อัลกอริทึมแบบรวมสำหรับการเลือกคุณสมบัติของข้อมูล

โดย

นายภูริพัทร์ ทองคำ

วิทยานิพนธ์นี้เป็นส่วนหนึ่งของการศึกษาตามหลักสูตร
วิทยาศาสตรมหาบัณฑิต (วิทยาการคอมพิวเตอร์)
สาขาวิชาวิทยาการคอมพิวเตอร์
คณะวิทยาศาสตร์และเทคโนโลยี มหาวิทยาลัยธรรมศาสตร์
ปีการศึกษา 2559

อัลกอริทึมแบบรวมสำหรับการเลือกคุณสมบัติของข้อมูล

โดย

นายภูริพัทร์ ทองคำ

วิทยานิพนธ์นี้เป็นส่วนหนึ่งของการศึกษาตามหลักสูตร
วิทยาศาสตรมหาบัณฑิต (วิทยาการคอมพิวเตอร์)
สาขาวิชาวิทยาการคอมพิวเตอร์
คณะวิทยาศาสตร์และเทคโนโลยี มหาวิทยาลัยธรรมศาสตร์
ปีการศึกษา 2559

ENSEMBLE ALGORITHM FOR FEATURE SELECTION

BY

MR. PURIPAT THONGKAM

A THESIS SUBMITTED IN PARTIAL FULFILLMENT OF THE REQUIREMENTS

FOR THE DEGREE OF MASTER OF SCIENCE (COMPUTER SCIENCE)

DEPARTMENT OF COMPUTER SCIENCE

FACULTY OF SCIENCE AND TECHNOLOGY

THAMMASAT UNIVERSITY

ACADEMIC YEAR 2016

มหาวิทยาลัยธรรมศาสตร์
คณะวิทยาศาสตร์และเทคโนโลยี

วิทยานิพนธ์

ของ

นายภูริพัทร์ ทองคำ

เรื่อง

อัลกอริทึมแบบรวมสำหรับการเลือกคุณสมบัติของข้อมูล

ได้รับการตรวจสอบและอนุมัติ ให้เป็นส่วนหนึ่งของการศึกษาตามหลักสูตร
วิทยาศาสตรมหาบัณฑิต (วิทยาการคอมพิวเตอร์)

เมื่อ วันที่ 31 กรกฎาคม พ.ศ. 2560

ประธานกรรมการสอบวิทยานิพนธ์ _____
(ผู้ช่วยศาสตราจารย์ ดร. วิรัตน์ จารีวงศ์ไพบูลย์)

กรรมการและอาจารย์ที่ปรึกษาวิทยานิพนธ์ _____
(ผู้ช่วยศาสตราจารย์ ดร. ปกรณ์ ลี้สุทธิพรชัย)

กรรมการสอบวิทยานิพนธ์ _____
(ผู้ช่วยศาสตราจารย์ ดร. เด่นดวง ประดับสุวรรณ)

กรรมการสอบวิทยานิพนธ์ _____
(อาจารย์ ดร. ขจรพงษ์ อัครจิตสกุล)

คณบดี _____
(รองศาสตราจารย์ ปกรณ์ เสริมสุข)

หัวข้อวิทยานิพนธ์	อัลกอริทึมแบบรวมสำหรับการเลือกคุณสมบัติของข้อมูล
ชื่อผู้เขียน	นายภูริพัทร์ ทองคำ
ชื่อปริญญา	วิทยาศาสตรมหาบัณฑิต (วิทยาการคอมพิวเตอร์)
สาขาวิชา/คณะ/มหาวิทยาลัย	สาขาวิชาวิทยาการคอมพิวเตอร์ คณะวิทยาศาสตร์และเทคโนโลยี มหาวิทยาลัยธรรมศาสตร์
อาจารย์ที่ปรึกษาวิทยานิพนธ์	ผศ.ดร. ปกรณ์ ลี้สุทธิพรชัย
ปีการศึกษา	2559

บทคัดย่อ

งานวิจัยนี้นำเสนอเทคนิคในการปรับปรุงอัลกอริทึมสำหรับ การคัดเลือกคุณสมบัติแบบรวม (Ensemble Feature Selection) ที่มีอยู่เดิม เพื่อให้มีประสิทธิภาพที่ดีขึ้นโดยพิจารณาความสำคัญ (Priority) ลำดับของคุณสมบัติ (Order) และคะแนนของแต่ละคุณสมบัติ ข้อมูลที่ใช้ในการวิจัยมาจากแหล่งข้อมูลที่มีชื่อว่า “Kent Ridge Bio - Medical” และ “Machine Learning Data Repository” ซึ่งเป็นแหล่งข้อมูลสาธารณะด้านการแพทย์โดยชุดข้อมูลที่นำมาใช้คือข้อมูลการเกิดโรคมะเร็งปอด, มะเร็งต่อมน้ำเหลือง, มะเร็งเต้านม, มะเร็งรังไข่และมะเร็งเม็ดเลือดขาว สำหรับอัลกอริทึมการคัดเลือกคุณสมบัติที่ใช้ในงานวิจัยมีดังนี้ Symmetrical Uncertainty, ReliefF, Information Gain , Gain Ratio และ OneR

ในการวัดผลจะใช้ค่า AUC (Area Under Curve), Recall และ Precision ที่คำนวณได้จากแบบจำลอง ซึ่งสร้างจากกระบวนการจำแนกประเภทข้อมูล (Classification) โดยเทคนิคการจำแนกประเภทข้อมูลที่นำมาใช้ในงานวิจัยนี้ได้แก่ อัลกอริทึมการแยกประเภทแบบการหาเพื่อนบ้านใกล้ที่สุด (K – Nearest Neighbor) , อัลกอริทึมการจำแนกประเภทแบบเบย์ (Naïve Bayes), อัลกอริทึมการจำแนกประเภทแบบซัพพอร์ตเวกเตอร์แมชชีน (Support Vector Machines) , อัลกอริทึมการจำแนกประเภทแบบการสุ่มป่าไม้ (Random Forest) , อัลกอริทึมการจำแนกประเภทแบบถดถอยโลจิสติก (Logistic Regression) และอัลกอริทึมการจำแนกประเภทแบบต้นไม้ตัดสินใจ (Decision Tree) ผลลัพธ์ของอัลกอริทึมที่นำเสนอจะถูกนำมาเปรียบเทียบกับอัลกอริทึมสำหรับการคัดเลือกคุณสมบัติแบบรวมที่มีอยู่เดิมกับแบบเดี่ยว(Individuals)

คำสำคัญ: คุณสมบัติ , การคัดเลือกคุณสมบัติแบบรวม , การจำแนกประเภทข้อมูล

Thesis Title	ENSEMBLE ALGORITHM FOR FEATURE SELECTION
Author	MR. PURIPAT THONGKAM
Degree	MASTER OF SCIENCE (COMPUTER SCIENCE)
Major Field/Faculty/University	COMPUTER SCIENCE FACULTY OF SCIENCE AND TECHNOLOGY THAMMASAT UNIVERSITY
Thesis Advisor	ASST. PROF. DR. PAKORN LEESUTTHIPORNCHAI
Academic Years	2016

ABSTRACT

This thesis proposed the technique to improve performance of the existing ensemble features selection algorithm by using priority, features rank order and feature score as considerate factors. Datasets were gathered from public research dataset repositories called “Kent Ridge bio-medical” and “Machine Learning Data”. The selected data sets are lung cancer, lymphoma, breast cancer, ovarian cancer and leukemia. The feature ranker algorithms that were used to evaluate the performance of ensemble algorithm are Symmetrical Uncertainty, ReliefF, Information Gain, Gain Ratio and OneR.

For performance evaluation, AUC (Area Under Curve), Precision, and Recall are considered as metrics to prove that the selected features from the proposed algorithm is better than those of the existing ensemble features selection algorithm and single feature ranker algorithms. The values of AUC, Precision and Recall are calculated from the classification results that are obtained from classification techniques. The selected classification techniques are K-Nearest Neighbor, Naïve Bayes, Support Vector Machines, Random Forest, Logistic Regression and Decision Tree.

Keywords: features, ensemble features selection , classification

กิตติกรรมประกาศ

วิทยานิพนธ์ฉบับนี้สำเร็จลุล่วงได้ด้วยความกรุณาและความช่วยเหลือเป็นอย่างดียิ่งจากอาจารย์ที่ปรึกษาวิทยานิพนธ์ ผศ.ดร.ปกรณ์ ลี้สุทธิพรชัย ที่กรุณาเสียสละเวลาให้คำแนะนำคำปรึกษา ตรวจแก้ไขข้อบกพร่องด้วยความเอาใจใส่ ผู้วิจัยรู้สึกซาบซึ้งในความกรุณาของอาจารย์และขอกราบขอบพระคุณเป็นอย่างสูงไว้ ณ ที่นี้

ขอกราบขอบพระคุณ ผศ.ดร.วิรัตน์ จารีวงศ์ไพบูลย์ ผศ.ดร.เด่นดวง ประดับสุวรรณ และ อาจารย์ ดร.ขจรพงษ์ อัครจิตสกุล กรรมการสอบวิทยานิพนธ์ ที่ให้คำแนะนำและเสียสละเวลาในการตรวจสอบวิทยานิพนธ์ และขอขอบคุณเจ้าหน้าที่ภาควิชาวิทยาการคอมพิวเตอร์ ที่ช่วยอำนวยความสะดวกในการศึกษาตลอดหลักสูตรให้เป็นไปด้วยความเรียบร้อย และขอขอบพระคุณ คณะวิทยาศาสตร์และเทคโนโลยีที่ได้มอบทุนการศึกษาให้แก่ผู้วิจัย ผู้วิจัยรู้สึกซาบซึ้งเป็นอย่างยิ่ง

สุดท้ายนี้ ผู้วิจัยขอกราบขอบพระคุณครอบครับทองคำ มารดาและบิดา ที่เป็นขวัญกำลังใจ สนับสนุน และอยู่เคียงข้างมาโดยตลอด และขอบคุณเพื่อนๆ ทุกคนที่ให้ความช่วยเหลือในการจัดทำวิทยานิพนธ์นี้สำเร็จลุล่วงไปด้วยดี

นายภูริพัทร์ ทองคำ

สารบัญ

สารบัญตาราง

สารบัญภาพ

รายการสัญลักษณ์และคำย่อ

สัญลักษณ์/คำย่อ	คำเต็ม/คำจำกัดความ
AUC	Area Under Curve
IG	Information Gain
GR	Gain Ratio
RFF	Relief F
SU	Symmetrical Uncertainty
OneR	One Rule
K-NN	K – Nearest Neighbor
LR	Logistic Regression
C4.5	Decision Tree

บทที่ 1
บทนำ

การคัดเลือกคุณสมบัติ (Feature Selection) เป็นขั้นตอนหนึ่งที่สำคัญในกระบวนทำเหมืองข้อมูล (Data Mining) เนื่องจากกระบวนการคัดเลือกคุณสมบัติจะคัดเลือกคุณสมบัติ ที่มีแนวโน้มทำให้โมเดล (Model) ผลลัพธ์ที่เกิดจากการจำแนกกลุ่มข้อมูล (Classification) มีความถูกต้องแม่นยำในการทำนายมากยิ่งขึ้น ซึ่งการคัดเลือกคุณสมบัติสามารถแบ่งออกเป็น 2 ลักษณะ คือการคัดเลือกคุณสมบัติแบบเดี่ยว (Individual Feature Selection) และการคัดเลือกคุณสมบัติแบบรวม (Ensemble Feature Selection) ซึ่งในงานวิจัยนี้พิจารณาการคัดเลือกคุณสมบัติแบบรวมเป็นหลัก

1.1 ความเป็นมาและความสำคัญของปัญหา

ในปัจจุบันมีงานวิจัยที่เกี่ยวข้องกับการคัดเลือกคุณสมบัติแบบรวมมากมายซึ่งแต่ละงานวิจัยก็ได้มีการนำเสนออัลกอริทึมในรูปแบบต่างๆ เพื่อให้ผลลัพธ์จากการคัดเลือกคุณสมบัติแบบรวมมีประสิทธิภาพ แต่บางอัลกอริทึมที่ถูกนำเสนอ ยังสามารถที่จะปรุงปรับให้มีประสิทธิภาพดียิ่งขึ้น ซึ่งหนึ่งในงานวิจัยที่ผู้วิจัยเลงเห็นว่าสามารถปรับปรุงการทำงานของอัลกอริทึม ให้มีประสิทธิภาพดียิ่งขึ้นได้ คืองานวิจัย “Ensemble of Feature Selection Techniques for High Dimensional Data” ของ Sri Harsha Vege ดังนั้นผู้วิจัยจึงจะใช้งานวิจัยนี้เป็นต้นแบบเพื่อทำการพัฒนาต่อยอด

1.2 วัตถุประสงค์การวิจัย

1.2.1 เพื่อพัฒนาอัลกอริทึมที่ช่วยเพิ่มประสิทธิภาพของการคัดเลือกคุณสมบัติแบบรวม (Ensemble Feature Selection) จาก (1) งานวิจัยที่มีอยู่เดิม

1.2.2 เพื่อศึกษาและเปรียบเทียบประสิทธิภาพของการคัดเลือกคุณสมบัติแบบรวมกับแบบเดี่ยว

1.3 ประโยชน์ที่คาดว่าจะได้รับ

1.3.1 ได้อัลกอริทึมสำหรับการคัดเลือกคุณสมบัติแบบรวมที่มีประสิทธิภาพดีกว่า (1) งานวิจัยต้นแบบ

1.4 ขอบเขตงานวิจัย

1.4.1 ข้อมูลมะเร็งปอด นำมาจากแหล่งข้อมูลที่มีชื่อว่า “Kent Ridge Bio - Medical” ซึ่งมีจำนวนคุณสมบัติ 57 คุณสมบัติและจำนวนกรณี 32 กรณี

1.4.2 ข้อมูลมะเร็งต่อมน้ำเหลือง นำมาจากแหล่งข้อมูลที่มีชื่อว่า “Kent Ridge Bio - Medical” ซึ่งมีจำนวนคุณสมบัติ 4,027 คุณสมบัติและจำนวนกรณี 96 กรณี

1.4.3 ข้อมูลมะเร็งเต้านม นำมาจากแหล่งข้อมูลที่มีชื่อว่า “Machine Learning Data Repository” ซึ่งมีจำนวนคุณสมบัติ 24,482 คุณสมบัติและจำนวนกรณี 78 กรณี

1.4.4 ข้อมูลมะเร็งรังไข่ นำมาจากแหล่งข้อมูลที่มีชื่อว่า “Machine Learning Data Repository” ซึ่งมีจำนวนคุณสมบัติ 15,155 คุณสมบัติและจำนวนกรณี 253 กรณี

1.4.5 ข้อมูลมะเร็งเม็ดเลือดขาว นำมาจากแหล่งข้อมูลที่มีชื่อว่า “Machine Learning Data Repository” ซึ่งมีจำนวนคุณสมบัติ 7,143 คุณสมบัติและจำนวนกรณี 49 กรณี

1.5 รายละเอียดของหัวข้อวิจัย

รายละเอียดของหัวข้อวิจัยสามารถแบ่งออกเป็น 5 บทดังนี้

บทที่ 1 บทนำ ความเป็นมาและความสำคัญของปัญหา วัตถุประสงค์ ประโยชน์ที่คาดว่าจะได้รับ ขอบเขตงานวิจัย

บทที่ 2 เอกสาร แนวทาง ทฤษฎี ที่เกี่ยวข้องกับการทำเหมืองข้อมูล ขั้นตอนการทำเหมืองข้อมูล ขั้นตอนการคัดเลือกคุณสมบัติและงานวิจัยประยุกต์ ที่เกี่ยวข้องกับงานวิจัยนี้

บทที่ 3 วิธีการดำเนินงานวิจัยประกอบด้วยรายละเอียดขั้นตอนของงานวิจัย และวิธีการวัดผลการวิจัย

บทที่ 4 ผลของค่า AUC Precision และ Recall ที่ได้จากการสร้างแบบจำลองในแต่ละชุดข้อมูล และอภิปรายผลลัพธ์การทดลอง

บทที่ 5 สรุปผลการดำเนินงานวิจัย ข้อเสนอแนะ และงานวิจัยในอนาคต

บทที่ 2
วรรณกรรมและงานวิจัยที่เกี่ยวข้อง

ผู้วิจัยได้ทำการศึกษาเอกสาร แนวทาง ทฤษฎี และงานวิจัยที่เกี่ยวข้องดังนี้

2.1 การทำเหมืองข้อมูล (Data Mining)

2.2 ชุดข้อมูลที่นำมาวิจัย (Dataset)

2.3 การคัดเลือกคุณสมบัติ (Feature Selection)

2.4 เทคนิคการจำแนกกลุ่มข้อมูล (Classification)

2.5 เครื่องมือที่ใช้ในการทำงานวิจัย (Tools)

2.6 งานวิจัยที่เกี่ยวข้อง

2.1 การทำเหมืองข้อมูล (Data Mining)

การทำเหมืองข้อมูล (3,4) เป็นกระบวนการที่คัดเลือกข้อมูลที่มีประโยชน์จากฐานข้อมูลหรือแหล่งข้อมูลที่มีขนาดใหญ่เพื่อให้ได้ข้อมูลที่มีเหตุผล และช่วยในการตัดสินใจอีกทั้งยังช่วยในการทำนายสิ่งที่อาจเกิดขึ้นในอนาคต โดยลักษณะการทำเหมืองข้อมูลคือ นำเอาข้อมูลที่เก็บในอดีตมาสร้างแบบจำลอง (Model) โดยการสร้างแบบจำลองประกอบด้วย 4 ขั้นตอน คือ 1) การทำความเข้าใจข้อมูล (Data Understanding) โดยการเก็บรวบรวมข้อมูล ทำความคุ้นเคยกับข้อมูล ระบุปัญหา คุณภาพของข้อมูลแบบเชิงลึก ตั้งสมมติฐานเกี่ยวกับข้อมูลในมุมมองต่างๆ 2) การเตรียมข้อมูล (Data Preparation) โดยการสร้างชุดข้อมูล การเลือกตัวแปรที่สำคัญ การเปลี่ยนแปลงข้อมูล การทำความสะอาดข้อมูล 3) การสร้างแบบจำลอง (Modeling) ด้วยเทคนิคแบบต่างๆเช่นการจำแนกข้อมูล 4) การประเมินผล (Evaluation) ประเมินผลลัพธ์ของแบบจำลองที่ได้จากการสร้างโดยดูจากค่าต่างๆเช่น ค่าความถูกต้องของแบบจำลอง (Model Accuracy) หรือค่าของพื้นที่ใต้ส่วนโค้ง (Area Under Curve : AUC) เป็นต้น รูปแบบของการทำเหมืองข้อมูลนั้นมีด้วยกันหลายลักษณะแต่สามารถแบ่งออกเป็น 4 รูปแบบหลักๆคือ

2.1.1 การค้นหากฎความสัมพันธ์ (Association Rule)

เป็นการค้นหากฎความสัมพันธ์ของข้อมูลโดยค้นหาความเชื่อมโยงของข้อมูลตั้งแต่ 2 ชุดขึ้นไป

2.1.2 การจำแนกประเภทและการทำนาย (Classification and Prediction)

เป็นวิธีการที่ใช้สำหรับจัดการประเภทข้อมูลโดยการสร้างแบบจำลองที่อธิบายข้อมูลแต่ละประเภท ซึ่งอาจแสดงแบบจำลองในลักษณะที่เป็นต้นไม้ตัดสินใจ (Decision Tree) โครงข่ายประสาท (Neural Network) เป็นต้น

2.1.3 การจัดกลุ่มข้อมูล (Cluster Analysis)

เป็นเทคนิคที่ใช้สำหรับจำแนกกรณี (Case) หรือแบ่งตัวแปรออกเป็นกลุ่มย่อย ตั้งแต่ 2 กลุ่มขึ้นไป โดยลักษณะการจำแนกจะดูที่ความคล้ายคลึงกันของกรณีหรือ ตัวแปรนั้นๆ หากกรณีหรือตัวแปรที่กำลังพิจารณามีลักษณะเหมือนหรือคล้ายกับกลุ่มใด ตัวแปรดังกล่าวก็จะถูกจัดให้อยู่ในกลุ่มนั้นๆ

2.1.4 การสืบค้นรูปแบบโดยลำดับเหตุการณ์ (Sequential Pattern)

เป็นเทคนิคการหาความสัมพันธ์ของข้อมูลระหว่างการติดต่อกัน (Transaction) ซึ่งทำให้มีเวลาเข้ามาปัจจัยที่เกี่ยวข้อง ด้วยรูปแบบของลำดับที่เด่นชัดหรือพบบ่อย จะแสดงให้เห็นว่าถ้าหากเกิดเหตุการณ์นี้แล้วหรือพบกลุ่มของข้อมูลชุดนี้แล้วจะมีแนวโน้มที่จะเกิดเหตุการณ์หรือพบกลุ่มของข้อมูลแบบใดตามมาในภายหลัง

2.2 ชุดข้อมูลที่นำมาวิจัย (Dataset)

ชุดข้อมูล (21,22) ที่ใช้ในการวิจัยครั้งนี้นำมาจากแหล่งข้อมูลที่มีชื่อว่า“Kent Ridge Bio - Medical” และ “Machine Learning Data” ซึ่ง เป็นแหล่งข้อมูลสาธารณะ ที่มีการนำเอาข้อมูลด้านการแพทย์ต่างๆเช่น สาเหตุการเกิดมะเร็งปอด (Lung Cancer) หรือ ลูคีเมีย (Leukemia) มาเผยแพร่ให้ผู้สนใจทำการดาวน์โหลด ไปใช้ประโยชน์ในด้านต่างๆ โดยในงานวิจัยนี้ผู้วิจัยได้นำเอาข้อมูลมา 5 ชุดเพื่อทำการวิจัย สาเหตุที่ทำการเลือก 5 ชุดข้อมูลดังกล่าวเนื่องจากมีความสอดคล้องกับชุดข้อมูลที่งานวิจัยต้นแบบ (1) นำมาใช้

2.2.1 สาเหตุการเกิดมะเร็งปอด (Lung Cancer)

ชุดข้อมูลของสาเหตุการเกิดมะเร็งปอด ประกอบด้วยจำนวนคุณสมบัติ (Features) จำนวน 57 คุณสมบัติ และจำนวนกรณี (Instances) 32 กรณี โดยค่าของคุณสมบัติ ทุกค่าเป็นตัวเลขและค่าของคลาสของคุณสมบัติประกอบด้วยค่าตัวเลข 3 ค่า

2.2.2 สาเหตุการเกิดมะเร็งต่อมน้ำเหลือง (Lymphoma)

ชุดข้อมูลของสาเหตุการเกิดมะเร็งต่อมน้ำเหลือง ประกอบด้วย จำนวนคุณสมบัติ

(Features) จำนวน 4,027 คุณสมบัติ และจำนวนกรณี (Instances) 96 กรณี โดยค่าของคุณสมบัติทุกค่าเป็นตัวอักษร และค่าของคลาสของคุณสมบัติประกอบด้วยค่าตัวอักษร 9 ค่า

2.2.3 สาเหตุการเกิดมะเร็งเต้านม (Breast Cancer)

ชุดข้อมูลของสาเหตุการเกิดมะเร็งเต้านม ประกอบด้วยจำนวนคุณสมบัติ (Features) จำนวน 24,482 คุณสมบัติ และจำนวนกรณี (Instances) 78 กรณี โดยค่าของคุณสมบัติ ทุกค่าเป็นตัวอักษร และค่าของคลาสของคุณสมบัติประกอบด้วยค่าตัวอักษร 2 ค่า

2.2.4 สาเหตุการเกิดมะเร็งรังไข่ (Ovarian Cancer)

ชุดข้อมูลของสาเหตุการเกิดมะเร็งรังไข่ ประกอบด้วยจำนวนคุณสมบัติ (Features) จำนวน 15,155 คุณสมบัติ และจำนวนกรณี (Instances) 253 กรณี โดยค่าของคุณสมบัติ ทุกค่าเป็นตัวอักษร และค่าของคลาสของคุณสมบัติประกอบด้วยค่าตัวอักษร 2 ค่า

2.2.5 สาเหตุการเกิดมะเร็งเม็ดเลือดขาว (Leukemia)

ชุดข้อมูลของสาเหตุการเกิดมะเร็งเม็ดเลือดขาว ประกอบด้วย จำนวนคุณสมบัติ (Features) จำนวน 7,143 คุณสมบัติและจำนวนกรณี (Instances) 49 กรณี โดยค่าของคุณสมบัติทุกค่าเป็นตัวอักษร และค่าของคลาสของคุณสมบัติประกอบด้วยค่าอักษร 2 ค่า

2.3 การคัดเลือกคุณสมบัติ (Feature Selection)

การคัดเลือกคุณสมบัติ (6,9) เป็นวิธีการที่ช่วยลดจำนวนตัวแปรที่ใช้ในการพยากรณ์แบบจำลอง โดยอาจทำเพื่อเลือกคุณสมบัติที่ดีที่สุดเพียงคุณสมบัติเดียว หรือเลือกกลุ่มคุณสมบัติที่มีความสำคัญต่อการพยากรณ์ กระบวนการคัดเลือกคุณสมบัติเป็นกระบวนการที่สำคัญในการเตรียมข้อมูลสำหรับ การทำเหมืองข้อมูลเพื่อให้การพยากรณ์ของแบบจำลองมีประสิทธิภาพมากขึ้น เนื่องจากมีการลดจำนวนคุณสมบัติที่ไม่จำเป็นต่อการพยากรณ์ หรือทำให้ค่าของการพยากรณ์ผิดพลาดออกไปการคัดเลือกคุณสมบัติสามารถแบ่งประเภทได้ 2 แบบคือ

2.3.1 การคัดเลือกสับเซตจากคุณสมบัติทั้งหมด (Feature Subset Selection)

การคัดเลือกเอาเพียงบางสับเซตของคุณสมบัติ จากจำนวนคุณสมบัติทั้งหมดโดยสับเซตที่เลือกมีแนวโน้มที่จะทำให้ประสิทธิภาพของการทำนายจากแบบจำลองดีขึ้น

2.3.2 การจัดลำดับคุณสมบัติ (Feature Ranking)

การคำนวณคะแนนของแต่ละคุณสมบัติ (Feature Score) แล้วทำการเรียงลำดับของแต่ละคุณสมบัติตามคะแนนที่ได้โดยเรียงจากมากไปน้อย

ในงานวิจัยนี้จะเน้นไปที่เทคนิคการจัดลำดับคุณสมบัติ เพื่อช่วยให้ประสิทธิภาพการทำนายของแบบจำลองดีขึ้น โดยเทคนิคการจัดลำดับคุณสมบัติที่ใช้มีดังนี้

2.3.2.1 การจัดลำดับคุณสมบัติแบบ Information Gain (IG)

การจัดลำดับคุณสมบัติแบบ Information Gain (17) มีหลักการพื้นฐานจากการสุ่มตัวอย่าง (Entropy) โดยค่าที่ได้จาก Information Gain คือค่าของความต่างระหว่างตัวแปร X ที่เป็น ตัวแปรเป้าหมาย (Target Variable) กับตัวแปรอิสระ A (Independent Variable) ซึ่งลักษณะของ Information Gain จะทำการลด Entropy ของตัวแปรเป้าหมาย X โดยการเรียนรู้จากสถานะของ ตัวแปรอิสระ A

Information Gain มีแนวโน้มที่จะเลือกคุณสมบัติ ที่มีจำนวนค่าที่แตกต่างกันสูงมากกว่าคุณสมบัติที่มีจำนวนค่าที่แตกต่างกันต่ำ ถึงแม้ว่าคุณสมบัติที่มีจำนวนค่าที่แตกต่างกันต่ำจะให้ข้อมูลที่เป็นประโยชน์

การคำนวณของเทคนิค Information Gain นั้นจะพิจารณาระหว่างคุณสมบัติ X ซึ่งก็คือตัวแปรเป้าหมายกับคุณสมบัติที่เป็นคลาสของข้อมูล Y หรือตัวแปรอิสระจากนั้นทำการดูจากความน่าจะเป็นที่เกิดขึ้นระหว่างค่าของคุณสมบัติ X กับค่าของคุณสมบัติ Y ว่ามีมากน้อยเพียงใด หากมีความน่าจะเป็นเกิดขึ้นน้อยก็จะทำการให้คะแนนของคุณสมบัติ X น้อยตามไปด้วยโดยสามารถเขียนเป็นสมการได้ดังนี้

$$Gain(Y;X) = H(Y) - H(Y|X) \quad (1)$$

โดย $H(Y)$ คือค่าความน่าจะเป็นจากการสุ่มตัวอย่างของ Y

$H(Y|X)$ คือค่าความน่าจะเป็นจากการสุ่มตัวอย่างของ Y เทียบกับ X

$Gain(Y;X)$ คือค่าของคะแนนจากการสุ่มตัวอย่างที่คำนวณได้โดยจะมีตั้งแต่ 0 ถึง 1

ค่าของ Y เป็นค่าคุณสมบัติที่เป็นคลาสของข้อมูลซึ่งคำนวณตั้งแต่ { $Y_1, Y_2, \dots, Y_n$}

ค่าของ X เป็นค่าคุณสมบัติอื่นๆที่ไม่ใช่คลาส ซึ่งคำนวณตั้งแต่ { $X_1, X_2, \dots, X_n$}

ค่าของ $H(Y)$ และ $H(Y|X)$ คำนวณมาจาก

$$\mathrm{H(Y)} = -\sum_{i=1}^{i=k} P(Y = y_i) \log_2 P(Y = y_i) \quad (2)$$

$$\mathrm{H(Y|X)} \ = -\sum_{i=1}^{i=k} P(X = x_i)\, \mathrm{H}(Y|X = x_i) \quad (3)$$

โดย $P(Y = y_i)$ คือ ค่าความน่าจะเป็นตั้งแต่ y_1 จนถึง y_k

$P(X = x_i)$ คือ ค่าความน่าจะเป็นตั้งแต่ x_1 จนถึง x_k

2.3.2.2 การจัดลำดับคุณสมบัติแบบ Gain Ratio (GR)

เป็นวิธีการจัดลำดับคุณสมบัติ (18) โดยมีหลักการทำงานในลักษณะเดียวกับการทำงานของต้นไม้ตัดสินใจ โดยทำการหาตัวแปรที่เป็นตัวแบ่งข้อมูลออกเป็นกลุ่มย่อยที่มีสมาชิกภายในกลุ่มเป็นชนิดเดียวกันมากที่สุดด้วยสิ่งที่เรียกว่า อัตราส่วนเกน (Gain Ration) ซึ่งอัตราส่วนเกนคือมาตรวัดการได้ประโยชน์จากการแบ่งกลุ่มย่อย โดยค่าของอัตราส่วนเกนมาจากอัตราส่วนระหว่างค่าเกน (Gain) กับค่าสารสนเทศการแบ่งกลุ่ม (Split Info) ซึ่งเป็นการลดปัจจัยของตัวแปรที่มีค่าหลากหลาย ผลลัพธ์ของอัตราส่วนเกนจะนำไปจัดลำดับคุณสมบัติต่อไป โดยคุณสมบัติใดมีค่าอัตราส่วนเกนสูงจะถือว่ามีอิทธิพลในการพยากรณ์ตัวแปรเป้าหมายมาก โดยสมการหาอัตราส่วนเกนสามารถเขียนได้ดังนี้

$$GainRatio(A) = \frac{Gain(A)}{SplitInfo(A)} \quad (4)$$

2.3.2.3 การจัดลำดับคุณสมบัติแบบ Symmetrical Uncertainty (SU)

เป็นวิธีการจัดลำดับคุณสมบัติ (18) โดยใช้ความสัมพันธ์ระหว่างคุณสมบัติย่อย กับคุณสมบัติหลักที่เป็นคลาส โดยมีสมมติฐานในการจัดลำดับคือคุณสมบัติย่อยที่ดี ต้องมีความสัมพันธ์กับคุณสมบัติหลักที่เป็นคลาสสูง แต่จะมีความสัมพันธ์กับคุณสมบัติย่อยด้วยกันต่ำ ดังนั้นจึงสามารถกล่าวโดยสรุปว่า Symmetrical Uncertainty ใช้ในการวัดระดับความสัมพันธ์ ของคุณสมบัติย่อยกับคุณสมบัติที่เป็นคลาส โดยสามารถเขียนเป็นสมการได้ดังนี้

$$SU = 2.0 * \frac{\mathrm{H(X)+H(Y)-H(X,Y)}}{H(Y)+H(X)} \quad (5)$$

โดย $\mathrm{H(X)}$ คือค่าความน่าจะเป็นจากการสุ่มตัวอย่างของ X

$\mathrm{H(Y)}$ คือค่าความน่าจะเป็นจากการสุ่มตัวอย่างของ Y

$\mathrm{H(Y,X)}$ คือค่าความน่าจะเป็นจากการสุ่มตัวอย่างของ X และ Y

SU ค่าของคะแนนความสัมพันธ์ที่คำนวณได้โดยจะมีตั้งแต่ 0 ถึง 1

2.3.2.4 การจัดลำดับคุณสมบัติแบบ ReliefF (RFF)

เป็นวิธีการจัดลำดับคุณสมบัติ (14) โดยอาศัยหลักการที่ว่าคุณสมบัติใดที่สามารถแยกกรณีที่คล้ายกันได้มาก ย่อมมีแนวโน้มที่คุณสมบัตินั้น จะทำให้ประสิทธิภาพของแบบจำลองดีขึ้นด้วย โดยขั้นตอนการจัดลำดับคุณสมบัติแบบ ReliefF เป็นดังนี้

