# การพัฒนาแบบจำลองเพื่อการพยากรณ์การรักษาซ้ำของผู้ป่วยโรคจิตเภทโดยเทคนิคเหมืองข้อมูล

# Developing the Prediction Models for Readmission of Schizophrenia Patients using Data Mining Techniques

วีระยุทธ มายุศิริ[1*], จารี ทองคำ[2], วาทินี สุขมาก[3]
Weerayut Mayusiri[1*], Jaree Thongkam[2], Vatinee Sukmak[3]

## บทคัดย่อ

จิตเภท (Schizophrenia) เป็นโรคทางจิตเวชที่พบมากกว่าการเจ็บป่วยด้วยโรคจิตประเภทอื่นๆ และพบว่าผู้ป่วยเหล่านี้ส่วนใหญ่มักกลับมาเป็นซ้ำ (relapse) ต้องกลับเข้ารักษาในโรงพยาบาลเป็นระยะ ทำให้ผู้ดูแลและรัฐบาลต้องสูญเสียงบประมาณในการดูแลรักษาเป็นจำนวนมาก งานวิจัยนี้มีวัตถุประสงค์พัฒนาแบบจำลองในการพยากรณ์ระยะเวลาการรักษาซ้ำของผู้ป่วยโรคจิตเภทโดยเทคนิคเหมืองข้อมูล จากฐานข้อมูลโรงพยาบาลพระศรีมหาโพธิ์ จังหวัดอุบลราชธานี ปี พ.ศ. 2550 ถึงปี พ.ศ. 2555 จำนวน 2831 เรคคอร์ด โดยการแบ่งข้อมูลออกเป็น 2 คลาส คือ คลาส 0 เป็นกลุ่มของผู้ป่วยมารับการรักษาซ้ำใน 1 ถึง 28 วัน ส่วนคลาส 1 เป็นกลุ่มของผู้ป่วยมารับการรักษาซ้ำตั้งแต่ 29 ถึง 90 วัน แบบจำลองนี้สามารถช่วยในการวางแผนการรักษาของแพทย์ ผลการศึกษาพบว่าข้อมูลมีค่าผิดปกติ และมีความไม่สมดุลของคลาสในข้อมูลโดยมีจำนวนคลาสหนึ่งมากกว่าอีกคลาสหนึ่งเป็นจำนวนมาก ผู้วิจัยจึงได้ทำการแก้ปัญหาโดยคัดกรองเอาค่าผิดปกติออกด้วยเทคนิค Support Vector Machine และปรับความสมดุลของข้อมูลด้วยวิธี Synthetic Minority Over-sampling Technique (SMOTE) แล้วพัฒนาแบบจำลองด้วยเทคนิค ต้นไม้ตัดสินใจ (C4.5) นาอีฟเบย์ (Naïve Bayes) และPART decision list นอกจากนั้นผู้วิจัยยังได้ใช้ 10-fold cross validation ในการแบ่งข้อมูลออกเป็นชุดข้อมูลสอน และชุดข้อมูลทดสอบและได้ใช้ค่าความถูกต้อง ค่าความไว และค่าความจำเพาะในการแสดงประสิทธิภาพในการพยากรณ์ของแบบจำลองอีกด้วย ผลการทดลองประสิทธิภาพในการพยากรณ์ของแบบจำลองพบว่า เทคนิค PART decision list สามารถพยากรณ์ได้ดีกว่าต้นไม้ตัดสินใจ (C4.5) และนาอีฟเบย์ (Naïve Bayes) ซึ่งมีค่าความถูกต้อง (Accuracy) ร้อยละ 92.98 ค่าความไว (Sensitivity) ร้อยละ 92.95 และค่าความจำเพาะ (Specificity) ร้อยละ 93.02 ตามลำดับ

---

[1] นิสิตปริญญาโท, [2] ผู้ช่วยศาสตราจารย์ คณะวิทยาการสารสนเทศ, [3] รองศาสตราจารย์ คณะพยาบาลศาสตร์ มหาวิทยาลัยมหาสารคาม อำเภอกันทรวิชัย จังหวัดมหาสารคาม 44150

[1] Master degree student, [2] Assist. Prof., Faculty of Informatics, Mahasarakham University, Kantharawichai District, Maha sarakham 44150, Thailand. [3] Assoc. Prof., Faculty of Nursing, Mahasarakham University, Kantharawichai District, Maha sarakham 44150, Thailand.

* Corresponding author; Weerayut Mayusiri, Faculty of Informatics, Mahasarakham University, Kantharawichai District, Maha sarakham 44150, Thailand. Aae606@gmail.com

**คำสำคัญ:** ข้อมูลไม่สมดุล เหมืองข้อมูล โรคจิตเภท การรักษาซ้ำ

## Abstract

Schizophrenia is a mental disorder that has been found more frequently than other mental illness, Most patients often relapsed and treated in hospital which increase cost and budget for care-giver and government. This research aims to develop a model for predicting a period of time before relapsing by using data-mining technique.

The 2831 data were obtained from the Prasrimahabhodi hospital database, Ubon Ratchathani from year 2550 to year 2555. The data were divided into two classes. The class 0 refers to the group of patients readmitted within 28 days of discharge while the class 1 refers to the group of patients readmitted between 29 and 90 days. The model can assist doctor to plan their treatment by using applied data-mining technique. After analyzing the data, to problems were found including outlier and class imbalanced. In order to improve quality of data, the support vector machine technique and SMOTE were used to filtering and increase the minority class. Then decision trees (C4.5), Naïve Bayes and PART decision list were employed to build the prediction models. Moreover, 10-fold cross validation ware utilized to split the data into the training and test set. In addition, accuracy, sensitivity and specificity of prediction models were exploited to examine the performance of the model. Experimental result demonstrated that PART decision list superior to **Decision tree (C4.5) Naïve Bayes** with accuracy 92.98%, sensitivity 92.95% and specificity 93.02% respectively.

