การเปรียบเทียบวิธีการเรียนรู้ด้วยเครื่องสำหรับการจำแนกระดับการเป็นมะเร็งปอดจากรหัสพันธุกรรม

A COMPARISON OF MACHINE LEARNING METHODS FOR CLASSIFICATION THE LEVEL OF LUNG CANCER FROM DNA

รักษ์ธรรม ลีพึ่งธรรม

วราพร โพธิโชติ

วิชญา เคารพ

ปัญหาพิเศษนี้เป็นส่วนหนึ่งของการศึกษาตามหลักสูตร

ปริญญาวิทยาศาสตรบัณฑิต (สถิติประยุกต์)

ภาควิชาสถิติ คณะวิทยาศาสตร์

สถาบันเทคโนโลยีพระจอมเกล้าเจ้าคุณทหารลาดกระบัง

ปีการศึกษา 2565

A COMPARISON OF MACHINE LEARNING METHODS FOR CLASSIFICATION THE LEVEL OF LUNG CANCER FROM DNA

RUKTHAM LEEPUNGTHAM

WARAPORN PHOTICHOT

WICHAYA KAOROP

A SPECIAL PROBLEM SUBMITTED IN PARTIAL FULFILLMENT OF

THE REQUIREMENT FOR

THE DEGREE OF BACHELOR OF SCIENCE (APPLIED STATISTICS)

DEPARTMENT OF STATISTICS, SCHOOL OF SCIENCE

KING MONGKUT'S INSTITUTE OF TECHNOLOGY LADKRABANG

ACADEMIC YEAR 2022

หัวข้อปัญหาพิเศษ	การเปรียบเทียบวิธีการเรียนรู้ด้วยเครื่องสำหรับการจำแนกระดับการเป็นมะเร็งปอดจากรหัสพันธุกรรม A Comparison of Machine Learning Methods for Classification the Level of Lung Cancer from DNA
ชื่อนักศึกษา	นายรักษ์ธรรม ลีพึ่งธรรม รหัสนักศึกษา 62050816 นางสาววราพร โพธิโชติ รหัสนักศึกษา 62050827 นางสาววิชญา เคารพ รหัสนักศึกษา 62050832
ปริญญา	วิทยาศาสตรบัณฑิต (สถิติประยุกต์)
ภาควิชา	สถิติ
ปีการศึกษา	2565
อาจารย์ที่ปรึกษา	รศ.ดร.อัชฌา อระวีพร

คณะวิทยาศาสตร์ สถาบันเทคโนโลยีพระจอมเกล้าเจ้าคุณทหารลาดกระบัง (สจล.) อนุมัติให้ปัญหาพิเศษนี้เป็นส่วนหนึ่งของการศึกษาตามหลักสูตรปริญญาวิทยาศาสตรบัณฑิต (สถิติประยุกต์) ประจำปีการศึกษา 2565

คณะกรรมการสอบ	ลายมือชื่อ
ผศ.ชูใจ คูหารัตนไชย ประธานกรรมการ	
ผศ.ดร.พรรณทิพา วาณิชย์จิรัฐติกาล กรรมการ	
รศ.ดร.อัชฌา อระวีพร กรรมการและอาจารย์ที่ปรึกษา	

หัวข้อปัญหาพิเศษ การเปรียบเทียบวิธีการเรียนรู้ด้วยเครื่องสำหรับการจำแนกระดับการเป็นมะเร็งปอดจากรหัสพันธุกรรม

ชื่อนักศึกษา นายรักษ์ธรรม ลีพึ่งธรรม รหัสนักศึกษา 62050816
นางสาววราพร โพธิโชติ รหัสนักศึกษา 62050827
นางสาววิชญา เคารพ รหัสนักศึกษา 62050832

ปริญญา วิทยาศาสตรบัณฑิต (สถิติประยุกต์)

ภาควิชา สถิติ

คณะ วิทยาศาสตร์

มหาวิทยาลัย สถาบันเทคโนโลยีพระจอมเกล้าเจ้าคุณทหารลาดกระบัง (สจล.)

ปีการศึกษา 2565

อาจารย์ที่ปรึกษา รศ.ดร.อัชฌา อระวีพร

บทคัดย่อ

ปัญหาพิเศษนี้ มีวัตถุประสงค์เพื่อศึกษาวิธีการจำแนกระดับของผู้ป่วยมะเร็งปอด และเปรียบเทียบประสิทธิภาพการจำแนกระดับการเป็นมะเร็งปอดจากรหัสพันธุกรรมด้วยวิธีการเรียนรู้ด้วยเครื่อง ประกอบด้วย 6 วิธี ดังนี้ วิธีต้นไม้ตัดสินใจ วิธีนาอีฟเบย์ วิธีเพื่อนบ้านใกล้เคียง k อันดับ วิธีซัพพอร์ตเวกเตอร์แมชชีน วิธีโครงข่ายประสาทเทียม และวิธีป่าสุ่ม โดยข้อมูลของผู้ป่วยมะเร็งปอดมีจำนวนทั้งหมด 197 คน ผู้ป่วยแต่ละคนมีรหัสพันธุกรรม 1,000 รหัส เป็นตัวแปรอิสระ และ ตัวแปรตาม คือ ประเภทของโรคมะเร็งปอด จำนวน 4 กลุ่ม คือ Small Cell Lung Cancer (SCLC) : Oat cell lung cancer , Non-Small Cell lung cancer (NSCLC) : Adenocarcinoma , Non-Small Cell lung cancer (NSCLC) : Squamous Cell carcinoma , Non-Small Cell lung cancer (NSCLC) : Large Cell lung cancer ซึ่งข้อมูลที่นำมาวิเคราะห์มีลักษณะเป็นข้อมูลมิติขั้นสูง ข้อมูลมีความสัมพันธ์เชิงเส้นพหุ ข้อมูลมีค่านอกเกณฑ์ และ ข้อมูลไม่สมดุล วิธีการดำเนินงานวิจัย เนื่องจากข้อมูลมีลักษณะเป็นข้อมูลไม่สมดุล จึงแบ่งชุดข้อมูลในการวิเคราะห์เป็น 2 ชุด คือ ข้อมูลชุดดั้งเดิม และ ชุดข้อมูลที่ปรับให้สมดุลแล้ว ในการวิเคราะห์จะแบ่งอัตราส่วนของชุดข้อมูลฝึกฝนและชุดข้อมูลทดสอบเป็น 70:30 โดยทำการสุ่มรหัสพันธุกรรมในจำนวน 200 400 600 800 และ 1,000 รหัส จำนวนละ 1,000 รอบ เพื่อหาค่าเฉลี่ยร้อยละความถูกต้อง (Accuracy) ค่าเฉลี่ยร้อยละความแม่นยำ (Precision) และ ค่าเฉลี่ยร้อยละความระลึก (Recall) โดยใช้โปรแกรม R Studio ในการวิเคราะห์ ผลการศึกษาพบว่า ในชุดข้อมูลดั้งเดิม วิธีซัพพอร์ตเวกเตอร์แมชชีน ที่จำนวนตัวแปรอิสระเท่ากับ 200 ให้ค่าเฉลี่ยร้อยละความถูกต้องสูงที่สุด คือ 96.6407 และ ในชุดข้อมูลที่ปรับให้สมดุลแล้ว วิธีป่าสุ่ม ที่จำนวนตัวแปรอิสระเท่ากับ 600 ให้ค่าเฉลี่ยร้อยละความถูกต้องสูงที่สุด คือ 99.6443

คำสำคัญ : ข้อมูลไม่สมดุล ซัพพอร์ตเวกเตอร์แมชชีน ป่าสุ่ม มะเร็งปอด รหัสพันธุกรรม วิธีการเรียนรู้ด้วยเครื่อง

Title	A Comparison of Machine Learning Methods for Classification the Level of Lung Cancer from DNA
Students	Mr. Ruktham Leepungtham Student ID 62050816
	Miss Waraporn Photichot Student ID 62050827
	Miss Wichaya Kaorop Student ID 62050832
Degree	Bachelor of Science (Applied Statistics)
Department	Statistics
School	Science
University	King Mongkut's Institute of Technology Ladkrabang (KMITL)
Academic Year	2022
Advisor	Assoc. Prof. Dr.Autcha Araveeporn

Abstract

The objectives of this special problem were to study methods for classifying lung cancer patients and to compare the efficiency of classifying lung cancer from genetic code using machine learning methods, consisting of 6 methods as follows: Decision tree, Naïve Bayes, K-Nearest Neighbors, Support Vector Machine, Neural Network, and Random Forest. Data of lung cancer patients, there were 197 patients in total; each patient had 1,000 genetic codes as independent variables, and dependent variables were the types of lung cancer in 4 groups: Small Cell Lung Cancer (SCLC): Oat cell lung cancer, Non-Small Cell lung cancer (NSCLC): Adenocarcinoma, Non-Small Cell lung cancer (NSCLC): Squamous Cell Carcinoma, Non-Small Cell lung cancer (NSCLC): Large Cell lung cancer. The data to be analyzed is characterized by high-dimensional, multicollinearity, outlier, and imbalanced data. For this research method, the data is an imbalanced pattern. Therefore, the data set for analysis was divided into two sets, the original and the balanced data set. In this analysis, the ratio of the training data set and the test data set was 70:30 by randomly the genetic codes in the amount of 200, 400, 600, 800, and 1,000 codes 1,000 times. The R Studio program analyzed the average percentage of accuracy, the average percentage of precision, and the average percentage of recall. The study found that the support vector machine is a performance method in the original dataset for 200 independent variables. The average percentage of accuracy is the highest at 96.6407. In the balanced dataset, the random forest at the number of independent variables equals 600, and the highest average percentage of accuracy is 99.6443.

Keywords : Imbalance data, Support Vector Machine, Random Forest, Lung Cancer, DNA, Machine Learning

กิตติกรรมประกาศ

ปัญหาพิเศษฉบับนี้สำเร็จลุล่วงไปได้ด้วยดี มีความละเอียดและความถูกต้องในเนื้อหา เนื่องด้วยได้รับความกรุณาจาก รศ.ดร.อัชฌา อระวีพร อาจารย์ที่ปรึกษาปัญหาพิเศษที่ให้คำปรึกษา คำแนะนำ และหนังสืออ้างอิง ที่ใช้ในการวิเคราะห์ข้อมูลและตรวจทานแก้ไขความถูกต้องตลอดจนติดตามผลงานทุกขั้นตอน และการดำเนินงานจนกระทั่งเสร็จสมบูรณ์จึงขอขอบพระคุณด้วยความเคารพเป็นอย่างสูงไว้ ณ ที่นี้ด้วย

ขอขอบพระคุณ ผศ.ชูใจ คูหารัตนไชย และ ดร.พรรณทิพา วาณิชย์จิรัฐติกาล ที่เป็นอาจารย์คณะกรรมการ ซึ่งได้กรุณาตรวจแก้ไขปัญหาพิเศษฉบับนี้ให้ถูกต้องสมบูรณ์ยิ่งขึ้น ให้คำแนะนำแนวคิด วิธีการและข้อมูลทุกอย่างที่เป็นประโยชน์เกี่ยวกับปัญหาพิเศษฉบับนี้ทั้งหมด

ขอขอบพระคุณคณาจารย์และบุคลากรภาควิชาสถิติประยุกต์ทุกท่าน ที่ได้ประสิทธิ์ประสาทวิชาความรู้และคำแนะนำที่มีประโยชน์ รวมถึงการให้ความช่วยเหลือในเรื่องต่างๆ มาโดยตลอด

สุดท้ายนี้ขอขอบคุณบิดา มารดา ของผู้จัดทำ ซึ่งสนับสนุนในด้านกำลังทรัพย์ และให้กำลังใจเสมอมา รวมถึงเพื่อนๆ ทุกคนที่ให้คำปรึกษา คำแนะนำ และคอยช่วยเหลือในการทำงานมาโดยตลอดจนปัญหาพิเศษฉบับนี้สำเร็จลุล่วงด้วยดี

รักษ์ธรรม ลีพึ่งธรรม
วราพร โพธิโชติ
วิชญา เคารพ

สารบัญ

หน้า

สารบัญ (ต่อ)

หน้า

สารบัญ (ต่อ)

หน้า

สารบัญตาราง

สารบัญตาราง (ต่อ)

ตารางที่ **หน้า**

สารบัญตาราง (ต่อ)

ตารางที่ **หน้า**

สารบัญรูป

สารบัญรูป (ต่อ)

รูปที่ **หน้า**

บทที่ 1
บทนำ

1.1 ความเป็นมาและความสำคัญของปัญหา

ในปัจจุบันเทคโนโลยีมีบทบาทสำคัญอย่างมากในการดำเนินชีวิตในทุกๆมิติ ไม่ว่าจะเป็นการติดต่อสื่อสาร การศึกษา การประกอบธุรกิจต่างๆ โดยเฉพาะในเรื่องของการรักษาโรคภัย ซึ่งจะต้องใช้ระบบคอมพิวเตอร์ในการช่วยวิเคราะห์ เก็บข้อมูลในปริมาณมาก ไม่ว่าจะเป็นข้อมูลประเภทข้อความ รูปภาพ วิดีโอ และเสียง ในการรักษาโรคภัยนั้นมีปัจจัยสำคัญที่ทำให้รู้สาเหตุของการเกิดโรคได้หลายอย่างเช่น รหัสพันธุกรรม (Genetic code) คือชุดข้อมูลทางพันธุกรรมที่ประกอบไปด้วยลำดับเบสบน DNA (Deoxyribonucleic acid) ทำหน้าที่ เก็บ ควบคุม และถ่ายทอดลักษณะทางพันธุกรรมของสิ่งมีชีวิตจากรุ่นสู่รุ่น การตรวจสอบรหัสพันธุกรรมนั้นใช้เพื่อวิเคราะห์ถึงลักษณะความคล้ายคลึงของสิ่งมีชีวิตว่ามีความสัมพันธ์กันในด้านพันธุกรรมอย่างไร เนื่องจากรหัสพันธุกรรมนั้นสามารถบ่งบอกได้ถึงลักษณะในส่วนที่เหมือนกัน จึงนำรหัสพันธุกรรมมาวิเคราะห์เกี่ยวกับการเป็นโรคภัยได้ ทำให้ต้องมีการเก็บข้อมูลอย่างต่อเนื่อง ข้อมูลจึงมีปริมาณที่เพิ่มขึ้นตลอดเวลา ทำให้เกิดข้อมูลที่มีจำนวนมากมายมหาศาล

การเรียนรู้ด้วยเครื่อง (Machine Learning) เป็นการสอนให้ระบบคอมพิวเตอร์ทำการเรียนรู้ได้ด้วยตนเอง จากการใช้ข้อมูลที่ป้อนให้ มีสองรูปแบบ คือ การเรียนรู้โดยไม่มีผู้สอน (Unsupervised) และ การเรียนรู้โดยมีผู้สอน (Supervised) การเรียนรู้โดยไม่มีผู้สอน คือการให้คอมพิวเตอร์ทำการเรียนรู้ด้วยตัวเอง โดยไม่ต้องตั้งค่ากำหนดเป้าหมายของแต่ละข้อมูล ระบบสามารถนำไปวิเคราะห์และสร้างแบบแผนจากข้อมูลที่ได้รับเข้าไป สามารถนำไปประยุกต์กับข้อมูลได้ 2 รูปแบบคือ การแบ่งกลุ่ม (Clustering) และการหารูปแบบความสัมพันธ์ (Association) ในความคล้ายคลึงหรือความแตกต่างของข้อมูล การเรียนรู้โดยมีผู้สอน คือการที่ให้คอมพิวเตอร์เรียนรู้ด้วยตัวเอง สามารถหาคำตอบได้เองหลังจากเรียนรู้ชุดข้อมูลตัวอย่างไปแล้วระยะหนึ่งแล้วจึงนำชุดข้อมูลอื่นมาทดสอบความแม่นยำในการทำนาย ซึ่งสามารถนำไปประยุกต์ใช้ในการทำนายได้ คือ การวิเคราะห์การถดถอย (Regression) โดยผลลัพธ์ที่ได้จะออกมาเป็นตัวเลข และการจำแนกประเภทของข้อมูล (Classification)

การจำแนกประเภทข้อมูล เป็นกระบวนการสร้างตัวแบบเพื่อจัดการข้อมูลให้อยู่ในกลุ่มที่กำหนดมาให้ เช่น การแบ่งประเภทข้อมูลของลูกค้าว่าเชื่อถือได้ หรือเชื่อถือไม่ได้ โดยพิจารณาจากข้อมูลที่มีอยู่ โดยการสร้างกฎเพื่อช่วยในการตัดสินใจจากข้อมูลที่มีอยู่ เพื่อใช้ในการทำนายแนวโน้มการเกิดขึ้นของข้อมูลที่ยังไม่เกิดขึ้น การจำแนกประเภทข้อมูลมี 2 ประเภทหลัก คือ การจำแนกแบบ

ทวิภาค (Binary Classification) เป็นตัวแบบที่มีการกำหนดหมวดหมู่ของตัวแปรเพียงสองหมวดหมู่ เช่น ผลลัพธ์แบบ “ใช่” หรือ “ไม่ใช่” เป็นต้น และอีกประเภทหนึ่ง คือ การจำแนกแบบพหุ (Multiple Classification) เป็นตัวแบบที่มีการกำหนดหมวดหมู่ของตัวแปรมากกว่าสองหมวดหมู่ วิธีการจำแนกประเภทข้อมูลของวิธีการเรียนรู้ด้วยเครื่อง สามารถทำได้หลายวิธี ได้แก่ วิธีต้นไม้ตัดสินใจ (Decision Tree) วิธีนาอีฟเบย์ (Naïve Bayes) วิธีเพื่อนบ้านใกล้เคียง k อันดับ (K-Nearest Neighbors หรือ KNN) วิธีซัพพอร์ตเวกเตอร์แมชชีน (Support Vector Machine) วิธีโครงข่ายประสาทเทียม (Artificial Neural Network) และวิธีป่าสุ่ม (Random Forest) ซึ่งในการวิเคราะห์ข้อมูลนั้นมีการกำหนดตัวแปร 2 ประเภท คือ ตัวแปรอิสระ (Independent Variable) เป็นตัวแปรที่ไม่ขึ้นอยู่กับตัวแปรตัวอื่นๆ เป็นตัวแปรที่เกิดขึ้นก่อน เป็นตัวเหตุที่ทำให้เกิดผลตามมา และ ตัวแปรตาม (Dependent Variable) เป็นตัวแปรที่แปรผันไปตามตัวแปรอิสระ หรือกล่าวได้ว่า เป็นตัวแปรที่เป็นผล เมื่อตัวแปรอิสระเป็นเหตุ

Agrawal et al. (2022) ได้ศึกษาวิธีการเรียนรู้ด้วยเครื่องสำหรับการจำแนกผู้ป่วยที่มีสุขภาพดีและกล้ามเนื้อตายโดยใช้ความแปรปรวนของอัตราการเต้นของหัวใจที่ได้มาจากขนาดเวกเตอร์ พบว่า วิธีการจำแนกประเภท มีค่าความไว (Sensitivity) ค่าความจำเพาะ (Specificity) และค่าความถูกต้อง (Accuracy) ในการจำแนกสูง ระบุผลการวิเคราะห์ว่า วิธีการเรียนรู้ด้วยเครื่อง ด้วยวิธีต้นไม้ตัดสินใจเป็นตัวทำนายที่ดีที่สุดสำหรับความแม่นยำในการจำแนก รองลงมาคือ วิธีซัพพอร์ตเวกเตอร์แมชชีน และ วิธีโครงข่ายประสาทเทียม ตามลำดับ

ประยุทธ์ศิลป์ (2562) ได้ศึกษาการสร้างแบบจำลองจำแนกกลุ่มผู้ป่วยโรคไตเรื้อรังโดยใช้เทคนิคเหมืองข้อมูลและการนำเสนอภาพข้อมูล (Visualization) จากผลการเปรียบเทียบความแม่นยำของแต่ละวิธี พบว่า วิธีนาอีฟเบย์มีค่าความแม่นยำ 98.1% และค่าความผิดพลาด 0.13736 วิธีต้นไม้ตัดสินใจและวิธีป่าสุ่มมีค่าความแม่นยำ 98% และค่าความผิดพลาด 0.1443 วิธีซัพพอร์ตเวกเตอร์แมชชีนมีค่าความแม่นยำ 49% และคำความผิดพลาด 0.54944 วิธีเพื่อนบ้านใกล้เคียง k อันดับ มีค่าความแม่นยำ 96% และค่าความผิดพลาด 0.2041 ทำให้วิธีนาอีฟเบย์เป็นแบบจำลองที่มีความแม่นยำและเหมาะสมสำหรับการพยากรณ์ภาวะโรคไตเรื้อรังมากที่สุดจากแบบจำลองทั้ง 5 แบบ

Majumder et al. (2022) ได้ศึกษาแบบจำลองการทำนายโรคหัวใจด้วยวิธีการวิเคราะห์ถดถอยโลจิสติก วิธีนาอีฟเบย์ และ วิธีเพื่อนบ้านใกล้เคียง k อันดับ พบว่า แบบจำลองนี้มีค่าความถูกต้อง 82.8%, 82.5% และ 83.2% ตามลำดับสำหรับ วิธีการวิเคราะห์ถดถอยโลจิสติก วิธีนาอีฟเบย์ และ วิธีเพื่อนบ้านใกล้เคียง k อันดับ ซึ่งมีค่าความถูกต้องมากที่สุด

Swain et al. (2022) การจำแนกโรคไตเรื้อรังโดยใช้วิธีการเรียนรู้ด้วยเครื่อง พบว่า วิธีซัพพอร์ตเวกเตอร์แมชชีน และ วิธีป่าสุ่ม มีอัตราการทำนายไม่ตรงกับสิ่งที่เกิดขึ้นจริง (False-Negative)

ต่ำที่สุดและมีค่าความถูกต้องในการทดสอบเท่ากับ 99.33% และ 98.67% ตามลำดับ อย่างไรก็ตาม วิธีซัพพอร์ตเวกเตอร์แมชชีนได้ผลลัพธ์ที่ดีกว่าวิธีป่าสุ่ม เมื่อตรวจสอบความถูกต้องด้วยการตรวจสอบ 10-fold cross-validation.

สายชล (2561) ได้ศึกษาการเปรียบเทียบประสิทธิภาพในการทำนายผลการเป็นโรคเบาหวานของโรงพยาบาลแห่งหนึ่ง ด้วยวิธีการจำแนกกลุ่มที่นำมาเปรียบเทียบ คือ วิธีเพื่อนบ้านใกล้เคียง k อันดับ วิธีต้นไม้ตัดสินใจ วิธีโครงข่ายประสาทเทียม วิธีซัพพอร์ตเวกเตอร์แมชชีน วิธีการถดถอยโลจิสติกส์แบบ 2 กลุ่ม และ วิธีนาอีฟเบย์ ในการเปรียบเทียบประสิทธิภาพในการทำนายผลของวิธีการจำแนกกลุ่มทั้ง 6 วิธี จะใช้ค่าความถูกต้อง ค่าความคลาดเคลื่อนสัมบูรณ์เฉลี่ย (Mean Absolute Error : MAE) และค่าความคลาดเคลื่อนกำลังสองเฉลี่ย (Mean Square Error : MSE) ผลการศึกษาพบว่า วิธีโครงข่ายประสาทเทียมมีค่าความถูกต้อง ค่าความคลาดเคลื่อนสัมบูรณ์เฉลี่ย และค่าความคลาดเคลื่อนกำลังสองเฉลี่ยดีที่สุด คือ 95.94 % , 0.0491 และ 0.0396 ตามลำดับ ดังนั้นวิธีโครงข่ายประสาทเทียมมีประสิทธิภาพในการทำนายผลดีที่สุด

Grandini et al. (2021) ได้ศึกษาการจำแนกโรคของต่อมไทรอยด์โดยใช้วิธีการเรียนรู้ด้วยเครื่องพบว่า วิธีป่าสุ่มให้ผลลัพธ์ที่มีค่าความถูกต้องมากที่สุดคือ 98.93% ซึ่งมีค่าความถูกต้องมากกว่าวิธีซัพพอร์ตเวกเตอร์แมชชีน วิธีต้นไม้ตัดสินใจ วิธีนาอีฟเบย์ วิธีการวิเคราะห์ถดถอยโลจิสติก วิธีเพื่อนบ้านใกล้เคียง k อันดับ วิธีโครงข่ายประสาทเทียมแบบหลายชั้น และ วิธีการวิเคราะห์การจำแนกประเภทเชิงเส้น

ลักษณะของข้อมูลมิติขั้นสูง (High dimensional data) มีลักษณะคือ มีจำนวนตัวแปรอิสระมากกว่าขนาดของตัวอย่าง ซึ่งข้อมูลมิติขั้นสูง มีจำนวนตัวแปรอิสระที่มากมายมหาศาล ดังนั้นมีโอกาสอย่างมากที่จะทำให้ตัวแปรอิสระเพียงบางส่วนมีความสัมพันธ์กับตัวแปรตาม ตัวแปรอิสระส่วนใหญ่ไม่มีความสัมพันธ์กับตัวแปรตาม หรือเมื่อมีจำนวนตัวแปรอิสระในจำนวนมากอาจเกิดปัญหาที่ตัวแปรอิสระนั้นมีความสัมพันธ์กันเองสูง เรียกว่า ความสัมพันธ์เชิงเส้นพหุ (Multicollinearity) นอกจากนั้นหากมีข้อมูลในจำนวนมากๆ ข้อมูลมีโอกาสมีค่านอกเกณฑ์ (Outlier) คือ ข้อมูลที่มีค่าผิดปกติ มีเป็นข้อมูลที่มีค่าแยกออกจากกลุ่มหรือข้อมูลที่มีค่าที่แตกต่างไปจากข้อมูลค่าอื่นๆ มีค่าสูงหรือต่ำไปกว่าข้อมูลปกติ ซึ่งอาจเกิดได้จากการเก็บรวบรวมบันทึกข้อมูลที่ผิดพลาดไปหรืออาจเกิดจากข้อมูลนั้นมีค่าที่ผิดปกติจริง และในชุดข้อมูลที่มีกลุ่มของข้อมูลจำนวนมากนั้น อาจมีข้อมูลไม่สมดุล (Imbalance data) หมายถึง ข้อมูลที่มีการกระจายตัวไม่เท่าเทียมกัน คือ ข้อมูลในกลุ่มหนึ่ง มีจำนวนมากกว่าข้อมูลของอีกกลุ่มในจำนวนมาก ซึ่งอาจเกิดจากธรรมชาติของข้อมูลที่มีความแตกต่างกันของจำนวน

ในแต่ละกลุ่ม หรืออาจเกิดจากข้อจำกัดในการเก็บข้อมูล เช่น ค่าใช้จ่าย ขอบเขตของพื้นที่ในการเก็บข้อมูล เป็นต้น

ในการศึกษาครั้งนี้ผู้วิจัยมีความสนใจศึกษาการจำแนกระดับการเป็นมะเร็งปอด เป็นตัวแปรตาม จากข้อมูลรหัสพันธุกรรม เป็นตัวแปรอิสระ โดยมีลักษณะเป็น ข้อมูลมิติขั้นสูง ข้อมูลมีความสัมพันธ์เชิงเส้นพหุ ข้อมูลมีค่านอกเกณฑ์ และ ข้อมูลไม่สมดุล ของผู้ที่เป็นมะเร็งปอด มาจำแนกระดับของการเป็นมะเร็งด้วยวิธีการจำแนกแบบพหุ เนื่องจากมีข้อมูลเป้าหมายที่ต้องการทราบเป็นผลลัพธ์มากกว่าสองหมวดหมู่ และใช้วิธีการจำแนกข้อมูลของวิธีการเรียนรู้ด้วยเครื่อง ด้วยวิธี 6 วิธี คือ วิธีต้นไม้ตัดสินใจ วิธีนาอีฟเบย์ วิธีเพื่อนบ้านใกล้เคียง k อันดับ วิธีซัพพอร์ตเวกเตอร์แมชชีน วิธีโครงข่ายประสาทเทียม และ วิธีป่าสุ่ม

1.2 วัตถุประสงค์ของงานวิจัย

1.2.1 ศึกษาวิธีการจำแนกระดับของผู้ป่วยมะเร็งปอดจากรหัสพันธุกรรม ด้วยวิธีการเรียนรู้ด้วยเครื่องประกอบด้วย 6 วิธี ดังนี้ วิธีต้นไม้ตัดสินใจ วิธีนาอีฟเบย์ วิธีเพื่อนบ้านใกล้เคียง k อันดับ วิธีซัพพอร์ตเวกเตอร์แมชชีน วิธีโครงข่ายประสาทเทียม และวิธีป่าสุ่ม

1.2.2 เปรียบเทียบประสิทธิภาพการจำแนกระดับการเป็นมะเร็งปอดจากรหัสพันธุกรรมด้วยวิธีการเรียนรู้ด้วยเครื่องประกอบด้วย 6 วิธี ดังนี้วิธีต้นไม้ตัดสินใจ วิธีนาอีฟเบย์ วิธีเพื่อนบ้านใกล้เคียง k อันดับ วิธีซัพพอร์ตเวกเตอร์แมชชีน วิธีโครงข่ายประสาทเทียม และวิธีป่าสุ่ม

1.3 ขอบเขตของงานวิจัย

1.3.1 ศึกษาข้อมูลที่มีมิติขั้นสูง เป็นข้อมูลระดับการเป็นมะเร็งปอด ที่ได้ข้อมูลจากผู้ป่วยจำนวน 197 คน มีตัวแปรอิสระ คือ รหัสพันธุกรรมของคนไข้ จำนวน 1,000 รหัส และตัวแปรตาม คือ ประเภทของโรคมะเร็งปอด จำนวน 4 กลุ่ม คือ Small Cell Lung Cancer (SCLC) : Oat cell lung cancer , Non-Small Cell lung cancer (NSCLC) : Adenocarcinoma , Non-Small Cell lung cancer (NSCLC) : Squamous Cell carcinoma , Non-Small Cell lung cancer (NSCLC) : Large Cell lung cancer

1.3.2 ทำการสุ่มรหัสพันธุกรรมในจำนวน 200 400 600 800 และ 1,000 รหัส และใช้โปรแกรม R Studio ในการจำแนกกลุ่มของระดับการเป็นมะเร็งปอด ด้วยการเรียนรู้ด้วยเครื่อง 6 วิธี ได้แก่ ต้นไม้ตัดสินใจ นาอีฟเบย์ เพื่อนบ้านใกล้เคียง k อันดับ ซัพพอร์ตเวกเตอร์แมชชีน โครงข่ายประสาทเทียม และ ป่าสุ่ม และ คำนวณหา ค่าความถูกต้อง (Accuracy) ค่าความแม่นยำ (Precision) และ ค่าความระลึก (Recall) จากตารางเมทริกซ์ความสับสน (Confusion Matrix)

1.3.3 ตารางเมทริกซ์ความสับสน

ค่าทำนาย $(\hat{y})$	ค่าจริง (y)			
	Class 1	Class 2	Class 3	Class 4
Class 1	A_{11}	A_{12}	A_{13}	A_{14}
Class 2	A_{21}	A_{22}	A_{23}	A_{24}
Class 3	A_{31}	A_{32}	A_{33}	A_{34}
Class 4	A_{41}	A_{42}	A_{43}	A_{44}

TP (True Positive)

คือ สิ่งที่ทำนาย ตรงกับสิ่งที่เกิดขึ้น ในกรณี ทำนายว่า จริง และสิ่งที่เกิดขึ้นคือ จริง

TN (True Negative)

คือ สิ่งที่ทำนาย ตรงกับสิ่งที่เกิดขึ้น ในกรณี ทำนายว่า ไม่จริง และสิ่งที่เกิดขึ้นคือ ไม่จริง

FP (False Positive)

คือ สิ่งที่ทำนาย ไม่ตรงกับสิ่งที่เกิดขึ้น ในกรณี ทำนายว่า จริง แต่สิ่งที่เกิดขึ้นคือ ไม่จริง

FN (False Negative)

คือ สิ่งที่ทำนาย ไม่ตรงกับที่เกิดขึ้นจริง ในกรณี ทำนายว่า ไม่จริง แต่สิ่งที่เกิดขึ้นคือ จริง

$$(TP_i) = A_{ii}$$

$$(FP_i) = \sum_{c=1}^{4} A_{cr} - TP_i$$

$$(FN_i) = \sum_{r=1}^{4} A_{cr} - TP_i$$

$$(TN_i) = \sum_{c=1}^{4}\sum_{r=1}^{4} A_{cr} - TP_i - FP_i - FN_i$$

โดยที่ i = 1, 2, 3, 4

1.3.4 คำนวณค่าร้อยละความถูกต้อง ร้อยละความแม่นยำ ร้อยละความระลึก จากตารางเมทริกซ์ความสับสน

ค่าร้อยละความถูกต้อง (Accuracy) $= \frac{\sum_{i=1}^{4} A_{ii}}{\sum_{c=1}^{4}\sum_{r=1}^{4} A_{cr}} \times 100$

ค่าร้อยละความแม่นยำ (Precision) $= \frac{TP_i}{TP_i + FP_i} \times 100$

ค่าร้อยละความระลึก (Recall) $= \frac{TP_i}{TP_i + FN_i} \times 100$

1.3.5 ทำการสุ่มรหัสพันธุกรรมจำนวน 1,000 รอบ และหาค่าเฉลี่ยร้อยละความถูกต้อง ค่าเฉลี่ยร้อยละความแม่นยำ และ ค่าเฉลี่ยร้อยละความระลึก ที่สูงที่สุด

1.4 ประโยชน์ที่คาดว่าจะได้รับ

1.4.1 นำผลเปรียบเทียบประสิทธิภาพที่ได้ไปใช้ในการเลือกวิธีการเรียนรู้ด้วยเครื่อง สำหรับการจำแนกกลุ่มมะเร็งอื่นๆ ที่เหมาะสมกับข้อมูลรหัสพันธุกรรม

1.4.2 สามารถนำผลการวิเคราะห์การจำแนกระดับการเป็นมะเร็งปอดไปประกอบการตัดสินใจในการรักษา และนำไปประยุกต์ใช้สำหรับการจำแนกระดับของการเป็นโรคอื่นๆ ด้วย วิธีการเรียนรู้ด้วยเครื่อง

1.5 นิยามศัพท์เฉพาะ

1.5.1 รหัสพันธุกรรม (Genetic code) หมายถึง ชุดข้อมูลทางพันธุกรรมที่ประกอบไปด้วยลำดับเบสบน DNA (Deoxyribonucleic acid) ทำหน้าที่ เก็บ ควบคุม และถ่ายทอดลักษณะทางพันธุกรรมของสิ่งมีชีวิตจากรุ่นสู่รุ่น

1.5.2 การเรียนรู้ของเครื่อง (Machine Learning) หมายถึง วิธีการวิเคราะห์ข้อมูลด้วย 6 วิธีดังนี้ วิธีต้นไม้ตัดสินใจ (Decision Tree) วิธีนาอีฟเบย์ (Naïve Bayes) วิธีเพื่อนบ้านใกล้เคียง k อันดับ (K-Nearest Neighbors หรือ KNN) วิธีซัพพอร์ตเวกเตอร์แมชชีน (Support Vector Machine) วิธีโครงข่ายประสาทเทียม (Artificial Neural Network) และ วิธีป่าสุ่ม (Random Forest)

1.5.3 การจำแนกประเภทของข้อมูล (Classification) หมายถึง เป็นเทคนิคในการจำแนกเทคนิคหนึ่งที่ใช้ในงานด้านการทำเหมืองข้อมูล (Data Mining) ในการสร้างแบบจำลองเพื่อ

ทำนายคำตอบที่มีลักษณะคำข้อมูลเป็นแบบเชิงคุณภาพ (Qualitative value) หรือเชิงกลุ่ม (Category data) เช่น จัดอยู่ในกลุ่ม เสี่ยง/ไม่เสี่ยง ระดับความเสี่ยงต่อการเป็นโรค เสี่ยงมาก/เสี่ยงน้อย/เสี่ยงปานกลาง เป็นต้น โดยหลักการทำงานของเทคนิค คือ จะเป็นการสร้างแบบจำลองจากชุดข้อมูลที่มีอยู่และนำมาประยุกต์ใช้ทำนายคำตอบของชุดข้อมูลใหม่

1.5.4 ชุดข้อมูลฝึกฝน (Training data set) หมายถึง ชุดข้อมูลที่ถูกสุ่มหรือเลือกจากข้อมูลที่ทำการวิเคราะห์เพื่อใช้ในการปรับปรุงตัวแบบให้ดีขึ้น โดยคัดเลือกตัวแปรที่เหมาะสมตามเกณฑ์ที่กำหนด

1.5.5 ชุดข้อมูลทดสอบ (Test data set) หมายถึง ชุดข้อมูลที่ถูกสุ่มจากชุดข้อมูลที่ทำการวิเคราะห์ซึ่งไม่ได้ใช้ในการปรับปรุงตัวแบบให้ดีขึ้นและตรวจสอบตัวแบบ เป็นชุดข้อมูลที่นำมาทดสอบตัวแบบที่ผ่านการตรวจสอบจากวิธีต่างๆ เพื่อคัดเลือกตัวแบบที่เหมาะสมไปใช้งาน

1.5.6 การแบ่งกลุ่ม (Clustering) คือ การวิเคราะห์กลุ่มเป็นเทคนิคที่ใช้ในการจัดกลุ่มโดยไม่ทราบมาก่อนว่าควรมีกี่กลุ่ม แต่จะแบ่งตามค่าของตัวแปรที่นำมาใช้ในการแบ่ง โดยให้หน่วยที่อยู่ในกลุ่มเดียวกัน มีความคล้ายกันในตัวแปรที่ศึกษา ซึ่งหน่วยที่อยู่ต่างกลุ่มกันจะมีความแตกต่างกัน ตัวอย่างเช่น คนที่อยู่ในกลุ่มเดียวกันมีอายุและรายได้ใกล้เคียงกัน

1.5.7 การจำแนกแบบพหุ (Multiple Classification) คือ เป็นตัวแบบที่มีการกำหนดหมวดหมู่ของตัวแปรมากกว่าสองหมวดหมู่

1.5.8 ข้อมูลมิติขั้นสูง (High dimensional data) คือ เป็นข้อมูลที่มีจำนวนตัวแปรอิสระมากกว่าขนาดของตัวอย่าง ซึ่งข้อมูลมิติสูง มีจำนวนตัวแปรอิสระจำนวนมากในหลักร้อยหรือหลักพันสำหรับการวิเคราะห์ข้อมูล

1.5.9 ความสัมพันธ์เชิงเส้นพหุ (Multicollinearity) คือ ตัวแปรอิสระนั้นมีความสัมพันธ์เชิงเส้นกันเองสูง

1.5.10 ค่านอกเกณฑ์ (Outlier) คือ ข้อมูลที่มีค่าผิดปกติ เป็นข้อมูลที่มีค่าแยกออกจากกลุ่มหรือข้อมูลที่มีค่าที่แตกต่างไปจากข้อมูลค่าอื่นๆ มีค่าสูงหรือต่ำไปกว่าข้อมูลปกติ

1.5.11 ข้อมูลไม่สมดุล (Imbalance data) หมายถึง ข้อมูลที่มีการกระจายตัวไม่เท่าเทียมกัน คือ ข้อมูลในกลุ่มหนึ่งมีจำนวนมากกว่าข้อมูลของอีกกลุ่มในจำนวนมาก

1.5.12 ค่าความถูกต้อง (Accuracy) หมายถึง ค่าที่ได้จากความถูกต้องของผลทำนายจากตัวแบบที่ทายถูก เทียบกับสิ่งที่เกิดขึ้นจริง

1.5.13 ค่าความแม่นยำ (Precision) หมายถึง ค่าที่ได้จากความน่าจะเป็นที่ตัวแบบทำนายในกรณี ทำนายว่า จริง และสิ่งที่เกิดขึ้นคือ จริง เทียบกับการทำนายว่าจริงทั้งหมด

1.5.14 ค่าความระลึก (Recall) หมายถึง ค่าที่ได้จากความน่าจะเป็นที่ตัวแบบทำนาย ในกรณี ทำนายว่า จริง และสิ่งที่เกิดขึ้นคือ จริง เทียบกับผลรวมส่วนที่ทายถูกในกรณีทำนายว่า จริง และสิ่งที่เกิดขึ้นคือ จริง และ ทายผิด ในกรณี ทำนายว่า ไม่จริง แต่สิ่งที่เกิดขึ้นคือ จริง

บทที่ 2
ทฤษฎีและงานวิจัยที่เกี่ยวข้อง

ในการวิจัยครั้งนี้ผู้วิจัยมีความสนใจที่จะศึกษาการเปรียบเทียบประสิทธิภาพวิธีการจำแนกกลุ่ม โดยมีวิธีที่จะนำมาศึกษา ได้แก่ วิธีต้นไม้ตัดสินใจ (Decision Tree) วิธีนาอีฟเบย์ (Naïve Bayes) วิธีเพื่อนบ้านใกล้เคียง k อันดับ (K-Nearest Neighbors หรือ KNN) วิธีซัพพอร์ตเวกเตอร์แมชชีน (Support Vector Machine) วิธีโครงข่ายประสาทเทียม (Artificial Neural Network) และวิธีป่าสุ่ม (Random Forest) ซึ่งรายละเอียดทฤษฎีและงานวิจัยที่เกี่ยวข้องของแต่ละวิธีดังต่อไปนี้

2.1 การทำเหมืองข้อมูล (Data Mining)

สุภาภรณ์ (2564) เป็นกระบวนการในการค้นหารูปแบบในชุดข้อมูลขนาดใหญ่ โดยใช้วิธีการของการเรียนรู้ของเครื่องมือทางสถิติในระบบฐานข้อมูล และการทำเหมืองข้อมูล เป็นขั้นตอนของวิธีการในกระบวนการสืบค้นข้อมูลในฐานระบบ (Knowledge Discovery in Databases: KDD) ซึ่งเป็นเทคนิคในการค้นหาอัตโนมัติในรูปแบบ และความสัมพันธ์ที่ซ่อนอยู่ในชุดข้อมูลนั้นจากข้อมูลจำนวนมหาศาล ดังนั้นการวิเคราะห์ หรือ การทำเหมืองข้อมูลที่มีความซับซ้อนมากจะต้องใช้เวลานานในการประมวลผลข้อมูลที่มีปริมาณมากนี้ มาสร้างแบบจำลองและความสัมพันธ์ทางสถิติ

Neli (2018) ได้แบ่งวิธีการเรียนรู้ด้วยเครื่อง ออกเป็น 2 ประเภท คือ กระบวนการสร้างตัวแบบการเรียนรู้แบบมีผู้สอน (Supervised Learning) และแบบไม่มีผู้สอน (Unsupervised Learning) ซึ่งปัญหาพิเศษนี้ได้ยกประเภท การเรียนรู้แบบมีผู้สอน มาใช้ในการสร้างแบบจำลอง โดยเป็นการเรียนรู้ด้วยวิธีการที่มีข้อมูลฝึกฝน (Training data set) ที่ประกอบไปด้วยตัวแปรที่ต้องการทำนายอยู่แล้ว ซึ่งวิธีการเรียนรู้แบบมีผู้สอน จำแนกออกไปอีก 2 วิธีหลัก ๆ คือ การจำแนกประเภทข้อมูล (Classification) และการถดถอยของข้อมูล (Regression) ซึ่งปัญหาพิเศษนี้จะขอกล่าวถึงเพียงวิธีการจำแนกประเภทข้อมูล ซึ่งเป็นการจำแนกประเภทข้อมูลที่ถูกกำหนดด้วยผลลัพธ์ที่มีอยู่แล้ว สำหรับกระบวนการจำแนกประเภทข้อมูล จะแบ่งข้อมูลออกเป็น 2 ชุด ได้แก่ ชุดข้อมูลฝึกฝน และ ชุดข้อมูลทดสอบ

2.2 วิธีต้นไม้ตัดสินใจ (Decision Tree)

