19th International Computer Science and Engineering Conference (ICSEC)
Chiang Mai, Thailand, 23-26 November, 2015

การเปรียบเทียบเทคนิคเหมืองข้อมูลสำหรับการสร้างตัวแบบคัดกรองความเสี่ยงที่จะเกิดความเจ็บป่วยและเสียชีวิตของมารดาที่ตั้งครรภ์จากข้อมูลองค์การอนามัยโลก 2007-2008

A comparison of data mining technique for risk classification model of maternal mortality and morbidity from WHO global survey 2007-2008

ศรุต แสงอรุณ[1], มัลลิกา วัฒนะ[2]
[1,2]ภาควิชาวิทยาการคอมพิวเตอร์ คณะวิทยาศาสตร์ มหาวิทยาลัยขอนแก่น ขอนแก่น
E-mail: s.sarut@kkumail.com[1], monlwa@kku.ac.th[2]

บทคัดย่อ

จากการสำรวจขององค์การอนามัยโลกพบว่าผู้หญิงมากกว่า 500,000 คน ทั่วโลกต้องเสียชีวิตจากภาวะแทรกซ้อนในขณะตั้งครรภ์ เนื่องจากการตั้งครรภ์นั้นมีความเสี่ยงที่ทำให้ร่างกายอ่อนแอ แล้วเกิดภาวะแทรกซ้อนได้ง่าย ดังนั้นงานวิจัยนี้ได้เสนอการเปรียบเทียบโมเดลเพื่อนำมาใช้ในการทำนายความเสี่ยงจากการตั้งครรภ์ โดยใช้ข้อมูลผลของการตั้งครรภ์จากองค์กรอนามัยโลกปี 2007 – 2008 ซึ่งเป็นข้อมูลที่ได้มีการจัดเก็บในประเทศไทย เพื่อใช้สร้างโมเดลการทำนายโดยได้ใช้เทคนิคการทำนาย 3 รูปแบบ คือ C4.5, Multilayer Perceptron และ Naïve Bayes ผลลัพธ์ที่ได้จากการทดสอบนั้นจะนำไปหาค่า G-mean เนื่องจากค่า G-mean จะใช้วัดประสิทธิภาพของข้อมูลที่มีความไม่สมดุล จากการทดลองพบว่าค่า G-mean ของ C4.5 เท่ากับ 50.85% ค่า G-mean ของ Multilayer Perceptron เท่ากับ 49.46% และค่า G-mean ของ Naïve Bayes เท่ากับ 59.32% ซึ่งเป็นผลลัพธ์ดีที่สุด ดังนั้นเทคนิคของ Naïve Bayes เป็นเทคนิคที่เหมาะสมในการนำไปประยุกต์เป็นเครื่องมือในการทำนายเพื่อป้องกันความเสี่ยงที่จะทำให้ผู้หญิงที่ตั้งครรภ์เจ็บป่วยหรือเสียชีวิต

คำสำคัญ: การจำแนกประเภท, โครงข่ายประสาทเทียมแบบเพอร์เซปตรอนหลายชั้น, C4.5, Naïve Bayes, การเจ็บป่วยและเสียชีวิตขณะตั้งครรภ์, ข้อมูลผลของการตั้งครรภ์จากองค์การอนามัยโลก 2007-2008

Abstract

The survey of WHO found that there were more than 500,000 women around the world who died from complications while in pregnancy, because the pregnancy risks from being physically weak more easily causes mortality. Therefore, this research proposed a model for classification and prediction of pregnancy morbid mortality. The data was applied in the research from GSA WHO in 2007-2008 that was collected from pregnant women in Thailand. The research compared the 3 techniques for classification: C4.5, Multilayer Perceptron and Naïve Bayes. The result of the experiment used the G-mean to analyze the performance measurement imbalance data. Results were as follows: G-mean of C4.5 was 50.85%, G-mean of Multilayer Perceptron was 49.46 %, and G-mean of Naïve Bayes was 59.32%. The G-Mean of Naïve Bayes was the best value. Therefore, Naïve Bayes could be the proper technique to be applied to protect risks or mortality of women in pregnancy.

Keywords: Classification, Multilayer Perceptron, C4.5, Naïve Bayes, maternal mortality and morbidity, WHO Global Survey

1. บทนำ

ทั่วโลกมีผู้หญิงมากกว่า 45 ล้านคนที่ตั้งครรภ์และมากกว่า 500,000 คนได้เสียชีวิต เนื่องจากเกิดภาวะแทรกซ้อนในขณะตั้งครรภ์และคลอดบุตร [1,2] นอกจากนี้ 90% มารดาที่เจ็บป่วยและเสียชีวิตจากการตั้งครรภ์มาจากประเทศที่กำลังพัฒนา เพราะกลุ่มประเทศที่กำลังพัฒนาขาดแคลนทรัพยากรและการบริการด้านสุขภาพที่ดี สำหรับประเทศไทยอยู่ในกลุ่มประเทศที่กำลังพัฒนา ซึ่งขาดข้อมูลผลของการตั้งครรภ์ที่น่าเชื่อถือ ขาดการวางแผนดูแลรักษาที่ทันสมัย และการประเมินผลที่ดี โดยข้อมูลผลของการตั้งครรภ์นี้จำเป็นต่อการบริหารนโยบายด้านการดูแลสุขภาพอนามัยของมารดาและทารกแรกคลอด

