ประสิทธิภาพของวิธีการจัดการข้อมูลไม่สมดุลสำหรับการจำแนกกลุ่มภายใต้เงื่อนไขที่แตกต่างกัน

น.ส.กาญธนา ลออสิริกุล

วิทยานิพนธ์นี้เป็นส่วนหนึ่งของการศึกษาตามหลักสูตรปริญญาครุศาสตรมหาบัณฑิต
สาขาวิชาสถิติและสารสนเทศการศึกษา ภาควิชาวิจัยและจิตวิทยาการศึกษา
คณะครุศาสตร์ จุฬาลงกรณ์มหาวิทยาลัย
ปีการศึกษา 2565

Effectiveness of Handling Imbalanced Data Methods for Classification under Varied Conditions

Miss Kantana La-orsirikul

A Thesis Submitted in Partial Fulfillment of the Requirements
for the Degree of Master of Education in Educational Statistics and Information
Department of Educational Research and Psychology
FACULTY OF EDUCATION
Chulalongkorn University
Academic Year 2022

หัวข้อวิทยานิพนธ์	ประสิทธิภาพของวิธีการจัดการข้อมูลไม่สมดุลสำหรับการจำแนกกลุ่มภายใต้เงื่อนไขที่แตกต่างกัน
โดย	น.ส.กาญธนา ลออสิริกุล
สาขาวิชา	สถิติและสารสนเทศการศึกษา
อาจารย์ที่ปรึกษาวิทยานิพนธ์หลัก	อาจารย์ ดร.ประภาศิริ รัชชประภาพรกุล
อาจารย์ที่ปรึกษาวิทยานิพนธ์ร่วม	อาจารย์ ดร.สุรศักดิ์ เก้าเอี้ยน

คณะครุศาสตร์ จุฬาลงกรณ์มหาวิทยาลัย อนุมัติให้นับวิทยานิพนธ์ฉบับนี้เป็นส่วนหนึ่งของการศึกษาตามหลักสูตรปริญญาครุศาสตรมหาบัณฑิต

..... คณบดีคณะครุศาสตร์
(รองศาสตราจารย์ ดร.ศิริเดช สุชีวะ)

คณะกรรมการสอบวิทยานิพนธ์

..... ประธานกรรมการ
(รองศาสตราจารย์ ดร.สังวรณ์ งัดกระโทก)

..... อาจารย์ที่ปรึกษาวิทยานิพนธ์หลัก
(อาจารย์ ดร.ประภาศิริ รัชชประภาพรกุล)

..... อาจารย์ที่ปรึกษาวิทยานิพนธ์ร่วม
(อาจารย์ ดร.สุรศักดิ์ เก้าเอี้ยน)

..... กรรมการ
(ผู้ช่วยศาสตราจารย์ ดร.สิวะโชติ ศรีสุทธิยากร)

..... กรรมการภายนอกมหาวิทยาลัย
(รองศาสตราจารย์ ดร.สังวรณ์ งัดกระโทก)

กาญธนา ลออสิริกุล : ประสิทธิภาพของวิธีการจัดการข้อมูลไม่สมดุลสำหรับการจำแนกกลุ่มภายใต้เงื่อนไขที่แตกต่างกัน. (Effectiveness of Handling Imbalanced Data Methods for Classification under Varied Conditions) อ.ที่ปรึกษาหลัก : อ. ดร.ประภาศิริ รัชชประภาพรกุล, อ.ที่ปรึกษาร่วม : อ. ดร.สุรศักดิ์ เก้าเอี้ยน

การวิจัยนี้มีจุดประสงค์เพื่อศึกษาปฏิสัมพันธ์ของวิธีการปรับสมดุลข้อมูลกับเงื่อนไขด้านขนาดตัวอย่าง เทคนิคการจำแนกข้อมูล จำนวนตัวแปรระหว่างกลุ่มตัวแปรจัดประเภทต่อกลุ่มตัวแปรต่อเนื่อง อัตราออด และร้อยละของจำนวนข้อมูลกลุ่มหลักต่อข้อมูลกลุ่มรองที่มีต่อประสิทธิภาพของการจำแนกกลุ่ม การปรับสมดุลของข้อมูลแบ่งออกเป็น 3 วิธี ได้แก่ (1) ไม่ปรับสมดุล (2) วิธี random oversampling และ (3) วิธีผสมผสานระหว่างรูปแบบสุ่มเกินและสุ่มลด (hybrid) โดยใช้แพคเกจ ROSE ส่วนเงื่อนไขด้านขนาดตัวอย่างแบ่งออกเป็น ขนาดตัวอย่างเท่ากับ 100 300 และ 500 หน่วย ด้านเทคนิคการจำแนกข้อมูล แบ่งออกเป็น 4 วิธี ได้แก่ (1) เคเนียร์เรสเนเบอร์ (2) การถดถอยโลจิสติก (3) แรนดอมฟอร์เรส และ (4) ซัพพอร์ตเวกเตอร์แมชชีน ตัวแปรจากการจำลองแบ่งออกเป็นตัวแปรตามซึ่งจำลองด้วยการถดถอยโลจิสติก ส่วนตัวแปรอิสระในการจำลองข้อมูลครั้งนี้จะกำหนดให้ใช้ตัวแปรอิสระจำลองทั้งหมด 8 ตัว โดยกำหนดให้มีจำนวนตัวแปรระหว่างกลุ่มตัวแปรจัดประเภทต่อกลุ่มตัวแปรต่อเนื่อง 3 กรณี คือ 4:4 5:3 และ 6:2 ในขณะที่ระดับของอัตราออด จะสุ่มค่าจากช่วง [1,2) หรือ [2,3) และร้อยละของข้อมูลระหว่างข้อมูลกลุ่มหลักต่อข้อมูลกลุ่มรอง แบ่งออกเป็น 2 กรณี ได้แก่ 60:40 และ 70:30 พิจารณาเกณฑ์ประสิทธิภาพของข้อมูลด้วยตัวชี้วัดความถูกต้องในการจำแนก ความไว และความจำเพาะ การจำลองแต่ละสถานการณ์จะทำซ้ำสถานการณ์ละ 500 รอบ การวิเคราะห์ปฏิสัมพันธ์ระหว่างวิธีการปรับสมดุลข้อมูลกับเงื่อนไขต่าง ๆ ใช้การวิเคราะห์ความแปรปรวนพหุคูณหลายทาง (n-way MANOVA)

ผลการวิจัยพบว่า วิธีการปรับสมดุลข้อมูลมีปฏิสัมพันธ์แบบสองทางกับเงื่อนไขด้านขนาดตัวอย่าง ร้อยละของข้อมูลระหว่างข้อมูลกลุ่มหลักต่อข้อมูลกลุ่มรอง อัตราออด และเทคนิคการจำแนกข้อมูล และพบปฏิสัมพันธ์แบบสามทางกับเงื่อนไขต่อไปนี้ (1) ขนาดตัวอย่างและจำนวนตัวแปรระหว่างกลุ่มตัวแปรจัดประเภทต่อกลุ่มตัวแปรต่อเนื่อง (2) ขนาดตัวอย่างและเทคนิคการจำแนกข้อมูล และ (3) ร้อยละของข้อมูลระหว่างข้อมูลกลุ่มหลักต่อข้อมูลกลุ่มรอง และเทคนิคการจำแนกข้อมูล ดังนั้นนักวิเคราะห์ข้อมูลควรเลือกใช้วิธีการปรับสมดุลข้อมูลโดยพิจารณาให้เหมาะสมกับสภาพของข้อมูลที่ใช้ในการวิเคราะห์

สาขาวิชา	สถิติและสารสนเทศการศึกษา	ลายมือชื่อนิสิต .....
ปีการศึกษา	2565	ลายมือชื่อ อ.ที่ปรึกษาหลัก .....
		ลายมือชื่อ อ.ที่ปรึกษาร่วม .....

6282002827 : MAJOR EDUCATIONAL STATISTICS AND INFORMATION

KEYWORD: Imbalance Data, Categorical Variable, Classification, Simulation, Over Sampling Methods, Hybrid Methods

Kantana La-orsirikul : Effectiveness of Handling Imbalanced Data Methods for Classification under Varied Conditions. Advisor: Ph.D. PRAPASIRI RATCHAPRAPAPORNKUL, Ph.D. Co-advisor: Prof. SURASAK KAO-IEAN

The research aimed to study interaction effects between data balancing methods and data conditions included sample sizes, classification techniques, the number of variables between categorical variables and continuous variables, odds ratio, and percentage of data between majority and minority class. Data balancing methods divided into 3 methods, consisted of (1) do nothing (2) random oversampling and (3) combination between oversampling and undersampling (hybrid), using ROSE package. Conditions of sample sizes were included 100, 300, and 500. Classification techniques used in the study were (1) K-nearest neighbor, (2) logistic regression, (3) random forest and (4) support vector machine. Variables for classification analysis consisted of a dependent variable, which was simulated using logistic regression model, and 8 simulated independent variables. The number of variables between categorical variables and continuous variables were 4:4, 5:3, and 6:2, while levels of odds ratio were randomized from [1,2) or [2,3). The percentage of data between majority and minority class consisted of 60:40 and 70:30. 3 criterion of classification modeling were considered in this study included accuracy, sensitivity, and specificity. Each simulation was repeated 500 times. Interaction effects between data balancing methods and any conditions were analyzed using n-way MANOVA.

The result revealed that data balancing methods had 2-way interaction effects with sample sizes, percentage of data between majority and minority class, odds ratio, and classification techniques. Moreover, it had 3-way interaction effects with following terms: (1) sample sizes and the number of variables between categorical variables and continuous variables, (2) sample sizes and classification techniques, and (3) percentage of data between majority and minority class and classification techniques. Therefore, the analyst should choose the appropriate data balancing methods with data conditions.

Field of Study:	Educational Statistics and Information	Student's Signature .....
Academic Year:	2022	Advisor's Signature .....
		Co-advisor's Signature .....

กิตติกรรมประกาศ

วิทยานิพนธ์ฉบับนี้สำเร็จลุล่วงได้ด้วยความช่วยเหลือ ความเมตตาและความกรุณาอย่างยิ่งจาก อาจารย์ ดร.ประภาศิริ รัชชประภาพรกุล อาจารย์ที่ปรึกษาหลัก และอาจารย์ ดร.สุรศักดิ์ เก้าเอี้ยน อาจารย์ที่ปรึกษาร่วม ที่ได้สละเวลาอันมีค่าให้คำปรึกษา คำแนะนำ แนวทางแก้ปัญหา และกำลังใจตลอดระยะเวลาที่ทำวิทยานิพนธ์ให้มีความถูกต้องสมบูรณ์

ขอกราบขอบพระคุณ รองศาสตราจารย์ ดร.สังวรณ์ งัดกระโทก ประธานและกรรมการสอบวิทยานิพนธ์และผู้ช่วยศาสตราจารย์ ดร.สิวะโชติ ศรีสุทธิยากร กรรมการสอบวิทยานิพนธ์ ที่กรุณาตรวจสอบและให้คำแนะนำอันเป็นประโยชน์ในการปรับปรุงแก้ไขวิทยานิพนธ์ให้มีความสมบูรณ์ และขอกราบขอบพระคุณ รองศาสตราจารย์ ดร.สุชาดา บวรกิติวงศ์ และคณาจารย์ภาควิชาวิจัยและจิตวิทยาการศึกษาทุกท่านที่ได้มอบความรู้อันมีค่าและอบรมสั่งสอนตลอดระยะเวลาการศึกษา

ขอกราบขอบพระคุณผู้อำนวยการสถานศึกษา บุคลากรครู และนักเรียนโรงเรียนสตรีสมุทรปราการที่ให้กำลังใจและเป็นแรงสนับสนุนที่ดีเสมอมา

ขอขอบคุณเพื่อน พี่ น้องนิสิตในสาขาวิชาสถิติการศึกษา พี่กานต์และเพื่อนมัธยมที่ได้ให้ความช่วยเหลือ ให้กำลังใจ และให้คำแนะนำในการแก้ปัญหาเสมอมา

ขอกราบขอบพระคุณครอบครัว ทั้งคุณยาย คุณพ่อ คุณแม่ คุณน้า คุณอา และน้องสาวที่คอยเลี้ยงดู ให้ความรัก ความเอาใจใส่ ให้ความช่วยเหลือ และเป็นกำลังใจสำคัญให้ผู้วิจัยตลอดเวลา

ท้ายที่สุดนี้ ขอกราบขอบพระคุณทุกท่านที่มิได้กล่าวนามมา ณ ที่นี้ ที่มีส่วนช่วยให้วิทยานิพนธ์เล่มนี้สำเร็จลุล่วงไปได้ด้วยดี

กาญธนา ลออสิริกุล

สารบัญ

หน้า

สารบัญตาราง

หน้า

สารบัญรูปภาพ

หน้า

บทที่ 1
บทนำ

ความเป็นมาและความสำคัญของปัญหา

ปัจจุบันเทคโนโลยีสารสนเทศและคอมพิวเตอร์มีความก้าวหน้ามากขึ้น ทำให้ข้อมูลที่เก็บรวบรวมได้มีความหลากหลาย นักวิจัยจึงให้ความสำคัญกับการจำแนกประเภทข้อมูลให้มีประสิทธิภาพสูง ซึ่งเทคนิคการจำแนกข้อมูล (classification) เป็นเทคนิคที่สร้างโมเดลเพื่อใช้ทำนายหมวดหมู่ของข้อมูลที่มีความสัมพันธ์เกี่ยวข้องกันจะอยู่ในกลุ่มเดียวกัน ถือเป็นเทคนิคที่สำคัญและถูกใช้อย่างแพร่หลาย โดยเทคนิคการจำแนกประเภทข้อมูลมีหลายเทคนิค ได้แก่ แรนดอมฟอร์เรส (random forest) การวิเคราะห์การถดถอยโลจิสติก (logistic regression analysis) ซัพพอร์ตเวกเตอร์แมชชีน (support vector machine) วิธีเคเนียร์เรสเนเบอร์ (k-nearest neighbor) เป็นต้น (Chokthanahirun Siraphong & Hengpraprohm Supoj, 2016; Gordon, 1999; Kesavaraj & Sukumaran, 2013; Phyu, 2009) โดยเทคนิคการจำแนกประเภทข้อมูลจะมีประสิทธิภาพในการจำแนกสูงขึ้น เมื่อเครื่องเรียนรู้จากข้อมูลที่สมดุล (balanced data) และประสิทธิภาพจะลดลงเมื่อข้อมูลที่ถูกจำแนกเป็นข้อมูลไม่สมดุล (imbalanced data)

ข้อมูลไม่สมดุล (imbalanced data) คือ การที่มีจำนวนของข้อมูลกลุ่มหนึ่งแตกต่างจากจำนวนข้อมูลกลุ่มอื่น ๆ มาก โดยข้อมูลที่มีจำนวนสมาชิกมากกว่ากลุ่มอื่น เรียกว่า ข้อมูลกลุ่มหลัก (majority) และข้อมูลที่มีจำนวนสมาชิกของข้อมูลน้อยกว่า เรียกว่า ข้อมูลกลุ่มรอง (minority) (Chawla et al., 2002; Chawla et al., 2004; Farquad & Bose, 2012; เบญจภรณ์ จันทรกองกุล และคณะ, 2557) ตัวอย่างข้อมูลไม่สมดุล ได้แก่ จำนวนของนักเรียนหรือนักศึกษาที่ออกกลางคัน พบว่ามีจำนวนของนักเรียนหรือนักศึกษาที่ออกกลางคันน้อยกว่าจำนวนนักเรียนหรือนักศึกษาที่เรียนจบหลักสูตรมาก (Christian & Ayub, 2014; Hussain et al., 2018; Mousa & Maghari, 2017; เจนวิทย์ วารีบ่อ, 2562; พงษ์ศักดิพล ทาแก้ว, 2563; สาโรจน์ ขอจ่วนเตี๋ยว, 2559; สุกัญญา ทารส, 2562; อริสฬา เตหลิ่ม, 2563) จำนวนเด็กและเยาวชนบนท้องถนน (children in street) พบว่าจำนวนเด็กและเยาวชนบนท้องถนนน้อยกว่าจำนวนเด็กและเยาวชนที่ไม่ได้เป็นเด็กและเยาวชนบนท้องถนนมาก (กองทุนเพื่อความเสมอภาคทางการศึกษา, 2563; นรรจชนก กริชฤทธิเศรษฐ์, 2562; สมพงษ์ จิตระดับ และคณะ, 2563, 31 มกราคม) จำนวนนักเรียนหรือนักศึกษาหรือบุคลากรทางการศึกษาที่มีภาวะเหนื่อยหน่ายหรือหมดไฟในการเรียนการทำงาน พบว่า มีจำนวนน้อยกว่าจำนวนนักเรียนหรือนักศึกษาหรือบุคลากรทางการศึกษาที่ไม่มีภาวะเหนื่อยหน่ายหรือหมดไฟในการเรียนการทำงานมาก (Manee & Pontanya, 2022; Pongpisanrat, 2022; ฉัตรชกรณ์ ระบิล และวิลาสินี

จินตลิขิตดี, 2564) ทั้งนี้จากข้อมูลส่วนใหญ่พบว่ามีจำนวนข้อมูลกลุ่มหลักอยู่ในช่วงร้อยละ 60-80 ถ้านำข้อมูลกลุ่มหลักและกลุ่มรองให้เครื่องเรียนรู้และจำแนกกลุ่มข้อมูล จะทำให้ผลการจำแนกกลุ่มข้อมูลมีความผิดพลาด เกิดปัญหาที่เรียกว่า “ปัญหาการจำแนกข้อมูลผิดกลุ่ม (misclassification)” ซึ่งเป็นปัญหาของการที่ข้อมูลในกลุ่มรองถูกครอบงำหรือจะถูกจัดให้ไปอยู่ในกลุ่มหลัก (เบญจภรณ์ จันทรกองกุล และคณะ, 2557) และในความเป็นจริงการจะกำหนดจำนวนสมาชิกข้อมูลกลุ่มรองและกลุ่มหลักให้เท่ากันเป็นเรื่องยากหรืออาจสูญเสียสารสนเทศ นอกจากนี้แล้วการเก็บรวบรวมข้อมูลทางสังคมศาสตร์ในบางครั้งก็ทำการสำรวจและเก็บรวบรวมได้ยาก เนื่องจากเป็นข้อมูลของกลุ่มที่หาได้ยาก (rare case) เป็นกลุ่มเฉพาะ หรือเป็นกลุ่มเปราะบาง เช่น ข้อมูลของเด็กและและเยาวชนบนท้องถนน ซึ่งผู้ให้ข้อมูลมีข้อจำกัดในเรื่องของการใช้ภาษาและการอ่านออกเขียนได้ ทำให้ตัวแปรส่วนใหญ่เป็นตัวแปรจัดประเภทและข้อคำถามที่ใช้ส่วนใหญ่เป็นคำถามแบบใช่ ไม่ใช่ เช่น เพศ การอยู่อาศัยกับครอบครัวหรืออยู่คนเดียว การมีหรือไม่มีเพื่อน การเข้าศึกษาต่อหรือไม่ศึกษาต่อ การเคยถูกทำร้ายหรือไม่เคยถูกทำร้าย เป็นต้น ข้อมูลที่ได้รับจึงเป็นข้อมูลนามบัญญัติมากกว่าข้อมูลแบบต่อเนื่อง (กองทุนเพื่อความเสมอภาคทางการศึกษา, 2563; พิชามญชุ์ โตโฉมงาม, 2553; มิรันตี วรอุไร, 2009; อิศรา ชอบขาย, 2550) แต่งานวิจัยที่มีการวิเคราะห์และจำแนกข้อมูลส่วนใหญ่มักศึกษาในข้อมูลที่มีตัวแปรส่วนใหญ่เป็นตัวแปรต่อเนื่องมากกว่าตัวแปรจัดประเภท (ศิรินทรา เสือพิทักษ์ และคณะ, 2564; สิริกุล รัตนมณี และคณะ, 2561) ดังนั้นจึงเป็นปัญหาที่ท้าทายและน่าสนใจในการหาขั้นตอนวิธีการที่เหมาะสมเพื่อใช้ในการจำแนกกลุ่มข้อมูลไม่สมดุลที่มีข้อมูลส่วนใหญ่เป็นข้อมูลนามบัญญัติ

จากที่กล่าวมาข้างต้นนักวิจัยได้คิดค้นวิธีการในการแก้ปัญหาข้อมูลไม่สมดุล โดยเป็นหลักการจัดการข้อมูลก่อนกระบวนการสร้างตัวแบบ ซึ่งหลักการแบ่งออกเป็น 3 ระดับคือ 1) การแก้ปัญหาข้อมูลไม่สมดุลที่ระดับข้อมูล (data level solutions) เป็นการใช้เทคนิคการสุ่มเลือกข้อมูลเพื่อปรับข้อมูลไม่สมดุลให้กลายเป็นข้อมูลสมดุล ซึ่งเป็นการแก้ปัญหาในขั้นตอนก่อนการประมวลผล ประกอบด้วย 3 วิธี ได้แก่ วิธีสุ่มเกิน (over sampling methods) วิธีสุ่มลด (under sampling methods) และวิธีผสมผสาน (hybrid methods) ซึ่งวิธีสุ่มลดจะทำให้สูญเสียข้อมูลที่เก็บรวบรวมมาได้ ดังนั้นในงานวิจัยนี้จึงเลือกใช้เฉพาะวิธี random oversampling และวิธีผสมผสานระหว่างรูปแบบสุ่มเกินและสุ่มลด 2) การแก้ปัญหาข้อมูลไม่สมดุลที่ระดับขั้นตอนวิธีการ (algorithmic level solutions) เป็นกระบวนการปรับการเรียนรู้ของอัลกอริทึมมาตรฐานในการจำแนกข้อมูลเดิมให้สามารถเรียนรู้ข้อมูลไม่สมดุลโดยให้มีการเอนเอียงไปทางข้อมูลกลุ่มรอง และ 3) การแก้ปัญหาข้อมูลไม่สมดุลด้วยการเรียนรู้แบบมีค่าใช้จ่าย (cost-sensitive training) (Teeranai, 2019; Yuanwei Zhu et al., 2020; ภาสพิชญ์ ชูใจ, 2557)

เทคนิคการปรับสมดุลข้อมูลและการจำแนกข้อมูลส่วนใหญ่จะมีความเหมาะสมและมีประสิทธิภาพสูง เมื่อใช้กับข้อมูลที่มีจำนวนตัวแปรอิสระที่เป็นตัวแปรต่อเนื่องมากกว่าตัวแปร

จัดประเภทหรือมีตัวแปรอิสระเป็นตัวแปรต่อเนื่องทั้งหมด เช่น เทคนิค SMOTE (synthetic minority over sampling) (Chawla et al., 2002) และในงานวิจัยส่วนใหญ่เปรียบเทียบประสิทธิภาพการจำแนกโดยศึกษาในชุดข้อมูลมีตัวแปรอิสระส่วนใหญ่เป็นตัวแปรต่อเนื่อง เช่น งานวิจัยของ อัจฉรา แผ้วบาง and สายชล สินสมบูรณ์ทอง (2020) ที่ศึกษาการปรับความไม่สมดุลของข้อมูลด้วยการจำแนก 5 วิธี กับชุดข้อมูล 3 ชุด ได้แก่ ชุดข้อมูลเคมีบาบัดมะเร็งลำไส้ใหญ่ระยะ B/C ชุดข้อมูลโรคที่มีความผิดปกติของโปรตีน และชุดข้อมูลการรักษาอาการปวดศีรษะขั้นรุนแรง และงานวิจัยของ (พัชรียา ทองพูล และคณะ, 2562) ที่ศึกษาการเปรียบเทียบประสิทธิภาพในการทำนายผลการปรับความไม่สมดุลของข้อมูลในการจำแนกด้วยเทคนิคการทำเหมืองข้อมูลกับชุดข้อมูล 3 ชุด คือ ชุดข้อมูลการรับรู้ทางหูของเด็กที่มีภาวะน้ำคั่งในหูชั้นกลางหรือหูน้ำหนวก ชุดข้อมูลยอดคงเหลือในบัตรเครดิตของลูกค้า และชุดข้อมูลคุณภาพไวน์แดง

จากปัญหาดังกล่าวนักวิจัยส่วนใหญ่ให้ความสำคัญกับประสิทธิภาพในการจำแนกข้อมูลที่อยู่ในกลุ่มรอง ให้มีประสิทธิภาพในการจำแนกสูงขึ้นภายใต้แนวคิดของการใช้เทคนิคการสุ่มซ้ำ (resampling) เช่น งานวิจัยของ Cateni et al. (2014) ที่เพิ่มจำนวนข้อมูลในกลุ่มรองให้มีจำนวนใกล้เคียงกับจำนวนข้อมูลในกลุ่มหลักด้วยวิธีสุ่มเกิน และลดจำนวนข้อมูลกลุ่มหลักด้วยวิธีสุ่มลด จากนั้นจำแนกข้อมูลกับชุดข้อมูลจริง หรืองานวิจัยของ Gameng et al. (2019) ที่เมื่อปรับข้อมูลให้สมดุล แล้วจึงจำแนกข้อมูลด้วยการถดถอยโลจิสติกและแรนดอมฟอรเรส หรืองานวิจัยของ Yan Zhu et al. (2020) ที่ใช้วิธีผสมผสานในการปรับข้อมูลให้สมดุล แล้วจำแนกข้อมูลด้วยแรนดอมฟอรเรส เคเนียร์เรสเนเบอร์ ซัพพอร์ตเวกเตอร์แมชชีน นาอีฟเบย์ และการวิเคราะห์แบบจำแนกประเภท หรืองานวิจัยของ Zhu et al. (2021) ที่ใช้การสุ่มเกินในการปรับสมดุลข้อมูลร่วมกับวิธีแรนดอมฟอร์เรส วิธีการเคเนียร์เรสเนเบอร์ (K-Nearest Neighbors) และวิธีซัพพอร์ตเวคเตอร์แมชชีน และจากงานวิจัยที่เก็บข้อมูลในสถานการณ์จริง เช่น ในงานวิจัยของ สุกัญญา ทารส (2562) หรืองานวิจัยของ พัชรียา ทองพูล และคณะ (2562) หรืองานวิจัยของ Prasad et al. (2016) หรืองานวิจัยของ Mousa and Maghari (2017) หรืองานวิจัยของ Hussain et al. (2018) หรืองานวิจัยของ Bach et al. (2019) ข้อมูลที่ได้เป็นข้อมูลที่ไม่สมดุล โดยมีจำนวนข้อมูลกลุ่มรองในช่วงร้อยละ 30 ถึง 40 ขนาดตัวอย่างอยู่ในช่วง 100 ถึง 500 ตัวอย่าง มีอัตราออดของตัวแปรอิสระตั้งแต่ 1 ถึง 3 และมีจำนวนตัวแปรจัดประเภทมากกว่าจำนวนตัวแปรต่อเนื่อง

ดังนั้นการศึกษาครั้งนี้จึงจำลองข้อมูลเพื่อศึกษาอิทธิพลปฏิสัมพันธ์ระหว่างวิธีการปรับสมดุลข้อมูล ได้แก่ วิธีการไม่ปรับสมดุลข้อมูล วิธี random oversampling และวิธีผสมผสานระหว่างรูปแบบสุ่มเกินและสุ่มลดในการปรับสมดุลข้อมูล (hybrid) ภายใต้เงื่อนไขที่แตกต่างกัน คือ (1) ขนาดตัวอย่าง (2) เทคนิคการจำแนกข้อมูล ได้แก่ แรนดอมฟอร์เรส (random forest) การถดถอยโลจิสติก (logistic regression) และซัพพอร์ตเวคเตอร์แมชชีน (support vector machine) (3) จำนวน

ตัวแปรระหว่างกลุ่มตัวแปรจัดประเภทต่อกลุ่มตัวแปรต่อเนื่อง (4) อัตราออด และ (5) ร้อยละของจำนวนข้อมูลระหว่างข้อมูลกลุ่มหลักต่อข้อมูลกลุ่มรองที่มีต่อประสิทธิภาพของการจำแนกกลุ่ม โดยการจำลองนี้จะกระทำซ้ำหลายรอบแล้วนำผลประสิทธิภาพ ได้แก่ ค่าความถูกต้อง (accuracy) ค่าความไว (sensitivity) และค่าความจำเพาะ (specificity) ไปวิเคราะห์เพื่อหาเงื่อนไขที่มีอิทธิพลปฏิสัมพันธ์กับวิธีการปรับสมดุลข้อมูล

คำถามวิจัย

1. ปฏิสัมพันธ์ของวิธีการปรับสมดุลข้อมูลกับเงื่อนไขขนาดตัวอย่าง เทคนิคการจำแนกข้อมูล จำนวนตัวแปรระหว่างกลุ่มตัวแปรจัดประเภทต่อกลุ่มตัวแปรต่อเนื่อง อัตราออด และร้อยละของจำนวนข้อมูลระหว่างข้อมูลกลุ่มหลักต่อข้อมูลกลุ่มรองที่มีต่อประสิทธิภาพของการจำแนกกลุ่มเป็นอย่างไร

วัตถุประสงค์ของการวิจัย

1. เพื่อศึกษาปฏิสัมพันธ์ของวิธีการปรับสมดุลข้อมูลกับเงื่อนไขขนาดตัวอย่าง เทคนิคการจำแนกข้อมูล จำนวนตัวแปรระหว่างกลุ่มตัวแปรจัดประเภทต่อกลุ่มตัวแปรต่อเนื่อง อัตราออด และร้อยละของจำนวนข้อมูลระหว่างข้อมูลกลุ่มหลักต่อข้อมูลกลุ่มรองที่มีต่อประสิทธิภาพของการจำแนกกลุ่ม

ขอบเขตของการวิจัย

การศึกษาประสิทธิภาพของการจำแนกข้อมูลทางการศึกษา ภายใต้เงื่อนไขที่ข้อมูลไม่สมดุล ข้อมูลที่ใช้ในการศึกษาได้จากการจำลองข้อมูลภายใต้โปรแกรม R โดยมีจำนวนตัวแปรอิสระ 8 ค่า ที่เป็นอิสระต่อกัน แบ่งเงื่อนไขออกเป็น 6 ประเด็น ได้แก่

1. ขนาดตัวอย่างที่ใช้ในการศึกษาครั้งนี้ ประกอบไปด้วย 3 ค่า ได้แก่ 100 300 และ 500

2. วิธีการปรับสมดุลข้อมูลที่ใช้ในการศึกษาครั้งนี้ ประกอบไปด้วย 2 วิธี ได้แก่ วิธี random oversampling และวิธีผสมผสานระหว่างรูปแบบสุ่มเกินและสุ่มลด

3. เทคนิคการจำแนกข้อมูลที่ใช้ในการศึกษาครั้งนี้ ประกอบไปด้วย 4 เทคนิค ได้แก่ การถดถอยโลจิสติก (logistic regression) แรนดอมฟอรเรส (random forest) ซัพพอร์ตเวคเตอร์แมชชีน (support vector machine) และวิธีเคเนียร์เรสเนเบอร์ (k-nearest neighbor)

4. จำนวนตัวแปรระหว่างกลุ่มตัวแปรจัดประเภทต่อกลุ่มตัวแปรต่อเนื่องที่ใช้ในการศึกษาครั้งนี้ ประกอบไปด้วย 5 กรณี ได้แก่ 4:4, 5:3 และ 6:2

5. ร้อยละของจำนวนข้อมูลระหว่างข้อมูลกลุ่มหลักต่อข้อมูลกลุ่มรองที่ใช้ในการศึกษาครั้งนี้ ประกอบไปด้วย 2 กรณี ได้แก่ 60:40 และ 70:30

6. อัตราออด (odds ratio) ของตัวแปรอิสระกลุ่มที่เป็นตัวแปรจัดประเภทและกลุ่มที่เป็นตัวแปรต่อเนื่องสุ่มจากช่วง [1, 2) หรือ [2, 3) โดยแบ่งออกเป็น 4 ดังนี้ คือ [1,2) กับ [1,2), [1,2) กับ [2,3), [2,3) กับ [1,2) และ [2,3) กับ [2,3)

คำจำกัดความที่ใช้ในการวิจัย

การปรับสมดุลข้อมูล หมายถึง กระบวนการปรับข้อมูลที่ไม่สมดุล ซึ่งเป็นข้อมูลที่มีการกระจายตัวที่ไม่เท่าเทียมกัน หรือข้อมูลซึ่งจำนวนสมาชิกในกลุ่มหลักและกลุ่มรองมีจำนวนไม่เท่ากัน ให้เป็นข้อมูลที่มีจำนวนสมาชิกในกลุ่มหลักและกลุ่มรองใกล้เคียงกันหรือเท่ากัน

ข้อมูลกลุ่มหลัก หมายถึง ข้อมูลกลุ่มที่มีจำนวนข้อมูลจำนวนมากเมื่อเทียบกับจำนวนข้อมูลทั้งหมด

ข้อมูลกลุ่มรอง หมายถึง ข้อมูลกลุ่มที่มีจำนวนข้อมูลจำนวนน้อยเมื่อเทียบกับจำนวนข้อมูลทั้งหมด

ประโยชน์ที่ได้รับ

1. สร้างองค์ความรู้ในการเพิ่มประสิทธิภาพการจำแนกข้อมูลที่ไม่สมดุลด้วยวิธีการปรับสมดุลข้อมูลแต่ละวิธี ที่เหมาะสมกับการนำไปใช้ภายใต้สถานการณ์ต่าง ๆ ที่มีข้อมูลที่สนใจศึกษาเป็นข้อมูลกลุ่มรองที่มีความเปราะบาง และส่วนใหญ่เป็นตัวแปรจัดประเภท ก่อนนำไปวิเคราะห์ด้วยโมเดลอื่น ๆ เพื่อให้ผลการวิเคราะห์มีประสิทธิภาพมากยิ่งขึ้น

2. หน่วยงานที่เกี่ยวข้องกับการศึกษาและนักวิจัยที่สนใจศึกษาการจำแนกข้อมูลที่ไม่สมดุลสามารถนำวิธีการปรับสมดุลและการจำแนกข้อมูลที่มีประสิทธิภาพสูงจากงานวิจัยนี้ไปใช้ในการวิเคราะห์ผลในการวิจัยต่าง ๆ

