภาพที่ดีที่สุดทั้งการจำแนกด้วยการถดถอยลอจิสติก และต้นไม้ตัดสินใจ ในเกือบทุกค่าวัดประสิทธิภาพ ยกเว้นอัตราความถูกต้องในการทำนายกลุ่มส่วนมาก ซึ่งการไม่เพิ่มข้อมูลให้ผลดีที่สุด อันเนื่องมาจากวิธีการจำแนกประเภทโดยทั่วไปจะมีความโน้มเอียงที่จะทำนายตัวอย่างใหม่เป็นกลุ่มส่วนมาก แต่เมื่อพิจารณาอัตราความถูกต้องในการทำนายกลุ่มส่วนน้อย พบว่าการไม่เพิ่มข้อมูลให้มีความสมดุลมากขึ้น ทำให้อัตราความถูกต้องในการทำนายกลุ่มส่วนน้อย มีค่าต่ำ ในส่วนของวิธี ROS ซึ่งเป็นการสุ่มตัวอย่างเพิ่มข้อมูลเริ่มต้น เป็นการทำซ้ำเพื่อเพิ่มจำนวนตัวอย่างของกลุ่มส่วนน้อยจากข้อมูลเดิมซึ่งไม่ได้ทำให้ได้สารสนเทศของข้อมูลเพิ่มขึ้นเนื่องจากไม่ใช่ข้อมูลใหม่ จึงอาจส่งผลให้เกิด overfitting ซึ่งแตกต่างจากวิธี SMOTE ที่สร้างข้อมูลใหม่ จึงให้ผลลัพธ์ที่ดีกว่า สอดคล้องกับงานของ Yen & Lee (2009) และ Cateni, Colla, & Vannucci (2014)

ในการเปรียบเทียบวิธีการจำแนกประเภทข้อมูลระหว่างการถดถอยลอจิสติก และต้นไม้ตัดสินใจ เมื่อพิจารณาในภาพรวมด้วยค่าความแม่นยำพบว่า ต้นไม้ตัดสินใจให้ค่าวัดประสิทธิภาพสูงกว่าการถดถอยลอจิสติก แต่เมื่อพิจารณาอัตราความถูกต้องในการทำนายกลุ่มส่วนน้อยซึ่งมักเป็นกลุ่มสำคัญที่ต้องการค้นหา พบว่า การถดถอยลอจิสติกให้ค่าที่สูงกว่าหรือเท่ากับต้นไม้ตัดสินใจเมื่อมีการสุ่มตัวอย่างซ้ำไม่ว่าจะเป็นวิธี ROS หรือ SMOTE

ข้อเสนอแนะ

1. ในการทำนายกลุ่มรายได้ของผู้ประกอบการร้านยาประเภท ข.ย.1 ควรมีการสุ่มตัวอย่างเพิ่มข้อมูลกลุ่มผู้มีรายได้เกิน 150,000 บาทต่อเดือน ซึ่งเป็นกลุ่มส่วนน้อย ให้มีจำนวนใกล้เคียงกับกลุ่มที่มีรายได้ไม่เกิน 150,000 บาทต่อเดือน ซึ่งเป็นกลุ่มส่วนมาก ด้วยวิธี SMOTE และใช้วิธีการจำแนกประเภทด้วยต้นไม้ตัดสินใจ เพื่อให้มีประสิทธิภาพในการจำแนกสูงสุด
2. ความไม่สมดุลของข้อมูลกลุ่มรายได้ของผู้ประกอบการร้านยาประเภท ข.ย.1 ยังไม่มีความรุนแรงมาก การเปรียบเทียบวิธีการต่างๆ อาจยังไม่ครอบคลุมสถานการณ์ที่มีความรุนแรงของความไม่สมดุล ดังนั้นจึงควรทำการทดลองในกรณีที่มีความรุนแรงของความไม่สมดุลที่มากขึ้น

