การเปรียบเทียบเทคนิคการสุ่มตัวอย่างเพื่อการจำแนกข้อมูลที่ไม่สมดุล

Comparison of sampling techniques for imbalanced data classification

กาญจน์ ณ ศรีธะ[1*], กิตติศักดิ์ เกิดประสพ[2] และนิตยา เกิดประสพ[2]

Karn Nasritha[1*], Kittisak Kerdprasop[2] and Nittaya Kerdprasop[2]

[1] Doctoral Degree Student, School of Information Technology, Institute of Social Technology, Suranaree University of Technology, Nakhonratchasrima, Thailand.

[2] Assoc. Prof., School of Computer Engineering, Institute of Engineering, Suranaree University of Technology, Nakhonratchasrima, Thailand.

karn.n@windowslive.com, kittisakthailand@gmail.com, nittaya.k@gmail.com

Received: 23 June 2017

Accepted: 10 Octoer 2017

Keywords:

ข้อมูลไม่สมดุล, การสุ่มเพิ่มตัวอย่างส่วนน้อย, การสุ่มตัวอย่างซ้ำ, การจำแนกข้อมูล, เทคนิคการรวมกลุ่ม

Imbalance Data, SMOTE, Resample, Classification, Ensemble Technique

บทคัดย่อ: ความไม่สมดุลของข้อมูลเป็นปัญหาในกระบวนการเรียนรู้ของเครื่องสำหรับการจำแนกข้อมูลซึ่งส่งผลให้ประสิทธิภาพในการจำแนกมีค่าต่ำ และพบว่ามีการนำเทคนิคการสุ่มเพิ่มตัวอย่างหลายวิธีสำหรับการแก้ปัญหาค่าประสิทธิภาพต่ำเนื่องจากความไม่สมดุลของข้อมูล งานวิจัยนี้จึงมีวัตถุประสงค์เพื่อเปรียบเทียบเทคนิคการสุ่มตัวอย่างเพื่อการจำแนกข้อมูลที่ไม่สมดุล ในงานวิจัยได้ทำการทดลองกับข้อมูลจำนวน 3 ชุดข้อมูล ใช้เทคนิคการสุ่มเพิ่มตัวอย่างส่วนน้อย เทคนิคการสุ่มลดตัวอย่างส่วนมาก และเทคนิคการสุ่มตัวอย่างซ้ำสำหรับการปรับปรุงข้อมูลที่ไม่สมดุล ใช้เทคนิคต้นไม้ตัดสินใจ คาร์ท เรนดอมฟรอเรส ซัพพอรท์เวกเตอร์แมชชีน และโครงข่ายประสาทเทียม ร่วมกับเทคนิครวมกลุ่มเอดาบูทและถุงจำแนก เพื่อสร้างแบบจำลองสำหรับจำแนกข้อมูล ใช้วิธี 10-fold cross validation เพื่อวัดประสิทธิภาพของแบบจำลอง วัดค่าประสิทธิภาพด้วยค่าความแม่นยำ ค่าความระลึก และค่าเอฟ ผลการวิจัยพบว่าเทคนิคการสุ่มตัวอย่างซ้ำสามารถปรับปรุงข้อมูลที่ไม่สมดุลได้ดีกว่าเทคนิคการสุ่มเพิ่มตัวอย่างส่วนน้อย นอกจากนี้ยังพบว่าแบบจำลองเรนดอมฟรอเรส แบบจำลองเอดาบูทร่วมกับเรนดอมฟรอเรส และแบบจำลองถุงจำแนกร่วมกับเรนดอมฟรอเรส มีค่าประสิทธิภาพการจำแนกที่ดีกับข้อมูลในงานวิจัย

Abstract: Imbalanced data is a problem in the machine learning process for data classification, which results in low classification efficiency. It has also been found that random sampling techniques are used in several ways for solving low performance problems due to data imbalances. This research aims to compare sampling techniques for imbalanced data classification. The research was conducted on three data sets, which are Synthetic minority over-sampling technique, under-sampling technique and resample techniques for Imbalanced data preprocessing. Decision Tree, cart, random forest, support vector machine and artificial neural network algorithms are ensembled with adaboost and bagging algorithms to create models for data classification. Ten-fold cross validation was used to measure model performance. Performance was measured with precision, recall and f-measure. The results showed that resample techniques could improve the imbalanced data better than synthetic minority over-sampling technique. In addition, it was found that the random forest model, the adaboost ensemble with random forest model and the bagging ensemble with random forest model were efficient for data classification in this research.

1. บทนำ

ในปัจจุบันพบว่ามีการใช้การเรียนรู้ของเครื่องในการวิเคราะห์ข้อมูลทางการแพทย์ เช่น งานวิจัยการสร้างแบบจำลองเพื่อพยากรณ์ผู้ป่วยเบาหวาน (Meng et al. 2013) ได้ทำการวิเคราะห์ตัวแปรพื้นฐานในคนไข้กลุ่มที่ยืนยันว่าเป็นเบาหวานจำนวน 735 คน และกลุ่มคนปกติจำนวน 752 คน เปรียบเทียบแบบจำลอง 3 รูปแบบ คือ โครงข่ายประสาทเทียม ต้นไม้ตัดสินใจ และการถดถอยโลจิสติกส์ ในการพยากรณ์การเกิดโรคเบาหวาน ผลสรุป คือ ต้นไม้ตัดสินใจ มีความแม่นยำในการจำแนกสูงสุด ในงานวิจัยระบบช่วยเหลืออัจฉริยะเพื่อการวางแผนการรักษาโรคเรื้อรังโดยใช้เหมืองข้อมูล (เสกสันติ จันทะมงคล และคณะ, 2555) ได้ทำการสร้างแบบจำลองวางแผนการรักษาโรคเรื้อรังประเภทเบาหวานเพื่อช่วยให้แพทย์วินิจฉัยได้แม่นยำขึ้นด้วยต้นไม้ตัดสินใจ วิเคราะห์ข้อมูลการรักษาผู้ป่วยเรื้อรัง จำนวน 6,864 ระเบียน ตัวแปร คือ เพศอายุ ดัชนีมวลกาย ผลการตรวจระดับน้ำตาลในเลือด และผลการตรวจการทำงานของไต มีค่าความถูกต้อง (Correctly) สูงสุดที่ 83.05% ต่ำสุดที่ 16.95% และวัดค่าความคลาดเคลื่อน (Root Mean Squared Error) ได้ 14.37% และงานวิจัยการสร้างแบบจำลองเพื่อพยากรณ์การเกิดแผลที่เท้าของผู้ป่วยโรคเบาหวานโดยใช้เทคนิคเหมืองข้อมูล (บรรจบ ดลกุล และคณะ, 2557) ได้ทำการสร้างแบบจำลองเพื่อพยากรณ์การเกิดแผลที่เท้าของผู้ป่วยโรคเบาหวานโดยใช้เทคนิคต้นไม้ตัดสินใจ กับข้อมูลการเกิดแผลที่เท้าของผู้ป่วยเบาหวาน จำนวน 758 ราย 30 ตัวแปร ผลการวิจัย พบว่า แบบจำลองมีประสิทธิภาพค่าความถูกต้อง 75.90% จากตัวอย่างงานวิจัยดังที่ได้กล่าวมานั้นทำให้ทราบว่าการประยุกต์ใช้การเรียนรู้ของเครื่องกับข้อมูลทางการแพทย์สามารถช่วยวิเคราะห์ข้อมูลที่มีอยู่ให้ทราบถึงปัจจัยที่สำคัญและพยากรณ์การเกิดโรคฯได้ หน่วยงานทางการแพทย์สามารถนำแบบจำลองที่ได้ไปวางแผนการรักษาล่วงหน้าได้อย่างรวดเร็วและแม่นยำยิ่งขึ้น แต่ยังพบว่าในข้อมูลทางการแพทย์มักเป็นข้อมูลที่ไม่สมดุล ซึ่งเป็นข้อมูลที่มีสัดส่วนของกลุ่มข้อมูลไม่สมดุลกัน โดยแบ่งออกเป็นสองส่วน คือ ข้อมูลส่วนน้อยและข้อมูลส่วนมาก ซึ่งเป็นปัญหาที่สำคัญกับข้อมูลทางการแพทย์ หากนำข้อมูลที่ไม่สมดุลนี้ไปสร้างแบบจำลองสำหรับการจำแนกผู้ป่วยและผู้ไม่ป่วย แบบจำลองที่สร้างขึ้นอาจจำแนกผู้ป่วยจริงได้ไม่เป็นที่น่าพอใจหรือไม่มีประสิทธิภาพเพียงพอที่จะนำไปใช้สำหรับการพยากรณ์การเป็นโรคได้ ซึ่งเป็นปัญหาอย่างหนึ่งสำหรับการเรียนรู้ของเครื่อง (ภรัณยา ปาลวิสุทธิ์, 2559)

วิธีการแก้ปัญหาสำหรับข้อมูลที่ไม่สมดุลนี้ นักวิจัยมักใช้วิธีการสุ่มตัวอย่างเพิ่มขึ้นของข้อมูลส่วนน้อยหรือลดข้อมูลส่วนมากลง ก่อนนำข้อมูลเข้าสู่กระบวนการเรียนรู้ของเครื่อง เช่น ในงานวิจัยการจำแนกต้นไม้ตัดสินใจสำหรับชุดข้อมูลไม่สมดุลโดยใช้น้ำหนักต่างกันบนข้อมูลสังเคราะห์ (สุรพงษ์ เชี่ยวสกุลวัฒนา และคณะ, 2557) ได้ใช้เทคนิคการสุ่มเพิ่มตัวอย่างกลุ่มน้อยกับข้อมูลไม่สมดุล 8 ชุด ผลการวิจัยพบว่า เทคนิคการสุ่มตัวอย่างกลุ่มน้อยสามารถเพิ่มประสิทธิภาพของแบบจำลองได้ดีขึ้น ในงานวิจัยการเพิ่มประสิทธิภาพเทคนิคต้นไม้ตัดสินใจบนชุดข้อมูลที่ไม่สมดุลโดยวิธีการสุ่มเพิ่มตัวอย่างกลุ่มน้อยสำหรับข้อมูลการเป็นโรคติดอินเทอร์เน็ต (ภรัณยา ปาลวิสุทธิ์, 2559) ได้ใช้เทคนิคการสุ่มเพิ่มตัวอย่างกลุ่มน้อย กับข้อมูลการเป็นโรคติดอินเทอร์เน็ตซึ่งเป็นข้อมูลที่ไม่สมดุล ผลการวิจัยพบว่า เทคนิคการสุ่มตัวอย่างกลุ่มน้อยสามารถเพิ่มประสิทธิภาพของแบบจำลองได้ดีขึ้นเช่นกัน และในงานวิจัยการพัฒนาแบบจำลองเพื่อการพยากรณ์การรักษาซ้ำของผู้ป่วยโรคจิตเภทโดยเทคนิคเหมืองข้อมูล (วีระยุทธ มายุศิริ และคณะ, 2557) ได้สร้างแบบจำลองเพื่อการพยากรณ์การรักษาซ้ำของผู้ป่วยโรคจิตเภทและใช้เทคนิคการสุ่มเพิ่มตัวอย่างกลุ่มน้อย กับข้อมูลในงานวิจัย พบว่า แบบจำลองมีประสิทธิภาพการพยากรณ์ ค่าความถูกต้องสูงถึง 91.78% ทั้งสามงานวิจัยนี้ พบว่า ใช้เทคนิคต้นไม้ตัดสินใจ สำหรับการจำแนกข้อมูล ประกอบกับการนำเทคนิคการสุ่มตัวอย่างส่วนน้อยเข้ามาช่วยปรับปรุงข้อมูลก่อนนำเข้าสู่กระบวนการเรียนรู้ของเครื่องสามารถช่วยเพิ่มค่าประสิทธิภาพของแบบจำลองได้ดียิ่งขึ้น นอกจากการสุ่มตัวอย่างกลุ่มน้อยแล้วยังพบงานวิจัยที่ใช้เทคนิคการสุ่มลดข้อมูลส่วนมากลงเพื่อให้ข้อมูลมีความสมดุลเท่ากับข้อมูลของกลุ่มข้อมูลส่วนน้อย เช่น งานวิจัยกระบวนการเตรียมข้อมูลสำหรับข้อมูลไม่สมดุลเพื่อเพิ่มประสิทธิภาพในการจำแนก (ภาสพิชญ์ ชูใจ และคณะ, 2557) และงานวิจัยเทคนิคการสุ่มเพิ่มตัวอย่างข้างน้อยสังเคราะห์และเทคนิคการสุ่มลดตัวอย่างข้างมากสำหรับปัญหาความไม่สมดุลระหว่างกลุ่ม (ปณต ทรงวัฒนศิริ, 2553) และยังพบว่ามีเทคนิคการสุ่มตัวอย่าง

ที่เป็นการสุ่มตัวอย่างข้อมูลชุดใหม่จากการสุ่มเลือกชุดข้อมูลเดิมในจำนวนเท่ากัน (Efron, 1987) งานวิจัยที่ใช้เทคนิคนี้ เช่น ข้อมูลเชิงเวลากับการจำแนกประเภทผู้เป็นโรคเบาหวานในประเทศไทย (สมภพ ปฐมนพ และคณะ 2555) จากงานวิจัยข้างต้น พบว่า การใช้เทคนิคการสุ่มตัวอย่างให้กับข้อมูลที่ไม่สมดุล สามารถช่วยเพิ่มประสิทธิภาพการจำแนกข้อมูลที่ไม่สมดุลได้ดี ซึ่งข้อมูลทางการแพทย์ที่ใช้นั้นมีความไม่สมดุลระหว่างผู้ป่วยและผู้ไม่ป่วย ซึ่งเป็นข้อมูลแบบ 2 คลาส และงานวิจัยข้างต้นพบว่าใช้วิธีการสุ่มตัวอย่างเพียงเทคนิคเดียวในงานวิจัยนั้น ซึ่งขาดการเปรียบเทียบเทคนิคการสุ่มตัวอย่างที่พบในงานวิจัยอื่น ๆ

ดังนั้น ผู้วิจัยจึงมีแนวคิดที่จะนำข้อมูลทางการแพทย์ แบบ 2 คลาส จำนวน 3 ชุดข้อมูล ที่มีปัญหาความไม่สมดุลกันระหว่างผู้ป่วยและผู้ไม่ป่วย มาเปรียบเทียบเทคนิคการสุ่มตัวอย่างดังนี้ การสุ่มเพิ่มตัวอย่างส่วนน้อย การลุ่มลดตัวอย่างกลุ่มมาก และการสุ่มตัวอย่างใหม่จากข้อมูลเดิม และใช้เทคนิคการจำแนกข้อมูลดังนี้ เทคนิคต้นไม้ตัดสินใจ, เทคนิคคาร์ท, เทคนิคเรนดอมฟรอเรส, เทคนิคซัพพอร์ทแวกเตอร์แมชชีน และเทคนิคโครงข่ายประสาทเทียม สำหรับเทคนิคการเพิ่มประสิทธิภาพแบบร่วมกัน คือ เทคนิคเอดาบูทและเทคนิคถุงจำแนก เพื่อเปรียบเทียบเทคนิคการสุ่มตัวอย่างและเทคนิควิธีการจำแนกข้อมูลที่มีประสิทธิภาพที่เหมาะสมกับข้อมูลในงานวิจัยนี้

