KKU RESEARCH JOURNAL (GRADUATE STUDY)
Vol. 18 NO. 3: July-September 2018

การแก้ปัญหาข้อมูลไม่สมดุลของข้อมูลสำหรับการจำแนกผู้ป่วยโรคเบาหวาน

Imbalanced Data Problem Solving in Classification of Diabetes Patients

วิชญ์วิสิฐ เกษรสิทธิ์ (Witwisit Kesornsit)[1]* ดร.วิชิต หล่อจีระชุณห์กุล (Dr.Vichit Lorchirachoonkul)**
ดร.จิราวัลย์ จิตรถเวช (Dr.Jirawan Jitthavech)***

บทคัดย่อ

การจำแนกโดยใช้ข้อมูลที่ไม่สมดุลเป็นปัญหาสำคัญในเทคนิคการจำแนก ซึ่งการจำแนกข้อมูลที่มีข้อมูลในกลุ่มมากและกลุ่มน้อยปะปนกันจะทำให้ข้อมูลในกลุ่มมากจะมีคุณสมบัติบางประการที่บดบังคุณสมบัติของกลุ่มน้อย ทำให้การจำแนกข้อมูลในกลุ่มน้อยไม่สามารถจำแนกได้อย่างมีประสิทธิภาพ การวิจัยครั้งนี้จึงมีวัตถุประสงค์เพื่อเปรียบเทียบประสิทธิภาพของวิธีการแก้ปัญหาข้อมูลไม่สมดุลของข้อมูลผู้ป่วยโรคเบาหวานโดยการแก้ปัญหาข้อมูลไม่สมดุลที่ระดับข้อมูลจำนวน 4 วิธีคือวิธีสุ่มเกิน วิธีสุ่มลด วิธีผสมผสาน และวิธีสังเคราะห์ข้อมูลใหม่ (SMOTE) โดยใช้เทคนิคการจำแนกคือวิธีการถดถอยโลจิสติกแบบมัลติโนเมียลและวิธีต้นไม้การตัดสินใจในการจำแนกผู้ป่วยโรคเบาหวาน จากการเปรียบเทียบประสิทธิภาพของสถิติและอัลกอริทึมในการจำแนก พบว่าข้อมูลที่แก้ปัญหาความไม่สมดุลด้วยวิธีวิธีสังเคราะห์ข้อมูลใหม่สามารถจำแนกผู้ป่วยโรคเบาหวานด้วยวิธีต้นไม้การตัดสินใจมีผลลัพธ์ดีที่สุด

ABSTRACT

Classification techniques when using imbalanced data is a challenging problem in the classification research area. The classification techniques of imbalanced data will cause the data in a majority class to have some features that obscure the characteristics of the minority class and make the classification performance of the minority class unacceptable. This research intends to compare the efficiency of solving the imbalanced data of diabetes patients using Data Level Solutions by 4 methods: Oversampling, Undersampling, Hybrid method and Synthetic Minority Oversampling TEchnique (SMOTE) in the classification using the multinomial logistic regression and decision tree techniques. By comparing the statistics and algorithms in the classification, it can be concluded that the classification by decision tree technique using SMOTE method to solve the imbalanced data by using decision tree technique yields the best result.

คำสำคัญ: การจำแนก ข้อมูลไม่สมดุล ต้นไม้การตัดสินใจ การถดถอยโลจิสติกแบบมัลติโนเมียล
Keywords: Classification, Imbalanced data, Decision tree, Multinomial logistic regression

---

[1] *Correspondent author: witwisit.kes@gmail.com*
** นักศึกษา หลักสูตรวิทยาศาสตรมหาบัณฑิต สาขาสถิติ คณะสถิติประยุกต์ สถาบันบัณฑิตพัฒนบริหารศาสตร์*
*** รองศาสตราจารย์ สาขาสถิติ คณะสถิติประยุกต์ สถาบันบัณฑิตพัฒนบริหารศาสตร์*
**** ศาสตราจารย์ สาขาสถิติ คณะสถิติประยุกต์ สถาบันบัณฑิตพัฒนบริหารศาสตร์*

บทนำ

การจำแนกเป็นเทคนิคหนึ่งที่สำคัญของการสืบค้นความรู้บนฐานข้อมูลขนาดใหญ่ (Knowledge Discovery from very large Database: KDD) และเป็นกระบวนการสร้างสมการทำนายจากข้อมูลการเรียนรู้ (training set) เพื่อนำสมการทำนายจำแนกข้อมูลใหม่ไปตามตัวแปรเป้าหมาย โดยอาศัยตัวแปรอิสระ (classifying attribute) ที่ให้สารสนเทศเกี่ยวกับตัวแปรเป้าหมาย [1] ซึ่งลักษณะของข้อมูลเป็นปัจจัยสำคัญที่มีผลต่อความถูกต้องของสมการทำนายตัวแปรเป้าหมาย เช่นการเก็บรวบรวมข้อมูลไว้อย่างครบถ้วน ไม่มีปัญหาข้อมูลสูญหาย การเก็บรวบรวมข้อมูลตรงตามคำอธิบายข้อมูลในตารางข้อมูลอภิพันธุ์ (metadata) การเก็บรวบรวมข้อมูลทันต่อเวลาและอยู่ในระยะเวลาที่เหมาะสมกับการสร้างสมการทำนาย ข้อมูลตรงตามวัตถุประสงค์ของการวิจัย และความไม่สมดุลของข้อมูลของตัวแปรเป้าหมายในการจำแนก เป็นต้น

ข้อมูลที่ไม่สมดุลของกลุ่มของตัวแปรเป้าหมายที่นำมาศึกษามีผลต่อความถูกต้องของสมการทำนาย จึงเป็นปัญหาที่นักวิจัยให้ความสนใจ ปัญหาความไม่สมดุลของข้อมูลพบได้อย่างบ่อยครั้งในข้อมูลจริง โดยเฉพาะข้อมูลทางการแพทย์ เช่นข้อมูลของตัวแปรที่เกี่ยวข้องกับความเจ็บป่วย ผู้ป่วยที่เข้ามารับการรักษาในโรงพยาบาลมีจำนวนมากแต่ผู้ป่วยที่ถูกวินิจฉัยว่าเป็นมะเร็งมีจำนวนน้อยเมื่อเทียบกับจำนวนผู้ป่วยทั้งหมด หรือบางโรคที่อาจพบได้ยากทางด้านการแพทย์อื่น ๆ เมื่อนำข้อมูลเหล่านี้มาใช้งานทางด้านการเรียนรู้ของเครื่องและการทำเหมืองข้อมูลจะส่งผลกระทบต่อการเรียนรู้ของอัลกอริทึม ซึ่งเมื่อทำการจำแนกข้อมูลด้วยวิธีการจำแนกข้อมูลแบบปกติที่ให้ความสำคัญกับข้อมูลทุกกลุ่มเป้าหมายพอ ๆ กันจะทำให้ประสิทธิภาพในการจำแนกประเภทข้อมูลส่วนน้อยมีความถูกต้องน้อยลง [2-3]

