Research Article

Received: July 3, 2018; Accepted: July 16, 2018

เทคนิคการจำแนกข้อมูลที่พัฒนาสำหรับชุดข้อมูลที่ไม่สมดุลของภาวะข้อเข่าเสื่อมในผู้สูงอายุ

Improved Classification Techniques for Imbalanced Data Sets of Elderly's Knee Osteoarthritis

พุทธิพร ธนธรรมเมธี และเยาวเรศ ศิริสถิตย์กุล*

หลักสูตรวิศวกรรมซอฟต์แวร์ สำนักวิชาสารสนเทศศาสตร์ มหาวิทยาลัยวลัยลักษณ์

ตำบลไทยบุรี อำเภอท่าศาลา จังหวัดนครศรีธรรมราช 80160

Putthiporn Thanathamathee and Yaowarat Sirisathitkul*

Software Engineering Program, School of Informatics, Walailak University,

Thaiburi, Thasala, Nakhon Si Thammarat Thailand 80160

บทคัดย่อ

งานวิจัยนี้มีวัตถุประสงค์เพื่อพัฒนาตัวแบบการพยากรณ์ภาวะข้อเข่าเสื่อมในผู้สูงอายุ จากข้อมูลแบบบันทึกการประเมินข้อเข่าเสื่อม โรงพยาบาลส่งเสริมสุขภาพตำบลบ้านหาร อำเภอท่าศาลา จังหวัดนครศรีธรรมราช จำนวน 370 เรคอร์ด และมีข้อมูล 4 คลาส ได้แก่ คลาส 0 ยังไม่พบอาการผิดปกติ 200 เรคอร์ด คลาส 1 เริ่มมีอาการข้อเข่าเสื่อม 115 เรคอร์ด คลาส 2 มีอาการโรคข้อเข่าเสื่อมระดับปานกลาง 39 เรคอร์ด และคลาส 3 เป็นโรคข้อเข่าเสื่อมระดับรุนแรง 16 เรคอร์ด สำหรับการวินิจฉัยทางการแพทย์ ข้อมูลกลุ่มน้อย คือ ข้อมูลที่สนใจและการจำแนกผิดพลาดเกิดขึ้นได้สูงกว่าข้อมูลกลุ่มมาก ซึ่งข้อมูลชุดนี้มีจำนวนรวมของคลาส 0 และคลาส 1 สูงกว่าคลาส 2 และคลาส 3 เป็นจำนวนมาก จึงเกิดความไม่สมดุลของข้อมูล ส่งผลให้การจำแนกข้อมูลผิดพลาดได้ การปรับความไม่สมดุลของข้อมูลคลาส 2 และคลาส 3 ทำได้ด้วยเทคนิคการปรับเพิ่มข้อมูลด้วยวิธีสุ่ม โดยใช้วิธี ADASYN และ SMOTE และใช้งานวิธีการตรวจสอบไขว้แบบ 10 กลุ่ม ในการแบ่งเป็นชุดข้อมูลสอนและชุดข้อมูลทดสอบ จากนั้นจำแนกข้อมูลด้วย multi-class imbalanced data classification ด้วยวิธี one-vs-one และ one-vs-all และเทคนิค Gentleboost ผลการทดสอบประสิทธิภาพพบว่าวิธี ADASYN และ one-vs-one ให้ค่าความถูกต้อง 97.31 % และทดสอบตัวแบบกับชุดข้อมูลจริงที่ไม่สมดุลจาก โรงพยาบาลส่งเสริมสุขภาพตำบลบ้านหัวคู อำเภอท่าศาลา จังหวัดนครศรีธรรมราชจำนวน 232 เรคอร์ด และมีข้อมูล 4 คลาส ได้แก่ คลาส 0 ยังไม่พบอาการผิดปกติ 141 เรคอร์ด คลาส 1 เริ่มมีอาการข้อเข่าเสื่อม 63 เรคอร์ด คลาส 2 มีอาการโรคข้อเข่าเสื่อมระดับปานกลาง 16 เรคอร์ด และคลาส 3 เป็นโรคข้อเข่าเสื่อมระดับรุนแรง 12 เรคอร์ด พบว่าจำแนกถูกต้อง 85.78 % และจำแนกคลาส 2 และคลาส 3 ได้ถูกต้องเพิ่มขึ้น โดยเฉพาะคลาส 3 เพิ่มขึ้นจาก 0 เป็น 75 % ซึ่งตัวแบบนี้สามารถนำมาใช้ในแผนส่งเสริมสุขภาพเพื่อวินิจฉัยและบำบัดผู้สูงอายุ

***ผู้รับผิดชอบบทความ :** syaowara@wu.ac.th

คำสำคัญ : ภาวะข้อเข่าเสื่อม; ความไม่สมดุลของข้อมูล; SMOTE; ADASYN; Gentleboost

Abstract

This research aimed to develop a predicting model for the assessment of elderly's knee osteoarthritis incidence. The 370 personal health records were collected from the osteoarthritis assessed from Ban Han Sub-district Health Promoting Hospital, Thasala district, Nakhon Si Thammarat province. The data are classified into 4 classes: class 0 refers to Excellent (200 records), class 1 refers to Good (115 records), class 2 refers to Moderate (39 records) and class 3 refers to Poor (16 records). In medical diagnosis application, the minority class was the class of primary interest and had a much higher misclassification than the majority class. The difference in total number of class 0 and class 1 from class 2 and class 3 indicated that these data were imbalanced. Thus the predicting model from these imbalanced data might limit the performance. The minority data of class 2 and class 3 were then adjusted by using the method of oversampling through ADASYN (adaptive synthetic sampling technique) and SMOTE (synthetic minority over-sampling technique). Subsequently, the data were divided into training data and testing data by using 10-fold cross validation. In addition, the multi-class imbalanced data classification algorithms; one-vs-one and one-vs-all; were employed in conjunction with the Gentleboost. The experimental results showed that the ADASYN and one-vs-one method achieved the best accuracy of 97.31 % on the imbalanced data. Moreover, our proposed predicting model was also tested with another imbalanced data from Ban Hua Ku Sub-district Health Promoting Hospital, Thasala district, Nakhon Si Thammarat province. The data are classified into 4 classes: class 0 refers to Excellent (141 records), class 1 refers to Good (63 records), class 2 refers to Moderate (16 records) and class 3 refers to Poor (12 records). It was found that the correct classification was 85.78 %. Furthermore, it also achieved the best performance in class 2 and class 3, especially in class 3 (Poor), the correct classification was increased from 0 to 75 %. In conclusion, the health promotion scheme can employ this model for diagnosis and plan the treatment for senior citizens.

Keywords: knee osteoarthritis; imbalanced data; SMOTE; ADASYN; Gentleboost

1. บทนำ

สำนักงานคณะกรรมการพัฒนาการเศรษฐกิจและสังคมแห่งชาติ [1] ระบุว่าประเทศไทยก้าวเข้าสู่สังคมผู้สูงอายุตั้งแต่ปี พ.ศ. 2548 โดยมีโครงสร้างสัดส่วนของประชากรวัยสูงอายุเพิ่มขึ้นมากกว่าร้อยละ 10 และในปี พ.ศ. 2583 จะเพิ่มขึ้นเป็นร้อยละ 32.1 เพื่อเป็นการเตรียมความพร้อม กรมอนามัย กระทรวงสาธารณสุข จึงมีการเตรียมความพร้อมโดยจัดทำสมุดปกฟ้าหรือสมุดบันทึกผู้สูงอายุ และแจกให้กับผู้สูงอายุตั้งแต่อายุ 60 ปีขึ้นไป เมื่อไปรับบริการในสถานพยาบาล

ทุกแห่ง รวมถึงโรงพยาบาลส่งเสริมสุขภาพตำบล (รพ.สต.) เพื่อช่วยในการติดตามการรักษาและประกอบการตัดสินใจในการรักษาผู้สูงอายุ