**Keyword:** Schizophrenia Readmission, Classification, Imbalanced data

## บทนำ

จิตเภท (Schizophrenia) เป็นโรคทางจิตเวชที่พบมากกว่าการเจ็บป่วยด้วยโรคจิตประเภทอื่นๆ จากสถิติของโรงพยาบาลจิตเวช สภาบันสุขภาพจิต พุทธศักราช 2541 พบว่าประเทศไทย มีผู้ป่วยจิตเภทมากกว่า 500,000 คนหรือร้อยละ 1 ของประชากรผู้ป่วยเป็นโรคจิตเภท ซึ่งมักเป็นการเจ็บป่วยเรื้อรังต้องได้รับการดูแลรักษาอย่างต่อเนื่อง แต่พบว่าผู้ป่วยเหล่านี้ส่วนใหญ่มักกลับมาเป็นซ้ำ (relapse) ต้องกลับเข้ารักษาในโรงพยาบาลเป็นระยะๆ ภายใน 28 หรือ 90 วัน ทำให้ผู้ดูแลและรัฐบาลต้องสูญเสียงบประมาณในการดูแลรักษาเป็นจำนวนมาก[1]

ปัจจุบันได้มีการประยุกต์ใช้เทคโนโลยีทางคอมพิวเตอร์ โดยใช้เทคนิคการทำเหมืองข้อมูลในการสร้างแบบจำลองเพื่อทำการวิเคราะห์ และพยากรณ์การเกิดโรคทางระบบประสาท ดังเช่นงานวิจัยของ Casey Bennett และคณะ[2] ได้ใช้การทำเหมืองข้อมูลบนฐานการศึกษาแบบสอบถามที่ใช้ในด้านสุขภาพจิต เพื่อช่วยในการสนับสนุนการตัดสินใจและการรักษาส่วนบุคคล โดยใช้เทคนิค Naïve Bayes, Neural Network, Random Forests, K-nearest Neighbors, Decision Trees และ Logistic Regression จากผลศึกษาพบว่า Naïve Bayes ให้ค่าความถูกต้องที่ดีที่สุดคิดเป็นร้อยละ 76.64 ส่วนวิจัยของ João Maroco และคณะ[3] ได้นำการทำเหมืองข้อมูลในการพยากรณ์

โรควิกลจริต โดยใช้เทคนิควิธี Neural Networks, Support Vector Machines และ Random Forest พบว่าการสร้างแบบจำลองด้วยเทคนิควิธี Support Vector Machines ให้ความถูกต้องดีกว่าวิธีอื่นๆ คิดเป็นร้อยละ76.0 และงานวิจัยของ Christine Howes และคณะ[4] ได้สร้างแบบจำลองเพื่อการพยากรณ์การติดตาม การรักษาโรคจิตเภทจากการรักษาโดยการทำจิตบำบัด โดยใช้เทคนิค Support Vector machine, Decision Tree พบว่า Decision Tree ให้ความถูกต้องที่ดีที่สุดคิดเป็นร้อยละ 90.0 งานวิจัยที่กล่าวมา อย่างไรก็ตามยังมีการศึกษาเกี่ยวกับการพัฒนาแบบจำลองเพื่อพยากรณ์ระยะเวลาที่จะกลับมารักษาซ้ำของผู้ป่วยโรคจิตเภทโดยใช้วิธีการทำเหมืองข้อมูลน้อย

ซึ่งเทคนิคเหล่านี้ไม่สามารถสร้างแบบจำลองที่มีประสิทธิภาพได้ถ้าข้อมูลมีค่าผิดปกติ และความไม่สมดุลของข้อมูล[5] โดยมีนักวิจัยหลายท่านได้ทำการแก้ปัญหา ค่าผิดปกติโดยการกรองเอาค่าผิดปกติออกโดยใช้วิธีการต่างๆ เช่น งานวิจัยของ Thongkam และคณะ[6] ได้ทำการวิเคราะห์กระบวนการก่อนการสร้างแบบจำลอง โดยใช้การกรองด้วย C-Support Vector Classification Filter (C-SVCF) และการปรับความสมดุลด้วย SMOTE เพื่อเพิ่มคุณภาพให้กับข้อมูล ผลปรากฏว่า การใช้การกรองด้วย SVM-C และSMOTE ทำให้ประสิทธิภาพในการพยากรณ์สูงขึ้น 23.31% โดยใช้ PART decision list ในการสร้างแบบจำลอง ส่วนงานวิจัยของเชาวนันท์ โสโท[7] ได้สร้างแบบจำลองการทำนายผลการรักษาผู้ป่วยมะเร็งปากมดลูกด้วยโครงข่ายประสาทเทียม ด้วยการนำวิธีการ Cost-Sensitive Learning: CSL และSynthetic Minority Over-sampling Technique: SMOTE มาทำการปรับสมดุลของข้อมูลแล้วทำการสร้างแบบจำลองด้วยโครงข่ายประสาทเทียมเปรียบเทียบวิธีถดถอยโลจิสติก และใช้ K-Fold Cross Validation ผลการทดลองว่าโครงข่ายประสาทเทียมที่มีการปรับความสมดุลของข้อมูลด้วยวิธีการ SMOTE มีประสิทธิภาพในการทำนายที่ดีกว่าวิธีอื่นโดยมีค่าความถูกต้องเท่ากับ 81.70% ค่าความไวเท่ากับ 94.47% และค่าความจำเพาะเท่ากับ 55.47%

