การพยากรณ์การรอดชีวิตของผู้ป่วยมะเร็งเต้านม
Predicting Breast Cancer Patient Survival

จารี ทองคำ[1*] และวาทินี สุขมาก[2]
Jaree Thomgkam[1*] and Vatinee Sukmak[2]

Received: June 30, 2020; Revised: December 4, 2020; Accepted: December 4, 2020

บทคัดย่อ

งานวิจัยฉบับนี้มีวัตถุประสงค์เพื่อพัฒนาแบบจำลองที่มีประสิทธิภาพในการพยากรณ์การรอดชีวิตของผู้ป่วยมะเร็งเต้านมซึ่งเป็นมะเร็งที่พบมากในเพศหญิงเป็นอันดับที่สองรองจากมะเร็งรังไข่ ข้อมูลเก็บรวบรวมจากฐานข้อมูล SEER ในปี ค.ศ. 2004 ถึง 2014 จำนวน 115,184 ระเบียน การวิจัยนี้ใช้เทคนิคเหมืองข้อมูลพื้นฐาน คือ เทคนิคนาอีฟเบย์ เทคนิคส่วนของรายการตัดสินใจ เทคนิคเพอร์เซปตรอนหลายชั้น และเทคนิคซัพพอร์ตเวกเตอร์แมชีน ในการสร้างแบบจำลองเปรียบเทียบกับแบบจำลองดังกล่าวร่วมกับเทคนิคการห่อเพื่อเพิ่มประสิทธิภาพการพยากรณ์ คณะผู้วิจัยใช้หลักการ 10-โฟดครอสวาลิเดชันในการแบ่งชุดข้อมูลเป็นชุดข้อมูลฝึกสอนและชุดข้อมูลทดสอบ โดยใช้ค่าความไว ความจำเพาะ และความถูกต้องวัดประสิทธิภาพของแบบจำลอง ผลทดลองพบว่า เทคนิคส่วนของรายการตัดสินใจร่วมกับเทคนิคการห่อสามารถสร้างแบบจำลองการพยากรณ์การรอดชีวิตของผู้ป่วยมะเร็งเต้านมที่มีความถูกต้องสูงสุดที่ร้อยละ 98.89

คำสำคัญ : มะเร็งเต้านม; การรอดชีวิต; เทคนิคแบบรวม; เหมืองข้อมูล

[1] *คณะวิทยาการสารสนเทศ มหาวิทยาลัยมหาสารคาม*
[2] *คณะพยาบาลศาสตร์ มหาวิทยาลัยมหาสารคาม*
[1] *Faculty of Informatics, Mahasarakham University*
[2] *Faculty of Nursing Mahasarakham University*
[*] *Corresponding Author E - mail Address: jaree.thongkam@gmail.com*

ISSN 2672-9369 (Online)

Abstract

The objective of this research is to develop the effective model for predicting the survival of patients with breast cancer. Breast cancer is the second most common cancer in women. Data were collected from the SEER database in 2004 to 2014. It has up to 115,184 records. The prediction models were modeled with the basic techniques including Naive Bayes, PART decision list, MutiLayer Perceptron and Support Vector Machine. Moreover, the research team adopted Bagging technique to combine with these basic techniques in order to increase performance of the built prediction models. 10-fold cross-validation has been used to divide the dataset into training and testing sets. Sensitivity, specificity and accuracy values were used to compare the performance of models. The experiment result shows that that PART combine with bagging technique can construct breast cancer survival models with the highest accuracy of 98.89 %.

