การเพิ่มประสิทธิภาพในการจำแนกข้อมูลโรคมะเร็งปอดด้วยวิธีการกลุ่มก้อนข้อมูล

อภิสิทธิ์ น้ำแก้ว[1] , ไก้รุ่ง เฮงพระพรหม[1*] และ สุพจน์ เฮงพระพรหม[1]

[1] สาขาวิชาวิทยาการข้อมูล คณะวิทยาศาสตร์และเทคโนโลยี มหาวิทยาลัยราชภัฏนครปฐม
*Kairung2011.heng@gmail.com

บทคัดย่อ

งานวิจัยนี้มีวัตถุประสงค์ของ1) เพื่อเปรียบเทียบประสิทธิภาพในการจำแนกข้อมูลโรคมะเร็งปอดด้วยเทคนิคเหมืองข้อมูล 4 วิธี ได้แก่ วิธีต้นไม้ตัดสินใจ, วิธีแบบเบย์, วิธีซัพพอร์ตเวกเตอร์แมทชีน, และโครงข่ายประสาทเทียม 2) เพื่อเพิ่มประสิทธิภาพโมเดลในการจำแนกข้อมูล โดยใช้ชุดข้อมูลโรคมะเร็งปอด ซึ่งประกอบด้วยจำนวนคอลัมน์ 16 คอลัมน์ จำนวนแถว 310 แถว และทดสอบการวัดประสิทธิภาพด้วยวิธี 10-Fold Cross Validation โดยใช้โปรแกรม RapidMiner Studio 9.10 ในการวิเคราะห์ข้อมูล ผลการวิจัย พบว่า เทคนิคโครงข่ายประสาทเทียมให้ประสิทธิภาพที่ดีที่สุดในการจำแนกข้อมูลโรคมะเร็งปอดด้วยค่าความถูกต้องในการจำแนกเท่ากับ 89.25 % ค่าความระลึกเท่ากับ 90.12 % และค่าความแม่นยำเท่ากับ 97.33 % หลังจากนั้นทำการเพิ่มประสิทธิภาพของการจำแนกข้อมูลของโครงข่ายประสาทเทียมด้วยวิธีการกลุ่มก้อนการจำแนกข้อมูลแบบรวมเสียงข้างมากซึ่งทำให้ประสิทธิภาพเพิ่มขึ้นด้วยค่าความถูกต้องในการจำแนกเท่ากับ 91.40 % ค่าความระลึกเท่ากับ 92.59 % และค่าความแม่นยำเท่ากับ 97.40 % ตามลำดับ

คำสำคัญ: เหมืองข้อมูล วิธีต้นไม้ตัดสินใจ วิธีแบบเบย์ วิธีซัพพอร์ตเวกเตอร์แมทชีน การเพิ่มประสิทธิภาพ

Improving the efficiency of lung cancer data classification using ensemble technique

Aphisit Namkaew[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 compare the efficiency of lung cancer data classification using four data mining techniques: Decision Tree, Naive Bayes, Support Vector Machine, and Artificial Neural Network; 2) to improve the data classification model for lung cancer datasets which consisted of 16 columns and 310 rows. The performance was tested by the 10-Fold Cross Validation method using Rapid Miner Studio 9.10 software. The results show that the Artificial Neural Network gave the best efficiency with an accuracy of 89.25 %, recall of 90.12 %, and precision of 97.33 %. After that, the efficiency of data classification of the Artificial Neural Network is improved by using the majority vote ensemble method. It improved the efficiency with a classification accuracy of 91.40 % and a recall of 92.59 % and a precision is 97.40 %, respectively.

Keywords: data mining, decision tree, naive bayes, support vector machine, data classification

1. บทนำ

มะเร็งปอดเป็นมะเร็งที่พบมากที่สุดในโลกและเป็นโรคที่มีผู้เสียชีวิตมากที่สุดในกลุ่มโรคมะเร็ง (โดยเฉลี่ยประมาณ 1 ใน 5 ของผู้ป่วยที่เป็นมะเร็ง) และมีอัตราส่วนระหว่างการเสียชีวิตต่อจำนวนผู้ป่วยสูงถึง 0.87 สำหรับในประเทศไทยมะเร็งปอดเป็นมะเร็งที่พบสูงสุดเป็นอันดับ 1 ในเพศชาย และเป็นอันดับต้นๆรองจากมะเร็งปากมดลูกและมะเร็งเต้านมในเพศหญิง นอกจากนี้มะเร็งปอดยังเป็นสาเหตุของการเสียชีวิตจากมะเร็งมากที่สุดทั้งในเพศชายและหญิง อัตราการรอดชีวิตของผู้ป่วยแตกต่างกันออกไปขึ้นกับระยะของโรค สุขภาพโดยรวม และปัจจัยอื่นๆ โดยผู้ป่วยส่วนมากที่ได้รับการวินิจฉัยว่าเป็นมะเร็งปอดจะเสียชีวิตภายในปีแรกเนื่องจากผู้ป่วยส่วนมากมักได้รับการวินิจฉัยเมื่ออยู่ในระยะท้ายเพราะอาการของโรคคล้ายโรคอื่น ทั้งนี้อัตราการรอดชีวิตที่หนึ่งปีของผู้ป่วยมะเร็งปอดในประเทศไทยโดยรวมน้อยกว่า 20 % การเจ็บป่วยด้วยโรคมะเร็งปอดส่งผลกระทบกับผู้ป่วยได้แก่ความเจ็บปวดที่สามารถเกิดขึ้นได้ทุกระยะของโรค ความกลัว ซึมเศร้า การปรับตัว และการแพร่กระจายของโรค ทำให้ไม่สามารถรักษาโรคให้หายขาดได้ ระยะเวลาการมีชีวิตอยู่ขอผู้ป่วยสั้นลง โดยระยะแพร่กระจายของมะเร็งเต้านมจะมีค่าเฉลี่ยของการรอดชีวิตประมาณ 18 ถึง 24 เดือน นอกจากนี้ยังมีผลของการรักษาที่ผู้ป่วยอาจจะมีอาการข้างเคียงจากการได้รับยาเคมีบำบัด การสูญเสียสัญลักษณ์ทางเพศในรายที่ผ่าตัด อีกทั้งยังส่งผลต่อสัมพันธภาพระหว่างผู้ป่วยและคู่สมรส รวมทั้งงบประมาณของประเทศในการรักษา