จากปัญหาดังกล่าวได้มีนักวิจัยคิดค้นวิธีการในการแก้ปัญหาความไม่สมดุลของข้อมูลซึ่งเป็นกระบวนการในการจัดการข้อมูลก่อนดำเนินการสร้างตัวแบบโดยใช้หลักการซึ่งแบ่งออกเป็น 3 ระดับคือ 1) การแก้ปัญหาข้อมูลไม่สมดุลที่ระดับข้อมูล (Data Level Solutions) ซึ่งเป็นการแก้ปัญหาในขั้นตอนก่อนการประมวลผลโดยจะปรับข้อมูลที่ไม่สมดุลให้กลายเป็นข้อมูลสมดุลด้วยเทคนิคการสุ่มเลือกข้อมูล 2) การแก้ปัญหาข้อมูลไม่สมดุลที่ระดับขั้นตอนวิธีการ (Algorithmic Level Solutions) เป็นการแก้ปัญหาโดยการปรับการเรียนรู้ของอัลกอริทึมมาตรฐานสำหรับการจำแนกข้อมูลที่มีอยู่เดิมให้สามารถเรียนรู้ข้อมูลไม่สมดุลโดยให้มีการเอนเอียงไปทางข้อมูลกลุ่มน้อย และ 3) การแก้ปัญหาข้อมูลไม่สมดุลด้วยการเรียนรู้แบบมีค่าใช้จ่าย (Cost-Sensitive Training) [3]

การวิจัยในครั้งนี้ผู้วิจัยใช้ข้อมูลผู้ป่วยโรคเบาหวานที่มีการแบ่งกลุ่มผู้ป่วยโรคเบาหวานให้กลับมารักษาซ้ำในโรงพยาบาลจำนวน 3 กลุ่มซึ่งเป็นการนำปัจจัยเสี่ยงของผู้ป่วยกับกระบวนการรักษาที่จัดให้กับผู้ป่วย รวมทั้งปัจจัยทางด้านกายภาพของโรงพยาบาลมาใช้ในการจำแนกการกลับมารักษาซ้ำในโรงพยาบาล โดยข้อมูลที่ใช้ในการวิจัยเกิดปัญหาข้อมูลไม่สมดุลและใช้วิธีการแก้ปัญหาข้อมูลไม่สมดุลที่ระดับข้อมูลจำนวน 4 วิธีคือวิธีการสุ่มเพิ่ม (Oversampling) การสุ่มลด (Undersampling) วิธีผสมผสาน (Hybrid method) และการสังเคราะห์ข้อมูลใหม่ (Synthetic Minority Oversampling TEchnique: SMOTE) และเปรียบเทียบความถูกต้องในการจำแนกกลุ่มของสมการทำนายหลังจากการแก้ปัญหาความไม่สมดุลของข้อมูลแล้ว

การสร้างสมการและกฎการทำนาย

การสร้างสมการทำนายเป็นวิธีการวิเคราะห์ทางสถิติที่เกี่ยวข้องกับการสร้างตัวแบบทางคณิตศาสตร์เพื่อแสดงความสัมพันธ์ระหว่างตัวแปรตามกับตัวแปรอิสระจำนวน 1 ตัวหรือมากกว่า 1 ตัวก็ได้ และการสร้างกฎการทำนายเป็นกระบวนการในการเรียนรู้ของอัลกอริทึมเพื่อสร้างสมการหรือกฎการจำแนกเพื่อช่วยในการตัดสินใจจากข้อมูลที่มีอยู่

โดยการเรียนรู้จากข้อมูลในอดีต (Supervised Learning) และใช้ทำนายแนวโน้มการเกิดขึ้นของข้อมูลที่ยังไม่เกิดขึ้น [1] สำหรับเทคนิคการจำแนกได้มีนักวิจัยคิดค้นและพัฒนาขึ้นมาจำนวนหลายประเภทซึ่งเทคนิคการจำแนกแต่ละประเภทมีความสามารถในการจำแนกแตกต่างกันและไม่สามารถระบุได้ชัดเจนว่าวิธีการใดเป็นวิธีที่ดีที่สุดในข้อมูลทุกประเภท เพราะข้อมูลแต่ละประเภทมีความเฉพาะตัวที่แตกต่างกัน สำหรับการวิจัยครั้งนี้เลือกใช้เทคนิคการวิเคราะห์การถดถอยโลจิสติกแบบมัลติโนเมียเป็นการศึกษาอิทธิพลของตัวแปรอิสระที่มีผลต่อตัวแปรตามทำให้ทราบว่าตัวแปรอิสระตัวใดบ้างที่มีความสัมพันธ์กับตัวแปรตามและสร้างสมการพยากรณ์โดยใช้ตัวแปรอิสระเป็นตัวพยากรณ์ตัวแปรตามซึ่งตัวแปรตามเป็นตัวแปรเชิงกลุ่มที่มีค่ามากกว่า 2 ค่า [4] และต้นไม้การตัดสินใจเป็นอัลกอริทึมที่จะทำการคัดเลือกตัวแปรที่มีความสัมพันธ์กับกลุ่มเป้าหมายมากที่สุดขึ้นมาเป็นโหนดบนสุดของต้นไม้เรียกว่าโหนดราก หลังจากนั้นก็จะหาตัวแปรถัดไปเรื่อย ๆ เพื่อสร้างใบของต้นไม้ [5]

ขั้นตอนการดำเนินการวิจัย

ในการศึกษาประสิทธิภาพการจำแนกผู้ป่วยโรคเบาหวานเมื่อใช้ข้อมูลไม่สมดุล ผู้วิจัยได้ดำเนินการวิจัยโดยศึกษางานวิจัยและทฤษฎีที่เกี่ยวข้อง ศึกษาโปรแกรมที่ใช้ในการวิเคราะห์ข้อมูล แล้วออกแบบการวิจัย การกำหนดตัวแปรที่ใช้ในการวิจัย การกำหนดเกณฑ์ในการเปรียบเทียบ หลังจากนั้นเตรียมข้อมูลสำหรับการทำวิจัย สร้างชุดข้อมูลสอน และชุดข้อมูลทดสอบโดยการสุ่มตัวอย่าง ทำการสร้างชุดข้อมูลใหม่โดยทำการแก้ปัญหาข้อมูลไม่สมดุลโดยวิธีสุ่มเพิ่ม วิธีสุ่มลด วิธีผสมผสาน และวิธีสังเคราะห์ข้อมูลใหม่ (SMOTE) หลังจากนั้นวิเคราะห์ข้อมูลเพื่อการจำแนกประเภทโดยใช้วิธีการวิเคราะห์การถดถอยโลจิสติกแบบมัลติโนเมียล และต้นไม้การตัดสินใจโดยใช้ข้อมูลเดิม และข้อมูลที่แก้ปัญหาความไม่สมดุลทั้ง 4 วิธีดังกล่าวข้างต้น และทดสอบประสิทธิภาพของแต่ละสถิติและอัลกอริทึมกับข้อมูลที่รวบรวมไว้ เพื่อบันทึกพฤติกรรมการสังเคราะห์ความรู้ของแต่ละสถิติและอัลกอริทึม หลังจากนั้นเปรียบเทียบประสิทธิภาพของการจำแนกของแต่ละสถิติและอัลกอริทึมที่สนใจศึกษาโดยใช้เกณฑ์ในการพิจารณาคือค่า 1) ผลบวกเท็จคือจำนวนข้อมูลที่มีประเภทเป็นลบแต่ถูกตัวแบบจำแนกว่าเป็นบวก 2) ผลลบเท็จคือจำนวนข้อมูลที่มีประเภทเป็นบวกแต่ถูกตัวแบบจำแนกว่าเป็นลบ 3) ความไวคือสมรรถภาพของการทดสอบในการจำแนกกลุ่มที่มีภาวะนั้น ๆ คำนวณจากสัดส่วนของจำนวนข้อมูลที่จำแนกได้ผลบวกจากข้อมูลทั้งหมด 4) ความถูกต้องคือความแม่นยำในการจำแนก 5) จำนวนตัวแปรที่ใช้ในการสร้างสมการหรือกฎการทำนาย 6) ROC Curve (The Receiver Operating Characteristic) เป็นกราฟที่แสดงความสัมพันธ์ระหว่างอัตราผลบวกจริงและผลลบจริงโดยค่า ROC ควรมีค่าเข้าใกล้ 1 และ 7) Lift Chart จะใช้ในการเปรียบเทียบเพื่อเลือกตัวแบบไปใช้งาน ซึ่งคำนวณจากสัดส่วนของจำนวนข้อมูลที่เป็นบวกกับจำนวนข้อมูลที่ตัวแบบทำนายว่าเป็นบวก (success rate) หารด้วยสัดส่วนที่จำนวนข้อมูลที่เป็นบวกเทียบกับจำนวนข้อมูลทั้งหมด (hit rate) ขั้นตอนสุดท้ายสรุปผลการทดลองและเสนอแนะวิธีการแก้ปัญหาข้อมูลไม่สมดุลที่เหมาะสมในการจำแนก

