การเปรียบเทียบประสิทธิภาพเทคนิคกลุ่มก้อนการจำแนกข้อมูล โดยใช้ชุดข้อมูลผู้ป่วยมะเร็งเต้านม

ธาดา ลิมกุลาคมน์[1*], ไก้รุ่ง เฮงพระพรหม[1] และ สุพจน์ เฮงพระพรหม[1]

[1]สาขาวิชาวิทยาการข้อมูล คณะวิทยาศาสตร์และเทคโนโลยี มหาวิทยาลัยราชภัฏนครปฐม
*614285003@webmail.npru.ac.th

บทคัดย่อ

วิจัยนี้มีวัตถุประสงค์ 1) เพื่อศึกษาเทคนิคกลุ่มก้อนข้อมูลแบบการโหวตเสียงส่วนใหญ่และบูตสตรัปแอกเกรเกตติ้ง 2) เปรียบเทียบประสิทธิภาพเทคนิคกลุ่มก้อนการจำแนกข้อมูลโดยนำเทคนิคเหมืองข้อมูล ขั้นตอนวิธีการเพื่อนบ้านใกล้ที่สุด โครงข่ายประสาทเทียม ต้นไม้การตัดสินใจ ซัพพอร์ตเวกเตอร์แมชชีน ด้วยการโหวตเสียงส่วนใหญ่และบูตสตรัปแอกเกรเกตติ้ง มาเปรียบเทียบประสิทธิภาพของแต่ละเทคนิคข้างต้นกับชุดข้อมูลผู้ป่วยมะเร็งเต้านม ผลจากการศึกษาพบว่าวิธีที่ให้ประสิทธิภาพที่ดีที่สุด คือ การโหวตเสียงส่วนใหญ่และบูตสตรัปแอกเกรเกตติ้งของโครงข่ายประสาทเทียมโดยให้ค่าความถูกต้องในการจำแนกเท่ากับ 97.66 ค่าระลึก 93.75 ค่าความแม่นยำเท่ากับ 96.4 รองลงมาคือ โครงข่ายประสาทเทียมให้ค่าความถูกต้องเท่ากับ 97.08 ค่าระลึก 93.75 ค่าความแม่นยำเท่ากับ 98.36

คำสำคัญ: เทคนิคกลุ่มก้อนการจำแนกข้อมูล ขั้นตอนวิธีการเพื่อนบ้านใกล้ที่สุด โครงข่ายประสาทเทียม ต้นไม้การตัดสินใจ ซัพพอร์ตเวกเตอร์แมชชีน แบบการโหวตเสียงส่วนใหญ่ บูตสตรัปแอกเกรเกตติ้ง

The Efficiency Comparison of Data Classification Ensemble Techniques for Breast Cancer Patients Dataset

Tada Limkulakhom[1*], Kairung Hengpraprohm[1], and Supojn Hengpraprohm[1]

[1] Program in data science, Faculty of Science and Technology, Nakhon Pathom Rajabhat University
*614285003@webmail.npru.ac.th

Abstract

The objectives of this research are to 1) study the majority vote ensemble and bootstrap aggregating techniques, and 2) compare the efficiency of the ensemble data classification using 4 data mining techniques including k-nearest neighbor, artificial neural network, decision tree, and support vector machine: with the majority vote ensemble and bootstrap aggregating techniques for classification of breast cancer patient data. The results of the study show that the method that gives the best performance is majority vote and bootstrap aggregating of the artificial neural network by giving a classification accuracy of 97.66, a recall of 93.75 and a precision of 96.4. Followed by the artificial neural network give an accuracy of 97.08, a recall of 93.75 and a precision of 98.36.

Keywords: ensemble classification, K-Nearest Neighbor, Neural Network, Decision tree, Support Vector Machine, Majority vote,Bagging

1. บทนำ

ปัจจุบัน นักวิจัยจำนวนมากได้ให้ความสำคัญกับปัญหาที่ต้องการวิธีการจำแนกข้อมูลที่แม่นยำ โดยการจำแนกข้อมูลนั้นมีด้วยกันหลายวิธี เช่น ขั้นตอนวิธีการเพื่อนบ้านใกล้ที่สุด (K-Nearest Neighbor), โครงข่ายประสาทเทียม (Neural Network), ต้นไม้การตัดสินใจ (decision tree), ซัพพอร์ตเวกเตอร์แมชชีน (Support Vector Machine : SVM) เป็นต้น

