การปรับปรุงประสิทธิภาพต้นไม้ตัดสินใจแบบจำแนกและแบบถดถอย ด้วยเทคนิคการสุ่มซ้ำ

นางสาวศุภจิตรี บุญอำนวย

วิทยานิพนธ์นี้เป็นส่วนหนึ่งของการศึกษาตามหลักสูตรปริญญาวิศวกรรมศาสตรมหาบัณฑิต
สาขาวิชาวิศวกรรมคอมพิวเตอร์
มหาวิทยาลัยเทคโนโลยีสุรนารี
ปีการศึกษา 2562

IMPROVING THE CLASSIFICATION AND REGRESSION TREE PERFORMANCE WITH RESAMPLING TECHNIQUE

Supajittree Boonamnuay

A Thesis Submitted in Partial Fulfillment of the Requirements for the

Degree of Master of Engineering in Computer Engineering

Suranaree University of Technology

Academic Year 2019

การปรับปรุงประสิทธิภาพต้นไม้ตัดสินใจแบบจำแนกและแบบถดถอย
ด้วยเทคนิคการสุ่มซ้ำ

มหาวิทยาลัยเทคโนโลยีสุรนารี อนุมัติให้นับวิทยานิพนธ์ฉบับนี้เป็นส่วนหนึ่งของการศึกษาตามหลักสูตรปริญญามหาบัณฑิต

คณะกรรมการสอบวิทยานิพนธ์

(รศ. ดร.กิตติศักดิ์ เกิดประสพ)
ประธานกรรมการ

(รศ. ดร.นิตยา เกิดประสพ)
กรรมการ (อาจารย์ที่ปรึกษาวิทยานิพนธ์)

(อ. ดร.นันทวุฒิ คะอังกุ)
กรรมการ

(รศ. ร.อ. ดร.กนต์ธร ชำนิประศาสน์)
รองอธิการบดีฝ่ายวิชาการและพัฒนาความเป็นสากล

(รศ. ดร. พรศิริ จงกล)
คณบดีสำนักวิชาวิศวกรรมศาสตร์

ศุภจิตรี บุญอำนวย : การปรับปรุงประสิทธิภาพต้นไม้ตัดสินใจแบบจำแนกและแบบถดถอยด้วยเทคนิคการสุ่มซ้ำ (IMPROVING THE CLASSIFICATION AND REGRESSION TREE PERFORMANCE WITH RESAMPLING TECHNIQUE) อาจารย์ที่ปรึกษา : รองศาสตราจารย์ ดร.นิตยา เกิดประสพ, 63 หน้า.

ปัจจุบันการทำเหมืองข้อมูลได้รับความนิยมอย่างมากในด้านการวิเคราะห์ข้อมูล โดยการนำโมเดลที่ได้จากกระบวนการเรียนรู้มาใช้ในการจำแนกข้อมูล แบ่งกลุ่มข้อมูล หรือทำนายข้อมูลในอนาคต ในอัลกอริทึมเหมืองข้อมูลที่มีหลากหลายอัลกอริทึมต้นไม้ตัดสินใจแบบจำแนก และแบบถดถอย เป็นอัลกอริทึมที่มีจุดเด่นคือสามารถนำมาวิเคราะห์ข้อมูลได้ทั้งตัวเลข และข้อความทั้งยังมีประสิทธิภาพค่อนข้างสูงเมื่อนำมาวิเคราะห์ข้อมูลที่มีลักษณะไม่สมดุล ซึ่งเป็นลักษณะของข้อมูลส่วนใหญ่ในชีวิตจริง งานวิจัยนี้จึงเสนอวิธีการเพิ่มประสิทธิภาพอัลกอริทึมต้นไม้แบบจำแนกและแบบถดถอยด้วยการนำเทคนิคการสุ่มซ้ำเข้ามาช่วยในขั้นตอนการเตรียมข้อมูลด้วยการหาอัตราส่วนระหว่างจำนวนข้อมูลคลาสส่วนมากต่อจำนวนข้อมูลคลาสส่วนน้อย ซึ่งจากการทดลองปรากฏว่าเมื่ออัตราส่วนของสองคลาสเริ่มใกล้เคียงกัน ค่าประสิทธิภาพของการจำแนกข้อมูลยิ่งสูงขึ้น โดยเฉพาะเมื่ออัตราส่วนเป็น 50 : 50 และเนื่องด้วยข้อมูลของการวิจัยนี้ค่าของแอททริบิวต์ต่าง ๆ นอกเหนือจากคลาสเป้าหมายเป็นข้อมูลตัวเลขทั้งหมด จึงได้ทดลองนำเทคนิคการจัดกลุ่มข้อมูลเข้ามาช่วย โดยจัดกลุ่มข้อมูลคลาสส่วนมากให้เท่ากับจำนวนข้อมูลคลาสส่วนน้อย แล้วใช้ค่าเฉลี่ยของข้อมูลแต่ละแอททริบิวต์ในแต่ละกลุ่มเป็นข้อมูลตัวแทนของข้อมูลชุดนั้น ๆ ซึ่งจากวิธีการดังกล่าว สามารถสร้างแบบจำลองที่จำแนกคลาสน้อยได้ดียิ่งขึ้น

สาขาวิชา วิศวกรรมคอมพิวเตอร์ ลายมือชื่อนักศึกษา ศุภจิตรี บุญอำนวย

ปีการศึกษา 2562 ลายมือชื่ออาจารย์ที่ปรึกษา

SUPAJITTREE BOONAMNUAY : IMPROVING THE CLASSIFICATION AND REGRESSION TREE PERFORMANCE WITH RESAMPLING TECHNIQUE. THESIS ADVISOR : ASSOC. PROF. NITTAYA KERDPRASOP, Ph.D., 63 PP.

CLASSIFICATION AND REGRESSION TREE/RESAMPLING TECHNIQUE

Data mining is a very popular method for data analysis by introducing models derived from the learning process for classification, clustering or predicting. From numerous data mining algorithms, Classification and Regression Tree (CART) algorithm is the prominent one with its advantage of being able to analyze data in both numeric and categorical forms. CART performance is also highly effective when analyze imbalanced data that is the normally found in real life. This research proposes to improve CART performance by using resampling technique to find the appropriate ratio between the number of majority class and the number of minority class that can best improve CART performance. When we try to randomly reduce the number of majority class to a close proportion of minority class, CART shows better discriminative performance, and the best ratio is 50:50. Since the value of all attributes other than the target class in this research is numeric, clustering technique is also used to help grouping majority class to be equal to minority class. The mean value of each attribute in each group is the representative data. From the applied technique, models can detect more of minority class and have good performance. Finally hopefully this research will be useful in the future.

School of Computer Engineering Student's Signature Supajittree Boonamnuay

Academic Year 2019 Advisor's Signature Nittaya

กิตติกรรมประกาศ

วิทยานิพนธ์นี้สำเร็จลุล่วงด้วยดี ผู้วิจัยขอกราบขอบพระคุณ บุคคล และกลุ่มบุคคลที่ได้กรุณาให้คำปรึกษา แนะนำ ช่วยเหลืออย่างดียิ่ง ทั้งในด้านวิชาการ และด้านการดำเนินงานวิจัย ดังต่อไปนี้

รองศาสตราจารย์ ดร.นิตยา เกิดประสพ อาจารย์ที่ปรึกษาวิทยานิพนธ์ และรองศาสตราจารย์ ดร.กิตติศักดิ์ เกิดประสพ ที่ให้คำปรึกษาคำแนะนำในการทำงานวิจัย ไปจนถึงการจัดรูปแบบวิทยานิพนธ์ และช่วยตรวจทานความถูกต้องของวิทยานิพนธ์

คุณอนุสรา หิรัญวนากุล คุณอนุพงษ์ บรรจงการ และคุณวรรธนะ พงษ์เสนา ที่ให้ความช่วยเหลือในการสอนใช้งานโปรแกรม SPSS Modeler รวมไปถึงช่วยตรวจทาน ติชม เพื่อปรับปรุงแก้ไขวิทยานิพนธ์ให้สมบูรณ์ยิ่งขึ้น

คุณปราณี กฐินใหม่ และคุณดารณี ทิพย์ทอง เลขานุการ และผู้ช่วยสอนประจำสาขาวิชาวิศวกรรมคอมพิวเตอร์ ที่ให้ความช่วยเหลือในการประสานงานระหว่างศึกษา

ขอขอบคุณนักศึกษาร่วมชั้นเรียนทั้งปริญญาโท และปริญญาเอก ที่ให้คำแนะนำคำปรึกษาด้านวิชาการ และช่วยสนับสนุนด้วยดีมาตลอด

ขอบคุณมหาวิทยาลัยเทคโนโลยีสุรนารี ที่ให้การสนับสนุนทุนการศึกษา ทุนวิจัย ทั้งยังค่าใช้จ่ายต่าง ๆ

นอกจากนี้ขอขอบคุณ ครู อาจารย์ ทั้งในอดีตและปัจจุบันที่ให้ความรู้แก่ผู้วิจัยอย่างมากมายจนประสบความสำเร็จ

ท้ายที่สุดขอกราบขอบพระคุณ บิดา มารดา ที่ให้กำเนิด อบรม เลี้ยงดู ส่งเสริมให้ผู้วิจัยมีความรู้ ความสามารถ ทั้งยังเป็นกำลังใจแก่ผู้วิจัยจนประสบความสำเร็จในชีวิต

ศุภจิตรี บุญอำนวย

สารบัญ

หน้า

สารบัญ (ต่อ)

หน้า

สารบัญตาราง

ตารางที่ **หน้า**

สารบัญรูป

รูปที่ หน้า

บทที่ 1
บทนำ

1.1 ความสำคัญและที่มาของปัญหาการวิจัย

การจำแนกประเภทข้อมูล (Data Classification) เป็นการทำเหมืองข้อมูลประเภทหนึ่ง (Han et al., 2011) ที่ได้รับความนิยมในการนำไปแก้ปัญหาทางวิทยาศาสตร์ อุตสาหกรรม เศรษฐศาสตร์ การแพทย์ และอื่น ๆ อีกมากมาย จากกระบวนการหารูปแบบ หรือหาความสัมพันธ์ของข้อมูล ซึ่งไม่ว่าศาสตร์ใดล้วนต้องการการวิเคราะห์ข้อมูลที่แม่นยำ และมีประสิทธิภาพที่สามารถนำไปใช้งานได้ในชีวิตจริง ซึ่งในปัจจุบันข้อมูลในฐานข้อมูลล้วนมีขนาดใหญ่ หลายตัวแปร หลายปัจจัย รวมไปถึงมีจำนวนหลายชุดข้อมูล ซึ่งหากทำการวิเคราะห์โดยประสบการณ์ของแต่ละบุคคล หรือวิเคราะห์ด้วยหลักทางสถิติศาสตร์ อาจก่อให้เกิดความล่าช้า และความผิดพลาดได้ อีกทั้งทำให้สิ้นเปลืองเวลาและงบประมาณในการวิเคราะห์ข้อมูล ดังนั้นการทำเหมืองข้อมูลจึงเป็นทางเลือกที่สามารถช่วยลดปัญหาในการวิเคราะห์ข้อมูลได้อย่างมีประสิทธิภาพ

จากเหตุผลข้างต้นการทำเหมืองข้อมูลจึงเป็นทางเลือกที่ดีในการช่วยวิเคราะห์ข้อมูล แต่ก็จำเป็นต้องเลือกวิธีการให้เหมาะสมกับประเภทของข้อมูลนั้น ๆ โดยในการวิจัยนี้เน้นไปที่การจำแนกประเภทข้อมูล เพื่อหาความสัมพันธ์ หรือกฎที่เป็นประโยชน์ในการนำไปใช้ อัลกอริทึมที่เป็นที่นิยมสำหรับการจำแนกข้อมูลนั้นคือ ต้นไม้ตัดสินใจ (Decision Tree) ซึ่งเป็นเทคนิคในการจำแนกข้อมูลออกเป็นโหนด โดยเริ่มต้นที่โหนดราก (Root Node) แล้วแตกโหนดใบ (Leaf Node) ออกไปคล้ายลักษณะของต้นไม้ ซึ่งแต่ละโหนดหมายถึงแอททริบิวต์ (Attribute) ที่นำมาเป็นเงื่อนไข และค่าที่เป็นเกณฑ์ในการจำแนก การนำอัลกอริทึมต้นไม้ตัดสินใจมาใช้จำแนกข้อมูลที่มีการกระจายแบบปกติ (Normal Distribution) หรือข้อมูลที่มีความสมดุล ปรากฏว่าสามารถวิเคราะห์ข้อมูลโดยให้ค่าความแม่นยำสูง รวมไปถึงกฎที่ได้นั้นสามารถทำความเข้าใจได้ง่าย เหมาะแก่การนำไปประยุกต์ใช้ในทุก ๆ ศาสตร์ แต่ปัจจุบันข้อมูลส่วนใหญ่ในชีวิตจริงส่วนมากมักมีลักษณะไม่สมดุล (Imbalanced) ข้อมูลที่ไม่สมดุลจะมีจำนวนข้อมูลในแต่ละคลาสเป้าหมายที่

แตกต่างกัน (Chawla et al., 2002) เช่น มีจำนวนข้อมูล 500 ข้อมูล แบ่งเป็นข้อมูลในคลาส Positive จำนวน 480 ข้อมูล และข้อมูลในคลาส Negative จำนวน 20 ข้อมูล จะเห็นได้ว่าจำนวนข้อมูลในคลาส Positive มีจำนวนมากกว่าจำนวนข้อมูลในคลาส Negative เป็นจำนวนมาก ซึ่งเรียกข้อมูลในคลาส Positive ว่า คลาสส่วนมาก (Majority Class) และเรียกข้อมูลในคลาส Negative ว่า คลาสส่วนน้อย (Minority Class) เมื่อนำข้อมูลไม่สมดุลไปทำการจำแนกประเภทข้อมูลด้วยอัลกอริทึมแบบมาตรฐาน อย่างเช่น ต้นไม้ตัดสินใจ จะส่งผลให้การจำแนกมีความเอนเอียง (Bias) ไปทางคลาสส่วนมากเนื่องจากมีจำนวนข้อมูลมากกว่า ทำให้คลาสส่วนน้อยจัดกลุ่มผิดประเภท (Misclassification)

อัลกอริทึมต้นไม้ตัดสินใจแบบจำแนกและแบบถดถอย (Classification and Regression Tree, CART) (Brieman et al., 1984) เป็นอัลกอริทึมที่ได้รับความนิยมเนื่องจากเป็นอัลกอริทึมที่สามารถใช้วิเคราะห์ได้ทั้งข้อมูลตัวเลข (Numeric) และข้อมูลจำพวกข้อความหรือข้อมูลเชิงกลุ่ม (Categorical) ทำให้มีความยืดหยุ่นต่อการนำไปใช้กับข้อมูลในหลาย ๆ ศาสตร์ โดยอัลกอริทึมต้นไม้ตัดสินใจแบบจำแนกและแบบถดถอยนี้ได้มีผู้วิจัยนำมาใช้ในการวิเคราะห์ข้อมูลเกี่ยวกับสารพันธุกรรมหรือชื่อแบบเต็มคือกรดดีออกซีไรโบนิวคลีอิก (Deoxyribonucleic Acid, DNA) รวมไปถึงการนำไปวินิจฉัยโรค ซึ่งให้ประสิทธิภาพค่อนข้างสูง (ดิษฐพล มั่นธรรม และ ลี่ลี อิงศรีสว่าง, 2010) จึงได้นำอัลกอริทึมต้นไม้ตัดสินใจแบบจำแนกและแบบถดถอยมาวิเคราะห์ข้อมูลที่ไม่สมดุลซึ่งพบว่านอกจากเป็นอัลกอริทึมที่ให้ความแม่นยำสูงแล้ว จุดเด่นอีกข้อหนึ่งของอัลกอริทึมคือการพยายามแบ่งโหนดให้แยกคลาสข้อมูลได้อย่างชัดเจน หรือมีความบริสุทธิ์ให้มากที่สุด ซึ่งเป็นการเพิ่มประสิทธิภาพในการวิเคราะห์ข้อมูลที่ไม่สมดุลได้อีกทางหนึ่ง

งานวิจัยนี้เสนอเทคนิคการสุ่มซ้ำ (Resampling) ซึ่งจะนำมาใช้ในขั้นตอนการเตรียมข้อมูล (Data preprocessing) โดยได้นำกระบวนการสุ่มซ้ำเข้ามาช่วยในการหาอัตราส่วนระหว่างคลาสส่วนมากและคลาสส่วนน้อยก่อนจะนำมาประมวลผลด้วยอัลกอริทึมต้นไม้ตัดสินใจแบบจำแนกและแบบถดถอย เพื่อให้สามารถจำแนกพบคลาสส่วนน้อยได้ดีขึ้น โดยหาอัตราส่วนการสุ่มซ้ำที่ให้ประสิทธิภาพดีที่สุด เทคนิคการสุ่มซ้ำแบ่งออกเป็น 2 ประเภทคือ การสุ่มเพิ่ม (Oversampling) ซึ่งคือการเพิ่มจำนวนคลาสส่วนน้อย และการสุ่มลด (Undersamplimg) ซึ่งคือการลดจำนวนคลาสส่วนมาก ซึ่งในงานวิจัยนี้ได้ทำการทดลองโดยการสุ่มข้อมูลให้มีอัตราส่วนระหว่างคลาสส่วนมากและคลาสส่วนน้อยเป็น 70:30 50:50 และ 30:70 เพื่อหาว่าอัตราส่วนระหว่างคลาสส่วนมากและคลาสส่วนน้อยใดเหมาะแก่การนำมาเพิ่มประสิทธิภาพสำหรับข้อมูลไม่สมดุลทั้ง 5 ชุดจากฐานข้อมูล KEEL-dataset repository มากที่สุด ขั้นตอนถัดมาได้นำวิธีการจัดกลุ่มข้อมูล (Clustering) เข้ามาช่วยในการจัดกลุ่มคลาสส่วนมากและคลาสส่วนน้อยให้เป็นไปตามอัตราส่วน

จากขั้นตอนก่อนหน้า หลังจากนั้นเลือกใช้ค่าเฉลี่ยมาเป็นตัวแทนของแต่ละแอททริบิวต์ในกลุ่มนั้น ๆ เพื่อสร้างเป็นชุดข้อมูลใหม่ ที่มีคุณสมบัติใกล้เคียงกับข้อมูลชุดเดิมมากที่สุด เพื่อนำมาเปรียบเทียบประสิทธิภาพการจำแนกโดยอัลกอริทึมที่ไม่ผ่านการปรับปรุงข้อมูล

1.2 วัตถุประสงค์ของการวิจัย

1. เพื่อศึกษาและปรับปรุงประสิทธิภาพของแบบจำลองอัลกอริทึมต้นไม้ตัดสินใจแบบจำแนกและแบบถดถอยกับข้อมูลที่ไม่สมดุล
2. ทดสอบและเปรียบเทียบประสิทธิภาพแบบจำลองด้วยมาตรวัดต่าง ๆ ได้แก่ Accuracy, Precision, Recall หรือ Sensitivity, Specificity และ F-measure
3. เพื่อศึกษาแก้ไขปัญหาข้อมูลไม่สมดุลโดยใช้เทคนิคการสุ่มซ้ำเข้ามาช่วยเพิ่มประสิทธิภาพให้กับแบบจำลองต้นไม้ตัดสินใจแบบจำแนกและแบบถดถอย
4. เพื่อศึกษาหาอัตราส่วนระหว่างคลาสส่วนมากและคลาสส่วนน้อยที่เหมาะสมที่สุด

1.3 ขอบเขตของการวิจัย

1. การเปรียบเทียบประสิทธิภาพของแบบจำลองต้นไม้ตัดสินใจแบบจำแนกและแบบถดถอย ก่อน และหลังจากนำเทคนิคการสุ่มซ้ำเข้ามาช่วยเพิ่มประสิทธิภาพ พิจารณาโดยใช้มาตรวัดประสิทธิภาพทั้งหมด 4 มาตรวัด ดังนี้ Accuracy, Precision, Recall, Specificity และ F-measure
2. ข้อมูลสำหรับการวิจัยนี้ เป็นข้อมูลที่มีคลาสเป้าหมายเพียง 2 คลาส และชนิดข้อมูลในทุกแอททริบิวต์ยกเว้นเป้าหมายเป็นตัวเลขทั้งหมด
3. ข้อมูลสำหรับวิจัยเป็นข้อมูลไม่สมดุลจากฐานข้อมูล KEEL-dataset repository โดยการวิจัยนี้ใช้ข้อมูลจากฐานข้อมูลดังกล่าวดังนี้

   ข้อมูล Pima จำนวน 1 ชุดข้อมูล

   (http://sci2s.ugr.es/keel/dataset.php?cod=155)

   ข้อมูล Yeast จำนวน 1 ชุดข้อมูล

   (http://sci2s.ugr.es/keel/dataset.php?cod=153)

   ข้อมูล Vehicle จำนวน 1 ชุดข้อมูล

   (http://sci2s.ugr.es/keel/dataset.php?cod=149)

   ข้อมูล Segment จำนวน 1 ชุดข้อมูล

   (http://sci2s.ugr.es/keel/dataset.php?cod=148)

และ ข้อมูล Page-Block จำนวน 1 ชุดข้อมูล

(http://sci2s.ugr.es/keel/dataset.php?cod=147)

4. งานวิจัยนี้ใช้โปรแกรม IBM SPSS Modeler 18.0 ในการแบ่งข้อมูลและสร้างโมเดลเพื่อจำแนกข้อมูล

1.4 ประโยชน์ที่ได้รับ

จากการศึกษาและพัฒนางานวิจัยนี้ มีความมุ่งหวังว่าอัลกอริทึมต้นไม้ตัดสินใจแบบจำแนกและแบบถดถอยที่นำมาผสานกับเทคนิคการสุ่มซ้ำจะมีประโยชน์ต่อการนำมาใช้ช่วยวิเคราะห์ข้อมูลที่มีความไม่สมดุล โดยเฉพาะอย่างยิ่งเหมาะจะเป็นจุดเริ่มต้นสำหรับผู้ที่สนใจศึกษาค้นคว้าในประเด็นดังต่อไปนี้

1. ผลลัพธ์ที่ได้สามารถทำความเข้าใจได้ง่าย
2. ประสิทธิภาพของโมเดลที่ได้อยู่ในเกณฑ์ที่ดีขึ้น
3. ช่วยลดปัญหาโมเดลที่ไม่สามารถทำนายคลาสน้อยได้
4. ช่วยลดเวลาในการจัดการข้อมูลในขั้นตอนการเตรียมข้อมูล

บทที่ 2
ปริทัศน์วรรณกรรม

เนื้อหาในบทนี้ประกอบด้วยการทบทวนวรรณกรรมและงานวิจัยที่เกี่ยวข้อง โดยมีรายละเอียดเกี่ยวกับข้อมูลไม่สมดุล (Imbalanced Data) เทคนิคการสุ่มซ้ำ (Resampling) เทคนิคการจัดกลุ่มข้อมูล (Clustering) ทฤษฎีต้นไม้ตัดสินใจแบบจำแนกและแบบถดถอย (CART) การประเมินประสิทธิภาพของการจำแนกข้อมูล (Classification Evaluation) และงานวิจัยที่เกี่ยวข้อง

2.1 ข้อมูลไม่สมดุล (Imbalanced Data)

ข้อมูลไม่สมดุล หมายถึง ข้อมูลที่มีจำนวนข้อมูลในกลุ่มหนึ่งมีจำนวนมากกว่าข้อมูลในอีกกลุ่มหนึ่งเป็นจำนวนมาก (Chawla et al., 2002; Chawla et al., 2004) ข้อมูลไม่สมดุลนั้นมีสาเหตุมาจากหลายปัจจัย เช่น ข้อมูลไม่สมดุลที่เกิดจากลักษณะทางธรรมชาติของข้อมูลเองดังที่พบในข้อมูลทางสาธารณสุขที่พบว่ามีผู้ป่วยที่ป่วยเป็นโรคน้อยกว่าผู้ที่มีสุขภาพแข็งแรงเป็นจำนวนมาก หรือข้อมูลการผลิตสินค้าในอุตสาหกรรมที่ผลิตคราวละจำนวนมาก ซึ่งจำนวนสินค้าดีมีมากกว่าสินค้าเสีย เป็นต้น นอกจากนี้ข้อมูลไม่สมดุลอาจเกิดจากการเก็บข้อมูลที่ผิดพลาดด้วยเช่นกัน

เนื่องมาจากขนาดข้อมูลของคลาสหนึ่งมีจำนวนแตกต่างกันมากกับอีกคลาสหนึ่ง ตัวอย่างดังตารางที่ 2.1 จะเห็นว่าข้อมูลคลาส positive มีเพียงชุดเดียว โดยมักเรียกแทนคลาสที่มีจำนวนมากกว่าว่าคลาสส่วนมาก (Majority Class) และเรียกแทนคลาสที่มีจำนวนน้อยกว่าว่าคลาสส่วนน้อย (Minority Class) (Boonchuay et al., 2011; Farquad and Bose, 2012; Gao et al., 2012)

ตารางที่ 2.1 แสดงตัวอย่างข้อมูลไม่สมดุล

A	B	C	D	Class
x	y	z	x	negative
y	z	x	x	negative
z	x	y	y	negative
x	y	z	y	negative
y	z	x	z	negative
z	x	y	z	negative
x	y	z	y	positive

ตารางที่ 2.1 แสดงตัวอย่างข้อมูลไม่สมดุล (ต่อ)