บทที่ 2
วรรณกรรมที่เกี่ยวข้อง

งานวิจัยนี้มุ่งศึกษาปฏิสัมพันธ์ของวิธีการปรับสมดุลข้อมูลกับเงื่อนไขขนาดตัวอย่าง เทคนิคการจำแนกข้อมูล จำนวนตัวแปรระหว่างกลุ่มตัวแปรจัดประเภทต่อกลุ่มตัวแปรต่อเนื่อง อัตราออดและร้อยละของจำนวนข้อมูลระหว่างข้อมูลกลุ่มหลักต่อข้อมูลกลุ่มรองที่มีต่อประสิทธิภาพของการจำแนกกลุ่ม โดยผู้วิจัยได้ศึกษาแนวคิด ทฤษฎี และงานวิจัยที่เกี่ยวข้องนำเสนอแบ่งได้เป็น 5 ตอน ดังนี้

ตอนที่ 1 การปรับสมดุลข้อมูล

ตอนที่ 2 เทคนิคการจำแนกข้อมูล

ตอนที่ 3 การแจกแจงความน่าจะเป็น

ตอนที่ 4 การวัดประสิทธิภาพการปรับสมดุลข้อมูล

ตอนที่ 5 กรอบแนวคิดการวิจัย

ตอนที่ 1 การปรับสมดุลข้อมูล

การวิเคราะห์จำแนกประเภทข้อมูลจะมีประสิทธิภาพเมื่อใช้กับข้อมูลสมดุล ในงานวิจัยนี้จึงทบทวนเอกสารและวรรณกรรมเกี่ยวกับการปรับสมดุลข้อมูล มีรายละเอียดดังต่อไปนี้

1.1 ความหมายและลักษณะของข้อมูลไม่สมดุล

ข้อมูลไม่สมดุล หมายถึง ข้อมูลที่มีจำนวนสมาชิกในกลุ่มหลักและกลุ่มรองไม่เท่ากันหรือไม่ใกล้เคียงกัน (เบญจภรณ์ จันทรกองกุล และคณะ, 2557)

ลักษณะโดยทั่วไปของข้อมูลไม่สมดุล คือ ข้อมูลที่มีจำนวนข้อมูลของกลุ่มหนึ่งมากกว่าหรือน้อยกว่าจำนวนข้อมูลของกลุ่มอื่น ๆ เป็นจำนวนมาก (Chawla et al., 2002; Chawla et al., 2004) โดยเรียกข้อมูลกลุ่มที่มีจำนวนมากว่า ข้อมูลกลุ่มหลัก (majority class หรือ negative class) และเรียกข้อมูลกลุ่มที่มีจำนวนน้อยว่า ข้อมูลกลุ่มรอง (minority class หรือ positive class) ทั้งนี้การที่ข้อมูลไม่สมดุลจะส่งผลต่อการจำแนกประเภทข้อมูล โดยจะสามารถจำแนกประเภทข้อมูลของกลุ่มที่มีข้อมูลจำนวนมากได้อย่างแม่นยำ แต่ความถูกต้องแม่นยำในการจำแนกประเภทข้อมูลของกลุ่มที่มีจำนวนข้อมูลน้อยจะลดลง (Farquad & Bose, 2012) ซึ่งงานวิจัยนี้ให้ความสำคัญกับข้อมูลกลุ่มรองมากกว่าข้อมูลกลุ่มหลัก

ตัวอย่างข้อมูลไม่สมดุล เช่น ข้อมูลชุดหนึ่งมีจำนวนข้อมูลทั้งหมด 1,000 ข้อมูล โดยแบ่งออกเป็น 2 กลุ่ม ได้แก่ กลุ่ม A มีข้อมูล 900 ข้อมูล และกลุ่ม B มีข้อมูล 100 ข้อมูล ซึ่งจะเรียกข้อมูลกลุ่ม A ว่าข้อมูลกลุ่มหลัก และจะเรียกข้อมูลกลุ่ม B ว่าข้อมูลกลุ่มรอง เมื่อนำข้อมูลทั้งสองกลุ่มเข้าสู่ขั้นตอนการจำแนกข้อมูลพร้อมกันทั้งหมด จะทำให้ผลการแบ่งกลุ่มข้อมูลเกิดความผิดพลาด กล่าวคือ ข้อมูลที่อยู่ในกลุ่มรองถูกครอบงำหรือจะถูกจัดให้ไปอยู่ในกลุ่มหลักทั้งหมด จึงนำไปสู่ปัญหาที่เรียกว่า “ปัญหาการจำแนกข้อมูลผิดกลุ่ม (misclassification)” (เบญจภรณ์ จันทรกองกุล และคณะ, 2557)

ข้อมูลไม่สมดุลนั้นสามารถพบเห็นได้ทั่วไป ซึ่งสาเหตุของการเกิดความไม่สมดุลนั้นอาจจะมาจากหลายสาเหตุ เช่น ข้อมูลไม่สมดุลที่เกิดจากธรรมชาติของข้อมูลเองซึ่งสามารถพบเจอได้ในข้อมูลการวินิจฉัยทางด้านการแพทย์ที่มีข้อมูลผู้ป่วยด้วยโรคร้ายแรงน้อยกว่าข้อมูลของผู้ที่มีสุขภาพดีเป็นจำนวนมาก ข้อมูลของบัตรเครดิตที่มีข้อมูลลูกค้าปกติมากกว่าลูกค้าที่ผิดปกติ ข้อมูลการตรวจจับผู้บุกรุกของเครือข่ายข้อมูล หรือข้อมูลไม่สมดุลอาจจะเกิดจากข้อจำกัดในการจัดเก็บ เช่น ค่าใช้จ่ายที่สูงมาก อันตรายที่เกิดจากการรวบรวมข้อมูล เป็นต้น (ภาสพิชญ์ ชูใจ, 2557)

1.2 ระดับความไม่สมดุลของข้อมูล

การจะระบุว่าข้อมูลใดเป็นข้อมูลที่สมดุลหรือเป็นข้อมูลที่ไม่สมดุลนั้นจะพิจารณาจากระดับความไม่สมดุลของข้อมูล โดยสามารถคำนวณได้จากอัตราส่วนระหว่างจำนวนข้อมูลของกลุ่มหลักต่อจำนวนข้อมูลของกลุ่มรอง (Orriols-Puig & Bernadó-Mansilla, 2009; Villar et al., 2011) ดังสมการ

$$\text{Imbalance Ratio (IR)} = \frac{n_{majority}}{n_{minority}}$$

โดยที่ $n_{majority}$ คือ จำนวนข้อมูลของกลุ่มหลัก

$n_{minority}$ คือ จำนวนข้อมูลของกลุ่มรอง

ถ้าระดับความไม่สมดุลของข้อมูลมากกว่า 1 หมายความว่าจำนวนข้อมูลของกลุ่มหลักมีจำนวนมากกว่าจำนวนข้อมูลของกลุ่มรอง ถ้าระดับความไม่สมดุลของข้อมูลเท่ากับ 1 หมายความว่าจำนวนข้อมูลของกลุ่มหลักมีจำนวนเท่ากับจำนวนข้อมูลของกลุ่มรอง และถ้าระดับความไม่สมดุลของข้อมูลน้อยกว่า 1 หมายความว่าจำนวนข้อมูลของกลุ่มหลักมีจำนวนน้อยกว่าจำนวนข้อมูลของกลุ่มรอง

1.3 วิธีการปรับสมดุลข้อมูล

นักวิจัยให้ความสนใจกับปัญหาข้อมูลไม่สมดุลเป็นอย่างมาก นักวิจัยเหล่านั้นจึงได้ศึกษาค้นคว้าวิธีการและนำเสนอการแก้ปัญหานี้ (López et al., 2012) โดยแบ่งการแก้ปัญหาออกเป็น 3 ระดับ คือ 1) การแก้ปัญหาข้อมูลไม่สมดุลที่ระดับข้อมูล (data level solutions) 2) การแก้ปัญหาข้อมูลไม่สมดุลที่ระดับขั้นตอนวิธีการ (algorithmic level solutions) และ 3) การแก้ปัญหาข้อมูลไม่สมดุลด้วยการเรียนรู้แบบมีค่าใช้จ่าย (cost-sensitive training) (ภาสพิชญ์ ชูใจ, 2557) ซึ่งทั้งสามระดับนี้มีเป้าหมายเดียวกัน คือ เพื่อเพิ่มประสิทธิภาพและความแม่นยำในการจำแนกประเภทข้อมูลกลุ่มหลักและกลุ่มรอง

1) การแก้ปัญหาระดับข้อมูล (data level solutions) เป็นการแก้ปัญหาในขั้นตอนก่อนการประมวลผล (preprocessing stage) ซึ่งจะเกี่ยวข้องกับข้อมูลโดยตรง โดยจะปรับข้อมูลที่ไม่สมดุลให้กลายเป็นข้อมูลสมดุลด้วยเทคนิคการสุ่มเลือกข้อมูล (data sampling technique) ซึ่งเทคนิคการสุ่มเลือกข้อมูลที่ได้รับความนิยมจะแบ่งออกเป็น 3 กลุ่ม คือ

1.1) วิธีสุ่มเกิน (over sampling methods) เป็นวิธีการเพิ่มจำนวนข้อมูลกลุ่มรองให้มีจำนวนใกล้เคียงหรือเท่ากันกับจำนวนข้อมูลกลุ่มหลัก ด้วยการใช้กระบวนการสุ่มตัวอย่างอย่างง่ายในการแทนที่ข้อมูลกลุ่มรองจนกระทั่งมีจำนวนข้อมูลเท่ากับจำนวนข้อมูลกลุ่มหลัก โดยข้อมูลใหม่ที่สมดุลแล้วจะเป็นข้อมูลที่มีองค์ประกอบตรงกันกับข้อมูลเดิมที่มีอยู่ (Bach et al., 2019; Lunardon et al., 2014; กิตติพงศ์ ชมบุญ, 2558; กีระชาติ สุขสุทธิ์, 2559; พัชรียา ทองพูล และคณะ, 2562; ภาสพิชญ์ ชูใจ, 2557) เช่น synthetic minority oversampling technique (SMOTE) (Chawla et al., 2002) borderline-SMOTE (Han et al., 2005) เป็นต้น โดยในงานวิจัยนี้จะใช้วิธี random oversampling ในการปรับสมดุลข้อมูล

1.2) วิธีสุ่มลด (under sampling methods) เป็นวิธีการลดจำนวนข้อมูลกลุ่มหลักให้มีจำนวนใกล้เคียงหรือเท่ากับจำนวนข้อมูลกลุ่มรอง โดยการลดจำนวนข้อมูลกลุ่มหลักให้น้อยลงเดิม (Bach et al., 2019; กิตติพงศ์ ชมบุญ, 2558; กีระชาติ สุขสุทธิ์, 2559; พัชรียา ทองพูล และคณะ, 2562; ภาสพิชญ์ ชูใจ, 2557) เช่น Wilson’s edited nearest neighbor (ENN) และ the one-sided selection (OSS) เป็นต้น

1.3) วิธีผสมผสาน (hybrid methods) เป็นวิธีการที่นำเทคนิควิธีสุ่มเกินและวิธีสุ่มลดมาทำงานร่วมกัน โดยข้อมูลในกลุ่มรองจะถูกเพิ่มจำนวนด้วยการสุ่มเกินแล้วแทนที่ และข้อมูลกลุ่มหลักจะถูกสุ่มลดโดยไม่มีการเพิ่มจำนวนแทนที่ ซึ่งในการสุ่มเกินของข้อมูลกลุ่มรองนั้นจะต้องให้ได้ตามขนาดของข้อมูลและความน่าจะเป็นที่กำหนด จึงส่งผลให้ไม่เกิดข้อมูลซ้ำจำนวนมากในข้อมูลกลุ่มรองและไม่ได้ลดจำนวนข้อมูลกลุ่มหลักลงมากนัก แต่ยังคงคุณสมบัติของข้อมูลกลุ่มรองและข้อมูลกลุ่มหลักไว้ และให้ผลลัพธ์ที่ได้มีจำนวนข้อมูลกลุ่มหลักเท่ากับหรือใกล้เคียงกับจำนวนข้อมูลกลุ่มรอง

(Bach et al., 2019; Lunardon et al., 2014; กิตติพงศ์ ชมบุญ, 2558; กีระชาติ สุขสุทธิ์, 2559; พัชรียา ทองพูล และคณะ, 2562; ภาสพิชญ์ ชูใจ, 2557) เช่น การนำเทคนิค SMOTE มาใช้ร่วมกับเทคนิค ENN กลายเป็นเทคนิค SMOTE-ENN (Batista et al., 2004) หรือนำเทคนิค SMOTE มาใช้ร่วมกับเทคนิค TomekLinks กลายเป็น SMOTE+TomekLinks (Batista et al., 2004) เป็นต้น โดยในงานวิจัยนี้จะใช้วิธีผสมผสานระหว่างรูปแบบสุ่มเกินและสุ่มลดซึ่งเป็นการนำวิธี random oversampling และ random undersampling มาใช้ร่วมกันในการปรับสมดุลข้อมูล

เมื่อนักวิจัยปรับข้อมูลให้มีความสมดุลด้วยเทคนิคการสุ่มเลือกข้อมูลแล้วจะนำข้อมูลสมดุลที่ได้ไปวิเคราะห์ร่วมกับเทคนิคอื่น ๆ เช่น งานวิจัยของ Dubey et al. (2014) ซึ่งศึกษาเกี่ยวกับการวิเคราะห์ข้อมูลไม่สมดุลด้วยเทคนิคการสุ่มข้อมูลกับชุดข้อมูลโรคอัลไซเมอร์ (Alzheimer's disease) โดยใช้เทคนิคการสุ่มเลือกข้อมูลด้วยวิธีสุ่มเกิน วิธีสุ่มลด และวิธีผสมผสาน ทำงานร่วมกับเทคนิคการเรียนรู้ร่วมกันและการคัดเลือกคุณลักษณะ (feature selection) และงานวิจัยของ Qian et al. (2014) ที่ศึกษาเกี่ยวกับกระบวนการสุ่มเพื่อจำแนกข้อมูลไม่สมดุล โดยใช้วิธีสุ่มเกินในการเพิ่มจำนวนข้อมูลกลุ่มรอง และใช้วิธีสุ่มลดในการลดข้อมูลที่อยู่ในกลุ่มหลัก และอัตราการสุ่มเลือกจะกำหนดจากอัตราส่วนของจำนวนข้อมูลกลุ่มหลักต่อจำนวนข้อมูลกลุ่มรอง จากนั้นจำแนกประเภทข้อมูลที่สมดุลแล้วด้วยวิธีการเรียนรู้ร่วมกันแบบแบ็กกิง เป็นต้น

2) การแก้ปัญหาระดับขั้นตอนวิธีการ (algorithmic level solutions) เป็นการแก้ปัญหาโดยการปรับการเรียนรู้ของอัลกอริทึมมาตรฐานสำหรับการจำแนกประเภทข้อมูลที่มีอยู่เดิมให้สามารถเรียนรู้ข้อมูลไม่สมดุลโดยให้มีการเอนเอียงไปทางข้อมูลกลุ่มรอง

3) การแก้ปัญหาด้วยวิธีการเรียนรู้แบบมีค่าใช้จ่าย (cost-sensitive training) เป็นวิธีการแก้ปัญหาที่นำทั้งการแก้ปัญหาที่ระดับข้อมูล และระดับอัลกอริทึมมาทำงานร่วมกัน โดยที่ระดับข้อมูลจะทำการเพิ่มค่าใช้จ่าย (cost) ที่พิเศษสำหรับกรณีที่มีการจำแนกประเภทผิดพลาด และที่ระดับอัลกอริทึมจะทำการปรับการเรียนรู้ของอัลกอริทึมมาตรฐานให้สอดคล้องกับการจำแนกประเภทข้อมูลผิดพลาด (ภาสพิชญ์ ชูใจ, 2557)

จากการศึกษางานวิจัยที่เกี่ยวข้องกับประสิทธิภาพของวิธีการปรับสมดุลข้อมูลและการจำแนกข้อมูลจากข้อมูลในสภาพจริง พบว่า จากงานวิจัยของ Bach et al. (2019) หรือ Cateni et al. (2014) หรือ Gameng et al. (2019) หรือ Hussain et al. (2018) หรือ Mousa & Maghari (2017) หรือ Prasad et al. (2016) หรือ Zhu et al. (2021) หรือ Yan Zhu et al. (2020) หรือ พัชรียา ทองพูล และคณะ (2562) หรือสุกัญญา ทารส (2562) มีขนาดตัวอย่างอยู่ในช่วง 100 ถึง 800 ตัวอย่าง แต่ขนาดตัวอย่างส่วนใหญ่อยู่ในช่วง 300 ถึง 500 ใช้เทคนิคการจำแนกข้อมูลด้วยวิธีต้นไม้ตัดสินใจ วิธีแรนดอมฟอร์เรส วิธีการถดถอยโลจิสติก วิธีซัพพอร์ตเวกเตอร์แมชชีน วิธีเคเนียร์เรสเนเบอร์ วิธีนาอีฟเบย์ วิธีโครงข่ายประสาทเทียม แต่เทคนิคการจำแนกข้อมูลที่งานวิจัยส่วนใหญ่นิยม

ใช้และมีประสิทธิภาพสูง ได้แก่ วิธีแรนดอมฟอร์เรส วิธีการถดถอยโลจิสติก วิธีซัพพอร์ตเวกเตอร์แมชชีน วิธีเคเนียร์เรสเนเบอร์ โดยถ้าข้อมูลมีตัวแปรอิสระที่เป็นตัวแปรต่อเนื่องเกือบทุกตัวแปรจะเลือกวิธีการปรับสมดุลข้อมูลได้หลากหลายวิธีมากกว่าการที่ข้อมูลมีตัวแปรอิสระที่เป็นตัวแปรจัดประเภทมากกว่าตัวแปรอิสระที่เป็นตัวแปรต่อเนื่อง อัตราออดของตัวแปรอิสระในข้อมูลส่วนใหญ่จะอยู่ในช่วง 0.5 ถึง 3 จำนวนตัวแปรอิสระที่เป็นตัวแปรจัดประเภทกับจำนวนตัวแปรอิสระที่เป็นตัวแปรต่อเนื่องมีผลต่อการเลือกวิธีการปรับสมดุลข้อมูล และร้อยละของจำนวนข้อมูลกลุ่มหลักมีค่าตั้งแต่ 60 ถึง 99 และร้อยละของจำนวนข้อมูลกลุ่มรอง มีค่าตั้งแต่ 1 ถึง 40 ซึ่งอัตราส่วนของจำนวนข้อมูลกลุ่มหลักต่อจำนวนข้อมูลกลุ่มรองส่วนใหญ่อยู่ในช่วง 1.5 ถึง 3 ดังนั้นการเลือกใช้วิธีการปรับสมดุลข้อมูลจะพิจารณาลักษณะของข้อมูลในด้านขนาดตัวอย่าง ด้านเทคนิคการจำแนกข้อมูล ด้านจำนวนตัวแปรระหว่างกลุ่มตัวแปรจัดประเภทต่อกลุ่มตัวแปรต่อเนื่อง ด้านอัตราออด และด้านร้อยละของจำนวนข้อมูลระหว่างข้อมูลกลุ่มหลักต่อข้อมูลกลุ่มรอง ซึ่งวิธีการปรับสมดุลข้อมูลที่เลือกใช้ ได้แก่ วิธีสุ่มเกิน (oversampling) วิธีสุ่มลด (undersampling) วิธีผสมผสาน (hybrid) และไม่ปรับสมดุลข้อมูล แต่ไม่นิยมที่จะไม่ปรับสมดุลข้อมูลเพราะทำให้ค่าประสิทธิภาพในการจำแนกข้อมูลต่ำ และเกิดปัญหาในการจำแนกข้อมูลผิดกลุ่ม งานวิจัยส่วนใหญ่จึงนิยมเลือกใช้วิธีสุ่มเกิน (oversampling) และวิธีผสมผสาน (hybrid) มากกว่าวิธีสุ่มลด (undersampling) ก่อนนำไปวิเคราะห์จำแนกกลุ่มข้อมูล เนื่องจากวิธีสุ่มลดจะทำให้สูญเสียข้อมูลกลุ่มหลักที่เก็บรวบรวมมาได้

ตอนที่ 2 เทคนิคการจำแนกข้อมูล

2.1 แรนดอมฟอร์เรส (random forest)

แรนดอมฟอรเรส (random forest) เป็นการเรียนรู้แบบมีผู้สอน เกิดจากการรวมกลุ่มกันของโครงสร้างต้นไม้ (Hartshorn, 2016; Leo, 2001) การเพิ่มจำนวนต้นไม้ในป่าเกิดจากการที่ค่าความคลาดเคลื่อนโดยรวมของป่าไม้ถูกเปลี่ยนเป็นค่าลิมิต โดยความมั่นคงของต้นไม้แต่ละต้นและความสัมพันธ์กันระหว่างต้นไม้แต่ละต้นจะส่งผลต่อค่าความคลาดเคลื่อนโดยรวม และลดค่าความผิดพลาด ด้วยหลักการใช้วิธีการสุ่มเลือกคุณสมบัติในการแบ่งแยกโหนด ซึ่งสามารถประมวลผลเพื่อตัดสินใจได้จากการสร้างแบบจำลองที่ใช้ต้นไม้หลาย ๆ ต้น

ขั้นตอนการทำงานของแรนดอมฟอรเรสจะจำแนกต้นไม้หลาย ๆ ต้น ซึ่งในต้นไม้แต่ละต้นจะมีการแบ่งเป็นคลาส ทั้งนี้ต้นไม้แต่ละต้นจะถูกสร้างจากตัวอย่างที่แตกต่างกันจากกระบวนการของต้นไม้ตัดสินใจซึ่งจะถูกสร้างขึ้นจนกลายเป็นป่า (forest) จนกระทั่งวิเคราะห์การตัดสินใจจากต้นไม้แต่ละต้นที่อยู่ในป่า ซึ่งแรนดอมฟอร์เรสเป็นต้นไม้ตัดสินใจที่มีลักษณะแบบไม่ตัดแต่งกิ่ง (unpruned) หรือต้นไม้ถดถอยถูกสร้างจากการนำข้อมูลฝึกสอนไปสุ่มเลือกตัวอย่างข้อมูลและคุณลักษณะข้อมูล

แล้วนำมาสร้างเป็นต้นไม้ตัดสินใจ สำหรับชุดทดสอบจะเป็นตัวอย่างที่ไม่ถูกเลือกจากชุดฝึกสอนจะนำมาใช้ในการทดสอบต้นไม้ตัดสินใจ เรียกข้อมูลชุดนี้ว่า out-of-bag (OOB) และเรียกวิธีการนี้ว่า แบ็กกิง ผลลัพธ์ที่ได้จากต้นไม้ตัดสินใจแต่ละต้นจะถูกนำมาคิดเป็นผลการโหวตที่มากที่สุด (กาญจน์ณ ศรีธะ และคณะ, 2561; นิเวศ จิระวิชิตชัย, 2563)

งานวิจัย เรื่อง A Modified Adaptive Synthetic SMOTE Approach in Graduation Success Rate Classification ทำกับชุดข้อมูลการรับสมัครเข้าเรียนวิทยาลัย Davao del Norte State College ของนักเรียนจำนวน 897 คน มีทั้งหมด 14 คุณลักษณะ ใช้เทคนิค SMOTE, ADASYN และ Modified ADASYN ในการปรับข้อมูลให้สมดุล แล้วจำแนกข้อมูลด้วยการถดถอยโลจิสติกและแรนดอมฟอรเรส พบว่าการจำแนกข้อมูลด้วยแรนดอมฟอรเรสมีประสิทธิภาพดีที่สุด โดยมีค่า accuracy, precision, recall และ F1 score อยู่ในช่วง 80% – 100% (Gameng et al., 2019)

งานวิจัย เรื่อง Inspector: a lysine succinylation predictor based on edited nearest-neighbor undersampling and adaptive synthetic oversampling ทำกับชุดข้อมูลโปรตีน Lysine succinylation จำนวน 2,322 ชุด ใช้วิธีผสมผสานในการปรับข้อมูลให้สมดุล แล้วจำแนกข้อมูลด้วยแรนดอมฟอรเรส เคเนียร์เรสเนเบอร์ ซัพพอร์ตเวกเตอร์แมชชีน นาอีฟเบย์ และการวิเคราะห์แบบจำแนกประเภท (discriminant analysis) พบว่าการจำแนกข้อมูลด้วยแรนดอมฟอรเรสมีประสิทธิภาพดีที่สุด โดยมีค่าความถูกต้อง 90.4% (Yan Zhu et al., 2020)

งานวิจัย เรื่อง Rapid identification of high-quality marine shale gas reservoirs based on the oversampling method and random forest algorithm ที่พิจารณาวิธีที่มีประสิทธิภาพสูงที่สุดในการระบุแหล่งกักเก็บก๊าซจากชั้นหินดินดานคุณภาพสูง โดยใช้การสุ่มเกินร่วมกับวิธีแรนดอมฟอร์เรส วิธีการเคเนียร์เรสเนเบอร์ (K-Nearest Neighbors) และวิธีซัพพอร์ตเวคเตอร์แมชชีน พบว่าการสุ่มเกินร่วมกับแรนดอมฟอร์เรสมีประสิทธิภาพสูงที่สุดสามารถเพิ่มค่าความถูกต้องในการทำนายจากร้อยละ 44 เป็นร้อยละ 78 (Zhu et al., 2021)

2.2 การถดถอยโลจิสติก (logistic regression)

การถดถอยโลจิสติก (logistic regression) (Kaiyawan, 2012; Kittithanusorn & Sa-ing, 2021; กัลยา วานิชย์บัญชา, 2015; กาญจน์เขจร ชูชีพ, 2018) เป็นเทคนิคการวิเคราะห์ที่มีวัตถุประสงค์เพื่อทำนายโอกาสในการเกิดเหตุการ์ที่สนใจ โดยตัวแปรอิสระเป็นได้ทั้งตัวแปรต่อเนื่อง (continuous variable) และตัวแปรจัดประเภท (categorical variable) ซึ่งการวิเคราะห์การถดถอยโลจิสติกแบ่งออกเป็น 2 ประเภทตามจำนวนกลุ่มของตัวแปรตาม ได้แก่ (1) การวิเคราะห์

การถดถอยโลจิสติกทวิ (binomial logistic regression) จะมีตัวแปรตามเป็นตัวแปรแบบทวินาม (dichotomous variable) ซึ่งเป็นตัวแปรที่มีได้ 2 ค่า ได้แก่ “ใช่” กับ “ไม่ใช่” หรือ “เป็น” กับ “ไม่เป็น” หรือ “กลุ่มที่ปรากฏเหตุการณ์ที่สนใจ” กับ “กลุ่มที่ไม่ปรากฏเหตุการณ์ที่สนใจ” เป็นต้น และ (2) การวิเคราะห์การถดถอยโลจิสติกพหุ (multinomial logistic regression) จะมีตัวแปรตามที่มีค่ามากกว่า 2 ค่า ได้แก่ ระดับการศึกษา แบ่งเป็น ประถมศึกษา มัธยมศึกษา และบัณฑิตศึกษา

การวิเคราะห์การถดถอยโลจิสติกทวิเป็นการประมาณค่าความน่าจะเป็นในการเกิดเหตุการณ์โดยมีฟังก์ชันโลจิสติกเป็นตัวแบบ

ในกรณีที่มีตัวแปรอิสระ 1 ตัว จะแสดงความน่าจะเป็นของเหตุการณ์ได้ดังสมการต่อไปนี้

$$P(y)=\frac{e^{(\beta_0+\beta_1 X)}}{1+e^{(\beta_0+\beta_1 X)}} \text{ หรือ } P(y)=\frac{1}{1+e^{-(\beta_0+\beta_1 X)}}$$

เมื่อ P(y) คือ ความน่าจะเป็นของการเกิดเหตุการณ์ y

β_0 คือ ค่าคงที่

β_1 คือ ค่าสัมประสิทธิ์ของตัวแปรอิสระ

X คือ ตัวแปรอิสระ

e คือ ลอการิทึมธรรมชาติ (มีค่าประมาณ 2.71828)

ในกรณีที่มีตัวแปรอิสระหลายตัว (n) จะแสดงความน่าจะเป็นของเหตุการณ์ได้ดังสมการต่อไปนี้

$$P(y)=\frac{e^{Z}}{1+e^{Z}} \text{ หรือ } P(y)=\frac{1}{1+e^{-Z}}$$

เมื่อ Z คือ linear combination ของตัวแปรตาม

$$Z=\beta_0+\beta_1 X_1+\beta_2 X_2+\cdots+\beta_n X_n$$

และความน่าจะเป็นของการไม่เกิดเหตุการณ์ คือ

$$Q_y=1-P_y$$

เมื่อ P_y คือ ความน่าจะเป็นของการเกิดเหตุการณ์ที่สนใจ

Q_y คือ ความน่าจะเป็นของการไม่เกิดเหตุการณ์ที่สนใจ

ความน่าจะเป็นจะมีค่าตั้งแต่ 0 ถึง 1 แต่ความสัมพันธ์ระหว่างตัวแปรอิสระและตัวแปรตามของการวิเคราะห์การถดถอยโลจิสติกไม่ใช่รูปแบบเชิงเส้น จึงมีการปรับความสัมพันธ์ให้อยู่ในรูปฟังก์ชันเชิงเส้น ในรูปแบบของ odds หรือ odds ratio ซึ่ง odds แทนความน่าจะเป็น โดย odds หรือ odds ratio คือ อัตราส่วนระหว่างโอกาสที่จะเกิดเหตุการณ์ที่สนใจกับโอกาสที่จะไม่เกิด

เหตุการณ์ที่สนใจ และหมายถึง โอกาสที่จะเกิดเหตุการณ์ที่สนใจเป็นกี่เท่าของโอกาสที่จะไม่เกิดเหตุการณ์ที่สนใจ เช่น การมีบุตร 1 ครั้ง มีความน่าจะเป็นที่จะได้บุตรสาว เท่ากับ 0.5 จะได้ว่า odds มีค่าเท่ากับ 0.5/0.5 เท่ากับ 1 เท่า หมายถึงโอกาสที่จะเกิดเหตุการณ์ที่สนใจเท่ากับโอกาสที่จะไม่เกิดเหตุการณ์ที่สนใจ ดังนั้นถ้า odds มีค่าน้อยกว่า 1 จะหมายถึงโอกาสที่จะเกิดเหตุการณ์ที่สนใจน้อยกว่าโอกาสที่จะไม่เกิดเหตุการณ์ที่สนใจ และถ้า odds มีค่ามากกว่า 1 จะหมายถึงโอกาสที่จะเกิดเหตุการณ์ที่สนใจมากกว่าโอกาสที่จะไม่เกิดเหตุการณ์ที่สนใจ ในการวิเคราะห์การถดถอยโลจิสติกแสดงค่า odds ได้ดังสมการ

$$\text{odds} = \frac{P_y}{Q_y} = \frac{P_y}{1-P_y}$$

เนื่องจากสมการดังกล่าวไม่ได้อยู่ในรูปสมการเชิงเส้น เมื่อแปลงให้เป็นฟังก์ชันเชิงเส้นจะได้ log ของ OR เรียกว่า logit ซึ่งมีสมการดังนี้

$$\log(\text{odds}) = \ln\left(\frac{P_y}{Q_y}\right)$$

สามารถแสดงในรูปสมการถดถอยเชิงเส้นได้ดังสมการ

$$\log(\text{odds}) = \text{logit} = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_n X_n$$

การสร้างสมการถดถอยโลจิสติกจะใช้วิธีการประมาณความน่าจะเป็นสูงสุด (maximum likelihood estimation) ซึ่งใช้กระบวนการวิเคราะห์แบบเวียนซ้ำ เริ่มจากการประมาณค่าสัมประสิทธิ์ของตัวแปรเพื่อให้ได้สมการตั้งต้น จากนั้นใช้สมการเพื่อทำนายค่าแล้วนำมาคำนวณซ้ำเพื่อหาค่าสัมประสิทธิ์ใหม่ที่ทำให้ได้ค่าความน่าจะเป็นสูงที่สุด เพื่อให้สามารถทำนายค่าตัวแปรตามได้ใกล้เคียงข้อมูลจริงมากที่สุด แสดงในรูปสมการประมาณค่าดังนี้

$$\hat{P}_y = \frac{e^{(\beta_0+\beta_1+\beta_2+\cdots+\beta_n)}}{1-e^{(\beta_0+\beta_1+\beta_2+\cdots+\beta_n)}}$$

ถ้ากระทำ antilog จะได้ OR ratio ดังสมการ

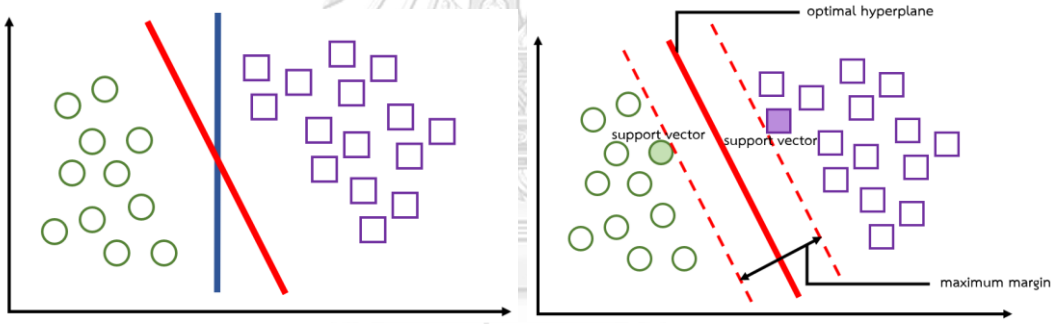
$$\frac{\hat{P}_y}{1-\hat{P}_y} = e^{(\beta_0+\beta_1+\beta_2+\cdots+\beta_n)}$$

งานวิจัย เรื่อง A Novel Imbalanced Data Classification Approach Based on Logistic Regression and Fisher Discriminant เกี่ยวกับการจำแนกลูกค้าด้วยการวิเคราะห์การถดถอยโลจิสติก การวิเคราะห์ด้วยฟิชเชอร์ดิสคริมิแนนต์ และซัพพอร์ตเวกเตอร์แมชชีน พบว่าการวิเคราะห์การถดถอยโลจิสติกและการวิเคราะห์ด้วยฟิชเชอร์ดิสคริมิแนนต์มีประสิทธิภาพดีกว่าและซัพพอร์ตเวกเตอร์แมชชีน (Shi et al., 2015)