แต่ถึงแม้ว่าเทคนิคข้างต้นจะให้ผลลัพธ์ในการจำแนกข้อมูลที่แม่นยำแต่โมเดลเดียว (Single Model) นั้นทำให้มีการกำหนดกลุ่มของข้อมูลที่ใช้ในการเรียนรู้ รวมทั้งมีการกำหนดค่าพารามิเตอร์ที่ตายตัว บางครั้งเกิด ปัญหาความโน้มเอียง (Bias) ทำให้ได้ประสิทธิภาพที่ ไม่ดีนัก หนทางหนึ่งที่จะสามารถลดค่าความโน้มเอียงได้ คือการใช้วิธีการร่วมกันตัดสินใจ (Ensemble) ซึ่งสามารถ สร้างความหลากหลายและลดค่าความผิดพลาดที่เกิดจาก ความแปรปรวนได้ แนวความคิดหลักของวิธีการร่วมกัน ตัดสินใจคือการรวมเอากลุ่มของตัวจำแนกข้อมูล เพื่อแก้ ปัญหาเดียวกันและผลลัพธ์ที่ได้จะมีความแม่นยำและความเที่ยงตรงมากกว่าการใช้โมเดลแบบเดี่ยว หากแต่วิธีการรวมกลุ่มถ้าจะให้มีประสิทธิภาพที่ดีได้นั้น โดยทั่วไปแล้วประสิทธิภาพของการรวมกลุ่มจะขึ้นอยู่กับความหลากหลาย และความแม่นยำของตัวแทนในการจำแนกข้อมูล

ดังนั้นสำหรับงานวิจัยในครั้งนี้ ผู้วิจัยได้ให้ความสำคัญในการใช้ตัวเทคนิคกลุ่มก้อนการจำแนกข้อมูล เพื่อมาทำการเปรียบเทียบหาวิธีที่ให้ค่าแม่นยำสูงที่สุด โดยใช้เกณฑ์การประเมินด้วยค่า Accuracy, Precision และ Recall เป็นเกณฑ์ในการเปรียบเทียบหาวิธีที่ให้ค่าแม่นยำสูงที่สุด สำหรับนำไปสร้างโมเดลใหม่ที่มีประสิทธิภาพด้านการจำแนกข้อมูล ได้ถูกต้องมากขึ้น

2.วัตถุประสงค์

2.1 เพื่อศึกษาเทคนิคกลุ่มก้อนการจำแนกข้อมูล

2.2 เพื่อเปรียบเทียบประสิทธิภาพเทคนิคกลุ่มก้อนการจำแนกข้อมูล

3.ทฤษฎีและงานวิจัยที่เกี่ยวข้อง

3.1.ทฤษฎีที่เกี่ยวข้อง

ขั้นตอนวิธีการเพื่อนบ้านใกล้ที่สุด (K-Nearest Neighbor : KNN) เป็นอัลกอริทึมในการจัดกลุ่มข้อมูล (Classification) โดยพิจารณาจากลักษณะเด่นของข้อมูล (Feature) โดยจะใช้วิธีการหาระยะห่าง (Distance) เพื่อหาข้อมูลที่ใกล้เคียงกันมากที่สุด โดยค่า K จะระบุจำนวนของเพื่อนบ้านที่ใกล้ที่สุดที่จะนำมาใช้ในการจัดกลุ่มข้อมูล.

โครงข่ายประสาทเทียม (Neural Network) เป็นโมเดลการเรียนรู้ของเครื่องจักรที่จำลองการทำงานของระบบประสาทเทียมในสมองของมนุษย์ โครงข่ายประสาทเทียมประกอบด้วยโนด (node) หรือเรียกว่าเซลล์ประสาท (neuron) ที่เชื่อมต่อกันเป็นระบบเพื่อประมวลผลข้อมูล โดยโนดแต่ละตัวมีหน้าที่คำนวณและส่งผลลัพธ์ต่อไปยังโนดต่อไปในโครงข่าย

ต้นไม้การตัดสินใจ (decision tree) เป็นเทคนิคการเรียนรู้ของเครื่องจักรที่นิยมใช้ในงานที่เกี่ยวข้องกับการจัดกลุ่ม (classification) และการทำนาย (prediction) โดยที่ผลลัพธ์ที่ได้จากการเรียนรู้นั้นจะเป็นต้นไม้แบบชั้นเปิด (open-layer) ที่แบ่งกลุ่มข้อมูลออกเป็นกลุ่มย่อยๆ โดยอิงตามลักษณะของข้อมูลแต่ละคุณลักษณะ (feature) ที่ใช้ในการเรียนรู้

ซัพพอร์ตเวกเตอร์แมชชีน (Support Vector Machine: SVM) เป็นอัลกอริทึมการเรียนรู้ของเครื่องจักรที่ใช้ในการจำแนกและประมวลผลข้อมูล โดยเฉพาะอย่างยิ่งในงานที่เกี่ยวข้องกับการจำแนกข้อมูล ตัวแบบของ SVM จะเรียนรู้โดยการใช้ตัวจำแนกหรือเรียกว่า hyperplane เพื่อแบ่งกลุ่มข้อมูลที่มีลักษณะเดียวกันหรือคล้ายคลึงกันในชุดข้อมูล โดย hyperplane ที่เลือกจะมีระยะห่าง (margin) จากกลุ่มข้อมูลทั้งสองฝั่งที่ใกล้ที่สุดมากที่สุดเท่าที่เป็นไปได้