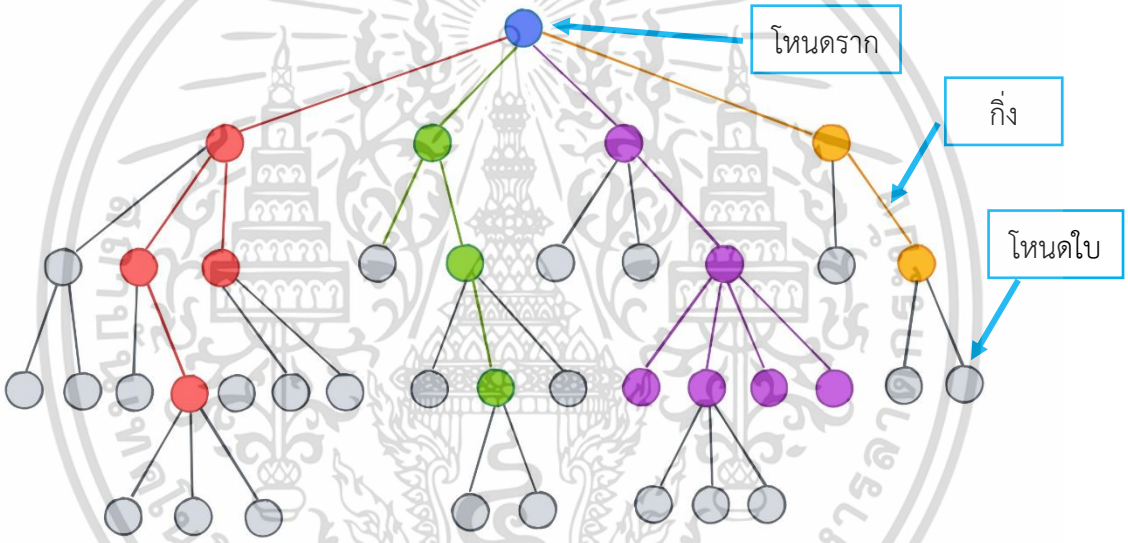
เป็นเครื่องมือที่ช่วยวิเคราะห์เหตุการณ์ หรือสถานการณ์ เพื่อการตัดสินใจได้อย่างเป็นระบบและรวดเร็ว ต้นไม้การตัดสินใจมีลักษณะเป็นกราฟรูปต้นไม้ ซึ่งแสดงที่ตั้งต้นที่มีรากและแขนงต่างๆ แตกออกมาจากต้นไม้ไปในทิศทางเดียว จนกระทั่งนำไปสู่ข้อสรุปสำหรับการตัดสินใจได้ เป็นการเชื่อมโยงเหตุการณ์ต่างๆเพื่อหาทางเลือกที่ดีที่สุด โดยนำข้อมูลมาสร้างตัวแบบการพยากรณ์ในรูปแบบของโครงสร้างต้นไม้ ซึ่งมีการเรียนรู้ข้อมูลแบบมีผู้สอน (Supervised Learning) สามารถ

สร้างตัวแบบการจัดกลุ่ม (Clustering) ได้จากกลุ่มตัวอย่างของข้อมูลฝึกฝน (Training Data Set) ได้โดยอัตโนมัติและสามารถพยากรณ์กลุ่มของรายการที่ยังไม่เคยนำมาจัดกลุ่มได้อีกด้วย (รุจิรา, 2554)

2.2.1 ส่วนประกอบของวิธีต้นไม้ตัดสินใจ

1. โหนดภายใน (Internal node) คือ คุณลักษณะต่างๆ ของข้อมูล ซึ่งเมื่อข้อมูลใดๆ ตกลงมาที่โหนด จะใช้คุณลักษณะนี้เป็นตัวตัดสินใจว่าข้อมูลจะไปในทิศทางใด โดยโหนดภายในที่เป็นจุดเริ่มต้นของต้นไม้ เรียกว่า โหนดราก (Root)
2. กิ่ง (Branch, Ink) เป็นค่าของคุณลักษณะในโหนดภายในที่แตกกิ่งออกมา ซึ่งโหนดภายในจะแตกกิ่งเป็นจำนวนเท่ากับจำนวนค่าของคุณลักษณะในโหนดภายในนั้น
3. โหนดใบ (Leaf node) คือกลุ่มต่างๆ ซึ่งเป็นผลลัพธ์ในการจำแนกประเภทข้อมูล

ดังภาพที่ 2.1 แสดงส่วนประกอบของวิธีต้นไม้ตัดสินใจ



รูปที่ 2.1 แสดงส่วนประกอบของวิธีต้นไม้ตัดสินใจ

2.2.2 การสร้างต้นไม้ตัดสินใจ

หลักการพื้นฐานของการสร้างต้นไม้ตัดสินใจเป็นการสร้างในลักษณะจากบนลงล่าง (Top- Down) คือเริ่มจากการสร้างรากของต้นไม้ก่อนแล้วจึงแตกกิ่งไปจนถึงใบ โดยแสดงขั้นตอนการสร้างต้นไม้ตัดสินใจได้ดังนี้ (Han and Kamber, 2001)

1. ต้นไม้เริ่มต้นโดยมีโหนดเพียงโหนดเดียวแสดงถึงชุดข้อมูลฝึกฝน (Training Data Set)

2. ถ้าข้อมูลทั้งหมดอยู่ในกลุ่มเดียวกันแล้ว ให้โหนดนั้นเป็นใบและตั้งชื่อแยกตามกลุ่มของข้อมูลนั้น

3. ถ้าในโหนดมีข้อมูลหลายกลุ่มปะปนอยู่ จะต้องวัดค่าเกนความรู้ (Gain) ของแต่ละคุณลักษณะ (Attribute) เพื่อที่จะเป็นเกณฑ์ในการคัดเลือกคุณลักษณะที่มีความสามารถในการ

แบ่งแยกข้อมูลออกเป็นกลุ่มต่างๆได้ดีที่สุด โดยมีคุณลักษณะที่มีค่าเกนความรู้ มากที่สุดจะถูกเลือกให้เป็น ตัวทดสอบหรือคุณลักษณะที่ใช้ในการตัดสินใจ โดยแสดงในรูปของโหนดบนต้นไม้

4. กิ่งของต้นไม้ถูกสร้างขึ้นจากค่าต่างๆ ที่เป็นไปได้ของโหนดทดสอบ ข้อมูลจะถูกทดสอบและถูกแบ่งออกตามกิ่งต่างๆที่สร้างขึ้น

5. ทำการวนซ้ำเพื่อหาคุณลักษณะที่มีค่าเกนความรู้มากที่สุด สำหรับข้อมูลที่ถูกแบ่งแยกออกมาในแต่ละกิ่งเพื่อนำคุณลักษณะนี้มาสร้างเป็นโหนดการตัดสินใจต่อไป โดยที่คุณลักษณะที่ถูกเลือกมาเป็นโหนดแล้ว จะไม่ถูกเลือกมาอีกสำหรับโหนดในระดับต่อไป

6. ทำการวนซ้ำเพื่อแบ่งข้อมูลและแตกกิ่งต้นไม้ไปเรื่อยๆ โดยการวนซ้ำจะสิ้นสุดก็ต่อเมื่อเงื่อนไขข้อใดข้อหนึ่งข้างบนนี้เป็นจริง

2.2.3 การคำนวณค่าเกนความรู้ (Information Gain)

ต้นไม้ตัดสินใจเป็นโครงสร้างที่ใช้แสดงกฎที่ได้จากเทคนิคการจำแนกประเภทข้อมูล โดยมีลักษณะคล้ายโครงสร้างต้นไม้ ซึ่งแต่ละโหนดแสดงคุณลักษณะ (Attribute) ในการสร้างต้นไม้ตัดสินใจ โดยค่านี้จะถูกประยุกต์ใช้ในอัลกอริทึม ID3 ที่ซึ่งจะทำการเลือกคุณลักษณะสำหรับแบ่งข้อมูลจากคุณลักษณะที่มีค่าเกนความรู้สูงที่สุด สิ่งสำคัญที่ควรพิจารณา คือ ควรจะตัดสินใจเลือกคุณลักษณะใดมาทำหน้าที่เป็นโหนดรากในแต่ละขั้นตอนของการสร้างต้นไม้และต้นไม้ย่อย (Subtree) ของต้นไม้ตัดสินใจ เกณฑ์ที่ใช้ช่วยประกอบการเลือกคุณลักษณะคือการคำนวณค่ามาตรฐานเกน (Gain Criterion) ซึ่งเป็นค่าที่บ่งบอกว่าคุณลักษณะนั้นสามารถจำแนกกลุ่มของข้อมูลได้ดีเพียงใด โดยทดลองเลือกแต่ละคุณลักษณะที่เป็นไปได้จากชุดข้อมูลมาทำหน้าที่เป็นโหนดราก หากคุณลักษณะใดให้ค่าเกนความรู้สูงที่สุด แสดงว่าคุณลักษณะนั้นสามารถจำแนกกลุ่มของข้อมูลได้ดีที่สุด การใช้ค่าเกนความรู้สูงจะช่วยลดจำนวนครั้งของการทดสอบในการแยกข้อมูล อีกทั้งยังรับประกันว่าต้นไม้ตัดสินใจที่ได้ไม่มีความซับซ้อนมากจนเกินไป (ขจรศักดิ์, 2552)

อัลกอริทึม ID3 (Iterative Dichotomiser3) เป็นอัลกอริทึมพื้นฐานที่ใช้ในการสร้างการตัดสินใจแบบโครงสร้างต้นไม้ที่ใช้หลักการของการใช้ทฤษฎีข่าวสาร (Information Theory) และค่าที่วัดได้จะนำมาใช้ตัดสินใจว่าจะใช้ตัวแปรใดใช้ในการทำนาย หรือแบ่งประเภทของข้อมูล โดยที่ชุดตัวอย่าง (Sample) คือชุดของข้อมูลที่ใช้ในการเรียนรู้ (Training Sample) ตัวแปรเป้าหมาย (Target Attribute) คือตัวแปรที่นำค่าไปใช้ในการทำนายผลในโครงสร้างต้นไม้และคุณลักษณะ (Attributes) คือตัวแปรอื่นๆที่ใช้ในการสร้างโหนดในต้นไม้ และไม่ใช่ตัวแปรเป้าหมาย (Target Attribute) ซึ่งค่าเกนความรู้ (Information Gain) หรือ ค่าเอ็นโทรปี (Entropy) ของชุดข้อมูลทั้งหมดนั้นสามารถคำนวณได้จากสมการ 2.1 (ศุภชัย, 2551)

$$E(D) = -\sum_{k=1}^{m} p_k \log_2(p_k) \quad (2.1)$$

เมื่อ D แทน ชุดข้อมูลที่สนใจ

p_k แทน ความน่าจะเป็นของจำนวนตัวแปรตาม k ต่อจำนวนตัวแปรตามทั้งหมด

k แทน กลุ่มของตัวแปรตาม ซึ่งมีทั้งหมด m กลุ่ม

m แทน จำนวนกลุ่มทั้งหมดของตัวแปรตาม

จากนั้นทำการคำนวณค่าเกนความรู้ของตัวแปรอิสระหรือตัวแปรอิสระแต่ละตัว ได้จากสมการ 2.2

$$E_A(D) = \sum_{j=1}^{p} \left| \frac{D_j}{D} \right| \times E(D_j) \quad (2.2)$$

เมื่อ D แทน ชุดข้อมูลที่สนใจ

D_j แทน ตัวแปรอิสระตัวที่ j

j แทน กลุ่มของตัวแปรอิสระ ซึ่งมีทั้งหมด p กลุ่ม

p แทน จำนวนกลุ่มทั้งหมดของตัวแปรอิสระ

ดังนั้นจะสามารถพิจารณาค่าเกนความรู้ของตัวแปรอิสระแต่ละตัว จากนั้นเลือกตัวแปรอิสระที่มีค่าเกนสูงที่สุดเป็นตัวจำแนกชุดข้อมูล ได้จากสมการ 2.3

$$Gain(A) = E(D) - E_A(D) \quad (2.3)$$

เมื่อ D แทน ชุดข้อมูลที่สนใจ

A แทน ตัวแปรอิสระที่สนใจ

2.3 วิธีนาอีฟเบย์ (Naïve Bayes)

นาอีฟเบย์เป็นเครื่องจักรเรียนรู้ที่อาศัยหลักการความน่าจะเป็น (Probability) ตามทฤษฎีของเบย์ (Bayes' theorem) เป็นการจำแนกข้อมูลโดยใช้ความน่าจะเป็นและคำนวณการแจกแจงความน่าจะเป็นตามสมมติฐานที่ตั้งให้กับข้อมูล จากการคำนวณตัวอย่างใหม่ที่ได้จะถูกนำมาปรับเปลี่ยนการแจกแจง ซึ่งมีผลต่อการเพิ่มหรือลดความน่าจะเป็นของข้อมูล จากข้อมูลเดิมตัวและตัวแบบจะถูกปรับเปลี่ยนความน่าจะเป็นใหม่ โดยผนวกกับข้อมูลเดิมที่มีหลักการของนาอีฟเบย์ใช้การคำนวณหาความน่าจะเป็นซึ่งถูกใช้ในการทำนายผลเป็นวิธีในการแก้ปัญหาแบบการจำแนกที่สามารถคาดการณ์ผลลัพธ์และวิเคราะห์ความสัมพันธ์ระหว่างตัวแปรเพื่อใช้ในการสร้างเงื่อนไขความน่าจะเป็นสำหรับแต่ละความสัมพันธ์ ทั้งนี้นาอีฟเบย์เป็นวิธีจำแนกข้อมูล ที่มีอัลกอริทึมในการทำงานที่ไม่ซับซ้อนและมีประสิทธิภาพ เนื่องจากสามารถเทรนโมเดลโดยใช้จำนวนชุดของ Training data ได้จำนวนน้อย เหมาะกับกรณีของเซตตัวอย่างที่มีจำนวนมากและคุณลักษณะ (Attribute) ของตัวอย่าง ไม่ขึ้นต่อกัน (พนิดา และคณะ, 2561)

โดยสามารถคำนวณความน่าจะเป็นแบบมีเงื่อนไขของเหตุการณ์ y เมื่อเหตุการณ์ x เกิดขึ้นแล้ว (กิตติศักดิ์ เพชรรุ่งนภา, 2561) ได้ดังสมการ 2.4

$$P(y \mid x) = \frac{P(x \mid y)P(y)}{P(x)} \tag{2.4}$$

เมื่อ $P(y \mid x)$ แทน ค่าความน่าจะเป็นที่เหตุการณ์ y เกิดขึ้น เมื่อเหตุการณ์ x เกิดขึ้นแล้ว

$P(x \mid y)$ แทน ค่าความน่าจะเป็นที่ชุดข้อมูลฝึกฝน ที่เหตุการณ์ x เมื่อเหตุการณ์ y เกิดขึ้น โดยที่ $x = x_1 \cap x_2 \cap x_3 ... \cap x_p$ โดยที่ p คือจำนวนตัวแปร/คุณลักษณะ (Attribute) ในชุดข้อมูลฝึกฝน

$P(y)$ แทน ค่าความน่าจะเป็นของที่เหตุการณ์ y เกิดขึ้น

$P(x)$ แทน ค่าความน่าจะเป็นของที่เหตุการณ์ x เกิดขึ้น

การจำแนกตัวแบบในทฤษฎีของเบย์มีการประยุกต์ใช้กับการวิเคราะห์ข้อมูลที่มีขนาดใหญ่และ ตัวแปรอิสระที่มีจำนวนมาก ดังสมการ 2.5

$$P(y_k \mid x_1, x_2, ..., x_j) = \frac{P(x_1, x_2, ..., x_j \mid y)P(y_k)}{P(x_1, x_2, ..., x_j)} \tag{2.5}$$

เมื่อ y_k แทน ตัวแปรตามกลุ่มที่ k โดยที่ $k = 1, 2, ..., m$

x_j แทน ตัวแปรตามอิสระ j โดยที่ $j = 1, 2, ..., p$

โดยการวิเคราะห์จำแนกกลุ่ม โดยพิจารณาทีจะกลุ่ม ว่ากลุ่มใดมีค่าความน่าจะเป็นสูงที่สุด ซึ่งสามารถคำนวณได้ดังสมการ 2.6

$$P(y_j \mid x_1, x_2, ..., x_p)P(y_j) = P(x_1 \mid y_k)P(x_2 \mid y_k)...P(x_p \mid y_k)P(y_k) \quad (2.6)$$

2.4 วิธีเพื่อนบ้านใกล้เคียง k อันดับ (K-Nearest Neighbors หรือ KNN)

วิธีเพื่อนบ้านใกล้เคียง k อันดับ เป็นวิธีที่ใช้ในการจัดแบ่งคลาส โดยเทคนิคนี้จะตัดสินใจว่าคลาสใดที่จะแทนเงื่อนไขหรือกรณีใหม่ๆ โดยการตรวจสอบจำนวนบางจำนวน “k” ในขั้นตอนวิธีการหาเพื่อนบ้านใกล้ที่สุดของกรณีหรือเงื่อนไขที่เหมือนกันหรือใกล้เคียงกันมากที่สุด โดยจะหาผลรวม (Count Up) ของจำนวนเงื่อนไข หรือกรณีต่างๆ สำหรับแต่ละกลุ่ม และในการใช้วิธีเพื่อนบ้านใกล้เคียง k อันดับนั้นเป็นการวิเคราะห์ข้อมูลใหม่จากข้อมูลเดิมที่อยู่ในบริเวณใกล้เคียงกันมากที่สุด โดยจำนวนที่เรากำหนด k ตัว หรืออธิบายได้ว่า ในการกำหนดค่า k คือการกำหนดว่าจะวิเคราะห์ข้อมูลที่ใกล้ข้อมูลที่ต้องการจำแนกที่สุดกี่ข้อมูล ซึ่งวิธีนี้เหมาะสำหรับข้อมูลแบบตัวเลข ส่วนตัวแปรที่เป็นค่าแบบไม่ต่อเนื่องนั้นสามารถทำได้แต่ต้องการการจัดการในรูปแบบพิเศษเพิ่มขึ้น (สมศักดิ์ และสมัย, 2563)

2.4.1 ขั้นตอนการทำงานของวิธีเพื่อนบ้านใกล้เคียง k อันดับ

1) กำหนดขนาดของ k ตัวอย่างเช่น k = 3 นั่นคือจะพิจารณาเฉพาะข้อมูล 3 ตัวแรกที่อยู่ใกล้บริเวณที่ต้องการทำนาย ซึ่งควรกำหนดให้เป็นเลขคี่ ได้แก่ 3, 5, 7 และ 9 เป็นต้น

2) คำนวณระยะห่างของข้อมูลที่ต้องการทำนายกับกลุ่มข้อมูลตัวอย่าง โดยใช้ระยะห่างยูคลิเดียน (Euclidian distance) หรือเรียกอีกอย่างว่า ระยะทางปกติระหว่างจุดสองจุดในแนวเส้นตรง (จิตกานต์ และคณะ, 2561) ดังสมการ 2.7

$$dist(x_i, x_j) = \sqrt{\sum_{k=1}^{m}(x_{i,k} - x_{j,k})^2} \quad (2.7)$$

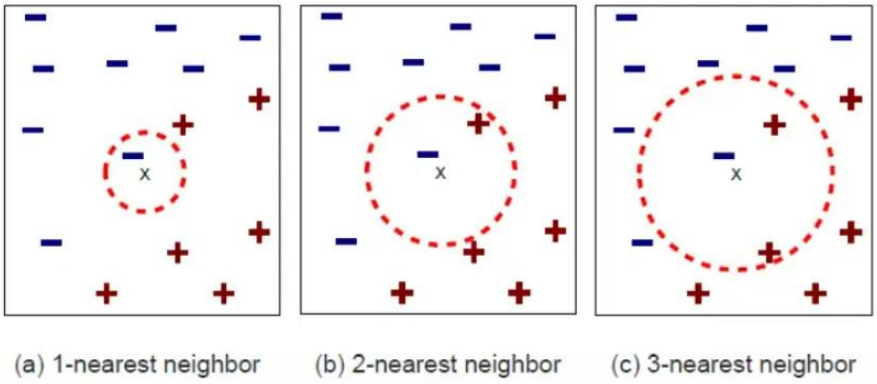
เมื่อ $dist(x_i, x_j)$ คือ ระยะห่างระหว่างตัวอย่าง x_i กับตัวอย่าง x_j

m คือ จำนวนคุณสมบัติทั้งหมดของตัวอย่าง

$x_{i,k}$ คือ คุณสมบัติที่ k ของตัวอย่าง x_i

$x_{j,k}$ คือ คุณสมบัติที่ k ของตัวอย่าง x_j

3) จัดเรียงลำดับของระยะห่าง และเลือกพิจารณาชุดข้อมูลที่ใกล้จุดที่ต้องการทำนายตามขนาดของ k ที่กำหนดไว้ เพื่อนำมาพิจารณาหาผลลัพธ์ ดังรูปที่ 2.2



รูปที่ 2.2 แสดงตัวอย่างวิธีเพื่อนบ้านใกล้ที่สุด k ตัว

เมื่อ (a) วิธีเพื่อนบ้านใกล้ที่สุด k ตัว โดยพิจารณาจากข้อมูล 1 ตัว

(b) วิธีเพื่อนบ้านใกล้ที่สุด k ตัว โดยพิจารณาจากข้อมูล 2 ตัว

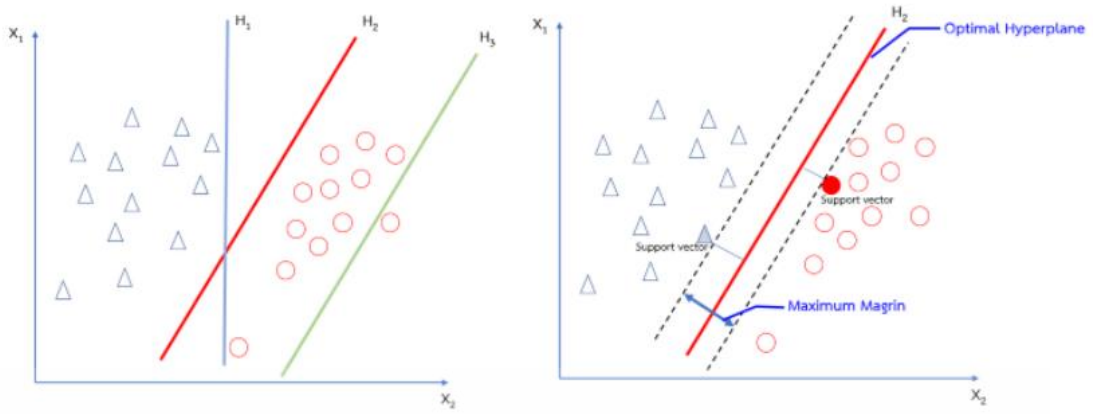
(c) วิธีเพื่อนบ้านใกล้ที่สุด k ตัว โดยพิจารณาจากข้อมูล 3 ตัว

4) พิจารณาข้อมูลจำนวน k ชุด และสังเกตว่ากลุ่มไหนที่ใกล้จุดที่ต้องการทำนายเป็นจำนวนมากที่สุด

5) กำหนดกลุ่มให้กับจุด หรือบริเวณที่ต้องการทำนาย กลุ่มที่ใกล้จุดหรือบริเวณที่ต้องการทำนายมากที่สุด

2.5 วิธีซัพพอร์ตเวกเตอร์แมชชีน (Support Vector Machine)

วิธีซัพพอร์ตเวกเตอร์แมชชีน เป็นวิธีการเรียนรู้ด้วยเครื่อง (Machine Learning) ที่มีพื้นฐานมาจากทฤษฎีการเรียนรู้จากสถิติ เป็นเทคนิคหนึ่งที่ได้รับความนิยมอย่างแพร่หลายในงานที่เกี่ยวข้องกับการจดจำรูปแบบตลอดจนการแก้ปัญหาการจัดกลุ่ม (Classification Problem) (Wang, et al. ,2009) โดยอาศัยหลักการของการหาสัมประสิทธิ์ของสมการเพื่อสร้างเส้นแบ่งแยกกลุ่มข้อมูลหรือไฮเปอร์เพลน (Hyperplane) ที่ถูกป้อนเข้าสู่กระบวนการสอนให้ระบบเรียนรู้ โดยเน้นจากเส้นแบ่งแยกกลุ่มของข้อมูลได้ดีที่สุด (Optimal Hyperplane) (Chen, et al. ,2009) วิธีซัพพอร์ตเวกเตอร์แมชชีน ถึงแม้จะเป็นวิธีที่ถูกออกแบบมาเพื่อการจำแนกแบบทวิภาค แต่สามารถนำไปประยุกต์ใช้กับการจำแนกแบบพหุได้ด้วย ดังรูปที่ 2.3



รูปที่ 2.3 ตัวอย่างวิธีซัพพอร์ตเวกเตอร์แมชชีน 2 มิติ

จากในรูปที่ 2.3 ข้อมูลเป็นลักษณะแบบทวิภาคเมื่อต้องการพิจารณาจำแนกข้อมูลเป็น 2 กลุ่ม โดยใช้ไฮเปอร์เพลนที่เป็นเส้นตรง จะเห็นว่าไฮเปอร์เพลน H_1 และ H_2 สามารถใช้ในการแบ่งแยกคลาสของข้อมูลได้เหมือนกัน แต่ไฮเปอร์เพลน H_2 จะถูกให้เลือกให้เป็นไฮเปอร์เพลนที่ดีที่สุด เรียกว่า (Optimal Line) เนื่องจากมีระยะในการแบ่งจากไฮเปอร์เพลนไปถึงเส้นที่ลากผ่านข้อมูลที่ใกล้ที่สุดนั้นกว้างกว่า H_1 ดังสมการที่ 2.8

$$D = \{(x_i, y_i; = 1, 2, ..., n)\} \tag{2.8}$$

เมื่อ $x_i = (x_{i1}, x_{i2}, ..., x_{im}) \in R^m$

$y_i \in \{1, -1\}$ โดย 1 คือ ข้อมูลกลุ่ม 1 และ -1 คือ ข้อมูลกลุ่ม 2

ซึ่งเป็นการกำหนดกลุ่มเป้าหมายให้ซัพพอร์ตเวกเตอร์แมชชีน โดยที่ ซัพพอร์ตเวกเตอร์แมชชีนนั้นมุ่งเป้าเพื่อหาฟังก์ชันการตัดสินใจที่สามารถแบ่งแยกค่าที่ไม่ทราบได้ดังสมการที่ 2.9

$$f(x) = sign\left\{\sum_{k=1}^{n_v} w_k \varphi_k(x)\varphi_k(x_k) + b\right\} \tag{2.9}$$

$$\varphi(x) = \left[\varphi_1(x_1), \varphi_2(x_2), ..., \varphi_n(x_{n_v})\right]^T \tag{2.10}$$

กลุ่มข้อมูล x จากสมการที่ 2.10 ไม่สามารถแบ่งแยกได้ด้วยสมการเส้นตรง แต่จะถูกแปลงให้อยู่ในรูปแบบที่สามารถใช้สมการเส้นตรงแบ่งแยกได้ โดยใช้เคอร์เนลฟังก์ชัน (Kernel Function) ดังสมการที่ 2.11

$$K(x, x_k) = \varphi(x)\varphi(x_k) \tag{2.11}$$

เมื่อ $\varphi(x)$ แทน ฟังก์ชันสำหรับแปลงข้อมูลที่ไม่เป็นเชิงเส้นให้เป็นข้อมูลที่อยู่ในรูปเชิงเส้นสามารถแบ่งแยกได้

w_k แทน ค่าน้ำหนักที่เชื่อมโยงจากพื้นที่ของคุณลักษณะ (Feature Space) ไปสู่พื้นที่ผลลัพธ์ (Output space)

b แทน ค่าโน้มเอียง (bias)

x_k แทน ซัพพอร์ตเวกเตอร์ โดย $k = 1, 2, ..., n_v$

n_v แทน จำนวนซัพพอร์ตเวกเตอร์

วิธีการที่ใช้ในการหาเส้นแบ่งที่ดีที่สุดคือการเพิ่มเส้นขอบ (Margin) ให้กับเส้นแบ่งทั้งสองข้าง และสร้างเส้นขอบที่สัมผัสกับค่าข้อมูลในพื้นที่ของคุณลักษณะ ที่ใกล้ที่สุดดังนั้นเส้นแบ่งที่มีเส้นขอบกว้างที่สุดจึง เป็นเส้นแบ่งที่ดีที่สุดและเรียกตำแหน่งการสัมผัสข้อมูลที่ใกล้ที่สุดจากการเพิ่มขอบนี้ว่า ซัพพอร์ตเวกเตอร์ (Support Vector) เนื่องจากในบางกรณีการแบ่งแยกกลุ่มไม่สามารถทำได้ถูกต้องโดยสมบูรณ์ ดังนั้นจึงต้องมีการกำหนดตัวแปรสำหรับยอมรับค่าความผิดพลาดโดยการเพิ่มตัวแปร ξ (Slack Variable) คือค่าคลาดเคลื่อนของข้อมูลที่ไม่อยู่บนระนาบ ซึ่งเป็นตัวแสดงความผิดพลาดจากการประมาณค่า ดังสมการที่ 2.12 และ 2.13 ดังนี้

$$w^T x + b \geq y - \xi_i \text{ เมื่อกำหนดให้ } y = 1 \quad (2.12)$$

$$w^T x + b \leq y + \xi_i \text{ เมื่อกำหนดให้ } y = -1 \quad (2.13)$$

โดยที่ w คือค่าความชัน

จากการกำหนดค่า $\xi > 0$ ทำให้โครงสร้างของซัพพอร์ตเวกเตอร์แมชชีนบรรลุวัตถุประสงค์ใน 2 ส่วน คือการเพิ่มระยะแบ่งแยกให้มากที่สุดและลดข้อผิดพลาดในการทำนายให้ต่ำที่สุด ดังสมการที่ 2.14

$$Minimize \frac{1}{2}\|w\|^2 + c\sum_{i=1}^{n}\xi_i \quad (2.14)$$

โดยที่ $y_i(w^T\varphi(x) + b) + \xi_i - 1 \geq 0$

$\xi_i \geq 0, i = 1, 2, ..., n$

นอกจากนี้สามารถใช้ วิธีซัพพอร์ตเวกเตอร์แมชชีน ในปัญหาที่ลักษณะข้อมูลไม่เป็นเชิงเส้นได้ ด้วยวิธีการ เคอร์เนล (Kernel) ในการหารูปแบบลักษณะและความสัมพันธ์ของข้อมูล ซึ่งเคอร์เนลฟังก์ชัน ที่นิยมใช้คือ

ฟังก์ชันพหุนาม (Polynomial) : $K(x_i, x_j) = (x_i^T x_j + r)^\gamma; \gamma > 0$

ฟังก์ชันรัศมีฐานหลัก (Radial Basis Function : RBF) : $K(x_i, x_j) = \exp(-\gamma\|x_i - x_j\|^2); \gamma > 0$

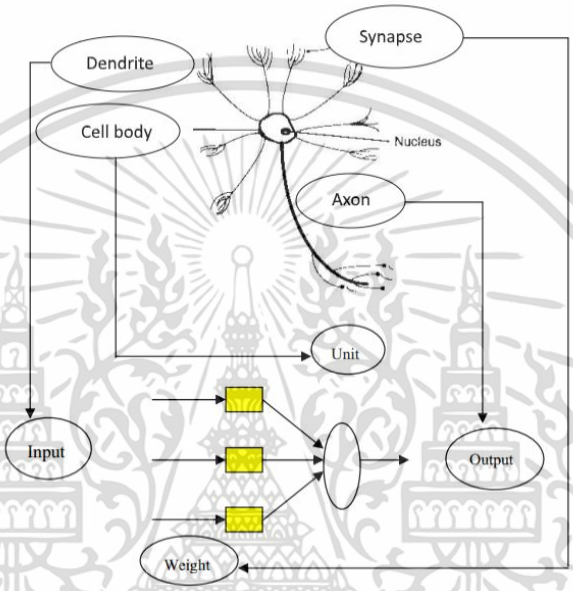
ฟังก์ชันซิกมอยด์ (Sigmoid) : $K(x_i, x_j) = \tanh(\gamma x_i^T x_j - r)$

สำหรับลักษณะข้อมูลที่เป็นเชิงเส้นตรง

เส้นตรง (Linear) : $K(x_i, x_j) = (x_i^T x_j)$

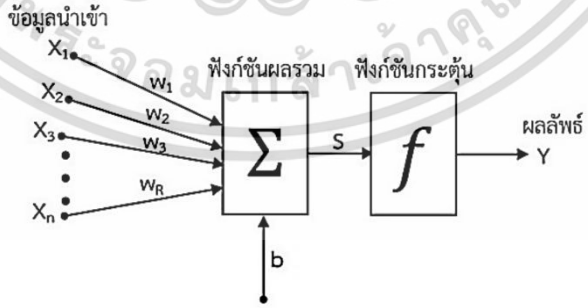
2.6 วิธีโครงข่ายประสาทเทียม (Artificial Neural Network)

วิธีโครงข่ายประสาทเทียม เป็นแนวคิดที่ถูกออกแบบให้ทำงานเช่นเดียวกับสมองมนุษย์ ซึ่งประกอบไปด้วยหน่วยประมวลผล (Processing Elements) ซึ่งมีเซลล์หลายๆ ตัวที่ทำหน้าที่คล้ายกับเซลล์สมองของมนุษย์ โดยที่แต่ละเซลล์จะโยงใยติดต่อกันโดยส่งสัญญาณออกเป็นเอาต์พุต (Output) ของส่วนที่เรียกว่า เดนไดรต์ (Dendrites) และเมื่อผ่านกระบวนการประมวลผลจะได้เอาต์พุตออกมาในส่วนที่เรียกว่า แอ็กซอน (Axon) ในแต่ละเซลล์จะรับรู้ข้อมูลจากหลายทาง แล้วส่งต่อไปยังเซลล์อื่นๆ โดยใช้หลักการ ของการเชื่อมโยงเซลล์สมอง (เสรี, 2544) ดังรูปที่ 2.4



รูปที่ 2.4 แบบจำลองการทำงานของโครงข่ายประสาทเทียม

โดยโครงข่ายประสาทเทียมจะมีการป้อนข้อมูลเข้าและการกำหนดค่าน้ำหนัก ซึ่งแบ่งชั้น (Layer) การทำงานออกเป็น 3 ชั้น (อกนิษฐ์ และคณะ, 2562)



รูปที่ 2.5 ตัวอย่างวิธีโครงข่ายประสาทเทียม

1. ชั้นข้อมูลนำเข้า (Input layer) ทำหน้าที่นำเข้าข้อมูลเข้าสู่โครงข่ายประสาทเทียม โดยข้อมูลนี้จะนำไปประมวลผลในแต่ละโหนด (Node) ของชั้นถัดไป ดังรูปที่ 2.5

2. ชั้นซ่อน (Hidden layer) ทำหน้าที่รับข้อมูลจากชั้นนำเข้าโดยการกำหนดค่าน้ำหนักของข้อมูลจากชั้นนำเข้าก่อนรับข้อมูลเข้าสู่ชั้นซ่อน ชั้นซ่อนจะทำหน้าที่เพิ่มประสิทธิภาพในการจัดกลุ่มข้อมูลก่อนจะส่งต่อข้อมูลไปยังชั้นต่อไป

3. ชั้นข้อมูลส่งออก (Output layer) ทำหน้าที่ส่งออกข้อมูลโดยผ่านการประมวลผลจากฟังก์ชันผลรวม (Summation function: S) เป็นผลรวมของข้อมูลป้อนเข้าที่ถ่วงด้วยค่าน้ำหนัก (W) กับค่าความเอนเอียง (b) และฟังก์ชันการแปลง (Transfer function) จนได้ผลลัพธ์ ดังสมการ

$$S = \sum_{i=1}^{n} W_i X_i + b \tag{2.15}$$

ฟังก์ชันกระตุ้น (Activation Function) เป็นส่วนที่ทำหน้าที่แปลงผลรวมของข้อมูลป้อนเข้าให้เป็นผลลัพธ์ในชั้นส่งออกโดยมีค่าอยู่ระหว่าง 0 ถึง 1 ซึ่งเป็นฟังก์ชันกระตุ้นซิกมอยด์ (Sigmoid Activation Function) สามารถคำนวณได้ ดังสมการ

$$f(S) = \frac{1}{1+e^{-S}} \tag{2.16}$$

โครงข่ายประสาทเทียมสามารถแบ่งได้ 3 แบบ ได้แก่

1. โครงข่ายประสาทแบบป้อนไปข้างหน้าชั้นเดียว (Single-layer feed forward neural networks) ประกอบด้วยชั้นสัญญาณประสาทขาเข้าและชั้นสัญญาณประสาทขาออกเท่านั้น

2. โครงข่ายประสาทแบบป้อนไปข้างหน้าหลายชั้น (Multi-layer feed forward neural networks) มีลักษณะเช่นเดียวกับโครงข่ายประสาทแบบป้อนไปข้างหน้าชั้นเดียว แต่จะมีชั้นซ่อนเพิ่มขึ้น โดยอยู่ระหว่างชั้นข้อมูลนำเข้าและชั้นส่งออก

3. โครงข่ายประสาทแบบแพร่ย้อนกลับ (Back propagation neural networks) มีลักษณะเช่นเดียวกับโครงข่ายประสาทแบบป้อนไปข้างหน้า ซึ่งมีชั้นซ่อนอย่างน้อย 1 ชั้น และมีการวนซ้ำแบบป้อนย้อนกลับอย่างน้อยหนึ่งครั้ง

โครงข่ายประสาทเทียมมีขั้นตอนการเรียนรู้ที่แตกต่างกัน สำหรับการวิจัยนี้ใช้โครงข่ายประสาทเทียมแบบแพร่ย้อนกลับโดยมีขั้นตอนการดำเนินการ ดังนี้

3.1 กำหนดชั้นข้อมูลนำเข้า ชั้นซ่อนและชั้นข้อมูลส่งออก อัตราการเรียนรู้ จำนวนรอบค่าความคลาดเคลื่อนและค่าน้ำหนักให้โครงข่ายประสาท

3.2 คำนวณค่าผลลัพธ์ของโครงข่าย และปรับค่าน้ำหนักในกรณีที่เพอร์เซ็ปตรอนจำแนกข้อมูลผิดพลาด

3.3 ทำซ้ำชุดข้อมูลการเรียนรู้ (Learning data set) จนกระทั่งเพอร์เซ็ปตรอนจำแนกข้อมูลได้ผลลัพธ์ตามค่าเป้าหมายที่กำหนดไว้

ประเภทของการเรียนรู้ของวิธีโครงข่ายประสาทเทียม

1. การเรียนรู้แบบมีผู้สอน (Supervised Learning)

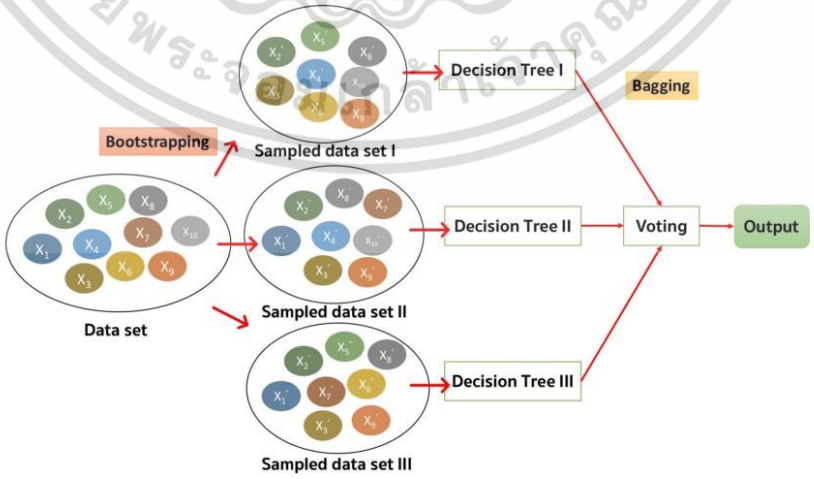
เป็นการเรียนแบบที่มีการตรวจคำตอบเพื่อให้โครงข่ายประสาทเทียมปรับตัว ชุดข้อมูลที่ใช้สอนโครงข่ายประสาทเทียมจะมีคำตอบไว้คอยตรวจดูว่าโครงข่ายประสาทเทียมให้คำตอบที่ถูกหรือไม่ ถ้าตอบไม่ถูก โครงข่ายประสาทเทียมก็จะปรับตัวเองเพื่อให้ได้คำตอบที่ดีขึ้น เปรียบเทียบกับคนเหมือนกับการสอนนักเรียนโดยมีครูผู้สอนคอยแนะนำ

2. การเรียนรู้แบบไม่มีผู้สอน (Unsupervised Learning)

เป็นการเรียนแบบไม่มีผู้แนะนำ ไม่มีการตรวจคำตอบว่าถูกหรือผิด โครงข่ายประสาทเทียมจะจัดเรียงโครงสร้างด้วยตัวเองตามลักษณะของข้อมูล ผลลัพธ์ที่ได้ โครงข่ายประสาทเทียมจะสามารถจัดหมวดหมู่ของข้อมูลได้ เปรียบเทียบกับคน เช่น การที่สามารถจำแนกพันธุ์พืช พันธุ์สัตว์ตามลักษณะรูปร่างของมันได้เองโดยไม่มีใครสอน

2.7 วิธีป่าสุ่ม (Random Forest)

สำหรับวิธีป่าสุ่ม เป็นวิธีที่นำแบบจำลองสำหรับการจำแนกประเภทข้อมูลหลายๆแบบจำลองมารวมกัน (Ensemble Model) ซึ่งจะนำเอาวิธีต้นไม้ตัดสินใจ ที่ได้หลายๆ Tree มาเทรนข้อมูลร่วมกัน โดยมีหลักการนำเอาข้อมูลมาเทรนกับแบบจำลองหลายๆครั้ง บนข้อมูลชุดเดียวกัน และใช้วิธีการตัดสินใจของอัลกอริทึมนั้นมาร่วมลงคะแนน หรือ โหวต (Vote) คลาสใดถูกเลือกมากที่สุดจากการสร้างวิธีต้นไม้ตัดสินใจ สำหรับกรณี การจำแนกประเภทข้อมูล หรือ ในกรณีของสมการเชิงเส้นกำหนดให้ทำการหาค่ากลางของข้อมูล (Mean) จากผลลัพธ์ของวิธีต้นไม้ตัดสินใจ แต่ละต้น ซึ่งภายในอัลกอริทึมของต้นไม้แต่ละต้น ต้องเรียนรู้กันอย่างเป็นอิสระต่อกันมากที่สุด ดังรูปที่ 2.6 (สุภาภรณ์, 2564)



รูปที่ 2.6 หลักการทำงานของวิธีป่าสุ่ม และเทคนิคการจำแนก

ประเภทข้อมูลแบบ Bagging (Bootstrap Aggregation) โดยต้นไม้แต่ละต้นที่นำมาเทรนในแบบจำลอง จะมีตัวแปรแต่ละตัวเป็นส่วนหนึ่งของตัวแปรหรือฟีเจอร์ ซึ่งจะนำมาเทรนในรูปแบบสุ่ม (Random) และในส่วนขั้นตอนการทำนายข้อมูลนั้น จะกำหนดให้แต่ละต้นไม้ ทำนายในต้นของตัวเองและคัดเลือกผลทำนายสุดท้ายจากค่าทำนายที่ได้รับการโหวตมากที่สุด เทคนิคดังกล่าวเรียกว่า "การสุ่มตัวอย่างข้อมูล" และเทคนิคการจำแนกประเภทข้อมูลแบบ Bagging (Bootstrap Aggregation)