งานวิจัย เรื่อง การจำแนกประเภทข่าวด้วยวิธีการเรียนรู้ด้วยเครื่อง โดยจำแนกด้วยมัลติโนเมียวนาอีฟเบย์ คอมพลีเมนต์นาอีฟเบย์ การวิเคราะห์การถดถอยโลจิสติก ลิเนียร์เอสวีซี และแรนดอมฟอร์เรส พบว่า การวิเคราะห์การถดถอยโลจิสติก มีประสิทธิภาพสูงสุดในการจำแนกประเภทข่าว โดยมีค่าความถูกต้อง 80.69 (Kittithanusorn & Sa-ing, 2021)

2.3 ซัพพอร์ตเวกเตอร์แมชชีน (support vector machine)

ซัพพอร์ตเวกเตอร์แมชชีน (support vector machine) (Cortes & Vapnik, 1995; ไกรศักดิ์ เกสร) เป็นวิธีหนึ่งในกลุ่มการเรียนรู้ของเครื่องที่ต้องมีผู้สอน (supervised learning) ที่มีความสามารถในการจำแนกข้อมูล ใช้หลักการสร้างไฮเปอร์เพลน (hyperplane) ซึ่งเป็นการสร้างเส้นตรงเพื่อจำแนกข้อมูล 2 กลุ่ม เส้นตรงดังกล่าวจะอยู่ตรงกลางระหว่างข้อมูล 2 กลุ่มและพยายามแบ่งข้อมูล 2 กลุ่มออกจากกันให้ดีที่สุด โดยไฮเปอร์เพลนที่จำแนกข้อมูลได้ดีที่สุดจะเรียกว่า optimal hyperplane และไฮเปอร์เพลนจะมีเส้นขอบทั้ง 2 ข้างเรียกว่า มาร์จิน (margin) โดยจะมีความกว้างที่มากที่สุดที่จะเป็นไปได้ (maximum margin)



(ก) (ข)

รูปที่ 2.1 แสดงตัวอย่างการแบ่งข้อมูล

จากรูปที่ 2.1 ถ้าต้องการจำแนกข้อมูลออกเป็น 2 กลุ่ม โดยใช้ไฮเปอร์เพลนที่เป็นเส้นตรงจะพบว่ามีเส้นตรงหลายเส้นที่สามารถจำแนกข้อมูลออกเป็น 2 กลุ่มอย่างชัดเจน แต่ซัพพอร์ตเวกเตอร์แมชชีนจะพิจารณาถึงมาร์จินที่กว้างที่สุดที่เป็นไปได้ และเรียกข้อมูลที่อยู่บนมาร์จินว่า support vector

งานวิจัย เรื่อง เทคนิคการจำแนกประเภทข้อมูลส่วนน้อยบนข้อมูลไม่สมดุลด้วยวิธีการแบ่งข้อมูล เกี่ยวกับการเพิ่มประสิทธิภาพในการจำแนกข้อมูลส่วนน้อยที่อยู่ในชุดข้อมูลไม่สมดุล โดยในงานวิจัยนี้แบ่งข้อมูลออกเป็น 2 ส่วน คือ ส่วนที่มีการซ้อนทับกัน และส่วนที่ไม่มีการซ้อนทับกัน พบว่าวิธีการจำแนกข้อมูลที่มีประสิทธิภาพสูงสุด คือ ซัพพอร์ตเวกเตอร์แมชชีน (กิตติพงศ์ ชมบุญ, 2558)

2.4 วิธีเคเนียร์เรสเนเบอร์ (k-nearest neighbor)

วิธีเคเนียร์เรสเนเบอร์ (k-nearest neighbor) (ชนาธิป หมั่นเพียรสุข และสุพจน์ เฮงพระพรหม, ; พงศกร ธีรรัศมี; อัจฉรา แผ้วบาง และสายชล สินสมบูรณ์ทอง, 2020) เป็นวิธีการที่ได้รับความนิยมในการใช้จำแนกข้อมูล โดยการจำแนกข้อมูลด้วยวิธีเคเนียร์เรสเนเบอร์จะไม่มีการสร้างโมเดลเพื่อจำแนกประเภทข้อมูลเตรียมไว้ล่วงหน้า เมื่อมีข้อมูลใหม่ที่ต้องการจำแนกประเภทจะนำมาเปรียบเทียบความเหมือนกันหรือใกล้เคียงกันของคุณสมบัติกับข้อมูลเดิมที่มี จากนั้นก็จำแนกประเภทของข้อมูลใหม่ออกมาให้เป็นประเภทเดียวกันกับข้อมูลเดิมที่อยู่ใกล้เคียงกัน

ขั้นตอนการทำงานของวิธีเคเนียร์เรสเนเบอร์มีวิธีดำเนินการดังนี้

1. กำหนดค่า k และวิธีการวัดระยะห่างระหว่างข้อมูลใหม่ที่สนใจกับข้อมูลเดิม โดยวิธีที่นิยมคือระยะทางยูคลิเดียน (Euclidean distance)
2. คำนวณระยะทางระหว่างข้อมูลใหม่ที่ต้องการจำแนกและข้อมูลเดิม
3. จัดลำดับระยะทาง และเลือกค่าข้อมูลที่มีค่าระยะทางน้อยที่สุด k ตัว
4. พิจารณาการจำแนกกลุ่มของข้อมูลเดิมทั้ง k ตัว
5. กำหนดกลุ่มให้กับข้อมูลใหม่ โดยพิจารณาจากประเภทกลุ่มเป้าหมายของเพื่อนบ้านที่มีจำนวนข้อมูลเดิมมากที่สุด

หลักการของวิธีเคเนียร์เรสเนเบอร์ (Dubey et al., 2014; ณัฏฐินี ดีแท้, 2016) เมื่อกำหนดให้ชุดข้อมูลเรียนรู้ ดังสมการ

$$T = \{(x_i, y_i)\} \quad ; i = 1, 2, 3, \ldots, n$$

เมื่อ x_i คือ ตัวแปรอิสระ

y_i คือ ตัวแปรจัดประเภท

ทำให้ได้ว่าข้อมูลตัวแปรอิสระ x เขียนแทนได้ดังสมการ

$$<a_1(x), a_2(x), \ldots, a_m(x)>$$

เมื่อ $a_b(m)$ คือ ค่าของ b^{th} ของตัวแปรอิสระ x ; b = 1, 2, 3, ..., m

ฟังก์ชันที่นิยมใช้วัดระยะห่างคือฟังก์ชันระยะทางยูคลิเดียน ดังสมการ

$$d_{euclidean} = (u, v) = \sqrt{\sum_{i=1}^{n} (u_i - v_i)^2}$$

เมื่อ $u = \{u_1, u_2, u_3, \ldots, u_m\}$ และ $v = \{v_1, v_2, v_3, \ldots, v_m\}$ คือการบันทึกค่าคุณลักษณะของทั้งสองกลุ่ม ระยะห่างระหว่าง x_0 ซึ่งเป็นค่าสังเกตใหม่ และ x_i ถูกกำหนดด้วยสมการ

$$d(x_0, x_i) = \sqrt{\sum_{i=1}^{n} [a_b(x_0) - a_b(x_i)]^2}$$

โดยค่าสังเกตค่าใหม่ในชุดข้อมูลทดสอบจะจัดเข้ากลุ่มที่มีค่า $\hat{p}_q = (x'_\ell \mid x_i)$ สูงสุด ซึ่งถูกกำหนดโดย

$$\hat{p}_q = (x'_\ell \mid x_i) = \frac{1}{k}\sum_{i=\ell}^{k} \delta_{(q)(x'_\ell)}$$

เมื่อ $\sum_{i=\ell}^{k} \delta_{(q)(x'_\ell)}$ คือ จำนวน x_i ของกลุ่ม q ภายในบริเวณใกล้เคียง k

$\delta_{(q)(x'_\ell)}$ คือ ฟังก์ชันดิแรก (dirac function) ซึ่งถูกกำหนดโดย

$$\delta_{(q)(x'_\ell)} = \begin{cases} 1, & q = x'_\ell \\ 0, & \text{otherwise} \end{cases}$$

งานวิจัย เรื่อง การเปรียบเทียบประสิทธิภาพในการทำนายผลการปรับความไม่สมดุลของข้อมูลในการจำแนกด้วยเทคนิคการทำเหมืองข้อมูล ซึ่งมีข้อมูลที่ไม่สมดุลในการศึกษา 3 ชุด คือ ชุดข้อมูลการรับรู้ทางหูของเด็กที่มีภาวะน้ำคั่งในหูชั้นกลางหรือหูน้ำหนวก ชุดข้อมูลยอดคงเหลือในบัตรเครดิตของลูกค้า และชุดข้อมูลคุณภาพไวน์แดง วิธีการปรับข้อมูลที่ไม่สมดุล 4 วิธี คือ วิธีการสุ่มเกิน วิธีการสุ่มเกินโดยเทคนิค SMOTE วิธีการสุ่มลด และวิธีการสุ่มผสมผสาน โดยวิธีการจำแนก 4 วิธี คือ วิธีเพื่อนบ้านใกล้สุด k ตัว วิธีต้นไม้ตัดสินใจ วิธีโครงข่ายประสาทเทียม และวิธีซัพพอร์ตเวกเตอร์แมชชีน ว่าวิธีใดมีประสิทธิภาพในการจำแนกดีที่สุด พบว่าในชุดข้อมูลยอดคงเหลือในบัตรเครดิตของลูกค้า วิธีที่มีประสิทธิภาพสูงสุด คือ วิธีเพื่อนบ้านใกล้สุด k ตัว โดยการปรับความไม่สมดุลด้วยวิธีการสุ่มเกินเทคนิค SMOTE (พัชรียา ทองพูล และคณะ, 2562)

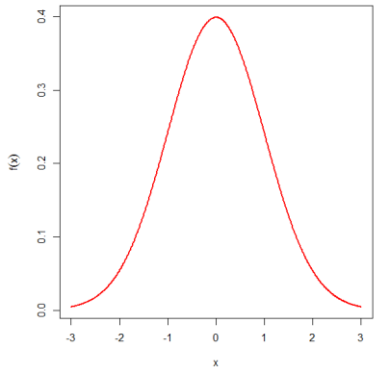
ตอนที่ 3 การแจกแจงความน่าจะเป็น

3.1 การแจกแจงปกติ (normal distribution)

การแจกแจงปกติ (normal distribution) เป็นการแจกแจงที่ตัวแปรมีข้อมูลเป็นข้อมูลต่อเนื่อง เมื่อทำการทดลองสุ่มที่เป็นอิสระต่อกันซ้ำ ๆ กันจำนวนหลาย ๆ ครั้ง แล้วหาค่าเฉลี่ย โดยข้อมูลส่วนใหญ่ที่ปรากฏในธรรมชาติมีการแจกแจงปกติ เช่น ส่วนสูง น้ำหนัก ผลผลิตทางการเกษตร ซึ่งการแจกแจงปกติขึ้นอยู่กับพารามิเตอร์ 2 ค่า คือ ค่าเฉลี่ย (μ) แสดงตำแหน่งศูนย์กลางของการแจกแจง และความแปรปรวน (σ^2) แสดงลักษณะการกระจายของการแจกแจง สามารถเขียนแทนด้วย $X \sim N(\mu, \sigma^2)$ (กิตติคุณ สุภาวณิชย์, 2564; นรารัตน์ เรืองชัยจตุพร, 2562b) โดยมีฟังก์ชันการแจกแจงความน่าจะเป็นดังนี้

$$f(x)=\frac{1}{\sqrt{2\pi}\sigma}\exp\left\{-\frac{1}{2\sigma^2}(x-\mu)^2\right\} \text{ โดยที่ } x\in R$$

และแสดงตัวอย่างฟังก์ชันความหนาแน่นความน่าจะเป็นของตัวแปรสุ่มแบบปกติได้ดังรูปที่ 2.2



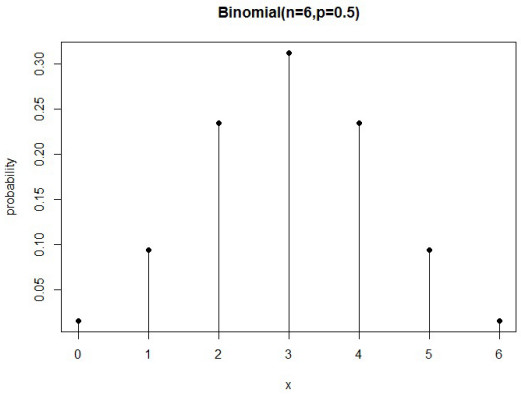
รูปที่ 2.2 ตัวอย่างฟังก์ชันความหนาแน่นความน่าจะเป็นของตัวแปรสุ่มแบบปกติ

3.2 การแจกแจงทวินาม (binomial distribution)

การแจกแจงทวินาม (binomial distribution) เป็นการแจกแจงที่ตัวแปรมีข้อมูลเป็นข้อมูลไม่ต่อเนื่อง หากทำการทดลองสุ่มที่เป็นอิสระต่อกัน n ครั้ง การทดลองแต่ละครั้งเกิดผลลัพธ์ได้เพียง 2 อย่าง คือ สำเร็จ (success) กับผิดหวัง (failure) เช่น การโยนเหรียญจะออกหัวหรือก้อยเสมอ ความน่าจะเป็นที่จะพบ 3 หัวพอดีในการโยนเหรียญซ้ำๆ กัน 10 ครั้ง ซึ่งการแจกแจงทวินามขึ้นอยู่กับพารามิเตอร์ 2 ค่า คือ จำนวนความสำเร็จที่ได้จากการทดลอง n ครั้ง (size) และความน่าจะเป็นที่จะสำเร็จ (p) สามารถเขียนแทนด้วย $Y\sim B(n,p)$ (จรรยา อ้นปันส์, 2564; นรารัตน์ เรืองชัยจตุพร, 2562a) โดยมีฟังก์ชันการแจกแจงความน่าจะเป็นดังนี้

$$p(y)=P(Y=y)=\binom{n}{y}p^y(1-p)^{n-y} \quad \text{โดยที่ } y = 0, 1, 2, 3, \ldots, n$$

และแสดงตัวอย่างการแจกแจงความน่าจะเป็นของจำนวนครั้งที่ขึ้นหัวจากการโยนเหรียญ 6 ครั้งได้ดังรูปที่ 2.3



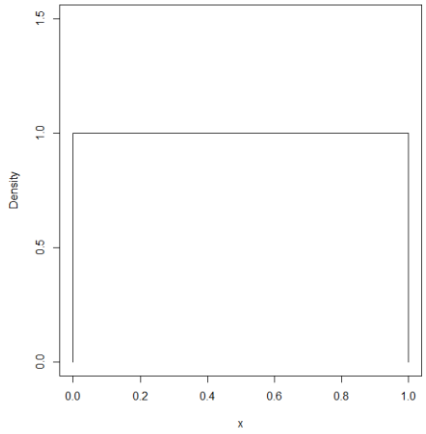
รูปที่ 2. 3 ตัวอย่างการแจกแจงความน่าจะเป็นของจำนวนครั้งที่ขึ้นหัวจากการโยนเหรียญ 6 ครั้ง
หมายเหตุ. จาก *บทที่ 5 การแจกแจงความน่าจะเป็นของตัวแปรสุ่มชนิดไม่ต่อเนื่อง* (หน้า 22), โดยจรรยา อ้นปันส์, 2564, มหาวิทยาลัยบูรพา.

3.3 การแจกแจงยูนิฟอร์ม (uniform distribution)

การแจกแจงยูนิฟอร์ม (uniform distribution) (RDocumentation; มหาวิทยาลัยราชภัฏบุรีรัมย์, 2563) เป็นการแจกแจงที่ตัวแปรมีข้อมูลเป็นข้อมูลต่อเนื่อง ความน่าจะเป็นของการเกิดค่าในช่วงที่ความกว้างเท่ากันใด ๆ มีค่าเท่ากัน ซึ่งค่าดังกล่าวอยู่ในช่วงของจำนวน 2 จำนวน คือ a และ b โดยที่ a น้อยกว่า b ค่าเฉลี่ย คือ $\frac{a+b}{2}$ และความแปรปรวน คือ $\frac{(a-b)^2}{12}$ เช่น ความน่าจะเป็นที่รถไฟฟ้าจะมาถึงสถานีในช่วงเวลา 7.00 ถึง 7.30 นาฬิกา ซึ่งการแจกแจงยูนิฟอร์มสามารถเขียนแทนด้วย $X \sim U(a,b)$ โดยมีฟังก์ชันการแจกแจงความน่าจะเป็นดังนี้

$$f(x) = \begin{cases} \frac{1}{b-a} & ; \ a \le x \le b \\ 0 & ; \quad \text{else} \end{cases}$$

และแสดงตัวอย่างฟังก์ชันความหนาแน่นความน่าจะเป็นของตัวแปรสุ่มแบบยูนิฟอร์มได้ดังรูปที่ 2.4



รูปที่ 2.4 ตัวอย่างฟังก์ชันความหนาแน่นความน่าจะเป็นของตัวแปรสุ่มแบบยูนิฟอร์ม

ตอนที่ 4 การวัดประสิทธิภาพการปรับสมดุลข้อมูล

เกณฑ์ที่ใช้ในการวัดประสิทธิภาพการปรับสมดุลข้อมูลได้มาจากการเปรียบเทียบกลุ่มที่ได้จากการทำนายกับกลุ่มที่แท้จริงของข้อมูล โดยแสดงผลที่เป็นไปได้จากการทำนายในลักษณะของเมทริกซ์ (กิตติพงศ์ ชมบุญ, 2558) ดังตารางที่ 2.1

ตารางที่ 2.1 แสดง confusion matrix ของค่าทำนาย (prediction) เปรียบเทียบกับค่าจริง (actual)

	Actually Positive	Actually Negative
Predicted Positive	True Positive (TP)	False Positive (FP)
Predicted Negative	False Negative (FN)	True Negative (TN)

โดย True Positive (TP) คือ จำนวนข้อมูลที่ทำนายว่าจริง และมีค่าเป็นจริง
False Positive (FP) คือ จำนวนข้อมูลที่ทำนายว่าจริง และมีค่าเป็นไม่จริง
False Negative (FN) คือ จำนวนข้อมูลที่ทำนายว่าไม่จริง และมีค่าเป็นจริง
True Negative (TN) คือ จำนวนข้อมูลที่ทำนายว่าไม่จริง และมีค่าเป็นไม่จริง

การวัดประสิทธิภาพการปรับสมดุลข้อมูลจะพิจารณาจากค่าความถูกต้อง ค่าความไว และค่าความจำเพาะ โดยค่าดังกล่าวจะคำนวณได้จากสมการ ต่อไปนี้

ค่าความถูกต้อง (accuracy) คือ การแสดงการวัดที่ได้มีความถูกต้องในรูปอัตราส่วน (สุรวัชร ศรีเปารยะ และสายชล สินสมบูรณ์ทอง, 2560)

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

ค่าความไว (sensitivity) คือ สัดส่วนของผลบวกที่เป็นจริงสำหรับภาวะนั้น ๆ (กีระชาติ สุขสุทธิ์, 2559)

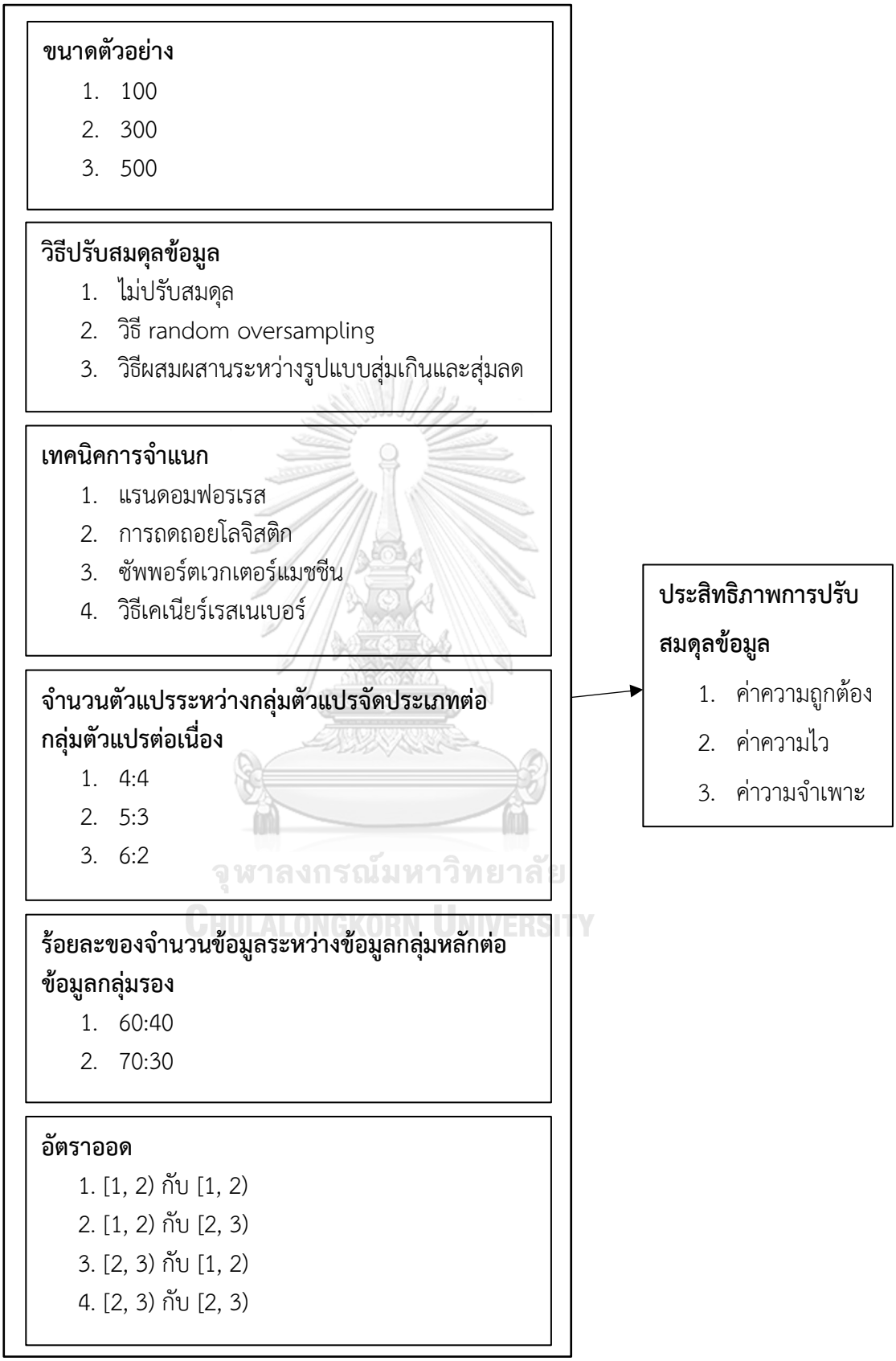
$$\text{sensitivity} = \frac{TP}{TP + FN}$$

ค่าความจำเพาะ (specificity) คือ สัดส่วนของผลลบที่เป็นจริงสำหรับภาวะนั้น ๆ (กีระชาติ สุขสุทธิ์, 2559)

$$\text{specificity} = \frac{TN}{TN + FP}$$

ตอนที่ 5 กรอบแนวคิดการวิจัย

ในงานวิจัยครั้งนี้จะเปรียบเทียบประสิทธิภาพของแต่ละวิธีการปรับสมดุลข้อมูลภายใต้สถานการณ์จำลองที่แตกต่างกัน รายละเอียดดังภาพที่ 2.2



รูปที่ 2.5 กรอบแนวคิดการวิจัย

บทที่ 3
วิธีดำเนินการวิจัย

การวิจัยนี้มีวัตถุประสงค์เพื่อศึกษาปฏิสัมพันธ์ของวิธีการปรับสมดุลข้อมูลกับเงื่อนไขขนาดตัวอย่าง เทคนิคการจำแนกข้อมูล จำนวนตัวแปรระหว่างกลุ่มตัวแปรจัดประเภทต่อกลุ่มตัวแปรต่อเนื่อง อัตราออด และร้อยละของจำนวนข้อมูลระหว่างข้อมูลกลุ่มหลักต่อข้อมูลกลุ่มรองที่มีต่อประสิทธิภาพของการจำแนกกลุ่ม โดยใช้การจำลองข้อมูลด้วยวิธีการมอนติคาร์โล (Monte Carlo simulation) ใช้การเขียนโปรแกรม R เวอร์ชั่น 4.2.3 ในการวิเคราะห์และประมวลผลข้อมูล กำหนดจำนวนตัวแปรอิสระ 8 ค่า ซึ่งตัวแปรแต่ละค่ามีการแจกแจงแบบปกติ โดยกำหนดจากการศึกษางานวิจัยที่เก็บข้อมูลเกี่ยวกับการศึกษาภายใต้สถานการณ์จริง พบว่า จำนวนตัวแปรอิสระส่วนใหญ่อยู่ในช่วง 7 ถึง 10 ตัวแปร จึงกำหนดเป็น 8 ตัวแปร เนื่องจากข้อจำกัดในการกำหนดจำนวนตัวแปรระหว่างกลุ่มตัวแปรจัดประเภทต่อกลุ่มตัวแปรต่อเนื่อง โดยทำการจำลองข้อมูลภายใต้เงื่อนไขที่แตกต่างกัน 6 เงื่อนไข ประกอบด้วย (1) วิธีปรับสมดุลข้อมูล 3 วิธี ได้แก่ ไม่ปรับสมดุล วิธี random oversampling และผสมผสานระหว่างรูปแบบสุ่มเกินและสุ่มลด (2) เทคนิคการจำแนก 4 เทคนิค (3) ขนาดตัวอย่าง 3 ขนาด (4) ตัวแปรอิสระ มีจำนวนตัวแปรระหว่างกลุ่มตัวแปรจัดประเภทต่อกลุ่มตัวแปรต่อเนื่อง 3 กรณี และอัตราออดของตัวแปรอิสระ 4 กรณี และ (5) ร้อยละของจำนวนข้อมูลระหว่างข้อมูลกลุ่มหลักต่อข้อมูลกลุ่มรอง 2 กรณี และจำลองข้อมูลตัวแปรตามจากการถดถอยโลจิสติก ดังนั้นการจำลองข้อมูลจากเงื่อนไขที่พิจารณาพบว่ามีทั้งสิ้น 864 เงื่อนไข (3x4x3x3x4x2) ทำซ้ำ 500 รอบ โดยจะนำเสนอเป็น 3 ตอน ได้แก่ 1) ข้อตกลงเบื้องต้นและเงื่อนไขที่ใช้ในการศึกษา 2) การเปรียบเทียบประสิทธิภาพของแต่ละวิธีการปรับสมดุลข้อมูล และ 3) ขั้นตอนการดำเนินงานวิจัย

ตอนที่ 1 ข้อตกลงเบื้องต้นและเงื่อนไขที่ใช้ในการศึกษา

1. วิธีปรับสมดุลข้อมูล

ในการศึกษาครั้งนี้จะทำการศึกษาแก้ปัญหาสมดุลข้อมูลที่ระดับข้อมูล ซึ่งจะใช้เทคนิคการสุ่มเลือกข้อมูลในการปรับสมดุลข้อมูล (López et al., 2012) ซึ่งเทคนิคการสุ่มเลือกข้อมูลที่นำมาใช้ประกอบไปด้วย 3 วิธี ดังนี้

1) การไม่ปรับสมดุล เพื่อใช้เป็นกรณีเบื้องต้นในการเปรียบเทียบประสิทธิภาพการแก้ปัญหาสมดุลข้อมูล ซึ่งถ้าไม่ปรับสมดุลให้กับข้อมูลที่ไม่สมดุลจะส่งผลให้ประสิทธิภาพแตกต่างกันหรือไม่มากน้อยเพียงใด

2) วิธี random oversampling จะเป็นการเพิ่มจำนวนข้อมูลกลุ่มรอง เพื่อทำให้จำนวนข้อมูลระหว่างกลุ่มรองและกลุ่มหลักมีจำนวนใกล้เคียงกัน ซึ่งการเพิ่มจำนวนของข้อมูลจะเป็นการสังเคราะห์และสร้างข้อมูลขึ้นมาใหม่จากกลุ่มรองหรือสุ่มเลือกข้อมูลจากกลุ่มรองให้มีจำนวนเพิ่มมากขึ้นจนใกล้เคียงหรือเท่ากับจำนวนข้อมูลในกลุ่มหลัก

3) วิธีผสมผสานระหว่างรูปแบบสุ่มเกินและสุ่มลด จะใช้วิธี random oversampling เพื่อเพิ่มจำนวนข้อมูลจากกลุ่มรอง และวิธี random undersampling เพื่อลดจำนวนข้อมูลจากกลุ่มหลัก เพื่อให้จำนวนข้อมูลกลุ่มรองและกลุ่มหลักมีจำนวนใกล้เคียงกัน กำหนดค่าความน่าจะเป็นเท่ากับ 0.5 ตามค่าเริ่มต้น

2. เทคนิคการจำแนกข้อมูล

จากการศึกษาเทคนิคการจำแนกที่นิยมใช้และมีประสิทธิภาพสูงจากงานวิจัยที่เก็บข้อมูลในสถานการณ์จริงที่สอดคล้องกับงานวิจัยซึ่งมีลักษณะของตัวแปรตามเป็นตัวแปรจัดประเภทที่มี 2 ค่า และมีตัวแปรอิสระส่วนใหญ่เป็นตัวแปรจัดประเภท ดังนั้นในการศึกษาครั้งนี้จึงกำหนดเทคนิคการจำแนก ดังนี้

1) การถดถอยโลจิสติก (logistic regression: LR) ใช้ในการวิเคราะห์เพื่อจำแนกข้อมูล โดยใช้การวิเคราะห์การถดถอยโลจิสติกทวิ เนื่องจากมีตัวแปรตามเป็นตัวแปรจัดประเภทที่มี 2 ค่า และจำแนกตัวแปรตามด้วยสมการดังนี้

$$\log(\text{odds}) = \ln\left(\frac{P_y}{Q_y}\right)$$

สามารถแสดงในรูปสมการถดถอยเชิงเส้นได้ดังสมการ

$$\log(\text{odds}) = \text{logit} = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_n X_n$$

เมื่อ P_y คือ ความน่าจะเป็นของการเกิดเหตุการณ์ที่สนใจ

Q_y คือ ความน่าจะเป็นของการไม่เกิดเหตุการณ์ที่สนใจ

β_0 คือ ค่าคงที่

β_1 คือ ค่าสัมประสิทธิ์ของตัวแปรอิสระ

X คือ ตัวแปรอิสระ

2) แรนดอมฟอรเรส (random forest: RF) จะจำแนกข้อมูลด้วยการสุ่มเลือกสุ่มเลือกคุณลักษณะและข้อมูลจากชุดข้อมูลทั้งหมด จากนั้นสร้างต้นไม้ตัดสินใจจากชุดข้อมูลตัวอย่างแต่ละชุด หาค่าทำนายจากต้นไม้แต่ละต้น และเลือกจำนวนต้นไม้ตัดสินใจที่ต้องการ แล้วทำซ้ำกระบวนการตั้งแต่เริ่มต้นในการสร้างต้นไม้ จากนั้นหาค่าพยากรณ์ ซึ่งค่าพยากรณ์ที่ได้จะเกิดจากการหาค่าพยากรณ์ของต้นไม้ตัดสินใจแต่ละต้น แล้วจะใช้วิธีผลโหวตมากที่สุด โดยผลลัพธ์จะเป็นค่าพยากรณ์ที่ได้จากต้นไม้ตัดสินใจที่มีค่าผลโหวตมากที่สุด

3) ซัพพอร์ตเวกเตอร์แมชชีน (support vector machine: SVM) จะสร้างเส้นตรงเพื่อจำแนกข้อมูล 2 กลุ่ม เส้นตรงดังกล่าวจะอยู่ตรงกลางระหว่างข้อมูล 2 กลุ่มและพยายามแบ่งข้อมูล 2 กลุ่มออกจากกันให้ดีที่สุด

4) วิธีเคเนียร์เรสเนเบอร์ (k-nearest neighbor: KNN) จะเปรียบเทียบความเหมือนกันหรือใกล้เคียงกันของคุณสมบัติระหว่างข้อมูลใหม่กับข้อมูลเดิมที่มี จากนั้นก็จำแนกประเภทของข้อมูลใหม่ออกมาให้เป็นประเภทเดียวกันกับข้อมูลเดิมที่อยู่ใกล้เคียงกัน โดยฟังก์ชันที่ใช้วัดระยะห่างคือฟังก์ชันระยะทางยูคลิเดียน ดังสมการ

$$d_{euclidean} = (u,v) = \sqrt{\sum_{i=1}^{n}(u_i - v_i)^2}$$

เมื่อ $u = \{u_1, u_2, u_3, ...,u_m\}$ และ $v = \{v_1, v_2, v_3, ...,v_m\}$ คือการบันทึกค่าคุณลักษณะของทั้งสองกลุ่ม

ระยะห่างระหว่าง x_0 ซึ่งเป็นค่าสังเกตใหม่ และ x_i ถูกกำหนดด้วยสมการ

$$d(x_0, x_i) = \sqrt{\sum_{i=1}^{n}[a_b(x_0) - a_b(x_i)]^2}$$

โดยค่าสังเกตค่าใหม่ในชุดข้อมูลทดสอบจะจัดเข้ากลุ่มที่มีค่า $\hat{p}_q = (x'_\ell \mid x_i)$ สูงสุด ซึ่งถูกกำหนดโดย

$$\hat{p}_q = (x'_\ell \mid x_i) = \frac{1}{k}\sum_{i=\ell}^{k}\delta_{(q)(x'_\ell)}$$

เมื่อ $\sum_{i=\ell}^{k}\delta_{(q)(x'_\ell)}$ คือ จำนวน x_i ของกลุ่ม q ภายในบริเวณใกล้เคียง k

$\delta_{(q)(x'_\ell)}$ คือ ฟังก์ชันดิแรก (dirac function) ซึ่งถูกกำหนดโดย

$$\delta_{(q)(x'_\ell)} = \begin{cases} 1, & q = x'_\ell \\ 0, & \text{otherwise} \end{cases}$$

3. ขนาดตัวอย่าง

ในการศึกษาครั้งนี้จะทำการศึกษาโดยการจำลองข้อมูลที่ใช้ขนาดตัวอย่างจากการศึกษางานวิจัยที่เก็บข้อมูลในสถานการณ์จริง อยู่ในช่วง 100 ถึง 500 ตัวอย่าง ดังนั้นในการศึกษาครั้งนี้จะทำการศึกษาโดยการจำลองข้อมูลจากขนาดตัวอย่าง 100, 300 และ 500 ตัวอย่าง