เทคนิคกลุ่มก้อนการจำแนกข้อมูล (ensemble classification) เป็นเทคนิคหนึ่งในการประมวลผลข้อมูลด้วยการผสมผสานผลลัพธ์จากหลายๆ โมเดลการจำแนกข้อมูลเข้าด้วยกัน เพื่อปรับปรุงประสิทธิภาพในการจำแนกข้อมูล การเลือกใช้เทคนิคกลุ่มก้อนการจำแนกข้อมูลทำให้มีความสามารถในการลดความผิดพลาดและเพิ่มความแม่นยำในการจำแนกข้อมูลที่มีความซับซ้อน

Bagging (Bootstrap Aggregating) เป็นเทคนิคในการสร้างโมเดล Machine Learning ที่มีวัตถุประสงค์เพื่อลดความเสี่ยงของโมเดลที่เกิดจาก overfitting หรือความเอาใจใส่กับข้อมูลส่วนตัว (overfitting) โดยใช้เทคนิคการสุ่มข้อมูลและสร้างโมเดลหลายๆ โมเดล โดยใช้ข้อมูลที่เหมือนกันแต่มีการสุ่มให้เป็นชุดข้อมูลย่อยๆ (bootstrap sample) เพื่อลดความผิดพลาดที่อาจเกิดจากความสัมพันธ์ที่แตกต่างกันของชุดข้อมูลแต่ละชุด

Majority vote เป็นเทคนิคในการจัดกลุ่มของโมเดลที่ใช้ในกระบวนการทำนาย (prediction) โดยใช้ผลลัพธ์ที่ได้จากการทำนายของโมเดลหลายๆ โมเดล เช่น โมเดลต่างๆ ที่ถูกสร้างขึ้นด้วยเทคนิคต่างๆ เช่น Decision Tree, Support Vector Machine, Random Forest, Neural Network เป็นต้น โดยการใช้ Majority vote เราจะเลือกผลลัพธ์ที่ได้จากโมเดลที่มีการทำนายถูกต้องสูงสุดเป็นผลลัพธ์สุดท้าย

3.2.งานวิจัยที่เกี่ยวข้อง

Vatinee Nuipian and Phayung Meesad [1] ปัญหาหนึ่งของการทำเหมืองข้อความคือข้อมูลมีปริมาณมาก นักวิจัยจำนวนมากใช้เทคนิคการคัดเลือก คุณลักษณะเพื่อได้คำทีเหมาะสมในการแทนเอกสารและเพิ่มประสิทธิภาพในการจำแนกเอกสารให้มีค่าความถูกต้องมากขึ้น เทคนิคที่ใช้แบ่งเป็น 2 วิธีได้แก่ การกรองและการควบรวม โดยเทคนิคการควบรวมสามารถใช้เทคนิคการทำเหมืองข้อความร่วมกับการค้นหาข้อมูล ในงานวิจัยนี้ได้ทำการเปรียบเทียบการคัดเลือกคุณลักษณะแบบการกรอง โดย เลือกใช้อินฟอร์เมชันเกน เกนเรโช และไคสแควร์ วิธีคัดเลือกแบบไคสแควร์ให้ผลดีที่สุดวัดประสิทธิภาพโดยรวม 92.2% และ การควบรวมใช้เทคนิคซัพพอร์ตเวกเตอร์แมชชีน (SVM) ร่วมกับการค้นหาด้วยวิธีเชิงพันธุกรรม

(SVMGA) และการค้นหาด้วยวิธีละโมบ (SVMGD) โดยวิธีคัดเลือกแบบ SVMGD ให้ผลดีที่สุดวัดประสิทธิภาพ โดยรวม 94% ซึ่งการจำแนกข้อความทั้งสองวิธีใช้ขันตอนวิธีแบบซัพพอร์ตเวกเตอร์แมชชีนโดยใช้เคอร์เนลแบบ เรเดียลเบสิสฟังก์ชัน (SVMR) เมือเปรียบเทียบประสิทธิภาพทังวิธีการกรองและการควบรวมสรุปได้ว่าประสิทธิภาพ โดยรวมของการควบรวมมีค่ามากกว่าการกรอง 1.8% ซึ่งทำให้นักวิจัยสามารถนำเทคนิคของการควบรวมไปใช้เพิ ม ประสิทธิภาพการจำแนกข้อความ