1. คำนวณค่าเหตุการณ์ที่เกือบพลาด (Nearest Miss) และ ค่าเหตุการณ์ที่เกือบโดน (Nearest Hit)
2. คำนวณน้ำหนัก (Weight) ของคุณสมบัติ
3. แสดงค่าผลลัพธ์ลำดับของคุณสมบัติที่ผ่านเกณฑ์ (Threshold)

การทำงานของ Relief ใช้เวลาเป็นเชิงเส้นตรง (Linear Time) ทั้งนี้ขึ้นอยู่กับจำนวนของคุณสมบัติ ด้วยสมการการทำงานของ ReliefF เป็นดังนี้

$$Relief_x = \mathrm{P}(\text{diff of X} \mid \text{diff class}) - \mathrm{P}(\text{diff of X} \mid \text{same class}) \quad (6)$$

โดย $\mathrm{P}(\text{diff of X} \mid \text{diff class})$ คือค่าเหตุการณ์ที่เกือบพลาด (Nearest Miss) ของคุณสมบัติ X

$\mathrm{P}(\text{diff of X} \mid \text{same class})$ คือค่าเหตุการณ์ที่เกือบโดน (Nearest Hit) ของคุณสมบัติ X

$Relief_x$ คือค่าน้ำหนักของคุณสมบัติ X

2.3.2.5 การจัดลำดับคุณสมบัติแบบ One Rule (OneR)

เป็นวิธีการจัดลำดับคุณสมบัติ (8) ที่ใช้ในการหากฎในการจำแนกกลุ่มข้อมูลโดยทำการสร้างกฎหนึ่งกฎขึ้นมาสำหรับแต่ละการทำนายในข้อมูลนั้นๆ จากนั้นทำการเลือกกฎหนึ่งกฎ ที่มีจำนวนข้อผิดพลาด (Error) น้อยที่สุดแล้วนำเอาคุณสมบัติที่อยู่ในกฎข้อนั้น มาใช้ในการทำนายต่อไป ลักษณะการทำงานของ One Rule เป็นดังนี้

1. ในทุกรอบของการทำนาย ทำการสร้างกฎโดย
   - นับจำนวนค่าของคลาสคุณสมบัติที่เกิดขึ้น
   - หาค่าของคลาสคุณสมบัติที่เกิดขึ้นบ่อยที่สุด
   - ทำการสร้างกฎสำหรับคลาสคุณสมบัตินั้นในการทำนายครั้งนี้
2. ทำการคำนวณค่าผิดพลาดของกฎแต่ละกฎ
3. เลือกกฎที่มีข้อผิดพลาดน้อยที่สุด

2.4 เทคนิคการจำแนกกลุ่มข้อมูล (Classification)

เป็นหนึ่งในเทคนิคที่ใช้ในการทำเหมืองข้อมูลโดยเทคนิคการจำแนกประเภทข้อมูล (3) คือการจัดแบ่งข้อมูลให้อยู่ในประเภทที่กำหนดคำตอบ (Class) ได้โดยกระบวนคือทำการแบ่งข้อมูลออกเป็น 2 ชุด คือชุดสอน (Train Data) และชุดทดสอบ (Test Data) โดยอาจแบ่งในอัตราส่วน 70 : 30 หรือ 75 : 25 ตามความเหมาะสม จากนั้นนำข้อมูลชุดสอนป้อนเข้าระบบ เพื่อให้ระบบทำการเรียนรู้โดย การสร้างเป็นต้นไม้แบบต่างๆ เช่นต้นไม้ตัดสินใจ (Decision Tree) หลังจากนั้นนำชุดข้อมูลทดสอบ ซึ่งเป็นชุดที่ไม่เหมือนกับชุดสอนทำการสร้างต้นไม้เช่นกันและเปรียบเทียบผลลัพธ์ที่ได้ โดยในงานวิจัยนี้ได้ใช้เทคนิคในการจำแนกกลุ่มข้อมูลดังนี้

2.4.1 ซัพพอร์ตเวกเตอร์แมชชีน (Support vector machine : SVM)

ซัพพอร์ตเวกเตอร์แมชชีน (4) เป็นเทคนิคที่ช่วยในการเรียนรู้ ซึ่งมีลักษณะการทำงานคล้ายกับ ANN (Artificial Neural Network) หรือโครงข่ายประสาทเทียม แต่ส่วนที่แตกต่างกัน คือ ในเทคนิคโครงข่ายประสาทเทียม ใช้หลักการลดความเสี่ยง เชิงการทดสอบให้มีค่าต่ำที่สุด (Empirical Risk Minimization : ERM) ส่วนซัพพอร์ตเวกเตอร์แมชชีน ใช้หลักการลดความเสี่ยงเชิงโครงสร้างให้ต่ำที่สุด (Structural Risk Minimization : SRM) โดยซัพพอร์ตเวกเตอร์แมชชีนประยุกต์การใช้งานได้ 2 รูปแบบคือการวิเคราะห์การถดถอย (Regression) หรือการประมาณค่าของฟังก์ชัน และการจำแนกประเภท (Classification) อีกทั้งซัพพอร์ตเวกเตอร์แมชชีน ยังถูกนำมาใช้ในการจำแนกข้อมูลในลักษณะหลายมิติ โดยการใช้ Hyper plane ในการแบ่งแยกคุณสมบัติออกจากกัน สำหรับการจำแนกในลักษณะไบนารี (Binary) นั้นมี Class A และ B โดย W คือ น้ำหนัก (Weight) , x คือคุณสมบัติต่างๆ และ b คือ Bias ดังสมการ

$$W^t x + b = 0 \tag{7}$$

โดยที่

$$W^t x + b \geq +1 \tag{8}$$

สำหรับทุก x ที่เป็นสมาชิก A

$$W^t x + b \leq -1 \tag{9}$$

สำหรับสมาชิก x ที่เป็นสมาชิกของ B ซึ่งมี Decision Function

$$f_{w,b} = sign(W^t\, x + b) \quad (10)$$

2.4.2 การหาเพื่อนบ้านใกล้ที่สุด (K – Nearest Neighbor : KNN)

เป็นเทคนิคที่ใช้กับปัญหาแบบการจำแนกกลุ่มข้อมูล (5) โดยเทคนิคการหาเพื่อนบ้านที่ใกล้ที่สุดมีความแตกต่างจากเทคนิคอื่นที่ไม่มีการแบ่งข้อมูลชุดสอน (Train Data) ในการสร้างแบบจำลองแต่จะใช้ข้อมูลทั้งชุดมาเป็นแบบจำลอง ในการใช้งานเทคนิคนี้จำเป็นจะต้องมีการระบุค่าตัวเลขจำนวนเพื่อนบ้านหรือค่า K ซึ่งค่านี้จะเป็นตัวบอกจำนวนของกรณี (Case) ที่จะต้องค้นหาในการทำนายกรณีใหม่ ซึ่งได้แก่ 1-NN , 2-NN , 3-NN , ... , K-NN โดยที่ K เป็นจำนวนเต็มบวก เช่น 2-NN หมายถึงจะมีการค้นหา 2 กรณีที่มีลักษณะใกล้เคียงกับกรณีใหม่ (Nearest Cases) ในการทำนายกรณีใหม่ ซึ่งสมการที่เกี่ยวข้องกับ KNN คือ

$$\mathrm{D} = \{(\mathrm{TV} - \mathrm{OB})/\mathrm{TV}\} * 100 \quad (11)$$

โดย D คือ ค่าระยะห่างของจุด 2 จุดของแบบ Percentage Disagreement

TV คือ ค่าทางทฤษฎี (Theoretical Value)

OB คือ ค่าที่สังเกตเห็น (Observed Value)

2.4.3 การจำแนกแบบเบย์ (Naïve-Bayes : NB)

เทคนิคที่ใช้ทฤษฎี Bayes Theorem (5) ในการคำนวณ อัลกอริทึม Naïve-Bayes เป็นเทคนิคสำหรับจัดการกับปัญหาประเภทจำแนกข้อมูล ผลลัพธ์ของเทคนิคนี้สามารถนำไปทำนายและช่วยอธิบายความหมายของข้อมูลได้อีกด้วย โดยการทำงานของอัลกอริทึม Naïve-Bayes คือทำการวิเคราะห์ความสัมพันธ์ระหว่างตัวแปรอิสระ แต่ละตัวกับตัวแปรตาม เพื่อใช้ในการสร้างเงื่อนไขความน่าจะเป็นสำหรับแต่ละความสัมพันธ์ ในทางทฤษฎีแล้วการทำนายผลของ Naïve-Bayes จะถูกต้อง ถ้าตัวแปรอิสระทั้งหมดเป็นอิสระต่อกัน โดยไม่ขึ้นกับตัวแปรอิสระตัวใดตัวหนึ่ง ซึ่งข้อมูลในปัจจุบันนั้นมีไม่มากนักที่ตัวแปรอิสระทั้งหมดจะเป็นอิสระต่อกัน ดังนั้นหากต้องใช้อัลกอริทึม Naïve-Bayes ผู้วิจัยควรคำนึงถึงประเด็นนี้ด้วย นอกจากนี้อัลกอริทึม Naïve-Bayes ยังไม่รองรับข้อมูลที่เป็นข้อมูลต่อเนื่อง (Continuous Data) ด้วยดังนั้นตัวแปรอิสระหรือตัวแปรตาม ที่มีค่าเป็นค่าต่อเนื่องจะต้องถูกแบ่งออกเป็นช่วง ซึ่งการแบ่งช่วงนั้นถ้ามีการแบ่งที่ไม่ดี ก็จะทำให้ผลลัพธ์ของแบบจำลองที่ได้มีคุณภาพไม่ดีตามไปด้วย ดังนั้นการที่จะใช้อัลกอริทึม Naïve-Bayes ควรจะคำนึงถึงลักษณะ

ของข้อมูลที่จะนำมาใช้ทำการทดสอบกับอัลกอริทึม Naïve-Bayes ด้วย

2.4.4 ต้นไม้ตัดสินใจ (C4.5)

เทคนิคที่ใช้ในสร้างต้นไม้ (6) จากชุดข้อมูลที่มีอยู่ โดยลักษณะของการสร้างต้นไม้จะใช้กฎในรูป แบบ “ถ้า เงื่อนไข แล้ว ผลลัพธ์” เช่น “ถ้าขยันอ่านหนังสือแล้วจะสอบผ่าน” เป็นต้น ต้นไม้ตัดสินใจเป็นเทคนิคที่ใช้กันอย่างแพร่หลาย เนื่องจากผลลัพธ์ที่ออกมาง่ายต่อการเข้าใจ เทคนิคต้นไม้ตัดสินใจ จะทำการจำกัดข้อมูลที่เป็นตัวแปรตาม (Dependent Variable) 1 ตัวต่อ 1 แบบจำลอง ซึ่งหากต้องการทำนายหลายตัว จะต้องทำการสร้างแบบจำลองสำหรับตัวแปรทุกตัว เช่นเดียวกันกับ Naïve-Bayes เทคนิคต้นไม้ตัดสินใจส่วนใหญ่จะไม่รองรับข้อมูลแบบต่อเนื่อง ดังนั้นจึงต้องมีการแบ่งข้อมูลให้เป็นแบบไม่ต่อเนื่องเสียก่อน อัลกอริทึมที่จัดอยู่ในต้นไม้ตัดสินใจได้แก่ Chi-squared Automatic Interaction Detection (CHAID) , Classification and Regression Tress (CART) , C4.5 เป็นต้น โดยในงานวิจัยนี้จะใช้ เทคนิคต้นไม้ตัดสินใจแบบ C4.5 ในการวิจัยนี้

2.4.5 การสุ่มป่าไม้ (Random Forest : RF)

เทคนิคที่ทำการสุ่มเลือกคุณสมบัติ ออกมาจากชุดข้อมูลหลายๆชุด (13) จากนั้นนำเอาชุดของคุณสมบัติเหล่านั้นมาสร้างแบบจำลองด้วยเทคนิคต้นไม้ตัดสินใจหลายๆต้น โดยเทคนิคการสุ่มป้าไม้ถูกนำเสนอครั้งแรกในปี ค.ศ. 1995 โดย Tin Kam ซึ่งต่อมาเทคนิคนี้ถูกต่อยอดโดย Leo Breiman ลักษณะของต้นไม้ที่อยู่ภายในป่าของเทคนิคการสุ่มป่าไม้ จะถูกควบคุมด้วย 3 ปัจจัยคือ

1. ต้นไม้แต่ละต้นจะถูกสอน (Train) โดยการใช้เซตย่อยจากข้อมูลตัวอย่าง
2. เมื่อต้นไม้โตขึ้น จะสามารถค้นหาโนด (Node) แต่ละโนดที่อยู่ในกิ่งที่ดีที่สุดของต้นไม้โดยใช้การสุ่มเลือกคุณสมบัติจาก N คุณสมบัติ
3. ต้นไม้แต่ละต้นจะไม่มีการตัดออก แต่จะปล่อยให้ต้นไม้โตขึ้นไปเรื่อยๆ จนได้ผลลัพธ์ที่ดีที่สุดหลังจากการสร้างป่า แล้วทำการให้คะแนน (Vote) โดยต้นไม้ภายในป่า หากต้นไม้ต้นใดได้คะแนนสูงสุดก็จะนำเอาต้นไม้นั้นออกมาสร้างเป็นโมเดล

2.4.6 การถดถอยโลจิสติก (Logistic Regression : LR)

การถดถอยโลจิสติก (2) เป็นเทคนิคที่ใช้ในการหาความสัมพันธ์ ระหว่างตัวแปรตามและตัวแปรอิสระ แล้วนำสมการที่ได้จากความความสัมพันธ์เหล่านั้นมาสร้างเป็นแบบจำลองจากนั้น นำแบบจำลองมาใช้ในการพยากรณ์ต่อไป ประเภทของการถดถอยแบบโลจิสติกมี 2 ประเภท คือ

1. Binary Logistic โดยวิธีนี้จะใช้เมื่อตัวแปรตามเป็นตัวแปรเชิงกลุ่มที่มีค่าได้เพียง 2 ค่า เช่น

$$\boldsymbol{Y} = \begin{cases} 1 \text{ ถ้าคนไข้เป็นมะเร็งปอด} \\ 0 \text{ ถ้าคนไข้ไม่ได้เป็นมะเร็งปอด} \end{cases}$$

2. Multinomial Logistic โดยวิธีนี้จะใช้เมื่อตัวแปรตามเป็นตัวแปรเชิงกลุ่มที่มีค่ามากกว่า 2 ค่าขึ้นไปเช่น

$$\boldsymbol{Y} = \begin{cases} 1 \text{ เป็นสีฟ้า} \\ 2 \text{ เป็นสีเขียว} \\ 3 \text{ เป็นสีส้ม} \end{cases}$$

ดังนั้นเมื่อต้องการจะใช้เทคนิคการถดถอยโลจิสติก ควรจะศึกษาชุดข้อมูลนั้นว่ามีลักษณะเป็นแบบใด แล้วเลือกวิธีการที่เหมาะสม เมื่อนำเอาข้อมูลมาผ่านเทคนิคการถดถอยโลจิสติกแบบใดก็ตาม จะได้ผลลัพธ์คือแบบจำลองที่ใช้ในการพยากรณ์ต่อไปได้

2.5 เครื่องมือที่ใช้ในการทำงานวิจัย (Tools)

ในงานวิจัยนี้ใช้เครื่องมือสำหรับทำกระบวนการเหมืองข้อมูล (1,6) ที่มีชื่อว่า Weka โดย Weka เป็นโปรแกรมที่มีเทคนิคต่างๆมากมายพัฒนาอยู่ภายในเช่น อัลกอริทึมการจัดหมวดหมู่: K-NN , LR , C4.5 , Support Vector Machine , การคัดเลือกคุณสมบัติ : Information Gain , Gain Ratio เป็นต้นซึ่ง Weka เป็นที่นิยมกันอย่างแพร่หลายในกลุ่มผู้วิจัยที่วิจัยเกี่ยวกับเหมืองข้อมูล Weka ไม่ได้มี เพียงตัวโปรแกรมที่สามารถใช้งานได้ด้วยตัวเอง (Standalone) เพียงอย่างเดียวแต่ยังมี Library สำหรับใช้คู่กับการเขียน โปรแกรมในภาษาจาวา (Java Programming) อีกด้วยซึ่งสะดวกมากสำหรับ ผู้วิจัยที่มีความจำเป็นจะต้องเขียนโปรแกรมเกี่ยวกับการทำเหมืองข้อมูล เช่น การคัดเลือกคุณสมบัติ , การหากฎความสัมพันธ์

2.6 งานวิจัยที่เกี่ยวข้อง

2.6.1 Osanaiye (12) นำเสนอวิธีการคัดเลือกคุณสมบัติแบบรวม เพื่อลดจำนวนคุณสมบัติ ให้น้อยที่สุดแต่ค่าความถูกต้องของแบบจำลองที่สร้างขึ้นยังคงมีค่าสูง ชุดข้อมูลที่นำมาใช้ในงานวิจัยนี้ คือบันทึกการโจมตีแบบ DDos (Distributed Denial-of-Service) บนระบบกลุ่มเมฆ (Cloud) ซึ่งมีที่มาจาก NSL-KDD และ KDD CUP อัลกอริทึมสำหรับการคัดเลือกคุณสมบัติ

แบบจัดลำดับที่นำมา ใช้ในงานวิจัยนี้มีด้วยกัน 4 อัลกอริทึมดังนี้ Information Gain, Gain Ratio Chi-Squared และ ReliefF จากนั้นทำการสร้างแบบจำลอง โดยใช้กระบวนการจัดหมวดหมู่ อัลกอริทึมสำหรับการจัดหมวดหมู่ ที่ใช้ในงานวิจัยนี้คือ C4.5 จากนั้นนำแบบจำลองที่สร้างจากการจัดหมวดหมู่ไปตรวจสอบค่าความถูกต้อง (Accuracy), อัตราการแจ้งเตือนจุดผิด (False Alarm Rate), อัตราการตรวจจับ (Detection Rate), เวลาที่ใช้ในการสร้างแบบจำลองและจำนวนของคุณสมบัติที่เหลือจากการคัดเลือก พบว่าวิธีที่ผู้วิจัยนำเสนอมีจำนวนของคุณสมบัติที่เหลือจากการคัดเลือกน้อยที่สุด เมื่อเปรียบเทียบกับการคัดเลือกคุณสมบัติแบบเดี่ยว อีกทั้งค่าความถูกต้อง อัตราการแจ้งเตือนจุดผิด , อัตราการตรวจจับ ยังมีค่าสูงกว่าอีกด้วย

2.6.2. Sujatha (13) ได้ทำการศึกษาเพื่อเปรียบเทียบประสิทธิภาพของอัลกอริทึมการคัดเลือกคุณสมบัติแบบจัดลำดับแต่ละวิธี โดยใช้ค่าของพื้นที่ใต้ส่วนโค้ง (Area Under Curve : AUC) ที่ได้จากแบบจำลองเป็นตัววัดผล สำหรับชุดข้อมูลที่ใช้ในงานวิจัยนี้คือ สาเหตุการเกิดมะเร็งปอดจาก แหล่งข้อมูลที่มีชื่อว่า “Kent Ridge Bio - Medical” อัลกอรึทึมสำหรับการคัดเลือกคุณสมบัติแบบจัดลำดับที่นำมาใช้ในการเปรียบเทียบมีด้วยกัน 5 อัลกอริทึมคือ Symmetrical Uncertainty, Relief, Information Gain, Gain Ratio และ OneR ลักษณะการทดลองคือเลือกคุณสมบัติมา N คุณสมบัติจากคุณสมบัติทั้งหมด จากนั้นนำ N คุณสมบัติมาผ่านทีละอัลกอริทึมแล้วนำผลลัพธ์ที่ได้จากคัดเลือกคุณสมบัติของแต่ละอัลกอริทึมมาผ่านกระบวนการจัดหมวดหมู่ เพื่อสร้างเป็นแบบจำลองแล้วดูผลลัพธ์จากค่าของ AUC ที่ได้จากแบบจำลอง อัลกอริทึมสำหรับการจัดหมวดหมู่ที่ใช้ในงานวิจัยนี้ได้แก่ K – Nearest Neighbor, Naïve Bayes, Support Vector Machines, Random Forest และ C4.5 ซึ่งจากผลการทดลอง ปรากฏว่าอัลกอรึทึมสำหรับการคัดเลือกคุณสมบัติแบบจัดลำดับที่มีค่า AUC โดยเฉลี่ยสูงสุดคือ Relief

2.6.3. Vege (1) นำเสนอวิธีการคัดเลือกคุณสมบัติแบบรวม (Ensemble Feature Selection) โดยใช้การคัดเลือกคุณสมบัติแบบจัดลำดับ ชุดข้อมูลที่ใช้ในงานวิจัยนี้ คือ สาเหตุการเกิดมะเร็งปอดและมะเร็งต่อมน้ำเหลือง จากแหล่งข้อมูลที่มีชื่อว่า “Kent Ridge Bio - Medical” อัลกอรึทึมสำหรับการคัดเลือกคุณสมบัติแบบจัดลำดับที่นำมาใช้ในงานวิจัยนี้มีด้วยกัน 5 อัลกอริทึมดังนี้ Information Gain , Gain Ratio , Symmetrical Uncertainty , Relief และ

OneR วิธีการคัดเลือกคุณสมบัติในงานวิจัยนี้ คือดูจากคะแนนของแต่ละคุณสมบัติโดยคะแนนของแต่ละคุณสมบัติในที่นี้คือ จำนวนความถี่ที่แต่ละคุณสมบัติปรากฏในผลลัพธ์ทั้ง 5 อัลกอริทึม โดยในงานวิจัยนี้ได้มีการคิดวิธีสำหรับป้องกันกรณีคุณสมบัติมีความถี่ซ้ำกันด้วยการใช้ลำดับโดยเฉลี่ยของแต่ละคุณสมบัติ จากนั้นทำการสร้างแบบจำลอง โดยใช้กระบวนการจัดหมวดหมู่โดยอัลกอริทึมสำหรับการจัดหมวดหมู่ที่ใช้ในงานวิจัยนี้มีด้วยกัน 5 อัลกอริทึมคือ K – Nearest Neighbor, Naïve Bayes, Support Vector Machines, Random Forest, Logistic Regression , C4.5 ซึ่งจากการทดลองพบว่า วิธีการคัดเลือกคุณสมบัติแบบรวม ทำให้ประสิทธิภาพของแบบจำลองที่สร้างจากกระบวนการจัดหมวดหมู่ดีกว่าวิธีการคัดเลือกคุณสมบัติแบบเดี่ยวสำหรับชุดข้อมูลที่มีหลายมิติ

จากงานวิจัยที่กล่าวมาทั้งหมดสามารถสรุปข้อแตกในส่วนของรูปแบบอัลกอริทึม และการวัดผลเป็นดังตารางที่ 2.1

ตารางที่ 2.1 : ตารางสรุปข้อแตกต่างระหว่างงานวิจัยนี้กับงานวิจัยที่เกี่ยวข้อง

ผู้วิจัย	รูปแบบอัลกอริทึม	Ranker ที่ใช้	การวัดผล	ข้อมูลที่ใช้ทดลอง
Puripat Thongkam (งานวิจัยนี้)	การคัดเลือกคุณสมบัติแบบรวม	Symmetrical Uncertainty , Relief , Information Gain , Gain Ratio และ OneR	เปรียบเทียบค่าของ AUC , Precision , Recall ที่ได้จากแบบจำลองซึ่งสร้างโดยกระบวนการจัดหมวดหมู่	สาเหตุการเกิดมะเร็งปอด , มะเร็งต่อมน้ำเหลือง , มะเร็งเต้านม , มะเร็งรังไข่ , มะเร็งเม็ดเลือดขาว
M.sujatha	การคัดเลือกคุณสมบัติแบบเดี่ยว	Symmetrical	เปรียบเทียบ	สาเหตุการเกิดมะเร็ง

		Uncertainty , Relief , Information Gain , Gain Ratio และ OneR	ค่าของ AUC ที่ได้จากแบบจำลองซึ่งสร้างโดยกระบวนการจัดหมวดหมู่	ปอด
Sri Harsha Vege	การคัดเลือกคุณสมบัติแบบรวม	Symmetrical Uncertainty , Relief , Information Gain , Gain Ratio และ OneR	เปรียบเทียบค่าของ AUC ที่ได้จากแบบจำลองซึ่งสร้างโดยกระบวนการจัดหมวดหมู่	สาเหตุการเกิดมะเร็งปอดและมะเร็งต่อมน้ำเหลือง
Opeyemi Osanaiye	การคัดเลือกคุณสมบัติแบบรวม	Information Gain, Gain Ratio, Chi-Squared และ ReliefF	จำนวนของคุณสมบัติที่เหลือจากการคัดเลือกน้อยที่สุด	บันทึกการโจมตีแบบ DDos (Distributed Denial-of-Service) บนระบบกลุ่มเมฆ (Cloud)

บทที่ 3
วิธีการวิจัย

ในงานวิจัยนี้จะทำการปรับปรุงอัลกอริทึมที่ใช้ใน การคัดเลือกคุณสมบัติแบบรวมจากวิทยานิพนธ์ของ Sri Harsha Vege (1) ซึ่งได้นำเสนอวิธีการคัดเลือกคุณสมบัติแบบรวม โดยใช้การคัดเลือกคุณสมบัติแบบจัดลำดับซึ่งผู้วิจัยได้ทำการศึกษาลักษณะของอัลกอรึทึม ที่ใช้ในการจัดลำดับพบว่าสามารถแก้ไขปรับปรุงให้อัลกอริทึมดีขึ้นได้ด้วยการให้ลำดับความสำคัญ (Priority) แต่ละอัลกอรึทึมที่ใช้ในการจัดลำดับ และเพิ่มน้ำหนักของลำดับให้กับคุณสมบัติที่เป็นผลลัพธ์ของการเรียงให้แตกต่างกันออกไป สำหรับหัวข้อในบทนี้จะกล่าวถึงด้วยกัน 3 ส่วนคือ

3.1 อัลกอริทึมต้นแบบที่นำมาต่อยอด

3.2 อัลกอริทึมที่นำเสนอ

3.3 การวัดผล

3.1 อัลกอริทึมต้นแบบที่นำมาต่อยอด (Based Algorithm)

อัลกอริทึมหลักที่นำมาต่อยอดคือ อัลกอริทึมจากวิจัยของ Sri Harsha Vege (1) ซึ่งในงานวิจัยนี้ได้ใช้ วิธีการคัดเลือกคุณสมบัติแบบรวม ซึ่งวิธีการคัดเลือกคุณสมบัติแบบรวม (Ensemble Feature Selection) แตกต่างจากการคัดเลือกคุณสมบัติแบบเดี่ยว (Individual Feature Selection) ตรงที่วิธีการคัดเลือกคุณสมบัติแบบรวม จะนำเอาเซตผลลัพธ์ของทุกอัลกอริทึม มาทำการประมวลผล ในงานวิจัยนี้สามารถแบ่งได้เป็น 4 ขั้นตอน

3.1.1 ชุดข้อมูล (Data Set)

แหล่งข้อมูลที่งานวิจัยต้นแบบนำมาใช้คือ “Kent Ridge Bio - Medical” ซึ่งข้อมูลที่นำมาวิจัยได้แก่

3.1.1.1 สาเหตุการเกิดโรคมะเร็งปอด (Lung Cancer)

ชุดข้อมูลของสาเหตุการเกิดโรคมะเร็งปอดอยู่ในรูปสกุล .arff ภายในชุดข้อมูลประกอบด้วยจำนวนคุณสมบัติ 57 คุณสมบัติ โดยคุณสมบัติที่เป็นคลาส คือคุณสมบัติลำดับที่ 57 ทุกคุณสมบัติเป็นตัวเลข (Numeric) ค่าของคุณสมบัติทุกค่าเป็นตัวเลขและค่าของคลาสคุณสมบัติประกอบด้วยค่าตัวเลข 3 ค่า และจำนวนกรณี (Instances) 32 กรณี

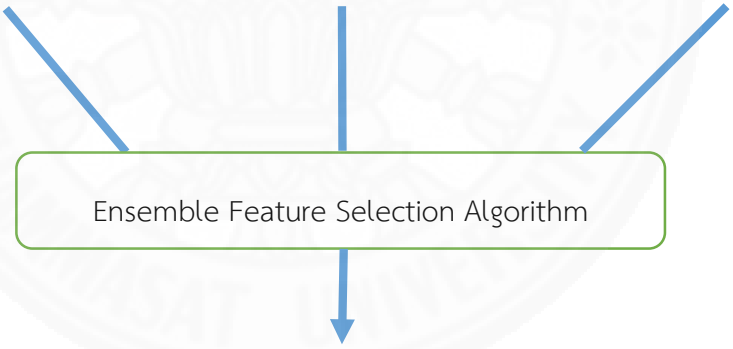
3.1.1.2 สาเหตุการเกิดมะเร็งต่อมน้ำเหลือง (Lymphoma)

ชุดข้อมูลของสาเหตุการเกิดมะเร็งต่อมน้ำเหลือง อยู่ในรูปสกุล .arff ชุดข้อมูลประกอบด้วยจำนวนคุณสมบัติ 4,027 คุณสมบัติ โดยคุณสมบัติที่เป็นคลาสคือคุณสมบัติลำดับที่ 4,027 ทุกคุณสมบัติเป็นตัวอักษร (Nominal) ค่าของคุณสมบัติทุกค่าเป็นตัวอักษร และค่าของคลาสคุณสมบัติประกอบด้วยค่าตัวอักษร 9 ค่า และจำนวนกรณี (Instances) 96 กรณี

3.1.2 การคัดเลือกคุณสมบัติแบบรวม (Ensemble Feature Selection)

การคัดเลือกคุณสมบัติที่ใช้ในงานวิจัยต้นแบบคือ การคัดเลือกคุณสมบัติแบบรวม (Ensemble Feature Selection) โดยการคัดเลือกคุณสมบัติแบบรวม คือการนำผลลัพธ์จากการคัดเลือกคุณสมบัติแบบเดี่ยวหลายๆเซตมาทำการผสมผสานกัน เพื่อให้ได้เซตผลลัพธ์ที่ดีขึ้นเพียงเซตเดียว

Ranker 1 : {$F1_1$,$F2_1$,$F3_1$, ... ,Fn_1} Ranker 2 : {$F1_2$,$F2_2$,$F3_2$, ... ,Fn_2} .....