Keywords: Breast Cancer; Survival; Ensemble Technique; Data Mining

บทนำ

มะเร็งเต้านมเป็นมะเร็งที่พบบ่อยที่สุด และเป็นสาเหตุอันดับสองของการเสียชีวิตด้วยโรคมะเร็งสำหรับเพศหญิง [1] การรอดชีวิตของผู้ป่วยมะเร็งเต้านมมีมากขึ้นเมื่อมีการตรวจพบแต่เนิ่น ๆ รวมถึงการรักษายังเป็นปัจจัยหนึ่งที่ทำให้โอกาสการรอดชีวิตสูงขึ้นอีกด้วย มะเร็งเต้านมสามารถแบ่งออกเป็น 4 ระยะ ได้แก่ ระยะที่ 1 เป็นระยะที่เซลล์มะเร็งมีขนาดค่อนข้างเล็กและมีอยู่ภายในอวัยวะที่มะเร็งเริ่มก่อตัว ระยะนี้เป็นระยะที่มะเร็งมีความรุนแรงน้อยที่สุด ระยะที่ 2 เป็นระยะที่มะเร็งเต้านมมีขนาดใหญ่กว่าในระยะที่ 1 แต่เซลล์มะเร็งเต้านมยังไม่เริ่มแพร่กระจายไปยังเนื้อเยื่อรอบ ๆ ระยะที่ 3 เป็นระยะที่มะเร็งเต้านมมีขนาดใหญ่ขึ้นอาจเริ่มแพร่กระจายไปยังเนื้อเยื่อรอบ ๆ และมีเซลล์มะเร็งเต้านมในต่อมน้ำเหลืองในบริเวณนั้น ระยะที่ 4 เป็นระยะที่เซลล์มะเร็งแพร่กระจายไปจากตำแหน่งที่เริ่มเป็นอวัยวะอื่น ๆ ซึ่งเรียกว่า มะเร็งระยะแพร่กระจายหรือมะเร็งระยะลุกลาม [2] อัตราการรอดชีวิตของผู้ป่วยมะเร็งเต้านมโดยทั่วไปใช้การพิจารณาจากการรอดชีวิตในระยะ 5 ปี โดยเฉพาะผู้ป่วยในระยะที่ 3 จะมีการเสียชีวิตในระยะก่อน 5 ปี ที่ร้อยละ 72 ดังนั้นอายุรแพทย์ด้านมะเร็งวิทยาได้อธิบายเกี่ยวกับอัตราการรอดชีวิตของมะเร็งเต้านมโดยใช้คำว่าอัตราการรอดชีวิตระยะ 5 ปี [3] ในทศวรรษที่ผ่านมาการรอดชีวิตระยะ 5 ปี ของผู้ป่วยมะเร็งเต้านมมีมากยิ่งขึ้น เนื่องจากผู้ป่วยมาพบแพทย์เร็วขึ้นและการรักษาที่มีประสิทธิภาพมากยิ่งขึ้น [1] ในอดีตการวิเคราะห์การรอดชีวิตของผู้ป่วยมะเร็งโดยส่วนใหญ่ใช้การวิเคราะห์ถดถอย (Regression Analysis) [4] ปัจจุบันการวิเคราะห์ข้อมูลด้วยเทคนิคเหมืองข้อมูลเข้ามามีบทบาทในการพยากรณ์และใช้กันอย่างแพร่หลาย ซึ่งเทคนิคเหมืองข้อมูลสามารถนำไปใช้ในการสร้างแบบจำลองเพื่อการพยากรณ์เกิดโรค [5] และการพยากรณ์การรอดชีวิตของผู้ป่วยมะเร็ง [6] ซึ่งแบบจำลองนี้สามารถนำไปใช้ในการพยากรณ์ทางการแพทย์และการเฝ้าระวังโรค

เทคนิคการทำเหมืองข้อมูล (Data Mining Techniques) นิยมนำมาใช้ในการสร้างแบบจำลองเพื่อการพยากรณ์ เช่น เทคนิคนาอีฟเบย์ (Naive Bayes: NB) เทคนิคส่วนของรายการตัดสินใจ (PART Decision List: PART) เทคนิคเพอร์เซปตรอนหลายชั้น (MutiLayer Perceptron: MLP) และเทคนิคซัพพอร์ตเวกเตอร์แมชีน (Support Vector Machine: SVM) เป็นต้น โดยเทคนิค NB เป็นเทคนิคที่มีประสิทธิภาพในการสร้างแบบจำลองเพื่อการพยากรณ์เมื่อข้อมูลมีลักษณะเป็น 2 ตัวเลือก (Binary) [7] และการทำเหมืองข้อความ (Text Mining) [8] - [9] เทคนิค PART เป็นเทคนิคที่ใช้ในการสร้างแบบจำลองแบบกฎ (Rule-Based) ซึ่งง่ายต่อการแปลความหมายและมีประสิทธิภาพในการพยากรณ์ [10] - [11] ส่วนเทคนิค MLP เป็นเทคนิคที่นิยมนำมาใช้ในการแก้ปัญหาเกี่ยวกับการรู้จำ หรือการพยากรณ์ข้อมูลรูปภาพที่มีจำนวนข้อมูลไม่มากนักแต่มีตัวแปรจำนวนมาก [12] และเทคนิค SVM เป็นเทคนิคที่มีประสิทธิภาพในการสร้างแบบจำลองเพื่อการพยากรณ์ที่ใช้กับข้อมูลที่มีขนาดเล็กและข้อมูลขนาดกลาง [13] [14]

เทคนิค Bagging เป็นเทคนิคแบบรวม (Ensemble) ที่ใช้ในการแก้ปัญหาแบบจำลองที่มีประสิทธิภาพต่ำจากแบบจำลองที่ใช้เทคนิคพื้นฐาน หรือเรียกอีกอย่างหนึ่งว่า Weak Classifier [15] โดยใช้หลักการ Bootstrap ในการสุ่มข้อมูลแบบกลับมาใช้ใหม่ให้แก่แบบจำลองที่ใช้เทคนิคพื้นฐาน [16]

ดังนั้นวัตถุประสงค์ของงานวิจัยนี้เพื่อพัฒนาแบบจำลองการพยากรณ์การรอดชีวิตของผู้ป่วยมะเร็งเต้านม โดยใช้เทคนิคพื้นฐานในการทำเหมืองข้อมูล 4 เทคนิค คือ NB, PART, MLP และ SVM ร่วมกับเทคนิค Bagging นอกจากนั้นคณะผู้วิจัยยังใช้หลักการ 10-Fold Cross-Validation ในการแบ่งข้อมูลออกเป็นชุดข้อมูลฝึกสอนใช้ในการสร้างแบบจำลองด้วยเทคนิคพื้นฐาน และชุดข้อมูลทดสอบเพื่อใช้ในการทดสอบประสิทธิภาพการพยากรณ์ของแบบจำลองด้วยค่าความไว (Sensitivity) ค่าความจำเพาะ (Specificity) และค่าความถูกต้อง (Accuracy)