4. ตัวแปรอิสระ

4.1 ในการศึกษาครั้งนี้กำหนดตัวแปรอิสระ 8 ค่า ประกอบด้วยตัวแปรอิสระที่เป็นตัวแปรจัดประเภทและตัวแปรตัวแปรอิสระที่เป็นตัวแปรต่อเนื่อง โดยตัวแปรอิสระที่เป็นตัวแปรจัดประเภทจำลองตัวแปรด้วยการแจกแจงแบบทวินาม (binomial distribution) กำหนดพารามิเตอร์จำนวนครั้งที่ทดลอง (size) เท่ากับ 1 และความน่าจะเป็นกำหนดจากการแจกแจงแบบยูนิฟอร์ม โดยมีจำนวนค่าสังเกต (n) เท่ากับ 1 ค่าต่ำสุด (minimum) คือ 0.2 และค่าสูงสุด (maximum) คือ 0.8 สำหรับตัวแปรอิสระที่เป็นตัวแปรต่อเนื่องจำลองตัวแปรด้วยการแจกแจงแบบปกติ (normal distribution) กำหนดพารามิเตอร์ค่าเฉลี่ยเท่ากับ 0 และส่วนเบี่ยงเบนมาตรฐานเท่ากับ 1

4.2 จำนวนตัวแปรระหว่างกลุ่มตัวแปรจัดประเภทต่อกลุ่มตัวแปรต่อเนื่อง 3 กรณี ได้แก่ 4:4, 5:3 และ 6:2 ซึ่งกำหนดจากจากการศึกษางานวิจัยที่เก็บข้อมูลในสถานการณ์จริง แล้วนำมาคิดเป็นอัตราส่วน

4.3 อัตราออดของตัวแปรอิสระ

เนื่องจากในการศึกษาครั้งนี้มีจำนวนตัวแปรอิสระ 8 ค่า แบ่งเป็น 2 กลุ่ม คือ ตัวแปรจัดประเภทและตัวแปรต่อเนื่อง และกำหนดอัตราออดของตัวแปรอิสระแต่ละกลุ่ม 2 ระดับ ด้วยการแจกแจงแบบยูนิฟอร์ม กำหนดจำนวนค่าสังเกต (n) เท่ากับ 1 ค่าต่ำสุด (minimum) มีค่าอยู่ในระดับต่ำ (ช่วง [1,2)) แทนด้วย L และค่าสูงสุด (maximum) มีค่าอยู่ในระดับสูง (ช่วง [2,3)) แทนด้วย H แบ่งได้ 4 กรณีดังตารางที่ 3.1 ดังนี้

ตารางที่ 3.1 แสดงอัตราออดของตัวแปรอิสระ

กรณี	ตัวแปรจัดประเภท	ตัวแปรต่อเนื่อง	เขียนแทนด้วย
1	L	L	LL
2	L	H	LH
3	H	L	HL
4	H	H	HH

จากตารางที่ 3. 1 เมื่อทำการจำลองข้อมูลจะได้ ดังนี้

กรณีที่ 1 ตัวแปรอิสระเป็นตัวแปรจัดประเภทจะมีอัตราออด อยู่ในระดับต่ำ (ช่วง [1,2)) และตัวแปรอิสระเป็นตัวแปรต่อเนื่องจะมีอัตราออด อยู่ในระดับต่ำ (ช่วง [1,2))

กรณีที่ 2 ตัวแปรอิสระเป็นตัวแปรจัดประเภทจะมีอัตราออด อยู่ในระดับต่ำ (ช่วง [1,2)) และตัวแปรอิสระเป็นตัวแปรต่อเนื่องจะมีอัตราออด อยู่ในระดับสูง (ช่วง [2,3))

กรณีที่ 3 ตัวแปรอิสระเป็นตัวแปรจัดประเภทจะมีอัตราออด อยู่ในระดับสูง (ช่วง [2,3)) และตัวแปรอิสระเป็นตัวแปรต่อเนื่องจะมีอัตราออด อยู่ในระดับต่ำ (ช่วง [1,2))

กรณีที่ 4 ตัวแปรอิสระเป็นตัวแปรจัดประเภทจะมีอัตราออด อยู่ในระดับสูง (ช่วง [2,3)) และตัวแปรอิสระเป็นตัวแปรต่อเนื่องจะมีอัตราออด อยู่ในระดับสูง (ช่วง [2,3))

5. ร้อยละของจำนวนข้อมูลระหว่างข้อมูลกลุ่มหลักต่อข้อมูลกลุ่มรอง

ในการศึกษาครั้งนี้จะทำการศึกษาโดยการจำลองข้อมูลจากร้อยละของจำนวนข้อมูลระหว่างข้อมูลกลุ่มหลักต่อข้อมูลกลุ่มรอง 2 กรณี คือ 60:40 และ 70:30 ซึ่งกำหนดจากจากการศึกษางานวิจัยที่เก็บข้อมูลในสถานการณ์จริง แล้วนำมาคิดเป็นอัตราส่วน

ตอนที่ 2 การเปรียบเทียบประสิทธิภาพของแต่ละวิธีการปรับสมดุลข้อมูล

การวัดประสิทธิภาพการปรับสมดุลข้อมูลจะพิจารณาจากค่าความถูกต้อง (accuracy) ค่าความไว (sensitivity) และค่าความจำเพาะ (specificity) ค่าดังกล่าวคำนวณได้จากสมการต่อไปนี้

ค่าความถูกต้อง (accuracy) คือ การแสดงการวัดที่ได้มีความถูกต้องในรูปอัตราส่วน

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

ค่าความไว (sensitivity) คือ สัดส่วนของผลบวกที่เป็นจริงสำหรับภาวะนั้น ๆ

$$\text{sensitivity} = \frac{TP}{TP + FN}$$

ค่าความจำเพาะ (specificity) คือ สัดส่วนของผลลบที่เป็นจริงสำหรับภาวะนั้น ๆ

$$\text{specificity} = \frac{TN}{TN + FP}$$

เมื่อ TP = True Positive, TN = True Negative, FN = False Negative และ FP = False Positive

ตอนที่ 3 ขั้นตอนการดำเนินงานวิจัย

1. ศึกษาเอกสารและงานวิจัยที่เกี่ยวข้องกับการปรับสมดุลข้อมูลและเทคนิคการจำแนกข้อมูล

2. กำหนดขนาดตัวอย่าง 3 ขนาด ได้แก่ 100 300 และ 500

3. จำลองตัวแปรอิสระ 8 ค่า ด้วยการแจกแจงปกติ แบ่งเป็น 2 กลุ่ม คือ ตัวแปรจัดประเภทและตัวแปรต่อเนื่อง โดยกำหนดอัจำนวนตัวแปรระหว่างกลุ่มตัวแปรจัดประเภทต่อกลุ่มตัวแปรต่อเนื่อง 3 กรณี ได้แก่ 4:4 5:3 และ 6:2 และกำหนดอัตราออดของตัวแปรอิสระแต่ละกลุ่ม 2 ระดับ ด้วยการแจกแจงยูนิฟอร์ม โดยมีค่าอยู่ในระดับต่ำ (ช่วง [1,2)) แทนด้วย และ [2,3) แบ่งได้ 4 กรณี ได้แก่ [1,2) กับ [1,2) , [1,2) กับ [2,3) , [2,3) กับ [1,2) และ [2,3) กับ [2,3)

4. จำลองตัวแปรตามด้วยการวิเคราะห์การถดถอยโลจิสติกจากการประมาณค่าความน่าจะเป็นในการเกิดเหตุการณ์ โดยมีฟังก์ชันโลจิสติกเป็นตัวแบบ ซึ่งในงานวิจัยนี้มีตัวแปรอิสระหลายตัว (n) จะแสดงความน่าจะเป็นของเหตุการณ์ได้ดังสมการต่อไปนี้

$$P(y) = \frac{e^{Z}}{1 + e^{Z}} \text{ หรือ } P(y) = \frac{1}{1 + e^{-Z}}$$

เมื่อ Z คือ linear combination ของตัวแปรตาม

$$Z = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_n X_n$$

และความน่าจะเป็นของการไม่เกิดเหตุการณ์ คือ

$$Q_y = 1 - P_y$$

เมื่อ P_y คือ ความน่าจะเป็นของการเกิดเหตุการณ์ที่สนใจ

Q_y คือ ความน่าจะเป็นของการไม่เกิดเหตุการณ์ที่สนใจ

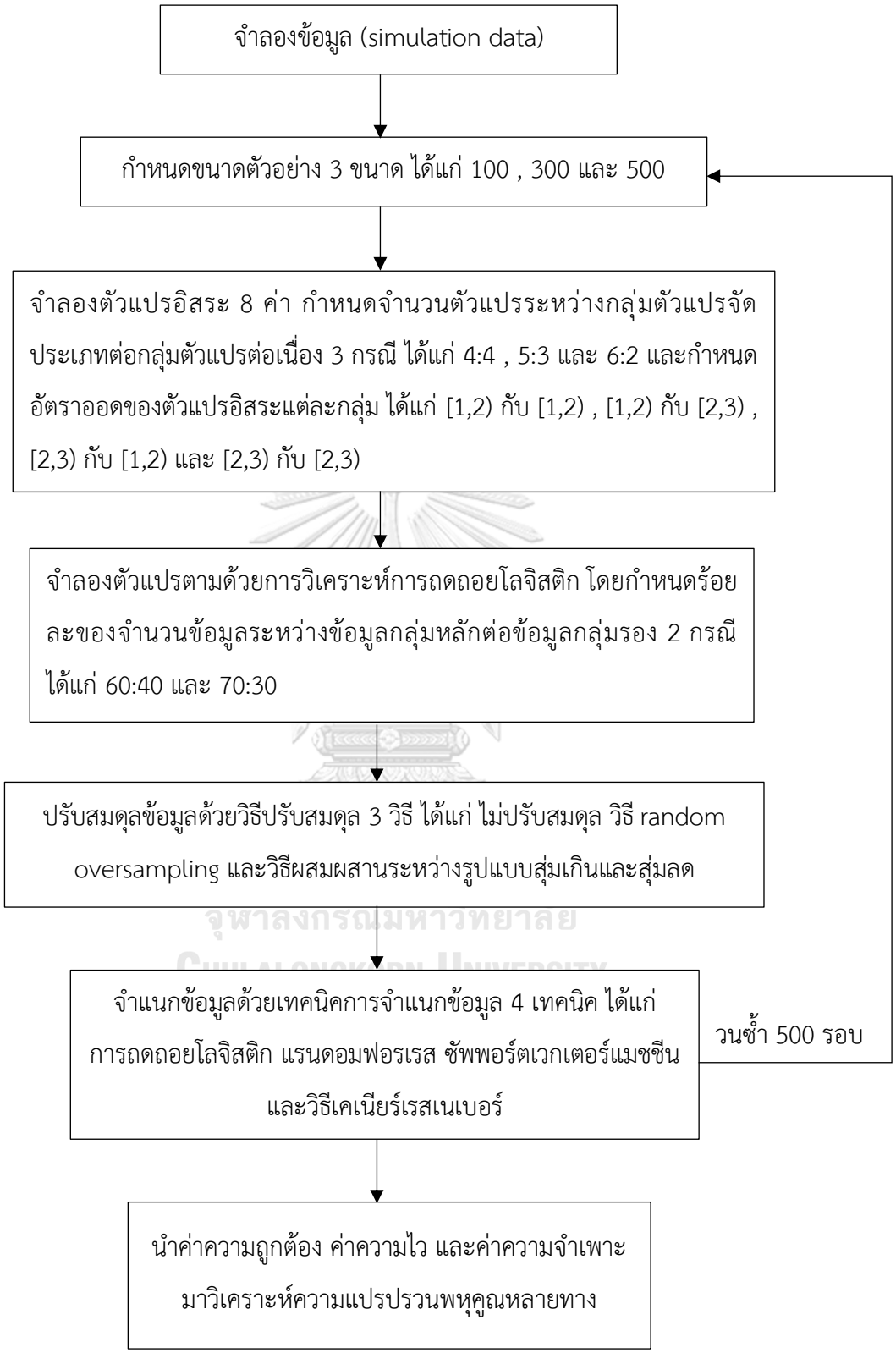
โดยกำหนดร้อยละของจำนวนข้อมูลระหว่างข้อมูลกลุ่มหลักต่อข้อมูลกลุ่มรอง 2 กรณี ได้แก่ 60:40 และ 70:30

5. ปรับสมดุลข้อมูลด้วยวิธีการปรับสมดุลข้อมูล 3 วิธี ได้แก่ ไม่ปรับสมดุล วิธี random oversampling และวิธีผสมผสานระหว่างรูปแบบสุ่มเกินและสุ่มลด

6. จำแนกข้อมูลด้วยเทคนิคการจำแนกข้อมูล 4 เทคนิค ได้แก่ การถดถอยโลจิสติก (LR) แรนดอมฟอรเรส (RF) ซัพพอร์ตเวกเตอร์แมชชีน (SVM) และวิธีเคเนียร์เรสเนเบอร์ (KNN)

7. นำค่าความถูกต้อง ค่าความไว และค่าความจำเพาะ ด้วยการวิเคราะห์ความพหุคูณหลายทาง (n-way multivariate analysis of variance: n-way MANOVA) แล้วสรุปผลข้อมูล

ซึ่งสามารถสรุปขั้นตอนการจำลองข้อมูลได้ดังรูปที่ 3.1



รูปที่ 3.1 ขั้นตอนการจำลองข้อมูล

บทที่ 4
ผลการวิเคราะห์ข้อมูล

การวิจัยนี้มีวัตถุประสงค์เพื่อศึกษาปฏิสัมพันธ์ของวิธีการปรับสมดุลข้อมูลกับเงื่อนไขขนาดตัวอย่าง เทคนิคการจำแนกข้อมูล จำนวนตัวแปรระหว่างกลุ่มตัวแปรจัดประเภทต่อกลุ่มตัวแปรต่อเนื่อง อัตราออด และจำนวนข้อมูลระหว่างข้อมูลกลุ่มหลักต่อข้อมูลกลุ่มรองที่มีต่อประสิทธิภาพของการจำแนกกลุ่ม โดยกำหนดจำนวนตัวแปรอิสระ 8 ค่า ประกอบด้วยตัวแปรจัดประเภทและตัวแปรต่อเนื่อง และทำการจำลองข้อมูลโดยมีเงื่อนไขดังนี้ วิธีปรับสมดุลข้อมูล 3 วิธี ได้แก่ ไม่ปรับสมดุล วิธี random oversampling และวิธีผสมผสานระหว่างรูปแบบสุ่มเกินและสุ่มลด เทคนิคการจำแนก 4 เทคนิค ได้แก่ การถดถอยโลจิสติก แรนดอมฟอรเรส ซัพพอร์ตเวกเตอร์แมชชีน และวิธีเคเนียร์เรสเนเบอร์ ขนาดตัวอย่าง 3 ขนาด ได้แก่ 100, 300 และ 500 จำนวนตัวแปรระหว่างกลุ่มตัวแปรจัดประเภทต่อกลุ่มตัวแปรต่อเนื่อง 3 กรณี ได้แก่ 4:4, 5:3 และ 6:2 และอัตราออดของตัวแปรอิสระ 4 กรณี และร้อยละของจำนวนข้อมูลระหว่างข้อมูลกลุ่มหลักต่อข้อมูลกลุ่มรอง 2 กรณี ได้แก่ 60:40 และ 70:30 และจำลองข้อมูลตัวแปรตามจากการถดถอยโลจิสติก โดยใช้การจำลองข้อมูลด้วยวิธีการมอนติคาร์โล (Monte Carlo simulation) รวมทั้งสิ้น 864 เงื่อนไขกำหนดการทำซ้ำ 500 รอบ โดยผู้วิจัยกำหนดสัญลักษณ์ดังต่อไปนี้ แทนความหมายต่าง ๆ ในตาราง รูปภาพ และสรุปผลการวิเคราะห์ข้อมูล

NS	หมายถึง ขนาดตัวอย่าง
BL	หมายถึง วิธีการปรับสมดุลข้อมูล
NONE	หมายถึง วิธีไม่ปรับสมดุลข้อมูล
OVER	หมายถึง วิธีปรับสมดุลข้อมูลด้วยวิธี random oversampling
HYBRID	หมายถึง วิธีปรับสมดุลข้อมูลด้วยวิธีผสมผสานระหว่างรูปแบบสุ่มเกินและสุ่มลด
OR	หมายถึง อัตราออด
PMM	หมายถึง ร้อยละของจำนวนข้อมูลระหว่างข้อมูลกลุ่มหลักต่อข้อมูลกลุ่มรอง
CATCON	หมายถึง จำนวนตัวแปรระหว่างกลุ่มตัวแปรจัดประเภทต่อกลุ่มตัวแปรต่อเนื่อง
CLS	หมายถึง เทคนิคการจำแนกข้อมูล

LR	หมายถึง การจำแนกข้อมูลด้วยการถดถอยโลจิสติก
RF	หมายถึง การจำแนกข้อมูลด้วยแรนดอมฟอร์เรส
SVM	หมายถึง การจำแนกข้อมูลด้วยซัพพอร์ตเวกเตอร์แมชชีน
KNN	หมายถึง การจำแนกข้อมูลด้วยวิธีเคเนียร์เรสเนเบอร์
LL	หมายถึง อัตราออดของตัวแปรอิสระที่เป็นตัวแปรจัดประเภท อยู่ในระดับต่ำ (ช่วง [1,2)) และอัตราออดของตัวแปรอิสระที่เป็นตัวแปรต่อเนื่อง อยู่ในระดับต่ำ (ช่วง [1,2))
LH	หมายถึง อัตราออดของตัวแปรอิสระที่เป็นตัวแปรจัดประเภท อยู่ในระดับต่ำ (ช่วง [1,2)) และอัตราออดของตัวแปรอิสระที่เป็นตัวแปรต่อเนื่อง อยู่ในระดับสูง (ช่วง [2,3))
HL	หมายถึง อัตราออดของตัวแปรอิสระที่เป็นตัวแปรจัดประเภท อยู่ในระดับสูง (ช่วง [2,3)) และอัตราออดของตัวแปรอิสระที่เป็นตัวแปรต่อเนื่อง อยู่ในระดับต่ำ (ช่วง [1,2))
HH	หมายถึง อัตราออดของตัวแปรอิสระที่เป็นตัวแปรจัดประเภท อยู่ในระดับสูง (ช่วง [2,3)) และอัตราออดของตัวแปรอิสระที่เป็นตัวแปรต่อเนื่อง อยู่ในระดับสูง (ช่วง [2,3))
Mean	หมายถึง ค่าเฉลี่ย
SD	หมายถึง ค่าส่วนเบี่ยงเบนมาตรฐาน
accuracy	หมายถึง ค่าความถูกต้องในการจำแนกข้อมูล
sensitivity	หมายถึง ค่าความไวในการจำแนกข้อมูล
specificity	หมายถึง ค่าความจำเพาะในการจำแนกข้อมูล

ผลการวิเคราะห์ข้อมูลมีรายละเอียดดังต่อไปนี้

จากวัตถุประสงค์การวิจัยซึ่งศึกษาปฏิสัมพันธ์ของวิธีการปรับสมดุลข้อมูล (BL) กับเงื่อนไขต่าง ๆ ได้แก่ ขนาดตัวอย่าง (NS) เทคนิคการจำแนกข้อมูล (CLS) จำนวนตัวแปรระหว่างกลุ่มตัวแปรจัดประเภทต่อกลุ่มตัวแปรต่อเนื่อง (CATCON) อัตราออด (OR) และร้อยละของจำนวนข้อมูลระหว่างข้อมูลกลุ่มหลักต่อข้อมูลกลุ่มรอง (PMM) ที่มีต่อประสิทธิภาพของการจำแนกกลุ่ม ได้แก่ ค่าความถูกต้อง (accuracy) ค่าความไว (sensitivity) และค่าความจำเพาะ (specificity) โดยใช้การวิเคราะห์ความแปรปรวนพหุคูณหลายทาง (n-way MANOVA) พบว่า วิธีการปรับสมดุลข้อมูล

มีปฏิสัมพันธ์สองทางกับเงื่อนไขด้านจำนวนตัวแปรระหว่างกลุ่มตัวแปรจัดประเภทต่อกลุ่มตัวแปรต่อเนื่อง ร้อยละของจำนวนข้อมูลระหว่างข้อมูลกลุ่มหลักต่อข้อมูลกลุ่มรอง อัตราออด เทคนิคการจำแนกข้อมูล อย่างมีนัยสำคัญทางสถิติที่ระดับ .05 ทั้งนี้ยังพบปฏิสัมพันธ์สามทาง 3 ชุด ได้แก่ (1) วิธีการปรับสมดุลข้อมูล*ขนาดตัวอย่าง*จำนวนตัวแปรระหว่างกลุ่มตัวแปรจัดประเภทต่อกลุ่มตัวแปรต่อเนื่อง (2) วิธีการปรับสมดุลข้อมูล*ขนาดตัวอย่าง*เทคนิคการจำแนกข้อมูล และ (3) วิธีการปรับสมดุลข้อมูล*ร้อยละของจำนวนข้อมูลระหว่างข้อมูลกลุ่มหลักต่อข้อมูลกลุ่มรอง*เทคนิคการจำแนกข้อมูล อย่างมีนัยสำคัญทางสถิติที่ระดับ .05 แสดงให้เห็นถึงการใช้วิธีการปรับสมดุลข้อมูลที่ต้องพิจารณาร่วมกับเงื่อนไขของสภาพข้อมูลด้วย ผลการวิเคราะห์ในตารางที่ 4.1

ตารางที่ 4.1 ผลการวิเคราะห์ความแปรปรวนพหุคูณหลายทาง (n-way MANOVA) ของวิธีการปรับสมดุลข้อมูลกับเงื่อนไขต่าง ๆ ที่มีนัยสำคัญที่ระดับ .05

Source	Pillai's Trace	F	Hypothesis df	Error df	p
BL*CATCON	0.037	2.725	12	2592	0.001
BL*PMM	0.034	4.905	6	1726	< .001
BL*OR	0.036	1.764	18	2592	0.024
BL*CLS	0.128	6.443	18	2592	< .001
BL*CATCON *NS	0.047	1.708	24	2592	0.017
BL*CLS*NS	0.091	2.262	36	2592	< .001
BL*PMM*CLS	0.036	1.742	18	2592	0.027

หมายเหตุ BL = วิธีการปรับสมดุลข้อมูล, CATCON = จำนวนตัวแปรระหว่างกลุ่มตัวแปรจัดประเภทต่อกลุ่มตัวแปรต่อเนื่อง, PMM = ร้อยละของจำนวนข้อมูลระหว่างข้อมูลกลุ่มหลักต่อข้อมูลกลุ่มรอง, OR = อัตราออด, CLS = เทคนิคการจำแนกข้อมูล, NS = ขนาดตัวอย่าง

เมื่อพิจารณารายละเอียดค่าคะแนนเฉลี่ยตัวบ่งชี้ค่าความถูกต้อง ค่าความไว และค่าความจำเพาะภายใต้สถานการณ์ปฏิสัมพันธ์ระหว่างวิธีการปรับสมดุลข้อมูลกับเงื่อนไขต่าง ๆ ได้ข้อค้นพบดังนี้

1. ผลการวิเคราะห์ปฏิสัมพันธ์สองทางของวิธีการปรับสมดุลข้อมูลกับเงื่อนไขต่าง ๆ ที่มีนัยสำคัญ

1.1 วิธีการปรับสมดุลข้อมูลกับจำนวนตัวแปรระหว่างกลุ่มตัวแปรจัดประเภทต่อกลุ่มตัวแปรต่อเนื่อง

ผลการวิเคราะห์ข้อมูลพบว่า เมื่อจำนวนตัวแปรระหว่างกลุ่มตัวแปรจัดประเภทต่อกลุ่มตัวแปรต่อเนื่องมีค่าต่างกัน จะใช้วิธีการปรับสมดุลข้อมูลที่แตกต่างกัน โดยตัวแปรระหว่างกลุ่มตัวแปรจัดประเภทต่อกลุ่มตัวแปรต่อเนื่องมีค่า 4:4 การไม่ปรับสมดุลข้อมูลจะมีค่าความถูกต้องและค่าความจำเพาะในการจำแนกข้อมูลสูงกว่าวิธีอื่น ๆ ในขณะที่การปรับสมดุลข้อมูลด้วยวิธีผสมผสานระหว่างรูปแบบสุ่มเกินและสุ่มลดจะมีค่าความไวในการจำแนกข้อมูลสูงกว่าวิธีอื่น ๆ

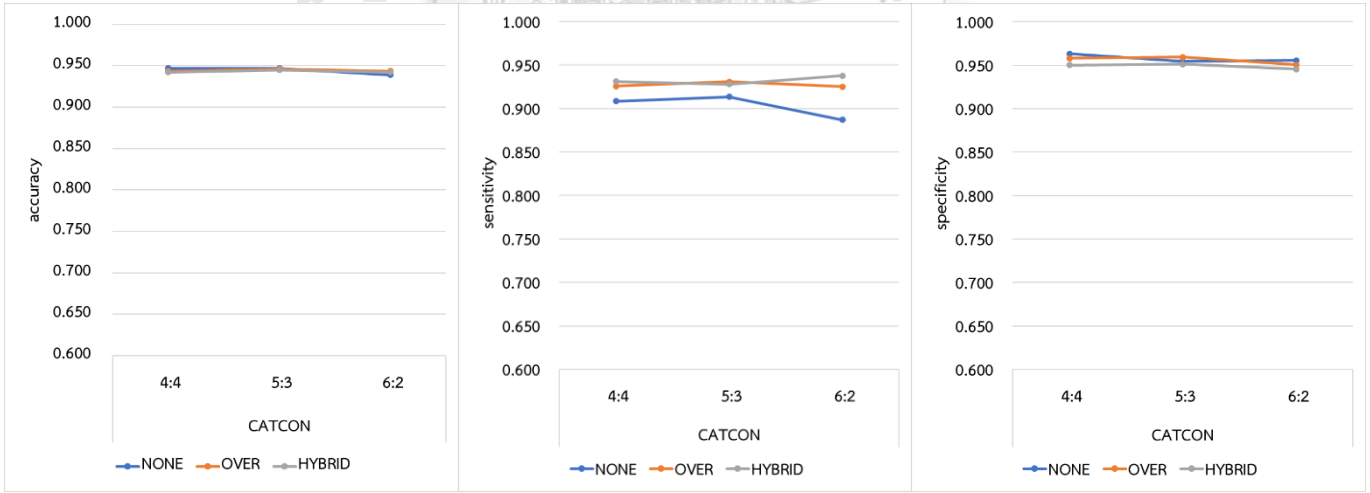
เมื่อพิจารณาสถานการณ์ที่มีจำนวนตัวแปรระหว่างกลุ่มตัวแปรจัดประเภทต่อกลุ่มตัวแปรต่อเนื่องเท่ากับ 5:3 พบว่า การปรับสมดุลข้อมูลด้วยวิธี random oversampling จะให้ค่าความไวและค่าความจำเพาะในการจำแนกข้อมูลสูงกว่าวิธีอื่น ๆ แต่มีค่าความถูกต้องในการจำแนกข้อมูลไม่แตกต่างจากวิธีอื่น ๆ

เมื่อพิจารณาสถานการณ์ที่มีจำนวนตัวแปรระหว่างกลุ่มตัวแปรจัดประเภทต่อกลุ่มตัวแปรต่อเนื่องเท่ากับ 6:2 พบว่า การปรับสมดุลข้อมูลด้วยวิธี random oversampling และวิธีผสมผสานระหว่างรูปแบบสุ่มเกินและสุ่มลดมีค่าความถูกต้องในการจำแนกข้อมูลไม่แตกต่างกัน แต่สูงกว่าการไม่ปรับสมดุลข้อมูล ในขณะที่การปรับสมดุลข้อมูลด้วยวิธีผสมผสานระหว่างรูปแบบสุ่มเกินและสุ่มลดจะมีค่าความไวในการจำแนกข้อมูลสูงกว่าวิธีอื่น ๆ ทั้งนี้การปรับสมดุลข้อมูลทั้งสามวิธีจะให้ค่าความจำเพาะในการจำแนกข้อมูลไม่แตกต่างกัน ดังตารางที่ 4.2 และรูปที่ 4.1

ตารางที่ 4.2 ผลการเปรียบเทียบค่าเฉลี่ยแต่ละวิธีการปรับสมดุลข้อมูลกับเงื่อนไขจำนวนตัวแปรระหว่างกลุ่มตัวแปรจัดประเภทต่อกลุ่มตัวแปรต่อเนื่อง

BL	จำนวนตัวแปรระหว่างกลุ่มตัวแปรจัดประเภทต่อกลุ่มตัวแปรต่อเนื่อง																	
	ค่าความถูกต้อง						ค่าความไว						ค่าความจำเพาะ					
	4:4		5:3		6:2		4:4		5:3		6:2		4:4		5:3		6:2	
	Mean	SD	Mean	SD	Mean	SD	Mean	SD	Mean	SD	Mean	SD	Mean	SD	Mean	SD	Mean	SD
NONE	**0.947**	0.047	**0.947**	0.036	0.939	0.054	0.909	0.100	0.914	0.095	0.887	0.118	**0.963**	0.026	0.954	0.028	**0.956**	0.037
OVER	0.944	0.057	0.946	0.048	**0.944**	0.056	0.926	0.078	**0.931**	0.069	0.925	0.082	0.958	0.042	**0.959**	0.035	0.951	0.051
HYBRID	0.942	0.059	0.945	0.047	0.942	0.061	**0.931**	0.070	0.928	0.075	**0.938**	0.067	0.950	0.054	0.951	0.048	0.946	0.058

หมายเหตุ BL = วิธีการปรับสมดุลข้อมูล, Mean = ค่าเฉลี่ย, SD = ค่าส่วนเบี่ยงเบนมาตรฐาน



รูปที่ 4.1 ผลการเปรียบเทียบค่าเฉลี่ยแต่ละวิธีการปรับสมดุลข้อมูลกับเงื่อนไขจำนวนตัวแปรระหว่างกลุ่มตัวแปรจัดประเภทต่อกลุ่มตัวแปรต่อเนื่อง

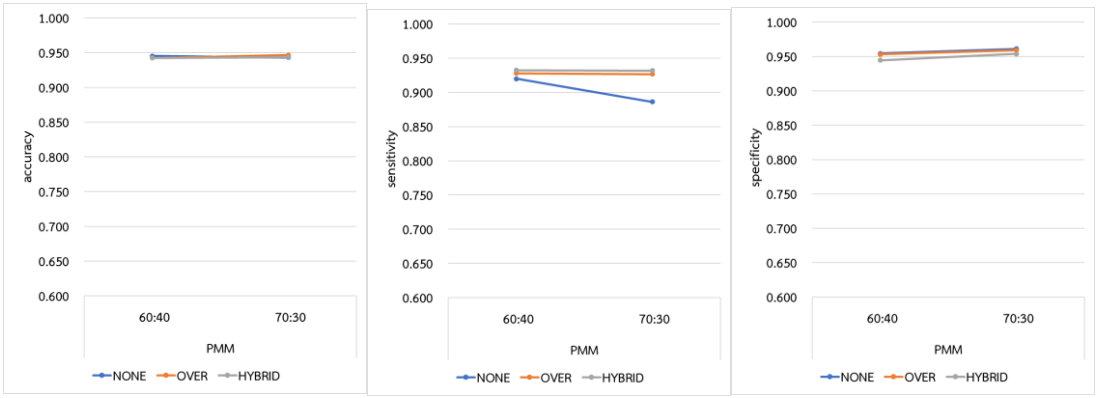
1.2 วิธีการปรับสมดุลข้อมูลกับร้อยละของจำนวนข้อมูลระหว่างข้อมูลกลุ่มหลักต่อข้อมูลกลุ่มรอง

ผลการวิเคราะห์ข้อมูลพบว่า เมื่อร้อยละของจำนวนข้อมูลระหว่างข้อมูลกลุ่มหลักต่อข้อมูลกลุ่มรองมีค่าต่างกัน จะใช้วิธีการปรับสมดุลข้อมูลที่แตกต่างกัน โดยร้อยละของจำนวนข้อมูลระหว่างข้อมูลกลุ่มหลักต่อข้อมูลกลุ่มรองมีค่า 60:40 การไม่ปรับสมดุลข้อมูลจะมีค่าความถูกต้องและค่าความจำเพาะในการจำแนกข้อมูลสูงกว่าวิธีอื่น ๆ ในขณะที่การปรับสมดุลข้อมูลด้วยวิธีผสมผสานระหว่างรูปแบบสุ่มเกินและสุ่มลดจะมีค่าความไวในการจำแนกข้อมูลสูงกว่าวิธีอื่น ๆ เมื่อพิจารณาสถานการณ์ที่มีร้อยละของจำนวนข้อมูลระหว่างข้อมูลกลุ่มหลักต่อข้อมูลกลุ่มรองเท่ากับ 70:30 พบว่าการปรับสมดุลข้อมูลด้วยวิธี random oversampling และวิธีผสมผสานระหว่างรูปแบบสุ่มเกินและสุ่มลดมีค่าความถูกต้องในการจำแนกข้อมูลไม่แตกต่างกัน แต่สูงกว่าการไม่ปรับสมดุลข้อมูล ในขณะที่การปรับสมดุลข้อมูลด้วยวิธีผสมผสานระหว่างรูปแบบสุ่มเกินและสุ่มลดจะมีค่าความไวในการจำแนกข้อมูลสูงกว่าวิธีอื่น ๆ ทั้งนี้การปรับสมดุลข้อมูลทั้งสามวิธีจะให้ค่าความจำเพาะในการจำแนกข้อมูลไม่แตกต่างกัน ดังตารางที่ 4.3 และรูปที่ 4.2

ตารางที่ 4.3 ผลการเปรียบเทียบค่าเฉลี่ยแต่ละวิธีการปรับสมดุลข้อมูลกับเงื่อนไขร้อยละของจำนวนข้อมูลระหว่างข้อมูลกลุ่มหลักต่อข้อมูลกลุ่มรอง