โรคข้อเข่าเสื่อม (knee osteoarthritis) เป็นโรคที่พบบ่อยในผู้สูงอายุ ส่งผลให้การทำกิจวัตรประจำวัน ได้แก่ การลุกยืน การนั่ง และการเดินเป็นไปด้วยความลำบาก [2] ซึ่งทำให้เกิดความเสี่ยงต่อการหกล้มและความทุพพลภาพ อาจเป็นสาเหตุให้ผู้สูงอายุต้องพึ่งพิงครอบครัว ผู้ดูแล และชุมชน เกิดความทุกข์ทรมาน ส่งผลต่อกิจวัตรประจำวัน ทำให้คุณภาพชีวิตลดลง และเกิดภาวะซึมเศร้าได้ [3] สาเหตุที่ทำให้เกิดโรคข้อเข่าเสื่อม 3 อันดับแรก คือ อายุที่มากขึ้น การไม่ออกกำลังกาย และการรับประทานอาหารไม่เหมาะสม ความเสี่ยงด้านสุขภาพที่จะทำให้เกิดโรคข้อเข่าเสื่อม คือ ภาวะน้ำหนักเกินและอ้วน และความเสี่ยงด้านลักษณะการทำกิจกรรมที่จะทำให้เกิดโรคข้อเข่าเสื่อม คือ การนั่งพับเพียบและการนั่งขัดสมาธิ [4] และปัจจัยที่เกี่ยวข้องของโรคข้อเข่าเสื่อม คือ เพศหญิงและผู้ที่มีดัชนีมวลกายมากกว่า 23.0 กิโลกรัมต่อตารางเมตร [5]

ผลการสำรวจสุขภาวะผู้สูงอายุไทยใน พ.ศ. 2556 พบว่าผู้สูงอายุส่วนใหญ่ได้รับบริการตรวจคัดกรองโรคข้อเข่าเสื่อมจากโรงพยาบาลส่งเสริมสุขภาพตำบลร้อยละ 37.80 แต่ผู้สูงอายุที่ได้รับการประเมินระดับความรุนแรงของโรคข้อเข่าเสื่อมโดยบุคลากรทางการแพทย์มีจำนวนเพียงร้อยละ 12.30 ของผู้เข้ารับการตรวจคัดกรองเท่านั้น ดังนั้นหากผู้ปฏิบัติหน้าที่ในรพ.สต. สามารถคัดกรองเพื่อวินิจฉัยความผิดปกติตั้งแต่ระยะแรกเริ่ม ระบุกรณีเสี่ยงและกรณีป่วยได้อย่างแม่นยำ รวดเร็ว และครอบคลุมกลุ่มเป้าหมาย [6] ก็จะให้บริการสุขภาพที่เหมาะสมสำหรับผู้ป่วยแต่ละกลุ่มได้

ข้อมูลจากแบบบันทึกการประเมินข้อเข่าเสื่อมปี พ.ศ. 2558-2559 จาก รพ.สต. บ้านหาร อำเภอท่าศาลา จังหวัดนครศรีธรรมราช ซึ่งเป็นแบบประเมินที่ใช้ Oxford knee score 12 ข้อ ดังตารางที่ 1 แต่ละข้อมีคะแนน 5 ระดับ โดย ระดับ 0 คือ รุนแรง และระดับ 4 คือ ปกติ ตามลำดับ และเกณฑ์ในสมุดบันทึกสุขภาพผู้สูงอายุ จำแนกความรุนแรงของโรคข้อเข่าเสื่อมเป็น 4

ตารางที่ 1 ข้อคำถามของแบบประเมินที่ใช้ Oxford knee score 12 ข้อ

ลำดับ	ข้อคำถาม
1	ลักษณะอาการเจ็บปวดเข่าของท่าน
2	ท่านมีปัญหาเรื่องเข่าในการทำกิจวัตรประจำวันด้วยตนเองหรือไม่ เช่น การยืนอาบน้ำ เป็นต้น
3	ท่านมีปัญหาเรื่องเข่า เมื่อก้าวขึ้นลงรถ หรือรถประจำทางหรือไม่
4	ระยะเวลานานเท่าไรที่ท่านเดินได้มากที่สุดก่อนที่ท่านจะมีอาการปวดเข่า
5	หลังทานอาหารเสร็จ ในขณะที่ลุกจากเก้าอี้นั่ง เข่าของท่านมีอาการอย่างไร
6	ท่านต้องเดินโยกตัว (เดินกระโผลกกระเผลก) เพราะอาการที่เกิดจากเข่าของท่านหรือไม่
7	ท่านสามารถนั่งลงคุกเข่าและลุกขึ้นได้หรือไม่
8	ท่านมีปัญหาปวดเข่าในขณะที่นอนกลางคืนหรือไม่
9	ในขณะที่ท่านทำงาน/ทำงานบ้าน ท่านมีอาการปวดเข่าหรือไม่
10	ท่านเคยมีความรู้สึกว่าเข่าของท่านทรุดลงทันทีหรือหมดแรงทันทีจนตัวทรุดลง
11	ท่านสามารถไปซื้อของใช้ต่าง ๆ ได้ด้วยตัวท่านเอง
12	ท่านสามารถเดินลงบันไดได้หรือไม่

ตารางที่ 2 ผลการประเมินระดับความรุนแรงของข้อมูลผู้สูงอายุ 370 เรคอร์ด

คลาส	จำนวนเรคอร์ด
คลาส 0 ยังไม่พบอาการผิดปกติ	200
คลาส 1 เริ่มมีอาการข้อเข่าเสื่อม	115
คลาส 2 มีอาการโรคข้อเข่าเสื่อมระดับปานกลาง	39
คลาส 3 เป็นโรคข้อเข่าเสื่อมระดับรุนแรง	16

ระดับ ได้แก่ ระดับ 0 คือ ยังไม่พบอาการผิดปกติ (40-48 คะแนน) ควรตรวจร่างกายเป็นประจำทุกปี ระดับ 1 คือ เริ่มมีอาการข้อเข่าเสื่อม (30-39 คะแนน) ควรได้รับคำแนะนำจากบุคลากรทางการแพทย์และการทำกิจกรรมอย่างเหมาะสม ระดับ 2 คือ มีอาการโรคข้อเข่าเสื่อมระดับปานกลาง (20-29 คะแนน) ควรปรึกษาศัลยแพทย์ผู้เชี่ยวชาญกระดูกและข้อเพื่อรับการตรวจรักษา และระดับ 3 คือ เป็นโรคข้อเข่าเสื่อมระดับรุนแรง (0-19 คะแนน) ควรรับการรักษาจากศัลยแพทย์ผู้เชี่ยวชาญกระดูกและข้อทันที และผลการประเมินระดับความรุนแรงของข้อมูลผู้สูงอายุ 370 เรคอร์ด ดังตารางที่ 2

งานวิจัยของ พุทธิพร และเยาวเรศ [7] ได้วิเคราะห์ข้อคำถามของแบบบันทึกการประเมินข้อเข่าเสื่อมในสมุดบันทึกสุขภาพผู้สูงอายุที่ใช้ Oxford knee score 12 ข้อ และเพิ่มเติมเพศ และ BMI รวมเป็น 14 ข้อคำถาม เพื่อค้นหาข้อคำถามที่ส่งผลต่อการทำนายภาวะข้อเข่าเสื่อม โดยใช้ข้อมูลจากแบบบันทึกการประเมินข้อเข่าเสื่อมในสมุดบันทึกสุขภาพผู้สูงอายุเพื่อประเมินระดับความรุนแรงของโรคข้อเข่าเสื่อมจำนวน 370 เรคอร์ด ตั้งแต่ปี พ.ศ. 2558-2559 จาก รพ.สต.บ้านหาร อำเภอท่าศาลา จังหวัดนครศรีธรรมราช และค้นพบข้อคำถาม 7 ข้อจาก 14 ข้อ ที่ส่งผลต่อการทำนายภาวะข้อเข่าเสื่อมในผู้สูงอายุ ดังตารางที่ 3

ตารางที่ 3 ข้อคำถาม 7 ข้อ ที่ส่งผลต่อการทำนายภาวะข้อเข่าเสื่อมในผู้สูงอายุ

ลำดับ	ข้อคำถาม
1	ท่านมีปัญหาเรื่องเข่าในการทำกิจวัตรประจำวันด้วยตนเองหรือไม่ เช่น การยืนอาบน้ำ เป็นต้น
2	ท่านมีปัญหาเรื่องเข่า เมื่อก้าวขึ้นลงรถ หรือรถประจำทางหรือไม่
3	ระยะเวลานานเท่าไรที่ท่านเดินได้มากที่สุด ก่อนที่ท่านจะมีอาการปวดเข่า
4	หลังทานอาหารเสร็จ ในขณะที่ลุกจากเก้าอี้นั่ง เข่าของท่านมีอาการอย่างไร
5	ท่านต้องเดินโยกตัว (เดินกระโผลกกระเผลก) เพราะอาการที่เกิดจากเข่าของท่านหรือไม่
6	ท่านสามารถนั่งลงคุกเข่าและลุกขึ้นได้หรือไม่
7	ท่านสามารถไปซื้อของใช้ต่าง ๆ ได้ด้วยตัวท่านเอง

