เปรียบเทียบประสิทธิภาพการจำแนกข้อมูลที่ไม่สมดุลโดยใช้การจำลอง
Comparison of Efficiency for Imbalanced Data Classification via Simulation

กาญธนา ลออสิริกุล (Kantana La-orsirikul)*, ประภาศิริ รัชชประภาพรกุล (Prapasiri Ratchaprapapornkul)*, และ สุรศักดิ์ เก้าเอี้ยน (Surasak Kao-Iean)*

Received: May 28, 2023
Revised: July 18, 2023
Accepted: July 25, 2023

*ผู้นิพนธ์ประสานงาน: ประภาศิริ รัชชประภาพรกุล (Prapasiri Ratchaprapapornkul) อีเมล: Prapasiri.r@chula.ac.th

DOI:10.14416/j.it.2024.v1.001

บทคัดย่อ

งานวิจัยนี้มีวัตถุประสงค์เพื่อเปรียบเทียบประสิทธิภาพการปรับสมดุลข้อมูลระหว่างวิธีสุ่มเกินและวิธีผสมผสาน และเปรียบเทียบการจำแนกข้อมูลปรับสมดุลแล้วด้วยเทคนิคแรนดอมฟอร์เรส การถดถอยโลจิสติก และซัพพอร์ตเวคเตอร์แมชชีน เมื่อข้อมูลมีความไม่สมดุลระดับสูงและตัวแปรทำนายส่วนใหญ่เป็นตัวแปร จัดประเภท ด้วยการจำลองข้อมูลโดยพิจารณาจากขนาดตัวอย่าง อัตราส่วนจำนวนตัวแปรจัดประเภทต่อตัวแปรต่อเนื่อง และอัตราออด ผลการศึกษาพบว่าการปรับสมดุลข้อมูลด้วยวิธีสุ่มเกินก่อนนำข้อมูลไปจำแนกจะมีค่าความถูกต้อง ค่าความไว และค่าความจำเพาะสูงกว่าการใช้วิธีผสมผสานในแต่ละขนาดตัวอย่าง ทั้งนี้เมื่อนำข้อมูลที่ปรับสมดุลข้อมูลแล้วมาจำแนกด้วยแรนดอมฟอร์เรสจะมีค่าความถูกต้อง ค่าความไว และค่าความจำเพาะสูงที่สุด โดยเฉลี่ยคือ 0.996, 0.999 และ 0.998 ตามลำดับ นอกจากนี้เมื่อชุดข้อมูลมีจำนวนตัวแปรจัดประเภทเพิ่มขึ้นแล้วจำแนกข้อมูลด้วยวิธีการถดถอยโลจิสติกจะทำให้ค่าความถูกต้องในการจำแนกลดลง ซึ่งผลการวิจัยดังกล่าวสามารถนำไปใช้เป็นแนวทางในการเลือกวิธีการปรับสมดุลข้อมูลให้เหมาะสมกับสภาพข้อมูลในสถานการณ์จริง

คำสำคัญ: ข้อมูลไม่สมดุล ข้อมูลจัดประเภท การจำแนกข้อมูล การจำลองข้อมูล วิธีสุ่มเกิน วิธีผสมผสาน

Abstract

The aims of this research are to compare the efficiency of imbalance techniques between over sampling and hybrid methods and to compare performance of classification techniques: random forest, logistic regression, and support vector machine, via simulation. The study is given by high imbalanced data and predicted variables which are mostly categorical data. The criteria of the simulation are sample sizes, ratio of the number of predicted variables between categorical variables and continuous variables, and odds ratio. The results shown that balancing data with over sampling method before classify had higher accuracy, sensitivity, and specificity than hybrid method in each sample sizes. In addition, the balanced data classified with random forest had the highest accuracy, sensitivity, and specificity, the average were 0.996, 0.999 and 0.998 respectively. Moreover, logistic regression technique yields less accurate classification when the number of categorical variables is higher. The result of research can be used as a guideline for choosing a data balancing method which appropriate to data conditions in real situations.

Keywords: Imbalance Data, Categorical Data, Classification, Simulation, Over Sampling Methods, Hybrid Methods.

1. บทนำ

ปัจจุบันการจำแนกประเภทของข้อมูล (Data Classification) ถูกใช้ในการจำแนกข้อมูลที่ไม่ทราบประเภท โดยเทคนิคที่นิยมนำมาใช้ในการจำแนกข้อมูลมีหลายเทคนิค เช่น การใช้โครงข่าย

* ภาควิชาวิจัยและจิตวิทยาการศึกษา คณะครุศาสตร์ จุฬาลงกรณ์มหาวิทยาลัย

* Department of Educational Research and Psychology, Faculty of Education, Chulalongkorn University