A	B	C	D	Class
y	z	x	y	negative
z	x	y	x	negative
x	y	z	x	negative

สำหรับดัชนีที่บอกถึงความไม่สมดุลมากหรือน้อยนั้นสามารถดูได้จาก Imbalance Ratio (IR) ดังสมการที่ 2.1 (Orriosls-Puig et al., 2009; Villar et al., 2011) ซึ่งคือการคำนวณอัตราส่วนระหว่างจำนวนคลาสส่วนมากกับคลาสส่วนน้อย ซึ่งหากค่าดัชนีความไม่สมดุลมีค่าเท่ากับ 1 หมายความว่าจำนวนข้อมูลในคลาสส่วนมากมีจำนวนเท่ากับคลาสส่วนน้อย หรือข้อมูลมีความสมดุลกัน แต่หากค่าดัชนีมีค่ามากกว่า 1 หมายความว่าจำนวนข้อมูลคลาสส่วนมากมีจำนวนมากกว่าคลาสส่วนน้อย และหากค่าดัชนีมีค่าน้อยกว่า 1 หมายความว่าจำนวนข้อมูลคลาสส่วนน้อยมีจำนวนมากกว่าคลาสส่วนมาก

$$\text{Imbalance Ratio (IR)} = \frac{\text{Number of Majority Class}}{\text{Number of Minority Class}} \tag{2.1}$$

จากตัวอย่างในตารางที่ 2.1 จำนวนข้อมูลคลาส positive หรือคลาสส่วนน้อย = 1 และจำนวนข้อมูลคลาส negative หรือคลาสส่วนมาก = 9 จึงสามารถหาดัชนีความไม่สมดุลของข้อมูลชุดนี้ได้เป็น IR = 9/1 = 9.0 นั่นหมายความว่าข้อมูลตัวอย่างนี้มีระดับความไม่สมดุลเท่ากับ 9.0

จากปัญหาของข้อมูลไม่สมดุลโดยทั่วไปแบ่งวิธีการในการจัดการกับข้อมูลไม่สมดุลออกเป็น 3 ระดับได้แก่

ระดับที่ 1 : ระดับจัดการข้อมูล เป็นระดับที่ทำในขั้นตอนการเตรียมข้อมูลก่อนที่จะนำไปเรียนรู้เพื่อสร้างแบบจำลอง วิธีการที่นิยมใช้ในการปรับสมดุลข้อมูลก็คือการสุ่มซ้ำ ซึ่งนักวิจัยจำนวนมากวิจัยมาก่อนแล้วว่าเทคนิคการสุ่มซ้ำสามารถช่วยเพิ่มประสิทธิภาพในการจำแนกข้อมูลไม่สมดุลได้อย่างแน่นอน (Batista et al., 2004; Galar et al., 2012) ซึ่งงานวิจัยของวิทยานิพนธ์นี้เน้นที่การจัดการข้อมูลไม่สมดุลในระดับนี้

ระดับที่ 2 : ระดับจัดการอัลกอริทึม ในระดับนี้มักเป็นการเสนออัลกอริทึมใหม่ หรือแก้ไขอัลกอริทึมเดิมให้สามารถทำงานกับข้อมูลไม่สมดุลได้ดียิ่งขึ้น ไม่ว่าจะเป็นการตั้งเกณฑ์จำกัดเฉพาะให้แต่ละคลาสเพื่อแยกข้อมูลออกจากกัน (Weiss, 2004) หรือแม้กระทั่งการเรียนรู้เฉพาะข้อมูลบางส่วนอย่างเฉพาะเจาะจง (Cohen, 1995; Raskutti, 2004)

ระดับที่ 3 : ระดับการตั้งค่าน้ำหนักให้ข้อมูลแต่ละชุด เพื่อที่จะช่วยให้อัลกอริทึมสนใจข้อมูลในส่วนที่มีค่าน้ำหนักมากก่อน และพยายามละทิ้งข้อมูลที่ทำนายไม่ถูกกลุ่ม หรือคาดคะเนว่าถูกจัดไว้ผิดกลุ่มซึ่งถูกกำหนดไว้ให้มีค่าน้ำหนักน้อย (Longadge et al., 2013)

2.2 เทคนิคการสุ่มซ้ำ (Resampling)

งานวิจัยนี้เน้นที่การแก้ปัญหาข้อมูลไม่สมดุลในขั้นตอนก่อนที่จะมีการประมวลผล (Preprocessing) โดยการแก้ไขในระดับนี้จะแก้ไขกับข้อมูลโดยตรง โดยจะทำการปรับปรุงข้อมูลที่มีความไม่สมดุลให้กลายเป็นข้อมูลที่มีความสมดุลด้วยเทคนิคการสุ่มซ้ำ โดยจะแบ่งออกเป็น 2 กลุ่ม ได้แก่การสุ่มเกิน (Over Sampling) และการสุ่มลด (Under Sampling)

วิธีสุ่มเกิน (Over Sampling)

วิธีสุ่มเกิน (กิตติพงศ์ ชมบุญ, 2016) เป็นเทคนิค หรือวิธีที่ใช้ในการเพิ่มข้อมูลที่อยู่ในคลาสส่วนน้อย โดยการสุ่มเกินเพื่อเพิ่มข้อมูลให้คลาสส่วนน้อย โดยการสุ่มเลือกข้อมูลจากข้อมูลเดิม หรือสร้างข้อมูลขึ้นมาใหม่จากตัวอย่างของข้อมูลเดิม

จากตารางที่ 2.1 ชุดข้อมูล ซึ่งมีจำนวนคลาสทั้งหมด 2 คลาส ได้แก่คลาส positive และคลาส negative โดยข้อมูลในคลาส positive มีจำนวนทั้งหมด 1 ข้อมูล และข้อมูลในคลาส negative มีจำนวนทั้งหมด 1 ข้อมูล ซึ่งงานวิจัยนี้ใช้วิธีการเพิ่มตัวเอง (Duplicate) ของข้อมูลในการเพิ่มจำนวนชุดข้อมูลในคลาสส่วนน้อย เช่นหากทำการเพิ่มคลาสส่วนน้อย 5 ข้อมูลจะได้ผลตามตารางที่ 2.2

ตารางที่ 2.2 ชุดข้อมูลหลังจากใช้เทคนิคการเพิ่มตัวเองของข้อมูลจากคลาสส่วนน้อยเพิ่ม 5 ข้อมูล

A	B	C	D	Class
x	y	z	x	negative
y	z	x	x	negative
z	x	y	y	negative
x	y	z	y	negative
y	z	x	z	negative
z	x	y	z	negative
x	y	z	y	positive
y	z	x	y	negative
z	x	y	x	negative
x	y	z	x	negative
x	**y**	**z**	**y**	**positive**

ตารางที่ 2.2 ชุดข้อมูลหลังจากใช้เทคนิคการเพิ่มตัวเองของข้อมูลจากคลาสส่วนน้อยเพิ่ม 5 ข้อมูล (ต่อ)

A	B	C	D	Class
x	**y**	**z**	**y**	**positive**
x	**y**	**z**	**y**	**positive**
x	**y**	**z**	**y**	**positive**
x	**y**	**z**	**y**	**positive**

วิธีการสุ่มลด (Under Sampling)

วิธีสุ่มลด (Japkowicz, 2000; Japkowicz and Stephen, 2002 ;กิตติพงศ์, 2016) เป็นเทคนิคหรือวิธีที่ใช้ในการลดจำนวนข้อมูลที่อยู่ในคลาสส่วนมาก

จากตารางที่ 2.1 ชุดข้อมูล ซึ่งมีจำนวนคลาสทั้งหมด 2 คลาส ได้แก่คลาส positive และคลาส negative โดยข้อมูลในคลาส positive มีจำนวนทั้งหมด 1 ข้อมูล และข้อมูลในคลาส negative มีจำนวนทั้งหมด 9 ข้อมูล ซึ่งงานวิจัยนี้ใช้วิธีการสุ่มลดข้อมูลในการลดจำนวนชุดข้อมูลในคลาสส่วนมาก เช่นหากทำการลดคลาสส่วนมาก 3 ข้อมูลจะได้ผลตามตารางที่ 2.3

ตารางที่ 2.3 ชุดข้อมูลหลังจากใช้เทคนิคการสุ่มลดของข้อมูลจากคลาสส่วนมากลง 3 ข้อมูล

A	B	C	D	Class
x	y	z	x	negative
z	x	y	y	negative
x	y	z	y	negative
z	x	y	z	negative
x	y	z	y	positive
y	z	x	y	negative
x	y	z	x	negative

2.3 เทคนิคการจัดกลุ่มข้อมูล (Clustering)

การวิเคราะห์กลุ่ม (Cluster Analysis) เป็นเทคนิคการแบ่งกลุ่มหน่วยข้อมูล หรือเป็นการแบ่งคน สัตว์ สิ่งของ องค์กร ฯลฯ ออกเป็นกลุ่มย่อยอย่างน้อย 2 กลุ่ม โดยมีหลักเกณฑ์ในการแบ่งดังนี้ ตัวแปรที่อยู่ในกลุ่มเดียวกันจะมีความสัมพันธ์กันมากกว่าตัวแปรที่อยู่ต่างกลุ่มกัน ส่วนตัวแปรที่อยู่ต่างกลุ่มกันจะมีความสัมพันธ์กันน้อยหรือไม่มีความสัมพันธ์กันเลย

สำหรับประเภทของเทคนิคการจัดกลุ่มที่ใช้กันมากมี 2 เทคนิค คือ

1. การวิเคราะห์กลุ่มแบบขั้นตอน (Hierarchical Cluster Analysis)

2. การวิเคราะห์กลุ่มแบบไม่เป็นขั้นตอน (Nonhierarchical Cluster Analysis) ซึ่งงานวิจัยนี้เน้นโดยการนำวิธีจัดกลุ่มข้อมูลนี้ไปใช้ร่วมกับการสุ่มซ้ำข้อมูล

การจัดกลุ่มแบบ K-Means Clustering

เป็นวิธีการวิเคราะห์กลุ่มแบบไม่เป็นขั้นตอน (Nonhierarchical Cluster Analysis) หรือ การแบ่งส่วน (Partioning) เป็นอัลกอริทึมเทคนิคการเรียนรู้ที่นิยมมากที่สุดวิธีหนึ่ง โดยอัลกอริทึม K-Means จะตัดแบ่ง (Partition) วัตถุออกเป็น K กลุ่ม แล้วแทนค่าแต่ละกลุ่มด้วยค่าเฉลี่ยของกลุ่ม ซึ่งใช้เป็นจุดศูนย์กลาง (centroid) ของกลุ่มในการวัดระยะห่างของข้อมูลในกลุ่มเดียวกัน โดยมีขั้นตอนการจัดกลุ่มดังนี้

1. กำหนดหรือสุ่มค่าเริ่มต้น จำนวน k ข้อมูล เพื่อทำหน้าที่เป็นจุดศูนย์กลางเริ่มต้น k จุด เรียกว่า Cluster Center หรือ Centroid

2. จัดข้อมูลทั้งหมดเข้ากลุ่ม โดยทำการหาระยะห่างระหว่างข้อมูลกับจุดศูนย์กลางทั้ง k จุด หากข้อมูลใดใกล้จุดศูนย์กลางใดมากที่สุดให้อยู่กลุ่มนั้น

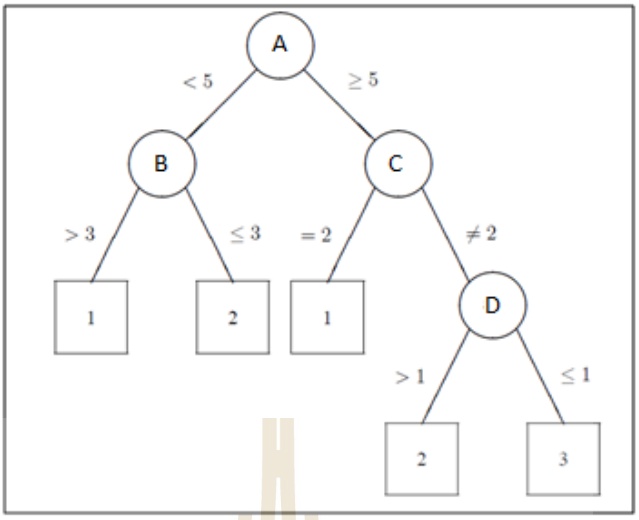
3. หาค่าเฉลี่ย (Mean) แต่ละกลุ่ม เพื่อใช้เป็นค่าจุดศูนย์กลางใหม่

4. ทำข้อ 2 และ 3 ซ้ำ จนกว่าค่าเฉลี่ย หรือจุดศูนย์กลางในแต่ละกลุ่มไม่มีการเปลี่ยนแปลง

2.4 ทฤษฎีของต้นไม้ตัดสินใจแบบจำแนกและแบบถดถอย (Classification and Regression Tree, CART)

เทคนิคต้นไม้ตัดสินใจแบบจำแนกและแบบถดถอย (Classification and Regression Trees, CART) คิดค้นโดย Brieman และคณะในปี ค.ศ.1984 โดยเน้นไปที่การแนะนำการสร้างแบบจำลองของต้นไม้ที่ใช้หลักสถิติ พร้อมทั้งทำการตรวจสอบเพื่อเลือกต้นไม้ที่ดีที่สุด ซึ่งผู้คิดค้นทั้ง 4 ท่านได้กล่าวถึงปรัชญาของพวกเขาไว้ว่า “ ปรัชญาของพวกเราในการวิเคราะห์ข้อมูล คือ การมองข้อมูลจากหลายมุมมองที่แตกต่างกัน ซึ่งต้นไม้ที่มีโครงสร้างแบบถดถอยเป็นอีกทางเลือกที่น่าสนใจสำหรับการนำมาแก้ปัญหาประเภทถดถอย ที่บางครั้งไม่สามารถแก้ไขปัญหาให้ชัดเจนได้ด้วยการวิเคราะห์การถดถอยเชิงเส้น แต่อย่างไรก็ตามเช่นเดียวกันกับเครื่องมือใด ๆ ผลประโยชน์ที่ได้จากวิธีการนี้ ขึ้นอยู่กับความเหมาะสมของเครื่องมือกับข้อมูล”

ต้นไม้ตัดสินใจแบบจำแนกและแบบถดถอย (Brieman et al., 1984) เป็นต้นไม้ตัดสินใจที่มีโครงสร้างแบบไบนารี แต่ละโหนดแสดงถึงกลุ่มย่อยของข้อมูลที่ถูกสร้างโดยการแยกโหนดเป็นสองโหนดลูกซ้ำแล้วซ้ำอีก การสร้างต้นไม้ตัดสินใจเริ่มต้นด้วยโหนดรากที่มีตัวอย่างการเรียนรู้ทั้งหมด โดยโมเดลทั่วไปมีลักษณะดังรูปที่ 2.1



รูปที่ 2.1 แสดงลักษณะของแบบจำลองที่ได้จากอัลกอริทึม
ต้นไม้ตัดสินใจแบบจำแนกและแบบถดถอย

จากรูปที่ 2.1 สามารถอธิบายได้ว่า หากข้อมูลที่กำลังพิจารณาแอททริบิวต์ A มีค่าน้อยกว่า 5 ข้อมูลจะถูกแยกไปยังโหนด B หรือหากมีค่ามากกว่าหรือเท่ากับ 5 ข้อมูลจะถูกแยกไปยังโหนด C ในกรณีที่ข้อมูลที่กำลังพิจาณามีค่าแอททริบิวต์ A มีค่าน้อยกว่า 5 จะถูกพิจารณาที่แอททริบิวต์ B ต่อ ซึ่งหากค่าแอททริบิวต์ B มีค่ามากกว่า 3 ค่าตัวแปรเป้าหมายของข้อมูลนี้จะมีค่าเท่ากับ 1 หรือหากค่าแอททริบิวต์ B น้อยกว่าหรือเท่ากับ 3 ค่าตัวแปรเป้าหมายของข้อมูลนี้จะมีค่าเท่ากับ 2 เป็นต้น

กระบวนการทำงานของต้นไม้ตัดสินใจแบบแจกแจงและแบบถดถอยสามารถเขียนรหัสเทียมได้ดังรูปที่ 2.2

1. คำนวณค่าดัชนีจีนีของทุกแอททริบิวต์
2. เลือกแอททริบิวต์ที่มีค่าดัชนีจีนีน้อยที่สุดมาเป็นโหนดที่ต้องการพิจารณา จากนั้นแยกข้อมูลออกเป็นกลุ่มตามค่าตัวแปร
3. ทำข้อ 1 และ 2 ซ้ำ จนกว่าค่าดัชนีจีนีไม่มีการเปลี่ยนแปลง

รูปที่ 2.2 รหัสเทียมของต้นไม้ตัดสินใจแบบจำแนกและแบบถดถอย

ค่าดัชนีจีนี (Gini Index)

เป็นค่าดัชนีที่อัลกอริทึมต้นไม้ตัดสินใจแบบจำแนกและแบบถดถอยใช้ในการตัดสินใจแตกโหนดของต้นไม้แบบจำลอง โดยจะเลือกแตกโหนดไปในทิศทางแอททริบิวต์ที่มีค่าดัชนีต่ำที่สุด ยิ่งค่าดัชนีต่ำเท่าไร แปลว่าโหนดนั้นจำแนกข้อมูลได้บริสุทธิ์มากเท่านั้น สำหรับค่าดัชนีจีนีสามารถคำนวณได้ตามสมการที่ 2.2

$$\text{Gini} = 1 - \sum (P_i)^2 \tag{2.2}$$

เมื่อ P_i คือ ความน่าจะเป็นของแต่ละคลาส

ตัวอย่างในการสร้างต้นไม้ตัดสินใจแบบจำแนกและแบบถดถอย

การอธิบายขั้นตอนการสร้างต้นไม้ตัดสินใจแบบจำแนกและแบบถดถอยจะใช้ข้อมูลเกี่ยวกับบุคคล 10 ท่าน เกี่ยวกับโรคอ้วน (Obesity) โดยจำแนกจากแอททริบิวต์ค่าดัชนีมวลกาย (BMI) ซึ่งแบ่งออกเป็น 3 ระดับ น้ำหนักต่ำกว่าเกณฑ์ (Underweight) สมส่วน (Healthy) และน้ำหนักเกินเกณฑ์ (Overweight) แอททริบิวต์การนอนหลับประจำวัน (Sleep) นอนหลับเพียงพอ 6 ถึง 8 ช.ม.หรือไม่ แอททริบิวต์รับประทานของหวานหรือไม่ (Desserts) และแอททริบิวต์ออกกำลังกายหรือไม่ (Exercise) ตัวอย่างตามตารางที่ 2.4 โดยตัวแปรเป้าหมายคือแอททริบิวต์ Obesity

ตารางที่ 2.4 ข้อมูลตัวอย่างในการสร้างต้นไม้ตัดสินใจแบบจำแนกและแบบถดถอย

Person	BMI	Sleep	Desserts	Exercise	Obesity
1	Overweight	No	Yes	No	Yes
2	Healthy	Yes	No	Yes	No
3	Healthy	Yes	No	No	No
4	Underweight	Yes	Yes	Yes	No
5	Overweight	Yes	No	Yes	No
6	Overweight	No	Yes	No	Yes
7	Underweight	Yes	Yes	Yes	No
8	Healthy	Yes	No	Yes	No
9	Underweight	No	Yes	No	No
10	Overweight	Yes	Yes	No	Yes

จากข้อมูลตัวอย่างในตารางที่ 2.4 เริ่มต้นโดยคำนวณค่าดัชนีจีนีของทุกแอททริบิวต์ เมื่อพิจารณาจากแอททริบิวต์ BMI จะคำนวณดัชนีจีนีดังนี้

Gini(BMI=Underweight) $= 1 - (0/3)^2 - (3/3)^2 = 1 - 0 - 1 = 0$

Gini(BMI=Healthy) $= 1 - (0/3)^2 - (3/3)^2 = 0$

Gini(BMI=Overweight) $= 1 - (3/4)^2 - (1/4)^2 = 1 - 0.5625 - 0.0625 = 0.375$

ดังนั้นดันชีจีนีรวมสำหรับแอททริบิวต์ BMI คือ

Gini(BMI) $= (3/10) \times 0 + (3/10) \times 0 + (4/10) \times 0.375 = 0.15$

เมื่อพิจารณาจากแอททริบิวต์ Sleep จะคำนวณดันชีจีนีได้ดังนี้

Gini(Sleep=Yes) $= 1 - (1/7)^2 - (6/7)^2 = 1 - 0.02 - 0.734 = 0.237$

Gini(Sleep=No) $= 1 - (2/3)^2 - (1/3)^2 = 1 - 0.44 - 0.11 = 0.45$

ดังนั้นดัชนีจีนีรวมสำหรับแอททริบิวต์ Sleep คือ

Gini(Sleep) $= (7/10) \times 0.237 + (3/10) \times 0.45$

$= 0.165 + 0.135 = 0.3$

เมื่อพิจารณาจากแอททริบิวต์ Desserts จะคำนวณดันชีจีนีได้ดังนี้

Gini(Desserts=Yes) $= 1 - (3/6)^2 - (3/6)^2 = 1 - 0.25 - 0.25 = 0.5$

Gini(Desserts=No) $= 1 - (0/4)^2 - (4/4)^2 = 1 - 0 - 1 = 0$

ดังนั้นดัชนีจีนีรวมสำหรับแอททริบิวต์ Desserts คือ

Gini(Desserts) $= (6/10) \times 0.5 + (4/10) \times 0 = 0.3$

เมื่อพิจารณาจากแอททริบิวต์ Exercise จะคำนวณดันชีจีนีได้ดังนี้

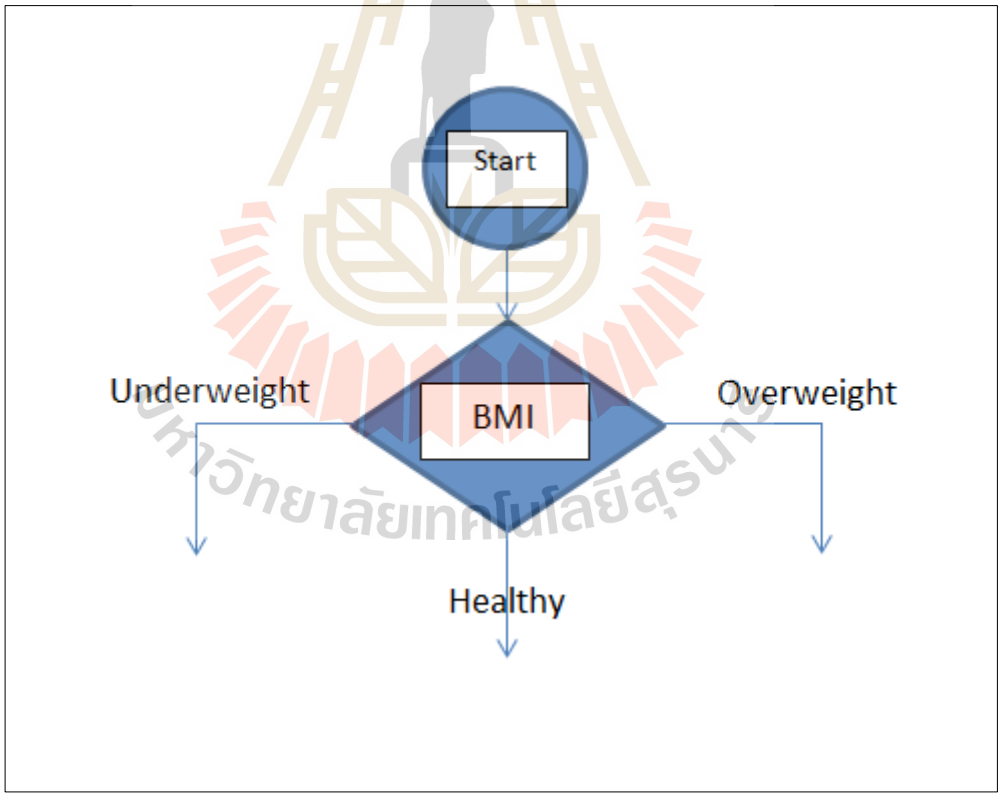
Gini(Exercise=Yes) $= 1 - (0/5)^2 - (5/5)^2 = 1 - 0 - 1 = 0.375$

Gini(Exercise=No) $= 1 - (2/5)^2 - (3/5)^2 = 1 - 0.16 - 0.36 = 0.48$

ดังนั้นดัชนีจีนีรวมสำหรับแอททริบิวต์ Exercise คือ

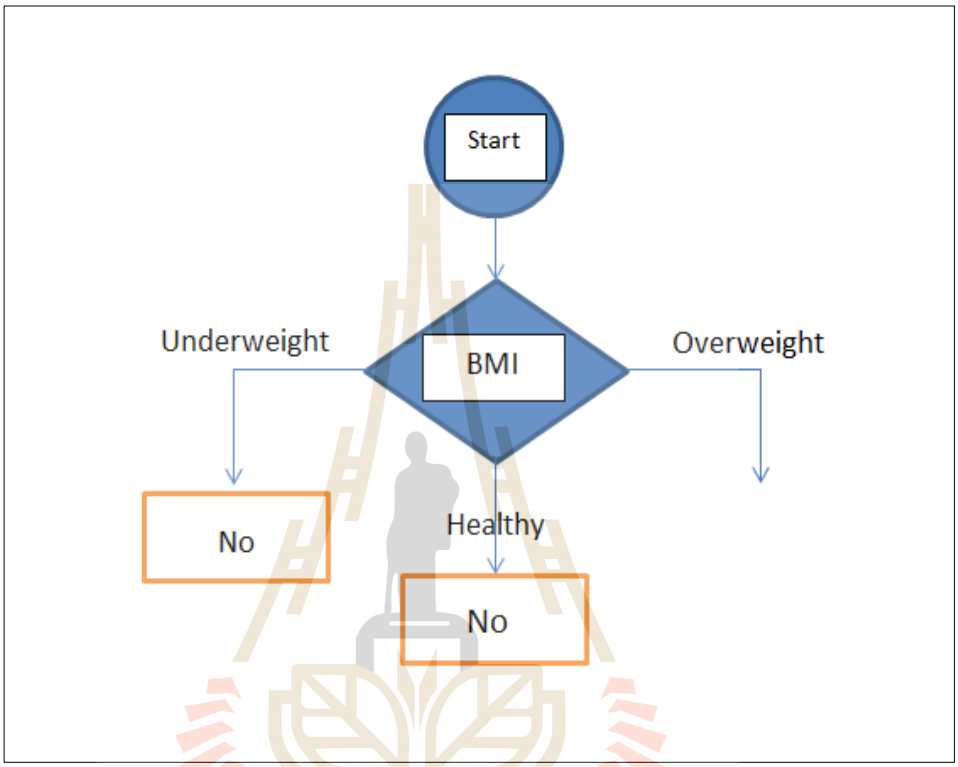
Gini(Exercise) $= (5/10) \times 0 + (5/10) \times 0.48 = 0.24$