ผู้วิจัยมีแนวคิดในการสร้างตัวแบบเพื่อใช้ในการพยากรณ์ภาวะข้อเข่าเสื่อมในผู้สูงอายุ เพื่อเพิ่มความถูกต้องในการคัดกรองความผิดปกติระดับปานกลางและระดับรุนแรง โดยใช้ข้อมูลจากแบบบันทึก

การประเมินข้อเข่าเสื่อมในสมุดบันทึกสุขภาพผู้สูงอายุ เพื่อประเมินระดับความรุนแรงของโรคข้อเข่าเสื่อม จำนวน 370 เรคอร์ด ตั้งแต่ปี พ.ศ. 2558-2559 จาก รพ.สต.บ้านหาร อำเภอท่าศาลา จังหวัดนครศรีธรรมราช และใช้คำตอบของข้อคำถาม 7 ข้อ ดังตารางที่ 3 ของ พุทธิพร และเยาวเรศ [7] และจากการวิเคราะห์ข้อมูลในตารางที่ 2 พบว่าข้อมูลชุดนี้เป็นข้อมูลไม่สมดุล ซึ่งเกิดขึ้นในข้อมูลทุกประเภท ข้อมูลไม่สมดุลจะแบ่งเป็นข้อมูลกลุ่มน้อย (minority class) ซึ่งเป็นข้อมูลที่มีจำนวนน้อยในชุดข้อมูล และข้อมูลกลุ่มมาก (majority class) ซึ่งเป็นข้อมูลที่มีจำนวนมากในชุดข้อมูล [8] ถ้านำข้อมูลทั้งสองชุดเข้าสู่ขั้นตอนการเรียนรู้พร้อมกันทั้งหมด จะทำให้ผลการแบ่งกลุ่มข้อมูลเกิดความผิดพลาด เนื่องจากการจำแนกข้อมูลมักให้ผลลัพธ์ในข้อมูลกลุ่มมาก ข้อมูลที่อยู่ในกลุ่มน้อยถูกครอบงำหรือถูกจัดให้อยู่ในกลุ่มมากทั้งหมด ซึ่งเรียกว่าปัญหาการแบ่งกลุ่มข้อมูลผิดกลุ่ม (misclassification) ส่งผลให้การจำแนกข้อมูลกลุ่มมากทำได้ถูกต้องและให้ผลลัพธ์ที่ไม่ถูกต้องกับข้อมูลกลุ่มน้อย ทำให้การจำแนกข้อมูลที่สนใจผิดพลาดได้ ดังนั้นงานวิจัยนี้จึงนำเทคนิคการปรับเพิ่มข้อมูลด้วยวิธีสุ่มมาใช้ปรับสมดุลของข้อมูล โดยใช้วิธี ADASYN และ SMOTE และเพิ่มประสิทธิภาพการจำแนกด้วยวิธี multi-class imbalanced data classification และ GentleBoost เพื่อให้จำแนกข้อมูลกลุ่มน้อยได้ดีขึ้น จากนั้นนำไปสร้างตัวแบบในการพยากรณ์ เพื่อเป็นประโยชน์ในการวินิจฉัยและวางแผนดูแลรักษาผู้สูงอายุต่อไป

2. การทบทวนวรรณกรรมที่เกี่ยวข้อง

2.1 การทำเหมืองข้อมูล

การทำเหมืองข้อมูลเป็นกระบวนการที่ดำเนินการกับข้อมูลขนาดใหญ่และมีจำนวนมาก เพื่อค้นหารูปแบบความสัมพันธ์ และความรู้ที่มีอยู่ในข้อมูลเป็นจำนวนมาก ซึ่งการค้นหาข้อมูลด้วยความรู้มี 5 รูปแบบ คือ (1) การค้นหากฎความสัมพันธ์ (2) การจำแนกประเภทและการพยากรณ์ (3) การจัดกลุ่มข้อมูล (4) การหาค่าผิดปกติที่เกิดขึ้น และ (5) การวิเคราะห์แนวโน้ม [9] และการสร้างตัวแบบในการจำแนกข้อมูลทำโดยการแบ่งข้อมูลเป็นชุดข้อมูลสอนและชุดข้อมูลทดสอบ จากนั้นประเมินความถูกต้องของตัวแบบ และสามารถนำตัวแบบไปใช้ทำนายข้อมูลชุดใหม่ต่อไป

ปัจจุบันมีการนำเทคนิคเหมืองข้อมูลมาใช้ในการจำแนกและพยากรณ์ในทางการแพทย์เพื่อจำแนกผู้ป่วย เช่น โรคระบบการหายใจ [10] โรคติดต่อทางเพศสัมพันธ์ [11] โรคตับอักเสบ [12] และในงานด้านอื่น ๆ เช่น งานวิจัยของ ธิดาภัทร และนิติ [13] ที่วิเคราะห์ปัจจัยที่มีอิทธิพลต่อการเกิดน้ำท่วมโดยใช้ต้นไม้การตัดสินใจ และงานวิจัยของ อนันต์ [14] ที่พัฒนาระบบสนับสนุนการตัดสินใจการเลือกสาขาวิชาเพื่อศึกษาต่อในระดับปริญญาตรีโดยใช้ต้นไม้การตัดสินใจและการเรียนรู้เบย์อย่างง่าย แล้วเปรียบเทียบประสิทธิภาพจากค่าความถูกต้อง (accuracy) และงานวิจัยของ จามรี และสมศรี [15] ที่พัฒนาตัวแบบเพื่อพยากรณ์คุณภาพผลิตภัณฑ์ฮาร์ดดิสก์ด้วยการถดถอยโลจิสติกส์และโครงข่ายประสาทเทียม โดยใช้การวิเคราะห์เหมืองข้อมูล ลดตัวแปรอิสระจาก 197 ตัว เหลือเพียง 25 ตัว ซึ่งพบว่าแบบจำลองจากโครงข่ายประสาทเทียมให้ความถูกต้องในการพยากรณ์สูงกว่าการถดถอยโลจิสติกส์

2.2 เทคนิคการปรับเพิ่มข้อมูลด้วยวิธีสุ่ม (oversampling)

การจัดการความไม่สมดุลของข้อมูล (imbalanced data) เกิดขึ้นเมื่อข้อมูลที่ใช้งานมีจำนวนข้อมูลในแต่ละคลาสแตกต่างกันมาก ทำให้ผลลัพธ์จากการจำแนกข้อมูลมีความโน้มเอียงไปทาง

ข้อมูลกลุ่มมาก ซึ่งเทคนิคการปรับเพิ่มข้อมูลด้วยวิธีสุ่ม เป็นการสุ่มเพิ่มจำนวนข้อมูลของข้อมูลกลุ่มน้อย ให้มีปริมาณข้อมูลใกล้เคียงกับข้อมูลกลุ่มมาก

2.2.1 SMOTE (synthetic minority over-sampling technique)

SMOTE เป็นเทคนิคที่ใช้ในการแก้ปัญหาการจำแนกข้อมูลที่ไม่สมดุล เนื่องจากข้อมูลในแต่ละคลาสมีจำนวนแตกต่างกันมาก ทำให้ผลลัพธ์ของการจำแนกอยู่ในข้อมูลกลุ่มมาก ดังนั้นวิธี SMOTE เป็นการเพิ่มจำนวนข้อมูลกลุ่มน้อย ให้มีจำนวนเพิ่มขึ้น [16] โดยการเพิ่มข้อมูลในกลุ่มน้อยนั้นทำให้การกระจายของกลุ่มข้อมูลมีความสมดุลมากขึ้น โดยทำการสุ่มค่าข้อมูลที่อยู่ในกลุ่มข้อมูลน้อยขึ้นมา 1 ค่า หลังจากนั้นพิจารณาค่าข้อมูลใกล้เคียงอีกจำนวน K ค่า (K-nearest neighbor) แล้วคำนวณค่าระยะทาง (Euclidean distance) ระหว่างค่าที่สุ่มกับค่าข้อมูลใกล้เคียงแต่ละค่า เพื่อหาค่าระยะทางที่น้อยที่สุด ระหว่างค่าที่สุ่มกับค่าข้อมูลใกล้เคียง จากนั้นจึงสร้างข้อมูลเทียมระหว่างค่าข้อมูลที่สุ่มกับค่าข้อมูลใกล้เคียงตัวที่ให้ค่าระยะทางที่น้อยที่สุด [17]