ดังนั้น งานวิจัยนี้จึงได้ทำการศึกษาพัฒนาแบบจำลองเพื่อพยากรณ์ระยะเวลาที่จะกลับมารักษาซ้ำ โดยการเพิ่มคุณภาพของข้อมูล ของผู้ป่วยโรคจิตเภทโดยใช้การทำการกรองข้อมูลที่ผิดปกติด้วยเทคนิค C-Support Vector Classification Filter (C-SVCF) และปรับความไม่สมดุลของข้อมูลด้วยเทคนิค SMOTE ทำการสร้างแบบจำลองด้วยเทคนิค Decision Tree, Naïve Bayes และPART decision list โดยมีการวัดประสิทธิภาพของแบบจำลองโดยการวัดค่าความถูกต้อง ค่าความไวและค่าความจำเพาะ เพื่อนำแบบจำลองไปพยากรณ์หาสาเหตุการกลับมารักษาซ้ำของผู้ป่วยโรคจิตเภทและช่วยในการวางแผนการรักษาของแพทย์

## ทฤษฎีและเทคนิคที่เกี่ยวข้อง

### โรคจิตเภท

โรคจิตเภท (Schizophrenia) เป็นความผิดปกติที่ยังไม่ทราบสาเหตุแน่ชัด ผู้ป่วยส่วนใหญ่เริ่มแสดงอาการในช่วงวัยรุ่น เมื่อเป็นแล้วมักไม่หายขาด ส่วนใหญ่มีอาการกำเริบเป็นช่วงๆ โดยยังมีอาการทางจิตหลงเหลืออยู่บ้างในระหว่างนั้น อาการในช่วงกำเริบจะเป็นอาการแสดงของโรคจิตเภทอาจแบ่งออกเป็นสองกลุ่ม คือ กลุ่มอาการด้านบวก และกลุ่มอาการด้านลบ กลุ่มอาการด้านบวก เช่น ประสาทหลอน หลงผิด และในระยะหลังส่วนใหญ่มีกลุ่มอาการด้านลบ เช่น พูดน้อย เฉื่อยชา และแยกตนเอง โรคจิตเภทเป็นโรคจิตที่พบมากที่สุดในบรรดาโรคจิตทั้งหมด เกิดขึ้นได้กับบุคคลทุกเพศ ทุกระดับการศึกษา อาชีพ และฐานะ แม้ว่างานค้นคว้าวิจัยในต่างประเทศจะแสดงว่าประชาชนในระดับเศรษฐกิจและสังคมต่ำเป็นโรค

จิตเภทมากกว่ากลุ่มบุคคลระดับเศรษฐกิจและสังคมสูงก็ตาม ผลจากการสำรวจของสำนักงานสถิติแห่งชาติ พ.ศ. 2552 พบว่าประชาชนอายุ 15 ปีขึ้นไป มีปัญหาสุขภาพจิตร้อยละ 12 หรือประมาณ 5 ล้านคน ซึ่งคนกลุ่มนี้ยังไม่ใช่คนป่วย หากได้รับการดูแลส่งเสริมสุขภาพจิตก็จะทุเลาและหายเป็นปกติ ส่วนผู้ป่วยทางจิตที่ได้รับการดูแลรักษาในปี พ.ศ. 2551 ทั่วประเทศมีจำนวน 1,668,041 ราย มากที่สุดคือโรคจิตเภทมีจำนวน 445,840 ราย รองลงมาได้แก่ โรควิตกกังวลมีจำนวน 375,035 ราย โรคซึมเศร้า 199,667 ราย และที่เหลือเป็นโรคอื่นๆ เช่น ติดสารเสพติด โรคลมชัก ปัญญาอ่อน โดยเฉลี่ยผู้ป่วยทางจิตจะใช้เวลาในการรักษาในโรงพยาบาลเป็นเวลา 41 วัน สูงกว่าผู้ป่วยโรคทางกาย 5-6 เท่าตัว [8, 9]

**Outlier filtering**

การกรองข้อมูลผิดปกติเป็นการ จำแนกความผิดปกติ โดยใช้ค่าพิสัยตรวจจับตัวอย่างที่ผิดปกติโดยการใช้ k-Nearest Neighbors (K-NN)[10] ซึ่งข้อมูลจะเป็นข้อมูลในลักษณะกลุ่ม ส่วนการกรองข้อมูลด้วยเทคนิคต่างๆ ในเหมืองข้อมูลก็ได้มีการทำกันอย่างแพร่หลาย และปรากฏผลที่ดี เช่น Thongkam และคณะ[11] ได้แสดงเทคนิคการกรองที่มีประสิทธภาพสูงสุดคือการใช้เทคนิคซัพพอร์ทเวกเตอร์แมทชีน ในการคัดแยกและจัดการค่าผิดปกติของข้อมูลการรอดชีวิตของผู้ป่วยมะเร็งเต้านม ผลปรากฏว่าค่าความถูกต้องเพิ่มขึ้นเฉลี่ย 17.35% และคะแนนเฉลี่ยค่า AUC เพิ่มขึ้น 20.28% หลังกรองค่าผิดปกติออก 20% ดังนั้นงานวิจัยฉบับนี้จึงใช้เทคนิค ซัพพอร์ทเวกเตอร์แมทชีน (Support Vector Machine : SVM)[12] ซึ่งเป็นเทคนิคที่ใช้ระนาบการตัดสินใจในการแบ่งข้อมูลออกเป็นสองส่วนตามคลาส เพื่อใช้ในการกรองข้อมูลที่อยู่ในคลาสเดียวกันได้อย่างมีประสิทธิภาพ

**SMOTE**

SMOTE (Synthetic Minority Over-sampling Technique) เป็นเทคนิคในการแก้ปัญหาข้อมูลไม่สมดุล (Imbalanced data)[7] วิธีนี้เป็นการเพิ่มคลาสที่มีกลุ่มน้อยให้เพิ่มมากขึ้นเพื่อให้ใกล้เคียงกันกับอีกคลาส โดยวิธีนี้จะสุ่มเพิ่มชุดข้อมูลของคลาสที่มีจำนวนน้อย มีวิธีการดังนี้

1) คำนวณหาผลต่างระหว่างชุดข้อมูลที่พิจารณาและชุดข้อมูลที่ใกล้เคียง

2) คำนวณหาค่าข้อมูลใหม่ด้วยการคูณผลลัพธ์จากข้อ 1) ด้วยค่าที่ได้จากการสุ่มตัวเลขที่อยู่ในช่วง 0 ถึง 1