Patharawut Saengsiri, Sageemas Na Wichian and Phayung Meesad [2] การค้นหากลุ่มย่อยของยีนที่มีอำนาจจำแนก เป็นปัญหาที่สำคัญสำหรับงานวิจัยทางด้านชีววิทยา เนื่องจากมีจำนวนยีนเพิ่มขึ้นเป็นจำนวนมาก ดังนั้นเทคนิคการลดมิติของข้อมูล จึงเป็นประโยชน์ในการช่วยค้นหากลุ่มย่อยของยีน สาเหตุเนื่องจากเมื่อข้อมูลมีจำนวนมิติหรือตัวแปรมาก ทำให้ข้อมูลเกิดการกระจาย (data sparse) และทำให้เกิดปัญหามิติข้อมูล (Curse of Dimensionality) งานวิจัยนี้จะนำเอาข้อมูลยีนของโรคมะเร็งเม็ดเลือดขาวแบบเฉียบพลัน (Acute Leukemia) ซึ่งมีจำนวนมิติของข้อมูล 7,129 มิติ แบ่งออกเป็น 2 กลุ่ม คือ ALL และ AML มาทำการทดลองและเพื่อเปรียบเทียบประสิทธิภาพของการลดมิติข้อมูล ระหว่างวิธี Correlation Based Feature Selection, Gain Ratio และ Information Gain โดยนำผลลัพธ์ที่ได้จากการลดมิติ มาเป็นข้อมูลอินพุตของซัพพอร์ตเวกเตอร์แมชชีน (Support Vector Machine) เพื่อคัดแยกประเภทของโรคมะเร็ง ซึ่งผลการทดลองแสดงให้เห็นว่าการลดข้อมูลโดยวิธี Gain Ratio และ Information Gain มีความเหมาะสม คือ สามารถลดมิติของข้อมูลเหลือ 36 มิติ และเพิ่มความแม่นยำจากเดิม 73.53% เป็น 88.24%

Anutchai Chutipascharoen and Charun Sanrach [3] การเปรียบเทียบประสิทธิภาพของอลักอริทึมในการทำนายและคุณลักษณะ ที่มีต่อโอกาสความสำเร็จในการโอนเงินข้ามประเทศของบุคคลทั่วไป โดยทำการศึกษาข้อมูลการโอนเงินของบุคคลทั่วไปจำนวน 51,901 ระเบียน ทา การเก็บข้อมูลตั้งแต่ปี 2559 -2560 โดยใช้เทคนิคการจำแนกข้อมูลทั้งหมด 3 เทคนิคได้แก่ เทคนิคต้นไม้ตัดสินใจ เทคนิคนาอีฟเบย์ และเทคนิคการค้นหาเพื่อนบ้านใกล้สุด ซึ่งทำการเปรียบเทียบประสิทธิภาพ รูปแบบเทคนิคการทา นายระหว่างการใช้คุณลักษณะทั้งหมด การทดสอบประสิทธิภาพตัวแบบทา นายด้วยวิธีการ Cross Validation โดยใช้โปรแกรม RapidMiner Studio 8 จากนั้นทำการทดลองเพื่อหาผลการทดสอบประสิทธิภาพที่มีค่าความถูกต้องที่สูงที่สุด ผลการศึกษาพบว่าการใช้เทคนิคต้นไม้ตัดสินใจด้วยการเลือกคุณลักษณะทั้งหมด มีค่าความถูกต้องเท่ากับ 99.90% เทคนิคการค้นหาเพื่อนบ้านใกล้สุดมีค่าความถูกต้องเท่ากับ 99.55% และเทคนิคนาอีฟเบย์มีค่า ความถูกต้องเท่ากับ 96.71% จากผลการเปรียบเทียบประสิทธิภาพในครั้งนี้สามารถนา เทคนิคต้นไม้ตัดสินใจ ที่มีค่าความถูกต้องสูงสุดไปใช้ในการพยากรณ์โอกาสความสำเร็จในการโอนเงินข้ามประเทศของบุคคลทั่วไปต่อไป

Jiraporn Jareanying [4] การทำเหมืองข้อมูลทางการศึกษา โดยการจำแนกประเภทข้อมูล การวิเคราะห์องค์ประกอบหลักของข้อมูล เพื่อหาความสัมพันธ์ของตัวแปร และเปรียบเทียบประสิทธิภาพของอัลกอริทึม ซึ่งเป็นการศึกษาด้วยเทคนิค ต้นไม้ตัดสินใจ เทคนิคป่าแห่งการทำนาย การเรียนรู้เบย์ และ K-NN โดยใช้ข้อมูลของนักเรียนระดับมัธยมศึกษาในโรงเรียนโปรตุเกส ประกอบด้วยข้อมูลด้านผลการเรียน ด้านความเป็นอยู่ และความเชื่อมโยงทางสังคมและโรงเรียน จาก UCI Machine Learning Repository มีข้อมูล 649 รายการ 31 แอตทริบิวต์ จากผลการวิจัย พบว่าเทคนิคที่ให้ค่าความถูกต้อง มากที่สุด คือเทคนิคป่าไม้ตัดสินใจ เท่ากับ 80.74 และเทคนิคที่ มีค่าความถ่วงดุลมากที่สุด คือ เทคนิคการเรียนรู้เบย์ เท่ากับ 55.52