2.8 งานวิจัยที่เกี่ยวข้อง

อัจฉราภรณ์ และคณะ (2563) ได้ทำการศึกษาแบบจำลองการวินิจฉัยอัตโนมัติสำหรับความเสี่ยงต่อการเกิดลิ่มเลือดอุดตันในหลอดเลือดดำตามอาการ โดยอาศัยวิธีการเรียนรู้ของเครื่อง งานวิจัยนี้ ได้สร้างแบบจำลองการวินิจฉัยความเสี่ยงต่อการเกิดลิ่มเลือดอุดตันในหลอดเลือดดำโดยอาศัยหลักการเรียนรู้ของเครื่อง จากการเก็บรวบรวมข้อมูลผู้ป่วยในหอผู้ป่วยอายุรศาสตร์ โรงพยาบาลจุฬาลงกรณ์ สภากาชาดไทย งานวิจัยนี้ได้เตรียมข้อมูลทั้งหมด 1,290 แถว 65 คอลัมน์ และตรวจสอบค่าที่ขาดหายไปพร้อมทั้งแปลงข้อมูลให้พร้อมสำหรับนำไปสร้างแบบจำลองการทำนายในลำดับต่อไป จากนั้นแบ่งข้อมูลสำหรับการฝึกสอนและข้อมูลสำหรับการทดสอบในอัตราส่วน 70:30 ผลการทดลองของงานวิจัยได้เปรียบเทียบประสิทธิภาพของแต่ละ 3 วิธี ประกอบด้วย ต้นไม้ตัดสินใจ การวิเคราะห์การถดถอยโลจิสติกส์ และ โครงข่ายประสาทเทียม จากผลการทดลองแบบจำลองต้นไม้ตัดสินใจมีประสิทธิภาพที่สุด มีค่าความถูกต้องสูงที่สุด 96.6% โดยการปรับสมดุลของข้อมูลด้วยวิธีการให้น้ำหนักในแต่ละคลาส (Class Weight)

ประยุทธ์ศิลป์ (2562) ได้ศึกษาการสร้างแบบจำลองจำแนกกลุ่มผู้ป่วยโรคไตเรื้อรังโดยใช้เทคนิคเหมืองข้อมูลและการนำเสนอภาพข้อมูล (Visualization) จากผลการเปรียบเทียบความแม่นยำของแต่ละวิธี พบว่า วิธีนาอีฟเบย์มีค่าความแม่นยำ 98.1% และค่าความผิดพลาด 0.13736 วิธีต้นไม้ตัดสินใจและวิธีป่าสุ่มมีค่าความแม่นยำ 98% และค่าความผิดพลาด 0.1443 วิธีซัพพอร์ตเวกเตอร์แมชชีนมีค่าความแม่นยำ 49% และค่าความผิดพลาด 0.54944 วิธีเพื่อนบ้านใกล้เคียง k อันดับ มีค่าความแม่นยำ 96% และค่าความผิดพลาด 0.2041 ทำให้วิธีนาอีฟเบย์เป็นแบบจำลองที่มีความแม่นยำและเหมาะสมสำหรับการพยากรณ์ภาวะโรคไตเรื้อรังมากที่สุดจากแบบจำลองทั้ง 5 แบบ

Tao et al. (2013) ได้ทำการศึกษาการเปรียบเทียบวิธีซัพพอร์ตเวกเตอร์แมชชีนสำหรับการใช้คอมพิวเตอร์ช่วยวินิจฉัยโรคมะเร็งปอด ในการศึกษานี้มีการตรวจสอบการใช้การจำแนกวิธีซัพพอร์ตเวกเตอร์แมชชีนสำหรับมะเร็งปอด โดยเปรียบเทียบกับวิธีต้นไม้ตัดสินใจ เพื่อนบ้านที่ใกล้เคียง k อันดับ โครงข่ายประสาทเทียม และป่าสุ่ม โดยลักษณะเฉพาะของผู้ป่วยและคุณลักษณะทางสัณฐานวิทยา ถูกนำมาใช้ในการฝึกตัวแยกประเภทและประเมินประสิทธิภาพ พบว่าค่าความถูกต้องของ วิธีซัพพอร์ตเวกเตอร์แมชชีนมีค่ามากที่สุด คือ 94% รองลงมาคือ โครงข่ายประสาทเทียม วิธีป่าสุ่ม วิธีต้นไม้ตัดสินใจ และ วิธีเพื่อนบ้านที่ใกล้เคียง k อันดับ ตามลำดับ

นรินทร์ และนิเวศ (2553) ได้ทำการศึกษาการจำแนกมะเร็งเม็ดเลือดขาวโดยใช้เทคนิคการลดมิติข้อมูลด้วยไคแสควร์ งานวิจัยนี้ได้เสนอวิธีการลดมิติของยีน ด้วยค่าสถิติไคแสควร์ (Chi-square) และโดยทำการทดสอบประสิทธิภาพ การจำแนกประเภทของมะเร็งเม็ดเลือดขาวกับ อัลกอริทึมซัพพอร์ตเวกเตอร์แมชชีน ต้นไม้ตัดสินใจ นาอีฟเบย์ วิธีเพื่อนบ้านใกล้เคียง k อันดับ จากการทดลองพบว่า การลดมิติ ของข้อมูลด้วยวิธีไคแสควร์แล้วส่งเข้าประมวลผลด้วยเครื่องจักรการเรียนรู้โดยวัดประสิทธิภาพจากค่าความถูกต้อง จากจำนวนมิติของข้อมูลที่ส่งผลดีที่สุดพบว่า วิธีซัพพอร์ตเวกเตอร์แมชชีน วิธีเพื่อนบ้านใกล้เคียง k อันดับ ที่จำนวน300 มิติ ค่าความถูกต้อง เท่ากับ 98.61% เท่ากัน วิธีนาอีฟเบย์ที่จำนวน 4000 มิติ ค่าความถูกต้อง เท่ากับ 98.61 % วิธีต้นไม้ตัดสินใจ ที่จำนวน 1 มิติ ค่าความถูกต้อง เท่ากับ 93.06%

Danish et al. (2022) ได้ทำการศึกษาวิธีการเรียนรู้ด้วยเครื่องการจำแนกแบบทวิภาคและการจำแนกแบบพหุของข้อมูลโรคหัวใจที่ไม่สมดุล สำหรับการศึกษานี้ หลังจากใช้เทคนิควิธีการเรียนรู้ของเครื่อง เช่น วิธีป่าสุ่ม ต้นไม้ตัดสินใจ วิธี Gradient Boosted Trees (GBT) ซัพพอร์ตเวกเตอร์แมชชีน การวิเคราะห์การถดถอยโลจิสติกส์ เมื่อรวมกับเทคนิค เพอร์เซ็ปตรอนแบบหลายชั้นทำงานได้อย่างเหมาะสมทั้งในแบบการจำแนกแบบทวิภาค (ค่าความถูกต้อง 94.8%) และการจำแนกแบบพหุ (ค่าความถูกต้อง 88.2%) เมื่อเปรียบเทียบกับวิธีการจำแนกประเภททวิภาคอื่นๆ แล้ว GBT ให้ผลลัพธ์ที่ถูกต้อง (ค่าความถูกต้อง 95.8%) อย่างไรก็ตาม วิธีโครงข่ายประสาทเทียมเป็นวิธีที่สามารถจำแนกประเภทได้ดีในแบบการจำแนกแบบพหุ

ธรรมนูญ และคณะ (2565) ได้ศึกษาการเพิ่มประสิทธิภาพจำแนกข้อมูลผลกระทบโควิด-19 ต่อผู้ป่วยมะเร็งตับด้วยเทคนิคการเรียนรู้ของเครื่อง โดยมีวิธีการจำแนกข้อมูลด้วยอัลกอริทึมประกอบด้วย วิธีซัพพอร์ตเวกเตอร์แมชชีน นาอีฟเบย์ เพื่อนบ้านใกล้เคียง k อันดับ ต้นไม้ตัดสินใจ โครงข่ายประสาทเทียม และการเพิ่มประสิทธิภาพการจำแนกข้อมูล ประกอบด้วย การโหวตป่าสุ่ม ชุดข้อมูลทดลองวิจัย ได้แก่ ข้อมูลโควิด-19 ผลกระทบต่อผู้ป่วยมะเร็งตับ แบ่งชุดข้อมูลการทดลองเป็น 2 วิธี คือ แบบแยกชุดข้อมูลเป็น 80:20 และแบ่งแบบ K-Fold Cross Validation เครื่องมือที่ใช้ในการทดลองภาษาไพทอนและโปรแกรม Visual Studio Code วัดประสิทธิภาพด้วยวิธี

ค่าความแม่นยำ ค่าความครบถ้วน F1-Score และค่าความถูกต้อง ผลการวิจัย พบว่า การจำแนกข้อมูลผลกระทบโควิด-19 ต่อผู้ป่วยมะเร็งตับ ด้วยการแบ่งชุดข้อมูล 80:20 วิธีการจำแนกที่ดีที่สุด ได้แก่ วิธีการต้นไม้ตัดสินใจ และวิธีการโครงข่ายประสาทเทียม มีความถูกต้อง 100% การแบ่งชุดข้อมูลทดสอบประสิทธิภาพ 5-Fold Cross Validation วิธีการจำแนกที่ดีที่สุด ได้แก่ วิธีการต้นไม้ตัดสินใจ มีความถูกต้อง 99.6% และ ผลการเพิ่มประสิทธิภาพการจำแนกข้อมูลด้วยการเรียนรู้แบบรวมกลุ่มทดสอบประสิทธิภาพ 5-Fold Cross Validation วิธีการป่าสุ่ม มีความถูกต้อง 99.8%

บทที่ 3
วิธีการดำเนินงานวิจัย

การวิจัยครั้งนี้จะทำการศึกษาการการเปรียบเทียบประสิทธิภาพวิธีการจำแนกการเป็นโรคมะเร็งปอดจากรหัสพันธุกรรมด้วยการเรียนรู้ด้วยเครื่อง ซึ่งจะนำชุดข้อมูลการจำแนกประเภทโรคมะเร็งปอด โดยใช้ วิธีต้นไม้ตัดสินใจ วิธีนาอีฟเบย์ วิธีเพื่อนบ้านใกล้เคียง k อันดับ วิธีซัพพอร์ตเวกเตอร์แมชชีน วิธีโครงข่ายประสาทเทียม และวิธีป่าสุ่ม ในการศึกษาแต่ละวิธีจะใช้โปรแกรม R Studio เวอร์ชัน 4.2.2 เข้ามาช่วยในการวิเคราะห์ข้อมูลและดำเนินการตามวัตถุประสงค์ของงานวิจัยต่อไป ในบทนี้กล่าวถึงการวางแผนการวิจัยและวิธีการดำเนินงานวิจัย

3.1 รายละเอียดของข้อมูล

ข้อมูลที่นำมาใช้ในการดำเนินงานครั้งนี้ ได้ทำการศึกษาจากข้อมูลทุติยภูมิ (Secondary Data) จากเว็บไซต์ https://www.broadinstitute.org/MPR/CNS โดยข้อมูลผู้ป่วยที่หาได้มีจำนวนข้อมูลทั้งหมด 197 คน ซึ่งประกอบด้วย ตัวแปรอิสระ คือ รหัสพันธุกรรมของผู้ป่วย จำนวน 1,000 รหัส และตัวแปรตาม คือประเภทของโรคมะเร็งปอด จำนวน 4 กลุ่ม ดังนี้

1. มะเร็งปอดประเภทเซลล์เล็ก (Small Cell Lung Cancer (SCLC) : Oat cell lung cancer) ในชุดข้อมูลแทนด้วยหมายเลข 0
2. เซลล์มะเร็งประเภทอะดีโนคาร์ซิโนมา (Non-Small Cell lung cancer (NACLC) : Adenocarcinoma) ในชุดข้อมูลแทนด้วยหมายเลข 1
3. เซลล์มะเร็งประเภทสะความัสคาร์ซิโนมา (Non-Small Cell lung cancer (NSCLC) : Squamous Cell carcinoma) ในชุดข้อมูลแทนด้วยหมายเลข 2
4. เซลล์มะเร็งปอดประเภทเซลล์ใหญ่ (Non-Small Cell lung cancer (NSCLC) : Large Cell lung cancer) ในชุดข้อมูลแทนด้วยหมายเลข 3

3.2 เครื่องมือที่ใช้ในการศึกษา

งานวิจัยนี้ผู้วิจัยเลือกใช้เครื่องมือที่ใช้ในการศึกษา ดังนี้

3.2.1 โปรแกรม Microsoft Excel

ใช้ในการจัดเก็บข้อมูลเพื่อนำไปวิเคราะห์ในโปรแกรม RStudio Version 4.2.2

3.2.2 โปรแกรม R Studio Version 4.2.2

เพื่อใช้ในการสร้างตัวแบบการจำแนกด้วยวิธีการเรียนรู้ด้วยเครื่อง 6 วิธี สร้างกราฟ และ คำนวณหาค่าเฉลี่ยความถูกต้อง ค่าเฉลี่ยความแม่นยำ ค่าเฉลี่ยความระลึก

1. แพ็คเกจ MASS เพื่อใช้ในการสนับสนุนฟังก์ชันและชุดข้อมูล

2. แพ็คเกจ neuralnet เพื่อใช้ในการสร้างตัวแบบของวิธีโครงข่ายประสาทเทียม

3. แพ็คเกจ rpart เพื่อใช้ในการสร้างตัวแบบของวิธีต้นไม้ตัดสินใจ

4. แพ็คเกจ class เพื่อใช้ในการจำแนกประเภทของข้อมูล (Classification)

5. แพ็คเกจ e1071 เพื่อใช้ในการวิเคราะห์ทางสถิติ และความน่าจะเป็น

6. แพ็คเกจ caTools เพื่อใช้ในการช่วยย้ายหน้าต่างทางสถิติ เช่น Receiver operating characteristic curve (ROC) และ Area under the curve (AUC) เป็นต้น

7. แพ็คเกจ randomForest เพื่อใช้ในการสร้างตัวแบบของวิธีป่าสุ่ม

8. แพ็คเกจ corrplot เพื่อใช้ในการสร้างตารางแสดงความสัมพันธ์ของข้อมูล

9. แพ็คเกจ RColorBrewer เพื่อใช้ในการเลือกชุดสี

3.3 วิธีการวิเคราะห์ข้อมูลและจัดเตรียมข้อมูล

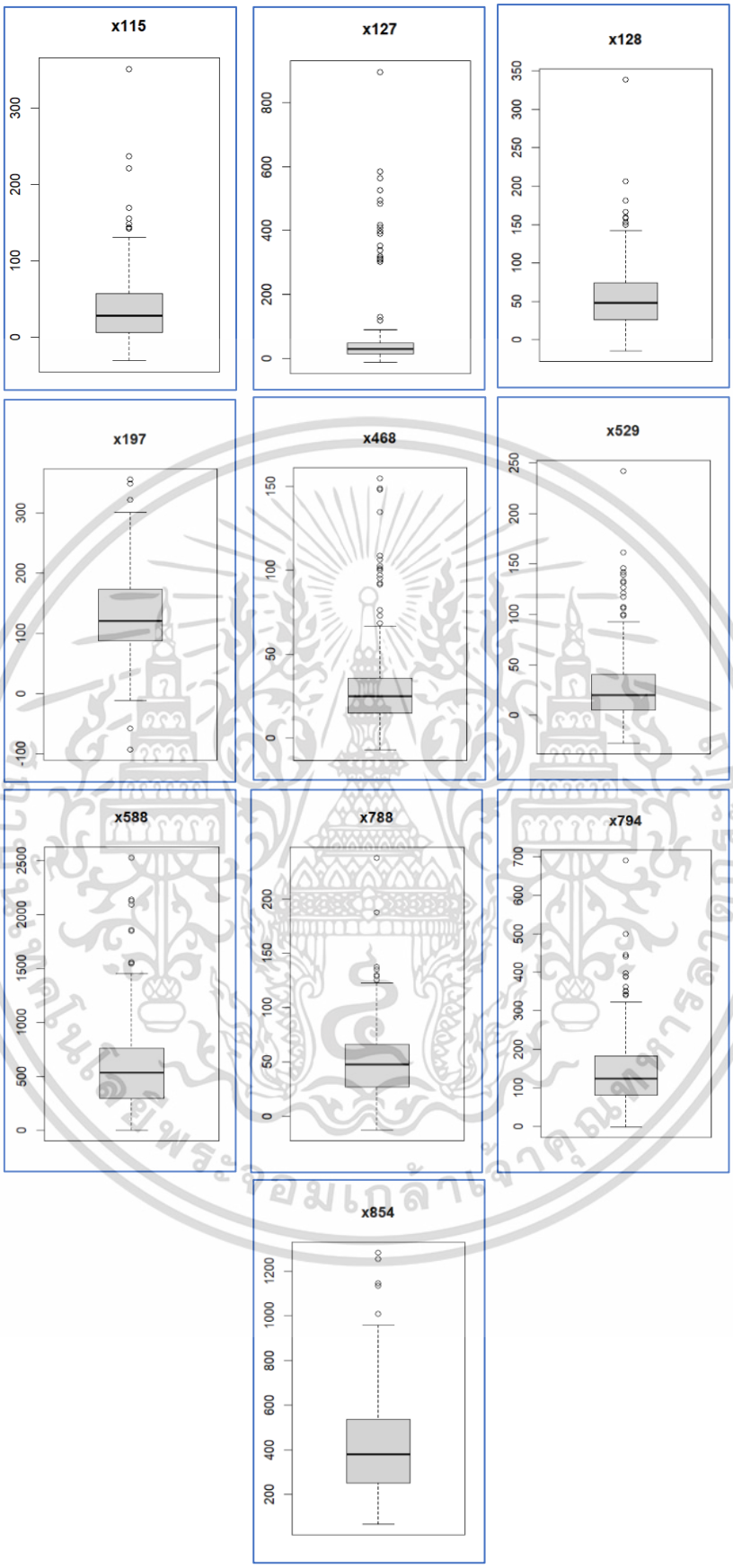
3.3.1 ศึกษารายละเอียดของข้อมูล

ผู้วิจัยได้นำเสนอตัวอย่างข้อมูลของตัวแปรอิสระจำนวน 10 ชุด ได้ดังตารางที่ 3.1

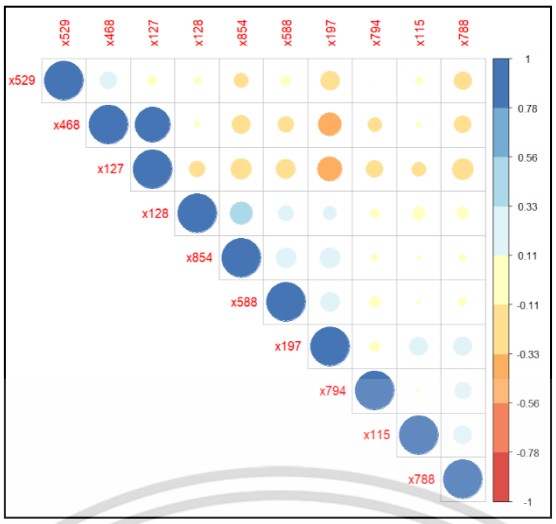
ตารางที่ 3.1 ตารางแสดงตัวอย่างข้อมูลจากตัวแปรอิสระ 10 ชุด

ลำดับ	y	x115	x127	x128	x197	x468	x529	x588	x788	x794	x854
1	0	16.21	35.64	20.25	60.77	20.25	107.82	1223.31	34.83	261.67	692.5
2	0	102.55	20.8	105.73	76.01	23.99	37.79	684.41	50.53	46.28	707.32
3	0	52.74	22.875	59.995	172.305	23.26	14.36	583.75	35.125	68.82	276.75
4	0	80.585	28.09	56.385	139.515	32.76	28.275	1150.59	72.18	155.63	491.275
5	0	56.74	18.27	79	183.1	11.18	17.26	2092.92	50.67	33.46	402.59
6	0	220.84	24.39	32.58	168.26	9.17	12.68	219.68	41.94	57.15	115.64
7	0	65.325	52.515	68.25	154.79	34.23	21.59	1051.93	50.135	141.175	525.25
8	0	52.97	20.73	58.01	153.98	5.63	23.75	529.73	75.16	132.73	534.88
9	0	94.89	27.28	134.77	135.085	14.31	24.64	747.125	49.395	145.795	441.58
10	0	142.31	54.12	149.37	208.17	37.67	41.19	308.14	60	219.93	608.82
.	.	.	.	.	.	.	.	.	.	.	.
.	.	.	.	.	.	.	.	.	.	.	.
.	.	.	.	.	.	.	.	.	.	.	.
195	3	11.3	484.58	31.23	31.23	60.7	36.43	299.45	15.63	65.9	129.22
196	3	17.58	318.09	59.38	68.28	54.04	14.02	234.1	27.36	68.28	171.63
197	3	-6.09	583.64	3.3	40.17	154.58	58.97	238.84	-8.26	223.56	360.58

จากตัวอย่างข้อมูลของตัวแปรอิสระจำนวน 10 ชุด นำมาสร้างเป็นแผนภาพกล่อง และรูปแสดงความสัมพันธ์ระหว่างตัวแปร เพื่อศึกษารายละเอียดของข้อมูล ดังรูปที่ 3.1 และ 3.2



รูปที่ 3.1 รูปแสดงตัวอย่างข้อมูลตัวแปรอิสระ 10 ชุด ด้วยแผนภาพกล่อง



รูปที่ 3.2 รูปแสดงความสัมพันธ์ระหว่างตัวแปร ตัวแปรอิสระ 10 ชุด

จากรูปที่ 3.1 ตัวอย่างของข้อมูลตัวแปรอิสระ 10 ตัว พบว่า ข้อมูลในแต่ละตัวแปรมีการกระจายของข้อมูลไม่เท่ากัน และข้อมูลมีค่านอกเกณฑ์ จากรูปที่ 3.2 มีโอกาสในการเกิดความสัมพันธ์เชิงเส้นแบบพหุ จึงแสดงความสัมพันธ์ของตัวแปรอิสระจำนวน 10 ชุด ที่ทำการสุ่มมา

3.3.2 การปรับปรุงชุดข้อมูลอสมดุลให้มีความสมดุล

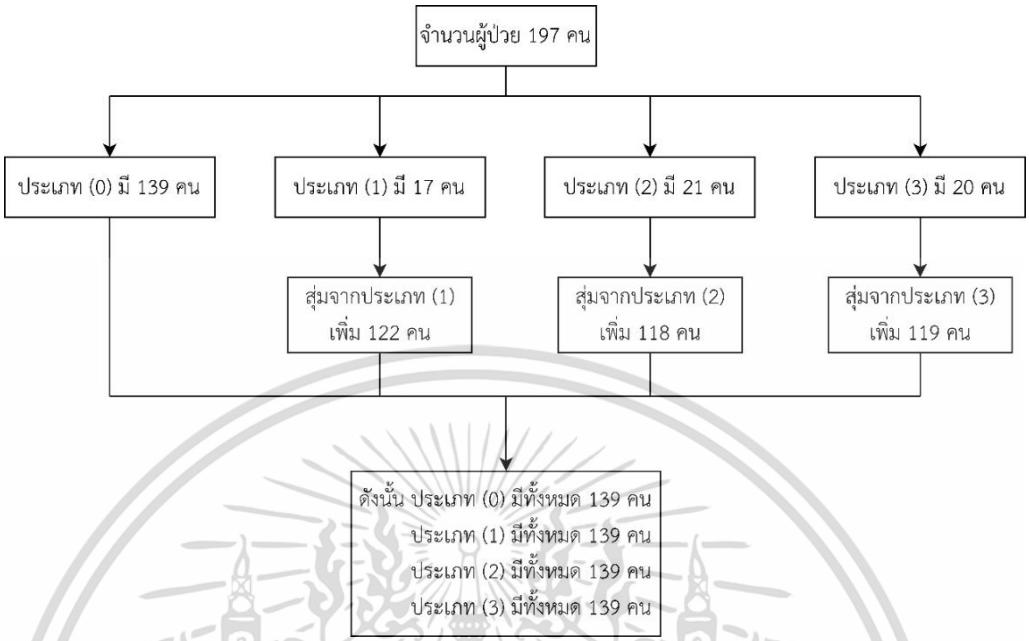
เนื่องจากข้อมูลที่นำมาวิเคราะห์นั้น มีตัวแปรตามในกลุ่มที่มีแทนด้วยหมายเลข 0 มีจำนวนมากกว่าตัวแปรตามของอีกกลุ่มในจำนวนมาก ซึ่งเป็นข้อมูลที่ไม่สมดุล ทำให้สัดส่วนของข้อมูลมีค่าที่แตกต่างกันสูง ดังตารางที่ 3.2

ตารางที่ 3.2 ตารางแสดงสัดส่วนกลุ่มข้อมูลของตัวแปรตาม

ประเภทตัวแปรตาม	จำนวน	สัดส่วน
0	139	1 : 1.417
1	17	1 : 11.588
2	21	1 : 9.381
3	20	1 : 9.85
รวม	197	

จากตารางที่ 3.2 พบว่าสัดส่วนของประเภทตัวแปรตามที่แทนด้วยหมายเลข 0,1,2 และ 3 มีสัดส่วนที่แตกต่างกัน เมื่อเทียบกับจำนวนตัวแปรตามทั้งหมด ซึ่งตัวแปรตามที่แทนด้วยหมายเลข 0 มีสัดส่วนที่มากที่สุด จึงต้องทำการสุ่มตัวอย่างภายในกลุ่มเพื่อทำให้จำนวนในแต่ละกลุ่มมีจำนวนข้อมูลที่เท่ากัน โดยใช้การสุ่มตัวอย่างแบบง่าย (Simple Random Sampling : SRS)

เป็นการสุ่มตัวอย่างที่ทุกๆหน่วยหรือทุกๆสมาชิกในประชากรมีโอกาสถูกเลือกเท่าๆ กันแบบแทนที่ ดังรูปที่ 3.3



รูปที่ 3.3 แผนภาพแสดงการปรับปรุงชุดข้อมูลอสมดุลให้มีความสมดุล

3.4 การประเมินประสิทธิภาพ

โดยเกณฑ์ที่ใช้พิจารณาประกอบด้วย

1. ค่าเฉลี่ยร้อยละความถูกต้อง (Accuracy) ดังสมการ

$$MeanAccuracy = \frac{\sum_{i=1}^{1000} Accuracy}{1000}$$

2. ค่าเฉลี่ยร้อยละความแม่นยำ (Precision)

2.1 ค่าเฉลี่ยร้อยละความแม่นยำของกลุ่มมะเร็งปอดประเภทเซลล์เล็ก (0) คำนวณดังสมการ

$$MeanPrecision(0) = \frac{\sum_{i=1}^{1000} Precision(0)}{1000}$$

2.2 ค่าเฉลี่ยร้อยละความแม่นยำของกลุ่มเซลล์มะเร็งประเภทอะดีโนคาร์ซิโนมา (1) คำนวณดังสมการ

$$MeanPrecision(1) = \frac{\sum_{i=1}^{1000} Precision(1)}{1000}$$

2.3 ค่าเฉลี่ยร้อยละความแม่นยำของกลุ่มเซลล์มะเร็งประเภทสะความัสคาร์ซิโนมา (2) คำนวณดังสมการ

$$MeanPrecision(2) = \frac{\sum_{i=1}^{1000} Precision(2)}{1000}$$

2.4 ค่าเฉลี่ยร้อยละความแม่นยำของกลุ่มเซลล์มะเร็งปอดประเภทเซลล์ใหญ่ (3) คำนวณดังสมการ

$$MeanPrecision(3) = \frac{\sum_{i=1}^{1000} Precision(3)}{1000}$$

3. ค่าเฉลี่ยร้อยละความระลึก (Recall)

3.1 ค่าเฉลี่ยร้อยละความระลึกของกลุ่มมะเร็งปอดประเภทเซลล์เล็ก (0) คำนวณดังสมการ

$$MeanRecall(0) = \frac{\sum_{i=1}^{1000} Recall(0)}{1000}$$

3.2 ค่าเฉลี่ยร้อยละความระลึกของกลุ่มเซลล์มะเร็งประเภทอะดีโนคาร์ซิโนมา (1) คำนวณดังสมการ

$$MeanRecall(1) = \frac{\sum_{i=1}^{1000} Recall(1)}{1000}$$

3.3 ค่าเฉลี่ยร้อยละความระลึกของกลุ่มเซลล์มะเร็งประเภทสะความัสคาร์ซิโนมา (2) คำนวณดังสมการ

$$MeanRecall(2) = \frac{\sum_{i=1}^{1000} Recall(2)}{1000}$$

3.4 ค่าเฉลี่ยร้อยละความระลึกของกลุ่มเซลล์มะเร็งปอดประเภทเซลล์ใหญ่ (3) คำนวณดังสมการ

$$MeanRecall(3) = \frac{\sum_{i=1}^{1000} Recall(3)}{1000}$$

3.5 ขั้นตอนในการดำเนินงานวิจัย

จากการใช้โปรแกรม R Studio ในการจำแนกประเภทมะเร็งปอด 4 ประเภท โดยการสุ่มจำนวนตัวแปรอิสระของแต่ละจำนวนมา แล้วสุ่มข้อมูลชุดฝึกฝนและ ชุดข้อมูลทดสอบ เพื่อนำไปวิเคราะห์ด้วยวิธีการเรียนรู้ด้วยเครื่องทั้งหมด 6 วิธี เพื่อหาค่าร้อยละความถูกต้อง ค่าร้อยละความแม่นยำ และค่าร้อยละความระลึก เนื่องจากวิธีเพื่อนบ้านใกล้เคียง k อันดับ วิธีโครงข่ายประสาทเทียม และ วิธีซัพพอร์ตเวกเตอร์แมชชีน ต้องมีการกำหนดค่าเฉพาะ เพื่อให้ได้ค่าเฉลี่ยร้อยละที่มากที่สุดของแต่ละจำนวนตัวแปรที่สุ่ม จึงแสดงตารางในการดำเนินการหาค่าเฉพาะดังตารางที่ 3.3 3.4 และ 3.5

ตารางที่ 3.3 ตารางหาค่า k ที่ให้ค่าเฉลี่ยร้อยละความถูกต้องที่สูงที่สุดของวิธีการเรียนรู้ด้วยเครื่องวิธีเพื่อนบ้านใกล้เคียง k อันดับ

ค่า k	จำนวนตัวแปรอิสระ (p)				
	200	400	600	800	1000
k = 1	90.6220	91.4051	91.3966	91.5271	91.4831
k = 2	90.3898	91.2881	91.5051	91.5881	91.7729
k = 3	91.7763	92.7288	92.8102	92.7678	92.8593
k = 4	91.7271	92.5356	92.7288	93.0797	92.9593
k = 5	91.9627	92.7729	93.4085	93.4831	**93.9695**
k = 6	**94.3661**	92.8237	94.4864	**93.5000**	-
k = 7	92.0085	**94.6220**	**94.5915**	93.1864	93.8322
k = 8	91.3356	92.4051	92.7475	92.9746	93.3763
k = 9	91.1492	92.0949	92.6305	92.6119	93.0644

หมายเหตุ ตัวเลขที่เป็นตัวหนา หมายถึง ค่าเฉลี่ยร้อยละความถูกต้องที่มากที่สุดตามจำนวนตัวแปรอิสระ และ สัญลักษณ์ (-) หมายถึง ไม่สามารถทำการวิเคราะห์ได้

จากตารางที่ 3.3 พบว่า ในวิธีเพื่อนบ้านใกล้เคียง k อันดับ ที่จำนวนตัวแปรอิสระเท่ากับ 200 และ 800 ควรกำหนดค่า k = 6 ที่จำนวนตัวแปรอิสระเท่ากับ 400 และ 600 ควรกำหนดค่า k = 7 และ ที่จำนวนตัวแปรอิสระเท่ากับ 1,000 ควรกำหนดค่า k = 5 เพื่อให้ได้ค่าเฉลี่ยร้อยละความถูกต้องที่สูงที่สุดตามจำนวนตัวแปร

ตารางที่ 3.4 ตารางหาค่า h ที่ให้ค่าเฉลี่ยร้อยละความถูกต้องที่สูงที่สุดของวิธีการเรียนรู้ด้วยเครื่องวิธีโครงข่ายประสาทเทียม

ค่า h	จำนวนตัวแปรอิสระ (p)				
	200	400	600	800	1000
h = 150	88.5390	88.2119	88.5627	88.6322	88.5932
h = 175	88.5881	88.7424	88.9407	88.4407	88.6288
h = 200	88.6559	88.9085	**89.0729**	88.9576	**89.0780**
h = 225	-	88.8407	88.8119	86.3831	88.8712
h = 250	88.8153	88.8441	88.9051	88.8695	88.8085
h = 275	**88.9763**	**89.0542**	89.0288	89.1017	88.9339
h = 300	88.5932	88.8034	89.0576	**89.3017**	88.9932
h = 325	88.7051	88.3475	88.7915	88.9271	89.0203
h = 350	88.6898	88.1356	88.6517	88.5661	88.2763

หมายเหตุ ตัวเลขที่เป็นตัวหนา หมายถึง ค่าเฉลี่ยร้อยละความถูกต้องที่มากที่สุดตามจำนวนตัวแปรอิสระ และ สัญลักษณ์ (-) หมายถึง ไม่สามารถทำการวิเคราะห์ได้

จากตารางที่ 3.4 พบว่า ในวิธีโครงข่ายประสาทเทียม ที่จำนวนตัวแปรอิสระเท่ากับ 200 และ 400 ควรกำหนดค่า h = 275 ที่จำนวนตัวแปรอิสระเท่ากับ 600 และ 1,000 ควรกำหนดค่า h = 200 และ ที่จำนวนตัวแปรอิสระเท่ากับ 800 ควรกำหนดค่า h = 300 เพื่อให้ได้ค่าเฉลี่ยร้อยละความถูกต้องที่สูงที่สุดตามจำนวนตัวแปร

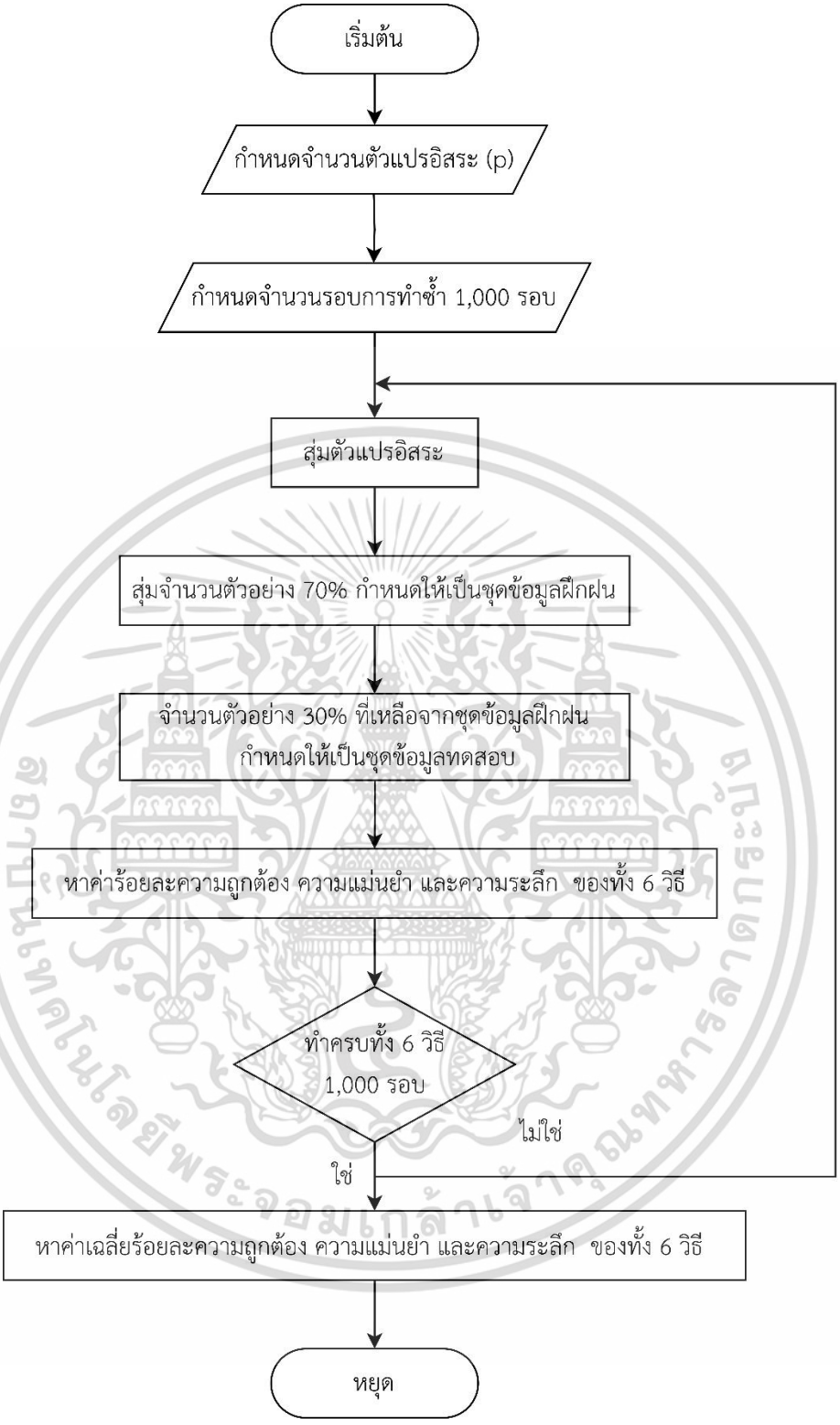
ตารางที่ 3.5 ตารางหา Kernel Function ที่ให้ค่าเฉลี่ยร้อยละความถูกต้องที่สูงที่สุดของวิธีการเรียนรู้ด้วยเครื่อง วิธีซัพพอร์ตเวกเตอร์แมชชีน

Kernel function	จำนวนตัวแปรอิสระ (p)				
	200	400	600	800	1000
Linear	95.7407	94.5170	95.9610	95.9898	96.0102
Polynomial	92.3661	**98.5246**	**98.5976**	**98.6737**	**98.5964**
Radial basis	-	-	-	-	-
Sigmoid	**96.6407**	96.7966	96.8712	96.9661	96.9593

หมายเหตุ ตัวเลขที่เป็นตัวหนา หมายถึง ค่าเฉลี่ยร้อยละความถูกต้องที่มากที่สุดตามจำนวนตัวแปรอิสระ และ สัญลักษณ์ (-) หมายถึง ไม่สามารถทำการวิเคราะห์ได้

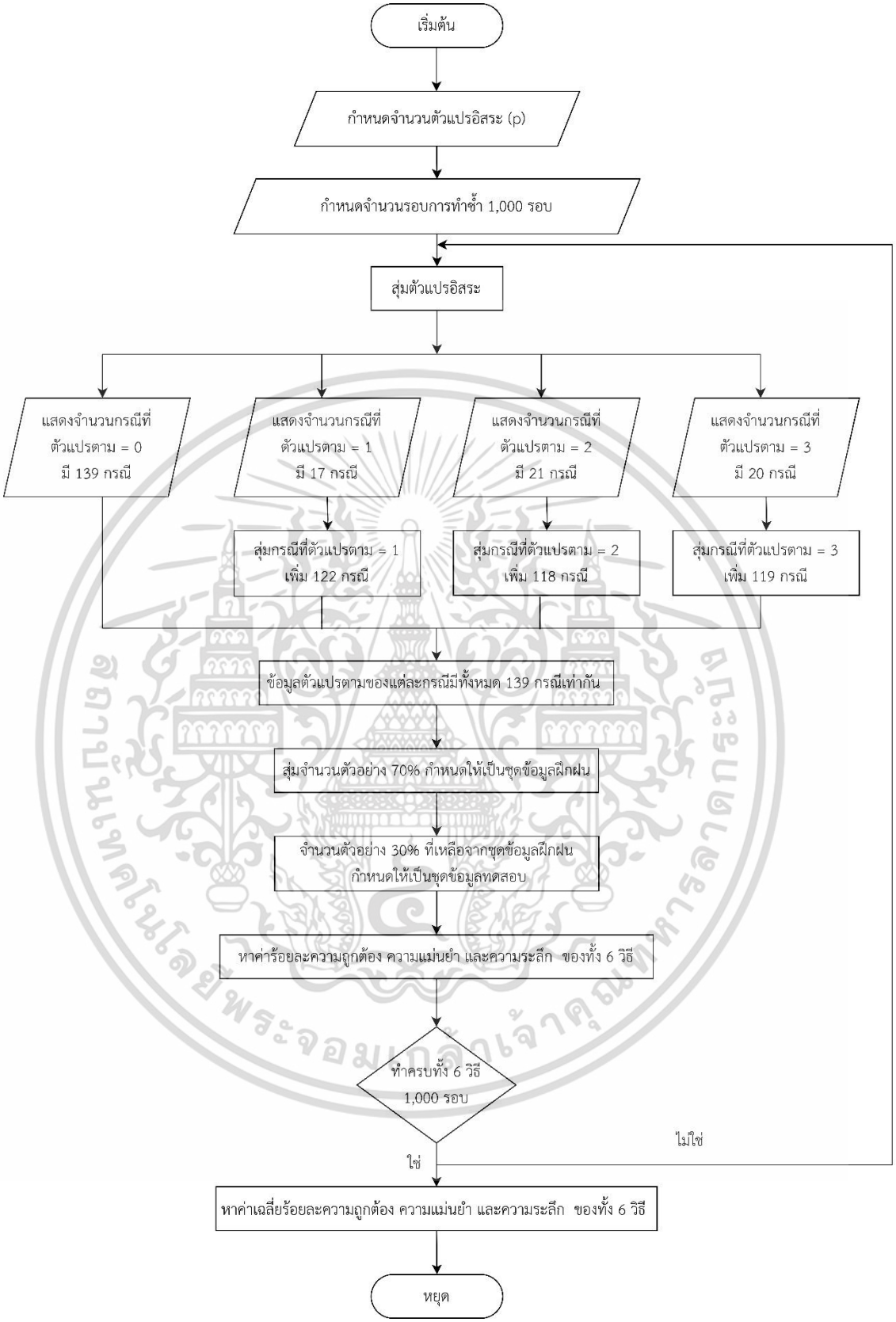
จากตารางที่ 3.5 พบว่า ในวิธีซัพพอร์ตเวกเตอร์แมชชีน ที่จำนวนตัวแปรอิสระเท่ากับ 200 ควรกำหนด Kernel function เป็น Sigmoid และ ที่จำนวนตัวแปรอิสระเท่ากับ 400 600 800 และ 1,000 ควรกำหนด Kernel function เป็น Polynomial เพื่อให้ได้ค่าเฉลี่ยร้อยละความถูกต้องที่สูงที่สุดตามจำนวนตัวแปร

หลังจากที่ได้ค่าเฉพาะของวิธีเพื่อนบ้านใกล้เคียง k อันดับ วิธีโครงข่ายประสาทเทียม และวิธีซัพพอร์ตเวกเตอร์แมชชีนแล้ว จึงนำมาหาค่าเฉลี่ยร้อยละความถูกต้อง ค่าเฉลี่ยร้อยละความแม่นยำ และค่าเฉลี่ยร้อยละความระลึก โดยทำซ้ำ 1,000 รอบ ของวิธีการเรียนรู้ด้วยเครื่องทั้ง 6 วิธี โดยแสดงขั้นตอนการดำเนินงาน ดังรูปที่ 3.4



รูปที่ 3.4 แผนภาพแสดงขั้นตอนการทำงานจากข้อมูลดั้งเดิม

เนื่องจากชุดข้อมูลดั้งเดิมนั้นเป็นข้อมูลที่ไม่สมดุล คือตัวแปรตามในกลุ่มหนึ่ง มีสัดส่วนที่มากกว่ากลุ่มอื่นในจำนวนมาก จึงต้องทำการปรับข้อมูลให้มีความสมดุลเพื่อนำมาใช้ในการวิเคราะห์ ดังรูปที่ 3.5



รูปที่ 3.5 แผนภาพแสดงขั้นตอนการทำงานในการปรับข้อมูลให้สมดุล

บทที่ 4
ผลการวิจัยและการอภิปรายผล

ในงานวิจัยนี้ผู้วิจัยได้ทำการนำข้อมูลรหัสพันธุกรรมของกลุ่มผู้ป่วยที่เกี่ยวกับโรคมะเร็งปอด มาใช้เป็นข้อมูลตัวอย่างเพื่อทำการเปรียบเทียบประสิทธิภาพวิธีการวิเคราะห์การจำแนกกลุ่มแต่ละวิธี ดังนี้ วิธีต้นไม้ตัดสินใจ (Decision Tree) วิธีนาอีฟเบย์ (Naïve Bayes) วิธีเพื่อนบ้านใกล้เคียง k อันดับ (K-Nearest Neighbors หรือ KNN) วิธีซัพพอร์ตเวกเตอร์แมชชีน (Support Vector Machine) วิธีโครงข่ายประสาทเทียม (Artificial Neural Network) และวิธีป่าสุ่ม (Random Forest) โดยใช้โปรแกรม R Studio เวอร์ชัน 4.2.2 ในการวิเคราะห์ ซึ่งจะมีผลลัพธ์ในการวิเคราะห์ดังนี้