BL	ร้อยละของจำนวนข้อมูลระหว่างข้อมูลกลุ่มหลักต่อข้อมูลกลุ่มรอง											
	ค่าความถูกต้อง				ค่าความไว				ค่าความจำเพาะ			
	60:40		70:30		60:40		70:30		60:40		70:30	
	Mean	SD	Mean	SD	Mean	SD	Mean	SD	Mean	SD	Mean	SD
NONE	**0.945**	0.045	0.943	0.048	0.920	0.083	0.886	0.121	**0.954**	0.033	**0.961**	0.028
OVER	0.943	0.058	**0.946**	0.049	0.928	0.076	0.927	0.077	0.953	0.048	0.959	0.037
HYBRID	0.942	0.055	0.944	0.058	**0.932**	0.071	**0.932**	0.072	0.944	0.058	0.954	0.049

หมายเหตุ BL = วิธีการปรับสมดุลข้อมูล, Mean = ค่าเฉลี่ย, SD = ค่าส่วนเบี่ยงเบนมาตรฐาน



รูปที่ 4.2 ผลการเปรียบเทียบค่าเฉลี่ยแต่ละวิธีการปรับสมดุลข้อมูลกับเงื่อนไขร้อยละของจำนวนข้อมูลระหว่างข้อมูลกลุ่มหลักต่อข้อมูลกลุ่มรอง

1.3 วิธีการปรับสมดุลข้อมูลกับอัตราออด

ผลการวิเคราะห์ข้อมูลพบว่า เมื่ออัตราออดมีค่าต่างกัน จะใช้วิธีการปรับสมดุลข้อมูลที่แตกต่างกัน โดยอัตราออดของตัวแปรอิสระที่เป็นตัวแปรจัดประเภท อยู่ในระดับสูง (ช่วง [2,3)) และอัตราออดของตัวแปรอิสระที่เป็นตัวแปรต่อเนื่อง อยู่ในระดับสูง (ช่วง [2,3)) การปรับสมดุลข้อมูลด้วยวิธี random oversampling จะมีค่าความถูกต้องและค่าความจำเพาะในการจำแนกข้อมูลสูงกว่าวิธีอื่น ๆ ในขณะที่การปรับสมดุลข้อมูลด้วยวิธี random oversampling และวิธีผสมผสานระหว่างรูปแบบสุ่มเกินและสุ่มลดมีค่าความไวในการจำแนกข้อมูลไม่แตกต่างกัน แต่สูงกว่าการไม่ปรับสมดุลข้อมูล

เมื่อพิจารณาสถานการณ์ที่มีอัตราออดของตัวแปรอิสระที่เป็นตัวแปรจัดประเภท อยู่ในระดับสูง (ช่วง [2,3)) และอัตราออดของตัวแปรอิสระที่เป็นตัวแปรต่อเนื่อง อยู่ในระดับต่ำ (ช่วง [1,2)) พบว่า การปรับสมดุลข้อมูลด้วยวิธี random oversampling และการไม่ปรับสมดุลข้อมูลมีค่าความถูกต้องและค่าความจำเพาะในการจำแนกข้อมูลไม่แตกต่างกัน แต่สูงกว่าการจำแนกข้อมูลด้วยวิธีผสมผสานระหว่างรูปแบบสุ่มเกินและสุ่มลด ในขณะที่การปรับสมดุลข้อมูลด้วยวิธีผสมผสานระหว่างรูปแบบสุ่มเกินและสุ่มลดจะมีค่าความไวในการจำแนกข้อมูลสูงกว่าวิธีอื่น ๆ

เมื่อพิจารณาสถานการณ์ที่มีอัตราออดของตัวแปรอิสระที่เป็นตัวแปรจัดประเภท อยู่ในระดับต่ำ (ช่วง [1,2)) และอัตราออดของตัวแปรอิสระที่เป็นตัวแปรต่อเนื่อง อยู่ในระดับสูง (ช่วง [2,3)) พบว่า การไม่ปรับสมดุลข้อมูลจะมีค่าความถูกต้องและค่าความจำเพาะในการจำแนกข้อมูลสูงกว่าวิธีอื่น ๆ ในขณะที่การปรับสมดุลข้อมูลด้วยวิธีผสมผสานระหว่างรูปแบบสุ่มเกินและสุ่มลดจะมีค่าความไวในการจำแนกข้อมูลสูงกว่าวิธีอื่น ๆ

เมื่อพิจารณาสถานการณ์ที่มีอัตราออดของตัวแปรอิสระที่เป็นตัวแปรจัดประเภท อยู่ในระดับต่ำ (ช่วง [1,2)) และอัตราออดของตัวแปรอิสระที่เป็นตัวแปรต่อเนื่อง อยู่ในระดับต่ำ (ช่วง [1,2)) พบว่า การปรับสมดุลข้อมูลด้วยวิธีผสมผสานระหว่างรูปแบบสุ่มเกินและสุ่มลดจะมีค่าความถูกต้องและค่าความไวในการจำแนกข้อมูลสูงกว่าวิธีอื่น ๆ ทั้งนี้การปรับสมดุลข้อมูลทั้งสามวิธีจะให้ค่าความจำเพาะในการจำแนกข้อมูลไม่แตกต่างกัน ดังตารางที่ 4.4 และรูปที่ 4.3

1.4 วิธีการปรับสมดุลข้อมูลกับเทคนิคการจำแนกข้อมูล

ผลการวิเคราะห์ข้อมูลพบว่า เมื่อเทคนิคการจำแนกข้อมูลต่างกัน จะใช้วิธีการปรับสมดุลข้อมูลที่แตกต่างกัน โดยจำแนกข้อมูลด้วยวิธีเคเนียร์เรสเนเบอร์และวิธีการถดถอยโลจิสติก การไม่ปรับสมดุลข้อมูลจะมีค่าความถูกต้องและค่าความจำเพาะในการจำแนกข้อมูลสูงกว่าวิธีอื่น ๆ ในขณะที่การปรับสมดุลข้อมูลด้วยวิธีผสมผสานระหว่างรูปแบบสุ่มเกินและสุ่มลดจะมีค่าความไวในการจำแนกข้อมูลสูงกว่าวิธีอื่น ๆ

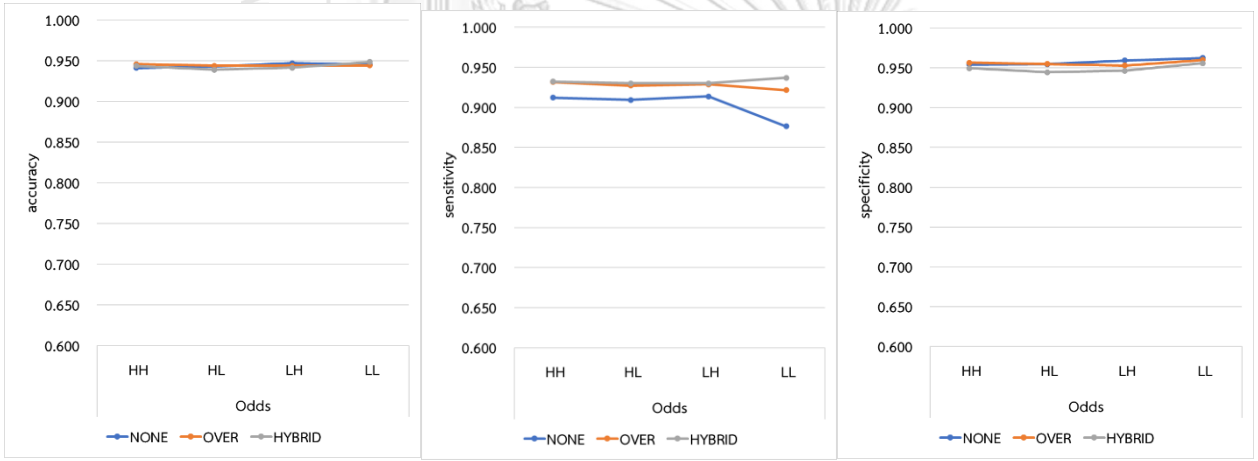
เมื่อพิจารณาสถานการณ์ที่มีจำแนกข้อมูลด้วยวิธีแรนดอมฟอร์เรส พบว่า การปรับสมดุลข้อมูลด้วยวิธี random oversampling มีค่าความถูกต้อง ค่าความไว และค่าความจำเพาะในการจำแนกข้อมูลสูงกว่าวิธีอื่น ๆ

เมื่อพิจารณาสถานการณ์ที่จำแนกข้อมูลด้วยวิธีซัพพอร์ตเวกเตอร์แมชชีน พบว่า การปรับสมดุลข้อมูลด้วยวิธี random oversampling มีค่าความถูกต้องและค่าความจำเพาะในการจำแนกข้อมูลสูงกว่าวิธีอื่น ๆ ในขณะที่การปรับสมดุลข้อมูลด้วยวิธีผสมผสานระหว่างรูปแบบสุ่มเกินและสุ่มลดจะมีค่าความไวในการจำแนกข้อมูลสูงกว่าวิธีอื่น ๆ ดังตารางที่ 4.5 และรูปที่ 4.4

ตารางที่ 4.4 ผลการเปรียบเทียบค่าเฉลี่ยแต่ละวิธีการปรับสมดุลข้อมูลกับเงื่อนไขอัตราออด

BL	อัตราออด																							
	ค่าความถูกต้อง								ค่าความไว								ค่าความจำเพาะ							
	HH		HL		LH		LL		HH		HL		LH		LL		HH		HL		LH		LL	
	Mean	SD	Mean	SD	Mean	SD	Mean	SD	Mean	SD	Mean	SD	Mean	SD	Mean	SD	Mean	SD	Mean	SD	Mean	SD	Mean	SD
NONE	0.941	0.047	0.943	0.052	**0.947**	0.041	0.946	0.045	0.912	0.091	0.910	0.098	0.914	0.086	0.877	0.135	0.954	0.031	**0.955**	0.033	**0.959**	0.027	**0.963**	0.030
OVER	**0.946**	0.049	**0.944**	0.059	0.944	0.048	0.944	0.059	**0.932**	0.068	0.927	0.081	0.929	0.066	0.922	0.090	**0.957**	0.040	**0.955**	0.048	0.953	0.040	0.960	0.044
HYBRID	0.944	0.050	0.939	0.065	0.941	0.054	**0.948**	0.053	**0.932**	0.070	**0.930**	0.073	**0.930**	0.069	**0.937**	0.073	0.950	0.050	0.944	0.062	0.946	0.055	0.956	0.045

หมายเหตุ BL = วิธีการปรับสมดุลข้อมูล, Mean = ค่าเฉลี่ย, SD = ค่าส่วนเบี่ยงเบนมาตรฐาน

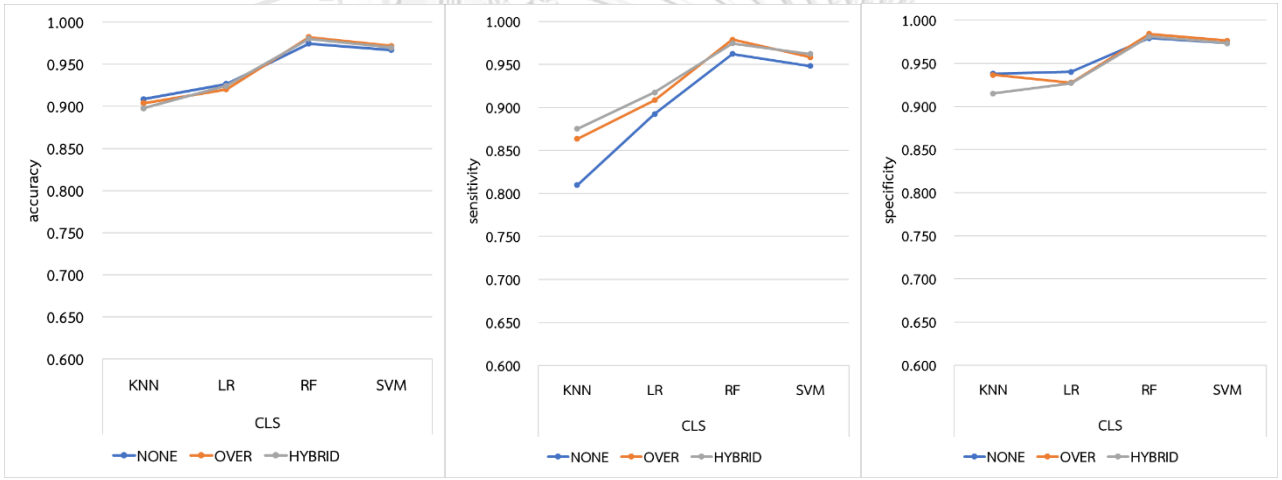


รูปที่ 4.3 ผลการเปรียบเทียบค่าเฉลี่ยแต่ละวิธีการปรับสมดุลข้อมูลกับเงื่อนไขอัตราออด

ตารางที่ 4.5 ผลการเปรียบเทียบค่าเฉลี่ยแต่ละวิธีการปรับสมดุลข้อมูลกับเงื่อนไขเทคนิคการจำแนกข้อมูล

BL	เทคนิคการจำแนกข้อมูล																							
	ค่าความถูกต้อง								ค่าความไว								ค่าความจำเพาะ							
	KNN		LR		RF		SVM		KNN		LR		RF		SVM		KNN		LR		RF		SVM	
	Mean	SD	Mean	SD	Mean	SD	Mean	SD	Mean	SD	Mean	SD	Mean	SD	Mean	SD	Mean	SD	Mean	SD	Mean	SD	Mean	SD
NONE	**0.909**	0.058	**0.927**	0.039	0.975	0.017	0.967	0.018	0.810	0.141	0.892	0.085	0.962	0.029	0.948	0.046	**0.938**	0.030	**0.940**	0.026	0.979	0.015	0.973	0.024
OVER	0.904	0.062	0.920	0.055	**0.982**	0.013	**0.972**	0.016	0.864	0.102	0.909	0.065	**0.979**	0.016	0.959	0.024	0.937	0.037	0.928	0.058	**0.984**	0.012	**0.976**	0.017
HYBRID	0.898	0.071	0.924	0.050	0.980	0.015	0.970	0.018	**0.875**	0.098	**0.917**	0.061	0.974	0.017	**0.962**	0.020	0.915	0.068	0.927	0.055	0.981	0.014	0.973	0.019

หมายเหตุ BL = วิธีการปรับสมดุลข้อมูล, Mean = ค่าเฉลี่ย, SD = ค่าส่วนเบี่ยงเบนมาตรฐาน



รูปที่ 4.4 ผลการเปรียบเทียบค่าเฉลี่ยแต่ละวิธีการปรับสมดุลข้อมูลกับเงื่อนไขเทคนิคการจำแนกข้อมูล

2. ผลการวิเคราะห์ปฏิสัมพันธ์สามทางของวิธีการปรับสมดุลข้อมูลกับเงื่อนไขต่าง ๆ ที่มีนัยสำคัญ

2.1 วิธีการปรับสมดุลข้อมูล จำนวนตัวแปรระหว่างกลุ่มตัวแปรจัดประเภทต่อกลุ่มตัวแปรต่อเนื่อง และขนาดตัวอย่าง

ผลการวิเคราะห์ข้อมูลพบปฏิสัมพันธ์สามทางระหว่างวิธีการปรับสมดุลข้อมูล จำนวนตัวแปรระหว่างกลุ่มตัวแปรจัดประเภทต่อกลุ่มตัวแปรต่อเนื่อง และขนาดตัวอย่าง ในกรณีที่ขนาดตัวอย่างเท่ากับ 100 หน่วย และจำนวนตัวแปรระหว่างกลุ่มตัวแปรจัดประเภทต่อกลุ่มตัวแปรต่อเนื่องมีจำนวนเท่ากัน (4:4) พบว่า การไม่ปรับสมดุลข้อมูลจะให้ค่าความถูกต้องและค่าความจำเพาะในการจำแนกข้อมูลสูงกว่าวิธีอื่น ๆ ส่วนการปรับสมดุลข้อมูลด้วยวิธีผสมผสานระหว่างรูปแบบสุ่มเกินและสุ่มลดจะให้ค่าความไวในการจำแนกข้อมูลสูงกว่าวิธีอื่น ๆ

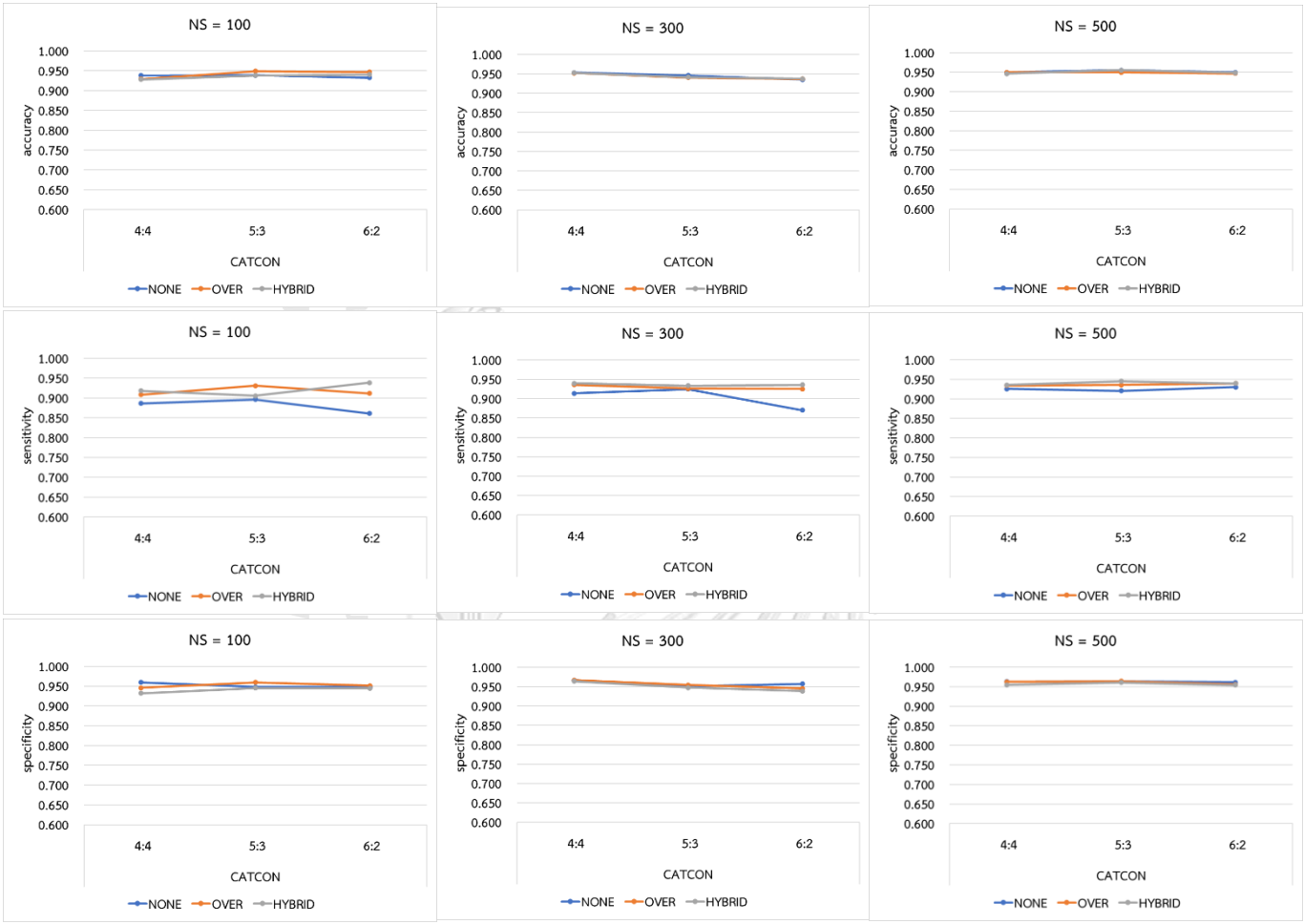
ทั้งนี้เมื่อจำนวนตัวแปรระหว่างกลุ่มตัวแปรจัดประเภทต่อกลุ่มตัวแปรต่อเนื่องมีจำนวนแตกต่างกัน (5:3 และ 6:2) การปรับสมดุลข้อมูลด้วยวิธี random oversampling จะให้ค่าความถูกต้องและค่าความจำเพาะสูงกว่าวิธีอื่น ๆ ในขณะที่การปรับสมดุลข้อมูลด้วยวิธีผสมผสานระหว่างรูปแบบสุ่มเกินและสุ่มลดจะมีค่าความไวในการจำแนกข้อมูลสูงกว่าวิธีอื่น ๆ ในกรณีที่จำนวนตัวแปรระหว่างกลุ่มตัวแปรจัดประเภทต่อกลุ่มตัวแปรต่อเนื่องเท่ากับ 6:2

เมื่อพิจารณากรณีที่มีขนาดตัวอย่างเท่ากับ 300 และ 500 หน่วย พบว่า การไม่ปรับสมดุลข้อมูลและการปรับสมดุลข้อมูลทั้งสองวิธีมีค่าความถูกต้องและค่าความจำเพาะในการจำแนกข้อมูลไม่แตกต่างกันทั้งในกรณีที่จำนวนตัวแปรระหว่างกลุ่มตัวแปรจัดประเภทต่อกลุ่มตัวแปรต่อเนื่องมีค่าเท่ากับ 4:4 5:3 และ 6:2 แต่เมื่อพิจารณาค่าความไวของการจำแนกข้อมูลพบว่า การปรับสมดุลข้อมูลด้วยวิธีผสมผสานระหว่างรูปแบบสุ่มเกินและสุ่มลดจะมีค่าความไวในการจำแนกข้อมูลสูงกว่าการไม่ปรับสมดุลข้อมูล แต่ไม่แตกต่างจากวิธี random oversampling ดังตารางที่ 4.6 และรูปที่ 4.5 และสามารถสรุปแนวทางการปรับสมดุลข้อมูลภายใต้เงื่อนไขด้านขนาดตัวอย่างและจำนวนตัวแปรระหว่างกลุ่มตัวแปรจัดประเภทต่อกลุ่มตัวแปรต่อเนื่องได้ดังตารางที่ 4.7

ตารางที่ 4.6 แสดงผลการเปรียบเทียบค่าเฉลี่ยแต่ละวิธีการปรับสมดุลข้อมูลกับเงื่อนไข จำนวนตัวแปรระหว่างกลุ่มตัวแปรจัดประเภทต่อกลุ่มตัวแปรต่อเนื่อง และขนาดตัวอย่าง

NS	BL	จำนวนตัวแปรระหว่างกลุ่มตัวแปรจัดประเภทต่อกลุ่มตัวแปรต่อเนื่อง																	
		ค่าความถูกต้อง						ค่าความไว						ค่าความจำเพาะ					
		4:4		5:3		6:2		4:4		5:3		6:2		4:4		5:3		6:2	
		Mean	SD	Mean	SD	Mean	SD	Mean	SD	Mean	SD	Mean	SD	Mean	SD	Mean	SD	Mean	SD
100	NONE	**0.938**	0.065	0.939	0.039	0.933	0.058	0.886	0.127	0.896	0.115	0.861	0.144	**0.960**	0.030	0.948	0.023	0.948	0.043
	OVER	0.929	0.084	**0.949**	0.043	**0.947**	0.064	0.908	0.100	**0.931**	0.061	0.911	0.101	0.946	0.063	**0.960**	0.037	**0.951**	0.041
	HYBRID	0.928	0.085	0.938	0.057	0.941	0.074	**0.918**	0.085	0.906	0.092	**0.938**	0.075	0.932	0.084	0.945	0.064	0.945	0.059
300	NONE	**0.953**	0.036	**0.946**	0.039	0.935	0.060	0.914	0.093	0.924	0.080	0.870	0.113	**0.967**	0.022	0.952	0.032	**0.957**	0.036
	OVER	0.952	0.038	0.940	0.060	0.937	0.062	0.936	0.063	0.926	0.083	0.925	0.082	**0.967**	0.022	**0.954**	0.039	0.946	0.067
	HYBRID	**0.953**	0.040	0.941	0.051	**0.938**	0.064	**0.939**	0.065	**0.934**	0.076	**0.936**	0.065	0.964	0.024	0.947	0.045	0.938	0.075
500	NONE	**0.938**	0.032	0.939	0.027	0.933	0.040	0.886	0.068	0.896	0.085	0.861	0.079	**0.960**	0.026	0.948	0.026	0.948	0.029
	OVER	0.929	0.033	**0.949**	0.037	**0.947**	0.039	0.908	0.063	**0.931**	0.063	0.911	0.056	0.946	0.026	**0.960**	0.028	**0.951**	0.039
	HYBRID	0.928	0.038	0.938	0.029	0.941	0.040	**0.918**	0.056	0.906	0.046	**0.938**	0.062	0.932	0.030	0.945	0.026	0.945	0.032

หมายเหตุ BL = วิธีการปรับสมดุลข้อมูล, NS = ขนาดตัวอย่าง, Mean = ค่าเฉลี่ย, SD = ค่าส่วนเบี่ยงเบนมาตรฐาน



รูปที่ 4.5 แสดงผลการเปรียบเทียบค่าเฉลี่ยแต่ละวิธีการปรับสมดุลข้อมูลกับเงื่อนไขจำนวนตัวแปรระหว่างกลุ่มตัวแปรจัดประเภทต่อกลุ่มตัวแปรต่อเนื่องและขนาดตัวอย่าง

ตารางที่ 4.7 แนวทางการปรับสมดุลข้อมูลภายใต้เงื่อนไขด้านจำนวนตัวแปรระหว่างกลุ่มตัวแปรจัดประเภทต่อกลุ่มตัวแปรต่อเนื่อง และขนาดตัวอย่าง

ขนาดตัวอย่าง	จำนวนตัวแปรระหว่างกลุ่มตัวแปรจัดประเภทต่อกลุ่มตัวแปรต่อเนื่อง		
	4:4	5:3	6:2
100	NONE	OVER	OVER
300	ANY	ANY	HYBRID
500	ANY	ANY	HYBRID

2.2 วิธีการปรับสมดุลข้อมูล เทคนิคการจำแนกข้อมูล และขนาดตัวอย่าง

ผลการวิเคราะห์ข้อมูลพบปฏิสัมพันธ์สามทางระหว่างวิธีการปรับสมดุลข้อมูล เทคนิคการจำแนกข้อมูล และขนาดตัวอย่าง ในกรณีที่ขนาดตัวอย่างเท่ากับ 100 หน่วย และจำแนกข้อมูลด้วยวิธีเคเนียร์เรสเนเบอร์ พบว่า การปรับสมดุลข้อมูลด้วยวิธี random oversampling และการไม่ปรับสมดุลข้อมูลจะให้ค่าความถูกต้องและค่าความจำเพาะในการจำแนกข้อมูลสูงกว่าวิธี random oversampling แต่เมื่อพิจารณาการจำแนกข้อมูลด้วยวิธีการถดถอยโลจิสติก พบว่า การไม่ปรับสมดุลข้อมูลจะให้ค่าความถูกต้องและค่าความจำเพาะในการจำแนกข้อมูลสูงกว่าวิธีอื่น ๆ ในขณะที่การปรับสมดุลข้อมูลด้วยวิธีแรนดอมฟอร์เรสและวิธีซัพพอร์ตเวกเตอร์แมชชีน เมื่อปรับสมดุลข้อมูลด้วยวิธี random oversampling จะมีค่าความถูกต้องและค่าความจำเพาะในการจำแนกข้อมูลสูงกว่าวิธีอื่น ๆ นอกจากนี้เมื่อพิจารณาการจำแนกข้อมูลด้วยวิธีเคเนียร์เรสเนเบอร์ วิธีการถดถอยโลจิสติก และวิธีซัพพอร์ตเวกเตอร์แมชชีน และปรับสมดุลข้อมูลด้วยวิธีผสมผสานระหว่างรูปแบบสุ่มเกินและสุ่มลดจะมีค่าความไวสูงกว่าวิธีอื่น ๆ แต่การจำแนกข้อมูลด้วยวิธีแรนดอมฟอร์เรส เมื่อปรับสมดุลด้วยวิธี random oversampling จะให้ค่าความไวสูงกว่าวิธีอื่น ๆ

เมื่อพิจารณากรณีที่มีขนาดตัวอย่างเท่ากับ 300 หน่วย พบว่า การจำแนกข้อมูลด้วยวิธีเคเนียร์เรสเนเบอร์และวิธีการถดถอยโลจิสติก เมื่อปรับสมดุลข้อมูลด้วยวิธี random oversampling จะมีค่าถูกต้องในการจำแนกข้อมูลสูงกว่าวิธีอื่น ๆ ในขณะเดียวกันการปรับสมดุลข้อมูลด้วยวิธีผสมผสานระหว่างรูปแบบสุ่มเกินและสุ่มลดจะมีค่าความไวในการจำแนกข้อมูลสูงกว่าวิธีอื่น ๆ แต่การจำแนกข้อมูลด้วยวิธีเคเนียร์เรสเนเบอร์ เมื่อปรับสมดุลข้อมูลด้วยวิธี random oversampling จะมีค่าความจำเพาะในการจำแนกข้อมูลสูงกว่าวิธีอื่น ๆ และการจำแนกข้อมูลด้วยวิธีการถดถอยโลจิสติก ร่วมกับการไม่ปรับสมดุลจะให้ค่าความจำเพาะในการจำแนกข้อมูลสูงกว่าวิธีอื่น ๆ

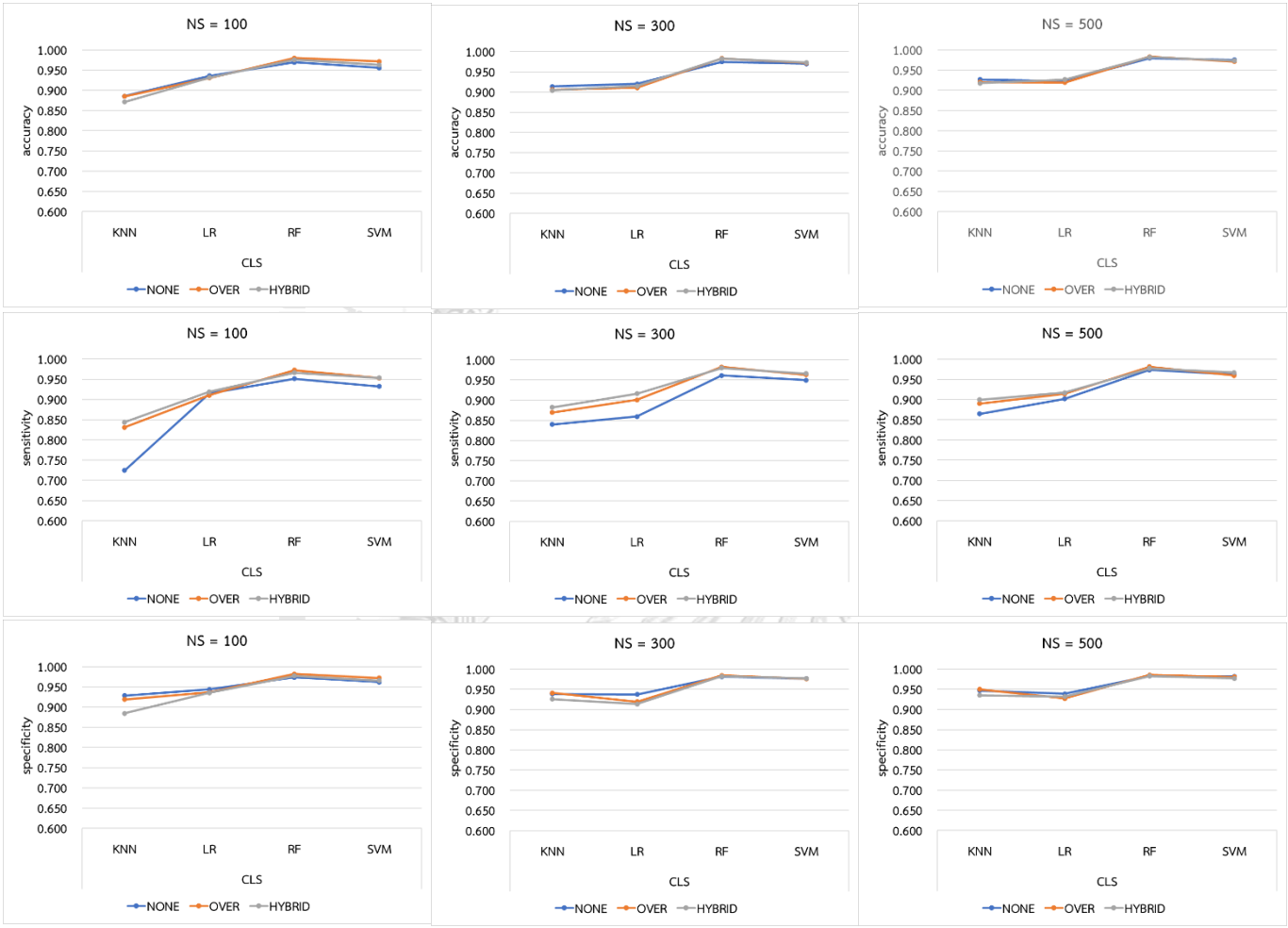
เมื่อจำแนกข้อมูลด้วยวิธีแรนดอมฟอร์เรส การปรับสมดุลด้วยวิธี random oversampling และวิธีผสมผสานระหว่างรูปแบบสุ่มเกินและสุ่มลดจะให้ค่าความถูกต้องในการจำแนกข้อมูลไม่แตกต่างกัน ในขณะที่การปรับสมดุลด้วยวิธี random oversampling จะให้ค่าความไวและความจำเพาะในการจำแนกข้อมูลสูงกว่าวิธีอื่น ๆ และเมื่อจำแนกข้อมูลด้วยวิธีซัพพอร์ตเวกเตอร์แมชชีน การปรับสมดุลด้วยวิธีผสมผสานระหว่างรูปแบบสุ่มเกินและสุ่มลดและการไม่ปรับสมดุลข้อมูลจะให้ค่าความจำเพาะในการจำแนกข้อมูลไม่แตกต่างกัน ในขณะที่การปรับสมดุลด้วยวิธีผสมผสานระหว่างรูปแบบสุ่มเกินและสุ่มลดจะให้ค่าความถูกต้องและความไวในการจำแนกข้อมูลสูงกว่าวิธีอื่น ๆ