Final Set : {F1 ,F2 ,F3 ,..... ,Fn }

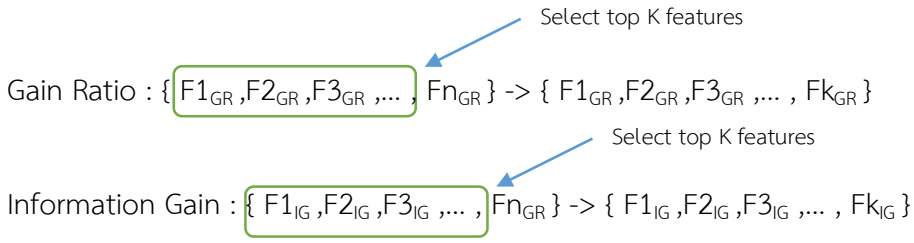
ภาพที่ 3.1 : ลักษณะการทำงานของการคัดเลือกคุณสมบัติแบบรวม

3.1.2.1 อัลกอริทึมต้นแบบ (Based Ensemble Feature Selection Algorithm)

อัลกอริทึมของงานวิจัยต้นแบบ (1) คือใช้การคัดเลือกคุณสมบัติแบบรวมโดยใช้เทคนิคแบบการจัดลำดับซึ่งสามารถแบ่งขั้นตอนการทำงานออกเป็น 3 ขั้นตอนคือ

1. สร้างเซตผลลัพธ์ของแต่ละการคัดเลือกคุณสมบัติ ด้วยเทคนิคแบบการจัดลำดับแบบต่างๆ โดยเทคนิคการจัดลำดับที่นำมาใช้ในงานวิจัยต้นแบบได้แก่ Symmetrical Uncertainty , Relief Information Gain , Gain Ratio และ OneR โดยนำเอาชุดข้อมูลที่มี N คุณสมบัติมาผ่าน

การคัดเลือกคุณสมบัติด้วยเทคนิคเหล่านี้จากนั้นทำการเลือกคุณสมบัติที่ดีที่สุด K คุณสมบัติ (Top K Features) จากผลลัพธ์ที่ได้ในแต่ละเทคนิค



หลังจากได้เซตของคุณสมบัติที่ทำการเลือกมา K คุณสมบัติแล้ว 2. คำนวณค่าความถี่ของแต่ละคุณสมบัติโดยดูจำนวนครั้งในปรากฏของแต่ละคุณสมบัติในทุกเซตที่เลือกมาตัวอย่างเช่นเลือก K = 5 ผลลัพธ์การเลือกคุณสมบัติของแต่ละเทคนิคเป็นดังนี้

Gain Ratio -> { A , B , C , D , E }

Information Gain -> { Z , J , A , C , B }

Symmetrical Uncertainty -> { A , G , B , F , H }

Relief -> { A , G , D , B , Y }

OneR -> {A , B , U , I , R }

นับจำนวนครั้งการปรากฏของคุณสมบัติจากทุกเซต ยกตัวอย่างการนับเช่นพิจารณาคุณสมบัติ A

Gain Ratio -> { A , B , C , D , E }

Information Gain -> { Z , J , A , C , B }

Symmetrical Uncertainty -> { A , G , B , F , H }

Relief -> { A , G , D , B , Y }

OneR -> {A , B , U , I , R }

ดังนั้นค่าของ $A_{frequency}$ = 5 เป็นต้น

ทำการกำหนดค่าความถี่ของแต่ละคุณสมบัติทั้งหมดด้วยวิธีแบบเดียวกันจนครบทุกคุณสมบัติ จากนั้นทำการคำนวณค่าลำดับโดยเฉลี่ยของแต่ละคุณสมบัติเพื่อแก้ปัญหากรณีที่ค่าความถี่ ของคุณสมบัติซ้ำจะใช้ค่าลำดับโดยเฉลี่ยเข้ามาช่วย ในการเรียงลำดับโดยค่าลำดับโดยเฉลี่ยคำนวณ มาจากลำดับของแต่ละคุณสมบัติในทุกเซตผลลัพธ์ที่ยังไม่มีการเลือก K คุณสมบัติซึ่งยกตัวอย่างเช่น จำนวนคุณสมบัติทั้งหมดในชุดข้อมูลหรือ N = 4 และผลลัพธ์การเลือกคุณสมบัติของแต่ละเทคนิค ก่อนจะเลือก K คุณสมบัติเป็นดังนี้

Gain Ratio -> { A , B , C , D , E }

Information Gain -> { B , C , A , D , E }

Symmetrical Uncertainty -> { A , C , B , E , D }

Relief -> { A , C , D , B , E }

OneR -> {A , B , C , E , D }

พิจารณาลำดับโดยเฉลี่ยของ A

Gain Ratio -> { A , B , C , D , E } -> A อยู่ในลำดับที่ 1

Information Gain -> { B , C , A , D , E } -> A อยู่ในลำดับที่ 3

Symmetrical Uncertainty -> { A , C , B , E , D } -> A อยู่ในลำดับที่ 1

Relief -> { A , C , D , B , E } -> A อยู่ในลำดับที่ 1

OneR -> {A , B , C , E , D } -> A อยู่ในลำดับที่ 1

ดังนั้นลำดับโดยเฉลี่ยของ A คือ (1 + 3 + 1 + 1 + 1) / 5 = 7 / 5 = 1.4 เป็นต้น

3. ทำการเรียงลำดับคุณสมบัติโดยใช้ค่าความถี่จากมากไปน้อย หากมีกรณีที่ค่าความถี่ของคุณสมบัติเท่ากันให้ใช้ค่าลำดับโดยเฉลี่ยในการเรียงโดยทำการเรียงจากน้อยไปมาก ซึ่ง 3 ขั้นตอนที่ได้กล่าวไปนั้นสามารถสรุปโดยเขียนเป็นอัลกอริทึมได้ดังนี้

Input: Dataset with N features

Output: Dataset with new order of K features

1: Initialize E as output variable & F[i] for temporary variable

2: For each ranker i //keep all ranker result to F array

3: F[i] = each ranker result

4: EndFor

5: For each feature in list F for 0 – (K-1) time //For get number of apprear and order

6: if(E don’t have F[i][K] add F[i][K] to E and set count = 1)

7: else set count++

8: E(F[i][k].rank) = K

9: EndFor

10: Sort the features in F based on their frequency, if same frequency, sort by mean rank; select the top k features and assign the features to list E

3.1.3 การจำแนกประเภทข้อมูล (Classification)

งานวิจัยต้นแบบได้มีการเอาผลลัพธ์ที่ได้จากการคัดเลือกคุณสมบัติแบบรวม มาทำการสร้างแบบจำลอง โดยใช้กระบวนการจัดหมวดหมู่ อัลกอริทึมสำหรับการจัดหมวดหมู่ ที่ใช้ในงานวิจัยนี้มี ด้วยกัน 5 อัลกอริทึมคือ K – Nearest Neighbor , Naïve Bayes , Support Vector Machines , Random Forest , Logistic Regression , C4.5 เพื่อเปรียบเทียบค่าพื้นที่ใต้ส่วนโค้ง (Area Under Curve: AUC) ของการคัดเลือกคุณสมบัติแบบรวมกับแบบเดี่ยวว่าเทคนิคใด จะให้ค่าประสิทธิภาพของพื้นที่ใต้ส่วนโค้งที่ได้จากแบบจำลองมากกว่ากัน

3.1.4 การวัดผล (Evaluation)

จากการวัดผลโดยทำการทดสอบกับข้อมูล 2 ชุดคือ สาเหตุการเกิดโรคมะเร็งปอด และสาเหตุการเกิดโรคมะเร็งต่อมน้ำเหลือง และค่า K ที่หลากหลายพบว่าการคัดเลือกคุณสมบัติแบบรวมให้ค่า AUC โดยเฉลี่ยซึ่งวัดจากแบบจำลองสูงกว่าแบบเดี่ยว โดยผลลัพธ์กรณีใช้ชุดข้อมูลสาเหตุการเกิดโรคมะเร็งปอดเป็นดังนี้

ตารางที่ 3.1 : ตารางเปรียบเทียบค่า AUC ของแบบจำลองโดยใช้การคัดเลือกคุณสมบัติแบบรวมและแบบเดี่ยวโดย K = 15

Ranker	KNN	C4.5	NB	RF	LR	SVM	Average
GR	0.7302	0.8043	0.8750	0.7772	0.6682	0.7813	0.7727
RFF	0.7061	0.8204	0.8297	0.7770	0.7366	0.7187	0.7647
SU	0.7986	0.7927	0.8693	0.7808	0.6932	0.7295	0.7773
OneR	0.5993	0.8038	0.8234	0.6446	0.5916	0.7157	0.6964
IG	0.7986	0.7927	0.8693	0.7808	0.6932	0.7295	0.7773
Ensemble	0.7504	0.8146	0.8707	0.7967	0.7150	0.7670	0.7857
Based Dataset	0.5970	0.6620	0.7130	0.6630	0.5810	0.6240	0.6400

ตารางที่ 3.2 : ตารางเปรียบเทียบค่า AUC ของแบบจำลองโดยใช้การคัดเลือกคุณสมบัติแบบรวมและแบบเดี่ยวโดย K = 20

Ranker	KNN	C4.5	NB	RF	LR	SVM	Average
GR	0.7583	0.7871	0.8678	0.8098	0.6386	0.7529	0.7690
RFF	0.7991	0.7848	0.8329	0.7636	0.7215	0.7530	0.7758
SU	0.7892	0.7883	0.8552	0.7671	0.7116	0.7676	0.7798
OneR	0.7290	0.7273	0.8288	0.7865	0.6332	0.7553	0.7433
IG	0.7296	0.7759	0.8449	0.7741	0.7770	0.7543	0.7759
Ensemble	0.7755	0.7883	0.8673	0.8070	0.7454	0.7676	0.7918
Base Dataset	0.5970	0.6620	0.7130	0.6630	0.5810	0.6240	0.6400

ผลลัพธ์กรณีใช้ชุดข้อมูลสาเหตุการเกิดโรคมะเร็งต่อมน้ำเหลืองเป็นดังนี้

ตารางที่ 3.3 : ตารางเปรียบเทียบค่า AUC ของแบบจำลองโดยใช้การคัดเลือกคุณสมบัติแบบรวมและแบบเดี่ยวโดย K = 25

Ranker	KNN	C4.5	NB	RF	LR	SVM	Average
GR	0.8714	0.8170	0.9412	0.9374	0.9523	0.8541	0.8955
RFF	0.8522	0.8360	0.9275	0.9723	0.9741	0.9660	0.9213
SU	0.9366	0.9037	0.9216	0.9758	0.9832	0.9276	0.9414
OneR	0.8357	0.8028	0.9045	0.9699	0.9535	0.9185	0.8974
IG	0.9388	0.8665	0.9260	0.9756	0.9755	0.9411	0.9372
Ensemble	0.9419	0.9303	0.9243	0.9828	0.9819	0.9381	0.9498
Base Dataset	0.8600	0.8920	0.7640	0.9640	0.9820	0.9800	0.9070

ตารางที่ 3.4 : ตารางเปรียบเทียบค่า AUC ของแบบจำลองโดยใช้การคัดเลือกคุณสมบัติแบบรวมและแบบเดี่ยวโดย K = 50

Ranker	KNN	C4.5	NB	RF	LR	SVM	Average
GR	0.9096	0.8829	0.9529	0.9705	0.9774	0.9656	0.9431
RFF	0.8707	0.8947	0.9221	0.9790	0.9730	0.9646	0.9340
SU	0.9341	0.9113	0.9235	0.9858	0.9943	0.9693	0.9510
OneR	0.8211	0.8192	0.9167	0.9679	0.9703	0.9548	0.9457
IG	0.9509	0.8830	0.9211	0.9636	0.9878	0.9680	0.9457
Ensemble	0.9579	0.9389	0.9261	0.9747	0.9630	0.9537	0.9523
Base Dataset	0.8600	0.8920	0.7640	0.9640	0.9820	0.9800	0.907

3.2 อัลกอริทึมที่นำเสนอ (Proposed Algorithm)

จากที่ได้ทำการศึกษาอัลกอริทึมของงานวิจัยต้นแบบ พบว่าสามารถปรับปรุงงานอัลกอริทึมที่ใช้สำหรับการคัดเลือกคุณสมบัติแบบรวมของงานวิจัยต้นแบบ ให้มีประสิทธิภาพมากขึ้นได้โดยการเพิ่มลำดับความสำคัญ (Priority) ให้แต่ละอัลกอรึทึมที่ใช้ในการจัดลำดับ เพิ่มน้ำหนักให้กับคุณสมบัติที่เป็นผลลัพธ์ของการเรียงให้แตกต่างกันออกไปนำเอาคะแนนที่ได้จากการจัดลำดับของแต่ละอัลกอริทึมมาคิดคำนวณด้วย เพราะฉะนั้นในงานวิจัยนี้จะเน้นไปที่การปรับปรุงอัลกอริทึมที่ใช้ในการคัดเลือกคุณสมบัติแบบรวม

3.2.1 ปัญหาที่เกิดขึ้นในอัลกอริทึมต้นแบบ (Based Algorithm Issue)

ผู้วิจัยได้ทดสอบอัลกอริทึมที่ใช้ สำหรับการคัดเลือกคุณสมบัติแบบรวมของงานวิจัยต้นแบบโดยทำการทดสอบกับชุดข้อมูล 2 ชุดคือ สาเหตุการเกิดโรคมะเร็งปอดและมะเร็งต่อมน้ำเหลือง โดยทำการเลือกจำนวนคุณสมบัติเพื่อทดสอบที่แตกต่างกัน พบว่ามีกรณีที่ผลลัพธ์ของค่าพื้นที่ใต้ส่วนโค้งจากการคัดเลือกคุณสมบัติแบบรวมต่ำกว่า การคัดเลือกคุณสมบัติแบบเดี่ยว เช่นผลการทดสอบชุดข้อมูลสาเหตุการเกิดโรคมะเร็งต่อมน้ำเหลือง โดยใช้ K = 100

ตารางที่ 3.5 : ตารางเปรียบเทียบค่า AUC ของแบบจำลองโดยใช้การคัดเลือกคุณสมบัติแบบรวมและแบบเดี่ยวโดย K = 100

Ranker	KNN	C4.5	NB	RF	LR	SVM	Average
GR	0.9331	0.9000	0.9273	0.9716	0.9881	0.9769	0.9495
RFF	0.9065	0.9127	0.9174	0.9664	0.9702	0.9632	0.9394
SU	0.9578	0.9083	0.9346	0.9865	0.9968	0.9692	0.9588
OneR	0.8438	0.8748	0.9037	0.9521	0.9800	0.9689	0.9205
IG	0.9643	0.9222	0.9198	0.9771	0.9949	0.9692	0.9579
Ensemble	0.9516	0.9095	0.9336	0.9760	0.9878	0.9638	0.9537
Base Dataset	0.8600	0.8920	0.7640	0.9640	0.9820	0.9800	0.9070

ดังนั้นผู้วิจัยจึงทำการเปรียบเทียบค่าเฉลี่ยของพื้นที่ใต้ส่วนโค้ง (AUC) ของแต่ละเทคนิคที่ใช้ในการเรียงพบว่าได้ผลดังนี้

ตารางที่ 3.6 : ตารางเปรียบเทียบค่า AUC ของแต่ละเทคนิคในการจัดลำดับโดยใช้ชุดข้อมูล สาเหตุการเกิดมะเร็งปอด

	5 Feature	10 Feature	15 Feature	20 Feature
Gain Ratio	0.7407	0.7567	0.7727	0.7690
ReliefF	0.8222	0.7919	0.7647	0.7758
Symmetrical	0.8097	0.7818	0.7773	0.7798
OneR	0.7592	0.7601	0.6964	0.7433
Information Gain	0.8097	0.7705	0.7773	0.7759

ตารางที่ 3.7 : ตารางเปรียบเทียบค่า AUC ของแต่ละเทคนิคในการจัดลำดับโดยใช้ชุดข้อมูล สาเหตุการเกิดมะเร็งต่อมน้ำเหลือง

	25 Feature	50 Feature	100 Feature	500 Feature
Gain Ratio	0.8955	0.9431	0.9495	0.9467
ReliefF	0.9213	0.9340	0.9394	0.9441
Symmetrical	0.9414	0.9510	0.9588	0.9553
OneR	0.8974	0.9083	0.9205	0.9213
Information Gain	0.0372	0.9457	0.9579	0.9550

สังเกตจากตารางเปรียบเทียบพบว่าการจัดลำดับแบบ Symmetrical Uncertainty มีค่า AUC โดยเฉลี่ยสูงสุดจากทั้ง 2 ชุดข้อมูลและรองลงคือ ReliefF , Information Gain , Gain Ratio และ OneR ตามลำดับ ดังนั้นผู้วิจัยจึงคิดว่าหากมีการให้ลำดับความสำคัญ (Priority) แก่ผลลัพธ์ที่ได้จากอัลกอริทึมการจัดลำดับจะทำให้ค่าของ AUC ที่ได้จากการสร้างแบบจำลอง มีแนวโน้มที่ดีขึ้น ผู้วิจัยจึงได้ทำการคิดค่าตัวเลขจำนวน 5 ตัวเลขคือ {1, 2 ,3 ,4 ,5} ซึ่งผู้วิจัยเรียกตัวเลขเหล่านี้ว่าค่าลำดับความสำคัญ โดย 5 เลขนี้มาจากจำนวน Ranker ที่ใช้ทดสอบ โดยผู้วิจัยทำการจับคู่ตัวเลขเหล่านี้ให้กับแต่ละ Ranker โดยหาก Ranker ใดให้ค่า AUC โดยเฉลี่ยสูงก็จะมีค่าลำดับความสำคัญสูงตามไปด้วย หลังจากทำการจับคู่ค่าลำดับความสำคัญจากผลการทดลองแล้วได้ข้อสรุปดังตาราง

ตารางที่ 3.8 : ตารางแสดงค่า Priority ของแต่ละเทคนิค

	Priority
Gain Ratio	2
Information Gain	3
Symmetrical Uncertainty	5
Relief	4
OneR	1

อีกหนึ่งปัญหาของอัลกอริทึมต้นแบบคือ การคำนวณค่าความถี่ของอัลกอริทึมต้นแบบ นั้น อัลกอริทึมต้นแบบ ไม่สนใจว่าคุณสมบัตินั้น จะอยู่ในตำแหน่งใดขอเพียงอยู่ภายในช่วงของ K ก็ถือว่ามีค่าเท่ากันทุกคุณสมบัติซึ่งความจริงแล้วลำดับแต่ละลำดับของคุณสมบัติมีผลต่อ การคำนวณค่า AUC ของแบบจำลองที่แตกต่างกัน ดังนั้นผู้วิจัยจึงได้เพิ่มอัลกอริทึมให้เอาปัจจัยของลำดับเข้ามาเป็นน้ำหนักในการคำนวณด้วย ปัญหาสุดท้ายคือถึงแม้ว่าอัลกอริทึมต้นแบบจะมีการนำเอาค่าของลำดับเฉลี่ยมาแก้ไขกรณีที่เกิดปัญหาค่าความถี่ซ้ำกัน แต่จากผลการทดสอบพบว่ามีบางกรณีที่ค่าของลำดับเฉลี่ยมีการซ้ำกันเกิดขึ้น ซึ่งอัลกอริทึมต้นแบบไม่ได้มีการป้องกันปัญหานี้เอาไว้ ทางผู้วิจัยจึงได้คิดวิธีป้องกันปัญหานี้โดยใช้คะแนนเฉลี่ย (Score) ของแต่ละคุณสมบัติ ที่ได้จากการคำนวณในแต่ละเทคนิคการจัดลำดับ ซึ่งค่าของคะแนนเฉลี่ยนี้เป็นทศนิยม 5 ตำแหน่งจึงลดโอกาสที่จะมีกรณีค่าซ้ำกันเกิดขึ้นอีก

3.2.2 รูปแบบอัลกอริทึม (Algorithm)

การทำงานในขั้นต้นลักษณะเหมือนอัลกอริทึมต้นแบบคือ สร้างเซตผลลัพธ์ของแต่ละการคัดเลือกคุณสมบัติด้วยเทคนิคการจัดลำดับแบบต่างๆ จากนั้นทำการเลือกคุณสมบัติที่ดีที่สุด K คุณสมบัติจากผลลัพธ์ที่ได้ในแต่ละเทคนิคจากนั้นหาค่าความถี่ของแต่ละคุณสมบัติ โดยคำนวณจากจำนวนครั้งในการปรากฏของแต่ละคุณสมบัติและลำดับของคุณสมบัติที่อยู่ในเซต แล้วนำเอาค่าที่ได้ไปคำนวณกับค่าความสำคัญสำหรับแต่ละอัลกอริทึมการจัดลำดับ ตัวอย่างเช่นเลือก K = 5 จากจำนวนคุณสมบัติทั้งหมด 10 คุณสมบัติและผลลัพธ์การเลือกคุณสมบัติของแต่ละเทคนิคเป็นดังนี้

Gain Ratio -> { A , B , C , D , E }

Information Gain -> { Z , J , A , C , B }

Symmetrical Uncertainty -> { A , G , B , F , H }

Relief -> { A , G , D , B , Y }

OneR -> {A , B , U , I , R }

นับจำนวนครั้งการปรากฏของคุณสมบัติจากทุกเซตและคำนวณค่าลำดับกับ ความสำคัญของแต่ละอัลกอริทึมเข้าไปด้วย ยกตัวอย่างการหาค่าความถี่ เช่นพิจารณาคุณสมบัติ B จาก

Gain Ratio -> { A , B , C , D , E } : B อยู่ในลำดับที่ 2

Information Gain -> { Z , J , A , C , B } : B อยู่ในลำดับที่ 5

Symmetrical Uncertainty -> { A , G , B , F , H } : B อยู่ในลำดับที่ 3

Relief -> { A , G , D , M , Y } : B ไม่ได้อยู่ในช่วง K คุณสมบัติ

OneR -> {A , B , U , I , R } : B อยู่ในลำดับที่ 2

ในการคำนวณหาค่าความถี่ของคุณสมบัติจะใช้สมการดังนี้

$$Feature_{frequency} = \frac{\left[\sum_{i=1}^{j}(N-F_r)*priority\right]}{j} \quad (12)$$

โดยค่า N คือจำนวนของคุณสมบัติทั้งหมด

F_r คือลำดับของคุณสมบัติในแต่ละเซต

j คือจำนวนครั้งในการปรากฏ

$priority$ คือค่าลำดับความสำคัญของแต่ละอัลกอริทึม

จากสมการจะสามารถหาค่าความถี่ของคุณสมบัติ B ได้ดังนี้

j = 4 , N = 10

จะได้ว่า $B_{frequency}$ = [(10 - 2) * 2 + (10 - 5) * 3 + (10 - 3) * 5 + (10 - 2) * 1] / 4

= [16 + 15 + 42 + 8] / 4 = 20.25

ทำการกำหนดค่าความถี่ของแต่ละคุณสมบัติทั้งหมดด้วยวิธีแบบเดียวกันจนครบทุกคุณสมบัติ จากนั้นกำหนดค่าคะแนนเฉลี่ยให้กับแต่ละคุณสมบัติ ซึ่งค่านี้ได้มาจากการคำนวณลำดับในแต่ละเซตของอัลกอริทึมจัดลำดับ จากนั้นทำการเรียงลำดับคุณสมบัติโดยเรียงค่าความถี่จากมากไปน้อย หากมีกรณีที่ค่าความถี่ของคุณสมบัติเท่ากันจะใช้ค่าคะแนนเฉลี่ยในการเรียง ซึ่งคะแนนเฉลี่ยที่ใช้จะต้องนำมาทำให้มีค่าอยู่ในช่วง 0 - 1 ก่อนเช่นคะแนนเฉลี่ยของคุณสมบัติหนึ่งที่เกิดจาก OneR คือ 89 ทำให้อยู่ใน ช่วง 0 - 1 คือ 0.89 เป็นต้น จากนั้นทำการเรียงลำดับจากคะแนนเฉลี่ย โดยเรียงจากมากไปน้อย จากขั้นตอนที่ได้กล่าวไปทั้งหมดนั้นสามารถสรุปโดยเขียนเป็นอัลกอริทึมได้ดังนี้

Input: Dataset with N features

Output: Dataset with new order of K features

1: Initialize E as output variable & F[i] for temporary variable , score[i] for score variable ,

2: For each ranker i //keep all ranker result to F array

3: F[i] = each ranker result

4: EndFor

5: For each ranker list F //Assign score of each feature to score array

6: For each feature in list F

7: score[F[i][k]] = (F[i, k].score)

8: EndFor

9: EndFor

10: For each feature in list F for 0 – (K-1) time //For get number of appear and order

11: if (E don't have F[i][K]) add F[i][K] to E and set count = 1

12: else set count ++

13: E[F[i][K]].rank = K

14: E[F[i][K]].score = score[F[i][k]]

15: EndFor

16: For each feature in list E for 0 – (K-1)

17: $$E[k].frequency = \frac{\left[\sum_{i=1}^{E[k].count}(N-E[k].rank)*priority\{IG,GR,SU,OneR,RF\}\right]}{jE[k].rank}$$

18: EndFor

19: If not same frequency, sort the features in E based on their frequency by descending order. Else, Normalize average score to have range 0 - 1 then sort average score by descending order.

20: Select Top K features.

3.3 การวัดผล (Evaluation)

งานวิจัยนี้ทำการวัดผลโดย นำเอาผลลัพธ์จากการคัดเลือกคุณสมบัติแบบรวมมาสร้างแบบจำลอง โดยใช้กระบวนการจัดหมวดหมู่อัลกอริทึม สำหรับการจัดหมวดหมู่ที่ใช้มีด้วยกัน 5 อัลกอริทึมคือ K – Nearest Neighbor , Naïve Bayes , Support Vector Machines , Random Forest , Logistic Regression และ C4.5 จากนั้นทำการเปรียบเทียบประสิทธิภาพของแบบจำลองที่สร้างด้วยการคัดเลือกคุณสมบัติที่ผู้วิจัยได้นำเสนอกับต้นแบบโดยใช้ 3 ค่าในการเปรียบเทียบ ดังนี้

3.3.1 ค่าพื้นที่ใต้ส่วนโค้ง (Area Under Curve : AUC)

ค่าพื้นที่ใต้ส่วนโค้ง คือ ค่าพื้นที่ใต้ส่วนโค้ง ROC ที่แสดงถึงค่าความสัมพันธ์ของคุณสมบัติที่เป็นคลาส (Class) ซึ่งได้จากการคำนวณค่าทำนายผลเชิงบวกจริง (True Positive) และค่าทำนายผลเชิงบวกเท็จ (False Positive) หรือสามารถกล่าวในลักษณะค่าแบบ Confusion Matrix ได้ดังนี้

$$AUC = \frac{True\ Positve}{False\ Positive} \tag{13}$$

3.3.2 ค่าความระลึก (Recall)

ค่าความระลึก คือ ค่าความครบถ้วนที่แสดงให้เห็นถึง ความสามารถของระบบในการเลือกข้อมูลหรือคำตอบที่เกี่ยวข้องได้จำนวนมากน้อยเพียงใด โดยคำนวณจากสมการ

$$Recall = \frac{Corre\ Doc}{All\ Corre\ Doc} \tag{14}$$

โดยค่า $Corre\ Doc$ คือจำนวนข้อมูลหรือคำตอบที่เกี่ยวข้องและถูกเลือกออกมา

$All\ Corre\ Doc$ คือจำนวนข้อมูลหรือคำตอบที่เกี่ยวข้องทั้งหมด

หรือสามารถกล่าวในลักษณะค่าแบบ Confusion Matrix ได้ดังนี้

$$Recall = \frac{True\ Positve}{True\ Positive + False\ Negative} \tag{15}$$

3.3.3 ค่าความถูกต้อง (Precision)

ค่าความถูกต้อง คือ การวัดความสามารถของระบบในการขจัดคำตอบหรือ ข้อมูลที่ไม่เกี่ยวข้องออกไป ถ้าระบบสามารถขจัดคำตอบหรือข้อมูลที่ไม่เกี่ยวข้องออกไปได้มาก แสดงถึงความแม่นยำของระบบสูง การหาค่าความถูกต้องสามารถคำนวณได้จากสมการ

$$Precision = \frac{Corre\ Doc}{All\ Doc} \tag{16}$$

โดยค่า $Corre\ Doc$ คือจำนวนข้อมูลหรือคำตอบที่เกี่ยวข้องและถูกเลือกออกมา

$All\ Doc$ คือจำนวนข้อมูลหรือคำตอบทั้งหมด

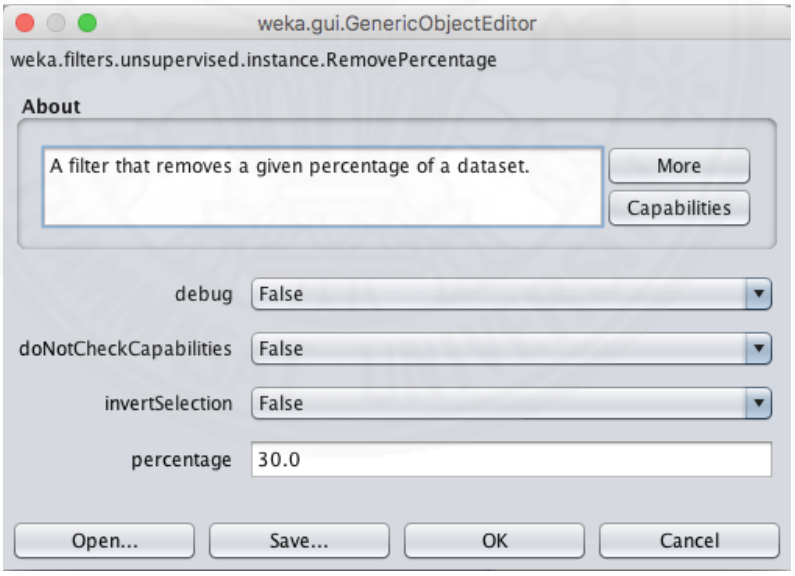
หรือสามารถกล่าวในลักษณะค่าแบบ Confusion Matrix ได้ดังนี้

$$Precision = \frac{True\ Positive}{True\ Positive + False\ Positve} \tag{17}$$

บทที่ 4

ผลการทดลอง

ในการดำเนินการทดลองผู้วิจัยได้นำข้อมูลสำหรับการทดลอง มาผ่านกระบวนการทำความสะอาด (Cleaning) เพื่อกำจัดรายการ (Record) ที่เป็นค่าว่าง (NULL) หรือไม่มีความหมายออก จากนั้นทำการแบ่งจำนวนคุณสมบัติทั้งหมด ที่ผ่านกระบวนการคัดเลือกคุณสมบัติแบบรวมและแบบเดี่ยวมาแล้ว ออกเป็น 4 ส่วนคือ 20% , 40% , 60% และ 80% ซึ่งการแบ่งในลักษณะนี้มีที่มาจากงานวิจัยของ K. Sutha และ J. Temilselvi (9) เรื่อง “ A Review of Feature Selection Algorithms for Data Mining Techniques ” ซึ่งเป็นงานวิจัยที่มีลักษณะเดียวกับงานวิจัยนี้ ผู้วิจัยจึงได้นำเอาวิธีการแบ่งในลักษณะดังกล่าว มาใช้ในการทดสอบ หลังจากทำการแบ่งจำนวนคุณสมบัติแล้ว ผู้วิจัยทำการแบ่งจำนวนรายการ (Instances) ทั้งหมดออกเป็น 2 ส่วน โดยแบ่งในลักษณะเรียงลำดับ (Serialize) ด้วยอัตราส่วน 70% สำหรับชุดฝึก (Training) และ 30% สำหรับชุดทดสอบ (Testing) เพื่อทำการสร้างโมเดลด้วยกระบวนการจัดหมวดหมู่



ภาพที่ 4.1 : หน้าจอการตั้งค่าจำนวนรายการภายในโปรแกรม Weka

และวัดผลค่า AUC, Recall และ Precision เนื่องจากผลที่ได้จากอัลกอริทึมการคัดเลือกคุณสมบัติแบบรวมและแบบเดี่ยวไม่ได้เป็นแบบสุ่ม (Stochastic Data) หากใช้ชุดข้อมูลเดิมในการทดลองหลายครั้งก็จะได้ผลลัพธ์เท่ากันทุกรอบ สำหรับหัวข้อในบทนี้จะกล่าวด้วยกัน 5 ส่วนคือ

4.1 ผลการทดลองสำหรับชุดข้อมูลมะเร็งปอด
4.2 ผลการทดลองสำหรับชุดข้อมูลมะเร็งต่อมน้ำเหลือง
4.3 ผลการทดลองสำหรับชุดข้อมูลมะเร็งเต้านม
4.4 ผลการทดลองสำหรับชุดข้อมูลมะเร็งรังไข่
4.5 ผลการทดลองสำหรับชุดข้อมูลมะเร็งเม็ดเลือดขาว

4.1 ผลการทดลองสำหรับชุดข้อมูลมะเร็งปอด

ในการทดลองโดยใช้ข้อมูลมะเร็งปอดนั้น ผู้วิจัยได้ทำการแบ่งจำนวนคุณสมบัติหลังจากที่ผ่านกระบวนการคัดเลือกคุณสมบัติแล้วทั้งหมด 57 คุณสมบัติออกเป็น 4 ส่วนคือ 20% , 40% , 60% และ 80% เพื่อทำการทดลองดังตารางที่ 4.1

ตารางที่ 4.1 : ตารางแสดงค่าจำนวนคุณสมบัติคิดเป็นเปอร์เซ็นต์สำหรับชุดข้อมูลมะเร็งปอด

20%	40%	60%	80%	ทั้งหมด
11	23	34	47	57

จากนั้นทำการแบ่งจำนวนรายการ (Instances) ทั้งหมดออกเป็น 2 ส่วนคือ ชุดฝึกจำนวน 70% และสำหรับชุดทดสอบจำนวน 30% ซึ่งอัตราส่วนของค่าคลาสคุณสมบัติของชุดข้อมูลมะเร็งปอด ซึ่งเป็นตัวเลข {1, 2, 3} ในส่วนของชุดฝึกและชุดทดสอบหลังจากการแบ่งจำนวนรายการแล้วเป็นตามตาราง ที่ 4.2 และ 4.3

ตารางที่ 4.2 : ตารางแสดงอัตราส่วนของค่าคลาสคุณสมบัติของชุดข้อมูลมะเร็งปอดสำหรับชุดฝึก

ค่าของคุณสมบัติคลาส	อัตราส่วน
1	39%
2	43%
3	18%

ตารางที่ 4.3: ตารางแสดงอัตราส่วนของค่าคลาสคุณสมบัติของชุดข้อมูลมะเร็งปอดสำหรับชุดทดสอบ

ค่าของคุณสมบัติคลาส	อัตราส่วน
1	27%
2	32%
3	41%

ผลการทดลองหลังจากการแบ่งจำนวนคุณสมบัติออกเป็น 4 ส่วนเพื่อทำการหาค่า AUC , Recall , Precision เป็นดังนี้

ตารางที่ 4.4 : ตารางแสดงค่า AUC ของแบบจำลองที่สร้างจากจำนวนคุณสมบัติ 20% สำหรับชุดข้อมูลมะเร็งปอด

Ranker	K-NN	C4.5	NB	RF	LR	SVM	Average
GR	0.757	0.764	0.851	0.789	0.718	0.700	0.763
RF	0.799	0.773	0.864	0.836	0.735	0.715	0.787
SU	0.777	0.768	0.855	0.823	0.684	0.761	0.778
OneR	0.784	0.79	0.830	0.810	0.653	0.748	0.769
IG	0.798	0.752	0.837	0.803	0.629	0.776	0.766
Proposed Ensemble	0.777	0.768	0.855	0.823	0.684	0.761	0.778
Based Ensemble	0.777	0.768	0.855	0.823	0.684	0.761	0.778
Original Dataset	0.597	0.662	0.713	0.644	0.581	0.624	0.637

ตารางที่ 4.5 : ตารางแสดงค่า AUC ของแบบจำลองที่สร้างจากจำนวนคุณสมบัติ 40% สำหรับชุดข้อมูลมะเร็งปอด

Ranker	K-NN	C4.5	NB	RF	LR	SVM	Average
GR	0.758	0.787	0.868	0.813	0.639	0.753	0.770
RF	0.7991	0.785	0.833	0.79	0.722	0.753	0.780
SU	0.7892	0.788	0.855	0.809	0.712	0.768	0.787
OneR	0.729	0.727	0.829	0.771	0.633	0.75β5	0.741
IG	0.7296	0.776	0.845	0.786	0.777	0.754	0.778
Proposed Ensemble	0.798	0.794	0.871	0.843	0.704	0.786	0.799
Based Ensemble	0.776	0.788	0.867	0.803	0.745	0.768	0.791
Original Dataset	0.597	0.662	0.713	0.644	0.581	0.624	0.637

ตารางที่ 4.6 : ตารางแสดงค่า AUC ของแบบจำลองที่สร้างจากจำนวนคุณสมบัติ 60% สำหรับชุดข้อมูลมะเร็งปอด