ข้อมูลที่ใช้ในการวิจัย

งานวิจัยนี้ทำการจำแนกประเภทของการกลับมารักษาซ้ำในโรงพยาบาลของผู้ป่วยโรคเบาหวานเมื่อข้อมูลไม่สมดุลใช้ข้อมูลจาก 130 โรงพยาบาลในประเทศสหรัฐอเมริกา จำนวน 101,766 ระเบียนของ Clinical Care at 130 US Hospitals and Integrated Delivery Networks [6] ซึ่งจัดเก็บอยู่ที่ UCI Machine Learning Repository[2] ตัวแปรที่ใช้ในการ

[2] *ชุดข้อมูลจัดเก็บอยู่ที่ [http://archive.ics.uci.edu/ml]. Irvine, CA: University of California, School of Information and Computer Science*

วิจัยจำนวน 33 ตัวแปรประกอบด้วย 1) เพศจำนวน 2 ระดับคือชายและหญิง 2) ช่วงอายุจำนวน 4 ระดับคือ 0-60 ปี 61-70 ปี 71-80 ปี และ 80 ปีขึ้นไป 3) ประเภทของการเข้ารับการรักษาในโรงพยาบาลจำนวน 3 ระดับคือฉุกเฉิน/เร่งด่วน เพิ่งปรากฏว่ามีแนวโน้มการเป็นโรค และประเภทอื่น ๆ 4) ประเภทของการออกจากโรงพยาบาลจำนวน 3 ระดับคือกลับบ้าน เข้ารับการรักษาในโรงพยาบาลเดิม และโอนไปยังสถานพยาบาลอื่น ๆ 5) ประเภทของแหล่งหรือแผนกที่เข้ารับการรักษาจำนวน 3 ระดับคือห้องฉุกเฉิน โรงพยาบาลหรือคลินิกอื่น และประเภทอื่น ๆ 6) จำนวนวันที่รักษาตัวในโรงพยาบาล 7) จำนวนของการสั่งตรวจทางห้องปฏิบัติการ 8) จำนวนของการสั่งตรวจรักษานอกเหนือจากการตรวจทางห้องปฏิบัติการ 9) จำนวนยาที่ใช้ในการรักษา 10) จำนวนของการเข้ารับการรักษาเป็นผู้ป่วยนอก 11) จำนวนของการเข้ารับการรักษาเป็นผู้ป่วยฉุกเฉิน 12) จำนวนของการเข้ารับการรักษาเป็นผู้ป่วยใน 13) จำนวนโรคของผู้ป่วยที่ถูกวินิจฉัยเข้ามาในระบบ 14) ผลการทดสอบซีรัมกลูโคส (Glucose serum) จำนวน 4 ระดับคือไม่ได้ทดสอบ ปกติ มากกว่าสองร้อย และมากกว่าสามร้อย 15) ผลการทดสอบ A1C จำนวน 4 ระดับคือไม่ได้ทดสอบ ปกติ มากกว่าเจ็ด และมากกว่าแปด 16) การเปลี่ยนแปลงในการใช้ยาเบาหวานทั้งปริมาณหรือชื่อทั่วไปจำนวน 2 ระดับคือไม่มีการเปลี่ยนแปลงและมีการเปลี่ยนแปลง 17) สถานะการใช้ยารักษาโรคเบาหวานจำนวน 2 ระดับคือใช้และไม่ใช้ 18-32) การเปลี่ยนแปลงปริมาณยาที่ใช้ต่อหนึ่งครั้งจำนวน 15 ชนิดได้แก่ Metformin, Repaglinide, Nateglinide, Insulin เป็นต้น จำนวน 3 ระดับคือไม่ได้ใช้ยาในการรักษา (No) ไม่มีการเปลี่ยนแปลงปริมาณยาที่ใช้ต่อหนึ่งครั้ง (Steady) และมีการเปลี่ยนแปลงปริมาณยาที่ใช้ต่อหนึ่งครั้ง (Up, Down) และ 33) สถานะการนัดกลับมารักษาซ้ำจำนวน 3 ระดับคือไม่นัดกลับมารักษาซ้ำ นัดกลับมารักษาซ้ำภายใน 30 วัน และนัดกลับมารักษาซ้ำมากกว่า 30 วัน และโปรแกรมที่ใช้ในการวิเคราะห์ข้อมูลการวิจัยคือ SAS (Statistical Analysis System) โดยใช้ SAS 9.4 และ SAS Enterprise Miner 14.2

จากการวิเคราะห์ข้อมูลตัวแปรสถานการณ์รักษาซึ่งแบ่งเป็น 3 กลุ่มคือไม่กลับมารักษาซ้ำ กลับมารักษาซ้ำภายใน 30 และกลับมารักษาซ้ำมากกว่า 30 วัน พบว่าข้อมูลของผู้ป่วยที่ถูกนัดให้กลับมารักษาซ้ำส่วนใหญ่จะอยู่ในประเภทไม่ถูกนัดให้กลับมารักษาซ้ำ คิดเป็นประมาณร้อยละ 54 จากของข้อมูลทั้งหมด ข้อมูลในกลุ่มกลับมารักษาซ้ำภายใน 30 วันคิดเป็นประมาณร้อยละ 11 จากข้อมูลทั้งหมด และข้อมูลในกลุ่มกลับมารักษาซ้ำมากกว่า 30 วันคิดเป็นประมาณร้อยละ 35 จากข้อมูลทั้งหมด

การแก้ปัญหาข้อมูลไม่สมดุล

กระบวนการวิจัยครั้งนี้ผู้วิจัยออกแบบการวิจัยออกเป็น 1) การใช้ข้อมูลเดิมทั้งชุดข้อมูลในการแก้ปัญหาข้อมูลไม่สมดุล และทำการวิเคราะห์ข้อมูลเพื่อสร้างสมการหรือกฎการจำแนก 2) การแบ่งข้อมูลเป็นข้อมูลทดสอบ 20% ของแต่ละกลุ่มเพื่อใช้ข้อมูลชุดเดียวกันเป็นข้อมูลทดสอบสำหรับการแก้ปัญหาข้อมูลไม่สมดุลทุกวิธีที่ใช้ในการวิจัยครั้งนี้ และนำข้อมูลอีก 80% ในการแก้ปัญหาข้อมูลไม่สมดุล หลังจากนั้นทำการวิเคราะห์ข้อมูลเพื่อสร้างสมการหรือกฎการจำแนก ซึ่งจากการทบทวนวรรณกรรมการแก้ปัญหาข้อมูลไม่สมดุลที่ระดับข้อมูลในการวิจัยครั้งนี้ใช้เทคนิควิธีดังนี้