ประสาทเทียม (Neural Network) หรือแรนดอมฟอร์เรส (Random Forest) หรือการวิเคราะห์การถดถอยโลจิสติก (Logistic Regression Analysis) หรือซัพพอร์ตเวคเตอร์แมชชีน (Support Vector Machine) ซึ่งเทคนิคเหล่านี้จะมีประสิทธิภาพและมีความถูกต้องแม่นยำในการจำแนกที่สูงขึ้นเมื่อนำไปใช้ในการจำแนกข้อมูลที่สมดุล (Balance Data) และประสิทธิภาพจะลดลงเมื่อเครื่องเรียนรู้จากข้อมูลไม่สมดุล (Imbalanced Data) โดยข้อมูลไม่สมดุล คือข้อมูลที่มีจำนวนข้อมูลในแต่ละกลุ่มแตกต่างกันมาก โดยมีข้อมูลในกลุ่มใดกลุ่มหนึ่งแตกต่างจากข้อมูลกลุ่มอื่น ๆ เป็นจำนวนมาก ปัญหาการจำแนกประเภทข้อมูลที่ไม่สมดุล (Imbalanced Datasets) เกิดจากการที่มีข้อมูล 2 กลุ่มหรือมากกว่า โดยข้อมูลที่เป็นกลุ่มหลัก (Majority) จะมีจำนวนสมาชิกของข้อมูลมากกว่า ในขณะที่ข้อมูลกลุ่มรอง (Minority) จะมีจำนวนสมาชิกของข้อมูลน้อยกว่า [1]-[4] เช่น การออกกลางคันของนักเรียนเด็กและเยาวชนบนท้องถนน และสภาวะหมดไฟในการทำงานของบุคลากรทางการศึกษา จากการศึกษาพบว่า จำนวนของนักเรียนที่ออกกลางคันมีน้อยกว่าจำนวนนักเรียนที่เรียนจนครบหลักสูตรมาก [5]-[7] จำนวนเด็กและเยาวชนบนท้องถนนมีน้อยกว่าจำนวนเด็กและเยาวชนที่ไม่ได้เป็นเด็กและเยาวชนบนท้องถนนมาก [8]-[9] และจำนวนบุคลากรทางการศึกษาที่มีสภาวะหมดไฟในการทำงานมีน้อยกว่าจำนวนบุคลากรทางการศึกษาที่ไม่ได้มีสภาวะหมดไฟในการทำงานมาก [10] จากข้อมูลดังกล่าวส่วนใหญ่พบว่า ข้อมูลกลุ่มหลักอยู่ในช่วงร้อยละ 70-85 ทั้งนี้ โดยธรรมชาติของความเป็นจริงการที่เราจะกำหนดให้จำนวนสมาชิกของข้อมูลในกลุ่มหลักและกลุ่มรองให้มีจำนวนที่เท่ากัน เพื่อให้เครื่องเกิดการเรียนรู้หรือการจัดกลุ่มข้อมูลนั้นเป็นเรื่องยากหรืออาจจะเป็นไปไม่ได้ ดังนั้น จึงเป็นปัญหาที่ท้าทายสำหรับการหาขั้นตอนวิธีที่เหมาะสมสำหรับการจัดการข้อมูลที่ไม่สมดุล เนื่องจากถ้านำข้อมูลทั้งสองชุดเข้าสู่ขั้นตอนการเรียนรู้ด้วยเครื่องพร้อมกันทั้งหมด จะทำให้ผลการแบ่งกลุ่มข้อมูลเกิดความผิดพลาดสูง กล่าวคือข้อมูลที่อยู่ในกลุ่มรองจะถูกครอบงำหรือถูกจัดให้ไปอยู่ในกลุ่มหลักทั้งหมด ซึ่งจะนำไปสู่ปัญหาที่เรียกว่า “ปัญหาการแบ่งกลุ่มข้อมูลผิดกลุ่ม (misclassification)” [1]

วิธีการแก้ปัญหาความไม่สมดุลของข้อมูลซึ่งเป็นกระบวนการในการจัดการข้อมูลก่อนดำเนินการสร้างตัวแบบ โดยใช้หลักการซึ่งแบ่งออกเป็น 3 ระดับคือ 1) การแก้ปัญหาข้อมูลไม่สมดุลที่ระดับข้อมูล (Data Level Solutions) เป็นการแก้ปัญหาในขั้นตอนก่อนการประมวลผลโดยจะปรับข้อมูลที่ไม่สมดุลให้กลายเป็นข้อมูลสมดุลด้วยเทคนิคการสุ่มเลือกข้อมูล ซึ่งวิธีที่นิยมใช้ คือ วิธีสุ่มเกิน (Over Sampling Methods) และวิธีผสมผสาน (Hybrid Methods) 2) การแก้ปัญหาข้อมูลไม่สมดุลที่ระดับขั้นตอนวิธีการ (Algorithmic Level Solutions) เป็นการแก้ปัญหาโดยการปรับการเรียนรู้ของอัลกอริทึมมาตรฐานสำหรับการจำแนกข้อมูลที่มีอยู่เดิมให้สามารถเรียนรู้ข้อมูลไม่สมดุลโดยให้มีการเอนเอียงไปทางข้อมูลกลุ่มรอง และ 3) การแก้ปัญหาข้อมูลไม่สมดุลด้วยการเรียนรู้แบบมีค่าใช้จ่าย (Cost-Sensitive Training) [11]

นอกจากนี้เทคนิควิธีในการจำแนกและปรับสมดุลข้อมูลส่วนใหญ่เหมาะที่จะใช้กับข้อมูลที่มีตัวแปรทำนายเป็นตัวแปรต่อเนื่องทั้งหมดหรือมีสัดส่วนของจำนวนตัวแปรทำนายที่เป็นตัวแปรต่อเนื่องมากกว่าตัวแปรจัดประเภท เช่น เทคนิค SMOTE (Synthetic Minority Over Sampling) [2] และการเปรียบเทียบประสิทธิภาพการจำแนกส่วนใหญ่ศึกษาในข้อมูลที่ตัวแปรทำนายส่วนใหญ่เป็นตัวแปรต่อเนื่อง [12], [13]

ดังนั้นการศึกษาครั้งนี้จึงจำลองข้อมูลเพื่อเปรียบเทียบประสิทธิภาพของวิธีการปรับสมดุลข้อมูลภายใต้สถานการณ์จำลองที่แตกต่างกัน ระหว่างวิธีสุ่มเกินและวิธีผสมผสานในการปรับสมดุลข้อมูล และจำแนกประเภทข้อมูลปรับสมดุลแล้วด้วยเทคนิค ได้แก่ แรนดอมฟอร์เรส (Random Forest) การถดถอยโลจิสติก (Logistic Regression) และซัพพอร์ตเวคเตอร์แมชชีน (Support Vector Machine) ภายใต้เงื่อนไขที่อัตราส่วนของจำนวนตัวแปรทำนายที่เป็นตัวแปรจัดประเภทมากกว่าตัวแปรต่อเนื่อง โดยการจำลองนี้จะกระทำซ้ำหลายรอบแล้วนำผลที่ได้ไปเปรียบเทียบประสิทธิภาพด้วยค่า Accuracy Sensitivity และ Specificity เพื่อใช้ตัดสินวิธีการปรับสมดุลข้อมูลที่เหมาะสมและมีประสิทธิภาพในการจำแนกกลุ่ม

2. ทฤษฎีและงานวิจัยที่เกี่ยวข้อง

2.1 การปรับสมดุลข้อมูล

ในการศึกษาที่ใช้การวิเคราะห์จำแนกประเภทข้อมูลมาสร้างโมเดล เทคนิคที่ใช้ในการจำแนกประเภทข้อมูลจะมีประสิทธิภาพเมื่อใช้ในการจำแนกประเภทของข้อมูลที่สมดุล ดังนั้นในงานวิจัยนี้จึงทบทวนทฤษฎีและงานวิจัยที่เกี่ยวกับ

ข้อมูลที่ไม่สมดุลและการปรับสมดุลข้อมูล มีรายละเอียดดังต่อไปนี้

2.1.1 ความหมายและลักษณะของข้อมูลไม่สมดุล

ข้อมูลไม่สมดุล หมายถึง ข้อมูลที่มีการกระจายตัวที่ไม่เท่าเทียมกัน หรือข้อมูลซึ่งจำนวนสมาชิกในกลุ่มหลักและกลุ่มรองมีจำนวนไม่เท่ากัน [1]