Ranker	K-NN	C4.5	NB	RF	LR	SVM	Average
GR	0.770	0.788	0.854	0.809	0.682	0.730	0.772
RF	0.768	0.785	0.825	0.764	0.692	0.754	0.765
SU	0.745	0.776	0.811	0.789	0.749	0.733	0.767
OneR	0.610	0.691	0.797	0.776	0.622	0.687	0.697
IG	0.745	0.768	0.809	0.755	0.627	0.729	0.739
Proposed Ensemble	0.756	0.788	0.813	0.811	0.738	0.753	0.777
Based Ensemble	0.692	0.788	0.830	0.809	0.715	0.747	0.764
Original Dataset	0.597	0.662	0.713	0.644	0.581	0.624	0.637

ตารางที่ 4.7 : ตารางแสดงค่า AUC ของแบบจำลองที่สร้างจากจำนวนคุณสมบัติ 80% สำหรับชุดข้อมูลมะเร็งปอด

Ranker	K-NN	C4.5	NB	RF	LR	SVM	Average
GR	0.766	0.785	0.810	0.773	0.647	0.754	0.756
RF	0.75	0.727	0.812	0.737	0.661	0.699	0.731
SU	0.824	0.785	0.805	0.776	0.671	0.754	0.769
OneR	0.625	0.676	0.782	0.784	0.635	0.656	0.693
IG	0.709	0.761	0.810	0.745	0.675	0.712	0.735
Proposed Ensemble	0.833	0.796	0.794	0.781	0.683	0.756	0.774
Based Ensemble	0.684	0.76	0.804	0.754	0.661	0.74	0.734
Original Dataset	0.597	0.662	0.713	0.644	0.581	0.624	0.637

ตารางที่ 4.8 : ตารางแสดงค่า Recall ของแบบจำลองที่สร้างจากจำนวนคุณสมบัติ 20% สำหรับชุดข้อมูลมะเร็งปอด

Ranker	K-NN	C4.5	NB	RF	LR	SVM	Average
GR	0.594	0.563	0.750	0.656	0.594	0.531	0.615
RF	0.625	0.656	0.750	0.750	0.531	0.594	0.651
SU	0.563	0.656	0.750	0.718	0.594	0.656	0.656
OneR	0.687	0.688	0.719	0.688	0.563	0.625	0.662
IG	0.656	0.656	0.750	0.688	0.500	0.625	0.646
Proposed Ensemble	0.563	0.656	0.750	0.750	0.438	0.656	0.636
Based Ensemble	0.563	0.656	0.750	0.781	0.594	0.656	0.667
Original Dataset	0.457	0.406	0.373	0.449	0.450	0.471	0.434

ตารางที่ 4.9 : ตารางแสดงค่า Recall ของแบบจำลองที่สร้างจากจำนวนคุณสมบัติ 40% สำหรับชุดข้อมูลมะเร็งปอด

Ranker	K-NN	C4.5	NB	RF	LR	SVM	Average
GR	0.563	0.625	0.750	0.719	0.500	0.688	0.641
RF	0.656	0.625	0.750	0.688	0.594	0.625	0.656
SU	0.625	0.625	0.750	0.656	0.531	0.656	0.641
OneR	0.469	0.563	0.719	0.625	0.500	0.625	0.584
IG	0.594	0.625	0.781	0.688	0.594	0.625	0.651
Proposed Ensemble	0.655	0.625	0.750	0.688	0.594	0.656	0.661
Based Ensemble	0.625	0.625	0.750	0.688	0.625	0.625	0.656
Original Dataset	0.457	0.406	0.373	0.449	0.450	0.471	0.434

ตารางที่ 4.10 : ตารางแสดงค่า Recall ของแบบจำลองที่สร้างจากจำนวนคุณสมบัติ 60% สำหรับชุดข้อมูลมะเร็งปอด

Ranker	K-NN	C4.5	NB	RF	LR	SVM	Average
GR	0.625	0.625	0.688	0.688	0.563	0.594	0.631
RF	0.563	0.625	0.750	0.656	0.563	0.625	0.630
SU	0.531	0.625	0.719	0.656	0.531	0.625	0.615
OneR	0.500	0.563	0.688	0.656	0.438	0.563	0.568
IG	0.594	0.625	0.750	0.719	0.500	0.594	0.630
Proposed Ensemble	0.620	0.625	0.719	0.688	0.531	0.625	0.635
Based Ensemble	0.531	0.625	0.750	0.688	0.531	0.656	0.630
Original Dataset	0.457	0.406	0.373	0.449	0.450	0.471	0.434

ตารางที่ 4.11 : ตารางแสดงค่า Recall ของแบบจำลองที่สร้างจากจำนวนคุณสมบัติ 80% สำหรับชุดข้อมูลมะเร็งปอด

Ranker	K-NN	C4.5	NB	RF	LR	SVM	Average
GR	0.563	0.625	0.719	0.719	0.438	0.625	0.615
RF	0.531	0.594	0.750	0.656	0.594	0.563	0.615
SU	0.594	0.625	0.719	0.625	0.469	0.625	0.610
OneR	0.438	0.500	0.656	0.688	0.531	0.531	0.557
IG	0.563	0.656	0.688	0.625	0.531	0.563	0.604
Proposed Ensemble	0.594	0.625	0.719	0.656	0.469	0.625	0.615
Based Ensemble	0.500	0.594	0.750	0.625	0.469	0.625	0.594
Original Dataset	0.457	0.406	0.373	0.449	0.450	0.471	0.434

ตารางที่ 4.12 : ตารางแสดงค่า Precision ของแบบจำลองที่สร้างจากจำนวนคุณสมบัติ 20% สำหรับชุดข้อมูลมะเร็งปอด

Ranker	K-NN	C4.5	NB	RF	LR	SVM	Average
GR	0.599	0.614	0.759	0.660	0.590	0.550	0.629
RF	0.629	0.661	0.750	0.763	0.531	0.594	0.655
SU	0.638	0.661	0.761	0.731	0.607	0.650	0.675
OneR	0.708	0.709	0.731	0.692	0.565	0.621	0.671
IG	0.678	0.650	0.761	0.718	0.528	0.635	0.662
Proposed Ensemble	0.638	0.661	0.761	0.749	0.607	0.650	0.678
Based Ensemble	0.638	0.661	0.761	0.789	0.607	0.650	0.684
Original Dataset	0.532	0.438	0.359	0.442	0.443	0.589	0.467

ตารางที่ 4.13 : ตารางแสดงค่า Precision ของแบบจำลองที่สร้างจากจำนวนคุณสมบัติ 40% สำหรับชุดข้อมูลมะเร็งปอด

Ranker	K-NN	C4.5	NB	RF	LR	SVM	Average
GR	0.602	0.650	0.781	0.732	0.497	0.682	0.657
RF	0.703	0.628	0.761	0.702	0.605	0.618	0.670
SU	0.658	0.634	0.781	0.679	0.518	0.662	0.655
OneR	0.490	0.561	0.738	0.653	0.524	0.625	0.599
IG	0.613	0.634	0.805	0.708	0.590	0.624	0.662
Proposed Ensemble	0.658	0.634	0.732	0.708	0.605	0.662	0.667
Based Ensemble	0.658	0.634	0.761	0.702	0.605	0.662	0.670
Original Dataset	0.532	0.438	0.359	0.442	0.443	0.589	0.467

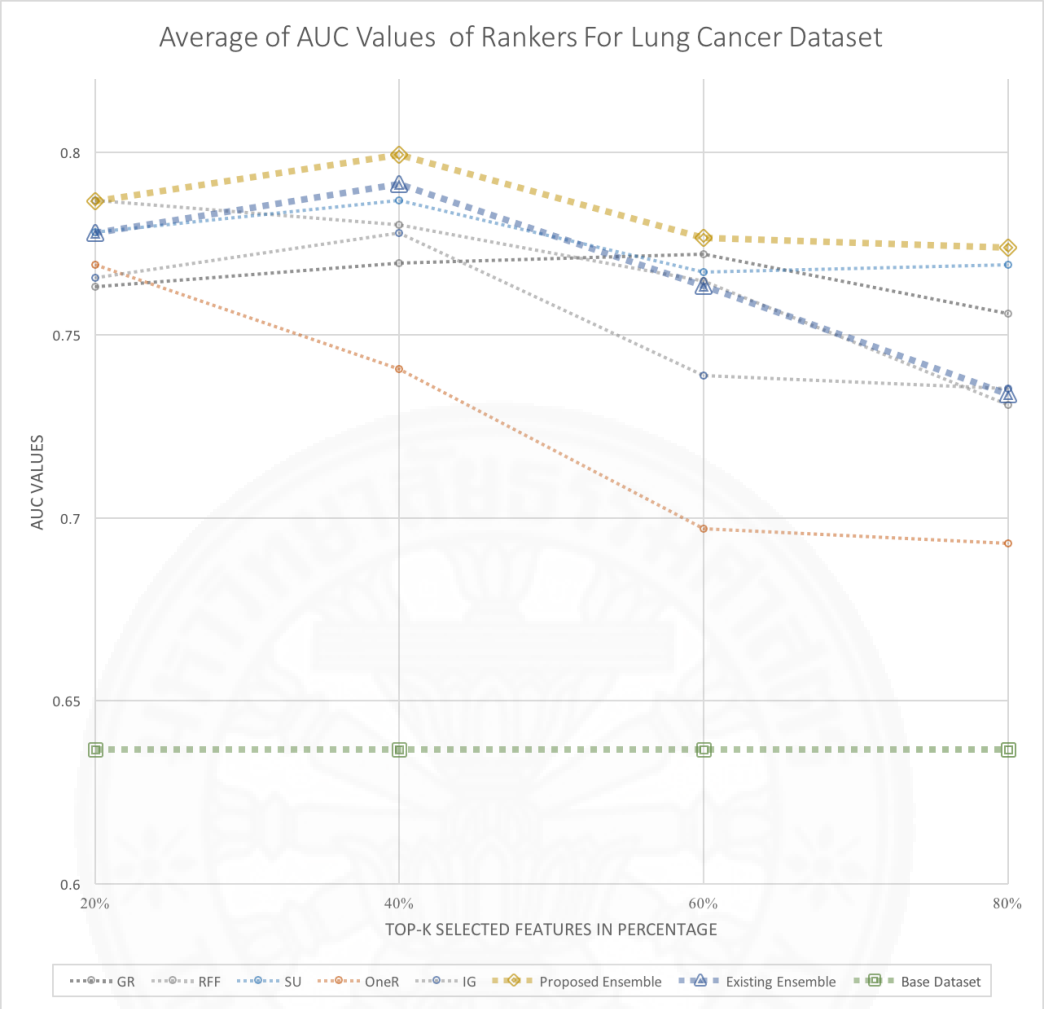
ตารางที่ 4.14 : ตารางแสดงค่า Precision ของแบบจำลองที่สร้างจากจำนวนคุณสมบัติ 60% สำหรับชุดข้อมูลมะเร็งปอด

Ranker	K-NN	C4.5	NB	RF	LR	SVM	Average
GR	0.658	0.634	0.736	0.702	0.576	0.589	0.649
RF	0.594	0.628	0.761	0.679	0.563	0.618	0.641
SU	0.540	0.634	0.732	0.679	0.530	0.613	0.621
OneR	0.539	0.578	0.705	0.684	0.423	0.563	0.582
IG	0.626	0.634	0.761	0.737	0.510	0.609	0.646
Proposed Ensemble	0.591	0.634	0.761	0.724	0.581	0.613	0.651
Based Ensemble	0.513	0.634	0.761	0.702	0.529	0.649	0.631
Original Dataset	0.532	0.438	0.359	0.442	0.443	0.589	0.467

ตารางที่ 4.15 : ตารางแสดงค่า Precision ของแบบจำลองที่สร้างจากจำนวนคุณสมบัติ 80% สำหรับชุดข้อมูลมะเร็งปอด

Ranker	K-NN	C4.5	NB	RF	LR	SVM	Average
GR	0.621	0.628	0.735	0.731	0.443	0.618	0.629
RF	0.533	0.596	0.773	0.679	0.596	0.555	0.622
SU	0.622	0.628	0.756	0.629	0.450	0.618	0.617
OneR	0.446	0.524	0.684	0.699	0.534	0.518	0.568
IG	0.570	0.663	0.702	0.651	0.555	0.563	0.617
Proposed Ensemble	0.622	0.628	0.773	0.712	0.450	0.618	0.634
Based Ensemble	0.473	0.596	0.781	0.632	0.464	0.618	0.594
Original Dataset	0.532	0.438	0.359	0.442	0.443	0.589	0.467

จากตารางผลลัพธ์สามารถสรุปเป็นกราฟโดยใช้ค่าเฉลี่ยของ AUC, Recall และ Precision ของแต่ละเทคนิคได้ดังนี้



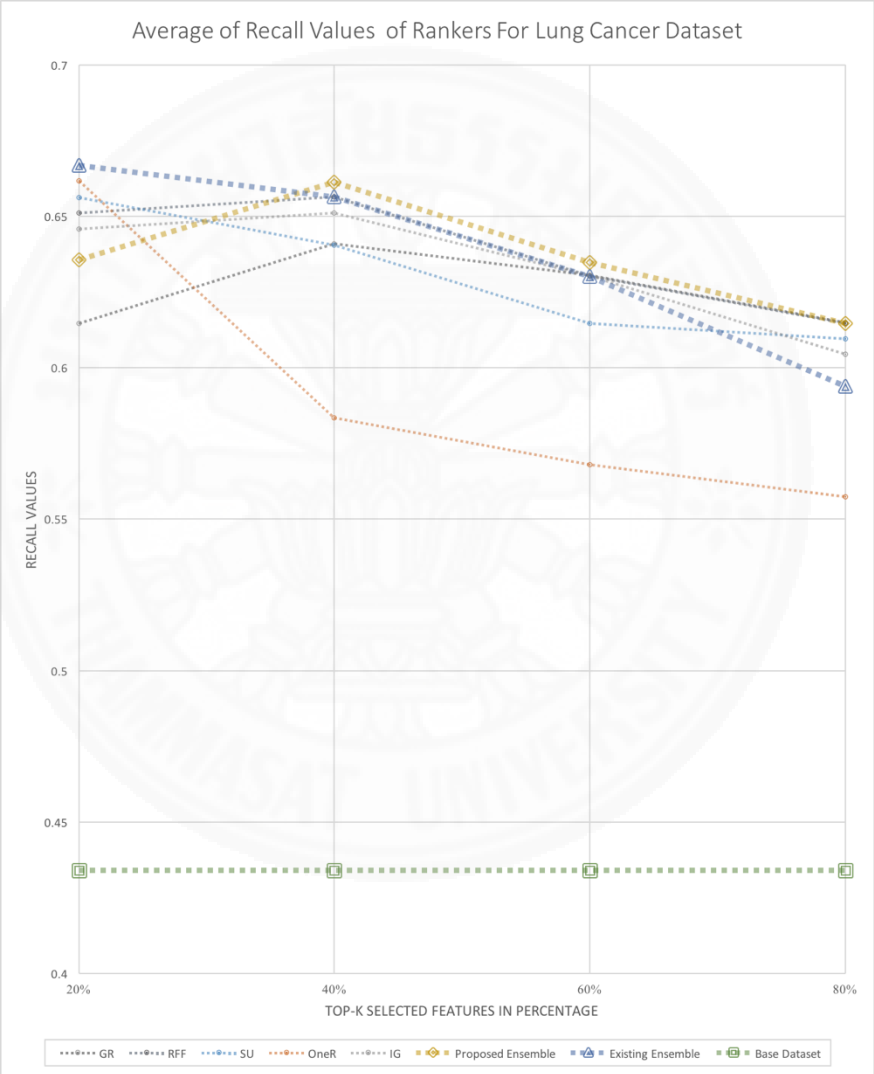
ภาพที่ 4.2 : กราฟแสดงค่า **AUC** โดยเฉลี่ยสำหรับชุดข้อมูลมะเร็งปอด

ผลลัพธ์จากกราฟค่า AUC โดยเฉลี่ยนั้น จะเห็นว่าหากนำข้อมูลมาผ่านอัลกอริทึมที่ใช้ในการคัดเลือกคุณสมบัติไม่ว่าจะเป็นแบบรวมหรือแบบเดี่ยวจะให้ผลลัพธ์ที่ดีกว่าแบบ Base Dataset ในทุกจำนวนคุณสมบัติ จำนวนคุณสมบัติที่แต่ละอัลกอริทึมมีค่า AUC โดยเฉลี่ยใกล้เคียงกันคือ 40% และ อัลกอริทึมที่มีแนวโน้มมีประสิทธิภาพดีที่สุด โดยพิจารณาจากทุกจำนวนคุณสมบัติ ที่ทำการทดสอบคือ อัลกอริทึมที่ผู้วิจัยนำเสนอ

จากกราฟค่า AUC โดยเฉลี่ยพบว่าอัลกอริทึมที่ผู้วิจัยนำเสนอมีประสิทธิภาพดีที่สุดในจำนวนคุณสมบัติ 40%, 60%, 80% โดยทำค่า AUC ได้สูงสุดที่จำนวนคุณสมบัติ 40% ซึ่งมีค่า AUC โดยเฉลี่ยเท่ากับ 0.799 สำหรับจำนวนคุณสมบัติ 20% อัลกอริทึมที่มีค่า AUC โดยเฉลี่ยสูงที่สุด คือ ReliefF ซึ่งมีค่า AUC โดยเฉลี่ยเท่ากับ 0.7868 ในขณะที่อัลกอริทึมที่นำเสนอมีค่า AUC โดยเฉลี่ยเท่ากับ 0.7865

สำหรับอัลกอริทึมต้นแบบมีค่า AUC โดยเฉลี่ยต่ำกว่าอัลกอริทึมที่นำเสนอในทุกๆจำนวนคุณสมบัติ โดยมีค่า AUC โดยเฉลี่ยสูงสุดที่จำนวนคุณสมบัติ 40% ซึ่งมีค่า AUC โดยเฉลี่ยเท่ากับ 0.791 อัลกอริทึมที่พบว่ามีค่า AUC โดยเฉลี่ยต่ำที่สุดในทุกจำนวนคุณสมบัติคือ OneR

สำหรับค่า Recall โดยเฉลี่ยของแต่ละเทคนิคสามารถสรุปเป็นกราฟได้ดังนี้



ภาพที่ 4.3 : กราฟแสดงค่า **Recall** โดยเฉลี่ยสำหรับชุดข้อมูลมะเร็งปอด

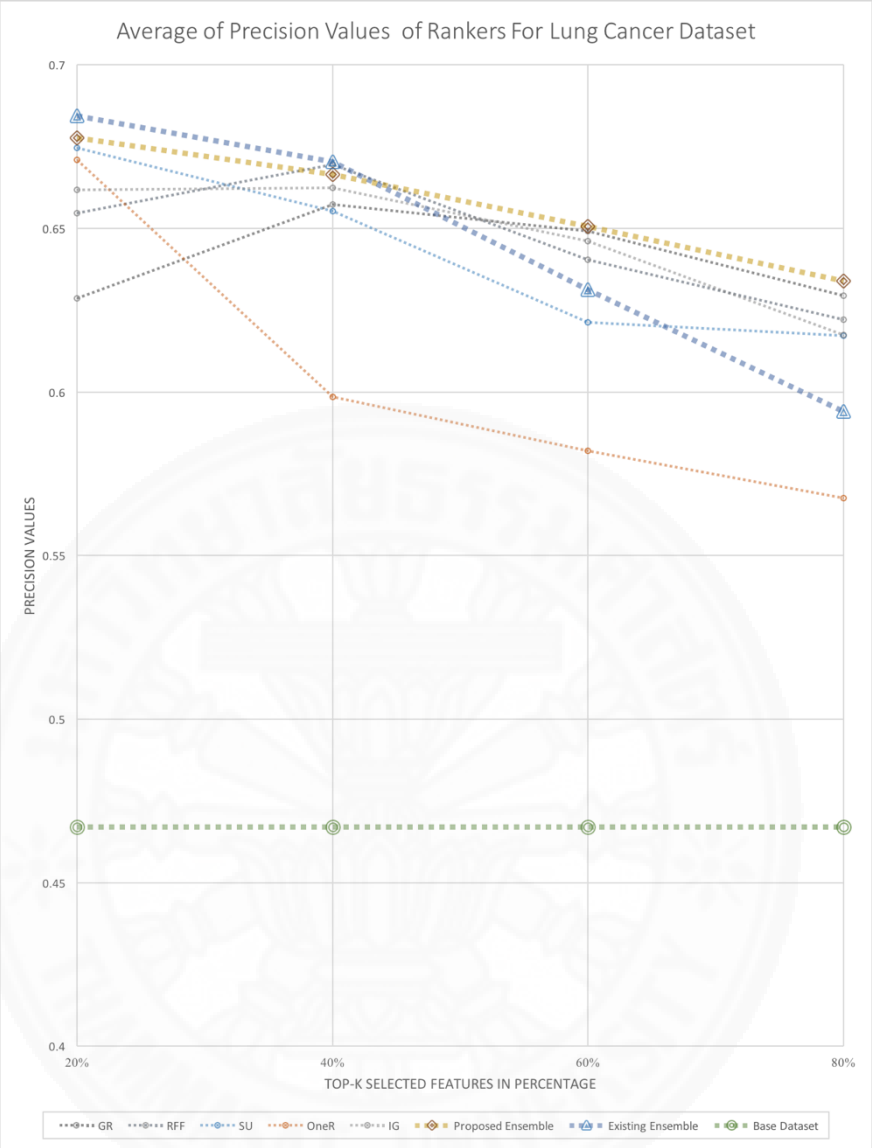
ผลลัพธ์จากกราฟค่า Recall โดยเฉลี่ยนั้น จะเห็นว่าหากนำข้อมูลมาผ่านอัลกอริทึมที่ใช้ในการคัดเลือกคุณสมบัติไม่ว่าจะเป็นแบบรวมหรือแบบเดี่ยวจะให้ผลลัพธ์ที่ดีกว่าแบบ Base Dataset ในทุกจำนวนคุณสมบัติเช่นกัน จำนวนคุณสมบัติที่แต่ละอัลกอริทึมมีค่า Recall โดยเฉลี่ย ใกล้เคียงกันคือ

20% และอัลกอริทึมที่มีแนวโน้มมีประสิทธิภาพดีที่สุดโดยพิจารณาจากทุกจำนวนคุณสมบัติที่ทำการทดสอบคือ อัลกอริทึมที่ผู้วิจัยนำเสนอ

จากกราฟค่า Recall โดยเฉลี่ยพบว่าอัลกอริทึมที่ผู้วิจัยนำเสนอมีประสิทธิภาพดีที่สุด ในจำนวนคุณสมบัติ 40%, 60% และ 80% โดยทำค่าได้สูงสุดที่จำนวนคุณสมบัติ 40% ซึ่งมีค่า Recall โดยเฉลี่ยเท่ากับ 0.661 สำหรับอัลกอริทึมที่มีค่า Recall โดยเฉลี่ยสูงที่สุด ที่จำนวนคุณสมบัติ 20% คือ อัลกอริทึมต้นแบบซึ่งมีค่า Recall โดยเฉลี่ยเท่ากับ 0.666 ในขณะที่อัลกอริทึมที่นำเสนอมีค่า Recall โดยเฉลี่ยที่จำนวนคุณสมบัติ 20% คือ 0.635

สำหรับอัลกอริทึมต้นแบบมีค่า Recall โดยเฉลี่ยต่ำกว่าอัลกอริทึมที่นำเสนอ ที่จำนวนคุณสมบัติ 40%, 60% และ 80% โดยมีค่า Recall โดยเฉลี่ยสูงสุดที่จำนวนคุณสมบัติ 20% ซึ่งมีค่า Recall โดยเฉลี่ยเท่ากับ 0.666 อัลกอริทึมที่พบว่ามีค่า Recall โดยเฉลี่ยต่ำที่สุดคือ OneR

สำหรับค่า Precision โดยเฉลี่ยของแต่ละเทคนิคสามารถสรุปเป็นกราฟได้ดังนี้



ภาพที่ 4.4 : กราฟแสดงค่า **Precision** โดยเฉลี่ยสำหรับชุดข้อมูลมะเร็งปอด

ผลลัพธ์จากกราฟค่า Precision โดยเฉลี่ยนั้น จะเห็นว่าหากนำข้อมูลมาผ่านอัลกอริทึมที่ใช้ในการคัดเลือกคุณสมบัติไม่ว่าจะเป็นแบบรวมหรือแบบเดี่ยวจะให้ผลลัพธ์ที่ดีกว่าแบบ Base Dataset ในทุกจำนวนคุณสมบัติเช่นกัน สำหรับค่า Precision นั้นพบว่า แต่ละอัลกอริทึมมีค่าใกล้เคียงกันในทุกจำนวนคุณสมบัติ แต่จะเริ่มมีความแตกต่างกันเล็กน้อย ตั้งแต่จำนวนคุณสมบัติ 60% และอัลกอริทึมที่มีแนวโน้มมีประสิทธิภาพดีที่สุด โดยพิจารณาจากทุกจำนวนคุณสมบัติที่ทำการทดสอบ คือ อัลกอริทึมที่ผู้วิจัยนำเสนอ

จากกราฟค่า Precision โดยเฉลี่ยพบว่าอัลกอริทึมที่ผู้วิจัยนำเสนอมีประสิทธิภาพดีที่สุดในจำนวนคุณสมบัติ 60% และ 80% โดยทำค่า Precision ได้สูงสุดที่จำนวนคุณสมบัติ 60% ซึ่งมีค่า Precision โดยเฉลี่ยเท่ากับ 0.650 สำหรับอัลกอริทึมที่มีค่า Precision โดยเฉลี่ยสูงที่สุดที่จำนวนคุณสมบัติ 20% และ 40% คือ อัลกอริทึมต้นแบบ ซึ่งมีค่า Precision โดยเฉลี่ยเท่ากับ 0.684 และ 0.670 ในขณะที่อัลกอริทึมที่นำเสนอมีค่า Precision โดยเฉลี่ยที่จำนวนคุณสมบัติ 20% และ 40% คือ 0.678 และ 0.667

สำหรับอัลกอริทึมต้นแบบมีค่า Precision โดยเฉลี่ยต่ำกว่าอัลกอริทึมที่นำเสนอ ที่จำนวนคุณสมบัติ 60% และ 80% โดยมีค่า Precision โดยเฉลี่ยสูงสุดที่จำนวนคุณสมบัติ 20% ซึ่งมีค่า Precision โดยเฉลี่ยเท่ากับ 0.678 อัลกอริทึมที่พบว่ามีค่า Precision โดยเฉลี่ยต่ำที่สุดคือ OneR

4.2 ผลการทดลองสำหรับชุดข้อมูลมะเร็งต่อมน้ำเหลือง

ในการทดลองโดยใช้ข้อมูลมะเร็งต่อมน้ำเหลือง ผู้วิจัยได้ทำการแบ่งจำนวนคุณสมบัติหลังจากที่ผ่านกระบวนการคัดเลือกคุณสมบัติแล้วทั้งหมด 4,027 คุณสมบัติออกเป็น 4 ส่วนคือ 20% ,40% ,60% และ 80% เพื่อทำการทดลองดังตารางที่ 4.16

ตารางที่ 4.16 : ตารางแสดงค่าจำนวนคุณสมบัติคิดเป็นเปอร์เซ็นต์สำหรับชุดข้อมูลมะเร็งต่อมน้ำเหลือง

20%	40%	60%	80%	ทั้งหมด
805	1,611	2,416	3,222	4,027

จากนั้นทำการแบ่งจำนวนรายการ (Instances) ทั้งหมดออกเป็น 2 ส่วนคือ ชุดฝึกจำนวน 70% และสำหรับชุดทดสอบจำนวน 30% ซึ่งอัตราส่วนของค่าคลาสคุณสมบัติของชุดข้อมูลมะเร็งต่อมน้ำเหลืองซึ่งเป็นตัวอักษร {DLBCL, GCD, NIL, ABB, RAT, TCL, FL, RBB, CLL} ในส่วนของชุดฝึกและชุดทดสอบหลังจากการแบ่งจำนวนรายการแล้วเป็นตามตารางที่ 4.17 และ 4.18

ตารางที่ 4.17 : ตารางแสดงอัตราส่วนของค่าคลาสคุณสมบัติของชุดข้อมูลมะเร็งต่อมน้ำเหลืองสำหรับชุดฝึก

ค่าของคุณสมบัติคลาส	อัตราส่วน	ค่าของคุณสมบัติคลาส	อัตราส่วน
DLBCL	12%	FL	8%
GCD	9%	RBB	12%
NIL	16%	CLL	14%
ABB	13%		
RAT	6%		
TCL	10%		

ตารางที่ 4.18 : ตารางแสดงอัตราส่วนของค่าคลาสคุณสมบัติของชุดมะเร็งต่อมน้ำเหลืองสำหรับชุดทดสอบ

ค่าของคุณสมบัติคลาส	อัตราส่วน	ค่าของคุณสมบัติคลาส	อัตราส่วน
DLBCL	4%	FL	9%
GCD	22%	RBB	16%
NIL	14%	CLL	12%
ABB	7%		
RAT	11%		
TCL	5%		

ผลการทดลองหลังจากทำการแบ่งจำนวนคุณสมบัติออกเป็น 4 ส่วนเป็นดังนี้

ตารางที่ 4.19 : ตารางแสดงค่า AUC ของแบบจำลองที่สร้างจากจำนวนคุณสมบัติ 20% สำหรับชุดข้อมูลมะเร็งต่อมน้ำเหลือง

Ranker	K-NN	C4.5	NB	RF	LR	SVM	Average
GR	0.871	0.818	0.941	0.971	0.952	0.854	0.901
RF	0.852	0.840	0.928	0.983	0.974	0.966	0.924
SU	0.937	0.897	0.922	0.979	0.983	0.928	0.941
OneR	0.814	0.892	0.927	0.969	0.949	0.911	0.910
IG	0.939	0.885	0.926	0.981	0.976	0.942	0.942
Proposed Ensemble	0.958	0.902	0.923	0.980	0.985	0.930	0.946
Based Ensemble	0.942	0.923	0.922	0.968	0.982	0.938	0.946
Original Dataset	0.860	0.892	0.764	0.816	0.982	0.980	0.882

ตารางที่ 4.20 : ตารางแสดงค่า AUC ของแบบจำลองที่สร้างจากจำนวนคุณสมบัติ 40% สำหรับชุดข้อมูลมะเร็งต่อมน้ำเหลือง

Ranker	K-NN	C4.5	NB	RF	LR	SVM	Average
GR	0.906	0.886	0.916	0.979	0.992	0.968	0.941
RF	0.887	0.885	0.907	0.984	0.983	0.974	0.937
SU	0.964	0.886	0.916	0.988	0.996	0.978	0.955
OneR	0.861	0.897	0.899	0.978	0.998	0.968	0.934
IG	0.964	0.897	0.921	0.987	0.992	0.979	0.957
Proposed Ensemble	0.967	0.897	0.925	0.988	0.997	0.976	0.958
Based Ensemble	0.951	0.886	0.921	0.986	0.997	0.976	0.953
Original Dataset	0.860	0.892	0.764	0.816	0.982	0.980	0.882

ตารางที่ 4.21 : ตารางแสดงค่า AUC ของแบบจำลองที่สร้างจากจำนวนคุณสมบัติ 60% สำหรับชุดข้อมูลมะเร็งต่อมน้ำเหลือง

Ranker	K-NN	C4.5	NB	RF	LR	SVM	Average
GR	0.935	0.909	0.913	0.985	0.995	0.975	0.952
RF	0.933	0.927	0.897	0.985	0.997	0.978	0.953
SU	0.964	0.880	0.916	0.987	0.996	0.981	0.954
OneR	0.851	0.850	0.892	0.981	0.993	0.959	0.921
IG	0.967	0.897	0.907	0.989	0.993	0.977	0.955
Proposed Ensemble	0.964	0.892	0.922	0.984	0.998	0.982	0.957
Based Ensemble	0.931	0.881	0.907	0.991	0.997	0.980	0.948
Original Dataset	0.860	0.892	0.764	0.816	0.982	0.980	0.882

ตารางที่ 4.22 : ตารางแสดงค่า AUC ของแบบจำลองที่สร้างจากจำนวนคุณสมบัติ 80% สำหรับชุดข้อมูลมะเร็งต่อมน้ำเหลือง

Ranker	K-NN	C4.5	NB	RF	LR	SVM	Average
GR	0.964	0.909	0.907	0.984	0.995	0.975	0.956
RF	0.930	0.920	0.901	0.983	0.994	0.978	0.951
SU	0.970	0.880	0.907	0.983	0.991	0.982	0.952
OneR	0.851	0.874	0.873	0.977	0.986	0.960	0.920
IG	0.977	0.897	0.907	0.989	0.994	0.982	0.958
Proposed Ensemble	0.970	0.920	0.914	0.985	0.996	0.989	0.962
Based Ensemble	0.944	0.868	0.902	0.984	0.992	0.983	0.946
Original Dataset	0.860	0.892	0.764	0.816	0.982	0.980	0.882

ตารางที่ 4.23 : ตารางแสดงค่า Recall ของแบบจำลองที่สร้างจากจำนวนคุณสมบัติ 20% สำหรับชุด ข้อมูลมะเร็งต่อมน้ำเหลือง

Ranker	K-NN	C4.5	NB	RF	LR	SVM	Average
GR	0.833	0.729	0.833	0.802	0.792	0.781	0.795
RF	0.781	0.750	0.802	0.844	0.823	0.896	0.816
SU	0.896	0.813	0.833	0.917	0.865	0.823	0.858
OneR	0.740	0.760	0.750	0.833	0.760	0.823	0.778
IG	0.906	0.813	0.833	0.885	0.844	0.875	0.859
Proposed Ensemble	0.927	0.813	0.833	0.906	0.854	0.823	0.859
Based Ensemble	0.906	0.844	0.823	0.906	0.854	0.854	0.865
Original Dataset	0.458	0.572	0.543	0.537	0.541	0.459	0.518