จากการคำนวณดัชนีจีนีครบทุกแอทริบิวต์ สรุปได้ว่าแอททริบิวต์ BMI มีค่าดัชนีจีนีน้อยที่สุด จึงเริ่มสร้างต้นไม้ตัดสินใจได้เป็นไปตามรูปที่ 2.3



รูปที่ 2.3 ต้นไม้ตัดสินใจแบบจำแนกและแบบถดถอยที่ได้จากการคำนวณดัชนีจีนีของข้อมูลตัวอย่าง (1)

จากรูปที่ 2.3 ในส่วนของข้อมูลที่แอททริบิวต์ BMI = Underweight และ BMI = Healthy ค่าตัวแปรที่ถูกจำแนกออกมานั้นเป็น No ทั้งหมด โหนดใบนี้จึงไม่จำเป็นต้องทำการวนคำนวณดัชนีจีนีซ้ำอีก จึงได้ต้นไม้ตัดสินใจเป็นไปตามรูปที่ 2.4



รูปที่ 2.4 ต้นไม้ตัดสินใจแบบจำแนกและแบบถดถอยที่ได้จากการคำนวณดัชนีจีนีของข้อมูลตัวอย่าง (2)

ลำดับถัดมาเมื่อพิจารณาข้อมูลของฝั่งที่แอททริบิวต์ BMI = Overweight เหลือข้อมูลที่ต้องพิจารณาดังตารางที่ 2.5

ตารางที่ 2.5 ข้อมูลตัวอย่างที่แอททริบิวต์ BMI = Overweight

Person	BMI	Sleep	Desserts	Exercise	Decision
1	Overweight	No	Yes	No	Yes
6	Overweight	Yes	No	Yes	No
7	Overweight	No	Yes	No	Yes
10	Overweight	Yes	Yes	No	Yes

จากตารางที่ 2.5 สามารถคำนวณดัชนีจีนีที่แอททริบิวต์ต่าง ๆ ได้ดังนี้

Gini(BMI=Overweight and Sleep=Yes) $= 1 - (1/2)^2 - (1/2)^2 = 1 - 0.25 - 0.25$

$= 0.5$

Gini(BMI=Overweight and Sleep=No) $= 1 - (2/2)^2 - (0/2)^2 = 0 - 1 - 0$

$= 0$

Gini(BMI=Overweight and Sleep) $= (2/4)x0.5 + (2/4)x0 = 0.25$

Gini(BMI=Overweight and Desserts=Yes) $= 1 - (3/3)^2 - (0/3)^2 = 1 - 1 - 0 = 0$

Gini(BMI=Overweight and Desserts=No) $= 1 - (0/1)^2 - (1/1)^2 = 1 - 0 - 1 = 0$

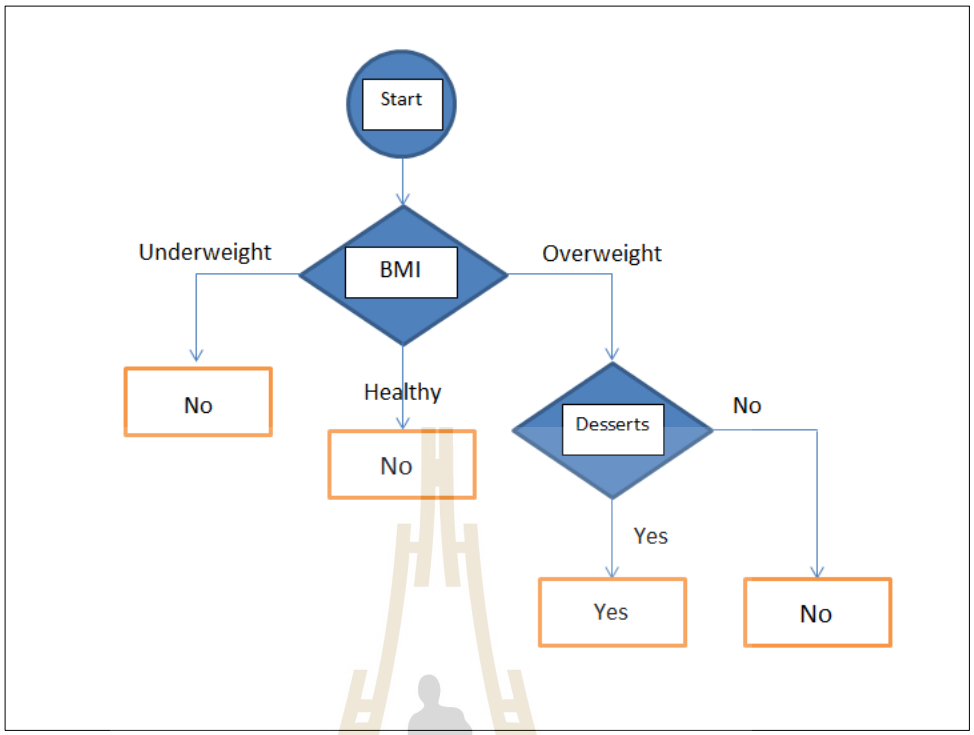
Gini(BMI=Overweight and Desserts) $= (3/4)x0 + (1/4)x0 = 0$

Gini(BMI=Overweight and Exercise=Yes) $= 1 - (0/1)^2 - (1/1)^2 = 1 - 0 - 1 = 0$

Gini(BMI=Overweight and Exercise=No) $= 1 - (3/3)^2 - (0/3)^2 = 1 - 1 - 0 = 0$

Gini(BMI=Overweight and Exercise) $= (1/4)x0 + (3/4)x0 = 0$

จากการคำนวณดัชนีจีนีทางฝั่งข้อมูลที่แอททริบิวต์ BMI = Overweight ครบทุก แอทริบิวต์ สรุปได้ว่าแอททริบิวต์ Desserts และ Exercise มีค่าดัชนีจีนีน้อยที่สุด โดยเราเลือก แอททริบิวต์ Desserts มาเป็นโหนดลำดับถัดไป เนื่องจากเป็นแอททริบิวต์ลำดับก่อนแอททริบิวต์ Exercise เมื่อพิจารณาข้อมูลที่แอททริบิวต์ Desserts = Yes ค่าตัวแปรที่ถูกจำแนกออกมานั้นเป็น Yes ทั้งหมด และเมื่อพิจารณาข้อมูลที่แอททริบิวต์ Desserts = No ค่าตัวแปรที่ถูกจำแนกออกมานั้นเป็น No ทั้งหมด โหนดใบนี้จึงไม่จำเป็นต้องทำการวนคำนวณดัชนีจีนีซ้ำอีก จึงได้ต้นไม้ตัดสินใจเป็นไปตามรูปที่ 2.5



รูปที่ 2.5 ต้นไม้ตัดสินใจแบบจำแนกและแบบถดถอยที่ได้จากการคำนวณดัชนีจีนีของข้อมูลตัวอย่าง (3)

2.5 การประเมินประสิทธิภาพของการจำแนกข้อมูล (Classification Evaluation)

สำหรับการประเมินประสิทธิภาพการจำแนกข้อมูลในการวิจัยนี้เนื่องด้วยข้อมูลที่สนใจเป็นข้อมูลไม่สมดุล หากใช้วิธีการประเมินประสิทธิภาพด้วยมาตรวัดพื้นฐาน เช่น การใช้ค่าความแม่นยำในการจำแนก (Accuracy) อาจจะไม่เพียงพอต่อการประเมินประสิทธิภาพที่แท้จริงของแบบจำลอง เนื่องจากข้อมูลไม่สมดุลนั้นมีจำนวนข้อมูลในแต่ละคลาสต่างกันเป็นจำนวนมาก หากจำแนกข้อมูลใหม่เป็นคลาสส่วนมากทั้งหมดก็จะทำให้ค่าความแม่นยำในการจำแนกสูงได้เช่นกัน ตัวอย่างเช่น ชุดข้อมูลหนึ่งมีจำนวนข้อมูล 1,000 ข้อมูล แบ่งเป็นคลาสส่วนมากจำนวน 950 ข้อมูล และคลาสส่วนน้อยจำนวน 50 ข้อมูล หากแบบจำลองจำแนกว่าข้อมูลทั้ง 1,000 ข้อมูลอยู่ในคลาสส่วนมาก แบบจำลองนี้จะมีค่าความแม่นยำในการจำแนกเท่ากับ 95% แม้จะไม่สามารถจำแนกข้อมูลในคลาสส่วนน้อยได้แม้แต่ข้อมูลเดียว ดังนั้นการจำแนกข้อมูลไม่สมดุลจึงจำเป็นต้องมีการประเมินประสิทธิภาพแบบจำลองด้วยมาตรวัดอื่น ๆ เพิ่มเติม

เมตริกซ์วัดประสิทธิภาพ (Confusion Matrix) คือเมตริกซ์ที่ใช้แสดงผลการจำแนกข้อมูลจากการทดสอบด้วยชุดข้อมูลทดสอบออกเป็นแต่ละคลาส เพื่อการประเมินประสิทธิภาพการทำนายคลาสส่วนน้อยเป็นหลัก โดยมีรูปแบบแสดงดังตารางที่ 2.6

ตารางที่ 2.6 Confusion Matrix

	Prediction	
Actual	Positive	Negative
Positive	TP	FN
Negative	FP	TN

จากตารางที่ 2.6 เมื่อแทนคลาสส่วนมากด้วยค่า Positive และคลาสส่วนน้อยด้วยค่า Negative แถวของเมตริกซ์จะแสดงจำนวนข้อมูลจริงของแต่ละคลาส และคอลัมน์ของเมตริกซ์จะแสดงจำนวนที่ทำนายได้ของแต่ละคลาส แบ่งผลการทำนายออกเป็น 4 กรณี ดังนี้

กรณีที่ 1 : True Positive (TP) หมายถึง จำนวนข้อมูลที่อยู่ในคลาสส่วนมาก และแบบจำลองจำแนกได้ถูกต้องว่าข้อมูลนั้นอยู่ในคลาสส่วนมาก

กรณีที่ 2 : True Negative (TN) หมายถึง จำนวนข้อมูลที่อยู่ในคลาสส่วนน้อย และแบบจำลองจำแนกได้ถูกต้องว่าข้อมูลนั้นอยู่ในคลาสส่วนน้อย

กรณีที่ 3 : False Positive (FP) หมายถึง จำนวนข้อมูลที่อยู่ในคลาสส่วนมาก แต่แบบจำลองจำแนกผิดพลาด โดยทำนายว่าข้อมูลนั้นอยู่ในคลาสส่วนน้อย

กรณีที่ 4 : False Negative (FN) หมายถึง จำนวนข้อมูลที่อยู่ในคลาสส่วนน้อย แต่แบบจำลองจำแนกผิดพลาด โดยทำนายว่าข้อมูลนั้นอยู่ในคลาสส่วนมาก

และจากกรณีต่าง ๆ ตามตาราง Confusion Matrix สามารถประเมินประสิทธิภาพแบบจำลองตามการคำนวณมาตรวัดทั้ง 4 ได้ดังต่อไปนี้

ค่าความแม่นยำ (Accuracy)

เป็นมาตรวัดความแม่นยำในการจำแนกข้อมูล เป็นการประเมินประสิทธิภาพการจำแนกโดยรวมของทุกคลาสของแบบจำลอง ดังสมการที่ 2.3

$$\text{Accuracy} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{FP} + \text{TN} + \text{FN}} \tag{2.3}$$

ค่าความเที่ยง (Precision)

เป็นมาตรวัดความแม่นยำในการจำแนกข้อมูลในคลาสส่วนน้อย โดยคำนวณจากจำนวนข้อมูลที่ถูกจำแนกเป็นคลาสส่วนน้อยได้ถูกต้องเทียบกับจำนวนข้อมูลที่ถูกทำนายเป็นคลาสส่วนน้อยทั้งหมด ดังสมการที่ 2.4

$$\text{Precision} = \frac{\text{TN}}{\text{TN} + \text{FN}} \tag{2.4}$$

ค่าระลึก หรือค่าความไว (Recall / Sensitivity)

เป็นมาตรวัดความแม่นยำในการจำแนกคลาสส่วนน้อยว่าสามารถจำแนกได้แม่นยำเพียงใด โดยคำนวณจากจำนวนข้อมูลที่ถูกจำแนกเป็นคลาสส่วนน้อยได้ถูกต้อง เทียบกับจำนวนข้อมูลจริงของคลาสส่วนน้อยทั้งหมด ดังสมการที่ 2.3.3

$$\text{Recall} \mid \text{Sensitivity} = \frac{\text{TN}}{\text{TN} + \text{FP}} \tag{2.5}$$

ค่าความจำเพาะ (Specificity)

เป็นมาตรวัดความแม่นยำในการจำแนกข้อมูลในคลาสส่วนมาก โดยคำนวณจากจำนวนข้อมูลที่ถูกจำแนกเป็นคลาสส่วนมากได้ถูกต้องเทียบกับจำนวนข้อมูลจริงของคลาสส่วนมากทั้งหมด ดังสมการที่ 2.6

$$\text{Specificity} = \frac{\text{TP}}{\text{TP} + \text{FN}} \tag{2.6}$$

ค่าการวัดเอฟ (F-Measure)

เป็นมาตรวัดความแม่นยำของการจำแนกคลาสส่วนน้อย โดยดูจากผลเฉลี่ยของ Precision และ Recall ดังสมการที่ 2.7

$$\text{F} - \text{Measure} = \frac{2 \text{ x Precision x Recall}}{\text{Precision} + \text{Recall}} \tag{2.7}$$

ตัวอย่างการคำนวณมาตรวัดประสิทธิภาพแบบจำลอง

สมมติให้การจำแนกข้อมูลไม่สมดุลที่มี 2 คลาส โดยมีข้อมูลทั้งหมด 100 ข้อมูล แบ่งเป็นข้อมูลในคลาสส่วนมากจำนวน 95 ข้อมูล และจำนวนข้อมูลในคลาสส่วนน้อย 5 ข้อมูล โดยมีผลลัพธ์การจำแนกแสดงดังตารางที่ 2.7

ตารางที่ 2.7 ผลลัพธ์การจำแนกข้อมูลไม่สมดุล

	Prediction	
Actual	Positive	Negative
Positive	85	1
Negative	4	10

จะได้ว่าแบบจำลองการจำแนกนี้ มีประสิทธิภาพในการจำแนกด้วยมาตรวัดต่าง ๆ ดังนี้

ค่า Accuracy = (85 + 10) / (100) = 0.95

ค่า Precision = (10) / (10 + 1) = 0.90

ค่า Recall = (10) / (10+4) = 0.71

ค่า Specificity = (85) / (85 + 1) = 0.98

ค่า F-measure = (2*0.90*0.71) / (0.90 + 0.71) = 0.79

โดยสามารถสรุปได้ว่า แบบจำลองนี้มีความสามารถในการจำแนกทั้งสองคลาสอยู่ที่ 95% สามารถจำแนกข้อมูลที่อยู่ในคลาสส่วนน้อยได้ถูกต้องเทียบกับการจำแนกข้อมูลว่าเป็นคลาสส่วนน้อยทั้งหมดอยู่ที่ 90% สามารถจำแนกข้อมูลของคลาสส่วนน้อยทั้งหมดได้แม่นยำอยู่ที่ 71% และสามารถจำแนกข้อมูลว่าเป็นคลาสส่วนมากได้ถูกต้องทั้งหมดได้แม่นยำอยู่ที่ 98% โดยรวมแล้วโมเดลนี้สามารถจำแนกเฉพาะข้อมูลคลาสส่วนน้อยมีความแม่นยำอยู่ที่ 79%

2.6 งานวิจัยที่เกี่ยวข้อง

งานวิจัยที่เกี่ยวข้องกับอัลกอริทึมต้นไม้ตัดสินใจแบบจำแนกและแบบถดถอย (CART) มีดังนี้

Grubinger et al. (2010) เสนอการสร้างต้นไม้ตัดสินใจแบบจำแนกและแบบถดถอย (CART) โดยนำวิธี Bootstrap เข้ามาประยุกต์ เพื่อหาโมเดลที่จะนำไปใช้กับข้อมูลสุขภาพของชาวออสเตรีย ซึ่งการทดลองแบ่งออกเป็น 2 ส่วน คือ การเลือกโมเดลจากสถิติทางการแพทย์ โดยเลือก

โมเดลจากความถูกต้องและแม่นยำ ผลจากการวิจัยสรุปได้ว่า เมื่อมีการนำ Bootstrap เข้ามาช่วยอัลกอริทึมต้นไม้ตัดสินใจแบบจำแนกและแบบถดถอยมีความแม่นยำเพิ่มขึ้น

Lemon et al. (2003) เสนอการนำอัลกอริทึมต้นไม้ตัดสินใจแบบจำแนกและแบบถดถอย (CART) ไปใช้ในงานสาธารณสุข โดยเปรียบเทียบกับวิธีการวิเคราะห์แบบถดถอยโลจิสติค ซึ่งอัลกอริทึมต้นไม้ตัดสินใจแบบจำแนกและแบบถดถอย (CART) ให้ค่าแม่นยำที่สูงกว่า แม้ไม่มากก็ตาม โดยอัลกอริทึมต้นไม้ตัดสินใจแบบจำแนกและแบบถดถอยเหมาะสมกับการจัดข้อมูลเป็นกลุ่มด้วยตัวแปรใดเพียงตัวเดียว

ดิษฐพล มั่นธรรม และ ลี่ลี อิงศรีสว่าง (2010) เสนอการประยุกต์ใช้วิธีอัลกอริทึมต้นไม้ตัดสินใจแบบจำแนกและแบบถดถอย (CART) กับการวินิจฉัยโรคระบบการหายใจ สามารถจำแนกผู้ป่วยออกเป็นกลุ่มผู้ป่วยที่ติดเชื้อทางหายใจส่วนบนแบบเฉียบพลัน และโรคปอดอักเสบ จากการใช้เพียง 7 ลักษณะอาการได้ประสิทธิภาพสูงถึงร้อยละ 92 และ 95 ตามลำดับ และในทำนองเดียวกันสามารถจำแนกผู้ป่วยโรคโพรงจมูกอักเสบเฉียบพลัน จากการใช้ 8 ลักษณะอาการ ได้แม่นยำถึงร้อยละ 95

Arimai-Dieguez et al. (2015) ได้เสนอการวิจัยที่มีการประยุกต์ใช้อัลกอริทึมต้นไม้ตัดสินใจแบบจำแนกและแบบถดถอย (CART) เปรียบเทียบกับการถดถอยแบบโลจิสติค ในการวิเคราะห์ข้อมูลทางด้านการเงิน และเศรษฐกิจ ซึ่งผลปรากฏออกมาว่าอัลกอริทึมต้นไม้ตัดสินใจแบบจำแนกและแบบถดถอยสามารถจำแนกข้อมูลได้ชัดเจนมากกว่า

Zhao (2016) ได้เสนอการวิจัยว่าด้วยการวิเคราะห์ความสัมพันธ์ระหว่างปริมาณน้ำบาดาลในรูปแบบที่ทำให้เกิดดินถล่ม กับข้อมูลอื่น ๆ เช่น การเปลี่ยนแปลงของปริมาณน้ำฝน เป็นต้น เพื่อทำการคาดการณ์ปัจจัยที่มีอิทธิพลต่อเหตุการณ์ดังกล่าว ซึ่งการวิจัยนี้เน้นไปที่โมเดลที่ได้จากอัลกอริทึมต้นไม้ตัดสินใจแบบจำแนกและแบบถดถอย (CART) ซึ่งเมื่อนำมาเปรียบเทียบกับโมเดลที่ได้จาก PSO-SVR โมเดลที่ได้จากอัลกอริทึมต้นไม้ตัดสินใจแบบจำแนกและแบบถดถอย สามารถทำนายได้แม่นยำ รวมไปถึงแสดงข้อมูลความสัมพันธ์ระหว่างตัวแปรที่ตัดความซ้ำซ้อนออกไปได้ดีกว่า

จากการศึกษางานวิจัยที่เกี่ยวข้องนั้นพบว่ามีหลากหลายสาขาที่นำอัลกอริทึมต้นไม้ตัดสินใจแบบจำแนกและแบบถดถอยไปใช้ในการวิจัย และได้รับผลลัพธ์ที่ดี จึงเป็นแรงบันดาลใจให้ผู้วิจัยเลือกใช้อัลกอริทึมนี้จากจุดเด่น และความยืดหยุ่นต่อข้อมูล พร้อมทั้งต้องการที่จะเพิ่มประสิทธิภาพด้วยการนำเทคนิคการสุ่มซ้ำเข้ามาช่วยเพื่อลดปัญหาในการจำแนกข้อมูลที่ไม่สมดุล ซึ่งงานวิจัยที่ผ่านมาไม่ได้มุ่งเน้น แต่เป็นปัญหาสำคัญสำหรับการจำแนกข้อมูล

จากงานวิจัยข้างต้น สามารถสรุปได้ดังตารางที่ 2.8

ตารางที่ 2.8 ตารางสรุปเปรียบเทียบงานวิจัยที่เกี่ยวข้อง

กระบวนการทำงาน	งานวิจัยที่เกี่ยวข้อง*					
	ก	ข	ค	ง	จ	ฉ
จุดประสงค์การวิจัย						
แสดงประสิทธิภาพในการจำแนกข้อมูล	✓	✓	✓	✓	✓	✓
เปรียบเทียบความแม่นยำกับอัลกอริทึมอื่น						
• CART	✓	✓	✓	✓	✓	✓
• Logistic Regression		✓		✓		
• PSO-SVR					✓	
เทคนิคที่นำมาช่วยในการเพิ่มประสิทธิภาพความแม่นยำของอัลกอริทึม						
• Resampling						✓
• Bootstrap	✓					
• Clustering						✓

*งานวิจัยที่เกี่ยวข้องประกอบด้วย

ก แทนงานวิจัยของ Grubinger et al. (2010)

ข แทนงานวิจัยของ Lemon et al. (2003)

ค แทนงานวิจัยของ ดิษฐพล มั่นธรรม และ ลี่ลี อิงศรีสว่าง (2010)

ง แทนงานวิจัยของ Irimia-Dieguez et al. (2015)

จ แทนงานวิจัยของ Zhao. et al. (2016)

ฉ แทนงานวิจัยของวิทยานิพนธ์ฉบับนี้

บทที่ 3
วิธีดำเนินการวิจัย

งานวิจัยนี้มีวัตถุประสงค์ที่จะพัฒนาประสิทธิภาพของแบบจำลองต้นไม้ตัดสินใจแบบจำแนกและแบบถดถอย ให้มีประสิทธิภาพในการจำแนกคลาสส่วนน้อยได้ดียิ่งขึ้น ด้วยการนำเทคนิคในการจัดการข้อมูลไม่สมดุลมาช่วยในขั้นตอนการจัดเตรียมข้อมูล ได้แก่เทคนิคการสุ่มซ้ำและเทคนิคการจัดกลุ่มข้อมูล โดยข้อมูลของงานวิจัยจะใช้ข้อมูลประเภทตัวเลข และมีตัวแปรเป้าหมายจำนวน 2 คลาส ซึ่งในบทนี้จะกล่าวถึง ข้อมูลที่ใช้ในการวิจัย เครื่องมือที่ใช้ในการวิจัย กรอบแนวคิดของการวิจัย และวิธีดำเนินงานวิจัย พร้อมขั้นตอนต่าง ๆ ในงานวิจัย รายละเอียดดังนี้

3.1 ข้อมูลที่ใช้ในการวิจัย

สำหรับข้อมูลที่ใช้ในการวิจัยเป็นข้อมูลไม่สมดุลจำนวน 5 ชุดข้อมูล ได้แก่ ข้อมูลโรคเบาหวานของอินเดียนแดงเผ่าพิม่าในประเทศสหรัฐอเมริกา (Pima) ข้อมูลตำแหน่งโปรตีนในเซลล์ร่างของแบคทีเรียแกรมลบ (Yeast) ข้อมูลยานพาหนะ (Vehicel) ข้อมูลการแบ่งส่วนภาพ (Segment) และข้อมูลการแบ่งบล็อกของโครงร่างหน้าเอกสาร (Page-Block) ซึ่งได้รับจากฐานข้อมูล KEEL (KEEL-dataset Repository) โดยรายละเอียดของแต่ละชุดข้อมูลแสดงดังตารางที่ 3.1

ตารางที่ 3.1 รายละเอียดชุดข้อมูลที่นำมาใช้ในงานวิจัย

ชุดข้อมูล	แอททริบิวต์	จำนวนตัวอย่าง			ระดับความไม่สมดุล
		คลาสส่วนมาก	คลาสส่วนน้อย	รวม	
Pima	9	500	268	768	1.87
Yeast	9	1055	429	1484	2.46
Vehicle	19	647	199	846	3.25
Segment	20	1979	329	2308	6.02
Page-Block	11	4913	559	5472	8.79

จากตารางที่ 3.1 สามารถอธิบายรายละเอียดของแต่ละชุดข้อมูลได้ดังนี้ สำหรับชุดข้อมูล Pima ประกอบไปด้วย 9 คอลัมน์ (8 คอลัมน์สำหรับใช้ในการจำแนก และ 1 คอลัมน์สำหรับคลาส) โดยมีจำนวนข้อมูลทั้งหมด 768 ข้อมูล มีจำนวนข้อมูลในคลาสส่วนมากทั้งหมด 500 ข้อมูล มีจำนวนข้อมูลในคลาสส่วนน้อย 268 ข้อมูล และมีระดับความไม่สมดุลอยู่ที่ 1.87 มีรายละเอียดแต่ละคอลัมน์แสดงดังตารางที่ 3.2