เมื่อพิจารณากรณีที่มีขนาดตัวอย่างเท่ากับ 500 หน่วย พบว่า การจำแนกข้อมูลด้วยวิธีเคเนียร์เรสเนเบอร์และวิธีการถดถอยโลจิสติก การไม่ปรับสมดุลข้อมูลจะให้ค่าความถูกต้องในการจำแนกข้อมูลสูงกว่าวิธีอื่น ๆ การปรับสมดุลด้วยวิธี random oversampling จะให้ค่าความจำเพาะในการจำแนกข้อมูลสูงกว่าวิธีอื่น ๆ ในขณะที่การปรับสมดุลข้อมูลด้วยวิธีผสมผสานระหว่างรูปแบบสุ่มเกินและสุ่มลดร่วมกับการจำแนกข้อมูลด้วยวิธีเคเนียร์เรสเนเบอร์ วิธีการถดถอยโลจิสติก และวิธีซัพพอร์ตเวกเตอร์แมชชีนจะให้ค่าความไวในการจำแนกข้อมูลสูงกว่าวิธีปรับสมดุลข้อมูลอื่น ๆ ในขณะที่การจำแนกข้อมูลด้วยวิธีแรนดอมฟอร์เรส การปรับสมดุลด้วยวิธี random oversampling จะให้ค่าความถูกต้อง ค่าความไวและความจำเพาะในการจำแนกข้อมูลสูงกว่าวิธีอื่น ๆ และการจำแนกข้อมูลด้วยวิธีซัพพอร์ตเวกเตอร์แมชชีน การไม่ปรับสมดุลข้อมูลจะมีค่าความถูกต้องและค่าความจำเพาะในการจำแนกข้อมูลสูงกว่าวิธีอื่น ๆ ดังตารางที่ 4.8 และรูปที่ 4.6 และสามารถสรุปแนวทางการปรับสมดุลข้อมูลภายใต้เงื่อนไขด้านขนาดตัวอย่างและเทคนิคการจำแนกข้อมูลได้ดังตารางที่ 4.6

ตารางที่ 4.8 ผลการเปรียบเทียบค่าเฉลี่ยแต่ละวิธีการปรับสมดุลข้อมูลกับเงื่อนไขขนาดตัวอย่างและเทคนิคการจำแนกข้อมูล

NS	BL	เทคนิคการจำแนกข้อมูล																							
		ค่าความถูกต้อง								ค่าความไว								ค่าความจำเพาะ							
		KNN		LR		RF		SVM		KNN		LR		RF		SVM		KNN		LR		RF		SVM	
		Mean	SD	Mean	SD	Mean	SD	Mean	SD	Mean	SD	Mean	SD	Mean	SD	Mean	SD	Mean	SD	Mean	SD	Mean	SD	Mean	SD
100	NONE	**0.885**	0.082	**0.936**	0.031	0.969	0.016	0.955	0.016	0.725	0.157	0.916	0.059	0.952	0.039	0.933	0.065	**0.929**	0.036	**0.944**	0.025	0.973	0.017	0.962	0.033
	OVER	**0.885**	0.088	0.931	0.063	**0.980**	0.014	**0.971**	0.016	0.831	0.118	0.911	0.072	**0.973**	0.018	0.953	0.032	0.919	0.044	0.936	0.066	**0.982**	0.013	**0.972**	0.020
	HYBRID	0.872	0.101	0.932	0.063	0.976	0.017	0.963	0.021	**0.844**	0.119	**0.919**	0.070	0.966	0.018	**0.954**	0.022	0.884	0.097	0.936	0.064	0.977	0.017	0.966	0.024
300	NONE	**0.914**	0.043	**0.920**	0.057	0.975	0.018	0.970	0.016	0.840	0.117	0.860	0.110	0.961	0.022	0.950	0.037	0.939	0.027	**0.938**	0.032	0.981	0.013	**0.977**	0.016
	OVER	0.906	0.054	0.911	0.061	**0.983**	0.012	0.972	0.017	0.869	0.092	0.901	0.079	**0.983**	0.014	0.963	0.020	**0.941**	0.036	0.920	0.067	**0.984**	0.012	0.976	0.016
	HYBRID	0.904	0.057	0.915	0.052	**0.983**	0.013	**0.973**	0.016	**0.882**	0.091	**0.916**	0.063	0.980	0.015	**0.966**	0.018	0.926	0.052	0.914	0.068	0.983	0.013	**0.977**	0.016
500	NONE	**0.926**	0.031	0.924	0.017	0.980	0.014	**0.976**	0.016	0.864	0.107	0.902	0.066	0.973	0.016	0.962	0.020	0.947	0.024	**0.939**	0.020	0.983	0.012	**0.982**	0.013
	OVER	0.920	0.016	0.919	0.036	**0.983**	0.012	0.971	0.015	0.890	0.088	0.914	0.035	**0.981**	0.015	0.960	0.016	**0.950**	0.020	0.927	0.033	**0.985**	0.011	0.981	0.014
	HYBRID	0.918	0.028	**0.927**	0.030	0.982	0.013	0.973	0.016	**0.900**	0.071	**0.917**	0.049	0.977	0.016	**0.967**	0.017	0.935	0.024	0.932	0.020	0.983	0.012	0.977	0.015

หมายเหตุ BL = วิธีการปรับสมดุลข้อมูล, NS = ขนาดตัวอย่าง, LR = การจำแนกข้อมูลด้วยการถดถอยโลจิสติก, RF = การจำแนกข้อมูลด้วยแรนดอมฟอร์เรส, SVM = การจำแนกข้อมูลด้วยซัพพอร์ตเวกเตอร์แมชชีน, KNN การจำแนกข้อมูลด้วยวิธีเคเนียร์เรสเนเบอร์, Mean = ค่าเฉลี่ย, SD = ค่าส่วนเบี่ยงเบนมาตรฐาน



รูปที่ 4.6 ผลการเปรียบเทียบค่าเฉลี่ยแต่ละวิธีการปรับสมดุลข้อมูลกับเงื่อนไขขนาดตัวอย่างและเทคนิคการจำแนกข้อมูล

ตารางที่ 4.9 แนวทางการปรับสมดุลข้อมูลภายใต้เงื่อนไขด้านเทคนิคการจำแนกข้อมูลและขนาดตัวอย่าง

NS	เทคนิคการจำแนกข้อมูล			
	KNN	LR	RF	SVM
100	OVER	NONE	OVER/ HYBRID	OVER
300	OVER	HYBRID	OVER/ HYBRID	ANY
500	OVER	HYBRID	OVER	ANY

หมายเหตุ NS = ขนาดตัวอย่าง, LR = การจำแนกข้อมูลด้วยการถดถอยโลจิสติก, RF = การจำแนกข้อมูลด้วยแรนดอมฟอร์เรส, SVM = การจำแนกข้อมูลด้วยซัพพอร์ตเวกเตอร์แมชชีน, KNN การจำแนกข้อมูลด้วยวิธีเคเนียร์เรสเนเบอร์, Mean = ค่าเฉลี่ย, SD = ค่าส่วนเบี่ยงเบนมาตรฐาน

2.3 วิธีการปรับสมดุลข้อมูล ร้อยละของจำนวนข้อมูลระหว่างข้อมูลกลุ่มหลักต่อข้อมูลกลุ่มรอง และเทคนิคการจำแนกข้อมูล

ผลการวิเคราะห์ข้อมูลพบปฏิสัมพันธ์สามทางระหว่างวิธีการปรับสมดุลข้อมูล ร้อยละของจำนวนข้อมูลระหว่างข้อมูลกลุ่มหลักต่อข้อมูลกลุ่มรอง และเทคนิคการจำแนกข้อมูล ในกรณีที่ร้อยละของจำนวนข้อมูลระหว่างข้อมูลกลุ่มหลักต่อข้อมูลกลุ่มรอง เท่ากับ 60:40 การจำแนกข้อมูลด้วยวิธีเคเนียร์เรสเนเบอร์และวิธีการถดถอยโลจิสติก และไม่ปรับสมดุลข้อมูลจะมีค่าความถูกต้องในการจำแนกข้อมูลสูงกว่าวิธีอื่น ๆ ในขณะที่การจำแนกข้อมูลด้วยวิธีเคเนียร์เรสเนเบอร์และปรับสมดุลข้อมูลด้วยวิธี random oversampling จะมีค่าความจำเพาะในการจำแนกข้อมูลสูงกว่าวิธีอื่น ๆ แต่การไม่ปรับสมดุลข้อมูลและจำแนกข้อมูลด้วยการถดถอยโลจิสติกจะมีค่าความจำเพาะในการจำแนกข้อมูลสูงกว่าวิธีอื่น ๆ

ในกรณีที่ร้อยละของจำนวนข้อมูลระหว่างข้อมูลกลุ่มหลักต่อข้อมูลกลุ่มรอง เท่ากับ 70:30 พบว่า เมื่อการจำแนกข้อมูลด้วยวิธีเคเนียร์เรสเนเบอร์ การไม่ปรับสมดุลข้อมูลและการปรับสมดุลข้อมูลด้วยวิธี random oversampling มีค่าความถูกต้องในการจำแนกข้อมูลไม่แตกต่างกัน ในขณะที่การจำแนกข้อมูลด้วยวิธีการถดถอยโลจิสติก การไม่ปรับสมดุลข้อมูลและการปรับสมดุลข้อมูลด้วยวิธีผสมผสานระหว่างรูปแบบสุ่มเกินและสุ่มลดมีค่าความถูกต้องในการจำแนกข้อมูลไม่แตกต่างกัน

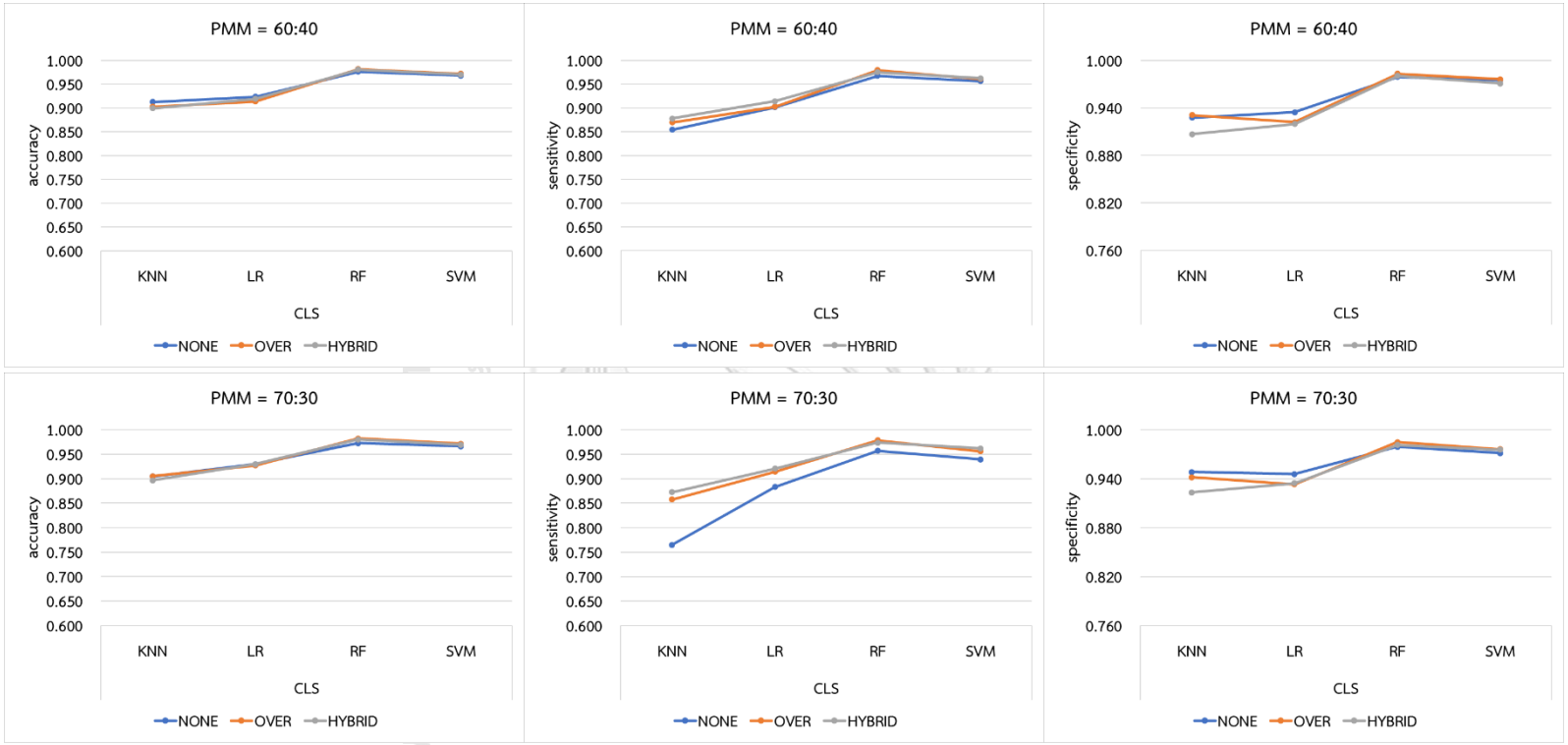
ทั้งนี้เมื่อร้อยละของจำนวนข้อมูลระหว่างข้อมูลกลุ่มหลักต่อข้อมูลกลุ่มรองเท่ากับ 60:40 และ 70:30 พบว่า การจำแนกข้อมูลด้วยวิธีการถดถอยโลจิสติก เมื่อไม่ปรับสมดุลข้อมูลจะมี

ค่าความจำเพาะในการจำแนกข้อมูลสูงกว่าวิธีอื่น ๆ ทั้งนี้การจำแนกข้อมูลด้วยวิธีแรนดอมฟอร์เรสและวิธีซัพพอร์ตเวกเตอร์แมชชีนและปรับสมดุลข้อมูลด้วยวิธี random oversampling จะมีค่าความถูกต้องและค่าความจำเพาะในการจำแนกข้อมูลสูงกว่าวิธีอื่น ๆ ในขณะเดียวกันการจำแนกข้อมูลด้วยวิธีเคเนียร์เรสเนเบอร์ วิธีการถดถอยโลจิสติก และวิธีซัพพอร์ตเวกเตอร์แมชชีนและการปรับสมดุลข้อมูลด้วยวิธีผสมผสานระหว่างรูปแบบสุ่มเกินและสุ่มลดจะมีค่าความไวในการจำแนกข้อมูลสูงกว่าวิธีอื่น ๆ แต่เมื่อปรับสมดุลข้อมูลด้วยวิธี random oversampling และจำแนกข้อมูลด้วยวิธีแรนดอมฟอร์เรสจะมีค่าความไวในการจำแนกข้อมูลสูงกว่าวิธีอื่น ๆ ดังตารางที่ 4.10 และรูปที่ 4.7 และสามารถสรุปแนวทางการปรับสมดุลข้อมูลภายใต้เงื่อนไขด้านขนาดตัวอย่างและจำนวนตัวแปรระหว่างกลุ่มตัวแปรจัดประเภทต่อกลุ่มตัวแปรต่อเนื่องได้ดังตารางที่ 4.11

ตารางที่ 4.10 แสดงผลการเปรียบเทียบค่าเฉลี่ยแต่ละวิธีการปรับสมดุลข้อมูลกับเงื่อนไขร้อยละของจำนวนข้อมูลระหว่างข้อมูลกลุ่มหลักต่อข้อมูลกลุ่มรอง และเทคนิคการจำแนกข้อมูล

PMM	BL	เทคนิคการจำแนกข้อมูล																							
		ค่าความถูกต้อง								ค่าความไว								ค่าความจำเพาะ							
		KNN		LR		RF		SVM		KNN		LR		RF		SVM		KNN		LR		RF		SVM	
		Mean	SD	Mean	SD	Mean	SD	Mean	SD	Mean	SD	Mean	SD	Mean	SD	Mean	SD	Mean	SD	Mean	SD	Mean	SD	Mean	SD
60:40	NONE	**0.913**	0.055	**0.924**	0.038	0.977	0.015	0.968	0.019	0.854	0.112	0.901	0.075	0.967	0.017	0.956	0.037	0.928	0.031	**0.935**	0.027	0.979	0.014	0.975	0.016
	OVER	0.903	0.063	0.913	0.066	**0.982**	0.012	**0.972**	0.016	0.869	0.097	0.903	0.074	**0.979**	0.015	0.961	0.017	**0.931**	0.042	0.922	0.066	**0.983**	0.012	**0.976**	0.015
	HYBRID	0.899	0.063	0.919	0.053	0.981	0.014	0.970	0.017	**0.878**	0.096	**0.914**	0.064	0.975	0.015	**0.963**	0.019	0.907	0.072	0.920	0.061	0.981	0.013	0.971	0.017
70:30	NONE	**0.905**	0.062	**0.930**	0.040	0.972	0.018	0.966	0.016	0.765	0.154	0.884	0.093	0.957	0.037	0.940	0.052	**0.948**	0.026	**0.946**	0.025	0.979	0.016	0.971	0.030
	OVER	**0.905**	0.061	0.927	0.041	**0.982**	0.014	**0.971**	0.016	0.858	0.108	0.915	0.054	**0.978**	0.017	0.956	0.030	0.942	0.031	0.933	0.047	**0.985**	0.013	**0.976**	0.019
	HYBRID	0.897	0.078	**0.930**	0.047	0.980	0.016	0.969	0.020	**0.872**	0.100	**0.921**	0.058	0.974	0.019	**0.962**	0.021	0.923	0.064	0.935	0.048	0.981	0.015	**0.976**	0.021

หมายเหตุ BL = วิธีการปรับสมดุลข้อมูล, PMM = ร้อยละของจำนวนข้อมูลระหว่างข้อมูลกลุ่มหลักต่อข้อมูลกลุ่มรอง, LR = การจำแนกข้อมูลด้วยการถดถอยโลจิสติก, RF = การจำแนกข้อมูลด้วยแรนดอมฟอร์เรส, SVM = การจำแนกข้อมูลด้วยซัพพอร์ตเวกเตอร์แมชชีน, KNN การจำแนกข้อมูลด้วยวิธีเคเนียร์เรสเนเบอร์, Mean = ค่าเฉลี่ย, SD = ค่าส่วนเบี่ยงเบนมาตรฐาน



รูปที่ 4.7 แสดงผลการเปรียบเทียบค่าเฉลี่ยแต่ละวิธีการปรับสมดุลข้อมูลกับเงื่อนไขร้อยละของจำนวนข้อมูลระหว่างข้อมูลกลุ่มหลักต่อข้อมูลกลุ่มรอง และเทคนิคการจำแนกข้อมูล

ตารางที่ 4.11 แนวทางการปรับสมดุลข้อมูลภายใต้เงื่อนไขด้านร้อยละของจำนวนข้อมูลระหว่างข้อมูลกลุ่มหลักต่อข้อมูลกลุ่มรอง และเทคนิคการจำแนกข้อมูล

PMM	เทคนิคการจำแนกข้อมูล			
	KNN	LR	RF	SVM
60:40	OVER	NONE	OVER	OVER
70:30	NONE	HYBRID	OVER	OVER/ HYBRID

หมายเหตุ PMM = ร้อยละของจำนวนข้อมูลระหว่างข้อมูลกลุ่มหลักต่อข้อมูลกลุ่มรอง, LR = การจำแนกข้อมูลด้วยการถดถอยโลจิสติก, RF = การจำแนกข้อมูลด้วยแรนดอมฟอร์เรส, SVM = การจำแนกข้อมูลด้วยซัพพอร์ตเวกเตอร์แมชชีน, KNN การจำแนกข้อมูลด้วยวิธีเคเนียร์เรสเนเบอร์, Mean = ค่าเฉลี่ย, SD = ค่าส่วนเบี่ยงเบนมาตรฐาน

บทที่ 5
สรุปผลการวิจัย อภิปรายผล และข้อเสนอแนะ

การวิจัยนี้มีวัตถุประสงค์เพื่อศึกษาปฏิสัมพันธ์ของวิธีการปรับสมดุลข้อมูล (ไม่ปรับสมดุล วิธี random oversampling และวิธีผสมผสานระหว่างรูปแบบสุ่มเกินและสุ่มลด) กับเงื่อนไขขนาดตัวอย่าง (100, 300 และ 500) เทคนิคการจำแนก (การถดถอยโลจิสติก แรนดอมฟอร์เรส ซัพพอร์ตเวกเตอร์แมชชีน และวิธีเคเนียร์เรสเนเบอร์) จำนวนตัวแปรระหว่างกลุ่มตัวแปรจัดประเภทต่อกลุ่มตัวแปรต่อเนื่อง (4:4, 5:3 และ 6:2) อัตราออด และร้อยละของจำนวนข้อมูลระหว่างข้อมูลกลุ่มหลักต่อข้อมูลกลุ่มรอง (60:40 และ 70:30) ที่มีต่อประสิทธิภาพของการจำแนกกลุ่ม (ค่าความถูกต้อง ค่าความไว และค่าความจำเพาะ) โดยใช้การจำลองข้อมูลด้วยวิธีการมอนติคาร์โล (Monte Carlo simulation) ศึกษาภายใต้สถานการณ์ต่าง ๆ ที่กำหนดขึ้น

เมื่อพิจารณาผลการวิเคราะห์มีประเด็นที่น่าสนใจและนำไปสู่การสรุปและอภิปรายผล โดยมีรายละเอียด ดังนี้

สรุปผลการวิจัย

การศึกษาปฏิสัมพันธ์ระหว่างวิธีการปรับสมดุลข้อมูลกับเงื่อนไขต่าง ๆ ได้แก่ ขนาดตัวอย่าง เทคนิคการจำแนกข้อมูล จำนวนตัวแปรระหว่างกลุ่มตัวแปรจัดประเภทต่อกลุ่มตัวแปรต่อเนื่อง อัตราออด และร้อยละของจำนวนข้อมูลระหว่างข้อมูลกลุ่มหลักต่อข้อมูลกลุ่มรอง ที่มีต่อประสิทธิภาพของการจำแนกกลุ่ม พบว่า วิธีการปรับสมดุลข้อมูลมีปฏิสัมพันธ์แบบสองทางกับจำนวนตัวแปรระหว่างกลุ่มตัวแปรจัดประเภทต่อกลุ่มตัวแปรต่อเนื่อง ร้อยละของจำนวนข้อมูลระหว่างข้อมูลกลุ่มหลักต่อข้อมูลกลุ่มรอง อัตราออด และเทคนิคการจำแนกข้อมูล อย่างมีนัยสำคัญทางสถิติที่ระดับ .05 และมีปฏิสัมพันธ์แบบสามทางกับเงื่อนไขต่าง ๆ ได้แก่ (1) วิธีการปรับสมดุลข้อมูล*ขนาดตัวอย่าง* จำนวนตัวแปรระหว่างกลุ่มตัวแปรจัดประเภทต่อกลุ่มตัวแปรต่อเนื่อง (2) วิธีการปรับสมดุลข้อมูล* ขนาดตัวอย่าง*เทคนิคการจำแนกข้อมูล และ (3) วิธีการปรับสมดุลข้อมูล*ร้อยละของจำนวนข้อมูลระหว่างข้อมูลกลุ่มหลักต่อข้อมูลกลุ่มรอง*เทคนิคการจำแนกข้อมูล อย่างมีนัยสำคัญทางสถิติที่ระดับ .05

ผลการวิเคราะห์รายละเอียดของปฏิสัมพันธ์แบบสองทางระหว่างวิธีการปรับสมดุลข้อมูลกับจำนวนตัวแปรระหว่างกลุ่มตัวแปรจัดประเภทต่อกลุ่มตัวแปรต่อเนื่อง พบว่า เมื่อจำนวนตัวแปร

ระหว่างกลุ่มตัวแปรจัดประเภทต่อกลุ่มตัวแปรต่อเนื่องเท่ากับ 4:4 จะไม่ปรับสมดุลข้อมูล แต่หากจำนวนตัวแปรระหว่างกลุ่มตัวแปรจัดประเภทต่อกลุ่มตัวแปรต่อเนื่องเท่ากับ 5:3 และ 6:2 จะปรับสมดุลข้อมูลด้วยวิธี random oversampling จึงจะมีค่าประสิทธิภาพในการจำแนกข้อมูลสูงที่สุด

ผลการวิเคราะห์รายละเอียดของปฏิสัมพันธ์แบบสองทางระหว่างวิธีการปรับสมดุลข้อมูลกับร้อยละของจำนวนข้อมูลระหว่างข้อมูลกลุ่มหลักต่อข้อมูลกลุ่มรอง พบว่า เมื่อร้อยละของจำนวนข้อมูลระหว่างข้อมูลกลุ่มหลักต่อข้อมูลกลุ่มรองเท่ากับ 60:40 จะปรับสมดุลข้อมูลด้วยวิธีผสมผสานระหว่างรูปแบบสุ่มเกินและสุ่มลด แต่หากร้อยละของจำนวนข้อมูลระหว่างข้อมูลกลุ่มหลักต่อข้อมูลกลุ่มรองเท่ากับ 70:30 จะปรับสมดุลข้อมูลด้วยวิธี random oversampling จึงจะมีค่าประสิทธิภาพในการจำแนกข้อมูลสูงที่สุด

ผลการวิเคราะห์รายละเอียดของปฏิสัมพันธ์แบบสองทางระหว่างวิธีการปรับสมดุลข้อมูลกับอัตราออด พบว่า กรณีตัวแปรอิสระเป็นตัวแปรจัดประเภทและตัวแปรต่อเนื่องมีอัตราออดอยู่ในระดับสูง (ช่วง [2,3)) และกรณีตัวแปรอิสระเป็นตัวแปรจัดประเภทมีอัตราออดอยู่ในระดับสูง (ช่วง [2,3)) และตัวแปรอิสระเป็นตัวแปรต่อเนื่องมีอัตราออดอยู่ในระดับต่ำ (ช่วง [1,2)) รวมถึงกรณีตัวแปรอิสระเป็นตัวแปรจัดประเภทมีอัตราออดอยู่ในระดับต่ำ (ช่วง [1,2)) และตัวแปรอิสระเป็นตัวแปรต่อเนื่องมีอัตราออดอยู่ในระดับสูง (ช่วง [2,3)) จะปรับสมดุลข้อมูลด้วยวิธี random oversampling แต่กรณีที่ตัวแปรอิสระเป็นตัวแปรจัดประเภทและตัวแปรต่อเนื่องมีอัตราออดอยู่ในระดับต่ำ (ช่วง [1,2)) จะปรับสมดุลข้อมูลด้วยวิธีผสมผสานระหว่างรูปแบบสุ่มเกินและสุ่มลด จึงจะมีค่าประสิทธิภาพในการจำแนกข้อมูลสูงที่สุด

ผลการวิเคราะห์รายละเอียดของปฏิสัมพันธ์แบบสองทางระหว่างวิธีการปรับสมดุลข้อมูลกับเทคนิคการจำแนกข้อมูล พบว่า เมื่อจำแนกข้อมูลด้วยวิธีเคเนียร์เรสเนเบอร์ วิธีแรนดอมฟอร์เรส และวิธีซัพพอร์ตเวกเตอร์แมชชีน จะปรับสมดุลข้อมูลด้วยวิธี random oversampling แต่หากจำแนกข้อมูลด้วยวิธีการถดถอยโลจิสติก จะไม่ปรับสมดุลข้อมูล จึงจะมีค่าประสิทธิภาพในการจำแนกข้อมูลสูงที่สุด

ผลการวิเคราะห์รายละเอียดของปฏิสัมพันธ์แบบสามทางระหว่างวิธีการปรับสมดุลข้อมูลกับขนาดตัวอย่างและจำนวนตัวแปรระหว่างกลุ่มตัวแปรจัดประเภทต่อกลุ่มตัวแปรต่อเนื่อง พบว่า เมื่อขนาดตัวอย่างเท่ากับ 100 หน่วยและจำนวนตัวแปรระหว่างกลุ่มตัวแปรจัดประเภทต่อกลุ่มตัวแปรต่อเนื่องเท่ากับ 4:4 จะไม่ปรับสมดุลข้อมูล แต่หากจำนวนตัวแปรระหว่างกลุ่มตัวแปร

จัดประเภทต่อกลุ่มตัวแปรต่อเนื่องเท่ากับ 5:3 และ 6:2 จะปรับสมดุลข้อมูลด้วยวิธี random oversampling เมื่อขนาดตัวอย่างเท่ากับ 300 และ 500 หน่วย จำนวนตัวแปรระหว่างกลุ่มตัวแปรจัดประเภทต่อกลุ่มตัวแปรต่อเนื่องเท่ากับ 6:2 จะปรับสมดุลข้อมูลด้วยวิธีผสมผสานระหว่างรูปแบบสุ่มเกินและสุ่มลด นอกจากนี้หากจำนวนตัวแปรระหว่างกลุ่มตัวแปรจัดประเภทต่อกลุ่มตัวแปรต่อเนื่องเท่ากับ 4:4 และ 5:3 จะไม่ปรับสมดุลข้อมูลหรือปรับสมดุลข้อมูลทั้งสองวิธีก็ได้ จึงจะมีค่าประสิทธิภาพในการจำแนกข้อมูลสูงที่สุด

ผลการวิเคราะห์รายละเอียดของปฏิสัมพันธ์แบบสามทางระหว่างวิธีการปรับสมดุลข้อมูลกับขนาดตัวอย่างและเทคนิคการจำแนกข้อมูล พบว่า การจำแนกข้อมูลด้วยวิธีเคเนียร์เรสเนเบอร์ในทุกขนาดตัวอย่าง จะปรับสมดุลข้อมูลด้วยวิธี random oversampling หากจำแนกข้อมูลด้วยวิธีการถดถอยโลจิสติก เมื่อขนาดตัวอย่างเท่ากับ 100 หน่วย จะไม่ปรับสมดุลข้อมูล แต่เมื่อขนาดตัวอย่างเท่ากับ 300 และ 500 หน่วย จะปรับสมดุลข้อมูลด้วยวิธีผสมผสานระหว่างรูปแบบสุ่มเกินและสุ่มลด ในขณะที่การจำแนกข้อมูลด้วยวิธีแรนดอมฟอร์เรส เมื่อขนาดตัวอย่างเท่ากับ 100 และ 300 หน่วย จะปรับสมดุลข้อมูลด้วยวิธี random oversampling หรือวิธีผสมผสานระหว่างรูปแบบสุ่มเกินและสุ่มลด แต่ถ้าขนาดตัวอย่างเท่ากับ 500 หน่วย จะปรับสมดุลข้อมูลด้วยวิธี random oversampling นอกจากนี้เมื่อจำแนกข้อมูลด้วยวิธีซัพพอร์ตเวกเตอร์แมชชีน และมีขนาดตัวอย่างเท่ากับ 100 หน่วย จะไม่ปรับสมดุลข้อมูล และเมื่อขนาดตัวอย่างเท่ากับ 300 และ 500 หน่วย จะไม่ปรับสมดุลข้อมูลหรือปรับสมดุลข้อมูลทั้งสองวิธีก็ได้ จึงจะมีค่าประสิทธิภาพในการจำแนกข้อมูลสูงที่สุด

ผลการวิเคราะห์รายละเอียดของปฏิสัมพันธ์แบบสามทางระหว่างวิธีการปรับสมดุลข้อมูลกับร้อยละของจำนวนข้อมูลระหว่างข้อมูลกลุ่มหลักต่อข้อมูลกลุ่มรองและเทคนิคการจำแนกข้อมูล พบว่า เมื่อร้อยละของจำนวนข้อมูลระหว่างข้อมูลกลุ่มหลักต่อข้อมูลกลุ่มรองเท่ากับ 60:40 และจำแนกข้อมูลด้วยวิธีเคเนียร์เรสเนเบอร์ วิธีแรนดอมฟอร์เรส และวิธีซัพพอร์ตเวกเตอร์แมชชีน จะปรับสมดุลข้อมูลด้วยวิธี random oversampling แต่หากจำแนกข้อมูลด้วยวิธีการถดถอยโลจิสติก จะไม่ปรับสมดุลข้อมูล ในขณะที่ร้อยละของจำนวนข้อมูลระหว่างข้อมูลกลุ่มหลักต่อข้อมูลกลุ่มรองเท่ากับ 70:30 การจำแนกข้อมูลด้วยวิธีเคเนียร์เรสเนเบอร์จะไม่ปรับสมดุลข้อมูล การจำแนกข้อมูลด้วยวิธีการถดถอยโลจิสติกจะปรับสมดุลข้อมูลด้วยวิธีผสมผสานระหว่างรูปแบบสุ่มเกินและสุ่มลด การจำแนกข้อมูลด้วยวิธีแรนดอมฟอร์เรสจะปรับสมดุลข้อมูลด้วยวิธี random oversampling และการจำแนกข้อมูลด้วยวิธีซัพพอร์ตเวกเตอร์แมชชีนจะปรับสมดุลข้อมูลด้วยวิธี random

oversampling หรือวิธีผสมผสานระหว่างรูปแบบสุ่มเกินและสุ่มลด จึงจะมีค่าประสิทธิภาพในการจำแนกข้อมูลสูงที่สุด

อภิปรายผลการวิจัย

การอภิปรายผลการวิจัยการศึกษาปฏิสัมพันธ์ของวิธีการปรับสมดุลข้อมูลกับเงื่อนไขขนาดตัวอย่าง เทคนิคการจำแนกข้อมูล จำนวนตัวแปรระหว่างกลุ่มตัวแปรจัดประเภทต่อกลุ่มตัวแปรต่อเนื่อง อัตราออด และร้อยละของจำนวนข้อมูลระหว่างข้อมูลกลุ่มหลักต่อข้อมูลกลุ่มรอง ที่มีต่อประสิทธิภาพของการจำแนกกลุ่มภายใต้สถานการณ์จำลองที่แตกต่างกัน พบว่า

เมื่อพิจารณาจำนวนตัวแปรระหว่างกลุ่มตัวแปรจัดประเภทต่อกลุ่มตัวแปรต่อเนื่อง หากจำนวนตัวแปรจัดประเภทเพิ่มขึ้นควรปรับสมดุลข้อมูลด้วยวิธี random oversampling หรือวิธีผสมผสานระหว่างรูปแบบสุ่มเกินและสุ่มลดก่อนนำไปวิเคราะห์ สอดคล้องกับงานวิจัยของ Cateni et al. (2014) ที่ปรับสมดุลข้อมูลด้วยวิธีผสมผสานระหว่างรูปแบบสุ่มเกินและสุ่มลดก่อนนำไปจำแนกข้อมูล