วิธีดำเนินการวิจัย

วิธีดำเนินการวิจัยในงานวิจัยนี้แบ่งออกเป็น 4 ขั้นตอนตามวิธีการทำเหมืองข้อมูลของ Han, J. W. and Kamber, M. [17] ดังนี้คือ ขั้นตอนการเตรียมข้อมูล ขั้นตอนก่อนการสร้างแบบจำลอง ขั้นตอนการสร้างแบบจำลอง และการวัดประสิทธิภาพของแบบจำลอง

1. **ขั้นตอนการเตรียมข้อมูล** เป็นขั้นตอนในการทำความเข้าใจข้อมูล คณะผู้วิจัยทำการรวบรวมข้อมูลผู้ป่วยในกลุ่มการวินิจฉัยโรคมะเร็งเต้านมจากฐานข้อมูล SEER ซึ่งข้อมูลอยู่ในรูปแบบแฟ้มข้อความ (Text File) โดยมีตัวแปรทั้งหมด 74 ตัวแปร จำนวน 547,920 ระเบียน เมื่อทำการตรวจสอบความสมบูรณ์ของข้อมูลพบว่า ข้อมูลมีปัญหาเกี่ยวกับข้อมูลซ้ำซ้อนและข้อมูลไม่สมบูรณ์ในอัตรามากกว่าร้อยละ 50 ของข้อมูลทั้งหมด โดยข้อมูลจากเดือนมกราคม ค.ศ. 2004 ถึงเดือนธันวาคม ค.ศ. 2014 เป็นข้อมูลที่มีความสมบูรณ์มากที่สุดจำนวน 193,053 ระเบียน

2. **ขั้นตอนก่อนการสร้างแบบจำลอง** เป็นขั้นตอนที่มีความสำคัญต่อการวัดประสิทธิภาพของแบบจำลอง ซึ่งประกอบด้วย 4 ขั้นตอนดังนี้

   2.1 กำจัดข้อมูลที่ซ้ำกันออกจากชุดข้อมูล

   2.2 กำจัดข้อมูลที่ไม่สมบูรณ์ออก และคำนึงถึงตัวแปรที่ใช้ในการสร้างแบบจำลอง

2.3 กำหนดระยะเวลาการเสียชีวิต แบ่งข้อมูลออกเป็น 2 กลุ่มหรือคลาส คลาส 0 คือ ข้อมูลของผู้ป่วยที่เสียชีวิต ก่อนระยะ 5 ปี และคลาส 1 คือ ข้อมูลของผู้ป่วยที่รอดชีวิตมากกว่าระยะ 5 ปี หลังจากวันที่ได้รับการวินิจฉัยโรค

2.4 ทำการตรวจจับข้อมูลที่ผิดปกติ (Outliers) ออกจากชุดข้อมูล ด้วยเทคนิค C4.5

หลังจากคณะผู้วิจัยได้ดำเนินการตามทั้ง 4 ขั้นตอนข้างต้นแล้วมีข้อมูลคงเหลือ 115,184 ระเบียน โดยประกอบด้วยคลาส 0 จำนวน 92,720 และคลาส 1 จำนวน 22,464 มีตัวแปรทั้งสิ้น จำนวน 10 ตัวแปร แบ่งออกเป็นตัวแปรต้นจำนวน 9 ตัวแปร และตัวแปรตามจำนวน 1 ตัวแปร ดังตารางที่ 1

ตารางที่ 1 ตัวแปรในงานวิจัย

ลำดับ	รายละเอียด	รหัสตัวแปร	ชนิดของข้อมูล
1	Marital Status	MAR_STAT	Nominal
2	Race	RAC_RECY	Nominal
3	Age at Diagnosis	AGE	Number
4	Primary Site	PRIMSITE	Nominal
5	Lateral	LATERAL	Nominal
6	Cancer Grade	GRADE	Nominal
7	Tumor Size	CSTUMSIZ	Nominal
8	Stage	DSS2000S	Nominal
9	Surgery of Primary Site	SURGPRIF	Nominal
10	Survival Class	Class	Nominal

3. **ขั้นตอนการสร้างแบบจำลอง** การสร้างแบบจำลองเพื่อการพยากรณ์การรอดชีวิตของผู้ป่วยมะเร็งเต้านม ด้วยเทคนิคเหมืองข้อมูลที่มีประสิทธิภาพในการพยากรณ์ที่เป็นที่ยอมรับกันอย่างแพร่หลาย ประกอบด้วยเทคนิคเหมืองข้อมูลพื้นฐาน คือ เทคนิค Naive Bayes, PART, MLP และ SVM ร่วมกับเทคนิค Bagging

3.1 เทคนิคนาอีฟเบย์ (Naive Bayes) [18] เป็นเทคนิคที่ใช้หลักการของความน่าจะเป็นมาแก้ปัญหา โดยการสร้างแบบจำลองจาก Bayes′ Theorem และสมมติฐานที่ถูกกำหนดขึ้นจากการเกิดของเหตุการณ์ต่าง ๆ ที่ใช้ในการพยากรณ์