โดยกำหนดสัญลักษณ์แทนวิธีการเรียนรู้ด้วยเครื่องดังนี้

DT	หมายถึง	วิธีต้นไม้ตัดสินใจ (Decision Tree)
Naïve	หมายถึง	วิธีนาอีฟเบย์ (Naïve Bayes)
KNN	หมายถึง	วิธีเพื่อนบ้านใกล้เคียง k อันดับ (K-Nearest Neighbors)
SVM	หมายถึง	วิธีซัพพอร์ตเวกเตอร์แมชชีน (Support Vector Machine)
ANN	หมายถึง	วิธีโครงข่ายประสาทเทียม (Artificial Neural Network)
RF	หมายถึง	วิธีป่าสุ่ม (Random Forest)

4.1 ผลการวิจัยค่าเฉลี่ยร้อยละความถูกต้อง ของชุดข้อมูลดั้งเดิม

ผลการวิจัยค่าเฉลี่ยร้อยละความถูกต้อง จากการวิเคราะห์โดยรวมของตัวแปรตามทั้ง 4 กลุ่ม ได้แก่ 1. Small Cell Lung Cancer (SCLC) : Oat cell lung cancer แทนด้วยหมายเลข 0 2. Non-Small Cell lung cancer (NACLC) : Adenocarcinoma แทนด้วยหมายเลข 1 3. Non-Small Cell lung cancer (NSCLC) : Squamous Cell carcinoma แทนด้วยหมายเลข 2 4. Non-Small Cell lung cancer (NSCLC) : Large Cell lung cancer แทนด้วยหมายเลข 3 ของชุดข้อมูลดั้งเดิม โดยมีจำนวนตัวแปรอิสระเป็น 200 400 600 800 และ 1,000 ของวิธีการเรียนรู้ด้วยเครื่องทั้ง 6 วิธี

ตารางที่ 4.1 ตารางแสดงค่าเฉลี่ยร้อยละความถูกต้องของวิธีการเรียนรู้ด้วยเครื่อง 6 วิธี จากชุดข้อมูลดั้งเดิม

วิธีการเรียนรู้ด้วยเครื่อง	จำนวนตัวแปรอิสระ (p)				
	200 (k=6, sigmoid, h=275)	400 (k=7, polynomial, h=275)	600 (k=7, polynomial, h=200)	800 (k=6, polynomial, h=300)	1,000 (k=5, polynomial, h=200)
1. DT	86.6153	87.3814	86.9475	86.9390	86.7271
2. Naïve	95.8831	**96.2881**	**96.4593**	**96.3848**	**96.5051**
3. KNN (k)	94.3661	94.6220	94.5915	93.5000	93.9695
4. SVM (Kernel Function)	**96.6407**	92.5017	92.4949	92.6390	92.5661
5. ANN (h)	88.9763	89.0542	89.0729	89.3017	89.0780
6. RF	94.4881	94.5170	94.6305	94.7576	94.7034

หมายเหตุ ตัวเลขที่เป็นตัวหนา หมายถึง ค่าเฉลี่ยร้อยละความถูกต้องที่มากที่สุดตามจำนวนตัวแปรอิสระ

จากตารางที่ 4.1 แสดงค่าเฉลี่ยร้อยละความถูกต้องในการจำแนกระดับของมะเร็งปอดด้วยวิธีการเรียนรู้ด้วยเครื่อง 6 วิธี จากชุดข้อมูลดั้งเดิม ซึ่งเป็นข้อมูลดั้งเดิมที่ไม่สมดุล พบว่าที่จำนวนตัวแปรอิสระมีค่าเท่ากับ 400 600 800 และ 1,000 ของวิธีการเรียนรู้ด้วยเครื่อง Naïve มีค่าเฉลี่ยร้อยละความถูกต้องที่มากที่สุด และ ที่จำนวนตัวแปรอิสระมีค่าเท่ากับ 200 ของวิธีการเรียนรู้ด้วยเครื่อง SVM มีค่าเฉลี่ยร้อยละความถูกต้องที่มากที่สุดเมื่อเทียบกับวิธีการเรียนรู้ด้วยเครื่องทั้ง 6 วิธี

4.2 ผลการวิจัยค่าเฉลี่ยร้อยละความถูกต้อง ของชุดข้อมูลที่ปรับให้สมดุลแล้ว

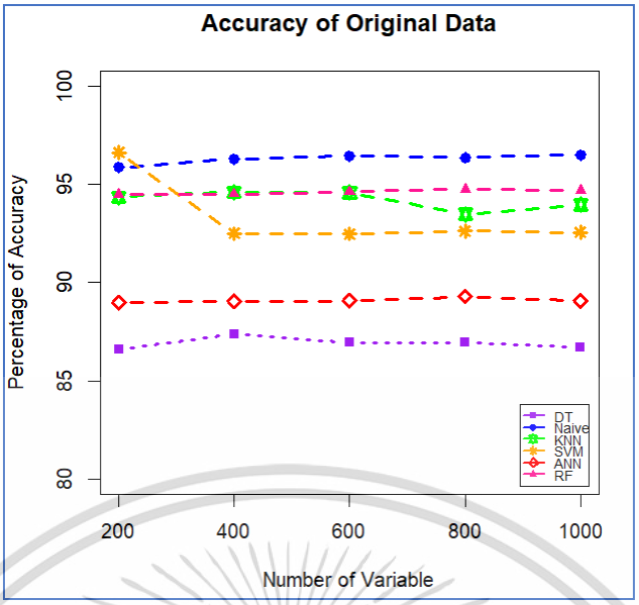
ผลการวิจัยค่าเฉลี่ยร้อยละความถูกต้อง จากการวิเคราะห์โดยรวมของตัวแปรตามทั้ง 4 กลุ่ม ได้แก่ 1. Small Cell Lung Cancer (SCLC) : Oat cell lung cancer แทนด้วยหมายเลข 0 2. Non-Small Cell lung cancer (NACLC) : Adenocarcinoma แทนด้วยหมายเลข 1 3. Non-Small Cell lung cancer (NSCLC) : Squamous Cell carcinoma แทนด้วยหมายเลข 2 4. Non-Small Cell lung cancer (NSCLC) : Large Cell lung cancer แทนด้วยหมายเลข 3 ของชุดข้อมูลที่ปรับให้สมดุลแล้ว โดยมีจำนวนตัวแปรอิสระเป็น 200 400 600 800 และ 1,000 ของวิธีการเรียนรู้ด้วยเครื่องทั้ง 6 วิธี

ตารางที่ 4.2 ตารางแสดงค่าเฉลี่ยร้อยละความถูกต้องของวิธีการเรียนรู้ด้วยเครื่อง 6 วิธี จากชุดข้อมูลที่ปรับให้สมดุลแล้ว

วิธีการเรียนรู้ด้วยเครื่อง	จำนวนตัวแปรอิสระ (p)				
	200 (k=6, sigmoid, h=275)	400 (k=7, polynomial, h=275)	600 (k=7, polynomial, h=200)	800 (k=6, polynomial, h=300)	1,000 (k=5, polynomial, h=200)
1. DT	96.2371	96.5683	96.6641	96.5773	96.5072
2. Naïve	98.3479	98.5599	98.6701	98.6425	98.6778
3. KNN (k)	94.1395	95.2114	95.6665	96.2635	96.7755
4. SVM (Kernel Function)	99.0784	98.5246	98.5976	98.6737	98.6737
5. ANN (h)	98.1108	98.1114	98.2976	98.2222	98.2419
6. RF	**99.5982**	**99.6042**	**99.6443**	**99.6084**	**99.6156**

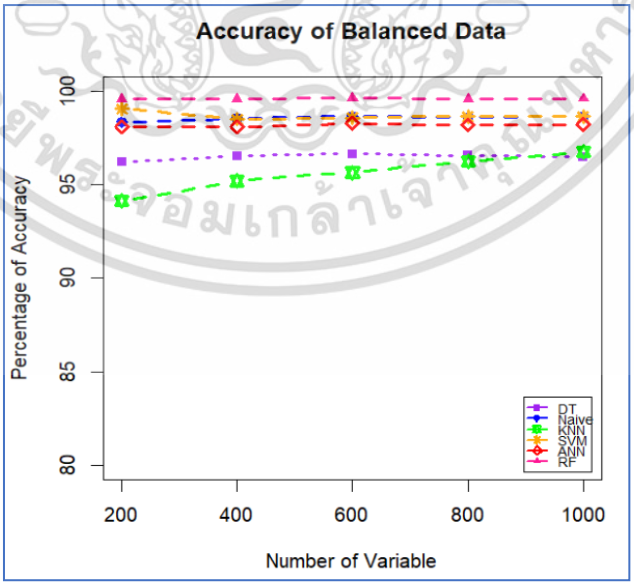
หมายเหตุ ตัวเลขที่เป็นตัวหนา หมายถึง ค่าเฉลี่ยร้อยละความถูกต้องที่มากที่สุดตามจำนวนตัวแปรอิสระ

จากตารางที่ 4.2 แสดงค่าเฉลี่ยร้อยละความถูกต้องในการจำแนกระดับของมะเร็งปอดด้วยวิธีการเรียนรู้ด้วยเครื่อง 6 วิธี จากชุดข้อมูลที่ปรับให้สมดุลแล้ว พบว่าที่จำนวนตัวแปรอิสระมีค่าเท่ากับ 200 400 600 800 และ 1,000 ของวิธีการเรียนรู้ด้วยเครื่อง RF มีค่าเฉลี่ยร้อยละความถูกต้องที่มากที่สุดเมื่อเทียบกับวิธีการเรียนรู้ด้วยเครื่องทั้ง 6 วิธี และนำค่าเฉลี่ยร้อยละความถูกต้องจากตารางที่ 4.1 และ 4.2 มาแสดงผลในรูปแบบของกราฟ ดังรูปที่ 4.1 และ 4.2



รูปที่ 4.1 กราฟแสดงค่าเฉลี่ยร้อยละความถูกต้องของชุดข้อมูลดั้งเดิม ของวิธีการเรียนรู้ด้วยเครื่อง 6 วิธี

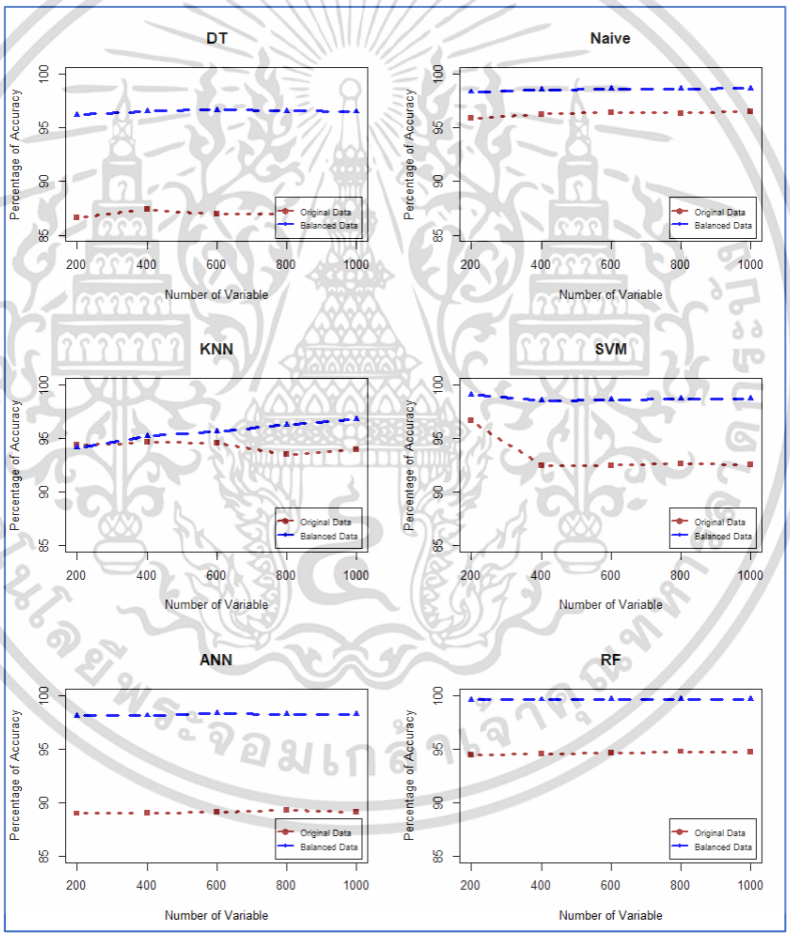
จากรูปที่ 4.1 พบว่าชุดข้อมูลดั้งเดิม หากมีจำนวนตัวแปรอิสระเพิ่มมากขึ้น ค่าเฉลี่ยร้อยละความถูกต้องของวิธีการเรียนรู้ด้วยเครื่อง Naïve ANN RF ก็มีค่าเฉลี่ยร้อยละความถูกต้องเพิ่มขึ้นด้วย เมื่อพิจารณาจำนวนตัวแปรอิสระมีค่าเท่ากับ 200 ของวิธีการเรียนรู้ด้วยเครื่องทั้ง 6 วิธี จะเห็นว่าวิธี SVM มีค่าเฉลี่ยร้อยละความถูกต้องสูงที่สุด และ เมื่อพิจารณาจำนวนตัวแปรอิสระมีค่าเท่ากับ 400 600 800 และ 1,000 ของวิธีการเรียนรู้ด้วยเครื่องทั้ง 6 วิธี จะเห็นว่าวิธี Naïve มีค่าเฉลี่ยร้อยละความถูกต้องสูงที่สุด



รูปที่ 4.2 กราฟแสดงค่าเฉลี่ยร้อยละความถูกต้องของชุดข้อมูลที่ปรับให้สมดุลแล้วของวิธีการเรียนรู้ด้วยเครื่อง 6 วิธี

จากรูปที่ 4.2 พบว่าชุดข้อมูลที่ปรับให้สมดุลแล้วมีค่าเฉลี่ยร้อยละความถูกต้องมากกว่าชุดข้อมูลดั้งเดิม หากมีจำนวนตัวแปรอิสระเพิ่มมากขึ้น ค่าเฉลี่ยร้อยละความถูกต้องของวิธีการเรียนรู้ด้วยเครื่อง Naïve ANN RF ก็มีค่าเฉลี่ยร้อยละความถูกต้องเพิ่มขึ้นเล็กน้อย แต่วิธี KNN จะสังเกตได้ว่าหากจำนวนตัวแปรอิสระเพิ่มมากขึ้น ค่าเฉลี่ยร้อยละความถูกต้องก็เพิ่มขึ้นไปด้วยอย่างเห็นได้ชัด เมื่อพิจารณาจำนวนตัวแปรอิสระมีค่าเท่ากับ 200 400 600 800 และ 1,000 ของวิธีการเรียนรู้ด้วยเครื่องทั้ง 6 วิธี จะเห็นว่าวิธี RF มีค่าค่าเฉลี่ยร้อยละความถูกต้องสูงที่สุด

จากรูปที่ 4.1 และ 4.2 เป็นเพียงการแสดงค่าเฉลี่ยร้อยละความถูกต้องของแต่ละชุดข้อมูล อาจไม่สามารถเปรียบเทียบความแตกต่างได้อย่างชัดเจน ระหว่างค่าเฉลี่ยร้อยละความถูกต้องของชุดข้อมูลดั้งเดิม และ ชุดข้อมูลที่ปรับให้สมดุลแล้ว จึงนำมาแสดงผลดังรูปที่ 4.3



รูปที่ 4.3 กราฟเปรียบเทียบความแตกต่างค่าเฉลี่ยร้อยละความถูกต้อง ระหว่างข้อมูลชุดดั้งเดิม และ ชุดข้อมูลที่ปรับให้สมดุลแล้ว ของวิธีการเรียนรู้ด้วยเครื่อง 6 วิธี

จากรูปที่ 4.3 พบว่า ค่าเฉลี่ยร้อยละความถูกต้องของชุดข้อมูลที่ปรับให้สมดุลแล้ว มีค่ามากกว่าชุดข้อมูลดั้งเดิมทั้ง 6 วิธี โดยสังเกตได้จากกราฟเส้นสีน้ำเงิน(ชุดข้อมูลที่ปรับให้สมดุลแล้ว) จะอยู่สูงกว่ากราฟเส้นสีแดง (ชุดข้อมูลดั้งเดิม) ในทั้งหมด 6 วิธี

4.3 ผลการวิจัยค่าเฉลี่ยร้อยละความแม่นยำ ของชุดข้อมูลดั้งเดิม

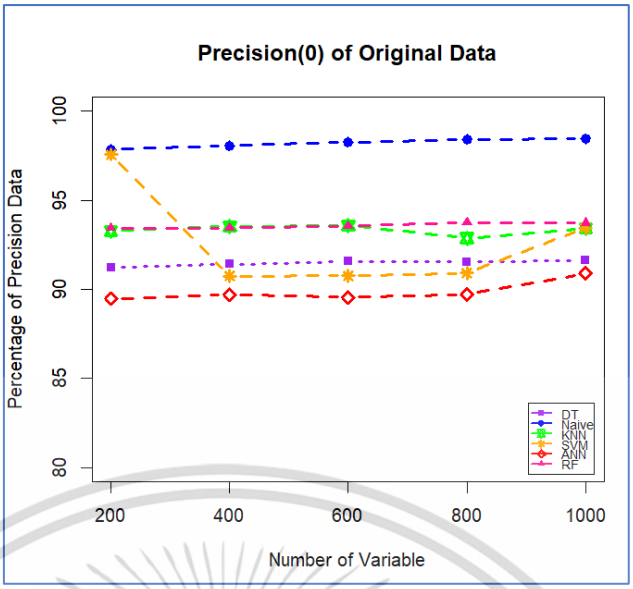
ผลการวิจัยค่าเฉลี่ยร้อยละความแม่นยำ แยกตามกลุ่มของตัวแปรตาม ทั้ง 4 กลุ่ม ได้แก่ 1. Small Cell Lung Cancer (SCLC) : Oat cell lung cancer แทนด้วยหมายเลข 0 2. Non-Small Cell lung cancer (NACLC) : Adenocarcinoma แทนด้วยหมายเลข 1 3. Non-Small Cell lung cancer (NSCLC) : Squamous Cell carcinoma แทนด้วยหมายเลข 2 4. Non-Small Cell lung cancer (NSCLC) : Large Cell lung cancer แทนด้วยหมายเลข 3 ของชุดข้อมูลดั้งเดิม โดยมีจำนวนตัวแปรอิสระเป็น 200 400 600 800 และ 1,000 ของวิธีการเรียนรู้ด้วยเครื่องทั้ง 6 วิธี

ตารางที่ 4.3 ตารางแสดงค่าเฉลี่ยร้อยละความแม่นยำของกลุ่มตัวแปรตาม ที่แทนด้วยหมายเลข 0 (Small Cell Lung Cancer (SCLC) : Oat cell lung cancer) ของวิธีการเรียนรู้ด้วยเครื่อง 6 วิธี จากชุดข้อมูลดั้งเดิม

วิธีการเรียนรู้ด้วยเครื่อง	จำนวนตัวแปรอิสระ (p)				
	200 (k=6, sigmoid, h=275)	400 (k=7, polynomial, h=275)	600 (k=7, polynomial, h=200)	800 (k=6, polynomial, h=300)	1,000 (k=5, polynomial, h=200)
1. DT	91.2062	91.4049	91.5592	91.5306	91.6221
2. Naïve	**97.8566**	**98.0657**	**98.2546**	**98.4329**	**98.4383**
3. KNN (k)	93.2766	93.5303	93.5866	92.8798	93.4263
4. SVM (Kernel Function)	97.5763	90.7383	90.7580	90.9208	90.8948
5. ANN (h)	89.4820	89.7130	89.5556	89.7095	89.9568
6. RF	93.4439	93.4502	93.5389	93.7562	93.7307

หมายเหตุ ตัวเลขที่เป็นตัวหนา หมายถึง ค่าเฉลี่ยร้อยละความแม่นยำที่มากที่สุดตามจำนวนตัวแปรอิสระ

จากตารางที่ 4.3 แสดงค่าเฉลี่ยร้อยละความแม่นยำของกลุ่มตัวแปรตาม ที่แทนด้วยหมายเลข 0 ในการจำแนกระดับของมะเร็งปอดด้วยวิธีการเรียนรู้ด้วยเครื่อง 6 วิธี จากชุดข้อมูลดั้งเดิม พบว่าที่จำนวนตัวแปรอิสระมีค่าเท่ากับ 200 400 600 800 และ 1,000 ของวิธีการเรียนรู้ด้วยเครื่อง Naïve มีค่าเฉลี่ยร้อยละความแม่นยำที่มากที่สุด และนำค่าเฉลี่ยร้อยละความแม่นยำจากตารางที่ 4.3 มาแสดงผลในรูปแบบของกราฟ ดังรูปที่ 4.4



รูปที่ 4.4 กราฟแสดงค่าเฉลี่ยร้อยละความแม่นยำกลุ่มตัวแปรตาม ที่แทนด้วยหมายเลข 0 ของชุดข้อมูลดั้งเดิม ของวิธีการเรียนรู้ด้วยเครื่อง 6 วิธี

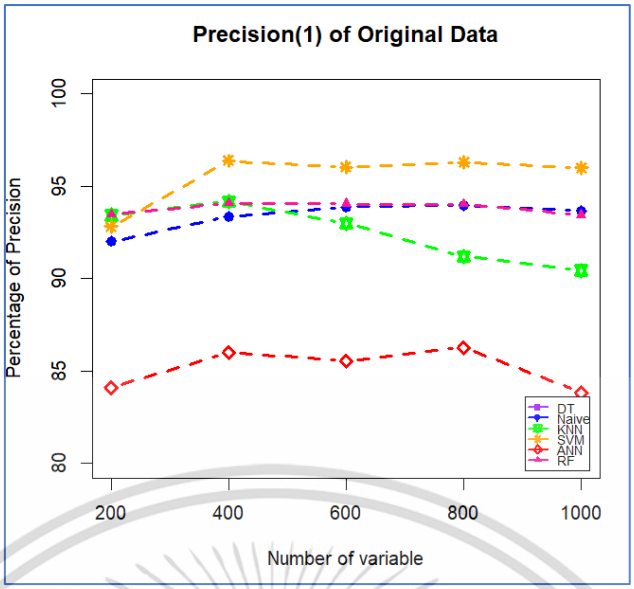
จากรูปที่ 4.4 พบว่าชุดข้อมูลดั้งเดิมของค่าเฉลี่ยร้อยละความแม่นยำกลุ่มตัวแปรตาม ที่แทนด้วยหมายเลข 0 หากมีจำนวนตัวแปรอิสระเพิ่มมากขึ้น ค่าเฉลี่ยร้อยละความแม่นยำของวิธีการเรียนรู้ด้วยเครื่อง DT Naïve ANN RF ก็มีค่าเฉลี่ยร้อยละความแม่นยำเพิ่มขึ้นด้วย เมื่อพิจารณาจำนวนตัวแปรอิสระมีค่าเท่ากับ 200 400 600 800 และ 1,000 ของวิธีการเรียนรู้ด้วยเครื่องทั้ง 6 วิธี จะเห็นว่า วิธี Naïve มีค่าเฉลี่ยร้อยละความแม่นยำสูงที่สุด

ตารางที่ 4.4 ตารางแสดงค่าเฉลี่ยร้อยละความแม่นยำของกลุ่มตัวแปรตาม ที่แทนด้วยหมายเลข 1 (Non-Small Cell lung cancer (NACLC) : Adenocarcinoma) ของวิธีการเรียนรู้ด้วยเครื่อง 6 วิธี จากชุดข้อมูลดั้งเดิม

วิธีการเรียนรู้ด้วยเครื่อง	จำนวนตัวแปรอิสระ (p)				
	200 (k=6, sigmoid, h=275)	400 (k=7, polynomial, h=275)	600 (k=7, polynomial, h=200)	800 (k=6, polynomial, h=300)	1,000 (k=5, polynomial, h=200)
1. DT	75.5899	76.5009	72.1625	69.3374	68.0688
2. Naïve	92.0261	93.3646	93.8911	93.9647	93.6858
3. KNN (k)	93.4427	94.1846	93.0025	91.2165	90.4587
4. SVM (Kernel Function)	92.8249	**96.3860**	**96.0414**	**96.3044**	**96.0036**
5. ANN (h)	84.1197	86.0035	85.5543	86.2750	83.8322
6. RF	**93.4826**	94.0788	94.0527	94.0314	93.4561

หมายเหตุ ตัวเลขที่เป็นตัวหนา หมายถึง ค่าเฉลี่ยร้อยละความแม่นยำที่มากที่สุดตามจำนวนตัวแปรอิสระ

จากตารางที่ 4.4 แสดงค่าเฉลี่ยร้อยละความแม่นยำของกลุ่มตัวแปรตาม ที่แทนด้วยหมายเลข 1 ในการจำแนกระดับของมะเร็งปอดด้วยวิธีการเรียนรู้ด้วยเครื่อง 6 วิธี จากชุดข้อมูลดั้งเดิม พบว่าที่จำนวนตัวแปรอิสระมีค่าเท่ากับ 200 ของวิธีการเรียนรู้ด้วยเครื่อง RF มีค่าเฉลี่ยร้อยละความแม่นยำที่มากที่สุด เมื่อพิจารณาจำนวนตัวแปรอิสระมีค่าเท่ากับ 400 600 800 และ 1,000 ของวิธีการเรียนรู้ด้วยเครื่องทั้ง 6 วิธี จะเห็นว่าวิธี SVM มีค่าเฉลี่ยร้อยละความแม่นยำสูงที่สุด และนำค่าเฉลี่ยร้อยละความแม่นยำจากตารางที่ 4.4 มาแสดงผลในรูปแบบของกราฟ ดังรูปที่ 4.5



รูปที่ 4.5 กราฟแสดงค่าเฉลี่ยร้อยละความแม่นยำกลุ่มตัวแปรตาม ที่แทนด้วยหมายเลข 1 ของชุดข้อมูลดั้งเดิม ของวิธีการเรียนรู้ด้วยเครื่อง 6 วิธี

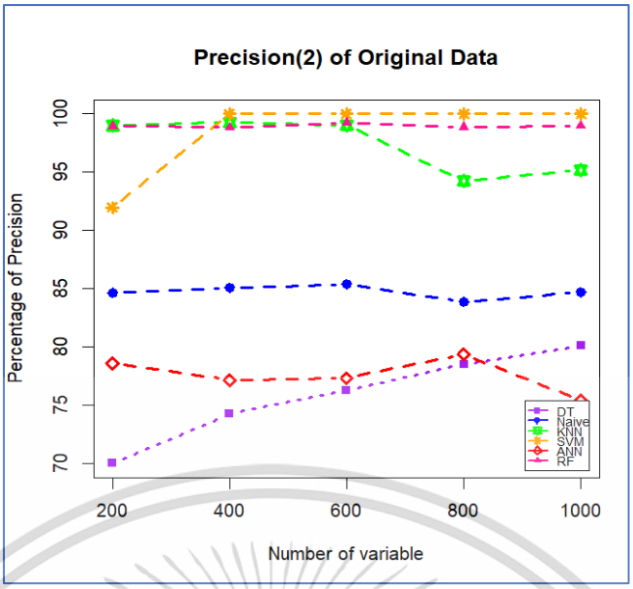
จากรูปที่ 4.5 พบว่าชุดข้อมูลดั้งเดิมของค่าเฉลี่ยร้อยละความแม่นยำกลุ่มตัวแปรตาม ที่แทนด้วยหมายเลข 1 หากมีจำนวนตัวแปรอิสระเพิ่มมากขึ้น ค่าเฉลี่ยร้อยละความแม่นยำของวิธีการเรียนรู้ด้วยเครื่อง Naïve ก็มีค่าเฉลี่ยร้อยละความแม่นยำเพิ่มขึ้นด้วย เมื่อพิจารณาที่จำนวนตัวแปรอิสระมีค่าเท่ากับ 200 ของวิธีการเรียนรู้ด้วยเครื่อง RF มีค่าเฉลี่ยร้อยละความแม่นยำที่มากที่สุด เมื่อพิจารณาจำนวนตัวแปรอิสระมีค่าเท่ากับ 400 600 800 และ 1,000 ของวิธีการเรียนรู้ด้วยเครื่องทั้ง 6 วิธี จะเห็นว่า วิธี SVM มีค่าเฉลี่ยร้อยละความแม่นยำสูงที่สุด

ตารางที่ 4.5 ตารางแสดงค่าเฉลี่ยร้อยละความแม่นยำของกลุ่มตัวแปรตาม ที่แทนด้วยหมายเลข 2 (Non-Small Cell lung cancer (NSCLC) : Squamous Cell carcinoma) ของวิธีการเรียนรู้ด้วยเครื่อง 6 วิธี จากชุดข้อมูลดั้งเดิม

วิธีการเรียนรู้ด้วยเครื่อง	จำนวนตัวแปรอิสระ (p)				
	200 (k=6, sigmoid, h=275)	400 (k=7, polynomial, h=275)	600 (k=7, polynomial, h=200)	800 (k=6, polynomial, h=300)	1,000 (k=5, polynomial, h=200)
1. DT	70.0483	74.2722	76.3169	78.5165	80.1495
2. Naïve	84.6909	85.0829	85.3978	83.8886	84.7228
3. KNN (k)	**99.0165**	99.2671	99.0565	94.2784	95.2032
4. SVM (Kernel Function)	91.9675	**100**	**100**	**100**	**100**
5. ANN (h)	78.6193	77.1698	77.3555	79.4051	75.4428
6. RF	98.9507	98.8368	99.2519	98.8649	98.9994

หมายเหตุ ตัวเลขที่เป็นตัวหนา หมายถึง ค่าเฉลี่ยร้อยละความแม่นยำที่มากที่สุดตามจำนวนตัวแปรอิสระ

จากตารางที่ 4.5 แสดงค่าเฉลี่ยร้อยละความแม่นยำของกลุ่มตัวแปรตาม ที่แทนด้วยหมายเลข 2 ในการจำแนกระดับของมะเร็งปอดด้วยวิธีการเรียนรู้ด้วยเครื่อง 6 วิธี จากชุดข้อมูลดั้งเดิม พบว่าที่จำนวนตัวแปรอิสระมีค่าเท่ากับ 200 ของวิธีการเรียนรู้ด้วยเครื่อง KNN มีค่าเฉลี่ยร้อยละความแม่นยำที่มากที่สุด เมื่อพิจารณาจำนวนตัวแปรอิสระมีค่าเท่ากับ 400 600 800 และ 1,000 ของวิธีการเรียนรู้ด้วยเครื่องทั้ง 6 วิธี จะเห็นว่าวิธี SVM มีค่าเฉลี่ยร้อยละความแม่นยำสูงที่สุด และนำค่าเฉลี่ยร้อยละความแม่นยำจากตารางที่ 4.5 มาแสดงผลในรูปแบบของกราฟ ดังรูปที่ 4.6



รูปที่ 4.6 กราฟแสดงค่าเฉลี่ยร้อยละความแม่นยำกลุ่มตัวแปรตาม ที่แทนด้วยหมายเลข 2 ของชุดข้อมูลดั้งเดิม ของวิธีการเรียนรู้ด้วยเครื่อง 6 วิธี

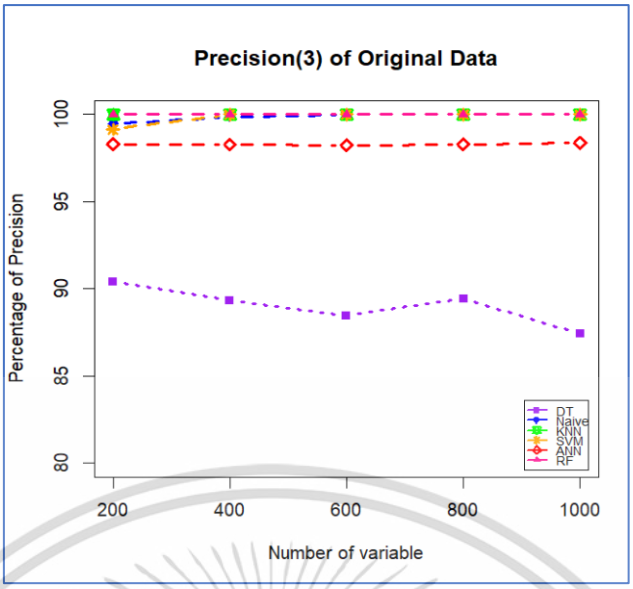
จากรูปที่ 4.6 พบว่าชุดข้อมูลดั้งเดิมของค่าเฉลี่ยร้อยละความแม่นยำกลุ่มตัวแปรตาม ที่แทนด้วยหมายเลข 2 หากมีจำนวนตัวแปรอิสระเพิ่มมากขึ้น ค่าเฉลี่ยร้อยละความแม่นยำของวิธีการเรียนรู้ด้วยเครื่อง DT และ RF ก็มีค่าเฉลี่ยร้อยละความแม่นยำเพิ่มขึ้นด้วย เมื่อพิจารณาที่จำนวนตัวแปรอิสระมีค่าเท่ากับ 200 ของวิธีการเรียนรู้ด้วยเครื่อง KNN มีค่าเฉลี่ยร้อยละความแม่นยำที่มากที่สุด เมื่อพิจารณาจำนวนตัวแปรอิสระมีค่าเท่ากับ 400 600 800 และ 1,000 ของวิธีการเรียนรู้ด้วยเครื่องทั้ง 6 วิธี จะเห็นว่า วิธี SVM มีค่าเฉลี่ยร้อยละความแม่นยำสูงที่สุด

ตารางที่ 4.6 ตารางแสดงค่าเฉลี่ยร้อยละความแม่นยำของกลุ่มตัวแปรตาม ที่แทนด้วยหมายเลข 3 (Non-Small Cell lung cancer (NSCLC) : Large Cell lung cancer) ของวิธีการเรียนรู้ด้วยเครื่อง 6 วิธี จากชุดข้อมูลดั้งเดิม

วิธีการเรียนรู้ด้วยเครื่อง	จำนวนตัวแปรอิสระ (p)				
	200 (k=6, sigmoid, h=275)	400 (k=7, polynomial, h=275)	600 (k=7, polynomial, h=200)	800 (k=6, polynomial, h=300)	1,000 (k=5, polynomial, h=200)
1. DT	90.4345	89.3565	88.4767	89.4363	87.4557
2. Naïve	99.4500	99.8813	**100**	**100**	**100**
3. KNN (k)	**100**	**100**	**100**	**100**	**100**
4. SVM (Kernel Function)	99.1505	**100**	**100**	**100**	**100**
5. ANN (h)	98.2938	98.2810	98.2416	98.2970	98.3717
6. RF	**100**	**100**	**100**	**100**	**100**

หมายเหตุ ตัวเลขที่เป็นตัวหนา หมายถึง ค่าเฉลี่ยร้อยละความแม่นยำที่มากที่สุดตามจำนวนตัวแปรอิสระ

จากตารางที่ 4.6 แสดงค่าเฉลี่ยร้อยละความแม่นยำของกลุ่มตัวแปรตาม ที่แทนด้วยหมายเลข 3 ในการจำแนกระดับของมะเร็งปอดด้วยวิธีการเรียนรู้ด้วยเครื่อง 6 วิธี จากชุดข้อมูลดั้งเดิม พบว่าที่จำนวนตัวแปรอิสระมีค่าเท่ากับ 200 ของวิธีการเรียนรู้ด้วยเครื่อง KNN และ RF มีค่าเฉลี่ยร้อยละความแม่นยำที่มากที่สุด เมื่อพิจารณาจำนวนตัวแปรอิสระมีค่าเท่ากับ 400 ของวิธีการเรียนรู้ด้วยเครื่อง KNN SVM และ RF มีค่าเฉลี่ยร้อยละความแม่นยำที่มากที่สุด เมื่อพิจารณาจำนวนตัวแปรอิสระมีค่าเท่ากับ 600 800 และ 1,000 ของวิธีการเรียนรู้ด้วยเครื่อง Naïve KNN SVM และ RF มีค่าเฉลี่ยร้อยละความแม่นยำสูงที่สุด และนำค่าเฉลี่ยร้อยละความแม่นยำจากตารางที่ 4.6 มาแสดงผลในรูปแบบของกราฟ ดังรูปที่ 4.7



รูปที่ 4.7 กราฟแสดงค่าเฉลี่ยร้อยละความแม่นยำกลุ่มตัวแปรตาม ที่แทนด้วยหมายเลข 3 ของชุดข้อมูลดั้งเดิม ของวิธีการเรียนรู้ด้วยเครื่อง 6 วิธี

จากรูปที่ 4.7 พบว่าชุดข้อมูลดั้งเดิมของค่าเฉลี่ยร้อยละความแม่นยำกลุ่มตัวแปรตาม ที่แทนด้วยหมายเลข 3 หากมีจำนวนตัวแปรอิสระเพิ่มมากขึ้น ค่าเฉลี่ยร้อยละความแม่นยำของวิธีการเรียนรู้ด้วยเครื่อง ANN ก็มีค่าเฉลี่ยร้อยละความแม่นยำเพิ่มขึ้นด้วย เมื่อพิจารณาที่จำนวนตัวแปรอิสระมีค่าเท่ากับ 200 ของวิธีการเรียนรู้ด้วยเครื่อง KNN และ RF มีค่าเฉลี่ยร้อยละความแม่นยำที่มากที่สุด เมื่อพิจารณาจำนวนตัวแปรอิสระมีค่าเท่ากับ 400 ของวิธีการเรียนรู้ด้วยเครื่อง KNN SVM และ RF มีค่าเฉลี่ยร้อยละความแม่นยำที่มากที่สุด เมื่อพิจารณาจำนวนตัวแปรอิสระมีค่าเท่ากับ 600 800 และ 1,000 ของวิธีการเรียนรู้ด้วยเครื่อง Naïve KNN SVM และ RF มีค่าเฉลี่ยร้อยละความแม่นยำสูงที่สุด

4.4 ผลการวิจัยค่าเฉลี่ยร้อยละความแม่นยำ ของชุดข้อมูลที่ปรับให้สมดุลแล้ว

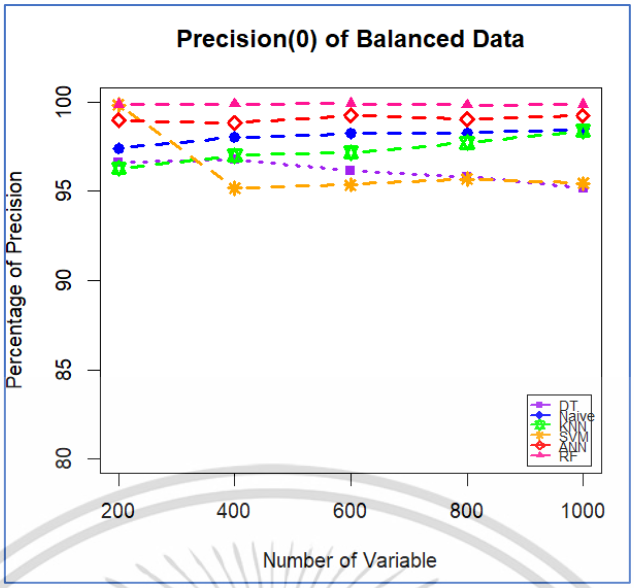
ผลการวิจัยค่าเฉลี่ยร้อยละความแม่นยำ แยกตามกลุ่มของตัวแปรตาม ทั้ง 4 กลุ่ม ได้แก่ 1. Small Cell Lung Cancer (SCLC) : Oat cell lung cancer แทนด้วยหมายเลข 0 2. Non-Small Cell lung cancer (NACLC) : Adenocarcinoma แทนด้วยหมายเลข 1 3. Non-Small Cell lung cancer (NSCLC) : Squamous Cell carcinoma แทนด้วยหมายเลข 2 4. Non-Small Cell lung cancer (NSCLC) : Large Cell lung cancer แทนด้วยหมายเลข 3 ของชุดข้อมูลที่ปรับให้สมดุลแล้ว โดยมีจำนวนตัวแปรอิสระเป็น 200 400 600 800 และ 1,000 ของวิธีการเรียนรู้ด้วยเครื่องทั้ง 6 วิธี

ตารางที่ 4.7 ตารางแสดงค่าเฉลี่ยร้อยละความแม่นยำของกลุ่มตัวแปรตาม ที่แทนด้วยหมายเลข 0 (Small Cell Lung Cancer (SCLC) : Oat cell lung cancer) ของวิธีการเรียนรู้ด้วยเครื่อง 6 วิธี จากชุดข้อมูลที่ปรับให้สมดุลแล้ว

วิธีการเรียนรู้ด้วยเครื่อง	จำนวนตัวแปรอิสระ (p)				
	200 (k=6, sigmoid, h=275)	400 (k=7, polynomial, h=275)	600 (k=7, polynomial, h=200)	800 (k=6, polynomial, h=300)	1,000 (k=5, polynomial, h=200)
1. DT	96.6050	96.7817	96.1580	95.8108	95.1796
2. Naïve	97.4144	98.0245	98.2319	98.2703	98.4580
3. KNN (k)	96.2780	97.0112	97.1806	97.7164	98.3849
4. SVM (Kernel Function)	99.8341	95.1866	95.3664	95.6764	95.4292
5. ANN (h)	98.9831	98.8247	99.2486	99.0333	99.2381
6. RF	**99.8565**	**99.8720**	**99.9036**	**99.8369**	**99.8685**

หมายเหตุ ตัวเลขที่เป็นตัวหนา หมายถึง ค่าเฉลี่ยร้อยละความแม่นยำที่มากที่สุดตามจำนวนตัวแปรอิสระ

จากตารางที่ 4.7 แสดงค่าเฉลี่ยร้อยละความแม่นยำของกลุ่มตัวแปรตาม ที่แทนด้วยหมายเลข 0 ในการจำแนกระดับของมะเร็งปอดด้วยวิธีการเรียนรู้ด้วยเครื่อง 6 วิธี จากชุดข้อมูลที่ปรับให้สมดุลแล้ว พบว่า ที่จำนวนตัวแปรอิสระมีค่าเท่ากับ 200 400 600 800 และ 1,000 ของวิธีการเรียนรู้ด้วยเครื่องทั้ง 6 วิธี จะเห็นว่าวิธี RF มีค่าเฉลี่ยร้อยละความแม่นยำสูงที่สุด และนำค่าเฉลี่ยร้อยละความแม่นยำจากตารางที่ 4.7 มาแสดงผลในรูปแบบของกราฟ ดังรูปที่ 4.8



รูปที่ 4.8 กราฟแสดงค่าเฉลี่ยร้อยละความแม่นยำกลุ่มตัวแปรตาม ที่แทนด้วยหมายเลข 0 ของชุดข้อมูลที่ปรับให้สมดุลแล้ว ของวิธีการเรียนรู้ด้วยเครื่อง 6 วิธี

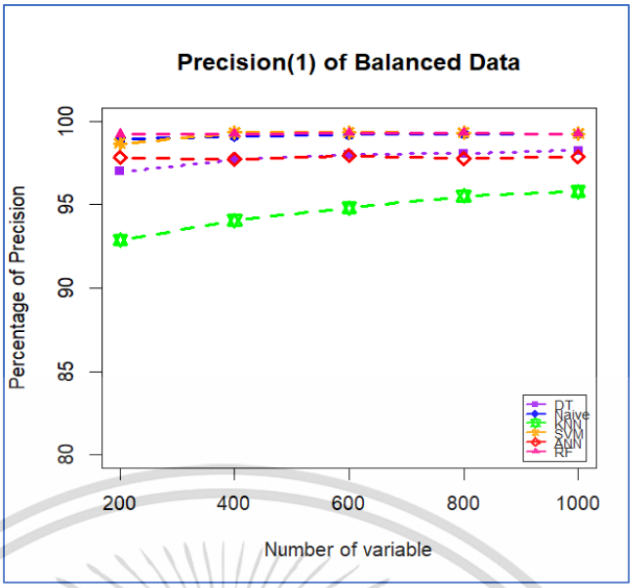
จากรูปที่ 4.8 พบว่าชุดข้อมูลที่ปรับให้สมดุลแล้ว ของค่าเฉลี่ยร้อยละความแม่นยำกลุ่มตัวแปรตาม ที่แทนด้วยหมายเลข 0 หากมีจำนวนตัวแปรอิสระเพิ่มมากขึ้น ค่าเฉลี่ยร้อยละความแม่นยำของวิธีการเรียนรู้ด้วยเครื่อง Naïve และ KNN ก็มีค่าเฉลี่ยร้อยละความแม่นยำเพิ่มขึ้นด้วย เมื่อพิจารณาที่จำนวนตัวแปรอิสระมีค่าเท่ากับ 200 400 600 800 และ 1,000 ของวิธีการเรียนรู้ด้วยเครื่องทั้ง 6 วิธี จะเห็นว่า วิธี RF มีค่าเฉลี่ยร้อยละความแม่นยำสูงที่สุด

