การเปรียบเทียบประสิทธิภาพการจำแนกประเภทชุดข้อมูลผู้ป่วยมะเร็งเต้านม โดยใช้เทคนิคเหมืองข้อมูล

อาทิตยา กะการดี[1] , ไก้รุ่ง เฮงพระพรหม[1*] และ สุพจน์ เฮงพระพรหม[1]

[1] สาขาวิชาวิทยาการข้อมูล คณะวิทยาศาสตร์และเทคโนโลยี มหาวิทยาลัยราชภัฏนครปฐม
*kairung2011.heng@gmail.com

บทคัดย่อ

งานวิจัยนี้มีวัตถุประสงค์เพื่อ 1) เพื่อศึกษาแนวทางในการจำแนกข้อมูลผู้ป่วยโรคมะเร็งเต้านม 2) เพื่อเปรียบเทียบประสิทธิภาพในการจำแนกข้อมูลผู้ป่วยโรคมะเร็งเต้านมด้วยเทคนิคการจำแนกข้อมูล 3 ได้แก่ วิธีเพื่อนบ้านที่ใกล้ที่สุด ต้นไม้ตัดสินใจ และนาอีฟเบย์ จากชุดข้อมูลข้อมูลผู้ป่วยโรคมะเร็งเต้านม ผลการวิจัยนี้ได้ทำการจำแนกข้อมูลผู้ป่วยโรคมะเร็งเต้านมด้วยวิธีการใช้เทคนิคเหมืองข้อมูล 3 วิธีการ ได้แก่ 3 เทคนิควิธี เพื่อนบ้านที่ใกล้ที่สุด ต้นไม้ตัดสินใจ และนาอีฟเบย์ ผลการทดลองพบว่า การจำแนกข้อมูลเทคนิคนาอีฟเบย์ให้ประสิทธิภาพที่ดีที่สุด โดยให้ผลค่าความถูกต้อง 60% , ค่าความแม่นยำ 70% , ค่าความระลึก 63% และค่าคะแนนเอฟหนึ่ง 57% รองลงมา คือ ต้นไม้ตัดสินใจโดยให้ผลค่าความถูกต้อง 60% , ค่าความแม่นยำ 61% , ค่าความระลึก 61% และค่าคะแนนเอฟหนึ่ง 60% สุดท้าย คือ วิธีเพื่อนบ้านที่ใกล้ที่สุดโดยให้ผลค่าความถูกต้อ 46% , ค่าความแม่นยำ 23% , ค่าความระลึก 50% และค่าคะแนนเอฟหนึ่ง 31% ตามลำดับ

คำสำคัญ: วิธีเพื่อนบ้านที่ใกล้ที่สุด ต้นไม้ตัดสินใจ นาอีฟเบย์ ชุดข้อมูลผู้ป่วยโรคมะเร็งเต้านม การจำแนกข้อมูล

A comparison of classification efficiency of breast cancer patient datasets using data mining techniques

Artitaya Kakandee[1], Kairung Hengpraprohm[1*], and Supojn Hengpraprohm[1]

[1]Program in data science, Faculty of Science and Technology, Nakhon Pathom Rajabhat University
*kairung2011.heng@gmail.com

Abstract

The objectives of this research are: 1) to study a method for classifying breast cancer patient data, and 2) to compare the efficiency in breast cancer patient classification. Three data classification techniques, namely the k-nearest neighbor, decision tree, and naive bayes are used to test with the breast cancer patient dataset. The experimental results show that the classification of the naive bayes technique gives the best performance with 60% accuracy, 70% precision, 63% recall, and 57% f1-score. Followed by a decision tree that yields 60% accuracy, 61% precision, 61% recall, and 60% f1-score. Finally, the k-nearest neighbor method yields 46% accuracy, 23% precision, 50% recall, and 31% f1-score, respectively.

Keywords: K-Nearest Neighbor, Decision Tree, Naive Bayes, Breast Cancer Dataset, Data Classification

1. บทนำ

เต้านม อวัยวะที่แสดงถึงลักษณะทางเพศหญิงอย่างหนึ่ง ประกอบไปด้วยต่อมน้ำนม ท่อน้ำนม ไขมัน เส้นเลือด ต่อมน้ำเหลือง เต้านมวางอยู่บนกล้ามเนื้อหน้าอกและซี่โครง โดยเต้านมจะขยายขนาดตอบสนองกับฮอร์โมนเอสโตรเจน (Estrogen) ซึ่งผลิตจากรังไข่เป็นหลัก ที่ทำหน้าที่รวบรวมน้ำนมทีผลิตจากต่อมน้ำนมมายังหัวนม เซลล์ต่างๆ เหล่านี้สามารถกลายพันธุ์เกิดเป็นมะเร็งได้ทั้งนั้น แต่ที่พบบ่อยที่สุดที่เกิดความผิดปกติ และทำให้เกิดมะเร็งเต้านม คือ เซลล์ท่อน้ำนม ดังนั้นมะเร็งเต้านมชนิดที่พบบ่อยที่สุด จึงมีชื่อเรียกว่า (invasive ductal carcinoma) และชนิดของมะเร็งที่พบน้อย คือ มะเร็งของต่อมน้ำนม ที่มีชื่อเรียกว่า (invasive lobular carcinoma) ซึ่งมะเร็งทั้งสองชนิดนี้มี วิธีการรักษาเหมือนกัน และอีกชนิดสุดท้าย ซึ่งพบเป็นก้อนที่เต้านมเกิดจากมะเร็งจากที่อื่นแพร่กระจายมา เรียกว่า (metastatic carcinoma)