ตารางที่ 4.24 : ตารางแสดงค่า Recall ของแบบจำลองที่สร้างจากจำนวนคุณสมบัติ 40% สำหรับชุดข้อมูลมะเร็งต่อมน้ำเหลือง

Ranker	K-NN	C4.5	NB	RF	LR	SVM	Average
GR	0.885	0.802	0.854	0.854	0.927	0.948	0.878
RF	0.854	0.802	0.833	0.875	0.938	0.948	0.875
SU	0.927	0.823	0.854	0.896	0.969	0.948	0.903
OneR	0.833	0.781	0.833	0.844	0.927	0.938	0.859
IG	0.927	0.844	0.854	0.875	0.969	0.958	0.905
Proposed Ensemble	0.927	0.844	0.854	0.906	0.979	0.948	0.910
Based Ensemble	0.917	0.823	0.844	0.865	0.969	0.948	0.894
Original Dataset	0.458	0.572	0.543	0.537	0.541	0.459	0.518

ตารางที่ 4.25 : ตารางแสดงค่า Recall ของแบบจำลองที่สร้างจากจำนวนคุณสมบัติ 60% สำหรับชุดข้อมูลมะเร็งต่อมน้ำเหลือง

Ranker	K-NN	C4.5	NB	RF	LR	SVM	Average
GR	0.917	0.833	0.844	0.906	0.927	0.958	0.898
RF	0.896	0.896	0.833	0.885	0.917	0.958	0.898
SU	0.927	0.813	0.844	0.896	0.938	0.969	0.898
OneR	0.833	0.729	0.823	0.833	0.906	0.927	0.842
IG	0.938	0.844	0.844	0.906	0.958	0.958	0.908
Proposed Ensemble	0.927	0.833	0.844	0.885	0.958	0.969	0.903
Based Ensemble	0.906	0.813	0.833	0.875	0.958	0.969	0.892
Original Dataset	0.458	0.572	0.543	0.537	0.541	0.459	0.518

ตารางที่ 4.26 : ตารางแสดงค่า Recall ของแบบจำลองที่สร้างจากจำนวนคุณสมบัติ 80% สำหรับชุดข้อมูลมะเร็งต่อมน้ำเหลือง

Ranker	K-NN	C4.5	NB	RF	LR	SVM	Average
GR	0.927	0.833	0.844	0.875	0.927	0.958	0.894
RF	0.896	0.875	0.833	0.906	0.927	0.958	0.899
SU	0.938	0.813	0.854	0.865	0.948	0.969	0.898
OneR	0.833	0.771	0.813	0.823	0.885	0.927	0.842
IG	0.948	0.844	0.865	0.885	0.938	0.969	0.908
Proposed Ensemble	0.938	0.872	0.854	0.899	0.958	0.969	0.915
Based Ensemble	0.906	0.781	0.833	0.896	0.948	0.969	0.889
Original Dataset	0.458	0.572	0.543	0.537	0.541	0.459	0.518

ตารางที่ 4.27 : ตารางแสดงค่า Precision ของแบบจำลองที่สร้างจากจำนวนคุณสมบัติ 20% สำหรับชุดข้อมูลมะเร็งต่อมน้ำเหลือง

Ranker	K-NN	C4.5	NB	RF	LR	SVM	Average
GR	0.823	0.741	0.764	0.804	0.781	0.746	0.777
RF	0.791	0.740	0.788	0.779	0.817	0.845	0.793
SU	0.865	0.795	0.784	0.904	0.866	0.720	0.822
OneR	0.729	0.745	0.695	0.804	0.766	0.717	0.743
IG	0.897	0.790	0.774	0.875	0.851	0.840	0.838
Proposed Ensemble	0.914	0.778	0.784	0.888	0.865	0.720	0.825
Based Ensemble	0.855	0.820	0.779	0.894	0.869	0.787	0.834
Original Dataset	0.347	0.489	0.537	0.613	0.589	0.361	0.489

ตารางที่ 4.28 : ตารางแสดงค่า Precision ของแบบจำลองที่สร้างจากจำนวนคุณสมบัติ 40% สำหรับชุดข้อมูลมะเร็งต่อมน้ำเหลือง

Ranker	K-NN	C4.5	NB	RF	LR	SVM	Average
GR	0.872	0.792	0.783	0.822	0.920	0.929	0.853
RF	0.835	0.773	0.779	0.822	0.923	0.931	0.844
SU	0.916	0.807	0.795	0.874	0.949	0.932	0.879
OneR	0.820	0.773	0.781	0.824	0.933	0.935	0.844
IG	0.910	0.825	0.794	0.837	0.950	0.940	0.876
Proposed Ensemble	0.916	0.828	0.795	0.870	0.959	0.932	0.883
Based Ensemble	0.903	0.818	0.787	0.802	0.950	0.932	0.865
Original Dataset	0.347	0.489	0.537	0.613	0.589	0.361	0.489

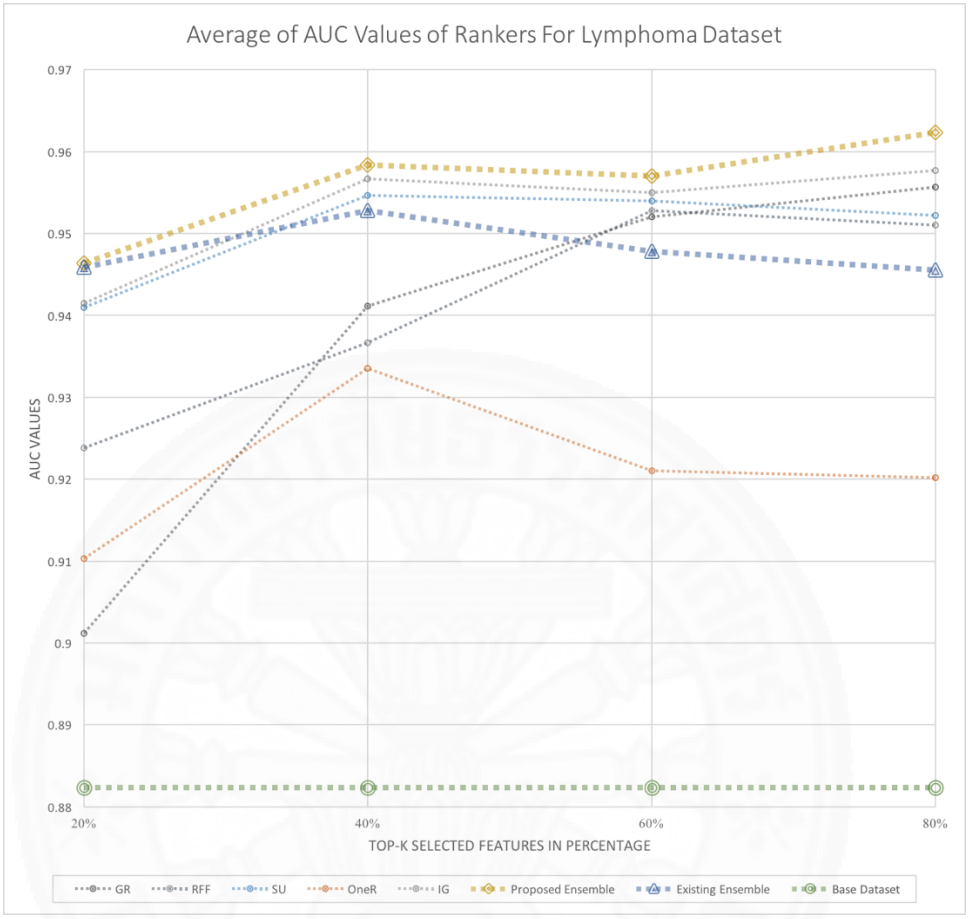
ตารางที่ 4.29 : ตารางแสดงค่า Precision ของแบบจำลองที่สร้างจากจำนวนคุณสมบัติ 60% สำหรับชุดข้อมูลมะเร็งต่อมน้ำเหลือง

Ranker	K-NN	C4.5	NB	RF	LR	SVM	Average
GR	0.899	0.821	0.788	0.876	0.914	0.939	0.873
RF	0.875	0.859	0.783	0.832	0.909	0.939	0.866
SU	0.910	0.797	0.787	0.842	0.924	0.950	0.868
OneR	0.825	0.733	0.776	0.795	0.882	0.902	0.819
IG	0.923	0.825	0.789	0.875	0.960	0.939	0.885
Proposed Ensemble	0.910	0.845	0.787	0.843	0.943	0.950	0.880
Based Ensemble	0.891	0.811	0.783	0.837	0.941	0.950	0.869
Original Dataset	0.347	0.489	0.537	0.613	0.589	0.361	0.489

ตารางที่ 4.30 : ตารางแสดงค่า Precision ของแบบจำลองที่สร้างจากจำนวนคุณสมบัติ 80% สำหรับชุดข้อมูลมะเร็งต่อมน้ำเหลือง

Ranker	K-NN	C4.5	NB	RF	LR	SVM	Average
GR	0.911	0.821	0.789	0.822	0.914	0.939	0.866
RF	0.872	0.851	0.779	0.897	0.919	0.939	0.876
SU	0.920	0.797	0.796	0.815	0.933	0.950	0.869
OneR	0.811	0.784	0.706	0.788	0.860	0.902	0.809
IG	0.931	0.825	0.804	0.882	0.922	0.950	0.886
Proposed Ensemble	0.920	0.813	0.796	0.872	0.961	0.950	0.885
Based Ensemble	0.892	0.779	0.783	0.867	0.933	0.950	0.867
Original Dataset	0.347	0.489	0.537	0.613	0.589	0.361	0.489

จากตารางผลลัพธ์สามารถสรุปเป็นกราฟโดยใช้ค่าเฉลี่ยของ AUC, Recall และ Precision ของแต่ละเทคนิคได้ดังนี้



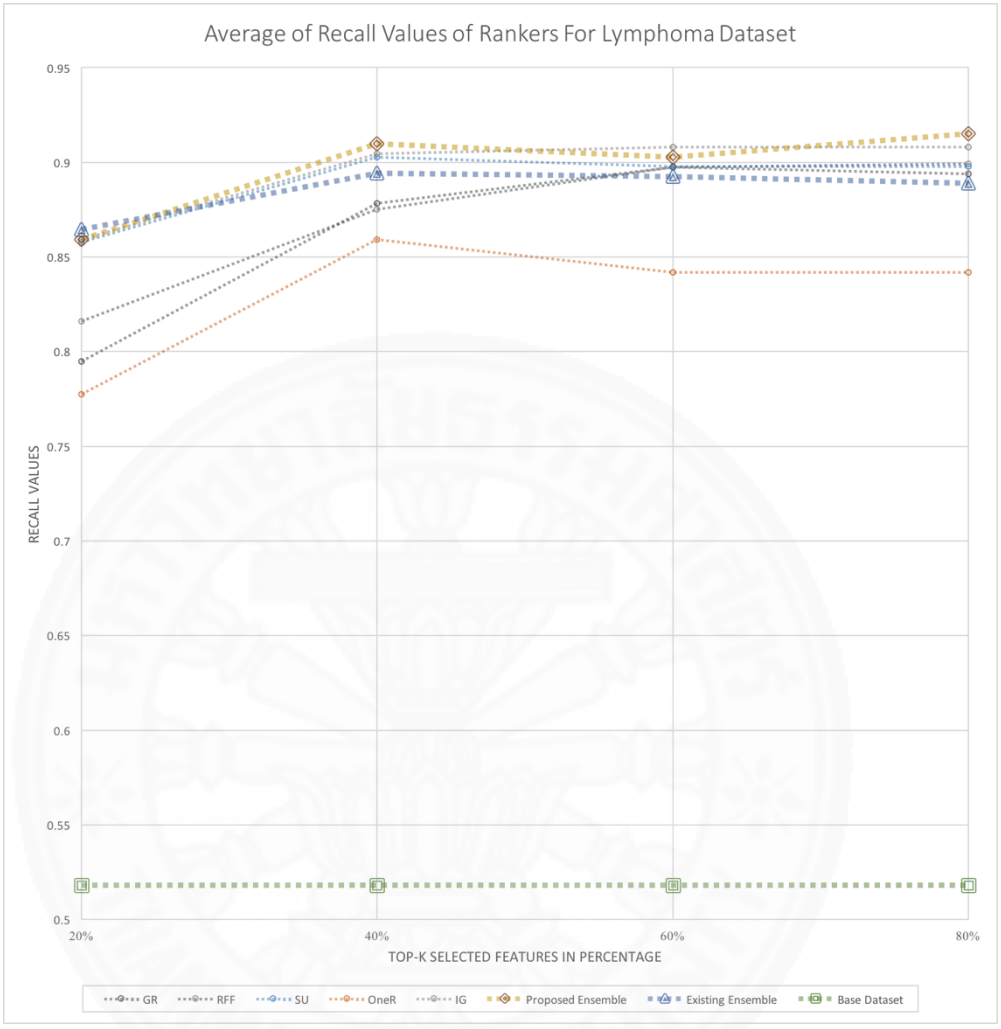
ภาพที่ 4.5 : กราฟแสดงค่า **AUC** โดยเฉลี่ยสำหรับชุดข้อมูลมะเร็งต่อมน้ำเหลือง

ผลลัพธ์จากกราฟค่า AUC โดยเฉลี่ยนั้น จะเห็นว่าหากนำข้อมูลมาผ่านอัลกอริทึมที่ใช้ในการคัดเลือกคุณสมบัติไม่ว่าจะเป็นแบบรวมหรือแบบเดี่ยว จะให้ผลลัพธ์ที่ดีกว่าแบบ Base Dataset ในทุกจำนวนคุณสมบัติ จำนวนคุณสมบัติที่แต่ละอัลกอริทึมมีค่า AUC โดยเฉลี่ยใกล้เคียงกันคือ 60% และ อัลกอริทึมที่มีแนวโน้มมีประสิทธิภาพดีที่สุด โดยพิจารณาจากทุกจำนวนคุณสมบัติที่ทำการทดสอบคือ อัลกอริทึมที่ผู้วิจัยนำเสนอ

จากกราฟค่า AUC โดยเฉลี่ย พบว่าอัลกอริทึมที่ผู้วิจัยนำเสนอมีประสิทธิภาพดีที่สุดในทุกจำนวนคุณสมบัติ โดยทำค่าได้สูงสุดที่จำนวนคุณสมบัติ 80% ซึ่งมีค่า AUC โดยเฉลี่ยเท่ากับ 0.962

สำหรับอัลกอริทึมต้นแบบมีค่า AUC โดยเฉลี่ยต่ำกว่าอัลกอริทึมที่นำเสนอ ในทุกจำนวนคุณสมบัติ โดยมีค่า AUC โดยเฉลี่ยสูงสุดที่จำนวนคุณสมบัติ 40% ซึ่งมีค่า AUC โดยเฉลี่ยเท่ากับ 0.953 อัลกอริทึมที่พบว่ามีค่า AUC โดยเฉลี่ยต่ำที่สุดในทุกจำนวนคุณสมบัติ คือ OneR

สำหรับค่า Recall โดยเฉลี่ยของแต่ละเทคนิคสามารถสรุปเป็นกราฟได้ดังนี้



ภาพที่ 4.6 : กราฟแสดงค่า **Recall** โดยเฉลี่ยสำหรับชุดข้อมูลมะเร็งต่อมน้ำเหลือง

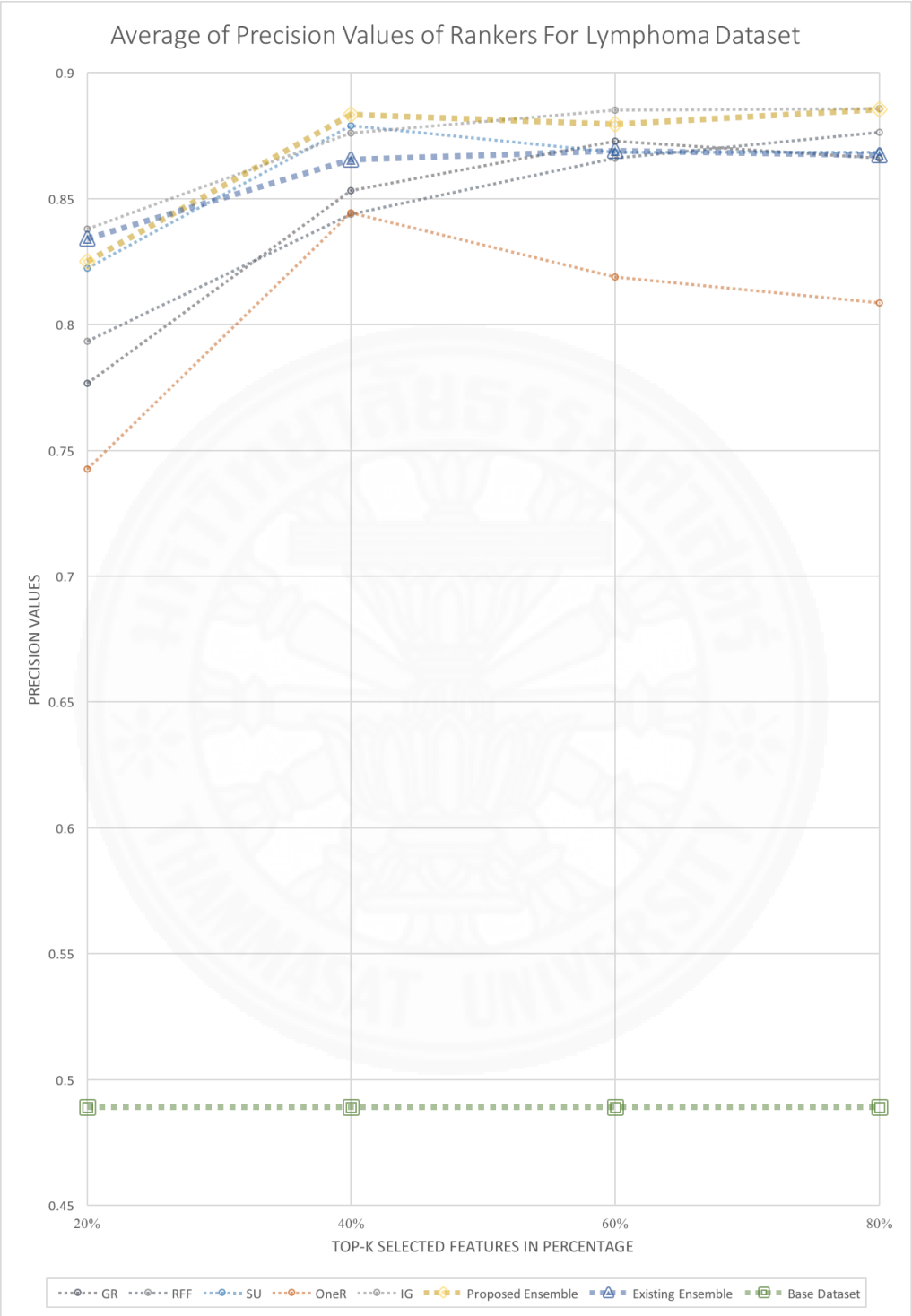
ผลลัพธ์จากกราฟค่า Recall โดยเฉลี่ยนั้น จะเห็นว่าหากนำข้อมูลมาผ่านอัลกอริทึมที่ใช้ในการคัดเลือกคุณสมบัติไม่ว่าจะเป็นแบบรวมหรือแบบเดี่ยว จะให้ผลลัพธ์ที่ดีกว่าแบบ Base Dataset ในทุกจำนวนคุณสมบัติเช่นกัน จำนวนคุณสมบัติที่แต่ละอัลกอริทึมมีค่า Recall โดยเฉลี่ยใกล้เคียงกันคือ 60% และอัลกอริทึมที่มีแนวโน้มมีประสิทธิภาพดีที่สุดโดยพิจารณาจากทุกจำนวนคุณสมบัติที่ทำการทดสอบคือ อัลกอริทึมที่ผู้วิจัยนำเสนอ

จากกราฟค่า Recall โดยเฉลี่ยพบว่าอัลกอริทึมที่ผู้วิจัยนำเสนอมีประสิทธิภาพดีที่สุด ในจำนวนคุณสมบัติ 40% และ 80% โดยทำค่าได้สูงสุดที่จำนวนคุณสมบัติ 80% ซึ่งมีค่า Recall โดยเฉลี่ยเท่ากับ 0.915 สำหรับอัลกอริทึมที่มีค่า Recall โดยเฉลี่ยสูงที่สุดที่จำนวนคุณสมบัติ 20% และ 60% คือ

อัลกอริทึมต้นแบบและ Information Gain ตามลำดับ ซึ่งมีค่า Recall โดยเฉลี่ยเท่ากับ 0.864 และ 0.908 ในขณะที่อัลกอริทึมที่นำเสนอมีค่า Recall โดยเฉลี่ยที่จำนวนคุณสมบัติ 60% คือ 0.903

สำหรับอัลกอริทึมต้นแบบมีค่า Recall โดยเฉลี่ยต่ำกว่าอัลกอริทึมที่นำเสนอ ในทุกๆจำนวนคุณสมบัติ โดยมีค่า Recall โดยเฉลี่ยสูงสุดที่จำนวนคุณสมบัติ 20% ซึ่งมีค่า Recall โดยเฉลี่ยเท่ากับ 0.894 อัลกอริทึมที่พบว่ามีค่า Recall โดยเฉลี่ยต่ำที่สุดคือ OneR

สำหรับค่า Precision โดยเฉลี่ยของแต่ละเทคนิคสามารถสรุปเป็นกราฟได้ดังนี้



ภาพที่ 4.7 : กราฟแสดงค่า **Precision** โดยเฉลี่ยสำหรับชุดข้อมูลมะเร็งต่อมน้ำเหลือง

ผลลัพธ์จากกราฟค่า Precision โดยเฉลี่ยนั้น จะเห็นว่าหากนำข้อมูลมาผ่านอัลกอริทึมที่ใช้ในการ

คัดเลือกคุณสมบัติไม่ว่าจะเป็นแบบรวมหรือแบบเดี่ยว จะให้ผลลัพธ์ที่ดีกว่าแบบ Base Dataset ในทุกจำนวนคุณสมบัติเช่นกัน สำหรับค่า Precision นั้นพบว่าแต่ละอัลกอริทึมมีค่าใกล้เคียงกันที่จำนวนคุณสมบัติ 40% และอัลกอริทึมที่มีแนวโน้มมีประสิทธิภาพดีที่สุด โดยพิจารณาจากทุกจำนวนคุณสมบัติที่ทำการทดสอบคือ Information Gain

จากกราฟค่า Precision โดยเฉลี่ยพบว่าอัลกอริทึมที่ผู้วิจัยนำเสนอ มีประสิทธิภาพดีที่สุดที่จำนวนคุณสมบัติ 40% ซึ่งมีค่า Precision โดยเฉลี่ยเท่ากับ 0.883 สำหรับอัลกอริทึมที่มีค่า Precision โดยเฉลี่ยสูงที่สุดที่จำนวนคุณสมบัติ 20%, 60% และ 80% คือ Information Gain ซึ่งมีค่า Precision โดยเฉลี่ยเท่ากับ 0.838, 0.885 และ 0.886 ในขณะที่อัลกอริทึมที่นำเสนอมีค่า Precision โดยเฉลี่ยที่จำนวนคุณสมบัติ 20%, 60% และ 80% คือ 0.824, 0.880 และ 0.885

สำหรับอัลกอริทึมต้นแบบมีค่า Precision โดยเฉลี่ยต่ำกว่าอัลกอริทึมที่นำเสนอ ที่จำนวนคุณสมบัติ 40%, 60% และ 80% โดยมีค่า Precision โดยเฉลี่ยสูงสุด ที่จำนวนคุณสมบัติ 60% ซึ่งมีค่า Precision โดยเฉลี่ยเท่ากับ 0.868 อัลกอริทึมที่พบว่ามีค่า Precision โดยเฉลี่ยต่ำที่สุดคือ OneR

4.3 ผลการทดลองสำหรับชุดข้อมูลมะเร็งเต้านม

ในการทดลองโดยใช้ข้อมูลมะเร็งเต้านม ผู้วิจัยได้ทำการแบ่งจำนวนคุณสมบัติหลังจากที่ผ่านกระบวนการคัดเลือกคุณสมบัติแล้วทั้งหมด 24,482 คุณสมบัติออกเป็น 4 ส่วนคือ 20% ,40% ,60% และ 80% เพื่อทำการทดลองดังตารางที่ 4.31

ตารางที่ 4.31 : ตารางแสดงค่าจำนวนคุณสมบัติคิดเป็นเปอร์เซ็นต์สำหรับชุดข้อมูลมะเร็งเต้านม

20%	40%	60%	80%	ทั้งหมด
4,896	9,793	146,89	19,586	24,482

จากนั้นทำการแบ่งจำนวนรายการ (Instances) ทั้งหมดออกเป็น 2 ส่วนคือ ชุดฝึกจำนวน 70% และสำหรับชุดทดสอบจำนวน 30% ซึ่งอัตราส่วนของค่าคลาสคุณสมบัติของชุดข้อมูลมะเร็งเต้านม ซึ่งเป็นตัวอักษร {relapse, non-relapse } ในส่วนของชุดฝึกและชุดทดสอบ หลังจากการแบ่งจำนวนรายการแล้ว เป็นตามตารางที่ 4.32 และ 4.33

ตารางที่ 4.32 : ตารางแสดงอัตราส่วนของค่าคลาสคุณสมบัติของชุดข้อมูลมะเร็งเต้านมสำหรับชุดฝึก

ค่าของคุณสมบัติคลาส	อัตราส่วน
relapse	62%
non-relapse	38%

ตารางที่ 4.33 : ตารางแสดงอัตราส่วนของค่าคลาสคุณสมบัติของชุดข้อมูลมะเร็งเต้านมสำหรับชุดทดสอบ

ค่าของคุณสมบัติคลาส	อัตราส่วน
relapse	53%
non-relapse	47%

ผลการทดลองหลังจากทำการแบ่งจำนวนคุณสมบัติออกเป็น 4 ส่วนเป็นดังนี้

ตารางที่ 4.34 : ตารางแสดงค่า AUC ของแบบจำลองที่สร้างจากจำนวนคุณสมบัติ 20% สำหรับชุดข้อมูลมะเร็งเต้านม

Ranker	K-NN	C4.5	NB	RF	LR	SVM	Average
GR	0.643	0.689	0.840	0.921	0.934	0.778	0.801
RF	0.792	0.655	0.859	0.856	0.993	0.781	0.823
SU	0.735	0.727	0.841	0.935	0.996	0.744	0.830
OneR	0.686	0.755	0.755	0.906	0.812	0.791	0.784
IG	0.690	0.710	0.715	0.938	0.912	0.733	0.783
Proposed Ensemble	0.735	0.781	0.838	0.931	0.995	0.744	0.837
Based Ensemble	0.705	0.720	0.711	0.933	0.921	0.759	0.792
Original Dataset	0.573	0.471	0.423	0.619	0.568	0.490	0.524

ตารางที่ 4.35 : ตารางแสดงค่า AUC ของแบบจำลองที่สร้างจากจำนวนคุณสมบัติ 40% สำหรับชุดข้อมูลมะเร็งเต้านม

Ranker	K-NN	C4.5	NB	RF	LR	SVM	Average
GR	0.697	0.631	0.701	0.954	0.920	0.796	0.783
RF	0.880	0.585	0.835	0.908	0.934	0.852	0.832
SU	0.736	0.645	0.845	0.957	0.967	0.862	0.835
OneR	0.732	0.579	0.621	0.936	0.901	0.769	0.756
IG	0.814	0.591	0.661	0.941	0.945	0.855	0.801
Proposed Ensemble	0.783	0.602	0.857	0.938	0.965	0.878	0.837
Based Ensemble	0.665	0.660	0.715	0.938	0.963	0.848	0.798
Original Dataset	0.573	0.471	0.423	0.619	0.568	0.490	0.524

ตารางที่ 4.36 : ตารางแสดงค่า AUC ของแบบจำลองที่สร้างจากจำนวนคุณสมบัติ 60% สำหรับชุดข้อมูลมะเร็งเต้านม

Ranker	K-NN	C4.5	NB	RF	LR	SVM	Average
GR	0.781	0.613	0.686	0.945	0.890	0.747	0.777
RF	0.848	0.584	0.850	0.892	0.921	0.826	0.820
SU	0.820	0.577	0.874	0.938	0.932	0.855	0.833
OneR	0.769	0.486	0.607	0.936	0.902	0.765	0.744
IG	0.823	0.614	0.658	0.943	0.933	0.826	0.800
Proposed Ensemble	0.807	0.634	0.853	0.945	0.934	0.840	0.836
Based Ensemble	0.841	0.590	0.722	0.928	0.921	0.803	0.801
Original Dataset	0.573	0.471	0.423	0.619	0.568	0.490	0.524

ตารางที่ 4.37 : ตารางแสดงค่า AUC ของแบบจำลองที่สร้างจากจำนวนคุณสมบัติ 80% สำหรับชุดข้อมูลมะเร็งเต้านม

Ranker	K-NN	C4.5	NB	RF	LR	SVM	Average
GR	0.783	0.540	0.675	0.927	0.890	0.773	0.765
RF	0.830	0.650	0.847	0.883	0.921	0.799	0.822
SU	0.793	0.561	0.801	0.938	0.932	0.811	0.806
OneR	0.802	0.473	0.608	0.927	0.902	0.747	0.743
IG	0.805	0.582	0.644	0.930	0.933	0.773	0.778
Proposed Ensemble	0.805	0.621	0.821	0.937	0.934	0.826	0.824
Based Ensemble	0.804	0.561	0.716	0.932	0.921	0.814	0.791
Original Dataset	0.573	0.471	0.423	0.619	0.568	0.490	0.524

ตารางที่ 4.38 : ตารางแสดงค่า Recall ของแบบจำลองที่สร้างจากจำนวนคุณสมบัติ 20% สำหรับชุดข้อมูลมะเร็งเต้านม

Ranker	K-NN	C4.5	NB	RF	LR	SVM	Average
GR	0.679	0.692	0.641	0.808	0.744	0.795	0.727
RF	0.808	0.615	0.782	0.769	0.808	0.795	0.763
SU	0.756	0.744	0.577	0.859	0.692	0.756	0.731
OneR	0.705	0.692	0.603	0.769	0.628	0.795	0.699
IG	0.731	0.731	0.564	0.859	0.744	0.744	0.729
Proposed Ensemble	0.803	0.746	0.577	0.859	0.744	0.795	0.754
Based Ensemble	0.718	0.731	0.551	0.821	0.705	0.769	0.716
Original Dataset	0.581	0.436	0.511	0.596	0.549	0.521	0.532

ตารางที่ 4.39 : ตารางแสดงค่า Recall ของแบบจำลองที่สร้างจากจำนวนคุณสมบัติ 40% สำหรับชุดข้อมูลมะเร็งเต้านม

Ranker	K-NN	C4.5	NB	RF	LR	SVM	Average
GR	0.731	0.654	0.654	0.885	0.795	0.808	0.755
RF	0.897	0.590	0.795	0.846	0.846	0.859	0.806
SU	0.756	0.615	0.654	0.859	0.795	0.846	0.754
OneR	0.756	0.577	0.590	0.833	0.756	0.769	0.714
IG	0.833	0.615	0.615	0.846	0.821	0.859	0.765
Proposed Ensemble	0.891	0.615	0.743	0.872	0.845	0.885	0.809
Based Ensemble	0.833	0.679	0.654	0.872	0.769	0.859	0.778
Original Dataset	0.581	0.436	0.511	0.596	0.549	0.521	0.532

ตารางที่ 4.40 : ตารางแสดงค่า Recall ของแบบจำลองที่สร้างจากจำนวนคุณสมบัติ 60% สำหรับชุดข้อมูลมะเร็งเต้านม

Ranker	K-NN	C4.5	NB	RF	LR	SVM	Average
GR	0.795	0.654	0.628	0.846	0.795	0.756	0.746
RF	0.872	0.590	0.795	0.885	0.782	0.833	0.793
SU	0.833	0.603	0.641	0.859	0.795	0.859	0.765
OneR	0.795	0.487	0.590	0.885	0.795	0.769	0.720
IG	0.846	0.615	0.603	0.885	0.833	0.833	0.769
Proposed Ensemble	0.821	0.603	0.654	0.872	0.795	0.846	0.765
Based Ensemble	0.859	0.615	0.667	0.821	0.718	0.808	0.748
Original Dataset	0.581	0.436	0.511	0.596	0.549	0.521	0.532

ตารางที่ 4.41 : ตารางแสดงค่า Recall ของแบบจำลองที่สร้างจากจำนวนคุณสมบัติ 80% สำหรับชุดข้อมูลมะเร็งเต้านม

Ranker	K-NN	C4.5	NB	RF	LR	SVM	Average
GR	0.795	0.577	0.590	0.821	0.705	0.782	0.712
RF	0.859	0.654	0.808	0.808	0.782	0.808	0.787
SU	0.808	0.590	0.615	0.846	0.756	0.821	0.739
OneR	0.821	0.487	0.590	0.846	0.744	0.756	0.707
IG	0.833	0.615	0.603	0.833	0.756	0.782	0.737
Proposed Ensemble	0.859	0.610	0.808	0.846	0.776	0.833	0.789
Based Ensemble	0.808	0.590	0.641	0.846	0.756	0.821	0.744
Original Dataset	0.581	0.436	0.511	0.596	0.549	0.521	0.532