วัตถุประสงค์และขอบเขตงานวิจัย

งานวิจัยการเปรียบเทียบเทคนิคการสุ่มตัวอย่างเพื่อการจำแนกข้อมูลที่ไม่สมดุลมีวัตถุประสงค์และขอบเขตงานวิจัย คือ เพื่อเปรียบเทียบวิธีการสุ่มเพิ่มตัวอย่างข้อมูลส่วนน้อย การสุ่มลดตัวอย่างข้อมูลส่วนมาก และการสุ่มตัวอย่างซ้ำ เพื่อการจำแนกข้อมูลทางการแพทย์ที่ไม่สมดุลแบบ 2 คลาส จำนวน 3 ชุดข้อมูล ดังนี้ Breast Cancer, Pima Diabetes และ Diabetic foot ใช้เทคนิคการจำแนก เทคนิคต้นไม้ตัดสินใจ, เทคนิคคาร์ท, เทคนิคเรนดอมฟรอเรส, เทคนิคซัพพอร์ทแวกเตอร์แมชชีน และเทคนิคโครงข่ายประสาทเทียม และเทคนิคช่วยเพิ่มประสิทธิภาพแบบร่วมกัน คือ เทคนิคเอดาทูบและเทคนิคถุงจำแนก เพื่อทดสอบการจำแนกแต่ละชุดข้อมูล และวัดค่าประสิทธิภาพของแบบจำลองที่สร้างขึ้นด้วยค่าความแม่นยำ ค่าความเที่ยงตรง และค่าเอฟ

2. ทฤษฎีที่เกี่ยวข้อง

2.1 การเรียนรู้ของเครื่อง

การเรียนรู้ของเครื่อง (Machine Learning) คือ การนำข้อมูลในอดีตมาเรียนรู้ แล้วสร้างกฎหรือแบบจำลองจากข้อมูลที่ได้เรียนรู้มานั้นไปพยากรณ์หรือจำแนกข้อมูลในอนาคตได้ การเรียนรู้ของเครื่องจึงถูกนำมาใช้อย่างแพร่หลายในปัจจุบัน แต่พบว่ายังมีปัญหาประสิทธิภาพของแบบจำลองที่ได้ยังไม่เป็นที่น่าพอใจ เนื่องจากการนำข้อมูลที่ไม่สมดุลมาใช้ในการเรียนรู้ข้องเครื่อง

2.2 ข้อมูลไม่สมดุล

ข้อมูลที่ไม่สมดุล คือ ข้อมูลที่มีจำนวนข้อมูลของบางกลุ่มมีมากกว่า จำนวนข้อมูลของอีกกลุ่มอยู่เป็นจำนวนมาก (Chawla, 2005) โดยแบ่งเป็นข้อมูลส่วนมาก (Majority Class) และข้อมูลส่วนน้อย (Minority Class) (Chawla et al., 2002) ข้อมูลที่ไม่สมดุลจะส่งผลต่อการจำแนกประเภทข้อมูลของกลุ่มข้อมูลส่วนน้อยได้ไม่ถูกต้อง หรือถูกต้องน้อย แต่ในขณะเดียวกันจะสามารถจำแนกประเภทข้อมูลของกลุ่มข้อมูลส่วนมากได้ถูกต้องกว่า ข้อมูลที่ไม่สมดุลนั้นเกิดขึ้นได้กับข้อมูลทั่วไป สาเหตุของการเกิดความไม่สมดุลนั้น เช่น ข้อมูลไม่สมดุลที่เกิดจากธรรมชาติของข้อมูลเอง เช่นข้อมูลการวินิจฉัยผู้ป่วยทางการแพทย์ ซึ่งข้อมูลผู้ป่วยจะมีน้อยกว่าข้อมูลของผู้ที่เป็นปกติ เป็นต้น (ภาสพิชญ์ ชูใจ, 2557)

2.3 เทคนิคการสุ่มตัวอย่าง

การสุ่มเพิ่มตัวอย่างกลุ่มส่วนน้อย (Synthetic Minority Over-sampling Technique: SMOTE) (Chawla et al., 2002) เป็นวิธีการเพิ่มจำนวนข้อมูลกลุ่มน้อย เพื่อให้มีจำนวนข้อมูลใกล้เคียงกับข้อมูลกลุ่มมาก โดยสุ่มข้อมูลขึ้นมาหนึ่งตัว และหาค่าระยะห่างระหว่างข้อมูลที่เลือกกับข้อมูลที่ใกล้เคียงที่สุด ดังสมการที่ 1

$$X_{new} = x_i + (\hat{x}_i - x_i) \times \delta \quad (1)$$

โดยที่

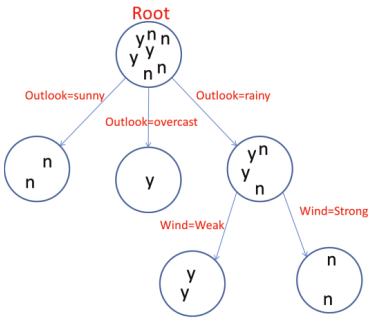
X_{new} คือ ข้อมูลใหม่
x_i คือ ข้อมูลที่สุ่มในตอนแรก
$\hat{x}_i$ คือ ข้อมูลที่สุ่มเพิ่มมาอีก
δ คือ ค่าสุ่มตั้งแต่ 0-1

การสุ่มตัวอย่างซ้ำ (Resampling) (Efron, 1987) เป็นวิธีการสุ่มข้อมูลตัวอย่างทางสถิติอย่างหนึ่ง เพื่อสร้างข้อมูลชุดใหม่จากข้อมูลเดิมที่มีเพียงชุดเดียว โดยการสุ่มตัวอย่างแบบคืนที่ (Resampling with Replacement) การสุ่มตัวอย่างจะทำการสุ่มตัวอย่างทีละ 1 ค่าจำนวน n ครั้ง จากชุดของตัวอย่าง X1, X2 ,X3,...,Xn โดยค่าที่ได้จะคืนกลับไปในชุดตัวอย่างก่อนจะมีการสุ่มตัวอย่างครั้งต่อไป และกำหนดให้ X1*, X2*, X3* ,...,Xn* เป็นชุดของข้อมูลตัวอย่างขนาด n ที่สุ่มได้ ซึ่งเรียกชุดของตัวอย่างนี้ว่าตัวอย่างบูทสแทรป (Bootstrap sample)

การสุ่มลดตัวอย่างกลุ่มส่วนมาก เป็นวิธีการสุ่มลดข้อมูลกลุ่มมากให้มีขนาดเท่ากับข้อมูลกลุ่มน้อย โดยการสุ่มลดแบบสุ่ม

2.4 เทคนิคการจำแนกข้อมูล

เทคนิคต้นไม้ตัดสินใจ (Decision Tree : J48) (Quinlan, 1986) เป็นเทคนิคการพยากรณ์ที่แสดงผลลัพธ์คล้ายต้นไม้จึงนิยมเรียกกันว่า ต้นไม้ตัดสินใจ โดยโหนดที่อยู่บนสุดของต้นไม้เรียกว่าโหนดราก (root node) ส่วนโหนดภายในต้นไม้เรียกว่าโหนด (node) ถัดมาคือ กิ่งของต้นไม้ (branch) และใบ (leaf) อยู่ล่างสุดของต้นไม้ตัดสินใจ แสดงถึงกลุ่มของข้อมูลที่ต้นไม้สามารถจำแนกได้ (class) หรือผลลัพธ์ที่ได้จากการพยากรณ์ ดังแสดงในภาพประกอบที่ 1 โดยในงานวิจัยนี้ได้เลือกใช้เทคนิค C4.5 (Quinlan, 1993) ในโปรแกรม Weka > Classify > Classifier > trees > J48



ภาพประกอบที่ 1 ต้นไม้ตัดสินใจ

ต้นไม้ตัดสินใจมีการคำนวณค่าของแต่ละคุณลักษณะหรือปัจจัย ดังค่าต่อไปนี้

ค่า Gini Index คือ ค่าที่บ่งบอกว่าคุณลักษณะหรือปัจจัยใดควรนำมาใช้เป็นคุณลักษณะในการแบ่งกลุ่มของ C4.5 ดังสมการที่ 2

$$Gini(t_i) = 1 - \sum_{i=1}^{N} [p(t_i)]^2 \quad (2)$$

ค่า Entropy คือ ค่าคาดคะเนของข้อมูล ดังสมการที่ 3

$$Entropy(t_1) = 1 - \sum_{i=1}^{N} [p(t_i)] \log_2 \quad (3)$$

โดยที่

t_i คือ คุณลักษณะที่นำมาวัดค่า Entropy
$p(t_i)$ คือ สัดส่วนของจำนวนสมาชิกของกลุ่ม i กับ จำนวน สมาชิกทั้งหมดของกลุ่มตัวอย่าง

เทคนิคคาร์ท (CART) (Breiman et al, 1984) เป็นเทคนิคที่สร้างต้นไม้ตัดสินใจโดยใช้หลักการแบบ Binary คือแต่ละโหนดจะมีกิ่งเพียง 2 กิ่ง ในงานวิจัยนี้เลือกใช้เทคนิค Simple Cart ในโปรแกรม Weka > Classify > Classifier > trees > SimpleCart

เทคนิคเรนดอมฟรอเรส (Random Forest : RF) (Breiman, 2001) เป็นเทคนิคแบบต้นไม้ตัดสินใจ โดยการสุ่มข้อมูลไปสร้างต้นไม้ขึ้นมาหลายๆชุด ส่วนข้อมูลที่ไม่ถูกสุ่มเลือก (Out of Bag) จะถูกนำไปใช้สำหรับทดสอบความถูกต้องของต้นไม้ที่สร้างขึ้น ในงานวิจัยนี้เลือกใช้เทคนิคเรนดอมฟรอเรสแบบ Random Forest ในโปรแกรม Weka > Classify > Classifier > trees > RandomForest

เทคนิคซัพพอรท์เวกเตอร์แมชชีน (Support Vector Machine : SVM) (Cortes and Vapnik, 1995) เป็นเทคนิคการจำแนกข้อมูลโดยใช้หลักการเชิงเส้น สร้างเส้นสมมุติขึ้น (Hyper Plane) เพื่อจำแนกประเภทของข้อมูลให้ชัดเจนมากที่สุดด้วยระยะห่างจากเส้นสมมุติ (Margin Width) ซึ่งงานวิจัยนี้เลือกใช้เทคนิคซัพพอร์ทเวกเตอร์แมชชีนแบบ LibSVM และใช้สมการสร้างเส้นสมมุติแบบ Radial Basis ในโปรแกรม Weka > Classify > Classifier > functions > LipSVM

โครงข่ายประสาทเทียม (Artificial Neural Network : MLP) (Larose, 2014) เป็นเทคนิคการเรียนรู้ของเครื่องด้วยการเลียนแบบการทำงานของสมองโดยมีชั้นรับข้อมูล (Input layer) ชั้นซ้อน (Hidden Layer) และ

ชั้นแปรผลข้อมูล (Output layer) โดยจะมีฟังก์ชั้นผลรวม (Summation Function) สำหรับการแปรผล ในงานวิจัยนี้เลือกใช้เทคนิคโครงข่ายประสาทเทียมแบบ Multilayer Perceptron กำหนดรอบการเรียนรู้ 500 รอบ และ กำหนดเป้าหมายการเรียนรู้ที่ 0.5 ในโปรแกรม Weka > Classify > Classifier > functions > MultilayerPerceptron

จากเทคนิคการจำแนกข้อมูลที่กล่าวมาข้างต้นเป็นเทคนิคที่มีค่าประสิทธิภาพที่ดีสำหรับการจำแนกข้อมูลของงานวิจัยที่กล่าวมาข้างต้น ผู้วิจัยจึงสนใจนำเทคนิควิธีที่กล่าวข้างต้นมาเปรียบเทียบกันเพื่อให้ทราบถึงเทคนิควิธีที่เหมาะสมกับข้อมูลในงานวิจัยนี้

2.5 เทคนิคช่วยเพิ่มประสิทธิภาพ

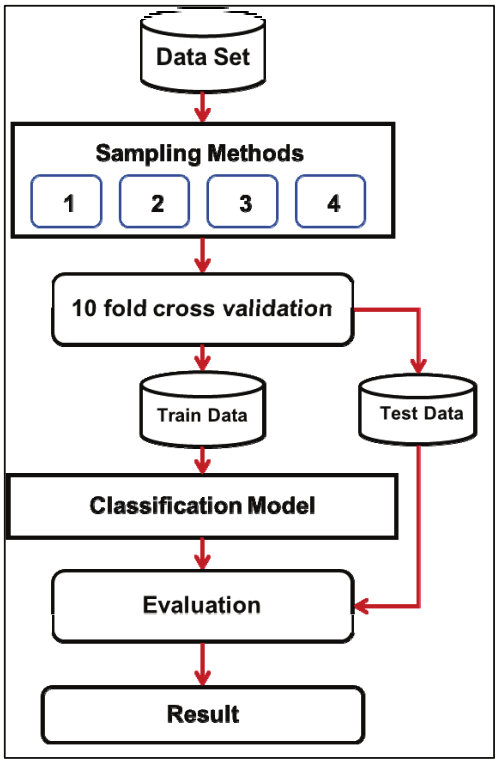
เทคนิคเอดาบูท (Adaboost : A) (Freund and Schapire, 1997) เป็นเทคนิคสำหรับช่วยเพิ่มประสิทธิภาพในการสร้างแบบจำลองโดยทำงานร่วมกับเทคนิคการเรียนรู้ต่าง ๆ เช่น เทคนิคเอดาบูทร่วมกับต้นไม้ตัดสินใจ เริ่มด้วยการสร้างชุดข้อมูลเรียนรู้ (Train data) จากการกำหนดค่าน้ำหนัก (Weight) ให้กับ ข้อมูลแต่ละตัว จากนั้นนำข้อมูลเรียนรู้ไปสร้างแบบจำลองกับเทคนิคการเรียนรู้ ซึ่งในแต่ละรอบที่มีการสร้างแบบจำลองใหม่ค่าน้ำหนักจะถูกเปลี่ยนไป คือ ถ้าแบบจำลองตอบผิดพลาด ข้อมูลตัวนั้นจะถูกเพิ่มค่าน้ำหนัก และถ้าแบบจำลองตอบถูกข้อมูลตัวนั้นก็จะถูกลดค่าน้ำหนักลง (เดช ธรรมศิริ และ พยุง มีสัจ, 2554) ซึ่งเทคนิคนี้นอกจากจะช่วยเพิ่มประสิทธิภาพของแบบจำลองแล้วยังสามารถใช้ได้ดีกับข้อมูลที่ไม่สมดุล ในโปรแกรม Weka > Classify > Classifier > meta > AdaboostM1