ตารางที่ 4.8 ตารางแสดงค่าเฉลี่ยร้อยละความแม่นยำของกลุ่มตัวแปรตาม ที่แทนด้วยหมายเลข 1 (Non-Small Cell lung cancer (NACLC) : Adenocarcinoma) ของวิธีการเรียนรู้ด้วยเครื่อง 6 วิธี จากชุดข้อมูลที่ปรับให้สมดุลแล้ว

วิธีการเรียนรู้ด้วยเครื่อง	จำนวนตัวแปรอิสระ (p)				
	200 (k=6, sigmoid, h=275)	400 (k=7, polynomial, h=275)	600 (k=7, polynomial, h=200)	800 (k=6, polynomial, h=300)	1,000 (k=5, polynomial, h=200)
1. DT	96.9923	97.7007	97.9952	98.0637	98.2507
2. Naïve	98.9373	99.1071	99.2066	99.2514	99.2336
3. KNN (k)	92.8773	94.0688	94.8064	95.5235	95.7901
4. SVM (Kernel Function)	98.6620	**99.3357**	**99.3325**	**99.3133**	99.2570
5. ANN (h)	97.8304	97.7237	97.9417	97.7789	97.8616
6. RF	**99.2132**	99.2454	99.3035	99.3027	**99.2586**

หมายเหตุ ตัวเลขที่เป็นตัวหนา หมายถึง ค่าเฉลี่ยร้อยละความแม่นยำที่มากที่สุดตามจำนวนตัวแปรอิสระ

จากตารางที่ 4.8 แสดงค่าเฉลี่ยร้อยละความแม่นยำของกลุ่มตัวแปรตาม ที่แทนด้วยหมายเลข 1 ในการจำแนกระดับของมะเร็งปอดด้วยวิธีการเรียนรู้ด้วยเครื่อง 6 วิธี จากชุดข้อมูลที่ปรับให้สมดุลแล้ว พบว่า ที่จำนวนตัวแปรอิสระมีค่าเท่ากับ 200 และ 1,000 ของวิธีการเรียนรู้ด้วยเครื่อง RF มีค่าเฉลี่ยร้อยละความแม่นยำที่มากที่สุด เมื่อพิจารณาจำนวนตัวแปรอิสระมีค่าเท่ากับ 400 600 และ 800 ของวิธีการเรียนรู้ด้วยเครื่องทั้ง 6 วิธี จะเห็นว่าวิธี SVM มีค่าเฉลี่ยร้อยละความแม่นยำสูงที่สุด และนำค่าเฉลี่ยร้อยละความแม่นยำจากตารางที่ 4.8 มาแสดงผลในรูปแบบของกราฟ ดังรูปที่ 4.9



รูปที่ 4.9 กราฟแสดงค่าเฉลี่ยร้อยละความแม่นยำกลุ่มตัวแปรตาม ที่แทนด้วยหมายเลข 1 ของชุดข้อมูลที่ปรับให้สมดุลแล้ว ของวิธีการเรียนรู้ด้วยเครื่อง 6 วิธี

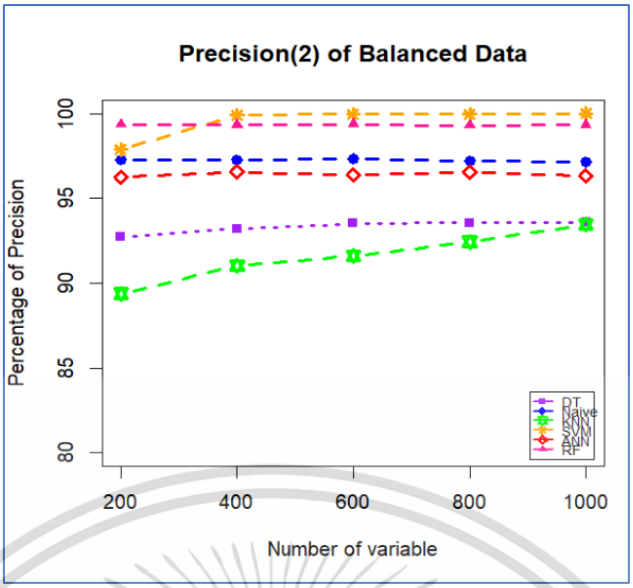
จากรูปที่ 4.9 พบว่าชุดข้อมูลที่ปรับให้สมดุลแล้ว ของค่าเฉลี่ยร้อยละความแม่นยำกลุ่มตัวแปรตาม ที่แทนด้วยหมายเลข 1 หากมีจำนวนตัวแปรอิสระเพิ่มมากขึ้น ค่าเฉลี่ยร้อยละความแม่นยำของวิธีการเรียนรู้ด้วยเครื่อง DT และ KNN ก็มีค่าเฉลี่ยร้อยละความแม่นยำเพิ่มขึ้นด้วย เมื่อพิจารณาที่จำนวนตัวแปรอิสระมีค่าเท่ากับ 200 และ 1,000 ของวิธีการเรียนรู้ด้วยเครื่อง RF มีค่าเฉลี่ยร้อยละความแม่นยำที่มากที่สุด เมื่อพิจารณาจำนวนตัวแปรอิสระมีค่าเท่ากับ 400 600 และ 800 ของวิธีการเรียนรู้ด้วยเครื่องทั้ง 6 วิธี จะเห็นว่าวิธี SVM มีค่าเฉลี่ยร้อยละความแม่นยำสูงที่สุด

ตารางที่ 4.9 ตารางแสดงค่าเฉลี่ยร้อยละความแม่นยำของกลุ่มตัวแปรตามที่แทนด้วยหมายเลข 2 (Non-Small Cell lung cancer (NSCLC) : Squamous Cell carcinoma)ของวิธีการเรียนรู้ด้วยเครื่อง 6 วิธี จากชุดข้อมูลที่ปรับให้สมดุลแล้ว

วิธีการเรียนรู้ด้วยเครื่อง	จำนวนตัวแปรอิสระ (p)				
	200 (k=6, sigmoid, h=275)	400 (k=7, polynomial, h=275)	600 (k=7, polynomial, h=200)	800 (k=6, polynomial, h=300)	1,000 (k=5, polynomial, h=200)
1. DT	92.7466	93.2014	93.5255	93.5660	93.5877
2. Naïve	97.2733	97.2496	97.3511	97.1953	97.1366
3. KNN (k)	89.3922	91.0341	91.6024	92.4433	93.4571
4. SVM (Kernel Function)	97.8900	**99.9049**	**99.9717**	**99.9814**	**99.9943**
5. ANN (h)	96.2493	96.5578	96.3918	96.5603	96.3255
6. RF	**99.3669**	99.3281	99.3847	99.2952	99.3343

หมายเหตุ ตัวเลขที่เป็นตัวหนา หมายถึง ค่าเฉลี่ยร้อยละความแม่นยำที่มากที่สุดตามจำนวนตัวแปรอิสระ

จากตารางที่ 4.9 แสดงค่าเฉลี่ยร้อยละความแม่นยำของกลุ่มตัวแปรตาม ที่แทนด้วยหมายเลข 2 ในการจำแนกระดับของมะเร็งปอดด้วยวิธีการเรียนรู้ด้วยเครื่อง 6 วิธี จากชุดข้อมูลที่ปรับให้สมดุลแล้ว พบว่า ที่จำนวนตัวแปรอิสระมีค่าเท่ากับ 200 ของวิธีการเรียนรู้ด้วยเครื่อง RF มีค่าเฉลี่ยร้อยละความแม่นยำที่มากที่สุด เมื่อพิจารณาจำนวนตัวแปรอิสระมีค่าเท่ากับ 400 600 800 และ 1,000 ของวิธีการเรียนรู้ด้วยเครื่องทั้ง 6 วิธี จะเห็นว่าวิธี SVM มีค่าเฉลี่ยร้อยละความแม่นยำสูงที่สุด และนำค่าเฉลี่ยร้อยละความแม่นยำจากตารางที่ 4.9 มาแสดงผลในรูปแบบของกราฟ ดังรูปที่ 4.10



รูปที่ 4.10 กราฟแสดงค่าเฉลี่ยร้อยละความแม่นยำกลุ่มตัวแปรตาม ที่แทนด้วยหมายเลข 2 ของชุดข้อมูลที่ปรับให้สมดุลแล้ว ของวิธีการเรียนรู้ด้วยเครื่อง 6 วิธี

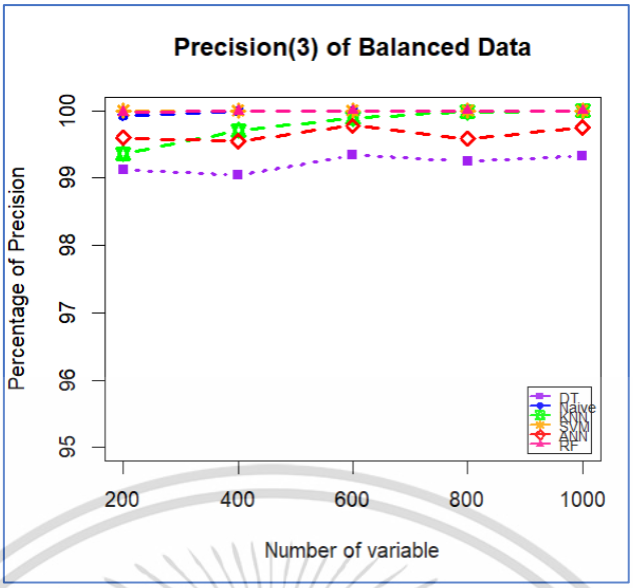
จากรูปที่ 4.10 พบว่าชุดข้อมูลที่ปรับให้สมดุลแล้ว ของค่าเฉลี่ยร้อยละความแม่นยำกลุ่มตัวแปรตาม ที่แทนด้วยหมายเลข 2 หากมีจำนวนตัวแปรอิสระเพิ่มมากขึ้น ค่าเฉลี่ยร้อยละความแม่นยำของวิธีการเรียนรู้ด้วยเครื่อง DT KNN และ SVM ก็มีค่าเฉลี่ยร้อยละความแม่นยำเพิ่มขึ้นด้วย เมื่อพิจารณาที่จำนวนตัวแปรอิสระมีค่าเท่ากับ 200 ของวิธีการเรียนรู้ด้วยเครื่อง RF มีค่าเฉลี่ยร้อยละความแม่นยำที่มากที่สุด เมื่อพิจารณาจำนวนตัวแปรอิสระมีค่าเท่ากับ 400 600 800 และ 1,000 ของวิธีการเรียนรู้ด้วยเครื่องทั้ง 6 วิธี จะเห็นว่าวิธี SVM มีค่าเฉลี่ยร้อยละความแม่นยำสูงที่สุด

ตารางที่ 4.10 ตารางแสดงค่าเฉลี่ยร้อยละความแม่นยำของกลุ่มตัวแปรตามที่แทนด้วยหมายเลข 3 (Non-Small Cell lung cancer (NSCLC) : Large Cell lung cancer) ของวิธีการเรียนรู้ด้วยเครื่อง 6 วิธี จากชุดข้อมูลที่ปรับให้สมดุลแล้ว

วิธีการเรียนรู้ด้วยเครื่อง	จำนวนตัวแปรอิสระ (p)				
	200 (k=6, sigmoid, h=275)	400 (k=7, polynomial, h=275)	600 (k=7, polynomial, h=200)	800 (k=6, polynomial, h=300)	1,000 (k=5, polynomial, h=200)
1. DT	99.1234	99.0410	99.3447	99.2437	99.3346
2. Naïve	99.9329	99.9875	99.9952	**100**	**100**
3. KNN (k)	99.3594	99.7058	99.8919	99.9832	**100**
4. SVM (Kernel Function)	**100**	**100**	**100**	**100**	**100**
5. ANN (h)	99.5972	99.5521	99.7762	99.5884	99.7465
6. RF	99.9723	99.9921	99.9948	**100**	**100**

หมายเหตุ ตัวเลขที่เป็นตัวหนา หมายถึง ค่าเฉลี่ยร้อยละความแม่นยำที่มากที่สุดตามจำนวนตัวแปรอิสระ

จากตารางที่ 4.10 แสดงค่าเฉลี่ยร้อยละความแม่นยำของกลุ่มตัวแปรตาม ที่แทนด้วยหมายเลข 3 ในการจำแนกระดับของมะเร็งปอดด้วยวิธีการเรียนรู้ด้วยเครื่อง 6 วิธี จากชุดข้อมูลที่ปรับให้สมดุลแล้ว พบว่า ที่จำนวนตัวแปรอิสระมีค่าเท่ากับ 200 400 และ 600 ของวิธีการเรียนรู้ด้วยเครื่อง SVM มีค่าเฉลี่ยร้อยละความแม่นยำที่มากที่สุด เมื่อพิจารณาจำนวนตัวแปรอิสระมีค่าเท่ากับ 800 วิธี Naïve SVM และ RF มีค่าเฉลี่ยร้อยละความแม่นยำที่มากที่สุด เมื่อพิจารณาจำนวนตัวแปรอิสระมีค่าเท่ากับ 1,000 จะเห็นว่าวิธี Naïve KNN SVM และ RF มีค่าเฉลี่ยร้อยละความแม่นยำสูงที่สุด และนำค่าเฉลี่ยร้อยละความแม่นยำจากตารางที่ 4.10 มาแสดงผลในรูปแบบของกราฟ ดังรูปที่ 4.11



รูปที่ 4.11 กราฟแสดงค่าเฉลี่ยร้อยละความแม่นยำกลุ่มตัวแปรตาม ที่แทนด้วยหมายเลข 3 ของชุดข้อมูลที่ปรับให้สมดุลแล้ว ของวิธีการเรียนรู้ด้วยเครื่อง 6 วิธี

จากรูปที่ 4.11 พบว่าชุดข้อมูลที่ปรับให้สมดุลแล้ว ของค่าเฉลี่ยร้อยละความแม่นยำกลุ่มตัวแปรตาม ที่แทนด้วยหมายเลข 3 หากมีจำนวนตัวแปรอิสระเพิ่มมากขึ้น ค่าเฉลี่ยร้อยละความแม่นยำของวิธีการเรียนรู้ด้วยเครื่อง KNN ก็มีค่าเฉลี่ยร้อยละความแม่นยำเพิ่มขึ้นด้วย เมื่อพิจารณาที่จำนวนตัวแปรอิสระมีค่าเท่ากับ ที่จำนวนตัวแปรอิสระมีค่าเท่ากับ 200 400 และ 600 ของวิธีการเรียนรู้ด้วยเครื่อง SVM มีค่าเฉลี่ยร้อยละความแม่นยำที่มากที่สุด เมื่อพิจารณาจำนวนตัวแปรอิสระมีค่าเท่ากับ 800 วิธี Naïve SVM และ RF มีค่าเฉลี่ยร้อยละความแม่นยำที่มากที่สุด เมื่อพิจารณาจำนวนตัวแปรอิสระมีค่าเท่ากับ 1,000 จะเห็นว่าวิธี Naïve KNN SVM และ RF มีค่าเฉลี่ยร้อยละความแม่นยำสูงที่สุด

4.5 ผลการวิจัยค่าเฉลี่ยร้อยละความระลึก ของชุดข้อมูลดั้งเดิม

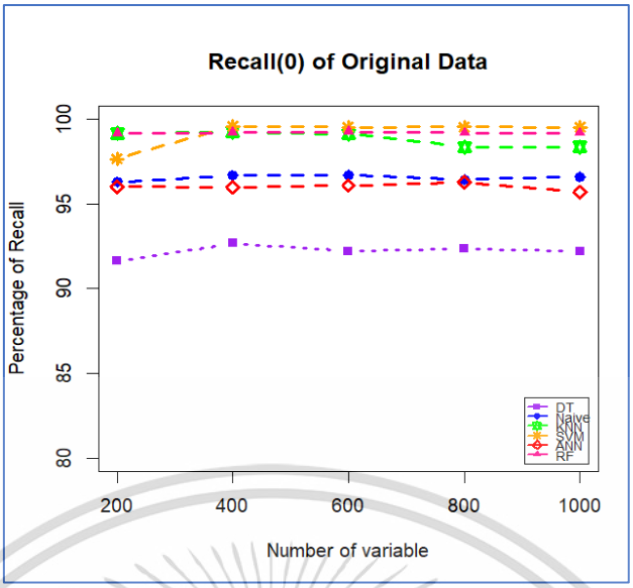
ผลการวิจัยค่าเฉลี่ยร้อยละความระลึก แยกตามกลุ่มของตัวแปรตาม ทั้ง 4 กลุ่ม ได้แก่ 1. Small Cell Lung Cancer (SCLC) : Oat cell lung cancer แทนด้วยหมายเลข 0 2. Non-Small Cell lung cancer (NACLC) : Adenocarcinoma แทนด้วยหมายเลข 1 3. Non-Small Cell lung cancer (NSCLC) : Squamous Cell carcinoma แทนด้วยหมายเลข 2 4. Non-Small Cell lung cancer (NSCLC) : Large Cell lung cancer แทนด้วยหมายเลข 3 ของชุดข้อมูลดั้งเดิม โดยมีจำนวนตัวแปรอิสระเป็น 200 400 600 800 และ 1,000 ของวิธีการเรียนรู้ด้วยเครื่องทั้ง 6 วิธี

ตารางที่ 4.11 ตารางแสดงค่าเฉลี่ยร้อยละความระลึกของกลุ่มตัวแปรตาม ที่แทนด้วยหมายเลข 0 (Small Cell Lung Cancer (SCLC) : Oat cell lung cancer) ของวิธีการเรียนรู้ด้วยเครื่อง 6 วิธี จากชุดข้อมูลดั้งเดิม

วิธีการเรียนรู้ด้วยเครื่อง	จำนวนตัวแปรอิสระ (p)				
	200 (k=6, sigmoid, h=275)	400 (k=7, polynomial, h=275)	600 (k=7, polynomial, h=200)	800 (k=6, polynomial, h=300)	1,000 (k=5, polynomial, h=200)
1. DT	91.6487	92.6817	92.2335	92.3492	92.2150
2. Naïve	96.3059	96.6888	96.7354	96.4463	96.6191
3. KNN (k)	99.1659	99.2715	99.1504	98.3716	98.3826
4. SVM (Kernel Function)	97.6745	**99.5905**	**99.5481**	**99.5634**	**99.5384**
5. ANN (h)	96.0242	96	96.1041	96.2948	95.6880
6. RF	**99.1680**	99.2238	99.2769	99.2171	99.1965

หมายเหตุ ตัวเลขที่เป็นตัวหนา หมายถึง ค่าเฉลี่ยร้อยละความระลึกที่มากที่สุดตามจำนวนตัวแปรอิสระ

จากตารางที่ 4.11 แสดงค่าเฉลี่ยร้อยละความระลึกของกลุ่มตัวแปรตาม ที่แทนด้วยหมายเลข 0 ในการจำแนกระดับของมะเร็งปอดด้วยวิธีการเรียนรู้ด้วยเครื่อง 6 วิธี จากชุดข้อมูลดั้งเดิม พบว่าที่จำนวนตัวแปรอิสระมีค่าเท่ากับ 200 ของวิธีการเรียนรู้ด้วยเครื่อง RF มีค่าเฉลี่ยร้อยละความระลึกที่มากที่สุด เมื่อพิจารณาจำนวนตัวแปรอิสระมีค่าเท่ากับ 400 600 800 และ 1,000 ของวิธีการเรียนรู้ด้วยเครื่องทั้ง 6 วิธี จะเห็นว่าวิธี SVM มีค่าเฉลี่ยร้อยละความระลึกสูงที่สุด และนำค่าเฉลี่ยร้อยละความระลึกจากตารางที่ 4.11 มาแสดงผลในรูปแบบของกราฟ ดังรูปที่ 4.12



รูปที่ 4.12 กราฟแสดงค่าเฉลี่ยร้อยละความระลึกกลุ่มตัวแปรตาม ที่แทนด้วยหมายเลข 0 ของชุดข้อมูลดั้งเดิม ของวิธีการเรียนรู้ด้วยเครื่อง 6 วิธี

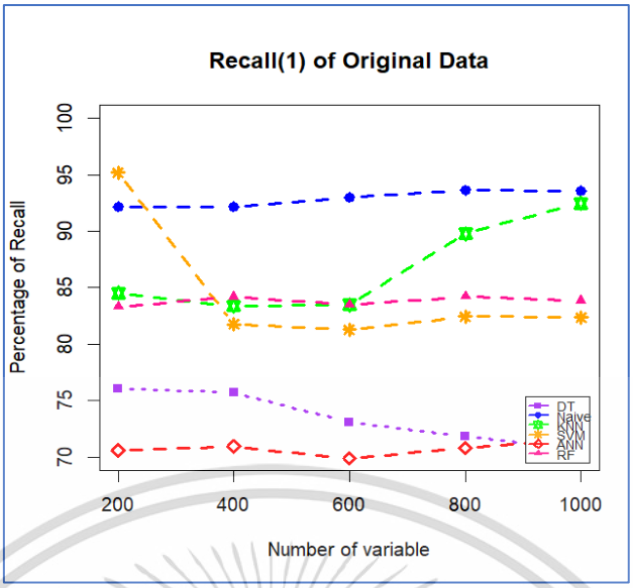
จากรูปที่ 4.12 พบว่าชุดข้อมูลดั้งเดิมของค่าเฉลี่ยร้อยละความระลึกกลุ่มตัวแปรตาม ที่แทนด้วยหมายเลข 0 เมื่อพิจารณาที่จำนวนตัวแปรอิสระมีค่าเท่ากับ ที่จำนวนตัวแปรอิสระมีค่าเท่ากับ 200 ของวิธีการเรียนรู้ด้วยเครื่อง RF มีค่าเฉลี่ยร้อยละความระลึกที่มากที่สุด เมื่อพิจารณาจำนวนตัวแปรอิสระมีค่าเท่ากับ 400 600 800 และ 1,000 ของวิธี SVM พบว่ามีค่าเฉลี่ยร้อยละความระลึกที่มากที่สุด

ตารางที่ 4.12 ตารางแสดงค่าเฉลี่ยร้อยละความระลึกของกลุ่มตัวแปรตาม ที่แทนด้วยหมายเลข 1 (Non-Small Cell lung cancer (NACLC) : Adenocarcinoma) ของวิธีการเรียนรู้ด้วยเครื่อง 6 วิธี จากชุดข้อมูลดั้งเดิม

วิธีการเรียนรู้ด้วยเครื่อง	จำนวนตัวแปรอิสระ (p)				
	200 (k=6, sigmoid, h=275)	400 (k=7, polynomial, h=275)	600 (k=7, polynomial, h=200)	800 (k=6, polynomial, h=300)	1,000 (k=5, polynomial, h=200)
1. DT	76.0164	75.6929	73.0683	71.8442	70.6766
2. Naïve	92.2014	**92.1884**	**93.0176**	**93.6422**	**93.6039**
3. KNN (k)	84.5057	83.3904	83.5062	89.7974	92.4683
4. SVM (Kernel Function)	**95.2077**	81.7506	81.2811	82.4608	82.3645
5. ANN (h)	70.5646	70.9262	69.8932	70.8001	71.5783
6. RF	83.3226	84.1995	83.4714	84.2162	83.8748

หมายเหตุ ตัวเลขที่เป็นตัวหนา หมายถึง ค่าเฉลี่ยร้อยละความระลึกที่มากที่สุดตามจำนวนตัวแปรอิสระ

จากตารางที่ 4.12 แสดงค่าเฉลี่ยร้อยละความระลึกของกลุ่มตัวแปรตาม ที่แทนด้วยหมายเลข 1 ในการจำแนกระดับของมะเร็งปอดด้วยวิธีการเรียนรู้ด้วยเครื่อง 6 วิธี จากชุดข้อมูลดั้งเดิม พบว่าที่จำนวนตัวแปรอิสระมีค่าเท่ากับ 200 ของวิธีการเรียนรู้ด้วยเครื่อง SVM มีค่าเฉลี่ยร้อยละความระลึกที่มากที่สุด เมื่อพิจารณาจำนวนตัวแปรอิสระมีค่าเท่ากับ 400 600 800 และ 1,000 ของวิธีการเรียนรู้ด้วยเครื่องทั้ง 6 วิธี จะเห็นว่าวิธี Naïve มีค่าเฉลี่ยร้อยละความระลึกสูงที่สุด และนำค่าเฉลี่ยร้อยละความระลึกจากตารางที่ 4.12 มาแสดงผลในรูปแบบของกราฟ ดังรูปที่ 4.13



รูปที่ 4.13 กราฟแสดงค่าเฉลี่ยร้อยละความระลึกกลุ่มตัวแปรตาม ที่แทนด้วยหมายเลข 1 ของชุดข้อมูลดั้งเดิม ของวิธีการเรียนรู้ด้วยเครื่อง 6 วิธี

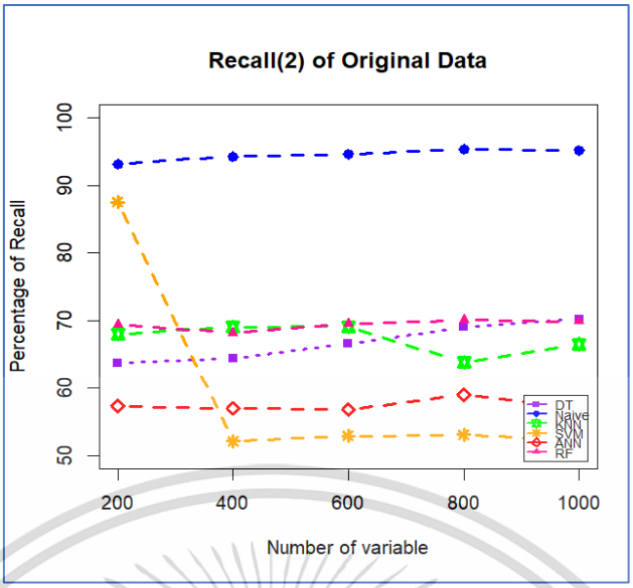
จากรูปที่ 4.13 พบว่าชุดข้อมูลดั้งเดิมของค่าเฉลี่ยร้อยละความระลึกกลุ่มตัวแปรตาม ที่แทนด้วยหมายเลข 1 หากมีจำนวนตัวแปรอิสระเพิ่มมากขึ้น ค่าเฉลี่ยร้อยละความระลึกของวิธีการเรียนรู้ด้วยเครื่อง DT จะมีค่าเฉลี่ยร้อยละความระลึกลดลง เมื่อพิจารณาที่จำนวนตัวแปรอิสระมีค่าเท่ากับที่จำนวนตัวแปรอิสระมีค่าเท่ากับ 200 ของวิธีการเรียนรู้ด้วยเครื่อง SVM มีค่าเฉลี่ยร้อยละความระลึกที่มากที่สุด เมื่อพิจารณาจำนวนตัวแปรอิสระมีค่าเท่ากับ 400 600 800 และ 1,000 ของวิธี Naïve พบว่ามีค่าเฉลี่ยร้อยละความระลึกที่มากที่สุด

ตารางที่ 4.13 ตารางแสดงค่าเฉลี่ยร้อยละความระลึกของกลุ่มตัวแปรตาม ที่แทนด้วยหมายเลข 2 (Non-Small Cell lung cancer (NSCLC) : Squamous Cell carcinoma)ของวิธีการเรียนรู้ด้วยเครื่อง 6 วิธี จากชุดข้อมูลดั้งเดิม

วิธีการเรียนรู้ด้วยเครื่อง	จำนวนตัวแปรอิสระ (p)				
	200 (k=6, sigmoid, h=275)	400 (k=7, polynomial, h=275)	600 (k=7, polynomial, h=200)	800 (k=6, polynomial, h=300)	1,000 (k=5, polynomial, h=200)
1. DT	63.6498	64.3628	66.5466	69.0353	70.2188
2. Naïve	**93.1983**	**94.3217**	**94.6309**	**95.3847**	**95.1671**
3. KNN (k)	67.9641	69.0520	69.0942	63.8188	66.4732
4. SVM (Kernel Function)	87.5462	52.1042	52.8796	53.0699	52.2524
5. ANN (h)	57.3580	56.9940	56.8159	59.0210	57.0807
6. RF	69.3907	68.1343	69.5379	70.0623	69.8520

หมายเหตุ ตัวเลขที่เป็นตัวหนา หมายถึง ค่าเฉลี่ยร้อยละความระลึกที่มากที่สุดตามจำนวนตัวแปรอิสระ

จากตารางที่ 4.13 แสดงค่าเฉลี่ยร้อยละความระลึกของกลุ่มตัวแปรตาม ที่แทนด้วยหมายเลข 2 ในการจำแนกระดับของมะเร็งปอดด้วยวิธีการเรียนรู้ด้วยเครื่อง 6 วิธี จากชุดข้อมูลดั้งเดิม พบว่าเมื่อพิจารณาจำนวนตัวแปรอิสระมีค่าเท่ากับ 200 400 600 800 และ 1,000 ของวิธีการเรียนรู้ด้วยเครื่องทั้ง 6 วิธี จะเห็นว่าวิธี Naïve มีค่าเฉลี่ยร้อยละความระลึกสูงที่สุด และนำค่าเฉลี่ยร้อยละความระลึกจากตารางที่ 4.13 มาแสดงผลในรูปแบบของกราฟ ดังรูปที่ 4.14



รูปที่ 4.14 กราฟแสดงค่าเฉลี่ยร้อยละความระลึกกลุ่มตัวแปรตาม ที่แทนด้วยหมายเลข 2 ของชุดข้อมูลดั้งเดิม ของวิธีการเรียนรู้ด้วยเครื่อง 6 วิธี

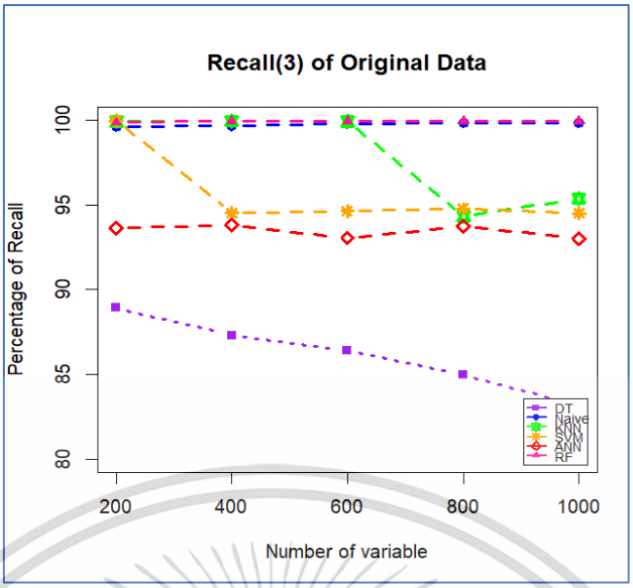
จากรูปที่ 4.14 พบว่าชุดข้อมูลดั้งเดิมของค่าเฉลี่ยร้อยละความระลึกกลุ่มตัวแปรตาม ที่แทนด้วยหมายเลข 2 หากมีจำนวนตัวแปรอิสระเพิ่มมากขึ้น ค่าเฉลี่ยร้อยละความระลึกของวิธีการเรียนรู้ด้วยเครื่อง DT ก็มีค่าเฉลี่ยร้อยละความระลึกเพิ่มขึ้นด้วย เมื่อพิจารณาที่จำนวนตัวแปรอิสระมีค่าเท่ากับ 200 400 600 800 และ 1,000 ของวิธีการเรียนรู้ด้วยเครื่อง จะเห็นว่าวิธี Naïve มีค่าเฉลี่ยร้อยละความระลึกที่มากที่สุด

ตารางที่ 4.14 ตารางแสดงค่าเฉลี่ยร้อยละความระลึกของกลุ่มตัวแปรตาม ที่แทนด้วยหมายเลข 3 (Non-Small Cell lung cancer (NSCLC) : Large Cell lung cancer) ของวิธีการเรียนรู้ด้วยเครื่อง 6 วิธี จากชุดข้อมูลดั้งเดิม

วิธีการเรียนรู้ด้วยเครื่อง	จำนวนตัวแปรอิสระ (p)				
	200 (k=6, sigmoid, h=275)	400 (k=7, polynomial, h=275)	600 (k=7, polynomial, h=200)	800 (k=6, polynomial, h=300)	1,000 (k=5, polynomial, h=200)
1. DT	88.9376	87.3127	86.4086	84.9567	83.1762
2. Naïve	99.6520	99.6827	99.8203	99.8574	99.8680
3. KNN (k)	99.9507	**99.9742**	99.9470	94.3431	95.3923
4. SVM (Kernel Function)	**99.9917**	94.5498	94.6417	94.7788	94.5218
5. ANN (h)	93.6369	93.8112	93.0629	93.7603	93.0183
6. RF	99.8766	99.9472	**99.9607**	**99.9392**	**99.9576**

หมายเหตุ ตัวเลขที่เป็นตัวหนา หมายถึง ค่าเฉลี่ยร้อยละความระลึกที่มากที่สุดตามจำนวนตัวแปรอิสระ

จากตารางที่ 4.14 แสดงค่าเฉลี่ยร้อยละความระลึกของกลุ่มตัวแปรตาม ที่แทนด้วยหมายเลข 3 ในการจำแนกระดับของมะเร็งปอดด้วยวิธีการเรียนรู้ด้วยเครื่อง 6 วิธี จากชุดข้อมูลดั้งเดิม พบว่า เมื่อพิจารณาจำนวนตัวแปรอิสระมีค่าเท่ากับ 200 วิธีการเรียนรู้ด้วยเครื่อง SVM มีค่าเฉลี่ยร้อยละความระลึกสูงที่สุด ที่จำนวนตัวแปรอิสระมีค่าเท่ากับ 400 วิธี KNN มีค่าเฉลี่ยร้อยละความระลึกสูงที่สุด เมื่อพิจารณาจำนวนตัวแปรอิสระมีค่าเท่ากับ 600 800 และ 1,000 ของวิธีการเรียนรู้ด้วยเครื่องทั้ง 6 วิธี จะเห็นว่าวิธี RF มีค่าเฉลี่ยร้อยละความระลึกสูงที่สุด และนำค่าเฉลี่ยร้อยละความระลึกจากตารางที่ 4.14 มาแสดงผลในรูปแบบของกราฟ ดังรูปที่ 4.15



รูปที่ 4.15 กราฟแสดงค่าเฉลี่ยร้อยละความระลึกกลุ่มตัวแปรตาม ที่แทนด้วยหมายเลข 3 ของชุดข้อมูลดั้งเดิม ของวิธีการเรียนรู้ด้วยเครื่อง 6 วิธี

จากรูปที่ 4.15 พบว่าชุดข้อมูลดั้งเดิมของค่าเฉลี่ยร้อยละความระลึกกลุ่มตัวแปรตาม ที่แทนด้วยหมายเลข 3 หากมีจำนวนตัวแปรอิสระเพิ่มมากขึ้น ค่าเฉลี่ยร้อยละความระลึกของวิธีการเรียนรู้ด้วยเครื่อง DT จะมีค่าเฉลี่ยร้อยละความระลึกลดลง และ หากมีจำนวนตัวแปรอิสระเพิ่มมากขึ้น ค่าเฉลี่ยร้อยละความระลึกของวิธีการเรียนรู้ด้วยเครื่อง Naïve ก็มีค่าเฉลี่ยร้อยละความระลึกเพิ่มขึ้นด้วย เมื่อพิจารณาที่จำนวนตัวแปรอิสระมีค่าเท่ากับ 200 วิธีการเรียนรู้ด้วยเครื่อง SVM มีค่าเฉลี่ยร้อยละความระลึกสูงที่สุด ที่จำนวนตัวแปรอิสระมีค่าเท่ากับ 400 วิธี KNN มีค่าเฉลี่ยร้อยละความระลึกสูงที่สุด เมื่อพิจารณาจำนวนตัวแปรอิสระมีค่าเท่ากับ 600 800 และ 1,000 วิธี RF มีค่าเฉลี่ยร้อยละความระลึกสูงที่สุด

4.6 ผลการวิจัยค่าเฉลี่ยร้อยละความระลึก ของชุดข้อมูลที่ปรับให้สมดุลแล้ว

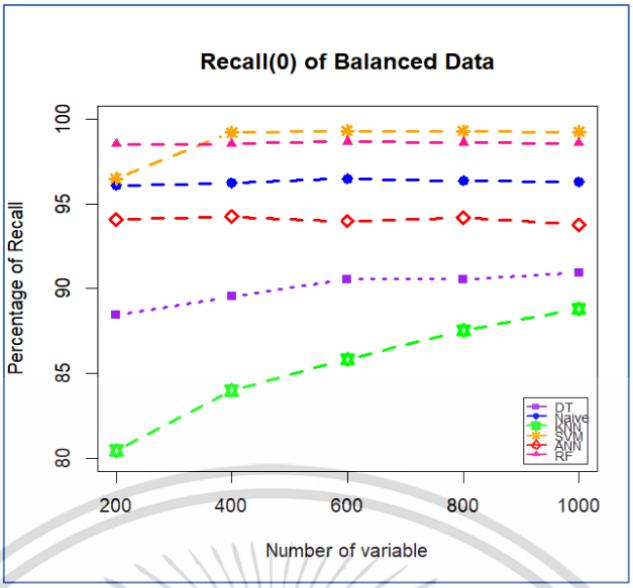
ผลการวิจัยค่าเฉลี่ยร้อยละความระลึก แยกตามกลุ่มของตัวแปรตาม ทั้ง 4 กลุ่ม ได้แก่ 1. Small Cell Lung Cancer (SCLC) : Oat cell lung cancer แทนด้วยหมายเลข 0 2. Non-Small Cell lung cancer (NACLC) : Adenocarcinoma แทนด้วยหมายเลข 1 3. Non-Small Cell lung cancer (NSCLC) : Squamous Cell carcinoma แทนด้วยหมายเลข 2 4. Non-Small Cell lung cancer (NSCLC) : Large Cell lung cancer แทนด้วยหมายเลข 3 ของชุดข้อมูลที่ปรับให้สมดุลแล้ว โดยมีจำนวนตัวแปรอิสระเป็น 200 400 600 800 และ 1,000 ของวิธีการเรียนรู้ด้วยเครื่องทั้ง 6 วิธี

ตารางที่ 4.15 ตารางแสดงค่าเฉลี่ยร้อยละความระลึกของกลุ่มตัวแปรตาม ที่แทนด้วยหมายเลข 0 (Small Cell Lung Cancer (SCLC) : Oat cell lung cancer) ของวิธีการเรียนรู้ด้วยเครื่อง 6 วิธี จากชุดข้อมูลที่ปรับให้สมดุลแล้ว

วิธีการเรียนรู้ด้วยเครื่อง	จำนวนตัวแปรอิสระ (p)				
	200 (k=6, sigmoid, h=275)	400 (k=7, polynomial, h=275)	600 (k=7, polynomial, h=200)	800 (k=6, polynomial, h=300)	1,000 (k=5, polynomial, h=200)
1. DT	88.4301	89.5415	90.5658	90.5630	90.9238
2. Naïve	96.1030	96.2477	96.5060	96.3839	96.3007
3. KNN (k)	80.4448	83.9950	85.8378	87.5334	88.8134
4. SVM (Kernel Function)	96.5070	**99.2450**	**99.3157**	**99.3086**	**99.2670**
5. ANN (h)	94.0938	94.2528	93.9964	94.2000	93.7723
6. RF	**98.5529**	98.5517	98.6841	98.6190	98.6014

หมายเหตุ ตัวเลขที่เป็นตัวหนา หมายถึง ค่าเฉลี่ยร้อยละความระลึกที่มากที่สุดตามจำนวนตัวแปรอิสระ

จากตารางที่ 4.15 แสดงค่าเฉลี่ยร้อยละความระลึกของกลุ่มตัวแปรตาม ที่แทนด้วยหมายเลข 0 ในการจำแนกระดับของมะเร็งปอดด้วยวิธีการเรียนรู้ด้วยเครื่อง 6 วิธี จากชุดข้อมูลที่ปรับให้สมดุลแล้ว พบว่า เมื่อพิจารณาจำนวนตัวแปรอิสระมีค่าเท่ากับ 200 วิธีการเรียนรู้ด้วยเครื่อง RF มีค่าเฉลี่ยร้อยละความระลึกสูงที่สุด เมื่อพิจารณาจำนวนตัวแปรอิสระมีค่าเท่ากับ 400 600 800 และ 1,000 ของวิธีการเรียนรู้ด้วยเครื่องทั้ง 6 วิธี จะเห็นว่าวิธี SVM มีค่าเฉลี่ยร้อยละความระลึกสูงที่สุด และนำค่าเฉลี่ยร้อยละความระลึกจากตารางที่ 4.15 มาแสดงผลในรูปแบบของกราฟ ดังรูปที่ 4.16



รูปที่ 4.16 กราฟแสดงค่าเฉลี่ยร้อยละความระลึกกลุ่มตัวแปรตาม ที่แทนด้วยหมายเลข 0 ของชุดข้อมูลที่ปรับให้สมดุลแล้ว ของวิธีการเรียนรู้ด้วยเครื่อง 6 วิธี

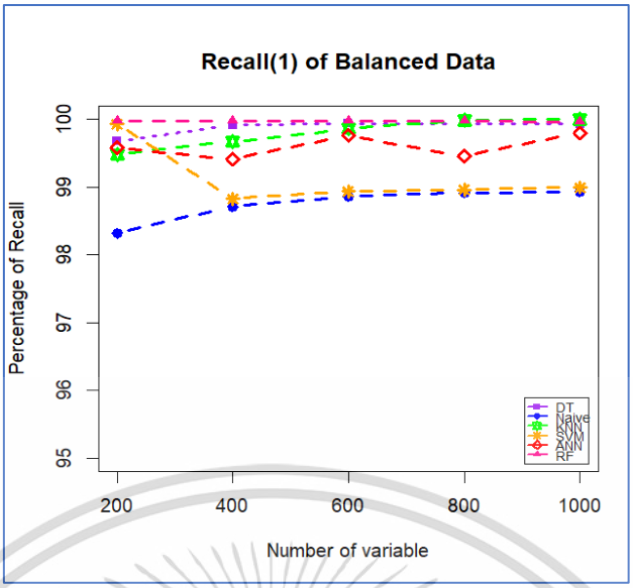
จากรูปที่ 4.16 พบว่าชุดข้อมูลที่ปรับให้สมดุลแล้วของค่าเฉลี่ยร้อยละความระลึกกลุ่มตัวแปรตาม ที่แทนด้วยหมายเลข 0 หากมีจำนวนตัวแปรอิสระเพิ่มมากขึ้น ค่าเฉลี่ยร้อยละความระลึกของวิธีการเรียนรู้ด้วยเครื่อง KNN จะมีค่าเฉลี่ยร้อยละความระลึกเพิ่มขึ้น เมื่อพิจารณาที่จำนวนตัวแปรอิสระมีค่าเท่ากับ 200 วิธีการเรียนรู้ด้วยเครื่อง RF มีค่าเฉลี่ยร้อยละความระลึกสูงที่สุด เมื่อพิจารณาจำนวนตัวแปรอิสระมีค่าเท่ากับ 400 600 800 และ 1,000 วิธี SVM มีค่าเฉลี่ยร้อยละความระลึกสูงที่สุด