กิตติกรรมประกาศ

ขอขอบคุณสมาคมร้านขายยา ในการให้ทุนสนับสนุนการดำเนินงานวิจัย และมหาวิทยาลัยหัวเฉียวเฉลิมพระเกียรติในการให้เวลาการทำงานวิจัย

เอกสารอ้างอิง

กีระชาติ สุขสุทธิ์. (2559). การจำแนกข้อมูลไม่สมดุลโดยใช้การปปรับปรุงข้อมูลร่วมกับการหาค่าพารามิเตอร์ที่เหมาะสมด้วยขั้นตอนวิธีเชิงพันธุกรรม (วิทยานิพนธ์ปริญญาวิศวกรรมศาสตรดุษฎีบัณฑิต). มหาวิทยาลัยเทคโนโลยีสุรนารี, นครราชสีมา.

นพมาศ อัครจันทโชติ ศิริวรรณ ตันตระวาณิชย์ พิมพ์ภัค ภัทรนาวิก และอุมา รัตนเทพี. (2561). การจัดกลุ่มธุรกิจร้านยาประเภท ข.ย.1. *วารสารวิจัยมหาวิทยาลัยเทคโนโลยีราชมงคลศรีวิชัย*, 10(2), 289-306.

ระบบสถิติทางการทะเบียน. (2562). จำนวนประชากรแยกรายอายุ ทั่วประเทศ. สืบค้นเมื่อ 3 กันยายน 2562, จาก http://stat.dopa.go.th/stat/statnew/upstat_age_disp.php

ศูนย์วิจัยและจัดการความรู้เพื่อการควบคุมยาสูบ. (2561). *รายงานสถิติการบริโภคยาสูบของประเทศไทย พ.ศ. 2561*. กรุงเทพฯ: มหาวิทยาลัยมหิดล.

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

Blagus, R., & Lusa, L. (2015). Joint use of over- and under-sampling techniques and cross-validation for the development and assessment of prediction models. *BMC Bioinformatics*, 16, 1-10.

Cateni, S., Colla, V., & Vannucci, M. (2014). A method for resampling imbalanced dataset in binary classification tasks for real-world problems. *Neurocomputing*, 135, 32-41.

Chawla, N.V. , Bowyer,K.W., Hall, L.O., & Kegelmeyer, W.P. (2002). SMOTE: Synthetic Minority Over-sampling Technique. *Journal of Artificial Intelligence Research*, 16, 321-357.

hfocus. (2562). ปี 61 ไทยมีผู้ติดเชื้อเอชไอวีรายใหม่วันละ 17 คน. สืบค้นเมื่อ 3 กันยายน 2562, จาก https://www.hfocus.org/content/2019/07/17321

Jagelid, M., & Movin, M. (2017). A Comparison of Resampling Techniques to Handle the Class Imbalance Problem in Machine Learning Conversion prediction of Spotify Users - A Case Study. (Bachelor thesis in Computer Science). KTH Royal Institute of Technology. Stockholm.

Kaur, P. & Gosain, A. (2018). Comparing the Behavior of Oversampling and Undersampling Approach of Class Imbalance Learning by Combining Class Imbalance Problem with Noise. *ICT Based Innovations*, 23-30.

Minh, H. (2018). How to Handle Imbalanced Data in Classification Problems. สืบค้นเมื่อ 3 กันยายน 2562, จาก https://medium.com/james-blogs/handling-imbalanced-data-in-classification-problems-7de598c1059f.

Mishra, S. (2017). Handling Imbalanced Data: SMOTE vs. Random Undersampling. International research *Journal of Engineering and Technology*, 4(8), 317-320.

Tan, P.-N., Steinbach, M., & Kumar, V. (2006). *Introduction to Data Mining*. Boston: Pearson Addison-Wesley.

Yen, S., & Lee, Y-S. (2009). Cluster-based under-sampling approaches for imbalanced data distributions. *Expert Systems with Application*, 36, 5718-5727.