งานวิจัยของ วิไลรัตน์ และคณะ [18] ใช้เทคนิคการทำเหมืองข้อมูล วิธีการจำแนกประเภทข้อมูลและวิธีต้นไม้ตัดสินใจด้วยโปรแกรม WEKA เพื่อหาปัจจัยต่าง ๆ ที่มีผลต่อการทำนายการคลอดก่อนกำหนดในหญิงตั้งครรภ์ที่คลอดมีชีพ โดยปรับเพิ่มข้อมูลด้วยวิธีสุ่มเพิ่มข้อมูล และใช้เทคนิค SMOTE และทดสอบด้วยวิธีการตรวจสอบไขว้แบบ 10 กลุ่ม (10-fold cross validation) สำหรับการวิเคราะห์ปัจจัยที่มีผลต่อการทำนายการคลอดก่อนกำหนดในหญิงตั้งครรภ์ในโรงพยาบาลตำรวจ ซึ่งตัวแบบให้ค่าความถูกต้องร้อยละ 90.37 และงานวิจัยของ ภรัณยา [19] ที่เพิ่มประสิทธิภาพวิธีต้นไม้ตัดสินใจบนชุดข้อมูลที่ไม่สมดุลโดยเทคนิค SMOTE สำหรับข้อมูลการเป็นโรคติดอินเทอร์เน็ต โดยทดลองใช้วิธีต้นไม้ตัดสินใจ J48, ID3, LMT, CART และ Random Forest และใช้วิธีการตรวจสอบไขว้แบบ 10 กลุ่ม ในการแบ่งข้อมูลออกเป็นชุดข้อมูลสอนและชุดข้อมูลทดสอบ และพบว่าตัวแบบที่พัฒนาโดยเทคนิค Random Forest สามารถพยากรณ์ได้ค่าความแม่นยำ 87.15 %

งานวิจัยของ Sukmak และ Thongkam [20] ที่วิเคราะห์กระบวนการก่อนสร้างแบบจำลองด้วยการคัดกรองข้อมูลที่มีค่าผิดปกติด้วยวิธี C-SVCF และปรับความสมดุลของข้อมูลด้วยวิธี SMOTE และพัฒนาแบบจำลองด้วยเทคนิค PART decision list ทำให้ประสิทธิภาพการพยากรณ์สูงขึ้น และงานวิจัยของเชาวนันท์ และคณะ [21] สร้างแบบจำลองการทำนายผลการ รักษาผู้ป่วยมะเร็งปากมดลูกด้วยโครงข่ายประสาทเทียม และปรับความสมดุลของข้อมูลด้วยวิธี CSL และ SMOTE พบว่าวิธี SMOTE ให้ค่าความถูกต้อง 81.70 % และงานวิจัยของ วีระยุทธ และคณะ [22] ที่พัฒนาแบบจำลองเพื่อการพยากรณ์การรักษาซ้ำของผู้ป่วยโรคจิตเภทโดยเทคนิคเหมืองข้อมูล ใช้งาน C-SVCF สำหรับการคัดกรองข้อมูลที่มีค่าผิดปกติ และใช้เทคนิค SMOTE ในการปรับความสมดุลของข้อมูล และพัฒนาแบบจำลองด้วยเทคนิคต้นไม้ตัดสินใจ นาอีฟเบย์ และ PART decision list และพบว่าเทคนิค PART decision list สามารถพยากรณ์ได้ดีกว่าเทคนิคต้นไม้ตัดสินใจ และนาอีฟเบย์

งานวิจัยของ Sakr และคณะ [23] เปรียบเทียบผลของการทำนายโดยใช้งานเทคนิค SMOTE ในการปรับข้อมูลที่ไม่สมดุลและใช้งานร่วมกับเทคนิคการจำแนกข้อมูลต่าง ๆ พบว่าการใช้งานเทคนิค SMOTE ทำให้ผลการทำนายมีความถูกต้องเพิ่มขึ้น ซึ่งสอดคล้องกับงานวิจัยของ Richardson และ Lidbury [24] ซึ่งใช้งานเทคนิค SMOTE ในการปรับข้อมูลที่ไม่สมดุลก่อนการจำแนกข้อมูล และงานวิจัยของ Wang

และคณะ [25] ปรับปรุงการพยากรณ์ของมะเร็งเต้านม โดยใช้เทคนิคการสุ่มและการเลือกคุณลักษณะในการแก้ปัญหาข้อมูลไม่สมดุล ซึ่งพบว่าเทคนิค CRC ส่งผลต่อการทำนายได้ดีกว่าวิธี SMOTE

2.2.2 ADAptive SYNthetic (ADASYN)

งานวิจัยของ He และคณะ [26] เสนอวิธีการสุ่มเพิ่มข้อมูลในกลุ่มที่เรียกว่า ADAptive SYNthetic (ADASYN) เป็นวิธีที่ปรับปรุงการทำงานของ SMOTE ให้ดีขึ้น ซึ่งในขั้นตอนการสร้างข้อมูลเทียม (synthetic data) นั้นไม่จำเป็นต้องพิจารณาข้อมูลทุกตัวที่อยู่ในกลุ่มน้อย โดย ADASYN จะใช้ค่าการแจกแจงแบบถ่วงน้ำหนัก (weight distribution) ของข้อมูลตัวอย่างในกลุ่มน้อย โดยการสร้างข้อมูลเทียมซึ่งขึ้นอยู่กับความสำคัญของข้อมูลนั้น ๆ ถ้าข้อมูลตัวใดยากต่อการแบ่งกลุ่มก็จะให้ค่าน้ำหนักข้อมูลนั้นมากและสร้างชุดข้อมูลเทียมขึ้นมาบริเวณนั้น ๆ ซึ่งจะทำให้มีการปรับขอบเขตของการตัดสินใจในการแบ่งกลุ่มดีขึ้น และงานวิจัยของ Aditsania และคณะ [27] เสนอการจัดการข้อมูลลูกค้าที่ไม่สมดุลโดยใช้หลักการ ADASYN สำหรับการสุ่มเพิ่มข้อมูลในข้อมูลกลุ่มน้อยเพื่อปรับปรุงประสิทธิภาพการทำนายกลุ่มลูกค้าที่อยู่ในประเภทลูกค้าชั้นแย่ (customer churn) ที่มีสัดส่วนเพียง 4 % ให้ดีขึ้น โดยใช้ Neural Network Backpropagation Algorithm ในการจำแนกข้อมูล ซึ่งตัวแบบให้ค่าความถูกต้อง 96.31 %

2.3 Multi-Class Imbalanced Data classification

การแบ่งกลุ่มข้อมูลหลายกลุ่ม โดยที่แต่ละกลุ่มมีปัญหาเรื่องข้อมูลไม่สมดุลนั้นเป็นอุปสรรคอย่างหนึ่งที่ทำให้ประสิทธิภาพในการแบ่งกลุ่มข้อมูลไม่ดี เนื่องจากขอบเขตของแต่ละกลุ่ม (boundary class) อาจมีการซ้อนทับกัน ซึ่งวิธีแก้ไขปัญหานี้ คือ เปลี่ยนปัญหาการจัดแบ่งข้อมูลแบบหลายกลุ่มให้อยู่ในลักษณะการจัดแบ่งข้อมูลแบบ binary (binarization classification techniques) โดยใช้วิธี pairwise learning เรียกว่า one-vs-one [28] หรืออีกวิธี คือ one-vs-all [29]

2.3.1 one-vs-one

วิธี one-vs-one แบ่งกลุ่มข้อมูล *m* กลุ่ม ไปเป็น *m(m-1)/2* กลุ่ม แบบ binary problem ซึ่งแต่ละกลุ่มข้อมูลจะถูกจัดแบ่งด้วย classifier แบบ binary ในแต่ละขั้นตอนการเรียนรู้ ตัว classifier จะพิจารณากลุ่มข้อมูลครั้งละ 2 กลุ่ม เท่านั้น ส่วนข้อมูลที่อยู่ในกลุ่มอื่น ๆ จะไม่นำมาพิจารณา และจะแบ่งข้อมูลแบบนี้ไปเรื่อย ๆ จนกว่าจะครบทุกกลุ่มข้อมูล ในขั้นตอนสุดท้ายเป็นการกำหนดกลุ่ม (class label) ในแต่ละตัวข้อมูลตาม Code-matrix M [30] ดังสมการ

$$M(i,j) = \begin{cases} 1, & output = i \\ 0, & otherwise \end{cases} \quad (1)$$