ตารางที่ 3.2 ข้อมูล Pima

ลำดับ	แอททริบิวต์	คำอธิบาย	ค่า
1	Preg	จำนวนครั้งที่ตั้งครรภ์	0-17
2	Plas	ความเข้มข้นของกลูโคสในพลาสมาช่วงเวลา 2 ชั่วโมงจากการทดสอบน้ำตาลในช่องปาก	0-199
3	Pres	ความดันโลหิต (มิลลิเมตรปรอท)	0-122
4	Skin	ความหนาผิวพับบริเวณไตรเซป (มิลลิเมตร)	0-99
5	Insu	ปริมาณอินซูลินช่วงเวลา 2 ชั่วโมง	0-846
6	Mass	ค่าดัชนีมวลกาย (Body Mass Index, BMI)	0.0-67.1
7	Pedi	ค่าบ่งชี้เบาหวานทางสายเลือด	0.078-2.42
8	Age	อายุ (ปี)	21-81
9	Class	ตัวแปร	Positive, Negative

สำหรับชุดข้อมูล Yeast ประกอบไปด้วย 9 คอลัมน์ (8 คอลัมน์สำหรับใช้ในการจำแนก และ 1 คอลัมน์สำหรับคลาส) โดยมีจำนวนข้อมูลในคลาสส่วนมากทั้งหมด 1055 ข้อมูล มีข้อมูลในคลาสส่วนน้อยทั้งหมด 429 ข้อมูล รวมทั้งหมด 1484 ข้อมูล และมีระดับความไม่สมดุลอยู่ที่ 2.46 มีรายละเอียดแต่ละคอลัมน์แสดงดังตารางที่ 3.3

ตารางที่ 3.3 ข้อมูล Yeast

ลำดับ	แอททริบิวต์	คำอธิบาย	ค่า
1	Mcg	ค่าการจำแนกลำดับสัญญาณด้วยวิธีของ McGeoch	0.11-1.0
2	Gvh	ค่าการจำแนกลำดับสัญญาณด้วยวิธีของ Heijne	0.13-1.0

ตารางที่ 3.3 ข้อมูล Yeast (ต่อ)

ลำดับ	แอททริบิวต์	คำอธิบาย	ค่า
3	Alm	ค่าการทำนายพื้นที่ครอบคลุมของเมมเบรน จากโปรแกรม ALOM	0.21-1.0
4	Mit	ค่าการวิเคราะห์ปริมาณกรดอะมิโนบนโปรตีนส่วน N-terminal ของไมโทคอนเดรีย	0.0-1.0
5	Erl	ค่าสตริงย่อยของ HDEL	0.5, 1.0
6	Pox	สัญญาณการกำหนดเป้าหมายเปอร์รือกซิโซมบน C-terminus	0.0, 0.5, 0.83
7	Vac	ค่าการวิเคราะห์ปริมาณกรดอะมิโนของโปรตีนแวคิวโอล่า	0.0-0.73
8	Nuc	ค่าการวิเคราะห์พื้นที่สัญญาณของโปรตีนนิวเคลียส	0.0-1.0
9	Class	ตัวแปร	Positive, Negative

สำหรับชุดข้อมูล Vehicle ประกอบไปด้วย 19 คอลัมน์ (18 คอลัมน์สำหรับใช้ในการจำแนก และ 1 คอลัมน์สำหรับคลาส) โดยมีจำนวนข้อมูลในคลาสส่วนมากทั้งหมด 647 ข้อมูล มีข้อมูลในคลาสส่วนน้อยทั้งหมด 199 ข้อมูล รวมทั้งหมด 846 ข้อมูล และมีระดับความไม่สมดุลอยู่ที่ 3.25 มีรายละเอียดแต่ละคอลัมน์แสดงดังตารางที่ 3.4

ตารางที่ 3.4 ข้อมูล Vehicle

ลำดับ	แอททริบิวต์	คำอธิบาย	ค่า
1	Compactness	ความกระชับกระทัดรัด	73-119
2	Circularity	วัฏจักร	33-59
3	Distance_circularity	ระยะห่างวัฏจักร	40-112
4	Radius_ratio	อัตราส่วนรัศมี	104-333
5	Praxis_aspect_ratio	อัตราส่วนภาพแกน Pr	47-138
6	Max_length_aspect_ratio	ความยาวสูงสุดของอัตราส่วนภาพ	2-55
7	Scatter_ratio	อัตราส่วนความกระจัดกระจาย	112-265

ตารางที่ 3.4 ข้อมูล Vehicle (ต่อ)

ลำดับ	แอททริบิวต์	คำอธิบาย	ค่า
8	Elongatedness	ความยืดขยาย	26-61
9	Praxis_rectangular	มุมฉากแกน Pr	17-29
10	Length_rectangular	ความยาวมุมฉาก	118-188
11	Major_variance	ความแปรปรวนสูงสุด	130-320
12	Minor_variance	ความแปรปรวนต่ำสุด	184-1018
13	Gyration_radius	รัศมีการหมุน	109-268
14	Major_skewness	ความเบ้สูงสุด	59-135
15	Minor_skewness	ความเบ้ต่ำสุด	0-22
16	Minor_kurtosis	ความโด่งต่ำสุด	0-41
17	Major_kurtosis	ความโด้งสูงสุด	176-206
18	Hollows_ratio	อัตราส่วนความกลวง	181-211
19	Class	ตัวแปร	Positive, Negative

สำหรับชุดข้อมูล Segment ประกอบไปด้วย 20 คอลัมน์ (19 คอลัมน์สำหรับใช้ในการจำแนก และ 1 คอลัมน์สำหรับคลาส) โดยมีจำนวนข้อมูลในคลาสส่วนมากทั้งหมด 1979 ข้อมูล มีข้อมูลในคลาสส่วนน้อยทั้งหมด 329 ข้อมูล รวมทั้งหมด 2308 ข้อมูล และมีระดับความไม่สมดุลอยู่ที่ 6.02 มีรายละเอียดแต่ละคอลัมน์แสดงดังตารางที่ 3.5

ตารางที่ 3.5 ข้อมูล Segment

ลำดับ	แอททริบิวต์	คำอธิบาย	ค่า
1	Region-centroid-col	คอลัมน์บริเวนกลางพิกเซล	1-254
2	Region-centroid-row	แถวบริเวณกลางพิกเซล	11-251
3	Region-pixel-count	จำนวนพิกเซลในพื้นที่	9
4	Short-line-density-5	จำนวนบรรทัดที่มีค่าคอนทราสต์น้อยกว่าหรือเท่ากับ 5 ทั่วภูมิภาค	0.0, 0.1, 0.2, 0.3
5	Short-line-density-2	จำนวนบรรทัดที่มีค่าคอนทราสต์น้อยกว่าหรือเท่ากับ 2 ทั่วภูมิภาค	0.0, 0.1, 0.2

ตารางที่ 3.5 ข้อมูล Segment (ต่อ)

ลำดับ	แอททริบิวต์	คำอธิบาย	ค่า
6	Vedge-mean	ค่าเฉลี่ยงความคมชัดของพิกเซลที่อยู่ติดกันในแนวนอน	0.0-29.2
7	Vedge-sd	ค่าเบี่ยงเบนมาตรฐานความคมชัดของพิกเซลที่อยู่ติดกันในแนวนอน	0.0-991.7
8	Hedge-mean	ค่าเฉลี่ยความคมชัดของพิกเซลที่อยู่ติดกันในแนวตั้ง	0.0-44.7
9	Hedge-sd	ค่าเบี่ยงเบนมาตรฐานความคมชัดของพิกเซลที่อยู่ติดกันในแนวตั้ง	-1.58E-8-1386.3
10	Intensity-mean	ค่าเฉลี่ยสี RGB ทั่วบริเวณ	0.0-143.4
11	Rawred-mean	ค่าเฉลี่ยสีแดงทั่วบริเวณ	0.0-137.1
12	Rawblue-mean	ค่าเฉลี่ยน้ำเงินทั่วบริเวณ	0.0-150.8
13	Rawgreen-mean	ค่าเฉลี่ยสีเขียวทั่วบริเวณ	0.0-142.5
14	Exred-mean	ค่าแสดงสีแดงส่วนเกิน	-49.6-9.8
15	Exblue-mean	ค่าแสดงสีน้ำเงินส่วนเกิน	-12.4-82.0
16	Exgreen-mean	ค่าแสดงสีเขียวส่วนเกิน	-33.8-24.6
17	Value-mean	ค่าการแปลงแบบไม่เชิงเส้นใน 3 มิติของ RGB	0.0-150.8
18	Saturation-mean	ค่าการแปลงแบบไม่เชิงเส้นใน 3 มิติของ RGB	0.0-1.0
19	Hue-mean	ค่าการแปลงแบบไม่เชิงเส้นใน 3 มิติของ RGB	-3.04-2.9
20	Class	ตัวแปร	Positive, Negative

สำหรับชุดข้อมูล Page-Block ประกอบไปด้วย 11 คอลัมน์ (10 คอลัมน์สำหรับใช้ในการจำแนก และ 1 คอลัมน์สำหรับคลาส) โดยมีจำนวนข้อมูลในคลาสส่วนมากทั้งหมด 4913 ข้อมูล มี

ข้อมูลในคลาสส่วนน้อยทั้งหมด 559 ข้อมูล รวมทั้งหมด 5472 ข้อมูล และมีระดับความไม่สมดุลอยู่ที่ 8.79 มีรายละเอียดแต่ละคอลัมน์แสดงดังตารางที่ 3.6

ตารางที่ 3.6 ข้อมูล Page-Block

ลำดับ	แอททริบิวต์	คำอธิบาย	ค่า
1	Height	ความสูงของบล็อก	1-804
2	Lenght	ความยาวของบล็อก	1-553
3	Area	พื้นที่ของบล็อก	7-143993
4	Eccen	ความผิดปกติของบล็อก	0.0070-537.0
5	P_black	เปอร์เซ็นของพิกเซลสีดำภายในบล็อก	0.052-1.0
6	P_and	เปอร์เซ็นของพิกเซลสีดำภายในบล็อกหลังจากใช้อัลกอริทึม RLSA	0.062-1.0
7	Mean_tr	ค่าเฉลี่ยของการเปลี่ยนสีขาว-ดำ	1.0-4955.0
8	Blackpix	จำนวนพิกเซลสีดำในบิตแมปเดิมของบล็อก	1-33017
9	Blackand	จำนวนพิกเซลสีดำในบิตแมปหลังจากใช้อัลกอริทึม RLSA	7-46133
10	Wb_trans	จำนวนการเปลี่ยนสีขาว-ดำในบิตแมป	1-3212
11	Class	ตัวแปร	Positive, Negative

3.2 เครื่องมือที่ใช้ในการวิจัย

เครื่องมือที่ใช้ในการพัฒนางานวิจัยนี้คือ ประกอบด้วย

1. เครื่องคอมพิวเตอร์สำหรับการพัฒนา มีรายละเอียดดังนี้
    - หน่วยประมวลผลกลาง : Intel Core i5
    - หน่วยความจำสำรอง : 1 TB
    - หน่วยความจำหลัก : 4 GB
2. ระบบปฏิบัติการและโปรแกรมประยุกต์สำหรับการพัฒนา ประกอบไปด้วย
    - ระบบปฏิบัติการ : Windows 8.1 Pro
    - เครื่องมือที่ใช้ในการพัฒนา : IBM SPSS Modeler version 18.0

3.3 กรอบแนวคิดของการวิจัย

แนวคิดหลักของงานวิจัยนี้คือ การจำแนกข้อมูลไม่สมดุล โดยอาศัยการปรับข้อมูลด้วยเทคนิคการสุ่มซ้ำ ที่ผนวกกับเทคนิคการจัดกลุ่มข้อมูลเพื่อสังเคราะห์สร้างข้อมูลเป็นข้อมูลตัวแทนสำหรับการเรียนรู้ที่จำนวนข้อมูลคลาสส่วนมากมีความสมดุลกับคลาสส่วนน้อยมากขึ้นตามอัตราส่วนที่เหมาะสม จากนั้นจึงทำการเรียนรู้ข้อมูลด้วยอัลกอริทึมต้นไม้ตัดสินใจแบบจำแนก และแบบถดถอย เพื่อหาแบบจำลองการเรียนรู้ จากนั้นจึงนำแบบจำลองดังกล่าวมาทำนายชุดข้อมูลทดสอบ เพื่อประเมินประสิทธิภาพของการจำแนก ระหว่างแบบจำลองที่ได้ก่อนการปรับสมดุลและหลังการปรับสมดุลข้อมูลแล้ว

กรอบแนวคิดของการวิจัยนี้ประกอบไปด้วย 4 ส่วนหลัก คือ

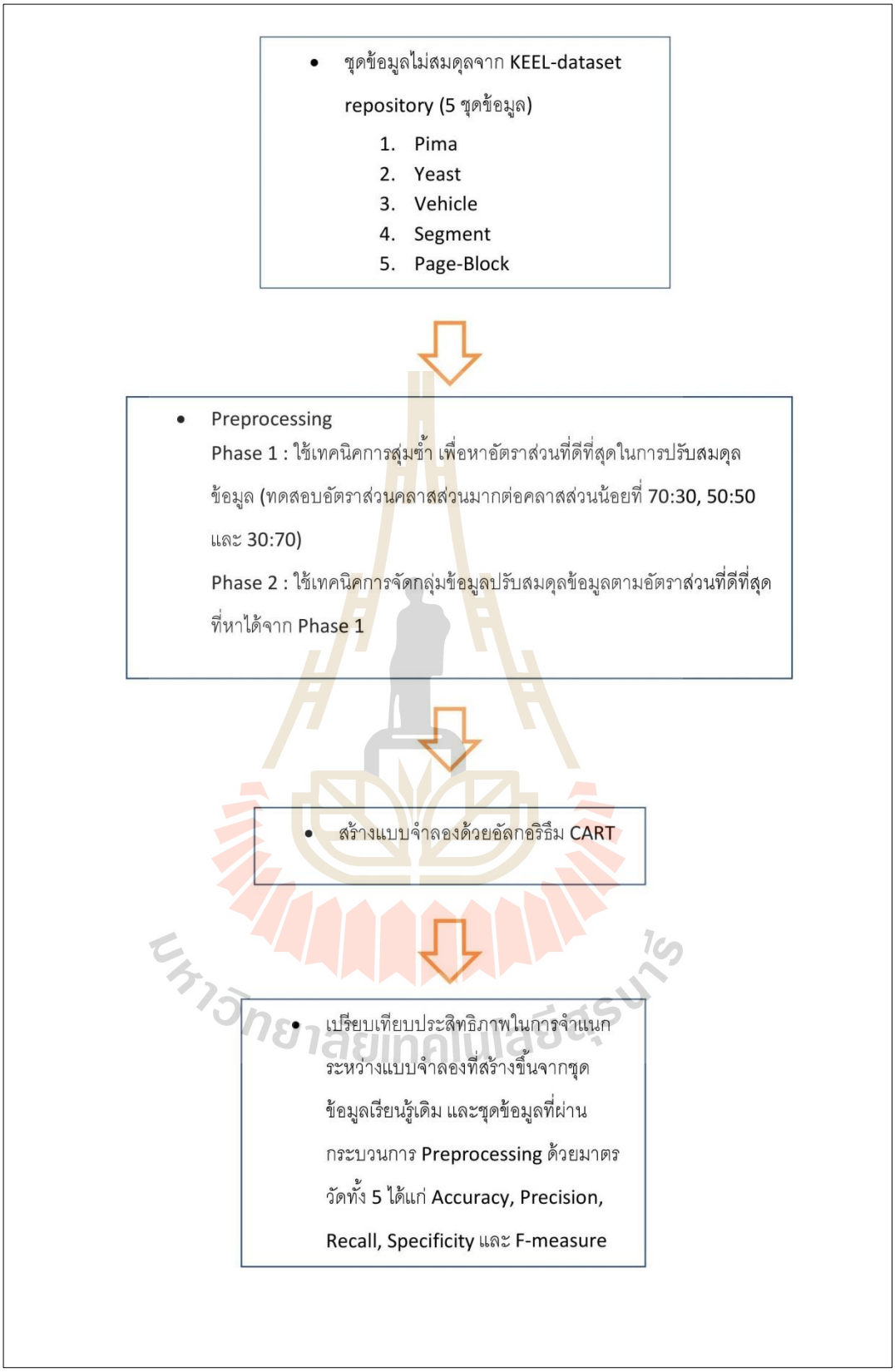
1. การรวบรวมชุดข้อมูลไม่สมดุล ซึ่งรวบรวมมาจาก KEEL-dataset repository จำนวน 5 ชุดข้อมูลได้แก่ Pima, Yeast, Vehicle, Segment และ Page-Block

2. การเตรียมข้อมูล สำหรับส่วนนี้แบ่งเป็น 2 ระยะ คือระยะการปรับข้อมูลให้มีความสมดุลระหว่างคลาสส่วนมากและคลาสน้อยใน 3 อัตราส่วนที่สนใจ คือ 70:30, 50:50 และ 30:70 เพื่อหาอัตราส่วนที่เหมาะสมสำหรับการปรับข้อมูลด้วยเทคนิคการสุ่มซ้ำ และระยะการปรับข้อมูลให้มีความสมดุลด้วยเทคนิคการสุ่มซ้ำ และจัดกลุ่มข้อมูลตามอัตราส่วนที่แบบจำลองมีประสิทธิภาพในการจำแนกดีที่สุดจากระยะที่ 1

3. การสร้างแบบจำลองเพื่อจำแนกข้อมูลไม่สมดุล ด้วยอัลกอริทึมต้นไม้ตัดสินใจแบบจำแนก และแบบถดถอย

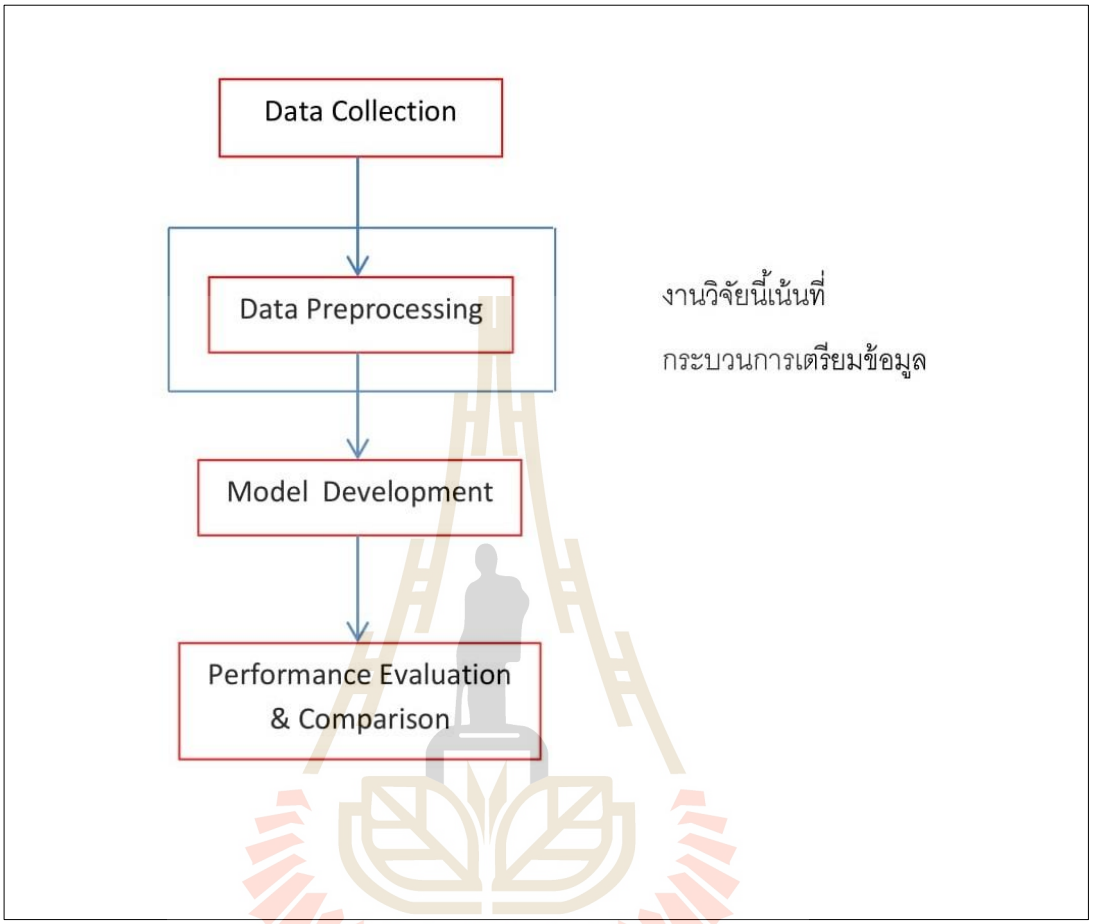
และ 4. การประเมินประสิทธิภาพของแบบจำลองการจำแนกข้อมูลไม่สมดุลด้วยมาตรวัดทั้ง 5 ได้แก่ ค่าความแม่นยำในการจำแนก (Accuracy) ค่าความเที่ยง (Precision) ค่าความไวหรือค่าระลึก (Recall) ค่าความจำเพาะ (Specificity) และค่าการวัดเอฟ (F-measure)

ดังแสดงกรอบแนวคิดของการวิจัยดังรูปที่ 3.1



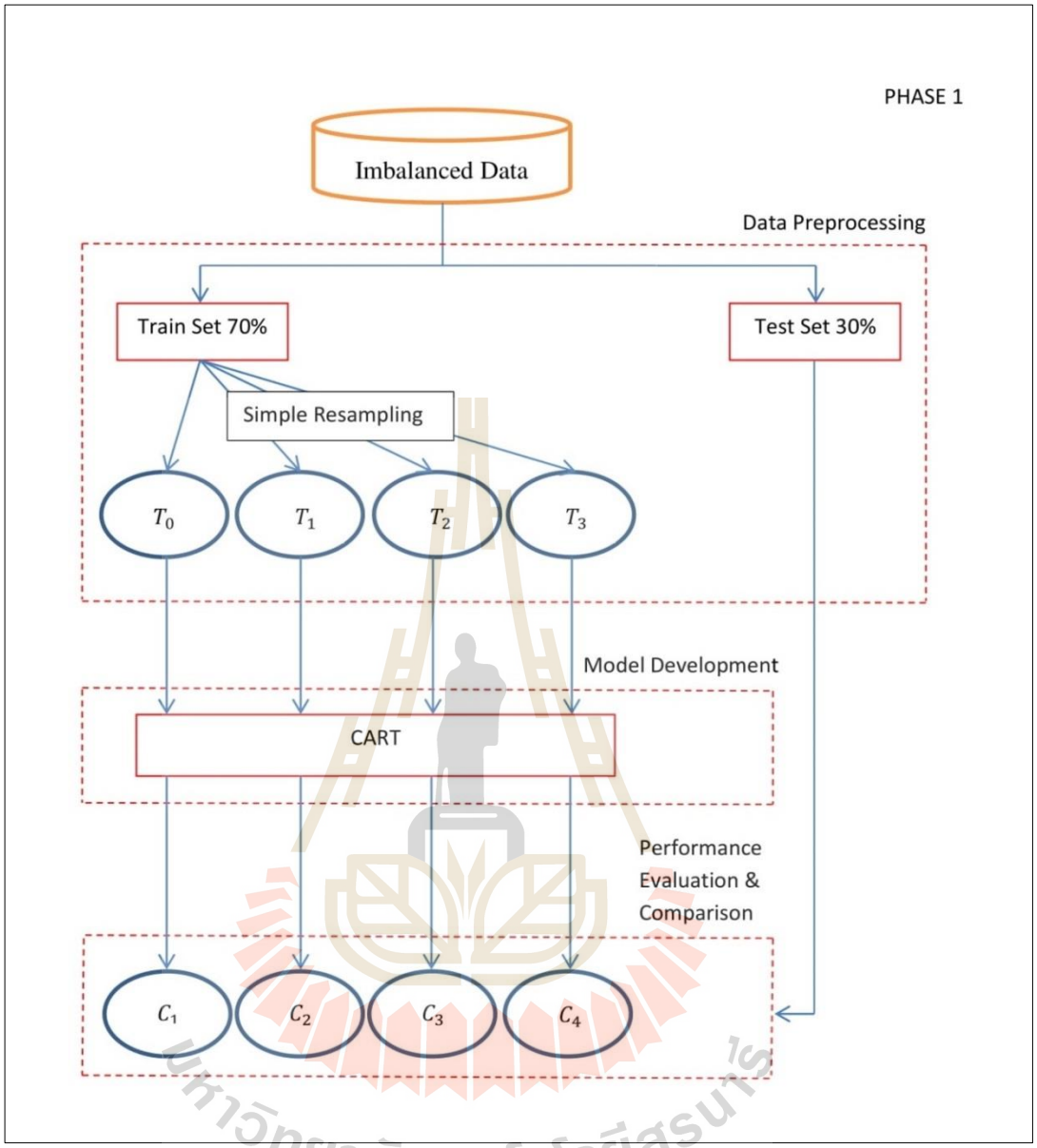
รูปที่ 3.1 กรอบแนวคิดของการวิจัย

3.4 วิธีดำเนินการวิจัย



รูปที่ 3.2 แสดงภาพรวมวิธีดำเนินการวิจัย

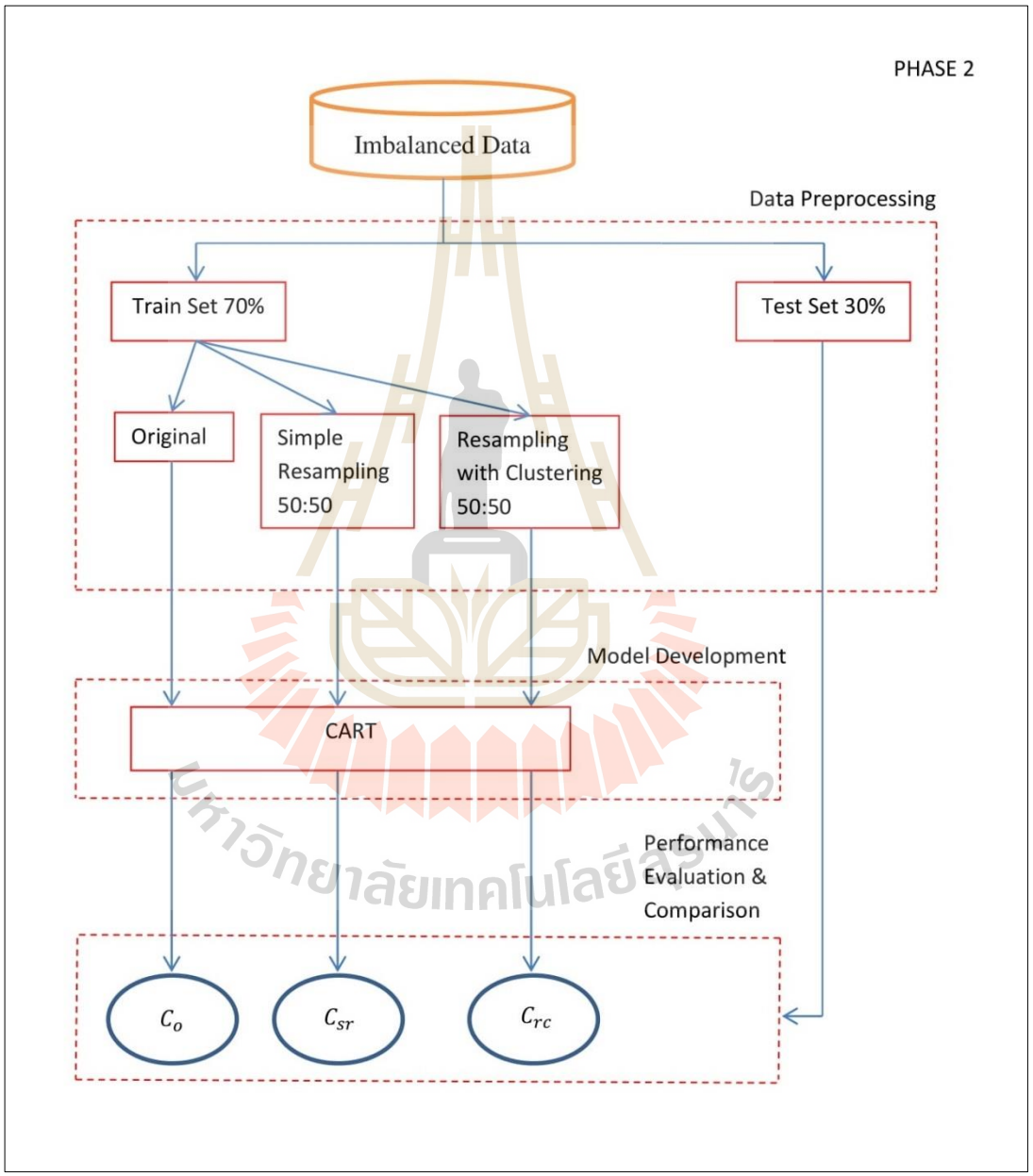
จากรูปที่ 3.2 วิธีดำเนินงานวิจัยแบ่งเป็น 4 ขั้นตอน โดยจุดประสงค์หลักของงานวิจัยนี้จะเน้นไปที่ขั้นตอนที่ 2 หรือขั้นตอนการเตรียมข้อมูล ซึ่งได้แบ่งขั้นตอนนี้เป็น 2 ระยะดังแสดงในรูปที่ 3.3 และ 3.4 ตามลำดับ