เทคนิคถุงจำแนก (Bagging : B) (Breiman, 1996) เป็นเทคนิคสำหรับช่วยเพิ่มประสิทธิภาพในการสร้างแบบจำลองโดยทำงานร่วมกับเทคนิคการเรียนรู้ต่าง ๆ เช่น ต้นไม้ตัดสินใจ โดยเริ่มจากการสุ่มเลือกข้อมูลด้วยหลักการสุ่มแบบ Bootstrap ขึ้นมาหลายๆ ชุด ในจำนวนเท่าข้อมูลเดิม แล้วนำข้อมูลที่สุ่มขึ้นมาไปเพื่อสร้างแบบจำลองทุกชุดข้อมูล จากนั้นใช้หลักการโหวด (Vote) เพื่อหาแบบจำลองที่สามารถตอบข้อมูลทดสอบได้ดีที่สุดมาเป็นแบบจำลองสำหรับการจำแนกข้อมูลชุดนั้น ในโปรแกรม Weka > Classify > Classifier > meta > Bagging

จากเทคนิคช่วยเพิ่มประสิทธิภาพข้างต้นเป็นเทคนิควิธีที่ใช้เพื่อเพิ่มประสิทธิภาพให้กับเทคนิคการเรียนรู้ต่าง ๆ ผู้วิจัยจึงสนใจนำเทคนิคช่วยเพิ่มประสิทธิภาพข้างต้นมาช่วยเพิ่มประสิทธิภาพให้กับให้กับเทคนิคการเรียนรู้ที่กล่าวข้างต้น และเปรียบเทียบกันเพื่อให้ทราบถึงเทคนิคช่วยเพิ่มประสิทธิภาพที่เหมาะสมกับข้อมูลในงานวิจัยนี้

3 ขั้นตอนการวิจัย

งานวิจัยนี้เป็นการเพิ่มประสิทธิภาพของการจำแนกข้อมูลที่ไม่สมดุลด้วยการใช้เทคนิคการสุ่มตัวอย่าง ในการทดลองได้ใช้ข้อมูลที่ไม่สมดุลจำนวน 3 ชุดข้อมูล มีขั้นตอนการวิจัยดังนี้ 1) การเตรียมชุดข้อมูล 2) การสุ่มตัวอย่างด้วยเทคนิควิธีที่ต่างกัน 3) การสร้างแบบจำลองด้วยเทคนิคที่ต่างกัน 4) การทดสอบประสิทธิภาพของแบบจำลอง และ5) การสรุปผลการทดลอง สามารถแสดงขั้นตอนได้ในภาพประกอบที่ 2



ภาพประกอบที่ 2 วิธีวิจัย

3.1 ข้อมูลวิจัย

การวิจัยนี้ใช้ข้อมูลทดสอบจำนวน 3 ชุดข้อมูล คือ ชุดข้อมูล Breast Cancer, Pima Diabetes และ Diabetic foot จากฐานข้อมูล UCI สามารถแสดงจำนวนตัวอย่างและจำนวนตัวแปรในตารางที่ 1

ตารางที่ 1 ชุดข้อมูลในงานวิจัย

ชุดข้อมูล	จำนวน				Imbalance ratio (maj/min)
	Instances	Attributes	Majority (N)	Minority (Y)	
Breast Cancer (BR)	277	10	196	81	2.419
Pima Diabetes (PI)	768	9	500	268	1.865
Diabetic foot (DM)	184	33	125	59	2.118

ตารางที่ 2 แสดงชุดข้อมูลหลังการเตรียมข้อมูล

ชุดข้อมูล	จำนวน		
	Instances	N	Y
BR-ORI	277	196	81
BR-SM	392	196	196
BR-SP	162	81	81
BR-RE	277	196	81
PI-ORI	768	500	268
ชุดข้อมูล	จำนวน		
	Instances	N	Y
PI-SM	1000	500	500
PI-SP	536	268	268
PI-RE	768	500	268
DM-ORI	184	125	59
DM-SM	250	125	125
DM-SP	108	59	59
DM-RE	184	125	59

จากตารางที่ 1 พบว่า ทั้ง 3 ชุดข้อมูลมีจำนวนข้อมูลในคลาส N มากกว่า จำนวนข้อมูลในคลาส Y หรือ มีจำนวนผู้เป็นปกติมากกว่าจำนวนผู้ป่วยจริง ประมาณหนึ่งเท่าตัว หมายความว่า ข้อมูลทั้ง 3 ชุด มีความไม่สมดุลกันระหว่างข้อมูลผู้ไม่ป่วย และผู้ป่วยจริง

3.2 การเตรียมข้อมูล

งานวิจัยนี้ได้เตรียมข้อมูลสำหรับนำไปทดสอบการสุ่มตัวอย่างด้วยวิธีที่แตกต่างกัน ดังนี้ 1.ข้อมูลเดิม (Ori) 2.สุ่มเพิ่มข้อมูลส่วนน้อย (SM) 3.สุ่มลดข้อมูลส่วนมาก (SP) และ 4.การสุ่มตัวอย่างแบบสุ่มซ้ำ (RE) หลังจากนั้นจึงใช้วิธี 10-fold Cross Validation (Kohavi, 1995) แบ่งข้อมูลสอนสำหรับนำไปสร้างแบบจำลองและข้อมูลทดสอบสำหรับนำไปทดสอบแบบจำลองที่สร้างขึ้น โดยสามารถแสดงชุดข้อมูลได้ในตารางที่ 2

จากตารางที่ 2 แสดงถึง ชุดข้อมูลหลังจากใช้เทคนิคการสุ่มตัวอย่างแล้ว พบว่ามีชุดข้อมูลทั้งหมด 12 ชุดข้อมูล ในชุดข้อมูล Breast Cancer ใช้เทคนิคการสุ่มเพิ่มตัวอย่างส่วนน้อย โดยสุ่มเพิ่มจำนวนคลาส Y จำนวน 115 ระเบียน คิดเป็นร้อยละ 141.98 ทำให้คลาส Y มีจำนวนข้อมูลทั้งหมด 196 ระเบียน ส่วนชุดข้อมูล Breast Cancer ที่ใช้เทคนิคสุ่มลดตัวอย่างส่วนมาก โดยสุ่มลดจำนานคลาส N จำนวน 115 ระเบียน คิดเป็นร้อยละ 141.98 ทำให้คลาส N มีจำนวนข้อมูลทั้งหมด 81 ระเบียน

ในชุดข้อมูล Pima Diabetes ใช้เทคนิคการสุ่มเพิ่มตัวอย่างส่วนน้อย โดยสุ่มเพิ่มจำนวนคลาส Y จำนวน 232 ระเบียน คิดเป็นร้อยละ 86.57 ทำให้คลาส Y มีจำนวนข้อมูลทั้งหมด 500 ระเบียน ส่วนชุดข้อมูล Pima Diabetes ที่ใช้เทคนิคสุ่มลดตัวอย่างส่วนมาก โดยสุ่มลดจำนานคลาส N จำนวน 232 ระเบียน คิดเป็นร้อยละ 86.57 ทำให้คลาส N มีจำนวนข้อมูลทั้งหมด 268 ระเบียน

ในชุดข้อมูล Diabetic foot ใช้เทคนิคการสุ่มเพิ่มตัวอย่างส่วนน้อย โดยสุ่มเพิ่มจำนวนคลาส Y จำนวน 66 ระเบียน คิดเป็นร้อยละ 111.86 ทำให้คลาส Y มีจำนวนข้อมูลทั้งหมด 125 ระเบียน ส่วนชุดข้อมูล Diabetic foot ที่ใช้เทคนิคสุ่มลดตัวอย่างส่วนมาก โดยสุ่มลดจำนานคลาส N จำนวน 66 ระเบียน คิดเป็นร้อยละ 111.86 ทำให้คลาส N มีจำนวนข้อมูลทั้งหมด 59 ระเบียน

3.3 การสร้างแบบจำลอง

งานวิจัยนี้สร้างแบบจำลองโดยนำข้อมูลที่เตรียมไว้รวม 12 ชุดข้อมูลไปสร้างแบบจำลองเพื่อโดยใช้เทคนิควิธีที่ต่างกันดังแสดงใน ตารางที่ 3

ตารางที่ 3 แสดงแบบจำลองในงานวิจัย

Model	Model	แบบจำลอง
J48	Decision Tree	ต้นไม้ตัดสินใจ
CART	CART	คาร์ท
RF	Random Forest	เรนดอมฟรอเรส
SVM	Support Vector Machine	ซัพพอร์ทเวกเตอร์แมชชีน
MLP	Multilayer Perceptron	โครงข่ายประสาทเทียม
A+J48	Adaboost ensemble with Decision Tree	เอดาบูทร่วมกับต้นไม้ตัดสินใจ
A+CART	Adaboost ensemble with CART	เอดาบูทร่วมกับคาร์ท
A+RF	Adaboost ensemble with Random Forest	เอดาบูทร่วมกับเรนดอมฟรอเรส
A+SVM	Adaboost ensemble with Support Vector Machine	เอดาบูทร่วมกับซัพพอร์ทแวกเตอร์แมชชีน
A+MLP	Adaboost ensemble with Multilayer Perceptron	เอดาบูทร่วมกับโครงข่ายประสาทเทียม
B+J48	Bagging ensemble with Decision Tree	ถุงจำแนกร่วมกับต้นไม้ตัดสินใจ
B+CART	Bagging ensemble with CART	ถุงจำแนกร่วมกับคาร์ท
B+RF	Bagging ensemble with Random Forest	ถุงจำแนกร่วมกับเรนดอมฟรอเรส
B+SVM	Bagging ensemble with Support Vector Machine	ถุงจำแนกร่วมกับซัพพอร์ทแวกเตอร์แมชชีน
B+MLP	Bagging ensemble with Multilayer Perceptron	ถุงจำแนกร่วมกับโครงข่ายประสาทเทียม

3.4 การทดสอบประสิทธิภาพของแบบจำลอง

งานวิจัยนี้ใช้การทดสอบประสิทธิภาพของแบบจำลองที่สร้างขึ้นด้วยวิธี 10-fold cross validation และใช้ค่าวัดประสิทธิภาพของแบบจำลอง ดังนี้

ค่าความแม่นยำ (Precision) คือ ค่าที่แสดงถึงประสิทธิภาพของแบบจำลองที่สามารถทำนาย ข้อมูลในคลาส Positive ได้ถูกต้อง จากจำนวนข้อมูลที่ถูกทำนายว่าอยู่ในคลาส Positive ทั้งหมด สามารถคำนวณได้จากสมการที่ 4

ค่าความระลึก (Recall) คือ ค่าที่แสดงถึงประสิทธิภาพของแบบจำลองที่สามารถทำนายข้อมูลในคลาส Positive ได้ถูกต้อง จากจำนวนข้อมูลจริงในคลาส Positive ทั้งหมด สามารถคำนวณได้จากสมการที่ 5

ค่าเอฟ (F-measure) คือ ค่าที่แสดงถึงประสิทธิภาพโดยเฉลี่ยของแบบจำลองจากค่าเฉลี่ยของค่าความแม่นยำและค่าความระลึก สามารถคำนวณได้จากสมการที่ 6

โดยสามารถคำนวณจากผลของแบบจำลองที่สามารถตอบข้อมูลทดสอบได้ซึ่งถูกจัดให้อยู่ในรูปแบบของ Confusion Matrix (Powers D.M.W., 2011) ดังแสดงในภาพประกอบที่ 3

	Predicted (Positive)	**Predicted** (Negative)
Actual (Positive)	True Positive	False Negative
Actual (Negative)	False Positive	True Negative

ภาพประกอบที่ 3 Confusion Matrix

$$\Pr ecision = \left(\frac{TP}{TP + FP} \right) \times 100 \quad (4)$$

ตารางที่ 4 ค่าความแม่นยำของแบบจำลองที่สามารถจำแนกข้อมูลผู้ไม่ป่วยได้

Data	Models															Average
	J48	CART	RF	SVM	MLP	A+J48	A+CART	A+RF	A+SVM	A+MLP	B+J48	B+CART	B+RF	B+SVM	B+MLP	
BR-ORI	75.72	73.98	77.19	71.17	79.34	76.08	77.03	76.23	77.33	79.25	75.42	76.44	77.54	73.11	**80.00**	76.39
BR-SM	71.36	75.28	79.89	70.05	**80.00**	77.47	79.88	79.78	70.90	79.78	79.06	78.21	78.19	70.00	78.69	76.57
BR-SP	61.96	59.14	61.45	65.17	59.76	60.76	58.67	62.96	**67.47**	63.41	60.92	62.65	64.56	65.59	64.56	62.60
BR-RE	78.06	81.55	88.78	70.91	87.75	87.44	**90.31**	88.18	77.38	87.32	78.85	83.80	86.67	71.48	86.26	*82.98*
PI-ORI	79.03	77.64	80.08	65.10	79.85	78.92	76.15	79.70	65.10	79.38	79.50	79.13	80.19	65.10	**80.61**	76.36
PI-SM	78.86	76.72	82.48	54.61	78.78	80.00	76.92	**82.56**	53.57	79.11	80.33	79.25	82.32	55.30	79.15	74.66
PI-SP	72.66	72.69	73.54	53.33	72.01	70.50	68.24	73.18	50.74	71.32	76.61	76.47	**76.71**	50.80	73.26	68.80
PI-RE	87.05	86.72	**91.55**	83.75	82.73	91.07	90.12	91.00	83.75	85.19	89.16	88.01	89.62	79.74	84.68	*86.94*
DM-ORI	73.28	70.06	70.97	67.80	74.80	71.97	**75.19**	69.74	68.52	76.86	71.01	71.83	70.63	67.61	70.99	71.42
DM-SM	74.38	79.31	76.74	68.53	78.05	76.98	74.38	77.69	68.61	78.05	76.03	78.99	**80.47**	67.10	79.53	75.66
DM-SP	50.85	55.56	59.38	47.73	56.67	58.73	52.38	59.32	44.44	55.00	59.32	57.14	59.02	50.77	**59.65**	55.06
DM-RE	82.81	83.94	85.51	80.00	87.10	88.19	85.82	86.23	85.51	**88.71**	80.99	80.00	85.21	79.74	85.38	*84.34*