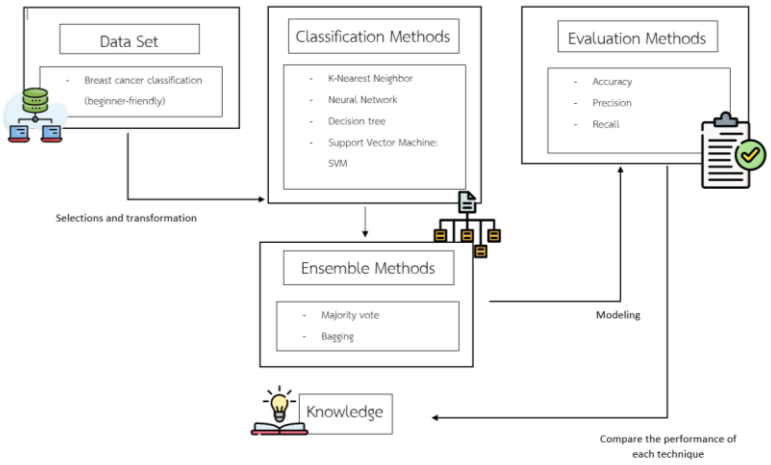
Ratiporn Chanklan [5] การศึกษาปัญหาการจำแนกข้อมูลไบโอเมตริกซ์ ซึ่งในการจำแนกข้อมูลไบโอเมตริกซ์นั้นนิยมใช้อัลกอริทึมซัพพอร์ตเวกเตอร์แมชชีน แต่การรู้จำเป็นต้องใช้เวลานาน จึงได้ใช้การลดมิติข้อมูลเข้ามาช่วยเพื่อลดเวลาการทำงาน การลดมิติข้อมูลเป็นขั้นตอนการเตรียมข้อมูลก่อนนำไปเข้า อัลกอริทึมที่ใช้ในการจำแนก ในอดีตได้มีหลายงานวิจัยที่ได้เสนอเทคนิคการจำแนกข้อมูลไบโอเมตริกซ์โดยใช้เทคนิคการลดมิติข้อมูลและส่วนใหญ่จะใช้กับข้อมูลภาพใบหน้า ซึ่งจะวัดประสิทธิภาพของโมเดลโดยการเปรียบเทียบค่าความแม่นตรงเพียงอย่างเดียว การลดมิติข้อมูลที่สามารถใช้กับข้อมูลไบโอเมตริกซ์หลายชนิดและนำเวลาที่ใช้ในการจำแนกมาเปรียบเทียบจะมี ความซับซ้อน จึงทำให้มีงานวิจัยด้านนี้ปรากฏค่อนข้างน้อย ผู้วิจัยได้เห็นความสำคัญในจุดนี้จึงได้ เสนอการปรับปรุงอัลกอริทึมซัพพอร์ตเวกเตอร์แมชชีนสำหรับการจำแนกข้อมูล

ไบโอเมตริกซ์ที่ เรียกว่า Bio-SVM เพื่อเพิ่มประสิทธิภาพและลดเวลาในการจำแนกข้อมูล โดยใช้เทคนิคการลดมิติ ข้อมูลได้แก่ การวิเคราะห์จำแนกประเภทเชิงเส้น (LDA) ร่วมกับอัลกอริทึมซัพพอร์ตเวกเตอร์แมชชีนเคอร์เนลเส้นตรงสำหรับการจำแนกข้อมูลภาพไบโอเมตริกซ์เชิงกายภาพ และใช้อัลกอริทึมซัพพอร์ตเวกเตอร์แมชชีนเคอร์เนลโพลิโนเมียลในการจำ แนกข้อมูลภาพไบโอเมตริกซ์เชิงพฤติกรรม โดยใช้ภาษาไพธอนในการทดลองงานวิจัยนี้ใช้ค่าความแม่นตรงและเวลาที่ใช้ในการจำแนกข้อมูลมาใช้ในการประเมินประสิทธิภาพการจำแนกข้อมูลไบโอเมตริกซ์