3.2 เทคนิคส่วนของรายการตัดสินใจ (PART) [11] เป็นเทคนิคที่ใช้ในการสร้างกฎการตัดสินใจ โดยการนำเอาหลักการของต้นไม้การตัดสินใจ และการสร้างกฎการตัดสินใจจากใบที่ดีที่สุด (Best Leave) ในการพยากรณ์ ถึงแม้ว่าเทคนิคนี้จะคล้ายกับต้นไม้ตัดสินใจ แต่เทคนิค PART ได้หลีกเลี่ยงการสร้างต้นไม้เต็มรูปแบบทำให้ลดเวลาในการสร้างกฎการตัดสินใจ (Decision Rule)

3.3 เทคนิคเพอร์เซปตรอนหลายชั้น (MLP) เป็นเทคนิคที่ใช้แนวคิดที่ได้มาจากการจำลองการทำงานของเซลล์สมองของมนุษย์ ซึ่งมีโครงสร้างประกอบด้วย ชั้นข้อมูลนำเข้า (Input Layer) ชั้นข้อมูลแฝง (Hidden Layer) และชั้นข้อมูลออก (Output Layer) โดยมีหน่วยย่อยเรียกว่า Perceptron ซึ่งเทียบเท่ากับเซลล์สมองของมนุษย์หนึ่ง Neuron โดยหลักการของ Neural Network จะมีการกำหนด

ค่าน้ำหนักและเกณฑ์ให้แก่ข้อมูลนำเข้าแต่ละตัวโดยใช้ Back-propagation Algorithm ในการคำนวณในการสร้างแบบจำลอง และให้ผลการพยากรณ์ได้อย่างแม่นยำ [19]

3.4 เทคนิคซัพพอร์ตเวกเตอร์แมชีน (SVM) [20] เป็นเทคนิคที่ได้จากทฤษฎีการเรียนรู้จากสถิติ เป็นการใช้หลักการลดค่าความเสี่ยงเชิงโครงสร้างให้ต่ำที่สุด (Structural Risk Minimized) เพื่อลดค่าความผิดพลาดของการทำนาย (Minimization Error) พร้อมกับเพิ่มระยะการแบ่งให้มากที่สุด (Maximized Margin) ในการพยากรณ์

3.5 เทคนิคการห่อ (Bagging) [21] เป็นเทคนิคแบบรวมที่ใช้หลักการของ Bootstrap ในการสุ่มข้อมูลแบบกลับไปใช้ใหม่ ซึ่งเป็นเทคนิคที่เพิ่มประสิทธิภาพให้กับเทคนิคพื้นฐาน ในงานวิจัยนี้ได้นำเอาเทคนิค Bagging รวมกับเทคนิค Naive Bayes เทคนิค PART เทคนิค MLP และเทคนิค SVM ตามลำดับ

4. **การวัดประสิทธิภาพของแบบจำลอง** ในงานวิจัยนี้เทคนิค 10-Fold Cross-Validation ได้ถูกนำมาใช้ในการแยกข้อมูลออกเป็นชุดข้อมูลฝึกสอน และชุดข้อมูลทดสอบ จำนวน 10 รอบ ซึ่งหลักการนี้จะช่วยลดความแตกต่าง และเพิ่มความน่าเชื่อถือของผลการทดลอง ในการวัดประสิทธิภาพของแบบจำลอง คณะผู้วิจัยใช้ค่าความไว (Sensitivity) ค่าจำเพาะ (Specificity) และค่าความถูกต้อง (Accuracy) ซึ่งคำนวณได้จากสมการที่ (1) - (3)

$$\text{Sensitivity} = \frac{\text{TP}}{\text{TP} + \text{FN}} \times 100 \tag{1}$$

$$\text{Specificity} = \frac{\text{TN}}{\text{TN} + \text{FP}} \times 100 \tag{2}$$

$$\text{Accuracy} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{FP} + \text{TN} + \text{FN}} \times 100 \tag{3}$$

เมื่อ

TP (True Positive) หมายถึง จำนวนข้อมูลที่แบบจำลองสามารถพยากรณ์ได้อย่างถูกต้องในกลุ่มผู้ป่วยที่เสียชีวิตก่อนระยะ 5 ปี นับจากวันที่ได้ตรวจพบมะเร็งเต้านมถึงวันที่เสียชีวิต

FP (False Positive) หมายถึง จำนวนข้อมูลที่แบบจำลองที่ไม่สามารถพยากรณ์ได้อย่างถูกต้องในกลุ่มผู้ป่วยที่เสียชีวิตก่อนระยะ 5 ปี นับจากวันที่ได้ตรวจพบมะเร็งเต้านมถึงวันที่เสียชีวิต

TN (True Negative) หมายถึง จำนวนข้อมูลที่แบบจำลองสามารถพยากรณ์ได้อย่างถูกต้องในกลุ่มผู้ป่วยที่เสียชีวิตหลังระยะ 5 ปี นับจากวันที่ได้ตรวจพบมะเร็งเต้านมถึงวันที่เสียชีวิต

FN (False Negative) หมายถึง จำนวนข้อมูลที่แบบจำลองที่ไม่สามารถพยากรณ์ได้อย่างถูกต้องในกลุ่มผู้ป่วยที่เสียชีวิตหลังระยะ 5 ปี นับจากวันที่ได้ตรวจพบมะเร็งเต้านมถึงวันที่เสียชีวิต