ตารางที่ 5 ค่าความระลึกของแบบจำลองที่สามารถจำแนกผู้ไม่ป่วยได้

Data	Models															Average
	J48	CART	RF	SVM	MLP	A+J48	A+CART	A+RF	A+SVM	A+MLP	B+J48	B+CART	B+RF	B+SVM	B+MLP	
BR-ORI	93.88	92.86	89.80	**99.49**	86.22	81.12	82.14	86.73	88.78	85.71	92.35	87.76	93.37	98.47	87.76	89.76
BR-SM	75.00	68.37	72.96	77.55	71.43	71.94	68.88	74.49	68.37	72.45	77.04	71.43	75.00	**78.57**	73.47	73.13
BR-SP	70.37	67.90	62.96	71.60	60.49	59.26	54.32	62.96	69.14	64.20	65.43	64.20	62.96	**75.31**	62.96	64.94
BR-RE	94.39	85.71	92.86	**99.49**	91.33	88.78	90.31	91.33	87.24	91.33	91.33	92.35	92.86	98.47	92.86	*92.04*
PI-ORI	81.40	86.80	83.60	**100.00**	83.20	78.60	79.80	84.80	**100.00**	82.40	82.20	83.40	85.00	**100.00**	84.80	86.40
PI-SM	74.60	73.80	77.20	**98.40**	75.00	76.80	74.00	78.60	90.00	75.00	77.60	76.40	78.20	80.40	74.40	78.69
PI-SP	**75.37**	67.54	70.52	23.88	72.01	68.66	64.93	71.27	51.49	68.66	70.90	72.76	71.27	70.90	70.52	66.04
PI-RE	91.40	91.40	95.40	**100.00**	86.20	93.80	93.00	95.00	**100.00**	88.60	93.80	94.00	95.00	**100.00**	86.20	*93.59*
DM-ORI	76.80	88.00	88.00	96.00	76.00	76.00	80.00	84.80	88.80	74.40	78.40	81.60	90.40	**95.20**	74.40	83.25
DM-SM	72.00	73.60	79.20	78.40	76.80	77.60	72.00	80.80	75.20	76.80	73.60	75.20	82.40	**83.20**	80.80	77.17
DM-SP	50.85	50.85	64.41	35.59	57.63	**62.71**	55.93	59.32	40.68	55.93	59.32	54.24	61.02	55.93	57.63	54.80
DM-RE	84.80	92.00	94.40	96.00	86.40	89.60	92.00	95.20	94.40	88.00	92.00	92.80	96.80	**97.60**	88.80	*92.05*

$$\text{Re}\,call = \left(\frac{TP}{TP + FN} \right) \times 100 \quad (5)$$

$$F - measure = \left(\frac{2 \times \text{Pr}\,ecision \times \text{Re}\,call}{\text{Pr}\,ecision + \text{Re}\,call} \right) \times 100 \quad (6)$$

โดยที่

TP คือ จำนวนข้อมูลที่เครื่องทำนายว่าจริง และ ถูกต้องตามข้อมูลจริง

TN คือ จำนวนข้อมูลที่เครื่องทำนายไม่จริง และ ถูกต้องตามข้อมูลจริง

FP คือ จำนวนข้อมูลที่เครื่องทำนายว่าจริง และ ไม่ถูกต้องตามข้อมูลจริง

FN คือ จำนวนข้อมูลที่เครื่องทำนายว่าไม่จริง และ ไม่ถูกต้องตามข้อมูลจริง

ค่าเฉลี่ยถ่วงน้ำหนัก (Weight Average) คือ ค่าเฉลี่ยของค่าประสิทธิภาพของทั้งสองคลาส สามารถคำนวณได้ดัง สมการที่ 7

$$Weight\_Average = \frac{(x_n\, n_n) + (x_y\, n_y)}{n_n + n_y} \quad (7)$$

โดยที่

x_n คือ ค่าประสิทธิถาพของคลาส N

x_y คือ ค่าประสิทธิภาพของคลาส Y

n_n คือ จำนวนข้อมูลคลาส N

n_y คือ จำนวนข้อมูลคลาส Y

4. ผลการทดลอง

การวิจัยนี้ใช้ Computer PC CPU : Intel core i7 7700, Ram : 16 Gb ใช้โปรแกรม Weka 3.8.1 (University of Waikato, 2017) ในการทดลอง ใช้ชุดข้อมูลสำหรับการทดลองจำนวน 12 ชุดข้อมูล เทคนิคการจำแนกจำนวน 15 เทคนิค วัดผลการทดลองได้จากค่าวัดประสิทธิภาพ คือ ค่าความแม่นยำ ค่าความระลึก และค่าเอฟ โดยแสดงผลของแบบจำลองที่สามารถจำแนกผู้ที่ไม่ป่วย (Class 'N') ผู้ที่ป่วย (Class 'Y') และแสดงผลรวมค่าประสิทธิภาพของแบบจำลองโดยแสดงในค่าเฉลี่ยถ่วงน้ำหนัก (Weight Average) ดังนี้

4.1 ผลการจำแนกผู้ไม่ป่วย

ผลการจำแนกผู้ไม่ป่วย แสดงถึงประสิทธิภาพของแบบจำลองในการจำแนกข้อมูลผู้ไม่ป่วยได้ถูกต้อง ซึ่งเป็นข้อมูลกลุ่มส่วนมากในชุดข้อมูลทดลอง โดยแสดงในค่าความแม่นยำ ค่าความระลึก และค่าเอฟ ได้ดังนี้

4.1.1 ค่าความแม่นยำของแบบจำลองที่สามารถจำแนกผู้ไม่ป่วยได้ค่าความแม่นยำ คือ ค่าที่แสดงถึงประสิทธิภาพของแบบจำลองที่สามารถจำแนกข้อมูลว่าเป็นผู้ไม่ป่วยได้ถูกต้อง จากจำนวนข้อมูลที่ถูกจำแนกทั้งหมด แสดงในรูปแบบร้อยละในตารางที่ 4

จากตารางที่ 4 สามารถแสดงค่าความแม่นยำของแบบจำลองที่สามารถจำแนกข้อมูลผู้ไม่ป่วยได้ตามกลุ่มชุดข้อมูล ดังนี้

กลุ่มชุดข้อมูล Breast Cancer พบว่า 1) เทคนิคการสุ่มเพิ่มตัวอย่างส่วนน้อย สามารถช่วยเพิ่มค่าความแม่นยำได้ดีขึ้นกว่าชุดข้อมูลเดิม โดยมีค่าเฉลี่ยทุกแบบจำลองเท่ากับ 76.57% ซึ่งแบบจำลองส่วนใหญ่มีค่าเพิ่มสูงขึ้น ยกเว้นแบบจำลอง J48, SVM, A+SVM และ B+SVM ที่พบว่ามีค่าลดลงเมื่อเปรียบเทียบกับแบบจำลองเดียวกันในชุดข้อมูลเดิม 2) เทคนิคการสุ่มลดตัวอย่างส่วนมาก ไม่สามารถช่วยเพิ่มค่าความแม่นยำได้ โดยมีค่าเฉลี่ยทุกแบบจำลองเท่ากับ 62.60% 3) เทคนิคการสุ่มตัวอย่างซ้ำ สามารถช่วยเพิ่มค่าความแม่นยำได้สูงที่สุดจากค่าเฉลี่ยทุกแบบจำลองเท่ากับ 82.98% ซึ่งแบบจำลองส่วนใหญ่มีค่าเพิ่มสูงขึ้น ยกเว้นแบบจำลอง SVM และ B+SVM พบว่ามีค่าลดลงเมื่อเปรียบเทียบกับแบบจำลองเดียวกันในชุดข้อมูลเดิม

กลุ่มชุดข้อมูล Pima Diabetes พบว่า 1) เทคนิคการสุ่มเพิ่มตัวอย่างส่วนน้อย สามารถช่วยเพิ่มค่าความแม่นยำได้ในบางแบบจำลอง โดยมีค่าเฉลี่ยทุกแบบจำลองเท่ากับ 74.66% ซึ่งแบบจำลองส่วนใหญ่มีค่าลดลง ยกเว้นแบบจำลอง J48, RF, A+J48, A+CART, A+RF, B+J48, B+CART และ B+RF พบว่ามีค่าสูงขึ้นเมื่อเปรียบเทียบกับแบบจำลองเดียวกันในชุดข้อมูลเดิม 2) เทคนิคการสุ่มลดตัวอย่างส่วนมาก ไม่สามารถช่วยเพิ่มค่าความแม่นยำได้ โดยมีค่าเฉลี่ยทุกแบบจำลองเท่ากับ 68.80% 3) เทคนิคการสุ่มตัวอย่างซ้ำ สามารถช่วยเพิ่มค่าความแม่นยำได้สูงที่สุดจากค่าเฉลี่ยทุกแบบจำลองเท่ากับ 86.94%

กลุ่มชุดข้อมูล Diabetic foot พบว่า 1) เทคนิคการสุ่มเพิ่มตัวอย่างส่วนน้อย สามารถช่วยเพิ่มค่าความแม่นยำได้ โดยมีค่าเฉลี่ยทุกแบบจำลองเท่ากับ 75.66% ซึ่งแบบจำลองส่วนใหญ่มีค่าเพิ่มสูงขึ้น ยกเว้นแบบจำลอง A+-CART และ B+SVM พบว่ามีค่าลดลงเมื่อเปรียบเทียบกับแบบจำลองเดียวกันในชุดข้อมูลเดิม 2) เทคนิคการสุ่มลดตัวอย่างส่วนมาก ไม่สามารถช่วยเพิ่มค่าความแม่นยำได้

โดยมีค่าเฉลี่ยทุกแบบจำลองเท่ากับ 55.06% 3) เทคนิคการสุ่มตัวอย่างซ้ำ สามารถช่วยเพิ่มค่าความแม่นยำได้สูงสุดจากค่าเฉลี่ยทุกแบบจำลองเท่ากับ 84.34%

4.1.2 ค่าความระลึกของแบบจำลองที่สามารถจำแนกผู้ไม่ป่วยได้ค่าความระลึก คือ ค่าที่แสดงถึงประสิทธิภาพของแบบจำลองที่สามารถจำแนกข้อมูลว่าเป็นผู้ไม่ป่วยได้ถูกต้อง จากจำนวนข้อมูลจริงที่เป็นผู้ไม่ป่วยทั้งหมด แสดงในรูปแบบร้อยละได้ในตารางที่ 5

จากตารางที่ 5 สามารถแสดงค่าความระลึกของแบบจำลองที่สามารถจำแนกข้อมูลผู้ไม่ป่วยได้ตามกลุ่มชุดข้อมูล ดังนี้

กลุ่มชุดข้อมูล Breast Cancer พบว่า 1) เทคนิคการสุ่มเพิ่มตัวอย่างส่วนน้อย ไม่สามารถช่วยเพิ่มค่าความระลึกได้ โดยมีค่าเฉลี่ยทุกแบบจำลองเท่ากับ 73.13% 2) เทคนิคการสุ่มลดตัวอย่างส่วนมาก ไม่สามารถช่วยเพิ่มค่าความระลึกได้ โดยมีค่าเฉลี่ยทุกแบบจำลองเท่ากับ 64.94% 3) เทคนิคการสุ่มตัวอย่างซ้ำ สามารถช่วยเพิ่มค่าความระลึกได้สูงที่สุดจากค่าเฉลี่ยทุกแบบจำลองเท่ากับ 92.04% ซึ่งแบบจำลองส่วนใหญ่มีค่าเพิ่มสูงขึ้น ยกเว้นแบบจำลอง CART, A+SVM และ B+RF พบว่ามีค่าลดลงเมื่อเปรียบเทียบกับแบบจำลองเดียวกันในชุดข้อมูลเดิม

กลุ่มชุดข้อมูล Pima Diabetes พบว่า 1) เทคนิคการสุ่มเพิ่มตัวอย่างส่วนน้อย ไม่สามารถช่วยเพิ่มค่าความระลึกได้ โดยมีค่าเฉลี่ยทุกแบบจำลองเท่ากับ 78.69% 2) เทคนิคการสุ่มลดตัวอย่างส่วนมาก ไม่สามารถช่วยเพิ่มค่าความระลึกได้ โดยมีค่าเฉลี่ยทุกแบบจำลองเท่ากับ 66.04% 3) เทคนิคการสุ่มตัวอย่างซ้ำ สามารถช่วยเพิ่มค่าความระลึกได้สูงที่สุดจากค่าเฉลี่ยทุกแบบจำลองเท่ากับ 93.59%

กลุ่มชุดข้อมูล Diabetic foot พบว่า 1) เทคนิคการสุ่มเพิ่มตัวอย่างส่วนน้อย สามารถช่วยเพิ่มค่าความระลึกได้ในบางแบบจำลอง โดยมีค่าเฉลี่ยทุกแบบจำลองเท่ากับ 77.17% ซึ่งแบบจำลองส่วนใหญ่มีค่าลดลง ยกเว้นแบบจำลอง MLP A+J48 A+MLP และ B+MLP พบว่ามีค่าเพิ่มสูงขึ้นเมื่อเปรียบเทียบกับแบบจำลองเดียวกันในชุดข้อมูลเดิม 2) เทคนิคการสุ่มลดตัวอย่างส่วนมาก ไม่สามารถช่วยเพิ่มค่าความระลึกได้ โดยมีค่าเฉลี่ยทุกแบบจำลองเท่ากับ 54.80% 3) เทคนิคการสุ่มตัวอย่างซ้ำ สามารถช่วยเพิ่มค่าความระลึกได้สูงที่สุดจากค่าเฉลี่ยทุกแบบจำลองเท่ากับ 92.05%

4.1.3 ค่าเอฟของแบบจำลองที่สามารถจำแนกผู้ไม่ป่วยได้ ค่าเอฟ คือ ค่าที่แสดงถึงประสิทธิภาพโดยรวมของแบบจำลองจากค่าความแม่นยำและค่าความระลึกที่สามารถจำแนกข้อมูลว่าเป็นผู้ไม่ป่วยได้ถูกต้อง แสดงในรูปแบบร้อยละได้ในตารางที่ 6

จากตารางที่ 6 สามารถแสดงค่าเอฟของแบบจำลองที่สามารถจำแนกข้อมูลผู้ไม่ป่วยได้ตามกลุ่มชุดข้อมูล ดังนี้