Orawan Voraaroon et al., [6] การศึกษาปัจจัยที่มีความสัมพันธ์กับการควบคุมตนเองในพฤติกรรมการสูบบุหรี่ในนักเรียนมัธยมศึกษาตอนต้น จังหวัดสุพรรณบุรี กลุ่มตัวอย่างคือ นักเรียนมัธยมศึกษา ที่ศึกษาอยู่ในชั้นมัธยมศึกษาปีที่ 1-3 ทั้งเพศชายและเพศหญิง ภาคเรียนที่ 1 ปีการศึกษา 2559 ในโรงเรียนเขตพื้นที่การศึกษามัธยมศึกษาเขต 9 จังหวัดสุพรรณบุรี จำนวน 404 คน เก็บรวบรวมข้อมูลโดยใช้แบบสอบถามตอบด้วยตนเอง สถิติที่ใช้ในการวิเคราะห์ข้อมูล ได้แก่ สถิติ เชิงพรรณนา การวิเคราะห์ค่าสัมประสิทธ์สหสัมพันธ์ของเพียร์สัน (Pearson’s Correlation Coefficient) การทดสอบไคสแควร์ (Chi-square Test) และการวิเคราะห์การถดถอยพหุคูณ แบบขั้นตอน (Stepwise Multiple Regression Analysis) ปัจจัยที่มีความสัมพันธ์อย่างมีนัยสำคัญทางสถิติกับการควบคุมตนเอง ในพฤติกรรมการสูบบุหรี่ในเด็กนักเรียนมัธยมศึกษาตอนต้น จังหวัดสุพรรณบุรี ได้แก่ เพศ พฤติกรรม การสูบบุหรี่ในปัจจุบัน ผลสัมฤทธิ์ทางการเรียน ค่าใช้จ่ายที่ได้รับจากผู้ปกครอง อิทธิพลของกลุ่มเพื่อน การรับรู้กฎระเบียบเรื่องบุหรี่ของโรงเรียน และ สภาพแวดล้อมของชุมชน นอกจากนี้ยังพบว่า เพศ พฤติกรรมการสูบบุหรี่ในปัจจุบัน ผลสัมฤทธิ์ทางการเรียน อิทธิพลของกลุ่มเพื่อน การรับรู้กฎระเบียบ เรื่องบุหรี่ของโรงเรียน สามารถร่วมกันทำนายการควบคุมพฤติกรรมการสูบบุหรี่ของตนเองได้ คิดเป็นร้อยละ 52

4. วิธีดำเนินการวิจัย

4.1 กรอบแนวคิดในการวิจัย



ภาพที่ 1 กรอบแนวคิดในการทำการวิจัย เริ่มจากการนำชุดข้อมูลมาคัดเลือกและแปลงข้อมูล จากนั้นนำข้อมูลมาสร้างแบบจำลองโมเดล ประเมินผลของข้อมูลจากเทคนิคจำแนกข้อมูล และเทคนิคกลุ่มก้อนการจำแนกข้อมูล เพื่อเปรียบเทียบหาวิธีที่ให้ค่าแม่นยำสูงที่สุด

4.2 ขั้นตอนการดำเนินงานวิจัย

1. การรวบรวมข้อมูล (Data Gathering) เป็นขั้นตอนการนำข้อมูลผู้ป่วยมะเร็งเต้านม จากเว็ปไซต์:

https://www.kaggle.com/code/schmoyote/breast-cancer-classification-beginner-friendly/data มาใช้ตัวอย่างชุดข้อมูลผู้ป่วยมะเร็งเต้านม บริจาคในปี 2538 โดยมหาวิทยาลัยวิสคอนซิน

id	diagnosis	radius_me...	texture_m...	perimeter_...	area_mean
842302	M	17.99	10.38	122.8	1001
842517	M	20.57	17.77	132.9	1326
84300903	M	19.69	21.25	130	1203
84348301	M	11.42	20.38	77.58	386.1
84358402	M	20.29	14.34	135.1	1297
843786	M	12.45	15.7	82.57	477.1
844359	M	18.25	19.98	119.6	1040
84458202	M	13.71	20.83	90.2	577.9
844981	M	13	21.82	87.5	519.8

ภาพที่ 2 ชุดข้อมูลผู้ป่วยมะเร็งเต้านม บริจาคในปี 2538 โดยมหาวิทยาลัยวิสคอนซิน

2. การเตรียมข้อมูลสำหรับการทำวิจัย (Data Pre - processing) เป็นขั้นตอนการเตรียมข้อมูลก่อนที่จะเข้าสู่การวิเคราะห์ข้อมูล

2.1.การคัดเลือกข้อมูล

การนำข้อมูลผู้ป่วยมะเร็งเต้านม ที่จะใช้ในการจำแนกข้อมูล มีลักษณะข้อมูลที่ใช้ตามข้อมูลในตารางที่ 1

ตารางที่ 1 ตารางหัวข้อการนำข้อมูลของผู้ป่วยมะเร็งเค้านมมาจำแนก

ลำดับ	ชื่อ	คำอธิบาย
1	ID Number	เลข ID ผู้ที่เป็นมะเร็ง
2	Diagnosis	ผมการวินิจฉัย (เป็น หรือ ไม่เป็น)
3	Radius	ค่าเฉลี่ยระยะทางจากจุดศูนย์กลางถึงเส้นขอบของเซลล์
4	Texture	ค่าเบี่ยงเบนมาตรฐานของค่าระดับสีเทาของเซลล์
5	Perimeter	เส้นรอบรูปของเซลล์
6	Area	พื้นที่ของเซลล์
7	Smoothness	ความผันแปรในความยาวรัศมีของเซลล์
8	Compactness	ค่าของ perimeter^2 / area - 1.0
9	Concavity	ความซัดของส่วนเว้าของรูปร่าง
10	concave points	จำนวนส่วนเว้าของรูปร่าง
11	Symmetry	ความสมมาตร
12	Fractal dimension	ค่าของ "coastline approximation" -1