ตารางที่ 4.42 : ตารางแสดงค่า Precision ของแบบจำลองที่สร้างจากจำนวนคุณสมบัติ 20% สำหรับชุดข้อมูลมะเร็งเต้านม

Ranker	K-NN	C4.5	NB	RF	LR	SVM	Average
GR	0.703	0.692	0.728	0.841	0.748	0.803	0.753
RF	0.809	0.622	0.790	0.769	0.809	0.798	0.766
SU	0.759	0.744	0.575	0.860	0.692	0.757	0.731
OneR	0.703	0.698	0.645	0.771	0.636	0.795	0.708
IG	0.735	0.730	0.539	0.860	0.750	0.743	0.726
Proposed Ensemble	0.759	0.732	0.575	0.863	0.797	0.757	0.747
Based Ensemble	0.718	0.732	0.505	0.820	0.704	0.769	0.708
Original Dataset	0.591	0.483	0.442	0.571	0.562	0.457	0.518

ตารางที่ 4.43 : ตารางแสดงค่า Precision ของแบบจำลองที่สร้างจากจำนวนคุณสมบัติ 40% สำหรับชุดข้อมูลมะเร็งเต้านม

Ranker	K-NN	C4.5	NB	RF	LR	SVM	Average
GR	0.773	0.653	0.785	0.889	0.794	0.81	0.784
RF	0.905	0.593	0.801	0.848	0.852	0.86	0.810
SU	0.764	0.613	0.785	0.861	0.798	0.846	0.778
OneR	0.756	0.589	0.655	0.836	0.756	0.771	0.727
IG	0.834	0.609	0.771	0.846	0.821	0.859	0.790
Proposed Ensemble	0.798	0.613	0.781	0.874	0.821	0.886	0.796
Based Ensemble	0.841	0.684	0.739	0.872	0.768	0.863	0.795
Original Dataset	0.591	0.483	0.442	0.571	0.562	0.457	0.518

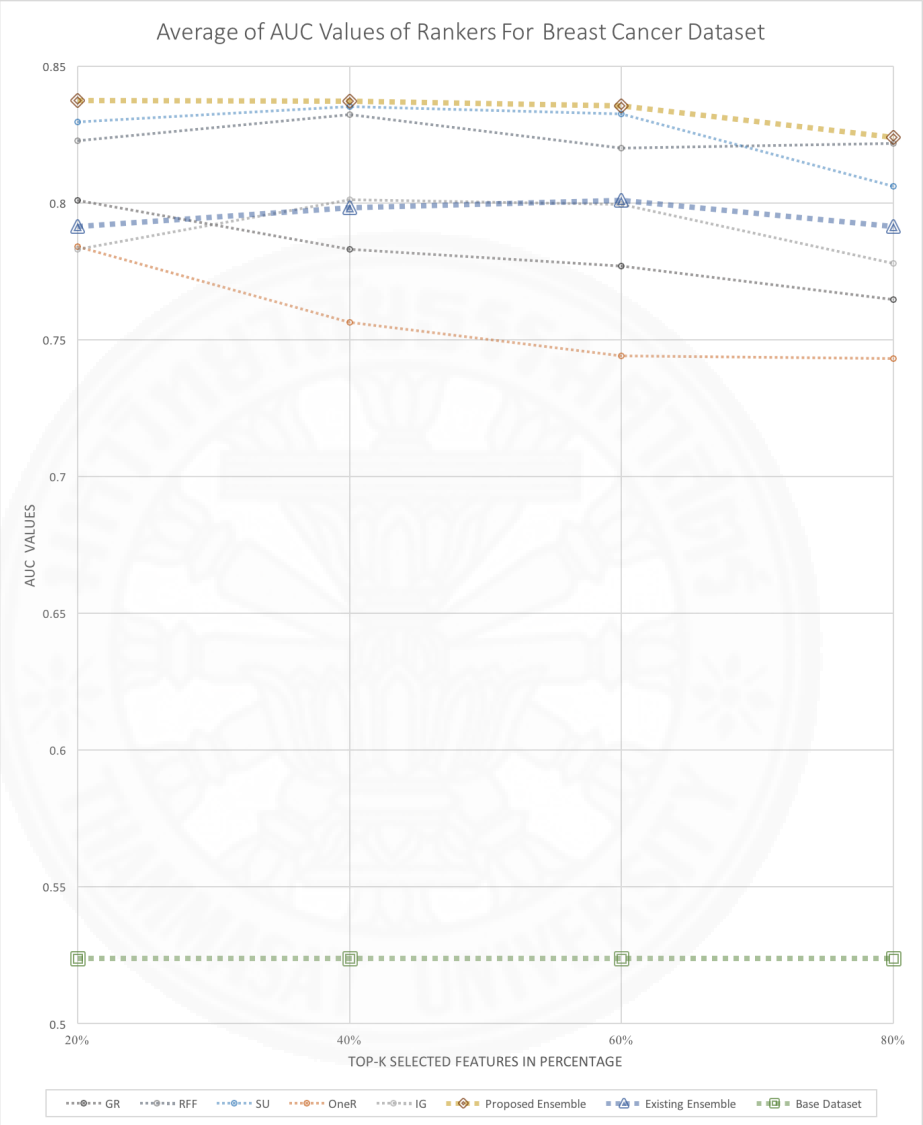
ตารางที่ 4.44 : ตารางแสดงค่า Precision ของแบบจำลองที่สร้างจากจำนวนคุณสมบัติ 60% สำหรับชุดข้อมูลมะเร็งเต้านม

Ranker	K-NN	C4.5	NB	RF	LR	SVM	Average
GR	0.803	0.651	0.776	0.848	0.795	0.755	0.771
RF	0.879	0.593	0.801	0.889	0.783	0.834	0.797
SU	0.834	0.601	0.781	0.859	0.794	0.859	0.788
OneR	0.798	0.484	0.655	0.886	0.797	0.769	0.732
IG	0.846	0.611	0.767	0.884	0.836	0.834	0.796
Proposed Ensemble	0.822	0.587	0.785	0.872	0.783	0.859	0.785
Based Ensemble	0.860	0.611	0.790	0.820	0.716	0.807	0.767
Original Dataset	0.591	0.483	0.442	0.571	0.562	0.457	0.518

ตารางที่ 4.45 : ตารางแสดงค่า Precision ของแบบจำลองที่สร้างจากจำนวนคุณสมบัติ 80% สำหรับชุดข้อมูลมะเร็งเต้านม

Ranker	K-NN	C4.5	NB	RF	LR	SVM	Average
GR	0.798	0.567	0.762	0.820	0.706	0.782	0.739
RF	0.860	0.651	0.811	0.807	0.782	0.808	0.787
SU	0.814	0.585	0.771	0.846	0.757	0.822	0.766
OneR	0.830	0.484	0.762	0.846	0.743	0.755	0.737
IG	0.837	0.610	0.767	0.833	0.756	0.782	0.764
Proposed Ensemble	0.825	0.629	0.771	0.848	0.744	0.834	0.775
Based Ensemble	0.814	0.585	0.781	0.848	0.757	0.820	0.768
Original Dataset	0.591	0.483	0.442	0.571	0.562	0.457	0.518

จากตารางผลลัพธ์สามารถสรุปเป็นกราฟโดยใช้ค่าเฉลี่ยของ AUC, Recall และ Precision ของแต่ละเทคนิคได้ดังนี้



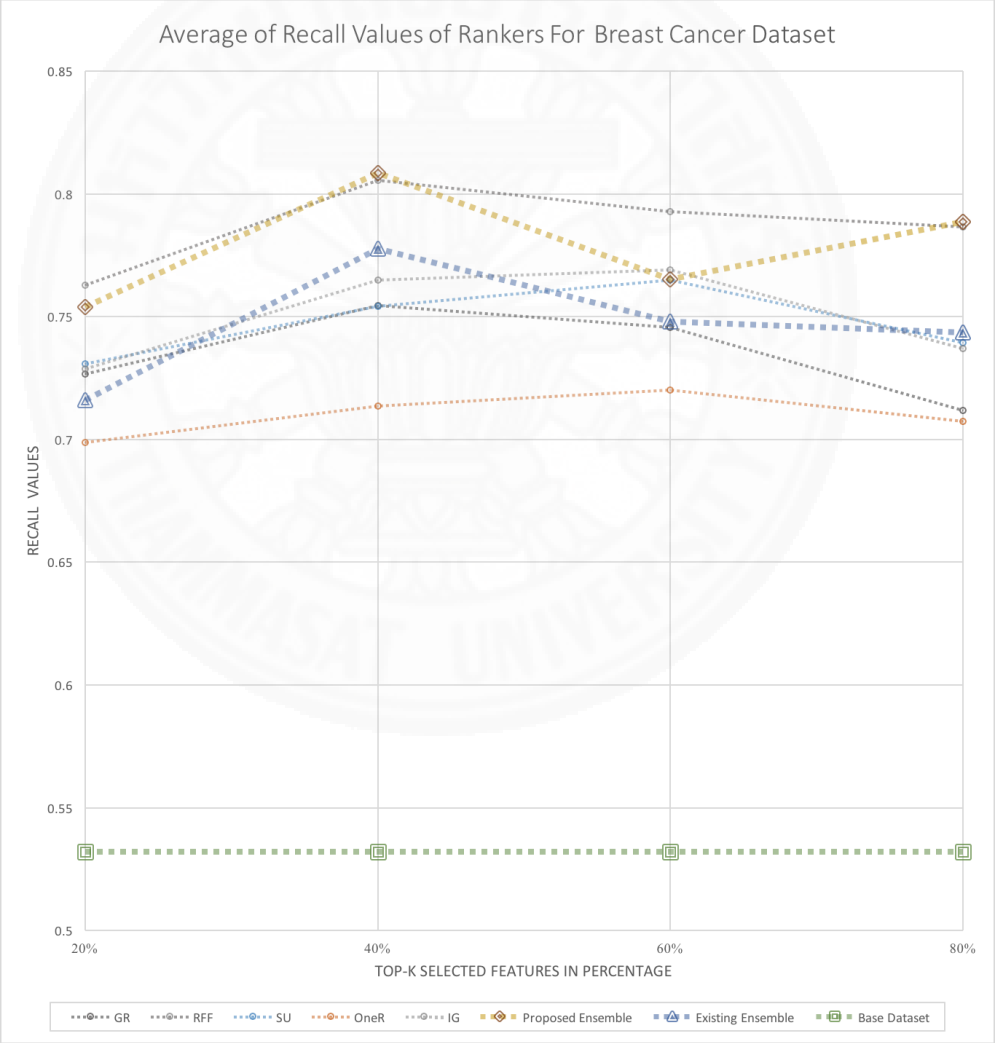
ภาพที่ 4.8 : กราฟแสดงค่า **AUC** โดยเฉลี่ยสำหรับชุดข้อมูลมะเร็งเต้านม

ผลลัพธ์จากกราฟค่า AUC โดยเฉลี่ยนั้น จะเห็นว่าหากนำข้อมูลมาผ่านอัลกอริทึมที่ใช้ในการคัดเลือกคุณสมบัติไม่ว่าจะเป็นแบบรวมหรือแบบเดี่ยวจะให้ผลลัพธ์ที่ดีกว่าแบบ Base Dataset ในทุกจำนวนคุณสมบัติ จำนวนคุณสมบัติที่หลายๆอัลกอริทึมมีค่า AUC โดยเฉลี่ยใกล้เคียงกันคือ 20% และ อัลกอริทึมที่มีแนวโน้มมีประสิทธิภาพดีที่สุดโดยพิจารณาจากทุกจำนวนคุณสมบัติที่ทำการทดสอบคือ อัลกอริทึมที่ผู้วิจัยนำเสนอ

จากกราฟค่า AUC โดยเฉลี่ยพบว่าอัลกอริทึมที่ผู้วิจัยนำเสนอมีประสิทธิภาพดีที่สุดในทุกจำนวนคุณสมบัติ โดยทำค่าได้สูงสุดที่จำนวนคุณสมบัติ 60% ซึ่งมีค่า AUC โดยเฉลี่ยเท่ากับ 0.836

สำหรับอัลกอริทึมต้นแบบมีค่า AUC โดยเฉลี่ยต่ำกว่าอัลกอริทึมที่นำเสนอในทุกๆจำนวนคุณสมบัติ โดยมีค่า AUC โดยเฉลี่ยสูงสุดที่จำนวนคุณสมบัติ 60% ซึ่งมีค่า AUC โดยเฉลี่ยเท่ากับ 0.800 อัลกอริทึมที่พบว่ามีค่า AUC โดยเฉลี่ยต่ำที่สุดคือ OneR

สำหรับค่า Recall โดยเฉลี่ยของแต่ละเทคนิคสามารถสรุปเป็นกราฟได้ดังนี้



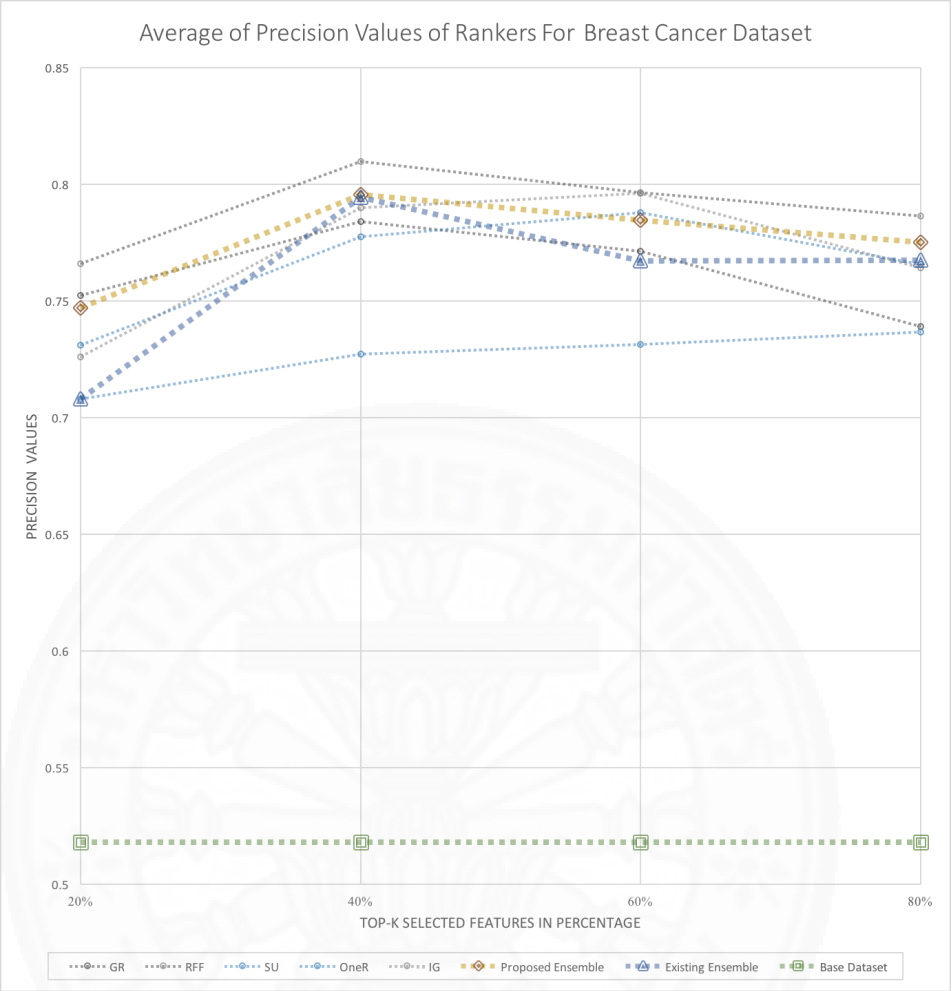
ภาพที่ 4.9 : กราฟแสดงค่า **Recall** โดยเฉลี่ยสำหรับชุดข้อมูลมะเร็งเต้านม

ผลลัพธ์จากกราฟค่า Recall โดยเฉลี่ยนั้น จะเห็นว่าหากนำข้อมูลมาผ่านอัลกอริทึมที่ใช้ในการคัดเลือกคุณสมบัติไม่ว่าจะเป็นแบบรวมหรือแบบเดี่ยวจะให้ผลลัพธ์ที่ดีกว่าแบบ Base Dataset ในทุกจำนวนคุณสมบัติเช่นกัน อัลกอริทึมที่มีแนวโน้มมีประสิทธิภาพดีที่สุดโดยพิจารณาจากทุกจำนวนคุณสมบัติที่ทำการทดสอบคือ Relief

จากกราฟค่า Recall โดยเฉลี่ยพบว่าอัลกอริทึมที่ผู้วิจัยนำเสนอมีประสิทธิภาพดีที่สุดในจำนวนคุณสมบัติ 40% และ 80% โดยทำค่าได้สูงสุดที่จำนวนคุณสมบัติ 40% ซึ่งมีค่า Recall โดยเฉลี่ยเท่ากับ 0.809 สำหรับอัลกอริทึมที่มีค่า Recall โดยเฉลี่ยสูงที่สุดที่จำนวนคุณสมบัติ 20% และ 60% คือ ReliefF ซึ่งมีค่า Recall โดยเฉลี่ยเท่ากับ 0.763 และ 0.793 ในขณะที่อัลกอริทึมที่นำเสนอมีค่า Recall โดยเฉลี่ยที่จำนวนคุณสมบัติ 20% และ 60% คือ 0.754 และ 0.765

สำหรับอัลกอริทึมต้นแบบมีค่า Recall โดยเฉลี่ยต่ำกว่าอัลกอริทึมที่นำเสนอที่ทุกจำนวนคุณสมบัติ โดยมีค่า Recall โดยเฉลี่ยสูงสุดที่จำนวนคุณสมบัติ 20% ซึ่งมีค่า Recall โดยเฉลี่ยเท่ากับ 0.778 อัลกอริทึมที่พบว่ามีค่า Recall โดยเฉลี่ยต่ำที่สุดคือ OneR

สำหรับค่า Precision โดยเฉลี่ยของแต่ละเทคนิคสามารถสรุปเป็นกราฟได้ดังนี้



ภาพที่ 4.10 : กราฟแสดงค่า Precision โดยเฉลี่ยสำหรับชุดข้อมูลมะเร็งเต้านม

ผลลัพธ์จากกราฟค่า Precision โดยเฉลี่ยนั้น จะเห็นว่าหากนำข้อมูลมาผ่านอัลกอริทึมที่ใช้ในการคัดเลือกคุณสมบัติไม่ว่าจะเป็นแบบรวมหรือแบบเดี่ยวจะให้ผลลัพธ์ที่ดีกว่าแบบ Base Dataset ในทุกจำนวนคุณสมบัติเช่นกัน สำหรับค่า Precision นั้นพบว่าแต่ละอัลกอริทึมมีค่าใกล้เคียงกันที่จำนวนคุณสมบัติ 80% และ อัลกอริทึมที่มีแนวโน้มมีประสิทธิภาพดีที่สุดโดยพิจารณาจากทุกจำนวนคุณสมบัติที่ทำการทดสอบคือ Relief

จากกราฟค่า Precision โดยเฉลี่ยพบว่าอัลกอริทึมที่มีประสิทธิภาพดีที่สุดในทุกจำนวนคุณสมบัติคือ ReliefF โดยอัลกอริทึมที่ผู้วิจัยนำเสนอมีค่า Precision โดยเฉลี่ยที่จำนวนคุณสมบัติ 20%, 40% , 60% และ 80% คือ 0.747, 0.796, 0.785 และ 0.775 ตามลำดับ อัลกอริทึมที่ผู้วิจัยนำเสนอ มีค่า Precision โดยเฉลี่ยสูงสุดที่จำนวนคุณสมบัติ 40% ซึ่งมีค่า Precision โดยเฉลี่ยเท่ากับ 0.796

สำหรับอัลกอริทึมต้นแบบมีค่า Precision โดยเฉลี่ยต่ำกว่าอัลกอริทึมที่นำเสนอทุกจำนวนคุณสมบัติ โดยมีค่า Precision โดยเฉลี่ยสูงสุดที่จำนวนคุณสมบัติ 40% ซึ่งมีค่า Precision โดยเฉลี่ยเท่ากับ 0.794 อัลกอริทึมที่พบว่ามีค่า Precision โดยเฉลี่ยต่ำที่สุดในทุกจำนวนคุณสมบัติคือ OneR

4.4 ผลการทดลองสำหรับชุดข้อมูลมะเร็งรังไข่

ในการทดลองโดยใช้ข้อมูลมะเร็งรังไข่ ผู้วิจัยได้ทำการแบ่งจำนวนคุณสมบัติหลังจากที่ผ่านกระบวนการคัดเลือกคุณสมบัติแล้วทั้งหมด 15,155 คุณสมบัติออกเป็น 4 ส่วนคือ 20% ,40% ,60% และ 80% เพื่อทำการทดลองดังตารางที่ 4.46

ตารางที่ 4.46 : ตารางแสดงค่าจำนวนคุณสมบัติคิดเป็นเปอร์เซ็นต์สำหรับชุดข้อมูลมะเร็งรังไข่

20%	40%	60%	80%	ทั้งหมด
3,031	6,062	9,093	12,124	15,155

จากนั้นทำการแบ่งจำนวนรายการ (Instances) ทั้งหมดออกเป็น 2 ส่วนคือชุดฝึกจำนวน 70% และสำหรับชุดทดสอบจำนวน 30% ซึ่งอัตราส่วนของค่าคลาสคุณสมบัติของชุดข้อมูลมะเร็งรังไข่ ซึ่งเป็นตัวอักษร { Cancer, Normal } ในส่วนของชุดฝึกและชุดทดสอบ หลังจากการแบ่งจำนวนรายการแล้ว เป็นตามตารางที่ 4.47 และ 4.48

ตารางที่ 4.47 : ตารางแสดงอัตราส่วนของค่าคลาสคุณสมบัติของชุดข้อมูลมะเร็งรังไข่สำหรับชุดฝึก

ค่าของคุณสมบัติคลาส	อัตราส่วน
Cancer	53%
Normal	47%

ตารางที่ 4.48 : ตารางแสดงอัตราส่วนของค่าคลาสคุณสมบัติของชุดข้อมูลมะเร็งรังไข่สำหรับชุดทดสอบ

ค่าของคุณสมบัติคลาส	อัตราส่วน
Cancer	44%
Normal	56%

ผลการทดลองหลังจากทำการแบ่งจำนวนคุณสมบัติออกเป็น 4 ส่วนเป็นดังนี้

ตารางที่ 4.49 : ตารางแสดงค่า AUC ของแบบจำลองที่สร้างจากจำนวนคุณสมบัติ 20% สำหรับชุดข้อมูลมะเร็งรังไข่

Ranker	K-NN	C4.5	NB	RF	LR	SVM	Average
GR	0.985	0.980	0.991	0.999	0.934	0.989	0.980
RF	0.989	0.980	0.988	0.999	0.993	0.977	0.988
SU	0.969	0.983	0.987	0.999	0.996	0.977	0.985
OneR	0.977	0.982	0.986	0.998	0.812	0.977	0.955
IG	0.974	0.983	0.994	0.999	0.912	0.977	0.973
Proposed Ensemble	0.980	0.987	0.989	0.999	0.995	0.977	0.988
Based Ensemble	0.969	0.983	0.987	0.999	0.921	0.977	0.973
Original Dataset	0.716	0.723	0.768	0.816	0.754	0.704	0.747

ตารางที่ 4.50 : ตารางแสดงค่า AUC ของแบบจำลองที่สร้างจากจำนวนคุณสมบัติ 40% สำหรับชุดข้อมูลมะเร็งรังไข่

Ranker	K-NN	C4.5	NB	RF	LR	SVM	Average
GR	0.993	0.980	0.992	1.000	0.920	1.000	0.981
RF	0.991	0.980	0.987	1.000	0.934	1.000	0.982
SU	0.993	0.989	0.987	1.000	0.962	1.000	0.989
OneR	0.993	0.983	0.986	1.000	0.901	1.000	0.977
IG	0.991	0.989	0.987	1.000	0.945	1.000	0.985
Proposed Ensemble	0.993	0.989	0.987	1.000	0.966	1.000	0.989
Based Ensemble	0.993	0.983	0.987	1.000	0.963	1.000	0.988
Original Dataset	0.716	0.723	0.768	0.816	0.754	0.704	0.747

ตารางที่ 4.51 : ตารางแสดงค่า AUC ของแบบจำลองที่สร้างจากจำนวนคุณสมบัติ 60% สำหรับชุดข้อมูลมะเร็งรังไข่

Ranker	K-NN	C4.5	NB	RF	LR	SVM	Average
GR	0.993	0.980	0.988	1.000	0.920	1.000	0.980
RF	0.991	0.980	0.991	1.000	0.934	1.000	0.983
SU	0.991	0.989	0.979	1.000	0.962	1.000	0.987
OneR	0.993	0.983	0.978	1.000	0.901	1.000	0.976
IG	0.991	0.983	0.979	1.000	0.945	1.000	0.983
Proposed Ensemble	0.991	0.983	0.983	1.000	0.966	1.000	0.987
Based Ensemble	0.991	0.983	0.980	1.000	0.963	1.000	0.986
Original Dataset	0.716	0.723	0.768	0.816	0.754	0.704	0.747

ตารางที่ 4.52 : ตารางแสดงค่า AUC ของแบบจำลองที่สร้างจากจำนวนคุณสมบัติ 80% สำหรับชุดข้อมูลมะเร็งรังไข่

Ranker	K-NN	C4.5	NB	RF	LR	SVM	Average
GR	0.993	0.974	0.987	1.000	0.920	1.000	0.979
RF	0.984	0.980	0.991	1.000	0.934	1.000	0.982
SU	0.991	0.983	0.978	1.000	0.962	1.000	0.986
OneR	0.984	0.983	0.975	1.000	0.901	1.000	0.974
IG	0.984	0.983	0.976	1.000	0.945	1.000	0.981
Proposed Ensemble	0.991	0.983	0.978	1.000	0.966	1.000	0.986
Based Ensemble	0.991	0.983	0.973	1.000	0.963	1.000	0.985
Original Dataset	0.716	0.723	0.768	0.816	0.754	0.704	0.747

ตารางที่ 4.53 : ตารางแสดงค่า Recall ของแบบจำลองที่สร้างจากจำนวนคุณสมบัติ 20% สำหรับชุดข้อมูลมะเร็งรังไข่

Ranker	K-NN	C4.5	NB	RF	LR	SVM	Average
GR	0.988	0.972	0.976	0.988	0.996	0.992	0.985
RFF	0.992	0.972	0.972	0.984	1.000	0.996	0.986
SU	0.976	0.976	0.976	0.976	0.988	0.980	0.979
OneR	0.980	0.968	0.976	0.976	0.976	0.980	0.976
IG	0.980	0.976	0.984	0.984	0.980	0.980	0.981
Proposed Ensemble	0.992	0.976	0.976	0.988	0.988	0.996	0.986
Based Ensemble	0.976	0.976	0.976	0.980	0.988	0.980	0.979
Original Dataset	0.727	0.731	0.587	0.751	0.774	0.747	0.720

ตารางที่ 4.54 : ตารางแสดงค่า Recall ของแบบจำลองที่สร้างจากจำนวนคุณสมบัติ 40% สำหรับชุดข้อมูลมะเร็งรังไข่

Ranker	K-NN	C4.5	NB	RF	LR	SVM	Average
GR	0.996	0.972	0.988	0.996	0.996	1.000	0.991
RFF	0.992	0.972	0.984	0.984	1.000	1.000	0.989
SU	0.996	0.980	0.980	0.996	1.000	1.000	0.992
OneR	0.996	0.976	0.972	0.996	0.996	1.000	0.989
IG	0.992	0.980	0.980	0.988	1.000	1.000	0.990
Proposed Ensemble	0.996	0.980	0.980	0.998	1.000	1.000	0.992
Based Ensemble	0.996	0.976	0.984	0.992	1.000	1.000	0.991
Original Dataset	0.727	0.731	0.587	0.751	0.774	0.747	0.720

ตารางที่ 4.55 : ตารางแสดงค่า Recall ของแบบจำลองที่สร้างจากจำนวนคุณสมบัติ 60% สำหรับชุดข้อมูลมะเร็งรังไข่

Ranker	K-NN	C4.5	NB	RF	LR	SVM	Average
GR	0.996	0.972	0.984	0.996	0.996	1.000	0.991
RFF	0.992	0.972	0.988	0.984	0.992	1.000	0.988
SU	0.992	0.980	0.964	0.992	1.000	1.000	0.988
OneR	0.996	0.976	0.964	0.992	1.000	1.000	0.988
IG	0.992	0.976	0.964	0.992	1.000	1.000	0.987
Proposed Ensemble	0.996	0.980	0.988	0.992	0.994	1.000	0.992
Based Ensemble	0.992	0.976	0.976	0.988	1.000	1.000	0.989
Original Dataset	0.727	0.731	0.587	0.751	0.774	0.747	0.720

ตารางที่ 4.56 : ตารางแสดงค่า Recall ของแบบจำลองที่สร้างจากจำนวนคุณสมบัติ 80% สำหรับชุดข้อมูลมะเร็งรังไข่

Ranker	K-NN	C4.5	NB	RF	LR	SVM	Average
GR	0.996	0.968	0.968	0.992	1.000	1.000	0.987
RFF	0.988	0.972	0.988	0.988	0.996	1.000	0.989
SU	0.992	0.976	0.960	0.988	1.000	1.000	0.986
OneR	0.988	0.976	0.957	0.988	0.996	1.000	0.984
IG	0.988	0.976	0.960	0.988	1.000	1.000	0.985
Proposed Ensemble	0.992	0.976	0.968	0.998	1.000	1.000	0.989
Based Ensemble	0.992	0.976	0.968	0.992	0.996	1.000	0.987
Original Dataset	0.727	0.731	0.587	0.751	0.774	0.747	0.720

ตารางที่ 4.57 : ตารางแสดงค่า Precision ของแบบจำลองที่สร้างจากจำนวนคุณสมบัติ 20% สำหรับชุดข้อมูลมะเร็งรังไข่

Ranker	K-NN	C4.5	NB	RF	LR	SVM	Average
GR	0.988	0.973	0.976	0.988	0.996	0.992	0.986
RFF	0.992	0.973	0.972	0.984	1.000	0.996	0.986
SU	0.977	0.977	0.977	0.976	0.988	0.980	0.979
OneR	0.981	0.968	0.977	0.976	0.977	0.980	0.977
IG	0.980	0.977	0.984	0.984	0.980	0.980	0.981
Proposed Ensemble	0.991	0.977	0.977	0.984	0.988	0.980	0.983
Based Ensemble	0.977	0.977	0.977	0.980	0.988	0.980	0.980
Original Dataset	0.761	0.731	0.656	0.761	0.732	0.751	0.732

ตารางที่ 4.58 : ตารางแสดงค่า Precision ของแบบจำลองที่สร้างจากจำนวนคุณสมบัติ 40% สำหรับชุดข้อมูลมะเร็งรังไข่

Ranker	K-NN	C4.5	NB	RF	LR	SVM	Average
GR	0.996	0.973	0.988	0.996	0.996	1.000	0.992
RFF	0.992	0.973	0.984	0.984	1.000	1.000	0.989
SU	0.996	0.981	0.980	0.996	1.000	1.000	0.992
OneR	0.996	0.977	0.972	0.996	0.996	1.000	0.990
IG	0.992	0.981	0.980	0.988	1.000	1.000	0.990
Proposed Ensemble	0.996	0.981	0.980	0.996	1.000	1.000	0.992
Based Ensemble	0.996	0.977	0.984	0.992	1.000	1.000	0.992
Original Dataset	0.761	0.731	0.656	0.761	0.732	0.751	0.732

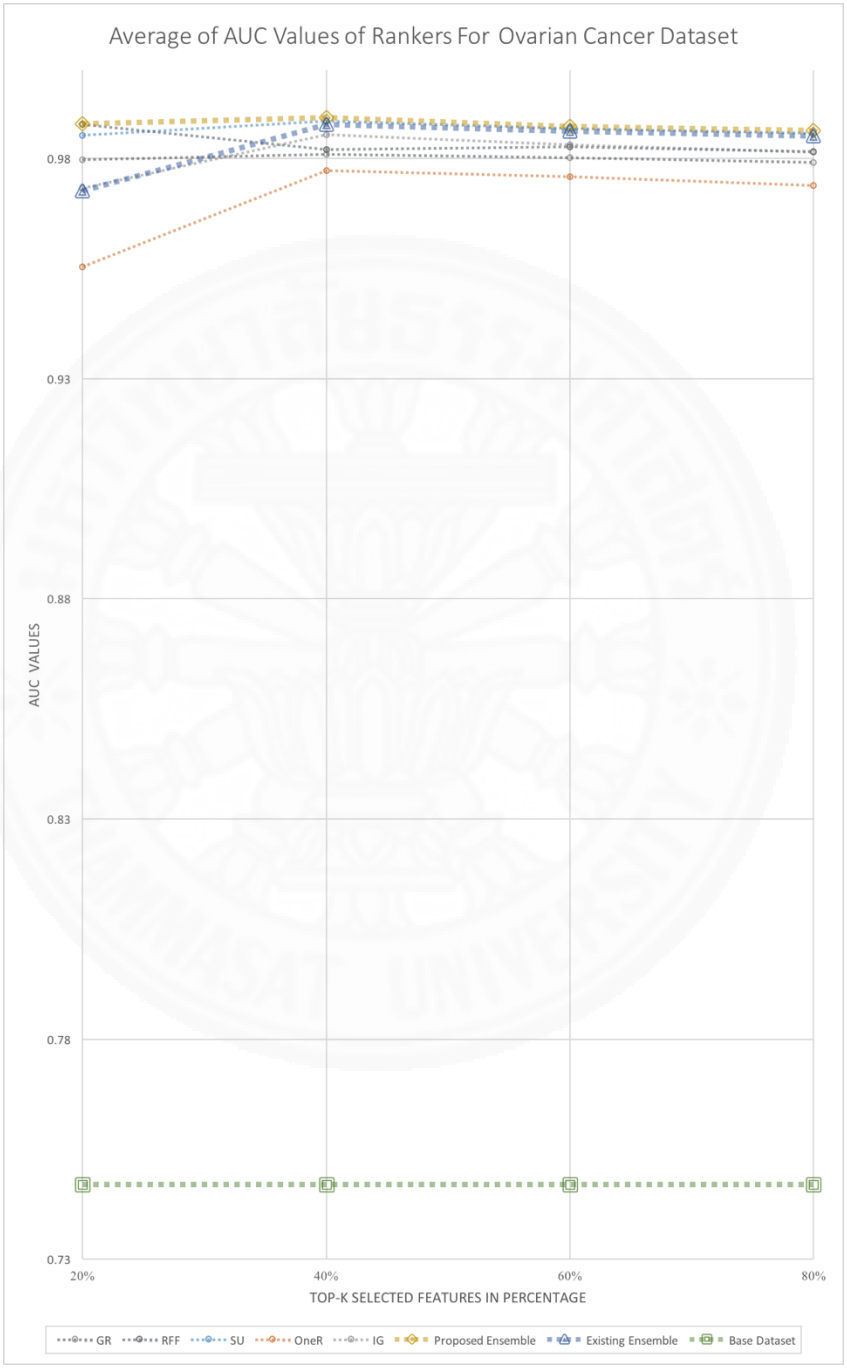
ตารางที่ 4.59 : ตารางแสดงค่า Precision ของแบบจำลองที่สร้างจากจำนวนคุณสมบัติ 60% สำหรับชุดข้อมูลมะเร็งรังไข่