กลุ่มชุดข้อมูล Breast Cancer พบว่า 1) เทคนิคการสุ่มเพิ่มตัวอย่างส่วนน้อย ไม่สามารถช่วยเพิ่มค่าเอฟได้ โดยมีค่าเฉลี่ยทุกแบบจำลองเท่ากับ 74.71% 2) เทคนิคการสุ่มลดตัวอย่างส่วนมาก ไม่สามารถช่วยเพิ่มค่าเอฟได้ โดยมีค่าเฉลี่ยทุกแบบจำลองเท่ากับ 63.68% 3) เทคนิคการสุ่มตัวอย่างซ้ำ สามารถช่วยเพิ่มค่าเอฟได้จากค่าเฉลี่ยทุกแบบจำลองเท่ากับ 87.06% ซึ่งแบบจำลองส่วนใหญ่มีค่าเพิ่มสูงขึ้นยกเว้นแบบจำลอง SVM, A+SVM และ B+SVM พบว่ามีค่าลดลงเมื่อเปรียบเทียบกับแบบจำลองเดียวกันในชุดข้อมูลเดิม

กลุ่มชุดข้อมูล Pima Diabetes พบว่า 1) เทคนิคการสุ่มเพิ่มตัวอย่างส่วนน้อย ไม่สามารถช่วยเพิ่มค่าเอฟได้ โดยมีค่าเฉลี่ยทุกแบบจำลองเท่ากับ 75.76% 2) เทคนิค

ตารางที่ 6 ค่าเอฟของแบบจำลองที่สามารถจำแนกผู้ไม่ป่วยได้

Data	Models															Average
	J48	CART	RF	SVM	MLP	A+J48	A+CART	A+RF	A+SVM	A+MLP	B+J48	B+CART	B+RF	B+SVM	B+MLP	
BR-ORI	83.83	82.35	83.02	82.98	82.64	78.52	79.51	81.15	82.66	82.35	83.03	81.71	**84.72**	83.91	83.70	82.40
BR-SM	73.13	71.66	76.27	73.61	75.47	74.60	73.97	77.04	69.61	75.94	**78.04**	74.67	76.56	74.04	75.99	74.71
BR-SP	65.90	63.22	62.20	68.24	60.12	60.00	56.41	62.96	68.29	63.80	63.10	63.41	63.75	**70.11**	63.75	63.68
BR-RE	85.45	83.58	90.77	82.80	89.50	88.10	**90.31**	89.72	82.01	89.28	84.63	87.86	89.66	82.83	89.43	*87.06*
PI-ORI	80.20	81.96	81.80	78.86	81.49	78.76	77.93	82.17	78.86	80.86	80.83	81.21	82.52	78.86	**82.65**	80.60
PI-SM	76.67	75.23	79.75	70.24	76.84	78.37	75.43	**80.53**	67.16	77.00	78.94	77.80	80.21	65.53	76.70	75.76
PI-SP	73.99	70.02	72.00	32.99	72.01	69.57	66.54	72.21	51.11	69.96	73.64	**74.57**	73.89	59.19	71.86	66.90
PI-RE	89.17	89.00	**93.44**	91.16	84.43	92.41	91.54	92.95	91.16	86.86	91.42	90.91	92.23	88.73	85.43	*90.06*
DM-ORI	75.00	78.01	78.57	**79.47**	75.40	73.93	77.52	76.53	77.35	75.61	74.52	76.40	79.30	79.07	72.66	76.62
DM-SM	73.17	76.35	77.95	73.13	77.42	77.29	73.17	79.22	71.76	77.42	74.80	77.05	81.42	74.29	**80.16**	76.31
DM-SP	50.85	53.10	**61.79**	40.78	57.14	60.66	54.10	59.32	42.48	55.46	59.32	55.65	60.00	53.23	58.62	54.83
DM-RE	83.79	87.79	89.73	87.27	86.75	88.89	88.80	90.49	89.73	88.35	86.14	85.93	**90.64**	87.77	87.06	*87.94*

ตารางที่ 7 ค่าความแม่นยำของแบบจำลองที่สามารถจำแนกผู้ป่วยได้

Data	Models															Average
	J48	CART	RF	SVM	MLP	A+J48	A+CART	A+RF	A+SVM	A+MLP	B+J48	B+CART	B+RF	B+SVM	B+MLP	
BR-ORI	64.71	54.84	59.18	66.67	57.81	45.59	48.53	51.85	57.69	56.92	59.46	53.85	68.29	**76.92**	61.29	58.91
BR-SM	73.66	71.03	75.12	74.86	74.19	73.81	72.65	76.08	69.46	74.77	**77.61**	73.71	75.98	75.58	75.12	*74.24*
BR-SP	65.71	62.32	62.03	68.49	60.00	60.24	57.47	62.96	68.35	63.75	62.67	63.29	63.86	**71.01**	63.86	63.73
BR-RE	72.50	60.56	**80.56**	50.00	76.71	71.79	76.54	77.03	55.36	76.39	66.00	75.41	79.10	57.14	78.79	70.26
PI-ORI	63.24	68.42	66.67	0.00	65.99	60.37	58.61	67.80	0.00	64.66	64.54	65.56	68.49	0.00	**68.60**	52.20
PI-SM	75.90	74.76	78.57	**91.92**	76.15	77.69	74.95	79.58	68.75	76.24	78.34	77.22	79.24	64.10	75.85	76.62
PI-SP	**74.42**	69.69	71.68	50.96	72.01	69.45	66.55	72.00	50.76	69.78	72.92	74.02	73.17	51.85	71.58	67.39
PI-RE	82.30	82.16	90.69	**100.00**	72.06	87.75	86.11	89.84	**100.00**	77.02	87.19	87.18	89.50	**100.00**	73.36	87.01
DM-ORI	45.28	44.44	48.28	28.57	47.37	42.31	**50.98**	40.63	36.36	49.21	41.30	45.24	50.00	25.00	39.62	42.31
DM-SM	72.87	75.37	78.51	74.77	77.17	77.42	72.87	80.00	72.57	77.17	74.42	76.34	**81.97**	77.89	80.49	76.65
DM-SP	50.85	54.69	**61.11**	48.65	56.90	60.00	52.73	59.32	45.31	55.17	59.32	56.45	59.65	50.94	59.02	55.34
DM-RE	66.07	78.72	84.78	85.29	71.67	77.19	80.00	86.96	84.78	75.00	76.19	76.92	**90.48**	90.32	74.07	79.90

การสุ่มลดตัวอย่างส่วนมาก ไม่สามารถช่วยเพิ่มค่าเอฟได้ โดยมีค่าเฉลี่ยทุกแบบจำลองเท่ากับ 66.90% 3) เทคนิคการสุ่มตัวอย่างซ้ำ สามารถช่วยเพิ่มค่าเอฟได้จากค่าเฉลี่ยทุกแบบจำลองเท่ากับ 90.06%

กลุ่มชุดข้อมูล Diabetic foot พบว่า 1) เทคนิคการสุ่มเพิ่มตัวอย่างส่วนน้อย สามารถช่วยเพิ่มค่าเอฟได้ในบางแบบจำลอง โดยมีค่าเฉลี่ยทุกแบบจำลองเท่ากับ 76.31% ซึ่งแบบจำลองบางส่วนมีค่าลดลง ยกเว้นแบบจำลอง MLP, A+J48, A+RF, A+MLP, B+J48, B+CART, B+RF และ B+MLP พบว่ามีค่าเพิ่มสูงขึ้นเมื่อเปรียบเทียบกับแบบจำลองเดียวกันในชุดข้อมูลเดิม 2) เทคนิคการสุ่มลดตัวอย่างส่วนมาก ไม่สามารถช่วยเพิ่มค่าเอฟได้ โดยมีค่าเฉลี่ยทุกแบบจำลองเท่ากับ 54.83% 3) เทคนิคการสุ่มตัวอย่างซ้ำ สามารถช่วยเพิ่มค่าเอฟได้สูงที่สุดจากค่าเฉลี่ยทุกแบบจำลองเท่ากับ 87.94%

4.2 ผลการจำแนกผู้ป่วย

ผลการจำแนกผู้ป่วย แสดงถึงประสิทธิภาพของแบบจำลองในการจำแนกกลุ่มผู้ป่วยได้ถูกต้อง ซึ่งเป็นข้อมูลกลุ่มน้อยในชุดข้อมูลทดลอง โดยแสดงในค่าความแม่นยำ ค่าความระลึก และค่าเอฟ ได้ดังนี้

4.2.1 ค่าความแม่นยำของแบบจำลองที่สามารถจำแนกผู้ป่วยได้

ค่าความแม่นยำ คือ ค่าที่แสดงถึงประสิทธิภาพของแบบจำลองที่สามารถจำแนกข้อมูลว่าเป็นผู้ป่วยได้ถูกต้อง จากจำนวนข้อมูลที่ถูกจำแนกทั้งหมด แสดงในรูปแบบร้อยละในตารางที่ 7

จากตารางที่ 7 สามารถแสดงค่าความแม่นยำของแบบจำลองที่สามารถจำแนกข้อมูลผู้ป่วยได้ตามกลุ่มชุดข้อมูล ดังนี้

กลุ่มชุดข้อมูล Breast Cancer พบว่า 1) เทคนิคการสุ่มเพิ่มตัวอย่างส่วนน้อย สามารถช่วยเพิ่มค่าความแม่นยำได้สูงสุด โดยมีค่าเฉลี่ยทุกแบบจำลองเท่ากับ 74.24% ซึ่งแบบจำลองส่วนใหญ่มีค่าเพิ่มสูงขึ้น ยกเว้นในแบบจำลอง B+SVM พบว่ามีค่าลดลงเมื่อเปรียบเทียบกับแบบจำลองเดียวกันในชุดข้อมูลเดิม 2) เทคนิคการสุ่มลดตัวอย่างส่วนมาก สามารถช่วยเพิ่มค่าความแม่นยำได้เล็กน้อย โดยมีค่าเฉลี่ยทุกแบบจำลองเท่ากับ 63.73% ซึ่งแบบจำลองส่วนใหญ่มีค่าเพิ่มสูงขึ้น ยกเว้นแบบจำลอง B+RF และ B+SVM พบว่ามีค่าลดลงเมื่อเปรียบเทียบกับแบบจำลองเดียวกันในชุดข้อมูลเดิม 3) เทคนิคการสุ่มตัวอย่างซ้ำ สามารถช่วยเพิ่มค่าความแม่นยำได้ดีขึ้น โดยมีค่าเฉลี่ยทุกแบบจำลองเท่ากับ 70.26% ซึ่งแบบจำลองส่วนใหญ่มีค่าเพิ่มสูงขึ้น ยกเว้นแบบจำลอง SVM, A+SVM และ B+SVM พบว่ามีค่าลดลงเมื่อเปรียบเทียบกับแบบจำลองเดียวกันในชุดข้อมูลเดิม

กลุ่มชุดข้อมูล Pima Diabetes พบว่า 1) เทคนิคการสุ่มเพิ่มตัวอย่างส่วนน้อย สามารถช่วยเพิ่มค่าความแม่นยำได้โดยมีค่าเฉลี่ยทุกแบบจำลองเท่ากับ 76.62% 2) เทคนิคการสุ่มลดตัวอย่างส่วนมาก สามารถช่วยเพิ่มค่าความแม่นยำได้ โดยมีค่าเฉลี่ยทุกแบบจำลองเท่ากับ 67.39% 3) เทคนิคการสุ่มตัวอย่างซ้ำ สามารถช่วยเพิ่มค่าความแม่นยำได้สูงสุดจากค่าเฉลี่ยทุกแบบจำลองเท่ากับ 87.01%

กลุ่มชุดข้อมูล Diabetic foot พบว่า 1) เทคนิคการสุ่มเพิ่มตัวอย่างส่วนน้อย สามารถช่วยเพิ่มค่าความแม่นยำได้โดยมีค่าเฉลี่ยทุกแบบจำลองเท่ากับ 76.65% 2) เทคนิคการ

ตารางที่ 8 ค่าความระลึกของแบบจำลองที่สามารถจำแนกผู้ป่วยได้

Data	Models															Average
	J48	CART	RF	SVM	MLP	A+J48	A+CART	A+RF	A+SVM	A+MLP	B+J48	B+CART	B+RF	B+SVM	B+MLP	
BR-ORI	27.16	20.99	35.80	2.47	45.68	38.27	40.74	34.57	37.04	45.68	27.16	34.57	34.57	12.35	**46.91**	32.26
BR-SM	69.90	77.55	81.63	66.84	82.14	79.08	**82.65**	81.12	71.94	81.63	79.59	80.10	79.08	66.33	80.10	*77.31*
BR-SP	56.79	53.09	60.49	61.73	59.26	61.73	61.73	62.96	**66.67**	62.96	58.02	61.73	65.43	60.49	65.43	61.23
BR-RE	35.80	53.09	**71.60**	1.23	69.14	69.14	76.54	70.37	38.27	67.90	40.74	56.79	65.43	4.94	64.20	52.35
PI-ORI	59.70	53.36	61.19	0.00	60.82	60.82	53.36	59.70	0.00	60.07	60.45	58.96	60.82	0.00	**61.94**	47.41
PI-SM	80.00	77.60	**83.60**	18.20	79.80	80.80	77.80	83.40	22.00	80.20	81.00	80.00	83.20	35.00	80.40	69.53
PI-SP	71.64	74.63	74.63	**79.10**	72.01	71.27	69.78	73.88	50.00	72.39	78.36	77.61	78.36	31.34	74.25	69.95
PI-RE	74.63	73.88	**83.58**	63.81	66.42	82.84	80.97	82.46	63.81	71.27	78.73	76.12	79.48	52.61	70.90	*73.43*
DM-ORI	40.68	20.34	23.73	3.39	**45.76**	37.29	44.07	22.03	13.56	52.54	32.20	32.20	20.34	3.39	35.59	28.47
DM-SM	75.20	**80.80**	76.00	64.00	78.40	76.80	75.20	76.80	65.60	78.40	76.80	80.00	80.00	59.20	79.20	*74.83*
DM-SP	50.85	59.32	55.93	**61.02**	55.93	55.93	49.15	59.32	49.15	54.24	59.32	59.32	57.63	45.76	**61.02**	55.59
DM-RE	62.71	62.71	66.10	49.15	72.88	74.58	67.80	67.80	66.10	**76.27**	54.24	50.85	64.41	47.46	67.80	63.39