2.2.การแปลงข้อมูล เนื่องจากข้อมูลมีทั้งที่เป็นตัวเลขและตัวอักษร อยู่ในรูปแบบที่ไม่สามารถวิเคราะห์ได้ จึงจำเป็นต้องมีการแปลงข้อมูลให้อยู่ในรูปแบบที่สามารถวิเคราะห์ได้

3. การสร้างแบบจำลองโมเดล (Modeling) ขั้นตอนนี้จะนำข้อมูลที่ใช้ในการวิเคราะห์เข้าไปวิเคราะห์ใน RapidMiner Studio โดยนำวิธี ขั้นตอนวิธีการเพื่อนบ้านใกล้ที่สุด (K-Nearest Neighbor), โครงข่ายประสาทเทียม (Neural Network), ต้นไม้การตัดสินใจ (decision tree), ซัพพอร์ตเวกเตอร์แมชชีน (Support Vector Machine : SVM), majority vote และ Bagging มาใช้ในการหาค่าความแม่นยำ

4. การประเมินผล (Evaluation) เป็นการประเมินแบบจำลองที่ใช้ในการวิเคราะห์ทั้งหมด เพื่อเปรียบเทียบประสิทธิภาพของแบบจำลอง หาค่าความแม่นยำที่ดีที่สุด โดยค่า Accuracy Recall Precision ที่มากที่สุด แสดงว่ามีผลการจำแนกดีที่สุด

5. ผลการเปรียบเทียบประสิทธิภาพของการจำแนกข้อมูลและการจำแนกข้อมูลด้วยวิธีกลุ่มก้อนข้อมูลแบบ Bagging กับ Majority Vote

ตารางที่ 2 ตารางผลความแม่นยำของการจำแนกข้อมูล และการจำแนกกลุ่มก้อนข้อมูลแบบ Bagging กับ Majority Vote

Technique		Accuracy	precision	Recall
Classification	K-nn	92.98	94.83	85.94
	Decision tree	90.06	91.23	81.25
	Neural net	97.08	98.36	93.75**
	SVM.	95.91	100	89.06
Bagging	K-nn	93.57	96.49	85.94
	Decision tree	92.40	91.8	87.5
	Neural net	97.66	100	93.75**
	SVM.	97.08	98.36	93.75
Majority Vote	K-nn	93.57	96.49	85.94
	Decision tree	90.06	92.73	79.69
	Neural net	97.66	100	93.75**
	SVM.	97.08	96.83	95.31

ผลการประเมินแบบจำลองที่ใช้ในการวิเคราะห์ จากตารางที่ 2 ผลจากการศึกษาพบว่าวิธีที่ให้ประสิทธิภาพที่ดีที่สุด คือ การโหวตเสียงส่วนใหญ่และบูตสตรัปแอกเกรเกตติ้งของโครงข่ายประสาทเทียมโดยให้ค่าความถูกต้องในการจำแนกเท่ากับ 97.66 ค่าค้นคืน 93.75 ค่าความแม่นยำเท่ากับ 96.4 รองลงมาคือ โครงข่ายประสาทเทียมให้ค่าความถูกต้องเท่ากับ 97.08 ค่าค้นคืน 93.75 ค่าความแม่นยำเท่ากับ 98.36 และบูตสตรัปแอกเกรเกตติ้งของซัพพอร์ตเวกเตอร์แมชชีนให้ค่าความถูกต้อง 97.08 ค่าค้นคืน 93.75 ค่าความแม่นยำเท่ากับ 98.36 การโหวตเสียงส่วนใหญ่ของซัพพอร์ตเวกเตอร์แมชชีนให้ค่าความถูกต้อง 97.08 ค่าค้นคืน 95.31 ค่าความแม่นยำ 96.83 ส่วนซัพพอร์ตเวกเตอร์แมชชีนให้ค่าความถูกต้อง 95.91 ค่าค้นคืน 89.06 ค่าความแม่นยำเท่ากับ 100 ส่วนการโหวตเสียงส่วนใหญ่และบูตสตรัปแอกเกรเกตติ้งของขั้นตอนวิธีการเพื่อนบ้านใกล้ที่สุดให้ค่าความถูกต้องเท่ากับ 93.57 ค่าระลึก 85.94 ค่าความแม่นยำเท่ากับ 96.49 ขั้นตอนวิธีการเพื่อนบ้านใกล้ที่สุดให้ค่าความถูกต้องเท่ากับ 92.98 ค่าระลึก 85.94 ค่าความแม่นยำเท่ากับ 94.83 บูตสตรัปแอกเกรเกตติ้งต้นไม้การตัดสินใจให้ค่าความถูกต้องเท่ากับ 92.4 ค่าระลึก 87.5 ค่าความแม่นยำเท่ากับ 91.8 การโหวตเสียงส่วนใหญ่ของต้นไม้การตัดสินใจให้ค่าความถูกต้องเท่ากับ 90.06 ค่าระลึก 79.69 ค่าความแม่นยำ 93.73 และต้นไม้การตัดสินใจให้ค่าความถูกต้อง 90.06 ค่าระลึก 81.25 ค่าความแม่นยำ 91.23