รูปที่ 3.3 แสดงวิธีดำเนินการวิจัยระยะที่ 1

จากรูปที่ 3.3 แสดงวิธีการดำเนินการวิจัยระยะที่ 1 เริ่มต้นด้วยการแบ่งชุดข้อมูลเป็น 2 ส่วน คือ ชุดข้อมูลเรียนรู้อัตราส่วน 70% และชุดข้อมูลทดสอบ 30% ขั้นถัดมาในส่วนชุดข้อมูลเรียนรู้ได้ใช้เทคนิคการสุ่มซ้ำปรับสมดุลระหว่างคลาสส่วนมาก และคลาสส่วนน้อยในอัตราส่วน 70:30 ที่ชุดข้อมูล T_1, อัตราส่วน 50:50 ที่ชุดข้อมูล T_2 และอัตราส่วน 30:70 ที่ชุดข้อมูล T_3 ถัดมาจึงทำการเรียนรู้ข้อมูลทั้ง 4 ชุด (รวม T_0 ชุดข้อมูลดั้งเดิม) ด้วยอัลกอริทึมต้นไม้ตัดสินใจแบบจำแนกและแบบถดถอยได้แบบจำลองออกมาเป็น C_1, C_2, C_3 และ C_4 ตามลำดับ สุดท้ายจึงนำแบบจำลองทั้ง 4

มาจำแนกชุดข้อมูลที่แยกไว้ตามรูปภาพแสดงข้างต้น แล้วเปรียบเทียบประสิทธิภาพในการจำแนกของแบบจำลองต่าง ๆ ด้วยมาตรวัดทั้ง 5 เพื่อสรุปอัตราส่วนที่ให้ประสิทธิภาพในการจำแนกที่ดีที่สุด (ซึ่งผลการวิจัยที่ได้คืออัตราส่วน 50:50 โดยผลการวิจัยอย่างละเอียดจะแสดงในบทที่ 4) เพื่อนำไปสู่การดำเนินการวิจัยระยะที่ 2 ซึ่งเน้นไปที่การปรับปรุงประสิทธิภาพในการจำแนกให้ดียิ่งขึ้น



รูปที่ 3.4 แสดงวิธีดำเนินงานวิจัยระยะที่ 2

จากรูปที่ 3.4 แสดงวิธีการดำเนินการวิจัยระยะที่ 2 เริ่มต้นด้วยการแบ่งชุดข้อมูลเป็น 2 ส่วน คือ ชุดข้อมูลเรียนรู้อัตราส่วน 70% และชุดข้อมูลทดสอบ 30% โดยในส่วนของชุดข้อมูลเรียนรู้จะแบ่งเป็น 3 ชุด คือชุดข้อมูลดั้งเดิม, ชุดข้อมูลที่ผ่านการปรับสมดุลระหว่างคลาสส่วนมาก และคลาสส่วนน้อยด้วยเทคนิคการสุ่มซ้ำในอัตราส่วน 50:50 และชุดข้อมูลที่ผ่านการปรับสมดุลระหว่างคลาสส่วนมาก และคลาสส่วนน้อยด้วยเทคนิคการสุ่มซ้ำ และการเทคนิคการจัดกลุ่มข้อมูลในอัตราส่วน 50:50 ซึ่งทำโดยการจัดกลุ่มข้อมูลในคลาสส่วนมากให้มีจำนวนกลุ่มเท่ากับจำนวนคลาสส่วนน้อย แล้วใช้ค่าเฉลี่ยของแต่ละแอททริบิวต์มาเป็นตัวแทนของชุดข้อมูลใหม่ ที่ผ่านการจัดกลุ่มข้อมูลแล้ว ถัดมาจึงทำการเรียนรู้ข้อมูลทั้ง 3 ชุด ด้วยอัลกอริทึมต้นไม้ตัดสินใจแบบจำแนกและแบบถดถอยได้แบบจำลองออกมาเป็น C_o, C_{sr} และ C_{rc} ตามลำดับ สุดท้ายจึงนำแบบจำลองทั้ง 3 มาจำแนกชุดข้อมูลที่แยกไว้ตามรูปภาพแสดงข้างต้น แล้วเปรียบเทียบประสิทธิภาพในการจำแนกของแบบจำลองต่าง ๆ ด้วยมาตรวัดทั้ง 5 เพื่อสรุปประสิทธิภาพในการจำแนก

บทที่ 4
ผลการทดสอบและอภิปรายผล

การทดสอบประสิทธิภาพของแบบจำลองนั้น จะทดสอบประสิทธิภาพด้วยค่าความแม่นยำในการจำแนก (Accuracy) ค่าความเที่ยง (Precision) ค่าความไวหรือค่าระลึก (Recall) ค่าความจำเพาะ (Specificity) และค่าการวัดเอฟ (F-measure) ในการจำแนกคลาสส่วนน้อยในข้อมูลไม่สมดุล โดยเปรียบเทียบการจำแนกระหว่างแบบจำลองที่ได้จากการเรียนรู้ชุดข้อมูลเรียนรู้ดั้งเดิม (ไม่ปรับสมดุลข้อมูล) กับแบบจำลองได้จากการปรับสมดุลข้อมูลด้วยเทคนิคการสุ่มซ้ำ และแบบจำลองที่ได้จากการปรับข้อมูลด้วยเทคนิคการสุ่มซ้ำ และการจัดกลุ่มข้อมูล สำหรับเนื้อหาในบทนี้จะประกอบด้วย ผลการทดสอบประสิทธิภาพแบบทดลองที่กล่าวข้างต้น และการอภิปรายผล

4.1 ผลการทดสอบประสิทธิภาพ

สำหรับการทดสอบประสิทธิภาพการจำแนกประเภทข้อมูลไม่สมดุลนั้น จะใช้ข้อมูลทั้งหมด 5 ชุดข้อมูล เป็นข้อมูลจริงจากฐานข้อมูลมาตรฐานจำนวน 5 ชุดข้อมูล เมื่อทำการเปรียบเทียบประสิทธิภาพการจำแนกข้อมูลไม่สมดุลระหว่างแบบจำลองที่สร้างจากชุดข้อมูลเรียนรู้แบบดั้งเดิม กับแบบจำลองที่สร้างจากชุดข้อมูลที่ปรับสมดุลในอัตราส่วนต่าง ๆ แล้วตามที่มีการนำเสนอไปในบทที่ 3 สำหรับวิธีดำเนินการวิจัยในระยะที่ 1 โดยมีรายละเอียดค่าประสิทธิภาพตามมาตรวัดต่าง ๆ แสดงดังตารางที่ 4.1

ตารางที่ 4.1 รายละเอียดค่าประสิทธิภาพตามมาตรวัดต่าง ๆ เมื่อทดสอบโมเดลด้วยชุดข้อมูลทดสอบในส่วนการวิจัยในระยะที่ 1

ชุดข้อมูล	อัตราส่วน	Accuracy	Precision	Recall	Specificity	F-measure
Pima (IR=1.87)	ดั้งเดิม	0.68	0.83	0.30	0.95	0.44
	70:30	0.76	0.78	0.60	0.88	0.68
	50:50	0.71	0.63	0.76	0.67	0.69
	30:70	0.59	0.50	0.96	0.31	0.66

ตารางที่ 4.1 รายละเอียดค่าประสิทธิภาพตามมาตรวัดต่าง ๆ เมื่อทดสอบโมเดลด้วยชุดข้อมูลทดสอบในส่วนการวิจัยในระยะที่ 1 (ต่อ)

ชุดข้อมูล	อัตราส่วน	Accuracy	Precision	Recall	Specificity	F-measure
Yeast (IR=2.46)	ดั้งเดิม	0.75	0.59	0.39	0.89	0.47
	70:30	0.75	0.59	0.48	0.86	0.53
	50:50	0.71	0.49	0.70	0.71	0.58
	30:70	0.50	0.36	0.93	0.32	0.52
Vehicle (IR=3.25)	ดั้งเดิม	0.89	0.78	0.54	0.96	0.64
	70:30	0.94	0.80	0.89	0.95	0.77
	50:50	0.89	0.63	1.00	0.86	0.85
	30:70	0.86	0.57	1.00	0.83	0.72
Segment (IR=6.02)	ดั้งเดิม	0.99	0.97	0.99	0.99	0.98
	70:30	0.98	0.86	1.00	0.97	0.93
	50:50	0.98	0.87	1.00	0.98	0.92
	30:70	0.95	0.71	1.00	0.94	0.83
Page-Block (IR=8.79)	ดั้งเดิม	0.96	0.81	0.83	0.97	0.82
	70:30	0.96	0.74	0.86	0.97	0.80
	50:50	0.92	0.57	0.95	0.92	0.71
	30:70	0.82	0.35	1.00	0.79	0.52

จากตารางที่ 4.1 สังเกตผลจากมาตรวัดค่าต่าง ๆ สรุปได้ดังนี้

1. เมื่อพิจารณามาตรวัดค่าความแม่นยำในการจำแนก จากข้อมูลทั้งหมด 5 ชุด แบบจำลองจากการปรับอัตราส่วนจำนวนคลาสส่วนมากต่อคลาสส่วนน้อยที่ 70:30 ให้ค่าความแม่นยำในการจำแนกดีที่สุด (ดีที่สุด หมายถึง ให้ค่ามาตรวัดที่พิจารณาสูงที่สุดเป็นจำนวนมากชุดข้อมูลที่สุดจากทั้ง 5 ชุดข้อมูล)

2. เมื่อพิจารณามาตรวัดค่าความเที่ยง จากข้อมูลทั้งหมด 5 ชุด แบบจำลองจากชุดข้อมูลดั้งเดิมให้ค่าความเที่ยงดีที่สุด
3. เมื่อพิจารณามาตรวัดค่าความไว หรือค่าระลึก จากข้อมูลทั้งหมด 5 ชุด แบบจำลองจากการปรับอัตราส่วนจำนวนคลาสส่วนมากต่อคลาสส่วนน้อยที่ 30:70 ให้ค่าความความไว หรือค่าระลึกดีที่สุด
4. เมื่อพิจารณามาตรวัดค่าความจำเพาะจากข้อมูลทั้งหมด 5 ชุด แบบจำลองจากชุดข้อมูลดั้งเดิมให้ค่าความจำเพาะดีที่สุด
5. เมื่อพิจารณามาตรวัดค่าการวัดเอฟ หรือค่าระลึก จากข้อมูลทั้งหมด 5 ชุด แบบจำลองจากการปรับอัตราส่วนจำนวนคลาสส่วนมากต่อคลาสส่วนน้อยที่ 50:50 ให้ค่าการวัดเอฟดีที่สุด

เมื่อพิจารณาครบทั้ง 5 มาตรวัดแล้ว เลือกการปรับอัตราส่วนจำนวนคลาสส่วนมากต่อคลาสส่วนน้อยที่ 50:50 เป็นอัตราส่วนที่ดีที่สุด เนื่องด้วยส่วนหนึ่งให้ค่าการวัดเอฟที่ดีสุดต่อจำนวนชุดข้อมูลทั้งหมด ซึ่งค่าการวัดเอฟเป็นค่าที่มีความเกี่ยวข้องกับค่าความเที่ยง และค่าความไว หรือค่าระลึกด้วย ซึ่งถึงแม้อัตราส่วนข้างต้นจะไม่ได้ให้ค่าทั้ง 2 ค่าดังกล่าวดีที่สุด แต่ก็สามารถบอกได้ว่าค่อนข้างให้ค่าไปในทางที่ดี และมีประสิทธิภาพไม่ขัดแย้งกันทั้ง ๆ ที่ค่าทั้ง 2 ควรจะเป็นไปในทิศทางเดียวกันหากแบบจำลองนั้นดีต่อการจำแนกคลาสส่วนน้อย และเมื่อพิจารณาค่าอื่น ๆ ประกอบก็พบว่าแบบจำลองจากอัตราส่วน 50:50 นั้นไม่ให้ค่าจากมาตรวัดใดเป็นแบบจำลองที่แย่ที่สุดเลย

เมื่อสรุปได้จากระยะที่ 1 แล้วว่าอัตราส่วนสำหรับการปรับจำนวนข้อมูลคลาสส่วนมากต่อคลาสส่วนน้อยคือ อัตราส่วน 50:50 จึงดำเนินการวิจัยในระยะที่ 2 และได้ผลการทดสอบแสดงดังตารางที่ 4.2

ตารางที่ 4.2 รายละเอียดค่าประสิทธิภาพตามมาตรวัดต่าง ๆ เมื่อทดสอบโมเดลด้วยชุดข้อมูลทดสอบในส่วนการวิจัยในระยะที่ 2

ชุดข้อมูล	อัตราส่วน	Accuracy	Precision	Recall	Specificity	F-measure
Pima (IR=1.87)	ดั้งเดิม	0.68	0.83	0.30	0.95	0.44
	50:50	0.71	0.63	0.76	0.68	0.69
	50:50 + จัดกลุ่มข้อมูล	0.73	0.57	0.67	0.76	0.61

ตารางที่ 4.2 รายละเอียดค่าประสิทธิภาพตามมาตรวัดต่าง ๆ เมื่อทดสอบโมเดลด้วยชุดข้อมูลทดสอบในส่วนการวิจัยในระยะที่ 2 (ต่อ)

ชุดข้อมูล	อัตราส่วน	Accuracy	Precision	Recall	Specificity	F-measure
Yeast (IR=2.46)	ดั้งเดิม	0.75	0.59	0.39	0.89	0.47
	50:50	0.71	0.49	0.70	0.71	0.58
	50:50 + จัดกลุ่มข้อมูล	0.59	0.38	0.74	0.54	0.50
Vehicle (IR=3.25)	ดั้งเดิม	0.89	0.78	0.54	0.97	0.64
	50:50	0.89	0.63	1.00	0.87	0.77
	50:50 + จัดกลุ่มข้อมูล	0.91	0.72	0.96	0.89	0.83
Segment (IR=6.02)	ดั้งเดิม	0.99	0.97	0.99	1.00	0.98
	50:50	0.98	0.87	1.00	0.98	0.93
	50:50 + จัดกลุ่มข้อมูล	0.99	0.95	0.97	0.99	0.96
Page-Block (IR=8.79)	ดั้งเดิม	0.96	0.81	0.83	0.98	0.82
	50:50	0.92	0.57	0.95	0.92	0.71
	50:50 + จัดกลุ่มข้อมูล	0.91	0.53	0.91	0.91	0.67

จากตารางที่ 4.2 เป็นผลการทดลองของทั้ง 5 ชุดข้อมูล เมื่อพิจารณาชุดข้อมูลเรียนรู้แต่ละแบบ แต่ละชุดให้ค่ามาตรวัดในการทดสอบประสิทธิภาพค่าต่าง ๆ ดีที่สุดในกรณีต่าง ๆ กันไป เพื่อให้สามารถพิจารณาได้ง่ายขึ้น จึงสรุปค่าเฉลี่ยของแต่ละวิธีมาแสดงที่ตารางที่ 4.3

ตารางที่ 4.3 รายละเอียดค่าประสิทธิภาพตามมาตรวัดต่าง ๆ แบบค่าเฉลี่ยทั้ง 5 ชุดข้อมูลจากตารางที่ 4.2

อัตราส่วน	Accuracy	Precision	Recall	Specificity	F-measure
ดั้งเดิม	0.85	0.80	0.61	0.96	0.67
50:50	0.84	0.64	0.88	0.83	0.74
50:50 + จัดกลุ่มข้อมูล	0.83	0.63	0.85	0.82	0.71

จากตารางที่ 4.3 เมื่อพิจารณาจากค่าการวัดเอฟ แบบจำลองจากชุดข้อมูลที่ปรับอัตราส่วนคลาสส่วนมากต่อคลาสส่วนน้อยเป็น 50:50 และไม่ได้ทำการจัดกลุ่มข้อมูลเพื่อใช้ค่าเฉลี่ยเป็นตัวแทนชุดข้อมูลนั้น ให้ประสิทธิภาพโดยรวมที่ดีที่สุด แต่การปรับข้อมูลด้วยวิธีการสุ่มธรรมดาอาจมีข้อกังขาในเรื่องการเลือกเฉพาะชุดข้อมูลที่เหมาะสมกับแบบจำลอง (Bias) เพื่อให้ได้ค่าประสิทธิภาพที่สูงกว่าที่ควรจะเป็น อย่างไรก็ตามเมื่อพิจารณาแบบจำลองจากชุดข้อมูลที่ปรับอัตราส่วนคลาสส่วนมากต่อคลาสส่วนน้อยเป็น 50:50 ด้วยวิธีการจัดกลุ่มข้อมูล แล้วใช้ค่าเฉลี่ยเป็นค่าตัวแทนนั้นให้ค่ามาตรวัดต่าง ๆ ไม่แตกต่างจากแบบจำลองจากข้อมูลชุดข้างต้นมากหนัก แต่ก็ดีกว่าแบบจำลองที่ได้จากชุดข้อมูลดั้งเดิมถึง 4% (ดูจากค่าการวัดเอฟ)

4.2 อภิปรายผล

จากผลการทดสอบประสิทธิภาพการจำแนกข้อมูลไม่สมดุล ได้ทำการทดสอบกับข้อมูลจำนวน 5 ชุดข้อมูล โดยแต่ละชุดข้อมูลประกอบไปด้วยคลาส 2 คลาส กระบวนการปรับสมดุลข้อมูลถูกนำมาใช้เพื่อปรับจำนวนข้อมูลในแต่ละคลาสให้มีขนาดใกล้เคียงกันเพื่อไม่ให้เกิดการเอนเอียงไปทางคลาสที่มีจำนวนข้อมูลมากกว่า (คลาสส่วนมาก) ก่อนนำข้อมูลไปสร้างแบบจำลองในการจำแนกประเภท และประเมินประสิทธิภาพการจำแนก สามารถสรุปผลการทดสอบเปรียบเทียบได้ดังนี้

1) การปรับสมดุลข้อมูลเรียนรู้ เพื่อเตรียมข้อมูลก่อนการนำไปสร้างแบบจำลองการจำแนกมีความสามารถในการจำแนกข้อมูลจากคลาสส่วนน้อยได้ดียิ่งขึ้น (สามารถจำแนกคลาสส่วนน้อยได้แม่นยำมากขึ้น)

2) การใช้อัลกอริทึมต้นไม้ตัดสินใจแบบจำแนกและแบบถดถอยแบบดั้งเดิม ให้ประสิทธิภาพที่ดีเมื่อวัดประสิทธิภาพด้วยค่าความแม่นยำในการจำแนก และค่าความจำเพาะ

เนื่องจากหากทำนายข้อมูลทดสอบทั้งหมดให้เป็นคลาสส่วนมากทั้งหมดก็ส่งผลให้ความแม่นยำในการจำแนกสูง แต่ในขณะที่ความสามารถในการจำแนกข้อมูลจากคลาสส่วนน้อยจะต่ำมาก

3) อัลกอริทึมต้นไม้ตัดสินใจแบบจำแนกและแบบถดถอยมีประสิทธิภาพที่ดีมากในด้านค่าความไว หรือค่าระลึก เนื่องจากจำนวนที่จำแนกประเภทข้อมูลเป็นคลาสส่วนน้อยมีการทำนายในปริมาณที่น้อยและในจำนวนที่ทำนายสามารถทำนายได้ถูกจึงส่งผลให้ค่าความไว หรือค่าความระลึกมีค่าสูง แต่เมื่อพิจารณาว่าสามารถจำแนกจำนวนข้อมูลที่อยู่ในคลาสส่วนน้อยได้ทั้งหมด พบว่ามีประสิทธิรูปที่ต่ำ (พิจารณาจากค่าความเที่ยง)

4) เมื่อพิจารณาความสามารถในการจำแนกประเภทข้อมูลจากคลาสส่วนน้อยโดยรวม พิจารณาจากค่าการวัดเอฟ (ค่าการวัดเอฟเป็นอัตราเฉลี่ยระหว่างจำนวนที่จำแนกข้อมูลว่าเป็นคลาสส่วนน้อยได้ถูกต้อง และจำนวนข้อมูลที่อยู่ในคลาสส่วนน้อยทั้งหมด โดยพิจารณาจากค่าความเที่ยงและค่าระลึกหรือค่าความไว) หากค่าการวัดเอฟมีค่าที่สูงหมายความว่าโมเดลมีจำนวนการจำแนกประเภทข้อมูลที่อยู่ในคลาสส่วนน้อยในปริมาณที่สูง และการจำแนกข้อมูลในคลาสส่วนน้อยนั้นแม่นยำสูงด้วย จึงสรุปได้ว่าการปรับสมดุลข้อมูล และการปรับสมดุลข้อมูลด้วยการจัดกลุ่มข้อมูลนั้นสามารถช่วยเพิ่มประสิทธิภาพโดยรวมของอัลกอริทึมจริง

5) เมื่อพิจารณาประสิทธิภาพการจำแนกข้อมูลไม่สมดุล พบว่าเทคนิคที่นำเสนอเหมาะสำหรับนำไปจำแนกประเภทข้อมูลไม่สมดุลที่มีระดับความไม่สมดุลในช่วง 3.25 - 6.02 ได้ดี (พิจารณาค่าต่าง ๆ อย่างละเอียดจากตารางที่ 4.2 พบว่าเมื่อปรับสมดุลด้วยการจัดกลุ่มข้อมูล แล้วใช้ค่าเฉลี่ยเป็นตัวแทนนั้น ให้ประสิทธิภาพที่ดีกับชุดข้อมูล Vehicle และ Segment)