ยกตัวอย่างงานวิจัยที่นำเทคนิค SMOTE ไปใช้ เช่น งานวิจัยของ M. Mostafizur Rahman และ D. N. Davis[13] การจัดการกับปัญหาเกี่ยวกับคลาสไม่สมดุลในข้อมูลทางการแพทย์ ได้ใช้เทคนิค FURIA[14], Decision Tree ปรับความสมดุลของข้อมูลด้วยวิธีการ SMOTEพบว่าเทคนิค Decision Tree มีประสิทธิภาพในการทำนายที่ดีกว่าวิธีอื่น โดยมีค่าความถูกต้อง 85.78% ค่าความไวเท่ากับ 84.21% และค่าความจำเพาะเท่ากับ 87.34 หลังจากปรับความไม่สมดุลของข้อมูล พบว่าค่าความถูกต้องเพิ่มขึ้น 6.19% ถัดมาค่าความไวเพิ่มขึ้น 64.21% และค่าความจำเพาะลดลง 2.42%

**SubSampling**

Subsampling[15] เป็นเทคนิคที่มีการผลิต subsample โดยการสุ่มของชุดข้อมูลที่ชุดเดิมต้องพอดีกันอย่างสิ้นเชิงในหน่วยความจำ ตัวกรองนี้จะช่วยให้สามารถระบุสูงสุด ของการแพร่กระจายของข้อมูล ระหว่างชั้นที่หาได้ยากและที่พบมากที่สุด เมื่อใช้ในโหมดแบทช์ กระบวนการที่ตามมาจะไม่มีการสุ่มใหม่

งานวิจัยที่นำเทคนิค subsample ไปใช้ เช่น งานวิจัยของ Andrew McDowell และคณะ[16] การจัดการปัญหาความท้าทายของความไม่สมดุลของ class นอนหลับหรือตื่นนอนบนเตียงซึ่งมาจากพื้นฐานของการบันทึกการเคลื่อนไหวของนอนหลับแบบโดยอ้อม เทคนิค Spread Subsample

and Synthetic Minority Oversampling ถูกนำมาเปรียบเทียบ ผลปรากฏว่า การพัฒนาของการตรวจพบการตื่นตัวมากถึง 28% เมื่อเทียบกับข้อมูลพื้นฐานขณะที่รักษาความถูกต้องของตัวจำแนกทั้งหมด 90%

**เทคนิคเพื่อการพยากรณ์**

**กฎของเบย์ (Bayes's Rule)**

นาอีฟเบย์(Naïve Bayesian) [17, 18] ทำนายผลโดยหลักการจำแนกประเภทโดยทฤษฎีของเบย์ (Bayes Theorem) การเรียนรู้เบย์อย่างง่าย (Naïve Bayesian) เป็นวิธีจำแนกประเภทข้อมูลที่มีประสิทธิภาพวิธีหนึ่งเหมาะกับกรณีของเซตตัวอย่างมีจำนวนมากและคุณสมบัติ (Attribute) ของตัวอย่างไม่ขึ้นต่อกัน มีการจำแนกประเภทเบย์อย่างง่ายไปประยุกต์ใช้ในงานด้าน การวินิจฉัย (Diagnosis) และพบว่าให้ประสิทธิภาพที่ดี

**1. เทคนิคต้นไม้ตัดสินใจ**

ต้นไม้ตัดสินใจ (Decision Tree : C4.5)[19] เป็นการนำข้อมูลมาสร้างแบบจำลอง มีลักษณะเป็นโครงสร้างเหมือนต้นไม้เป็นการเรียนรู้แบบมีผู้สอน (Supervised learning) การสร้างโมเดล decision tree จะทำการคัดเลือกแอตทริบิวต์ที่มีความสัมพันธ์กับคลาสมากที่สุดขึ้นมาเป็นโหนดบนสุดของ tree (root node) หลังจากนั้นก็จะหาแอตทริบิวต์ถัดไปเรื่อยๆในการหาความสัมพันธ์ของแอตทริบิวต์นี้จะใช้ตัววัด ที่เรียกว่า GainRatio

**2. PART decision list**

PART decision list[20] เป็นกระบวนการใหม่สำหรับการนำกฎการตัดสินใจโดยการรวมทั้ง C4.5 และ RIPPER เพื่อ 1) หาและสร้างเซ็ตของกฎโดยการใช้ประโยชน์การเข้าถึงกฎการนำเข้า 2) เลือกใช้วิธีการแบ่งแยกและการได้มาเพื่อสร้างต้นไม้ย่อย และ 3) สร้างกฎจากต้นไม้ย่อย ถึงแม้กระบวนการนี้จะคล้ายกับกฎของ C4.5 แต่กระบวนการนี้ได้เลี่ยงการสร้างต้นไม้ตัดสินใจแบบสมบูรณ์ และนำไปสู่การปรับปรุง ระยะเวลาการสอน ซึ่งต่างจาก RIPPER ดังนั้น PART decision list จึงสามารถสร้างแต่ละกฎการตัดสินใจที่ซึ่งสอดคล้องกับใบไม้ด้วยการครอบคลุมที่ใหญ่ที่สุดในต้นไม้ตัดสินใจบางส่วน ด้วยวิธีการนี้ PART decision list สร้างและตัดกิ่งต้นไม้ตัดสินใจบางส่วน จัดการกับข้อมูลที่หายไป คุณลักษณะทางตัวเลขและที่ต่างกันและจัดให้มี เซ็ต กฎที่แม่นยำ[21]

## วิธีดำเนินการวิจัย

เหมืองข้อมูล คือ กระบวนการค้นหาวิธีการสร้างแบบจำลองและความสัมพันธ์ที่ซ่อนอยู่ในชุดข้อมูลจำนวนมหาศาลโดยอัตโนมัติ ซึ่งใช้ขั้นตอนวิธีการทางสถิติ การเรียนรู้ของเครื่อง และการรู้จำแบบ ซึ่งการทำเหมืองข้อมูลในการสร้างแบบจำลอง