Ranker	K-NN	C4.5	NB	RF	LR	SVM	Average
GR	0.996	0.973	0.984	0.996	0.996	1.000	0.991
RFF	0.992	0.973	0.988	0.984	0.992	1.000	0.988
SU	0.992	0.981	0.964	0.992	1.000	1.000	0.988
OneR	0.996	0.977	0.964	0.992	1.000	1.000	0.988
IG	0.992	0.977	0.964	0.992	1.000	1.000	0.988
Proposed Ensemble	0.992	0.981	0.968	0.996	1.000	1.000	0.990
Based Ensemble	0.992	0.977	0.976	0.988	1.000	1.000	0.989
Original Dataset	0.761	0.731	0.656	0.761	0.732	0.751	0.732

ตารางที่ 4.60 : ตารางแสดงค่า Precision ของแบบจำลองที่สร้างจากจำนวนคุณสมบัติ 80% สำหรับชุดข้อมูลมะเร็งรังไข่

Ranker	K-NN	C4.5	NB	RF	LR	SVM	Average
GR	0.996	0.969	0.968	0.992	1.000	1.000	0.988
RFF	0.988	0.973	0.988	0.988	0.996	1.000	0.989
SU	0.992	0.977	0.960	0.988	1.000	1.000	0.986
OneR	0.988	0.977	0.957	0.988	0.996	1.000	0.984
IG	0.988	0.977	0.960	0.988	1.000	1.000	0.986
Proposed Ensemble	0.992	0.977	0.964	0.992	1.000	1.000	0.988
Based Ensemble	0.992	0.977	0.968	0.992	0.996	1.000	0.988
Original Dataset	0.761	0.731	0.656	0.761	0.732	0.751	0.732

จากตารางผลลัพธ์สามารถสรุปเป็นกราฟโดยใช้ค่าเฉลี่ยของ AUC, Recall และ Precision ของแต่ละเทคนิคได้ดังนี้



ภาพที่ 4.11 : กราฟแสดงค่า **AUC** โดยเฉลี่ยสำหรับชุดข้อมูลมะเร็งรังไข่

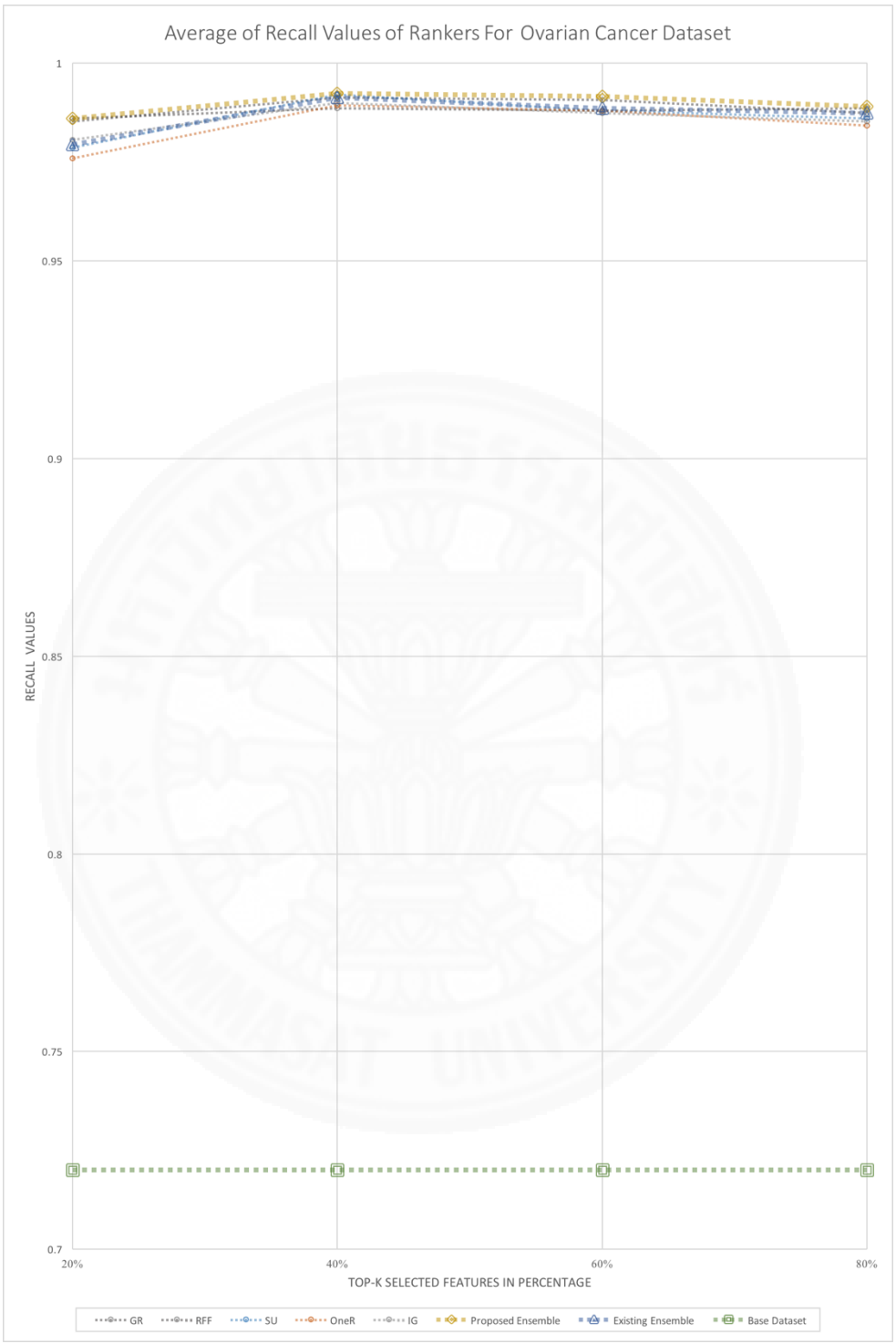
ผลลัพธ์จากกราฟค่า AUC โดยเฉลี่ยนั้น จะเห็นว่าหากนำข้อมูลมาผ่านอัลกอริทึมที่ใช้ในการคัด

เลือกคุณสมบัติไม่ว่าจะเป็นแบบรวมหรือแบบเดี่ยว จะให้ผลลัพธ์ที่ดีกว่าแบบ Base Dataset ในทุกจำนวนคุณสมบัติ จำนวนคุณสมบัติที่แต่ละอัลกอริทึมมีค่า AUC โดยเฉลี่ยใกล้เคียงกันตั้งแต่ 40% ขึ้นไป และอัลกอริทึมที่มีแนวโน้มมีประสิทธิภาพดีที่สุดโดยพิจารณาจากทุกจำนวนคุณสมบัติที่ทำการทดสอบคือ อัลกอริทึมที่ผู้วิจัยนำเสนอ

จากกราฟค่า AUC โดยเฉลี่ยพบว่าอัลกอริทึมที่ผู้วิจัยนำเสนอมีประสิทธิภาพดีที่สุดในทุกจำนวนคุณสมบัติโดยทำค่าได้สูงสุดที่จำนวนคุณสมบัติ 40% ซึ่งมีค่า AUC โดยเฉลี่ยเท่ากับ 0.989

สำหรับอัลกอริทึมต้นแบบมีค่า AUC โดยเฉลี่ยต่ำกว่าอัลกอริทึมที่นำเสนอในทุกจำนวนคุณสมบัติ โดยมีค่า AUC โดยเฉลี่ยสูงสุดที่จำนวนคุณสมบัติ 40% ซึ่งมีค่า AUC โดยเฉลี่ยเท่ากับ 0.987 อัลกอริทึม ที่พบว่ามีค่า AUC โดยเฉลี่ยต่ำที่สุดในทุกจำนวนคุณสมบัติ คือ OneR

สำหรับค่า Recall โดยเฉลี่ยของแต่ละเทคนิคสามารถสรุปเป็นกราฟได้ดังนี้



ภาพที่ 4.12 : กราฟแสดงค่า **Recall** โดยเฉลี่ยสำหรับชุดข้อมูลมะเร็งรังไข่

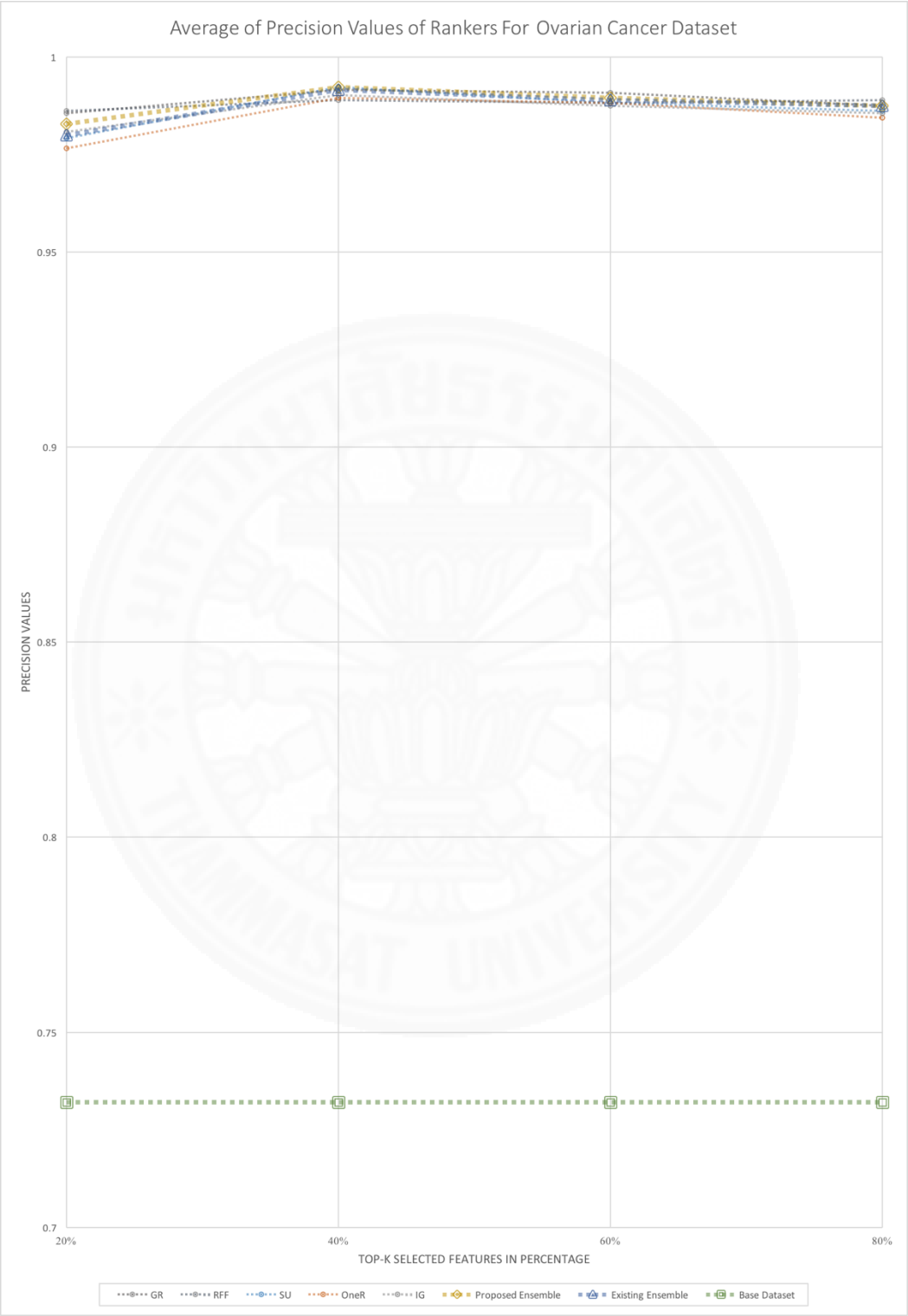
ผลลัพธ์จากกราฟค่า Recall โดยเฉลี่ยนั้น จะเห็นว่าหากนำข้อมูลมาผ่านอัลกอริทึมที่ใช้ในการคัดเลือกคุณสมบัติไม่ว่าจะเป็นแบบรวมหรือแบบเดี่ยวจะให้ผลลัพธ์ที่ดีกว่าแบบ Base Dataset ในทุก

จำนวนคุณสมบัติ จำนวนคุณสมบัติที่แต่ละอัลกอริทึมมีค่า Recall โดยเฉลี่ยใกล้เคียงกันคือ 40% และอัลกอริทึมที่มีแนวโน้มมีประสิทธิภาพดีที่สุด โดยพิจารณาจากทุกจำนวนคุณสมบัติ ที่ทำการทดสอบคือ อัลกอริทึมที่ผู้วิจัยนำเสนอ

จากกราฟค่า Recall โดยเฉลี่ยพบว่าอัลกอริทึมที่ผู้วิจัยนำเสนอมีประสิทธิภาพดีที่สุดในทุกจำนวนคุณสมบัติ โดยทำค่าได้สูงสุดที่จำนวนคุณสมบัติ 40% ซึ่งมีค่า Recall โดยเฉลี่ยเท่ากับ 0.992

สำหรับอัลกอริทึมต้นแบบมีค่า Recall โดยเฉลี่ยต่ำกว่าอัลกอริทึมที่นำเสนอในทุกจำนวนคุณสมบัติ โดยมีค่า Recall โดยเฉลี่ยสูงสุดที่จำนวนคุณสมบัติ 20% ซึ่งมีค่า Recall โดยเฉลี่ยเท่ากับ 0.991 อัลกอริทึมที่พบว่ามีค่า Recall โดยเฉลี่ยต่ำที่สุดคือ OneR

สำหรับค่า Precision โดยเฉลี่ยของแต่ละเทคนิคสามารถสรุปเป็นกราฟได้ดังนี้



ภาพที่ 4.13 : กราฟแสดงค่า **Precision** โดยเฉลี่ยสำหรับชุดข้อมูลมะเร็งรังไข่

ผลลัพธ์จากกราฟค่า Precision โดยเฉลี่ยนั้น จะเห็นว่าหากนำข้อมูลมาผ่านอัลกอริทึมที่ใช้ในการคัดเลือกคุณสมบัติไม่ว่าจะเป็นแบบรวมหรือแบบเดี่ยว จะให้ผลลัพธ์ที่ดีกว่าแบบ Base Dataset ในทุกจำนวนคุณสมบัติเช่นกัน จำนวนคุณสมบัติที่แต่ละอัลกอริทึมมีค่า Precision โดยเฉลี่ยใกล้เคียงกันคือตั้งแต่ 20% ขึ้นไป และอัลกอริทึมที่มีแนวโน้มมีประสิทธิภาพดีที่สุดโดยพิจารณาจากทุกจำนวนคุณสมบัติที่ทำการทดสอบคือ ReliefF

จากกราฟค่า Precision โดยเฉลี่ยพบว่าอัลกอริทึมที่ผู้วิจัยนำเสนอมีประสิทธิภาพดีที่สุดในจำนวนคุณสมบัติ 40% ซึ่งเท่ากับ Symmetrical Uncertainty ซึ่งมีค่า Precision โดยเฉลี่ยเท่ากับ 0.992 สำหรับอัลกอริทึมที่มีค่า Precision โดยเฉลี่ยสูงที่สุด ที่จำนวนคุณสมบัติ 20%, 60% และ 80% คือ ReliefF, Gain Ratio และ ReliefF ตามลำดับ ซึ่งมีค่า Precision โดยเฉลี่ยเท่ากับ 0.986, 0.991 และ 0.989 ในขณะที่อัลกอริทึมที่นำเสนอมีค่า Precision โดยเฉลี่ยที่จำนวนคุณสมบัติ 20%, 60% และ 80% คือ 0.983, 0.990 และ 0.988

สำหรับอัลกอริทึมต้นแบบมีค่า Precision โดยเฉลี่ยต่ำกว่าอัลกอริทึมที่นำเสนอ ในทุกจำนวนคุณสมบัติ โดยมีค่า Precision โดยเฉลี่ยสูงสุดที่จำนวนคุณสมบัติ 40% ซึ่งมีค่า Precision โดยเฉลี่ยเท่ากับ 0.992 อัลกอริทึมที่พบว่ามีค่า Precision โดยเฉลี่ยต่ำที่สุดคือ OneR

4.5 ผลการทดลองสำหรับชุดข้อมูลมะเร็งเม็ดเลือดขาว

ในการทดลองโดยใช้ข้อมูลมะเร็งเม็ดเลือดขาว ผู้วิจัยได้ทำการแบ่งจำนวนคุณสมบัติหลังจากที่ผ่านกระบวนการคัดเลือกคุณสมบัติแล้วทั้งหมด 7,130 คุณสมบัติออกเป็น 4 ส่วนคือ 20% ,40% ,60% และ 80% เพื่อทำการทดลองดังตารางที่ 4.61

ตารางที่ 4.61 : ตารางแสดงค่าจำนวนคุณสมบัติคิดเป็นเปอร์เซ็นต์สำหรับชุดข้อมูลมะเร็งเม็ดเลือดขาว

20%	40%	60%	80%	ทั้งหมด
1,426	2,852	4,278	5,704	7,130

จากนั้นทำการแบ่งจำนวนรายการ (Instances) ทั้งหมดออกเป็น 2 ส่วนคือชุดฝึกจำนวน 70% และสำหรับชุดทดสอบจำนวน 30% ซึ่งอัตราส่วนของค่าคลาสคุณสมบัติของชุดข้อมูลมะเร็งเม็ดเลือดขาวซึ่งเป็นตัวอักษร { ALL, AML } ในส่วนของชุดฝึกและชุดทดสอบ หลังจากการแบ่งจำนวนรายการแล้ว เป็นตามตารางที่ 4.62 และ 4.63

ตารางที่ 4.62 : ตารางแสดงอัตราส่วนของค่าคลาสคุณสมบัติของชุดข้อมูลมะเร็งเม็ดเลือดขาวสำหรับชุดฝึก

ค่าของคุณสมบัติคลาส	อัตราส่วน
ALL	42%
AML	58%

ตารางที่ 4.63 : ตารางแสดงอัตราส่วนของค่าคลาสคุณสมบัติของชุดข้อมูลมะเร็งเม็ดเลือดขาวสำหรับชุดทดสอบ

ค่าของคุณสมบัติคลาส	อัตราส่วน
ALL	46%
AML	54%

ผลการทดลองหลังจากทำการแบ่งจำนวนคุณสมบัติออกเป็น 4 ส่วนเป็นดังนี้

ตารางที่ 4.64 : ตารางแสดงค่า AUC ของแบบจำลองที่สร้างจากจำนวนคุณสมบัติ 20% สำหรับชุดข้อมูลมะเร็งเม็ดเลือดขาว

Ranker	K-NN	C4.5	NB	RF	LR	SVM	Average
GR	0.877	0.845	0.990	1.000	0.920	0.981	0.936
RF	0.916	0.872	0.992	1.000	0.934	0.955	0.945
SU	0.877	0.872	0.992	1.000	0.962	1.000	0.951
OneR	0.916	0.963	1.000	1.000	0.901	1.000	0.963
IG	0.877	0.872	0.992	1.000	0.945	1.000	0.948
Proposed Ensemble	0.877	0.872	0.992	1.000	0.966	1.000	0.951
Based Ensemble	0.916	0.872	0.992	1.000	0.963	1.000	0.957
Original Dataset	0.613	0.746	0.651	0.914	0.734	0.682	0.723

ตารางที่ 4.65 : ตารางแสดงค่า AUC ของแบบจำลองที่สร้างจากจำนวนคุณสมบัติ 40% สำหรับชุดข้อมูลมะเร็งเม็ดเลือดขาว

Ranker	K-NN	C4.5	NB	RF	LR	SVM	Average
GR	0.916	0.845	1.000	1.000	0.920	1.000	0.947
RF	0.877	0.872	0.980	1.000	0.934	0.936	0.933
SU	0.916	0.872	1.000	1.000	0.962	1.000	0.958
OneR	0.975	0.953	1.000	1.000	0.901	1.000	0.972
IG	1.000	0.872	1.000	1.000	0.945	1.000	0.970
Proposed Ensemble	1.000	0.872	1.000	1.000	0.966	1.000	0.973
Based Ensemble	0.916	0.872	1.000	1.000	0.963	1.000	0.959
Original Dataset	0.613	0.746	0.651	0.914	0.734	0.682	0.723

ตารางที่ 4.66 : ตารางแสดงค่า AUC ของแบบจำลองที่สร้างจากจำนวนคุณสมบัติ 60% สำหรับชุดข้อมูลมะเร็งเม็ดเลือดขาว

Ranker	K-NN	C4.5	NB	RF	LR	SVM	Average
GR	0.877	0.845	1.000	1.000	0.920	1.000	0.940
RF	0.877	0.872	0.980	0.993	0.934	0.955	0.935
SU	0.916	0.872	1.000	1.000	0.962	1.000	0.958
OneR	0.961	0.963	1.000	1.000	0.901	1.000	0.971
IG	1.000	0.872	1.000	1.000	0.945	1.000	0.970
Proposed Ensemble	1.000	0.872	1.000	0.995	0.966	1.000	0.972
Based Ensemble	0.877	0.872	1.000	1.000	0.963	1.000	0.952
Original Dataset	0.613	0.746	0.651	0.914	0.734	0.682	0.723

ตารางที่ 4.67 : ตารางแสดงค่า AUC ของแบบจำลองที่สร้างจากจำนวนคุณสมบัติ 80% สำหรับชุดข้อมูลมะเร็งเม็ดเลือดขาว

Ranker	K-NN	C4.5	NB	RF	LR	SVM	Average
GR	0.877	0.845	1.000	1.000	0.920	1.000	0.940
RF	0.877	0.872	0.980	0.997	0.934	0.955	0.936
SU	0.961	0.872	1.000	1.000	0.962	1.000	0.966
OneR	0.961	0.963	1.000	1.000	0.901	1.000	0.971
IG	1.000	0.872	1.000	1.000	0.945	1.000	0.970
Proposed Ensemble	0.961	0.872	1.000	1.000	0.973	1.000	0.968
Based Ensemble	0.877	0.872	1.000	1.000	0.963	1.000	0.952
Original Dataset	0.613	0.746	0.651	0.914	0.734	0.682	0.723

ตารางที่ 4.68 : ตารางแสดงค่า Recall ของแบบจำลองที่สร้างจากจำนวนคุณสมบัติ 20% สำหรับชุดข้อมูลมะเร็งเม็ดเลือดขาว

Ranker	K-NN	C4.5	NB	RF	LR	SVM	Average
GR	0.947	0.895	0.974	0.974	0.974	0.974	0.956
RFF	0.974	0.895	0.947	0.947	0.974	0.974	0.952
SU	0.947	0.895	0.974	0.974	0.974	1.000	0.961
OneR	0.974	0.947	1.000	0.974	1.000	1.000	0.983
IG	0.947	0.895	0.974	0.974	0.974	1.000	0.961
Proposed Ensemble	0.947	0.895	0.974	1.000	0.974	1.000	0.965
Based Ensemble	0.974	0.895	0.974	0.974	0.974	1.000	0.965
Original Dataset	0.741	0.773	0.724	0.708	0.692	0.747	0.731

ตารางที่ 4.69 : ตารางแสดงค่า Recall ของแบบจำลองที่สร้างจากจำนวนคุณสมบัติ 40% สำหรับชุดข้อมูลมะเร็งเม็ดเลือดขาว

Ranker	K-NN	C4.5	NB	RF	LR	SVM	Average
GR	0.974	0.895	1.000	0.974	1.000	1.000	0.974
RFF	0.947	0.895	0.974	0.974	0.974	0.947	0.952
SU	0.974	0.895	1.000	0.974	0.974	1.000	0.970
OneR	1.000	0.947	1.000	0.974	1.000	1.000	0.987
IG	1.000	0.895	1.000	0.974	0.947	1.000	0.969
Proposed Ensemble	1.000	0.920	1.000	0.974	0.984	1.000	0.980
Based Ensemble	0.974	0.895	1.000	0.974	1.000	1.000	0.974
Original Dataset	0.741	0.773	0.724	0.708	0.692	0.747	0.731

ตารางที่ 4.70 : ตารางแสดงค่า Recall ของแบบจำลองที่สร้างจากจำนวนคุณสมบัติ 60% สำหรับชุดข้อมูลมะเร็งเม็ดเลือดขาว

Ranker	K-NN	C4.5	NB	RF	LR	SVM	Average
GR	0.947	0.895	1.000	0.974	1.000	1.000	0.969
RFF	0.947	0.895	0.974	0.947	0.974	0.974	0.952
SU	0.974	0.895	1.000	0.974	1.000	1.000	0.974
OneR	0.974	0.947	1.000	0.974	1.000	1.000	0.983
IG	1.000	0.895	1.000	0.974	0.947	1.000	0.969
Proposed Ensemble	1.000	0.951	1.000	0.974	1.000	1.000	0.988
Based Ensemble	0.947	0.895	1.000	1.000	1.000	1.000	0.974
Original Dataset	0.741	0.773	0.724	0.708	0.692	0.747	0.731

ตารางที่ 4.71 : ตารางแสดงค่า Recall ของแบบจำลองที่สร้างจากจำนวนคุณสมบัติ 80% สำหรับชุดข้อมูลมะเร็งเม็ดเลือดขาว

Ranker	K-NN	C4.5	NB	RF	LR	SVM	Average
GR	0.947	0.895	1.000	0.947	1.000	1.000	0.965
RFF	0.947	0.895	0.974	0.974	0.974	0.974	0.956
SU	0.974	0.895	1.000	0.974	0.947	1.000	0.965
OneR	0.974	0.947	1.000	0.974	1.000	1.000	0.983
IG	1.000	0.895	1.000	0.974	0.947	1.000	0.969
Proposed Ensemble	1.000	0.951	1.000	0.974	1.000	1.000	0.988
Based Ensemble	0.947	0.895	1.000	0.947	1.000	1.000	0.965
Original Dataset	0.741	0.773	0.724	0.708	0.692	0.747	0.731

ตารางที่ 4.72 : ตารางแสดงค่า Precision ของแบบจำลองที่สร้างจากจำนวนคุณสมบัติ 20% สำหรับชุดข้อมูลมะเร็งเม็ดเลือดขาว

Ranker	K-NN	C4.5	NB	RF	LR	SVM	Average
GR	0.951	0.894	0.976	0.975	0.976	0.976	0.958
RFF	0.975	0.895	0.955	0.951	0.976	0.975	0.955
SU	0.951	0.895	0.976	0.975	0.976	1.000	0.962
OneR	0.975	0.955	1.000	0.975	1.000	1.000	0.984
IG	0.951	0.895	0.976	0.975	0.976	1.000	0.962
Proposed Ensemble	0.951	0.895	0.976	1.000	0.976	1.000	0.966
Based Ensemble	0.975	0.895	0.976	0.975	0.976	1.000	0.966
Original Dataset	0.718	0.764	0.732	0.661	0.721	0.738	0.722

ตารางที่ 4.73 : ตารางแสดงค่า Precision ของแบบจำลองที่สร้างจากจำนวนคุณสมบัติ 40% สำหรับชุดข้อมูลมะเร็งเม็ดเลือดขาว

Ranker	K-NN	C4.5	NB	RF	LR	SVM	Average
GR	0.975	0.894	1.000	0.975	1.000	1.000	0.974
RFF	0.951	0.895	0.976	0.975	0.976	0.947	0.953
SU	0.975	0.895	1.000	0.975	0.976	1.000	0.970
OneR	1.000	0.955	1.000	0.975	1.000	1.000	0.988
IG	1.000	0.895	1.000	0.975	0.955	1.000	0.971
Proposed Ensemble	1.000	0.920	1.000	0.975	0.984	1.000	0.980
Based Ensemble	0.975	0.895	1.000	0.975	1.000	1.000	0.974
Original Dataset	0.718	0.764	0.732	0.661	0.721	0.738	0.722

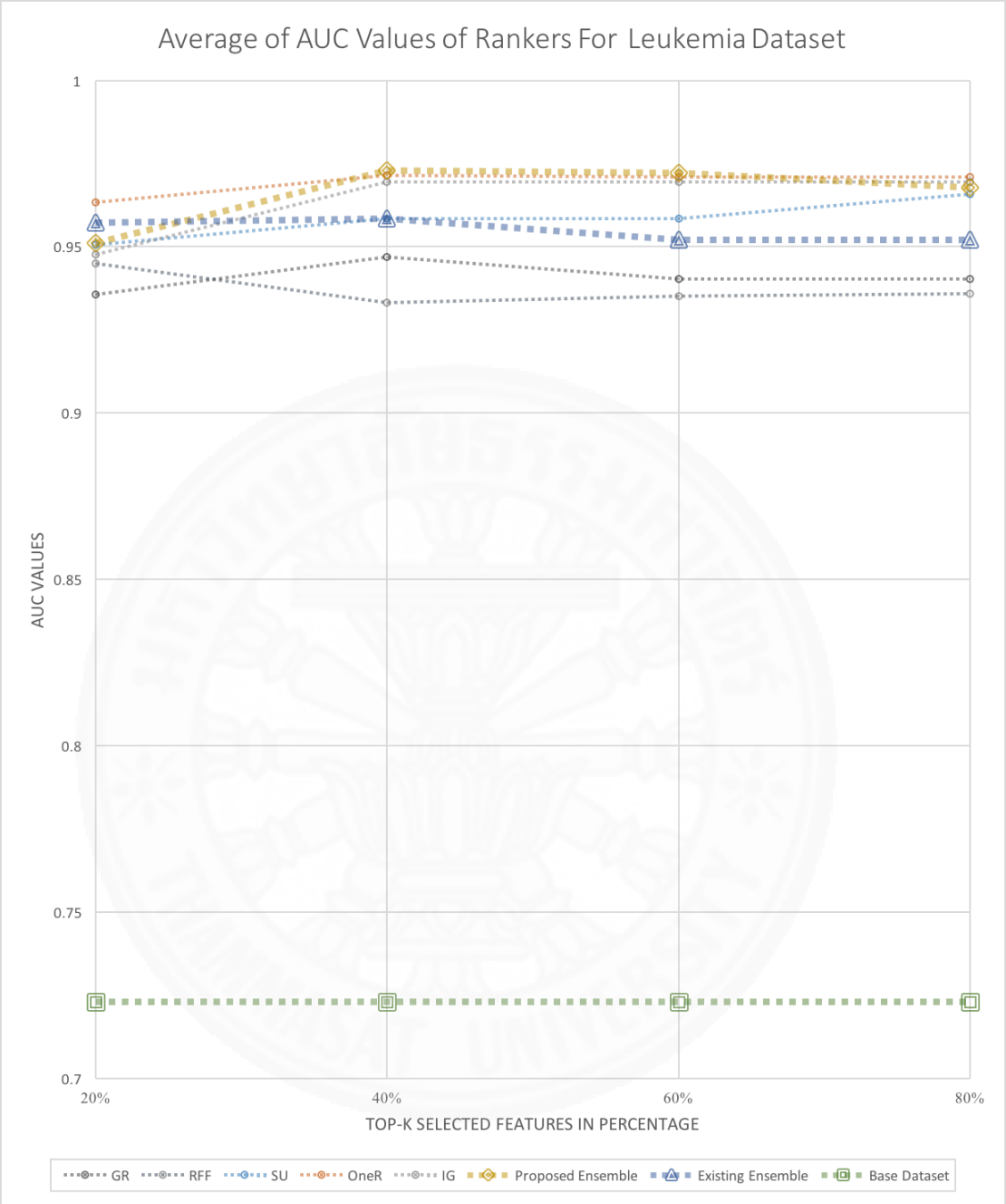
ตารางที่ 4.74 : ตารางแสดงค่า Precision ของแบบจำลองที่สร้างจากจำนวนคุณสมบัติ 60% สำหรับชุดข้อมูลมะเร็งเม็ดเลือดขาว