สำหรับความสำคัญของการวัดการเสียชีวิตของมารดา[3] ได้บ่งบอกถึงการดูแลและการบริการด้านสุขภาพ โดยข้อมูลการเสียชีวิตของมารดาจะแสดงให้ทราบถึงปัจจัยเสี่ยงของการตั้งครรภ์และคลอดบุตร การเก็บข้อมูลจึงจำเป็นต้องใช้การเก็บข้อมูลของผู้ที่อยู่ในช่วงวัยเจริญพันธุ์ โดยจากวิธีการสำรวจข้อมูลนั้นได้ใช้วิธีการสำรวจข้อมูลแบบ Reproductive Age Mortality Survey (RAMOS)[4,5] ซึ่งเป็นวิธีสำรวจที่ดีที่สุดในการวัดการเสียชีวิตของมารดา คือการสอบสวนหาสาเหตุจากวัยเจริญพันธุ์ทั้งหมด โดยเป็นการหาข้อมูลจากหลายแหล่งข้อมูล เช่น ทะเบียนราษฎร์ บันทึกของสถานบริการ ผู้นำชุมชน เด็กนักเรียน เพื่อนำมาหาข้อมูลของหญิงวัยเจริญพันธุ์ที่เสียชีวิตทั้งหมด วิธีนี้ใช้เวลาและความชำนาญเป็นอย่างมาก เนื่องจากการเก็บข้อมูลนั้นจะต้องเข้าใจปัจจัยที่เกี่ยวข้องกับความเสี่ยงของสตรีที่ตั้งครรภ์ในระยะเจ็บคลอด[6] ซึ่งการประเมินภาวะแทรกซ้อน โดย

การประเมินเบื้องต้นจากข้อมูล และการตรวจร่างกายเพื่อคัดกรองของภาวะเสี่ยง

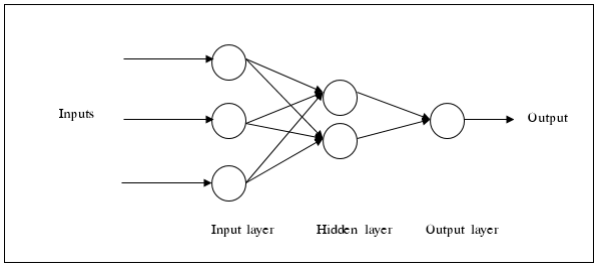
จากข้อมูลที่ได้ศึกษาเกี่ยวกับสุขภาพมารดาและทารกแรกเกิดทั้งในประเทศและระดับโลก สุขภาพมารดานั้นจะทำให้ทราบถึงสถานการณ์ในขณะตั้งครรภ์ ซึ่งข้อมูลการตรวจสุขภาพมารดานั้น เราสามารถนำไปใช้ในการพยากรณ์หรือดูแนวโน้มความเสี่ยงที่จะเกิดภาวะแทรกซ้อน เพื่อนำไปสู่การดูแลงานด้านอนามัยแม่และเด็กต่อไป

งานวิจัยนี้ จึงใช้เทคนิคเหมืองข้อมูล ในการวิเคราะห์ความเสี่ยงที่ทำให้มารดาเจ็บป่วยหรือเสียชีวิตขณะตั้งครรภ์หรือหลังคลอดบุตร โดยข้อมูลที่นำมาใช้ในการวิเคราะห์เพื่อสร้างโมเดล ได้จากองค์การอนามัยโลก ปี 2007-2008 เป็นข้อมูลที่ได้มีการจัดเก็บในประเทศไทย เนื่องจากข้อมูลที่ได้มานั้นเกิดความไม่สมดุล คือจำนวนข้อมูลที่อยู่ในแต่ละคลาสแตกต่างกันมาก ดังนั้นผู้วิจัยจึงใช้การแก้ปัญหาความไม่สมดุลของข้อมูล โดยใช้เทคนิค Synthetic Minority Over-sampling Technique - Tomek Links (SMOTE-TL)[14] ในการเตรียมข้อมูลเพื่อเพิ่มประสิทธิภาพผลการคัดกรองความเสี่ยงให้ถูกต้องมากขึ้น

2. วรรณกรรม

2.1 โครงข่ายประสาทเทียมแบบเพอร์เซปตรอนหลายชั้น (Multilayer Perceptron)

โครงข่ายประสาทเทียม[7,8]เป็นการทำงานแบบเครือข่ายเซลล์ประสาทที่มีหน้าที่รู้จำ ประมวลผล หรือใช้ช่วยในการตัดสินใจ โดยโครงข่ายประสาทเทียมประกอบด้วย 3 ชั้นคือ ชั้นข้อมูลเข้า ชั้นซ่อน และชั้นผลลัพธ์



รูปที่ 1 สถาปัตยกรรมโครงข่ายประสาทเทียมแบบเพอร์เซปตรอนหลายชั้น [7]

โครงข่ายประสาทเทียมแบบเพอร์เซปตรอนหลายชั้นนี้เหมาะกับการใช้ข้อมูลที่ซับซ้อนได้เป็นอย่างดี โดยมีกระบวนการฝึกฝนเป็นแบบมีการเรียนรู้ แบบมีการควบคุม และใช้ขั้นตอนการส่งค่าย้อนกลับ การตัดสินใจจะมีการเปลี่ยนแปลงค่าให้เหมาะสมและแก้ไขข้อผิดพลาดเพื่อให้ได้ค่าที่เข้าใกล้เป้าหมายที่แท้จริง