เมื่อพิจารณาขนาดตัวอย่าง 100 300 และ 500 หน่วย หากจำแนกข้อมูลด้วยวิธีเคเนียร์เรสเนเบอร์ควรปรับสมดุลข้อมูลด้วยวิธี random oversampling เมื่อจำแนกข้อมูลด้วยวิธีการถดถอยโลจิสติกควรปรับสมดุลข้อมูลด้วยวิธีผสมผสานระหว่างรูปแบบสุ่มเกินและสุ่มลด ถ้าจำแนกข้อมูลด้วยวิธีแรนดอมฟอร์เรสควรปรับสมดุลข้อมูลด้วยวิธี random oversampling หรือวิธีผสมผสานระหว่างรูปแบบสุ่มเกินและสุ่มลด และการจำแนกข้อมูลด้วยวิธีซัพพอร์ตเวกเตอร์แมชชีนควรปรับสมดุลข้อมูลด้วยวิธี random oversampling หรือไม่ปรับสมดุลข้อมูล โดยมีข้อสังเกตว่า เมื่อขนาดตัวอย่างเพิ่มขึ้นควรปรับสมดุลข้อมูลด้วยวิธี random oversampling หรือวิธีผสมผสานระหว่างรูปแบบสุ่มเกินและสุ่มลด และไม่ควรที่จะไม่ปรับสมดุลข้อมูล สอดคล้องกับงานวิจัยของ Zhu et al. (2021) ที่พิจารณาเลือกวิธีที่มีประสิทธิภาพสูงที่สุดในการระบุแหล่งกักเก็บก๊าซจากชั้นหินดินดานคุณภาพสูงโดยใช้การปรับสมดุลข้อมูลด้วยการสุ่มเกินร่วมกับการจำแนกข้อมูลด้วยวิธีแรนดอมฟอร์เรส วิธีการเคเนียร์เรสเนเบอร์ และวิธีซัพพอร์ตเวกเตอร์แมชชีน พบว่าการปรับสมดุลข้อมูลด้วยวิธีการสุ่มเกินร่วมกับการจำแนกข้อมูลด้วยวิธีแรนดอมฟอร์เรสมีประสิทธิภาพสูงที่สุด

เมื่อพิจารณาเทคนิคการจำแนกข้อมูล โดยร้อยละของจำนวนข้อมูลระหว่างข้อมูลกลุ่มหลักต่อข้อมูลกลุ่มรองเท่ากับ 60:40 และจำแนกข้อมูลด้วยวิธีแรนดอมฟอร์เรส วิธีซัพพอร์ตเวกเตอร์แมชชีน และวิธีเคเนียร์เรสเนเบอร์ควรปรับสมดุลข้อมูลด้วยวิธี random oversampling

แต่ถ้าจำแนกข้อมูลด้วยวิธีการถดถอยโลจิสติกไม่ควรปรับสมดุลข้อมูล ในขณะที่ร้อยละของจำนวนข้อมูลระหว่างข้อมูลกลุ่มหลักต่อข้อมูลกลุ่มรองเท่ากับ 70:30 จำแนกข้อมูลด้วยวิธีเคเนียร์เรสเนเบอร์ไม่ควรปรับสมดุลข้อมูล แต่การจำแนกข้อมูลด้ยวิธีการถดถอยโลจิสติก วิธีแรนดอมฟอร์เรส และวิธีซัพพอร์ตเวกเตอร์แมชชีนควรปรับสมดุลข้อมูลก่อนทำการวิเคราะห์ข้อมูล สอดคล้องกับงานวิจัยของ กาญจน์ ณ ศรีธะ และคณะ (2561) ซึ่งพิจารณาเลือกใช้เทคนิคการสุ่มตัวอย่างในการจำแนกข้อมูลที่ไม่สมดุล พบว่าการจำแนกข้อมูลด้วยวิธีแรนดอมฟอร์เรส มีค่าประสิทธิภาพจากการจำแนกข้อมูลที่ดี และงานวิจัยของ Yan Zhu et al. (2020) จำแนกข้อมูลโปรตีน Lysine succinylation ด้วยวิธีแรนดอมฟอรเรส เคเนียร์เรสเนเบอร์ ซัพพอร์ตเวกเตอร์แมชชีน นาอีฟเบย์ และการวิเคราะห์แบบจำแนกประเภท (discriminant analysis) พบว่าการจำแนกข้อมูลด้วยแรนดอมฟอร เรสมีประสิทธิภาพดีที่สุด

ข้อเสนอแนะ

1. ข้อเสนอแนะในการนำผลการวิจัยไปใช้

1. ผลการวิจัยแสดงให้เห็นถึงประสิทธิภาพในการจำแนกข้อมูลที่มีลักษณะไม่สมดุลภายใต้สถานการณ์ต่าง ๆ โดยใช้วิธีการปรับสมดุลข้อมูล 3 วิธี อย่างไรก็ตามผลการวิจัยดังกล่าวเป็นเพียงแนวทางที่ผู้วิเคราะห์ข้อมูลสามารถนำไปใช้วิเคราะห์ข้อมูลที่เก็บรวบรวมได้ตามสถานการณ์ที่เกิดขึ้นจริง ซึ่งผู้วิเคราะห์ข้อมูลควรพิจารณาสภาพของข้อมูลที่จะใช้ในการวิเคราะห์นั้นประกอบการเลือกใช้วิธีการปรับสมดุลข้อมูลตามแนวทางที่เป็นข้อค้นพบจากการวิจัยนี้ด้วย

2. จากข้อค้นพบในการวิจัยที่พบปฏิสัมพันธ์ระหว่างวิธีการปรับสมดุลข้อมูลกับเทคนิคการจำแนกข้อมูลร่วมกับคุณลักษณะของข้อมูล ได้แก่ ขนาดตัวอย่าง และจำนวนตัวแปรระหว่างกลุ่มตัวแปรจัดประเภทต่อกลุ่มตัวแปรต่อเนื่อง ทำให้ผู้วิเคราะห์ข้อมูลสามารถออกแบบการจัดเก็บข้อมูลและการวิเคราะห์ข้อมูลโดยคำนึงถึงขนาดตัวอย่างที่ต้องจัดเก็บและรูปแบบของตัวแปร (ตัวแปรจัดประเภท/ ตัวแปรต่อเนื่อง) อย่างไรก็ตามภายหลังการจัดเก็บข้อมูล หากพบว่าข้อมูลที่จัดเก็บมานั้นมีลักษณะข้อมูลไม่สมดุล อาจเลือกใช้วิธีการปรับสมดุลตามที่งานวิจัยนี้แนะนำให้สอดคล้อง เหมาะสมกับสภาพข้อมูลที่ใช้ในการวิเคราะห์

2. ข้อเสนอแนะในการทำการวิจัยครั้งต่อไป

ค่าเฉลี่ยตัวบ่งชี้ประสิทธิภาพการจำแนกข้อมูลที่ใช้ในการวิเคราะห์ปฏิสัมพันธ์ระหว่างวิธีการปรับสมดุลข้อมูลกับเงื่อนไขต่าง ๆ เป็นค่าเฉลี่ยที่เกิดขึ้นจากการจำแนกข้อมูลด้วยเทคนิคการวิเคราะห์ทั้ง 4 วิธีร่วมกัน จึงทำให้ค่าประสิทธิภาพของเกณฑ์ทั้ง 3 ประเด็น คือ ความถูกต้อง ความไว และความจำเพาะในการจำแนกข้อมูล ภายใต้สถานการณ์ต่าง ๆ มีค่าใกล้เคียงกันสูง ดังนั้นในการวิจัยครั้งต่อไปอาจศึกษาปฏิสัมพันธ์ของวิธีการปรับสมดุลกับเงื่อนไขต่าง ๆ ด้วยการกำหนดให้ใช้เทคนิคการจำแนกข้อมูลวิธีใดวิธีหนึ่งแทนการใช้ค่าเฉลี่ยจากทุกวิธี

บรรณานุกรม

Bach, M., Werner, A., & Palt, M. (2019). The proposal of undersampling method for learning from imbalanced datasets. Procedia Computer Science, 159, 125-134.

Batista, G. E., Monard, M. C., & Bazzan, A. L. (2004). Improving rule induction precision for automated annotation by balancing skewed data sets. International Symposium on Knowledge Exploration in Life Science Informatics,

Cateni, S., Colla, V., & Vannucci, M. (2014). A method for resampling imbalanced datasets 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.

Chawla, N. V., Japkowicz, N., & Kotcz, A. (2004). Special issue on learning from imbalanced data sets. ACM SIGKDD explorations newsletter, 6(1), 1-6.

Chokthanahirun Siraphong, & Hengpraprohm Supoj. (2016). The data classification using Genetic Algorithm with k-length chromosome. Interdisciplinary Research Review, 11(1), 43-49.

Christian, T. M., & Ayub, M. (2014). Exploration of classification using NBTree for predicting students' performance. 2014 international conference on data and software engineering (ICODSE),

Cortes, C., & Vapnik, V. (1995). Support-vector networks. Machine learning, 20, 273-297.

Dubey, R., Zhou, J., Wang, Y., Thompson, P. M., Ye, J., & Initiative, A. s. D. N. (2014). Analysis of sampling techniques for imbalanced data: An n= 648 ADNI study. NeuroImage, 87, 220-241.

Farquad, M. A. H., & Bose, I. (2012). Preprocessing unbalanced data using support vector machine. Decision Support Systems, 53(1), 226-233.

Gameng, H. A., Gerardo, B. D., & Medina, R. P. (2019). A modified adaptive synthetic smote approach in graduation success rate classification. International Journal, 8(6).

Gordon, A. D. (1999). Classification. CRC Press.

Han, H., Wang, W.-Y., & Mao, B.-H. (2005). Borderline-SMOTE: a new over-sampling method in imbalanced data sets learning. International conference on intelligent computing,

Hartshorn, S. (2016). Machine learning with random forests and decision trees: A Visual guide for beginners. Kindle edition.

Hussain, S., Dahan, N. A., Ba-Alwib, F. M., & Ribata, N. (2018). Educational data mining and analysis of students' academic performance using WEKA. Indonesian Journal of Electrical Engineering and Computer Science, 9(2), 447-459.

Kaiyawan, Y. (2012). Principle and using logistic regression analysis for research. RMUTSV Research Journal, 4(1), 1-12.

Kesavaraj, G., & Sukumaran, S. (2013). A study on classification techniques in data mining. 2013 fourth international conference on computing, communications and networking technologies (ICCCNT),

Kittithanusorn, K., & Sa-ing, V. (2021). NEWS CATEGORY CLASSIFICATION WITH MACHINE LEARNING METHOD.

Leo, B. (2001). Random forests. Machine learning, 45, 5-23.

López, V., Fernández, A., Moreno-Torres, J. G., & Herrera, F. (2012). Analysis of preprocessing vs. cost-sensitive learning for imbalanced classification. Open problems on intrinsic data characteristics. Expert Systems with Applications, 39(7), 6585-6608.

Lunardon, N., Menardi, G., & Torelli, N. (2014). ROSE: a package for binary imbalanced learning. R journal, 6(1).

Manee, T., & Pontanya, A. (2022). รูปแบบความสัมพันธ์เชิงสาเหตุของปัจจัยที่ส่งผลต่อภาวะหมดไฟในการเรียนของนักเรียนชั้นมัธยมศึกษาตอนปลาย: CAUSAL RELATIONSHIP MODEL OF FACTORS AFFECTING STUDY-RELATED BURNOUT AMONG HIGH SCHOOL STUDENTS. JOURNAL OF EDUCATION NARESUAN UNIVERSITY, 24(2), 136-147.

Mousa, H., & Maghari, A. (2017). School student's performance prediction using data mining classification. International Journal of Advanced Research in Computer and Communication Engineering, 6(8), 136-141.

Orriols-Puig, A., & Bernadó-Mansilla, E. (2009). Evolutionary rule-based systems for imbalanced data sets. Soft Computing, 13, 213-225.

Phyu, T. N. (2009). Survey of classification techniques in data mining. Proceedings of the international multiconference of engineers and computer scientists,

Pongpisanrat, C. (2022). การศึกษาความเหนื่อยหน่ายในการเรียนของนิสิตมหาวิทยาลัยมหาสารคาม (Retracted). Journal of Educational Measurement Mahasarakham University, 28(2), 95-110.

Prasad, K., Vaidya, R., & Kumar, V. A. (2016). Teacher's performance as a function of occupational stress and coping with reference to CBSE affiliated school teachers in and around Hyderabad: a multinomial regression approach. Psychology, 7(13), 1700-1718.

Qian, Y., Liang, Y., Li, M., Feng, G., & Shi, X. (2014). A resampling ensemble algorithm for classification of imbalance problems. Neurocomputing, 143, 57-67.

RDocumentation. Uniform: The Uniform Distribution. Retrieved from https://www.rdocumentation.org/packages/stats/versions/3.6.2/topics/Uniform

Shi, B., Wang, J., Qi, J., & Cheng, Y. (2015). A novel imbalanced data classification approach based on logistic regression and Fisher discriminant. Mathematical Problems in Engineering, 2015.

Teeranai, S. (2019). Classification of risk attitudes from customer behavior with machine learning. 2019 23rd International Computer Science and Engineering Conference (ICSEC),

Villar, P., Fernández, A., & Herrera, F. (2011). Studying the behavior of a multiobjective genetic algorithm to design fuzzy rule-based classification systems for imbalanced data-sets. 2011 IEEE International Conference on Fuzzy Systems (FUZZ-IEEE 2011),

Zhu, L., Zhou, X., & Zhang, C. (2021). Rapid identification of high-quality marine shale gas reservoirs based on the oversampling method and random forest algorithm. Artificial Intelligence in Geosciences, 2, 76-81.

Zhu, Y., Jia, C., Li, F., & Song, J. (2020). Inspector: a lysine succinylation predictor based on edited nearest-neighbor undersampling and adaptive synthetic oversampling. Analytical biochemistry, 593, 113592.

Zhu, Y., Yan, Y., Zhang, Y., & Zhang, Y. (2020). EHSO: Evolutionary Hybrid Sampling in overlapping scenarios for imbalanced learning. Neurocomputing, 417, 333-346.

เจนวิทย์ วารีบ่อ. (2562). ปรากฏการณ์การลาออกกลางคันและการคงอยู่ของนิสิตระดับปริญญาตรี มหาวิทยาลัยบูรพา โดยการสำรวจ การอธิบาย และการทำนาย [ดุษฎีนิพนธ์ปรัชญาดุษฎีบัณฑิต, มหาวิทยาลัยบูรพา]. ชลบุรี.

เบญจภรณ์ จันทรกองกุล, สุวรรณา รัศมีขวัญ, สุนิสา ริมเจริญ, ภูสิต กุลเกษม, กฤษณะ ชินสาร, อัณณ์นุพันธ์ รอดทุกข์, ปิยนุช วรบุตร และจรรยา อ้นปันส์. (2557). วิธีการที่เหมาะสมสำหรับการแบ่งกลุ่มข้อมูลที่ไม่สมดุลสูง.

ไกรศักดิ์ เกสร. Data Science : วิทยาศาสตร์ข้อมูล. มหาวิทยาลัยนเรศวร.

กองทุนเพื่อความเสมอภาคทางการศึกษา. (2563). โครงการพัฒนาองค์ความรู้และแม่แบบเพื่อการดูแลและสนับสนุนด้านการศึกษาของเด็กบนท้องถนน (Children in Street) ในกรุงเทพมหานคร.

กัลยา วานิชย์บัญชา. (2015). การใช้ SPSS for windows ในการวิเคราะห์ข้อมูล (ฉบับปรับปรุง). ภาควิชาสถิติ คณะพาณิชยศาสตร์และการบัญชี จุฬาลงกรณ์มหาวิทยาลัย.

กาญจน์ ณ ศรีธะ, กิตติศักดิ์ เกิดประสพ และนิตยา เกิดประสพ. (2561). การเปรียบเทียบเทคนิคการสุ่มตัวอย่างเพื่อการจำแนกข้อมูลที่ไม่สมดุล. Journal of Applied Informatics and Technology, 1(1), 20-37.

กาญจน์เขจร ชูชีพ. (2018). การถดถอยโลจีสติก (Logistic Regression). คณะวนศาสตร์ มหาวิทยาลัยเกษตรศาสตร์.

กิตติคุณ สุภาวณิชย์. (2564). การเปรียบเทียบการประมาณค่าความแปรปรวนของการแจกแจงปรกติและปรกติปลอมปน [ปริญญานิพนธ์ปริญญามหาบัณฑิต, สถาบันเทคโนโลยีพระจอมเกล้าเจ้าคุณทหารลาดกระบัง]. กรุงเทพฯ.

กิตติพงศ์ ชมบุญ. (2558). เทคนิคการจำแนกประเภทข้อมูลส่วนน้อยบนข้อมูลไม่สมดุลด้วยวิธีการแบ่งข้อมูล [ดุษฎีนิพนธ์ปรัชญาดุษฎีบัณฑิต, มหาวิทยาลัยเทคโนโลยีสุรนารี]. นครราชสีมา.

กีระชาติ สุขสุทธิ์. (2559). การจำแนกข้อมูลไม่ สมดุลโดยใช้การปรับปรุงข้อมูลร่วมกับการหาค่าพารามิเตอร์ที่เหมาะสมด้วยขั้นตอนวิธีเชิงพันธุกรรมที่มีการเริ่มต้นใหม่ [ดุษฎีนิพนธ์ปรัชญาดุษฎีบัณฑิต, มหาวิทยาลัยเทคโนโลยีสุรนารี]. นครราชสีมา.

จรรยา อ้นปันส์. (2564). บทที่ 5 การแจกแจงความน่าจะเป็นของตัวแปรสุ่มชนิดไม่ต่อเนื่อง. มหาวิทยาลัยบูรพา.

ฉัตรชกรณ์ ระบิล และวิลาสินี จินตลิขิตดี. (2564). ภาวะหมดไฟในการทำงานของบุคลากรในสำนักงานคณะกรรมการการศึกษาขั้นพื้นฐาน. วารสารสวนสุนันทาวิชาการ และวิจัย, 15(2), 60-79.

ชนาธิป หมั่นเพียรสุข และสุพจน์ เฮงพระพรหม. ประสิทธิภาพของฟังก์ชันความเหมือนต่อขั้นตอนวิธีเพื่อนบ้านใกล้ที่สุดเค สำหรับการจำแนกประเภทข้อมูล.

ณัฏฐินี ดีแท้. (2016). ประสิทธิภาพการจัดกลุ่มของวิธีซัพพอร์ตเวกเตอร์แมทชีนและวิธีเคเนียเรสเนเบอร์ เมื่อข้อมูลมีการแจกแจงเบ้. Life Sciences and Environment Journal, 17(1), 44-53.

นรรจชนก กริชฤทธิเศรษฐ์. (2562). กระบวนการกลายเป็นชายขอบของการเข้าไม่ถึงสิทธิการศึกษาขั้นพื้นฐานของเด็กบนท้องถนนในบริเวณสะพานพุทธ [ปริญญาศิลปศาสตร์บัณฑิต, มหาวิทยาลัยศิลปากร]. นครปฐม.

นรารัตน์ เรืองชัยจตุพร. (2562a). บทที่ 3 การแจกแจงความน่าจะเป็นแบบไม่ต่อเนื่อง (Discrete Probability Distributions). มหาวิทยาลัยขอนแก่น.

นรารัตน์ เรืองชัยจตุพร. (2562b). บทที่ 4 การแจกแจงความน่าจะเป็นแบบต่อเนื่อง (Continuous Probability Distributions). มหาวิทยาลัยขอนแก่น.

นิเวศ จิระวิชิตชัย. (2563). แบบจำลองการตรวจสอบการทุจริตสำหรับข้อมูลที่ไม่สมดุลโดยใช้เทคนิคการลดมิติข้อมูลร่วมกับการเรียนรู้ของเครื่อง.

พงศกร ธีรรัศมี. วิธีการหาค่าเคที่เหมาะสมในการจำแนกแบบเคเนียร์เรสเนเบอร์กับข้อมูลทางการแพทย์ [ปริญญานิพนธ์ปริญญามหาบัณฑิต, มหาวิทยาลัยเทคโนโลยีสุรนารี]. นครราชสีมา.

พงษ์ศักดิพล ทาแก้ว. (2563). รูปแบบการติดตามนักเรียน นักศึกษาของวิทยาลัยเทคนิคเดชอุดมเพื่อป้องกันปัญหาการออกกลางคันโดยนำระบบสารสนเทศมาใช้ในการบริหารจัดการ. E-Journal of Education Studies, Burapha University, 2(3), 34-51.

พัชรียา ทองพูล, พิมพ์ชนก จำเรือง, รมย์นลิน บุญฤทธิ์ และสายชล สินสมบูรณ์ทอง. (2562). การเปรียบเทียบประสิทธิภาพในการทำนายผลการปรับความไม่สมดุลของข้อมูลในการจำแนกด้วยเทคนิคการทำเหมืองข้อมูล. Thai Journal of Science and Technology, 8(6), 565-584.

พิชามญชุ์ โตโฉมงาม. (2553). ปัจจัยทางจิตและการสนับสนุนทางสังคมที่สัมพันธ์กับพฤติกรรมการยอมรับ

นวัตกรรมด้านการป้องกันโรคเอดส์ของเด็กเร่ร่อนในศูนย์สร้างโอกาสเด็กกรุงเทพมหานคร. วารสารพฤติกรรมศาสตร์เพื่อการพัฒนา, 2(1), 18-28.

ภาสพิชญ์ ชูใจ. (2557). การเรียนรู้ร่วมกันสำหรับปัญหาการจำแนกข้อมูลไม่สมดุล [ดุษฎีนิพนธ์ปรัชญาดุษฎีบัณฑิต, มหาวิทยาลัยเทคโนโลยีสุรนารี]. นครราชสีมา.

มหาวิทยาลัยราชภัฏบุรีรัมย์. (2563). การแจกแจงของตัวแปรสุ่มต่อเนื่อง (Continuous Distributions). สืบค้นจาก https://dspace.bru.ac.th/xmlui/bitstream/handle/123456789/7124/4.%20%E0%B8%81%E0%B8%B2%E0%B8%A3%E0%B9%81%E0%B8%88%E0%B8%81%E0%B9%81%E0%B8%88%E0%B8%87%E0%B8%82%E0%B8%AD%E0%B8%87%E0%B8%95%E0%B8%B1%E0%B8%A7%E0%B9%81%E0%B8%9

B%E0%B8%A3%E0%B8%AA%E0%B8%B8%E0%B9%88%E0%B8%A1%E0%B8%95%E0%B9%88%E0%B8%AD%E0%B9%80%E0%B8%99%E0%B8%B7%E0%B9%88%E0%B8%AD%E0%B8%87%2063.pdf?sequence=1&isAllowed=y

มิรันตี วรอุไร. (2009). วิถีชีวิตของเด็กเร่ร่อนในกรุงเทพมหานคร [ปริญญานิพนธ์ปริญญามหาบัณฑิต, มหาวิทยาลัยธรรมศาสตร์]. กรุงเทพฯ.

ศิรินทรา เสือพิทักษ์, บุญมี นิสสัยดี และวิภาวรรณ บัวทอง. (2564). การเปรียบเทียบเทคนิคการจำแนกประเภทข้อมูลสำหรับทำนายแนวโน้มการสำเร็จการศึกษาของนักเรียน. PKRU SciTech Journal, 5(2), 42-50.

สมพงษ์ จิตระดับ, สุอังคะ วาทิน และ ดานา โมหะหมัดรักษาผล. (2563, 31 มกราคม). ความรู้สู่ความเข้าใจเด็กบนท้องถนน. มติชนออนไลน์.

สาโรจน์ ขอจ่วนเตี๋ยว. (2559). รายงานการสังเคราะห์งานวิจัย 100 เรื่อง เพื่อลดจำนวนผู้เรียนที่ออกกลางคันของสถานศึกษาอาชีวศึกษาช่างอุตสาหกรรม.

สิริกุล รัตนมณี, เอกวิทย์ โทปุรินทร์ และสมพงษ์ ปั้นหุ่น. (2561). การวิเคราะห์จำแนก ปัจจัยการออกกลางคัน นิสิตระดับปริญญาตรีมหาวิทยาลัยบูรพา. วารสารวิจัยรำไพพรรณี, 12(3), 124-134.

สุกัญญา ทารส. (2562). ปัจจัยจำแนกการออกกลางคันของนิสิตปริญญาตรี มหาวิทยาลัยมหาสารคาม. Journal of Educational Measurement Mahasarakham University, 26(1), 273-287.

สุรวัชร ศรีเปารยะ และสายชล สินสมบูรณ์ทอง. (2560). การเปรียบเทียบประสิทธิภาพวิธีการจำแนกกลุ่มการเป็นโรคไตเรื้อรัง: กรณีศึกษาโรงพยาบาลแห่งหนึ่งในประเทศอินเดีย. Thai Science and Technology Journal, 839-853.

อริสฬา เตหลิ่ม. (2563). การวิเคราะห์ความอยู่รอดจากการออกกลางคันของนิสิตระดับปริญญาตรี มหาวิทยาลัยบูรพา. วารสารวิจัยรำไพพรรณี, 14(3), 72-83.

อัจฉรา แผ้วบาง และสายชล สินสมบูรณ์ทอง. (2020). การปรับความไม่สมดุลของข้อมูลด้วยการจำแนก 5 วิธี. Thai Journal of Science and Technology, 9(4), 418-435.

อิศรา ชอบขาย. (2550). วิถีชีวิตของเด็กเร่ร่อนในเขตเทศบาลนครขอนแก่น [ปริญญานิพนธ์ปริญญามหาบัณฑิต, มหาวิทยาลัยราชภัฏมหาสารคาม]. มหาสารคาม.

ภาคผนวก

ตารางที่ 1 แสดงผลการเปรียบเทียบประสิทธิภาพของเทคนิคการจำแนกข้อมูลที่ถูกปรับสมดุลแต่ละวิธี เมื่อร้อยละของจำนวนข้อมูลระหว่างข้อมูลกลุ่มหลักต่อข้อมูลกลุ่มรอง คือ 60:40 และจำนวนตัวแปรระหว่างกลุ่มตัวแปรจัดประเภทต่อกลุ่มตัวแปรต่อเนื่อง คือ 4:4 โดยจำแนกตามขนาดตัวอย่าง อัตราออด และเทคนิคการปรับสมดุลข้อมูล

N	OR		NONE				OVER				HYBRID			
	CAT	CON	LR	RF	SVM	KNN	LR	RF	SVM	KNN	LR	RF	SVM	KNN
accuracy														
100	L	L	0.953	0.990	0.979	0.893	0.869	0.991	0.978	0.891	0.875	0.989	0.982	0.884
	L	H	0.938	0.985	0.976	0.908	0.824	0.989	0.976	0.906	0.821	0.985	0.983	0.908
	H	L	0.961	0.984	0.980	0.860	0.872	0.994	0.983	0.867	0.872	0.990	0.983	0.851
	H	H	0.945	0.982	0.979	0.892	0.821	0.987	0.973	0.843	0.834	0.987	0.961	0.871
	ค่าเฉลี่ย		0.949	0.985	0.979	0.888	0.847	0.990	0.978	0.877	0.851	0.988	0.977	0.879
300	L	L	0.956	0.992	0.983	0.931	0.953	0.995	0.986	0.923	0.960	0.991	0.985	0.917
	L	H	0.943	0.984	0.983	0.936	0.935	0.990	0.983	0.929	0.939	0.989	0.984	0.921
	H	L	0.955	0.985	0.981	0.889	0.962	0.991	0.981	0.896	0.950	0.992	0.990	0.905
	H	H	0.942	0.980	0.979	0.918	0.950	0.990	0.981	0.926	0.947	0.985	0.984	0.912
	ค่าเฉลี่ย		0.949	0.985	0.982	0.919	0.950	0.992	0.983	0.919	0.949	0.989	0.986	0.914
500	L	L	0.942	0.984	0.983	0.924	0.940	0.993	0.981	0.916	0.938	0.990	0.984	0.917
	L	H	0.923	0.985	0.983	0.941	0.927	0.990	0.981	0.935	0.923	0.989	0.978	0.927
	H	L	0.953	0.980	0.979	0.902	0.956	0.995	0.982	0.905	0.954	0.988	0.979	0.898
	H	H	0.934	0.986	0.985	0.930	0.937	0.991	0.984	0.917	0.931	0.992	0.987	0.908
	ค่าเฉลี่ย		0.938	0.984	0.983	0.924	0.940	0.992	0.982	0.918	0.937	0.990	0.982	0.913

N	OR		NONE				OVER				HYBRID			
	CAT	CON	LR	RF	SVM	KNN	LR	RF	SVM	KNN	LR	RF	SVM	KNN
sensitivity														
100	L	L	0.932	0.978	0.960	0.806	0.862	0.980	0.967	0.852	0.873	0.991	0.985	0.872
	L	H	0.936	0.979	0.970	0.886	0.807	0.978	0.972	0.891	0.826	0.980	0.976	0.920
	H	L	0.950	0.977	0.969	0.775	0.868	0.983	0.978	0.829	0.873	0.979	0.974	0.837
	H	H	0.943	0.981	0.976	0.846	0.800	0.980	0.962	0.807	0.831	0.970	0.938	0.898
	ค่าเฉลี่ย		0.940	0.979	0.969	0.828	0.834	0.980	0.970	0.845	0.851	0.980	0.968	0.882
300	L	L	0.927	0.977	0.969	0.863	0.945	0.996	0.980	0.875	0.948	0.988	0.980	0.880
	L	H	0.929	0.981	0.972	0.897	0.922	0.988	0.977	0.901	0.932	0.982	0.974	0.889
	H	L	0.953	0.979	0.969	0.813	0.959	0.987	0.975	0.845	0.940	0.990	0.988	0.866
	H	H	0.931	0.973	0.967	0.882	0.940	0.989	0.978	0.900	0.940	0.980	0.979	0.899
	ค่าเฉลี่ย		0.935	0.978	0.969	0.864	0.942	0.990	0.978	0.880	0.940	0.985	0.980	0.884
500	L	L	0.921	0.981	0.971	0.877	0.932	0.986	0.973	0.875	0.928	0.987	0.977	0.904
	L	H	0.915	0.982	0.975	0.926	0.921	0.981	0.974	0.912	0.913	0.981	0.976	0.925
	H	L	0.937	0.979	0.971	0.876	0.952	0.989	0.972	0.869	0.950	0.979	0.976	0.876
	H	H	0.921	0.983	0.981	0.900	0.934	0.983	0.978	0.894	0.910	0.986	0.984	0.893
	ค่าเฉลี่ย		0.924	0.981	0.975	0.895	0.935	0.985	0.974	0.888	0.925	0.983	0.978	0.900

N	OR		NONE				OVER				HYBRID			
	CAT	CON	LR	RF	SVM	KNN	LR	RF	SVM	KNN	LR	RF	SVM	KNN
specificity														
100	L	L	0.963	0.991	0.988	0.938	0.876	0.995	0.990	0.929	0.878	0.987	0.977	0.890
	L	H	0.939	0.983	0.979	0.922	0.842	0.990	0.980	0.920	0.815	0.991	0.989	0.891
	H	L	0.969	0.990	0.987	0.910	0.877	0.995	0.989	0.906	0.870	0.992	0.990	0.859
	H	H	0.945	0.989	0.981	0.922	0.843	0.991	0.982	0.878	0.836	0.990	0.985	0.842
	ค่าเฉลี่ย		0.954	0.988	0.984	0.923	0.860	0.993	0.985	0.908	0.850	0.990	0.985	0.871
300	L	L	0.972	0.994	0.990	0.965	0.960	0.993	0.992	0.970	0.973	0.993	0.990	0.951
	L	H	0.954	0.993	0.990	0.963	0.947	0.996	0.988	0.957	0.945	0.995	0.992	0.951
	H	L	0.955	0.989	0.987	0.936	0.966	0.992	0.987	0.946	0.959	0.995	0.992	0.942
	H	H	0.950	0.990	0.988	0.945	0.962	0.997	0.984	0.952	0.954	0.991	0.989	0.926
	ค่าเฉลี่ย		0.958	0.992	0.989	0.952	0.959	0.995	0.988	0.956	0.958	0.994	0.991	0.943
500	L	L	0.955	0.993	0.992	0.955	0.948	0.995	0.990	0.957	0.948	0.993	0.991	0.930
	L	H	0.929	0.991	0.989	0.953	0.933	0.992	0.987	0.958	0.932	0.990	0.979	0.927
	H	L	0.964	0.988	0.985	0.922	0.960	0.997	0.992	0.939	0.957	0.990	0.982	0.919
	H	H	0.945	0.990	0.988	0.955	0.940	0.994	0.990	0.940	0.951	0.995	0.990	0.923
	ค่าเฉลี่ย		0.948	0.991	0.989	0.946	0.945	0.995	0.990	0.949	0.947	0.992	0.986	0.925

ตารางที่ 2 แสดงผลการเปรียบเทียบประสิทธิภาพของเทคนิคการจำแนกข้อมูลที่ถูกปรับสมดุลแต่ละวิธี เมื่อร้อยละของจำนวนข้อมูลระหว่างข้อมูลกลุ่มหลักต่อข้อมูลกลุ่มรอง คือ 60:40 และจำนวนตัวแปรระหว่างกลุ่มตัวแปรจัดประเภทต่อกลุ่มตัวแปรต่อเนื่อง คือ 5:3 โดยจำแนกตามขนาดตัวอย่าง อัตราออด และเทคนิคการปรับสมดุลข้อมูล

N	OR		NONE				OVER				HYBRID			
	CAT	CON	LR	RF	SVM	KNN	LR	RF	SVM	KNN	LR	RF	SVM	KNN
accuracy														
100	L	L	0.943	0.978	0.954	0.915	0.967	0.990	0.978	0.935	0.971	0.997	0.982	0.873
	L	H	0.943	0.982	0.971	0.929	0.922	0.993	0.988	0.933	0.943	0.988	0.971	0.957
	H	L	0.957	0.988	0.985	0.928	0.977	0.993	0.988	0.853	0.971	0.990	0.985	0.802
	H	H	0.929	0.973	0.971	0.915	0.941	0.989	0.976	0.941	0.929	0.975	0.948	0.859
	ค่าเฉลี่ย		0.943	0.980	0.970	0.922	0.952	0.991	0.983	0.916	0.954	0.988	0.972	0.873
300	L	L	0.919	0.992	0.990	0.927	0.875	0.989	0.975	0.917	0.914	0.996	0.995	0.914
	L	H	0.914	0.996	0.995	0.972	0.934	0.990	0.967	0.938	0.924	0.989	0.976	0.900
	H	L	0.933	0.978	0.966	0.915	0.930	0.991	0.982	0.886	0.914	0.985	0.971	0.909
	H	H	0.919	0.992	0.990	0.918	0.902	0.994	0.986	0.935	0.919	0.993	0.990	0.909
	ค่าเฉลี่ย		0.921	0.990	0.985	0.933	0.910	0.991	0.978	0.919	0.918	0.991	0.983	0.908
500	L	L	0.925	0.994	0.980	0.972	0.899	0.993	0.991	0.939	0.965	0.996	0.994	0.934
	L	H	0.925	0.972	0.968	0.970	0.927	0.986	0.973	0.937	0.943	0.983	0.971	0.914
	H	L	0.948	0.991	0.988	0.961	0.932	0.989	0.983	0.927	0.945	0.990	0.980	0.914
	H	H	0.928	0.990	0.980	0.960	0.915	0.989	0.983	0.939	0.945	0.981	0.977	0.928
	ค่าเฉลี่ย		0.932	0.987	0.979	0.966	0.918	0.989	0.983	0.936	0.950	0.988	0.981	0.923