ลักษณะโดยทั่วไปของข้อมูลไม่สมดุล คือ ข้อมูลที่มีจำนวนข้อมูลของกลุ่มหนึ่งมากกว่าจำนวนข้อมูลของกลุ่มที่เหลือเป็นจำนวนมาก [2], [3] ซึ่งข้อมูลไม่สมดุลนี้จะส่งผลกระทบต่อการจำแนกประเภทข้อมูลทำให้ไม่สามารถจำแนกประเภทข้อมูลของกลุ่มที่มีจำนวนข้อมูลน้อยได้ถูกต้องแม่นยำ ในขณะเดียวกันจะสามารถจำแนกประเภทข้อมูลของกลุ่มที่มีข้อมูลจำนวนมากได้อย่างแม่นยำ โดยข้อมูลกลุ่มที่มีจำนวนมากจะถูกเรียกว่า ข้อมูลกลุ่มหลัก (Majority Class หรือ Negative Class) และข้อมูลกลุ่มที่มีจำนวนน้อยจะถูกเรียกว่า ข้อมูลกลุ่มรอง (Minority Class หรือ Positive Class) [4] ซึ่งข้อมูลที่อยู่ในกลุ่มรองจะเป็นข้อมูลที่งานวิจัยนี้ให้ความสำคัญมากกว่าข้อมูลที่อยู่ในกลุ่มหลัก

2.1.2 วิธีการปรับสมดุลข้อมูล

ปัญหาข้อมูลไม่สมดุลเป็นปัญหาที่นักวิจัยให้ความสนใจเป็นอย่างมาก ซึ่งนักวิจัยเหล่านั้นได้นำเสนอเทคนิควิธีการต่าง ๆ เพื่อนำมาใช้สำหรับแก้ปัญหานี้ [14] เพื่อเพิ่มประสิทธิภาพและความแม่นยำในการจำแนกประเภทข้อมูลของทั้งสองกลุ่ม การแก้ปัญหาข้อมูลไม่สมดุลที่ระดับข้อมูล (Data Level Solutions) จะเป็นการแก้ปัญหาในขั้นตอนก่อนการประมวลผล (Preprocessing Stage) ซึ่งจะเกี่ยวข้องกับข้อมูลโดยตรง โดยจะปรับข้อมูลที่ไม่สมดุลให้กลายเป็นข้อมูลสมดุลด้วยเทคนิคการสุ่มเลือกข้อมูล (Data Sampling Technique) ซึ่งเทคนิคการสุ่มเลือกที่ใช้ในงานวิจัยนี้ ได้แก่

วิธีสุ่มเกิน (Over Sampling Methods) เป็นวิธีที่ใช้ในการเพิ่มจำนวนข้อมูลที่อยู่ในกลุ่มรองให้มีจำนวนใกล้เคียงหรือเท่ากับจำนวนข้อมูลที่อยู่ในกลุ่มหลัก ซึ่งการเพิ่มข้อมูลนั้นจะเพิ่มโดยการสุ่มเลือกจากข้อมูลเดิม หรือสร้างข้อมูลขึ้นมาใหม่จากตัวอย่างที่มีอยู่เดิม ได้แก่ วิธี Random Over Sampling วิธี SMOTE และวิธี ADASYN เป็นต้น [11], [12], [15], [16]

วิธีผสมผสาน (Hybrid Methods) เป็นวิธีการที่จะสุ่มลดจำนวนข้อมูลจากกลุ่มหลัก และสุ่มเพิ่มข้อมูลจากกลุ่มรองเพื่อให้มีจำนวนข้อมูลของกลุ่มหลักใกล้เคียงหรือเท่ากับจำนวนข้อมูลในกลุ่มรอง ได้แก่ วิธี SMOTE-ENN วิธี SMOTE+TomekLinks เป็นต้น [11], [12], [15], [16]

นักวิจัยได้นำเทคนิคการสุ่มเลือกข้อมูลมาใช้เพื่อปรับข้อมูลให้มีความสมดุล หลังจากนั้นจะนำข้อมูลที่สมดุลนั้นไปทำงานร่วมกับเทคนิควิธีอื่น ๆ ซึ่งจะเห็นได้จากงานวิจัยของ Qian และคณะ [17] ที่ได้ทำการเพิ่มข้อมูลที่อยู่ในกลุ่มรองด้วยวิธีสุ่มเกิน และลดข้อมูลที่อยู่ในกลุ่มหลักด้วยวิธีการสุ่มลดโดยที่อัตราการสุ่มเลือกจะกำหนดด้วยอัตราส่วนของจำนวนข้อมูลกลุ่มรองและจำนวนข้อมูลกลุ่มหลัก หลังจากนั้นจะนำข้อมูลที่มีความสมดุลไปทำการจำแนกประเภทข้อมูลด้วยวิธีการเรียนรู้ร่วมกันแบบแบ็กกิง และงานวิจัยของ Dubey และคณะ [18] ที่นำเทคนิคการสุ่มเลือกข้อมูลทั้งแบบสุ่มเกิน สุ่มลด และวิธีผสมผสานมาใช้งานร่วมกับเทคนิคการเรียนรู้ร่วมกันและการคัดเลือกคุณลักษณะ (Feature Selection) ด้วยชุดข้อมูลโรคอัลไซเมอร์ (Alzheimer's Disease) เป็นต้น

2.2 เทคนิคที่ใช้ในการจำแนกข้อมูล

แรนดอมฟอร์เรส (Random Forest) เกิดจากการรวมกลุ่มกันของโครงสร้างต้นไม้ [19], [20] ซึ่งค่าความคลาดเคลื่อนโดยรวมของป่าไม้จะถูกเปลี่ยนเป็นค่าลิมิต ทำให้จำนวนต้นไม้ในป่ามีเพิ่มขึ้น ค่าความคลาดเคลื่อนโดยรวมจะขึ้นกับความมั่นคงของต้นไม้แต่ละต้นและความสัมพันธ์กันระหว่างต้นไม้แต่ละต้น มีหลักการคือจะใช้วิธีการสุ่มเลือกคุณสมบัติเพื่อการแบ่งแยกโหนด ทำให้ค่าความผิดพลาดลดลง ซึ่งสามารถนำการสร้างแบบจำลองที่ใช้ต้นไม้หลาย ๆ ต้นมาใช้ในการประมวลผลเพื่อตัดสินใจ