2.2 Naïve Bayes

Naïve Bayes[9] เป็นเทคนิคที่ใช้ทฤษฎีของ Bayes Theorem โดยจะสามารถแสดงเหตุการณ์ต่าง ๆ ที่ใช้ในการจัดกลุ่มที่มีอิสระต่อกัน โดยจะวิเคราะห์ความสัมพันธ์ระหว่างตัวแปรอิสระและตัวแปรตาม เพื่อสร้างเงื่อนไขความน่าจะเป็นของเหตุการณ์

$$P(H|E) = \frac{P(E|H) \times P(H)}{P(E)} \quad (1)$$

เมื่อ P(H) คือ ความน่าจะเป็นที่เกิดเหตุการณ์ H

P(E) คือ ความน่าจะเป็นที่เกิดเหตุการณ์ E

P(H|E) คือ ความน่าจะเป็นที่เกิดเหตุการณ์ H เมื่อเกิดเหตุการณ์ E

P(E|H) คือ ความน่าจะเป็นที่เกิดเหตุการณ์ E เมื่อเกิดเหตุการณ์ H

2.3 วิธีการต้นไม้ตัดสินใจ C4.5

C4.5 คือ การทำต้นไม้ตัดสินใจวิธีการหนึ่ง[10] ได้พัฒนามาจาก ID3[11] ขั้นตอนการสร้างต้นไม้ตัดสินใจนั้น จะนำค่าที่ได้มาใช้ในการตัดสินใจว่าจะใช้ตัวแปรใดเพื่อใช้แบ่งข้อมูลเพื่อนำมาใช้ในการสร้างต้นไม้ตัดสินใจโดยใช้ค่าเกนมาตรฐาน เลือกตัวแปรเพื่อกำหนดให้เป็นค่าของโหนดหรือค่าของราก ซึ่งวิธีนี้เหมาะสมกับข้อมูลที่มีตัวแปรที่ใช้ในการตัดสินใจเป็นขั้นตอนหลายขั้นตอน โดยการเลือกแอตทริบิวต์จะใช้มาตรฐานอัตราส่วนเกน (Gain raio) มาใช้ในการคัดเลือก โดยถ้าชุดข้อมูล M ประกอบด้วยค่าที่เป็นไปได้ คือ $\{m_1, m_2, ..., m_n\}$ และความน่าจะเป็นที่จะเกิดค่า m_1 มีค่าเท่ากับ $P(m_1)$ สามารถเขียนได้ดังสมการ 2

$$Info(M) = -\sum_{i=1}^{n} P(m_1) \times \log_2 P(m_1) \quad (2)$$

และถ้า T เป็นข้อมูลที่นำมาฝึกสอน และ x คือ คุณลักษณะที่เป็นโหนด โดยโหนดปัจจุบันจะแบ่งค่า T ออกเป็นกิ่ง ดังสมการ 3

$$Info_x(T) = \sum_{i=1}^{n} \frac{|t_i|}{|T|} \times Info(t_i) \quad (3)$$

ดังนั้น ค่าเกน (Data Gain) ที่ได้จากการแยกข้อมูลดังสมการ 4

$$Gain(x) = Info(T) - Info_x(T) \quad (4)$$

การแก้ไขความลำเอียงโดยการการปรับค่าเกนให้ถูกต้องโดยใช้ค่า Split information ของแต่ละแอตทริบิวต์เพื่อให้คำนวณค่ามาตรฐานดังสมการที่ 5

$$SplitInfo(T) = \sum_{i=1}^{n} \frac{|t_i|}{|T|} \times \log\left(\frac{|t_i|}{T}\right) \quad (5)$$

คำนวณค่ามาตรฐานเกนได้จากสมการ 6

$$GainRatio(T) = \frac{Gain(x)}{SplitInfo(x)} \quad (6)$$

2.4 การแก้ปัญหาความไม่สมดุลของข้อมูล

ปัญหาการจำแนกข้อมูลที่พบส่วนใหญ่ คือ ปัญหาที่เกิดจากความไม่สมดุลของกลุ่มข้อมูล ซึ่งจะเรียกปัญหานี้ว่า ปัญหาความไม่สมดุล ซึ่งข้อมูลที่มีความไม่สมดุลนั้นจะมีค่าของตัวแปรตัวหนึ่งน้อยกว่าอีกค่าของตัวแปรเดียวกัน โดยการแก้ปัญหานั้นจะใช้วิธีการ Synthetic Minority Over-sampling Technique (SMOTE)[12] เพื่อทำการเพิ่มค่าของตัว

แปรที่น้อยกว่าโดยการสังเคราะห์การแทนที่ชุดข้อมูลที่มีน้อย ให้ใกล้เคียงกับชุดข้อมูลที่มีมากในตัวแปรเดียวกัน ทั้งนี้การเพิ่มค่าที่ต้องการเลือกค่าของเพื่อนบ้านที่ใกล้ที่สุดเพื่อทำการสังเคราะห์

วิธีของ Tomek links[13] เป็นวิธีหนึ่งที่ใช้ในการแก้ปัญหาความไม่สมดุลของข้อมูล โดยวิธีการนี้จะเป็นการลดค่าของตัวที่มีค่ามากกว่าให้มีค่าใกล้เคียงอีกค่าที่มีค่าน้อยกว่า โดยมีวิธีการดังนี้

```
Algorithm TomekLinks (d is Dataset)
Data: d is the training data set
Result: Collection of Tomek links represented as pairs of examples
var
      setTomek is PairsExamples
      exMin, exMaj, ex is TrainingExamples
      dst is double
forall example of the majority class exMaj in d do
      forall example of the minority class exMin in d do
            dst = dist(exMin, exMaj)
            if ¬∃ex ∈ d|dist(ex, exMin)<dst ∨dist(ex, exMaj) <
dist then
                  cjtTomek := addLink(cltTomek;<exMin, exMaj>)
            end if
      end for
end for
```