บทที่ 5
สรุปและข้อเสนอแนะ

5.1 สรุปผลการวิจัย

ข้อดีจากการจำแนกข้อมูลด้วยอัลกอริทึมต้นไม้ตัดสินใจแบบจำแนกและแบบถดถอยคือความแม่นยำที่ค่อนข้างสูงแล้ว ยังสามารถใช้งานได้หลากหลายขึ้นอยู่กับข้อมูลที่เราสนใจ โดยอัลกอริทึมต้นไม้แบบจำแนกและแบบถดถอยใช้ได้กับทั้งข้อมูลแบบตัวเลขและแบบข้อความ ทั้งนี้ยังสามารถแสดงผลการทำงานได้ใน 2 ลักษณะขึ้นกับตัวแปรที่เป็นเป้าหมาย โดยหากตัวแปรเป้าหมายเป็นข้อมูลแบบตัวเลข อัลกอริทึมนี้สามารถทำการพยากรณ์ตัวเลขเป้าหมายจากแนวโน้มข้อมูล และหากข้อมูลเป็นข้อความ อัลกอริทึมนี้ก็สามารถจัดกลุ่มข้อมูลได้ตามจำนวนกลุ่มข้อมูลพร้อมทั้งบอกแนวโน้มของข้อมูลอื่น ๆ ของแต่ละกลุ่มเช่นกัน ทว่าปัจจุบันนี้ข้อมูลส่วนใหญ่ในชีวิตจริงนั้นเป็นข้อมูลไม่สมดุล ทำให้อัลกอริทึมพื้นฐานโดยทั่วไปไม่สามารถจำแนกข้อมูลได้ดีพอ โดยมักจะละทิ้งข้อมูลคลาสส่วนน้อยเพื่อให้จำแนกแล้วมีความผิดพลาดน้อยที่สุด ซึ่งก่อให้เกิดปัญหาต่อการนำแบบจำลองไปใช้งานจริงในลักษณะของการทำงานที่คลาสส่วนน้อยมีความสำคัญสูง ยกตัวอย่างเช่นสาธารณสุขศาสตร์ที่มีจำนวนคนเป็นโรคน้อยกว่าคนปกติ แต่อัลกอริทึมต้นไม้ตัดสินใจแบบจำแนกและแบบถดถอยนั้น จำแนกข้อมูลด้วยการหาจุดแบ่งเพื่อแบ่งข้อมูลให้ได้มีความบริสุทธิ์มากที่สุด ซึ่งก็คือในคลาสส่วนมากจะพยายามไม่มีคลาสส่วนน้อยมาเจือปน หรือเจือปนให้น้อยที่สุด และคลาสส่วนน้อยก็ให้มีคลาสส่วนมากเจือปนน้อย หรือไม่เจือปนเลยเช่นเดียวกัน ด้วยเหตุผลนี้จึงสามารถนำมาช่วยจำแนกข้อมูลที่ไม่มีความสมดุลได้ดี

ต่อมาหลังจากนำเทคนิคการสุ่มซ้ำข้อมูลเข้ามาใช้ ปรากฏว่าผลความแม่นยำของอัลกอริทึมต้นไม้ตัดสินใจแบบจำแนกและแบบถดถอยในการจำแนกข้อมูลไม่สมดุลเพิ่มขึ้น ซึ่งจากผลการทดสอบที่แสดงไปในบทที่ 4 นั้นพบว่าเมื่อทำการปรับสมดุลอัตราส่วนข้อมูลคลาสส่วนมากและคลาสส่วนน้อยเป็น 50:50 นั้นแบบจำลองที่ได้จากอัลกอริทึมต้นไม้ตัดสินใจแบบจำแนกและแบบถดถอยมีประสิทธิภาพโดยรวมสูงขึ้นถึง 7% แต่เพื่อไม่ให้มีข้อกังขาในการเลือกข้อมูล จึงใช้เทคนิคการจัดกลุ่มข้อมูลมาทำงานในส่วนของการปรับอัตราส่วนข้อมูลระหว่างคลาสส่วนมากและคลาสส่วนน้อย จากนั้นจึงใช้ค่าเฉลี่ยของแต่ละกลุ่มข้อมูลเป็นตัวแทนกลุ่มข้อมูลนั้น ๆ ซึ่งก็ยังช่วยเพิ่มประสิทธิภาพให้กับอัลกอริทึมให้จำแนกคลาสส่วนน้อยได้ดีขึ้นถึง 4%

5.2 การประยุกต์ผลการวิจัย

จากผลการวิจัยแสดงให้เห็นว่าสามารถนำอัลกอริทึมต้นไม้ตัดสินใจแบบจำแนกและแบบถดถอยร่วมกับเทคนิคการสุ่มซ้ำไปใช้ในการทำงานจริงได้อย่างมีประสิทธิภาพ โดยเฉพาะการนำไปใช้งานทางสาธารณสุขในการช่วยจำแนกกลุ่มผู้ป่วยโรคต่าง ๆ รวมไปถึงแสดงแนวโน้มของปัจจัยที่ทำให้เป็นโรคใดโรคหนึ่ง หรือกระทั่งการคาดการณ์จำนวนผู้ป่วยโรค เพื่อเตรียมการรับมือล่วงหน้า ซึ่งต้องการทั้งความแม่นยำ และความรวดเร็ว ทั้งนี้อัลกอริทึมต้นไม้ตัดสินใจแบบจำแนกและแบบถดถอยยังตอบสนองต่อความต้องการในการลดค่าใช้จ่าย ลดจำนวนข้อมูลที่ต้องเก็บรวบรวมเพื่อนำมาใช้ในการวิเคราะห์อีกด้วย และช่วยแก้ปัญหาหากข้อมูลบางส่วนขาดหายไป

5.3 ข้อเสนอแนะ

แม้อัลกอริทึมต้นไม้ตัดสินใจแบบจำแนกและแบบถดถอยจะมีประสิทธิภาพในการนำไปใช้จำแนกข้อมูลไม่สมดุล แต่การวิจัยในที่นี้ยังเป็นการวิจัยเบื้องต้นด้วยการกำหนดอัตราส่วนคร่าว ๆ ในการสุ่มซ้ำ ซึ่งอาจไม่ได้มีผลลัพธ์ที่ดีต่อทุกชุดข้อมูล ในอนาคตจึงมีแผนที่จะปรับใช้ความรู้ที่ได้จากงานวิจัยนี้เพื่อหาสมการ หรือค่าดัชนีที่ดีต่อการนำวิธีการนี้ไปใช้กับข้อมูลทุก ๆ ชุดในอนาคต

รายการอ้างอิง

กิตติพงษ์ ชมบุญ. (2016). เทคนิคการค้นหาคลาสที่ค้นพบได้ยากสำหรับข้อมูลที่ขนาดแตกต่างกันมาก. **วิทยานิพนธ์วิศวกรรมศาสตร์ดุษฎีบัณฑิต.** สาขาวิชาวิศวกรรมคอมพิวเตอร์ มหาวิทยาลัยเทคโนโลยีสุรนารี.

ดิษฐพล มั่นธรรม และ ลี่ลี อิงศรีสว่าง. (2010). การประยุกต์ขั้นตอนวิธีต้นไม้ตัดสินใจกับการวินิจฉัยโรคระบบการหายใจ : กรณีศึกษาที่โรงพยาบาลพระนครศรีอยุธยา. **วารสารวิจัยระบบสาธารณสุข.** 4(1): 73-81.

Batista G. E. A. P. A., Prati R. C., and Monard M. C. (2004). A study of the behavior of several methods for balancing machine learning training data. **SIGKDD Explorations.** 6(1).

Boonchuay K., Sinapiromsaran K. and Lursinsap C. (2011). Minority Split and Gain Ratio for a Class Imbalance. **Eight International Conference on Fuzzy Systems and Knowledge Discovery (FSKD) 2011**. pp. 2060-3064.

Breiman L., Friedman J.H., Olshen R. and Stone C.J. (1984). Classification and Regression Tree. **Wadsworth & Brooks/Cole Advanced Books & Software**, Pacific California.

Chawla N.V., Bowyer K.W., Hall L.O., and Kegelmeyer W.P. (2002). SMOTE: synthetic minority over-sampling technique. **Journal of Artificial Intelligence Research.** 16:321–357.

Chawla N.V., Japkowicz N., and Kotcz A. (2004). Editorial: special issue on learning from imbalanced data sets. **ACM Sigkdd Explorations Newsletter.** 6(1): 1-6.

Cortes C, and Vapnik V. (1995). Support vector network. **Machine Learning.** 20(3):273–297.

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

Gao M., Hong X., Chen S., and Harris, C. J. (2012). Probability density function estimation based over-sampling for imbalanced two-class problems. In Neural Networks (IJCNN). **The 2012 International Joint Conference on IEEE.** 1-8.

Grubinger T., Kobel C. and Pfeiffer K.P. (2010). Regression tree construction by bootstrap: Model search for DRG-Systems applied to Austrian health-data. **BMC Medical Informatics and Decision Making**. 10: 9-19.

Han J., Pei J., and Kamber M. (2011). Data mining: concepts and techniques. **Elsevier.**

Irimia-Dieguez A.I., Blanco-Oliver A. and Vazquez-Cueto M.J. (2015). A Comparison of Classification/Regression Trees and Logistic Regression in Failure Models. **Procedia Economics and Finance.** 23: 9-14.

Japkowicz N. (2000a). Learning from imbalanced data sets: a comparison of various strategies. **AAAI Tech Report WS-00-05. AAAI.**

Japkowicz N. (2000b). The Class Imbalance Problem: Significance and Strategies. **In Proceedings of the 2000 International Conference on Artificial Intelligence (IC-A1'2000): Special Track on Inductive Learning.** Las Vegas. Nevada.

Japkowicz N., and Stephen S. (2002). The class imbalance problem: A systematic study. **Intelligent Data Analysis.** 6(5):203-231.

Lemon S.C., Roy J. and Clark M.A. (2003). Classification and Regression Tree Analysis in Public Health: Methodological Review and Comparison With Logistic Regression. **Annals of Behavioral Medicine.** 26(3): 172-181.

Orriols-Puig A., and Bernadó-Mansilla E. (2009). Evolutionary rule-based systems for imbalanced data sets. **Soft Computing.** 13(3):213-225.

Weiss G. and Provost F. (2003). Learning when Training Data are Costly: The Effect of Class Distribution on Tree Induction. **Journal of ArtiJicial Intelligence Research.** 19:315-354.

Zhao Y., Li Y., Zhang L. and Wang Q. (2016). Groundwater level prediction of landslide based on classification and regression tree. **Geodesy and Geodynamics.** 7(5): 348-355

ภาคผนวก

บทความวิชาการที่ได้รับการตีพิมพ์เผยแพร่ในระหว่างศึกษา

รายชื่อบทความที่ได้รับการตีพิมพ์เผยแพร่ในระหว่างการศึกษา

ศุภจิตรี บุญอำนวย, รติพร จันทร์กลั่น, กิตติศักดิ์ เกิดประสพ และ นิตยา เกิดประสพ. (2016). **การพยากรณ์จำนวนผู้ป่วยโรคไข้เลือดออกด้วยต้นไม้เชิงจำแนก และเชิงถดถอย.** การประชุมวิชาการระดับชาติ เครือข่ายวิจัยสถาบันอุดมศึกษาทั่วประเทศ ครั้งที่ 11, มหาวิทยาลัยเทคโนโลยีสุรนารี, ประเทศไทย. 19-20 ธันวาคม 2016.

Supajittree Boonamnuay, Nittaya Kerdprasop, and Kittisak Kerdprasop. (2018). **Classification and Regression Tree with Resampling for Classifying Imbalanced Data.** International Journal of Machine Learning and Computing, volume 8, issue 4, pages 336-340. (Indexing: Scopus)

การพยากรณ์จำนวนผู้ป่วยโรคไข้เลือดออกด้วยต้นไม้เชิงจำแนกและเชิงถดถอย

Forecasting number of dengue cases with classification and regression tree

นางสาวศุภจิตรี บุญอำนวย, นางสาวรติพร จันทร์กลั่น, รศ.ดร.กิตติศักดิ์ เกิดประสพ และ รศ.ดร.นิตยา เกิดประสพ

สาขาวิชาวิศวกรรมคอมพิวเตอร์ คณะวิศวกรรมศาสตร์ มหาวิทยาลัยเทคโนโลยีสุรนารี

Email: Eternity_Faith@windowslive.com

บทคัดย่อ

โรคไข้เลือดออกเป็นโรคระบาดที่ในแต่ละปีมีผู้ป่วยจำนวนมาก และเป็นโรคที่อันตรายถึงชีวิต หากไม่ได้รับการดูแลอย่างถูกต้อง โดยการแพร่กระจายของเชื้อไข้เลือดออกเกิดจากพาหะของโรคคือยุง ที่ต้องอาศัยแหล่งน้ำในการขยายจำนวนประชากร ทั้งนี้การวิจัยจึงมุ่งเน้นไปที่การนำข้อมูลสถิติค่าปริมาณน้ำฝนรายเดือนของแต่ละเดือนมาเป็นค่าตั้งต้นในการทำนายหาจำนวนผู้ป่วยในช่วงเดือนนั้น ๆ ว่ามีความสัมพันธ์กันมากน้อยเพียงใด เนื่องด้วยน้ำฝนเป็นสาเหตุทำให้เกิดแหล่งน้ำที่เราควบคุมได้ยาก จึงน่าจะเป็นแหล่งน้ำที่ทำให้เกิดประชากรยุงจำนวนมากที่สุด โดยงานวิจัยนี้ได้นำอัลกอริทึมต้นไม้เชิงจำแนก และเชิงถดถอยมาพยากรณ์จำนวนผู้ป่วยเฉลี่ย จากการแบ่งกลุ่มค่าปริมาณน้ำฝนรายเดือน ซึ่งสามารถคำนวณจำนวนผู้ป่วยเฉลี่ยได้ค่อนข้างแม่นยำกับจำนวนผู้ป่วยจริง

คำสำคัญ: ต้นไม้เชิงจำแนกและเชิงถดถอย, โรคไข้เลือดออก, การพยากรณ์

Abstract

Dengue fever is an epidemic that have many patients each year and the disease is fatal if not taken care of properly. The spread of dengue fever caused by a mosquito which requires water to grow the population. The research is focused on bringing the rainfall statistics of each month data to predict the related number of patients during that month. Cause the rain made it hard to control our water sources. It seems to be a cause the largest number of mosquito population. This research has led algorithms classification and regression tree to classify monthly rainfall to predict the average number of patients each month. At the end the number of patients from prediction actually quite with actual number of patients.

Key word: classification and regression tree, dengue, forecasting

คำนำ

โรคไข้เลือดออกเป็นโรคระบาดที่เป็นปัญหาสาธารณสุขของประเทศไทย ทั้งยังเป็นปัญหาสาธารณสุขทั่วโลก เนื่องจากเป็นโรคระบาดที่มีถึง 4 ชนิด และถึงแม้จะมีวัคซีนป้องกัน หรือการที่ผู้ป่วยเคยเป็นโรคไข้เลือดออกชนิดใด ๆ มาแล้วก็ยังมีภูมิคุ้มกันถาวรเพียงแค่ 1 ชนิด และมีภูมิคุ้มกันต่อชนิดอื่น ๆ เพียงชั่วคราวเท่านั้น โรคไข้เลือดออกจึงเป็นโรคที่ไม่สามารถละเลยที่จะให้ความสำคัญได้ เนื่องจากเป็นโรคระบาดที่มีความรุนแรงถึงขั้นทำให้ผู้ป่วยเสียชีวิตได้ หากไม่ได้รับการดูแลอย่างถูกต้อง ทั้งนี้สามารถเห็นได้จากความนิยมในการทำวิจัยเกี่ยวกับโรคไข้เลือดออกที่มีอย่างต่อเนื่อง ไม่ว่าจะเป็น การนำแผนที่ความเสี่ยงมาช่วยในการคาดการณ์การกระจายตัวทั่วโลกของไข้เลือดออก (Rogers et al., 2014) หรือจะเป็นการสำรวจการใช้ที่ดินในประเทศมาเลเซีย ว่าแต่ละพื้นที่ พบการระบาดของโรคไข้เลือดออกในระดับใด (Cheong et al., 2014) ไปจนกระทั่งการศึกษาว่าประชาชนที่ประเทศออสเตรเลียมีความตระหนักเกี่ยวกับโรคไข้เลือดออก หรือมีความรู้ในการรับมือต่อโรคไข้เลือดออกมากน้อยหรือไม่โดยการสุ่มโทรศัพท์สัมภาษณ์ประชากร (Gyawali et al., 2016) นอกจากนี้ยังมีการศึกษาหลังจากประชากรได้รับวัคซีนไปแล้ว สามารถลดจำนวนผู้ป่วยลงได้จริงหรือไม่ เพื่อที่จะนำไปวางแผนการกระจายวัคซีนต่อไปในอนาคตให้มีประสิทธิภาพมากขึ้น ลดจำนวนผู้ป่วยลงได้จริงหรือไม่ เพื่อที่จะนำไปวางแผนการกระจายวัคซีนต่อไปในอนาคตให้มีประสิทธิภาพมากขึ้น (Gessner et al., 2016) และอีกตัวอย่างงานวิจัยที่ข้อมูลที่ได้มานั้นสำคัญมากคืองานการวิจัยที่ได้ไปสำรวจยอดผู้ป่วยจริง ในประเทศแทบลาตินอเมริกา เพื่อเปรียบเทียบกับรายงานที่องค์กรควบคุมเกี่ยวกับระบาดวิทยาได้รับนั้นพบว่าตัวเลขที่รายงานมีจำนวนน้อยกว่าผู้ป่วยจริงค่อนข้างมาก (Sarti et al., 2016) จึงเกิดเป็นความสนใจขึ้นมาว่าการไปสำรวจ ณ พื้นที่จริงค่อนข้างลำบาก จึงมีความสนใจที่จะศึกษาว่ามีปัจจัยใดที่น่าจะมีผลต่อการระบาด ซึ่งในงานวิจัยนี้ได้นำมาศึกษาคือ ปริมาณน้ำฝน ซึ่งเป็นปัจจัยสำคัญต่อการเพิ่มประชากรของยุงที่ควบคุมได้ยาก โดยหากสามารถนำมาพยากรณ์จำนวนผู้ป่วยล่วงหน้าได้จะมีประโยชน์ในอนาคตต่อการช่วยป้องกันในการเกิดไข้เลือดออก และเตรียมความพร้อมแก่ประชาชน ทั้งนี้ยังเป็นการเตรียมการรองรับและทำการรักษาผู้ป่วยได้อย่างรวดเร็ว เพื่อลดจำนวนผู้ติดเชื้อและผู้เสียชีวิตจากโรคไข้เลือดออก

สำหรับอัลกอริทึมที่นำมาพยากรณ์ในงานวิจัยนี้คือ อัลกอริทึมต้นไม้เชิงจำแนกและเชิงถดถอย (Brieman et al, 1984) เป็นต้นไม้ตัดสินใจแบบไบนารี ที่ถูกสร้างโดยการแยกโหนดเป็นสองโหนดลูกซ้ำแล้วซ้ำอีก โดยเริ่มต้นด้วยโหนดรากที่มีตัวอย่างการเรียนรู้ทั้งหมด สำหรับสัญลักษณ์ตัวแปรต่าง ๆ ที่ใช้ในทฤษฎีเกี่ยวกับอัลกอริทึมนี้มีความหมายดังตารางที่ 1

ตารางที่ 1 สัญลักษณ์ตัวแปร และความหมาย

สัญลักษณ์	ความหมาย
Y	ตัวแปรตาม หรือตัวแปรเป้าหมาย อาจเป็นตัวแปรแบบมีลำดับ หรือไม่มีลำดับ หรือแบบต่อเนื่อง ถ้าหาก Y เป็นข้อมูลประเภทเดียวกันกับคลาส J ค่า C = { 1 , … , J }
X_m , m = 1 , … , M	เซตของตัวแปรทั้งหมด (อาจเรียกเป็นตัวแปรต้น หรือตัวแปรทำนาย) อาจเป็นตัวแปรแบบมีลำดับ หรือไม่มีลำดับ หรือแบบต่อเนื่อง
$\bar{h} = \{ X_n , y_n \}_{n=1}^{N}$	ตัวอย่างการเรียนรู้ทั้งหมด
$\overline{h(t)}$	ตัวอย่างการเรียนรู้ในโหนด t
W_n	น้ำหนักของกรณีที่เกี่ยวข้องกับกรณี n
f_n	น้ำหนักความถี่ของกรณีที่เกี่ยวข้องกับกรณี n
$\pi(j), j = 1 , \dots , J$	ความน่าจะเป็นก่อน Y = j
$p(j,t) , j = 1 , \dots , J$	ความน่าจะเป็นของกรณีคลาส j และโหนด t
$p(t)$	ความน่าจะเป็นของกรณีในโหนด t
$p(j\|t), j = 1 , \dots , J$	ความน่าจะเป็นของกรณีคลาส j ในโหนด t
$C(i\|j)$	ค่าความผิดพลาดของการจำแนกกรณีคลาส j เป็นคลาส i ถ้าจำแนกได้อย่างชัดเจน ค่า $C(i\|j) = 0$

กระบวนการเจริญเติบโตของต้นไม้ คือ การเลือกแยกจากแยกที่เป็นไปได้ทั้งหมดในแต่ละโหนด เพื่อให้เกิดโหนดลูกที่มี "ความบริสุทธิ์มากที่สุด" สำหรับอัลกอริทึมนี้ใช้เพียงตัวแปรทำนายเพียงตัวเดียวในการหาแยกแต่ละครั้ง ถ้า X เป็นตัวแปรอิสระแบบไม่มีลำดับในกลุ่ม I จะมีแยกที่เป็นไปได้ 2I -1 -1 สำหรับตัวแปรทำนายนี้ ถ้า X เป็นตัวแปรอิสระแบบมีลำดับ หรือแบบต่อเนื่อง ซึ่งมีค่าแตกต่างกัน K ค่า จะมีแยกที่เป็นไปได้ K – 1 สำหรับตัวแปรทำนาย X นี้

ต้นไม้จะเติบโตโดยเริ่มจากโหนดราก แล้วทำตามขั้นตอนต่อไปนี้ซ้ำ ๆ ในโหนดแต่ละโหนด

1. ค้นหาแยกที่ดีที่สุดของตัวแปรทำนายแต่ละตัว โดยสำหรับตัวแปรทำนายแบบมีลำดับ หรือแบบต่อเนื่อง ให้ทำการเรียงลำดับจากค่าน้อยที่สุดไปหาค่าที่มากที่สุดก่อนทำการแยกกรณีตามค่าจุดแยก (แทนจุดแยกด้วย v ถ้า $X \leq v$ กรณีนั้นจะไปอยู่โหนดลูกทางด้านซ้าย มิฉะนั้น จะไปทางขวา)
2. ค้นหาแยกที่ดีสุดของแต่ละโหนด จากแยกที่ดีที่สุดที่หาได้จากข้อ 1 เพื่อเลือกอันใดอันหนึ่งที่มีเกณฑ์การแยกสูงที่สุด
3. ทำการแยกโหนดอีกครั้งที่แยกที่ดีที่สุดที่ได้จากข้อ 2 ถ้ายังไม่เป็นตามกฎหยุดที่ระบุไว้

เกณฑ์การแยก และมาตรวัดความไม่บริสุทธิ์ ณ โหนด t แยกที่ดีสุด s เลือกจากแยกที่มีเกณฑ์การแยกสูงที่สุด ซึ่งเมื่อวัดความไม่ บริสุทธิ์แล้วเกณฑ์การแยกมีความสอดคล้องกับการลดลงของความไม่บริสุทธิ์ หรือเรียกว่าค่าการปรับปรุง $\Delta i(s,t)$มีเกณฑ์การแยกที่นิยมใช้อยู่ 3 เกณฑ์ ได้แก่ $Gini , Twoing$ และ $Ordered\ Twoing$ โดยที่โหนด t สามารถหาความน่าจะเป็นของ $p(j,t) , p(t)$ และ $p(j|t)$ได้ดังสมการที่ 1 , 2 และ 3 ตามลำดับ

$$p(j,t) = \frac{\pi(j)N_{w,j}(t)}{N_{w,j}} \quad (1)$$

$$p(t) = \sum_j p(j,t) \quad (2)$$

$$p(j|t) = \frac{p(j,t)}{p(t)} = \frac{p(j,t)}{\sum_j p(j,t)} \quad (3)$$

เมื่อ $N_{w,j} = \sum_{n\in\bar{h}} w_n f_n I(y_n = j)$, $N_{w,j}(t) = \sum_{n\in\overline{h(t)}} w_n f_n I(y_n = j)$

ด้วย I(a = b) คือฟังก์ชั่นการบ่งชี้ค่า โดยให้ค่าเป็น 1 เมื่อ a = b และเป็น 0 ในกรณีอื่น ๆ

เกณฑ์การแยก Gini สามารถวัดความไม่บริสุทธิ์แบบ Gini ที่โหนด t ได้ดังสมการที่ 4

$$i(t) = \sum_{i,j} C(i|j)p(i|t)p(j|t) \quad (4)$$

เกณฑ์การแยก Gini มีค่าการปรับปรุงดังสมการที่ 5

$$\Delta i(s,t) = i(t) - p_L i(t_L) - p_R i(t_R) \quad (5)$$

โดย p_L และ p_R คือความน่าจะเป็นของการส่งกรณีไปยังโหนดลูกทางซ้ายสำหรับ L และทางขวาสำหรับ R ซึ่งประมาณค่าได้ดังนี้ $p_L = p(t_L)/p(t)$ และ $p_R = p(t_R)/p(t)$

เกณฑ์การแยก Twoing มีค่าการปรับปรุงดังสมการที่ 6

$$\Delta i(s,t) = p_L p_R[\sum_j |p(j|t_L) - p(j|t_R)|]^2 \quad (6)$$

เกณฑ์การวัดแบบ Ordered Twoing เกณฑ์การวัดนี้ใช้กับตัวแปร Y ที่เป็นตัวแปรแบบเรียงลำดับเท่านั้น ซึ่งมีขั้นตอนดังนี้

1. แยกคลาส $C = \{1,\dots,J\}$ ของตัวแปร Y เป็น 2 คลาสย่อย คือ C_1 และ C_2 โดย $C_1 = \{1,\dots,j1\}$, $j1 = 1,\dots,J-1$ และ $C_2 = C - C_1$
2. ใช้ทั้ง 2 คลาสเพื่อหาค่า $i(t) = p(C_1|t)p(C_2|t)$ จากนั้นจึงหาค่า $s^*(C_1)$ ที่ทำให้ค่า $\Delta i(s,t)$ มีค่ามากที่สุด ซึ่งเป็นไปตามสมการที่ 7

$$\Delta i(s,t) = i(t) - p_L i(t_L) - p_R i = p_L p_R[\sum_{j\in C_1}\{p(j|t_L) - p(j|t_R)\}]^2 \quad (7)$$