มะเร็งเต้านมคือ ความผิดปกติของเซลล์เต้านมที่ไม่สามารถควบคุมได้และมีการ แพร่กระจายของเซลล์มะเร็งไปยังเนื้อเยื่อเต้านมและแพร่กระจายไปยังบริเวณอื่น ๆ ของร่างกาย โดย ส่วนมากจะพบในเพศหญิง ส่วนเพศชายก็พบได้เหมือนกัน (American Cancer Society, 2011) มะเร็งเต้านมเป็นปัญหาสุขภาพที่สำคัญของสตรีทั่วโลก เป็นสาเหตุการเสียชีวิตด้วยโรคมะเร็งเป็นอันดับสองรองจากมะเร็งปอดจากรายงานการสาธารณสุขไทย (Thailand Health Profile 2005-2007) มะเร็งเต้านมที่พบในสตรีไทยมีแนวโน้มสูงขึ้น ตั้งแต่ปีพ.ศ. 2533 พบอัตราป่วย 13.5 ต่อ ประชากร 1 แสนคน, ปี พ.ศ. 2536 พบอัตราป่วย 16.3 ต่อประชากร 1 แสนคน, ปี พ.ศ. 2539 พบอัตรา ป่วย 17.2 ต่อประชากร 1 แสนคน, ปี พ.ศ. 2542 พบอัตราป่วย 19.9 ต่อประชากร 1 แสนคน และปี พ.ศ. 2553 พบอัตราป่วย 20.5 ต่อประชากร 1 แสนคนตามล าดับอัตราการเป็นมะเร็งเต้านมพบมากขึ้นเรื่อย ๆ ประมาณว่า 1ใน 10ของสตรีมีโอกาสที่จะเป็นมะเร็งเต้านมในช่วงหนึ่งของชีวิต ดังนั้นจึงมีการ

ตื่นตัวในการตรวจหาและรักษาปัญหาก่อนที่เต้านมเพื่อให้ได้การวินิจฉัยมะเร็งเต้านมในระยะแรก และรักษาก่อนที่จะมีการแพร่กระจายของโรคออกไป

ผู้วิจัยได้ศึกษาหาข้อมูลงานวิจัยที่เกี่ยวกับมะเร็งเต้านมด้วยเทคนิคการทำเหมืองข้อมูลของหลายๆมหาวิทยาลัย และสถาบันอื่นๆมา พบว่ามีการใช้วิธีการจำแนกประเภทข้อมูลด้วยเทคนิคเหมืองข้อมูลในโดเมนที่แตกต่างกัน และมีวิธีการทำ การคิด และการวิเคราะห์ที่แตกต่างกัน จากนั้นได้มีการทำการเปรียบเทียบเทคนิคต่างๆ ในงานวิจัยที่เคยทำมาแล้วเพื่อนำมาเลือกเทคนิคในการทำเหมืองข้อมูลที่เหมาะสมในการทำวิจัยของโรคมะเร็งเต้านม ดังนั้นในการใช้เทคนิคในวิจัยนี้ คือ Naive Bayes, Decision tree และ k – nearest neighbor

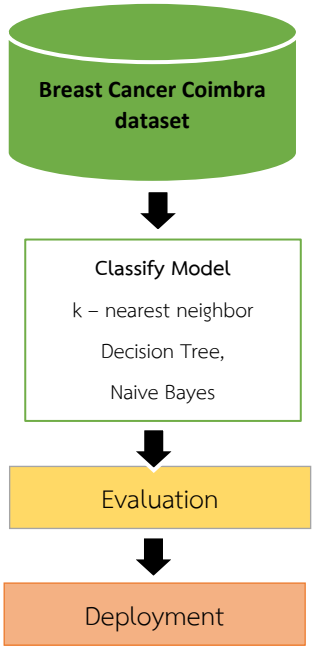
2. วัตถุประสงค์การวิจัย

2.1 เพื่อศึกษาเทคนิคเหมืองข้อมูลสำหรับการจำแนกชุดข้อมูลผู้ป่วยโรคมะเร็งเต้านม

2.2 เพื่อเปรียบเทียบประสิทธิภาพของเทคนิคการจำแนกประเภทผู้ป่วยโรคมะเร็งเต้านม ได้แก่ Naive Bayes, Decision tree และ k – nearest neighbor