Ranker	K-NN	C4.5	NB	RF	LR	SVM	Average
GR	0.951	0.894	1.000	0.975	1.000	1.000	0.970
RFF	0.951	0.895	0.976	0.951	0.976	0.975	0.954
SU	0.975	0.895	1.000	0.975	1.000	1.000	0.974
OneR	0.975	0.955	1.000	0.975	1.000	1.000	0.984
IG	1.000	0.895	1.000	0.975	0.955	1.000	0.971
Proposed Ensemble	1.000	0.951	1.000	0.975	1.000	1.000	0.988
Based Ensemble	0.951	0.895	1.000	1.000	1.000	1.000	0.974
Original Dataset	0.718	0.764	0.732	0.661	0.721	0.738	0.722

ตารางที่ 4.75 : ตารางแสดงค่า Precision ของแบบจำลองที่สร้างจากจำนวนคุณสมบัติ 80% สำหรับชุดข้อมูลมะเร็งเม็ดเลือดขาว

Ranker	K-NN	C4.5	NB	RF	LR	SVM	Average
GR	0.951	0.894	1.000	0.951	1.000	1.000	0.966
RFF	0.951	0.895	0.976	0.975	0.976	0.975	0.958
SU	0.975	0.895	1.000	0.975	0.955	1.000	0.967
OneR	0.975	0.955	1.000	0.975	1.000	1.000	0.984
IG	1.000	0.895	1.000	0.975	0.955	1.000	0.971
Proposed Ensemble	1.000	0.951	1.000	0.975	1.000	1.000	0.988
Based Ensemble	0.951	0.895	1.000	0.951	1.000	1.000	0.966
Original Dataset	0.718	0.764	0.732	0.661	0.721	0.738	0.722

จากตารางผลลัพธ์สามารถสรุปเป็นกราฟโดยใช้ค่าเฉลี่ยของ AUC, Recall และ Precision ของแต่ละเทคนิคได้ดังนี้



ภาพที่ 4.14 : กราฟแสดงค่า **AUC** โดยเฉลี่ยสำหรับชุดข้อมูลมะเร็งเม็ดเลือดขาว

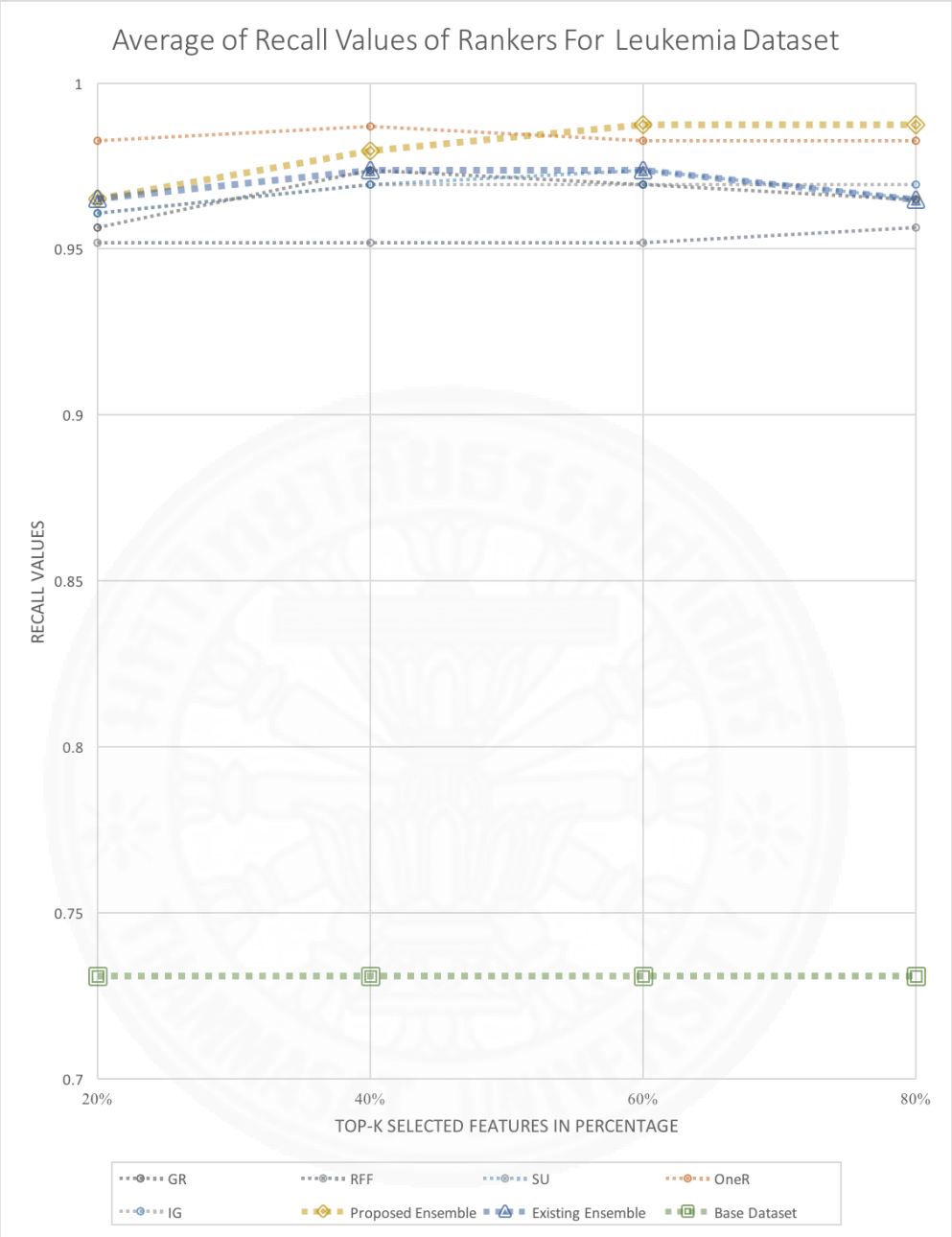
ผลลัพธ์จากกราฟค่า AUC โดยเฉลี่ยนั้น จะเห็นว่าหากนำข้อมูลมาผ่านอัลกอริทึมที่ใช้ในการคัดเลือกคุณสมบัติไม่ว่าจะเป็นแบบรวมหรือแบบเดี่ยว จะให้ผลลัพธ์ที่ดีกว่าแบบ Base Dataset ในทุกจำนวนคุณสมบัติ จำนวนคุณสมบัติที่แต่ละอัลกอริทึมมีค่า AUC โดยเฉลี่ย ใกล้เคียงกันตั้งแต่ 40%

ขึ้นไปและอัลกอริทึมที่มีแนวโน้มมีประสิทธิภาพดีที่สุดโดยพิจารณาจากทุกจำนวนคุณสมบัติที่ทำการทดสอบคือ อัลกอริทึมที่ผู้วิจัยนำเสนอ

จากกราฟค่า AUC โดยเฉลี่ยพบว่าอัลกอริทึมที่ผู้วิจัยนำเสนอมีประสิทธิภาพดีที่สุดในจำนวนคุณสมบัติ 40% และ 60% โดยทำค่าได้สูงสุดที่จำนวนคุณสมบัติ 40% ซึ่งมีค่า AUC โดยเฉลี่ยเท่ากับ 0.973 สำหรับจำนวนคุณสมบัติ 20% และ 80% อัลกอริทึมที่มีค่า AUC โดยเฉลี่ยสูงที่สุดคือ OneR ซึ่งมีค่า AUC โดยเฉลี่ยเท่ากับ 0.963 และ 0.971 ในขณะที่อัลกอริทึมที่นำเสนอมีค่า AUC โดยเฉลี่ยเท่ากับ 0.951 และ 0.968

สำหรับอัลกอริทึมต้นแบบมีค่า AUC โดยเฉลี่ยต่ำกว่าอัลกอริทึมที่นำเสนอที่จำนวนคุณสมบัติ 40%, 60% และ 80% โดยมีค่า AUC โดยเฉลี่ยสูงสุดที่จำนวนคุณสมบัติ 40% ซึ่งมีค่า AUC โดยเฉลี่ยเท่ากับ 0.958 อัลกอริทึมที่พบว่ามีค่า AUC โดยเฉลี่ยต่ำที่สุดในทุกจำนวนคุณสมบัติ คือ ReliefF

สำหรับค่า Recall โดยเฉลี่ยของแต่ละเทคนิคสามารถสรุปเป็นกราฟได้ดังนี้



ภาพที่ 4.15 : กราฟแสดงค่า **Recall** โดยเฉลี่ยสำหรับชุดข้อมูลมะเร็งเม็ดเลือดขาว

ผลลัพธ์จากกราฟค่า Recall โดยเฉลี่ยนั้น จะเห็นว่าหากนำข้อมูลมาผ่านอัลกอริทึมที่ใช้ในการคัดเลือกคุณสมบัติไม่ว่าจะเป็นแบบรวมหรือแบบเดี่ยว จะให้ผลลัพธ์ที่ดีกว่าแบบ Base Dataset ในทุกจำนวนคุณสมบัติเช่นกัน เมื่อดูจากกราฟจะเห็นว่าอัลกอริทึมที่ผู้วิจัยนำเสนอ กับ ReliefF เป็นเพียง 2 อัลกอริทึมที่มีแนวโน้มของค่า Recall โดยเฉลี่ยสูงขึ้น อัลกอริทึมที่มีแนวโน้มมีประสิทธิภาพดีที่สุดโดยพิจารณาจากทุกจำนวนคุณสมบัติที่ทำการทดสอบคือ OneR

จากกราฟค่า Recall โดยเฉลี่ยพบว่าอัลกอริทึมที่ผู้วิจัยนำเสนอมีประสิทธิภาพดีที่สุดในจำนวนคุณสมบัติ 60% และ 80% โดยทำค่า Recall ได้สูงสุดที่จำนวนคุณสมบัติ 60% ซึ่งมีค่า Recall โดยเฉลี่ยเท่ากับ 0.988 สำหรับอัลกอริทึมที่มีค่า Recall โดยเฉลี่ยสูงที่สุดที่จำนวนคุณสมบัติ 20% และ 40% คือ OneR ซึ่งมีค่า Recall โดยเฉลี่ยเท่ากับ 0.983 และ 0.987 ในขณะที่อัลกอริทึมที่นำเสนอมีค่า Recall โดยเฉลี่ยที่จำนวนคุณสมบัติ 20% และ 40% คือ 0.965 และ 0.980

สำหรับอัลกอริทึมต้นแบบมีค่า Recall โดยเฉลี่ยต่ำกว่าอัลกอริทึมที่นำเสนอ ในทุกจำนวนคุณสมบัติ โดยมีค่า Recall โดยเฉลี่ยสูงสุดที่จำนวนคุณสมบัติ 20% ซึ่งมีค่า Recall โดยเฉลี่ยเท่ากับ 0.9738 อัลกอริทึมที่พบว่ามีค่า Recall โดยเฉลี่ยต่ำที่สุดคือ ReliefF

สำหรับค่า Precision โดยเฉลี่ยของแต่ละเทคนิคสามารถสรุปเป็นกราฟได้ดังนี้

ภาพที่ 4.16 : กราฟแสดงค่า **Precision** โดยเฉลี่ยสำหรับชุดข้อมูลมะเร็งเม็ดเลือดขาว

ผลลัพธ์จากกราฟค่า Precision โดยเฉลี่ยนั้น จะเห็นว่าหากนำข้อมูลมาผ่านอัลกอริทึมที่ใช้ในการคัดเลือกคุณสมบัติไม่ว่าจะเป็นแบบรวมหรือแบบเดี่ยว จะให้ผลลัพธ์ที่ดีกว่าแบบ Base Dataset ในทุกจำนวนคุณสมบัติเช่นกัน เมื่อดูจากกราฟจะเห็นว่าอัลกอริทึมที่ผู้วิจัยนำเสนอกับ ReliefF เป็นเพียง 2 อัลกอริทึมที่มีแนวโน้มของค่า Precision โดยเฉลี่ยสูงขึ้น อัลกอริทึมที่มีแนวโน้มมีประสิทธิภาพดีที่สุดโดยพิจารณาจากทุกจำนวนคุณสมบัติที่ทำการทดสอบคือ OneR

จากกราฟค่า Precision โดยเฉลี่ยพบว่าอัลกอริทึมที่ผู้วิจัยนำเสนอมีประสิทธิภาพดีที่สุดในจำนวนคุณสมบัติ 60% และ 80% โดยทำค่าได้สูงสุดที่จำนวนคุณสมบัติ 60% ซึ่งมีค่า Precision โดยเฉลี่ยเท่ากับ 0.988 สำหรับอัลกอริทึมที่มีค่า Precision โดยเฉลี่ยสูงสุดที่จำนวนคุณสมบัติ 20% และ 40% คือ OneR ซึ่งมีค่า Precision โดยเฉลี่ยเท่ากับ 0.984 และ 0.988 ในขณะที่อัลกอริทึมที่นำเสนอมีค่า Precision โดยเฉลี่ยที่จำนวนคุณสมบัติ 20% และ 40% คือ 0.966 และ 0.980

สำหรับอัลกอริทึมต้นแบบมีค่า Precision โดยเฉลี่ยต่ำกว่าอัลกอริทึมที่นำเสนอ ในทุกจำนวนคุณสมบัติ โดยมีค่า Precision โดยเฉลี่ยสูงสุดที่จำนวนคุณสมบัติ 60% ซึ่งมีค่า Precision โดยเฉลี่ยเท่ากับ 0.9743 อัลกอริทึมที่พบว่ามีค่า Precision โดยเฉลี่ยต่ำที่สุดคือ ReliefF

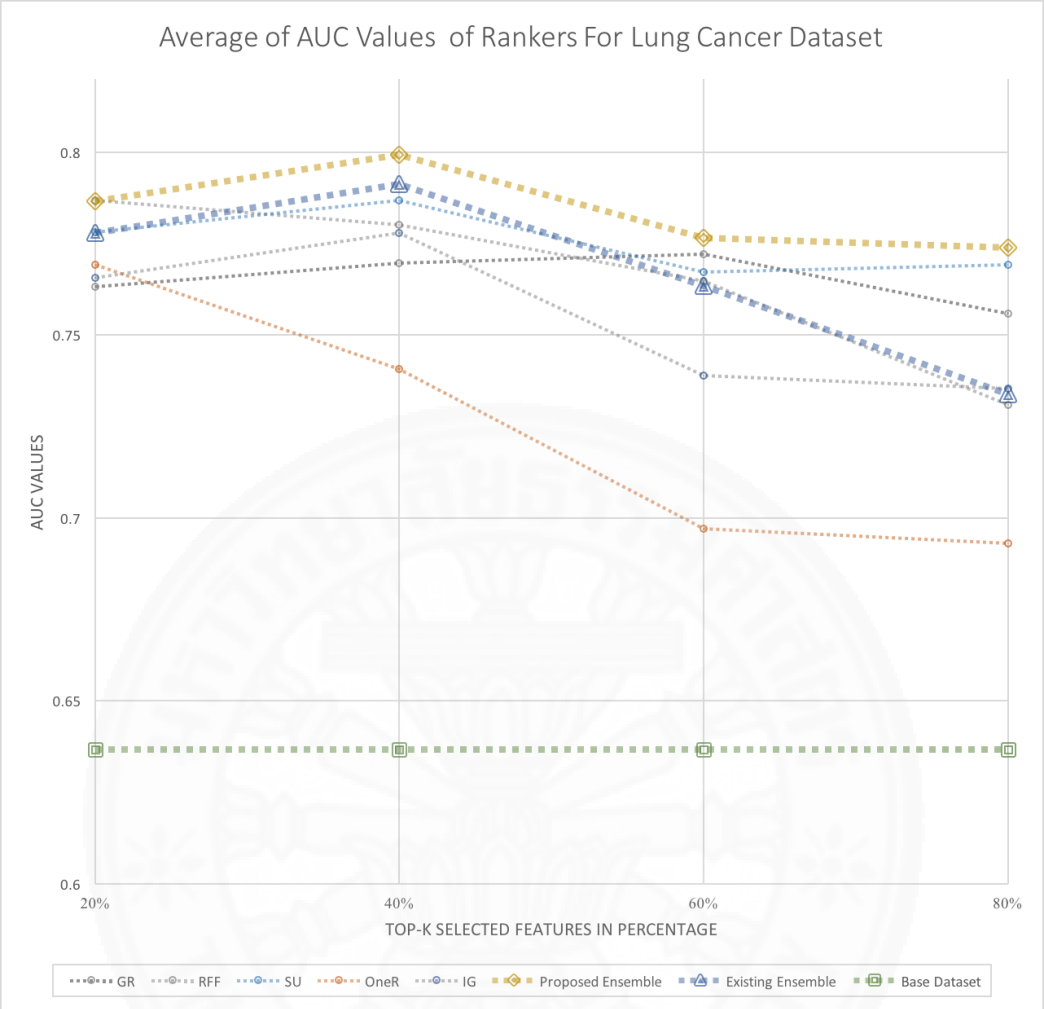
บทที่ 5
สรุปผลงานวิจัยและข้อเสนอแนะ

งานวิจัยนี้มีวัตถุประสงค์ เพื่อนำเสนอเทคนิคในการปรับปรุงอัลกอริทึมสำหรับการคัดเลือกคุณสมบัติแบบรวม (Ensemble Feature Selection) ที่มีอยู่ในปัจจุบันเพื่อให้มีประสิทธิภาพมากขึ้น โดยใช้หลักการ การพิจารณาความสำคัญ (Priority) ลำดับของคุณสมบัติ (Order) และคะแนนของแต่ละคุณสมบัติ ข้อมูลที่ใช้ในการวิจัยมาจากแหล่งข้อมูลที่มีชื่อว่า “Kent Ridge Bio - Medical” และ “Machine Learning Data Repository” ซึ่งเป็นแหล่งข้อมูลสาธารณะด้านการแพทย์ โดยชุดข้อมูลที่นำมาใช้คือ ข้อมูลการเกิดโรคมะเร็งปอด, มะเร็งต่อมน้ำเหลือง, มะเร็งเต้านม, มะเร็งรังไข่ และ มะเร็งเม็ดเลือดขาว สำหรับอัลกอริทึมการคัดเลือกคุณสมบัติที่ใช้ในงานวิจัยมีดังนี้ Symmetrical Uncertainty , ReliefF , Information Gain , Gain Ratio และ OneR ในการวัดผลจะใช้ค่า AUC, Recall และ Precision ซึ่งคำนวณได้จากแบบจำลองที่สร้างจากกระบวนการจำแนกประเภทข้อมูล (Classification) โดยเทคนิคการจำแนกประเภทข้อมูลที่นำมาใช้ในงานวิจัยนี้ได้แก่ อัลกอริทึมการจำแนกประเภทแบบการหาเพื่อนบ้านใกล้ที่สุด (K – Nearest Neighbor) , อัลกอริทึมการจำแนกประเภทแบบเบย์ (Naïve Bayes) , อัลกอริทึมการจำแนกประเภทแบบซัพพอร์ตเวกเตอร์แมชชีน (Support Vector Machines) , อัลกอริทึมการจำแนกประเภทแบบการสุ่มป่าไม้ (Random Forest) , อัลกอริทึมการจำแนกประเภทแบบถดถอยโลจิสติก (Logistic Regression) และ อัลกอริทึมการจำแนกประเภทแบบต้นไม้ตัดสินใจ (Decision Tree) ผลลัพธ์ของอัลกอริทึมที่นำเสนอจะถูกนำมาเปรียบเทียบกับอัลกอริทึมสำหรับการคัดเลือกคุณสมบัติแบบรวมที่มีอยู่เดิมที่มีอยู่ในปัจจุบัน (1) และแบบเดี่ยว (Individuals)

5.1 สรุปผลการวิจัย

จากผลการทดสอบประสิทธิภาพของอัลกอริทึมที่ผู้วิจัยนำเสนอ โดยใช้ชุดข้อมูลทดสอบจำนวน 5 ชุดพบว่าผลลัพธ์ของค่า AUC ,Recall และ Precision ของอัลกอริทึมที่ผู้วิจัยนำเสนอมีทั้งกรณีที่ดีกว่าอัลกอริทึมต้นแบบและอัลกอริทึมการคัดเลือกคุณสมบัติแบบเดี่ยว และกรณีที่มีค่าต่ำกว่าอัลกอริทึมต้นแบบและอัลกอริทึมการคัดเลือกคุณสมบัติแบบเดี่ยว

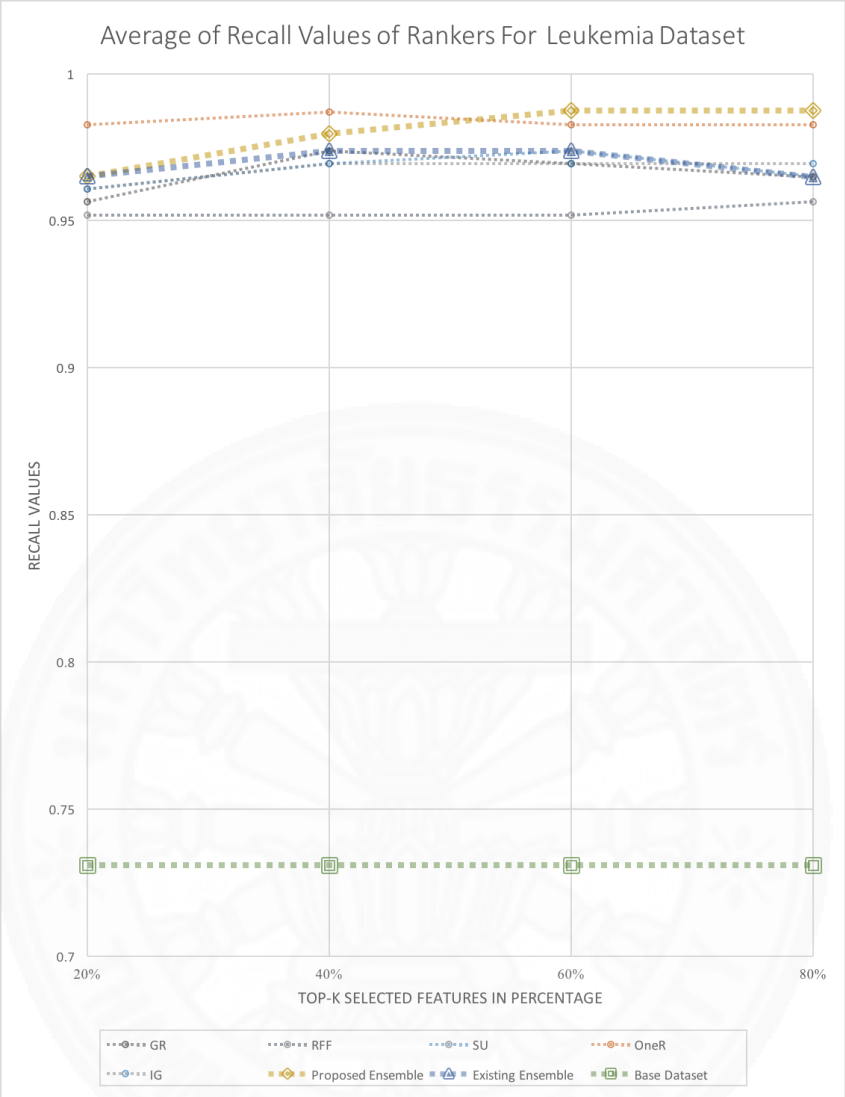
สำหรับกรณีให้ผลลัพธ์ที่ดีกว่าอัลกอริทึมต้นแบบ และอัลกอริทึมการคัดเลือกคุณสมบัติแบบเดี่ยว นั้นพบว่าส่วนมากจะเกิดขึ้นสำหรับค่า AUC โดยชุดข้อมูลที่เห็นได้ค่อนข้างชัดเจนคือชุดข้อมูลมะเร็งปอด และ มะเร็งต่อมน้ำเหลือง



ภาพที่ 5.1 : กราฟแสดงค่า **AUC** โดยเฉลี่ยสำหรับชุดข้อมูลมะเร็งปอด

ซึ่งถ้าหากพิจารณาที่ลักษณะของข้อมูลทั้ง 2 ชุดพบว่าจำนวนคุณสมบัติของมะเร็งปอดมีค่อนข้างน้อย และมีปริมาณค่าของคุณสมบัติคลาสน้อยเช่นกัน ในทางกลับกันจำนวนคุณสมบัติของมะเร็งต่อมน้ำเหลืองมีค่อนข้างมากและ มีปริมาณค่าของคุณสมบัติคลาสเยอะเช่นกัน ซึ่งถ้าหากดูที่ลักษณะของข้อมูลก็จะสังเกตเห็นว่า หากลักษณะข้อมูลที่นำมาใช้ในการทดลอง มีลักษณะจำนวนคุณสมบัติและปริมาณค่าของคุณสมบัติคลาสเป็นไปในทางเดียวกัน จะทำให้ค่าของ AUC ที่ได้จากอัลกอริทึมที่ผู้วิจัยนำเสนอมีแนวโน้มที่ดีกว่าอัลกอริทึมแบบอื่นที่นำมาทดลอง

สำหรับกรณีให้ผลลัพธ์ต่ำกว่าอัลกอริทึมต้นแบบและอัลกอริทึมการคัดเลือกคุณสมบัติแบบเดี่ยว นั้นพบว่าโดยส่วนมากจะเกิดขึ้นสำหรับค่า Recall และ Precision เช่นกรณีของชุดข้อมูลค่า Precision ของชุดข้อมูลมะเร็งรังไข่ และค่า Recall ของชุดข้อมูลมะเร็งเม็ดเลือดขาว



ภาพที่ 5.2 : กราฟแสดงค่า **Recall** โดยเฉลี่ยสำหรับชุดข้อมูลมะเร็งเม็ดเลือดขาว

ซึ่งถ้าหากพิจารณาในลักษณะเดียวกันกับกรณีให้ผลลัพธ์ที่ดีกว่าอัลกอริทึมต้นแบบ และ อัลกอริทึมการคัดเลือกคุณสมบัติแบบเดี่ยว พบว่าจำนวนคุณสมบัติของชุดข้อมูลมะเร็งรังไข่ และมะเร็งเม็ดเลือดขาวมีปริมาณมาก แต่ปริมาณค่าของคุณสมบัติคลาสน้อย จึงเป็นส่วนที่ทำให้ ค่าของ Precision และ Recall จากผลลัพธ์ของอัลกอริทึมที่ผู้วิจัยนำเสนอมีค่าน้อยกว่าอัลกอริทึมต้นแบบ และอัลกอริทึมการคัดเลือกคุณสมบัติแบบเดี่ยวในบางจำนวนคุณสมบัติ (Top – K Selected Features) แต่เมื่อจำนวนคุณสมบติที่ใช้ในการพิจารณามีมากขึ้น ก็จะทำให้ค่าของ Precision และ Recall จากผลลัพธ์ของอัลกอริทึมที่ผู้วิจัยนำเสนอมีค่าสูงขึ้น

หากพิจารณาเปรียบเทียบประสิทธิภาพระหว่าง อัลกอริทึมที่ผู้วิจัยนำเสนอ กับอัลกอริทึมที่นำมาต่อยอด (1) ในทุกชุดข้อมูล พบว่าโดยส่วนมาก อัลกอริทึมที่ผู้วิจัยนำเสนอ

จะให้ผลลัพธ์ในส่วนของค่า AUC , Precision และ Recall มากกว่า จึงทำให้เห็นว่าการนำเอาค่าความสำคัญ (Priority Value) มาเป็นปัจจัยหนึ่ง ในการทำงานของอัลกอริทึมการคัดเลือกคุณสมบัติแบบรวม มีผลทำให้ประสิทธิภาพในการจัดลำดับคุณสมบัติของชุดข้อมูลที่ผู้วิจัยนำมาทดสอบดีขึ้น

5.2 ข้อเสนอแนะ

ข้อมูลที่นำมาใช้สำหรับอัลกอริทึมการคัดเลือกคุณสมบัติแบบรวม โดยพิจารณาความสำคัญจากลำดับของคุณสมบัติ ควรเป็นข้อมูลที่มีลักษณะเป็นธรรมชาติ ไม่มีการจัดเรียงหรือตัดแต่งข้อมูลมาก่อน เนื่องจากหากมีการจัดเรียงข้อมูลมาก่อน จะทำให้ไม่สามารถเห็นประสิทธิภาพของอัลกอรึทึมอย่างเด่นชัดและข้อมูลที่นำมาใช้กับอัลกอรึทึม ไม่ควรมีค่าว่างหรือค่าขยะมากจนเกิดไป เพราะอาจทำให้เกิดข้อผิดพลาดในการทำงานของอัลกอริทึม

5.3 งานวิจัยในอนาคต

งานวิจัยนี้ได้เลือกใช้ชุดข้อมูลการทดลอง 5 ชุดได้แก่ ชุดข้อมูลการเกิดโรคมะเร็งปอด มะเร็งต่อมน้ำเหลือง, มะเร็งเต้านม, มะเร็งรังไข่และมะเร็งเม็ดเลือดขาว ซึ่งชุดข้อมูลเหล่านี้เป็นชุดข้อมูลที่เกี่ยวข้องทางด้านการแพทย์เพียงอย่างเดียว ดังนั้นจึงควรนำเอาชุดข้อมูลที่เกี่ยวข้องกับด้านอื่นๆ เช่นตลาดหุ้นหรือการศึกษา มาทำการทดสอบเพิ่มเติม

การทำงานของอัลกอริทึมที่นำเสนอ ยังไม่ได้มีการนำค่าทางสถิติของชุดข้อมูลเพื่อใช้ในการคำนวณหาค่าความถี่ ดังนั้นงานวิจัยในอนาคต อาจนำค่าทางสถิติของชุดข้อมูลที่ใช้ในการทดสอบเช่น ส่วนเบี่ยงเบนมาตรฐาน, ค่ามัธยฐาน เป็นต้น มาเป็นปัจจัยในการคำนวณหาค่าความถี่สำหรับใช้ในการเรียงลำดับด้วย

รายการอ้างอิง

วิทยานิพนธ์

1. Vege H. Ensemble of Feature Selection Techniques for High Dimensional Data [master's thesis]. [Hamiltion]: Waikato University; 2012.

วารสาร

2. Silwattananusarn T, Kanarkard W, Tuamsuk K. Enhanced Classification Accuracy for Cardiotocogram Data with Ensemble Feature Selection and Classifier Ensemble. J Comput Commun. 2016; (6):20-35.

3. Zilin Z, Hongjun Z, Rui Z, Youlian Z. Hybrid Feature Selection Method based on Rough Conditional Mutual Information and Naïve Bayesian Classifier. J Comput Commun. 2014; (12):11-17.

4. Kashif J, Haroon AB, Mehreen S. Feature Selection based on Class-Dependent Densities for High Dimensional Binary Data. IEEE Trans Knowl Data Eng. 2012; (24): 45-49.

5. Qinbao S, Jingjie N, Guangtao W. Fast Clustering - Based Feature Subset Selection Algorithm for High - Dimensional Data. IEEE Trans Knowl Data Eng. 2013; (22):31- 37.

6. Dainotti A, Pescapé A, Sansone C. Early Classification of Network Traffic through Multi - classification. Traffic Monitoring and Analysis. Springer : Berlin / Heidelberg. 2011; (13):122–135.

7. Wang H, Taghi M, Gao K. High - Dimensional Software Engineering Data and Feature Selection. 21st IEEE International Conference on Tools with Artificial Intelligence. 2009; (25):7-19.

8. Kexin Z, Jian Y. A Cluster-Based Sequential Feature Selection Algorithm. IEEE Trans Knowl Data Eng. 2013; (2):18-36.

9. Sutha K, Temilselvi J. A Review of Feature Selection Algorithms for Data Mining Techniques. IJCSE. 2016; (7):65-68.

10. Wang P, Sanin C, Szczerbicki E. Prediction based on Integration Decisional DNA and a Feature Selection Algorithm relief-F. Cyber System. 2013; 173–183.

11. Fahad A, Tari Z, Khalil I, Habib I, Alnuwiri H. Toward efficient and scalable feature selection approach for internet traffic classification. IEEE Comput Netw Conference. 2013; (57):2040–2057.

12. Osanaiye O, Raymond C, Dehghantanha A, Zheng X, Mqhele D. Ensemble - based multi - filter feature selection method for DDoS detection in cloud computing. J Wirel Commum Netw. 2016; (13):131-136.

13. Sujatha M, Devi L. Feature Selection Techniques using for High Dimensional Data in Machine Learning. IJERT. 2013; (2):97-102

14. Zhao Z, Liu H. On Similarity Preserving Feature Selection. IEEE Trans Knowl Data Eng. 2013; (25):36-41.

15. Srirama SN, Jakovits P, Vainikko E. Adapting scientific computing problems to clouds using MapReduce. Future Gener. J Comput Syst. 2012; (28):184–192.

16. Szabó G, Veres A, Malomsoky S, Gódor I, Molnár S. Traffic Classification over Gbit Speed with Commodity Hardware. IEEE J Commun Syst Software. 2010; (17):198-206.

17. DeDonato W, Pescapé A, Dainotti A. Traffic Identification Engine : An Open Platform for Traffic Classification. IEEE Netw. 2014; (28):56–64.

18. Wang Y. Fisher Scoring: An interpolation family and its Monte Carlo implementations. J Comput Stat Data Anal. 2010; (54):170–175.

19. Dahiya S, Singh NP. A Rank Aggregation Algorithm for Ensemble of Multiple

Feature Selection Techniques in Credit Risk Evaluation. IJARAI. 2016; (5):26-31.

ฐานข้อมูลบนอินเตอร์เน็ต

20. Weka Software Tools & Documentation [Internet]. University of Waikato. [1991] - [cited 2017 Jan 1]. Available from: http://www.cs.waikato.ac.nz/ml/ weka/

21. Kent Ridge Bio – Medical Repository Data & Documentation [Internet]. ELVIRA. [2005] - [cited 2017 Jan 3]. Available from: http://datam.i2r.astar.edu.sg/data sets/krbd/

22. Machine Learning Repository Data [Internet]. Mldata. [2012] - [cited 2017 Jan 3]. Available from: http://mldata.org/repository/data/