1) วิธีสุ่มเกิน (Oversampling) เป็นการเพิ่มจำนวนข้อมูลที่อยู่ในกลุ่มส่วนน้อยให้มีจำนวนใกล้เคียงหรือเท่ากับจำนวนข้อมูลที่อยู่ในกลุ่มส่วนมาก ซึ่งการเพิ่มข้อมูลนั้นจะเพิ่มโดยการสุ่มเลือกจากข้อมูลเดิม [7-8] ในการวิจัยครั้งนี้จะใช้วิธีการสุ่มแบบเป็นระบบ ซึ่งผลการเพิ่มจำนวนข้อมูลที่อยู่ในกลุ่มส่วนน้อยให้มีจำนวนใกล้เคียงหรือเท่ากับจำนวนข้อมูลที่อยู่ในกลุ่มส่วนมากซึ่งมีจำนวนข้อมูล 54,850 กรณีสำหรับชุดข้อมูลเดิมทั้งชุดข้อมูลและมีจำนวนข้อมูล 43,880 กรณีสำหรับชุดข้อมูลที่แบ่งเป็นข้อมูลทดสอบ 20% แสดงได้ดังตารางที่ 1

2) วิธีสุ่มลด (Undersampling) เป็นการลดจำนวนข้อมูลที่อยู่ในกลุ่มส่วนมากให้มีจำนวนใกล้เคียงหรือเท่ากับจำนวนข้อมูลที่อยู่ในกลุ่มส่วนน้อย [7-8] ในการวิจัยครั้งนี้จะใช้วิธีการสุ่มแบบเป็นระบบ ซึ่งผลการลดจำนวนข้อมูลที่

อยู่ในกลุ่มส่วนมากให้มีจำนวนใกล้เคียงหรือเท่ากับจำนวนข้อมูลที่อยู่ในกลุ่มส่วนน้อยซึ่งมีจำนวน 11,356 กรณีสำหรับชุดข้อมูลเดิมทั้งชุดข้อมูลและมีจำนวนข้อมูล 9,085 กรณีสำหรับชุดข้อมูลที่แบ่งเป็นข้อมูลทดสอบ 20% แสดงได้ดังตารางที่ 1

3) วิธีผสมผสาน (Hybrid Methods) เป็นวิธีการที่นำเทคนิควิธีสุ่มเกินและวิธีสุ่มลดมาทำงานร่วมกัน โดยพยายามหาค่ากลางในการจะชักตัวอย่างให้ได้ตามจำนวนที่อยู่ตรงกลางระหว่างข้อมูลในกลุ่มส่วนมากกับข้อมูลในกลุ่มส่วนน้อย [8-11] ซึ่งจำนวนที่อยู่ตรงกลางมีจำนวน 43,494 กรณีสำหรับชุดข้อมูลเดิมทั้งชุดข้อมูลและมีจำนวนข้อมูล 34,795 กรณีสำหรับชุดข้อมูลที่แบ่งเป็นข้อมูลทดสอบ 20% และในการวิจัยในครั้งนี้จะใช้วิธีการสุ่มแบบเป็นระบบซึ่งผลการแก้ปัญหาแสดงได้ดังตารางที่ 1

4) วิธีสังเคราะห์ข้อมูลเพิ่ม (Synthetic Minority Oversampling TEchnique: SMOTE) เป็นเทคนิคการสุ่มตัวอย่างแบบพิเศษของการสุ่มเพิ่ม แทนที่จะสุ่มเพิ่มโดยใช้ข้อมูลเดิมแต่จะทำการสังเคราะห์ข้อมูลขึ้นมาใหม่จากข้อมูลเดิมที่มีอยู่หลักการเพื่อนบ้านที่อยู่ใกล้ที่สุดในการขยายขอบเขตการตัดสินใจของตัวแบบ [8, 10, 12-13] ซึ่งขั้นตอนในการสังเคราะห์ข้อมูลใหม่มีขั้นตอนดังนี้คือระบุเพื่อนบ้านที่ใกล้เคียงที่สุด k ค่าของข้อมูลเดิม สำหรับข้อมูลเดิม M หา $k=l$ ที่มีระยะทางใกล้เคียงกับข้อมูลเดิมโดย l คือจำนวนเพื่อนบ้านที่ใกล้เคียงกับจุด M แล้วสุ่มเลือกจุดระหว่างสองจุดและสร้างกรณีใหม่ ในการวิจัยครั้งนี้ทำการสังเคราะห์ข้อมูลใหม่สำหรับข้อมูลกลุ่มที่กลับมารักษาซ้ำภายใน 30 วันโดยกำหนดค่า $l=4$ ซึ่งทำการสังเคราะห์ข้อมูลใหม่จำนวน 43,239 กรณีสำหรับชุดข้อมูลเดิมทั้งชุดข้อมูลและทำการสังเคราะห์ข้อมูลใหม่จำนวน 36,340 กรณีสำหรับชุดข้อมูลที่แบ่งเป็นข้อมูลทดสอบ 20% และสำหรับข้อมูลกลุ่มที่กลับมารักษาซ้ำมากกว่า 30 วันทำการสังเคราะห์ข้อมูลใหม่โดยกำหนดค่า $l=1$ ซึ่งทำการสังเคราะห์ข้อมูลใหม่จำนวน 20,062 กรณีสำหรับชุดข้อมูลเดิมทั้งชุดข้อมูลและทำการสังเคราะห์ข้อมูลใหม่จำนวน 15,457 กรณีสำหรับชุดข้อมูลที่แบ่งเป็นข้อมูลทดสอบ 20% ตัวอย่างเช่นสร้างจุด $m_1(c_1,c_2,...,c_n)$ ระหว่าง $M(a_1,a_2,...,a_n)$ และ $M_1(b_1,b_2,...,b_n)$ เมื่อ[3]

$$c_1 = a_1 + (b_1 - a_1) \times rand('UNIFORM')$$
$$c_2 = a_2 + (b_2 - a_2) \times rand('UNIFORM')$$
$$.$$
$$.$$
$$c_n = a_n + (b_n - a_n) \times rand('UNIFORM')$$

โดยที่ m_1 คือจุดที่สังเคราะห์ขึ้นมาใหม่ระหว่าง $M(a_1,a_2,...,a_n)$ และ $M_1(b_1,b_2,...,b_n)$
$a_1,a_2,...,a_n$ คือข้อมูลในค่าสังเกตที่จุด M และ $b_1,b_2,...,b_n$ คือข้อมูลในค่าสังเกตที่จุด M_1

สำหรับข้อมูลหลายมิตินั้นอัลกอริทึมจะทำงานในลักษณะเดียวกัน สำหรับตัวแปรช่วง (interval variable) สามารถคำนวณระยะทางแบบยุคลิดได้โดยตรง แต่สำหรับตัวแปรที่มีตัวแปรแบ่งประเภท (categorical variable) ต้องใช้การแปลงตัวแปรหุ่น (dummy variable) ก่อนเพื่อให้มีรูปแบบตัวเลขและคำนวณระยะทาง ถ้าตัวแปรช่วงอยู่ในระดับที่แตกต่างกันวิธีที่ดีที่สุดคือการกำหนดค่าเป็นค่ามาตรฐาน (standardize) ก่อนที่จะคำนวณระยะทาง [14]

---

[3] *ผู้วิจัยใช้อัลกอริทึมจากบทความของ Wang, R., N. Lee, and Y. Wei, A Case Study: Improve Classification of Rare Events with SAS® Enterprise Miner™. SAS and all other SAS Institute Inc, 2015. 3282: p. 1-12.*