ตารางที่ 4.16 ตารางแสดงค่าเฉลี่ยร้อยละความระลึกของกลุ่มตัวแปรตาม ที่แทนด้วยหมายเลข 1 (Non-Small Cell lung cancer (NACLC) : Adenocarcinoma) ของวิธีการเรียนรู้ด้วยเครื่อง 6 วิธี จากชุดข้อมูลที่ปรับให้สมดุลแล้ว

วิธีการเรียนรู้ด้วยเครื่อง	จำนวนตัวแปรอิสระ (p)				
	200 (k=6, sigmoid, h=275)	400 (k=7, polynomial, h=275)	600 (k=7, polynomial, h=200)	800 (k=6, polynomial, h=300)	1,000 (k=5, polynomial, h=200)
1. DT	99.6839	99.9219	99.9392	99.9390	99.9343
2. Naïve	98.3188	98.7144	98.8635	98.9198	98.9319
3. KNN (k)	99.4835	99.6745	99.8657	**99.9867**	**100**
4. SVM (Kernel Function)	99.9296	98.8291	98.9390	98.9618	98.9970
5. ANN (h)	99.5811	99.4155	99.7654	99.4561	99.7985
6. RF	**99.9834**	**99.9802**	**99.9784**	99.9806	99.9682

หมายเหตุ ตัวเลขที่เป็นตัวหนา หมายถึง ค่าเฉลี่ยร้อยละความระลึกที่มากที่สุดตามจำนวนตัวแปรอิสระ

จากตารางที่ 4.16 แสดงค่าเฉลี่ยร้อยละความระลึกของกลุ่มตัวแปรตาม ที่แทนด้วยหมายเลข 1 ในการจำแนกระดับของมะเร็งปอดด้วยวิธีการเรียนรู้ด้วยเครื่อง 6 วิธี จากชุดข้อมูลที่ปรับให้สมดุลแล้ว พบว่า เมื่อพิจารณาจำนวนตัวแปรอิสระมีค่าเท่ากับ 200 400 และ 600 วิธีการเรียนรู้ด้วยเครื่อง RF มีค่าเฉลี่ยร้อยละความระลึกสูงที่สุด เมื่อพิจารณาจำนวนตัวแปรอิสระมีค่าเท่ากับ 800 และ 1,000 ของวิธีการเรียนรู้ด้วยเครื่องทั้ง 6 วิธี จะเห็นว่าวิธี KNN มีค่าเฉลี่ยร้อยละความระลึกสูงที่สุด และนำค่าเฉลี่ยร้อยละความระลึกจากตารางที่ 4.16 มาแสดงผลในรูปแบบของกราฟ ดังรูปที่ 4.17



รูปที่ 4.17 กราฟแสดงค่าเฉลี่ยร้อยละความระลึกกลุ่มตัวแปรตาม ที่แทนด้วยหมายเลข 1 ของชุดข้อมูลที่ปรับให้สมดุลแล้ว ของวิธีการเรียนรู้ด้วยเครื่อง 6 วิธี

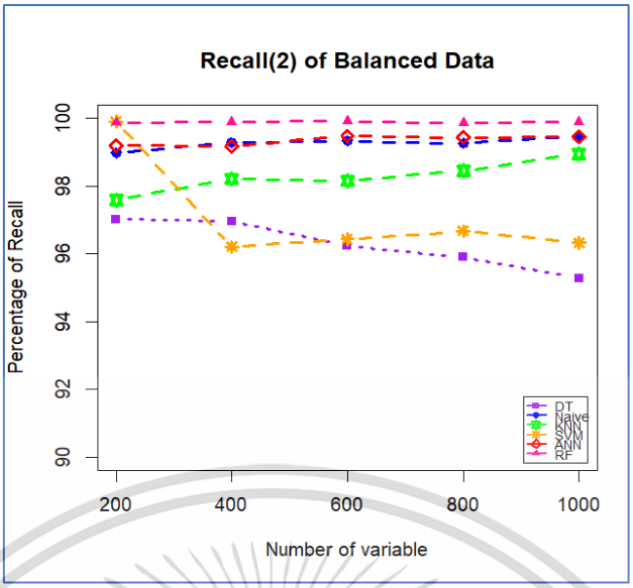
จากรูปที่ 4.17 พบว่าชุดข้อมูลที่ปรับให้สมดุลแล้วของค่าเฉลี่ยร้อยละความระลึกกลุ่มตัวแปรตาม ที่แทนด้วยหมายเลข 1 หากมีจำนวนตัวแปรอิสระเพิ่มมากขึ้น ค่าเฉลี่ยร้อยละความระลึกของวิธีการเรียนรู้ด้วยเครื่อง Naïve และ KNN จะมีค่าเฉลี่ยร้อยละความระลึกเพิ่มขึ้นด้วย เมื่อพิจารณาที่จำนวนตัวแปรอิสระมีค่าเท่ากับ 200 400 และ 600 วิธีการเรียนรู้ด้วยเครื่อง RF มีค่าเฉลี่ยร้อยละความระลึกสูงที่สุด เมื่อพิจารณาจำนวนตัวแปรอิสระมีค่าเท่ากับ 800 และ 1,000 วิธี KNN มีค่าเฉลี่ยร้อยละความระลึกสูงที่สุด

ตารางที่ 4.17 ตารางแสดงค่าเฉลี่ยร้อยละความระลึกของกลุ่มตัวแปรตามที่แทนด้วยหมายเลข 2 (Non-Small Cell lung cancer (NSCLC) : Squamous Cell carcinoma) ของวิธีการเรียนรู้ด้วยเครื่อง 6 วิธี จากชุดข้อมูลที่ปรับให้สมดุลแล้ว

วิธีการเรียนรู้ด้วยเครื่อง	จำนวนตัวแปรอิสระ (p)				
	200 (k=6, sigmoid, h=275)	400 (k=7, polynomial, h=275)	600 (k=7, polynomial, h=200)	800 (k=6, polynomial, h=300)	1,000 (k=5, polynomial, h=200)
1. DT	97.0247	96.9675	96.2478	95.9030	95.2682
2. Naïve	98.9778	99.2918	99.3377	99.2718	99.4852
3. KNN (k)	97.5891	98.2123	98.1666	98.4550	98.9506
4. SVM (Kernel Function)	**99.9069**	96.2025	96.4392	96.6704	96.3331
5. ANN (h)	99.2048	99.1777	99.4887	99.4394	99.4675
6. RF	99.8741	**99.8906**	**99.9179**	**99.8591**	**99.8918**

หมายเหตุ ตัวเลขที่เป็นตัวหนา หมายถึง ค่าเฉลี่ยร้อยละความระลึกที่มากที่สุดตามจำนวนตัวแปรอิสระ

จากตารางที่ 4.17 แสดงค่าเฉลี่ยร้อยละความระลึกของกลุ่มตัวแปรตาม ที่แทนด้วยหมายเลข 2 ในการจำแนกระดับของมะเร็งปอดด้วยวิธีการเรียนรู้ด้วยเครื่อง 6 วิธี จากชุดข้อมูลที่ปรับให้สมดุลแล้ว พบว่า เมื่อพิจารณาจำนวนตัวแปรอิสระมีค่าเท่ากับ 200 วิธีการเรียนรู้ด้วยเครื่อง SVM มีค่าเฉลี่ยร้อยละความระลึกสูงที่สุด เมื่อพิจารณาจำนวนตัวแปรอิสระมีค่าเท่ากับ 400 600 800 และ 1,000 ของวิธีการเรียนรู้ด้วยเครื่องทั้ง 6 วิธี จะเห็นว่าวิธี RF มีค่าเฉลี่ยร้อยละความระลึกสูงที่สุด และนำค่าเฉลี่ยร้อยละความระลึกจากตารางที่ 4.17 มาแสดงผลในรูปแบบของกราฟ ดังรูปที่ 4.18



รูปที่ 4.18 กราฟแสดงค่าเฉลี่ยร้อยละความระลึกกลุ่มตัวแปรตามที่แทนด้วยหมายเลข 2 ของชุดข้อมูลที่ปรับให้สมดุลแล้ว ของวิธีการเรียนรู้ด้วยเครื่อง 6 วิธี

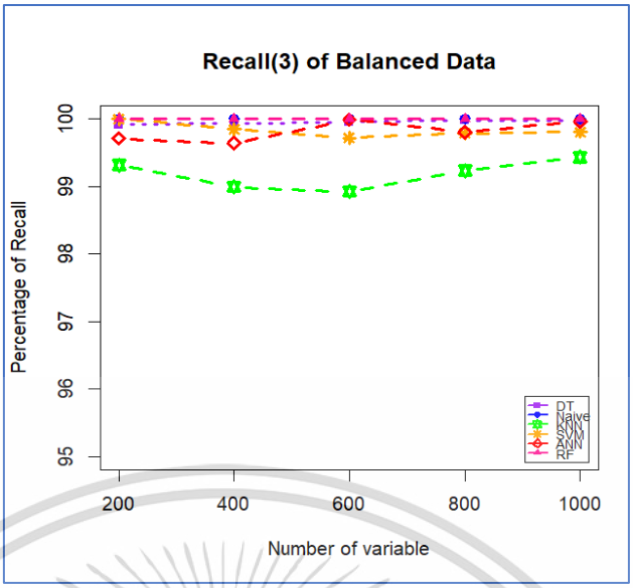
จากรูปที่ 4.18 พบว่าชุดข้อมูลที่ปรับให้สมดุลแล้วของค่าเฉลี่ยร้อยละความระลึกกลุ่มตัวแปรตาม ที่แทนด้วยหมายเลข 2 หากมีจำนวนตัวแปรอิสระเพิ่มมากขึ้น ค่าเฉลี่ยร้อยละความระลึกของวิธีการเรียนรู้ด้วยเครื่อง DT จะมีค่าเฉลี่ยร้อยละความระลึกลดลง หากมีจำนวนตัวแปรอิสระเพิ่มมากขึ้น ค่าเฉลี่ยร้อยละความระลึกของวิธีการเรียนรู้ด้วยเครื่อง KNN จะมีค่าเฉลี่ยร้อยละความระลึกเพิ่มขึ้นด้วย เมื่อพิจารณาที่จำนวนตัวแปรอิสระมีค่าเท่ากับ 200 วิธีการเรียนรู้ด้วยเครื่อง SVM มีค่าเฉลี่ยร้อยละความระลึกสูงที่สุด เมื่อพิจารณาจำนวนตัวแปรอิสระมีค่าเท่ากับ 400 600 800 และ 1,000 วิธี RF มีค่าเฉลี่ยร้อยละความระลึกสูงที่สุด

ตารางที่ 4.18 ตารางแสดงค่าเฉลี่ยร้อยละความระลึกของกลุ่มตัวแปรตามที่แทนด้วยหมายเลข 3 (Non-Small Cell lung cancer (NSCLC) : Large Cell lung cancer) ของวิธีการเรียนรู้ด้วยเครื่อง 6 วิธี จากชุดข้อมูลที่ปรับให้สมดุลแล้ว

วิธีการเรียนรู้ด้วยเครื่อง	จำนวนตัวแปรอิสระ (p)				
	200 (k=6, sigmoid, h=275)	400 (k=7, polynomial, h=275)	600 (k=7, polynomial, h=200)	800 (k=6, polynomial, h=300)	1,000 (k=5, polynomial, h=200)
1. DT	99.9134	99.9410	99.9587	99.9828	99.9769
2. Naïve	**100**	**100**	**100**	**100**	**100**
3. KNN (k)	99.3214	98.9978	98.9254	99.2384	99.4353
4. SVM (Kernel Function)	**100**	99.8516	99.7176	99.7873	99.8112
5. ANN (h)	99.7150	99.6423	99.9915	99.8043	99.9651
6. RF	**100**	**100**	**100**	**100**	**100**

หมายเหตุ ตัวเลขที่เป็นตัวหนา หมายถึง ค่าเฉลี่ยร้อยละความระลึกที่มากที่สุดตามจำนวนตัวแปรอิสระ

จากตารางที่ 4.18 แสดงค่าเฉลี่ยร้อยละความระลึกของกลุ่มตัวแปรตามที่แทนด้วยหมายเลข 3 ในการจำแนกระดับของมะเร็งปอดด้วยวิธีการเรียนรู้ด้วยเครื่อง 6 วิธี จากชุดข้อมูลที่ปรับให้สมดุลแล้ว พบว่า เมื่อพิจารณาจำนวนตัวแปรอิสระมีค่าเท่ากับ 200 วิธีการเรียนรู้ด้วยเครื่อง Naïve SVM และ RF มีค่าเฉลี่ยร้อยละความระลึกสูงที่สุด เมื่อพิจารณาจำนวนตัวแปรอิสระมีค่าเท่ากับ 400 600 800 และ 1,000 ของวิธีการเรียนรู้ด้วยเครื่องทั้ง 6 วิธี จะเห็นว่าวิธี Naïve และ RF มีค่าเฉลี่ยร้อยละความระลึกสูงที่สุด และนำค่าเฉลี่ยร้อยละความระลึกจากตารางที่ 4.18 มาแสดงผลในรูปแบบของกราฟ ดังรูปที่ 4.19



รูปที่ 4.19 กราฟแสดงค่าเฉลี่ยร้อยละความระลึกกลุ่มตัวแปรตามที่แทนด้วยหมายเลข 3 ของชุดข้อมูลที่ปรับให้สมดุลแล้ว ของวิธีการเรียนรู้ด้วยเครื่อง 6 วิธี

จากรูปที่ 4.19 พบว่าชุดข้อมูลที่ปรับให้สมดุลแล้วของค่าเฉลี่ยร้อยละความระลึกกลุ่มตัวแปรตาม ที่แทนด้วยหมายเลข 2 เมื่อพิจารณาที่จำนวนตัวแปรอิสระมีค่าเท่ากับ 200 วิธีการเรียนรู้ด้วยเครื่อง Naïve SVM และ RF มีค่าเฉลี่ยร้อยละความระลึกสูงที่สุด เมื่อพิจารณาจำนวนตัวแปรอิสระมีค่าเท่ากับ 400 600 800 และ 1,000 วิธี Naïve และ RF มีค่าเฉลี่ยร้อยละความระลึกสูงที่สุด

4.7 ผลสรุปวิธีที่มีค่าเฉลี่ยร้อยละความถูกต้องที่สูงที่สุด

ผลสรุปวิธีที่มีค่าเฉลี่ยร้อยละความถูกต้องสูงที่สุด จากการวิเคราะห์โดยรวมตามกลุ่มของตัวแปรตามทั้ง 4 กลุ่ม ของชุดข้อมูลดั้งเดิม และชุดข้อมูลที่ปรับให้สมดุลแล้ว โดยมีจำนวนตัวแปรอิสระเป็น 200 400 600 800 และ 1,000 ดังตารางที่ 4.19

ตารางที่ 4.19 ตารางสรุปวิธีที่มีค่าเฉลี่ยร้อยละความถูกต้องที่สูงที่สุด ตามจำนวนตัวแปรอิสระของชุดข้อมูลดั้งเดิม และชุดข้อมูลที่ปรับให้สมดุลแล้ว

ประเภทของชุดข้อมูล	จำนวนตัวแปรอิสระ (p)				
	200	400	600	800	1,000
ชุดข้อมูลดั้งเดิม	**SVM**	Naïve	Naïve	Naïve	Naïve
ชุดข้อมูลที่ปรับให้สมดุลแล้ว	RF	RF	**RF**	RF	RF

หมายเหตุ ตัวหนังสือที่เป็นตัวหนา หมายถึง วิธีที่มีค่าเฉลี่ยร้อยละความถูกต้องที่มากที่สุด ตามประเภทของชุดข้อมูล

จากตารางที่ 4.19 พบว่าชุดข้อมูลดั้งเดิม วิธี SVM มีประสิทธิภาพในการทำนายสูงที่สุด เนื่องจากให้ค่าเฉลี่ยร้อยละความถูกต้องที่สูงที่สุด และวิธี Naïve มีประสิทธิภาพในการทำนายรองลงมา และที่ชุดข้อมูลที่ปรับให้สมดุลแล้ว วิธี RF มีประสิทธิภาพในการทำนายสูงที่สุด เนื่องจากให้ค่าเฉลี่ยร้อยละความถูกต้องที่สูงที่สุด

4.8 ผลสรุปวิธีที่มีค่าเฉลี่ยร้อยละความแม่นยำที่สูงที่สุด

ผลสรุปวิธีที่มีค่าเฉลี่ยร้อยละความแม่นยำสูงที่สุด จากการวิเคราะห์ตามกลุ่มของตัวแปรตาม ทั้ง 4 กลุ่ม ได้แก่

1. Small Cell Lung Cancer (SCLC) : Oat cell lung cancer แทนด้วยหมายเลข 0
2. Non-Small Cell lung cancer (NACLC) : Adenocarcinoma แทนด้วยหมายเลข 1
3. Non-Small Cell lung cancer (NSCLC) : Squamous Cell carcinoma แทนด้วยหมายเลข 2
4. Non-Small Cell lung cancer (NSCLC) : Large Cell lung cancer แทนด้วยหมายเลข 3

ของชุดข้อมูลดั้งเดิม และชุดข้อมูลที่ปรับให้สมดุลแล้ว โดยมีจำนวนตัวแปรอิสระเป็น 200 400 600 800 และ 1,000 ดังตารางที่ 4.20

ตารางที่ 4.20 ตารางสรุปวิธีที่มีค่าเฉลี่ยร้อยละความแม่นยำที่สูงที่สุด ตามจำนวนตัวแปรอิสระของชุดข้อมูลดั้งเดิม และชุดข้อมูลที่ปรับให้สมดุลแล้ว โดยแบ่งตามประเภทของตัวแปรตาม

	ประเภทของตัวแปรตาม	จำนวนตัวแปรอิสระ (p)				
		200	400	600	800	1,000
ชุดข้อมูลดั้งเดิม	0	Naïve	Naïve	Naïve	Naïve	**Naïve**
	1	RF	**SVM**	SVM	SVM	SVM
	2	KNN	**SVM**	**SVM**	**SVM**	**SVM**
	3	**KNN ,RF**	**KNN ,SVM, RF**	**Naïve, KNN, SVM, RF**	**Naïve, KNN, SVM, RF**	**Naïve, KNN, SVM, RF**
ชุดข้อมูลที่ปรับให้สมดุลแล้ว	0	RF	RF	**RF**	RF	RF
	1	RF	**SVM**	SVM	SVM	RF
	2	RF	SVM	SVM	SVM	**SVM**
	3	**SVM**	**SVM**	**SVM**	**Naïve, SVM, RF**	**Naïve, KNN, SVM, RF**

หมายเหตุ ตัวหนังสือที่เป็นตัวหนา หมายถึง วิธีที่มีค่าเฉลี่ยร้อยละความแม่นยำที่มากที่สุด ตามประเภทของตัวแปรตาม

จากตารางที่ 4.20 พบว่าชุดข้อมูลดั้งเดิม มีวิธีที่ให้ค่าเฉลี่ยร้อยละความแม่นยำที่สูงที่สุด และมากที่สุด คือ วิธี Naïve ที่ประเภทของตัวแปรตามที่แทนด้วยหมายเลข 0 วิธี SVM ที่ประเภทของ

ตัวแปรตามที่แทนด้วยหมายเลข 1 วิธี SVM ที่ประเภทของตัวแปรตามที่แทนด้วยหมายเลข 2 วิธี Naïve KNN SVM และ RF ที่ประเภทของตัวแปรตามที่แทนด้วยหมายเลข 3

ชุดข้อมูลที่ปรับให้สมดุลแล้ว มีวิธีที่ให้ค่าเฉลี่ยร้อยละความแม่นยำที่สูงที่สุด และมากที่สุด คือ วิธี RF ที่ประเภทของตัวแปรตามที่แทนด้วยหมายเลข 0 วิธี SVM ที่ประเภทของตัวแปรตามที่แทนด้วยหมายเลข 1 วิธี SVM ที่ประเภทของตัวแปรตามที่แทนด้วยหมายเลข 2 วิธี Naïve KNN SVM และ RF ที่ประเภทของตัวแปรตามที่แทนด้วยหมายเลข 3

4.9 ผลสรุปวิธีที่มีค่าเฉลี่ยร้อยละความระลึกที่สูงที่สุด

ผลสรุปวิธีที่มีค่าเฉลี่ยร้อยละความระลึกสูงที่สุด จากการวิเคราะห์ตามประเภทของตัวแปรตาม ทั้ง 4 กลุ่ม ได้แก่

1. Small Cell Lung Cancer (SCLC) : Oat cell lung cancer แทนด้วยหมายเลข 0
2. Non-Small Cell lung cancer (NACLC) : Adenocarcinoma แทนด้วยหมายเลข 1
3. Non-Small Cell lung cancer (NSCLC) : Squamous Cell carcinoma แทนด้วยหมายเลข 2
4. Non-Small Cell lung cancer (NSCLC) : Large Cell lung cancer แทนด้วยหมายเลข 3

ของชุดข้อมูลดั้งเดิม และชุดข้อมูลที่ปรับให้สมดุลแล้ว โดยมีจำนวนตัวแปรอิสระเป็น 200 400 600 800 และ 1,000 ดังตารางที่ 4.21

ตารางที่ 4.21 ตารางสรุปวิธีที่มีค่าเฉลี่ยร้อยละความระลึกที่สูงที่สุด ตามจำนวนตัวแปรอิสระของชุดข้อมูลดั้งเดิม และชุดข้อมูลที่ปรับให้สมดุลแล้ว โดยแบ่งตามประเภทของตัวแปรตาม

	ประเภทของตัวแปรตาม	จำนวนตัวแปรอิสระ (p)				
		200	400	600	800	1,000
ชุดข้อมูลดั้งเดิม	0	RF	**SVM**	SVM	SVM	SVM
	1	**SVM**	Naïve	Naïve	Naïve	Naïve
	2	Naïve	Naïve	Naïve	**Naïve**	Naïve
	3	**SVM**	KNN	RF	RF	RF
ชุดข้อมูลที่ปรับให้สมดุลแล้ว	0	RF	SVM	**SVM**	SVM	SVM
	1	RF	RF	RF	KNN	**KNN**
	2	SVM	RF	**RF**	RF	RF
	3	**Naïve, SVM, RF**	**Naïve, RF**	**Naïve, RF**	**Naïve, RF**	**Naïve, RF**

หมายเหตุ ตัวหนังสือที่เป็นตัวหนา หมายถึง วิธีที่มีค่าเฉลี่ยร้อยละความระลึกที่มากที่สุด ตามประเภทของตัวแปรตาม

จากตารางที่ 4.21 พบว่าชุดข้อมูลดั้งเดิม มีวิธีที่ให้ค่าเฉลี่ยร้อยละความระลึกที่สูงที่สุด และมากที่สุด คือ วิธี SVM ที่ประเภทของตัวแปรตามที่แทนด้วยหมายเลข 0 วิธี SVM มีค่าเฉลี่ยร้อยละความระลึกที่สูงที่สุด และรองลงมาคือวิธี Naïve ที่ประเภทของตัวแปรตามที่แทนด้วยหมายเลข 1 วิธี Naïve ที่ประเภทของตัวแปรตามที่แทนด้วยหมายเลข 2 และวิธี SVM ที่ประเภทของตัวแปรตามที่แทนด้วยหมายเลข 3

ชุดข้อมูลที่ปรับให้สมดุลแล้ว มีวิธีที่ให้ค่าเฉลี่ยร้อยละความระลึกที่สูงที่สุด และมากที่สุด คือ วิธี SVM ที่ประเภทของตัวแปรตามที่แทนด้วยหมายเลข 0 วิธี KNN ที่ประเภทของตัวแปรตามที่แทนด้วยหมายเลข 1 วิธี RF ที่ประเภทของตัวแปรตามที่แทนด้วยหมายเลข 2 และวิธี Naïve SVM และ RF ที่ประเภทของตัวแปรตามที่แทนด้วยหมายเลข 3

บทที่ 5
สรุปผลการวิจัยและข้อเสนอแนะ

ในการทำงานวิจัยครั้งนี้มีวัตถุประสงค์เพื่อเปรียบเทียบวิธีการเรียนรู้ด้วยเครื่องสำหรับการจำแนกระดับการเป็นมะเร็งปอดจากรหัสพันธุกรรม เมื่อมีการจำแนก 6 วิธีคือ วิธีต้นไม้ตัดสินใจ วิธีนาอีฟเบย์ วิธีเพื่อนบ้านใกล้เคียง k อันดับ วิธีซัพพอร์ตเวกเตอร์แมชชีน วิธีโครงข่ายประสาทเทียม และ วิธีป่าสุ่ม โดยจะวัดประสิทธิภาพของวิธีการจำแนกกลุ่มการเป็นโรคมะเร็งปอด จากค่าความถูกต้อง (Accuracy) ค่าความแม่นยำ (Precision) และ ค่าความระลึก (Recall) ซึ่งสรุปผลได้ดังต่อไปนี้

5.1 สรุปผลการวิจัย

จากการเปรียบเทียบประสิทธิภาพในการจำแนกกลุ่มการเป็นโรคมะเร็งปอด โดยวิเคราะห์จากค่าเฉลี่ยร้อยละความถูกต้อง ค่าเฉลี่ยร้อยละความแม่นยำ ค่าเฉลี่ยร้อยละความระลึกของชุดข้อมูลดั้งเดิม และชุดข้อมูลที่ปรับให้สมดุลแล้ว เพื่อให้ได้วิธีการเรียนรู้ด้วยเครื่องที่เหมาะสมตามชุดข้อมูลที่นำมาวิเคราะห์

เมื่อทำการเปรียบเทียบวิธีการเรียนรู้ด้วยเครื่อง โดยการใช้ค่าเฉลี่ยร้อยละความถูกต้องเป็นเกณฑ์ในการวัดประสิทธิภาพการทำนายของตัวแบบ จากการวิเคราะห์โดยรวมของระดับโรคมะเร็งปอดทั้ง 4 กลุ่ม ของชุดข้อมูลดั้งเดิม และชุดข้อมูลที่ปรับให้สมดุลแล้ว สามารถสรุปผลได้ดังตารางที่ 5.1

ตารางที่ 5.1 ตารางสรุปผลการวิจัยของวิธีที่มีค่าเฉลี่ยร้อยละความถูกต้องที่สูงที่สุด

ประเภทของชุดข้อมูล	วิธีการเรียนรู้ด้วยเครื่อง	จำนวนรหัสพันธุกรรม	ค่าเฉลี่ยร้อยละความถูกต้อง
ชุดข้อมูลดั้งเดิม	SVM	200	96.6407
ชุดข้อมูลที่ปรับให้สมดุลแล้ว	RF	600	99.6443

จากตารางที่ 5.1 พบว่า ที่ชุดข้อมูลดั้งเดิม ค่าเฉลี่ยร้อยละความถูกต้องของวิธีซัพพอร์ตเวกเตอร์แมชชีน มีค่าเฉลี่ยร้อยละความถูกต้องสูงที่สุด คือ 96.6407 ที่จำนวนรหัสพันธุกรรมเท่ากับ 200 และที่ชุดข้อมูลที่ปรับให้สมดุลแล้ว วิธีป่าสุ่ม มีค่าเฉลี่ยร้อยละความถูกต้องสูงที่สุด คือ 99.6443 ที่จำนวนรหัสพันธุกรรมเท่ากับ 600 เมื่อนำค่าเฉลี่ยร้อยละความถูกต้องของชุดข้อมูลดั้งเดิมและชุดข้อมูลที่ปรับให้สมดุลแล้วมาเปรียบเทียบผล พบว่าค่าเฉลี่ยร้อยละความถูกต้องของชุดข้อมูลที่ปรับให้สมดุลแล้ว มีค่าสูงขึ้นจากชุดข้อมูลดั้งเดิม นั่นคือ ประสิทธิภาพในการทำนายของตัวแบบดีขึ้น

เมื่อทำการเปรียบเทียบวิธีการเรียนรู้ด้วยเครื่อง โดยการใช้ค่าเฉลี่ยร้อยละความแม่นยำเป็นเกณฑ์ในการวัดประสิทธิภาพในการทำนายว่าเป็นโรคเทียบกับผลการทำนายว่าเป็นโรคจากการวิเคราะห์ในแต่ละระดับของโรคมะเร็งปอดทั้ง 4 กลุ่ม ของชุดข้อมูลดั้งเดิม และชุดข้อมูลที่ปรับให้สมดุลแล้ว สามารถสรุปผลได้ดังตารางที่ 5.2

ตารางที่ 5.2 ตารางสรุปผลการวิจัยของวิธีที่มีค่าเฉลี่ยร้อยละความแม่นยำที่สูงที่สุด

	ประเภทของตัวแปรตาม	วิธีการเรียนรู้ด้วยเครื่อง	จำนวนรหัสพันธุกรรม	ค่าเฉลี่ยร้อยละความแม่นยำ
ชุดข้อมูลดั้งเดิม	Small Cell Lung Cancer (SCLC) : Oat cell lung cancer (0)	Naïve	1,000	98.4383
	Non-Small Cell lung cancer (NSCLC) : Adenocarcinoma (1)	SVM	400	96.386
	Non-Small Cell lung cancer (NSCLC) : Squamous Cell carcinoma (2)	SVM	400 600 800 และ 1,000	100
	Non-Small Cell lung cancer (NSCLC) : Large Cell lung cancer (3)	KNN และ RF KNN SVM และ RF Naïve KNN SVM และ RF	200 400 600 800 และ 1,000	100
ชุดข้อมูลที่ปรับให้สมดุลแล้ว	Small Cell Lung Cancer (SCLC) : Oat cell lung cancer (0)	RF	600	99.9036
	Non-Small Cell lung cancer (NSCLC) : Adenocarcinoma (1)	SVM	400	99.3357
	Non-Small Cell lung cancer (NSCLC) : Squamous Cell carcinoma (2)	SVM	1,000	99.9943
	Non-Small Cell lung cancer (NSCLC) : Large Cell lung cancer (3)	SVM Naïve SVM และ RF Naïve KNN SVM และ RF	200 400 และ 600 800 1,000	100

จากตารางที่ 5.2 วิธีที่มีค่าเฉลี่ยร้อยละความแม่นยำที่สูงที่สุดของชุดข้อมูลดั้งเดิม พบว่า

- กลุ่ม Small Cell Lung Cancer (SCLC) : Oat cell lung cancer

คือวิธีนาอีฟเบย์ มีค่าเฉลี่ยร้อยละความแม่นยำสูงที่สุด คือ 98.4383 ที่จำนวนรหัสพันธุกรรมเท่ากับ 1,000

- กลุ่ม Non-Small Cell lung cancer (NACLC) : Adenocarcinoma

คือวิธีซัพพอร์ตเวกเตอร์แมชชีน มีค่าเฉลี่ยร้อยละความแม่นยำสูงที่สุด คือ 96.3860 ที่จำนวนรหัสพันธุกรรมเท่ากับ 400

- กลุ่ม Non-Small Cell lung cancer (NSCLC): Squamous Cell carcinoma

คือวิธีซัพพอร์ตเวกเตอร์แมชชีน มีค่าเฉลี่ยร้อยละความแม่นยำสูงที่สุด คือ 100 ที่จำนวนรหัสพันธุกรรมเท่ากับ 400 600 800 และ 1,000

- กลุ่ม Non-Small Cell lung cancer (NSCLC) : Large Cell lung cancer มีค่าเฉลี่ยร้อยละความแม่นยำที่สูงที่สุด คือ 100 ที่จำนวนรหัสพันธุกรรมเท่ากับ 200 คือวิธีเพื่อนบ้านใกล้เคียง-k อันดับ และ วิธีป่าสุ่ม ที่จำนวนรหัสพันธุกรรมเท่ากับ 400 คือวิธีเพื่อนบ้านใกล้เคียง k อันดับ วิธีซัพพอร์ตเวกเตอร์แมชชีน และ วิธีป่าสุ่ม ที่จำนวนรหัสพันธุกรรมเท่ากับ 600 800 และ 1,000 คือ วิธีนาอีฟเบย์ วิธีเพื่อนบ้านใกล้เคียง k อันดับ วิธีซัพพอร์ตเวกเตอร์แมชชีน และ วิธีป่าสุ่ม

วิธีที่มีค่าเฉลี่ยร้อยละความแม่นยำที่สูงที่สุดของชุดข้อมูลที่ปรับให้สมดุลแล้ว พบว่า

- กลุ่ม Small Cell Lung Cancer (SCLC) : Oat cell lung cancer คือ วิธีป่าสุ่ม มีค่าเฉลี่ยร้อยละความแม่นยำสูงที่สุด คือ 99.9036 ที่จำนวนรหัสพันธุกรรมเท่ากับ 600

- กลุ่ม Non-Small Cell lung cancer (NACLC) : Adenocarcinoma คือวิธีซัพพอร์ตเวกเตอร์แมชชีน มีค่าเฉลี่ยร้อยละความแม่นยำสูงที่สุด คือ 99.3357 ที่จำนวนรหัสพันธุกรรมเท่ากับ 400

- กลุ่ม Non-Small Cell lung cancer (NSCLC): Squamous Cell คือวิธีซัพพอร์ต-เวกเตอร์แมชชีน มีค่าเฉลี่ยร้อยละความแม่นยำสูงที่สุด คือ 99.9943 ที่จำนวนรหัสพันธุกรรมเท่ากับ 1,000

- กลุ่ม Non-Small Cell lung cancer (NSCLC) : Large Cell lung cancer มีค่าเฉลี่ยร้อยละความแม่นยำที่สูงที่สุด คือ 100 ที่จำนวนรหัสพันธุกรรมเท่ากับ 200 400 และ 600 คือ วิธีซัพพอร์ตเวกเตอร์แมชชีน ที่จำนวนรหัสพันธุกรรมเท่ากับ 800 คือวิธีนาอีฟเบย์ วิธีซัพพอร์ตเวกเตอร์แมชชีน และวิธีป่าสุ่ม ที่จำนวนรหัสพันธุกรรมเท่ากับ 1,000 คือวิธีนาอีฟเบย์ วิธีเพื่อนบ้านใกล้เคียง k อันดับ วิธีซัพพอร์ตเวกเตอร์แมชชีน และวิธีป่าสุ่ม

เมื่อทำการเปรียบเทียบวิธีการเรียนรู้ด้วยเครื่อง โดยการใช้ค่าเฉลี่ยร้อยละความระลึกเป็นเกณฑ์ในการวัดประสิทธิภาพในการทำนายว่าเป็นโรคเทียบกับความจริงทั้งหมด จากการวิเคราะห์ในแต่ละระดับของโรคมะเร็งปอดทั้ง 4 กลุ่ม ของชุดข้อมูลดั้งเดิม และชุดข้อมูลที่ปรับให้สมดุลแล้ว สามารถสรุปผลได้ดังตารางที่ 5.3

ตารางที่ 5.3 ตารางสรุปผลการวิจัยของวิธีที่มีค่าเฉลี่ยร้อยละความระลึกที่สูงที่สุด

	ประเภทของตัวแปรตาม	วิธีการเรียนรู้ด้วยเครื่อง	จำนวนรหัสพันธุกรรม	ค่าเฉลี่ยร้อยละความระลึก
ชุดข้อมูลดั้งเดิม	Small Cell Lung Cancer (SCLC) : Oat cell lung cancer (0)	SVM	400	99.5905
	Non-Small Cell lung cancer (NSCLC) : Adenocarcinoma (1)	SVM	200	95.2077
	Non-Small Cell lung cancer (NSCLC) : Squamous Cell carcinoma (2)	Naïve	800	95.3847
	Non-Small Cell lung cancer (NSCLC) : Large Cell lung cancer (3)	SVM	200	99.9917
ชุดข้อมูลที่ปรับให้สมดุลแล้ว	Small Cell Lung Cancer (SCLC) : Oat cell lung cancer (0)	SVM	600	99.3157
	Non-Small Cell lung cancer (NSCLC) : Adenocarcinoma (1)	KNN	1,000	100
	Non-Small Cell lung cancer (NSCLC) : Squamous Cell carcinoma (2)	RF	600	99.9179
	Non-Small Cell lung cancer (NSCLC) : Large Cell lung cancer (3)	Naïve SVM และ RF Naïve และ RF	200 400 600 800 และ 1,000	100

จากตารางที่ 5.3 วิธีที่มีค่าเฉลี่ยร้อยละความระลึกที่สูงที่สุดของชุดข้อมูลดั้งเดิม พบว่า

- กลุ่ม Small Cell Lung Cancer (SCLC) : Oat cell lung cancer

คือวิธีซัพพอร์ตเวกเตอร์แมชชีน มีค่าเฉลี่ยร้อยละความระลึกสูงที่สุด คือ 99.5905 ที่จำนวนรหัสพันธุกรรมเท่ากับ 400

- กลุ่ม Non-Small Cell lung cancer (NACLC) : Adenocarcinoma

คือวิธีซัพพอร์ตเวกเตอร์แมชชีน มีค่าเฉลี่ยร้อยความระลึกสูงที่สุด คือ 95.2077 ที่จำนวนรหัสพันธุกรรมเท่ากับ 200

- กลุ่ม Non-Small Cell lung cancer (NSCLC): Squamous Cell carcinoma

คือวิธีนาอีฟเบย์ มีค่าเฉลี่ยร้อยละความระลึกสูงที่สุด คือ 95.3847 ที่จำนวนรหัสพันธุกรรมเท่ากับ 800

- กลุ่ม Non-Small Cell lung cancer (NSCLC) : Large Cell lung cancer
คือวิธีซัพพอร์ตเวกเตอร์แมชชีน มีค่าเฉลี่ยร้อยความระลึกสูงที่สุด คือ 99.9917 ที่จำนวนรหัสพันธุกรรมเท่ากับ 200

วิธีที่มีค่าเฉลี่ยร้อยละความระลึกที่สูงที่สุดของชุดข้อมูลที่ปรับให้สมดุลแล้ว พบว่า

- กลุ่ม Small Cell Lung Cancer (SCLC) : Oat cell lung cancer
คือวิธีซัพพอร์ตเวกเตอร์แมชชีน มีค่าเฉลี่ยร้อยละความระลึกสูงที่สุด คือ 99.3157 ที่จำนวนรหัสพันธุกรรมเท่ากับ 600

- กลุ่ม Non-Small Cell lung cancer (NACLC) : Adenocarcinoma
คือวิธีเพื่อนบ้านใกล้เคียง k อันดับ มีค่าเฉลี่ยร้อยความระลึกสูงที่สุด คือ 100 ที่จำนวนรหัสพันธุกรรมเท่ากับ 1,000

- กลุ่ม Non-Small Cell lung cancer (NSCLC): Squamous Cell carcinoma
คือวิธีป่าสุ่ม มีค่าเฉลี่ยร้อยละความระลึกสูงที่สุด คือ 99.9179 ที่จำนวนรหัสพันธุกรรมเท่ากับ 600

- กลุ่ม Non-Small Cell lung cancer (NSCLC) : Large Cell lung cancer
มีค่าเฉลี่ยร้อยละความระลึกที่สูงที่สุด คือ 100 ที่จำนวนรหัสพันธุกรรมเท่ากับ 200 คือวิธีนาอีฟเบย์ วิธีซัพพอร์ตเวกเตอร์แมชชีน และวิธีป่าสุ่ม ที่จำนวนรหัสพันธุกรรมเท่ากับ 400 600 800 และ 1,000 คือวิธีนาอีฟเบย์ และวิธีป่าสุ่ม

5.2 ข้อเสนอแนะ

1. ผู้ที่มีความประสงค์นำไปใช้ในการวิจัย ควรทำการเปรียบเทียบประสิทธิภาพโดยใช้โปรแกรมอื่นๆ นอกเหนือจากโปรแกรมอาร์ในการวิเคราะห์ข้อมูลและการจำแนกข้อมูล เช่น Python เป็นต้น

2. ควรมีการใช้วิธีการจำแนกข้อมูลอื่นๆในการเปรียบเทียบประสิทธิภาพการจำแนกกลุ่มของข้อมูล เช่น Logistic Regression และRule-Based Classification เป็นต้น

3. จำนวนของผู้ป่วยโรคมะเร็งปอด ที่นำมาวิเคราะห์ในปัญหาพิเศษนี้เป็นเพียงส่วนหนึ่งของผู้ป่วยที่เป็นโรคมะเร็งปอด ดังนั้นเพื่อให้ผลการวิเคราะห์ข้อมูลมีประสิทธิภาพมากขึ้น ควรมีการเพิ่มจำนวนของข้อมูลให้มากขึ้น

4. ควรศึกษาวิธีการปรับความไม่สมดุลของข้อมูลด้วยวิธีอื่นๆ เช่น วิธีการสุ่มเกินโดยเทคนิค SMOTE (Synthetic Minority Oversampling Technique) วิธีการสุ่มลด (Under Sampling) และวิธีการสุ่มผสมผสาน (Hybrid Method) เป็นต้น

เอกสารอ้างอิง

กิตติศักดิ์ เพชรรุ่งนภา. 2561. การจำแนกประเภทบทวิจารณ์ของผู้ใช้โมไบล์แอปพลิเคชันเพื่อการสร้างทิคเก็ตสำหรับระบบติดตามปัญหา. วิทยาศาสตรมหาบัณฑิต สาขาวิศวกรรมซอฟต์แวร์. จุฬาลงกรณ์มหาวิทยาลัย

ไกรศักดิ์ เกสร. 2564. วิทยาศาสตร์ข้อมูล. พิมพ์ครั้งที่1. พิษณุโลก. ภาควิชาวิทยาการคอมพิวเตอร์และเทคโนโลยีสารสนเทศ

จิตกานต์ จันทราช, มนทิราลัย ชัยมงคล, รัตนชัย แซ่โง้ว และสายทิพย์ พลอยสัมฤทธิ์. 2561. การเปรียบเทียบประสิทธิภาพในการทำนายผลการจำแนกในกรณีที่มีข้อมูลสูญหายด้วยเทคนิคการทำเหมืองข้อมูล. วิทยาศาสตร์บัณฑิต สาขาสถิติประยุกต์. สถาบันเทคโนโลยีพระจอมเกล้าเจ้าคุณทหารลาดกระบัง

ธรรมนูญ ปัญญาทิพย์, ปนัดดา โพธินาม และคมกริช อ่อนประสงค์. 2565. การเพิ่มประสิทธิภาพจำแนกข้อมูลผลกระทบโควิด-19 ต่อผู้ป่วยมะเร็งตับ. วารสารวิชาการการจัดการเทคโนโลยีมหาวิทยาลัยราชภัฏมหาสารคาม. 2: 66-78.

ธิษณ์ปัณฑา คนโทฉิมพลี, กอบเกียรติ ผ่องพุฒิ และณัฐ มาแจ้ง. 2561. การพยากรณ์ปริมาณน้ำไหลเข้าอ่างเก็บน้ำโดยใช้โครงข่ายประสาทเทียม. 68-91. ในการประชุมวิชาการระดับชาติแห่งประเทศไทย ครั้งที่ 11. นนทบุรี.

นพมาศ อัครจันทโชติ และ ดิเรก พนิตสุภากมล. 2562. การเปรียบเทียบวิธีการแก้ปัญหาข้อมูลไม่สมดุลสำหรับการจำแนกกลุ่มรายได้ของผู้ประกอบการร้านยาประเภท ข.ย.1. 1577-1586. ในการประชุมเสนอผลงานระดับชาติ มหาวิทยาลัยสุโขทัยธรรมาธิราช ครั้งที่ 9. นนทบุรี.

นรินทร์ พนาวาส และนิเวศ จิระวิชิตชัย. 2553. การจำแนกมะเร็งเม็ดเลือดขาวโดยใช้เทคนิคการลดมิติข้อมูลด้วย Chi-square. 7-12. ในการประชุมวิชาการระดับประเทศด้านเทคโนโลยีสารสนเทศ ครั้งที่ 3.

ประยุทธ์ศิลป์ ชัยนาม. 2562. การสร้างแบบจำลองจำแนกกลุ่มผู้ป่วยโรคไตเรื้อรังโดยใช้เทคนิคเหมืองข้อมูลและวิชวลไลเซชั่น. ในการประชุมวิชาการ มหาวิทยาลัยเทคโนโลยีราชมงคลรัตนโกสินทร์ครั้งที่ 4 และการประชุมระดับนานาชาติ มหาวิทยาลัยเทคโนโลยีราชมงคลรัตนโกสินทร์ครั้งที่ 1 “การยกระดับงานวิจัยเพื่อขับเคลื่อนเศรษฐกิจและสังคมอย่างยั่งยืน”. กรุงเทพมหานคร.