3. วิธีดำเนินการวิจัย

3.1 กรอบแนวคิดในการวิจัย



ภาพที่ 1 กรอบแนวคิดในการวิจัย

3.2 ข้อมูลสำหรับการวิจัย

ชุดข้อมูล Breast Cancer Coimbra ที่นำมาใช้ในงานวิจัยได้นำมาจากแหล่งข้อมูลจากเว็บไซต์ของ UCI Machine Learning [1] ในชุดของ Breast Cancer Coimbra Dataset เป็นชุดข้อมูลเกี่ยวกับโรคมะเร็งเต้านม ภายในจะมี 10 คอลัมน์ และ 116 แถว โดยข้อมูลมีคุณลักษณะประกอบด้วยรายละเอียดคือ อายุ (Age), ค่าดัชนีมวลกาย (BMI), ค่าน้ำตาล (Glucose), ค่าฮอร์โมน (Insulin), ค่าคะแนนที่ใช้วัดความดื้อต่อฮอร์โมนอินซูลิน (HOMA หรือ HOMA-IR), ค่าฮอร์โมนที่ช่วยในการควบคุมความหิว (Leptin), ค่าฮอร์โมนจากเนื้อเยื่อไขมัน1 (Adiponectin), ค่าฮอร์โมนจากเนื้อเยื่อไขมัน

2 (Resistin), ค่าการแสดงออกของยีนส์ (MCP.1 หรือ monocyte chemoattractant protein-1) และการจำแนกกลุ่มข้อมูล (Classification)

ตารางที่ 1 แสดงโครงสร้างและข้อมูลโรคมะเร็งเต้านม (Breast Cancer Coimbra Dataset)

Age	BMI	Glucose	Insulin	HOMA	Leptin	Adiponectin	Resistin	MCP.1	Classification
48	23.5	70	2.707	0.467409	8.8071	9.7024	7.99585	417.114	1
83	20.69049	92	3.115	0.706897	8.8438	5.429285	4.06405	468.786	1
45	21.30395	102	13.852	3.485163	7.6476	21.05663	23.03408	552.444	2
45	20.83	74	4.56	0.832352	7.7529	8.237405	28.0323	382.955	2
49	20.95661	94	12.305	2.853119	11.2406	8.412175	23.1177	573.63	2
45	37.03561	83	6.76	1.383997	39.9802	4.617125	8.70448	586.173	1

3.3 ขั้นตอนในการดำเนินการวิจัย (ขั้นตอนการทำ Data mining-KDD)

การดำเนินการวิจัย ประกอบด้วยกรอบแนวคิดในการวิจัยดังนี้

3.3.1 ทำการศึกษาลักษณะชุดข้อมูล Breast Cancer Coimbra

3.3.2 จำแนกประเภทกลุ่มโรค (Classify) จากนั้นนำข้อมูลที่ผ่านการแปลงข้อมูลนำเข้าสู่วิธีการจำแนก ข้อมูลด้วยการใช้ Data Mining Techniques คือ k – nearest neighbor, Decision Tree และ Naive Bayes เพื่อเปรียบเทียบประสิทธิภาพการจำแนกที่มีค่า Accuracy, Precision, Recall, F1-score ด้วยการใช้ซอฟต์แวร์ Google Collaboratory ในการจำแนกดังกล่าว

3.3.3 ประเมินผล (Evaluation) เพื่อนำผลที่ได้จากการใช้ Google Collaboratory มาพิจารณาว่าเทคนิคเหมืองข้อมูลไหนมีประสิทธิภาพที่เหมาะสมที่สุด

3.3.4 นำแบบจำลองใปใช้งาน (Deployment) นำผลการวิเคราะห์ของแบบจำลองที่ทำการเปรียบเทียบประสิทธิภาพเพื่อได้แบบจำลองเทคนิคที่มีความน่าเชื่อถือ และสามารถนำผลที่ได้มาใช้ให้เกิดประโยชน์เกี่ยวกับทางการแพทย์ได้ต่อไป

3.4 การประเมิลผลการวิจัย

การวิจัยนี้ใช้วิธีการประเมินประสิทธิภาพในการจำแนกข้อมูล [2] ด้วยค่าความถูกต้อง (Accuracy), ค่าความแม่นยำ (precision), ค่าความระลึก (recall) และค่าเฉลี่ย (F1-score)

3.4.1 ค่าความถูกต้อง (Accuracy) คือ มาตรวัดค่าความแม่นตรง คือ ค่าที่บอกถึงความแม่นยำในการจำแนกข้อมูล จากสมการ

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

โดยที่ TP คือค่า True Positive, TN คือค่า True Negative,

FP คือค่า False Positive, FN คือค่า False Negative

3.4.2 ค่าความแม่นยำ (precision) คือ ความสามารถของเครื่องมือวัดที่จะบอกค่าที่วัดได้เป็นค่าใดค่าหนึ่งซ้ำเสมอ ซึ่งค่าที่อ่านได้อาจไม่ใช้ค่าที่แท้จริงก็ได้ จากสมการ

$$\text{Precision} = 1 - \left| \frac{x_n - \bar{x}_n}{\bar{x}_n} \right|$$

โดยที่

$$\bar{x}_n = \frac{\sum x}{n}$$

3.4.3 ค่าความระลึก(recall) คือ การวัดความสามารถของระบบในการค้นพบข้อมูลที่เกี่ยวข้อง โดยค่าความระลึก เป็นอัตราส่วนของจำนวนข้อมูลที่เกี่ยวข้องและถูกดึงออกมากับจำนวนข้อมูลที่เกี่ยวข้องทั้งหมด จากสมการ