N	OR		NONE				OVER				HYBRID			
	CAT	CON	LR	RF	SVM	KNN	LR	RF	SVM	KNN	LR	RF	SVM	KNN
sensitivity														
100	L	L	0.960	0.971	0.967	0.760	0.956	0.984	0.956	0.913	0.968	0.977	0.971	0.781
	L	H	0.928	0.974	0.964	0.892	0.837	0.985	0.979	0.860	0.875	0.990	0.937	0.937
	H	L	0.928	0.989	0.987	0.892	0.976	0.986	0.976	0.833	0.937	0.980	0.953	0.812
	H	H	0.900	0.970	0.966	0.933	0.902	0.990	0.975	0.951	0.938	0.964	0.946	0.687
	ค่าเฉลี่ย		0.929	0.976	0.971	0.869	0.918	0.986	0.972	0.889	0.930	0.978	0.952	0.804
300	L	L	0.896	0.986	0.977	0.942	0.879	0.983	0.967	0.887	0.898	0.991	0.990	0.907
	L	H	0.922	0.997	0.996	0.922	0.898	0.991	0.962	0.925	0.907	0.983	0.972	0.898
	H	L	0.924	0.969	0.956	0.913	0.915	0.994	0.983	0.872	0.944	0.987	0.972	0.879
	H	H	0.881	0.990	0.990	0.881	0.881	0.990	0.972	0.900	0.888	0.992	0.990	0.870
	ค่าเฉลี่ย		0.906	0.986	0.980	0.915	0.893	0.990	0.971	0.896	0.909	0.988	0.981	0.889
500	L	L	0.888	0.979	0.955	0.843	0.898	0.989	0.986	0.898	0.976	0.998	0.995	0.947
	L	H	0.906	0.967	0.953	0.893	0.905	0.984	0.950	0.895	0.953	0.985	0.970	0.906
	H	L	0.929	0.993	0.992	0.915	0.918	0.988	0.971	0.885	0.918	0.979	0.959	0.883
	H	H	0.899	0.980	0.966	0.892	0.910	0.981	0.965	0.891	0.930	0.976	0.970	0.906
	ค่าเฉลี่ย		0.906	0.980	0.967	0.886	0.908	0.986	0.968	0.892	0.944	0.985	0.974	0.911

N	OR		NONE				OVER				HYBRID			
	CAT	CON	LR	RF	SVM	KNN	LR	RF	SVM	KNN	LR	RF	SVM	KNN
specificity														
100	L	L	0.934	0.980	0.948	0.923	0.978	0.991	0.987	0.957	0.974	0.994	0.988	0.948
	L	H	0.953	0.983	0.976	0.953	0.935	0.995	0.990	0.942	0.952	0.987	0.981	0.974
	H	L	0.976	0.980	0.976	0.952	0.978	0.995	0.990	0.872	0.974	0.991	0.984	0.794
	H	H	0.951	0.975	0.975	0.902	0.977	0.987	0.977	0.933	0.923	0.981	0.947	0.861
	ค่าเฉลี่ย		0.954	0.980	0.969	0.933	0.967	0.992	0.986	0.926	0.956	0.988	0.975	0.894
300	L	L	0.935	0.994	0.992	0.933	0.871	0.991	0.982	0.948	0.932	0.997	0.996	0.922
	L	H	0.907	0.992	0.990	0.947	0.971	0.986	0.971	0.952	0.941	0.990	0.980	0.902
	H	L	0.940	0.980	0.974	0.914	0.946	0.988	0.982	0.901	0.883	0.981	0.970	0.941
	H	H	0.954	0.993	0.990	0.900	0.924	0.996	0.990	0.971	0.951	0.994	0.990	0.951
	ค่าเฉลี่ย		0.934	0.990	0.987	0.924	0.928	0.990	0.981	0.943	0.927	0.991	0.984	0.929
500	L	L	0.949	0.996	0.995	0.923	0.900	0.997	0.995	0.978	0.955	0.993	0.988	0.921
	L	H	0.940	0.981	0.980	0.937	0.948	0.997	0.995	0.976	0.932	0.981	0.972	0.921
	H	L	0.961	0.986	0.985	0.943	0.945	0.997	0.995	0.968	0.972	0.993	0.977	0.944
	H	H	0.950	0.991	0.990	0.931	0.920	0.990	0.987	0.985	0.960	0.990	0.983	0.949
	ค่าเฉลี่ย		0.950	0.989	0.988	0.934	0.928	0.995	0.993	0.977	0.955	0.989	0.980	0.934

ตารางที่ 3 แสดงผลการเปรียบเทียบประสิทธิภาพของเทคนิคการจำแนกข้อมูลที่ถูกปรับสมดุลแต่ละวิธี เมื่อร้อยละของจำนวนข้อมูลระหว่างข้อมูลกลุ่มหลักต่อข้อมูลกลุ่มรอง คือ 60:40 และจำนวนตัวแปรระหว่างกลุ่มตัวแปรจัดประเภทต่อกลุ่มตัวแปรต่อเนื่อง คือ 6:2 โดยจำแนกตามขนาดตัวอย่าง อัตราออด และเทคนิคการปรับสมดุลข้อมูล

N	OR		NONE				OVER				HYBRID			
	CAT	CON	LR	RF	SVM	KNN	LR	RF	SVM	KNN	LR	RF	SVM	KNN
accuracy														
100	L	L	0.957	0.989	0.985	0.901	0.945	0.993	0.989	0.858	0.985	0.992	0.989	0.871
	L	H	0.957	0.983	0.928	0.914	0.942	0.986	0.965	0.896	0.957	0.982	0.971	0.900
	H	L	0.943	0.968	0.956	0.859	0.990	0.992	0.988	0.860	0.975	0.990	0.987	0.942
	H	H	0.929	0.981	0.941	0.859	0.975	0.990	0.987	0.902	0.957	0.980	0.965	0.857
	ค่าเฉลี่ย		0.947	0.980	0.953	0.883	0.963	0.990	0.982	0.879	0.969	0.986	0.978	0.893
300	L	L	0.881	0.976	0.971	0.924	0.864	0.990	0.964	0.904	0.905	0.987	0.966	0.928
	L	H	0.886	0.987	0.985	0.909	0.884	0.988	0.978	0.901	0.905	0.990	0.985	0.928
	H	L	0.909	0.980	0.976	0.900	0.933	0.986	0.966	0.854	0.938	0.985	0.976	0.872
	H	H	0.886	0.981	0.976	0.919	0.932	0.997	0.995	0.918	0.881	0.995	0.971	0.928
	ค่าเฉลี่ย		0.891	0.981	0.977	0.913	0.903	0.990	0.976	0.894	0.907	0.989	0.975	0.914
500	L	L	0.911	0.990	0.988	0.960	0.930	0.993	0.985	0.930	0.914	0.987	0.985	0.962
	L	H	0.911	0.987	0.985	0.954	0.888	0.992	0.979	0.937	0.900	0.990	0.980	0.925
	H	L	0.937	0.995	0.988	0.897	0.920	0.989	0.975	0.947	0.928	0.988	0.977	0.922
	H	H	0.914	0.992	0.991	0.911	0.924	0.994	0.992	0.942	0.934	0.996	0.994	0.920
	ค่าเฉลี่ย		0.918	0.991	0.988	0.931	0.916	0.992	0.983	0.939	0.919	0.990	0.984	0.932

N	OR		NONE				OVER				HYBRID			
	CAT	CON	LR	RF	SVM	KNN	LR	RF	SVM	KNN	LR	RF	SVM	KNN
sensitivity														
100	L	L	0.884	0.991	0.988	0.769	0.888	0.987	0.977	0.777	0.971	0.993	0.991	0.857
	L	H	0.964	0.977	0.821	0.857	0.928	0.983	0.952	0.833	0.942	0.973	0.971	0.914
	H	L	0.931	0.955	0.943	0.862	0.984	0.990	0.976	0.809	0.974	0.991	0.980	0.971
	H	H	0.870	0.936	0.900	0.741	0.950	0.994	0.989	0.875	0.971	0.987	0.976	0.800
	ค่าเฉลี่ย		0.912	0.965	0.913	0.807	0.938	0.989	0.974	0.824	0.965	0.986	0.980	0.886
300	L	L	0.787	0.968	0.950	0.875	0.847	0.986	0.938	0.847	0.946	0.964	0.955	0.884
	L	H	0.844	0.977	0.966	0.900	0.892	0.980	0.966	0.859	0.902	0.989	0.982	0.911
	H	L	0.848	0.974	0.965	0.872	0.968	0.979	0.960	0.840	0.946	0.990	0.982	0.867
	H	H	0.873	0.970	0.968	0.884	0.905	0.998	0.991	0.913	0.902	0.997	0.973	0.955
	ค่าเฉลี่ย		0.838	0.972	0.962	0.883	0.903	0.986	0.964	0.865	0.924	0.985	0.973	0.904
500	L	L	0.893	0.985	0.980	0.933	0.910	0.989	0.985	0.905	0.897	0.992	0.988	0.954
	L	H	0.899	0.983	0.981	0.937	0.906	0.988	0.984	0.942	0.885	0.987	0.982	0.931
	H	L	0.953	0.993	0.980	0.860	0.905	0.985	0.970	0.925	0.925	0.980	0.977	0.914
	H	H	0.912	0.989	0.987	0.856	0.921	0.990	0.989	0.937	0.951	0.995	0.994	0.902
	ค่าเฉลี่ย		0.914	0.988	0.982	0.897	0.911	0.988	0.982	0.927	0.915	0.989	0.985	0.925

N	OR		NONE				OVER				HYBRID			
	CAT	CON	LR	RF	SVM	KNN	LR	RF	SVM	KNN	LR	RF	SVM	KNN
specificity														
100	L	L	0.967	0.980	0.977	0.977	0.951	0.994	0.988	0.936	0.988	0.990	0.979	0.885
	L	H	0.952	0.991	0.985	0.952	0.955	0.989	0.977	0.955	0.961	0.985	0.971	0.885
	H	L	0.952	0.970	0.961	0.857	0.991	0.995	0.990	0.909	0.972	0.984	0.975	0.914
	H	H	0.975	0.984	0.977	0.950	0.978	0.985	0.976	0.928	0.942	0.970	0.942	0.914
	ค่าเฉลี่ย		0.962	0.981	0.975	0.934	0.969	0.991	0.983	0.932	0.966	0.982	0.967	0.900
300	L	L	0.938	0.989	0.984	0.954	0.882	0.994	0.991	0.966	0.857	0.988	0.979	0.979
	L	H	0.917	0.990	0.988	0.917	0.875	0.992	0.987	0.946	0.908	0.991	0.989	0.948
	H	L	0.952	0.987	0.984	0.920	0.895	0.987	0.973	0.869	0.928	0.983	0.969	0.877
	H	H	0.896	0.985	0.982	0.948	0.962	0.990	0.977	0.924	0.857	0.989	0.969	0.897
	ค่าเฉลี่ย		0.926	0.988	0.985	0.935	0.904	0.991	0.982	0.926	0.888	0.988	0.977	0.925
500	L	L	0.925	0.996	0.995	0.980	0.950	0.997	0.985	0.955	0.931	0.985	0.982	0.971
	L	H	0.921	0.990	0.989	0.968	0.871	0.993	0.974	0.932	0.914	0.994	0.977	0.920
	H	L	0.925	0.997	0.995	0.925	0.935	0.995	0.980	0.970	0.931	0.994	0.977	0.931
	H	H	0.916	0.995	0.994	0.958	0.927	0.997	0.995	0.948	0.920	0.997	0.994	0.937
	ค่าเฉลี่ย		0.922	0.995	0.993	0.958	0.921	0.996	0.984	0.951	0.924	0.993	0.983	0.940

ตารางที่ 4 แสดงผลการเปรียบเทียบประสิทธิภาพของเทคนิคการจำแนกข้อมูลที่ถูกปรับสมดุลแต่ละวิธี เมื่อร้อยละของจำนวนข้อมูลระหว่างข้อมูลกลุ่มหลักต่อข้อมูลกลุ่มรอง คือ 70:30 และจำนวนตัวแปรระหว่างกลุ่มตัวแปรจัดประเภทต่อกลุ่มตัวแปรต่อเนื่อง คือ 4:4 โดยจำแนกตามขนาดตัวอย่าง อัตราออด และเทคนิคการปรับสมดุลข้อมูล

N	OR		NONE				OVER				HYBRID			
	CAT	CON	LR	RF	SVM	KNN	LR	RF	SVM	KNN	LR	RF	SVM	KNN
accuracy														
100	L	L	0.973	0.977	0.970	0.873	0.974	0.996	0.991	0.899	0.978	0.989	0.987	0.876
	L	H	0.936	0.980	0.978	0.899	0.937	0.992	0.982	0.881	0.936	0.980	0.975	0.867
	H	L	0.980	0.983	0.978	0.853	0.973	0.987	0.980	0.857	0.987	0.990	0.975	0.829
	H	H	0.941	0.975	0.971	0.867	0.940	0.988	0.976	0.879	0.947	0.981	0.975	0.883
	ค่าเฉลี่ย		0.958	0.979	0.974	0.873	0.956	0.991	0.982	0.879	0.962	0.985	0.978	0.864
300	L	L	0.960	0.980	0.975	0.927	0.952	0.995	0.983	0.921	0.959	0.989	0.988	0.910
	L	H	0.946	0.987	0.984	0.946	0.943	0.994	0.986	0.930	0.951	0.989	0.988	0.933
	H	L	0.968	0.990	0.983	0.907	0.960	0.989	0.977	0.900	0.966	0.999	0.984	0.882
	H	H	0.946	0.991	0.983	0.925	0.946	0.992	0.984	0.914	0.956	0.984	0.983	0.922
	ค่าเฉลี่ย		0.955	0.987	0.981	0.926	0.950	0.993	0.983	0.916	0.958	0.990	0.986	0.912
500	L	L	0.946	0.990	0.982	0.930	0.948	0.995	0.982	0.916	0.944	0.990	0.983	0.931
	L	H	0.928	0.987	0.980	0.936	0.936	0.985	0.980	0.933	0.920	0.990	0.982	0.931
	H	L	0.951	0.986	0.975	0.905	0.954	0.999	0.979	0.906	0.954	0.990	0.977	0.900
	H	H	0.931	0.991	0.984	0.927	0.922	0.989	0.981	0.913	0.921	0.989	0.979	0.907
	ค่าเฉลี่ย		0.939	0.989	0.980	0.925	0.940	0.992	0.981	0.917	0.935	0.990	0.980	0.917

N	OR		NONE				OVER				HYBRID			
	CAT	CON	LR	RF	SVM	KNN	LR	RF	SVM	KNN	LR	RF	SVM	KNN
sensitivity														
100	L	L	0.927	0.931	0.868	0.486	0.972	0.990	0.985	0.826	0.971	0.983	0.982	0.815
	L	H	0.907	0.964	0.951	0.844	0.921	0.981	0.974	0.849	0.925	0.978	0.975	0.855
	H	L	0.952	0.964	0.952	0.633	0.969	0.977	0.975	0.773	0.979	0.980	0.968	0.827
	H	H	0.924	0.953	0.948	0.775	0.919	0.970	0.961	0.834	0.950	0.967	0.962	0.854
	ค่าเฉลี่ย		0.928	0.953	0.930	0.685	0.945	0.980	0.974	0.821	0.956	0.977	0.972	0.838
300	L	L	0.897	0.973	0.922	0.750	0.933	0.990	0.968	0.857	0.949	0.982	0.980	0.856
	L	H	0.919	0.985	0.969	0.888	0.931	0.983	0.980	0.906	0.945	0.986	0.984	0.909
	H	L	0.946	0.988	0.965	0.793	0.949	0.988	0.963	0.831	0.957	0.998	0.974	0.814
	H	H	0.932	0.974	0.965	0.865	0.935	0.990	0.979	0.876	0.948	0.979	0.976	0.900
	ค่าเฉลี่ย		0.924	0.980	0.955	0.824	0.937	0.988	0.973	0.868	0.950	0.986	0.979	0.870
500	L	L	0.873	0.968	0.948	0.805	0.931	0.984	0.972	0.862	0.929	0.987	0.978	0.899
	L	H	0.905	0.980	0.972	0.896	0.915	0.981	0.974	0.903	0.908	0.985	0.973	0.919
	H	L	0.923	0.965	0.949	0.823	0.940	0.999	0.965	0.853	0.939	0.979	0.963	0.859
	H	H	0.910	0.984	0.978	0.889	0.918	0.982	0.968	0.884	0.913	0.980	0.968	0.892
	ค่าเฉลี่ย		0.903	0.974	0.962	0.853	0.926	0.987	0.970	0.876	0.922	0.983	0.971	0.892

N	OR		NONE				OVER				HYBRID			
	CAT	CON	LR	RF	SVM	KNN	LR	RF	SVM	KNN	LR	RF	SVM	KNN
specificity														
100	L	L	0.987	0.988	0.974	0.987	0.977	0.998	0.997	0.972	0.984	0.994	0.993	0.937
	L	H	0.952	0.995	0.994	0.928	0.952	0.993	0.991	0.911	0.945	0.982	0.975	0.872
	H	L	0.991	0.998	0.997	0.941	0.977	0.989	0.986	0.940	0.996	0.997	0.981	0.833
	H	H	0.949	0.985	0.984	0.922	0.962	0.994	0.991	0.925	0.945	0.996	0.989	0.912
	ค่าเฉลี่ย		0.970	0.992	0.987	0.945	0.967	0.994	0.991	0.937	0.968	0.992	0.985	0.889
300	L	L	0.983	0.993	0.991	0.980	0.970	0.999	0.998	0.984	0.968	0.998	0.996	0.962
	L	H	0.962	0.994	0.987	0.979	0.955	0.996	0.990	0.954	0.959	0.994	0.991	0.956
	H	L	0.978	0.992	0.991	0.959	0.971	0.997	0.991	0.967	0.974	0.999	0.994	0.946
	H	H	0.954	0.996	0.995	0.961	0.956	0.993	0.989	0.953	0.964	0.991	0.989	0.945
	ค่าเฉลี่ย		0.969	0.994	0.991	0.970	0.963	0.996	0.992	0.965	0.966	0.996	0.993	0.952
500	L	L	0.979	0.997	0.996	0.980	0.964	0.997	0.993	0.969	0.957	0.994	0.987	0.962
	L	H	0.942	0.989	0.984	0.962	0.957	0.990	0.986	0.963	0.931	0.993	0.992	0.942
	H	L	0.967	0.989	0.988	0.948	0.967	0.999	0.994	0.958	0.969	0.992	0.991	0.939
	H	H	0.946	0.992	0.989	0.953	0.925	0.996	0.993	0.942	0.929	0.991	0.990	0.920
	ค่าเฉลี่ย		0.959	0.992	0.989	0.961	0.953	0.996	0.992	0.958	0.947	0.993	0.990	0.941

ตารางที่ 5 แสดงผลการเปรียบเทียบประสิทธิภาพของเทคนิคการจำแนกข้อมูลที่ถูกปรับสมดุลแต่ละวิธี เมื่อร้อยละของจำนวนข้อมูลระหว่างข้อมูลกลุ่มหลักต่อข้อมูลกลุ่มรอง คือ 70:30 และจำนวนตัวแปรระหว่างกลุ่มตัวแปรจัดประเภทต่อกลุ่มตัวแปรต่อเนื่อง คือ 5:3 โดยจำแนกตามขนาดตัวอย่าง อัตราออด และเทคนิคการปรับสมดุลข้อมูล

N	OR		NONE				OVER				HYBRID			
	CAT	CON	LR	RF	SVM	KNN	LR	RF	SVM	KNN	LR	RF	SVM	KNN
accuracy														
100	L	L	0.929	0.958	0.957	0.915	0.942	0.997	0.990	0.932	0.957	0.979	0.938	0.957
	L	H	0.957	0.974	0.971	0.873	0.946	0.963	0.968	0.893	0.929	0.971	0.971	0.873
	H	L	0.957	0.974	0.971	0.873	0.978	0.981	0.978	0.915	0.929	0.952	0.927	0.901
	H	H	0.915	0.981	0.971	0.943	0.967	0.975	0.970	0.934	0.985	0.985	0.985	0.943
	ค่าเฉลี่ย		0.940	0.972	0.968	0.901	0.958	0.979	0.977	0.919	0.950	0.972	0.955	0.919
300	L	L	0.886	0.990	0.981	0.909	0.875	0.999	0.960	0.875	0.919	0.988	0.971	0.895
	L	H	0.928	0.980	0.976	0.957	0.909	0.995	0.983	0.946	0.872	0.995	0.981	0.914
	H	L	0.943	0.991	0.990	0.905	0.903	0.992	0.992	0.868	0.943	0.994	0.985	0.895
	H	H	0.914	0.983	0.971	0.919	0.926	0.996	0.982	0.947	0.919	0.999	0.985	0.919
	ค่าเฉลี่ย		0.918	0.986	0.980	0.923	0.903	0.996	0.979	0.909	0.913	0.994	0.981	0.906
500	L	L	0.957	0.987	0.982	0.911	0.952	0.998	0.988	0.930	0.948	0.996	0.991	0.943
	L	H	0.937	0.991	0.982	0.940	0.923	0.996	0.982	0.951	0.965	0.998	0.985	0.931
	H	L	0.945	0.979	0.977	0.917	0.942	0.983	0.976	0.928	0.948	0.985	0.974	0.945
	H	H	0.948	0.998	0.988	0.945	0.925	0.999	0.972	0.947	0.943	0.994	0.988	0.934
	ค่าเฉลี่ย		0.947	0.989	0.982	0.928	0.936	0.994	0.980	0.939	0.951	0.993	0.985	0.938

N	OR		NONE				OVER				HYBRID			
	CAT	CON	LR	RF	SVM	KNN	LR	RF	SVM	KNN	LR	RF	SVM	KNN
sensitivity														
100	L	L	0.833	0.843	0.833	0.666	0.924	0.998	0.994	0.943	0.937	0.972	0.947	0.968
	L	H	0.958	0.959	0.916	0.666	0.936	0.941	0.936	0.851	0.843	0.939	0.937	0.781
	H	L	0.958	0.986	0.984	0.750	0.957	0.963	0.957	0.914	0.906	0.947	0.938	0.906
	H	H	0.923	0.992	0.989	0.884	0.933	0.966	0.955	0.888	0.968	0.968	0.968	0.875
	ค่าเฉลี่ย		0.918	0.945	0.931	0.742	0.938	0.967	0.961	0.899	0.914	0.957	0.948	0.883
300	L	L	0.779	0.973	0.955	0.808	0.832	1.000	0.944	0.804	0.888	1.000	0.953	0.870
	L	H	0.930	0.962	0.953	0.906	0.896	0.993	1.000	0.936	0.870	0.997	0.981	0.870
	H	L	0.937	0.990	0.987	0.850	0.893	0.998	0.984	0.816	0.962	0.990	0.972	0.888
	H	H	0.892	0.987	0.982	0.903	0.923	1.000	0.966	0.898	0.907	1.000	1.000	0.925
	ค่าเฉลี่ย		0.885	0.978	0.969	0.867	0.886	0.998	0.974	0.864	0.907	0.997	0.977	0.888
500	L	L	0.926	0.975	0.972	0.733	0.942	0.995	0.975	0.892	0.953	0.990	0.988	0.947
	L	H	0.915	0.983	0.969	0.869	0.918	0.992	0.968	0.927	0.953	1.000	0.976	0.895
	H	L	0.942	0.966	0.942	0.827	0.960	0.972	0.956	0.882	0.941	0.972	0.965	0.895
	H	H	0.926	0.999	0.992	0.919	0.911	1.000	0.972	0.920	0.941	0.989	0.988	0.906
	ค่าเฉลี่ย		0.927	0.981	0.969	0.837	0.933	0.990	0.968	0.905	0.947	0.988	0.979	0.911

N	OR		NONE				OVER				HYBRID			
	CAT	CON	LR	RF	SVM	KNN	LR	RF	SVM	KNN	LR	RF	SVM	KNN
specificity														
100	L	L	0.962	0.971	0.965	0.931	0.960	0.997	0.980	0.921	0.974	0.988	0.927	0.948
	L	H	0.957	0.979	0.982	0.978	0.957	0.984	0.978	0.936	0.936	0.989	0.986	0.948
	H	L	0.957	0.969	0.957	0.936	0.998	0.999	0.982	0.916	0.948	0.956	0.914	0.897
	H	H	0.911	0.980	0.955	0.977	0.997	0.998	0.988	0.978	0.989	0.997	0.994	0.965
	ค่าเฉลี่ย		0.947	0.975	0.965	0.956	0.978	0.995	0.982	0.938	0.962	0.983	0.955	0.940
300	L	L	0.937	0.996	0.993	0.958	0.920	1.000	0.978	0.949	0.951	0.983	0.990	0.922
	L	H	0.928	0.994	0.992	0.992	0.924	1.000	0.966	0.957	0.873	0.992	0.980	0.961
	H	L	0.946	0.993	0.992	0.938	0.914	0.982	1.000	0.921	0.922	0.982	1.000	0.902
	H	H	0.932	0.966	0.949	0.932	0.928	0.993	1.000	1.000	0.932	0.996	0.970	0.912
	ค่าเฉลี่ย		0.936	0.987	0.982	0.955	0.922	0.994	0.986	0.957	0.920	0.988	0.985	0.924
500	L	L	0.971	0.994	0.987	0.991	0.961	1.000	1.000	0.965	0.944	0.995	0.994	0.938
	L	H	0.950	0.997	0.990	0.981	0.927	0.998	0.995	0.974	0.977	0.991	0.994	0.966
	H	L	0.947	0.999	0.995	0.965	0.926	1.000	0.995	0.971	0.955	1.000	0.983	0.994
	H	H	0.962	0.992	0.986	0.962	0.938	0.999	0.973	0.973	0.944	1.000	0.988	0.960
	ค่าเฉลี่ย		0.958	0.996	0.990	0.975	0.938	0.999	0.991	0.971	0.955	0.997	0.990	0.965

ตารางที่ 6 แสดงผลการเปรียบเทียบประสิทธิภาพของเทคนิคการจำแนกข้อมูลที่ถูกปรับสมดุลแต่ละวิธี เมื่อร้อยละของจำนวนข้อมูลระหว่างข้อมูลกลุ่มหลักต่อข้อมูลกลุ่มรอง คือ 70:30 และจำนวนตัวแปรระหว่างกลุ่มตัวแปรจัดประเภทต่อกลุ่มตัวแปรต่อเนื่อง คือ 6:2 โดยจำแนกตามขนาดตัวอย่าง อัตราออด และเทคนิคการปรับสมดุลข้อมูล

N	OR		NONE				OVER				HYBRID			
	CAT	CON	LR	RF	SVM	KNN	LR	RF	SVM	KNN	LR	RF	SVM	KNN
accuracy														
100	L	L	0.985	0.987	0.971	0.845	0.972	0.991	0.990	0.879	0.985	0.997	0.993	0.914
	L	H	0.971	0.972	0.958	0.915	0.958	0.981	0.961	0.927	0.942	0.984	0.971	0.885
	H	L	0.894	0.976	0.971	0.915	0.979	0.992	0.989	0.885	0.971	0.981	0.971	0.757
	H	H	0.914	0.974	0.971	0.871	0.954	0.989	0.988	0.873	0.971	0.987	0.973	0.914
	ค่าเฉลี่ย		0.941	0.977	0.968	0.887	0.966	0.988	0.982	0.891	0.967	0.987	0.977	0.868
300	L	L	0.919	0.977	0.966	0.938	0.879	0.999	0.980	0.909	0.895	0.989	0.981	0.867
	L	H	0.895	0.993	0.985	0.924	0.915	0.997	0.963	0.912	0.853	0.995	0.985	0.886
	H	L	0.962	0.991	0.971	0.890	0.901	0.999	0.981	0.923	0.914	0.996	0.985	0.933
	H	H	0.919	0.935	0.966	0.886	0.881	0.992	0.996	0.917	0.905	0.999	0.976	0.943
	ค่าเฉลี่ย		0.924	0.974	0.972	0.910	0.894	0.997	0.980	0.915	0.892	0.995	0.982	0.907
500	L	L	0.917	0.998	0.997	0.945	0.920	0.997	0.984	0.945	0.928	0.998	0.988	0.931
	L	H	0.908	0.989	0.985	0.943	0.901	0.996	0.978	0.941	0.897	0.985	0.982	0.942
	H	L	0.920	0.997	0.988	0.917	0.920	0.992	0.975	0.907	0.920	0.994	0.980	0.925
	H	H	0.937	0.981	0.977	0.923	0.930	0.999	0.975	0.913	0.911	0.993	0.988	0.920
	ค่าเฉลี่ย		0.921	0.991	0.987	0.932	0.918	0.996	0.978	0.927	0.914	0.993	0.985	0.930

N	OR		NONE				OVER				HYBRID			
	CAT	CON	LR	RF	SVM	KNN	LR	RF	SVM	KNN	LR	RF	SVM	KNN
sensitivity														
100	L	L	0.986	0.989	0.888	0.744	0.962	0.993	0.877	0.754	0.971	0.993	0.989	0.885
	L	H	0.916	0.921	0.933	0.750	0.914	0.979	0.957	0.893	0.942	0.986	0.971	0.971
	H	L	0.876	0.979	0.958	0.750	0.957	0.983	0.978	0.829	0.971	0.978	0.942	0.685
	H	H	0.892	0.957	0.928	0.750	0.904	0.991	0.976	0.809	0.971	0.974	0.967	0.885
	ค่าเฉลี่ย		0.918	0.962	0.927	0.674	0.934	0.987	0.947	0.821	0.964	0.983	0.967	0.857
300	L	L	0.750	0.907	0.865	0.846	0.874	1.000	0.962	0.861	0.884	0.977	0.964	0.849
	L	H	0.797	0.964	0.956	0.797	0.894	1.000	0.950	0.901	0.849	1.000	0.982	0.884
	H	L	0.913	0.951	0.913	0.826	0.922	0.999	0.985	0.873	0.946	0.994	0.991	0.911
	H	H	0.835	0.933	0.936	0.797	0.886	0.988	0.992	0.871	0.911	0.998	0.955	0.911
	ค่าเฉลี่ย		0.824	0.939	0.918	0.817	0.894	0.997	0.972	0.877	0.898	0.992	0.973	0.889
500	L	L	0.857	0.999	0.998	0.880	0.920	0.995	0.968	0.897	0.914	1.000	0.982	0.914
	L	H	0.848	0.980	0.978	0.906	0.896	1.000	0.962	0.929	0.885	0.973	0.965	0.925
	H	L	0.880	0.992	0.984	0.841	0.924	1.000	0.950	0.843	0.902	0.996	0.971	0.874
	H	H	0.931	0.972	0.958	0.882	0.917	1.000	0.970	0.888	0.897	1.000	0.988	0.897
	ค่าเฉลี่ย		0.879	0.986	0.980	0.877	0.914	0.999	0.963	0.889	0.900	0.992	0.977	0.903

N	OR		NONE				OVER				HYBRID			
	CAT	CON	LR	RF	SVM	KNN	LR	RF	SVM	KNN	LR	RF	SVM	KNN
specificity														
100	L	L	0.981	0.980	0.885	0.981	0.981	0.990	0.981	0.892	0.986	0.998	0.995	0.942
	L	H	0.973	0.982	0.964	0.916	0.947	0.989	0.967	0.959	0.942	0.977	0.971	0.800
	H	L	0.900	0.952	0.978	0.916	0.886	0.994	0.925	0.938	0.971	0.965	0.986	0.828
	H	H	0.928	0.986	0.979	0.952	0.947	0.973	0.958	0.933	0.971	0.991	0.974	0.942
	ค่าเฉลี่ย		0.946	0.975	0.952	0.941	0.940	0.987	0.958	0.931	0.968	0.983	0.982	0.878
300	L	L	0.974	0.984	0.968	0.968	0.885	0.998	1.000	0.959	0.908	0.990	1.000	0.887
	L	H	0.943	0.996	0.987	0.985	0.938	0.996	0.977	0.923	0.857	0.995	0.989	0.887
	H	L	0.985	0.993	0.973	0.922	0.877	1.000	0.977	0.977	0.877	0.998	0.979	0.959
	H	H	0.969	0.996	0.984	0.939	0.877	0.999	1.000	0.967	0.897	1.000	1.000	0.979
	ค่าเฉลี่ย		0.968	0.992	0.978	0.954	0.894	0.998	0.989	0.957	0.885	0.996	0.992	0.928
500	L	L	0.951	0.996	0.995	0.982	0.921	1.000	1.000	0.991	0.942	1.000	0.994	0.948
	L	H	0.948	0.991	0.990	0.966	0.907	0.993	0.995	0.953	0.908	0.992	1.000	0.960
	H	L	0.941	0.992	0.991	0.959	0.917	0.986	1.000	0.969	0.937	0.990	0.986	0.977
	H	H	0.941	0.991	0.990	0.951	0.942	1.000	0.980	0.937	0.925	0.989	0.988	0.942
	ค่าเฉลี่ย		0.945	0.993	0.992	0.965	0.922	0.995	0.994	0.963	0.928	0.993	0.992	0.957

ประวัติผู้เขียน

ชื่อ-สกุล	นางสาวกาญธนา ลออสิริกุล
วัน เดือน ปี เกิด	20 ตุลาคม 2537
สถานที่เกิด	กรุงเทพมหานคร
วุฒิการศึกษา	ปริญญาตรี: ครุศาสตรบัณฑิต จุฬาลงกรณ์มหาวิทยาลัย
ที่อยู่ปัจจุบัน	1295 หมู่ 3 ต.สำโรงเหนือ อ.เมือง จ.สมุทรปราการ 10270