พนิดา สมบัติมาก, ภัสสร จันทร์หอม, ศุภกร รัศมี, โอฬาร รุ่งมณีธรรมคุณ และสายชล สินสมบูรณ์ทอง. 2562. การเปรียบเทียบประสิทธิภาพในการจำแนกเมื่อข้อมูลมีค่านอกเกณฑ์ในการทำเหมืองข้อมูล. วารสารวิทยาศาสตร์และเทคโนโลยี. 6: 975-988.

พัชรียา ทองพูล, พิมพ์ชนก จำเรือง และรมย์นลิน บุญฤทธิ์. 2561. การเปรียบเทียบประสิทธิภาพในการทำนายผลการปรับความไม่สมดุลของข้อมูลในการจำแนกด้วยเทคนิคการทำเหมืองข้อมูล. วิทยาศาสตรบัณฑิต สาขาสถิติประยุกต์. สถาบันเทคโนโลยีพระจอมเกล้าเจ้าคุณทหารลาดกระบัง

รักถิ่น เหลาหา. 2553. การพยากรณ์ความเสี่ยงการเกิดโรคมะเร็งปอด โดยใช้ทฤษฎีการทำเหมืองข้อมูล. วิทยาศาสตรมหาบัณฑิต สาขาวิชาเทคโนโลยีสารสนเทศ. มหาวิทยาลัยขอนแก่น

รุ่งโรจน์ บุญมา และนิเวศ จิระวิชิตชัย. 2562. การจำแนกประเภทผู้ป่วยโรคเบาหวานโดยใช้เทคนิคเหมืองข้อมูลและการเลือกคุณลักษณะจากความสัมพันธ์ของข้อมูล. วารสารวิชาซายน์เทคมรภูเก็ต. 2 : 11-19.

รุจิรา ธรรมสมบัติ. 2555. ระบบสนับสนุนการตัดสินใจในการเลือกใช้แพคเกจอินเทอร์เน็ตมือถือโดยใช้ต้นไม้ตัดสินใจ. บริหารธุรกิจมหาบัณฑิต สาขาคอมพิวเตอร์ธุรกิจ. มหาวิทยาลัยราชพฤกษ์

วทัญญุตา นีลาภาตระกูล และชุติมา เบี้ยวไข่มุข. 2562. การศึกษาปัจจัยที่สัมพันธ์กับการตัดสินใจลาออกและการเปรียบเทียบประสิทธิภาพตัวแบบพยากรณ์การลาออกของพนักงานกรณีศึกษาบริษัทประกันภัย. วารสารวิชาการสมาคมสถาบันอุดมศึกษาเอกชนแห่งประเทศไทยฉบับวิทยาศาสตร์และเทคโนโลยี. 1: 46-63.

สมศักดิ์ ศรีสวการย์ และสมัย ศรีสวย. 2563. การวิเคราะห์เหมืองความคิดเห็นโดยใช้เทคนิคการสกัดคำ. วารสารวิชาการประยุกต์ใช้เทคโนโลยีสารสนเทศ. 2: 95-104.

สายชล สินสมบูรณ์ทอง. 2560. การทำเหมืองข้อมูล. เล่มที่ 1. พิมพ์ครั้งที่ 2. กรุงเทพฯ : จามจุรีโปรดักส์.

สายชล สินสมบูรณ์ทอง. 2561. การเปรียบเทียบประสิทธิภาพในการทำนายผลการเป็นโรคเบาหวาน. วารสารวิทยาศาสตร์และเทคโนโลยี. 2: 195-207

สุพัตรา ปัญญาคณ, ธัญลักษณ์ คล้ายสงคราม และธิปไตย พงษ์ศาสตร์. 2560. การศึกษาข่ายงานและการประยุกต์. วิทยาศาสตรบัณฑิต. สารสนเทศสถิติ. มหาวิทยาลัยขอนแก่น

สุภาภรณ์ พัฒนวงศ์ปราการ. 2563. การวิเคราะห์เทคนิคการจำแนกประเภทข้อมูล กรณีศึกษาการทำนายระดับชั้นผู้รับเหมาก่อสร้างสำหรับโครงการก่อสร้างของภาครัฐ. วิทยาศาสตรมหาบัณฑิต สาขาวิชาระบบสารสนเทศเพื่อการจัดการ. มหาวิทยาลัยธรรมมศาสตร์

อกนิษฐ์ ทองจิตร, พูลพงศ์ สุขสว่าง และจตุภัทร เมฆพายัพ. 2562. การพัฒนาวิธีจำแนกประเภทข้อมูลโดยใช้โครงข่ายประสาทเทียมแบบปรับเหมาะผสมผสานการหาค่าเหมาะสมที่สุดแบบกลุ่มอนุภาค สำหรับการจำแนกประเภทกลุ่มเสี่ยงในการเป็นโรคเบาหวาน. วิทยาการวิจัยและวิทยาการปัญญา. 2: 83-97

อัจฉรา แผ้วบาง และสายชล สินสมบูรณ์ทอง. 2563. การปรับความไม่สมดุลของข้อมูลด้วยการจำแนก 5 วิธี. วารสารวิทยาศาสตร์และเทคโนโลยี. 4: 418-435

อัจฉราภรณ์ สุขเพิ่ม, พลภัทร โรจน์นครินทร์, เบญจพร อัคควัฒน์ และวีระ สอิ้ง. 2563. แบบจำลองการวินิจฉัยอัตโนมัติสำหรับความเสี่ยงต่อการเกิดลิ่มเลือดอุดตันในหลอดเลือดดำตามอาการโดยอาศัยการเรียนรู้ของเครื่อง. วิทยาศาสตรมหาบัณฑิต สาขาวิชาวิทยาการข้อมูล. มหาวิทยาลัยศรีนครินทรวิโรฒ

อุกฤษฎ์ ศรีสุข และจารี ทองคำ. 2564. การเปรียบเทียบประสิทธิภาพของเทคนิคเหมืองข้อมูลสำหรับพยากรณ์การเกิดโรค. วารสารวิทยาศาสตร์และเทคโนโลยีมหาวิทยาลัยมหาสารคาม. 2: 157-163.

Agrawal, R, Sewani, R, Delen, D. and Benjamin, B. 2022. A machine learning approach for classifying healthy and infarcted patients using heart rate variabilities derived vector magnitude. Healthcare Analytics. 2: 1-11.

Grandini, M, Bagil, E. and Visani, G. 2020. Metrics for multi-class classification: an overview. A white paper. 2: 1-17.

Hamid, H, Ullah, S , Jawaid, L , Hussain , Saddam , Anwar, C. and Hassan, U. 2022. A Machine Learning in Binary and Multiclassification Results on Imbalanced Heart Disease Data Stream. Journal of Sensors. 2022: 8400622.

Han, J. and M. Kamber. 2001. Data Mining: Concepts and Techniques. San Francisco. CA: Morgan Kaufmann

Khalid, S. and Emrullah, S. 2021. Thyroid Disease Classification Using Machine Learning Algorithms. Journal of Physics: Conference Series. 1: 1-12.

Majumder, A, Gupta, S. and Singh, D. 2022. An Ensemble Heart Disease Prediction Model Bagged with Logistic Regression, Naïve Bayes and K Nearest Neighbour. Journal of Physics: Conference Series. 1: 1-10.

Sun, T, Wang, J, Li, X, Lv, P, Liu, F, Luo, Y, Gao, Q, Zhu, H. and Guo, X. 2013. Comparative Evaluation of Support Vector Machines for Computer Aided Diagnosis of Lung Cancer in CT Based on a Multi-Dimensional Data Set. Computer Methods And Programs in Biomedicine. 3: 519-524.

Swain, D, Mehta, U, Bhatt, A, Patel, H, Patel, K, Mehta, D, Acharya, B, Gerogiannis, V, Kanavos, A. and Manika, S. 2022. A Robust Chronic Kidney Disease Classifier Using Machine Learning. Electronics. 1: 1-14.

ภาคผนวก

คำสั่งโปรแกรม R Studio ที่ใช้ในงานวิจัย การวิเคราะห์ค่าเฉลี่ยร้อยละความถูกต้อง ค่าเฉลี่ยร้อยละความแม่นยำ ค่าเฉลี่ยร้อยละความระลึก ของวิธีต้นไม้ตัดสินใจ วิธีนาอีฟเบย์ วิธีเพื่อนบ้านใกล้เคียง k อันดับ วิธีซัพพอร์ตเวกเตอร์แมชชีน วิธีโครงข่ายประสาทเทียม และวิธีป่าสุ่ม ของชุดข้อมูลดั้งเดิม

```
install.packages("MASS")
install.packages("neuralnet")
install.packages("rpart")
install.packages("class")
install.packages("e1071")
install.packages("randomForest")
##########
set.seed(99)
m=1000  #จำนวน loop
p=200  #จำนวนตัวแปรอิสระ
##########
library("MASS")
library(neuralnet)
library(rpart)
library(class)
library(e1071)
library(caTools)
library(randomForest)
#####################
################################################
#getwd()
lung <- read.csv("C:/Users/USER/OneDrive/เดสก์ท็อป/
lung_Data.CSV",header=TRUE,sep=",",fill=TRUE)
attach(lung)
names(lung)
#View(lung)
##########################################
##########################################
```

```
accuracy1 = c(); precision1_1 = c(); precision1_2 = c(); precision1_3 = c();precision1_4
= c()
recall1_1 = c(); recall1_2  = c(); recall1_3 = c(); recall1_4 = c()
accuracy2 = c(); precision2_1 = c(); precision2_2 = c(); precision2_3 = c();precision2_4
= c()
recall2_1 = c(); recall2_2  = c(); recall2_3 = c(); recall2_4 = c()
accuracy3 = c(); precision3_1 = c(); precision3_2 = c(); precision3_3 = c();precision3_4
= c()
recall3_1 = c(); recall3_2  = c(); recall3_3 = c(); recall3_4 = c()
accuracy4 = c(); precision4_1 = c(); precision4_2 = c(); precision4_3 = c();precision4_4
= c()
recall4_1 = c(); recall4_2  = c(); recall4_3 = c(); recall4_4 = c()
accuracy5 = c(); precision5_1 = c(); precision5_2 = c(); precision5_3 = c();precision5_4
= c()
recall5_1 = c(); recall5_2  = c(); recall5_3 = c(); recall5_4 = c()
accuracy6 = c(); precision6_1 = c(); precision6_2 = c(); precision6_3 = c();precision6_4
= c()
recall6_1 = c(); recall6_2  = c(); recall6_3 = c(); recall6_4 = c()
############################################################
xm2 = data.frame(lung)
xm2
################################
for (j in 1:m) {
  ################################
  xm1 <- xm2[, -1]
  xm = sample(xm1,p,replace = FALSE)
  x = as.matrix(xm)
  y = as.factor(Class)
  lung$Class = as.factor(lung$Class)
    #str(lung)
  n = length(y)
  data = data.frame(y,xm)
```

```
################## KNN
#####################################################################
 a = round(0.7*n)
 train_data1 = sample(1:nrow(data), a)
 train_x1 = xm[train_data1,]
 test_x1  = xm[-train_data1,]
 train_y1 = y[train_data1]
 test_y1 = y[-train_data1]
 knn.pred = knn(train_x1,test_x1, train_y1, k=2) ###### K
 table(knn.pred, test_y1) #เทเบิ้ล
 a1 = as.matrix(table(knn.pred, test_y1))
 diag1 = diag(a1)
 nn = sum(a1)
 rowsums1 = apply(a1,1,sum)
 colsums1 = apply(a1,2,sum)
 #accuracy1 = sum(diag1)/nn
 accuracy1[j]  = sum(diag1)/nn
 #precision1 = diag1/rowsums1
 precision1_1[j] = diag1[1]/rowsums1[1]
 precision1_2[j] = diag1[2]/rowsums1[2]
 precision1_3[j] = diag1[3]/rowsums1[3]
 precision1_4[j] = diag1[4]/rowsums1[4]
 #recall1 = diag1/colsums1
 recall1_1[j] = diag1[1]/colsums1[1]
 recall1_2[j] = diag1[2]/colsums1[2]
 recall1_3[j] = diag1[3]/colsums1[3]
 recall1_4[j] = diag1[4]/colsums1[4]


###################### neural network
################################################
 train_x2 = as.matrix(train_x1)
 test_x2 =  as.matrix(test_x1)
 #train_y1 = as.factor(train$y)
```

```
  train_y2 = as.factor(train_y1)
  test_y2 = as.factor(test_y1)
  ################################################
  t_x1 = train_x2[,1] ;t_x2 = train_x2[,2] ; t_x3 = train_x2[,3];t_x4 = train_x2[,4];t_x5 =
train_x2[,5]
  t_x6 = train_x2[,6] ;t_x7 = train_x2[,7] ; t_x8 = train_x2[,8];t_x9 = train_x2[,9];t_x10 =
train_x2[,10]
  t_x11 = train_x2[,11] ;t_x12 = train_x2[,12] ; t_x13 = train_x2[,13];t_x14 =
train_x2[,14];t_x15 = train_x2[,15]
  t_x16 = train_x2[,16] ;t_x17 = train_x2[,17] ; t_x18 = train_x2[,18];t_x19 =
train_x2[,19];t_x20 = train_x2[,20]
  t_x21 = train_x2[,21] ;t_x22 = train_x2[,22] ; t_x23 = train_x2[,23];t_x24 =
train_x2[,24];t_x25 = train_x2[,25]
  t_x26 = train_x2[,26] ;t_x27 = train_x2[,27] ; t_x28 = train_x2[,28];t_x29 =
train_x2[,29];t_x30 = train_x2[,30]
  ################################################
  data3 = data.frame(train_y2,t_x1, t_x2, t_x3,t_x4, t_x5, t_x6, t_x7, t_x8, t_x9, t_x10,
                     t_x11,t_x12, t_x13,t_x14, t_x15, t_x16, t_x17, t_x18, t_x19, t_x20,
                     t_x21,t_x22, t_x23,t_x24, t_x25, t_x26, t_x27, t_x28, t_x29, t_x30)
  #plot(fit)
  fit = neuralnet(train_y2 ~ t_x1 + t_x2+t_x3+t_x4+t_x5 + t_x6+t_x7+t_x8+t_x9 +
t_x10+
                    t_x11 + t_x12+t_x13+t_x14+t_x15 + t_x16+t_x17+t_x18+t_x19 +
t_x20+
                    t_x21 + t_x22+t_x23+t_x24+t_x25 + t_x26+t_x27+t_x28+t_x29 + t_x30
                  , data = data3, hidden = 200, act.fct = "logistic",linear.output = FALSE)

#####################################################################
################
  te_x1 = test_x2[,1] ;te_x2 = test_x2[,2] ; te_x3 = test_x2[,3];te_x4 = test_x2[,4];te_x5
= test_x2[,5]
  te_x6 = test_x2[,6] ;te_x7 =test_x2[,7] ; te_x8 = test_x2[,8];te_x9 = test_x2[,9];te_x10
= test_x2[,10]
```

```
  te_x11 = test_x2[,11] ;te_x12 = test_x2[,12] ; te_x13 = test_x2[,13];te_x14 =
test_x2[,14];te_x15 = test_x2[,15]
  te_x16 = test_x2[,16] ;te_x17 = test_x2[,17] ; te_x18 = test_x2[,18];te_x19 =
test_x2[,19];te_x20 = test_x2[,20]
  te_x21 = test_x2[,21] ;te_x22 = test_x2[,22] ; te_x23 = test_x2[,23];te_x24 =
test_x2[,24];te_x25 = test_x2[,25]
  te_x26 = test_x2[,26] ;te_x27 = test_x2[,27] ; te_x28 = test_x2[,28];te_x29 =
test_x2[,29];te_x30 = test_x2[,30]
##########################################################################
################
   test2 = data.frame(te_x1, te_x2,te_x3,te_x4, te_x5, te_x6,te_x7,te_x8, te_x9, te_x10,
                      te_x11, te_x12,te_x13,te_x14, te_x15, te_x16,te_x17,te_x18, te_x19,
te_x20,
                      te_x21, te_x22,te_x23,te_x24, te_x25, te_x26,te_x27,te_x28, te_x29,
te_x30)
  mypredict = compute(fit, test2)$net.result
  maxidx <- function(arr){return(which(arr== max(arr)))
  }
  ########################################################
  idx <- apply(mypredict, c(1), maxidx)
  prediction <- c('0', '1', '2', '3')[idx]
  table(prediction, test_y2)
  ########################################################
  ########################################################
  a2 = as.matrix(table(prediction, test_y2))
  a2
  diag2 = diag(a2)
  rowsums2 = apply(a2,1,sum)
  colsums2 = apply(a2,2,sum)
  accuracy2[j]  = mean(test_y2==prediction)
  #accuracy2  = mean(test_y2==prediction)
  #precision2= diag2/rowsums2
  precision2_1[j] = diag2[1]/rowsums2[1]
```

```
precision2_2[j] = diag2[2]/rowsums2[2]
precision2_3[j] = diag2[3]/rowsums2[3]
precision2_4[j] = diag2[4]/rowsums2[4]
#recall2  = diag2/colsums2
recall2_1[j] = diag2[1]/colsums2[1]
recall2_2[j] = diag2[2]/colsums2[2]
recall2_3[j] = diag2[3]/colsums2[3]
recall2_4[j] = diag2[4]/colsums2[4]

### กำหนด y เป็น factor ก่อน Decision Tree
y <- as.factor(data$y)
#str(lung)
train_X_Y <- data[train_data1, ]
test_X_Y <- data[-train_data1, ]

###################### Decision Tree
##############################################
treemodel <- rpart(y ~ ., data = train_X_Y, method = "class")
pred3 <- predict(treemodel, newdata = test_X_Y, type = "class")
table(pred3, test_y1, dnn = c("prediction","actual"))
a3 = as.matrix(table(pred3, test_y1, dnn = c("prediction","actual")))
diag3 = diag(a3)
rowsums3 = rowSums(a3)
colsums3 = colSums(a3)
accuracy3[j] = mean(pred3 == test_y1)
#accuracy3 = mean(pred3 == test_y1)
#precision3 = diag3/rowsums3
precision3_1[j] = diag3[1]/rowsums3[1]
precision3_2[j] = diag3[2]/rowsums3[2]
precision3_3[j] = diag3[3]/rowsums3[3]
precision3_4[j] = diag3[4]/rowsums3[4]
#recall3 = diag3/colsums3
recall3_1[j] = diag3[1]/colsums3[1]
```

```
  recall3_2[j] = diag3[2]/colsums3[2]
  recall3_3[j] = diag3[3]/colsums3[3]
  recall3_4[j] = diag3[4]/colsums3[4]

###################### Naive bayes
#########################################################
  naivebayes1 <- naiveBayes(y~., data = train_X_Y,method = "class")
  pred4 <- predict(naivebayes1, newdata = test_X_Y,type = "class")
  table(pred4, test_y1, dnn = c("prediction","actual"))
  a4 = as.matrix(table(pred4, test_y1, dnn = c("prediction","actual")))
  diag4 = diag(a4)
  rowsums4 = rowSums(a4)
  colsums4 = colSums(a4)
  accuracy4[j] = mean(pred4== test_y1)
  #accuracy4 = mean(pred4== test_y1)
  #precision4 = diag4/rowsums4
  precision4_1[j] = diag4[1]/rowsums4[1]
  precision4_2[j] = diag4[2]/rowsums4[2]
  precision4_3[j] = diag4[3]/rowsums4[3]
  precision4_4[j] = diag4[4]/rowsums4[4]
  #recall4  = diag4/colsums4
  recall4_1[j] = diag4[1]/colsums4[1]
  recall4_2[j] = diag4[2]/colsums4[2]
  recall4_3[j] = diag4[3]/colsums4[3]
  recall4_4[j] = diag4[4]/colsums4[4]

###################### SVM
#############################################################
  support  = svm(y~.,data = train_X_Y,type = 'C-classification',kernel = 'sigmoid')
  #?svm
  pred5 = predict(support, newdata = test_X_Y,type = "class")
  table(pred5, test_y1)
  a5 = as.matrix(table(pred5, test_y1))
```

```
diag5 = diag(a5)
rowsums5 = rowSums(a5)
colsums5 = colSums(a5)
accuracy5[j] = mean(pred5== test_y1)
#accuracy5 = mean(pred5== test_y1)
#precision5 = diag5/rowsums5
precision5_1[j] = diag5[1]/rowsums5[1]
precision5_2[j] = diag5[2]/rowsums5[2]
precision5_3[j] = diag5[3]/rowsums5[3]
precision5_4[j] = diag5[4]/rowsums5[4]
#recall5  = diag5/colsums5
recall5_1[j] = diag5[1]/colsums5[1]
recall5_2[j] = diag5[2]/colsums5[2]
recall5_3[j] = diag5[3]/colsums5[3]
recall5_4[j] = diag5[4]/colsums5[4]

############################ Random Forest
#########################################
train_X_Y$y <- as.factor(train_X_Y$y)
test_X_Y$y <- as.factor(test_X_Y$y)
#str(lung)
RFM = randomForest(y~.,data = train_X_Y,type = 'C-classification')
y_Pred = predict(RFM,test_X_Y)
table(y_Pred,test_y1)
a6 = as.matrix(table(y_Pred,test_y1 ))
diag6 = diag(a6)
rowsums6 = rowSums(a6)
colsums6 = colSums(a6)
accuracy6[j] = mean(y_Pred== test_y1)
#accuracy6 = mean(y_Pred== test_y1)
#precision6 = diag6/rowsums6
precision6_1[j] = diag6[1]/rowsums6[1]
precision6_2[j] = diag6[2]/rowsums6[2]
```

```
  precision6_3[j] = diag6[3]/rowsums6[3]
  precision6_4[j] = diag6[4]/rowsums6[4]
  #recall6  = diag6/colsums6
  recall6_1[j] = diag6[1]/colsums6[1]
  recall6_2[j] = diag6[2]/colsums6[2]
  recall6_3[j] = diag6[3]/colsums6[3]
  recall6_4[j] = diag6[4]/colsums6[4]

#########################################################################
##############
  cat(c("loop :",j),fill=T)
}
#########################################################################
################

cat("KNN:mean_accuracy1 =",mean(na.omit(accuracy1)),"\n",
    "KNN:mean_precision1_1 =",mean(na.omit(precision1_1)),"\n",
    "KNN:mean_precision1_2 =",mean(na.omit(precision1_2)),"\n",
    "KNN:mean_precision1_3 =",mean(na.omit(precision1_3)),"\n",
    "KNN:mean_precision1_4 =",mean(na.omit(precision1_4)),"\n",
    "KNN:mean_recall1_1 =",mean(na.omit(recall1_1)),"\n",
    "KNN:mean_recall1_2 =",mean(na.omit(recall1_2)),"\n",
    "KNN:mean_recall1_3 =",mean(na.omit(recall1_3)),"\n",
    "KNN:mean_recall1_4 =",mean(na.omit(recall1_4)),"\n")

cat("ANN:mean_accuracy2 =",mean(na.omit(accuracy2)),"\n",
    "ANN:mean_precision2_1 =",mean(na.omit(precision2_1)),"\n",
    "ANN:mean_precision2_2 =",mean(na.omit(precision2_2)),"\n",
    "ANN:mean_precision2_3 =",mean(na.omit(precision2_3)),"\n",
    "ANN:mean_precision2_4 =",mean(na.omit(precision2_4)),"\n",
    "ANN:mean_recall2_1 =",mean(na.omit(recall2_1)),"\n",
    "ANN:mean_recall2_2 =",mean(na.omit(recall2_2)),"\n",
    "ANN:mean_recall2_3 =",mean(na.omit(recall2_3)),"\n",
```

```
    "ANN:mean_recall2_4 =",mean(na.omit(recall2_4)),"\n")

cat("Decision Tree:mean_accuracy3 =",mean(na.omit(accuracy3)),"\n",
    "Decision Tree:mean_precision3_1 =",mean(na.omit(precision3_1)),"\n",
    "Decision Tree:mean_precision3_2 =",mean(na.omit(precision3_2)),"\n",
    "Decision Tree:mean_precision3_3 =",mean(na.omit(precision3_3)),"\n",
    "Decision Tree:mean_precision3_4 =",mean(na.omit(precision3_4)),"\n",
    "Decision Tree:mean_recall3_1 =",mean(na.omit(recall3_1)),"\n",
    "Decision Tree:mean_recall3_2 =",mean(na.omit(recall3_2)),"\n",
    "Decision Tree:mean_recall3_3 =",mean(na.omit(recall3_3)),"\n",
    "Decision Tree:mean_recall3_4 =",mean(na.omit(recall3_4)),"\n")

cat("Naive Bayes:mean_accuracy4 =",mean(na.omit(accuracy4)),"\n",
    "Naive Bayes:mean_precision4_1 =",mean(na.omit(precision4_1)),"\n",
    "Naive Bayes:mean_precision4_2 =",mean(na.omit(precision4_2)),"\n",
    "Naive Bayes:mean_precision4_3 =",mean(na.omit(precision4_3)),"\n",
    "Naive Bayes:mean_precision4_4 =",mean(na.omit(precision4_4)),"\n",
    "Naive Bayes:mean_recall4_1 =",mean(na.omit(recall4_1)),"\n",
    "Naive Bayes:mean_recall4_2 =",mean(na.omit(recall4_2)),"\n",
    "Naive Bayes:mean_recall4_3 =",mean(na.omit(recall4_3)),"\n",
    "Naive Bayes:mean_recall4_4 =",mean(na.omit(recall4_4)),"\n")

cat("SVM:mean_accuracy5 =",mean(na.omit(accuracy5)),"\n",
    "SVM:mean_precision5_1 =",mean(na.omit(precision5_1)),"\n",
    "SVM:mean_precision5_2 =",mean(na.omit(precision5_2)),"\n",
    "SVM:mean_precision5_3 =",mean(na.omit(precision5_3)),"\n",
    "SVM:mean_precision5_4 =",mean(na.omit(precision5_4)),"\n",
    "SVM:mean_recall5_1 =",mean(na.omit(recall5_1)),"\n",
    "SVM:mean_recall5_2 =",mean(na.omit(recall5_2)),"\n",
    "SVM:mean_recall5_3 =",mean(na.omit(recall5_3)),"\n",
    "SVM:mean_recall5_4 =",mean(na.omit(recall5_4)),"\n")
```



```
cat("Random Forest:mean_accuracy6 =",mean(na.omit(accuracy6)),"\n",
```

```
    "Random Forest:mean_precision6_1 =",mean(na.omit(precision6_1)),"\n",
    "Random Forest:mean_precision6_2 =",mean(na.omit(precision6_2)),"\n",
    "Random Forest:mean_precision6_3 =",mean(na.omit(precision6_3)),"\n",
    "Random Forest:mean_precision6_4 =",mean(na.omit(precision6_4)),"\n",
    "Random Forest:mean_recall6_1 =",mean(na.omit(recall6_1)),"\n",
    "Random Forest:mean_recall6_2 =",mean(na.omit(recall6_2)),"\n",
    "Random Forest:mean_recall6_3 =",mean(na.omit(recall6_3)),"\n",
    "Random Forest:mean_recall6_4 =",mean(na.omit(recall6_4)),"\n")

#####################################################################
```

คำสั่งโปรแกรม R Studio ที่ใช้ในงานวิจัย การวิเคราะห์ค่าเฉลี่ยร้อยละความถูกต้อง ค่าเฉลี่ยร้อยละความแม่นยำ ค่าเฉลี่ยร้อยละความระลึก ของวิธีต้นไม้ตัดสินใจ วิธีนาอีฟเบย์ วิธีเพื่อนบ้านใกล้เคียง k อันดับ วิธีซัพพอร์ตเวกเตอร์แมชชีน วิธีโครงข่ายประสาทเทียม และวิธีป่าสุ่ม ของชุดข้อมูลที่ปรับให้สมดุลแล้ว

```
set.seed(99)
m=1000  #loop
p=200  #จำนวนตัวแปรอิสระ
######################
library("MASS")
library(neuralnet)
library(rpart)
library(class)
library(e1071)
library(caTools)
library(randomForest)
################################################
################################################
#getwd()
lung <- read.csv("C:/Users/USER/OneDrive/เดสก์ท็อป/
lung_Data.CSV",header=TRUE,sep=",",fill=TRUE)
attach(lung)
```

```
names(lung)
#View(lung)
##########################################
##########################################
accuracy1 = c(); precision1_1 = c(); precision1_2 = c(); precision1_3 = c();precision1_4
= c()
recall1_1 = c(); recall1_2  = c(); recall1_3 = c(); recall1_4 = c()
accuracy2 = c(); precision2_1 = c(); precision2_2 = c(); precision2_3 = c();precision2_4
= c()
recall2_1 = c(); recall2_2  = c(); recall2_3 = c(); recall2_4 = c()
accuracy3 = c(); precision3_1 = c(); precision3_2 = c(); precision3_3 = c();precision3_4
= c()
recall3_1 = c(); recall3_2  = c(); recall3_3 = c(); recall3_4 = c()
accuracy4 = c(); precision4_1 = c(); precision4_2 = c(); precision4_3 = c();precision4_4
= c()
recall4_1 = c(); recall4_2  = c(); recall4_3 = c(); recall4_4 = c()
accuracy5 = c(); precision5_1 = c(); precision5_2 = c(); precision5_3 = c();precision5_4
= c()
recall5_1 = c(); recall5_2  = c(); recall5_3 = c(); recall5_4 = c()
accuracy6 = c(); precision6_1 = c(); precision6_2 = c(); precision6_3 = c();precision6_4
= c()
recall6_1 = c(); recall6_2  = c(); recall6_3 = c(); recall6_4 = c()
##########################################################
xm2 = data.frame(lung)
xm2
y1 = xm2$Class
table(y1)
# y1
#  0   1   2   3
#139  17  21  20
#################################
for (j in 1:m) {
  ################################
```

```
xm1 <- xm2[, -1]
xm = sample(xm1,p,replace = FALSE)
x = as.matrix(xm)
#y = Class
y = as.factor(Class)
n = nrow(data)
lung$Class = as.factor(lung$Class) #กำหนด Factor ********
#str(lung)
data = data.frame(y,xm)
data0 = data[y==0,]
data1 = data[y==1,]
data2 = data[y==2,]
data3 = data[y==3,]
#####################################################
data1 = data[y==1,]
data11 =   sample(nrow(data1), 122, replace = T)
select1 = data1[data11,]
#nrow(select1)
total1 = rbind(data1,select1)
#nrow(total1)
#####################################################
data2 = data[y==2,]
data22 =   sample(nrow(data2), 118, replace = T)
select2 = data2[data22,]
#nrow(select2)
total2 = rbind(data2,select2)
#nrow(total2)
#####################################################
data3 = data[y==3,]
data33 =   sample(nrow(data3), 119, replace = T)
select3 = data3[data33,]
total3 = rbind(data3,select3)
##################KNN#################################
```

```
  data = rbind(data0, total1, total2, total3)
  #nrow(data)
  data5 <- data[, -1]
  data_x = as.matrix(data5)
  #nrow(data_x)
  data_y = data$y
  #length(data_y)
################## KNN
#####################################################################
  a = round(0.7*n)
  train_data1 = sample(1:nrow(data), a)
  train_x1 = data_x[train_data1,]
  #length(train_x1)
  test_x1  = data_x[-train_data1,]
  #length(test_x1)
  train_y1 = data_y[train_data1]
  #length(train_y1)
  test_y1 = data_y[-train_data1]
  #length(test_y1)
#######################################################
  knn.pred = knn(train_x1,test_x1,train_y1, k=2) ###### K
  table(knn.pred, test_y1) #เทเบิ้ล
  a1 = as.matrix(table(knn.pred, test_y1))
  diag1 = diag(a1)
  nn = sum(a1)
  rowsums1 = apply(a1,1,sum)
  colsums1 = apply(a1,2,sum)
  accuracy1[j]  = sum(diag1)/nn
  #accuracy1  = sum(diag1)/nn
  #precision1 = diag1/rowsums1
  precision1_1[j] = diag1[1]/rowsums1[1]
  precision1_2[j] = diag1[2]/rowsums1[2]
  precision1_3[j] = diag1[3]/rowsums1[3]
```

```
  precision1_4[j] = diag1[4]/rowsums1[4]
  #recall1   = diag1/colsums1
  recall1_1[j] = diag1[1]/colsums1[1]
  recall1_2[j] = diag1[2]/colsums1[2]
  recall1_3[j] = diag1[3]/colsums1[3]
  recall1_4[j] = diag1[4]/colsums1[4]
```

```
###################### neural network
################################################
```

```
  train_x2 = as.matrix(train_x1)
  test_x2 =  as.matrix(test_x1)
  #train_y1 = as.factor(train$y)
  train_y2 = as.factor(train_y1)
  test_y2 = as.factor(test_y1)
  ################################################
  t_x1 = train_x2[,1] ;t_x2 = train_x2[,2] ; t_x3 = train_x2[,3];t_x4 = train_x2[,4];t_x5 =
train_x2[,5]
  t_x6 = train_x2[,6] ;t_x7 = train_x2[,7] ; t_x8 = train_x2[,8];t_x9 = train_x2[,9];t_x10 =
train_x2[,10]
  t_x11 = train_x2[,11] ;t_x12 = train_x2[,12] ; t_x13 = train_x2[,13];t_x14 =
train_x2[,14];t_x15 = train_x2[,15]
  t_x16 = train_x2[,16] ;t_x17 = train_x2[,17] ; t_x18 = train_x2[,18];t_x19 =
train_x2[,19];t_x20 = train_x2[,20]
  t_x21 = train_x2[,21] ;t_x22 = train_x2[,22] ; t_x23 = train_x2[,23];t_x24 =
train_x2[,24];t_x25 = train_x2[,25]
  t_x26 = train_x2[,26] ;t_x27 = train_x2[,27] ; t_x28 = train_x2[,28];t_x29 =
train_x2[,29];t_x30 = train_x2[,30]
  ################################################
  data3 = data.frame(train_y2,t_x1, t_x2, t_x3,t_x4, t_x5, t_x6, t_x7, t_x8, t_x9, t_x10,
                     t_x11,t_x12, t_x13,t_x14, t_x15, t_x16, t_x17, t_x18, t_x19, t_x20,
                     t_x21,t_x22, t_x23,t_x24, t_x25, t_x26, t_x27, t_x28, t_x29, t_x30)
```

```
  #plot(fit)
  fit = neuralnet(train_y2 ~ t_x1 + t_x2+t_x3+t_x4+t_x5 + t_x6+t_x7+t_x8+t_x9 +
t_x10+
                  t_x11 + t_x12+t_x13+t_x14+t_x15 + t_x16+t_x17+t_x18+t_x19 +
t_x20+
                  t_x21 + t_x22+t_x23+t_x24+t_x25 + t_x26+t_x27+t_x28+t_x29 + t_x30
                , data = data3, hidden = 200, act.fct = "logistic",linear.output = FALSE)

#####################################################################
################
  te_x1 = test_x2[,1] ;te_x2 = test_x2[,2] ; te_x3 = test_x2[,3];te_x4 = test_x2[,4];te_x5
= test_x2[,5]
  te_x6 = test_x2[,6] ;te_x7 =test_x2[,7] ; te_x8 = test_x2[,8];te_x9 = test_x2[,9];te_x10
= test_x2[,10]
  te_x11 = test_x2[,11] ;te_x12 = test_x2[,12] ; te_x13 = test_x2[,13];te_x14 =
test_x2[,14];te_x15 = test_x2[,15]
  te_x16 = test_x2[,16] ;te_x17 = test_x2[,17] ; te_x18 = test_x2[,18];te_x19 =
test_x2[,19];te_x20 = test_x2[,20]
  te_x21 = test_x2[,21] ;te_x22 = test_x2[,22] ; te_x23 = test_x2[,23];te_x24 =
test_x2[,24];te_x25 = test_x2[,25]
  te_x26 = test_x2[,26] ;te_x27 = test_x2[,27] ; te_x28 = test_x2[,28];te_x29 =
test_x2[,29];te_x30 = test_x2[,30]

#####################################################################
######
  #test_y2 = as.factor(test$y)
  #test_y2 = as.factor(test_data1[,1])

  test2 = data.frame(te_x1, te_x2,te_x3,te_x4, te_x5, te_x6,te_x7,te_x8, te_x9, te_x10,
                     te_x11, te_x12,te_x13,te_x14, te_x15, te_x16,te_x17,te_x18, te_x19,
te_x20,
```

```
                    te_x21, te_x22,te_x23,te_x24, te_x25, te_x26,te_x27,te_x28, te_x29,
te_x30)
  mypredict = compute(fit, test2)$net.result
  maxidx <- function(arr){return(which(arr== max(arr)))
  }
  ##################################################
  idx <- apply(mypredict, c(1), maxidx)
  prediction <- c('0', '1', '2', '3')[idx]
  table(prediction, test_y2)
  ###################################3
  ############################################################
  a2 = as.matrix(table(prediction, test_y2))
  a2
  diag2 = diag(a2)
  rowsums2 = apply(a2,1,sum)
  colsums2 = apply(a2,2,sum)
  accuracy2[j]  = mean(test_y2==prediction)
  #accuracy2  = mean(test_y2==prediction)
  #precision2= diag2/rowsums2
  precision2_1[j] = diag2[1]/rowsums2[1]
  precision2_2[j] = diag2[2]/rowsums2[2]
  precision2_3[j] = diag2[3]/rowsums2[3]
  precision2_4[j] = diag2[4]/rowsums2[4]
  #recall2  = diag2/colsums2
  recall2_1[j] = diag2[1]/colsums2[1]
  recall2_2[j] = diag2[2]/colsums2[2]
  recall2_3[j] = diag2[3]/colsums2[3]
  recall2_4[j] = diag2[4]/colsums2[4]

  ### กำหนด y เป็น factor ก่อน Decision Tree ###############
  y <- as.factor(data$y)
  #str(lung)
  train_X_Y <- data[train_data1, ]
```

```
  test_X_Y <- data[-train_data1, ]

###################### Decision Tree
################################################
  treemodel <- rpart(y ~ ., data = train_X_Y, method = "class")
  pred3 <- predict(treemodel, newdata = test_X_Y, type = "class")
  table(pred3, test_y1, dnn = c("prediction","actual"))
  a3 = as.matrix(table(pred3, test_y1, dnn = c("prediction","actual")))
  diag3 = diag(a3)
  rowsums3 = rowSums(a3)
  colsums3 = colSums(a3)
  accuracy3[j] = mean(pred3 == test_y1)
  #accuracy3 = mean(pred3 == test_y1)
  #precision3 = diag3/rowsums3
  precision3_1[j] = diag3[1]/rowsums3[1]
  precision3_2[j] = diag3[2]/rowsums3[2]
  precision3_3[j] = diag3[3]/rowsums3[3]
  precision3_4[j] = diag3[4]/rowsums3[4]
  #recall3 = diag3/colsums3
  recall3_1[j] = diag3[1]/colsums3[1]
  recall3_2[j] = diag3[2]/colsums3[2]
  recall3_3[j] = diag3[3]/colsums3[3]
  recall3_4[j] = diag3[4]/colsums3[4]

###################### Naive bayes
################################################
  naivebayes1 <- naiveBayes(y~., data = train_X_Y,method = "class")
  pred4 <- predict(naivebayes1, newdata = test_X_Y,type = "class")
  table(pred4, test_y1, dnn = c("prediction","actual"))
  a4 = as.matrix(table(pred4, test_y1, dnn = c("prediction","actual")))
  diag4 = diag(a4)
  rowsums4 = rowSums(a4)
  colsums4 = colSums(a4)
```

```
accuracy4[j] = mean(pred4== test_y1)
#accuracy4 = mean(pred4== test_y1)
#precision4 = diag4/rowsums4
precision4_1[j] = diag4[1]/rowsums4[1]
precision4_2[j] = diag4[2]/rowsums4[2]
precision4_3[j] = diag4[3]/rowsums4[3]
precision4_4[j] = diag4[4]/rowsums4[4]
#recall4  = diag4/colsums4
recall4_1[j] = diag4[1]/colsums4[1]
recall4_2[j] = diag4[2]/colsums4[2]
recall4_3[j] = diag4[3]/colsums4[3]
recall4_4[j] = diag4[4]/colsums4[4]

###################### SVM
############################################################
support  = svm(y~.,data = train_X_Y,type = 'C-classification',kernel = 'sigmoid')
#??svm
pred5 = predict(support, newdata = test_X_Y,type = "class")
table(pred5, test_y1)
a5 = as.matrix(table(pred5, test_y1))
diag5 = diag(a5)
rowsums5 = rowSums(a5)
colsums5 = colSums(a5)
accuracy5[j] = mean(pred5== test_y1)
#accuracy5 = mean(pred5== test_y1)
#precision5 = diag5/rowsums5
precision5_1[j] = diag5[1]/rowsums5[1]
precision5_2[j] = diag5[2]/rowsums5[2]
precision5_3[j] = diag5[3]/rowsums5[3]
precision5_4[j] = diag5[4]/rowsums5[4]
#recall5  = diag5/colsums5
recall5_1[j] = diag5[1]/colsums5[1]
recall5_2[j] = diag5[2]/colsums5[2]
```

```
  recall5_3[j] = diag5[3]/colsums5[3]
  recall5_4[j] = diag5[4]/colsums5[4]

########################## Random Forest
##########################################
  train_X_Y$y <- as.factor(train_X_Y$y)
  test_X_Y$y <- as.factor(test_X_Y$y)
  #str(lung)
  RFM = randomForest(y~.,data = train_X_Y,type = 'C-classification')
  y_Pred = predict(RFM,test_X_Y)
  table(y_Pred,test_y1)
  a6 = as.matrix(table(y_Pred,test_y1 ))
  diag6 = diag(a6)
  rowsums6 = rowSums(a6)
  colsums6 = colSums(a6)
  accuracy6[j] = mean(y_Pred== test_y1)
  #accuracy6 = mean(y_Pred== test_y1)
  #precision6 = diag6/rowsums6
  precision6_1[j] = diag6[1]/rowsums6[1]
  precision6_2[j] = diag6[2]/rowsums6[2]
  precision6_3[j] = diag6[3]/rowsums6[3]
  precision6_4[j] = diag6[4]/rowsums6[4]
  #recall6  = diag6/colsums6
  recall6_1[j] = diag6[1]/colsums6[1]
  recall6_2[j] = diag6[2]/colsums6[2]
  recall6_3[j] = diag6[3]/colsums6[3]
  recall6_4[j] = diag6[4]/colsums6[4]

##########################################################################
##############
  cat(c("loop :",j),fill=T)
}
##########################################################
```

```
cat("KNN:mean_accuracy1 =",mean(na.omit(accuracy1)),"\n",
    "KNN:mean_precision1_1 =",mean(na.omit(precision1_1)),"\n",
    "KNN:mean_precision1_2 =",mean(na.omit(precision1_2)),"\n",
    "KNN:mean_precision1_3 =",mean(na.omit(precision1_3)),"\n",
    "KNN:mean_precision1_4 =",mean(na.omit(precision1_4)),"\n",
    "KNN:mean_recall1_1 =",mean(na.omit(recall1_1)),"\n",
    "KNN:mean_recall1_2 =",mean(na.omit(recall1_2)),"\n",
    "KNN:mean_recall1_3 =",mean(na.omit(recall1_3)),"\n",
    "KNN:mean_recall1_4 =",mean(na.omit(recall1_4)),"\n")

cat("ANN:mean_accuracy2 =",mean(na.omit(accuracy2)),"\n",
    "ANN:mean_precision2_1 =",mean(na.omit(precision2_1)),"\n",
    "ANN:mean_precision2_2 =",mean(na.omit(precision2_2)),"\n",
    "ANN:mean_precision2_3 =",mean(na.omit(precision2_3)),"\n",
    "ANN:mean_precision2_4 =",mean(na.omit(precision2_4)),"\n",
    "ANN:mean_recall2_1 =",mean(na.omit(recall2_1)),"\n",
    "ANN:mean_recall2_2 =",mean(na.omit(recall2_2)),"\n",
    "ANN:mean_recall2_3 =",mean(na.omit(recall2_3)),"\n",
    "ANN:mean_recall2_4 =",mean(na.omit(recall2_4)),"\n")

cat("Decision Tree:mean_accuracy3 =",mean(na.omit(accuracy3)),"\n",
    "Decision Tree:mean_precision3_1 =",mean(na.omit(precision3_1)),"\n",
    "Decision Tree:mean_precision3_2 =",mean(na.omit(precision3_2)),"\n",
    "Decision Tree:mean_precision3_3 =",mean(na.omit(precision3_3)),"\n",
    "Decision Tree:mean_precision3_4 =",mean(na.omit(precision3_4)),"\n",
    "Decision Tree:mean_recall3_1 =",mean(na.omit(recall3_1)),"\n",
    "Decision Tree:mean_recall3_2 =",mean(na.omit(recall3_2)),"\n",
    "Decision Tree:mean_recall3_3 =",mean(na.omit(recall3_3)),"\n",
    "Decision Tree:mean_recall3_4 =",mean(na.omit(recall3_4)),"\n")

cat("Naive Bayes:mean_accuracy4 =",mean(na.omit(accuracy4)),"\n",
    "Naive Bayes:mean_precision4_1 =",mean(na.omit(precision4_1)),"\n",
```

```
    "Naive Bayes:mean_precision4_2 =",mean(na.omit(precision4_2)),"\n",
    "Naive Bayes:mean_precision4_3 =",mean(na.omit(precision4_3)),"\n",
    "Naive Bayes:mean_precision4_4 =",mean(na.omit(precision4_4)),"\n",
    "Naive Bayes:mean_recall4_1 =",mean(na.omit(recall4_1)),"\n",
    "Naive Bayes:mean_recall4_2 =",mean(na.omit(recall4_2)),"\n",
    "Naive Bayes:mean_recall4_3 =",mean(na.omit(recall4_3)),"\n",
    "Naive Bayes:mean_recall4_4 =",mean(na.omit(recall4_4)),"\n")

cat("SVM:mean_accuracy5 =",mean(na.omit(accuracy5)),"\n",
    "SVM:mean_precision5_1 =",mean(na.omit(precision5_1)),"\n",
    "SVM:mean_precision5_2 =",mean(na.omit(precision5_2)),"\n",
    "SVM:mean_precision5_3 =",mean(na.omit(precision5_3)),"\n",
    "SVM:mean_precision5_4 =",mean(na.omit(precision5_4)),"\n",
    "SVM:mean_recall5_1 =",mean(na.omit(recall5_1)),"\n",
    "SVM:mean_recall5_2 =",mean(na.omit(recall5_2)),"\n",
    "SVM:mean_recall5_3 =",mean(na.omit(recall5_3)),"\n",
    "SVM:mean_recall5_4 =",mean(na.omit(recall5_4)),"\n")

cat("Random Forest:mean_accuracy6 =",mean(na.omit(accuracy6)),"\n",
    "Random Forest:mean_precision6_1 =",mean(na.omit(precision6_1)),"\n",
    "Random Forest:mean_precision6_2 =",mean(na.omit(precision6_2)),"\n",
    "Random Forest:mean_precision6_3 =",mean(na.omit(precision6_3)),"\n",
    "Random Forest:mean_precision6_4 =",mean(na.omit(precision6_4)),"\n",
    "Random Forest:mean_recall6_1 =",mean(na.omit(recall6_1)),"\n",
    "Random Forest:mean_recall6_2 =",mean(na.omit(recall6_2)),"\n",
    "Random Forest:mean_recall6_3 =",mean(na.omit(recall6_3)),"\n",
    "Random Forest:mean_recall6_4 =",mean(na.omit(recall6_4)),"\n")
```