ผลการดำเนินการวิจัย

เนื่องจากข้อมูลผู้ป่วยโรคเบาหวานที่ใช้ในการวิจัยเกิดปัญหาความไม่สมดุลของกลุ่มเป้าหมายโดยมีอัตราส่วนของข้อมูลระหว่างกลุ่มไม่กลับมารักษาซ้ำ กลับมารักษาซ้ำภายใน 30 วัน และกลับมารักษาซ้ำมากกว่า 30 วันเป็นร้อยละ 53.92 : 11.16 : 34.92 ตามลำดับ ซึ่งข้อมูลไม่สมดุลนี้จะส่งผลกระทบต่อการจำแนกข้อมูลทำให้ไม่สามารถจำแนกข้อมูลของกลุ่มที่มีจำนวนข้อมูลน้อยได้ถูกต้องแม่นยำ ในขณะเดียวกันจะสามารถจำแนกข้อมูลของกลุ่มที่มีจำนวนมากได้อย่างแม่นยำ ดังนั้นจึงได้ทำการปรับความสมดุลของข้อมูลก่อนการนำมาสร้างสมการทำนายโดยได้ผลการปรับสมดุลข้อมูลทั้ง 4 วิธีตามที่ได้กล่าวมาแล้วดังตารางที่ 1

หลังจากนั้นนำข้อมูลที่ได้ปรับสมดุลแล้วทั้ง 4 วิธีคือวิธีการสุ่มเพิ่ม วิธีการสุ่มลด วิธีการผสมผสาน และวิธีการสังเคราะห์ข้อมูลใหม่การจำแนกผู้ป่วยโรคเบาหวานออกเป็น 3 กลุ่มคือไม่กลับมารักษาซ้ำ กลับมารักษาซ้ำภายใน 30 วัน และกลับมารักษาซ้ำมากกว่า 30 วัน โดยใช้วิธีการวิเคราะห์การถดถอยโลจิสติกแบบมัลติโนเมียล และต้นไม้การตัดสินใจ ได้ผลการศึกษาดังตารางที่ 2

จะเห็นได้ว่าเมื่อทำการจำแนกจากข้อมูลที่มีปัญหาความไม่สมดุลสำหรับการวิเคราะห์การถดถอยและต้นไม้การตัดสินใจจะทำให้สามารถจำแนกข้อมูลของกลุ่มที่มีข้อมูลน้อยได้เพียงเล็กน้อยซึ่งเห็นได้ชัดเจนจากผลการจำแนกผู้ป่วยในกลุ่มของการกลับมารักษาซ้ำภายใน 30 วันซึ่งจะมีความไวหรือโอกาสที่ผู้ที่ควรถูกนัดมารักษาซ้ำภายใน 30 วันจะได้รับนัดมารักษาซ้ำภายใน 30 วันในร้อยละที่น้อย โดยสำหรับข้อมูลเดิมทั้งชุดข้อมูลคิดเป็นร้อยละ 1.06 และ 1.41 ตามลำดับ และสำหรับข้อมูลหลังจากแบ่งเป็นข้อมูลทดสอบ 20% ของแต่ละกลุ่มคิดเป็นร้อยละ 0.42 และ 0.51 ตามลำดับ ในกรณีที่ตัวแปรตามมีจำนวน 3 กลุ่มนั้นสัดส่วนของความไวในการจำแนกแต่ละกลุ่มควรมีค่าใกล้เคียงกันซึ่งสัดส่วนร้อยละของความไวที่คาดหวังควรจะเป็นร้อยละ 30-40 ของแต่ละกลุ่ม เมื่อพิจารณาความไวในกรณีนี้พบว่าสมรรถภาพของการทดสอบในการจำแนกกลุ่มที่มีสถานะของการกลับมารักษาซ้ำภายใน 30 วันสามารถจำแนกได้ในอัตราที่น้อยมากหากผู้วิจัยนำตัวแบบดังกล่าวนี้ไปใช้ก็จะมีผลต่อผู้ป่วยที่จะต้องนัดให้กลับมารักษาซ้ำภายใน 30 วันที่จะมีการจำแนกที่ผิดพลาดค่อนข้างสูง และจะมีค่าผลลบเท็จหรือผลลบที่ไม่เป็นจริงที่สูงกว่าชุดข้อมูลอื่นดังแสดงในตารางที่ 2

เมื่อแก้ปัญหาข้อมูลไม่สมดุลเพื่อลดผลกระทบต่อการจำแนกข้อมูลส่วนน้อยพบว่าประสิทธิภาพของการจำแนกผู้ป่วยโรคเบาหวานหลังการแก้ปัญหาข้อมูลไม่สมดุลสามารถลดปัญหาการจำแนกที่ไม่สามารถจำแนกข้อมูลของกลุ่มที่มีจำนวนข้อมูลน้อยได้ถูกต้องแม่นยำได้ดีมากขึ้นซึ่งสำหรับในการวิจัยนี้ข้อมูลในกลุ่มน้อยคือกลุ่มกลับมารักษาซ้ำภายใน 30 วันและกลับมารักษาซ้ำมากกว่า 30 วัน เมื่อแก้ปัญหาข้อมูลไม่สมดุลจะทำให้สามารถเพิ่มความไวในการจำแนกผู้ป่วยในกลุ่มกลับมารักษาซ้ำภายใน 30 วันให้มีค่าสูงขึ้นจากการจำแนกโดยใช้ข้อมูลเดิมส่งผลให้ตัวแบบในการทำนายสามารถนำไปใช้ในการจำแนกผู้ป่วยได้ในทุก ๆ กลุ่ม นอกจากนั้นแล้วยังสามารถลดค่าผลลบเท็จให้มีค่าน้อยลงอีกด้วยดังแสดงในตารางที่ 2

สำหรับการแก้ปัญหาข้อมูลที่ไม่สมดุลที่มีประสิทธิภาพดีที่สุดในการวิจัยครั้งนี้คือการใช้วิธีสังเคราะห์ข้อมูลเพิ่มโดยจะให้ค่าความถูกต้อง ค่า ROC Index และค่า Lift มากที่สุด รวมทั้งค่าผลลบเท็จน้อยที่สุดจากทุก ๆ วิธี และเทคนิคการจำแนกที่มีประสิทธิภาพดีที่สุดในการจำแนกผู้ป่วยโรคเบาหวานที่แก้ปัญหาข้อมูลไม่สมดุลด้วยวิธีสังเคราะห์ข้อมูลเพิ่มคือเทคนิคต้นไม้การตัดสินใจ โดยผลการวิจัยเป็นไปในทางเดียวกันทั้งการใช้ข้อมูลเดิมทั้งชุดข้อมูลและข้อมูลหลังจากแบ่งเป็นข้อมูลทดสอบ 20% ของแต่ละกลุ่มดังแสดงในตารางที่ 2