ตารางที่ 9 ค่าเอฟของแบบจำลองที่สามารถจำแนกผู้ป่วยได้

Data	Models															Average
	J48	CART	RF	SVM	MLP	A+J48	A+CART	A+RF	A+SVM	A+MLP	B+J48	B+CART	B+RF	B+SVM	B+MLP	
BR-ORI	38.26	30.36	44.62	4.76	51.03	41.61	44.30	41.48	45.11	50.68	37.29	42.11	45.90	21.28	**53.15**	39.46
BR-SM	71.73	74.15	78.24	70.62	77.97	76.35	77.33	78.52	70.68	78.05	**78.59**	76.77	77.50	70.65	77.53	*75.64*
BR-SP	60.93	57.33	61.25	64.94	59.63	60.98	59.52	62.96	**67.50**	63.35	60.26	62.50	64.63	65.33	64.63	62.38
BR-RE	47.93	56.58	75.82	2.41	72.73	70.44	**76.54**	73.55	45.26	71.90	50.38	64.79	71.62	9.09	70.75	57.32
PI-ORI	61.42	59.96	63.81	0.00	63.30	60.59	55.86	63.49	0.00	62.28	62.43	62.08	64.43	0.00	**65.10**	49.65
PI-SM	77.90	76.15	81.01	30.38	77.93	79.22	76.35	**81.45**	33.33	78.17	79.65	78.59	81.17	45.28	78.06	70.31
PI-SP	73.00	72.07	73.13	61.99	72.01	70.35	68.12	72.93	50.38	71.06	75.54	75.77	**75.68**	39.07	72.89	68.27
PI-RE	78.28	77.80	**86.99**	77.90	69.13	85.22	83.46	85.99	77.90	74.03	82.75	81.27	84.19	68.95	72.11	*79.06*
DM-ORI	42.86	27.91	31.82	6.06	46.55	39.64	47.27	28.57	19.75	**50.82**	36.19	37.62	28.92	5.97	37.50	32.50
DM-SM	74.02	77.99	77.24	68.97	77.78	77.11	74.02	78.37	68.91	77.78	75.59	78.13	**80.97**	67.27	79.84	*75.60*
DM-SP	50.85	56.91	58.41	54.14	56.41	57.89	50.88	59.32	47.15	54.70	59.32	57.85	58.62	48.21	**60.00**	55.38
DM-RE	64.35	69.81	74.29	62.37	72.27	75.86	73.39	**76.19**	74.29	75.63	63.37	61.22	75.25	62.22	70.80	70.09

สุ่มลดตัวอย่างส่วนมาก สามารถช่วยเพิ่มค่าความแม่นยำได้โดยมีค่าเฉลี่ยทุกแบบจำลองเท่ากับ 55.34% 3) เทคนิคการสุ่มตัวอย่างซ้ำ สามารถช่วยเพิ่มค่าความแม่นยำได้สูงที่สุดจากค่าเฉลี่ยทุกแบบจำลองเท่ากับ 79.90%

4.2.2 ค่าความระลึกของแบบจำลองที่สามารถจำแนกผู้ป่วยได้

ค่าความระลึก คือ ค่าที่แสดงถึงประสิทธิภาพของแบบจำลองที่สามารถจำแนกข้อมูลว่าเป็นผู้ป่วยได้ถูกต้อง จากจำนวนข้อมูลจริงที่เป็นผู้ป่วยทั้งหมด แสดงในรูปแบบร้อยละได้ในตารางที่ 8

จากตารางที่ 8 สามารถแสดงค่าความระลึกของแบบจำลองที่สามารถจำแนกข้อมูลผู้ป่วยได้ตามกลุ่มชุดข้อมูล ดังนี้

กลุ่มชุดข้อมูล Breast Cancer พบว่า 1) เทคนิคการสุ่มเพิ่มตัวอย่างส่วนน้อย สามารถช่วยเพิ่มค่าความระลึกได้สูงสุด โดยมีค่าเฉลี่ยทุกแบบจำลองเท่ากับ 77.31% 2) เทคนิคการสุ่มลดตัวอย่างส่วนมาก สามารถช่วยเพิ่มค่าความระลึกได้ โดยมีค่าเฉลี่ยทุกแบบจำลองเท่ากับ 61.23% 3) เทคนิคการสุ่มตัวอย่างซ้ำ สามารถช่วยเพิ่มค่าความระลึกได้ โดยมีค่าเฉลี่ยทุกแบบจำลองเท่ากับ 52.35% ซึ่งแบบจำลองส่วนใหญ่มีค่าเพิ่มสูงขึ้น ยกเว้นแบบจำลอง SVM และ B+SVM พบว่ามีค่าลดลงเมื่อเปรียบเทียบกับแบบจำลองเดียวกันในชุดข้อมูลเดิม

กลุ่มชุดข้อมูล Pima Diabetes พบว่า 1) เทคนิคการสุ่มเพิ่มตัวอย่างส่วนน้อย สามารถช่วยเพิ่มค่าความระลึกได้โดยมีค่าเฉลี่ยทุกแบบจำลองเท่ากับ 69.53% 2) เทคนิคการสุ่มลดตัวอย่างส่วนมาก สามารถช่วยเพิ่มค่าความระลึกได้ โดยมีค่าเฉลี่ยทุกแบบจำลองเท่ากับ 69.95% 3) เทคนิคการสุ่มตัวอย่างซ้ำ สามารถช่วยเพิ่มค่าความระลึก

ตารางที่ 10 ค่าความแม่นยำเฉลี่ยถ่วงน้ำหนัก

Data	Model															Average
	J48	CART	RF	SVM	MLP	A+J48	A+CART	A+RF	A+SVM	A+MLP	B+J48	B+CART	B+RF	B+SVM	B+MLP	
BR-ORI	72.50	68.39	71.93	69.85	73.05	67.16	68.70	69.10	71.59	72.72	70.75	69.84	**74.84**	74.22	74.53	71.28
BR-SM	72.51	73.15	77.50	72.45	77.10	75.64	76.26	77.93	70.18	77.27	**78.33**	75.96	77.09	72.79	76.90	75.40
BR-SP	63.84	60.73	61.74	66.83	59.88	60.50	58.07	62.96	67.91	63.58	61.79	62.97	64.21	**68.30**	64.21	63.17
BR-RE	76.43	75.42	**86.38**	64.79	84.52	82.86	86.28	84.92	70.94	84.12	75.10	81.34	84.46	67.29	84.07	*79.26*
PI-ORI	73.52	74.42	75.40	42.39	75.01	72.44	70.02	75.55	42.39	74.25	74.28	74.39	76.11	42.39	**76.42**	67.93
PI-SM	77.38	75.74	80.53	73.26	77.46	78.85	75.94	**81.07**	61.16	77.67	79.33	78.24	80.78	59.70	77.50	75.64
PI-SP	73.54	71.19	72.61	52.15	72.01	69.98	67.39	72.59	50.75	70.55	74.76	**75.25**	74.94	51.33	72.42	68.10
PI-RE	85.39	85.13	**91.25**	89.42	79.01	89.91	88.72	90.59	89.42	82.34	88.47	87.72	89.58	86.81	80.73	*86.97*
DM-ORI	64.30	61.85	63.69	55.22	66.01	62.46	67.43	60.40	58.21	**67.99**	61.49	63.30	64.01	53.95	60.93	62.08
DM-SM	73.62	77.34	77.63	71.65	77.61	77.20	73.62	78.85	70.59	77.61	75.23	77.66	**81.22**	72.50	80.01	76.16
DM-SP	50.85	55.12	**60.24**	48.19	56.78	59.37	52.55	59.32	44.88	55.09	59.32	56.80	59.33	50.86	59.33	55.20
DM-RE	77.44	82.27	85.27	81.70	82.15	84.66	83.95	86.46	85.27	84.31	79.45	79.01	**86.90**	83.13	81.76	*82.92*

ได้สูงสุดจากค่าเฉลี่ยทุกแบบจำลองเท่ากับ 73.43%

กลุ่มชุดข้อมูล Diabetic foot พบว่า 1) เทคนิคการสุ่มเพิ่มตัวอย่างส่วนน้อย สามารถช่วยเพิ่มค่าความระลึกได้สูงสุด โดยมีค่าเฉลี่ยทุกแบบจำลองเท่ากับ 74.83% 2) เทคนิคการสุ่มลดตัวอย่างส่วนมาก สามารถช่วยเพิ่มค่าประสิทธิภาพการจำแนกได้ โดยมีค่าเฉลี่ยทุกแบบจำลองเท่ากับ 55.59% 3) เทคนิคการสุ่มตัวอย่างซ้ำสามารถช่วยเพิ่มค่าประสิทธิภาพได้ โดยมีค่าเฉลี่ยทุกแบบจำลองเท่ากับ 63.39%

4.2.3 ค่าเอฟของแบบจำลองที่สามารถจำแนกผู้ป่วยได้ค่าเอฟ คือ ค่าที่แสดงถึงประสิทธิภาพโดยรวมของแบบจำลองจากค่าความแม่นยำและค่าความระลึกที่สามารถจำแนกข้อมูลว่าเป็นผู้ป่วยได้ถูกต้อง แสดงในรูปแบบร้อยละได้ในตารางที่ 9

จากตารางที่ 9 สามารถแสดงค่าเอฟของแบบจำลองที่สามารถจำแนกข้อมูลผู้ป่วยได้ตามกลุ่มชุดข้อมูล ดังนี้

กลุ่มชุดข้อมูล Breast Cancer พบว่า 1) เทคนิคการสุ่มเพิ่มตัวอย่างส่วนน้อย สามารถช่วยเพิ่มค่าเอฟได้สูงสุด โดยมีค่าเฉลี่ยทุกแบบจำลองเท่ากับ 75.64% 2) เทคนิคการสุ่มลดตัวอย่างส่วนมากสามารถช่วยเพิ่มค่าเอฟได้ โดยมีค่าเฉลี่ยทุกแบบจำลองเท่ากับ 62.38% 3) เทคนิคการสุ่มตัวอย่างซ้ำ สามารถช่วยเพิ่มค่าเอฟได้โดยมีค่าเฉลี่ยทุกแบบจำลอง เท่ากับ 57.32% ซึ่งแบบจำลองส่วนใหญ่มีค่าเพิ่มสูงขึ้นยกเว้นแบบจำลอง SVM และ B+SVM ที่พบว่ามีค่าลดลงเมื่อเปรียบเทียบกับแบบจำลองเดียวกันในชุดข้อมูลเดิม

กลุ่มชุดข้อมูล Pima Diabetes พบว่า 1) เทคนิคการสุ่มเพิ่มตัวอย่างส่วนน้อย สามารถช่วยเพิ่มค่าเอฟได้ โดยมีค่าเฉลี่ยทุกแบบจำลองเท่ากับ 70.31% 2) เทคนิคการสุ่มลดตัวอย่างส่วนมาก สามารถช่วยเพิ่มค่าเอฟได้ โดยมีค่าเฉลี่ยทุกแบบจำลองเท่ากับ 68.27% 3) เทคนิคการสุ่มตัวอย่างซ้ำ สามารถช่วยเพิ่มค่าเอฟได้สูงสุดจากค่าเฉลี่ยทุกแบบจำลองเท่ากับ 79.06%

กลุ่มชุดข้อมูล Diabetic foot พบว่า 1) เทคนิคการสุ่มเพิ่มตัวอย่างส่วนน้อย สามารถช่วยเพิ่มค่าเอฟได้สูงสุด โดยมีค่าเฉลี่ยทุกแบบจำลองเท่ากับ 75.60% 2) เทคนิคการสุ่มลดตัวอย่างส่วนมาก สามารถช่วยเพิ่มค่าเอฟได้ โดยมีค่าเฉลี่ยทุกแบบจำลองเท่ากับ 55.38% 3) เทคนิคการสุ่มตัวอย่างซ้ำ สามารถช่วยเพิ่มค่าเอฟได้ดี โดยมีค่าเฉลี่ยทุกแบบจำลองเท่ากับ 70.09%

4.3 ค่าเฉลี่ยถ่วงน้ำหนักของแบบจำลอง

ค่าเฉลี่ยถ่วงน้ำหนักของแบบจำลอง แสดงถึงประสิทธิภาพโดยรวมของแบบจำลองในการจำแนกผู้ไม่ป่วยและผู้ป่วยได้ถูกต้อง โดยแสดงในค่าความแม่นยำ ค่าความระลึก และค่าเอฟ ได้ดังนี้

4.3.1 ค่าความแม่นยำเฉลี่ยถ่วงน้ำหนัก

ค่าความแม่นยำเฉลี่ยถ่วงน้ำหนัก คือ ค่าที่แสดงถึงประสิทธิภาพโดยรวมของแบบจำลองที่สามารถจำแนกข้อมูลว่าเป็นผู้ไม่ป่วยและผู้ป่วยได้ถูกต้อง จากจำนวนข้อมูลที่ถูกจำแนกทั้งหมด แสดงในรูปแบบร้อยละในตารางที่ 10

จากตารางที่ 10 สามารถแสดงค่าความแม่นยำเฉลี่ยถ่วงน้ำหนักของแบบจำลองได้ตามกลุ่มชุดข้อมูล ดังนี้

กลุ่มชุดข้อมูล Breast Cancer พบว่า 1) เทคนิคการสุ่มเพิ่มตัวอย่างส่วนน้อย สามารถช่วยเพิ่มค่าความแม่นยำได้ โดยมีค่าเฉลี่ยทุกแบบจำลองเท่ากับ 75.40%

ตารางที่ 11 ค่าความระลึกเฉลี่ยถ่วงน้ำหนัก

Data	Models															Average
	J48	CART	RF	SVM	MLP	A+J48	A+CART	A+RF	A+SVM	A+MLP	B+J48	B+CART	B+RF	B+SVM	B+MLP	
BR-ORI	74.37	71.84	74.01	71.12	74.37	68.59	70.04	71.48	73.65	74.01	73.29	72.20	**76.17**	73.29	75.81	72.95
BR-SM	72.45	72.96	77.30	72.19	76.79	75.51	75.77	77.81	70.15	77.04	**78.32**	75.77	77.04	72.45	76.79	75.22
BR-SP	63.58	60.49	61.73	66.67	59.88	60.49	58.02	62.96	**67.90**	63.58	61.73	62.96	64.20	67.90	64.20	63.09
BR-RE	77.26	76.17	**86.64**	70.76	84.84	83.03	86.28	85.20	72.92	84.48	76.53	81.95	84.84	71.12	84.48	*80.43*
PI-ORI	73.83	75.13	75.78	65.10	75.39	72.40	70.57	76.04	65.10	74.61	74.61	74.87	76.56	65.10	**76.82**	72.80
PI-SM	77.30	75.70	80.40	58.30	77.40	78.80	75.90	**81.00**	56.00	77.60	79.30	78.20	80.70	57.70	77.40	74.11
PI-SP	73.51	71.08	72.57	51.49	72.01	69.96	67.35	72.57	50.75	70.52	74.63	**75.19**	74.81	51.12	72.39	68.00
PI-RE	85.55	85.29	**91.28**	87.37	79.30	89.97	88.80	90.63	87.37	82.55	88.54	87.76	89.58	83.46	80.86	*86.55*
DM-ORI	65.22	66.30	67.39	66.30	66.30	63.59	**68.48**	64.67	64.67	67.39	63.59	65.76	67.93	65.76	61.96	65.69
DM-SM	73.60	77.20	77.60	71.20	77.60	77.20	73.60	78.80	70.40	77.60	75.20	77.60	81.20	71.20	**80.00**	76.00
DM-SP	50.85	55.08	**60.17**	48.31	56.78	59.32	52.54	59.32	44.92	55.08	59.32	56.78	59.32	50.85	59.32	55.20
DM-RE	77.72	82.61	85.33	80.98	82.07	84.78	84.24	**86.41**	85.33	84.24	79.89	79.35	**86.41**	81.52	82.07	*82.86*