$$Recall = TP(TP + FN)$$

โดยที่ TP คือค่า True Positive

FN คือค่า False Negative

3.4.4 ค่าเฉลี่ย (F1-score) คือ ค่าเฉลี่ยแบบ harmonic mean ระหว่าง precision และ recall นักวิจัยสร้าง F1 ขึ้นมาเพื่อเป็น single metric ที่วัดความสามารถของโมเดล (ไม่ต้องเลือกระหว่าง precision, recall เพราะเฉลี่ยให้แล้ว) จากสมการ

$$F1 = 2 * \frac{(P*R)}{(P+R)}$$

โดยที่ P คือค่า precision

R คือค่า recall

4. ทฤษฎีและงานวิจัยที่เกี่ยวข้อง

4.1 ทฤษฎีที่เกี่ยวข้อง

4.1.1 เหมืองข้อมูล

เหมืองข้อมูล (data Mining) เป็นเทคนิคในการวิเคราะห์ข้อมูลอย่างหนึ่ง ซึ่งมาจากคำว่า เหมืองข้อมูล นั่นคือเป็นการค้นหาสิ่งที่มีประโยชน์จากฐานข้อมูลที่มีขนาดใหญ่ เช่น ข้อมูลการซื้อขายสินค้าในซุปเปอร์มาร์เก็ตต่าง ๆ โดยข้อมูลเหล่านี้จะเก็บจากรายการสินค้าที่ลูกค้าซื้อในแต่ละครั้ง โดยเมื่อทำการวิเคราะห์ข้อมูลด้วยเทคนิค Data Mining แล้วจะได้สิ่งที่เป็นประโยชน์

4.1.2 วิธีการเพื่อนบ้านใกล้ที่สุด

ขั้นตอนวิธีการเพื่อนบ้านใกล้ที่สุด (K-Nearest Neighbor Algorithm : Knn) เป็นวิธีที่ใช้ในการจัดแบ่งคลาส โดยเทคนิคนี้จะตัดสินใจว่า คลาสใดที่จะแทนเงื่อนไขหรือกรณีใหม่ๆ ได้บ้าง โดยการตรวจสอบจำนวนบางจำนวน ในขั้นตอนวิธีการเพื่อนบ้านใกล้ที่สุด ของกรณีหรือเงื่อนไขที่เหมือนกันหรือใกล้เคียงกันมากที่สุด โดยจะหาผลรวม (Count Up) ของจำนวนเงื่อนไข หรือกรณีต่างๆ สำหรับแต่ละคลาส และกำหนดเงื่อนไขใหม่ๆ ให้คลาสที่เหมือนกันกับคลาสที่ใกล้เคียงกันมากที่สุด

4.1.3 ต้นไม้การตัดสินใจ

ต้นไม้การตัดสินใจ (decision tree) เป็นเครื่องมือที่ช่วยให้วิเคราะห์เหตุการณ์ หรือสถานการณ์เพื่อการตัดสินใจได้อย่างเป็นระบบและรวดเร็ว ต้นไม้การตัดสินใจมีลักษณะเป็นกราฟรูปต้นไม้ ซึ่งแสดงที่ตั้งต้นที่มีรากและแขนงต่างๆ แตกออกมาจากต้นไม้ไปในทิศทางเดียว จนกระทั่งนำไปสู่ข้อสรุปสำหรับการตัดสินใจได้ ต้นไม้การตัดสินใจมีประโยชน์ในการ

สรุปการตัดสินใจที่มีความซับซ้อนให้ง่ายต่อความเข้าใจ ปัจจุบันต้นไม้การตัดสินใจเป็นที่นิยมใช้ในงานหลายอย่าง เช่น การแพทย์ ธุรกิจ การเขียนโปรแกรม การสร้างเครื่องที่เรียนรู้ได้เอง การสร้างระบบผู้เชี่ยวชาญ ฯลฯ

4.1.4 เทคนิคเนอีฟเบย์

อัลกอริทึมนาอีฟเบย์ (Naive Bayes) หมายถึง เครื่องจักรเรียนรู้ที่อาศัยหลักการความน่าจะเป็น ตามทฤษฎีของเบย์(Bayes Theorem) ซึ่งมีอลักอริทึมที่ไม่ซับซ้อน เป็นขั้นตอนวิธีในการจำแนกข้อมูลโดยการเรียนรู้ปัญหาที่เกิดขึ้น เพื่อนำมาสร้างเงื่อนไขการจำแนกข้อมูลใหม่ หลักการของนาอีฟเบย์ใช้การคำนวณหาความน่าจะเป็นในการทำนายผลเป็นเทคนิคในการแก้ปัญหาแบบจำแนกประเภทที่สามารถคาดการณ์ผลลัพธ์ได้จะทำการวิเคราะห์ความสัมพันธ์ระหว่างตัวแปรเพื่อใช้ในการสร้างเงื่อนไขความน่าจะเป็นสำหรับแต่ละความสัมพันธ์เหมาะกับกรณีของเซตตัวอย่างที่มีจำนวนมาก และคุณสมบัติ (Attribute)