จากปัญหาดังกล่าว ผู้วิจัยจึงได้พยากรณ์ การเกิดโรคมะเร็งปอดขึ้นโดยนำเทคนิคเหมืองข้อมูล มาประยุกต์ใช้วิเคราะห์ในการเกิดโรคมะเร็งปอดนอกจากนี้ ยังศึกษาถึงปัจจัยที่ส่งผลต่อความเสี่ยงการเกิดโรคมะเร็งปอดเพื่อเป็นแนวทางในการป้องกันรักษาโรคมะเร็งปอดได้อีกทางหนึ่ง

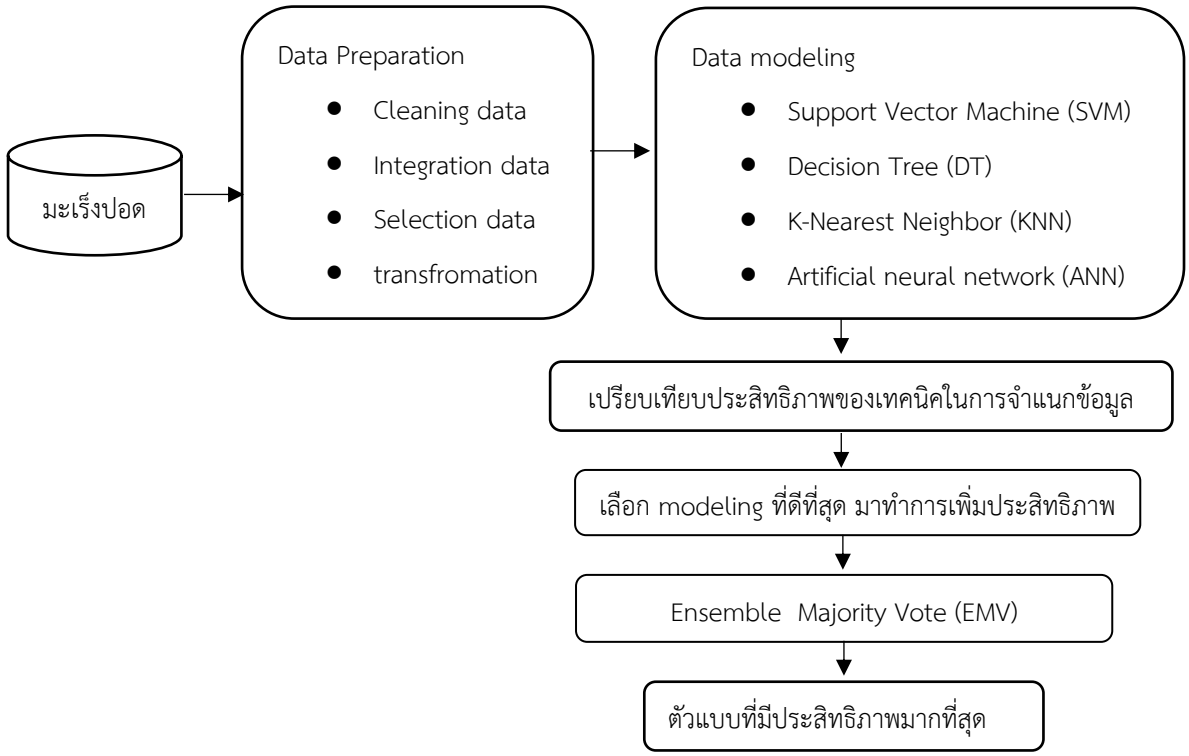
2. วัตถุประสงค์การวิจัย

2.1 เพื่อเปรียบเทียบประสิทธิภาพในการจำแนกข้อมูลด้วยเทคนิคเหมืองข้อมูล ได้แก่ วิธีต้นไม้ตัดสินใจ วิธีแบบเบย์ วิธีซัพพอร์ตเวกเตอร์แมทชีน และโครงข่ายประสาทเทียม

2.2 เพื่อเพิ่มประสิทธิภาพในการจำแนกข้อมูลด้วยวิธีการกลุ่มก้อนการจำแนกข้อมูลแบบรวมเสียงข้างมาก

3. วิธีดำเนินการวิจัย

3.1 กรอบแนวคิดในการวิจัย แสดงเป็นรูปภาพขั้นตอนการออกแบบการวิจัย



ภาพที่ 1 กรอบแนวคิดการวิจัย

3.2 ข้อมูลสำหรับการวิจัย

ชุดข้อมูลโรคมะเร็งเต้านม ประกอบด้วยจำนวนคอลัมน์ 16 คอลัมน์ จำนวนแถว 310 แถว จำนวนคลาสคำตอบ 2 คลาส ดังมีตัวอย่างรายละเอียดข้อมูลดังนี้