การวิเคราะห์การถดถอยโลจิสติก (Logistic Regression Analysis) [21] เป็นการวิเคราะห์ที่มีเป้าหมายเพื่อทำนายโอกาสที่จะเกิดเหตุการณ์ที่สนใจ โดยเป็นเทคนิคการวิเคราะห์สถิติเชิงคุณภาพ ซึ่งมีตัวแปรตามเป็นตัวแปรเชิงคุณภาพ การวิเคราะห์การถดถอยโลจิสติกแบ่งเป็น 2 ประเภท คือ การวิเคราะห์การถดถอยโลจิสติกทวิ (Binary Logistic Regression Analysis) และการวิเคราะห์การถดถอยโลจิสติกพหุกลุ่ม (Multinomial Logistic Regression Analysis) โดยในงานวิจัยนี้ใช้การวิเคราะห์การถดถอยโลจิสติกทวิ เนื่องจากมีตัวแปรตามแบ่งออกเป็น 2 กลุ่มย่อย ซึ่งมี 2 ค่า คือ 0 และ 1

ซัพพอร์ตเวคเตอร์แมชชีน (Support Vector Machine) [22] จัดเป็นการเรียนรู้ของเครื่องที่ต้องมีตัวอย่างในการเรียนรู้ (Supervised Learning) ที่มีความสามารถในการจัดกลุ่มข้อมูล

(Classification) โดยอาศัยหลักการหาสัมประสิทธิ์ของสมการเพื่อสร้างเส้นแบ่งแยกกลุ่มข้อมูลที่ถูกป้อนเข้าสู่กระบวนการสอนให้ระบบเรียนรู้โดยเน้นไปยังเส้นที่แบ่งแยกกลุ่มข้อมูลได้ดีที่สุด (Optimal Separating Hyperplane)

2.3 การจำลองมอนติคาโล

การจำลองมอนติคาโล (Monte Calo Simulation) เป็นเทคนิคการสร้างแบบจำลองโดยใช้หลักการสุ่มตัวอย่างให้เหมาะสมกับปัญหา และทำซ้ำจำนวนมาก โดยยิ่งทำซ้ำมากขึ้นผลลัพธ์ที่ได้จะยิ่งมีความเป็นมาตรฐาน [23] ในงานวิจัยนี้ใช้การสุ่มแบบแจกแจงปกติ (Normal Distribution) การสุ่มแบบแจกแจงเส้น (Uniform Distribution) และการสุ่มแบบทวินาม (Binomial Distribution) และทำซ้ำแต่ละกรณีจำนวน 500 รอบ

2.4 อัตราออด

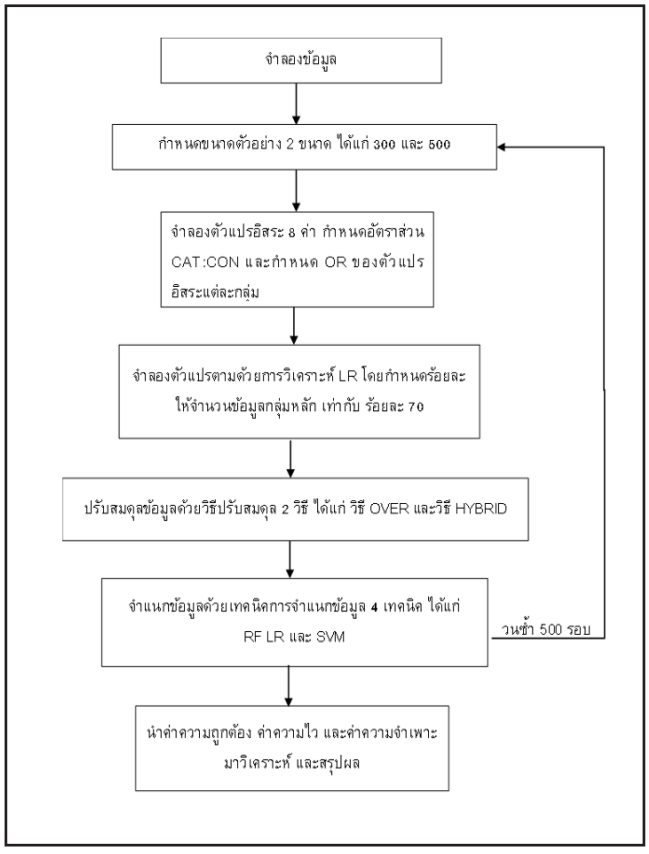
อัตราออด (Odds Ratio) คือ อัตราส่วนระหว่างโอกาสที่จะเกิดเหตุการณ์ที่สนใจกับโอกาสที่จะไม่เกิดเหตุการณ์ที่สนใจ และหมายถึง โอกาสที่จะเกิดเหตุการณ์ที่สนใจเป็นกี่เท่าของโอกาสที่จะไม่เกิดเหตุการณ์ที่สนใจ [24]

3. วิธีดำเนินการวิจัย

ข้อมูลที่ใช้ในการวิจัยได้จากการจำลองด้วยวิธีการมอนติคาร์โล (Monte Carlo Simulation) ใช้การเขียนโปรแกรม R เวอร์ชั่น 4.2.3 ในการวิเคราะห์ข้อมูลและประมวลผลตามขั้นตอนในภาพที่ 1 โดยทำการจำลองข้อมูลด้วยโมเดลการถดถอยโลจิสติกที่กำหนดให้จำนวนกลุ่มหลักมีร้อยละ 70 และจำนวนตัวแปรทำนายมี 8 ตัวแปร โดยจะแบ่งออกเป็น 2 กลุ่ม คือตัวแปรจัดประเภทและตัวแปรต่อเนื่อง นอกจากนี้การจำลองข้อมูลทำภายใต้สถานการณ์ที่แตกต่างกัน 5 เงื่อนไข ประกอบด้วย (1) วิธีปรับสมดุลข้อมูล 2 วิธี ได้แก่ วิธีสุ่มเกิน (OVER) และวิธีผสมผสาน (HYBRID) (2) เทคนิคการจำแนก 3 เทคนิค ได้แก่ แรนดอมฟอร์เรส (RF) การถดถอยโลจิสติก (LR) และซัพพอร์ตเวคเตอร์แมชชีน (SVM) (3) ขนาดตัวอย่าง 2 ระดับ ได้แก่ 300 และ 500 (4) อัตราส่วนของจำนวนตัวแปรทำนายที่เป็นตัวแปรจัดประเภท (CAT) ต่อตัวแปรต่อเนื่อง (CON) 2 ระดับ คือ 5:3 และ 6:2 เนื่องจากในงานวิจัยนี้สนใจศึกษาในขอบเขตของการมีจำนวนตัวแปรทำนายที่เป็นตัวแปรจัดประเภทมากกว่าตัวแปรต่อเนื่อง ทั้งนี้จำนวนของตัวแปรจัดประเภทเป็นปัจจัยหนึ่งที่มีผลต่อการพิจารณาเลือกวิธีการปรับสมดุลข้อมูล โดยวิธีการปรับสมดุลข้อมูลบางวิธีไม่สามารถใช้ปรับสมดุลข้อมูลที่มีจำนวนตัวแปรจัดประเภทจำนวนมากได้ และ (5) อัตราออด (Odds Ratio: OR) ของตัวแปรทำนาย โดยพิจารณาจาก 4 กรณี ดังนั้นการจำลองข้อมูลจากเงื่อนไขที่พิจารณาพบว่ามีทั้งสิ้น 96 เงื่อนไข (2x3x2x2x4) ทำซ้ำ 500 รอบ