ในงานวิจัยนี้ได้ดำเนินการมีขั้นตอนประกอบในการทำเหมืองข้อมูลมี 4 ขั้นตอน ดังต่อไปนี้ 1) รวบรวมข้อมูล 2) ปรับปรุงคุณภาพของข้อมูล 3) สร้างแบบจำลอง และ 4) การตรวจสอบประสิทธิภาพของ ผลลัพธ์แบบจำลอง แสดงใน Figure 1

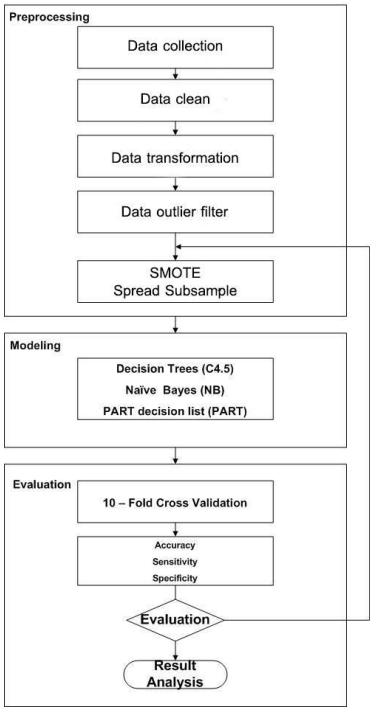


**Figure 1** Research Method

**การเตรียมข้อมูล (Data pre-processing)**

ขั้นตอนการเตรียมข้อมูลถือเป็นขั้นตอนที่สำคัญในการทำเหมืองข้อมูล ในงานวิจัยนี้ได้ใช้วิธีดังนี้

จำนวนข้อมูลทั้งสิ้นก่อนการปรับความสมดุลของข้อมูลแสดงให้เห็นดัง Table 1

**Table 1** Number of instances in the raw data set.

| Data set | | Total | Radio | |
|---|---|---|---|---|
| Class 0 | Class 1 | | Class 0 | Class 1 |
| 1927 | 901 | 2828 | 68.14 | 31.86 |

1) ข้อมูลผู้ป่วย ในฐานข้อมูลผู้ป่วยในโรงพยาบาลพระศรีมหาโพธิ์ จังหวัดอุบลราชธานี ระหว่าง ปี พ.ศ. 2550 ถึงปี พ.ศ. 2555 จำนวน 2831 เรคคอร์ด ใช้ข้อมูลผู้ป่วยที่มารับรักษาไว้เป็นผู้ป่วยในโรงพยาบาลครั้งแรกและจำนวนวันที่กลับมารักษาซ้ำเป็นผู้ป่วยในครั้งที่ 2

2) ตรวจสอบความถูกต้องของข้อมูล (Data cleaning)

3) ตรวจสอบข้อมูลที่ไม่สมบูรณ์ (Missing value) ในงานวิจัยนี้ได้ทำการลบเรคคอร์ดที่ข้อมูลไม่ครบสมบูรณ์ไปออกจำนวน 3 เรคคอร์ด เหลือข้อมูล จำนวน 2828 เรคคอร์ด

4) ทำการกรองข้อมูลอัตโนมัติด้วย C-SVCF ออกด้วยจำนวน 683 เรคคอร์ด เหลือข้อมูลจำนวน 2145 เรคคอร์ด ดัง Table 2

**Table 2** Number of instances in the data set after svm outlier filter.

| | Data set | | Total | Radio | |
|---|---|---|---|---|---|
| | Class 0 | Class 1 | | Class 0 | Class 1 |
| Raw+SVM | 1897 | 248 | 2145 | 88.44 | 11.56 |

5) ปรับความไม่สมดุลของข้อมูลด้วยวิธี SMOTE และ Spread Subsample (subSampling) ปรับค่าการเพิ่มจำนวนชุดข้อมูลของคลาสที่น้อย โดยปรับค่าพารามิเตอร์ผลการทดลองพบว่าขนาดชุดข้อมูลที่สามารถเพิ่มประสิทธิภาพของแบบจำลองได้ดีที่สุดคือร้อยละ 650 เพิ่มขึ้นจำนวน 1612 เรคคอร์ด รวมเป็นจำนวน 3757 เรคคอร์ด และปรับสมดุลของข้อมูลด้วยวิธี subsampling พบว่าขนาดชุดข้อมูลที่สามารถเพิ่มประสิทธิภาพของแบบจำลองได้ดีที่สุดคือร้อยละ 248 ลดลงจำนวน 1649 รวมเป็นจำนวน 496 เรคคอร์ด

ดังนั้นจำนวนข้อมูลทั้งสิ้นก่อนการสร้างแบบจำลองเป็น 3757 เรคคอร์ด และ 496 เรคคอร์ด ข้อมูลทั้งสอง ประกอบด้วย 2 คลาส คือ คลาส 0 คือข้อมูลผู้ป่วยกลับมารักษาซ้ำภายใน 28 วัน และคลาส 1 คือ กลับมารักษาซ้ำระหว่าง 29-90 วัน ซึ่งเป็นการกำหนดตามตัวชี้วัดด้านความเป็นเลิศในด้านการให้บริการของโรงพยาบาล โดยแสดงให้เห็นถึงคุณภาพในการรักษาผู้ป่วยของโรงพยาบาล ดัง Table 3

**Table 3** Number of instances in the data set.

| | Data set | | Total | Radio | |
|---|---|---|---|---|---|
| | Class 0 | Class 1 | | Class 0 | Class 1 |

| | | | | | |
|---|---|---|---|---|---|
| Raw+SVM+subSampling | 248 | 248 | 496 | 50 | 50 |
| Raw+SVM+SMOTE | 1897 | 1860 | 3757 | 50.49 | 49.51 |