และ final class label สามารถคำนวณได้จากการทำ maximum vote ดังสมการ

$$Class = arg \max_{i=1,\dots,c} \{\textstyle\sum_{j=1}^{c} M(i,j)\} \quad (2)$$

2.3.2 one-vs-all

วิธี one-vs-all แบ่งกลุ่มข้อมูล *m* กลุ่ม ไปเป็น *m* กลุ่ม แบบ binary problem ในแต่ละขั้นตอนการเรียนรู้จะแบ่งกลุ่มข้อมูลที่สนใจเป็น 1 กลุ่ม ซึ่งเรียกว่า positive examples และกลุ่มอื่น ๆ จะรวมกันเรียกว่า negative examples แต่ละ model ($F_1,\dots,F_C$) ที่ได้จากการแบ่งกลุ่มแต่ละครั้ง จะคำนวณ final decision function *F* [30] ดังสมการ

$$F(F_1,\dots,F_C) = \arg \underset{i=1,\dots,C}{max} (F_i) \quad (3)$$

2.4 Gentle AdaBoost Algorithm

Gentle Adaboost หรือ GentleBoost เป็นเทคนิคที่ปรับปรุงมาจากเทคนิคการบูทส่งเสริม (Adaboost algorithm) โดยลด exponential loss function ของ Adaboost ด้วยหลักการ Newton steps [31] ซึ่ง Gentle Adaboost นี้มีประสิทธิภาพในการจัดการกับข้อมูลที่มี noise ได้อย่างดี

Gentle AdaBoost algorithm [31]

Given: $(x_1, y_1), \ldots, (x_m, y_m)$ where $x_i \in X, y_i \in \{-1, +1\}$.

Start with weight $D_t(i) = \frac{1}{m}, i = 1, \ldots, m$

For $t = 1, \ldots, T$

(1) Estimate $h_t(x)$ by weighted least squares fitting of y to x.

(2) Update $H(x) \to H(x) + h_t(x)$

(3) Set $D_t(i) \to D_t(i) exp[-y_i . h_t(x_i)], i = 1, \ldots, m$ and renormalize so that $\sum_i D_i(t) = 1$

Output the classifier $H(x) = sign[\sum_{t=1}^{T} h_t(x)]$

งานวิจัยของ Mekhalfa และ Nacereddine [32] ได้นำเทคนิค Gentle AdaBoost มาใช้ในการแบ่งกลุ่มสำหรับการตรวจสอบด้วยรังสีถึงจุดบกพร่องในงานเชื่อม ผลการวิจัยพบว่าเทคนิคนี้ให้มีประสิทธิภาพในการทำนายที่ดีกว่าและเร็วกว่าวิธี Support Vector Machine โดยมีค่า AUC อยู่ที่ 0.9852 และ 0.9862

2.5 การวัดประสิทธิภาพ

ขั้นตอนการวัดประสิทธิภาพของแบบจำลอง โดยใช้ค่าความถูกต้อง (accuracy) ซึ่งคำนวณจากสมการ

$$\text{ค่าความถูกต้อง} = \frac{TP+TN}{TP+FP+FN+TN} \quad (4)$$

โดยที่ TP เป็นจำนวนข้อมูลที่ทำนายถูกต้องในกลุ่มบวก (positive); TN เป็นจำนวนข้อมูลที่ทำนายถูกต้องในกลุ่มลบ (negative); FP เป็นจำนวนข้อมูลที่ทำนายผิดว่าอยู่ในกลุ่มบวก (positive); FN เป็นจำนวนข้อมูลที่ทำนายผิดว่าอยู่ในกลุ่มลบ (negative)

3. วิธีดำเนินการวิจัย

3.1 การรวบรวมข้อมูลจากแบบประเมิน

งานวิจัยของ พุทธิพร และเยาวเรศ [7] ซึ่งใช้เทคนิคการทำเหมืองข้อมูล โดยใช้วิธีการตรวจสอบไขว้แบบ 10 กลุ่ม ใช้เทคนิคการสกัดคุณลักษณะแบบ DISR และจำแนกด้วยวิธีบูธส่งเสริม (AdaboostM2) และต้นไม้การตัดสินใจ (Decision tree) ในการวิเคราะห์ข้อถามที่มีคุณลักษณะที่ส่งผลต่อการทำนายภาวะข้อเข่าเสื่อมในผู้สูงอายุ จากคุณลักษณะ 12 ข้อ ดังตารางที่ 1 และคุณลักษณะของเพศ และ BMI รวมเป็น 14 ข้อ พบว่าสามารถสกัดข้อคำถาม 7 ข้อ ที่ส่งผลต่อการทำนายภาวะข้อเข่าเสื่อมในผู้สูงอายุ ดังตารางที่ 3

ผลการประเมินข้อมูลผู้สูงอายุ 370 เรคอร์ด ดังตารางที่ 2 พบว่าข้อมูลอยู่ในคลาส 0 ยังไม่พบอาการผิดปกติจำนวน 200 เรคอร์ด คลาส 1 เริ่มมีอาการข้อเข่าเสื่อม 115 เรคอร์ด คลาส 2 มีอาการโรคข้อเข่าเสื่อมระดับปานกลาง 39 เรคอร์ด และคลาส 3 เป็นโรคข้อเข่าเสื่อมระดับรุนแรง 16 เรคอร์ด ซึ่งพบว่าข้อมูลที่ได้มาปัญหาข้อมูลไม่สมดุล (imbalanced data) ดังนั้นผู้วิจัยจึงนำเทคนิคการปรับเพิ่มข้อมูลด้วยวิธีสุ่มด้วยวิธี SMOTE และ ADASYN มาใช้ปรับสมดุลข้อมูล เป็นการเพิ่มจำนวนข้อมูลกลุ่มน้อย ซึ่งก็คือจำนวนข้อมูลของคลาสที่ 2 และคลาสที่ 3 ก่อนนำเข้าสู่กระบวนการจำแนกข้อมูล พบว่าขนาดชุดข้อมูลที่สามารถเพิ่มประสิทธิภาพของแบบจำลองของคลาสที่ 2 และคลาสที่ 3 ด้วยวิธี SMOTE รวมเป็น 535 เรคอร์ด และด้วยวิธี ADASYN รวมเป็น 668 เรคอร์ด ดังตารางที่ 4 โดยในงานวิจัยนี้กำหนดพารามิเตอร์ที่สำคัญใน SMOTE คือ ค่า K-nearest neighbors มีค่าเท่ากับ 5 และพารามิเตอร์ที่สำคัญใน ADASYN กำหนดให้ β มีค่าเท่ากับ 1 ซึ่งค่า β มีค่าอยู่ระหว่าง 0-1 เป็นตัวกำหนดจำนวนข้อมูลเทียมที่สร้างในกลุ่มข้อมูลกลุ่มน้อยที่สนใจ ถ้าค่า β เท่ากับ 0 หมายความว่าไม่มีการสร้างข้อมูลเทียม และถ้าค่า β เท่ากับ 1 หมายถึงการสร้างข้อมูลเทียมในกลุ่มที่สนใจโดยมีจำนวนใกล้เคียงมากที่สุดกับจำนวนข้อมูลในกลุ่มมาก

ตารางที่ 4 จำนวนข้อมูลที่เพิ่มขึ้นด้วยวิธี SMOTE และ ADASYN

คลาส	ข้อมูลตั้งต้น	SMOTE	ADASYN
0	200	200	200
1	115	115	115
2	39	156	161
3	16	64	192
รวม	370	535	668

3.2 การพัฒนาตัวแบบการพยากรณ์

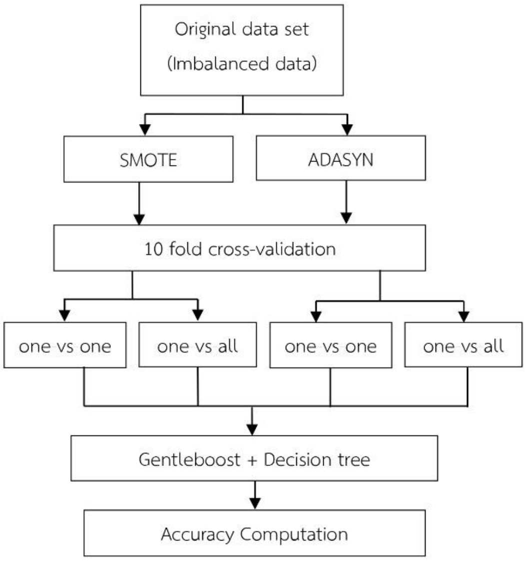
การพัฒนาตัวแบบการพยากรณ์โดยใช้งานโปรแกรม MATLAB มีขั้นตอนการดำเนินการดังนี้ ใช้วิธีการตรวจสอบไขว้แบบ 10 กลุ่ม โดยใช้งานคุณลักษณะ 7 ข้อ ในตารางที่ 3 ที่ได้เพิ่มขนาดชุดข้อมูลของคลาสที่ 2 และคลาสที่ 3 ด้วยวิธี SMOTE และ ADASYN จากนั้นจำแนกข้อมูลด้วยวิธี one-vs-one และ one-vs-all และ Gentleboost ดังรูปที่ 1