รูปที่ 2 ขั้นตอนวิธี Tomek Links[13]

จากรูปที่ 2 สามารถอธิบายอัลกอริทึม ได้ดังนี้

(1) ตัวอย่างให้ E_i และ E_j อยู่ในคลาสที่ต่างกันและ $d(E_i,E_j)$ คือระยะห่างของ E_i และ E_j

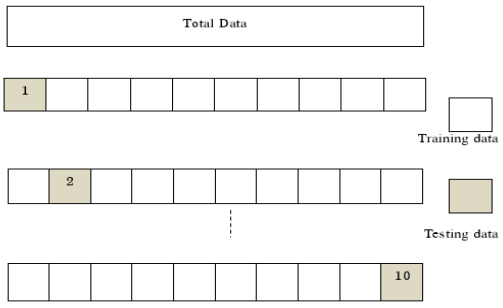
(2) การจับคู่ของ (E_i, E_j) จะถูกเรียกว่า Tomek Links ไม่ได้ถ้ายังไม่มีการเปรียบเทียบด้วยตัวอย่าง E_l กล่าวคือ $d(E_i, E_l) < d(E_i, E_j)$ หรือ $d(E_j , E_l) < d(E_i, E_j)$

(3) ถ้าการเปรียบเทียบตัวอย่างจากขั้นตอนที่ 2 แล้ว Tomek Links ที่ได้จะได้ตัวใดตัวหนึ่งเป็นค่า noise หรือเป็น borderline ทั้ง 2 ตัว

สำหรับวิธี SMOTE-TL[14] เป็นการรวมทั้ง 2 วิธีที่กล่าวมาแล้วเข้าด้วยกันคือ วิธีการของ SMOTE และ วิธีการของ Tomel links ซึ่งแทนที่จะเป็นการเพิ่มข้อมูลหรือลดข้อมูลเพียงอย่างเดียว วิธีการนี้ได้นำการเพิ่มข้อมูลที่น้อยกว่าและลดข้อมูลที่มากกว่า เพื่อให้ได้ข้อมูลที่ใกล้เคียงกันโดยไม่เป็นการเพิ่มหรือลดอย่างเดียว

2.5 การวัดประสิทธิภาพของเทคนิคการจำแนก

เทคนิค k-fold cross-validation[15] คือการวัดประสิทธิภาพตัวอย่างโมเดล ด้วยการแบ่งชุดข้อมูลออกเป็นส่วน ๆ และนำข้อมูลที่ได้ทำการแบ่งมาทดสอบเพื่อดูผลลัพธ์ของโมเดล โดยการแบ่งข้อมูลจะแบ่งออกเป็น 2 ชุด คือ ชุดข้อมูลฝึกสอนและชุดข้อมูลทดสอบ



รูปที่ 3 การแบ่งข้อมูลแบบ 10-fold cross-validation[19]

จากรูปที่ 3 เป็นการแบ่งข้อมูลออกเป็นชุด ทั้งหมด 10 ชุด โดยการทำงานทั้งหมด 10 รอบ โดยที่รอบที่ 1 ให้ข้อมูลชุดที่ 2 ถึง 10 เป็นชุดข้อมูลฝึกสอน จากนั้นจึงนำข้อมูลชุดที่ 1 มาทดสอบ เมื่อถึงรอบที่ 2 ให้นำข้อมูล ชุดที่ 1 และ ชุดที่ 3 ถึง ชุดที่ 10 มาเป็นชุดข้อมูลฝึกสอน จึงนำข้อมูลชุดที่ 2 เป็นชุดข้อมูลทดสอบ ทำตามขั้นตอนเช่นนี้จนครบ 10 รอบ

2.6 ตัววัดประสิทธิภาพ

การวัดความน่าเชื่อถือของแบบจำลองที่มีปัญหาความไม่สมดุลของข้อมูลนั้น ทำให้การวัดค่าความถูกต้องเอนเอียงไปในส่วนของข้อมูลที่มีค่ามากกว่า ซึ่งการวัดค่าความถูกต้องที่ได้จึงไม่เหมาะสมแก่การใช้วัดประสิทธิภาพของเทคนิคเหล่านั้น ดังนั้นจึงได้มีการนำเทคนิคที่ใช้สำหรับการวัดประสิทธิภาพของข้อมูลที่เกิดความไม่สมดุลโดยใช้ค่า G-Mean(Geometric Mean)[17] ซึ่งหาได้จากตาราง confusion matrix[16]

ตารางที่ 1 confusion matrix สำหรับ two-class classification[16]

		True class	
		Positive prediction	Negative prediction
Prediction class	Actual Positive	True Positive	False Negative
	Actual Negative	False Positive	True Negative

โดยคำนวณได้ตามสูตรต่าง ๆ ดังนี้

$$\text{Accuracy} = \frac{TP+TN}{TP+TN+FP+FN} \quad (7)$$

$$\text{sensitivity} = \frac{TP}{TP+FN} \quad (8)$$

$$\text{specificity} = \frac{TN}{TN+FP} \quad (9)$$

เพื่อวัดประสิทธิภาพของอัลกอริทึมของข้อมูลทั้ง 2 คลาสนั้น [19] ต้องวัดค่าการสร้างความสมดุลของข้อมูลเพื่อให้การทำนายไม่เอนเอียงไปทางฝั่งที่มีข้อมูลมากกว่า เพื่อให้การทำนายนั้นสามารถใช้ข้อมูลจากทั้ง 2 คลาสมาสร้างรูปแบบการเรียนรู้ได้ดี จากค่าที่ได้จากตาราง confusion matrix จึงนำมาหาค่า G-Mean ตามสมการที่ 10 เพื่อใช้ในการวัดความสมดุลของข้อมูลในชุดข้อมูลฝึกสอนและชุดข้อมูลทดสอบ ซึ่งในหลายงานวิจัยได้นำวิธีการดังกล่าวมาใช้วัดความสมดุล [20][21][22][23] โดยที่ตัวชี้วัดจะแสดงถึงความสมดุลทั้งชุดข้อมูลฝึกสอนและชุดข้อมูลทดสอบเมื่อนำกลุ่มข้อมูลที่มากและกลุ่มของข้อมูลที่น้อยมาเปรียบเทียบ

$$\text{G-Mean} = \sqrt{\text{sensitivity} \times \text{specificity}} \quad (10)$$