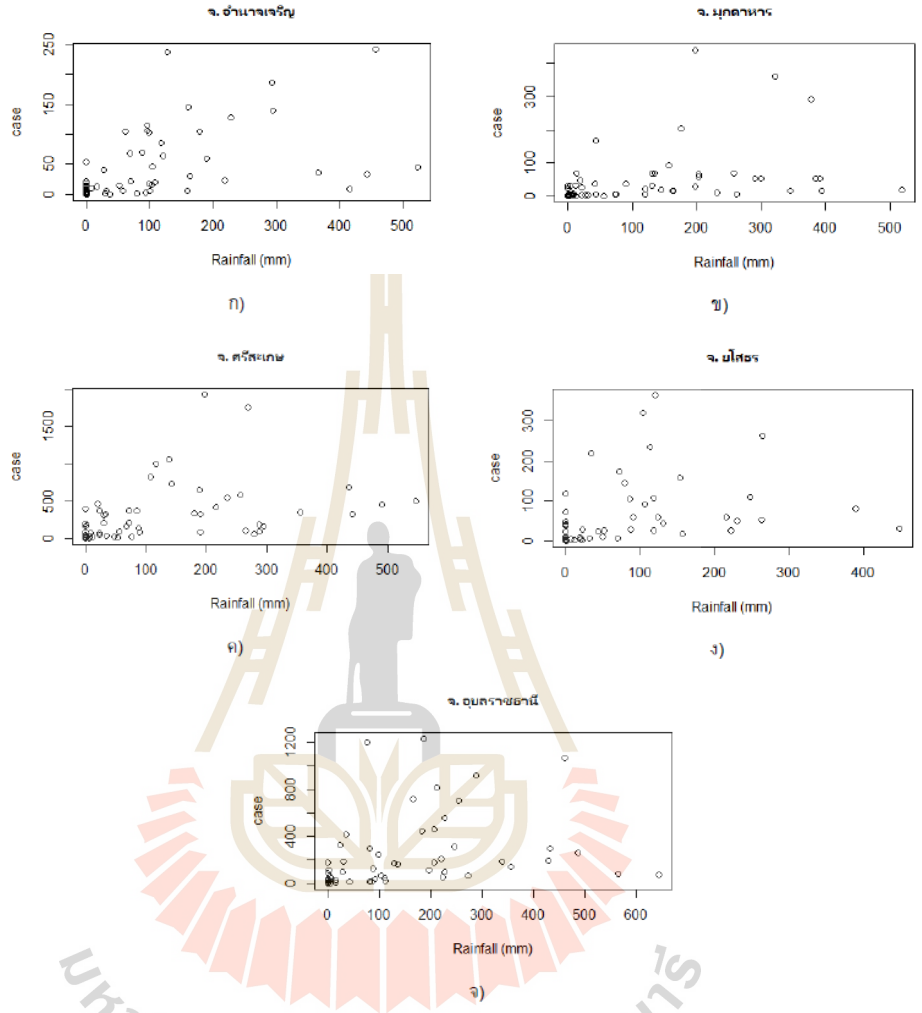
3. หา C_1^* ของคลาส C_1 ที่ทำให้ค่า $\Delta i(S^*(C_1), t$ มีค่ามากที่สุด

กฎการหยุดคอยควบคุมกระบวนการเจริญเติบโตของต้นไม้ โดยมีขั้นตอนดังต่อไปนี้

1. หากโหนดนั้นบริสุทธ์ ซึ่งก็คือทุกกรณีมีค่าขึ้นกับตัวแปรอิสระเหมือนกัน โหนดจะไม่ถูกแยก
2. หากทุกโหนดมีค่าเท่ากันสำหรับทุกตัวแปรทำนาย โหนดจะไม่ถูกแยก
3. หากความลึกของต้นไม้มีค่าถึงขีดจำกัดความลึกสูงสุดที่ระบุไว้ ต้นไม้จะหยุดกระบวนการเติบโต
4. ถ้าขนาดของโหนดน้อยกว่าขนาดขั้นต่ำที่ระบุไว้ โหนดจะไม่ถูกแยก
5. หากแยกโหนดผลลัพธ์ในโหนดลูก แล้วโหนดนั้นมีขนาดเล็กกว่าขนาดโหนดลูกที่ระบุไว้ โหนดนั้นจะไม่ถูกแยก
6. หากแยกที่ดีที่สุด s* ของโหนด t มีค่าการปรับปรุง
   $\Delta I(s^*, t) = p(t)\Delta i(s^*, t)$ น้อยกว่าขั้นต่ำที่ระบุไว้ โหนดนั้นจะไม่ถูกแยก

อุปกรณ์และวิธีการ

สำหรับข้อมูลในการวิจัยประกอบด้วย 2 ชุดข้อมูลคือ ข้อมูลสถิติปริมาณน้ำฝนรายเดือนของรายจังหวัดภาคตะวันออกเฉียงเหนือ จากศูนย์ประมวลวิเคราะห์สถานการณ์น้ำ สำนักงานชลประทานที่ 6 โดยค่าปริมาณน้ำฝนเป็นข้อมูลตัวเลขมีหน่วยเป็นมิลลิเมตร และข้อมูลผู้ป่วยโรคไข้เลือดออกในจังหวัดอำนาจเจริญ จังหวัดมุกดาหาร จังหวัดศรีสะเกษ จังหวัดยโสธร และจังหวัดอุบลราชธานี ในการวิจัยนี้ใช้เพียงจำนวนผู้ป่วยรายเดือนเท่านั้น โดยข้อมูลทั้ง 2 ชุด เป็นข้อมูลประจำปี พ.ศ. 2554 – 2558 รวมแล้วแต่ละจังหวัดจะมีข้อมูลทั้งสิ้น 60 records เมื่อนำข้อมูลมาพล็อตแสดงการกระจายจะได้ดังรูปที่ 1



รูปที่ 1 แสดงการกระจายข้อมูลที่พล็อตแสดงความสัมพันธ์ระหว่างปริมาณน้ำฝนกับจำนวนผู้ป่วยรายเดือนของทั้ง 5 จังหวัด: ก) จังหวัดอำนาจเจริญ ข) จังหวัดมุกดาหาร ค) จังหวัดศรีสะเกษ ง) จังหวัดยโสธร และ จ) จังหวัดอุบลราชธานี

โดยมีขั้นตอนการวิจัยงานวิจัยเป็นดังนี้

1. นำข้อมูลน้ำฝนรายเดือนของจังหวัดต่าง ๆ มาจัดเป็นคู่ลำดับกับจำนวนผู้ป่วยทั้งหมดในเดือนนั้นๆ ที่ได้จากการนับจำนวนผู้ป่วยจากข้อมูลผู้ป่วยโรคไข้เลือดออกจริง
2. จากนั้นนำข้อมูลในข้อ 1 ไปประมวลผลด้วยอัลกอริทึมต้นไม้เชิงจำแนก และเชิงถดถอย เพื่อหาโมเดลการพยากรณ์ ซึ่งในขั้นตอนนี้ได้ทำการแบ่งข้อมูลออกเป็นข้อมูลชุด Train 70% และชุดข้อมูลสำหรับ Test 30% เพื่อนำมาประเมินความถูกต้องของโมเดลในภายหลัง พร้อมทั้งนำผลมาเปรียบเทียบกับการประเมินผลด้วยโดยวิธี K-fold Cross Validation ซึ่งวิธีการนี้เป็นที่นิยมและเหมาะสำหรับความต้องการข้อมูลทดสอบที่มีความหลากหลาย โดยจะแบ่งข้อมูลออกเป็น k ชุดข้อมูล โดยแต่ละชุดมีข้อมูลเป็นจำนวนเท่า ๆ กัน ซึ่งในงานวิจัยนี้กำหนด k เท่ากับ 10 คือแบ่งข้อมูลออกเป็น 10 ชุดข้อมูล และแต่ละชุดข้อมูลมีข้อมูลจำนวน 6 ข้อมูล ว่าโมเดลที่ได้จากการประเมินทั้ง 2 แบบเป็นไปในทิศทางเดียวกันหรือไม่
3. ทำการเขียนผังต้นไม้จากโมเดล แล้วเปรียบเทียบจำนวนผู้ป่วยจริงกับแนวโน้มจำนวนผู้ป่วยที่ได้จากการแบ่งช่วงตามค่าปริมาณน้ำฝน เพื่อดูความสัมพันธ์ของปริมาณน้ำฝนกับจำนวนผู้ป่วยโรคไข้เลือดออก แล้วทำการวัดประสิทธิภาพของโมเดลด้วยค่าทางสถิติ 3 ชนิด ได้แก่ ค่าสัมประสิทธิ์สหสัมพันธ์ (Correlation coefficient, CC, r) , ค่าคลาดเคลื่อนสัมบูรณ์เฉลี่ย (Mean Absolute Error ,MAE) และ ค่ารากที่สองของค่าความคลาดเคลื่อนกำลังสองเฉลี่ย (Root mean Squared Error, RMSE) ซึ่งสามารถคำนวณได้ดังสมการที่ 8 , 9 และ 10 ตามลำดับ

$$r_{xy} = \frac{N\sum XY - (\sum X)(\sum Y)}{\sqrt{[N\sum X^2 - (\sum X)^2][N\sum Y^2 - (\sum Y)^2]}} \quad (8)$$

เมื่อ r_{xy} เป็น ค่าสัมประสิทธิ์สหสัมพันธ์

$\sum X$ เป็น ผลรวมของข้อมูลที่วัดได้จากตัวแปรตัวที่ 1 (X)

$\sum Y$ เป็น ผลรวมของข้อมูลที่วัดได้จากตัวแปรตัวที่ 2 (Y)

$\sum XY$ เป็น ผลรวมของผลคูณระหว่างข้อมูลตัวแปรที่ 1 และ 2

$\sum X^2$ เป็น ผลรวมของกำลังสองของข้อมูลที่วัดได้จากตัวแปรตัวที่ 1

$\sum Y^2$ เป็น ผลรวมของกำลังสองของข้อมูลที่วัดได้จากตัวแปรตัวที่ 2

N เป็น ขนาดของข้อมูล

การบอกระดับหรือขนาดของความสัมพันธ์ จะใช้ตัวเลขของค่าสัมประสิทธ์สหสัมพันธ์ หากค่าสัมประสิทธ์สหสัมพันธ์มีค่าเข้าใกล้ -1 หรือ 1 แสดงถึงการมีความสัมพันธ์กันในระดับสูง แต่หากมีค่าเข้าใกล้ 0 แสดงถึงการมีความสัมพันธ์กันในระดับน้อย หรือไม่มีเลย ส่วนเครื่องหมาย + , - หน้าตัวเลขสัมประสิทธิ์สหสัมพันธ์จะบอกถึงทิศทางของความสัมพันธ์ โดยที่หาก r มีเครื่องหมาย + หมายถึง การมีความสัมพันธ์กันไปในทิศทางเดียวกัน (ตัวแปรหนึ่งมีค่าสูง อีกตัวหนึ่งจะมีค่าสูงไปด้วย) และหาก r มีเครื่องหมาย - หมายถึง การมีความสัมพันธ์กันไปในทิศทางตรงกันข้าม (ตัวแปรหนึ่งมีค่าสูง ตัวแปรอีกตัวหนึ่งจะมีค่าต่ำ)

$$MAE = \frac{1}{N}\sum_{i=1}^{N}|f_i - y_i| \quad (9)$$

โดยที่ f_i = ค่าที่ได้จากการพยากรณ์
y_i = ค่าที่แท้จริงของข้อมูล
N = จำนวนข้อมูล

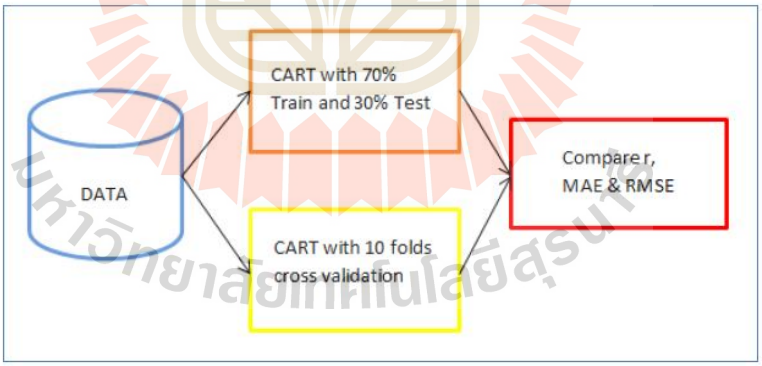
โดยหากค่า MAE มีค่าน้อย แสดงว่าแบบจำลองสามารถประมาณค่าได้ใกล้เคียงกับค่าจริง ดังนั้นหากค่านี้มีค่าเท่ากับศูนย์แล้วจะหมายความว่า ไม่เกิดความคลาดเคลื่อนในแบบจำลองนี้เลย

$$RMSE = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(x_i - \hat{x}_i)^2} \quad (10)$$

โดยที่ x_i = ค่าที่ได้จากการพยากรณ์
$\hat{x}_i$ = ค่าที่แท้จริงของข้อมูล
N = จำนวนข้อมูล

โดยหากค่า RMSE มีค่าน้อยแสดงว่าแบบจำลองสามารถประมาณค่าได้ใกล้เคียงกับค่าจริง ดังนั้นหากค่านี้มีค่าเข้าใกล้ 0 หมายถึง จะมีความแม่นยำมากยิ่งขึ้น

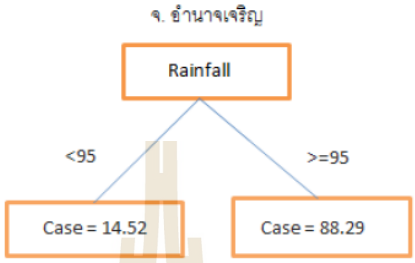
ซึ่งสามารถสรุปขั้นตอนการวิจัยเป็นกรอบแนวคิดการวิจัยโดยสังเขปได้ดังรูปที่ 2



รูปที่ 2 แสดงกรอบแนวคิดการวิจัยโดยสังเขป

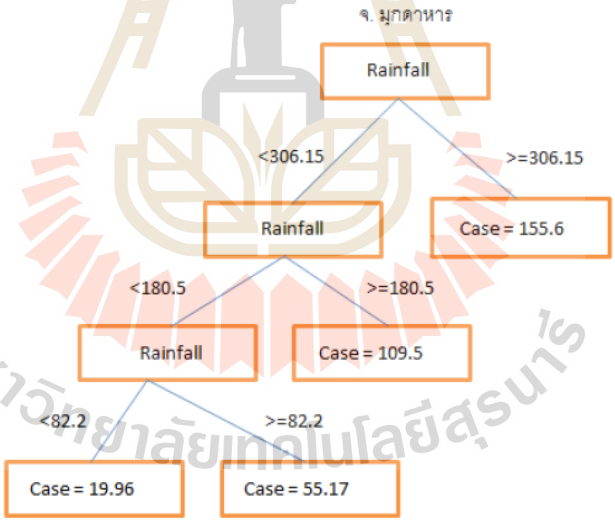
ผลการวิจัย

จากการนำข้อมูลชุด Train มาประมวลผลด้วยอัลกอริทึม ต้นไม้เชิงจำแนกและเชิงถดถอย ได้โมเดลการพยากรณ์ของแต่ละจังหวัดดังนี้



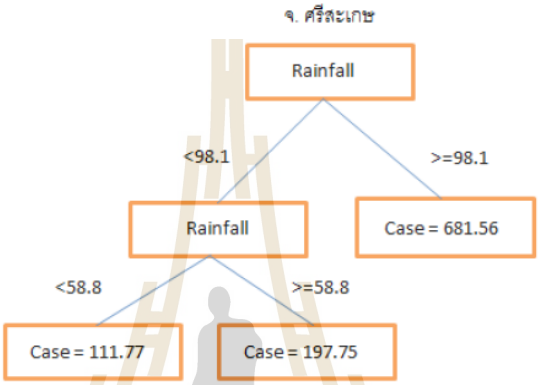
รูปที่ 3 โมเดลต้นไม้เชิงจำแนกและเชิงถดถอยของข้อมูลจังหวัดอำนาจเจริญ

จากรูปที่ 3 แสดงโมเดลต้นไม้เชิงจำแนกและเชิงถดถอยของข้อมูลจังหวัดอำนาจเจริญ ซึ่งจากภาพสามารถบอกได้ว่าหากค่าปริมาณน้ำฝนในเดือนนั้น ๆ มีค่ามากกว่าหรือเท่ากับ 95 มิลลิเมตร จำนวนผู้ป่วยที่พยากรณ์คือ 88.29 ราย และหากค่าปริมาณน้ำฝนมีค่าน้อยกว่า 95 มิลลิเมตร จำนวนผู้ป่วยที่พยากรณ์คือ 14.52 ราย



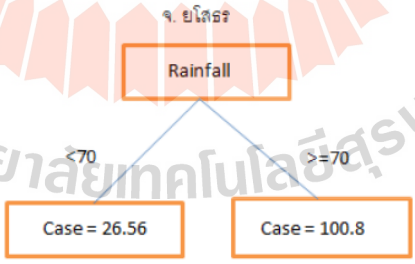
รูปที่ 4 โมเดลต้นไม้เชิงจำแนกและเชิงถดถอยของข้อมูลจังหวัดมุกดาหาร

จากรูปที่ 4 แสดงโมเดลต้นไม้เชิงจำแนกและเชิงถดถอยของข้อมูลจังหวัดมุกดาหาร ซึ่งจากภาพสามารถบอกได้ว่าหากค่าปริมาณน้ำฝนในเดือนนั้น ๆ มีค่ามากกว่าหรือเท่ากับ 306.15 มิลลิเมตร จำนวนผู้ป่วยที่พยากรณ์คือ 155.6 ราย หากค่าปริมาณน้ำฝนอยู่ในช่วงมากกว่าหรือเท่ากับ 180.5 ถึงน้อยกว่า 306.15 มิลลิเมตร จำนวนผู้ป่วยที่พยากรณ์คือ 109.5 ราย หากค่าปริมาณน้ำฝนอยู่ในช่วงมากกว่าหรือเท่ากับ 82.2 ถึงน้อยกว่า 180.5 มิลลิเมตร จำนวนผู้ป่วยที่พยากรณ์คือ 55.17 ราย และหากค่าปริมาณน้ำฝนมีค่าน้อยกว่า 82.2 มิลลิเมตร จำนวนผู้ป่วยที่พยากรณ์คือ 19.96 ราย



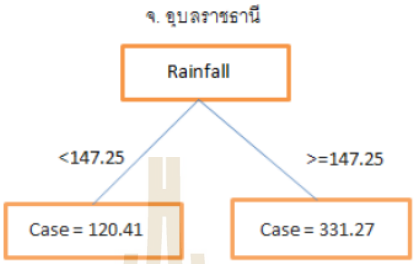
รูปที่ 5 โมเดลต้นไม้เชิงจำแนกและเชิงถดถอยของข้อมูลจังหวัดศรีสะเกษ

จากรูปที่ 5 แสดงโมเดลต้นไม้เชิงจำแนกและเชิงถดถอยของข้อมูลจังหวัดศรีสะเกษ ซึ่งจากภาพสามารถบอกได้ว่าหากค่าปริมาณน้ำฝนในเดือนนั้น ๆ มีค่ามากกว่าหรือเท่ากับ 98.1 มิลลิเมตร จำนวนผู้ป่วยที่พยากรณ์คือ 681.56 ราย หากค่าปริมาณน้ำฝนอยู่ในช่วงมากกว่าหรือเท่ากับ 58.8 ถึงน้อยกว่า 98.1 มิลลิเมตร จำนวนผู้ป่วยที่พยากรณ์คือ 197.75 ราย และหากค่าปริมาณน้ำฝนมีค่าน้อยกว่า 58.8 มิลลิเมตร จำนวนผู้ป่วยที่พยากรณ์คือ 111.77 ราย



รูปที่ 6 โมเดลต้นไม้เชิงจำแนกและเชิงถดถอยของข้อมูลจังหวัดยโสธร

จากรูปที่ 6 แสดงโมเดลต้นไม้เชิงจำแนกและเชิงถดถอยของข้อมูลจังหวัดยโสธร ซึ่งจากภาพสามารถบอกได้ว่าหากค่าปริมาณน้ำฝนในเดือนนั้น ๆ มีค่ามากกว่าหรือเท่ากับ 70 มิลลิเมตร จำนวนผู้ป่วยที่พยากรณ์คือ 100.8 ราย และหากค่าปริมาณน้ำฝนมีค่าน้อยกว่า 70 มิลลิเมตร จำนวนผู้ป่วยที่พยากรณ์คือ 26.56 ราย



รูปที่ 7 โมเดลต้นไม้เชิงจำแนกและเชิงถดถอยของข้อมูลจังหวัดอุบลราชธานี

จากรูปที่ 7 แสดงโมเดลต้นไม้เชิงจำแนกและเชิงถดถอยของข้อมูลจังหวัดอุบลราชธานี ซึ่งจากภาพสามารถบอกได้ว่าหากค่าปริมาณน้ำฝนในเดือนนั้น ๆ มีค่ามากกว่าหรือเท่ากับ 147.25 มิลลิเมตร จำนวนผู้ป่วยที่พยากรณ์คือ 331.27 ราย และหากค่าปริมาณน้ำฝนมีค่าน้อยกว่า 147.25 มิลลิเมตร จำนวนผู้ป่วยที่พยากรณ์คือ 120.41 ราย

ซึ่งเมื่อโมเดลจากชุดข้อมูล Train ของแต่ละจังหวัดมาเปรียบเทียบกับค่าข้อมูลชุด Test ได้ค่าความสัมพันธ์, ค่าคลาดเคลื่อนสัมบูรณ์เฉลี่ย และค่ารากที่สองของค่าความคลาดเคลื่อนกำลังสองเฉลี่ย พร้อมทั้งแสดงค่าทั้ง 3 จากการทำ Cross Validation ดังตารางที่ 2

ตารางที่ 2 แสดงค่าความสัมพันธ์, ค่าคลาดเคลื่อนสัมบูรณ์เฉลี่ย และค่ารากที่สองของค่าความคลาดเคลื่อนกำลังสองเฉลี่ยหลังเปรียบเทียบกับชุดข้อมูล Test ของโมเดลที่ได้จากชุดข้อมูล Train และค่าทั้ง 3 ที่ได้จากการทำ Cross Validation

จังหวัด / ค่าทางสถิติของโมเดล	Train Model			Cross Validation		
	r	MAE	RMSE	r	MAE	RMSE
จ. อำนาจเจริญ	0.4953	33.4054	42.8924	0.1671	42.5833	59.1886
จ. มุกดาหาร	0.2911	39.9162	57.4958	0.273	49.7204	80.6653
จ. ศรีสะเกษ	0.4556	224.2655	292.8118	0.5167	212.4759	334.1284
จ. ยโสธร	0.6489	45.7691	76.6033	0.3488	52.7591	77.1125
จ. อุบลราชธานี	0.5855	233.8387	318.4543	0.2643	199.6204	295.9478

จากตารางที่ 2 จะเห็นได้ว่าโมเดลที่ได้จากการแบ่งข้อมูลเป็น 2 ชุด คือ ชุดข้อมูล Train และชุดข้อมูล Test ให้ประสิทธิภาพในการทำนายเป็นไปตามแนวโน้มเดียวกับโมเดลที่ได้จากการทำ Cross Validation ที่เกิดจากการใช้ชุดข้อมูลที่หลากหลาย ซึ่งแสดงให้เห็นว่าโมเดลที่ได้จากการทำ Train-Test มีความน่าเชื่อถือ ทนทานต่อข้อมูลที่ผิดปกติ สามารถนำมาใช้ในการพยากรณ์ได้จริง

อภิปรายและสรุปผลการวิจัย

จากผลการวิจัยข้างต้นอัลกอริทึมสามารถพยากรณ์จำนวนผู้ป่วยจากค่าปริมาณน้ำฝนได้ ซึ่งผลการวิจัยก็เป็นไปตามแนวโน้มที่ตั้งสมมติฐานไว้ว่าปริมาณน้ำฝนมีผลต่อการระบาดของโรคไข้เลือดออก โดยค่าปริมาณน้ำฝนที่เพิ่มมากขึ้นนั้น ทำให้จำนวนผู้ป่วยโรคไข้เลือดออกเพิ่มมากขึ้น และจากการสังเกตค่าคลาดเคลื่อนสัมบูรณ์เฉลี่ย และค่ารากที่สองของค่าความคลาดเคลื่อนกำลังสองเฉลี่ยของของข้อมูลบางจังหวัดสูงมาก เนื่องจากข้อมูลที่ได้ทำการรวบรวมมานั้น บางปีมีผู้ป่วยเยอะกว่าปีอื่น ๆ ที่ได้เก็บสถิติมาแบบผิดปกติ ทำให้โมเดลเกิดค่าความคลาดเคลื่อนมาก ด้วยโมเดลพยายามที่จะคลอบคลุมข้อมูลให้ได้มากที่สุด ทั้งนี้หวังว่างานวิจัยนี้จะเป็นแนวคิดต่อยอดงานวิจัยอื่น ๆ ในอนาคตอย่างเช่นการนำค่าดัชนีอื่น ๆ มาร่วมพยากรณ์ เช่น อุณหภูมิ เพื่อให้ได้ความแม่นยำ และหลากหลายต่อโมเดลมากขึ้น

เอกสารอ้างอิง

Breiman, L., Friedman, J.H., Olshen, R., & Stone, C.J. (1984). Classification and Regression Tree. Wadsworth & Brooks/Cole Advanced Books & Software, Pacific California.

Cheong, Y.L., Leitao, P.J., & Lakes, T. (2014). Assessment of land use factors associated with dengue cases in Malaysia using Boosted Regression Trees. Spatial and Spatio-temporal Epidemiology, 10, 75-84.

Gessner, B.D., & Wilder-Smith, A. (2016). Estimating the public health importance of the CYD-tetravalent dengue vaccine: Vaccine preventable disease incidence and numbers needed to vaccinate. Vaccine, 34(20), 2397-2401.

Gyawali, N., Bradbury, R.S., & Taylor-Robinson, A.W. (2016). Knowledge, attitude and recommendations for practice regarding dengue among the resident population of Queensland, Australia. Asian Pacific Journal of Tropical Biomedicine, 6(4), 360-366.

Rogers, D.J., Suk, J.E., & Semenza, J.C. (2014). Using global maps to predict the risk of dengue in Europe. Acta Tropical, 129, 1-14.