ในการกำหนดค่า OR ของตัวแปรทำนายแต่ละกลุ่มจะศึกษาใน 2 ระดับ คือ ให้ค่า OR อยู่ระหว่าง [1, 2) แทนด้วย L และ [2, 3) แทนด้วย H สามารถแบ่งได้ 4 กรณีดังนี้

กรณี	CAT	CON	เขียนแทนด้วย
1	L	L	LL
2	L	H	LH
3	H	L	HL
4	H	H	HH



***ภาพที่ 1** ขั้นตอนการจำลองข้อมูล*

ชุดข้อมูลที่ได้จะมีตัวแปรทำนาย 8 ตัวแปร แบ่งเป็น 2 กลุ่ม คือกลุ่มตัวแปรจัดประเภทและกลุ่มตัวแปรต่อเนื่องในอัตราส่วน 5:3 และ 6:2 ตัวแปรทำนายแต่ละกลุ่มจะมีค่า OR ในช่วง [1,2) หรือ [2,3) และมีตัวแปรตามเป็นตัวแปรจัดประเภทซึ่งมีผลลัพธ์ที่เป็นไปได้ 2 ค่า โดยมีจำนวนข้อมูลกลุ่มหลักคิดเป็นร้อยละ 70 ดังในภาพที่ 2

การวัดประสิทธิภาพการปรับสมดุลข้อมูลจะพิจารณาจากค่าความถูกต้อง (accuracy) ค่าความไว (sensitivity) และค่าความจำเพาะ (specificity) ค่าดังกล่าวจะคำนวณได้จากสมการ ต่อไปนี้

ค่าความถูกต้อง (accuracy) คือ การแสดงการวัดที่ได้มีความถูกต้องในรูปอัตราส่วน [25]

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

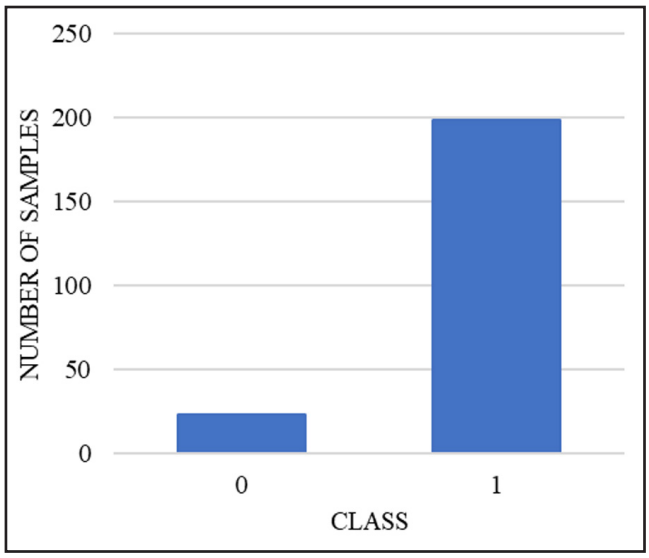
ค่าความไว (sensitivity) คือ สัดส่วนของผลบวกที่เป็นจริงสำหรับภาวะนั้น ๆ [16]

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

ค่าความจำเพาะ (specificity) คือ สัดส่วนของผลลบที่เป็นจริงสำหรับภาวะนั้น ๆ [16]

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

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



ภาพที่ 2 *การกระจายข้อมูลของตัวแปรตาม*

ตารางที่ 1 *ผลการเปรียบเทียบประสิทธิภาพของเทคนิคการจำแนกข้อมูลที่ถูกปรับสมดุลแต่ละวิธี เมื่ออัตราส่วนของจำนวนตัวแปรจัดประเภทและตัวแปรต่อเนื่อง คือ 5:3*

ขนาดตัวอย่าง	อัตราออด		วิธีสุ่มเกิน			วิธีผสมผสาน		
	ตัวแปรจัดประเภท	ตัวแปรต่อเนื่อง	LR	RF	SVM	LR	RF	SVM
ค่าความถูกต้อง								
300	L	L	0.875	**0.999**	0.960	0.919	**0.988**	0.971
	L	H	0.909	**0.995**	0.983	0.872	**0.995**	0.981
	H	L	0.903	**0.992**	**0.992**	0.943	**0.994**	0.985
	H	H	0.926	**0.996**	0.982	0.919	**0.999**	0.985
	ค่าเฉลี่ย		0.903	**0.996**	0.979	0.913	**0.994**	0.981
500	L	L	0.952	**0.998**	0.988	0.948	**0.996**	0.991
	L	H	0.923	**0.996**	0.982	0.965	**0.998**	0.985
	H	L	0.942	**0.983**	0.976	0.948	**0.985**	0.974
	H	H	0.925	**0.999**	0.972	0.943	**0.994**	0.988
	ค่าเฉลี่ย		0.936	**0.994**	0.980	0.951	**0.993**	0.985

ตารางที่ 1 *ผลการเปรียบเทียบประสิทธิภาพของเทคนิคการจำแนกข้อมูลที่ถูกปรับสมดุลแต่ละวิธี เมื่ออัตราส่วนของจำนวนตัวแปรจัดประเภทและตัวแปรต่อเนื่อง คือ 5:3 (ต่อ)*