สรุปผลการดำเนินการวิจัย

สำหรับการจำแนกผู้ป่วยโรคเบาหวานที่ตัวแปรตามมีจำนวน 3 กลุ่มคือไม่กลับมารักษาซ้ำ กลับมารักษาซ้ำภายใน 30 วัน และกลับมารักษาซ้ำมากกว่า 30 วันซึ่งพบว่ากฎการทำนายโดยใช้เทคนิคต้นไม้การตัดสินใจจะจำแนกการกลับมารักษาซ้ำของผู้ป่วยที่มีร้อยละของผลบวกเท็จ ร้อยละผลลบเท็จต่ำกว่า และมีร้อยละของการจำแนกกลุ่มได้ถูกต้องสูงกว่าสมการทำนายโดยใช้วิธีการถดถอยโลจิสติกแบบพหุสำหรับทุกชุดข้อมูลที่ใช้การแก้ปัญหาข้อมูลไม่สมดุลด้วยวิธีการสังเคราะห์ข้อมูลใหม่ (SMOTE) ซึ่งสอดคล้องกับงานวิจัยของ Wang (2015) ที่ได้ทำวิจัยเรื่อง A Case Study: Improve Classification of Rare Events with SAS® Enterprise Miner™ [14] และสอดคล้องกับงานวิจัยของ Damodaran (2015) ที่ได้ทำวิจัยเรื่อง Predicting Rare Events Using Specialized Sampling Techniques in SAS® [15] รวมทั้งสอดคล้องกับงานวิจัยของ Guzman (2015) ที่ได้ทำการวิจัยเรื่อง Data sampling improvement by developing SMOTE technique in SAS [12] ซึ่งงานวิจัยทั้งสามเรื่องได้วิจัยเกี่ยวกับเทคนิคการแก้ปัญหาข้อมูลไม่สมดุลและผลการวิจัยพบว่าการแก้ปัญหาข้อมูลไม่สมดุลด้วยวิธีสังเคราะห์ข้อมูลใหม่จะมีประสิทธิภาพดีที่สุดในการจำแนกข้อมูล

จากผลการวิจัยจะเห็นได้ว่าการแก้ปัญหาข้อมูลไม่สมดุลด้วยวิธีการสุ่มเพิ่ม วิธีการสุ่มลด และวิธีผสมผสานนั้นจะทำให้ผลการจำแนกมีประสิทธิภาพไม่ค่อยมีความแตกต่างกันมากนัก เนื่องจากการแก้ปัญหาด้วยวิธีดังกล่าวจะเป็นการเพิ่มและลดข้อมูลจากข้อมูลเดิมทำให้ข้อมูลเดิมที่ถูกปรับเปลี่ยนโครงสร้างมีปัญหาตามข้อจำกัดของแต่ละวิธีคือ 1) วิธีการสุ่มลดเป็นการลดจำนวนข้อมูลที่อยู่ในกลุ่มส่วนมากให้มีจำนวนใกล้เคียงหรือเท่ากับจำนวนข้อมูลที่อยู่ในกลุ่มส่วนน้อยซึ่งการใช้วิธีนี้จะทำให้ตัวแบบที่ได้สามารถใช้ในการจำแนกข้อมูลได้ทุกกลุ่ม แต่วิธีนี้จะทำให้สูญเสียข้อมูลในการวิเคราะห์จนทำให้มีตัวแบบที่ได้สูญเสียสารสนเทศในการอธิบายตัวแปรตามไปส่วนหนึ่งซึ่งจะส่งผลให้ผลการจำแนกมีความผิดพลาดมากขึ้น [7-8, 15] 2) วิธีการสุ่มเพิ่มเป็นการเพิ่มจำนวนข้อมูลที่อยู่ในกลุ่มส่วนน้อยให้มีจำนวนใกล้เคียงหรือเท่ากับจำนวนข้อมูลที่อยู่ในกลุ่มส่วนมากซึ่งการใช้วิธีนี้จะเป็นการหลีกเลี่ยงการสูญเสียข้อมูลในการวิเคราะห์แต่การสร้างตัวแบบจะมีความเสี่ยงที่จะได้ตัวอย่างเดียวกันในข้อมูลสอนและข้อมูลทดสอบ ทำให้ข้อมูลทดสอบไม่ได้เป็นอิสระจากข้อมูลสอนอีกต่อไปทำให้ผลการจำแนกมีความเอนเอียง [7-8, 15] 3) วิธีผสมผสานเป็นเทคนิคการชักตัวอย่างแบบเพิ่มและแบบลดร่วมกันก่อนการสร้างตัวแบบ ซึ่งสามารถใช้ทั้งสองวิธีร่วมกันในการจัดการข้อมูลเพื่อแก้ปัญหาความไม่สมดุลระหว่างกลุ่ม ซึ่งเป็นวิธียังคงดำเนินการเพิ่มข้อมูลจากข้อมูลเดิมซึ่งจะมีผลต่อการสร้างความเอนเอียงในการจำแนก และต้องลดข้อมูลจากข้อมูลเดิมทำให้สูญเสียข้อมูลส่วนหนึ่งในการสร้างตัวแบบ [8-9, 11]

นอกจากนั้นผลการวิจัยพบว่าการแก้ปัญหาข้อมูลไม่สมดุลด้วยวิธีการสังเคราะห์ข้อมูลใหม่จะมีประสิทธิภาพดีที่สุด และเป็นวิธีการที่ได้รับความนิยมเนื่องจากเป็นวิธีที่ค่อนข้างง่าย ซึ่งสามารถแก้ปัญหาความเอนเอียงและข้อมูลไม่เพียงพอในการสร้างตัวแบบได้เนื่องจากเป็นการสังเคราะห์ข้อมูลของในกลุ่มส่วนน้อยให้เพิ่มขึ้นมาจนได้ชุดข้อมูลใหม่โดยไม่ได้ใช้ข้อมูลเดิมในการเพิ่มจำนวนข้อมูลที่จะส่งผลให้เกิดความเอนเอียงในการจำแนกของตัวแบบ [7, 12, 14]

วิธีการแก้ปัญหาข้อมูลไม่สมดุลด้วยเทคนิคการแก้ปัญหาข้อมูลไม่สมดุลที่ระดับข้อมูลนี้จะเป็นวิธีที่เป็นการเปลี่ยนแปลงโครงสร้างและคุณสมบัติเดิมของข้อมูล โดยการเพิ่มเข้าหรือเอาออกอาจทำให้เกิดการสูญเสียคุณสมบัติและโครงสร้างเดิมของข้อมูล แต่เมื่อมีความจำเป็นต้องจัดการข้อมูลด้วยวิธีเหล่านี้ วิธีที่มีประสิทธิภาพดีที่สุดในการจำแนกคือวิธีการสังเคราะห์ข้อมูลใหม่โดยจะสามารถแก้ปัญหาการสูญเสียข้อมูลในการวิเคราะห์จนทำให้มีข้อมูลไม่เพียงพอในการสร้างตัวแบบ และลดความเสี่ยงที่จะได้ตัวอย่างเดียวกันในข้อมูลสอนและข้อมูลทดสอบ แม้ว่าการแก้ปัญหาข้อมูลไม่สมดุลด้วยวิธีที่ผู้วิจัยเลือกอาจไม่สามารถเพิ่มค่าความถูกต้องในการจำแนกประเภทให้มีค่าเกิน

ร้อยละ 80 ซึ่งเป็นค่าที่นิยมในการยอมรับได้ในการนำตัวแบบที่ใช้งาน [16] แต่การแก้ปัญหาข้อมูลไม่สมดุลสามารถปรับปรุงประสิทธิภาพการจำแนกให้มีประสิทธิภาพขึ้นกว่าข้อมูลเดิม และยังสามารถทำให้ตัวแบบสามารถจำแนกได้ทั้งข้อมูลกลุ่มมากและกลุ่มน้อยได้เท่าเทียมหรือใกล้เคียงกัน