Sarti, E., L'Azou, M., Mercado, M., Kuri, P., Siqueira, J.B., Jr., Solis, E., et al. (2016). A comparative study on active and passive epidemiological surveillance for dengue in five countries of Latin America. International Journal of Infections Diseases, 44, 44-49.

Classification and Regression Tree with Resampling for Classifying Imbalanced Data

Supajittree Boonamnuay, Nittaya Kerdprasop, and Kittisak Kerdprasop

***Abstract*— Data mining is the automatic process to find from data interesting and useful patterns for specific tasks such as predicting future data or classifying label or group of the new data items. Many data mining algorithms successfully applied to several real-life data are in a tree group. Among the tree-based algorithms, decision tree is the most popular and renowned one for its high accuracy on classifying data in general cases in which data in each class are quite equally distributed. But many datasets in real applications are imbalanced; amount of data in some group outnumber those in other group. Such uneven distribution among classes is a main reason why classification accuracy is not excellent even when using decision tree algorithm. Inefficiency is due to the case that in the tree growing phase, the algorithm tends to favor the majority data and ignores the minority data to be incorrectly classified. In the past many researchers try to solve this data imbalanced problem with many ways like over-sampling, under-sampling, cost-sensitive classification, or even ensemble of cost-sensitive decision tree. In this paper, we introduce a simplified method of learning classification and regression tree (CART) with resampling technique for classifying imbalanced datasets. We compare our proposed method with other methods based on several metrics including the precision on classifying the minority data as opposed to the classification on majority data, the overall accuracy regardless of minority nor majority classes, and the Matthews Correlation Coefficient (MCC). The use of MCC is suitable for imbalanced data because it takes into account all four classifying metrics: true positive, true negative, false positive, and false negative. The performance of our proposed method to combine resampling with CART is satisfied based on the MCC metric. From all five experimental imbalanced datasets, our method performs the best.**

***Index Terms*— Classification and regression tree, CART, Resampling technique, Imbalanced data, Matthews Coefficient Correlation.**

I. Introduction

Data mining is a kind of data-oriented discovery science to find the patterns, important indexes or relationships from the existing databases [1]. There are many types of data mining tasks such as data classification, association rule mining, clustering, and forecasting. Among numerous potential tasks of knowledge discovery, data classification is the majority of data mining task that has been widely applied in many real applications and it is also the main mining task of our focus.

We especially aim at studying a classification problem of accurately partitioning and predicting data with imbalanced distribution among classes. This is called learning from imbalanced data [2], [3]. Typically, imbalanced data classification refers to a class of classification problems where the class distributions are not represented equally [2]. For example, suppose we have a 2-class (or binary) classification problem with 100 data instances (or rows) in total. Among these, 80 instances are labeled with class *a*, while the remaining 20 instances are labeled with class *b*. This is an example of imbalanced dataset and the ratio of data instances in class *a* to those in class *b* is 80:20, or the imbalanced ratio equals 4:1.

To deal with class imbalance, the most intuitive solution is to rebalance data with either under sampling the majority data, or over sampling data in the minority class [4], [5]. In this work, we are interested in balancing data with the bootstrapping method using resampling technique.

Our specific emphasis is on balancing data for a tree-based learning method. Tree learning is widely accepted for its easy interpretation nature. There exist several research trying to improve accuracy of tree-based learning over imbalanced data such as the work of Bartosz Krawczyk and teammates [6]. Their research introduced cost-sensitive decision tree ensembles for effective classification of imbalanced data. They tried to improve decision tree accuracy by assigning different weight to data in different imbalanced ratios.

Efficient learning from imbalanced data is still a challenging problem because imbalance is common in many applications [7], [8], [9]. Most classification datasets do not have exactly equal number of instances in each class, but a small difference often does not matter. There are, however, problems where a class imbalance is not just ignorance; it is the main concern of the application. For example, in commerce datasets like those that characterize fraudulent transactions are imbalanced. The vast majority of the transactions will be in the "Not-Fraud" class and a very small but important minority will be in the "Fraud" class.

We focus our concern regarding imbalanced data with the classification and regression tree (CART) method. Our special interest in CART algorithm is because this algorithm can classify all types of target data including both categorical and numeric. This algorithm has also been reported by many researchers that it yields good results. In medical domain [10], this algorithm can help efficient diagnosis based on patients' symptom. In economy [11], CART algorithm can help deciding the way to manage

Manuscript received September 27, 2017. This work was supported by grants from Suranaree University of Technology through the funding of Data Engineering Research Unit and the Knowledge Engineering Research Unit.

S. Boonamnuay is a master student with the School of Computer Engineering, Suranaree University of Technology (SUT), 111 University Avenue, Muang, Nakhon Ratchasima 30000, Thailand. (corresponding author: +66892865318; e-mail: eternity_faith@windowslive.com).

N. Kerdprasop is an associate professor with the School of Computer Engineering and head of Data Engineering Research Unit, SUT. (e-mail: nittaya@sut.ac.th).

K. Kerdprasop is an associate professor with the School of Computer Engineering and head of Knowledge Engineering Research Unit, SUT. (e-mail: kerdpras@sut.ac.th).

business plans. Also in environmental application [12], this algorithm can help to predict rainfall and groundwater level.

II. Background Theories

A. Classification and Regression Tree

Classification and regression tree, or CART, is a classification method that builds a model from historical data. CART was firstly developed by Breiman, Freidman, Olshen and Stone in 1984 [13]. A CART tree is a binary decision tree in that in constructs a tree by splitting a node in half repeatedly resulting in two child nodes for each split. Tree construction begins with the root node that contains the whole learning samples. If data in the node are of mixing classes, that node has to be split. Splitting strategy is that the algorithm will search for all possible variables and all possible values in order to find the best split such that data in child nodes are of maximum homogeneity, or sometimes called purity.

The key idea of CART is recursive partitioning. The process of CART begins by taking all data for the consideration of all possible values of all variables for growing a tree. So it will select on variable or value that produces the best separation in the target attribute. If the value in focus is lower than the value at the separate point, that value will be placed on the left side of tree. For the value greater than or equal to the value at separate point, it will be sent to right side of tree like, as shown by example on Fig. 1. The tree will repeat this splitting process until it cannot find another best separate point the give the increase purity greater than the last separate point. The pseudocode of this tree growing process is illustrated in Fig. 2.

The index that used for checking the best separate point in the pseudocode is Gini index that can be computed as in equation (1). Gini is a measure of impurity computed by counting the frequency of events that how often a randomly chosen data instance is wrongly labeled, given that that instance is to be randomly labeled based on distribution of class labels. For a binary classification with class positive and negative, p_{pos} is the probability that data instance in class positive being chosen, and $(1\text{-}p_{pos})$ is the probability that that instance is incorrectly labeled as negative. The other term can be interpreted in the same way with the class label negative, instead of positive.

$$Gini\ index = p_{pos}(1\text{-}p_{pos}) + p_{neg}(1\text{-}p_{neg}) \quad (1)$$

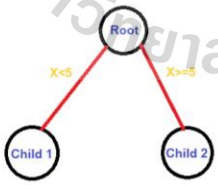


Fig. 1. Example CART model.

Classification and Regression Tree

1. Start at the root node.
2. For each ordered variable X,
   convert it to an unordered variable X' by grouping its values in the node into a small number of intervals
   if X is unordered, then set X' = X.
3. Perform a chi-squared test of independence of each X' variable versus Y on the data in the node and compute its significance probability.
4. Choose the variable X∗ associated with the X' that has the smallest significance probability.
5. Find the split set {X∗ ∈ S∗} that minimizes the sum of Gini indexes and use it to split the node into two child nodes.
6. If a stopping criterion is reached, exit.
   Otherwise, apply steps 2–5 to each child node.
7. Prune the tree with the CART method.

Fig. 2. The pseudocode of classification and regression tree.

B. Resampling

Bootstrap is a general purpose resampling technique for obtaining estimates of properties of statistical estimators without making assumptions about the distribution of the data [14]. This resampling method is often used to find

(1) standard errors of estimates,
(2) confidence intervals for unknown parameters, or
(3) p values for test statistics under a null hypothesis

Suppose Y has a cumulative distribution function, then F(y) = P(Y<=y). If we have a sample of size n from F(y), Y1, Y2, . . . , Yn, then the steps in computing resamples are as follows:

Step 1. Repeatedly simulate sample of size n from F.
Step 2. Compute statistic of interest.
Step 3. Study behavior of statistic over B repetitions.

Pretend that Fn(y) is the original distribution of F(y), sampling from Fn(y) is thus equivalent to sampling with replacement from originally observed Y1, ..., Yn.

TABLE I: Confusion Matrix for Two Class Classification

		Actual Data	
		Positive	Negative
Predicted Data	Positive	*TP*	*FP*
	Negative	*FN*	*TN*

C. Classification Performance Evaluation

To evaluate each classification technique, we use the accuracy metric for assessing their overall performance. The computation of this metric is based on the values in confusion matrix [15] as shown in Table I and accuracy can compute as in equation (2).

$$Accuracy = \frac{(TP + TN)}{(TP+FN+FP+TN)} \quad (2)$$

where:

TP is the number of actual data from positive class and the model can correctly predict that data to be in a positive class,

TN is the number of actual data from negative class and the model can correctly predict that data to be in a negative class,

FP is the number of actual data from negative class but the model incorrectly predicts that data to be in a positive class,

FN is the number of actual data from positive class but the model predicts that the data incorrectly as in a negative class.

In our experiments, we also evaluate classification by class. This measurement is called precision and its computation is in equation (3). For the case of classifying data with imbalanced distribution among classes, many researchers [16], [17] use Matthews Correlation Coefficient (MCC) as an effective metric for fair comparison because the MCC computes performance based on all values in the confusion matrix. MCC computation is shown in equation (4).

$$Precision_{Positive} = \frac{TP}{TP + FP}$$

$$Precision_{Negative} = \frac{TN}{TN + FN} \quad (3)$$

$$MCC = \frac{TP \times TN - FP \times FN}{\sqrt{(TP + FP)(TP + FN)(TN + FP)(TN + FN)}} \quad (4)$$

III. Materials and Experimentation

A. Datasets and Methods

In this research, we use standard imbalanced datasets that publicly available for download from KEEL Repository [18]. The five datasets and their details are summarized in Table II. All datasets are two classes. We show class distribution as proportion of data instances in minority class to those in the majority class. The imbalanced ratios (IR) are computed as the fraction of instances in majority to instances in minority class. IR equals to 1 means the data are well balance. The higher IR infers the more imbalanced situation among classes.

Our research methodology is that firstly classifying the selected datasets using the tree-based algorithms including decision tree induction, CART, AdaBoost, and bagging of decision trees. Then perform on the same datasets our proposed method of decision tree learning using CART with the applied resampling technique.

TABLE II: Data sets used in the experiments

Dataset	Features	Objects	No. Classes	Class distribution	IR
Pima	8	768	2	268:500	1.87
Yeast	8	1484	2	429:1055	2.46
Vehicle	18	846	2	199:647	3.25
Segment	19	2308	2	329:1979	6.02
Page-blocks	10	5472	2	559:4913	8.79

B. Experimental Setup

To test the performance of the proposed method (CART + resampling), we compare it against the other four techniques (decision tree, CART, AdaBoost, and Bagging). Comparative performance metrics are precision by class, overall accuracy, and MCC. These metrics are computed from the confusion matrix that to be obtained by running a classifier model ten times using the 10-fold cross validation method, which is conceptually shown in Fig. 3. We use 10-folds cross validation because we want to fairly compare the models using every data instance as train and test data in every model comparison. To apply 10-fold cross validation, we have to separate our dataset into 10 parts and repeat the experiment 10 times. At round one, we use parts 1-9 as training set and use part 10 as test set. Then in round two, we use parts 1-8 and 10 as training set and use part 9 as test set. We repeatedly do it in such manner 10 times and average, precision, accuracy, and MCC values from that 10 rounds to compare performance of each classification technique.

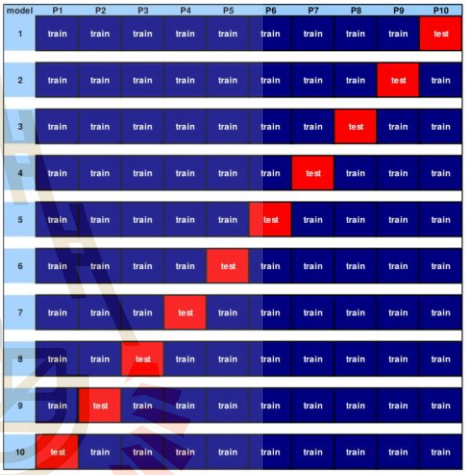
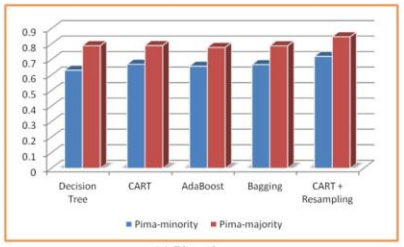


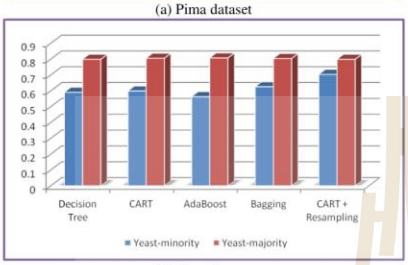
Fig. 3. Building and testing a model with10-folds cross validation.

IV. Experimental Results and Discussions

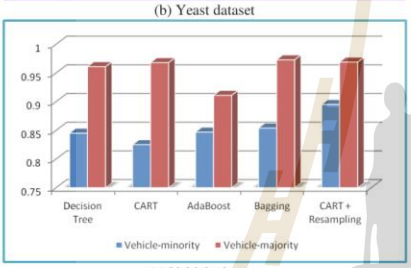
To observe classification performance of the proposed CART + resampling method on imbalanced data, we firstly analyze its precision on classifying minority cases as compared to the classification on the majority cases. The results are shown in Fig. 4. It can be noticed from the results that resampling can help improving classifying data in the minority class as well as yielding good precision on the majority class. It is true in almost all datasets, except the segment dataset that even though resampling can improve the precision of classifying majority cases, the precision on classifying minority cases is still low, comparative to the bagging technique. From observing precision on predicting minority and majority classes, we can conclude that CART + resampling performs well on four out of five datasets.



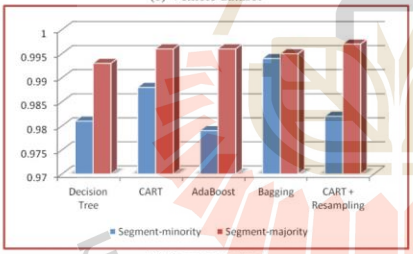
(a) Pima dataset



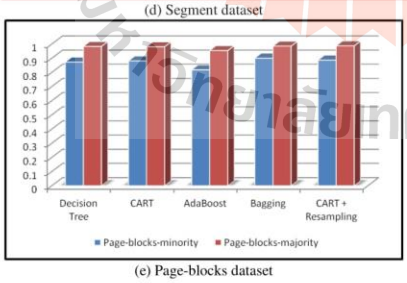
(b) Yeast dataset



(c) Vehicle dataset



(d) Segment dataset



(e) Page-blocks dataset

Fig. 4. Precision by class and by dataset of the CART + Resampling method comparative to other classification methods.

TABLE III: OVERALL ACCURACY FOR EACH CLASSIFICATION TECHNIQUE ON DIFFERENT DATASETS

Dataset	Decision Tree	CART	AdaBoost	Bagging	CART + Resampling
Pima	73.8281	75.3906	74.349	75.2604	**78.5156**
Yeast	75.2022	75.7412	74.3935	76.5499	**80.6604**
Vehicle	93.2624	93.1442	89.8345	94.3262	**95.3901**
Segment	99.1334	**99.4801**	99.3934	**99.4801**	**99.4801**
Page-blocks	97.2222	97.1491	94.3896	**97.6608**	97.6425

TABLE IV: MATTHEWS CORRELATION COEFFICIENTS FOR EACH CLASSIFIER ON EACH DATASET

Dataset	Decision Tree	CART	AdaBoost	Bagging	CART + Resampling
Pima	0.417	0.444	0.417	0.441	**0.566**
Yeast	0.360	0.379	0.360	0.391	**0.507**
Vehicle	0.815	0.817	0.704	0.847	**0.866**
Segment	0.964	**0.979**	0.975	**0.979**	**0.979**
Page-blocks	0.847	0.841	0.660	0.870	**0.871**

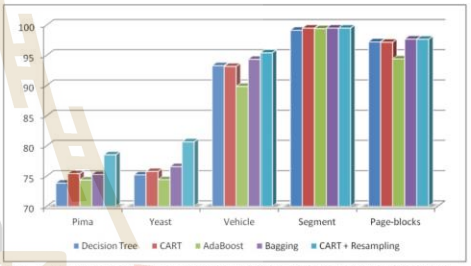


Fig. 5. Overall classification accuracy of the studied CART + Resampling method compared against other four methods

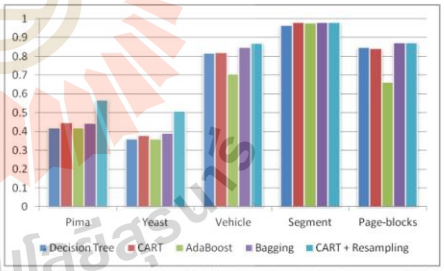


Fig. 6. Class imbalanced classification based on the Matthew correlation coefficient metric of each classification method on different datasets.

We then check the overall accuracy performance of the CART + resampling method. The results are summarized and illustrated in Table III. The graphical comparison is also shown in Fig. 5. By ignoring importance of minority versus majority and just evaluate the overall predictive accuracy, we find that our proposed method performs almost the best in every dataset, except in the page-blocks dataset that

bagging method is a little bit better with insignificant difference.

To take into account both precision on predicting minority and majority classes as well as penalty on misclassification, we compare the models' performance with the MCC metric. The results are illustrated in Table IV and graphically shown in Fig. 6. When consider both correct and incorrect classification cases, we can now clearly see the power of CART + resampling method as it performs the best in every dataset.

V. Conclusion

To improve the performance of imbalanced data classification using tree learning algorithms, we can apply the method or technique to improve accuracy by resample the datasets. For classifying categorical and numerical data, classification and regression tree (CART) algorithm is renowned for better classifying imbalanced data than normal decision tree. We propose in this work that we can further improve the performance of CART by handling the imbalanced data through the resampling technique.

The experimental results show that our proposed technique can help improving classification performance when several measurements including precision by class, overall accuracy, and Matthews Correlation Coefficient (MCC). The MCC metric is the most discriminative measurement confirming the power of resampling when applied to the CART algorithm. This method has been proven work well on datasets with numerous imbalanced ratios (IR); in our experiments the IR ranges from 1.87 up to 8.79. We notice that the CART + resampling is extremely powerful when IR is lower than 4.

For future work, we plan to further our investigation that how much data is enough to effectively represent the whole dataset. Such knowledge can facilitate our application of bootstrapping for under-sampling and over-sampling as well.

References

[1] J. Han and M. Kamber, *Data Mining: Concepts and Techniques*, Morgan Kaufmann, 2006.

[2] N. Chawla. "Data mining for imbalanced datasets: An overview," in O. Maimon and L. Rokash (Eds) *Data Mining and Knowledge Discovery Handbook*, pp. 875-886, Springer, 2010.

[3] B. Krawczyk, "Learning from imbalanced data: open challenges and future directions," Progress in Artificial Intelligence, vol. 5, no. 4, pp. 221-232, 2016.

[4] N. Chawla, K. Bowyer, L. Hall, and W. Kegelmeyer, "SMOTE: Synthetic minority over-sampling technique," *J Artif Int Res*, vol. 16, no. 1, pp. 321-357, 2002.

[5] R. Dubey, J. Zhou, Y. Wang, P. Thompson, and J. Ye, "Analysis of sampling techniques for imbalanced data: An N=648 ADNI study," *Neuroimage*, vol. 87, pp. 220-241, 2014.

[6] B. Krawczyk, M. Wozniak, and G. Schaefer, "Cost-sensitive decision tree ensembles for effective imbalanced classification," *Applied Soft Computing*, vol. 14, pp. 554-562, 2014.

[7] P. Gutierrez, M. Lastra, J. Benitez, and F. Herrera, "SMOTE-GPU: Big Data preprocessing on commodity hardware for imbalanced classification," *Progress in Artificial Intelligence*, 2017, Available: https://doi.org/10.1007/s13748-017-0128-2.

[8] S. Pouyanfar and S. Chen, "Automatic video event detection for imbalanced data using enhanced ensemble deep learning," *Int J of Semantic Computing*, vol. 11, no. 1, 2017, Available: https://doi.org/10.1142/S1793351X17400050.

[9] F. Li, S. Li, C. Zhu, X. Lan, and H. Chang, "Cost-effective class-imbalance aware CNN for vehicle localization and categorization in high resolution aerial images," *Remote Sensing*, vol. 9, no. 6, 2017, Avaliable: https://doi.org/10.3390/rs9050494.

[10] S. C. Lemon, J. Roy, and M. A. Clark, "Classification and regression tree analysis in public health: Methodological review and comparison with logistic regression," *Annals of Behavioral Medicine*, vol. 26, no. 3, pp. 172-181, 2003.

[11] A. I. Irimia-Dieguez, A. Blanco-Oliver, and M. J. Vazques-Cueto, "A comparison of classification/regression trees and logistic regression in failure models," *Procedia Economies and Finance*, vol. 23, pp. 9-14, 2015.

[12] Y. Zhao, Y. Li, L. Zhang, and Q. Wang, "Groundwater level prediction of landslide based on classification and regression tree," *Geodesy and Geodynamics*, vol. 7, no. 5, pp. 348-355, 2016.

[13] L. Breiman, J. H. Friedman, R. Olshen, and C. J. Stone, *Classification and Regression Tree*. Wadsworth & Brooks/Cole Advanced Books & Software, Pacific California, 1984.

[14] A. Benner. (October 2016). *Resampling and the Bootstrap* [online]. Available: http://www.bioconductor.org/workshop/203/NGFN03/resamplig.pdf.

[15] D. M. Powers, "Evaluation: from precision, recall and F-measure to ROC, informedness, markedness and correlation," *Journal of Machine Learning Technologies*, vol. 2, no. 1, pp. 37-63, 2011.

[16] B. W. Matthews, "Comparison of the predicted and observed secondary structure of T4 phage lysozyme," *Biochimica et Biophysica Acta (BBA) – Protein Structure*, vol. 405, no. 2, pp. 442-451, 1975.

[17] S. Boughorbel, F. Jarray, and M. El-Anbari, "Optimal classifier for imbalanced data using Matthews Correlation Coefficient metric," *PLoS ONE*, vol. 12, no. 6, article e0177678, June 2017. Available: https://doi.org/10.1371/journal.pone.0177678.

[18] J. Alcala-Fdez, A. Ferndedez, J. Luengo, J. Derrac, S. Garcia, L.Sanchez, and F. Herrera, "Keel data-mining software tool: data set repository integration of algorithms and experimental analysis framework," *Journal of Multiple-Valued Logic and Soft Computing*, vol. 17, no. 2-3, pp. 255-287, 2011.

S. Boonamnuay is currently a master student with the School of Computer Engineering, Suranaree University of Technology, Thailand. She received her bachelor degree in Computer Engineering from Suranaree University of Technology, Thailand, in 2015. Her current research of interest includes data mining, classification and regression tree, and imbalanced data classification.

Kittisak Kerdprasop is an associate professor at the School of Computer Engineering and Chair of the School. He is also the head of Knowledge Engineering Research Unit, Suranaree University of Technology, Thailand. He received his bachelor degree in Mathematics from Srinakarinwirot University, Thailand, in 1986, master degree in computer science from the Prince of Songkla University, Thailand, in 1991 and doctoral degree in computer science from Nova Southeastern University, U.S.A., in 1999. His current research includes Machine Learning, Artificial Intelligence and Probabilistic Knowledge Bases.

Nittaya Kerdprasop is an associate professor and the head of Data Engineering Research Unit, School of Computer Engineering, Suranaree University of Technology. She received her B.S. in radiation techniques from Mahidol University, Thailand, in 1985, M.S. in computer science from the Prince of Songkla University, Thailand, in 1991 and Ph.D. in computer science from Nova Southeastern University, U.S.A., in 1999. Her research of interest includes Knowledge Discovery in Databases, Data Mining, Artificial Intelligence, Logic and Constraint Programming, Deductive and Active Databases.

ประวัติผู้เขียน

นางสาวศุภจิตรี บุญอำนวย เกิดเมื่อวันที่ 7 พฤศจิกายน 2535 ที่อำเภอเมือง จังหวัดนครราชสีมา เริ่มเข้าศึกษาชั้นประถมศึกษาปีที่ 1 ที่โรงเรียนพระหฤทัย อำเภอเมือง จังหวัดเชียงใหม่ แล้วจึงย้ายมาศึกษาต่อชั้นประถมศึกษาปีที่ 3 ถึง 6 ที่โรงเรียนอนุบาลลำปลายมาศ อำเภอลำปลายมาศ จังหวัดบุรีรัมย์ จากนั้นศึกษาต่อในระดับมัธยมศึกษาตอนต้นและตอนปลายที่โรงเรียนบุรีรัมย์พิทยาคม อำเภอเมือง จังหวัดบุรีรัมย์ ในปีการศึกษา 2554 ได้เข้าศึกษาระดับปริญญาตรีในสาขาวิชาวิศวกรรมคอมพิวเตอร์ มหาวิทยาลัยเทคโนโลยีสุรนารี และสำเร็จการศึกษาเมื่อปี 2558 หลังจากนั้นในปีการศึกษา 2558 ได้เข้าศึกษาต่อในระดับปริญญาโทในสาขาวิชาวิศวกรรมคอมพิวเตอร์ มหาวิทยาลัยเทคโนโลยีสุรนารี

ในระหว่างการศึกษาได้รับการอนุเคราะห์เป็นอย่างดีจากอาจารย์ที่ปรึกษาและอาจารย์ประจำวิชาต่าง ๆ และได้รับการตีพิมพ์เผยแพร่บทความวิชาการซึ่งรายละเอียดสามารถดูได้ที่ภาคผนวก