ขนาดตัวอย่าง	อัตราออด		วิธีสุ่มเกิน			วิธีผสมผสาน		
	ตัวแปรจัดประเภท	ตัวแปรต่อเนื่อง	LR	RF	SVM	LR	RF	SVM
ค่าความไว								
300	L	L	0.832	**1.000**	0.944	0.888	**1.000**	0.953
	L	H	0.896	**0.993**	**1.000**	0.870	**0.997**	0.981
	H	L	0.893	**0.998**	0.984	0.962	**0.99**	0.972
	H	H	0.923	**1.000**	0.966	0.907	**1.000**	**1.000**
	ค่าเฉลี่ย		0.886	**0.998**	0.974	0.907	**0.997**	0.9765
500	L	L	0.942	**0.995**	0.975	0.953	**0.990**	0.988
	L	H	0.918	**0.992**	0.968	0.953	**1.000**	0.976
	H	L	0.960	**0.972**	0.956	0.941	**0.972**	0.965
	H	H	0.911	**1.000**	0.972	0.941	**0.989**	0.988
	ค่าเฉลี่ย		0.933	**0.990**	0.968	0.947	**0.988**	0.979
ค่าความจำเพาะ								
300	L	L	0.920	**1.000**	0.978	0.951	0.983	**0.990**
	L	H	0.924	**1.000**	0.966	0.873	**0.992**	0.980
	H	L	0.914	0.982	**1.000**	0.922	0.982	**1.000**
	H	H	0.982	0.993	**1.000**	0.932	**0.996**	0.970
	ค่าเฉลี่ย		0.923	**0.994**	0.986	0.920	**0.988**	0.985
500	L	L	0.961	**1.000**	**1.000**	0.944	**0.995**	0.994
	L	H	0.927	**0.998**	0.995	0.977	**0.991**	0.994
	H	L	0.926	**1.000**	0.995	0.955	**1.000**	0.983
	H	H	0.983	**0.999**	0.973	0.944	**1.000**	0.988
	ค่าเฉลี่ย		0.938	**0.999**	0.911	0.955	**0.977**	0.990

ตัวทึบ หมายถึง ค่าสูงสุดของค่าความถูกต้อง ค่าความไว และค่าความจำเพาะ

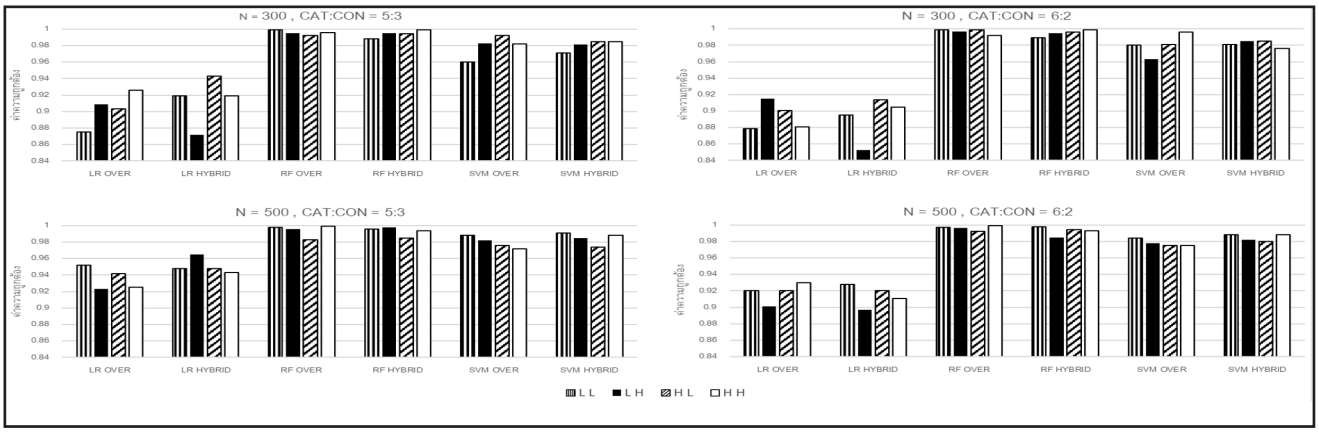
ตารางที่ 2 *ผลการเปรียบเทียบประสิทธิภาพของเทคนิคการจำแนกข้อมูลที่ถูกปรับสมดุลแต่ละวิธี เมื่ออัตราส่วนของจำนวนตัวแปรจัดประเภทและตัวแปรต่อเนื่อง คือ 6:2*

ขนาดตัวอย่าง	อัตราออด		วิธีสุ่มเกิน			วิธีผสมผสาน		
	ตัวแปรจัดประเภท	ตัวแปรต่อเนื่อง	LR	RF	SVM	LR	RF	SVM
ค่าความถูกต้อง								
300	L	L	0.879	**0.999**	0.98	0.895	**0.989**	0.981
	L	H	0.915	**0.997**	0.963	0.853	**0.995**	0.985
	H	L	0.901	**0.999**	0.981	0.914	**0.996**	0.985
	H	H	0.881	0.992	**0.996**	0.905	**0.999**	0.976
	ค่าเฉลี่ย		0.894	**0.997**	0.98	0.892	**0.995**	0.982
500	L	L	0.92	**0.997**	0.984	0.928	**0.998**	0.988
	L	H	0.901	**0.996**	0.978	0.897	**0.985**	0.982
	H	L	0.920	**0.992**	0.975	0.920	**0.994**	0.98
	H	H	0.930	**0.999**	0.975	0.911	**0.993**	0.988
	ค่าเฉลี่ย							
ค่าความไว								
300	L	L	0.874	**1.000**	0.962	0.884	**0.977**	0.964
	L	H	0.894	**1.000**	0.950	0.849	**1.000**	0.982
	H	L	0.922	**0.999**	0.985	0.946	**0.994**	0.991
	H	H	0.886	**0.988**	0.992	0.911	**0.998**	0.955
	ค่าเฉลี่ย		0.894	**0.997**	0.972	0.8975	**0.992**	0.973
500	L	L	0.920	**0.995**	0.968	0.914	**1.000**	0.982
	L	H	0.896	**1.000**	0.962	0.885	**0.973**	0.965
	H	L	0.924	**1.000**	0.950	0.902	**0.996**	0.971
	H	H	0.917	**1.000**	0.970	0.897	**1.000**	0.988
	ค่าเฉลี่ย		0.914	**0.999**	0.962	0.899	**0.992**	0.976

ตารางที่ 2 *ผลการเปรียบเทียบประสิทธิภาพของเทคนิคการจำแนกข้อมูลที่ถูกปรับสมดุลแต่ละวิธี เมื่ออัตราส่วนของจำนวนตัวแปรจัดประเภทและตัวแปรต่อเนื่อง คือ 6:2 (ต่อ)*