งานวิจัยนี้ประกอบด้วยตัวแปรทั้งหมด 11 ตัวแปรดัง Table 4

**Table 4** The attribute list.

| No. | Attributes | Description | Data type |
|---|---|---|---|
| 1 | Gender | Gender | Nominal |
| 2 | Age | Age | Numeric |
| 3 | Marital_Status | Marital_status | Nominal |
| 4 | Occupation | Occupation | Nominal |
| 5 | Province | Province | Nominal |
| 6 | Cause_readmis | Cause_readmis | Nominal |
| 7 | Diag | diagnosis | Nominal |
| 8 | Edu | Education | Nominal |
| 9 | R | Right | Nominal |
| 10 | Admis_no | Day in Admission | Nominal |
| 11 | Come_no | Classes (0,1) | Nominal |

**สร้างแบบจำลอง**

งานวิจัยนี้ได้ใช้เครื่องมือคือโปรแกรม Weka 3.7.1 ในการทำงานโดยเลือกใช้เทคนิควิธีดังนี้ 1) ต้นไม้ตัดสินใจ 2) การเรียนรู้แบบเบย์ 3) PART decision list

**การวัดประสิทธิภาพ**

ในงานวิจัยครั้งนี้ 10 – Fold Cross Validation ได้ถูกนำมาใช้เพื่อให้ข้อมูลทุกตัวมีโอกาสเป็นชุดทดสอบและชุดสอน โดยแบ่งข้อมูลออกเป็นสองชุด คือ ชุดข้อมูลสอน (Training Data) กับชุดข้อมูลทดสอบ (Testing Data) โดยแบ่งข้อมูลให้มีจำนวนเท่ากันออกเป็น 10 ส่วน ใช้ 9 ส่วนเป็นชุดข้อมูลสอน และที่เหลือ 1 ส่วน เป็นชุดข้อมูลทดสอบ จะทำทั้งหมด 10 รอบ จากนั้นสลับกันจนครบ 10 รอบ ซึ่งในการคำนวณประสิทธิภาพของแบบจำลองได้ใช้ Confusion Matrix คือ ตารางสรุปจำนวนข้อมูลที่ตัวแบบมีการจำแนกได้อย่างถูกต้องและไม่ถูกต้อง ผลการทดลองสามารถแสดงใน Figure 2

| | | Predicted | |
|---|---|---|---|
| | | Positive | Negative |
| Actual | Positive | TP | FN |
| | Negative | FP | TN |

**Figure 2** The confusion matrix

จาก Figure 2 สามารถคำนวณหาค่าความแม่นยำ (Accuracy) ค่าความอ่อนไหว (Sensitivity) และค่าความจำเพาะ (Specificity) ตามสมการที่ 1 สมการที่ 2 และสมการที่ 3 ตามลำดับ

$$Accuracy = \frac{TP + TN}{TP + FP + FN + TN} x100 \quad (1)$$

$$Sensitivity = \frac{TP}{TP + FN} x100 \quad (2)$$

$$Specificity = \frac{TN}{TN + FP} x100 \quad (3)$$

## ผลการดำเนินงานวิจัย

ในงานวิจัยนี้ได้นำการกรองข้อมูล SVM และวิธีการ SMOTE มาทำการปรับสมดุลของข้อมูลและสร้างแบบจำลองด้วยเทคนิค ต้นไม้ตัดสินใจ การเรียนรู้แบบเบย์ และ PART decision list เพื่อการพยากรณ์ระยะเวลาการกลับมารักษาซ้ำของผู้ป่วยโรคจิตเภทโดยเทคนิคเหมืองข้อมูลโดยข้อมูลจากโรงพยาบาลพระศรีมหาโพธิ์ จังหวัดอุบลราชธานี ในปี พ.ศ.2550 ถึงปี พ.ศ.2555 โดยใช้ 10 – Fold Cross Validation และค่าความถูกต้อง ค่าความไว และค่าความจำเพาะ เพื่อวัดประสิทธิภาพของแบบจำลอง

1. ค่าความถูกต้อง (Accuracy) เป็นการวัดประสิทธิภาพของแบบจำลองเพื่อพยากรณ์ ใช้ในการแสดงประสิทธิภาพในการพยากรณ์ของแบบจำลองโดยรวม เทคนิคที่นำมาใช้ในการสร้างแบบจำลองโดยรวม คือ 1) C4.5 2) Naïve Bayes และ 3) PART decision list ซึ่งสามารถสรุปผลค่าความถูกต้องได้ดัง Figure 3

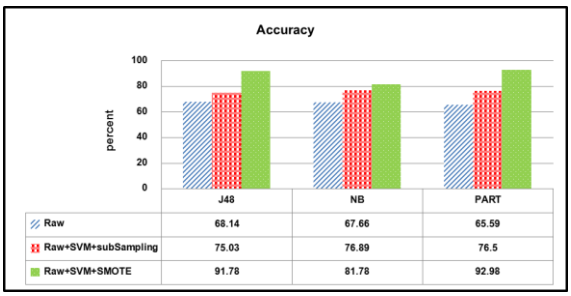


**Figure 3** Accuracy Comparisons

จาก Figure 3 แสดงค่าความถูกต้องของแบบจำลองที่ได้จากเทคนิค C4.5 Naïve Bayes และ PART decision list จากผลการทดลองพบว่าหลังจากกรอง และปรับความสมดุลของข้อมูลด้วย SMOTE แล้วทำให้ค่าความถูกต้องของแบบจำลองสูงขึ้นถึง C4.5 สูงขึ้นถึง 23.64% ส่วน Naïve Bayes สูงขึ้น 14.12% และ PART decision list สูงขึ้น 27.39% เห็นได้ว่าความแม่นยำของแบบจำลองเทคนิค PART decision list ให้ความถูกต้องสูงสุดร้อยละ 92.98 ตามมาด้วยเทคนิคต้นไม้ตัดสินใจ (C4.5) ให้ความถูกต้องคิดเป็นร้อยละ 91.78 และแบบจำลองที่สร้างด้วยเทคนิคนาอีฟเบย์ (NB) ให้ค่าความถูกต้องน้อยสุดคิดเป็นร้อยละ 81.78 ซึ่งเป็นการปรับสมดุลด้วย SMOTE จะสามารถทำให้ประสิทธิภาพของแบบจำลองสูงกว่าการปรับสมดุลด้วย subSampling