สำหรับปัจจัยในการนัดหมายผู้ป่วยให้กลับมารักษาซ้ำที่ได้จากการวิจัยครั้งนี้นั้นขึ้นอยู่กับปัจจัยเสี่ยงที่ตรวจพบจากผู้ป่วยและกระบวนการรักษาที่จัดให้กับผู้ป่วยหรือเรียกว่าปัจจัยทางชีวภาพ เช่นจำนวนโรคของผู้ป่วยที่ถูกวินิจฉัยเข้ามาในระบบ ผลการทดสอบ A1C ผลการทดสอบซีรั่มกลูโคส เป็นต้น รวมทั้งปัจจัยทางด้านกายภาพของโรงพยาบาลที่จะสามารถรองรับผู้ป่วยให้เข้ารับการรับการรักษาตัวในโรงพยาบาลตามระดับความสำคัญของอาการของผู้ป่วยที่เข้ารับการรักษาในขณะนั้น เช่นจำนวนเตียง จำนวนผู้ป่วยนอกที่จะต้องกลับมารักษาซ้ำ จำนวนผู้ป่วยฉุกเฉินที่โรงพยาบาลต้องดูแล เป็นต้นซึ่งสอดคล้องกับงานวิจัยของ Strack (2014) ที่ได้ศึกษาเรื่อง Impact of HbA1c measurement on hospital readmission rates: Analysis of 70,000 clinical database patient records และ Rubin (2015) ที่ได้ศึกษาเรื่อง Hospital Readmission of Patients with Diabetes [6, 17]

ปัจจุบันโรคเบาหวานเป็นโรคไม่ติดต่อเรื้อรังชนิดหนึ่งที่เป็นสาเหตุหลักของการเสียชีวิตของผู้ป่วยและมีแนวโน้มของผู้ป่วยเพิ่มขึ้นในทุก ๆ ปี ทำให้โรงพยาบาลต้องแบกรับค่าใช้จ่ายในการรักษาพยาบาลเพิ่มขึ้น จากการศึกษาของ Dogan (2013) พบว่าปัจจัยเสี่ยงต่าง ๆ ที่มีผลต่อการกลับมารักษาซ้ำของผู้ป่วยโรคเบาหวานซึ่งเป็นองค์ประกอบสำคัญที่มีผลต่อค่าใช้จ่ายทางด้านการแพทย์ (medical expenditures) และคุณภาพของการรักษา (quality of care) ในการเข้ารับการรักษาของผู้ป่วยนั้นทางโรงพยาบาลคาดหวังให้ผู้ป่วยมีอาการที่ดีขึ้น ซึ่งจะส่งผลต่อการลดอัตราการกลับมารักษาซ้ำที่จะเป็นกระบวนจัดการในการช่วยลดค่าใช้จ่ายของโรงพยาบาลและช่วยลดค่าใช้จ่ายของภาครัฐในการดูแลรักษาผู้ป่วย [17-19]

การนำปัจจัยเสี่ยงของผู้ป่วยกับกระบวนการรักษาที่จัดให้กับผู้ป่วย รวมทั้งปัจจัยทางด้านกายภาพของโรงพยาบาลมาใช้ในการจำแนกการกลับมารักษาซ้ำจะเป็นกระบวนการที่สำคัญในการบริหารจัดการเรื่องค่าใช้จ่ายของโรงพยาบาล รวมทั้งสามารถใช้ในการวางแผนการรักษาและวางแผนบริหารจัดการภายในโรงพยาบาลให้เหมาะสม นอกจากนี้ยังเป็นกระบวนการในการบริหารจัดการเพื่อลดผลกระทบต่อผู้ป่วยทั้งด้านร่างกาย จิตใจ อารมณ์ สังคม และเศรษฐกิจ รวมทั้งสะท้อนถึงคุณภาพการดูแลรักษาของโรงพยาบาล การจัดการเพื่อลดการกลับเข้ามารักษาซ้ำในโรงพยาบาลของผู้ป่วยจึงเป็นเรื่องสำคัญที่จะช่วยให้คุณภาพชีวิตของผู้ป่วยดีขึ้นและพัฒนาคุณภาพการพยาบาล [20-22]

เอกสารอ้างอิง

1. Ammaruekarat P, et al., A Comparative Efficiency of Feature Selection and Neural Network Classification, in The 5th National Conference on Computing and Information Technology. 2552. Thai
2. Chomboon K. Classification Technique for Minority Class on Imbalanced Dataset with Data Partitioning Method [PhD Thesis], Nakhon Ratchasima: Suranaree University of Technology; 2015. Thai
3. Chujai P. Ensemble Learning for Imbalanced Data Classification Problem [PhD Thesis], Nakhon Ratchasima: Suranaree University of Technology; 2014. Thai
4. Starkweather J, Moske AK. Multinomial Logistic Regression. 2011 [cited 2017 27 March]; Available from: https://it.unt.edu/sites/default/files/mlr_jds_aug2011.pdf.
5. Anderson E, et al., Data Mining in Readmission Problem. Technical Report, 2014.

6. Strack B, et al., Impact of HbA1c measurement on hospital readmission rates: Analysis of 70,000 clinical database patient records. BioMed Research International, 2014: p. 1-11
7. Fawcett T. Learning from Imbalanced Classes. 2016 [cited 2017 15 April]; Available from: https://www.svds.com/learning-imbalanced-classes/?utm_source=kdnuggets&utm_medium=blog&utm_campaign=learning+from+imbalanced+classes.
8. Garbled. Class Imbalance Problem. 2013 [cited 2016 14 May]; Available from: http://www.chioka.in/class-imbalance-problem/.
9. Gazzah S, Hechkel A, Amara NEB. A hybrid sampling method for imbalanced data. in Systems, Signals & Devices (SSD). The 12th International Multi-Conference. 2015. IEEE.
10. Songwattanasiri P. Synthetic Minority Over-Sampling and Majority Under-Sampling Techniques for Class Imbalanced Problems [MSc Thesis], Bangkok: Chulalongkorn University; 2010. Thai
11. Lu Y, Cheung Y, Tang Y. Hybrid Sampling with Bagging for Class Imbalance Learning. in Pacific-Asia Conference on Knowledge Discovery and Data Mining. 2016. Springer.
12. Guzman L. Data sampling improvement by developing SMOTE technique in SAS. SAS and all other SAS Institute Inc, 2015. 3483-2015.
13. He H, et al. ADASYN: Adaptive synthetic sampling approach for imbalanced learning. in Neural Networks, 2008. IJCNN 2008. (IEEE World Congress on Computational Intelligence). IEEE International Joint Conference. 2008. IEEE. p. 1322-1328.
14. Wang R, Lee N, Wei Y. A Case Study: Improve Classification of Rare Events with SAS® Enterprise Miner™. SAS and all other SAS Institute Inc, 2015. 3282: p. 1-12.
15. Damodaran R, et al., Predicting Rare Events Using Specialized Sampling Techniques in SAS®. SAS and all other SAS Institute Inc, 2015. 11140-2016.
16. Edell A. Understand these 5 basic concepts to sound like a machine learning expert. 2017 [cited 2017 20 June]; Available from: https://medium.com/towards-data-science/understand-these-5-basic-concepts-to-sound-like-a-machine-learning-expert-6221ec0fe960.
17. Rubin DJ. Hospital Readmission of Patients with Diabetes. Current Diabetes Reports, 2015. 15(4).
18. Dogan N, Tanrikulu Z. A comparative analysis of classification algorithms in data mining for accuracy, speed and robustness. Information Technology and Management, 2013. 14(2): p. 105-124.
19. Zhu, Q., Akkati, A. and Hongwattanakul, P. Risk feature assessment of readmission for diabetes. in 2016 IEEE International Conference on Bioinformatics and Biomedicine (BIBM). 2016. IEEE.
20. Soomhirun R, Monkong S, Khuwawatanasamrit K. A Literature Review Related to the Management for Reducing Readmission in Patients with Heart Failure. Thai Journal of Cardio-Thoracic Nursing, 2009. 20(1): p. 17-32. Thai
21. Enomoto LM, et al., Risk factors associated with 30-day readmission and length of stay in patients with type 2 diabetes. Journal of Diabetes and its Complications, 2017. 31(1): p. 122-127.
22. Ching HY, YAP FKP. Paediatric hospital readmissions with diabetes mellitus. 2012.