ตารางที่ 12 ค่าเอฟเฉลี่ยถ่วงน้ำหนัก

Data	Models															Average
	J48	CART	RF	SVM	MLP	A+J48	A+CART	A+RF	A+SVM	A+MLP	B+J48	B+CART	B+RF	B+SVM	B+MLP	
BR-ORI	70.50	67.15	71.79	60.11	73.40	67.73	69.21	69.55	71.68	73.09	69.65	70.13	73.37	65.60	**74.76**	69.85
BR-SM	72.43	72.90	77.25	72.11	76.72	75.48	75.65	77.78	70.14	76.99	**78.31**	75.72	77.03	72.35	76.76	75.18
BR-SP	63.41	60.28	61.72	66.59	59.88	60.49	57.97	62.96	**67.90**	63.58	61.68	62.96	64.19	67.72	64.19	63.03
BR-RE	74.48	75.69	**86.40**	59.29	84.60	82.94	86.28	84.99	71.27	84.19	74.62	81.12	84.38	61.27	83.97	*78.37*
PI-ORI	73.64	74.28	75.52	51.34	75.14	72.42	70.23	75.65	51.34	74.38	74.41	74.53	76.21	51.34	**76.53**	69.80
PI-SM	77.28	75.69	80.38	50.31	77.39	78.79	75.89	**80.99**	50.25	77.58	79.29	78.19	80.69	55.40	77.38	73.03
PI-SP	73.50	71.05	72.56	47.49	72.01	69.96	67.33	72.57	50.74	70.51	74.59	**75.17**	74.78	49.13	72.38	67.59
PI-RE	85.37	85.09	**91.19**	86.53	79.09	89.90	88.72	90.53	86.53	82.39	88.39	87.55	89.43	81.83	80.78	*86.22*
DM-ORI	64.69	61.95	63.58	55.93	66.15	62.93	**67.82**	61.15	58.88	67.66	62.23	63.97	63.14	55.63	61.38	62.47
DM-SM	73.59	77.17	77.59	71.05	77.60	77.20	73.59	78.79	70.33	77.60	75.19	77.59	81.20	70.78	**80.00**	75.95
DM-SP	50.85	55.00	**60.10**	47.46	56.78	59.28	52.49	59.32	44.82	55.08	59.32	56.75	59.31	50.72	59.31	55.11
DM-RE	77.56	82.02	84.78	79.29	82.10	84.71	83.86	**85.91**	84.78	84.27	78.84	78.01	85.70	79.58	81.84	*82.22*

ซึ่งแบบจำลองส่วนใหญ่มีค่าเพิ่มสูงขึ้น ยกเว้นแบบจำลอง A+SVM และ B+SVM ที่มีค่าลดลงเมื่อเปรียบเทียบกับแบบจำลองเดียวกันในชุดข้อมูลเดิม 2) เทคนิคการสุ่มลดตัวอย่างส่วนมาก ไม่สามารถช่วยเพิ่มค่าความแม่นยำได้ โดยมีค่าเฉลี่ยทุกแบบจำลองเท่ากับ 63.17% 3) เทคนิคการสุ่มตัวอย่างซ้ำ สามารถช่วยเพิ่มค่าความแม่นยำได้สูงสุด โดยมีค่าเฉลี่ยทุกแบบจำลอง เท่ากับ 79.26% ซึ่งแบบจำลองส่วนใหญ่มีค่าเพิ่มสูงขึ้นยกเว้นแบบจำลอง SVM, A+SVM และ B+SVM ที่พบว่ามีค่าลดลงเมื่อเปรียบเทียบกับแบบจำลองเดียวกันในชุดข้อมูลเดิม

กลุ่มชุดข้อมูล Pima Diabetes พบว่า 1) เทคนิคการสุ่มเพิ่มตัวอย่างส่วนน้อย สามารถช่วยเพิ่มค่าความแม่นยำได้ โดยมีค่าเฉลี่ยทุกแบบจำลองเท่ากับ 75.64% 2) เทคนิคการสุ่มลดตัวอย่างส่วนมาก สามารถช่วยเพิ่มค่าความแม่นยำได้ในบางแบบจำลอง โดยมีค่าเฉลี่ยทุกแบบจำลองเท่ากับ 68.10% ซึ่งในแบบจำลอง J48, SVM, A+SVM, B+J48, B+CART และ B+SVM มีค่าเพิ่มสูงขึ้นเล็กน้อยเมื่อเปรียบเทียบกับแบบจำลองเดียวกันในชุดข้อมูลเดิม 3) เทคนิคการสุ่มตัวอย่างซ้ำ สามารถช่วยเพิ่มค่าความแม่นยำได้สูงที่สุดจากค่าเฉลี่ยทุกแบบจำลองเท่ากับ 86.92%

กลุ่มชุดข้อมูล Diabetic foot พบว่า 1) เทคนิคการสุ่มเพิ่มตัวอย่างส่วนน้อย สามารถช่วยเพิ่มค่าความแม่นยำได้ โดยมีค่าเฉลี่ยทุกแบบจำลองเท่ากับ 76.16% 2) เทคนิคการสุ่มลดตัวอย่างส่วนมาก ไม่สามารถช่วยเพิ่มค่าความแม่นยำได้ โดยมีค่าเฉลี่ยทุกแบบจำลองเท่ากับ 55.20% 3) เทคนิคการสุ่มตัวอย่างซ้ำ สามารถช่วยเพิ่ม

ค่าความแม่นยำได้สูงสุด โดยมีค่าเฉลี่ยทุกแบบจำลองเท่ากับ 82.92%

4.3.2 ค่าความระลึกเฉลี่ยถ่วงน้ำหนัก

ค่าความระลึกเฉลี่ยถ่วงน้ำหนัก คือ ค่าที่แสดงถึงประสิทธิภาพของแบบจำลองที่สามารถจำแนกข้อมูลว่าเป็นผู้ไม่ป่วยและผู้ป่วยได้ถูกต้อง จากจำนวนข้อมูลจริงที่เป็นผู้ป่วยทั้งหมด แสดงในรูปแบบร้อยละได้ในตารางที่ 11

จากตารางที่ 11 สามารถแสดงค่าความระลึกเฉลี่ยถ่วงน้ำหนักของแบบจำลองได้ตามกลุ่มชุดข้อมูล ดังนี้

กลุ่มชุดข้อมูล Breast Cancer พบว่า 1) เทคนิคการสุ่มเพิ่มตัวอย่างส่วนน้อย สามารถช่วยเพิ่มค่าความระลึกได้ โดยมีค่าเฉลี่ยทุกแบบจำลองเท่ากับ 75.22% ซึ่งแบบจำลองส่วนใหญ่มีค่าเพิ่มสูงขึ้น ยกเว้นแบบจำลอง J48, A+SVM และ B+SVM พบว่ามีค่าลดลงเมื่อเปรียบเทียบกับแบบจำลองเดียวกันในชุดข้อมูลเดิม 2) เทคนิคการสุ่มลดตัวอย่างส่วนมาก ไม่สามารถช่วยเพิ่มค่าความระลึกได้ โดยมีค่าเฉลี่ยทุกแบบจำลองเท่ากับ 63.09% 3) เทคนิคการสุ่มตัวอย่างซ้ำ สามารถช่วยเพิ่มค่าความระลึกได้สูงสุด โดยมีค่าเฉลี่ยทุกแบบจำลอง เท่ากับ 80.43% ซึ่งแบบจำลองส่วนใหญ่มีค่าเพิ่มสูงขึ้นยกเว้นแบบจำลอง SVM, A+SVM และ B+SVM พบว่ามีค่าลดลงเมื่อเปรียบเทียบกับแบบจำลองเดียวกันในชุดข้อมูลเดิม

กลุ่มชุดข้อมูล Pima Diabetes พบว่า 1) เทคนิคการสุ่มเพิ่มตัวอย่างส่วนน้อย สามารถช่วยเพิ่มค่าความระลึกได้ โดยมีค่าเฉลี่ยทุกแบบจำลองเท่ากับ 74.11% ซึ่งแบบจำลองส่วนใหญ่มีค่าเพิ่มสูงขึ้น ยกเว้นแบบจำลอง SVM, A+SVM, B+SVM และ B+MLP พบว่ามีค่าลดลงเมื่อเปรียบเทียบกับแบบจำลองเดียวกันในชุดข้อมูลเดิม 2) เทคนิคการสุ่มลดตัวอย่างส่วนมาก ไม่สามารถช่วยเพิ่มค่าความความระลึกได้ โดยมีค่าเฉลี่ยทุกแบบจำลองเท่ากับ 68.00% ซึ่งมีเพียงแบบจำลอง B+J48 และ B+CART พบว่ามีค่าเพิ่มสูงขึ้นเล็กน้อยเมื่อเปรียบเทียบกับแบบจำลองเดียวกันในชุดข้อมูลเดิม 3) เทคนิคการสุ่มตัวอย่างซ้ำสามารถช่วยเพิ่มค่าความระลึกได้สูงที่สุดจากค่าเฉลี่ยทุกแบบจำลอง เท่ากับ 86.55%

กลุ่มชุดข้อมูล Diabetic foot พบว่า 1) เทคนิคการสุ่มเพิ่มตัวอย่างส่วนน้อย สามารถช่วยเพิ่มค่าความระลึกได้ โดยมีค่าเฉลี่ยทุกแบบจำลองเท่ากับ 76.00% 2) เทคนิคการสุ่มลดตัวอย่างส่วนมาก ไม่สามารถช่วยเพิ่มค่าความระลึกได้ โดยมีค่าเฉลี่ยทุกแบบจำลองเท่ากับ 55.20% 3) เทคนิคการสุ่มตัวอย่างซ้ำ สามารถช่วยเพิ่มค่าความระลึกได้สูงสุด โดยมีค่าเฉลี่ยทุกแบบจำลอง เท่ากับ 82.86%

4.3.3 ค่าเอฟเฉลี่ยถ่วงน้ำหนัก

ค่าเอฟเฉลี่ยถ่วงน้ำหนัก คือ ค่าที่แสดงถึงประสิทธิภาพโดยรวมของแบบจำลองจากค่าความแม่นยำและค่าความระลึกที่สามารถจำแนกข้อมูลว่าเป็นผู้ไม่ป่วยและผู้ป่วยได้ถูกต้อง แสดงในรูปแบบร้อยละได้ในตารางที่ 12

จากตารางที่ 12 สามารถแสดงค่าเอฟเฉลี่ยถ่วงน้ำหนักของแบบจำลองได้ตามกลุ่มชุดข้อมูล ดังนี้

กลุ่มชุดข้อมูล Breast Cancer พบว่า 1) เทคนิคการสุ่มเพิ่มตัวอย่างส่วนน้อย สามารถช่วยเพิ่มค่าเอฟได้โดยมีค่าเฉลี่ยทุกแบบจำลองเท่ากับ 75.18% 2) เทคนิคการสุ่มลดตัวอย่างส่วนมาก ไม่สามารถช่วยเพิ่มค่าเอฟได้โดยมีค่าเฉลี่ยทุกแบบจำลองเท่ากับ 63.03% ซึ่งมีเพียงแบบจำลอง SVM และ B+SVM พบว่ามีค่าเพิ่มขึ้นเล็กน้อยเมื่อเปรียบเทียบกับแบบจำลองเดียวกันในชุดข้อมูลเดิม 3) เทคนิคการสุ่มตัวอย่างซ้ำ สามารถช่วยเพิ่มค่าเอฟได้สูงสุด โดยมีค่าเฉลี่ยทุกแบบจำลอง เท่ากับ 78.37% ซึ่งแบบจำลองส่วนใหญ่มีค่าเพิ่มสูงขึ้นยกเว้นแบบจำลอง SVM, A+SVM และ B+SVM พบว่ามีค่าลดลงเมื่อเปรียบเทียบกับแบบจำลองเดียวกันในชุดข้อมูลเดิม

กลุ่มชุดข้อมูล Pima Diabetes พบว่า 1) เทคนิคการสุ่มเพิ่มตัวอย่างส่วนน้อยสามารถช่วยเพิ่มค่าเอฟได้โดยมีค่าเฉลี่ยทุกแบบจำลองเท่ากับ 73.03% ซึ่งแบบจำลองส่วนใหญ่มีค่าเพิ่มสูงขึ้น ยกเว้นแบบจำลอง SVM และ A+SVM พบว่ามีค่าลดลงเมื่อเปรียบเทียบกับแบบจำลองเดียวกันในชุดข้อมูลเดิม 2) เทคนิคการสุ่มลดตัวอย่างส่วนมาก ไม่สามารถช่วยเพิ่มค่าเอฟได้โดยมีค่าเฉลี่ยทุกแบบจำลองเท่ากับ 67.59% ซึ่งมีเพียงแบบจำลอง B+J48 และ B+CART พบว่ามีค่าเพิ่มสูงขึ้นเล็กน้อยเมื่อเปรียบเทียบกับแบบจำลองเดียวกันในชุดข้อมูลเดิม 3) เทคนิคการสุ่มตัวอย่างซ้ำ สามารถช่วยเพิ่มค่าเอฟได้สูงที่สุดจากค่าเฉลี่ยทุกแบบจำลอง เท่ากับ 86.22%

กลุ่มชุดข้อมูล Diabetic foot พบว่า 1) เทคนิคการสุ่มเพิ่มตัวอย่างส่วนน้อย สามารถช่วยเพิ่มค่าเอฟได้ โดยมีค่าเฉลี่ยทุกแบบจำลองเท่ากับ 75.95% 2) เทคนิคการสุ่มลดตัวอย่างส่วนมาก ไม่สามารถช่วยเพิ่มค่าเอฟได้ โดยมีค่าเฉลี่ยทุกแบบจำลองเท่ากับ 55.11% 3) เทคนิคการสุ่มตัวอย่างซ้ำ สามารถช่วยเพิ่มค่าเอฟได้สูงสุด โดยมีค่าเฉลี่ยทุกแบบจำลอง เท่ากับ 82.22%

5. วิจารณ์และสรุปผลการทดลอง

งานวิจัยนี้มีจุดประสงค์เพื่อเปรียบเทียบเทคนิคการสุ่มตัวอย่างเพื่อการจำแนกข้อมูลที่ไม่สมดุล จากชุดข้อมูลที่ไม่สมดุล 3 ชุดข้อมูล ใช้เทคนิควิธีการสุ่มตัวอย่างซ้ำ การสุ่มเพิ่มตัวอย่างกลุ่มน้อย และการสุ่มลดตัวอย่างกลุ่มมาก รวมทั้งหมด 12 ชุดข้อมูล จากนั้นจึงนำชุดข้อมูลทั้งหมดไปสร้างแบบจำลองด้วยเทคนิควิธีจำนวน 15 เทคนิค จากผลการทดลองที่ได้สามารถวิเคราะห์และสรุปผลการทดลองได้ดังนี้