คอลัมน์	ข้อมูล	อธิบาย
Gender:เพศ	M(male), F(female)	
Age:อายุ	Age of the patient	
Smoking:การสูบบุหรี่	2 , 1	ใช่=2,ไม่ใช่=1
Yellow fingers:นิ้วเหลือง	2 ,1.	ใช่=2,ไม่ใช่=1
Anxiety:ความวิตกกังวล	2 , 1.	ใช่=2,ไม่ใช่=1
Peer_pressure:แรงกดดันจากเพื่อน	2 , 1.	ใช่=2,ไม่ใช่=1
Chronic Disease:โรคเรื้อรัง	2 , 1.	ใช่=2,ไม่ใช่=1
Fatigue:ความเหนื่อยล้า	2 , 1.	ใช่=2,ไม่ใช่=1
Allergy:ภูมิแพ้	2 , 1	ใช่=2,ไม่ใช่=1
Wheezing:หายใจดังเสียงฮืด ๆ	2 ,1.	ใช่=2,ไม่ใช่=1
Alcohol:แอลกอฮอล์	2 , 1.	ใช่=2,ไม่ใช่=1
Coughing:อาการไอ	2 , 1.	ใช่=2,ไม่ใช่=1
Shortness of Breath:หายใจถี่	2 , 1.	ใช่=2,ไม่ใช่=1
Swallowing Difficulty:ความยากในการกลืน	2 , 1.	ใช่=2,ไม่ใช่=1
Chest pain:อาการเจ็บหน้าอก	2 , 1.	ใช่=2,ไม่ใช่=1
Lung Cancer:มะเร็งปอด	YES , NO.	ใช่,ไม่ใช่

3.3 ขั้นตอนในการดำเนินการวิจัย

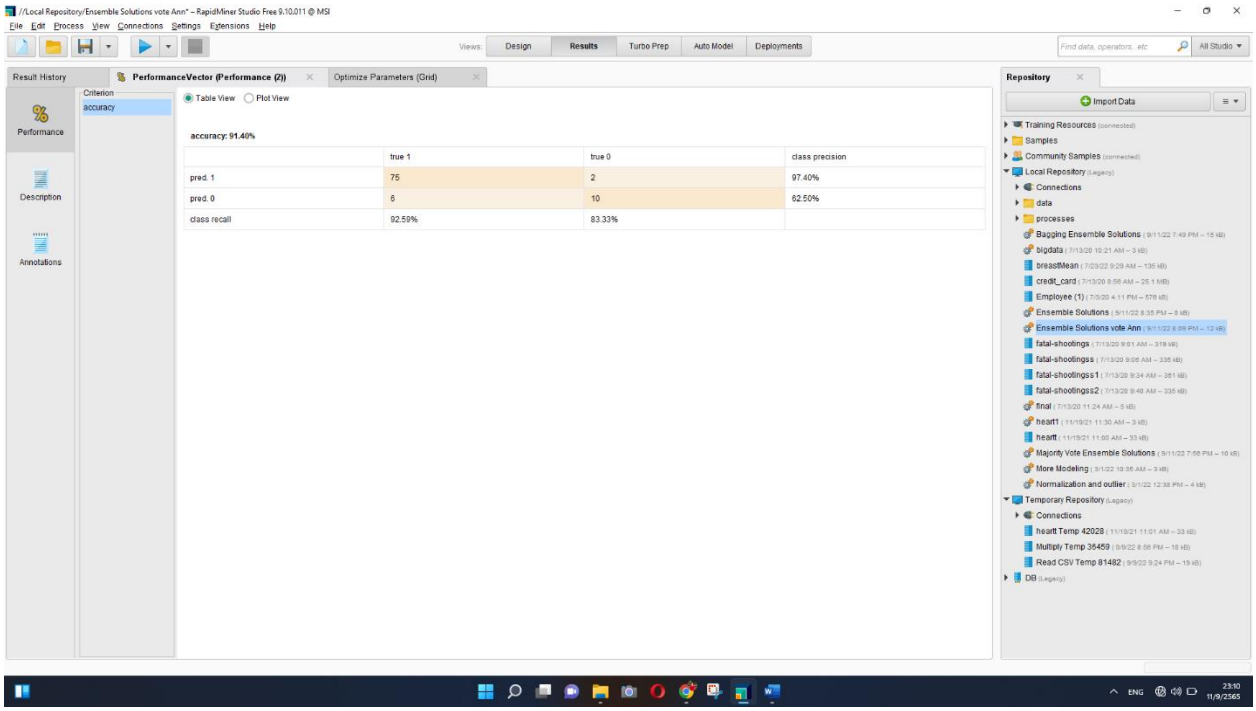
3.3.1 ทำความเข้าใจและศึกษาโรคมะเร็งปอด และศึกษาเทคนิคการจำแนกในการทำเหมืองข้อมูลเพื่อหาเทคนิคที่เหมาะสมในการจำแนกโรคมะเร็งปอด โดยศึกษาจากเอกสารและงานวิจัยที่เกี่ยวข้อง

3.3.2 ทำความเข้าใจข้อมูล ผู้วิจัยได้คัดเลือกข้อมูลจากชุดข้อมูล Kaggle ซึ่งเป็นชุดข้อมูลที่สำหรับโรคมะเร็งปอด มีข้อมูลจำนวน 310 เรคคอร์ด แอทริบิว 16 แอทริบิว คลาส 2 คลาส

3.3.3 การคัดเลือกข้อมูลผู้วิจัยได้ลบข้อมูลที่เป็นค่าว่างในชุดข้อมูลออกทั้งหมดและได้ทำการเปลี่ยนชื่อแอททริบิวจากเดิม Lung Cancer เป็น Class และแปลงข้อมูลจากตัวอักษรให้เป็นตัวเลขในแอททริบิว Class และ Gender

3.3.4 สร้างโมเดลโดยใช้โปรแกรม Rapid Miner และเลือกใช้เทคนิค 5 เทคนิค คือ Decision Tree, support vector machine (SVM), K-Nearest Neighbor (KNN), Artificial Neural Network (ANN) และ Ensemble majority vote