ขนาดตัวอย่าง	อัตราออด		วิธีสุ่มเกิน			วิธีผสมผสาน		
	ตัวแปรจัดประเภท	ตัวแปรต่อเนื่อง	LR	RF	SVM	LR	RF	SVM
ค่าความจำเพาะ								
300	L	L	0.885	0.998	**1.000**	0.908	**0.990**	**1.000**
	L	H	0.938	**0.996**	0.977	0.857	**0.995**	0.989
	H	L	0.877	**1.000**	0.977	0.877	**0.998**	0.979
	H	H	0.877	0.999	**1.000**	0.897	**1.000**	**1.000**
	ค่าเฉลี่ย		0.894	**0.998**	0.989	0.885	**0.996**	0.992
500	L	L	0.921	**1.000**	**1.000**	0.942	**1.000**	0.994
	L	H	0.907	0.993	**0.995**	0.908	0.992	**1.000**
	H	L	0.917	0.986	**1.000**	0.937	**0.990**	0.986
	H	H	0.942	**1.000**	0.980	0.925	**0.989**	0.988
	ค่าเฉลี่ย		0.922	**0.995**	0.994	0.928	**0.993**	0.992

ตัวทึบ หมายถึง ค่าสูงสุดของค่าความถูกต้อง ค่าความไว และค่าความจำเพาะ



ภาพที่ 3 *ค่าความถูกต้องของเทคนิคการจำแนกข้อมูลที่ถูกปรับสมดุลแต่ละวิธี*

4. ผลการดำเนินงาน

ผลการวิเคราะห์เปรียบเทียบประสิทธิภาพแสดงดังตารางที่ 1 และ 2

จากตารางที่ 1 อัตราส่วนของจำนวนตัวแปรจัดประเภทและตัวแปรต่อเนื่อง คือ 5:3 พบว่า ขนาดตัวอย่าง 300 ซึ่งจำแนกข้อมูลด้วยแรนดอมฟอร์เรสโดยการปรับสมดุลข้อมูลด้วยวิธีสุ่มเกิน ให้ค่าความถูกต้องสูงสุด คือ 0.996 และค่าความไวสูงสุด คือ 0.998 ส่วนขนาดตัวอย่าง 500 ซึ่งจำแนกข้อมูลด้วยแรนดอมฟอร์เรสโดยการปรับสมดุลข้อมูลด้วยวิธีสุ่มเกินค่าความจำเพาะสูงสุด คือ 0.999

จากตารางที่ 2 อัตราส่วนของจำนวนตัวแปรจัดประเภทและตัวแปรต่อเนื่อง คือ 6:2 พบว่า ขนาดตัวอย่าง 300

ซึ่งจำแนกข้อมูลด้วยแรนดอมฟอร์เรสโดยการปรับสมดุลข้อมูลด้วยวิธีสุ่มเกิน ให้ค่าความถูกต้องสูงสุด คือ 0.997 และค่าความจำเพาะสูงสุด คือ 0.998 ส่วนขนาดตัวอย่าง 500 ซึ่งจำแนกข้อมูลด้วยแรนดอมฟอร์เรสโดยการปรับสมดุลข้อมูลด้วยวิธีสุ่มเกินค่าความไวสูงสุด คือ 0.998

จากภาพที่ 3 พบว่าที่ขนาดตัวอย่างและระดับของอัตราออด (Odds Ratio) เท่ากัน การจำแนกด้วยวิธีถดถอยโลจิสติกส่วนใหญ่มีค่าความถูกต้องลดลง เมื่ออัตราส่วนของจำนวนตัวแปรจัดประเภทและตัวแปรต่อเนื่องเพิ่มจาก 5:3 เป็น 6:2 และถ้าพิจารณาจากขนาดตัวอย่าง พบว่าเมื่อขนาดตัวอย่างเพิ่มขึ้นจะทำให้ค่าความถูกต้องจะเพิ่มขึ้นเล็กน้อยในเกือบทุกวิธีการจำแนกโดยเฉพาะวิธีถดถอยโลจิสติก

5. สรุป

การวิจัยการเปรียบเทียบประสิทธิภาพของแต่ละวิธีการปรับสมดุลข้อมูลภายใต้สถานการณ์จำลองที่แตกต่างกัน พบว่าวิธีที่มีประสิทธิภาพสูงสุด คือ ปรับสมดุลข้อมูลด้วยวิธีสุ่มเกิน (Over Sampling Methods) และใช้การจำแนกด้วยเทคนิคแรนดอมฟอร์เรส (Random Forest) ถ้าพิจารณาเฉพาะเทคนิคการจำแนกข้อมูลจะพบว่า วิธีที่มีประสิทธิภาพสูงสุด คือ แรนดอมฟอร์เรส (Random Forest) รองลงมา คือ ซัพพอร์ตเวคเตอร์แมชชีน (Support Vector Machine) และการถดถอยโลจิสติก (Logistic Regression) ตามลำดับ สอดคล้องกับงานวิจัยของ Zhu, Zhou และ Zhang [26] ที่พิจารณาวิธีที่มีประสิทธิภาพสูงที่สุดในการระบุแหล่งกักเก็บก๊าซจากชั้นหินดินดานคุณภาพสูง โดยใช้การสุ่มเกินร่วมกับวิธีแรนดอมฟอร์เรส วิธีการเพื่อนบ้านใกล้ที่สุด (K-Nearest Neighbors) และวิธีซัพพอร์ตเวคเตอร์แมชชีน พบว่าการสุ่มเกินร่วมกับแรนดอมฟอร์เรส มีประสิทธิภาพสูงที่สุดสามารถเพิ่มค่าความถูกต้องในการทำนายจากร้อยละ 44 เป็นร้อยละ 78 นอกจากนั้นพบว่าค่าความถูกต้องลดลงสำหรับอัตราออดของตัวแปรทำนายส่วนใหญ่ที่มีจำนวนตัวแปรทำนายที่เป็นตัวแปรจัดประเภทเพิ่มขึ้น และค่าความถูกต้องจะเพิ่มขึ้นและค่าความถูกต้องจะเพิ่มขึ้นเล็กน้อย เมื่ออัตราส่วนของจำนวนตัวแปรจัดประเภทและตัวแปรต่อเนื่องเป็นอัตราส่วนเดียวกันแต่ขนาดตัวอย่างเพิ่มขึ้น

ข้อเสนอแนะในการวิจัย การปรับสมดุลของข้อมูลสามารถใช้วิธีอื่น ๆ ได้แก่ SMOTE-ENN และ ADASYN เป็นต้น การวิเคราะห์ข้อมูลอาจเพิ่มเทคนิคการจำแนกให้ครอบคลุมมากขึ้น ได้แก่ วิธีการเพื่อนบ้านใกล้ที่สุด (K-Nearest Neighbors) โครงข่ายประสาทเทียม (Neural Network) เป็นต้น นอกจากนี้สามารถนำโมเดลไปประยุกต์ใช้กับชุดข้อมูลจริงเพื่อทดสอบและเปรียบเทียบต่อไป

6. เอกสารอ้างอิง

[1] B. Jantarakongkul, S. Rasmequan, S. Rimcharoen, P. Kulsasem, K.Chinnasarn, A. Rodtook, P.Voraboot, and J. Onpans. *Optimal Methods for Classification of Highly Imbalanced Datasets,* 2557.