2. ค่าความไว (Sensitivity) เป็นการวัดประสิทธิภาพของแบบจำลองเพื่อการพยากรณ์การกลับมารักษาซ้ำภายใน 28 วัน โดยใช้เทคนิค C4.5, Naïve Bayes และ PART decision list ซึ่งสามารถสรุปผลค่าความไวได้ดัง Figure 5

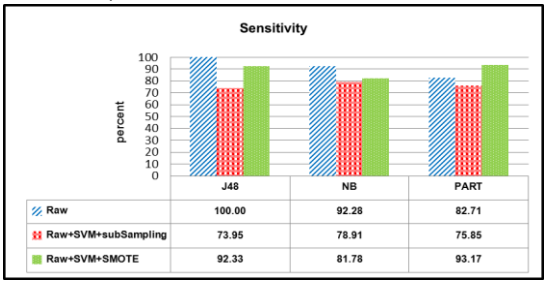


**Figure 4** Sensitivity Comparisons

จาก Figure 4 แสดงถึงค่า ความไว (Sensitivity) ของแบบจำลองที่ได้จากเทคนิค C4.5 Naïve Bayes และ PART decision list จากผลการทดลองพบว่าหลังจากกรอง และปรับความสมดุลของข้อมูลด้วย SMOTE จะเห็นว่าค่าความไวของแบบจำลองจะพยากรณ์เป็นการกลับมารักษาซ้ำ 1-28 วัน เพราะคลาส 1-28 มีจำนวนคลาสมากกว่าคลาส 29-90 ในอัตราส่วน 68.14% เห็นได้ว่าเทคนิคส่วนใหญ่มีความลำเอียงไปทางข้อมูลที่มาก แล้วทำให้ประสิทธิภาพของแบบจำลองลดลง C4.5 ลดลง 7.67% ส่วน Naïve Bayes ลดลง 10.51% และ PART decision list สูงขึ้น 10.45% เห็นได้ว่าแบบจำลอง เทคนิค PART decision list ให้ค่าความไวสูงสุดที่ร้อยละ 93.17 ลำดับถัดมาคือเทคนิคต้นไม้ตัดสินใจ (C4.5) ให้ค่าความไวที่ร้อยละ 91.17 และถัดมาได้แก่ เทคนิคนาอีฟเบย์ (NB) ให้ค่าความไวน้อยสุดที่ร้อยละ 81.78

3. ค่าความจำเพาะ (Specificity) เป็นการวัดประสิทธิภาพของแบบจำลองเพื่อพยากรณ์การกลับมารักษาซ้ำจาก 29-90 วัน โดยใช้เทคนิค C4.5, Naïve Bayes และ PART decision list ซึ่งสามารถสรุปผลค่าความจำเพาะได้ดัง Figure 5

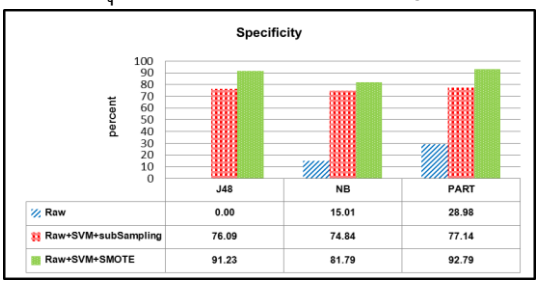


**Figure 5** Specificity Comparisons

จาก Figure 5 แสดงค่าความจำเพาะ (Specificity) ของแบบจำลองที่ได้จากเทคนิค C4.5 Naïve Bayes และ PART decision list จากผลการทดลองพบว่าหลังจากกรอง และปรับความสมดุลของข้อมูลด้วย SMOTE แล้วทำให้ค่าความจำเพาะของแบบจำลองสูงขึ้นถึง C4.5 สูงขึ้นถึง 91.23% ส่วน Naïve Bayes สูงขึ้น 66.78%

และ PART decision list สูงขึ้น 63.81% จะเห็นว่า PART decision list ให้ค่าความจำเพาะสูงสุดที่ร้อยละ 92.79 ถัดมาคือเทคนิค ต้นไม้ตัดสินใจ (C4.5) ให้ค่าความจำเพาะที่ร้อยละ 91.23 และเทคนิคนาอีฟเบย์ (NB) ให้ค่าความจำเพาะน้อยสุดที่ร้อยละ 81.79 ซึ่งเป็นการปรับสมดุลด้วย SMOTE จะสามารถทำให้ประสิทธิภาพของแบบจำลองสูงกว่าการปรับสมดุลด้วย subSampling

## สรุปผลการวิจัย

งานวิจัยฉบับนี้ มีวัตถุประสงค์เพื่อสร้างแบบจำลองเพื่อการพยากรณ์ระยะเวลาการกลับมารักษาซ้ำข้อมูลผู้ป่วยโรคจิตเภทโดยเทคนิคเหมืองข้อมูล จากผลการวิจัยสามารถสรุปได้ว่า ข้อมูลการรักษาของผู้ป่วยโรคจิตเภททางการแพทย์มีข้อมูลที่มีความผิดปกติและมีความไม่สมดุลของข้อมูล (Imbalanced Data Set) ซึ่งผู้วิจัยได้ใช้วิธีการกรองและ SMOTE ในการแก้ปัญหา จากการทดลองผู้วิจัยพบว่า