การจำแนกข้อมูลผู้ไม่ป่วยซึ่งเป็นข้อมูลส่วนมากในชุดข้อมูล แบบจำลองที่สร้างขึ้นสามารถจำแนกข้อมูลได้ดีขึ้นโดยเฉพาะชุดข้อมูลที่ใช้เทคนิคการสุ่มซ้ำ แต่ในชุดข้อมูลที่ใช้เทคนิคการสุ่มเพิ่มตัวอย่างส่วนน้อยและการสุ่มลดตัวอย่างส่วนมากกลับมีค่าประสิทธิภาพที่ลดลง

การจำแนกข้อมูลผู้ป่วยซึ่งเป็นข้อมูลส่วนน้อยในชุดข้อมูล แบบจำลองที่สร้างขึ้นสามารถจำแนกข้อมูลได้ดีขึ้นทั้งสามชุดข้อมูล

เทคนิคการสุ่มซ้ำ จากการสุ่มข้อมูลตัวอย่างขนาดเท่าเดิมของชุดข้อมูล และเมื่อนำแบบจำลองที่ได้มาทดสอบ พบว่า สามารถช่วยเพิ่มคุณภาพของข้อมูลและช่วยเพิ่มประสิทธิภาพของแบบจำลองได้ดีขึ้นทุกชุดข้อมูลในงานวิจัยสอดคล้องกับงานวิจัยของ (สมภพ ปฐมนพ และคณะ, 2555) เนื่องจากข้อมูลที่สุ่มขึ้นมาใหม่นั้นมีคุณภาพที่ดีขึ้น จึงส่งผลให้แบบจำลองได้เรียนรู้ข้อมูลที่มีคุณภาพและสามารถสร้างแบบจำลองที่สามารถจำแนกข้อมูลได้ดียิ่งขึ้น เห็นได้จากค่าประสิทธิภาพการจำแนกข้อมูลผู้ไม่ป่วยซึ่งเป็นข้อมูลส่วนมากในชุดข้อมูล การจำแนกข้อมูลผู้ป่วยซึ่งเป็นข้อมูลส่วนน้อยในชุดข้อมูล และผลรวมเฉลี่ยถ่วงน้ำหนัก

เทคนิคการสุ่มเพิ่มตัวอย่างส่วนน้อย ด้วยการสุ่มเพิ่มข้อมูลส่วนน้อยให้มีจำนวนเท่าข้อมูลส่วนมาก จากผลการทดลองที่ได้ พบว่า สามารถช่วยเพิ่มประสิทธิภาพการจำแนกของแบบจำลองได้ดี ซึ่งมีความสอดคล้องกับงานวิจัยของ (สุรพงษ์ เชี่ยวสกุลวัฒนาและสุกรี สินธุภิญโญ, 2557) (ภรัณยา ปาลวิสุทธิ์, 2559) และ (วีระยุทธ มายุศิริ และคณะ, 2557) ที่ได้นำเทคนิคการสุ่มเพิ่มตัวอย่างส่วนน้อยมาช่วยเพิ่มคุณภาพให้กับข้อมูลและสามารถสร้างแบบจำลองที่มีประสิทธิภาพดีขึ้นให้กับงานวิจัยได้ เห็นได้จากค่าประสิทธิภาพการจำแนกข้อมูลผู้ไม่ป่วยซึ่งเป็นข้อมูลส่วนมากในชุดข้อมูล การจำแนกข้อมูลผู้ป่วยซึ่งเป็นข้อมูลส่วนน้อยในชุดข้อมูล และผลรวมเฉลี่ยถ่วงน้ำหนัก

เทคนิคการสุ่มลดตัวอย่างส่วนมาก ด้วยการสุ่มลดข้อมูลส่วนมากลงให้มีจำนวนเท่ากับข้อมูลส่วนน้อย จากผลการทดลองที่ได้ พบว่า เมื่อนำข้อมูลที่ใช้เทคนิคนี้ไปสร้างแบบจำลองแล้วกลับพบว่าแบบจำลองที่ได้มีประสิทธิภาพต่ำกว่าข้อมูลชุดเดิม ซึ่งในการลดข้อมูลส่วนมากลงอาจลดข้อมูลที่มีความสำคัญต่อการสร้างแบบจำลองออกไปประกอบกับชุดข้อมูลที่ใช้ในงานวิจัยมีจำนวนระเบียนที่น้อย จึงส่งผลให้แบบจำลองที่ได้มีค่าประสิทธิภาพที่ลดลง

การสร้างแบบจำลองด้วยเทคนิควิธีจำนวน 15 เทคนิค ในผลการทดลอง ค่าเอฟ พบว่า เทคนิคเรนดอมฟอรเรส (RF) สามารถสร้างแบบจำลองที่มีประสิทธิภาพโดยรวมสูงสุดในงานวิจัย และเมื่อนำเทคนิครวมกลุ่มมาช่วย คือ เอดาบูทและถุงจำแนก ยิ่งสามารถช่วยเพิ่มประสิทธิภาพให้กลับแบบจำลองได้ดีขึ้น จึงสามารถสรุปงานวิจัยนี้ได้ว่า เทคนิคเรนดอมฟอรเรส (RF), เอดาบูทร่วมกับเรนดอมฟอรเรส (A+RF) และถุงจำแนกร่วมกับเรนดอมฟอรเรส (B+RF) เป็นเทคนิคที่มีความเหมาะสมที่จะนำไปสร้างแบบจำลองกับข้อมูลที่ใช้เทคนิคสุ่มข้อมูลซ้ำและเทคนิคสุ่มเพิ่มข้อมูลส่วนน้อยเพื่อให้แบบจำลองที่ได้มีประสิทธิภาพสูงขึ้น

งานวิจัยนี้สามารถนำเทคนิคการสุ่มตัวอย่างที่เสนอไปใช้แก้ไขกับข้อมูลที่ไม่สมดุล และช่วยเพิ่มค่าประสิทธิภาพในการจำแนกของแบบจำลองได้ผลดีขึ้น โดยเฉพาะเทคนิคการสุ่มซ้ำที่มีผลการวิจัยดีที่สุดในทุกชุดข้อมูลจึงเป็นเทคนิคที่มีความเหมาะสมต่อการนำไปปรับปรุงข้อมูลก่อนนำข้อมูลไปสร้างแบบจำลอง

เนื่องจากงานวิจัยนี้เลือกใช้วิธีการทดสอบประสิทธิภาพแบบจำลองด้วยเทคนิค 10-fold cross validation จึงอาจส่งผลให้ข้อมูลที่ถูกสุ่มเพิ่มขึ้นหรือสุ่มลดลงมีโอกาสถูกนำไปเป็นทั้งข้อมูลสอนและข้อมูลทดสอบได้ จึงอาจส่งผลให้แบบจำลองของงานวิจัยมีความเข้ากันเกินไป (Over Fitting) ผู้วิจัยจึงมีแนวคิดที่จะเลือกใช้วิธีการทดสอบประสิทธิภาพแบบจำลองด้วยวิธี Hold Out พร้อมทั้งปรับปรุงแบบจำลองให้มีค่าประสิทธิภาพที่ดียิ่งขึ้นด้วยการเพิ่มเทคนิควิธีทั้งเทคนิคการปรับปรุงข้อมูล เทคนิคการสุ่มข้อมูล เทคนิคการจำแนกข้อมูล และการปรับแต่งค่าพารามิเตอร์ให้ดียิ่งขึ้น รวมไปถึงการเพิ่มข้อมูลให้มากขึ้นในงานวิจัยต่อไป

เอกสารอ้างอิง

เดช ธรรมศิริ และพยุง มีสัจ. (2554). การเรียนรู้แบบรวมกลุ่มด้วยโครงข่ายประสาทเทียมเอดาบูทสำหรับการจำแนกข้อมูล. วารสารเทคโนโลยีสารสนเทศ, 7(14), 7-12.

บรรจบ ดลกุล, จารี ทองคำ และวาทินี สุขมาก. (2557). การสร้างแบบจำลองเพื่อพยากรณ์การเกิดแผลที่เท้าของผู้ป่วยโรคเบาหวานโดยใช้เทคนิคเหมืองข้อมูล. วารสารวิทยาศาสตร์และเทคโนโลยี มหาวิทยาลัยมหาสารคาม, 33(6), 703-710.

ปณต ทรงวัฒนศิริ. (2553). เทคนิคการสุ่มเพิ่มตัวอย่างข้างน้อยสังเคราะห์และเทคนิคการสุ่มลดตัวอย่างข้างมากสำหรับปัญหาความไม่สมดุลระหว่างกลุ่ม. วิทยานิพนธ์วิทยาศาสตรมหาบัณฑิต คณะวิทยาศาสตร์ จุฬาลงกรณ์มหาวิทยาลัย.

ภรัณยา ปาลวิสุทธิ์. (2559). การเพิ่มประสิทธิภาพเทคนิคต้นไม้ตัดสินใจบนชุดข้อมูลที่ไม่สมดุลโดยวิธีการสุ่มเพิ่มตัวอย่างกลุ่มน้อยสำหรับข้อมูลการเป็นโรคติดอินเทอร์เน็ต. วารสารเทคโนโลยีสารสนเทศ, 12(1), 54-63.

ภาสพิชญ์ ชูใจ. (2557). การเรียนรู้ร่วมกันสำหรับปัญหาการจำแนกข้อมูลไม่สมดุล. วิทยานิพนธ์วิศวกรรมศาสตรดุษฎีบัณฑิต, สาขาวิชาวิศวกรรมคอมพิวเตอร์, สำนักวิศวกรรมศาสตร์, มหาวิทยาลัยเทคโนโลยีสุรนารี.

ภาสพิชญ์ ชูใจ, นิตยา เกิดประสพ และกิตติศักดิ์ เกิดประสพ. (2557). กระบวนการเตรียมข้อมูลสำหรับข้อมูลไม่สมดุลเพื่อเพิ่มประสิทธิภาพในการจำแนก. การประชุมวิชาการระดับชาติมหาวิทยาลัยเทคโนโลยีราชมงคล ครั้งที่ 6, อาคารเทคโนโลยีสารสนเทศ มหาวิทยาลัยเทคโนโลยีราชมงคลสุวรรณภูมิ, จังหวัดพระนครศรีอยุธยา, 23-25 กรกฎาคม 2557.

วีระยุทธ มายุศิริ, จารี ทองคำ และวาทินี สุขมาก. (2557). การพัฒนาแบบจำลองเพื่อการพยากรณ์การรักษาซ้ำของผู้ป่วยโรคจิตเภทโดยเทคนิคเหมืองข้อมูล. วารสารฉบับพิเศษ ประชุมวิชาการ มหาวิทยาลัยมหาสารคาม, 10, 144-153.

สมภพ ปฐมนพ, กฤษฎา ศรีแผ้ว และกุลธร เกษมสันต์,ม.ล. (2555). ข้อมูลเชิงเวลากับการจำแนกประเภทผู้เป็นโรคเบาหวานในประเทศไทย. Journal of Information Science and Technology, 3(2), 14-21.

สุรพงษ์ เชี่ยวสกุลวัฒนา และสุกรี สินธุภิญโญ. (2557). การจำแนกต้นไม้ตัดสินใจสำหรับชุดข้อมูลไม่สมดุลโดยใช้น้ำหนักต่างกันบนข้อมูลสังเคราะห์. การประชุมวิชาการระดับชาติด้านคอมพิวเตอร์และเทคโนโลยีสารสนเทศ ครั้งที่ 10, โรงแรมอังสนาลากูนา จังหวัดภูเก็ต, 8-9 พฤษภาคม 2557.

เสกสันติ จันทะมงคล, สมจิตร อาจอินทร์, งามนิจ อาจอินทร์ และวัชชพล เดชขันธ์. (2555). ระบบช่วยเหลืออัจฉริยะเพื่อการวางแผนการรักษาโรคเรื้อรังโดยใช้เหมืองข้อมูล. การประชุมวิชาการระดับชาติด้านคอมพิวเตอร์และเทคโนโลยีสารสนเทศ ครั้งที่ 8, โรงแรมดุสิตธานีพัทยา จังหวัดชลบุรี, 9-10 พฤษภาคม 2555.

Breiman, L. (1996). Bagging predictors. Machine Learning, 24(2), 123-140.

Breiman, L. (2001). Random Forests. Machine Learning, 45(1), 5-32.

Breiman, L., Friedman, J., Stone, C. J., & Olshen, R. A. (1984). Classification and Regression Trees : Wadsworth Inc, Pacific Grove, CA.

Chawla, N. V. (2005). Data Mining for Imbalanced Datasets : An Overview. In O. Maimon & L. Rokach (Eds.), Data Mining and Knowledge Discovery Handbook (pp. 853-867). Boston, MA: Springer US.

Chawla, N. V., Bowyer, K. W., Hall, L. O., & Kegelmeyer, W. P. (2002). SMOTE: synthetic minority over-sampling technique. J. Artif. Int. Res., 16(1), 321-357.

Cortes, C., & Vapnik, V. (1995). Support-vector networks. Machine Learning, 20(3), 273-297.

Efron, B. (1987). Better Bootstrap Confidence Intervals. Journal of the American Statistical Association, 82(397), 171-185.

Freund, Y., & Schapire, R. E. (1997). A Decision-Theoretic Generalization of On-Line Learning and an Application to Boosting. Journal of Computer and System Sciences, 55(1), 119-139.

Kohavi, R. (1995). A study of cross-validation and

bootstrap for accuracy estimation and model selection. Paper presented at the Proceedings of the 14th international joint conference on Artificial intelligence - Volume 2, Montreal, Quebec, Canada.

Larose, D. T., & Larose, C. D. (2014). Neural Networks Discovering Knowledge in Data (pp. 187-208) : John Wiley & Sons Inc, Canada.

Meng, X.-H., Huang, Y.-X., Rao, D.-P., Zhang, Q., & Liu, Q. (2013). Comparison of three data mining models for predicting diabetes or prediabetes by risk factors. The Kaohsiung Journal of Medical Sciences, 29(2), 93-99.

Powers D.M.W. (2011). Evaluation : From Precision, Recall and F-measure To ROC, Informedness & Correlation Journal of Machine Learning Technologies, 2(1), pp. 37-63.

Quinlan, J. R. (1986). Induction of Decision Trees. Machine Learning., 1(1), 81-106.

Quinlan, J. R. (1993). C4.5: programs for machine learning : Morgan Kaufmann Publishers Inc, California.

University of Waikato. (2017). Weka Machine Learning. Retrieved from http://www.cs.waikato.ac.nz/ml/weka/