ตารางที่ 1 การแจกแจงความถี่ของการกลับมารักษาซ้ำในโรงพยาบาล 3 สถานะของผู้ป่วย

สถานะของผู้ป่วย	ข้อมูลเดิม		วิธีสุ่มเพิ่ม		วิธีสุ่มลด		วิธีผสมผสาน		วิธีสังเคราะห์ข้อมูลใหม่	
	ความถี่	ร้อยละ	ความถี่	ร้อยละ	ความถี่	ร้อยละ	ความถี่	ร้อยละ	ความถี่	ร้อยละ
ข้อมูลเดิมทั้งชุดข้อมูล										
ไม่กลับมารักษาซ้ำ (ไม่มีภาวะโรค)	**54,850**	**53.92**	**54,850**	**32.81**	**11,356**	**33.33**	**43,494**	**32.85**	**54,850**	**33.24**
กลับมารักษาซ้ำภายใน 30 วัน	**11,356**	**11.16**	**56,780**	**33.97**	**11,356**	**33.33**	**45,424**	**34.31**	**54,595**	**33.08**
กลับมารักษาซ้ำมากกว่า 30 วัน	**35,528**	**34.92**	**55,528**	**33.22**	**11,356**	**33.33**	**43,494**	**32.85**	**55,590**	**33.68**
ข้อมูลหลังจากแบ่งเป็นข้อมูลทดสอบ 20% ของแต่ละกลุ่ม										
ไม่กลับมารักษาซ้ำ (ไม่มีภาวะโรค)	**43,880**	**53.92**	**43,880**	**32.95**	**9085**	**33.33**	**34,795**	**32.85**	**43,880**	**32.95**
กลับมารักษาซ้ำภายใน 30 วัน	**9085**	**11.16**	**45,425**	**34.10**	**9085**	**33.33**	**36,340**	**34.30**	**45,425**	**34.10**
กลับมารักษาซ้ำมากกว่า 30 วัน	**28,423**	**34.92**	**43,880**	**32.95**	**9085**	**33.33**	**34,795**	**32.85**	**43,880**	**32.95**

ตารางที่ 2 ผลการจำแนกการกลับมารักษาซ้ำของผู้ป่วยทั้ง 3 สถานะของผู้ป่วยของสมการทำนาย

การวิเคราะห์ข้อมูล	ผลบวกเท็จ	ผลลบเท็จ	ความไว		ความถูกต้อง	จำนวนตัวแปร	ROC	Lift*
			≤ 30 วัน	>30 วัน				
ข้อมูลเดิมทั้งชุดข้อมูล								
ข้อมูลเดิม (Original Data)								
การถดถอยโลจิสติก	**0.16**	**34.95**	**1.06**	**24.00**	**57.09**	**4**	**0.61**	**1.41**
ต้นไม้การตัดสินใจ	**0.19**	**34.96**	**1.41**	**36.20**	**57.60**	**16**	**0.62**	**1.44**
ข้อมูลที่แก้ปัญหาความไม่สมดุลด้วยวิธีสุ่มเพิ่ม (Oversampling)								
การถดถอยโลจิสติก	**21.04**	**23.39**	**42.06**	**33.86**	**43.67**	**12**	**0.58**	**1.26**
ต้นไม้การตัดสินใจ	**23.31**	**23.20**	**46.75**	**39.19**	**45.69**	**17**	**0.59**	**1.32**
ข้อมูลที่แก้ปัญหาความไม่สมดุลด้วยวิธีสุ่มลด (Undersampling)								
การถดถอยโลจิสติก	**20.52**	**24.47**	**40.81**	**37.46**	**43.88**	**6**	**0.57**	**1.22**
ต้นไม้การตัดสินใจ	**24.27**	**23.17**	**47.17**	**29.96**	**45.17**	**14**	**0.58**	**1.26**
ข้อมูลที่แก้ปัญหาความไม่สมดุลด้วยวิธีผสมผสาน (Hybrid)								
การถดถอยโลจิสติก	**21.66**	**25.43**	**42.41**	**35.18**	**43.97**	**14**	**0.58**	**1.26**
ต้นไม้การตัดสินใจ	**23.56**	**24.44**	**46.29**	**38.03**	**45.62**	**21**	**0.59**	**1.32**

การวิเคราะห์ข้อมูล	ผลบวกเท็จ	ผลลบเท็จ	ความไว		ความถูกต้อง	จำนวนตัวแปร	ROC	Lift*
			≤ 30 วัน	>30 วัน				
ข้อมูลที่แก้ปัญหาความไม่สมดุลด้วยวิธีสังเคราะห์ข้อมูลเพิ่ม (SMOTE)								
การถดถอยโลจิสติก	**22.94**	**22.80**	**48.35**	**45.88**	**62.96**	**22**	**0.63**	**1.55**
ต้นไม้การตัดสินใจ	**22.59**	**22.76**	**46.67**	**37.93**	**65.07**	**15**	**0.62**	**1.53**
ข้อมูลหลังจากแบ่งเป็นข้อมูลทดสอบ 20% ของแต่ละกลุ่ม								
ข้อมูลเดิม (Original Data)								
การถดถอยโลจิสติก	**0.18**	**34.77**	**0.42**	**13.54**	**63.80**	**10**	**0.54**	**1.15**
ต้นไม้การตัดสินใจ	**0.37**	**34.74**	**0.51**	**19.56**	**64.84**	**14**	**0.55**	**1.28**
ข้อมูลที่แก้ปัญหาความไม่สมดุลด้วยวิธีสุ่มเพิ่ม (Oversampling)								
การถดถอยโลจิสติก	**11.34**	**26.05**	**25.39**	**17.93**	**63.91**	**16**	**0.57**	**1.35**
ต้นไม้การตัดสินใจ	**7.84**	**29.34**	**15.97**	**26.57**	**64.00**	**14**	**0.57**	**1.42**
ข้อมูลที่แก้ปัญหาความไม่สมดุลด้วยวิธีสุ่มลด (Undersampling)								
การถดถอยโลจิสติก	**8.91**	**30.28**	**18.24**	**25.08**	**63.98**	**7**	**0.56**	**1.37**
ต้นไม้การตัดสินใจ	**7.07**	**28.55**	**13.29**	**28.40**	**64.40**	**15**	**0.57**	**1.45**
ข้อมูลที่แก้ปัญหาความไม่สมดุลด้วยวิธีผสมผสาน (Hybrid)								
การถดถอยโลจิสติก	**9.97**	**28.56**	**18.21**	**25.08**	**63.98**	**15**	**0.56**	**1.34**
ต้นไม้การตัดสินใจ	**8.83**	**27.46**	**21.35**	**20.31**	**64.55**	**21**	**0.57**	**1.44**
ข้อมูลที่แก้ปัญหาความไม่สมดุลด้วยวิธีสังเคราะห์ข้อมูลเพิ่ม (SMOTE)								
การถดถอยโลจิสติก	**13.16**	**27.73**	**29.63**	**21.49**	**65.32**	**18**	**0.61**	**1.42**
ต้นไม้การตัดสินใจ	**9.98**	**24.57**	**20.61**	**31.79**	**65.33**	**14**	**0.62**	**1.44**

* ค่า Lift คือค่าสูงสุดจาก Lift chart