[2] N. V. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer. "SMOTE: synthetic minority over-sampling technique." *Journal of artificial intelligence research,* Vol. 16, pp. 321-357, 2002.

[3] N. V. Chawla, N. Japkowicz, and A. Kotcz. "Special issue on learning from imbalanced data sets." *ACM SIGKDD explorations newsletter,* Vol. 6, No. 1, pp. 1-6, 2004.

[4] M. A. H. Farquad and I. Bose. "Preprocessing unbalanced data using support vector machine." *Decision Support Systems,* Vol. 53, No. 1, pp. 226-233, 2012.

[5] W. Janewit, P. Saran, and T. Sakesan. "Dropout and Persistence Phenomena of Undergraduate Students of Burapha University: The Causal Relationship Model." *Journal of Graduate School Sakon Nakhon Rajabhat University,* Vol. 17, No. 78, pp. 20-29, 2020.

[6] A. Talim. "A Survival Analysis of Dropping Out of Undergraduate Students, Burapha University." *Rajabhat Rambhai Barni Research Journal,* Vol. 14, No. 3, pp. 72-83, 2020.

[7] S. Hussain, N. A. Dahan, F. M. Ba-Alwib, and N. Ribata. "Educational data mining and analysis of students' academic performance using WEKA." *Indonesian Journal of Electrical Engineering and Computer Science,* Vol. 9, No. 2, pp. 447-459, 2018.

[8] Equitable Education Fund, *complete report: Project*

to develop knowledge and role model for caring and supporting the education of children in the street (Children in Street) in Bangkok, 2020.

[9] The Secretariat of the Prime Minister Government House, *Teacher O moves forward to push stability for "Children's teachers on the road" across the country hope to build morale in their work passing on faith and trust to the target group.* Available Online at https://www.thaigov.go.th/news/contents/details/37497, accessed on 12 December 2022.

[10] T. Srisawat, and P. Ruengtip. "A Causal Relationship of Job Burnout Syndrome and Internal Factors to Personnel Performance in Burapha University." *KKBS Journal of Business Administration and Accountancy,* Vol. 5, No. 1, pp. 151-166, 2021.

[11] P. Chujai. *Ensemble Learning for Imbalanced Data Classification Problem.* A Thesis Submitted in Partial Fulfillment of the Requirements for the Degree of Doctor of Philosophy in Computer Engineering, Suranaree University of Technology, 2014.

[12] P. Thongpool, P. Jamrueng, R. Boonrit and S. Sinsomboonthong. "Performance Comparison in Prediction of Imbalanced Datain Data Mining Classification." *Thai Journal of Science and Technology,* Vol. 8, No. 6, pp. 565-584, 2019.

[13] A. Phaeobang and S. Sinsomboonthong. "Adjusting the Imbalanced Data with 5 Classification Methods." *Thai Journal of Science and Technology,* Vol. 9, No. 4, pp. 418-435, 2020.

[14] V. López, A. Fernández, J. G. Moreno-Torres and F. Herrera. "Class imbalance methods for translation initiation site recognition in DNA sequences." *Knowledge-Based Systems,* Vol. 25, No. 1, pp. 22-34, 2012.

[15] K. Chomboon. *Rare Class Discovery Techniques for Highly Imbalanced Data*. A Thesis Submitted in Partial Fulfillment of the Requirements for the Degree of Master of Engineering in Computer Engineering, Suranaree University of Technology, 2012.

[16] K. Suksut. *Imbalanced Data Classification Using Data Improvement and Parameter Optimization with Restarting Genetic Algorithm.* A Thesis Submitted in Partial Fulfillment of the Requirements for the Degree of Doctor of Philosophy in Computer Engineering, Suranaree University of Technology, 2016.

[17] Y. Qian, Y. Liang, M. Li, G. Feng and X. Shi. "A resampling ensemble algorithm for classification of imbalance problems." *Neurocomputing*, Vol. 143, pp. 57-67, 2014.

[18] R. Dubey, J. Zhou, Y. Wang, P. M. Thompson, J. Ye, and Alzheimer's Disease Neuroimaging Initiative. "Analysis of sampling techniques for imbalanced data: An n= 648 ADNI study." *NeuroImage,* Vol. 87, pp. 220-241, 2014.

[19] B. Leo. "Random Forests." *Journal Machine Learning,* Vol. 45, No. 1, pp. 5-32, 2001.

[20] S. Hartshorn. "Machine learning with random forests and decision trees: A Visual guide for beginners." *Kindle edition,* 2016.

[21] Y. Kaiyawan. "Principleand Using Logistic Regression Analysis for Research." *Research and Development Institute, Rajamangala University of Technology Srivijaya,* Vol. 4, No. 1, pp. 1-12, 2012.

[22] C. Cortes and V. Vapnik. "Support-vector networks." *Machine learning,* Vol. 20, pp. 273-297, 1995.

[23] C. Chaiyaphan and K. Ransikarbum. "Study of Factors and Market Layout Using the Analytic Hierarchy Process and Monte Carlo Simulation: A Comparative Study Between Private and Public Markets." *KKU Research Journal (Graduate Studies),* Vol. 21, No. 4, pp. 48-60, 2021.

[24] K. Kittithanusorn and V. Sa-ing. *News Category Classification with Machine Learning Method.* A Master's Project Submitted in Partial Fulfillment of the Requirements for the Degree of Master of Science Data Science), Srinakharinwirot University, 2020.

[25] S. Sripaoraya and S. Sinsomboonthong. "Efficiency Comparison of Data Mining Classification Methods for Chronic Kidney Disease: A Case Study of a Hospital in India." *Thai Journal of Science and Technology (TJST),* Vol. 25, No. 5, pp. 839-853, September-October, 2017.

[26] L. Zhu, X. Zhou and C. Zhang. "Rapid identification of high-quality marine shale gas reservoirs based on the oversampling method and random forest algorithm." Artificial Intelligence in Geosciences, Vol. 2, pp. 76-81, 2021.