4.2 งานวิจัยที่เกี่ยวข้อง

การวิจัยครั้งนี้ [3] มีวัตถุประสงค์ 1) เพื่อศึกษาการรับรู้เกี่ยวกับมะเร็งเต้านมและการตรวจเต้านมด้วย ตนเองของประชากรเพศหญิงในจังหวัดกรุงเทพมหานคร 2) เพื่อศึกษาความรู้เกี่ยวทักษะการตรวจเต้านมด้วยตนเองและพฤติกรรมเกี่ยวกับการตรวจเต้านมด้วยตนเอง 3) เพื่อศึกษาเกี่ยวกับปัจจัยที่ส่งผลต่อการตัดสินใจเข้ารับการตรวจคัดกรองมะเร็งเต้านมของประชากรเพศหญิง ในจังหวัดกรุงเทพมหานคร และ4) เพื่อรวบรวมข้อเสนอแนะความคิดเห็นเกี่ยวกับการเข้ารับการตรวจคัดกรองมะเร็งเต้านมประชากรเพศหญิงในจังหวัดกรุงเทพมหานคร กำหนดกลุ่มตัวอย่างที่ใช้ในการวิจัยครั้งนี้ คือ สตรีในเขตกรุงเทพมหานครจำนวน 400 คน โดยผู้วิจัยเลือกเทคนิคการสุ่มตัวอย่างแบบการสุ่มตัวอย่างแบบเฉพาะเจาะจง (Purposive Sampling) เครื่องมือที่ใช้ในการวิจัยครั้งนี้ คือ แบบสอบถาม (Questionnaire) ซึ่งผู้วิจัยสร้างขึ้นเพื่อเป็น เครื่องมือในการเก็บรวบรวมข้อมูลจากกลุ่มตัวอย่าง การวิเคราะห์ข้อมูลใช้ค่าสถิติโดยวิธีหาค่าร้อยละ ค่าเฉลี่ยเลขคณิต และส่วนเบี่ยงเบนมาตรฐาน การทดสอบสมมติฐานวิเคราะห์ข้อมูลโดยการหาค่าสถิติ ทดสอบที (t-test) เพื่อทดสอบตัวแปรสองกลุ่ม โดยใช้การวิเคราะห์ความแปรปรวนทางเดียว (One Way ANOVA) เมื่อพบความแตกต่างจะทำการทดสอบความแตกต่างเป็นรายคู่ ด้วยวิธีการของ LSD และการวิเคราะห์ถดถอยพหุคูณ (Regression analysis) โดยกำหนดการทดสอบ นัยสำคัญทางสถิติที่ระดับ 0.5

งานวิจัยฉบับนี้ [4] มีวัตถุประสงคเพื่อพัฒนาแบบจำลองที่มีประสิทธิภาพในการพยากรณการรอดชีวิตของผู้ป่วยมะเร็งเต้านมซึ่งเป็นมะเร็งที่พบมากในเพศหญิงเป็นอันดับที่สองรองจากมะเร็งรังไข่ ข้อมูลเก็บรวบรวมจากฐานขอมูล SEER ในปี ค.ศ. 2004 ถึง 2014 จำนวน 115,184 ระเบียน การวิจัยนี้ใชเทคนิคเหมืองข้อมูลพื้นฐาน คือ เทคนิคนาอีฟเบย์ เทคนิคส่วนของรายการตัดสินใจ เทคนิคเพอร์เซปตรอนหลายชั้น และเทคนิคซัพพอรตเวกเตอรแมชชีน ในการสรางแบบจำลองเปรียบเทียบกับแบบจำลองดังกลาวร่วมกับเทคนิคการหอเพื่อเพิ่มประสิทธิภาพการพยากรณ คณะผู้วิจัยใชหลักการ 10-โฟดครอสวาลิเดชันในการ แบ่งชุดข้อมูลเป็นชุดข้อมูลฝึกสอนและชุดข้อมูลทดสอบ โดยใช้ค่าความไว ความจำเพาะ และความถูกต้อง วัดประสิทธิภาพของแบบจำลอง ผลทดลองพบว่า เทคนิคส่วนของรายการตัดสินใจร่วมกับเทคนิคการห่อสามารถสรางแบบจำลองการพยากรณ์ การรอดชีวิตของผู้ป่วยมะเร็งเต้านมที่มีความถูกต้องสูงสุดที่ร้อยละ 98.89