ผลการทดลอง

จากการสร้างแบบจำลองด้วยเทคนิคเหมืองข้อมูลพื้นฐาน ผลการวัดประสิทธิภาพของแบบจำลอง ดังนี้

1. ค่าความไว (Sensitivity) ของแบบจำลองการพยากรณ์การรอดชีวิตของผู้ป่วยมะเร็งเต้านม ระยะ 5 ปี จากการทดลอง 10 รอบ ด้วยเทคนิคเหมืองข้อมูลพื้นฐาน และเทคนิค Bagging ร่วมกับเทคนิคเหมืองข้อมูลพื้นฐาน แสดงดังตารางที่ 2

ตารางที่ 2 ค่าความไวของแบบจำลอง

Rounds	Basic (%)				Bagging (%)			
	NB	**PART**	**MLP**	**SVM**	**NB**	**PART**	**MLP**	**SVM**
1	97.37	99.26	95.25	96.24	97.30	99.33	95.49	96.37
2	97.55	99.10	95.05	96.10	97.40	99.29	95.77	96.24
3	97.22	99.25	95.70	96.62	97.16	99.42	95.97	96.75
4	97.82	99.10	96.07	96.31	97.50	99.23	95.86	96.47
5	97.04	99.35	95.65	96.52	97.27	99.36	95.72	96.69
6	97.39	99.10	95.98	96.53	97.48	99.42	96.01	96.78
7	97.72	99.19	95.46	96.17	97.58	99.36	95.79	96.40
8	97.55	99.26	95.61	96.34	97.63	99.61	95.85	96.47
9	97.56	99.37	96.35	96.58	97.62	99.49	95.80	96.74
10	97.88	99.08	96.04	96.04	97.69	99.37	96.10	96.14
Average	97.51	99.21	95.72	96.34	97.46	**99.39**	95.84	96.50
SD.	0.25	0.10	0.38	0.20	0.17	0.10	0.16	0.22

จากตารางที่ 2 แสดงค่าความไวของแบบจำลองเพื่อการพยากรณ์การรอดชีวิตของผู้ป่วยมะเร็งเต้านม ที่สร้างจากเทคนิคพื้นฐาน NB PART MLP และ SVM พบว่า ทุกเทคนิคสามารถสร้างแบบจำลองที่มีประสิทธิภาพได้ในระดับที่ดีมากกล่าวคือ มากกว่าร้อยละ 90.00 โดยเทคนิค PART จะให้ค่าความไวสูงสุดถึงร้อยละ 99.21 รองลงมาคือเทคนิค NB SVM และ MLP ตามลำดับ เมื่อนำเทคนิค Bagging ร่วมกับเทคนิคพื้นฐานพบว่า ค่าความไวของแบบจำลอง Bagging ร่วมกับ PART MLP และ SVM มีค่าความไวเพิ่มขึ้นเล็กน้อย (น้อยกว่า 1 %) แต่ค่าความไวของแบบจำลอง Bagging ร่วมกับ NB มีค่าลดลงเล็กน้อย

2. **ค่าจำเพาะ (Specificity)** ของแบบจำลองการพยากรณ์การรอดชีวิตของผู้ป่วยมะเร็งเต้านม ระยะ 5 ปี พบว่า มีค่าจำเพาะของแบบจำลอง ดังตารางที่ 3

ตารางที่ 3 ค่าจำเพาะของแบบจำลอง

Rounds	Basic (%)				Bagging (%)			
	NB	PART	MLP	SVM	NB	PART	MLP	SVM
1	28.39	96.88	87.18	71.78	29.02	97.33	89.59	71.65
2	29.19	95.91	89.85	71.07	30.13	96.57	89.36	70.72
3	32.53	97.20	88.16	71.38	32.80	97.15	89.54	71.25
4	29.82	97.20	87.54	72.41	32.27	97.37	90.25	71.74
5	32.55	96.71	88.56	71.86	31.30	97.02	89.76	71.59
6	30.63	96.88	89.31	72.71	29.16	96.71	90.74	72.04
7	29.74	96.17	87.31	71.42	31.70	95.99	88.56	70.44
8	31.61	96.30	88.60	72.17	30.99	96.57	89.18	71.64
9	30.37	97.02	85.26	71.59	30.81	96.35	90.69	71.06
10	27.29	96.75	85.57	74.00	28.54	97.46	89.54	73.86
Average	30.21	96.70	87.74	72.04	30.67	**96.85**	89.72	71.60
SD.	1.62	0.42	1.41	0.81	1.36	0.46	0.64	0.89

จากตารางที่ 3 แสดงค่าจำเพาะของแบบจำลองสร้างจากเทคนิคพื้นฐาน คือ NB PART MLP และ SVM และแบบจำลองที่สร้างจากเทคนิคพื้นฐานร่วมกับเทคนิค Bagging พบว่าเทคนิค PART ร่วมกับเทคนิค Bagging ให้ค่าจำเพาะสูงสุดถึงร้อยละ 96.85 อย่างไรก็ตามเมื่อตรวจสอบรายละเอียดพบว่าเทคนิค Bagging สามารถเพิ่มค่าจำเพาะของแบบจำลองที่ใช้เทคนิค NB และ PART ขึ้นเพียงเล็กน้อย และสามารถเพิ่มประสิทธิภาพของแบบจำลองที่ใช้เทคนิค MLP และ SVM มากกว่า 1 % นอกจากนั้นเทคนิค Bagging ยังลดค่า SD. ให้แก่เทคนิค NB และ MLP ซึ่งหมายถึงมีความเสถียรของแบบจำลองในการพยากรณ์เพิ่มขึ้น