คำสั่งที่ใช้สร้างกราฟความสัมพันธ์ของข้อมูล และ แผนภาพกล่อง (Boxplot)

```
install.packages("corrplot")
install.packages("RColorBrewer")
############################
library(corrplot)
library(RColorBrewer)
#getwd()
lung <- read.csv("C:/Users/USER/OneDrive/เดสก์ท็อป/
lung_Data.CSV",header=TRUE,sep=",",fill=TRUE)
attach(lung)
names(lung)
############################
xm1 = data.frame(lung)
############################
p=10
xm = sample(xm1,p,replace = FALSE)
M = cor(xm)
#########################################################
corrplot(M, type = "upper", order = "hclust", col = brewer.pal(n=9, name = "RdYlBu"))
#แสดงความสัมพันธ์ของตัวแปรอิสระ
boxplot(xm) #แสดงBoxplot โดยรวม
```

คำสั่งที่ใช้สร้าง แผนภาพกล่อง (Boxplot) แบบเจาะจงตัวแปร

```
install.packages("corrplot")
install.packages("RColorBrewer")
install.packages("dplyr") #เพิ่มBoxplotแต่ละอัน

library(corrplot)
library(RColorBrewer)
library(dplyr)
#getwd()
```

```
lung <- read.csv("C:/Users/USER/OneDrive/เดสก์ท็อป/
lung_Data.CSV",header=TRUE,sep=",",fill=TRUE)
attach(lung)
names(lung)
##########################################
par(mfrow=c(2,5))
##########################################
xm1 = data.frame(lung)
########## Boxplot แยกกล่อ ตัวอย่างตัวแปรอิสระ 10 ตัว ###############
#View(xm1)
sample1 = xm1 %>% select(x115)
sample2 = xm1 %>% select(x127)
sample3 = xm1 %>% select(x128)
sample4 = xm1 %>% select(x197)
sample5 = xm1 %>% select(x468)
sample6 = xm1 %>% select(x529)
sample7 = xm1 %>% select(x588)
sample8 = xm1 %>% select(x788)
sample9 = xm1 %>% select(x794)
sample10 = xm1 %>% select(x854)
### boxplot ###################################
boxplot(sample1,main="x115")
boxplot(sample2,main="x127")
boxplot(sample3,main="x128")
boxplot(sample4,main="x197")
boxplot(sample5,main="x468")
boxplot(sample6,main="x529")
boxplot(sample7,main="x588")
boxplot(sample8,main="x788")
boxplot(sample9,main="x794")
boxplot(sample10,main="x854")
##########################################################################
```

คำสั่งที่ใช้สร้างกราฟเปรียบเทียบค่าเฉลี่ยร้อยละความถูกต้อง ของวิธีการเรียนรู้ด้วยเครื่อง 6 วิธี ของชุดข้อมูลดั้งเดิม และชุดข้อมูลที่ปรับให้สมดุลแล้ว

```
par(mfrow=c(1,2))
########################################## Acc Original ##############
DT = c(86.6153,87.3814,86.9475,86.9390,86.7271)
Naive = c(95.8831,96.2881,96.4593,96.3848,96.5051)
KNN = c(94.366,94.6220,94.5915,93.5000,93.9695)
SVM = c(96.6407,92.5017,92.4949,92.6390,92.5661)
ANN = c(88.9763,89.0542,89.0729,89.3017,89.0780)
RF = c(94.4881,94.5170,94.6305,94.7576,94.7034)
no = c(200,400,600,800,1000)

plot(no, DT , type = 'b',main = "Accuracy of Original Data", lty=3,lwd=2.9,col=
"purple", ylim=c(80,100),
    xlab = 'Number of Variable', ylab = 'Percentage of Accuracy',pch = 15,cex=1)

lines(no,Naive, type = "b",lty=2,lwd=2.9,col = "blue",pch = 10,cex=1)
lines(no,KNN, type = "b",lty=2,lwd=2.9,col = "green",pch = 11,cex=1)
lines(no,SVM, type = "b",lty=2,lwd=2.9,col = "orange",pch = 8,cex=1)
lines(no,ANN, type = "b",lty=2,lwd=2.9,col = "red",pch = 5,cex=1)
lines(no,RF, type = "b",lty=2,lwd=2.9,col = "deeppink",pch = 17,cex=1)

labels = c("DT","Naive","KNN","SVM","ANN","RF")
colors = c("purple","blue","green","orange","red","deeppink")
pchh = c(15,10,11,8,5,17)
legend("bottomright",inset = .02,labels,lwd=2,lty=1,col=colors,pch=pchh, cex=0.7)


########################################## Acc Balance#################
DT = c(96.2371,96.5683,96.6641,96.5773,96.5072)
Naive = c(98.3479,98.5599,98.6701,98.6425,98.6778)
KNN = c(94.1395,95.2114,95.6665,96.2635,96.7755)
```

```
SVM = c(99.0784,98.5246,98.5976,98.6737,98.6737)
ANN = c(98.1108,98.1114,98.2976,98.2222,98.2419)
RF = c(99.5982,99.6042,99.6443,99.6084,99.6156)
no = c(200,400,600,800,1000)

plot(no, DT , type = 'b',main =
       "Accuracy of Balanced Data", lty=3,lwd=2.9,col= "purple", ylim=c(80,100),
     xlab = 'Number of Variable', ylab = 'Percentage of Accuracy',pch = 15,cex=1)
lines(no,Naive, type = "b",lty=2,lwd=2.9,col = "blue",pch = 10,cex=1)
lines(no,KNN, type = "b",lty=2,lwd=2.9,col = "green",pch = 11,cex=1)
lines(no,SVM, type = "b",lty=2,lwd=2.9,col = "orange",pch = 8,cex=1)
lines(no,ANN, type = "b",lty=2,lwd=2.9,col = "red",pch = 5,cex=1)
lines(no,RF, type = "b",lty=2,lwd=2.9,col = "deeppink",pch = 17,cex=1)

labels = c("DT","Naive","KNN","SVM","ANN","RF")
colors = c("purple","blue","green","orange","red","deeppink")
pchh = c(15,10,11,8,5,17)
legend("bottomright",inset = .02,labels,lwd=2,lty=1,col=colors,pch=pchh, cex=0.7)
#####################################################################
```

คำสั่งที่ใช้สร้างกราฟเปรียบเทียบค่าเฉลี่ยร้อยละความถูกต้องของชุดข้อมูลดั้งเดิม และชุดข้อมูลที่ปรับให้สมดุลแล้ว ของวิธีการเรียนรู้ด้วยเครื่อง 6 วิธี

```
par(mfrow=c(3,2))
################## DT ##############################################
Original = c(86.6153,87.3814,86.9475,86.9390,86.7271)
Balance = c(96.2371,96.5683,96.6641,96.5773,96.5072)
no = c(200,400,600,800,1000)

plot(no, Original, type = 'b',main =
       "DT", lty=3,lwd=2.9,col= "brown", ylim=c(85,100),
       xlab = 'Number of Variable', ylab = 'Percentage of Accuracy',pch = 15,cex=1)
lines(no,Balance, type = "b",lty=2,lwd=2.9,col = "blue",pch = 17,cex=1)
```

```
labels = c("Original Data","Balanced Data")
colors = c("brown","blue")
pchh = c(8,16)
legend("bottomright",inset = .02,labels,lwd=2,lty=1,col=colors,pch=pchh, cex=0.7)

############################ Naïve ###################################
Original = c(95.8831,96.2881,96.4593,96.3848,96.5051)
Balance = c(98.3479,98.5599,98.6701,98.6425,98.6778)
no = c(200,400,600,800,1000)

plot(no, Original, type = 'b',main =
       "Naive", lty=3,lwd=2.9,col= "brown", ylim=c(85,100),
     xlab = 'Number of Variable', ylab = 'Percentage of Accuracy',pch = 15,cex=1)
lines(no,Balance, type = "b",lty=2,lwd=2.9,col = "blue",pch = 17,cex=1)
labels = c("Original Data","Balanced Data")
colors = c("brown","blue")
pchh = c(8,16)
legend("bottomright",inset = .02,labels,lwd=2,lty=1,col=colors,pch=pchh, cex=0.7)

######################## KNN ######################################
Original = c(94.366,94.6220,94.5915,93.5000,93.9695)
Balance = c(94.1395,95.2114,95.6665,96.2635,96.7755)
no = c(200,400,600,800,1000)

plot(no, Original, type = 'b',main =
       "KNN", lty=3,lwd=2.9,col= "brown", ylim=c(85,100),
     xlab = 'Number of Variable', ylab = 'Percentage of Accuracy',pch = 15,cex=1)
lines(no,Balance, type = "b",lty=2,lwd=2.9,col = "blue",pch = 17,cex=1)
labels = c("Original Data","Balanced Data")
colors = c("brown","blue")
pchh = c(8,16)
legend("bottomright",inset = .02,labels,lwd=2,lty=1,col=colors,pch=pchh, cex=0.7)
```

```
######################## SVM ######################################
Original = c(96.6407,92.5017,92.4949,92.6390,92.5661)
Balance = c(99.0784,98.5246,98.5976,98.6737,98.6737)
no = c(200,400,600,800,1000)

plot(no, Original, type = 'b',main =
       "SVM", lty=3,lwd=2.9,col= "brown", ylim=c(85,100),
     xlab = 'Number of Variable', ylab = 'Percentage of Accuracy',pch = 15,cex=1)
lines(no,Balance, type = "b",lty=2,lwd=2.9,col = "blue",pch = 17,cex=1)
labels = c("Original Data","Balanced Data")
colors = c("brown","blue")
pchh = c(8,16)
legend("bottomright",inset = .02,labels,lwd=2,lty=1,col=colors,pch=pchh, cex=0.7)

######################## ANN ######################################
Original = c(88.9763,89.0542,89.0729,89.3017,89.0780)
Balance = c(98.1108,98.1114,98.2976,98.2222,98.2419)
no = c(200,400,600,800,1000)

plot(no, Original, type = 'b',main =
       "ANN", lty=3,lwd=2.9,col= "brown", ylim=c(85,100),
     xlab = 'Number of Variable', ylab = 'Percentage of Accuracy',pch = 15,cex=1)
lines(no,Balance, type = "b",lty=2,lwd=2.9,col = "blue",pch = 17,cex=1)
labels = c("Original Data","Balanced Data")
colors = c("brown","blue")
pchh = c(8,16)
legend("bottomright",inset = .02,labels,lwd=2,lty=1,col=colors,pch=pchh, cex=0.7)

############################## RF ######################################
Original = c(94.4881,94.5170,94.6305,94.7576,94.7034)
Balance = c(99.5982,99.6042,99.6443,99.6084,99.6156)
no = c(200,400,600,800,1000)
```

```
plot(no, Original, type = 'b',main =
      "RF", lty=3,lwd=2.9,col= "brown", ylim=c(85,100),
    xlab = 'Number of Variable', ylab = 'Percentage of Accuracy',pch = 15,cex=1)
lines(no,Balance, type = "b",lty=2,lwd=2.9,col = "blue",pch = 17,cex=1)
labels = c("Original Data","Balanced Data")
colors = c("brown","blue")
pchh = c(8,16)
legend("bottomright",inset = .02,labels,lwd=2,lty=1,col=colors,pch=pchh, cex=0.7)
```

คำสั่งที่ใช้สร้างกราฟเปรียบเทียบค่าเฉลี่ยร้อยละความแม่นยำ ของวิธีการเรียนรู้ด้วยเครื่อง 6 วิธี ของชุดข้อมูลดั้งเดิม และชุดข้อมูลที่ปรับให้สมดุลแล้ว โดยแบ่งตามระดับของมะเร็งปอดทั้ง 4 กลุ่ม

```
par(mfrow=c(4,2))
# 0 ######################################## Pre Original ##############
DT = c(91.20618,91.40487,91.55918,91.5306,91.6221)
Naive = c(97.85656,98.06565,98.25463,98.4329,98.43828)
KNN = c(93.2766,93.53034,93.58662,92.8798,93.42632)
SVM = c(97.57632,90.7383,90.75804,90.92084,93.42632)
ANN = c(89.48197,89.71302,89.55555,89.70945,90.89482)
RF = c(93.44385,93.45016,93.5389,93.75616,93.73066)
no = c(200,400,600,800,1000)

plot(no, DT , type = 'b',main = "Precision(0) of Original Data", lty=3,lwd=2.9,col=
"purple", ylim=c(80,100),
    xlab = 'Number of Variable', ylab = 'Percentage of Precision Data',pch = 15,cex=1)

lines(no,Naive, type = "b",lty=2,lwd=2.9,col = "blue",pch = 10,cex=1)
lines(no,KNN, type = "b",lty=2,lwd=2.9,col = "green",pch = 11,cex=1)
lines(no,SVM, type = "b",lty=2,lwd=2.9,col = "orange",pch = 8,cex=1)
lines(no,ANN, type = "b",lty=2,lwd=2.9,col = "red",pch = 5,cex=1)
lines(no,RF, type = "b",lty=2,lwd=2.9,col = "deeppink",pch = 17,cex=1)
```

```
labels = c("DT","Naive","KNN","SVM","ANN","RF")
colors = c("purple","blue","green","orange","red","deeppink")
pchh = c(15,10,11,8,5,17)
legend("bottomright",inset = .02,labels,lwd=2,lty=1,col=colors,pch=pchh, cex=0.7)

# 0 ######################################## Pre Balance ##############
DT = c(96.60503,96.78168,96.1580,95.81077,95.17955)
Naive = c(97.41438,98.02453,98.23191,98.27029,98.45798)
KNN = c(96.27803,97.01117,97.18056,97.71644,98.38488)
SVM = c(99.83407,95.18664,95.36644,95.67638,95.42915)
ANN = c(98.98308,98.82467,99.2486,99.03328,99.23813)
RF = c(99.8565,99.87201,99.90356,99.83691,99.86852)

no = c(200,400,600,800,1000)

plot(no, DT , type = 'b',main =
       "Precision(0) of Balanced Data", lty=3,lwd=2.9,col= "purple", ylim=c(80,100),
     xlab = 'Number of Variable', ylab = 'Percentage of Precision',pch = 15,cex=1)
lines(no,Naive, type = "b",lty=2,lwd=2.9,col = "blue",pch = 10,cex=1)
lines(no,KNN, type = "b",lty=2,lwd=2.9,col = "green",pch = 11,cex=1)
lines(no,SVM, type = "b",lty=2,lwd=2.9,col = "orange",pch = 8,cex=1)
lines(no,ANN, type = "b",lty=2,lwd=2.9,col = "red",pch = 5,cex=1)
lines(no,RF, type = "b",lty=2,lwd=2.9,col = "deeppink",pch = 17,cex=1)

labels = c("DT","Naive","KNN","SVM","ANN","RF")
colors = c("purple","blue","green","orange","red","deeppink")
pchh = c(15,10,11,8,5,17)
legend("bottomright",inset = .02,labels,lwd=2,lty=1,col=colors,pch=pchh, cex=0.7)

# 1 ######################################### Pre Original #############
DT = c(75.5899,76.5009,72.1625,69.3374,68.0688)
Naive = c(92.0261,93.3646,93.8911,93.9647,93.6858)
KNN = c(93.4427,94.1846,93.0025,91.2165,90.4587)
```

```
SVM = c(92.8249,96.3860,96.0414,96.3044,96.0036)
ANN = c(84.1197,86.0035,85.5543,86.2750,83.8322)
RF = c(93.4826,94.0788,94.0527,94.0314,93.4561)
no = c(200,400,600,800,1000)

plot(no, DT , type = 'b',main = "Precision(1) of Original Data", lty=3,lwd=2.9,col= "purple", ylim=c(80,100),
    xlab = 'Number of variable', ylab = 'Percentage of Precision',pch = 15,cex=1)

lines(no,Naive, type = "b",lty=2,lwd=2.9,col = "blue",pch = 10,cex=1)
lines(no,KNN, type = "b",lty=2,lwd=2.9,col = "green",pch = 11,cex=1)
lines(no,SVM, type = "b",lty=2,lwd=2.9,col = "orange",pch = 8,cex=1)
lines(no,ANN, type = "b",lty=2,lwd=2.9,col = "red",pch = 5,cex=1)
lines(no,RF, type = "b",lty=2,lwd=2.9,col = "deeppink",pch = 17,cex=1)

labels = c("DT","Naive","KNN","SVM","ANN","RF")
colors = c("purple","blue","green","orange","red","deeppink")
pchh = c(15,10,11,8,5,17)
legend("bottomright",inset = .02,labels,lwd=2,lty=1,col=colors,pch=pchh, cex=0.7)

# 1 ##################################### Pre Balance ##############
DT = c(96.9923,97.7007,97.9952,98.0637,98.2507)
Naive = c(98.9373,99.1071,99.2066,99.2514,99.2336)
KNN = c(92.8773,94.0688,94.8064,95.5235,95.7901)
SVM = c(98.6620,99.3357,99.3325,99.3133,99.2570)
ANN = c(97.8304,97.7237,97.9417,97.7789,97.8616)
RF = c(99.2132,99.2454,99.3035,99.3027,99.2586)

no = c(200,400,600,800,1000)

plot(no, DT , type = 'b',main =
    "Precision(1) of Balanced Data", lty=3,lwd=2.9,col= "purple", ylim=c(80,100),
    xlab = 'Number of variable', ylab = 'Percentage of Precision',pch = 15,cex=1)
```

```
lines(no,Naive, type = "b",lty=2,lwd=2.9,col = "blue",pch = 10,cex=1)
lines(no,KNN, type = "b",lty=2,lwd=2.9,col = "green",pch = 11,cex=1)
lines(no,SVM, type = "b",lty=2,lwd=2.9,col = "orange",pch = 8,cex=1)
lines(no,ANN, type = "b",lty=2,lwd=2.9,col = "red",pch = 5,cex=1)
lines(no,RF, type = "b",lty=2,lwd=2.9,col = "deeppink",pch = 17,cex=1)


labels = c("DT","Naive","KNN","SVM","ANN","RF")
colors = c("purple","blue","green","orange","red","deeppink")
pchh = c(15,10,11,8,5,17)
legend("bottomright",inset = .02,labels,lwd=2,lty=1,col=colors,pch=pchh, cex=0.7)

# 2 ######################################## Pre Original ############
DT = c(70.0483,74.2722,76.3169,78.5165,80.1495)
Naive = c(84.6909,85.0829,85.3978,83.8886,84.7228)
KNN = c(99.0165,99.2671,99.0565,94.2784,95.2032)
SVM = c(91.9675,100.0000,100.0000,100.0000,100.0000)
ANN = c(78.6193,77.1698,77.3555,79.4051,75.4428)
RF = c(98.9507,98.8368,99.2519,98.8649,98.9994)
no = c(200,400,600,800,1000)

plot(no, DT , type = 'b',main = "Precision(2) of Original Data", lty=3,lwd=2.9,col=
"purple", ylim=c(70,100),
    xlab = 'Number of variable', ylab = 'Percentage of Precision',pch = 15,cex=1)

lines(no,Naive, type = "b",lty=2,lwd=2.9,col = "blue",pch = 10,cex=1)
lines(no,KNN, type = "b",lty=2,lwd=2.9,col = "green",pch = 11,cex=1)
lines(no,SVM, type = "b",lty=2,lwd=2.9,col = "orange",pch = 8,cex=1)
lines(no,ANN, type = "b",lty=2,lwd=2.9,col = "red",pch = 5,cex=1)
lines(no,RF, type = "b",lty=2,lwd=2.9,col = "deeppink",pch = 17,cex=1)

labels = c("DT","Naive","KNN","SVM","ANN","RF")
colors = c("purple","blue","green","orange","red","deeppink")
```

```
pchh = c(15,10,11,8,5,17)
legend("bottomright",inset = .02,labels,lwd=2,lty=1,col=colors,pch=pchh, cex=0.7)

# 2 ######################################## Pre Balance ##############
DT = c(92.7466,93.2014,93.5255,93.5660,93.5877)
Naive = c(97.2733,97.2496,97.3511,97.1953,97.1366)
KNN = c(89.3922,91.0341,91.6024,92.4433,93.4571)
SVM = c(97.8900,99.9049,99.9717,99.9814,99.9943)
ANN = c(96.2493,96.5578,96.3918,96.5603,96.3255)
RF = c(99.3669,99.3281,99.3847,99.2952,99.3343)

no = c(200,400,600,800,1000)

plot(no, DT , type = 'b',main =
       "Precision(2) of Balanced Data", lty=3,lwd=2.9,col= "purple", ylim=c(80,100),
     xlab = 'Number of variable', ylab = 'Percentage of Precision',pch = 15,cex=1)
lines(no,Naive, type = "b",lty=2,lwd=2.9,col = "blue",pch = 10,cex=1)
lines(no,KNN, type = "b",lty=2,lwd=2.9,col = "green",pch = 11,cex=1)
lines(no,SVM, type = "b",lty=2,lwd=2.9,col = "orange",pch = 8,cex=1)
lines(no,ANN, type = "b",lty=2,lwd=2.9,col = "red",pch = 5,cex=1)
lines(no,RF, type = "b",lty=2,lwd=2.9,col = "deeppink",pch = 17,cex=1)

labels = c("DT","Naive","KNN","SVM","ANN","RF")
colors = c("purple","blue","green","orange","red","deeppink")
pchh = c(15,10,11,8,5,17)
legend("bottomright",inset = .02,labels,lwd=2,lty=1,col=colors,pch=pchh, cex=0.7)

# 3 ######################################### Pre Original #############
DT = c(90.4345,89.3565,88.4767,89.4363,87.4557)
Naive = c(99.4500,99.8813,100.0000,100.0000,100.0000)
KNN = c(100.0000,100.0000,100.0000,100.0000,100.0000)
SVM = c(99.1505,100.0000,100.0000,100.0000,100.0000)
ANN = c(98.2938,98.2810,98.2416,98.2970,98.3717)
```

```
RF = c(100.0000,100.0000,100.0000,100.0000,100.0000)

no = c(200,400,600,800,1000)

plot(no, DT , type = 'b',main = "Precision(3) of Original Data", lty=3,lwd=2.9,col= "purple", ylim=c(80,100),
    xlab = 'Number of variable', ylab = 'Percentage of Precision',pch = 15,cex=1)

lines(no,Naive, type = "b",lty=2,lwd=2.9,col = "blue",pch = 10,cex=1)
lines(no,KNN, type = "b",lty=2,lwd=2.9,col = "green",pch = 11,cex=1)
lines(no,SVM, type = "b",lty=2,lwd=2.9,col = "orange",pch = 8,cex=1)
lines(no,ANN, type = "b",lty=2,lwd=2.9,col = "red",pch = 5,cex=1)
lines(no,RF, type = "b",lty=2,lwd=2.9,col = "deeppink",pch = 17,cex=1)

labels = c("DT","Naive","KNN","SVM","ANN","RF")
colors = c("purple","blue","green","orange","red","deeppink")
pchh = c(15,10,11,8,5,17)
legend("bottomright",inset = .02,labels,lwd=2,lty=1,col=colors,pch=pchh, cex=0.7)

# 3 ###################################### Pre Balance ###############
DT = c(99.1234,99.0410,99.3447,99.2437,99.3346)
Naive = c(99.9329,99.9875,99.9952,100.0000,100.0000)
KNN = c(99.3594,99.7058,99.8919,99.9832,100.0000)
SVM = c(100.0000,100.0000,100.0000,100.0000,100.0000)
ANN = c(99.5972,99.5521,99.7762,99.5884,99.7465)
RF = c(99.9723,99.9921,99.9948,100.0000,100.0000)

no = c(200,400,600,800,1000)

plot(no, DT , type = 'b',main =
      "Precision(3) of Balanced Data", lty=3,lwd=2.9,col= "purple", ylim=c(95,100),
    xlab = 'Number of variable', ylab = 'Percentage of Precision',pch = 15,cex=1)
lines(no,Naive, type = "b",lty=2,lwd=2.9,col = "blue",pch = 10,cex=1)
```

```
lines(no,KNN, type = "b",lty=2,lwd=2.9,col = "green",pch = 11,cex=1)
lines(no,SVM, type = "b",lty=2,lwd=2.9,col = "orange",pch = 8,cex=1)
lines(no,ANN, type = "b",lty=2,lwd=2.9,col = "red",pch = 5,cex=1)
lines(no,RF, type = "b",lty=2,lwd=2.9,col = "deeppink",pch = 17,cex=1)

labels = c("DT","Naive","KNN","SVM","ANN","RF")
colors = c("purple","blue","green","orange","red","deeppink")
pchh = c(15,10,11,8,5,17)
legend("bottomright",inset = .02,labels,lwd=2,lty=1,col=colors,pch=pchh, cex=0.7)
```

คำสั่งที่ใช้สร้างกราฟเปรียบเทียบค่าเฉลี่ยร้อยละความระลึก ของวิธีการเรียนรู้ด้วยเครื่อง 6 วิธี ของชุดข้อมูลดั้งเดิม และชุดข้อมูลที่ปรับให้สมดุลแล้ว โดยแบ่งตามระดับของมะเร็งปอดทั้ง 4 กลุ่ม

```
par(mfrow=c(4,2))
# 0 ######################################## Recall Original ##########
DT = c(91.6487,92.6817,92.2335,92.3492,92.2150)
Naive = c(96.30588,96.68884,96.73544,96.4463,96.6191)
KNN = c(99.16589,99.27145,99.15039,98.3716,98.3826)
SVM = c(97.67454,99.59047,99.54806,99.5634,99.5384)
ANN = c(96.02419,95.99997,96.1041,96.2948,95.6880)
RF = c(99.16801,99.22375,99.2769,99.2171,99.1965)
no = c(200,400,600,800,1000)

plot(no, DT , type = 'b',main = "Recall(0) of Original Data", lty=3,lwd=2.9,col= "purple",
ylim=c(80,100),
    xlab = 'Number of variable', ylab = 'Percentage of Recall',pch = 15,cex=1)

lines(no,Naive, type = "b",lty=2,lwd=2.9,col = "blue",pch = 10,cex=1)
lines(no,KNN, type = "b",lty=2,lwd=2.9,col = "green",pch = 11,cex=1)
lines(no,SVM, type = "b",lty=2,lwd=2.9,col = "orange",pch = 8,cex=1)
lines(no,ANN, type = "b",lty=2,lwd=2.9,col = "red",pch = 5,cex=1)
lines(no,RF, type = "b",lty=2,lwd=2.9,col = "deeppink",pch = 17,cex=1)
```

```
labels = c("DT","Naive","KNN","SVM","ANN","RF")
colors = c("purple","blue","green","orange","red","deeppink")
pchh = c(15,10,11,8,5,17)
legend("bottomright",inset = .02,labels,lwd=2,lty=1,col=colors,pch=pchh, cex=0.7)

# 0 ###################################### Recall Balance ###########
DT = c(88.4301,89.5415,90.5658,90.5630,90.9238)
Naive = c(96.1030,96.2477,96.5060,96.3839,96.3007)
KNN = c(80.4448,83.9950,85.8378,87.5334,88.8134)
SVM = c(96.5070,99.2450,99.3157,99.3086,99.2670)
ANN = c(94.0938,94.2528,93.9964,94.2000,93.7723)
RF = c(98.5529,98.5517,98.6841,98.6190,98.6014)

no = c(200,400,600,800,1000)

plot(no, DT , type = 'b',main =
       "Recall(0) of Balanced Data", lty=3,lwd=2.9,col= "purple", ylim=c(80,100),
     xlab = 'Number of variable', ylab = 'Percentage of Recall',pch = 15,cex=1)
lines(no,Naive, type = "b",lty=2,lwd=2.9,col = "blue",pch = 10,cex=1)
lines(no,KNN, type = "b",lty=2,lwd=2.9,col = "green",pch = 11,cex=1)
lines(no,SVM, type = "b",lty=2,lwd=2.9,col = "orange",pch = 8,cex=1)
lines(no,ANN, type = "b",lty=2,lwd=2.9,col = "red",pch = 5,cex=1)
lines(no,RF, type = "b",lty=2,lwd=2.9,col = "deeppink",pch = 17,cex=1)

labels = c("DT","Naive","KNN","SVM","ANN","RF")
colors = c("purple","blue","green","orange","red","deeppink")
pchh = c(15,10,11,8,5,17)
legend("bottomright",inset = .02,labels,lwd=2,lty=1,col=colors,pch=pchh, cex=0.7)

# 1 ###################################### Recall Origina l###########
DT = c(76.0164,75.6929,73.0683,71.8442,70.6766)
Naive = c(92.2014,92.1884,93.0176,93.6422,93.6039)
```

```
KNN = c(84.5057,83.3904,83.5062,89.7974,92.4683)
SVM = c(95.2077,81.7506,81.2811,82.4608,82.3645)
ANN = c(70.5646,70.9262,69.8932,70.8001,71.5783)
RF = c(83.3226,84.1995,83.4714,84.2162,83.8748)

no = c(200,400,600,800,1000)

plot(no, DT , type = 'b',main = "Recall(1) of Original Data", lty=3,lwd=2.9,col= "purple",
ylim=c(70,100),
    xlab = 'Number of variable', ylab = 'Percentage of Recall',pch = 15,cex=1)

lines(no,Naive, type = "b",lty=2,lwd=2.9,col = "blue",pch = 10,cex=1)
lines(no,KNN, type = "b",lty=2,lwd=2.9,col = "green",pch = 11,cex=1)
lines(no,SVM, type = "b",lty=2,lwd=2.9,col = "orange",pch = 8,cex=1)
lines(no,ANN, type = "b",lty=2,lwd=2.9,col = "red",pch = 5,cex=1)
lines(no,RF, type = "b",lty=2,lwd=2.9,col = "deeppink",pch = 17,cex=1)

labels = c("DT","Naive","KNN","SVM","ANN","RF")
colors = c("purple","blue","green","orange","red","deeppink")
pchh = c(15,10,11,8,5,17)
legend("bottomright",inset = .02,labels,lwd=2,lty=1,col=colors,pch=pchh, cex=0.7)

# 1 ######################################## Recall Balance ############
DT = c(99.6839,99.9219,99.9392,99.9390,99.9343)
Naive = c(98.3188,98.7144,98.8635,98.9198,98.9319)
KNN = c(99.4835,99.6745,99.8657,99.9867,100.0000)
SVM = c(99.9296,98.8291,98.9390,98.9618,98.9970)
ANN = c(99.5811,99.4155,99.7654,99.4561,99.7985)
RF = c(99.9834,99.9802,99.9784,99.9806,99.9682)

no = c(200,400,600,800,1000)
```



```
plot(no, DT , type = 'b',main =
```

```
      "Recall(1) of Balanced Data", lty=3,lwd=2.9,col= "purple", ylim=c(95,100),
    xlab = 'Number of variable', ylab = 'Percentage of Recall',pch = 15,cex=1)
lines(no,Naive, type = "b",lty=2,lwd=2.9,col = "blue",pch = 10,cex=1)
lines(no,KNN, type = "b",lty=2,lwd=2.9,col = "green",pch = 11,cex=1)
lines(no,SVM, type = "b",lty=2,lwd=2.9,col = "orange",pch = 8,cex=1)
lines(no,ANN, type = "b",lty=2,lwd=2.9,col = "red",pch = 5,cex=1)
lines(no,RF, type = "b",lty=2,lwd=2.9,col = "deeppink",pch = 17,cex=1)

labels = c("DT","Naive","KNN","SVM","ANN","RF")
colors = c("purple","blue","green","orange","red","deeppink")
pchh = c(15,10,11,8,5,17)
legend("bottomright",inset = .02,labels,lwd=2,lty=1,col=colors,pch=pchh, cex=0.7)

# 2 ######################################### Recall Original ##########
DT = c(63.6498,64.3628,66.5466,69.0353,70.2188)
Naive = c(93.1983,94.3217,94.6309,95.3847,95.1671)
KNN = c(67.9641,69.0520,69.0942,63.8188,66.4732)
SVM = c(87.5462,52.1042,52.8796,53.0699,52.2524)
ANN = c(57.3580,56.9940,56.8159,59.0210,57.0807)
RF = c(69.3907,68.1343,69.5379,70.0623,69.8520)
no = c(200,400,600,800,1000)

plot(no, DT , type = 'b',main = "Recall(2) of Original Data", lty=3,lwd=2.9,col= "purple",
ylim=c(50,100),
    xlab = 'Number of variable', ylab = 'Percentage of Recall',pch = 15,cex=1)

lines(no,Naive, type = "b",lty=2,lwd=2.9,col = "blue",pch = 10,cex=1)
lines(no,KNN, type = "b",lty=2,lwd=2.9,col = "green",pch = 11,cex=1)
lines(no,SVM, type = "b",lty=2,lwd=2.9,col = "orange",pch = 8,cex=1)
lines(no,ANN, type = "b",lty=2,lwd=2.9,col = "red",pch = 5,cex=1)
lines(no,RF, type = "b",lty=2,lwd=2.9,col = "deeppink",pch = 17,cex=1)
```



```
labels = c("DT","Naive","KNN","SVM","ANN","RF")
```

```
colors = c("purple","blue","green","orange","red","deeppink")
pchh = c(15,10,11,8,5,17)
legend("bottomright",inset = .02,labels,lwd=2,lty=1,col=colors,pch=pchh, cex=0.7)

# 2 ######################################## Recall Balance ############
DT = c(97.0247,96.9675,96.2478,95.9030,95.2682)
Naive = c(98.9778,99.2918,99.3377,99.2718,99.4852)
KNN = c(97.5891,98.2123,98.1666,98.4550,98.9506)
SVM = c(99.9069,96.2025,96.4392,96.6704,96.3331)
ANN = c(99.2048,99.1777,99.4887,99.4394,99.4675)
RF = c(99.8741,99.8906,99.9179,99.8591,99.8918)

no = c(200,400,600,800,1000)

plot(no, DT , type = 'b',main =
       "Recall(2) of Balanced Data", lty=3,lwd=2.9,col= "purple", ylim=c(90,100),
     xlab = 'Number of variable', ylab = 'Percentage of Recall',pch = 15,cex=1)
lines(no,Naive, type = "b",lty=2,lwd=2.9,col = "blue",pch = 10,cex=1)
lines(no,KNN, type = "b",lty=2,lwd=2.9,col = "green",pch = 11,cex=1)
lines(no,SVM, type = "b",lty=2,lwd=2.9,col = "orange",pch = 8,cex=1)
lines(no,ANN, type = "b",lty=2,lwd=2.9,col = "red",pch = 5,cex=1)
lines(no,RF, type = "b",lty=2,lwd=2.9,col = "deeppink",pch = 17,cex=1)

labels = c("DT","Naive","KNN","SVM","ANN","RF")
colors = c("purple","blue","green","orange","red","deeppink")
pchh = c(15,10,11,8,5,17)
legend("bottomright",inset = .02,labels,lwd=2,lty=1,col=colors,pch=pchh, cex=0.7)

# 3 ######################################### Recall Original ##########
DT = c(88.9376,87.3127,86.4086,84.9567,83.1762)
Naive = c(99.6520,99.6827,99.8203,99.8574,99.8680)
KNN = c(99.9507,99.9742,99.9470,94.3431,95.3923)
SVM = c(99.9917,94.5498,94.6417,94.7788,94.5218)
```

```
ANN = c(93.6369,93.8112,93.0629,93.7603,93.0183)
RF = c(99.8766,99.9472,99.9607,99.9392,99.9576)
no = c(200,400,600,800,1000)

plot(no, DT , type = 'b',main = "Recall(3) of Original Data", lty=3,lwd=2.9,col= "purple",
ylim=c(80,100),
     xlab = 'Number of variable', ylab = 'Percentage of Recall',pch = 15,cex=1)

lines(no,Naive, type = "b",lty=2,lwd=2.9,col = "blue",pch = 10,cex=1)
lines(no,KNN, type = "b",lty=2,lwd=2.9,col = "green",pch = 11,cex=1)
lines(no,SVM, type = "b",lty=2,lwd=2.9,col = "orange",pch = 8,cex=1)
lines(no,ANN, type = "b",lty=2,lwd=2.9,col = "red",pch = 5,cex=1)
lines(no,RF, type = "b",lty=2,lwd=2.9,col = "deeppink",pch = 17,cex=1)

labels = c("DT","Naive","KNN","SVM","ANN","RF")
colors = c("purple","blue","green","orange","red","deeppink")
pchh = c(15,10,11,8,5,17)
legend("bottomright",inset = .02,labels,lwd=2,lty=1,col=colors,pch=pchh, cex=0.7)

# 3 ###################################### Recall Balance ############
DT = c(99.9134,99.9410,99.9587,99.9828,99.9769)
Naive = c(100.0000,100.0000,100.0000,100.0000,100.0000)
KNN = c(99.3214,98.9978,98.9254,99.2384,99.4353)
SVM = c(100.0000,99.8516,99.7176,99.7873,99.8112)
ANN = c(99.7150,99.6423,99.9915,99.8043,99.9651)
RF = c(100.0000,100.0000,100.0000,100.0000,100.0000)

no = c(200,400,600,800,1000)

plot(no, DT , type = 'b',main =
       "Recall(3) of Balanced Data", lty=3,lwd=2.9,col= "purple", ylim=c(95,100),
     xlab = 'Number of variable', ylab = 'Percentage of Recall',pch = 15,cex=1)
lines(no,Naive, type = "b",lty=2,lwd=2.9,col = "blue",pch = 10,cex=1)
```

```
lines(no,KNN, type = "b",lty=2,lwd=2.9,col = "green",pch = 11,cex=1)
lines(no,SVM, type = "b",lty=2,lwd=2.9,col = "orange",pch = 8,cex=1)
lines(no,ANN, type = "b",lty=2,lwd=2.9,col = "red",pch = 5,cex=1)
lines(no,RF, type = "b",lty=2,lwd=2.9,col = "deeppink",pch = 17,cex=1)

labels = c("DT","Naive","KNN","SVM","ANN","RF")
colors = c("purple","blue","green","orange","red","deeppink")
pchh = c(15,10,11,8,5,17)
legend("bottomright",inset = .02,labels,lwd=2,lty=1,col=colors,pch=pchh, cex=0.7)
#####################################################################
```

งานทะเบียนคณะวิทยาศาสตร์
สถาบันเทคโนโลยีพระจอมเกล้าเจ้าคุณทหารลาดกระบัง
คำรับรองเล่มปัญหาพิเศษ

วันที่ 19 เดือน พฤษภาคม พ.ศ. 2566

ข้าพเจ้า นายรักษ์ธรรม ลีพึ่งธรรม รหัสประจำตัว 62050816
นางสาววราพร โพธิโชติ รหัสประจำตัว 62050827
นางสาววิชญา เคารพ รหัสประจำตัว 62050832
นักศึกษาหลักสูตรวิทยาศาสตรบัณฑิต สาขาวิชา สถิติประยุกต์ ภาควิชา สถิติ
ขอรับรองว่าปัญหาพิเศษ เรื่อง

การเปรียบเทียบวิธีการเรียนรู้ด้วยเครื่องสำหรับการจำแนกระดับ
การเป็นมะเร็งปอดจากรหัสพันธุกรรม
A COMPARISON OF MACHINE LEARNING METHODS FOR
CLASSIFICATION THE LEVEL OF LUNG CANCER FROM DNA

ปีการศึกษา 2565
เป็นผลงานวิจัยที่มิได้คัดลอกหรือละเมิดลิขสิทธิ์ของผู้อื่นและได้ผ่านการตรวจสอบความซ้ำซ้อนเรียบร้อยแล้วและได้แนบเอกสารการตรวจสอบการลอกเลียนงานวรรณกรรมที่ตรวจสอบจากเล่มปัญหาพิเศษฉบับสมบูรณ์แล้ว
โปรแกรมอักขราวิสุทธิ์ 4.29%

ลงชื่อ.....รักษ์ธรรม ลีพึ่งธรรม..... ลงชื่อ.....วราพร โพธิโชติ..... ลงชื่อ.....วิชญา เคารพ.....

(นายรักษ์ธรรม ลีพึ่งธรรม) (นางสาววราพร โพธิโชติ) (นางสาววิชญา เคารพ)
นักศึกษา นักศึกษา นักศึกษา

ข้าพเจ้า ร.ศ.ดร.อัชฌา อระวีพร อาจารย์ที่ปรึกษาปัญหาพิเศษ ได้ตรวจสอบปัญหาพิเศษ ของนักศึกษาข้างต้นแล้ว ขอรับรองว่าเป็นผลงานวิจัยของนักศึกษาจริงและมีเนื้อหาสมบูรณ์ จึงลงชื่อไว้เป็นหลักฐาน

ลงชื่อ.....อัชฌา อระวีพร.....

(ร.ศ.ดร.อัชฌา อระวีพร)
อาจารย์ที่ปรึกษา