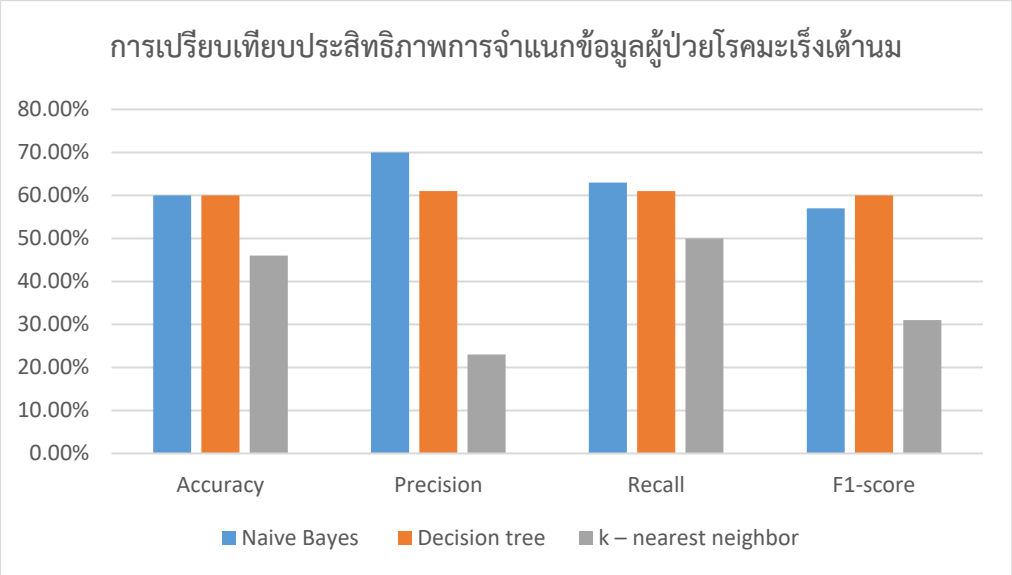
การสร้างแบบจำลองเพื่อการพยากรณ์ตามกฎให้มีประสิทธิภาพสูงเป็นงานวิจัยที่ท้าท้าย งานวิจัยนี้ [5] จึงมีวัตถุประสงค์ เพื่อสร้างแบบจำลองที่มีประสิทธิภาพในการพยากรณ์โดยใช้แบบจำลองพื้นฐาน คือ FURIA, MODLEM และ RIPPER และเทคนิคแบบรวมที่เป็นที่นิยม คือ Bagging และWeighted Instances Handler Wrapper (WI) โดยใช้ข้อมูลผู้ป่วยโรคมะเร็งเต้านม จำนวน 699 คน และข้อมูลผู้ป่วยโรคเบาหวานจำนวน 768 คน การวัดประสิทธิภาพแบบจำลองได้ทดลองแบ่งชุดข้อมูลออกเป็นชุดข้อมูลเรียนรู้และชุดข้อมูลทดสอบด้วยการใช้หลักการ10-fold cross validation และ ทำการทดลอง 10 รอบ เพื่อลดความลำเอียงของการทดลองในการวัดประสิทธิภาพการพยากรณ์ของแบบจำลองที่สร้างจากแต่ละเทคนิคด้วยค่า Sensitivity, Specificity และ Accuracy จากการศึกษาพบว่า เทคนิค Bagging สามารถเพิ่มค่า Accuracy ในการพยากรณ์ การเกิดโรคมะเร็งเต้านมได้สูงถึง 4.91%

งานวิจัยนี้ [6] มีวัตถุประสงค์เพื่อเปรียบเทียบประสิทธิภาพในการจำแนก 5 วิธี คือ วิธีนาอีฟเบส์ วิธีเพื่อนบ้าน ใกล้สุด k ตัว วิธีต้นไม้ตัดสินใจ วิธีโครงข่ายประสาทเทียม และวิธีซัพพอร์ตเวกเตอร์แมชชีน โดยพิจารณาจากค่า ความถูกต้อง ค่าคลาดเคลื่อนกำลังสองเฉลี่ยและค่าส่วนเบี่ยงเบนสัมบูรณ์เฉลี่ย และเพื่อเปรียบเทียบวิธีการสุ่ม ตัวอย่างระหว่างโปรแกรม SPSS และ WEKA โดยแบ่งข้อมูลเป็นชุดข้อมูลเรียนรู้ ชุดข้อมูลตรวจสอบความถูกต้อง และชุดข้อมูลทดสอบ ในอัตราส่วน 70, 20 และ 10 ตามลำดับ สำหรับการค้นคว้าและศึกษาค่านอกเกณฑ์ได้ใช้ ข้อมูลมีข้อมูล 3 ชุด คือ โรคมะเร็งเต้านมของรัฐวิสคอนซิน เป็นชุดข้อมูลที่มีค่านอกเกณฑ์อยู่ในระดับต่ำ โรคเบาหวานของชาวพม่า ประเทศอินเดีย เป็นชุดข้อมูลที่มีค่านอกเกณฑ์อยู่ในระดับปานกลาง และการชำระเงิน ด้วยบัตรเครดิตของลูกค้า เป็นชุดข้อมูลที่มีค่านอกเกณฑ์อยู่ในระดับสูง โดยใช้เครื่องมือ Highlight Exceptions ใน การตรวจจับค่านอกเกณฑ์จากการเปรียบเทียบข้อมูลโรคมะเร็งเต้านมของรัฐวิสคอนซิน วิธีที่มีประสิทธิภาพสูงสุด คือ วิธีโครงข่ายประสาทเทียม โดยการสุ่มของโปรแกรม SPSS โรคเบาหวานของชาวพม่า ประเทศอินเดีย วิธีที่มีประสิทธิภาพสูงสุด คือ วิธีเพื่อนบ้านใกล้สุด k ตัว โดยการสุ่มของโปรแกรม SPSS และ WEKA และการชำระเงินด้วยบัตรเครดิตของลูกค้า วิธีที่มีประสิทธิภาพสูงสุด คือ วิธีเพื่อนบ้านใกล้สุด k ตัว โดยการสุ่มของโปรแกรม SPSS และ WEKA ชุดข้อมูลที่มีค่านอกเกณฑ์อยู่ในระดับปานกลางและสูงให้ผลการจำแนกที่เหมือนกัน ซึ่งแตกต่างจากชุด ข้อมูลที่มีค่านอกเกณฑ์ในระดับที่ต่ำ