3.วิธีดำเนินงานวิจัย

3.1 ศึกษาชุดข้อมูลสำหรับการทดลอง

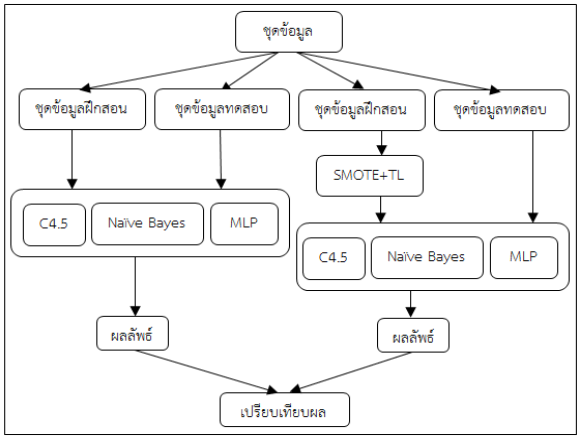
การดำเนินงานวิจัยเริ่มต้นจากการศึกษาข้อมูลและผลที่เกิดจากภาวะตั้งครรภ์จนถึงระยะของการตั้งครรภ์ เพื่อที่จะเป็นแนวทางการดำเนินงานวิจัย โดยได้ใช้ข้อมูลผลของการตั้งครรภ์จากองค์การอนามัยโลก ปี 2007-2008 ของประเทศไทยทั้งหมด 9,745 ระเบียน เนื่องจากในช่วงระยะเวลานี้ ได้มีการสำรวจผลการตั้งครรภ์ของมารดาในประเทศไทย ซึ่งได้บอกถึงลักษณะอาการเจ็บป่วยจนถึงการเสียชีวิต โดยทางด้าน Pisek Lumbiganon และคณะ[18] ทำการวิเคราะห์ปัจจัยและตัวแปรด้วยทางสถิติ พบว่าปัจจัยที่ทำให้เกิดอัตราการเจ็บป่วยและการเสียชีวิตของมารดาในขณะที่ตั้งครรภ์และหลังคลอด จากการคัดเลือกปัจจัยดังตารางที่ 2

ตารางที่ 2 ปัจจัยที่ทำให้เกิดการเจ็บป่วยและเสียชีวิตของมารดาขณะตั้งครรภ์และหลังคลอด

ชื่อแอตทริบิวต์	ค่าตัวแปร	จำนวน
1. อายุของมารดา(Maternal age)	"16-35" "<=16"<br">=35"	8205 1307 233
2. จำนวนปีการศึกษา (Year of education)	" <7" "7-12"<br">12"	2597 5187 1961
3. การคลอดบุตรครั้งแรก (Primiparaous)	Yes No	4698 5047
4. น้ำหนักทารกแรกเกิด(Birth weight)	"2500-4000"<br"<2500"<br">4000"	7074 991 207
5. ประวัติแท้งบุตรหรือเสียชีวิต(History of neonatal death or stillbirth)	No Yes	9649 96
6. มีเชื้อภาวะภูมิคุ้มกันบกพร่อง(HIV)	No Yes	9629 116
7. มีความดันโลหิตสูงก่อนการตั้งครรภ์ (Chronic hypertension)	No Yes	9685 60
8. มีภาวะโรคหัวใจหรือโรคไต (Cardiac/Renal diseases)	No Yes	9717 28
9. เม็ดเลือดแดงรูปเคียว(Sickle cell anemia)	No Yes	9715 30
10. มีโรคประจำตัว(Other medical conditions)	No Yes	9555 190
11. มีภาวะน้ำเดินหรือถุงน้ำคร่ำแตก (Prelabour rupture of membranes)	No Yes	9019 726
12. ความดันสูงระยะตั้งครรภ์ (Pregnancy induced hypertension)	No Yes	9511 234
13. ภาวะครรภ์เป็นพิษ (Pre-eclampsia)	No Yes	9535 210
14. ภาวะครรภ์เป็นพิษแล้วมีอาการชัก (Eclampsia)	No Yes	9733 12
15. มีการตกเลือดระหว่างตั้งครรภ์ในระยะที่ 2 (Vaginal bleeding in 2nd half of pregnancy)	No Yes	9692 53
16. มีการใช้ยาปฏิชีวนะในขณะตั้งครรภ์ (Any antenatal antibiotic treatment)	No Yes	9549 196
17. มีภาวะแทรกซ้อนในขณะตั้งครรภ์และคลอด(Referred for complication related to pregnancy or delivery)	No Yes	9097 648
18. ดัชนีชี้วัดความเสี่ยงที่เกิดการป่วยหรือเสียชีวิต (maternal mortality and morbidity)	No Yes	9569 176