6. สรุปผลการวิจัย ข้อเสนอแนะ

6.1 สรุปผลการวิจัย

งานวิจัยนี้ได้ทำการเปรียบเทียบประสิทธิภาพเทคนิคกลุ่มก้อนการจำแนกข้อมูล ด้วยข้อมูลผู้ป่วยมะเร็งเต้านม มาทำการเปรียบเทียบหาวิธีที่ให้ค่าแม่นยำสูงที่สุด โดยใช้วิธี ขั้นตอนวิธีการเพื่อนบ้านใกล้ที่สุด (K-Nearest Neighbor) โครงข่ายประสาทเทียม (Neural Network) ต้นไม้การตัดสินใจ (decision tree) ซัพพอร์ตเวกเตอร์แมชชีน (Support Vector Machine : SVM) majority vote และ Bagging โดยใช้เกณฑ์การประเมินด้วยค่า Accuracy, Precision และ Recall

ผลการวิจัยพบว่าวิธีที่ให้ค่าความแม่นยำมากที่สุด คือ Majority vote และ Bagging ของโครงข่ายประสาทเทียม (Neural Network) ให้ค่า Accuracy 97.66 Recall 93.75 กับ 100 Precision 100 กับ 96.4 รองลงมาคือ โครงข่ายประสาทเทียม (Neural Network) ให้ค่า Accuracy 97.08 Recall 93.75 กับ 99.07 Precision 98.36 กับ 98.36 , Bagging ของซัพพอร์ตเวกเตอร์แมชชีน (Support Vector Machine : SVM) ให้ค่า Accuracy 97.08 Recall 93.75 กับ 99.07 Precision 98.36 กับ 96.36

6.2 ข้อเสนอแนะ

เนื่องจากการใช้เทคนิควิธีกลุ่มก้อนการจำแนกข้อมูล และเทคนิคการจำแนกข้อมูลที่นำมาใช้ในการวิจัยนั้น อาจจะไม่ใช่วิธีที่ให้ค่าความแม่นยำมากที่สุด ดังนั้น ควรทดลองใช้วิธีการอื่น ๆ ที่ยังไม่ได้ถูกนำมาใช้เพื่อที่จะหาวิธีที่ให้ค่าความแม่นยำมากที่สุด ของเทคนิควิธีกลุ่มก้อนการจำแนกข้อมูล

7. เอกสารอ้างอิง

[1] Vatinee Nuipian and Phayung Meesad. (2013). A Comparison of Filter and Wrapper Approaches with Text Mining for Text Classification. **The Journal of Industrial Technology.** Vol. 9, No. 3 (P.118-129). (In Thai)

[2] Patharawut Saengsiri, Sageemas Na Wichian and Phayung Meesad. (2010) Classification of Leukemia Data Using Ranking and SupportVector Machine, **KKU Res J (GS)**. Vol.10, No.2, P.10-17. (In Thai)

[3] Anutchai Chutipascharoen and Charun Sanrach. (2018). A Comparison of the Efficiency of Algorithms and Feature Selection Methods for Predicting the Success of Personal Overseas Money Transfer. **KKU RESEARCH JOURNAL OF HUMANITIES AND SOCIAL SCIENCES (GRADUATE STUDIES)** .Vol. 6 NO. 3: September-December 2018. Vol. 6, No. 3, P.105-113. (In Thai)

[4] Jiraporn Jareanying Werayuth Charoenruengkit, THE PREDICTION OF STUDENT PERFORMANCEUSING DATA MINING TECHNIQUES WITH RAPID MINER , **A Master's Project Submitted in Partial Fulfillment of the Requirements for the Degree of MASTER OF SCIENCE (Information Technology),** Faculty of Science, Srinakharinwirot University.(In Thai)

[5] Ratiporn Chanklan (2014). THE IMPROVEMENT OF SUPPORT VECTOR MACHINE ALGORITHM FOR BIOMETRIC IMAGE IDENTIFICATION, **School of Computer Engineering**, Suranaree University of Technology. (In Thai)

[6] Orawan Voraaroon, Piyatida Khajornchaikul, Chardsumon Prutipinyo,Pimsurang Taechaboonsermsak, and Supachai Pitikultang.(2017). Factors Related to Smoking Self-Control Among Lower Secondary School Students, Suphanburi Province. **The Public Health Journal of Burapha University**. Vol.12 No.2, P. 75-85. (In Thai)