“มะเร็งเต้านม” เป็นโรคในกลุ่มโรคไม่ติดต่อเรื้อรังที่พบมากเป็นอันดับหนึ่งในผู้หญิงทั้งในประเทศไทยและต่างประเทศ สถิติของสถาบันมะเร็งแห่งชาติระบุว่า ผู้หญิงไทยที่ป่วยเป็นโรคมะเร็งเต้านมมีแนวโน้มที่จะเพิ่มมากขึ้นทุกปี ซึ่งถ้าหากตรวจพบเร็วและสามารถรักษาได้ถูกวิธีกจะสามารถลดอัตราความเสี่ยงในการเสียชีวิตได้อย่างมาก งานวิจัยนี้จึงได้นำเสนออัลกอริทึมการจำแนกประเภทแบบเคมีนร่วมกับค่าถ่วงน้ำหนักแบบปรับตัวเอง รวมทั้งทำการพัฒนาโปรแกรมพยากรณ์โรคมะเร็งเต้านมด้วยภาษาไพทอน โดยมีจุดประสงค์เพื่อช่วยในการคัดกรองผู้ป่วยโรคมะเร็งเต้านมในเบื้องต้น เพื่อให้กระบวนการการวินิจฉัยเป็นไปได้อย่างรวดเร็วและสามารถรักษาได้อย่างทันเวลา โดยอัลกอริทึมที่นำเสนอเป็น อัลกอริทึมที่ถูกพัฒนาต่อยอดมาจากอัลกอริทึมการแบ่งกลุ่มแบบเคมีนให้มีความสามารถในการจำแนกประเภท และปรับเปลี่ยนค่าถ่วงน้ำหนักของคุณลักษณะเด่นในสมการหาระยะห่างระหว่างข้อมูลกับจุดศูนย์กลางของประเภทแต่ละ ประเภทให้เหมาะสมได้ด้วยตัวเองในระหว่างการเรียนรู้ชุดข้อมูลงานวิจัยนี้ได้ทำการทดสอบประสิทธิภาพของอัลกอริทึม และโปรแกรมการพยากรณ์โรคมะเร็งเต้านมที่นำเสนอโดยใช้ชุดข้อมูล Breast Cancer Coimbra ผลการทดสอบแสดงให้ เห็นถึงความถูกต้องในการพยากรณ์โรคมะเร็งเต้านมของอัลกอริทึมที่นำเสนอที่สูงกว่าอัลกอริทึมทางปัญญาประดิษฐ์อื่นๆ

5. ผลการวิจัย

ผลการจำแนกประเภทข้อมูลจากชุดข้อมูลผู้ป่วยโรคมะเร็งเต้านมผ่านการจำแนกข่อมูลด้วยเทคนิคเหมืองข้อมูล 3 วิธี ได้แก่ได้แก่ เพื่อนบ้านที่ใกล้ที่สุด ต้นไม้ตัดสินใจ และนาอีฟเบย์ โดยแบงข่อมูลในการเรียนรู (Training Data) และข่อมูลในการทดสอบ (Testing Data) ตามซอฟต์แวร์ Google Collaboratory ได้แก่ ค่าความแม่นยำ (Accuracy) ค่าความถูกต้อง (precision) และค่าความระลึก (recall) แสดงผลในรูปแบบตารางที่ 2 และภาพที่ 2



ภาพที่ 2 การเปรียบเทียบประสิทธิภาพการจำแนกข้อมูลผู้ป่วยโรคมะเร็งเต้านม

ตารางที่ 2 การเปรียบเทียบแสดงค่าความแม่นยำ ความถูกต้อง และค่าความระลึกในการจำแนกข้อมูลผู้ป่วยโรคมะเร็งเต้านม

Techniques	Accuracy	Precision	Recall	F1-score
Naive Bayes	60.00%	70.00%	63.00%	57.00%
Decision tree	60.00%	61.00%	61.00%	60.00%
k – nearest neighbor	46.00%	23.00%	50.00%	31.00%