จากตารางที่ 2 คือตัวแปรที่ทำให้เกิดอัตราการเจ็บป่วยและเสียชีวิตจากการตั้งครรภ์ ซึ่งตัวแปรเหล่านี้จะถูกนำมาใช้เพื่อสร้างโมเดลการทำนายด้วยการเปรียบเทียบจาก 3 เทคนิคคือ C4.5, Naïve Bayes และ Multilayer Perceptron เพื่อหาเทคนิคที่เหมาะสมกับข้อมูลชุดนี้

3.2 ขั้นตอนการทดลอง



รูปที่ 4 ขั้นตอนการทดลอง

จากรูปที่ 4 คือ ขั้นตอนการดำเนินการ จากที่ได้มีการคัดกรองข้อมูลที่เป็นปัจจัยทำให้เกิดความเสี่ยงที่จะป่วยหรือหรือเสียชีวิต จากนั้นนำข้อมูลที่ได้รับการคัดแยกมาจัดแบ่งข้อมูลเป็นชุดข้อมูลฝึกสอน และชุดข้อมูลทดสอบ เพื่อเข้าสู่ขั้นตอนการทดสอบเพื่อวัดประสิทธิภาพของข้อมูลด้วยการแบ่งออกเป็น 2 ส่วนในการทดลอง คือการทดสอบระหว่างชุดข้อมูลที่ยังไม่มีการแก้ไขปัญหาความไม่สมดุลของข้อมูลและข้อมูลที่มีการ

แก้ไขปัญหาความไม่สมดุลของข้อมูล เพื่อเปรียบเทียบหาค่าที่มีประสิทธิภาพที่ดีที่สุด โดยการทดลองส่วนที่ 1 เป็นการเปรียบเทียบระหว่างอัลกอริทึม C4.5, Naïve Bayes และ Multilayer perceptron เพื่อเปรียบเทียบหาอัลกอริทึมที่มีประสิทธิภาพมากที่สุด และในการทดลองส่วนที่ 2 ใช้วิธีการแก้ปัญหาความไม่สมดุลด้วยการใช้ SMOTE-TL จากนั้นใช้อัลกอริทึม C4.5, Naïve Bayes, และ Multilayer perceptron เพื่อเปรียบเทียบหาประสิทธิภาพ เมื่อได้ผลลัพธ์จากการทดลองจากทั้ง 2 ส่วนมาเปรียบเทียบเพื่อหาอัลกอริทึมที่มีประสิทธิภาพมากที่สุดโดยจากการวัดค่า G-mean

4. ผลการดำเนินงาน

จากการแบ่งการทดลองเป็น 2 ส่วน ได้แก่ การทดลองส่วนที่ 1 คือ การทดลองที่ไม่ได้แก้ไขปัญหาความไม่สมดุล และการทดลองส่วนที่ 2 คือ การทดลองที่แก้ไขปัญหาความไม่สมดุล และนำมาเปรียบเทียบเพื่อหาขั้นตอนที่มีประสิทธิภาพของชุดข้อมูลผลของการตั้งครรภ์จากองค์การอนามัยโลก 2007 – 2008 โดยการทดลองส่วนที่ 1 ที่ไม่มีการแก้ไขปัญหาความไม่สมดุลของข้อมูล ได้ค่าความถูกต้องที่ดีที่สุดคือวิธีการ C4.5 เท่ากับ 98.19% เนื่องจากข้อมูลมีความไม่สมดุลจึงต้องนำการวัดประสิทธิภาพของข้อมูลมาใช้ จึงได้อัลกอริทึมจากการวัดประสิทธิภาพ G-mean คือ Naïve Bayes เท่ากับ 16.84% ซึ่งทำให้ทราบว่าอัลกอริทึมของ Naïve Bayes เหมาะสมกับข้อมูลชุดนี้ตามตารางที่ 3

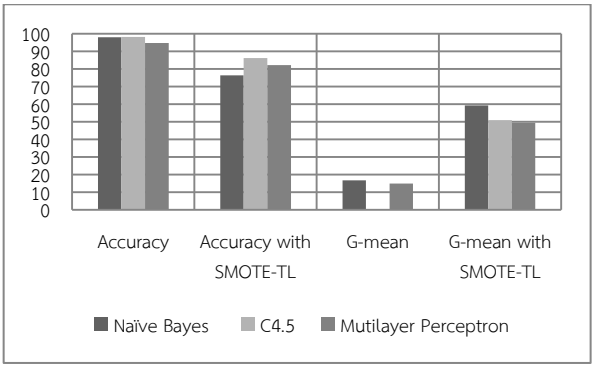
ตารางที่ 3 การวัดประสิทธิภาพของอัลกอริทึมที่ไม่มีการแก้ไขปัญหาความไม่สมดุลของข้อมูล

	Accuracy (%)	G-mean (%)
Naïve Bayes	98.03	16.84
C4.5	98.19	0
MLP	94.74	15.02

สำหรับการทดลองส่วนที่ 2 เมื่อเพิ่มการแก้ไขปัญหาความไม่สมดุลของข้อมูล จะได้ค่าความถูกต้องของวิธีการ C4.5 มากที่สุดเท่ากับ 86.08% แต่เมื่อนำมาเปรียบเทียบในการหาค่าของ G-mean ทำให้ทราบว่าอัลกอริทึม Naïve Bayes มีค่า G-mean มากที่สุดเมื่อใช้วิธีการแก้ไขปัญหาความไม่สมดุลของข้อมูลเท่ากับ 59.32% ตามตารางที่ 4