3.3.5 ประเมินผลและเปรียบเทียบประสิทธิภาพแบบจำลอง ค่าที่ใช้ในการวัดประสิทธิภาพคือค่าทางสถิติ ได้แก่ Accuracy, Precision และ Recall



ภาพที่ 3 การวิเคราะห์ตัวแบบ

3.4 เครื่องมือการวิจัย

3.4.1 โปรแกรม Rapid Miner ใช้ในการจำแนกข้อมูลรูปแบบการบุกรุกบนระบบเครือข่าย

3.4.2 เทคนิคที่ใช้ในการจำแนกข้อมูล 5 เทคนิค ได้แก่ เทคนิคการจำแนกข้อมูล Decision Tree เทคนิคการจำแนกข้อมูล support vector machine (SVM) เทคนิคการจำแนกข้อมูล K-Nearest Neighbor (KNN) และเทคนิคการจำแนกข้อมูล Artificial Neural Network (ANN) และ เทคนิคกลุ่มก้อนข้อมูล Ensemble majority vote

4. ทฤษฎีและงานวิจัยที่เกี่ยวข้อง

4.1 ทฤษฎีที่เกี่ยวข้อง

4.1.1 การจำแนกประเภทข้อมูล (Data Classification) หมายถึง เป็นการเรียนรู้แบบมีผู้สอน (Supervised Learning)โดยเป็นการสร้างโมเดลจำแนกประเภทเพื่อทำนายกลุ่มของข้อมูลใหม่การสร้างโมเดลจำแนกประเภทข้อมูลเกิดจากการหาความสัมพันธ์ของข้อมูลโดยข้อมูลทั้งหมดจะแบ่งเป็น 2 ส่วน คือ ส่วนแรกข้อมูลเรียนรู้ (Training Data) เป็นชุดข้อมูลที่ใช้ในการสร้างโมเดลจำแนกประเภทข้อมูลขึ้นมาใหม่ เพื่อให้โมเดลที่สร้างได้เรียนรู้ข้อมูลและส่วนทีสองของข้อมูลที่ใช้ในการทดสอบโมเดลที่สร้างขึ้นมา (Testing Data) เป็นชุดข้อมูลประเมินความถูกต้องของโมเดลจำแนกประเภทข้อมูลเทคนิคพื้นฐานที่นิยมนำมาใช้สำหรับการจำแนกประเภทข้อมูล

4.1.2 วิธีซัพพอร์ตเวกเตอร์แมทชีน (Support Vector Machaine : SVM) เป็นเทคนิคหนึ่งที่ได้รับความนิยมอย่างแพร่หลายในงานที่เกี่ยวข้องกับการจดจำรูปแบบตลอดจนการแก้ปัญหาการจัดกลุ่ม(classification problem) โดยอาศัยหลักการของการหาสัมประสิทธิ์ของสมการเพื่อสร้างเส้นแบ่งแยกกลุ่มข้อมูลที่ถูกป้อนเข้าสู่กระบวนการสอนให้ระบบเรียนรู้โดยเน้นไปยังเส้นแบ่งแยกกลุ่มข้อมูลได้ดีที่สุด (optimal separating hyperplane) เมื่อเราพิจารณาข้อมูลที่ประกอบด้วยข้อมูล 2 กลุ่ม

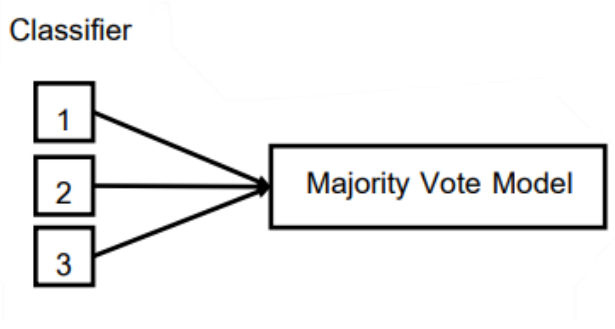
4.1.3 วิธีต้นไม้ตัดสินใจ (Decision Tree) เป็นเทคนิคที่ค่อนข้างแพร่หลายเนื่องจากการตัดสินใจเป็นแบบโครงสร้างต้นไม้ หรืออัลกอริทึม Decision Tree C4.5 ซึ่งเป็นส่วนขยายเพิ่มเติมมาจากอัลกอริทึม ID3 ที่ใช้หลักการของ Information

Entropy โดย Decision Tree C4.5 ใช้ความถูกต้องของ คุณลักษณะของข้อมูลเพื่อใช้เป็นการตัดสินใจแบ่งข้อมูลไปยังกลุ่มย่อย Decision Tree C4.5 จะพิจารณาตรวจสอบ Normalized Information Gain (ความแตกต่างใน Entropy) ผลลัพธ์จากการเลือกคุณลักษณะ Normalized Information Gain ที่สูงสุดนั้นคือการสร้างการตัดสินใจ ส่วน NBTree เป็นเทคนิคการผสมผสานระหว่าง Naïve Bayes และ Decision Tree โดยที่แต่ละโหนดจะใช้เทคนิคการเรียนรู้แบบต้นไม้ตัดสินใจในการสร้างโหนด และสำหรับโหนดใบจะใช้เทคนิคการเรียนรู้แบบ Naïve Bayes