3.3 การวัดประสิทธิภาพของตัวแบบการพยากรณ์

การทดสอบความแม่นยำของตัวแบบการพยากรณ์ด้วยวิธีการตรวจสอบไขว้แบบ 10 กลุ่ม เพื่อให้ข้อมูลทุกตัวมีโอกาสเป็นชุดทดสอบและชุดสอน และการวัดประสิทธิภาพโดยรวมของตัวแบบการพยากรณ์ ได้จากการคำนวณค่าความถูกต้อง ดังสมการที่ 4

3.4 การวัดความถูกต้องของตัวแบบการพยากรณ์กับข้อมูลจริงที่ไม่สมดุล

การทดสอบความถูกต้องของตัวแบบการพยากรณ์ดำเนินการกับชุดข้อมูลจริงที่ไม่สมดุลของ รพ.สต.บ้านหัวคู อำเภอท่าศาลา จังหวัดนครศรีธรรมราช ซึ่งเป็นข้อมูลจากแบบบันทึกการประเมินข้อเข่าเสื่อมในสมุดบันทึกสุขภาพผู้สูงอายุเพื่อประเมินระดับความรุนแรงของโรคข้อเข่าเสื่อมที่ใช้ Oxford knee score 12 ข้อ ดังตารางที่ 1 ซึ่งมีข้อมูลรวมทั้งสิ้น 232 เรคอร์ด และแบ่งข้อมูลออกเป็น 4 คลาส ได้แก่ คลาส 0 จำนวน 141 เรคอร์ด คลาส 1 จำนวน 63 เรคอร์ด คลาส 2 จำนวน 16 เรคอร์ด และคลาส 3 จำนวน 12 เรคอร์ด และคำนวณความถูกต้องในการทำนายของตัวแบบ



รูปที่ 1 ขั้นตอนการดำเนินงานวิจัย

4. ผลการวิจัย

ผลของการนำข้อมูลมาปรับสมดุลด้วยวิธี SMOTE และ ADASYN และนำเข้าสู่กระบวนการจำแนกข้อมูลด้วยวิธี one-vs-one และ one-vs-all และ Gentleboost เพื่อพยากรณ์ภาวะโรคข้อเข่าเสื่อม โดยเปรียบเทียบค่าความถูกต้องดังนี้ ข้อมูลที่ไม่ได้ปรับสมดุลด้วยวิธีจำแนกข้อมูลที่นำเสนอในงานวิจัยของ พุทธิพร และเยาวเรศ [7] ข้อมูลที่ผ่านการปรับสมดุลด้วยวิธี SMOTE และ ADASYN และจำแนกข้อมูลด้วยวิธี one-vs-one และ one-vs-all และ AdaboostM2 + Decision tree และข้อมูลที่

ผ่านการปรับสมดุลด้วยวิธี SMOTE และ ADASYN และจำแนกข้อมูลด้วยวิธี one-vs-one และ one-vs-all และ Gentleboost + Decision tree โดยใช้ค่าความถูกต้อง เพื่อระบุประสิทธิภาพของตัวแบบ ดังตารางที่ 5

ตารางที่ 5 ค่าความถูกต้องของตัวแบบทั้ง 9 วิธี

Approach	Number of data	Accuracy (%)
AdaboostM2 + Decision tree [7]	370	**87.32**
SMOTE + one-vs-one + AdaboostM2 + Decision tree [7]	535	93.27
SMOTE + one-vs-all + AdaboostM2 + Decision tree [7]	535	91.22
ADASYN + one-vs-one + AdaboostM2 + Decision tree [7]	668	**96.46**
ADASYN + one-vs-all + AdaboostM2 + Decision tree [7]	668	45.96
SMOTE + one-vs-one + Gentleboost + Decision tree	535	92.34
SMOTE + one-vs-all + Gentleboost + Decision tree	535	91.59
ADASYN + one-vs-one + Gentleboost + Decision tree	668	97.31
ADASYN + one-vs-all + Gentleboost + Decision tree	668	**97.01**

ตารางที่ 6 ผลการทำนายของตัวแบบในชุดข้อมูลจริงจาก รพ.สต.บ้านหัวคู อำเภอท่าศาลา จังหวัดนครศรีธรรมราช

คลาส	ชุดข้อมูลจริงที่ไม่สมดุล	การทำนายที่ถูกต้องด้วยวิธี AdaboostM2 + Decision tree[7]		การทำนายที่ถูกต้องด้วยวิธี ADASYN + one-vs-one + AdaboostM2 + Decision tree [7]		การทำนายที่ถูกต้องด้วยวิธี ADASYN + one-vs-one + Gentleboost + Decision tree	
		จำนวนข้อมูล	%	จำนวนข้อมูล	%	จำนวนข้อมูล	%
0	141	139	98.58	138	97.87	137	97.16
1	63	30	47.62	42	68.25	42	66.67
2	16	8	50	0	0	11	68.75
3	12	0	0	8	66.66	9	75
รวม	232	177	76.29	188	81.47	199	85.78

จากนั้นนำตัวแบบของวิธี AdaboostM2 + Decision tree [7] วิธี ADASYN + one-vs-one + AdaboostM2 + Decision tree [7] และ ADASYN + one-vs-one + Gentleboost + Decision tree ซึ่งให้ค่าความถูกต้องสูงสุดในแต่ละกลุ่มไปสร้างตัวแบบการทำนายภาวะข้อเข่าเสื่อม เพื่อทดสอบกับชุดข้อมูลจริงที่ไม่สมดุลจาก รพ.สต.บ้านหัวคู โดยขนาดชุดข้อมูลของทั้ง 4 คลาส และผลการทำนายของทั้ง 3 ตัวแบบ ในชุดข้อมูลจริง ดังตารางที่ 6

5. วิจารณ์และสรุปผล

การประเมินเทคนิคการปรับเพิ่มข้อมูลด้วยวิธีสุ่มโดยนำวิธี SMOTE และ ADASYN มาใช้ปรับสมดุลข้อมูล ซึ่งเป็นการเพิ่มจำนวนข้อมูลกลุ่มน้อย พบว่าวิธี ADASYN สามารถเพิ่มจำนวนข้อมูลกลุ่มน้อยของคลาส 2 และคลาส 3 ได้สูงกว่าวิธี SMOTE เนื่องจากในการเพิ่มข้อมูลเทียมนั้น วิธี ADASYN จะพิจารณาจากข้อมูลจริงที่ยากต่อการแบ่งกลุ่มข้อมูล ซึ่งไม่จำเป็นต้องพิจารณาข้อมูลทุกตัวที่อยู่ในกลุ่มน้อย ถ้าข้อมูลตัวใดยากต่อการแบ่งกลุ่มก็ให้ค่าน้ำหนักข้อมูลนั้นมากและสร้างชุดข้อมูลเทียมขึ้นมาบริเวณนั้น ๆ ซึ่งทำให้มีการปรับขอบเขตของเส้นการตัดสินใจในการแบ่งกลุ่มให้ดี [27] และพบว่าการปรับเพิ่มข้อมูลด้วยวิธีสุ่มด้วยวิธี ADASYN ทำให้ประสิทธิภาพของตัวแบบสูงกว่าวิธี SMOTE และสอดคล้องกับ Wang และคณะ [25] ซึ่งพบว่าวิธี SMOTE ส่งผลต่อการทำนายต่ำกว่าวิธี CSC และในงานของ Shoorangiz และคณะ [34] ได้นำวิธี SMOTE และ ADASYN มาใช้ในการทำนาย EEG ของภาวะหลับในหรือการหลับระยะสั้น ๆ ผลการทำนายแสดงค่า ROC ของทั้ง 2 วิธี อยู่ในระดับ 90 % เท่ากัน วิธี ADASYN ให้ค่า sensitivity ที่ 76 % แต่วิธี SMOTE อยู่ที่ 70 % ดังนั้น วิธี SMOTE จึงไม่ได้เป็นวิธีที่มีประสิทธิภาพดีที่สุดในการเพิ่มจำนวนข้อมูลสำหรับการทำนาย