1.การกรองและการทำ SMOTE สามารถเพิ่มประสิทธิภาพให้แบบจำลองเพิ่มขึ้น โดยค่าความถูกต้องเพิ่มขึ้นเฉลี่ยร้อยละ 46.36 ถัดมาค่าความไวเพิ่มขึ้นเฉลี่ยร้อยละ 20.05 และค่าความจำเพาะเพิ่มขึ้นร้อยละ 32.69 ซึ่งมากกว่า subSampling

2. PART decision list สามารถเพิ่มประสิทธิภาพให้แบบจำลองเพิ่มขึ้น โดยค่าความถูกต้องเพิ่มขึ้นเฉลี่ยร้อยละ 27.39 ถัดมาค่าความไวเพิ่มขึ้นเฉลี่ยร้อยละ 10.45 และค่าความจำเพาะเพิ่มขึ้นร้อยละ 63.81 ทำให้เทคนิค PART decision list มีประสิทธิภาพในการทำนายได้สูงที่สุด โดยมีค่าความถูกต้อง (Accuracy) เท่ากับร้อยละ 92.98 ค่าความไว (Sensitivity) ที่ร้อยละ 93.17 และค่าความจำเพาะ (Specificity) อยู่ที่ร้อยละ 92.79

## กิตติกรรมประกาศ

ขอขอบคุณโรงพยาบาลพระศรีมหาโพธิ์จังหวัดอุบลราชธานี ที่ให้ความอนุเคราะห์ข้อมูลในการศึกษาวิจัยในครั้งนี้

## เอกสารอ้างอิง

1. บุญวดี เพชรรัตน์. "งานวิจัยเรื่องปัจจัยที่สัมพันธ์กับพฤติกรรมการดูแลผู้ป่วยจิตเภทเรื้อรังที่บ้านของผู้ดูแล". 2552.
2. Bennett C, Doub T, Bragg A, Luellen J, Van C, Regenmorter, et al. "Data Mining Session-Based Patient Reported Outcomes (PROs) in a Mental Health Setting: Toward Data-Driven Clinical Decision Support and Personalized Treatment". IEEE; San Jose, CA; 229-36, 2011.
3. Maroco J, Silva D, Rodrigues A, Guerreiro M, Santana I, Mendonça Ad. "Data mining methods in the prediction of Dementia: A real-data comparison of the accuracy, sensitivity and specificity of linear discriminant analysis, logistic regression, neural networks, support vector machines, classification trees and random forests". BMC Research Notes, 4(1):pp.1-14 2011.
4. Howes C, Purver M, McCabe R, Healey PGT, Lavelle M. "ACM". ACM; pp.79-83, Predicting Adherence to Treatment for Schizophrenia from Dialogue Transcripts.
5. Chawla NV, Bowyer KW, Hall LO, Kegelmeyer WP. "SMOTE: Synthetic Minority Over-sampling Technique". Journal of Artificial Intelligence Research, 16:321–57 2002.
6. Thongkam J, Sukmak V. "Improving quality of breast cancer data through pre-

processing ". KKU Engineering Journal, 40:493-504 2013.

7. เชาวนันท์ โสโท. "แบบจำลองการทำนายผลการรักษาผู้ป่วยมะเร็งปากมดลูกด้วยโครงข่ายประสาทเทียม". KKU Res J (GS), 13:39-49 2013.
8. Matichon online. "เว็บไซต์". ได้จาก: http://www.matichon.co.th/news_detail.php?newsid=1282457462&catid=04
9. มานิต ศรีสุรภานนท์. "ปัจจัยเสี่ยงของโรคจิตเภท : การทบทวนวรรณกรรมทางระบาดวิทยา". ได้จาก: http://www.dmh.go.th/abstract/details.asp?id=3053
10. AHA DW, KIBLER D, ALBERT MK. "Instance-Based Learning Algorithms". Macine Learning, 6:37-66 1991.
11. Thongkam J, Xu G, Zhang Y, Huang F. "Support Vector Machine for Outlier Detection in Breast Cancer Survivability Prediction". Springer-Verlag:99-109 2008.
12. เจษฎา เทโวขัติ. "เว็บสไปเดอร์แบบจำเพาะกลุ่มเป้าหมายโดยอาศัยหลักทางเอสวีเอ็ม". 2553.
13. Rahman MM, Davis DN. "Addressing the Class Imbalance Problem in Medical Datasets". International Journal of Machine Learning and Computing, 3(2):224-8 2013.
14. Hühn J, Hüllermeier E. "FURIA: an algorithm for unordered fuzzy rule induction". 19(3):pp 293-319 2009.
15. Pooja, Ratnoo S. "A Comparative Study of Instance Reducetion Techniques". International Journal of Advances in Engineering Sciences Vol3 (3), July, 2013, 3(3):7-13 2013.
16. Andrew McDowell, Mark P. Donnelly, Chris D. Nugent, Member, IEEE, Leo Galway , et al. "Addressing the challenges of sleep/wake class imbalance in bed based non-contact actigraphic recordings of sleep". Engineering in Medicine and Biology Society (EMBC), 2013 35th Annual International Conference of the IEEE Page(s): 4654 - 7, 2013.
17. Liu Jie, Bo S. "Naive Bayesian Classifier Based on Genetic Simulated Annealing Algorithm". 504–9, 2011.
18. ธนกร เจริญเชาว์. "ระบบสนับสนุนการเสนอขายคอนโดมิเนียม โดยใช้เทคนิคนาอีฟเบย์เซียน". 2554.
19. ศักดิ์ชาย ตั้งประเสริฐ. "การพัฒนาระบบจำแนกประเภทแบบทดสอบสำหรับผู้ทดสอบสุขภาพจิตด้วยเทคนิค Decision Tree ผ่าน Web Application แบบ AJAX". 2550.
20. Frank E, Witten IH. "Generating Accurate Rule Sets Without Global Optimization". Fifteenth International Conference on Machine Learning; 144-51, 1998.
21. Thabtah F, Cowling P. "Mining the data from a hyperheuristic approach using associative classification". Expert Systems with Applications, 34:1093–101 2008.