4.1.4 เคเนียเรสเนเบอร์ **(**K-Nearest Neighbor) หลักการของวิธีการนี้จะจำแนกประเภทข้อมูลโดยขึ้นกับข้อมูลที่มีคุณสมบัติใกล้เคียงที่สุด K ตัวจากข้อมลบนชุดข้อมลตัวอย่างทำงานโดยขึ้นกับระยะทางน้อยสุดจากสมาชิกใหม่ หรือข้อมูลที่ป้อนถาม (input query instance) กับข้อมูลตัวอย่างฝึกฝน จะคำนวณหาเพื่อนบ้านที่ใกล้ที่สุด K ตัว หลงจากนั้นเราจะรวบรวมสมาชิกที่ใกล้เคียงที่สุด K ตัวแล้วเลือกคลาสที่สมาชิกส่วนใหญ่ที่ในกลุ่ม K ดังกล่าวสังกัดอยู่มากที่สุดให้กับสมาชิกใหม่ ข้อมูลการจำแนกโดยใช้ข้อมูลข้างเคียง K ตัว ประกอบด้วยแอททริบิวต์หลายตัวแปร Xi ซึ่งจะนำมาใช้ในการแบงกลุ่ม Yi โดยระบุค่าตัวเลขจำนวนเต็มบวกให้กับ K ซึ่งค่านี้จะเป็นตัวบอกจำนวนของกรณี (case) ที่จะต้องค้นหาในการทำนายกรณีใหม่ อัลกอริทึมแบบ KNN ได้แก่ 1-NN , 2-NN , 3-NN , ... K-NN

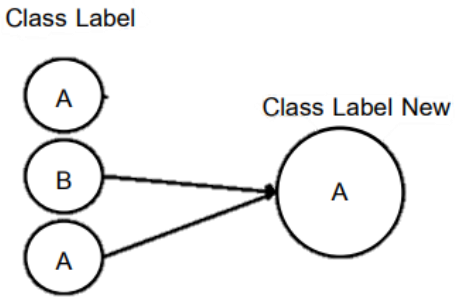
4.1.5 โครงข่ายประสาทเทียม **(Artificial Neural Networks)** เป็นศาสตร์แขนงหนึ่งทางด้านปัญญาประดิษฐ์ (Artificial Intelligence : AI) ที่สามารถนำไปประยุกต์ใช้กับงานหลายด้านได้อย่างมีประสิทธิภาพ หลักการสำคัญของโครงข่าย ประสาทเทียม คือ ความพยายามที่จะลอกเลียนแบบการทำงานของเซลล์ประสาทในสมองมนุษย์เพื่อทำงานได้อย่างมีประสิทธิภาพ ลักษณะทั่วไปของโครงข่ายประสาทเทียม คือ การที่โหนด (node) ต่าง ๆ จำลองมาจากไซแนป (synapse) ของเซลล์ประสาทระหว่าง เดนไดรท์ (dendrite) และแอกซอน (axon) โดยมีฟังก์ชันเป็นตัวกำหนด สัญญาณส่งออก (activation function or transfer function) นั่นเอง ลักษณะของโครงข่ายประสาทเทียม สามารถแบ่งได้ 2 แบบ คือ 1) โครงข่ายประสาทเทียมแบบ ชั้นเดียว (single layer) ซึ่งจะมีเพียงชั้นสัญญาณ ประสาทขาเข้า และชั้นสัญญาณประสาทขาออก เท่านั้น เช่น โครงข่ายเพอเซบตรอนอย่างง่าย (simple perceptron) และโครงข่ายโฮปฟิลด์ (hopfield networks) เป็นต้น และ 2) โครงข่ายประสาทเทียมแบบหลายชั้น (multilayer) ซึ่งมีลักษณะเช่นเดียวกับโครงข่ายประสาทเทียมแบบชั้นเดียว แต่จะมีชั้นแอบแฝง (hidden) เพิ่มขึ้น โดยอยู่ส่วนกลางระหว่างชั้นนำข้อมูลป้อนเข้าและชั้นส่งขอมูลออก ทั้งนี้ชั้นแอบแฝงอาจมีมากกว่า 1 ชั้น ในงานวิจัย นี้ได้ใช้ multilayer ในการสร้างแบบจำลอง

4.1.6 วิธีการกลุ่มก้อน Majority Vote Classifier เป็นการใช้ผลของการทำตัวจำแนกประเภท 3 วิธีมาทำการโหวต โดยถ้าผลของ คำตอบไหนเหมือนกันก็ให้ใช้คำตอบนั้นเป็นคำตอบที่ถูกต้องที่สุด



ภาพที่ 4 Diagram แสดง Majority Vote 1 2 3 Majority Vote Model

จากภาพที่ 2 จะเห็นได้ว่าเป็นภาพการทำงานของวิธี Majority Vote คือ การนำ ตัวจำแนกประเภทจำนวน 3 วิธีมาทำการโหวตหาค่าคำตอบที่ตอบมากที่สุดและเลือกคำตอบนั้น เพื่อมาสร้างโมเดล Majority Vote



ภาพที่ 5 แสดงวิธีการเลือกคำตอบ

4.2 งานวิจัยที่เกี่ยวข้อง

4.2.1 Supachai Prakongsilp [1] ได้นำเสนอการพัฒนาระบบสนับสนุนการตัดสินใจในการอนุมัติลูกบ้านเข้าโครงการ โดยใช้ เทคนิคต้นไม้ตัดสินใจและในอัลกอริธึม ID3 ในการเรียนรู้จากข้อมูลลูกบ้านในอดีตเพื่อสร้างโมเดลสำหรับตัดสินใจอนุมัติลูกบ้าน เข้าโครงการ โดยทำการทดสอบความแม่นยำของอัลกอริธึม ID3 เปรียบเทียบกับอัลกอริธึมอื่นที่เป็นเทคนิคต้นไม้ตัดสินใจ เหมือนกันด้วยการนำข้อมูลในอดีตของบุคคลที่เคยเขาร่วมโครงการทั้งหมด 963 รายการและ วัดความแม่นยำ ของอัลกอริธึม ID3 โดยใช้ Confusion Matrix ได้ผลความแม่นยา คือ 93.67 % ซึ่งมากกว่าอัลกอริธึมตัวอื่นๆ และผู้วิจัยได้วัด ประสิทธิภาพโมเดลของระบบที่พัฒนาขึ้น เปรียบเทียบกับแกรม Weka 3.5.8 โดยใช้ข้อมูลสำหรับการสอน (Training Data) ที่เหมือนกัน ผลที่ได้คือสามารถสร้างโมเดลได้เหมือนกับระบบที่ผู้วิจัยพัฒนา จึงสรุปได้ว่า โมเดลที่พัฒนาขึ้นมีความถูกต้องและ แม่นยำอยู่ในระดับดีมาก สามารถนำไปสนับสนุนงานได้จริง