สำหรับกระบวนการจำแนกข้อมูลด้วยวิธี one-vs-one และ one-vs-all และ Gentleboost + Decision tree พบว่าวิธี one-vs-one ให้ค่าความถูกต้องสูงกว่า one-vs-all สำหรับชุดข้อมูลที่เพิ่มข้อมูลด้วยวิธีสุ่มทั้งวิธี SMOTE และ ADASYN เนื่องจากในแต่ละรอบการเรียนรู้ด้วยวิธี one-vs-one มีการพิจารณาเฉพาะ 2 กลุ่มข้อมูลที่เกี่ยวข้องเท่านั้น ส่วนกลุ่มอื่น ๆ ที่ไม่เกี่ยวข้องจะทำการตัดออกจากกระบวนการเรียนรู้ ซึ่งเมื่อพิจารณาแล้ว วิธีการนี้เป็นวิธีจัดการกับข้อมูลที่สมดุลมากกว่าการเรียนรู้ด้วยวิธี one-vs-all ซึ่งเป็นวิธีที่พิจารณาข้อมูลที่สนใจ 1 กลุ่ม (positive) ส่วนข้อมูลอื่น ๆ ที่เหลือจะถูกจัดให้อยู่ในกลุ่ม negative ซึ่งอาจทำให้แต่ละรอบการเรียนรู้มีจำนวนข้อมูลในสองกลุ่มต่างกันมาก [33]

การเปรียบเทียบประสิทธิภาพการพยากรณ์ระหว่าง ADASYN + one-vs-one + Gentleboost + Decision tree ซึ่งมีการปรับจำนวนของชุดข้อมูลกลุ่มน้อย และเป็นวิธีที่ให้ค่าความถูกต้องสูงสุด และ AdaboostM2 + Decision tree [7] ซึ่งเป็นวิธีที่ไม่มีการปรับจำนวนของชุดข้อมูลกลุ่มน้อย พบว่าตัวแบบของวิธี ADASYN + one-vs-one + Gentleboost + Decision tree สามารถทำนายคลาส 1 คลาส 2 และคลาส 3 ได้ถูกต้องมากกว่า AdaboostM2 + Decision tree [7] และสามารถทำนายคลาส 2 และคลาส 3 ถูกต้องเพิ่มขึ้น ซึ่งเป็นข้อมูลที่มีอยู่จำนวนน้อยในชุดข้อมูลจริง แต่เป็นคลาสของโรคข้อเข่าเสื่อมที่อยู่ในระดับปานกลางและรุนแรง ดังนั้นการแก้ปัญหาข้อมูลไม่สมดุลก่อนกระบวนการจำแนกข้อมูลจึงเป็นการเพิ่มประสิทธิภาพให้กับตัวแบบการพยากรณ์

กรณีของการทำนายคลาส 3 ซึ่งหากทำนายด้วยตัวแบบของวิธี AdaboostM2 + Decision tree [7] จะไม่สามารถทำนายได้ถูกต้อง (ค่าความถูกต้อง 0%) แต่เมื่อใช้ตัวแบบ ADASYN + one-vs-one + Gentleboost + Decision tree จะทำนายได้ถูกต้องสูงถึง 75 % และการทำนายคลาส 2 และคลาส 3 ซึ่งเป็นข้อมูลที่มีอยู่จำนวนน้อยในชุดข้อมูลจริงได้ถูกต้องเพิ่มขึ้น ดังนั้นวิธี ADASYN + one-vs-one + Gentle-boost + Decision tree จึงเป็นตัวแบบที่เหมาะสมต่อการนำไปประยุกต์ใช้สำหรับการพยากรณ์ภาวะโรคข้อเข่าเสื่อม จะทำให้สามารถวินิจฉัยภาวะโรคข้อเข่าเสื่อมในระดับปานกลางและในระดับรุนแรงได้ถูกต้องเพิ่มขึ้น ทั้งนี้หากต้องการเพิ่มประสิทธิภาพของการพยากรณ์

สามารถประยุกต์ใช้เทคนิค Random forest [19] หรือ Deep learning [35]

6. ข้อเสนอแนะ

ตัวแบบการพยากรณ์ที่นำเสนอในงานวิจัยนี้เป็นหนึ่งในเทคนิคของการทำเหมืองข้อมูลที่มีจุดเด่น คือ ค้นหาตัวแบบที่เหมาะสมต่อการนำไปประยุกต์ใช้สำหรับการพยากรณ์ภาวะโรคข้อเข่าเสื่อมในผู้สูงอายุ เพื่อทำให้สามารถวินิจฉัยภาวะโรคข้อเข่าเสื่อมในระดับปานกลางและในระดับรุนแรงได้ถูกต้องเพิ่มขึ้นและรวดเร็ว โดยใช้ข้อคำถามเพียง 7 ข้อ จาก 14 ข้อ (ข้อคำถามในแบบประเมินข้อเข่าเสื่อม 12 ข้อ รวมเพศ และ BMI) ซึ่งการลดจำนวนข้อคำถามยังคงทำให้พยากรณ์ภาวะโรคข้อเข่าเสื่อมได้ถูกต้อง และมีข้อดี คือ ทำให้ลดจำนวนข้อมูลที่ต้องเก็บรวบรวมจากผู้ป่วย ลดเวลาการทำงานของหน่วยพยาบาลปฐมภูมิ และลดเวลาการกรอกข้อมูลของผู้ป่วย แต่มีข้อจำกัดของจำนวนชุดข้อมูล ดังนั้นการพัฒนางานวิจัยให้มีประสิทธิภาพเพิ่มขึ้นจะต้องเพิ่มปริมาณข้อมูลที่นำมาใช้สร้างตัวแบบเพื่อให้มีข้อมูลที่หลากหลาย ซึ่งจะส่งผลให้ตัวแบบมีประสิทธิภาพเพิ่มขึ้น

7. กิตติกรรมประกาศ

ขอขอบคุณ สำนักงานคณะกรรมการวิจัยแห่งชาติ (วช.) ที่สนับสนุนทุนการวิจัย สัญญาทุนเลขที่ WU60114 ขอขอบคุณ โรงพยาบาลส่งเสริมสุขภาพตำบลบ้านหาร และโรงพยาบาลส่งเสริมสุขภาพตำบลบ้านหัวคู อำเภอท่าศาลา จังหวัดนครศรีธรรมราช ที่จัดเตรียมข้อมูล ทำให้การศึกษาวิจัยสำเร็จลุล่วงด้วยดี

8. รายการอ้างอิง

[1] สำนักงานคณะกรรมการพัฒนาการเศรษฐกิจและสังคมแห่งชาติ, 2554, แผนพัฒนาเศรษฐกิจและสังคมแห่งชาติ ฉบับที่ 11 พ.ศ. 2555-2559. แหล่งที่มา : http://www.nesdb.go.th/Portals/0/news/plan/p11/plan11.pdf, 20 สิงหาคม 2558.

[2] ปภัสรา หาญมนตรี, พรรณี ปึงสุวรรณ, ภาวินี เสริมชีพ, วิชัย อึงพินิจพงศ์, อุไรวรรณ ชัชวาล และรุ้งทิพย์ พันธุเมธากุล, 2557, ความเที่ยงในการทดสอบซ้ำและความสัมพันธ์ของแบบประเมิน Western Ontario and McMaster Universities Osteoarthritis ฉบับภาษาไทยกับคะแนนปวดในผู้สูงอายุที่มีภาวะข้อเข่าเสื่อม, ว.เทคนิคการแพทย์และกายภาพบำบัด 6: 84-92.

[3] รังสิยา นารินทร์, วิลาวัณย์ เตือนราษฎร์ และวราภรณ์ บุญเชียง, 2558, การพัฒนาโปรแกรมดูแลผู้สูงอายุข้อเข่าเสื่อมโดยการมีส่วนร่วมของชุมชน, พยาบาลสาร 42: 170-181.

[4] ยุวดี สารบูรณ์, สุภาพ อารีเอื้อ และสุจินดา จารุพัฒน์ มารุโอ, 2557, อาการ ความรู้ และการรับรู้ความเจ็บป่วยด้วยโรคข้อเข่าเสื่อมของผู้สูงอายุในชุมชน : การศึกษานำร่อง, ว.วิทยาลัยพยาบาลบรมราชชนนี 30: 12-24.

[5] จันทร์จิรา เกิดวัน, จิราภรณ์ บุญอินทร์, ชุติมา ธีระสมบัติ และวิไล คุปต์นิรัติศัยกุล, 2559, การสำรวจความชุกของโรคข้อเข่าเสื่อมผู้สูงอายุในชุมชน, ว.กายภาพบำบัด 38: 59-70.