3. **ค่าความถูกต้อง (Accuracy)** ของแบบจำลองการพยากรณ์การรอดชีวิตของผู้ป่วยมะเร็งเต้านมระยะ 5 ปี ซึ่งสามารถแสดงค่าความถูกต้องได้ดังตารางที่ 4

ตารางที่ 4 ค่าความถูกต้องของแบบจำลอง

Rounds	Basic				Bagging			
	NB	PART	MLP	SVM	NB	PART	MLP	SVM
1	83.91	98.79	93.68	91.47	83.98	98.94	94.34	91.54
2	84.22	98.48	94.04	91.21	84.28	98.76	94.52	91.26
3	84.60	98.85	94.23	91.70	84.61	98.98	94.71	91.78
4	84.56	98.73	94.41	91.65	84.77	98.87	94.77	91.65
5	84.47	98.84	94.27	91.71	84.41	98.91	94.56	91.80
6	84.37	98.67	94.68	91.88	84.16	98.89	94.98	91.95

ตารางที่ 4 ค่าความถูกต้องของแบบจำลอง (ต่อ)

Rounds	Basic				Bagging			
	NB	PART	MLP	SVM	NB	PART	MLP	SVM
7	84.47	98.60	93.87	91.34	84.74	98.71	94.38	91.34
8	84.69	98.68	94.24	91.63	84.63	99.02	94.55	91.63
9	84.46	98.91	94.19	91.71	84.59	98.88	94.81	91.73
10	84.11	98.63	94.00	91.74	84.21	99.00	94.82	91.80
Average	84.39	98.72	94.16	91.60	84.44	**98.89**	94.64	91.65
SD.	0.23	0.12	0.27	0.19	0.26	0.10	0.20	0.21

จากตารางที่ 4 แสดงค่าความถูกต้องของแบบจำลองสร้างจากเทคนิคพื้นฐาน NB PART MLP และ SVM และแบบจำลองที่สร้างจากเทคนิคพื้นฐานร่วมกับเทคนิค Bagging พบว่า เทคนิค PART ให้ค่าความถูกต้องสูงสุดถึงร้อยละ 98.72 รองลงมาคือ เทคนิค MLP ให้ค่าความถูกต้องสูงสุดร้อยละ 94.16 เทคนิคที่ให้ค่าความถูกต้องต่ำที่สุด คือ เทคนิค NB เมื่อนำเทคนิค Bagging มาร่วมกับเทคนิคพื้นฐาน สามารถเพิ่มค่าความถูกต้องของแบบจำลองพื้นฐานขึ้นเพียงเล็กน้อย ยิ่งไปกว่านั้น เทคนิค Bagging สามารถเพิ่มความเสถียรให้แก่การพยากรณ์ด้วยเทคนิค PART และเทคนิค MLP มากขึ้นด้วย

จากผลการทดลองข้างต้น เทคนิค PART เป็นเทคนิคที่พยากรณ์การรอดชีวิตของผู้ป่วยมะเร็งเต้านมที่มีประสิทธิภาพมากที่สุด สามารถนำมาสร้างกฎการตัดสินใจได้ถึง 810 กฎ ซึ่งคณะผู้วิจัยได้แสดงตัวอย่างของกฎการตัดสินใจการพยากรณ์การรอดชีวิตของผู้ป่วยมะเร็งเต้านมในรูปของกฎการตัดสินใจ 6 กฎ ดังตารางที่ 5

ตารางที่ 5 ตัวอย่างกฎการตัดสินใจจากการพยากรณ์การรอดชีวิตของผู้ป่วยมะเร็งเต้านมด้วยเทคนิค PART

Rule Number	PART Models	Prediction Result
Rule 1	AGE > 62 AND GRADE = 3 AND CSTUMSIZ <= 989: 0	0
Rule 2	AGE > 62 AND GRADE = 2 ANDCSTUMSIZ <= 820:	0
Rule 3	AGE > 62 AND GRADE = 1 AND CSTUMSIZ <= 23: 0	0
Rule 4	GRADE = 4 AND AGE <= 85 AND CSTUMSIZ <= 20 AND SURGPRIF = 2 AND MAR_STAT = 2	1
Rule 5	RAC_RECY = 2 AND GRADE = 4 AND SURGPRIF = 2 AND AGE <= 64 ANDMAR_STAT = 1	1
Rule 6	RAC_RECY = 2 AND GRADE = 4 AND SURGPRIF = 2 AND AGE <= 64 AND MAR_STAT = 1	1