4.2.2 Pichaya Promla and Charun Sanrach [2] ปัจจุบันสถาบันการศึกษาให้ความสำคัญ กับการประเมินผลการจดัการเรียนรู้ของครูผู้สอนเป็นอย่างมาก นอกจากการประเมินทางตรงยังมีการประเมินทางอ้อมด้วยการสำรวจความพึงพอใจของผู้เรียนโดยใช้แบบสอบถาม ซึ่ง การวิเคราะห์ข้อมูลคำถามปลายปิดในแบบสอบถามนั้นสามารถทำได้ง่ายแต่คำถามปลายเปิดจะทำได้ยาก และซับซ้อน รวมทั้งอาจเกิดความไม่แม่นยำเนื่องจากอคติจากผู้วิเคราะห์ข้อมูลได้ในงานวิจัยนี้ได้ใช้กระบวนการวิเคราะห์ความรู้สึกเพื่อวิเคราะห์ความคิดเห็นที่ได้จากแบบสอบถามจำนวน 1,577 ข้อความให้เป็นขั้วความพึงพอใจ และเปรียบเทียบ ประสิทธิภาพการจำแนกโดยใช้เทคนิคการรวมกลุ่มเพื่อจำแนกข้อมูล ได้แก่ Vote, Bagging และ Random Forest กับเทคนิควิธีมาตรฐาน ได้แก่Decision Tree, Naïve Bayes และ K-NN พบว่าเทคนิควิธีการรวมกลุ่มเพื่อจำแนกข้อมูลแบบ Vote มีประสิทธิภาพมากที่สุด

4.2.3 Ekkasit Patcharawongsakda [3] กล่าวว่า เทคนิคในการทำเหมืองข้อมูลเป็นขั้นตอนแรก ในการจำแนกประเภทข้อมูล โดยการนำข้อมูลเทรนนิ่งดาต้า (Training Data) หรือข้อมูล ที่ใช้ในการเรียนรู้มาสร้างเป็นตัวแบบขึ้นมาด้วยเทคนิคการจำแนกประเภทข้อมูลแบบต่าง ๆ เช่น Decision Tree, Naive Bayes, K-Nearest Neighbors และ Neural Network เป็นต้น

4.2.4 Thiptida Wongpipan [4] การใช้เหมืองข้อมูลช่วยในการตัดสินใจการใช้สินเชื่อ เป็นงานวิจัยเกี่ยวกับให้บริการรถเชาระยะยาว สัญญาเชาระยะเวลา 3-5 ปและระยะสั้น สัญญาเชาระยะเวลาเป็นรายวัน รายสัปดาห์และรายเดือน โดยใช้เทคนิคเหมืองข้อมูล 3 วิธีคือ ต้นไม้ตัดสินใจ วิธีแบบเบย์เพื่อการวิเคราะห์ข้อมูลลูกค้าที่มาเชารถ เพื่อชวยการตัดสินใจพิจารณาข้อมูลสินเชื่อของลูกค้าให้มีประสิทธิภาพมากยิ่งขึ้น และลดความเสี่ยงด้านหนี้สูญในการรออนุมัติให้เชารถยนต์ต่อ

ลูกค่าที่ไม่เหมาะสม ผลการเปรียบเทียบประสิทธิภาพ พบวาการจำแนกวิธีต่นไม้ตัดสินใจให่ผลลัพธ์ดีที่สุด มีคาความถูกต่อง (90.47)