ตารางที่ 4 การวัดประสิทธิภาพของอัลกอริทึมที่มีการแก้ไขปัญหาความไม่สมดุลของข้อมูล

	Accuracy with SMOTE-TL (%)	G-mean with SMOTE-TL (%)
Naïve Bayes	76.29	59.32
C4.5	86.08	50.85
MLP	82.15	49.46



รูปที่ 5 กราฟการเปรียบเทียบค่าความถูกต้องและการวัดประสิทธิภาพของข้อมูลใน 3 เทคนิค

จากรูปที่ 5 คือกราฟการเปรียบเทียบให้เห็นว่า เมื่อไม่มีการแก้ไขปัญหาความไม่สมดุล ความถูกต้องจะมีค่าสูงเนื่องจากจำนวนข้อมูลของคลาสจะมีจำนวนที่แตกต่าง ส่งผลให้การวัดค่าความถูกต้องเอนเอียงไปในส่วนข้อมูลที่มีจำนวนมาก ทำให้ค่าความถูกต้องสูง แต่เมื่อการวัดประสิทธิภาพของข้อมูล G-mean ต่ำ เพราะข้อมูลไม่มีความสมดุลกัน ซึ่งข้อมูลลักษณะนี้ไม่เหมาะสมในการนำไปสร้างเป็นโมเดลการทำนาย นอกจากนี้จากกราฟได้แสดงการเปรียบเทียบค่าความถูกต้องและการวัดประสิทธิภาพ ของ G-mean ที่ได้ใช้วิธีการแก้ไขปัญหาของ SMOTE-TL โดยผลที่ได้ คือค่าความถูกต้องลดลงเมื่อเปรียบเทียบกับค่าความถูกต้องที่ไม่ได้ใช้วิธีการแก้ไขปัญหาความไม่สมดุล แต่ค่าการวัดประสิทธิภาพของ G-mean มีค่าเพิ่มสูงขึ้น เนื่องจากการแก้ไขปัญหาความไม่สมดุลของ SMOTE-TL ได้เพิ่มข้อมูลของคลาสที่มีจำนวนน้อยและลดข้อมูลของคลาสที่จำนวนมากกว่าให้ใกล้เคียงกัน จึงทำให้ค่า G-mean ที่แสดงถึงความสมดุลของข้อมูลที่มีประสิทธิภาพเพิ่ม และจากกราฟแสดงให้เห็นว่าเมื่อใช้วิธีการแก้ไขปัญหาความไม่สมดุลของข้อมูล ทำให้การวัดประสิทธิภาพของอัลกอริทึม Naïve Bayes เหมาะสมสำหรับชุดข้อมูลผลการตั้งครรภ์จากองค์การอนามัยโลก 2007 -2008 เนื่องจากอัลกอริทึม Naïve Bayes นั้นมีค่า เปรียบเทียบข้อมูลโดยเฉลี่ยของความถูกต้องที่ในการใช้การทำนายสูงที่สุด

5. สรุป

งานวิจัยนี้นำเสนอการเปรียบเทียบโมเดลเพื่อใช้ในการคัดกรองความเสี่ยงของมารดาในขณะตั้งครรภ์ ซึ่งอาจมีโอกาสเกิดภาวะแทรกซ้อนที่ทำให้เกิดการเจ็บป่วยจนถึงเสียชีวิต เพื่อป้องกันความเสี่ยงในการเจ็บป่วยจนถึงเสียชีวิตกับมารดาที่ตั้งครรภ์ ดังนั้นในงานวิจัยนี้จึงได้นำเสนอการเปรียบเทียบอัลกอริทึมเพื่อนำมาใช้ในการสร้างโมเดลการทำนายความเสี่ยงจากการตั้งครรภ์ โดยใช้ข้อมูลผลการตั้งครรภ์จากองค์กรอนามัยโลก 2007-2008 ที่ได้เก็บและศึกษาข้อมูลผลการตั้งครรภ์ในประเทศไทย ทั้งหมด 9,745 ระเบียน ซึ่งพบว่าข้อมูลชุดนี้มีความไม่สมดุลของข้อมูล ดังนั้นในงานวิจัยนี้จึงได้นำเสนอการแก้ไขปัญหาความไม่สมดุลของข้อมูล ด้วยวิธี SMOTE-TL จากการทดลองพบว่าอัลกอริทึมที่มี

ประสิทธิภาพที่เหมาะสมกับข้อมูลชุดนี้คือ Naïve Bayes โดยข้อมูลได้ผ่านการแก้ไขปัญหาความไม่สมดุลของข้อมูลของ SMOTE-TL ดังนั้นอัลกอริทึม Naïve Bayes ที่ข้อมูลได้ผ่านการแก้ไขปัญหาความไม่สมดุลของข้อมูลของ SMOTE-TL จึงเหมาะสมเป็นโมเดล เพื่อสร้างเครื่องมือคัดกรองมารดาที่ตั้งครรภ์ ที่อาจจะมีความเสี่ยงที่จะเกิดการบาดเจ็บจนถึงเสียชีวิตขณะตั้งครรภ์ เพื่อให้สามารถดูแลและรักษามารดาที่ตั้งครรภ์อย่างใกล้ชิด

สำหรับการพัฒนาในอนาคต คือการศึกษาและพัฒนาเทคนิควิธีในการแก้ปัญหาความไม่สมดุลของข้อมูล เพื่อเพิ่มประสิทธิภาพความถูกต้องมากขึ้น ซึ่งจะทำให้ได้โมเดลที่ดีและทำนายได้แม่นยำมากขึ้น

6. กิตติกรรมประกาศ

ขอขอบคุณ ศ.นพ. ภิเศก ลุมพิกานนท์ ศ.ดร. มาลินี เหล่าไพบูลย์ และองค์กรอนามัยโลก ที่เอื้อเฟื้อข้อมูลผลการตั้งครรภ์ขององค์การอนามัยโลกและคำแนะนำ ซึ่งให้งานวิจัยนี้สำเร็จลุล่วงไปได้ด้วยดี

เอกสารอ้างอิง

[1] World Health Organization (WHO),International Statistical Classification of Diseases and Related Health Problem : Tenth Revision, volume 1. Geneva: WHO, 1993.