จากตารางที่ 2 และภาพที่ 2 แสดงการเปรียบเทียบแสดงค่าความแม่นยำ ความถูกต้อง และค่าความระลึกในการจำแนกข้อมูลผู้ป่วยโรคมะเร็งเต้านม ผลการจำแนกข้อมูล เทคนิค Naive Bayes ให้ประสิทธิภาพที่ดีที่สุด โดยให้ผลค่าความแม่นยำ (accuracy) 60% , ค่าความถูกต้อง (precision) 70% , ค่าความระลึก (recall) 63% และค่าเฉลี่ย (f1-score) 57% รองลงมา คือ Decision Tree โดยให้ผลค่าความแม่นยำ (accuracy) 60% , ค่าความถูกต้อง (precision) 61% , ค่าความระลึก (recall) 61% และค่าเฉลี่ย (f1-score) 60% สุดท้าย คือ k-nearest neighbor โดยให้ผลค่าความแม่นยำ (accuracy) 46% , ค่าความถูกต้อง (precision) 23% , ค่าความระลึก (recall) 50% และค่าเฉลี่ย (f1-score) 31% ตามลำดับ

6. สรุปผล

ผลการวิจัยนี้ได้ทำการจำแนกข้อมูลผู้ป่วยโรคมะเร็งเต้านมด้วยวิธีการใช้เทคนิคเหมืองข้อมูล 3 วิธีการ ได้แก่ 3 เทคนิควิธี เพื่อนบ้านที่ใกล้ที่สุด ต้นไม้ตัดสินใจ และนาอีฟเบย์ ผลการประเมินประสิทธิภาพตัวแบบ คือ ผลการจำแนกข้อมูล เทคนิค Naive Bayes ให้ประสิทธิภาพที่ดีที่สุด โดยให้ผลค่าความแม่นยำ (accuracy) 60% , ค่าความถูกต้อง (precision) 70% , ค่าความระลึก (recall) 63% และค่าเฉลี่ย (f1-score) 57% รองลงมา คือ Decision Tree โดยให้ผลค่าความแม่นยำ (accuracy) 60% , ค่าความถูกต้อง (precision) 61% , ค่าความระลึก (recall) 61% และค่าเฉลี่ย (f1-score) 60% สุดท้าย คือ k-nearest neighbor โดยให้ผลค่าความแม่นยำ (accuracy) 46% , ค่าความถูกต้อง (precision) 23% , ค่าความระลึก (recall) 50% และค่าเฉลี่ย (f1-score) 31% ตามลำดับ จึงสรุปได้ว่า วิธีนาอีฟเบย์เป็นเทคนิคที่เหมาะสมที่สุดที่จะนำมาใช้จำแนกข้อมูลผู้ป่วยโรคมะเรงเต้านมชุดนี้

7. ข้อเสนอแนะ

ในการเลือกข้อมูลเพื่อลดความคลาดเคลื่อนในการนำข้อมูลไปวิเคราะห์ หากชุดข้อมูลเป็นตัวเลขและเป็นข้อมูลที่เป็นการจำแนกประเภท ข้อมูลสามารถนำไปพัฒนาโปรแกรมเพื่อทำการอ่านชุดข้อมูลและแปลงข้อมูลให้อยู่ในรูปแบบตามเทคนิคเหมืองข้อมูลต่างๆ ได้ และในส่วนการจำแนกประเภทข้อมูลสำหรับการวิจัยนี้ด้วยการใช้ซอฟต์แวร์ Google Collaboratory เพื่อนำมาประยุกต์ใช้ในการจำแนกสามารถทดลองด้วยเทคนิควิธีการอื่นเพื่อเปรียบเทียบกับ Naive Bayes จากชุดข้อมูลชุดนี้เพื่อจะได้เทคนิควิธีการจำแนกที่มีหาค่าความถูกต้อง (Accuracy), ค่าความแม่นยำ (precision), ค่าความระลึก (recall) และค่าเฉลี่ย (F1-score) ที่ดีที่สุด

8. เอกสารอ้างอิง

[1] UCI Machine Learning**.** Breast Cancer Coimbra Dataset**.** Retrieved March 21, 2023 from https://archive.ics.uci.edu/dataset/451/breast+cancer+coimbra**.**

[2] Kasidis Satangmongkol. (2019). วิธีการประเมินประสิทธิภาพในการจำแนกข้อมูล. Retrieved March 21, 2023 from https://datarockie.com/blog/top-ten-machine-learning-metrics/ (In Thai)

[3] M.B.A. For Modern Managers.(2001). Factors affecting female population's decision to undergo breast cancer screening in Bangkok Bangkok: Ramkhamhaeng University. Retrieved March 21, 2023 from https://mmm.ru.ac.th/MMM/IS/twin92/6214155565.pdf. (In Thai)

[4] Jaree thongkam and Vatinee Sukmak. (2019). Predicting Breast Cancer Patient Survival. RMUTI JOURNAL Science and Technology, Vol. 14, Issue 1. P.44-54. (In Thai)

[5] Papichaya Klangnok and Jaree Thongkum. (2018). Applying Ensemble Techniques for Improving the Performance of Rule-based Models in Data Mining. Journal of Industrial Technology Ubon Ratchathani Rajabhat University. Vol.9,No.1, P. 97-108. (In Thai)

[6] Panida Sombatmak, Passorn Janhom, Supakorn Rasamee, Oran Rungmaneethamkun and Saichon Sinsomboonthong.(2017).Performance Comparison of Data Mining’s Classification Methods on Data Set with Outliers. Journal of Science and Technology. Vol.27, No.6, P.975-988. (In Thai)