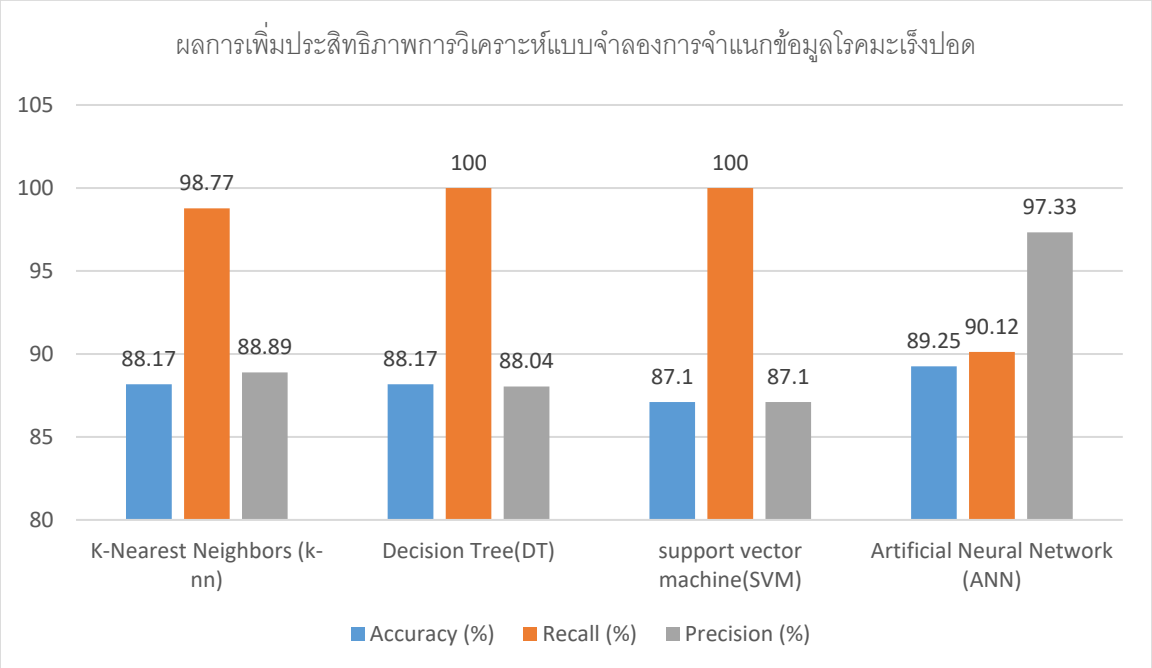
4.2.5 Kamalas Udomlamlert and Arnon Rungsawang [5] การคัดแยกหมวดหมู่เว็บเพจภาษาไทยอัตโนมัติ เป็นงานวิจัยเกี่ยวกับระบบแบบเว็บเพจออกเป็นหมวดหมู่ตามเนื้อหาที่ปรากฏ หรือที่เรียกว่า “เว็บไดเร็กทอรี่” (web directory) จากการทดลองกับข่อมูลทดสอบ 2 ชุด ซึ่งคัดเลือกมาจากเว็บไดเร็กทอรี่ภาษาไทย 5 หมวดหมู่หลักทำให้เห็นว่า การลดคุณลักษณะโดยวิธีการคัดเลือกคุณลักษณะที่เหมาะสมจะไม่ส่งผลต่อความถูกตองในการจำแนกหมวดหมู่และการลดคุณลักษณะ ประกอบกับการใช้ตัวคัดแยก Naïve Bayes จะให้ค่าประสิทธิผลดีที่สุด โดยมีค่าความถูกต้อง (83.4)

5. ผลการวิจัย

ตารางที่ 1 แสดงการเปรียบเทียบค่าทางสถิติของแบบจำลองทั้ง 4 เทคนิค

Techniques	Accuracy (%)	Recall (%)	Precision (%)
K-Nearest Neighbors (k-nn)	88.17	98.77	88.89
Decision Tree(DT)	88.17	100	88.04
support vector machine(SVM)	87.10	100	87.10
Artificial Neural Network (ANN)	89.25	90.12	97.33

จากตารางที่ 1 ผลการเพิ่มประสิทธิภาพการวิเคราะห์แบบจำลองการจำแนกข้อมูลโรคมะเร็งปอดของแต่ละ เทคนิค พบว่าแบบจำลองที่ใช้เทคนิค Artificial Neural Network (ANN) ให้ค่าความแม่นยำมากที่สุด จากนั้นจึงได้ทำการเพิ่มประสิทธิภาพโดยใช้โมเดล Ensemble majority vote (EMV)



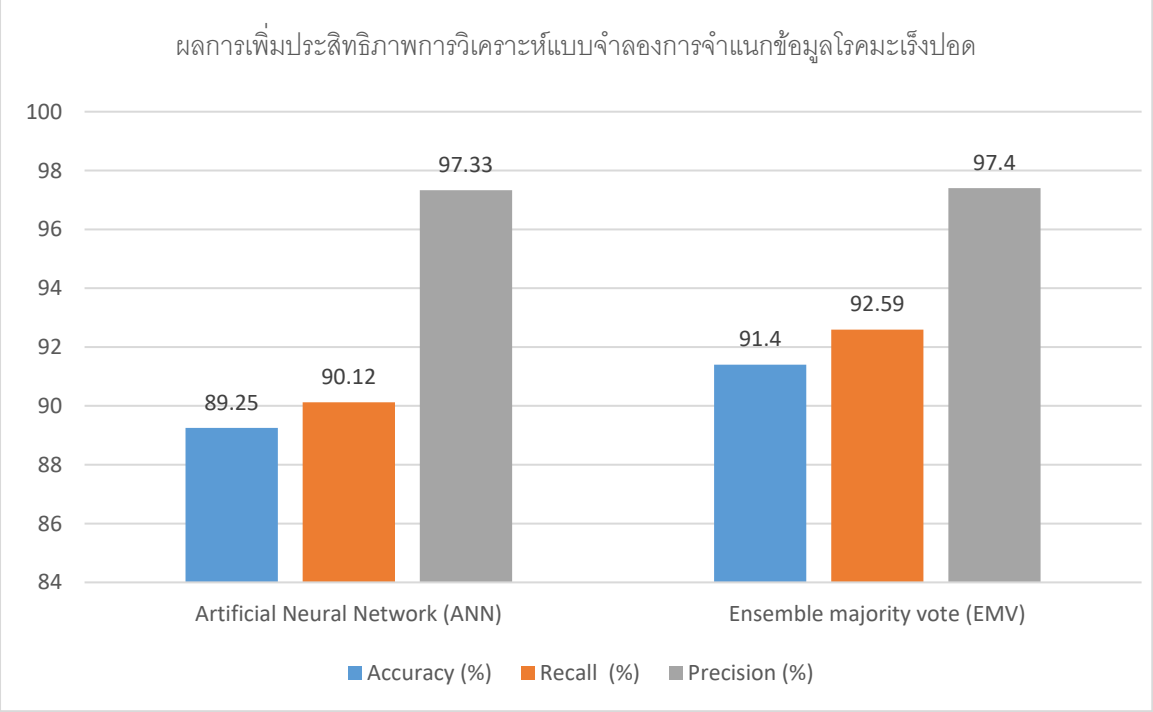
ภาพที่ 6 ผลการเพิ่มประสิทธิภาพการวิเคราะห์แบบจำลองการจำแนกข้อมูลโรคมะเร็งปอดของแต่ละเทคนิค