[2] World Health Organization,Maternal Health and Safe Motherhood Program. Revised 1990 estimates of Maternal Mortality: A New Approach by WHO and UNICEF. Geneva: WHO, 1996.

[3] ยงเจือ เหล่าศิริถาวร. (2557, ตุลาคม 20). การวิเคราะห์ระบบรายงานและสถานการณ์การตายของมารดาในประเทศไทยปี พ.ศ. 2538 – 2539[ออนไลน์]. จาก http://www.hiso.or.th/hiso/proReport/pro2_report4.php

[4] G.J. Walker, D.E. Ashley, A.M. McCaw, and G.W. Bernard, “Maternal mortality in Jamaica,” The Lancet, 327(8479), pp.486-488, 1986.

[5] G.J. Walker, A.M. McCaw, D.E. Ashley, and G.W. Bernard, “Identifying maternal deaths in developing countries: experience in Jamaica,” International Journal of Epidemiology, 19(3), pp.599-605, 1990.

[6] กรมการแพทย์. คู่มือเวชปฏิบัติการคลอดมาตรฐาน. พิมพ์ครั้งที่ 2. กรุงเทพฯ: กรมการแพทย์ กระทรวงสาธารณสุข, 2556.

[7] สิรภัทร เชี่ยวชาญวัฒนา. เอกสารประกอบการสอนวิชาข่ายงานประสาทเทียม (Artificial Neural Networks) 322 752 (ฉบับปรับปรุง พ.ศ. 2552). ขอนแก่น: ภาควิชาวิทยาการคอมพิวเตอร์ มหาวิทยาลัยขอนแก่น, 2552.

[8] J. Padmavathi. “A Comparative study on Breast Cancer Prediction Using RBF and MLP,” International Journal of Scientific & Engineering Research, 2(1), pp.14-18, 2011.

[9] G.H. John, P. Langley. “Estimating continuous distributions in Bayesian classifiers,” In P. Besnard, S. Hanks. Proceedings of the Eleventh Conference on Uncertainty in Artificial Intelligence, Montreal, Quebec, Canada, Aug. 1995, pp. 338-345.

[10] J.R Quinlan. “Induction of decision trees,” Machine Learning, 1(1): pp. 81-106, 1986.

[11] H.E. Haibo, editor. “Self-adaptive systems for machine intelligence,” Hoboken, New Jersey, Canada: John Wiley & Sons, Inc., 2011.

[12] N.V. Chawla, K.W. Bowyer, L.O. Hall, and W.P. Kegelmeyer, “SMOTE: Synthetic Minority Over-sampling Technique,” Journal of Artificial Intelligence Research, 16, pp.321-357, 2002.

[13] I. Tomek. “Two modifications of CNN. Systems, Man and Cybernetics,” IEEE Transaction, 6(11), pp. 769-772, 1976.

[14] H.E. Haibo, editor. “Self-adaptive systems for machine intelligence,” Hoboken, New Jersey, Canada: John Wiley & Sons, Inc., 2011.

[15] R. Kohavi. “A study of cross-validation and bootstrap for accuracy estimation and model selection,” Proceedings of the Fourteenth International Joint Conference on Artificial Intelligence, 2(12), pp. 1137–1143, 1995.

[16] R. Kohavi, P. Foster. “Glossary of Terms,” Machine Learning, 30, pp. 271-274, 1998.

[17] M. Bekkar, H.K. Djemaa, and T.A. Alitouche. “Evaluation Measures for Models Assessment over Imbalanced Data Sets,” Journalof Information Engineering and Applications, 3(10), pp. 27-38, 2013.

[18] L. Pisake, MA. Laopaiboon, M. Gülmezoglu, J.P. Souza, S. Taneepanichskul, P. Ruyan, D.E. Attygalle et al. “Method of delivery and pregnancy outcomes in Asia: the WHO global survey on maternal and perinatal health 2007–08,” The Lancet, 375(9713), pp. 490-499, 2010.

[19] M. Kubat, and S. Matwin, "Addressing the curse of imbalanced training sets: One-sided selection,” In Douglas H. Fisher, editor, ICML, pp 179–186, 1997.

[20] M.G. Karagiannopoulos, D.S. Anyfantis, S.B. Kotsiantis, and P.E. Pintelas, “Local cost sensitive learning for handling imbalanced data sets,” Mediterranean Conf on Control & Automation, MED '07, 2007.

[21] C. Su, and Y. Hasio, “ An evaluation of the robustness of MTS for imbalanced data,” IEEE transactions on knowledge and data engineering, 19(10), pp. 1321-1332, 2007.

[22] B. Sukarna, Md, I. Monirul. Y. Xin, and M. Kazuyuki, “MWMOTE - Majority Weighted Minority Oversampling Technique for Imbalanced Data Set Learning,” IEEE Transactions on Knowledge and Data Engineering, 2012.

[23] Z. Yong, and W. Dapeng, “A Cost-Sensitive Ensemble Method for Class-Imbalanced Datasets,” Abstract and Applied Analysis; Article ID 196256, 2013.