[6] นงพิมล นิมิตอานันท์, 2557, สถานการณ์ทางระบาดวิทยาและการประเมินความเสี่ยงโรคข้อเข่าเสื่อมในคนไทย, ว.พยาบาลทหารบก 15: 185-194.

[7] พุทธิพร ธนธรรมเมธี และเยาวเรศ ศิริสถิตย์กุล, 2561, การวิเคราะห์ข้อคำถามที่มีผลต่อการทำนายภาวะเข่าเสื่อมในผู้สูงอายุโดยใช้เทคนิคเหมืองข้อมูล, ว.วิทยาศาสตร์และเทคโนโลยี มหา

วิทยาลัยมหาสารคาม 37: 726-735.

[8] Boonchuay, K., Sinapiromsaran, K. and Lursinsap, C., 2011, Minority split and gain ratio for a class imbalance, pp. 2060-2064, 8th International Conference on Fuzzy Systems and Knowledge Discovery.

[9] Han, J. and Kamber, M., 2000, Data Mining Concept and Techniques, Morgan Kaufmann Publishers, Canada.

[10] ดิษฐพล มั่นธรรม และลี่ลี อิงศรีสว่าง, 2553, การประยุกต์ขั้นตอนวิธีต้นไม้ตัดสินใจกับการวินิจฉัยโรคระบบการหายใจ : กรณีศึกษาที่โรงพยาบาลพระนครศรีอยุธยา, ว.วิจัยระบบสาธารณสุข 4: 73-81.

[11] Boonlue, S., Kammanat, T. and Kawsan, K., 2009, Diagnosis preliminary for leucorrhea system using back-propagation neural network, p. 163, 5th National Conference on Computing and Information Technology (NCCIT'09), Thailand.

[12] Kim, D.H., Uhmn, S., Ko, Y.W., Cho, S.W., Cheong, J.Y. and Kim, J., 2007, Chronic hepatitis and cirrhosis classification using SNP data, decision tree and decision rule, pp. 585-596, International Conference on Computational Science and Its Applica tions.

[13] ธิดาภัทร อนุชาญ และนิติ เอี่ยมชื่น, 2561, การวิเคราะห์ความเสี่ยงพื้นที่น้ำท่วมโดยใช้แบบจำลองต้นไม้การตัดสินใจ บริเวณลุ่มน้ำทะเลสาบสงขลา, ว.วิทยาศาสตร์และเทคโนโลยี มหาวิทยาลัยมหาสารคาม 37: 98-107.

[14] อนันต์ ปินะเต, 2560, การใช้เทคนิคเหมืองข้อมูลในการเลือกสาขาวิชาเพื่อโอกาสในการเข้าศึกษาต่อระดับปริญญาตรี, ว.วิทยาศาสตร์และเทคโนโลยี มหาวิทยาลัยมหาสารคาม 36: 704-712.

[15] จามรี ชูบัวทอง และสมศรี บัณฑิตวิไล, 2560, การพัฒนาตัวแบบเพื่อพยากรณ์คุณภาพผลิตภัณฑ์ฮาร์ดดิสก์ด้วยการถดถอยโลจิสติกส์และโครงข่ายประสาทเทียม, ว.วิทยาศาสตร์และเทคโนโลยี 25: 1-13.

[16] Chawla, N.V., Bowyer, K.W., Hall, L.O. and Kegelmayer, W.P., 2002, SMOTE: Synthetic minority over-sampling technique, J. Artif. Intell. Res. 16: 321-357.

[17] He, H. and Garcia, E.A., 2009, Learning from Imbalanced Data, IEEE T. Knowl. Data. En. 21: 1263-1284.

[18] วีไลรัตน์ วิศวไพศาล, บุญช่วย ศรีธรรมศักดิ์ และสาธิษฐ์ นากกระแสร์, 2559, ปัจจัยที่มีผลต่อการทำนายการคลอดก่อนกำหนดในหญิงตั้งครรภ์ในโรงพยาบาลตำรวจ, ว.พยาบาลตำรวจ 8: 83-90.

[19] ภรัณยา ปาลวิสุทธิ์, 2559, การเพิ่มประสิทธิภาพเทคนิคต้นไม้ตัดสินใจบนชุดข้อมูลที่ไม่สมดุลโดยวิธีการสุ่มเพิ่มตัวอย่างกลุ่มน้อยสำหรับข้อมูลการเป็นโรคติดอินเทอร์เน็ต, ว.เทคโนโลยีสารสนเทศ 12: 54-63.

[20] Sukmak, V. and ThongKam, J., 2013, Improving quality of breast cancer data through pre-processing, KKU Eng. J. 40: 493-504.

[21] เชาวนันท์ โสโท, พุธษดี ศิริแสงตระกูล และวรชัย ตั้งวรพงศ์ชัย, 2556, แบบจำลองการทำนายผลการรักษาผู้ป่วยมะเร็งปากมดลูกด้วยโครงข่ายประสาทเทียม, ว.วิจัย มข. (ฉบับบัณฑิตศึกษา)

13: 39-49.

[22] วีระยุทธ มายุศิริ, จารี ทองคำ และวาทินี สุขมาก , 2557, การพัฒนาแบบจำลองเพื่อการพยากรณ์การรักษาซ้ำของผู้ป่วยโรคจิตเภทโดยเทคนิคเหมืองข้อมูล, ว.วิทยาศาสตร์และเทคโนโลยี มหาวิทยาลัยมหาสารคาม ฉบับพิเศษ: 144-153.

[23] Sakr, S., Elshawi, R., Ahmed, A.M., Qureshi, W.T., Brawner, C.A., Keteyian, S.J., Blaha, M.J. and Al-Mallah, M.H., 2017, Comparison of machine learning techniques to predict all-cause mortality using fitness data: The Henry ford exercise testing (FIT) project, BMC Med. Inform. Decis. 17: 174-188.

[24] Richardson, A.M and Lidbury, B.A., 2017, Enhancement of hepatitis virus immune assay outcome predictions in imbalanced routine pathology data by data balancing and feature selection before the application of support vector machines, BMC Med. Inform. Decis. 17: 121-131.

[25] Wang, K.J., Makond, B. and Wang, K.M., 2013, An improved survivability prognosis of breast cancer by using sampling and feature selection technique to solve imbalanced patient classification data, BMC Med. Inform. Decis. 13: 124-137.

[26] He, H., Bai, Y., Garcia, E.A. and Li, S., 2008, ADASYN: Adaptive synthetic sampling approach for imbalanced, pp. 1322-1328, 2008 IEEE International Joint Conference on Neural Networks (IEEE World Congress on Computational Intelligence).

[27] Aditsania, A., Adiwijaya and Saonard A.L., 2017, Handling imbalanced data in churn prediction using ADASYN and back propagation algorithm, pp. 533-536, 3rd International Conference on Science in Information Technology.

[28] Hastie, T. and Tibshirani, R., 1998, Classification by pairwise coupling, Ann. Stat. 26: 451-471.

[29] Rifkin, R. and Klautau, A., 2004, In defense of one-vs-all classification, J. Mach. Learn. Res. 5: 101-141.

[30] Galar, M., Fernandez, A., Barrenechea, E., Bustince, H. and Herrera, F., 2011, An overview of ensemble methods for binary classifier in multi-class problems: Experimental study on one-vs-one and one-vs-all schemes, Pattern Recogn. 44: 1761-1776.

[31] Friedman, J., Hastie, T. and Tibshirani, R., 2000, Additive logistic regression: A statistical view of boosting, Ann. Stat. 28: 337-374.

[32] Mekhalfa, F. and Nacereddine, N., 2017, Gentle adaboost algorithm for weld defect classification, pp. 301-306, 21th Signal Processing: Algorithms, Architec tures, Arrangements, and Applications (SPA 2017).

[33] Fernandez, A., Lopez, V., Galar, M., Jesus, M.J. and Herrera, F., 2013, Analysing the classification of imbalanced data-sets with multiple classes: Binarization techniques and ad-hoc approaches, Knowl-

Based. Syst. 42: 97-110.

[34] Shoorangiz, R., Weddell, S.J. and Jones, R.D., 2016, Prediction of microsleeps from EEG: Preliminary results, pp. 4650-4653, 38th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC).

[35] Fausta, O., Hagiwarab, Y., Hong, T.J., Lih, O.H. and Acharya, U.R., 2018, Deep learning for healthcare applications based on physiological signals: A review, Comp. Methods Programs Biomed. 161: 1-13.