จากตัวอย่างกฎการตัดสินใจการพยากรณ์การรอดชีวิตของผู้ป่วยมะเร็งเต้านม ในกฎข้อที่ 1 ถ้าผู้ป่วยมีอายุมากกว่า 62 ปี มีเซลล์มะเร็งอยู่ใน GRADE = 3 ซึ่งหมายถึง ระยะที่เซลล์มะเร็งมีการกระจายเร็วมากแต่ไม่สามารถแยกความแตกต่างของเซลล์มะเร็งได้ และไม่ทราบขอบเขตของขนาดก้อนมะเร็ง แบบจำลองกฎการตัดสินใจจะพยากรณ์ว่าผู้ป่วยจะเสียชีวิตก่อนระยะ 5 ปี หลังจากวินิจฉัยว่าเป็นมะเร็งเต้านม ร้อยละ 100 ซึ่งหมายความว่าถ้าผู้ป่วยเข้าเกณฑ์กฎข้อที่ 1 ผู้ป่วยจะเสียชีวิตทั้งหมดในก่อนระยะ 5 ปี กฎข้อที่ 4 สามารถนำไปพยากรณ์ได้คือ ถ้าผู้ป่วยเป็นมะเร็งใน GRADE = 4 ซึ่งหมายถึงเซลล์มะเร็งมีการเปลี่ยนรูปร่างแตกต่างจากเซลล์ต้นกำเนิด ผู้ป่วยมีอายุน้อยกว่า 85 ปี ขนาดของมะเร็งน้อยกว่า 20 มิลลิเมตร ได้รับการรักษาด้วยการผ่าตัดเอาเฉพาะเซลล์มะเร็งออกไปแล้ว และสถานภาพสมรส กฎการตัดสินใจจะพยากรณ์ว่าผู้ป่วยสามารถมีชีวิตรอดในระยะ 5 ปี ดังนั้นจากแบบจำลองและกฎการตัดสินใจการพยากรณ์การรอดชีวิตของผู้ป่วยมะเร็งเต้านม สามารถนำไปพัฒนาเป็นระบบเฝ้าระวังโรคให้แก่ผู้ป่วยและแพทย์ได้นำการพยากรณ์มาใช้ในการคัดกรองความเสี่ยงต่อการเสียชีวิตของผู้ป่วยในเบื้องต้นต่อไป

สรุปผลและการอภิปราย

งานวิจัยฉบับนี้ได้นำข้อมูลผู้ป่วยที่ได้รับการวินิจฉัยด้วยโรคมะเร็งเต้านมและเสียชีวิตด้วยโรคมะเร็งเต้านมจากฐานข้อมูล SEER ระหว่างปี ค.ศ. 2004 - 2014 จำนวน 115,184 ราย มาทำการสร้างแบบจำลองด้วยเทคนิคพื้นฐาน NB PART MLP และ SVM และแบบจำลองที่สร้างร่วมกับเทคนิค Bagging

จากผลการทดลองพบว่า เทคนิค PART มีประสิทธิภาพในการพยากรณ์สูงสุดโดยให้ค่าความไวเฉลี่ยร้อยละ 99.21 ค่าจำเพาะเฉลี่ยร้อยละ 96.70 และค่าความถูกเฉลี่ยร้อยละ 96.72 เมื่อนำเทคนิค Bagging มาเพิ่มประสิทธิภาพให้แก่เทคนิคพื้นฐานพบว่า เทคนิค Bagging สามารถเพิ่มค่าความไวให้กับเทคนิค PART MLP และ SVM โดยเพิ่มค่าจำเพาะให้กับเทคนิค NB PART และ MLP และสามารถเพิ่มค่าความถูกต้องให้กับแบบจำลอง NB PART MLP และ SVM นอกจากนั้นเทคนิค Bagging ยังสามารถเพิ่มความเสถียรให้แก่แบบจำลองโดยให้ค่าความถูกต้องในการพยากรณ์การรอดชีวิตของผู้ป่วยมะเร็งเต้านมด้วยเทคนิค PART และเทคนิค MLP เพิ่มขึ้นอีกด้วย

ผลการวิเคราะห์ข้างต้นแสดงให้เห็นว่าเทคนิค PART ร่วมกับ Bagging มีความเหมาะสมในการนำไปใช้ในการสร้างแบบจำลองเนื่องจากมีประสิทธิภาพสูงที่สุดโดยมีค่าความไว ค่าจำเพาะ และค่าความถูกต้องสูงถึงร้อยละ 99.39 ร้อยละ 96.85 และร้อยละ 98.89 ตามลำดับ และจากกฎที่สร้างจากเทคนิค PART สามารถที่นำไปพัฒนาเป็นระบบเฝ้าระวังโรคให้แก่ผู้ป่วย และแพทย์ได้นำการพยากรณ์ในการคัดกรองโรคเบื้องต้น

กิตติกรรมประกาศ

คณะผู้วิจัยขอขอบคุณเว็บไซต์ SEER ที่ให้ข้อมูลมาใช้ในการวิเคราะห์แบบจำลอง เพื่อการพยากรณ์การรอดชีวิตของผู้ป่วยมะเร็งเต้านม และขอขอบคุณคณะวิทยาการสารสนเทศ มหาวิทยาลัยมหาสารคามที่ให้ทุนในการศึกษางานวิจัยในครั้งนี้

References

[1] Siegel, R., Miller, K., and Jemal, A. (2018). Cancer Statistics, 2018. **CA: A Cancer Journal Clinic**. Vol. 68, Issue 1, pp. 7-30. DOI: 10.3322/caac.21442