พบว่าเทคนิค K-Nearest Neighbors (k-nn) ให้ค่า Accuracy 88.17%, Recall 98.77 %, Precision 88.89%, เทคนิค Decision Tree (DT) ให้ค่า Accuracy 88.17 %, Recall 100 %, Precision 88.0.4 %, เทคนิค support vector machine (SVM) ให้ค่า Accuracy 87.10 %, Recall 100 %, Precision 87.10 %, เทคนิค Artificial Neural Network (ANN) ให้ค่า Accuracy 88.17 %, Recall 90.12 %, Precision 97.33 % พบว่าแบบจำลองที่ใช้เทคนิค Artificial Neural Network (ANN) ให้ค่าความ Accuracy ที่สุด จากนั้นจึงได้ทำการเพิ่มประสิทธิภาพโดยใช้โมเดล Ensemble majority vote (EMV)

ตารางที่ 2 แสดงการเพิ่มประสิทธิภาพ

Techniques	Accuracy (%)	Recall (%)	Precision (%)
Artificial Neural Network (ANN)	89.25	90.12	97.33
Ensemble majority vote (EMV)	91.40	92.59	97.40

จากตารางที่ 2 ผลการเพิ่มประสิทธิภาพการวิเคราะห์แบบจำลองการจำแนกข้อมูลโรคมะเร็งปอดของแต่ละ เทคนิค พบว่าแบบจำลองที่ใช้เทคนิค Ensemble majority vote (EMV) สามารถเพิ่มประสิทธิภาพได้จริง



ภาพที่ 7 ผลการเพิ่มประสิทธิภาพการวิเคราะห์แบบจำลองการจำแนกข้อมูลโรคมะเร็งปอด

จากตารางที่ 2 ผลการเพิ่มประสิทธิภาพการวิเคราะห์แบบจำลองการจำแนกข้อมูลโรคมะเร็งปอดของแต่ละ เทคนิค พบว่าแบบจำลองที่ใช้เทคนิค Ensemble majority vote (EMV) สามารถเพิ่มประสิทธิภาพได้จริง

6. สรุปผล

งานวิจัยฉบับนี้มีวัตถุประสงค์เพื่อเปรียบเทียบและเพิ่มประสิทธิภาพในการจำแนกข้อมูลโรคมะเร็งปอดด้วยเทคนิคเหมืองข้อมูล จากนั้นได้นำเอาเทคนิควิธีการวิเคราะห์เหมืองข้อมูลมาทำการวิเคราะห์ Class จำนวน 5 เทคนิคด้วยกัน คือ เทคนิค K-Nearest Neighbors (k-nn), เทคนิค Decision Tree, เทคนิค support vector machine (SVM), เทคนิค Artificial Neural Network (ANN), เทคนิค Ensemble majority vote สำหรับขั้นตอนวิธีการวัดประสิทธิภาพของทั้ง 5 โมเดล ได้ใช้หลักการ 10- Cross-validation Folds ในการแบ่งกลุ่ม ข้อมูลเป็นชุดข้อมูลเรียนรู้และชุดข้อมูลทดสอบ และวัดประสิทธิภาพของแบบจำลองด้วยค่า Accuracy , Precision และ Recall ผลการวิจัยพบว่าเทคนิค Artificial Neural Network (ANN) ที่สุด จากนั้นนำมาเพิ่มประสิทธิภาพด้วย เทคนิค Ensemble majority vote (EMV) ให้ค่าประสิทธิภาพของ Accuracy 91.40 %, recall 92.59 %, precision 97.40 % ดังนั้นจึงสรุปได้ว่า เทคนิค Ensemble majority vote (EMV) สามารถเพิ่มประสิทธิภาพของการจำแนกข้อมูลได้จริง

7. ข้อเสนอแนะ

การวิจัยครั้งนี้เป็นการวิจัยที่ใช้โปรแกรม RapidMiner เพื่อนำมาประยุกต์ใช้ในการเปรียบเทียบประสิทธิภาพ โดยที่มีเครื่องมือต่าง ๆ ให้เลือกใช้อาทิเช่น Decision Tree, support vector machine(SVM) Artificial Neural Network (ANN) และ Ensemble majority vote โดยสามารถนำเครื่องมือเหล่านี้มาเปรียบเทียบประสิทธิภาพเพื่อหาค่าความถูกต้องมากที่สุด

8. เอกสารอ้างอิง

[1] Supachai Prakongsilp. (2008). Design and Development of Decision Support Systems for Approving Residents to Join the Project Using Decision Tree Techniques: A Case Study of the Habitat for Humanity Foundation. Special Issue Master of Science in Information Technology, Faculty of Information Technology, Graduate School, King Mongkut's Institute of Technology North Bangkok. (In Thai)

[2] Pichaya Promla and Charun Sanrach. (2020). The Comparison of Efficiency on The Analysis of Satisfaction on Teaching Performance using Sentiment Analysis by Ensemble Technique. KKU Research Journal (Graduate Studies). Vol. 20, No. 4, P. 140-149.(In Thai)

[3] Ekkasit Patcharawongsakda (2014). An Introduction to Data Mining Techniques. Bangkok: Asia Digital Printing Co., Ltd. (In Thai)

[4] Thiptida Wongpipan. (2013). Using Data Mining to Help Make Credit Decisions: A Case Study: Krungthai Car Rent and Lease Public Company Limited. Dhurakij Pundit University/Bangkok. (In Thai)

[5] Kamalas Udomlamlert and Arnon Rungsawang (2010). Automatic Thai web page categorization using Naive Bayes. Proceedings of 48th Kasetsart University Annual Conference: Architecture and Engineering. P.310-317.