[2] Cancer Research UK. **Stages of Cancer**. Access (12 April 2020). Avaliable (http://www.cancerresearchuk.org/about-cancer/what-is-cancer/stages-of-cancer#types)

[3] HD Editorial Department. **Breast Cancer, Stage 3, Treatment and Survival Rate**. Access (13 January 2020). Avaliable (https://www.honestdocs.co/stage-3-breast-cancer)

[4] Delen, D. and Patil, N. (2006). Knowledge Extraction from Prostate Cancer Data. In **Proceeding of the 39th Annual Hawaii International Conference on System Sciences**. pp. 92b-92b. USA: IEEE Publisher

[5] Umezu, T., Shibata, K., Kajiyama, H., Yamamoto, E., Mizuno, M., and Kikkawa, F. (2012). Prognostic Factors in Stage IA-IIA Cervical Cancer Patients Treated Surgically: Does the Waiting Time to the Operation Affect Survival?. **Archives of Gynecology and Obstetrics**. Vol. 285, No. 2, pp. 493-497. DOI: 10.1007/s00404-011-1966-y

[6] Poum, A., Kamsa-ard, S., and Promthet, S. (2012). Survival Rates of Breast Cancer: a Hospital-Based Study from Northeast of Thailand. **Asian Pacific Journal of Cancer Prevention**. Vol. 13, Issue 3, pp. 791-794. DOI: 10.7314/APJCP.2012.13.3.791

[7] Zhang, Y.-C. and Sakhanenko, L. (2019). The Naive Bayes Classifier for Functional Data. **Statistics & Probability Letters**. Vol. 152, pp. 137-146. DOI: 10.1016/j.spl.2019.04.017

[8] Liu, B., Blasch, E., Chen, Y., Shen, D., and Chen, G. (2013). Scalable Sentiment Classification for Big Data Analysis using Naïve Bayes Classifier. In **Proceeding of the International Conference on Big Data**. pp. 99-104. USA: IEEE Publisher

[9] Khunsuk, T. and Thongkam, J. (2020). Feature Selection Method for Improving Customer Reviews Classification. **RMUTI JOURNAL Science and Technology**. Vol. 13, No. 1, pp. 132-145

[10] Mazid, M. M., Ali, A. B. M. S., and Tickle, K. S. (2009). A Comparison Between Rule Based and Association Rule Mining Algorithms. In **Proceeding of the Third International Conference on Network and System Security**. pp. 452-455. Australia: IEEE Publisher

[11] Frank, E. and Witten, I. H. (1998). Generating Accurate Rule Sets Without Global Optimization. In **Proceeding of the 15th International Conference on Machine Learning**. pp. 144-151. USA: DBLP Publisher

[12] Mendes Souza, G. C. and Moreno, R. L. (2018). Netlab MLP - Performance Evaluation for Pattern Recognition in Myoletric Signal. **Procedia Computer Science**. Vol. 130, pp. 932-938. DOI: 10.1016/j.procs.2018.04.092

[13] Sun, N., Sun, B., Lin, J., and Wu, M. Y. -C. (2018). Lossless Pruned Naive Bayes for Big Data Classifications. **Big Data Research**. Vol. 14, pp. 27-36. DOI: DOI: 10.1016/j.bdr.2018.05.007

[14] Tapak, L., Shirmohammadi-Khorram, N., Amini, P., Alafchi, B., Hamidi, O., and Poorolajal, J. (2019). Prediction of Survival and Metastasis in Breast Cancer Patients using Machine Learning Classifiers. **Clinical Epidemiology and Global Health**. Vol. 7, Issue 3, pp. 293-299. DOI: 10.1016/ j.cegh.2018.10.003

[15] Traganitis, P. A., Pagès-Zamora, A., and Giannakis, G. B. (2017). Learning from Unequally Reliable Blind Ensembles of Classifiers. In **2017 IEEE Global Conference on Signal and Information Processing (GlobalSIP)**. Montreal, QC, Canada. pp. 106-110. DOI: 10.1109/ GlobalSIP.2017.8308613

[16] Zhang, D., Jiao, L., Bai, X., Wang, S., and Hou, B. (2018). A Robust Semi-Supervised SVM via Ensemble Learning. **Applied Soft Computing**. Vol. 65, No. C, pp. 632-643. DOI: 10. 1016/ j.asoc. 2018.01.038

[17] Han, J. W. and Kamber, M. (2006). **Data Miming Concepts and Techniques**. New York: Morgan Kaufmann

[18] John, G. H. and Langley, P. (1995). Estimating Continuous Distributions in Bayesian Classifiers. In **Proceeding of the 7th Conference on Uncertainty in Artificial Intelligence**. San Mateo: ACM Publisher. pp. 338-345

[19] Zhan, Q., Motlicek, P., Du, S., Shan, Y., Ma, S., and Xie, X. (2019). Cross-lingual Automatic Speech Recognition Exploiting Articulatory Features. In **Proceedings of APSIPA Annual Summit and Conference 2019**. pp. 1912-1916

[20 Chang, C.-C. and Lin, C.-J. (2001). **LIBSVM - A Library for Support Vector Machines**. Access (20 August 2019) Avaiable (http://www.csie.ntu.edu.tw/~cjlin/libsvm/)

[21] Breiman, L. (1996). Bagging Predictors. **Machine Learning**. Vol. 24, pp. 123-140